Fuzzy Clustering of Extremes: An Application to Sea-Level Time Series

Pierpaolo D'Urso, Dipartimento di Scienze Sociali ed Economiche, Sapienza - Universitá di Roma, Italy
Elizabeth A. Maharaj, Department of Econometrics and Business Statistics, Monash Business School, Monash University, Australia
Andrés M. Alonso, Department of Statistics, IFL, Universidad Carlos III de Madrid, Spain

Abstract Rising sea levels are of great concern to coastal communities around the world and studies have been undetaken 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. To this end, fuzzy clustering combined with extreme value analysis is applied to sea-level time series gathered from a number of tide gauges around the coast of Australia. Input features into the fuzzy clustering methods are parameter estimates of location, scale and shape obtained from fitting the generalised extreme value (GEV) distribution to block maxima of the time series. New generalised procedures for fuzzy clustering taking into account weights are developed, and iterative solutions based on the GEV parameter estimators are obtained. Simulation studies conducted to evaluate the methods, reveal fairly good performance, while outcomes from the application can be meaningfully interpreted and are well validated.

Keywords Fuzzy Clustering · Block Maxima · 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 extreme value analysis of shorter-term data may acceptable from an enginnering perpective in terms of designing, say for example, sea-protection walls on particular coastal areas.

Authors such as Tsimpis and Blackman [20], Unikrishnan et al. [21], Méndez et al. [14] and Scotto et al. [18] have used extreme value analysis to study sea level extremes, while Scotto et al. [19] and Alonso et al. [1] are amongst others who have applied extreme analysis to temperature extremes. In particular, Scotto et al. [18] combined a Bayesian analysis of extreme sea levels to estimate predictive distributions with hierarchical cluster analysis to distinguish groups of North Atlantic sea locations. Scotto et al. [19] 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 focussed on using clustering methods to group together locations based on predictive distributions while in a recent study, Maharaj

Submitted version
Accepted at Fuzzy Sets and Systems
https://doi.org/10.1016/j.fss.2016.10.006

et al. [11] considered non-hierarchical clustering methods and classifications 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. Here, we extend the study of Maharaj et al. [11] to group the series across the available record using fuzzy clustering methods.

In this study, we simulate seasonal times series analogous to sea levels and we obtain maxima of blocks of observations. To each series of block maxima, we fit a generalised extreme value (GEV) distribution, estimate the parameters of shape, location and scale, 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. New generalised procedures for fuzzy clustering taking into account weights are developed, and iterative solutions based on the GEV parameter estimators are obtained. We then apply these fuzzy clustering methods to the sea level time series.

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

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

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

provide $1+\xi\left(\frac{x_i-\mu}{\sigma}\right) > 0$ 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, 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 [3], a possible solution is to consider the r-largest values over fixed intervals. For m years, the log-likelihood function for the annual r-largest values is given by

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

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 values per year, r, should be chosen carefully since small values of it will produce likelihood estimators 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 [16] 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

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

and

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

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.

While in most applications of extreme value theory, attention is focussed on the extreme quantiles of the GEV distribution, our focus is on fitting GEV distributions to the block maxima of seasonal times series, estimating the parameters and using these parameters as the features for fuzzy clustering .

2.2 Fuzzy Clustering Models

While the traditional non-hierarchial 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 1 and that of an observation not belonging to a cluster is 0.

In the literature, several authors have given different reasons for adopting fuzzy clustering approach (D'Urso [6]). As remarked by Hwang et al. [9], 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, [13]. Secondly, fuzzy clustering has been shown to be less affected by local optima problems (Heiser and Groenen, [8]). 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. [7]).

