Available online at www.sciencedirect.com

ScienceDirect

Fuzzy Sets and Systems 318 (2017) 56–79

FUZZY sets and systems

www.elsevier.com/locate/fss

Fuzzy clustering of time series using extremes

Pierpaolo D'Urso [a], Elizabeth A. Maharaj [b,*], Andrés M. Alonso [c]

[a] *Department of Social Sciences and Economics, Sapienza – Universitá di Roma, Italy*
[b] *Department of Econometrics and Business Statistics, Monash University, Australia*
[c] *Department of Statistics and IFL, Universidad Carlos III de Madrid, Spain*

Received 21 February 2016; received in revised form 19 September 2016; accepted 15 October 2016

Available online 20 October 2016

Abstract

In this study we explore the grouping together of time series with similar seasonal patterns using extreme value analysis with fuzzy clustering. Input features into the fuzzy clustering methods are parameter estimates of time varying location, scale and shape obtained from fitting the generalised extreme value (GEV) distribution to annual maxima or the r-largest order statistics per year of the time series. An innovative contribution of the study is the development of new generalised fuzzy clustering procedures taking into account weights, and the derivation of iterative solutions based on the GEV parameter estimators. Simulation studies conducted to evaluate the methods, reveal good performance. An application is made to a set of daily sea-level time series from around the coast of Australia where the identified clusters are well validated and they can be meaningfully interpreted.

Keywords: Fuzzy c-means clustering; Fuzzy c-medoids clustering; Time series data; r-largest order statistics; Generalised extreme value distribution

1. Introduction

Extreme value analysis of seasonal time series such as that of temperatures and sea levels is of much relevance to research in areas such as climatology, oceanography, environmental science and engineering. In particular, the analysis of extreme sea levels could be useful for planning long-term coastal protection and development. Even if long-term sea-level records are not available, long-term projections of extremes resulting from the extreme value analysis of shorter-term data may be acceptable from an engineering perspective in terms of designing, say for example, sea-protection walls in particular coastal areas.

Authors such as Tsimpis and Blackman [26], Unikrishnan et al. [28], Méndez et al. [20] and Scotto et al. [24] have used extreme value analysis to study sea level extremes, while Scotto et al. [25] and Alonso et al. [1] are amongst others who have applied extreme analysis to temperature extremes. In particular, Scotto et al. [24] combined a Bayesian

* Corresponding author at: Department of Econometrics and Business Statistics, Monash University, Caulfield Campus, 900 Dandenong Road, Caulfield East, Victoria 3145, Australia. Fax: +61 3 9903 2007.
E-mail address: ann.maharaj@monash.edu (E.A. Maharaj).

http://dx.doi.org/10.1016/j.fss.2016.10.006
0165-0114/

analysis of extreme sea levels to estimate predictive distributions, with hierarchical cluster analysis to distinguish groups of North Atlantic sea locations. Scotto et al. [25] applied the same methodology to European daily temperature series to group together similar locations, while Alonso et al. [1] compared Generalised Pareto models fitted to extreme temperature observations. These above-mentioned studies focus on using clustering methods to group together locations based on predictive distributions, while in a recent study, Maharaj et al. [18] considered non-hierarchical clustering methods and classification methods to group together seasonal time series across the available record, and in an application to regional temperature time series revealed realistic groupings. Given that the dynamics of a time series may change over time, a time series might display patterns that may enable it to belong to one cluster over one period while over another period its pattern may be more consistent with those in another cluster. The traditional clustering (crisp clustering) procedures are unable to identify the changing patterns in a time. However clustering based on fuzzy logic will be able to detect the switching patterns from one time period to another, thus enabling some time series to simultaneously belong to more than one cluster.

In particular, with respect to the dynamic peculiarity of the time series, we have two motivations justifying the fuzzy approach for the clustering time series (D'Urso, [9,10]; D'Urso, Maharaj [11]):

- Greater sensitivity in capturing the features characterising the time series. In many cases, since the dynamics of the time series are drifting or switching, the hard (crisp) clustering approaches are likely to miss this underlying structure. The switches, which are usually vague, can be naturally treated by means of fuzzy clustering. Notice that the switching behaviour of the time series can be related not only to the observed time series but also to their suitable parametric or non-parametric representations, such as autocorrelation functions, periodograms, cepstrals, wavelets, density functions, forecast densities and so on.
- Greater adaptivity in defining the prototype time series. This can be better appreciated when the observed time patterns or their suitable parametric and non-parametric representations do not differ too much from each other. In this case, the fuzzy definition of the clusters allows us to single out underlying structures, if these are likely to exist in the given set of time series.

Here, we extend the study of Maharaj et al. [18] to group the series across the available record using fuzzy clustering methods. New generalised procedures for fuzzy clustering taking into account weights are developed, and iterative solutions based on the GEV parameter estimators are derived.

In a simulation study, we consider seasonal times series analogous to daily temperatures or sea levels and to each series, we fit a generalised extreme value (GEV) distribution to maxima from one and two blocks per year. We estimate the parameters of location, scale and shape, and use these parameters as features for fuzzy clustering of the series. The methods considered are fuzzy c-means, weighted fuzzy c-means, fuzzy c-medoids and weighted fuzzy c-medoids. We then apply these fuzzy clustering methods to daily sea level time series where the r-largest order statistics per year are used to estimate the GEV parameters.

In Section 2, we provide a brief description of the generalised extreme value distribution and we describe and develop the fuzzy clustering methods and provide iterative solutions when the estimated GEV parameters are used as features. In Sections 3 and 4, we describe and analyse the simulation study and the application, respectively. Finally, in Section 5, some concluding remarks are given.

2. Methods

2.1. Generalised extreme value distribution

The generalised extreme value (GEV) distribution is a family of continuous probability distributions developed within extreme value theory to combine the Gumbel, Fréchet and Weibull families also known as Type I, II and III extreme value distributions. As a result of the extreme value theorem, the GEV distribution is the limiting distribution of normalised maxima of a sequence of independent and identically distributed random variables. Hence, the GEV distribution is used as an approximation to model the maxima of long finite sequences of random variables. The GEV distribution has the following form:

$$G(x) = \exp\left\{-\left[1+\xi\left(\frac{X-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}, \tag{1}$$

defined on $\{x : 1+\xi(\frac{x-\mu}{\sigma}) > 0\}$ where $-\infty < \mu < \infty, \sigma > 0$, and $-\infty < \xi < \infty$. The three parameters μ, σ and ξ are the location, scale and shape parameters, respectively. The shape parameter determines the three extreme value types. When $\xi < 0$, $\xi > 0$ or $\xi = 0$, the GEV distribution is the negative Weibull, the Fréchet or the Gumbel distribution, respectively. This is assumed to be the case by taking the limit of Eq. (1) as $\xi \to 0$.

For m years, the log-likelihood function for the annual maxima is given by

$$\ell(\mu,\sigma,\xi) = -m\log(\sigma) - (1+1/\xi)\sum_{i=1}^{m}\log\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right] - \sum_{i=1}^{m}\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right]^{-1/\xi}, \tag{2}$$

provided $1+\xi(\frac{x_i-\mu}{\sigma}) > 0$ for $i = 1, 2, \ldots, m$. Eq. (2) is valid for $\xi \neq 0$. For $\xi = 0$, the log-likelihood function for the annual maxima is given by

$$\ell(\mu,\sigma) = -m\log(\sigma) - \sum_{i=1}^{m}\left(\frac{x_i-\mu}{\sigma}\right) - \sum_{i=1}^{m}\exp\left[-\left(\frac{x_i-\mu}{\sigma}\right)\right]. \tag{3}$$

The above log-likelihood expression creates a common difficulty in extreme value analysis when the number of extreme events is small. This is particularly severe when the method of maxima over fixed intervals is used. As mentioned in Coles [4], a possible solution is to consider the r-largest order statistics over fixed intervals. For m years, the log-likelihood function for the annual r-largest order statistics is given by

$$\ell(\mu,\sigma,\xi) = -mr\log(\sigma) - (1+1/\xi)\sum_{i=1}^{m}\sum_{k=1}^{r}\log\left[1+\xi\left(\frac{x_i^{(k)}-\mu}{\sigma}\right)\right] - \sum_{i=1}^{m}\left[1+\xi\left(\frac{x_i^{(r)}-\mu}{\sigma}\right)\right]^{-1/\xi}, \tag{4}$$

where $x_i^{(r)} \leq x_i^{(r-1)} \leq \ldots \leq x_i^{(1)}$ are the r-largest values of the i-th year and the $x_i^{(k)}$ satisfy the following restriction $1+\xi\left(\frac{x_i^{(k)}-\mu}{\sigma}\right) > 0$ for $i = 1, 2, \ldots, m$ and $k = 1, 2, \ldots, r$. For $\xi = 0$, the log-likelihood function for the annual r-largest values is given by

$$\ell(\mu,\sigma) = -m\log(\sigma) - \sum_{i=1}^{m}\sum_{k=1}^{r}\left(\frac{x_i^{(k)}-\mu}{\sigma}\right) - \sum_{i=1}^{m}\exp\left[-\left(\frac{x_i^{(r)}-\mu}{\sigma}\right)\right]. \tag{5}$$

The number of largest order statistics per year, r, should be chosen carefully since small values of it will produce likelihood estimates with high variance, whereas large values of r will produce biased estimates. In practice, r is selected as large as possible, subject to adequate model diagnostics. The validity of the models can be checked through the application of graphical methods, in particular, the probability plot, the quantile plot and the return level plot; for further details, see Reiss and Thomas [21] and references therein.

The implications of a fitted extreme value model are usually made with reference to extreme quantiles. By inversion of the GEV distribution function, the quantile, x_p, for a specified exceedance probability p is

$$x_p = \mu - \frac{\sigma}{\xi}\left[1-\left(-\log(1-p)^{-\xi}\right)\right] \quad \text{for } \xi \neq 0, \tag{6}$$

and

$$x_p = \mu - \sigma\log\left[-\log(1-p)\right] \quad \text{for } \xi = 0. \tag{7}$$

x_p is referred to the return level associated with a return period $1/p$. It is expected to be exceeded by the annual maximum in any particular year with probability p.

Notice that, in applications to sea level and temperature series, the set of maxima or the r-largest order statistics may not always be stationary. In that case, the location parameter is time-varying and may take the form of a regression equation, namely

$$\mu_t = \beta_0 + \beta_1 Y_{1t} + \beta_1 Y_{1t} + \ldots + \beta_k Y_{1k}, \tag{8}$$

where β_0 is the level, and $\beta_1, \beta_2, \ldots, \beta_k$ are the regression coefficients. In our application to sea-level time series, we will consider the model $\mu_t = \beta_0 + \beta_1 Y_{1t}$ for the location parameter, where β_1 is the trend component. Refer to Coles [4] for more details about GEV distributions with time-varying components.

2.2. Fuzzy clustering models

While the traditional non-hierarchical clustering methods such as k-means and k-medoids generate mutually exclusive clusters, a fuzzy clustering method allows an observation to belong to more than one cluster simultaneously based on the minimisation of an objective function. Each observation belonging to a particular cluster has a membership degree which lies between 0 and 1. The k-means and k-medoids methods which are referred to as crisp clustering methods can be regarded as special cases of the fuzzy c-means and fuzzy c-medoids methods, respectively, where the membership degree of an observation belonging to a cluster is one and that of an observation not belonging to a cluster is zero.