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 develop 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 they or 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 minimization algorithm; we get suitable weights such that the loss function is minimized 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 formalized as follow:

$$\min : \sum_{i=1}^{I}\sum_{c=1}^{C} u_{ic}^{m}\tilde{d}_{ic}^{2} = \sum_{i=1}^{I}\sum_{c=1}^{C} u_{ic}^{m}\sum_{s=1}^{3}(w_s \cdot d_{ics})^2 = \sum_{i=1}^{I}\sum_{c=1}^{C} u_{ic}^{m}\sum_{s=1}^{3}[w_s \cdot (x_{is} - h_{cs})]^2 \tag{8}$$

subject to the constraints

$$\sum_{c=1}^{C} u_{ic} = 1,\ u_{ic} \geq 0, \tag{9}$$

$$\sum_{s=1}^{3} w_s = 1,\ w_s \geq 0 \tag{10}$$

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; $\tilde{d}_{ic}^{2} = \sum_{s=1}^{3}(w_s \cdot d_{ics})^2 = \sum_{s=1}^{3}[w_s \cdot (x_{is} - h_{cs})]^2 = [w_1 \cdot (x_{i1} - h_{c1})]^2 + [w_s \cdot (x_{i2} - h_{c2})]^2 + [w_3 \cdot (x_{i3} - h_{c3})]^2$ 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 Generalized 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 shape, location and scale parameters of the Generalized 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. 8–10 are:

$$u_{ic} = \frac{1}{\sum\limits_{c'=1}^{C}\left[\frac{\sum\limits_{s=1}^{3}(w_s \cdot d_{ics})^2}{\sum\limits_{s=1}^{3}(w_s \cdot d_{ic's})^2}\right]^{\frac{1}{m-1}}}, \quad w_s = \frac{1}{\sum\limits_{s'=1}^{3}\left[\frac{\sum\limits_{i=1}^{I}\sum\limits_{c=1}^{C}(u_{ic}^{m} \cdot d_{ics}^{2})}{\sum\limits_{i=1}^{I}\sum\limits_{c=1}^{C}(u_{ic}^{m} \cdot d_{ics'}^{2})}\right]}, \quad h_{cs} = \frac{\sum\limits_{i=1}^{I} u_{ic}^{m} x_{is}}{\sum\limits_{i=1}^{I} u_{ic}^{m}}. \tag{11}$$

Proof of Preposition 1
First, fix h_{cs} and w_s, to determine the membership degrees u_{ic}. The solution of Eq. 8–10 is found by means of Lagrange multipliers. Thus, we consider the following Lagrangian function:

$$L_m(\mathbf{u}_i, \lambda) = \sum_{c=1}^{C} u_{ic}^{m}\sum_{s=1}^{3}(w_s \cdot d_{ics})^2 - \lambda\left(\sum_{c=1}^{C} u_{ic} - 1\right) \tag{12}$$

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

$$\frac{\partial L_m(\mathbf{u}_i, \lambda)}{\partial u_{i'c'}} = 0 \Leftrightarrow m \cdot u_{i'c'}^{m-1} \sum_{s=1}^{3} (w_s \cdot d_{i'c's})^2 - \lambda = 0 \tag{13}$$

$$\frac{\partial L_m(\mathbf{u}_i, \lambda)}{\partial \lambda} = 0 \Leftrightarrow \sum_{c=1}^{C} u_{ic} - 1 = 0 \tag{14}$$

From Eq. 13 and by considering Eq. 14 we obtain u_{ic}. Now, fixing u_{ic} and h_{cs}, we calculate the weights w_s. By considering the Lagrangian function:

$$L_m(\mathbf{w}, \xi) = \sum_{i=1}^{I} \sum_{c=1}^{C} u_{ic}^m \sum_{s=1}^{3} (w_s \cdot d_{ics})^2 - \xi \left(\sum_{s=1}^{3} w_s - 1 \right) \tag{15}$$

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

$$\frac{\partial L_m(\mathbf{w}, \xi)}{\partial w_{s'}} = 0 \Leftrightarrow 2 \cdot w_{s'} \sum_{i=1}^{I} \sum_{c=1}^{C} u_{ic}^m \cdot d_{ics'}^2 - \xi = 0 \tag{16}$$

$$\frac{\partial L_m(\mathbf{w}, \xi)}{\partial \xi} = 0 \Leftrightarrow \sum_{s=1}^{3} w_s - 1 = 0 \tag{17}$$

From Eq. 16 we have:

$$w_{s'} = \frac{\xi}{2 \sum_{i=1}^{I} \sum_{c=1}^{C} u_{ic}^m \cdot d_{ics'}^2} \tag{18}$$

and using Eq. 17:

$$\frac{\xi}{2} = \frac{1}{\sum_{s''=1}^{3} \left(\frac{1}{\sum_{i=1}^{I} \sum_{c=1}^{C} u_{ic}^m \cdot d_{ics''}^2} \right)}. \tag{19}$$

Then, replacing ξ in Eq. 18 by ξ from Eq. 19, by we obtain w_s.
For computing h_{cs}, 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.

Remark 1
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.11, 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 optimization 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. 8, we obtain the un-weighted 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

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

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/synthesize 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 formalize the so-called *WGEV–FcMd* model as follows:

$$\min : \sum_{i=1}^{I}\sum_{c=1}^{C} u_{ic}^{m} \tilde{d}_{ic}^{2} = \sum_{i=1}^{I}\sum_{c=1}^{C} u_{ic}^{m} \sum_{s=1}^{3} (w_s \cdot d_{ics})^2 = \sum_{i=1}^{I}\sum_{c=1}^{C} u_{ic}^{m} \sum_{s=1}^{3} [w_s \cdot (x_{is} - \tilde{x}_{cs})]^2 \tag{20}$$

$$\sum_{c=1}^{C} u_{ic} = 1,\ u_{ic} \geq 0, \tag{21}$$

$$\sum_{s=1}^{3} w_s = 1,\ w_s \geq 0 \tag{22}$$

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; $\tilde{d}_{ic}^2 = \sum_{s=1}^{3}[w_s \cdot (x_{is} - \tilde{x}_{cs})]^2 = [w_1 \cdot (x_{i1}\tilde{x}_{c1})]^2 + [w_2 \cdot (x_{i2}\tilde{x}_{c2})]^2 + [w_3 \cdot (x_{i3}\tilde{x}_{c3})]^2$ 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 suitable weights associate to each parameter of the GEV distribution.

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

$$u_{ic} = \frac{1}{\sum_{c'=1}^{C} \left[\frac{\sum_{s=1}^{3} (w_s \cdot d_{ics})^2}{\sum_{s=1}^{3} (w_s \cdot d_{ic's})^2} \right]^{\frac{1}{m-1}}}, \quad w_s = \frac{1}{\sum_{s'=1}^{3} \left[\frac{\sum_{i=1}^{I}\sum_{c=1}^{C} u_{ic}^m \cdot d_{ics}^2}{\sum_{i=1}^{I}\sum_{c=1}^{C} u_{ic}^m \cdot d_{ics'}^2} \right]}. \tag{23}$$

Notice that the objective function in Eq. 20 cannot be minimized by means of the alternating optimization 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 minimizes objective function in Eq. 20 can be built up (refer to Krishnapuram et al. [10]).

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. 20, we obtain the un-weighted 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

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

Remark 2

- 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 [6]). We use the fuzzy silhouette criterion (Campello and Hruschka [4]).
- In our fuzzy clustering models, the fuzziness parameter m plays an important role. The value of m should be chosen in advance. Different heuristic strategies are recommended in the literature. See D'Urso [6] for a detailed discussion.
- The k-means and k-medoids methods are special cases of the fuzzy c-means and fuzzy c-medoids methods, respectively, when m, which controls the degree of fuzziness is set to one.
- Similarly the weighted k-means and weighted k-medoids models are special cases of the weighted fuzzy c-means and weighted fuzzy c-medoids models respectively, when m is set to one.