In the literature, several authors have given different reasons for adopting fuzzy clustering approach (D'Urso [9]). As remarked by Hwang et al. [16], the fuzzy clustering approach offers other major advantages over the traditional clustering approach. Firstly, the fuzzy clustering models are computationally more efficient because dramatic changes in the value of cluster membership are less likely to occur in estimation procedures (McBratney and Moore, [19]). Secondly, fuzzy clustering has been shown to be less affected by local optima problems (Heiser and Groenen, [15]). Finally, the memberships for any given set of observations indicate whether there is a second-best cluster almost as good as the best cluster; a result which traditional clustering methods cannot uncover (Everitt et al. [13]).

We also consider weighted fuzzy c-means and weighted fuzzy c-medoids models of which the non-weighted fuzzy c-means and non-weighted fuzzy c-medoids models are special cases, respectively. In what follows, we describe the weighted fuzzy models and derive iterative solutions when the GEV estimates are used as the clustering features. In the weighted versions of the fuzzy clustering models, the weights could be fixed subjectively, a priori, by considering external or subjective conditions, or they could be computed objectively within a suitable clustering procedure. In particular, we can adopt either:

- An internal weighting system using an objective criterion where the weight values are not fixed a priori, but are computed via the minimisation algorithm; we get suitable weights such that the loss function is minimised with respect to the optimal values of the weights (refer to the iterative solutions that follow).
- An external weighting system where the weights can be fixed subjectively a priori, by taking into account external conditions.

2.2.1. Weighted fuzzy c-means clustering model based on GEV parameters of location, shape and scale (WGEV-FcM model)

The *WGEV-FcM* model is formalised as follows:

$$\min : \sum_{i=1}^{I}\sum_{c=1}^{C} u_{ic}^{m}\tilde{d}_{ic}^{2} = \sum_{i=1}^{I}\sum_{c=1}^{C} u_{ic}^{m}\sum_{s=1}^{3}(w_s \cdot d_{ics})^2 = \sum_{i=1}^{I}\sum_{c=1}^{C} u_{ic}^{m}\sum_{s=1}^{3}\left[w_s \cdot (x_{is} - h_{cs})\right]^2 \tag{9}$$

subject to the constraints

$$\sum_{c=1}^{C} u_{ic} = 1, \quad u_{ic} \geq 0, \tag{10}$$

$$\sum_{s=1}^{3} w_s = 1, \quad w_s \geq 0 \tag{11}$$

where: u_{ic} indicates the membership degree of the i-th time series to the c-th cluster; $m > 1$ is a weighting exponent that controls the fuzziness of the obtained partitions (for different heuristic methods, see D'Urso [9]); $\tilde{d}_{ic}^2 = \sum_{s=1}^{3}(w_s \cdot d_{ics})^2 = \sum_{s=1}^{3}[w_s \cdot (x_{is} - h_{cs})]^2$ represents the "weighted" Euclidean distance between the i-time series and the c-th prototype (centroid) time series based on the three parameters of the Generalised Extreme Value (GEV) distribution in which $x_{i1} \in (-\infty, +\infty)$, $x_{i2} \in (-\infty, +\infty)$, $x_{i3} \in [0, +\infty)$ represent, respectively, the observed location, shape and scale parameters of the Generalised Extreme Value (GEV) distribution; $h_{c1} \in (-\infty, +\infty)$, $h_{c2} \in (-\infty, +\infty)$, $h_{c3} \in [0, +\infty)$ indicate, respectively, the prototype (centroid) location, shape and scale parameters of the GEV distribution; and w_1, w_2 and w_3 are suitable weights associated with each parameter of the GEV distribution.

Proposition 1. *The iterative solutions to Eq. (9)–(11) satisfy (see the proof in the Appendix):*

$$u_{ic} = \frac{1}{\sum_{c'=1}^{C}\left[\frac{\sum_{s=1}^{3}(w_s \cdot d_{ics})^2}{\sum_{s=1}^{3}(w_s \cdot d_{ic's})^2}\right]^{\frac{1}{m-1}}}, \qquad w_s = \frac{1}{\sum_{s'=1}^{3}\left[\frac{\sum_{i=1}^{I}\sum_{c=1}^{C}(u_{ic}^m \cdot d_{ics}^2)}{\sum_{i=1}^{I}\sum_{c=1}^{C}(u_{ic}^m \cdot d_{ics'}^2)}\right]}, \qquad h_{cs} = \frac{\sum_{i=1}^{I} u_{ic}^m x_{is}}{\sum_{i=1}^{I} u_{ic}^m}. \tag{12}$$

Notice that, the weight w_s is intrinsically associated with the distance d_{ics} for the GEV parameter, s, while the overall dissimilarity is just a sum of the squares of these weighted distances. This allows us to appropriately tune the influence of the different GEV parameters when computing the dissimilarity between time series. Looking at the solution in Eq. (12), we observe that the weights w_s ($s = 1, 2, 3$) have a statistical meaning. In fact, they appear to mirror the heterogeneity of the total intra-cluster deviances, i.e., $\sum_{i=1}^{I}\sum_{c=1}^{C} u_{ic}^m d_{ic}^2$ across the different GEV parameters. In particular, weight w_s increases as long as the total intra-cluster deviance for the s-th GEV parameter decreases (compared with the remaining GEV parameters). Thus, the optimisation procedure tends to place more emphasis to the GEV parameters that are capable of increasing the within cluster similarity among the time series.

2.2.2. Fuzzy c-means clustering model based on GEV parameters (GEV-FcM)

By assuming the weights are determined a priori and fixing $w_s = 1$ for $s = 1, 2, 3$ in Eq. (9), we obtain the unweighted version of *WGEV-FcM* model, i.e., the *GEV-FcM* model:

$$\min \sum_{i=1}^{I}\sum_{c=1}^{C} u_{ic}^m \sum_{s=1}^{3}(x_{is} - h_{cs})^2$$

subject to the constraints

$$\sum_{c=1}^{C} u_{ic} = 1, \quad u_{ic} \geq 0.$$

2.2.3. Weighted fuzzy c-medoids clustering model based on location, shape and scale parameters (WGEV-FcMd model)

By applying the *WGEV-FcM* model we simultaneously obtain fuzzy partitions of a set of time series (by means of the corresponding three parameters of the GEV distribution) and estimate the prototype time series (prototypes of the three parameters of the GEV distribution), i.e. centroid time series (centroid parameters of the GEV distribution) that synthetically represent the features of the time series belonging to the corresponding clusters. However, there are several real cases where it is more realistic to represent/synthesise the cluster with a prototype time series belonging to the set of the observed time series, the so-called *medoid time series*. Then in our case, since the time series are represented by means of the three parameters of the respective GEV distributions, each cluster is represented by the medoid parameters of the GEV distribution. Then, we can formalise the so-called *WGEV-FcMd* model as follows:

$$\min : \sum_{i=1}^{I}\sum_{c=1}^{C} u_{ic}^m \tilde{d}_{ic}^2 = \sum_{i=1}^{I}\sum_{c=1}^{C} u_{ic}^m \sum_{s=1}^{3}(w_s \cdot d_{ics})^2 = \sum_{i=1}^{I}\sum_{c=1}^{C} u_{ic}^m \sum_{s=1}^{3}\left[w_s \cdot (x_{is} - \tilde{x}_{cs})\right]^2 \tag{13}$$

subject to the constraints

$$\sum_{c=1}^{C} u_{ic} = 1, \quad u_{ic} \geq 0, \tag{14}$$

$$\sum_{s=1}^{3} w_s = 1, \quad w_s \geq 0 \tag{15}$$

where u_{ic} indicates the membership degree of the i-th time series to the c-th cluster; $m > 1$ is a weighting exponent that controls the fuzziness of the obtained partition; $\tilde{d}_{ic}^2 = \sum_{s=1}^{3}[w_s \cdot (x_{is} - \tilde{x}_{cs})]^2$ represents the "weighted" Euclidean distance between the i-time series and the c-th medoid time series based on the three parameters of the GEV

distribution in which $\tilde{x}_{c1}, \tilde{x}_{c2}, \tilde{x}_{c3}$ indicate, respectively, the medoid location, shape and scale parameters of the GEV distribution and w_1, w_2 and w_3 are suitable weights associated with each parameter of the GEV distribution.

The membership degrees and the weights can be calculated in a heuristic manner in many different ways. For instance, we can adopt the membership degrees obtained by means of the *WGEV-FcM* model:

$$u_{ic} = \frac{1}{\sum_{c'=1}^{C}\left[\frac{\sum_{s=1}^{3}(w_s \cdot d_{ics})^2}{\sum_{s=1}^{3}(w_s \cdot d_{ic's})^2}\right]^{\frac{1}{m-1}}}, \quad w_s = \frac{1}{\sum_{s'=1}^{3}\left[\frac{\sum_{i=1}^{I}\sum_{c=1}^{C} u_{ic}^m \cdot d_{ics}^2}{\sum_{i=1}^{I}\sum_{c=1}^{C} u_{ic}^m \cdot d_{ics'}^2}\right]}. \tag{16}$$

Notice that the objective function in Eq. (13) cannot be minimised by means of the alternating optimisation algorithm, because the necessary conditions cannot be derived by differentiating it with respect to the medoids. Nonetheless, following Fu's heuristic algorithm a fuzzy clustering algorithm that minimises objective function in Eq. (13) can be built up (refer to Krishnapuram et al. [17]).

2.2.4. Fuzzy c-medoids clustering model based on GEV parameters (GEV-FcMd)

By assuming the weights are determined a priori and fixing $w_s = 1$, $s = 1, 2, 3$ in Eq. (13), we obtain the unweighted version of *WGEV-FcMd* model, i.e., the *GEV-FcMd* model:

$$\min : \sum_{i=1}^{I}\sum_{c=1}^{C} u_{ic}^m \sum_{s=1}^{3}(x_{is} - \tilde{x}_{cs})^2$$

subject to the constraints

$$\sum_{c=1}^{C} u_{ic} = 1, \quad u_{ic} \geq 0.$$

Notice that, setting $m = 1$ in the WGEV-FcM, GEV-FcM, WGEV-FcMd, GEV-FcMd models, we obtain their corresponding crisp versions.

2.2.5. Fuzzy clustering methods based on GEV parameters with time-varying components

When we have GEV distributions with time-varying components, i.e., the location parameter is time-varying and may take the form of a regression equation (see Section 2.1), we can reformulate the WGEV-FcM model (9)–(11) as follows (we call the new clustering model TV-WGEV-FcM model, i.e. WGEV-FcM model with time-varying components):

$$\min : \sum_{i=1}^{I}\sum_{c=1}^{C} u_{ic}^m \left\{ \left[\sum_{j=0}^{k} w_{1j} \cdot (\beta_{i1j} - {}_{\beta}h_{c1j}) \right]^2 + \left[w_2 \cdot (x_{i2} - h_{c2}) \right]^2 + \left[w_3 \cdot (x_{i3} - h_{c3}) \right]^2 \right\}$$

subject to the constraints

$$\sum_{c=1}^{C} u_{ic} = 1, \quad u_{ic} \geq 0, \qquad \sum_{j=0}^{k} w_{1j} + \sum_{s=2}^{3} w_s = 1, \quad w_{1j} \geq 0, \; w_s \geq 0,$$

where $\boldsymbol{\beta}_{i1} = (\beta_{i10}, \ldots, \beta_{i1k})'$ and ${}_{\beta}\boldsymbol{h}_{c1} = ({}_{\beta}h_{c10}, \ldots, {}_{\beta}h_{c1k})'$ indicate respectively, the regression location parameters vector (time-varying location parameters vector) of the i-th time series and of the c-th centroid; w_{1j} $(j = 0, \ldots, k)$ represents the weights associate to each regression location parameters.

Thus, from (9)–(11), it is straightforward to obtain the time-varying version of the iterative solutions (12). Similarly, we can reformulate the other clustering models shown in Sections 2.2.2, 2.2.3 and 2.2.4, in a time-varying framework.

2.2.6. Some cluster validity criteria

In our fuzzy clustering models, before computing the iterative solutions we have to fix a suitable number of clusters C. In the body of literature, many cluster-validity criteria have been suggested (D'Urso [12]). In particular, we consider for our models some cluster validity criteria based on the following indices:

Partition coefficient PC (Dunn [7]):

$$PC = \frac{1}{I}\sum_{i=1}^{I}\sum_{c=1}^{C} u_{ic}^2,$$

$\frac{1}{C} \leq PC \leq 1$: If $PC = 1/C$ we have maximum fuzziness; if $PC = 1$ we have a hard partition.

Partition entropy PE (Bezdek [3]):

$$PE = -\frac{1}{I}\sum_{i=1}^{I}\left(\sum_{c=1}^{C} u_{ic}\log u_{ic}\right).$$

$0 \leq PE \leq \log C$: If $PE = 0$ we have a hard partition; if $PE = \log C$ we have maximum fuzziness.

Modified partition coefficient MPC (Dunn [8]):

$$MPC = 1 - \frac{C}{C-1}(1 - PC)$$

$0 \leq MPC \leq 1$: If $MPC = 0$ we have maximum fuzziness; if $MPC = 1$ we have a hard partition.

Silhouette SIL (Rousseeuw [22]) and *Fuzzy Silhouette* SIL.F (Campello, Hruschka [4]):

The *Silhouette* (SIL) proposed by Rousseeuw [22] can be described as follows: consider an object $i \in \{1, \ldots, I\}$ belonging to cluster $c \in \{1, \ldots, C\}$. In the context of fuzzy partitions, this means that the membership of i-th time series to c-th cluster, u_{ic}, is higher than the membership of this time series to any other cluster. Let the average (squared Euclidean) distance of i-th time series to all other time series belonging to cluster c be denoted by a_{ic}. Also, let the average distance of this time series to all time series belonging to another cluster c', $c' \neq c$, be called $d_{ic'}$. Finally, let b_{ic} be the minimum $d_{ic'}$ computed over $c' = 1, \ldots, C$, $c' \neq c$, which represents the dissimilarity of i-th time series to its closest neighbouring cluster. Then, the silhouette of i-th time series is defined as follows: $s_i = \frac{b_{ic} - a_{ic}}{\max\{a_{ic}, b_{ic}\}}$. Clearly, the higher s_i is the better the assignment of i-th time series to c-th cluster. The *Silhouette* (SIL) is defined as the average of s_i over $i = 1, \ldots, I$:

$$SIL = \frac{1}{I}\sum_{i=1}^{I} s_i.$$

The best partition is achieved when the Silhouette SIL is maximised, which implies minimising the intra-cluster distance (a_{ic}) while maximising the inter-cluster distance (b_{ic}).

The *Fuzzy Silhouette* (SIL.F) (Campello, Hruschka [5]) makes explicit use of the fuzzy partition matrix $\mathbf{U} \equiv \{u_{ic} : i = 1, \ldots, I; c = 1, \ldots, C\}$. It may be able to discriminate between overlapped data clusters even if these clusters have their own distinct regions with higher data densities, since it considers the information contained in the fuzzy partition matrix $\mathbf{U}$ based on the degrees to which clusters overlap one another. The *Fuzzy Silhouette* (SIL.F) is defined as follows:

$$SIL.F = \frac{\sum_{i=1}^{I}(u_{ic} - u_{ic'})^{\alpha} s_i}{\sum_{i=1}^{n}(u_{ic} - u_{ic'})^{\alpha}}$$

where u_{ic} and $u_{ic'}$ are the first and second largest elements of the i-th row of the fuzzy partition matrix, respectively, and $\alpha \geq 0$ is a weighting coefficient.

Notice that with respect to other validity criteria based uniquely upon the fuzzy partition matrix (such as PC, PE and MPC), the *Fuzzy Silhouette* (SIL.F) takes into account, through the term s_i the geometrical information related to the data distribution.

3. Simulation study

In this simulation study we will generate seasonal time series that could represent daily temperatures or daily sea levels. We follow a similar procedure used in Maharaj et al. [18] in that we use a dynamic factor model that has been proposed by Safadi and Peña [23]. They used this model to generate air pollution series. The model has the following form:

$$y_t = Lf_t + e_t \tag{17}$$

$$f_t = \sum_{i=1}^{p} \rho_i f_{t-1} + w_t \tag{18}$$

where y_t is a $q \times 1$ vector of time series, L is a $q \times k$ matrix of factor loadings, $e_t \sim N(0, \Gamma)$, Γ is a $q \times q$ diagonal matrix. The factors f_t are represented by a $k \times 1$ vector which follows a multivariate autoregressive model where the AR matrices ρ_i are diagonal matrices with $\rho_i = \text{diag}(\rho_{i1}, \rho_{i2}, \ldots, \rho_{ik})$, $i = 1, 2, \ldots, p$ and $\{\rho_{1j}, \rho_{2j}, \ldots, \rho_{pj}\}$, $j = 1, 2, \ldots, k$ satisfy the stationary conditions and $w_t : N(0, I_k)$, where I_k is the identity matrix, and e_t and w_{t+h} are independent for all t and h.

In order to introduce seasonality into this dynamic factor model, a harmonic component is added to each factor in Eq. (18) as follows:

$$f_{t,k} = \sum_{i=1}^{p} \rho_{i,k} f_{t-1,k} + A_k \sin\left(\frac{2\pi t}{s}\right) + B_k \cos\left(\frac{2\pi t}{s}\right) + w_t, \tag{19}$$

where s is the length of the cycle. $A_k = R_k \cos\theta_k$ and $B_k = -R_k \sin\theta_k$. For each factor $f_{t,k}$, R_k is the amplitude or height of the cycle peaks, θ_k is the phase or the location of the peaks relative to time zero. Each factor can have different autoregressive dynamics, different seasonal dynamics, i.e., different amplitudes and phases.

We simulate three different scenarios with two groups of five series each to evaluate the fuzzy clustering methods when using the GEV parameters. Scenario 1: Series with different amplitudes but with the same phases; Scenario 2: Series with the same amplitudes but with different phases; Scenario 3: Series with different amplitudes and different phases.

For Scenario 1, we simulate five series with amplitude 10, and another five with amplitude 20. The phase for each of the ten series is set to 0.5. This is equivalent to having a single common factor, $f_{t,1}$, with $R_1 = 10$, $\theta_1 = 0.5$, and the factor loading matrix L being of dimension 10×1, i.e.,

$$L = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 2 & 2 & 2 & 2 & 2 \end{bmatrix}'. \tag{20}$$

For Scenario 2, we simulate five series with phase 0.5 and five series with phase 1. The amplitude for each of the ten series is set at 10. This is equivalent to having two common factors $f_{t,1}$ and $f_{t,2}$, with $R_1 = R_2 = 10$, $\theta_1 = 0.5$, $\theta_2 = 1$, and the factor loading matrix L being of dimension 10×2, i.e.,

$$L = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}'. \tag{21}$$

Scenario 3 is also a two common factor situation, but in this case we set $R_1 = 10$ and $\theta_1 = 0.5$ for the first set of five series, and $R_1 = 20$ and $\theta_1 = 1$ for the second set of five series. The factor loading matrix L is

$$L = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 & 2 & 2 & 2 & 2 \end{bmatrix}'. \tag{22}$$

For all three scenarios, AR matrices from a vector autoregressive model of order 1, VAR(1), are used to generate the factors with $\rho_1 = [0.5]$ for the first scenario and

$$\rho_1 = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix} \tag{23}$$

for both the second and third scenarios. For all three scenarios, the error series are generated from a $N(0, \Gamma)$ process with $\Gamma = I_{10}$, a 10×10 identity matrix. Figs. 1 to 3 show sections of series of length $T = 366$ generated for each scenario.

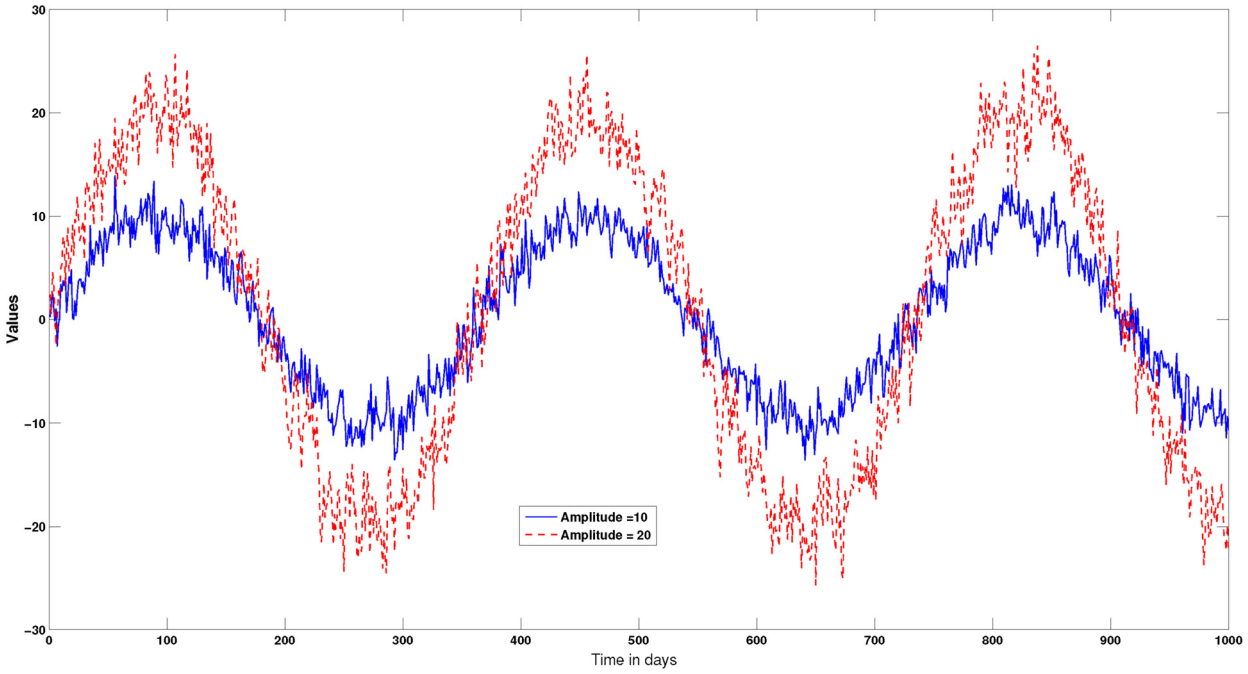

Fig. 1. Time series with different amplitudes.