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. [11] in that we use a dynamic factor model that has been proposed by Safadi and Pena [17]. They used this model to generate air pollution series. The model is of the form:

$$y_t = Lf_t + e_t \tag{24}$$

$$f_t = \sum_{i=1}^{p} \rho_i f_{t-1} + w_t \tag{25}$$

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 $\rho_i = \text{diag}(\rho_{i1}, \rho_{i2}, \cdots, \rho_{ik})$, $i = 1, 2, \cdots, p$ and $\{\rho_{1j}, \rho_{2j}, \cdots, \rho_{pj}$, $j = 1, 2, \cdots, k$ satisfy the stationary conditions and $w_t \sim N(0, I_k)$, where I_k is the identity matrix, and e_t and w_t are independent for all t and s.

In order to introduce seasonality to this dynamic factor model a harmonic component is added to each factor in Eq. 25 as follows:

$$f_{t,k} = \sum_{i=1}^{p} \rho_{i,k} f_{t-1,k} + A_k \sin\left(\frac{2\pi t}{s}\right) + B_k \cos\left(\frac{2\pi t}{s}\right) + w_t. \tag{26}$$

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 fives series each to evaluate the clustering and classification methods when using the GEV parameters.

– Scenario 1: Series with different amplitudes but with 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 simulated five series with amplitude 10, and another five with amplitude 20 . The phase for each of the ten series was set at 0.5 . This is equivalent to a 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.,

$$L = \begin{bmatrix} 1\ 1\ 1\ 1\ 1\ 2\ 2\ 2\ 2\ 2 \end{bmatrix}'. \tag{27}$$

For Scenario 2, we simulated five series with phase 0.5 and five series with phase 1. The amplitude for each of the ten series was set at 10. This is equivalent for 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.,

$$L = \begin{bmatrix} 1\ 1\ 1\ 1\ 1\ 0\ 0\ 0\ 0\ 0 \\ 0\ 0\ 0\ 0\ 0\ 1\ 1\ 1\ 1\ 1 \end{bmatrix}'. \tag{28}$$

Scenario 3 is also a two common factor situation, but in the 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

$$L = \begin{bmatrix} 1\ 1\ 1\ 1\ 1\ 0\ 0\ 0\ 0\ 0 \\ 0\ 0\ 0\ 0\ 0\ 2\ 2\ 2\ 2\ 2 \end{bmatrix}'. \tag{29}$$

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

$$\rho_1 = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix} \tag{30}$$

for both the second and third scenarios. For all three scenarios, the error series were generated from a $N(0, \Gamma)$ process with $\Gamma = I_{10}$, an identity matrix.

Figures 1 to 3 show sections of series of length $T = 366$ generated for each scenario.

Daily-type series for 10 and 20 years were simulated for the three scenarios, and GEV estimates of shape, location and scale were obtained for one and two blocks per year, and were used as the fuzzy clustering features. The performance of the fuzzy clustering methods were evaluated over 100 simulations and in all cases m was set to 2. Note that Bezdek [2] 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. [12] for more details on the cut-off value of 0.7 for non fuzzy classification.