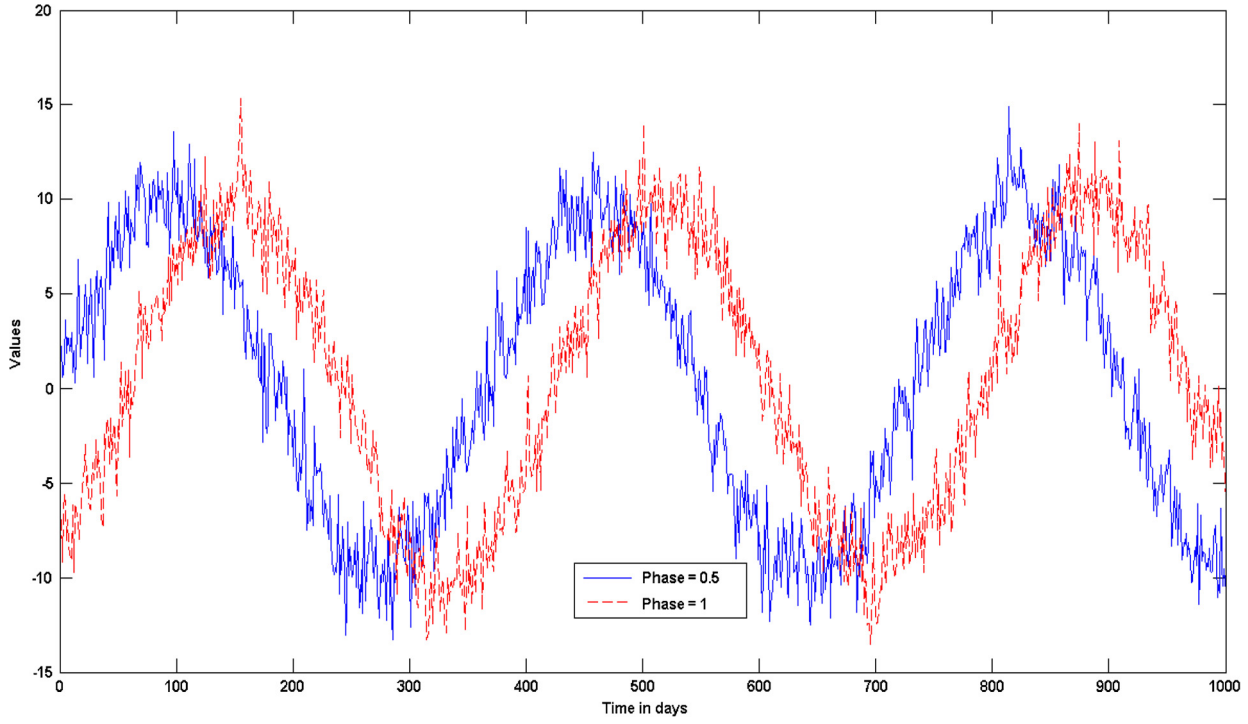

Fig. 2. Time series with different phases.

Daily-type series for 10 and 20 years are simulated for the three scenarios, and GEV estimates of shape, location and scale are obtained for one and two blocks per year, and are used as the fuzzy clustering features. The performance of the fuzzy clustering methods are evaluated over 100 simulations and in all cases m, the fuzziness parameter is set

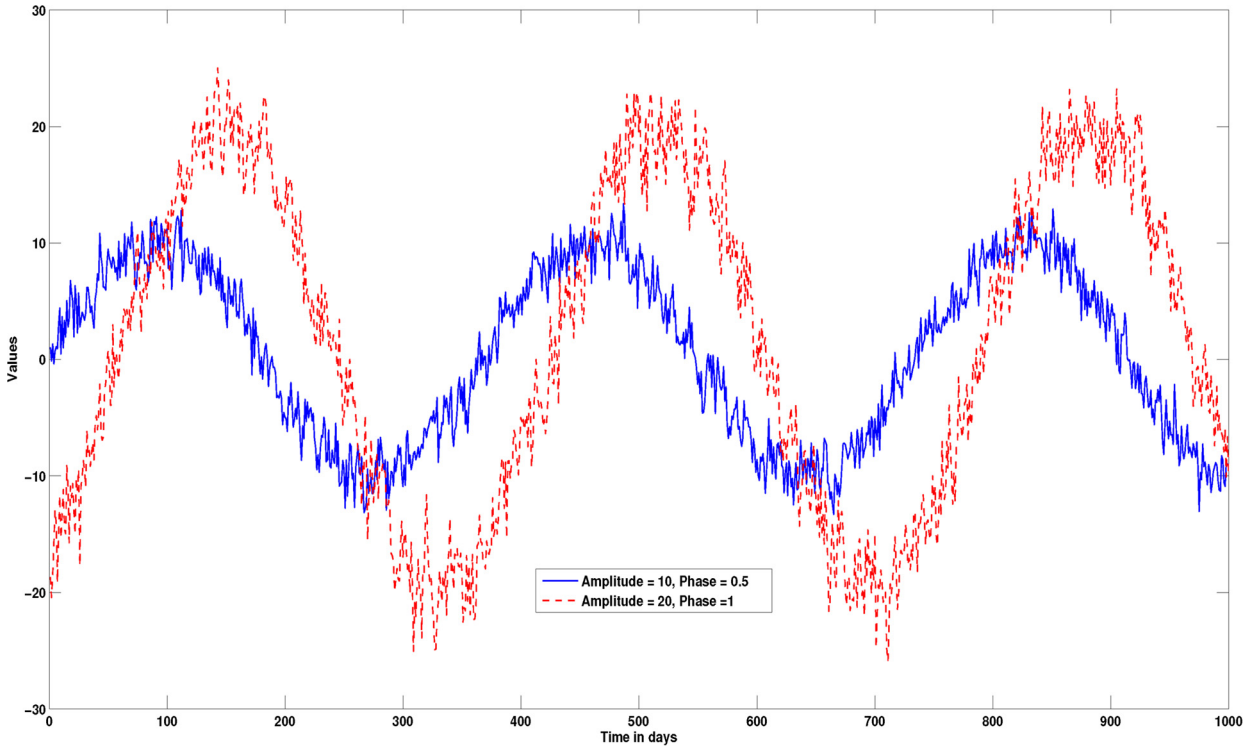

Fig. 3. Time series with different amplitudes and phases.

to 2 (similar results are observed for $m = 1.8$). Note that Bezdek [3] showed that fuzzy c-means clustering algorithm works well when $1.5 < m < 2.5$. We determined the percentage of correct classifications and whether they were fuzzy or not, based of a membership degrees of between 0.5 and 0.7 for fuzziness. Refer to Maharaj et al. [19] for more details on the cut-off value of 0.7 for non-fuzzy classification.

Tables 1 and 2 show the results of the fuzzy clustering methods based on the estimated GEV parameters for daily-type data over 10 and 20 years and for one and two blocks.

For time series of both 10 and 20 years,

- for the 1-block scenario, the weighted fuzzy methods outperform the un-weighted methods when distinguishing between series of different amplitudes only, and between series of both different amplitudes and phases; however for the different phase-only scenario, it is clear that none of the methods are able to distinguish between series of different phases; this is clearly indicated by 0% correct classification in Tables 1 and 2.
- for the 2-block scenario, the fuzzy c-means method is the best performer followed by the weighted fuzzy c-means method when distinguishing between series of different amplitudes only; for the different phase-only and for the different amplitude and phase scenarios, all methods perform to a high degree of accuracy.
- for both the 1-block and 2-block scenarios, the fuzzy c-means method reveals the highest proportion for fuzzy classification when distinguishing between series of different amplitude only, and between series of both different amplitudes and phases.

It is interesting to note that for the 1-block scenario, for the different amplitude-only, and the different amplitude and phase scenarios, the fuzzy c-means and fuzzy c-medoids methods perform much better for series over 20 years than over 10 years while the weighed methods perform just as well for both scenarios. Since 2 blocks per year provide double the number of maxima, it is clear that for the 2-block scenario, these methods perform well in distinguishing seasonal patterns for all three scenarios regardless of the different series lengths under consideration.

Fig. 4 shows the boxplots of the GEV parameter estimates for one simulation for the two groups of series with different amplitudes with one block per year over 20 years. It is clear from this boxplot that the estimated location

Table 1
Daily time series for 10 years: percentage of correct, non-fuzzy and fuzzy classification using GEV parameter estimates.

10 years, $T = 3660$	1 block			2 blocks		
	Correct	Non-fuzzy	Fuzzy	Correct	Non-fuzzy	Fuzzy
Different amplitudes						
Fuzzy c-means	0.54	0.10	0.44	0.99	0.42	0.57
Weighted fuzzy c-means	0.96	0.96	0.00	0.87	0.81	0.06
Fuzzy c-medoids	0.57	0.57	0.00	0.67	0.49	0.18
Weighted fuzzy c-medoids	1.00	1.00	0.00	0.68	0.54	0.14
Different phases						
Fuzzy c-means	0.00	0.00	0.00	1.00	0.98	0.02
Weighted fuzzy c-means	0.00	0.00	0.00	0.97	0.93	0.04
Fuzzy c-medoids	0.00	0.00	0.00	1.00	0.97	0.03
Weighted fuzzy c-medoids	0.00	0.00	0.00	0.96	0.96	0.00
Different amplitudes and phases						
Fuzzy c-means	0.57	0.14	0.43	1.00	0.89	0.11
Weighted fuzzy c-means	0.92	0.92	0.00	1.00	1.00	0.00
Fuzzy c-medoids	0.58	0.58	0.00	0.95	0.95	0.00
Weighted fuzzy c-medoids	1.00	1.00	0.00	1.00	1.00	0.00

Table 2
Daily time series for 20 years: percentage of correct, non-fuzzy and fuzzy classification using GEV parameter estimates.

10 years, $T = 7320$	1 block			2 blocks		
	Correct	Non-fuzzy	Fuzzy	Correct	Non-fuzzy	Fuzzy
Different amplitudes						
Fuzzy c-means	0.83	0.30	0.53	0.94	0.33	0.61
Weighted fuzzy c-means	0.97	0.97	0.00	0.70	0.62	0.08
Fuzzy c-medoids	0.75	0.75	0.00	0.64	0.53	0.11
Weighted fuzzy c-medoids	1.00	1.00	0.00	0.53	0.45	0.08
Different phases						
Fuzzy c-means	0.00	0.00	0.00	1.00	0.97	0.03
Weighted fuzzy c-means	0.00	0.00	0.00	1.00	1.00	0.00
Fuzzy c-medoids	0.00	0.00	0.00	1.00	0.99	0.01
Weighted fuzzy c-medoids	0.01	0.00	0.01	1.00	0.99	0.01
Different amplitudes and phases						
Fuzzy c-means	0.85	0.35	0.50	1.00	0.99	0.01
Weighted fuzzy c-means	0.99	0.99	0.00	1.00	1.00	0.00
Fuzzy c-medoids	0.83	0.83	0.00	1.00	1.00	0.00
Weighted fuzzy c-medoids	1.00	1.00	0.00	1.00	1.00	0.00

parameters make a greater contribution to group separation than the estimated shape and scale parameters. Similar observations were made when series of both different amplitude and phases were generated to those above.

4. Application

Rising sea levels are of great concern to coastal communities around the world and studies have been undertaken by relevant authorities in various countries to assess the impact of rising sea levels. Identifying areas with similar sea levels could contribute useful information to authorities to help develop common strategies to address rising sea levels that might occur in these areas, rather than having them focus on each coastal area, individually. Hence, resources could be used in an efficient manner to address concerns of future sea-level rises. To this end, fuzzy clustering com-

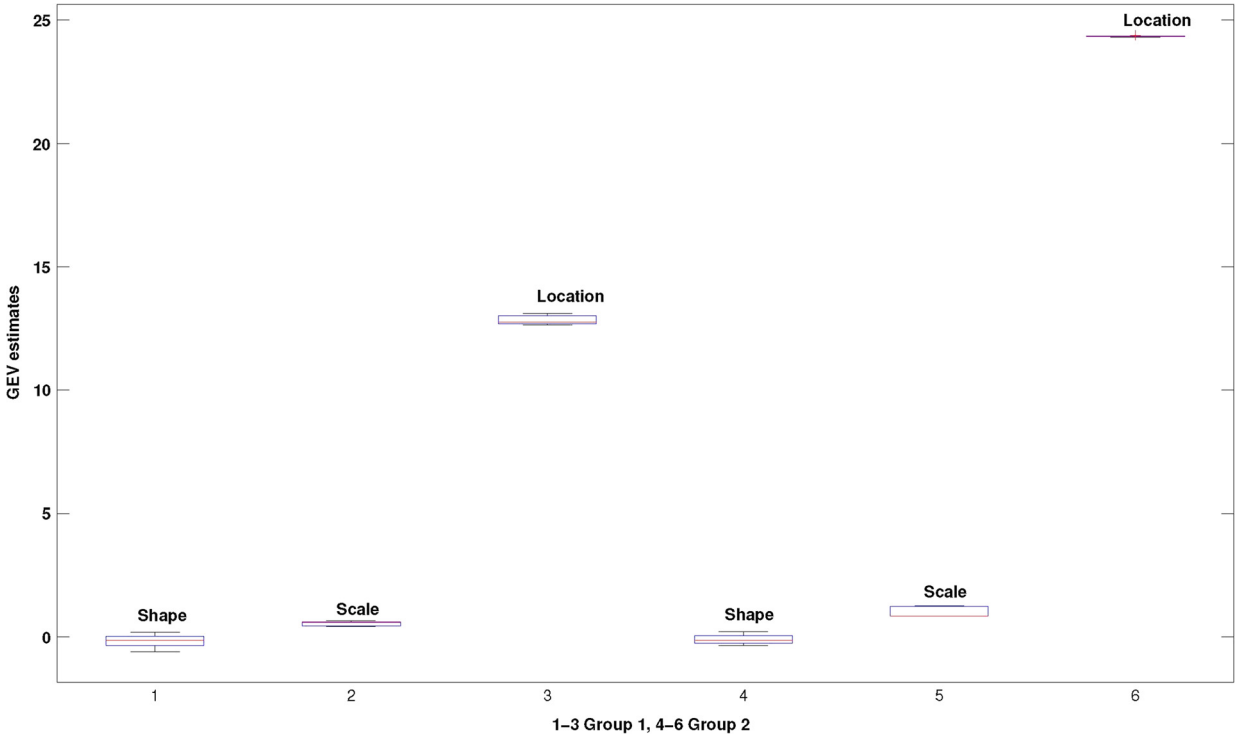

Fig. 4. Boxplot of GEV estimates for series with different amplitudes.

bined with extreme value analysis is applied to sea-level time series gathered from a number of tide gauge sites around the coast of Australia.

We consider time series of daily sea levels collected at 17 such tide gauge stations that have been obtained from the research quality database of the *University of Hawaii Sea Level Centre* (UHSLC). The measurement of sea levels is with reference to the zero point of the respective tide benchmarks. Refer to UHSCL [27] for more details about these sea level series. These 17 series for the period January 1993 to December 2012 were considered because they contained a manageable number of missing values. Each of the missing values was replaced by the mean of two values either before or after it. In many cases, the series record before January 1993 contained long tracks of missing values. Likewise, all the other Australian sea level time series available on this database contained long tracks of missing values or were too short to enable a useful analysis. Table 3 lists the tide gauge stations, their coastal directions, latitudes and longitudes.

A report by the Australian Government's Department of Climate Change in 2009 [6] provides findings of the first national assessment of the risks of climate change on Australia's coastal areas. In particular, the report discusses the possible impact of rising sea levels on these coastal areas in the coming decades. The aim of this application is to determine if the fuzzy clustering methods can group together time series of similar sea levels in a meaningful way and if one or more series could belong to more than one group.

Fig. 5 shows sea level series from the tide gauges sites. The seasonal patterns in the series are apparent with the Darwin and Wyndham displaying very higher sea levels, and Portland and Freemantle displaying much lower sea levels. Most of these series display gentle slopes.

Fig. 6 shows the values above the 95% percentile of each of four series mentioned above, where the differences in the sea levels between the higher and lower sea level series are clearly apparent.

It should be noted that the GEV distribution with constant location should be fitted to block maxima or the set of r-largest order statistics that are stationary. Since we have a large amount of daily data, fitting the GEV distribution to annual maxima will result in the loss of useful information. Hence, we fit the GEV to the set of r largest order statistics for each year over the 20 years of data. Furthermore, we model the trend of each series by fitting the GEV distribution with a time-varying location parameter since most of the series are not stationary.

Table 3
Tide gauge stations.

Coastal direction	Tide Gauge site	Latitude	Longitude
N	Booby Island	−10.600	141.920
SE	Brisbane	−27.367	153.167
NW	Broome	−18.001	122.219
S	Burnie	−41.052	145.907
NE	Cape Ferguson	−19.283	147.067
NW	Carnarvon	−24.900	113.650
NW	Cocos Island	−34.018	151.112
N	Darwin	−12.467	130.850
SW	Esperance	−33.871	121.895
SW	Freemantle	−32.050	115.733
SW	Hillarys	−31.798	115.745
S	Portland	−38.343	141.613
S	Spring Bay	−42.546	147.933
SE	Sydney	−33.850	151.233
S	Thevenard	−32.149	133.641
NE	Townsville	−19.270	147.060
N	Wyndham	−15.450	128.100

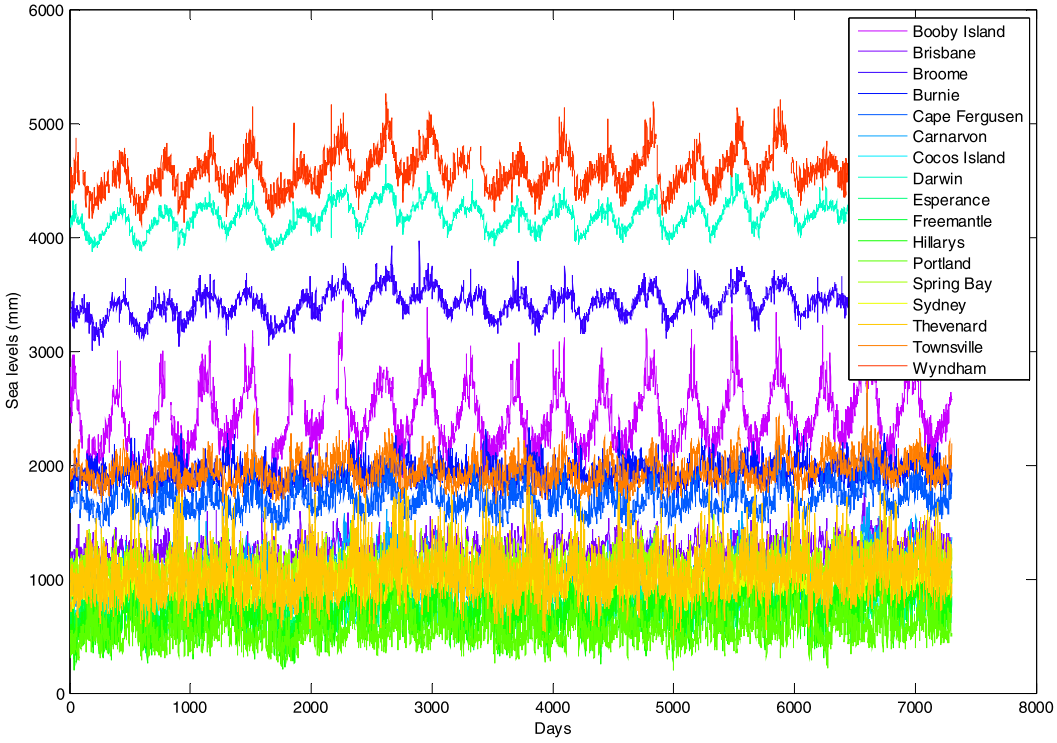

Fig. 5. Daily sea level series.

In order to extract the r-largest observations within each year, we follow the procedure adopted by Guedes Soares and Scotto [14], i.e., first we obtain the maximum of the year and we exclude a week of observations around this maximum, then we obtain the second largest value among the non-excluded observations. This exclusion guaranties that the first and second largest observations could be considered as independent. For the next largest value, we proceed in a similar way. We select an appropriate value of r based on the goodness of fit of the model, that is, by examining diagnostic plots, namely the residual probability and quantile plots. The fit is considered to be reasonable if the points are close to the diagonal in the plots. Refer to Coles [4] for more details.

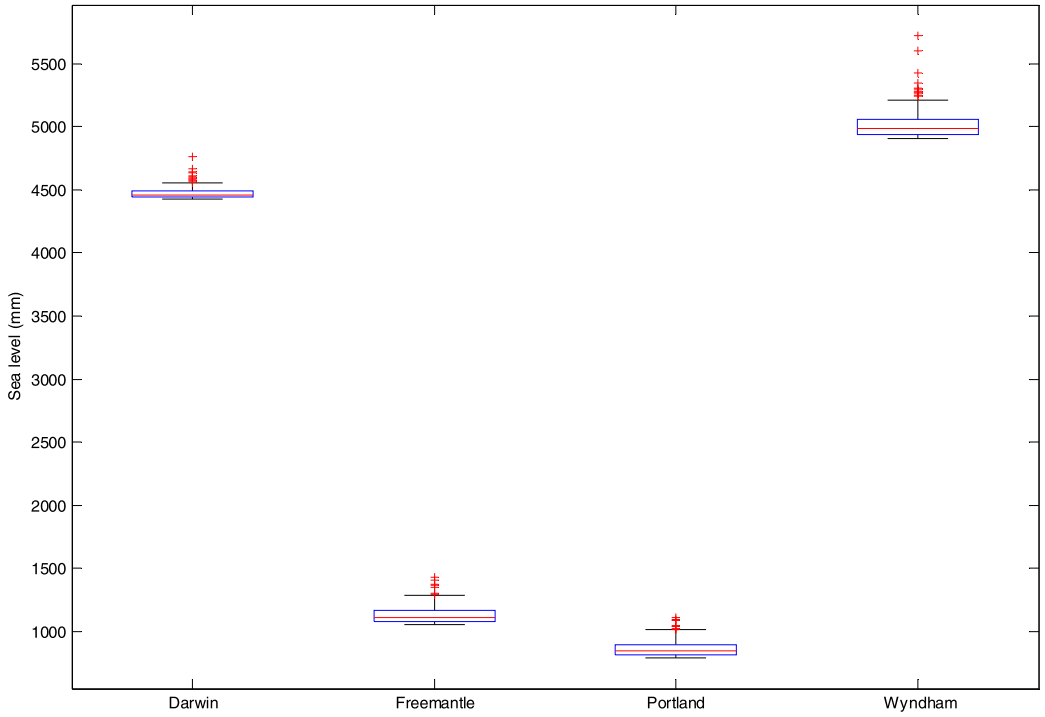

Fig. 6. Box-plot of the exceedances above the 95th percentile.

Figs. 7–9 show the best fitting residual probability plots which were the basis upon which the r-largest order statistics for each series were selected. For each plot, the x-axis represents the empirical probability while the y-axis represents the model probability. In most cases, the points are either on the diagonals or they do not deviate too far away from diagonals, indicating that the GEV distribution with time varying location fits the relevant time series reasonably well.

Table 4 shows the estimates of location (made up of level and slope), shape and scale with standard errors (below each estimate), obtained from fitting the GEV distribution to the set of the r-largest order statistics for each of the series.

We observe from Table 4, with the exception of the series from the tide gauge sites of Sydney and Thevenard, the slopes of the other sea level series are significantly different from zero at either the 5% ($**$) or 1% ($***$) levels. All the shape parameter estimates are greater than -0.5 which is the range within which the maximum likelihood estimates will have the usual asymptotic properties (Coles [4]).

It should be noted that the input for the fuzzy clustering methods are the extracted features of the time series, not the selected values of the sea levels. In particular, we use the estimates of the GEV parameters fitted to the set of r-largest sea level values, assuming that the location parameter follows a linear model. That is, we have four variables related to GEV distribution fitted to the set of r-largest sea levels. The algorithm could be summarise in the following steps:

- Given a set of n time series: $\boldsymbol{\mathcal{X}} = (\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_n})$, extract the selected set of features related to the highest sea levels.
- Obtain the fuzzy cluster solution using the extracted features as inputs.

We first apply the fuzzy c-means and fuzzy c-medoids method to the GEV estimates to determine the appropriate number of clusters using a number of optimal fuzzy cluster indices. We consider the cluster validity indices shown in Section 2.2.6.

For all of the indices under consideration it was found that a 2-cluster solution appeared to be the most appropriate when m was set to 2 or 1.8. Note that $1.5 < m < 2.5$ is an acceptable range for producing fuzzy clusters (Bezdek

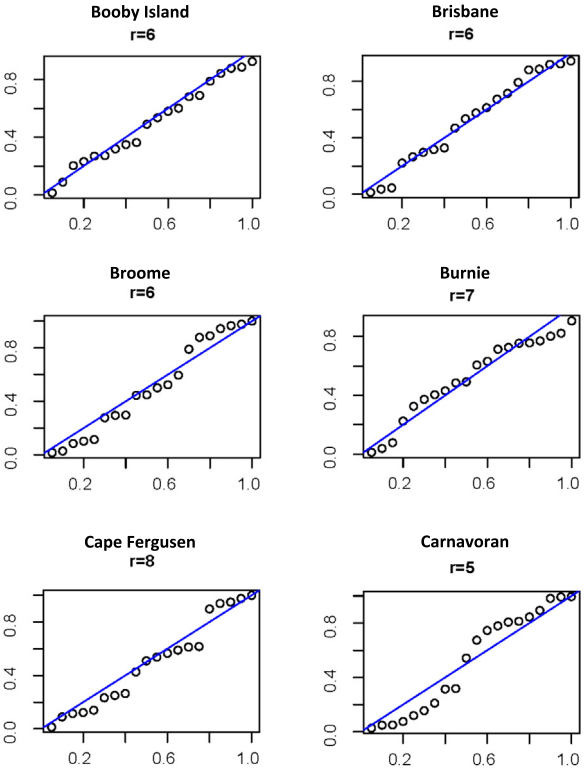

Fig. 7. Residual probability plots – Part 1.

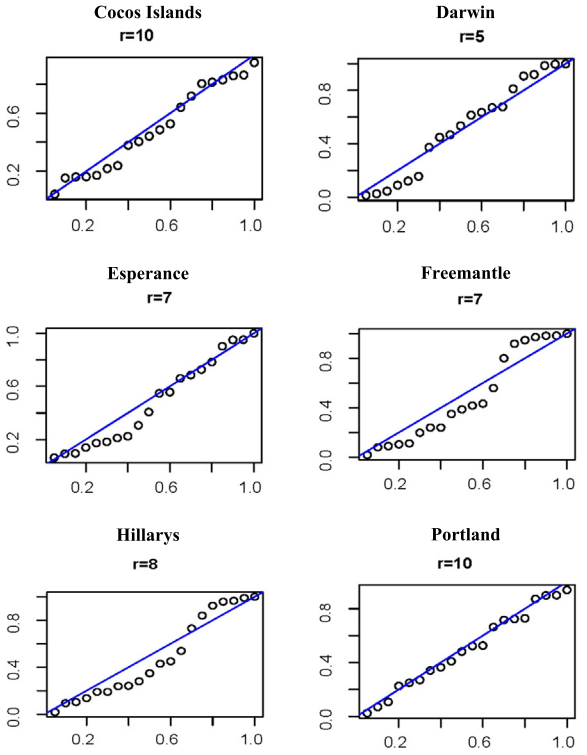

Fig. 8. Residual probability plots – Part 2.

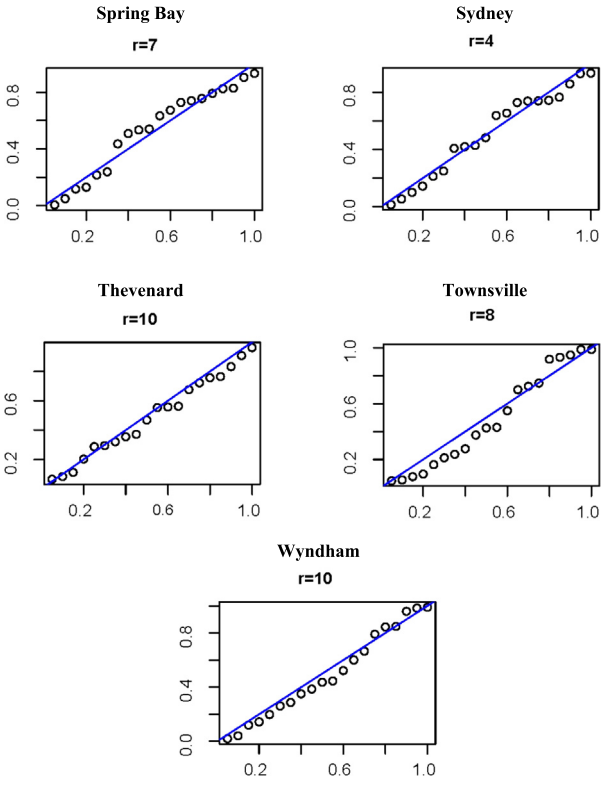

Fig. 9. Residual probability plots – Part 3.

[3]). Table 5 shows the values of these indices for 2, 3 and 4 clusters for $m = 2$ for both the fuzzy c-means and fuzzy c-medoids methods.

Tables 6 and 7 show the 2-cluster solutions for fuzzy c-means and c-medoids as well as the fuzzy weighed c-means and c-medoids methods for $m = 2$ and the equivalent hard cluster membership. A time series belongs to a particular hard cluster if its membership degree is the highest for that cluster. Table 7 also shows the weights associated with the level, slope, scale and shape which contributed to producing the weighted fuzzy cluster solutions.

All four methods produce the same cluster solution and it is clear, from the weights obtained from the weighted fuzzy cluster methods (as shown in Table 7), that the level parameter estimate is the dominating clustering feature. Table 8 shows clusters with the estimated location (level and slope), shape and scale parameters. Cluster 1 is associated with high sea levels while Cluster 2 is associated with lower sea levels and this is apparent from Table 9 which shows the mean, maximum and minimum location values for each of the clusters.

While most of the series generally have low fuzzy membership in a cluster other than the hard cluster they belong too, for all methods except the fuzzy c-medoids method, the series associated with Booby Island clearly has fuzzy membership in both clusters. This is also evident from its level estimate (in Table 8) why this would be the case. Furthermore, we observe from the daily sea-level series in Fig. 5 that the Booby Island series is clearly between the very high and very low sea levels.

It should be noted that two of the high sea level series in Cluster 1 (Darwin and Wyndham) as well as the series with fuzzy membership in both clusters (Booby Island) are from tide gauge sites on the north coast of Australia, while the other high sea level series in Cluster 1 (Broome) is on the north west coast.

Fig. 10 shows the results obtained by means of the WGEV-FcMd models. In particular the *scatterplot* (a) depicts the membership degrees of each time series to the two clusters, the *bar chart* (b) shows the weights obtained for each feature of the GEV distributions in the clustering process and in the *violin plot* (c) the different structural characteristics of the two clusters are shown with respect to the features of the GEV distributions. As we can see, the results shown in the *violin plot* corroborate the choice of the clusters and the fuzzy nature of some of the series.

Table 4
GEV estimates and standard errors.

	r	Level	Slope		Scale	Shape
Booby Island	6	2999.858	9.288	**	151.757	−0.287
		49.450	3.823		10.673	0.060
Brisbane	6	1457.435	5.685	***	57.492	−0.027
		14.576	0.985		6.343	0.074
Broome	6	3601.749	10.820	***	76.421	−0.004
		19.255	1.245		9.210	0.083
Burnie	7	2184.704	3.103	**	50.079	−0.270
		15.155	1.163		3.958	0.059
Cape Fergusen	8	1971.479	6.332	***	87.444	−0.013
		20.398	1.269		8.559	0.053
Carnarvon	5	1360.031	7.869	***	85.589	−0.216
		27.194	2.076		8.410	0.089
Cocos Island	10	1023.560	7.734	***	50.005	−0.401
		15.364	1.205		4.111	0.062
Darwin	5	4360.564	11.090	***	71.560	−0.128
		21.287	1.632		7.639	0.088
Esperance	7	1145.650	5.839	***	79.500	−0.195
		23.293	1.719		8.826	0.091
Freemantle	7	1154.500	7.005	***	84.448	−0.236
		26.046	1.967		6.999	0.064
Hillarys	8	1036.351	10.496	***	85.707	−0.212
		24.795	1.847		7.345	0.061
Portland	10	919.408	3.706	**	67.175	−0.305
		19.930	1.539		5.072	0.054
Spring Bay	7	1424.784	4.039	***	51.359	−0.208
		14.999	1.111		4.535	0.067
Sydney	4	1205.972	1.264		44.063	−0.364
		17.343	1.368		3.312	0.080
Thevenard	10	1697.502	3.926		95.554	−0.453
		33.716	2.754		9.819	0.078
Townsville	8	2225.723	5.558	***	87.471	−0.031
		21.184	1.371		7.838	0.046
Wyndham	10	4979.650	8.366	***	141.367	−0.102
		34.633	2.229		13.060	0.054

Fig. 11 shows the GEV density function for each of the sets of r-largest order statistic under consideration. The broken-line red density curves towards the right-hand side of the x-axis are those associated with Cluster 1 (mostly higher sea levels), and the solid blue line density curves towards the left-hand side are those associated with Cluster 2 (lower sea levels). Both the separation of Clusters 1 and 2 as well as fuzzy nature of the series associated with Booby Island (green dotted line) in these clusters are apparent from the positioning of the density curves on the x-axis.

4.1. Validation of the cluster solution

As a means of validating the fuzzy cluster solutions for the four methods, we applied the k-means and k-medoids methods with two clusters to the GEV estimates and found the crisp clusters had a 100% agreement with the hard

Table 5
Indices to determine optimal number of fuzzy clusters.

		clus = 2	clus = 3	clus = 4
Fuzzy c-means: $m = 2$				
Partition coefficient	PC	**0.914**	0.859	0.860
Partition entropy	PE	**0.156**	0.261	0.290
Modified partition coefficient	MPC	**0.829**	0.789	0.813
Silhouette	SIL	**0.873**	0.787	0.813
Fuzzy silhouette	SIL.F	**0.904**	0.835	0.845
Fuzzy -medoids: $m = 2$				
Partition coefficient	PC	**0.896**	0.859	0.778
Partition entropy	PE	**0.181**	0.261	0.424
Modified partition coefficient	MPC	**0.792**	0.788	0.704
Silhouette	SIL	**0.873**	0.787	0.638
Fuzzy silhouette	SIL.F	**0.889**	0.832	0.671

Table 6
Unweighted fuzzy cluster solutions.

		Fuzzy *c*-means			Fuzzy *c*-medoids		
		Cluster 1	Cluster 2	Hard cluster	Cluster 1	Cluster 2	Hard cluster
1	Booby Island	***0.671***	***0.329***	1	0.880	0.120	1
2	Brisbane	0.000	1.000	2	0.002	0.998	2
3	Broome	0.950	0.050	1	1.000	0.000	1
4	Burnie	0.135	0.865	2	0.253	0.747	2
5	Cape Fergusen	0.061	0.939	2	0.123	0.877	2
6	Carnarvon	0.001	0.999	2	0.000	1.000	2
7	Cocos Island	0.017	0.983	2	0.017	0.983	2
8	Darwin	0.992	0.008	1	0.940	0.060	1
9	Esperance	0.009	0.991	2	0.008	0.992	2
10	Freemantle	0.009	0.991	2	0.007	0.993	2
11	Hillarys	0.016	0.984	2	0.016	0.984	2
12	Portland	0.025	0.975	2	0.026	0.974	2
13	Spring Bay	0.000	1.000	2	0.001	0.999	2
14	Sydney	0.006	0.994	2	0.004	0.996	2
15	Thevenard	0.012	0.988	2	0.030	0.970	2
16	Townsville	0.153	0.847	2	0.284	0.716	2
17	Wyndham	0.942	0.058	1	0.873	0.127	1

clusters identified from the four fuzzy methods. This compatibility in cluster solutions would appear to validate the fuzzy solutions.

We also applied the k-nearest neighbour classification method (k-NN) to check on the validity of the hard cluster solution. k-NN is a non-parametric method that predicts class memberships of observations based on the k closest training examples in the feature space. Refer to Altman [2] for more details.

We ran the k-NN classification algorithm with one to six neighbours on the GEV parameter estimates of the 17 regions with the groups designated according to the hard cluster solutions. The leave-out-one cross-validation method was used to evaluate the quality of the classification. The algorithm was run 17 times and the mean classification error obtained for each nearest neighbour. The results are shown in Table 10 from where it is clear that for one and two nearest neighbours with 0% error rates, the GEV estimates provide good separation features. This further validates our fuzzy clustering solutions.

4.2. Return levels

One of the advantages of using the GEV features for clustering is that we can interpret the fuzzy cluster solutions using the N-years returns levels (extreme quantiles), that is, the values that can be exceeded once every N-years.

Table 7
Weighted fuzzy cluster solutions.

		Fuzzy weighted *c*-means			Fuzzy weighted *c*-medoids		
		Cluster 1	Cluster 2	Hard cluster	Cluster 1	Cluster 2	Hard cluster
1	Booby Island	***0.697***	***0.303***	1	***0.548***	***0.452***	1
2	Brisbane	0.026	0.974	2	0.037	0.963	2
3	Broome	0.935	0.065	1	0.700	0.300	1
4	Burnie	0.136	0.864	2	0.130	0.870	2
5	Cape Fergusen	0.102	0.898	2	0.100	0.900	2
6	Carnarvon	0.016	0.984	2	0.018	0.982	2
7	Cocos Island	0.041	0.959	2	0.047	0.953	2
8	Darwin	0.972	0.028	1	0.861	0.139	1
9	Esperance	0.010	0.990	2	0.000	1.000	2
10	Freemantle	0.015	0.985	2	0.006	0.994	2
11	Hillarys	0.062	0.938	2	0.061	0.939	2
12	Portland	0.031	0.969	2	0.023	0.977	2
13	Spring Bay	0.013	0.987	2	0.026	0.974	2
14	Sydney	0.054	0.946	2	0.078	0.922	2
15	Thevenard	0.058	0.942	2	0.090	0.910	2
16	Townsville	0.181	0.819	2	0.148	0.852	2
17	Wyndham	0.925	0.075	1	1.000	0.000	1
		Weights			Weights		
		level	0.547		level	0.440	
		slope	0.173		slope	0.226	
		scale	0.157		scale	0.162	
		shape	0.123		shape	0.172	

Table 8
Parameters estimates according to the cluster solutions.

	Level	Slope	Scale	Shape	Hard cluster
Booby Island	2999.858	9.288	151.757	−0.287	1
Broome	3601.749	10.820	76.421	−0.004	1
Darwin	4360.564	11.090	71.560	−0.128	1
Wyndham	4979.650	8.366	141.367	−0.102	1
Brisbane	1457.435	5.685	57.492	−0.027	2
Burnie	2184.704	3.103	50.079	−0.270	2
Cape Ferguson	1971.479	6.332	87.444	−0.013	2
Carnarvon	1360.031	7.869	85.589	−0.216	2
Cocos Island	1023.560	7.734	50.005	−0.401	2
Esperance	1145.650	5.839	79.500	−0.195	2
Freemantle	1154.500	7.005	84.448	−0.236	2
Hillarys	1036.351	10.496	85.707	−0.212	2
Portland	919.408	3.706	67.175	−0.305	2
Spring Bay	1424.784	4.039	51.359	−0.208	2
Sydney	1205.972	1.264	44.063	−0.364	2
Thevenard	1697.502	3.926	95.554	−0.453	2
Townsville	2225.723	5.558	87.471	−0.031	2

Table 9
Means, maxima and minima of the 2-cluster solution.

	Mean	Maximum	Minimum
Cluster 1	4269	5126	3269
Cluster 2	1556	2323	909

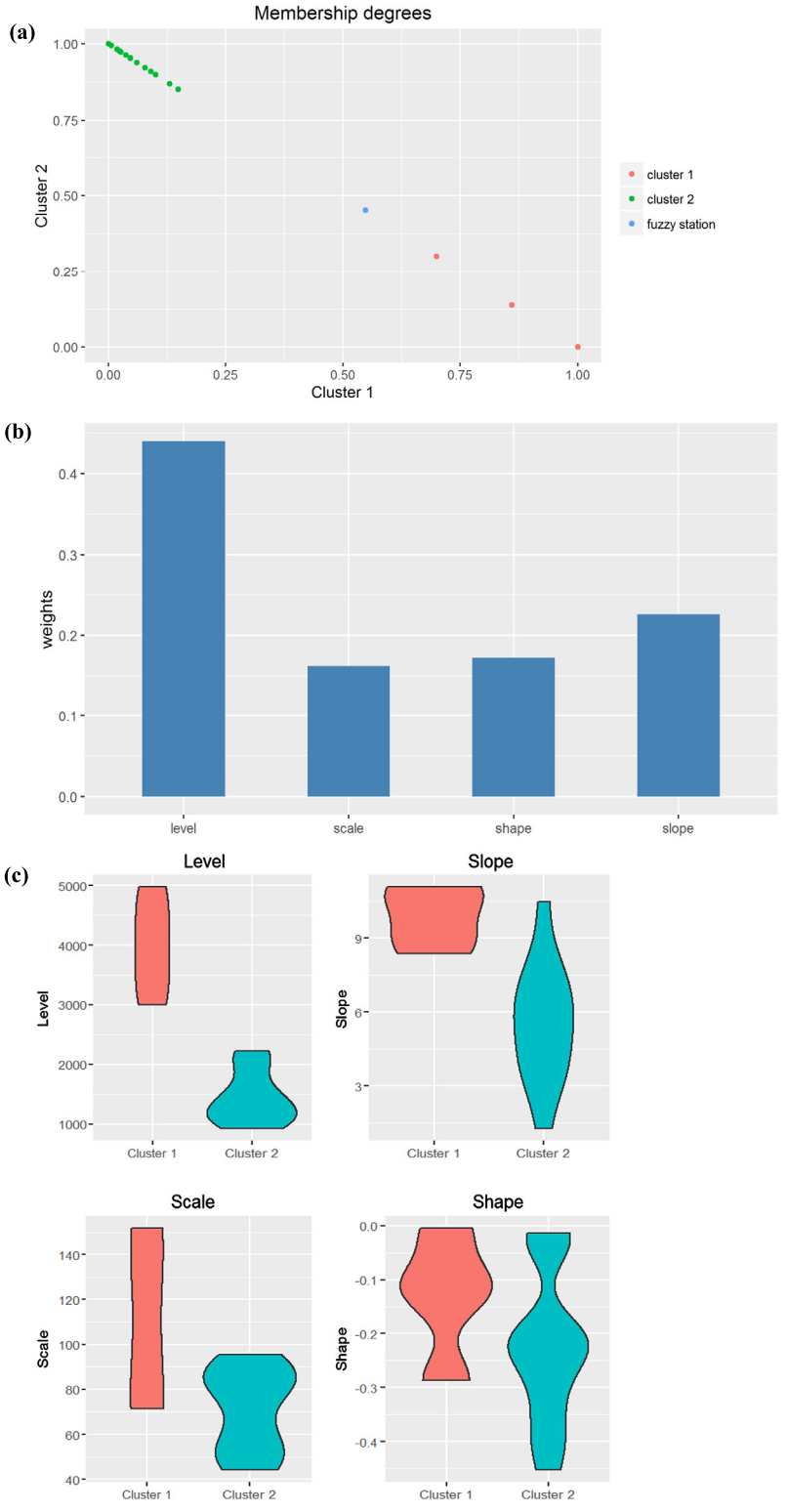

Fig. 10. Some plots on the cluster results.

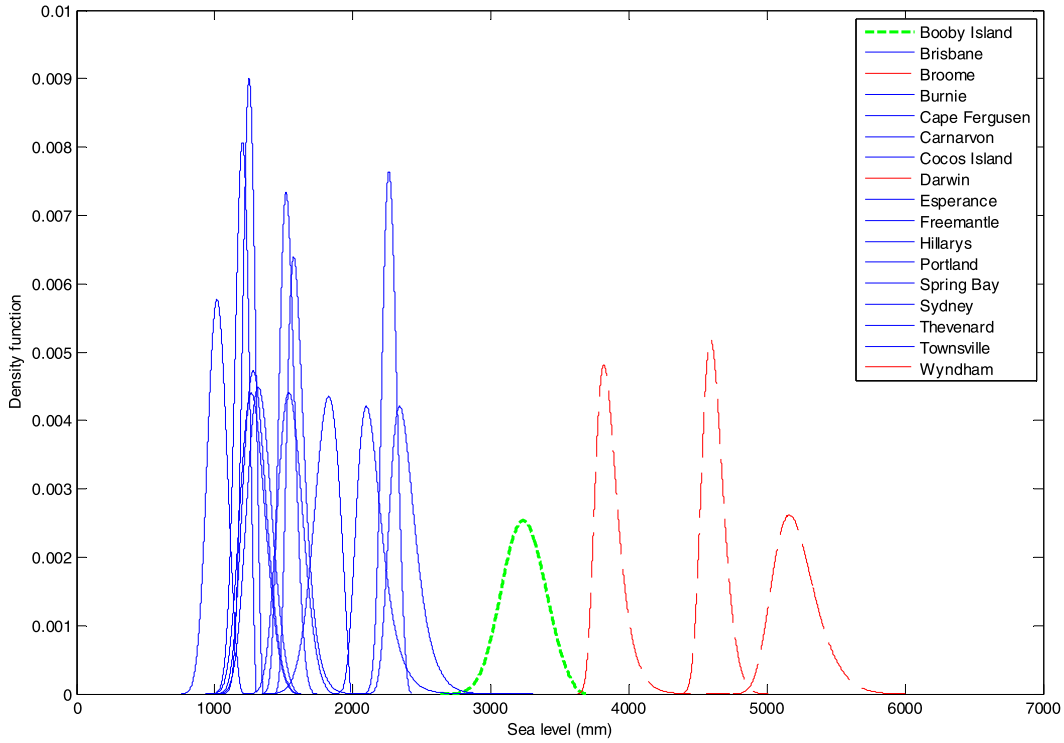

Fig. 11. GEV density functions: two clusters. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 10
k-nearest neighbours classification.

Number of neighbours	Classification error
1	0%
2	0%
3	6%
4	6%
5	6%
6	6%

We use the expressions in Eq. (6) or in Eq. (7) to obtain 25, 50 and 100 years in order to gain some insight into the 2-cluster solution obtained from the fuzzy clustering methods. The cluster mean returns with 95% confidence intervals for the fuzzy *c*-means method are presented in Table 11 and they confirm and complement our previous interpretation, that is, (1) the first cluster corresponds to localities having high sea levels that could be no more than 4638 millimetres in periods of 100 years, (2) the second cluster corresponds to localities having lower sea levels that could be no more than 1820 millimetres in periods of 100 years. Compared to the 2012 Cluster 1 and Cluster 2 mean maxima of approximately 4269 millimetres and 1556 millimetres, respectively, these returns appear to be realistic.

In particular, the difference between the 25 year cluster mean returns and the corresponding cluster mean location values (obtained from Table 8) based on the GEV fit to the 1993 to 2012 series of *r*-largest order statistics, in Clusters 1 and 2, are 0.198 and 0.171 metres, respectively. While these projections correspond to projected sea level rises in 2037, they do fall within the range of sea level projections of 0.132, 0.146 and 0.200 developed under three different scenarios by the Commonwealth Science and Industry Research Organisation of Australia (CSIRO) for 2030, relative to 1990 (refer to Table 2.1 in [6]).

Similar results were obtained for returns using the other fuzzy clustering methods.

Table 11
2012 cluster mean maxima and 25, 50 and 100 year mean returns levels.

Cluster	1	2
2012 cluster mean maxima	4268.750	1556.154
25-yr	4466.939	1726.870
95% C.I.	4413.471	1691.560
	4520.407	1762.179
50-yr	4515.189	1752.873
95% C.I.	4451.737	1714.469
	4578.642	1791.278
100-yr	4559.439	1776.148
95% C.I.	4481.181	1732.385
	4637.696	1819.911

5. Concluding remarks

New generalised procedures for fuzzy clustering taking into account weights have been developed, and iterative solutions based on the GEV parameter estimators have been derived. It is clear from the simulation study, the GEV location estimates, in particular, are good separation features for the clustering of seasonal time series. However, it has been observed from outcomes of the application to real world data, namely, the sea-level time series, the level components of the locations estimates are mainly responsible for contributing to cluster separation. For the application, we also noted that the fuzzy clustering solutions can be meaningfully interpreted and validated. It should be noted that if only crisp clustering methods were used to identify similar sea-level series, the useful information about overlapping clusters would be lost.

An added advantage of using GEV modelling to analyse seasonal time series is that return level statements can be made about long-term extremes. In this case, concerning groups of similar sea-level time series, this information could contribute to economic and technical planning decisions to help address likely long-term sea levels rises. Of course, other variables (coastal human activity, atmospheric ocean processes, greenhouse gas concentrations, etc.) can also be analysed with the proposed procedure.

The future direction that we will be embarking on in analysing real time series extremes is examining the fuzzy behaviour of the series by incorporating their spatial features as added sources of information.

Acknowledgements

The authors thank the two referees for their useful comments and suggestions which helped improve the quality and presentation of this paper.

Appendix A

Proof of Proposition 1. First, fix h_{cs} and w_s, to determine the membership degrees u_{ic}. The solution of Eqs. (9)–(11) is found by means of Lagrange multipliers. Thus, we consider the following Lagrangian function:

$$L_m(\boldsymbol{u}_i, \lambda) = \sum_{c=1}^{C} u_{ic}^m \sum_{s=1}^{3} (w_s \cdot d_{ics})^2 - \lambda \left(\sum_{c=1}^{C} u_{ic} - 1 \right) \tag{A.1}$$

where $\boldsymbol{u}_i = (u_{i1}, \ldots, u_{ic}, \ldots, u_{iC})'$ and λ is the Lagrange multiplier. Therefore, setting the first derivatives with respect to u_{ic} and λ equal to zero, yields

$$\frac{\partial L_m(\boldsymbol{u}_i, \lambda)}{\partial u_{i'c'}} = 0 \Leftrightarrow m \cdot u_{i'c'}^{m-1} \sum_{s=1}^{3} (w_s \cdot d_{i'c's})^2 - \lambda = 0 \tag{A.2}$$

$$\frac{\partial L_m(\boldsymbol{u}_i,\lambda)}{\partial\lambda}=0\Leftrightarrow\sum_{c=1}^{C}u_{ic}-1=0 \tag{A.3}$$

From Eq. (A.2) and by considering Eq. (A.3) we obtain u_{ic} which satisfies Eq. (12). Now, fixing u_{ic} and h_{cs}, we calculate the weights w_s. By considering the Lagrangian function:

$$L_m(\boldsymbol{w},\xi)=\sum_{i=1}^{I}\sum_{c=1}^{C}u_{ic}^m\sum_{s=1}^{3}(w_s\cdot d_{ics})^2-\xi\left(\sum_{s=1}^{3}w_s-1\right) \tag{A.4}$$

where $\boldsymbol{w}=(w_1,w_2,w_3)'$ and ξ is the Lagrange multiplier; by setting the first derivatives with respect to $w_{s'}$ and ξ equal to zero, we obtain:

$$\frac{\partial L_m(\boldsymbol{w},\xi)}{\partial w_{s'}}=0\Leftrightarrow 2\cdot w_{s'}\sum_{i=1}^{I}\sum_{c=1}^{C}u_{ic}^m\cdot d_{ics'}^2-\xi=0 \tag{A.5}$$

$$\frac{\partial L_m(\boldsymbol{w},\xi)}{\partial\xi}=0\Leftrightarrow\sum_{s=1}^{3}w_s-1=0 \tag{A.6}$$

From Eq. (A.5) we have:

$$w_{s'}=\frac{\xi}{2\sum_{i=1}^{I}\sum_{c=1}^{C}u_{ic}^m\cdot d_{ics'}^2} \tag{A.7}$$

and using Eq. (A.6):

$$\frac{\xi}{2}=\frac{1}{\sum_{s''=1}^{3}\left(\frac{1}{\sum_{i=1}^{I}\sum_{c=1}^{C}u_{ic}^m\cdot d_{ics''}^2}\right)}. \tag{A.8}$$

Then, replacing ξ in Eq. (A.7) by ξ from Eq. (A.8), we obtain w_s which satisfies Eq. (12). For computing h_{cs}, given u_{ic}, we have to solve an unconstrained minimisation problem. In particular, since

$$\min\sum_{i=1}^{I}\sum_{c=1}^{C}u_{ic}^m\sum_{s=1}^{3}(w_s\cdot d_{ics})^2=\sum_{c=1}^{C}\sum_{s=1}^{3}w_s^2\left[\min\sum_{i=1}^{I}u_{ic}^m(x_{is}-h_{cs})^2\right]$$

we have, putting $V_m(h_{cs})=\sum_{i=1}^{I}u_{ic}^m(x_{is}-h_{cs})^2$, the solution h_{cs}, setting the first derivatives of $V_m(h_{cs})$ with respect to h_{cs} equal to zero. □

References

[1] A.M. Alonso, P. De Zea Bermudez, M.G. Scotto, Comparing generalized Pareto models fitted to extreme observations: an application to the largest temperatures in Spain, Stoch. Environ. Res. Risk Assess. 28 (2014) 1221–1233.
[2] N.S. Altman, An introduction to kernel and nearest-neighbour nonparametric regression, Am. Stat. 463 (1992) 175–185.
[3] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York, 2001.
[4] S. Coles, An Introduction to Statistical Modeling of Extreme Values, Springer-Verlag, London, 2001.
[5] R.J. Campello, E.R. Hruschka, A fuzzy extension of the silhouette width criterion for cluster analysis, Fuzzy Sets Syst. 157 (21) (2006) 2858–2875.
[6] Department of Climate Change, Commonwealth of Australia, www.climatechange.gov.au, ISBN 978-1-921298-71-4, 2009.
[7] J.C. Dunn, Well separated clusters and optimal fuzzy partitions, J. Cybern. 4 (1) (1974) 95–104.
[8] J.C. Dunn, Indices of partition fuzziness and the detection of clusters in large data sets, in: M. Gupta, G. Saridis (Eds.), Fuzzy Automata and Decision Processes, Elsevier, New York, 1977.
[9] P. D'Urso, Fuzzy c-means clustering models for multivariate time-varying data: different approaches, Int. J. Uncertain. Fuzziness Knowl.-Based Syst. 12 (3) (2004) 287–326.
[10] P. D'Urso, Fuzzy clustering for data time arrays with inlier and outlier time trajectories, IEEE Trans. Fuzzy Syst. 13 (5) (2005) 583–604.
[11] P. D'Urso, E.A. Maharaj, Autocorrelation-based fuzzy clustering of time series, Fuzzy Sets Syst. 160 (2009) 3565–3589.
[12] P. D'Urso, Fuzzy clustering, in: C. Hennig, M. Meila, F. Murtagh, R. Rocci (Eds.), Handbook of Cluster Analysis, in: Chapman and Hall/CRC Handbooks of Modern Statistical Methods, 2015.
[13] B.S. Everitt, S. Landau, M. Leese, Cluster Analysis, 4th ed., Arnold Press, London, 2001.

[14] C. Guedes Soares, M.G. Scotto, Application of the r-order statistics for long-term predictions of significant wave heights, Coast. Eng. 51 (2004) 387–394.
[15] W.J. Heiser, P.J.F. Groenen, Cluster differences scaling with a within-clusters loss component and a fuzzy successive approximation strategy to avoid local minima, Psychometrika 62 (1997) 63–83.
[16] H. Hwang, W.S. De Sarbo, Y. Takane, Fuzzy clusterwise generalized structured component analysis, Psychometrika 72 (2007) 181–198.
[17] R. Krishnapuram, A. Joshi, O. Nasraoui, L. Yi, Low-complexity fuzzy relational clustering algorithms for web mining, IEEE Trans. Fuzzy Syst. 9 (4) (2001) 595–607.
[18] E.A. Maharaj, A.M. Alonso, P. D'Urso, Clustering seasonal time series using extreme value analysis: an application to Spanish temperature time series, Commun. Stat. Case Stud. Data Anal. 1 (2015) 175–191.
[19] E.A. Maharaj, P. D'Urso, D.U.A. Galagedera, Wavelets-based fuzzy clustering of time series, J. Classif. 27 (2) (2010) 231–275.
[20] F.J. Méndez, M. Menéndez, A. Luceño, I.J. Losada, Analysing monthly extreme sea levels with a time-dependent GEV model, J. Atmos. Ocean. Technol. 24 (2007) 894–911.
[21] R-D. Reiss, M. Thomas, Statistical Analysis of Extreme Values with Applications to Insurance, Finance, Hydrology and Other Fields, 3rd edition, Birkhäuser, Basel, Boston, Berlin, 2000.
[22] P.J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. 20 (1987) 53–65.
[23] T. Safadi, D. Peña, Bayesian analysis of dynamic factor models: an application to air pollution and mortality in Sao Paulo, Brazil, Environmetrics 19 (2008) 582–601.
[24] M.G. Scotto, S.M. Barbosa, A.M. Alonso, Clustering time series of sea levels: extreme value approach, J. Waterw. Port Coast. Ocean Eng. 136 (2010) 2793–2804.
[25] M.G. Scotto, S.M. Barbosa, A.M. Alonso, Extreme value and cluster analysis of European daily temperature series, J. Appl. Stat. 38 (12) (2011) 215–225.
[26] M.N. Tsimplis, D.L. Blackman, Extreme sea-level distribution and return periods in the Aegean and Ionian seas, Estuar. Coast. Shelf Sci. 44 (1997) 79–89.
[27] University of Hawaii Sea Level Centre UHSLC, http://uhslc.soest.hawaii.edu/data/download/rq.
[28] A.S. Unikrishnan, D. Sundar, D.L. Blackman, Analysis of extreme sea level along the east coast of India, J. Geophys. Res. 190 (2004) C06023.