Fig. 1 Time Series with Different Amplitudes

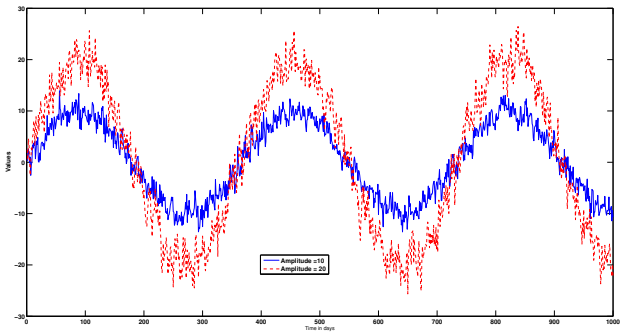


Fig. 2 Time Series with Different Phases

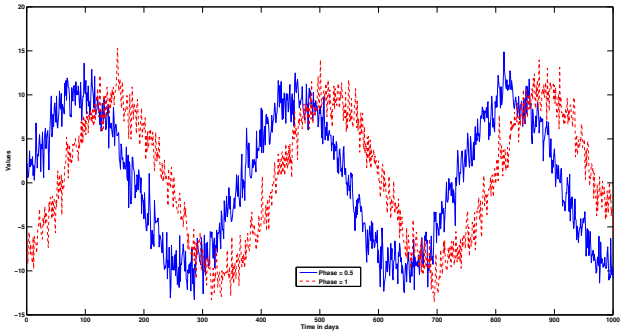
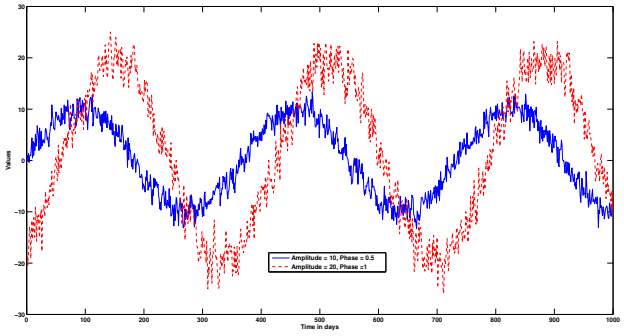


Fig. 3 Time Series with Different Amplitudes and Phases



Tables 1 and 2 show the results of the fuzzy clustering methods for daily-type data over 10 and 20 years and for block sizes one and two, using the estimated GEV parameters. 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.
- 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 amplitude and phases.

Since the 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. The simulations were repeated, this time setting the exponent of fuzziness, m to 1.8 and it was observed that similar results to those above were obtained.

Table 1 Daily Time Series for 10 years: Percentage of Correct, Non-fuzzy and Fuzzy Classification using GEV Parameter Estimates

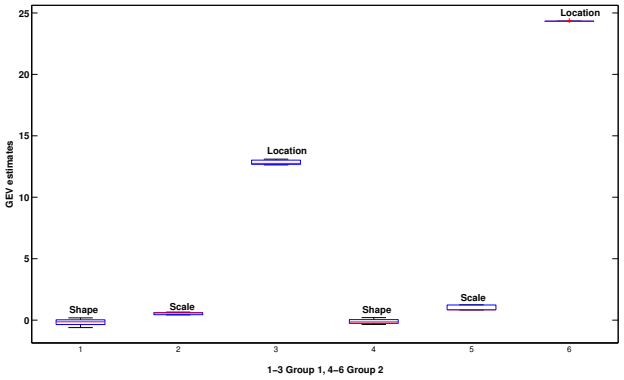
10 years, T=3660	**1 block**			**2 blocks**		
	correct	**non-fuzzy**	**fuzzy**	**correct**	**non-fuzzy**	**fuzzy**
different amplitudes						
fuzzy c-means	0.54	0.10	0.44	0.99	0.42	0.57
weighted fuzzy c-means	0.96	0.96	0.00	0.87	0.81	0.06
fuzzy c-medoids	0.57	0.57	0.00	0.67	0.49	0.18
weighted fuzzy c-mediods	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-mediods	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-mediods	1.00	1.00	0.00	1.00	1.00	0.00

Figure 4 show 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 parameters make a greater contribution to group separation than the the estimated shape and

Table 2 Daily Time Series for 20 years: Percentage of Correct, Non-fuzzy and Fuzzy Classification using GEV Parameter Estimates

20 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-mediods	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-mediods	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-mediods	1.00	1.00	0.00	1.00	1.00	0.00

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



scale parameters. Similar observations were made when series of both different amplitude and phases were generated.

4 Application

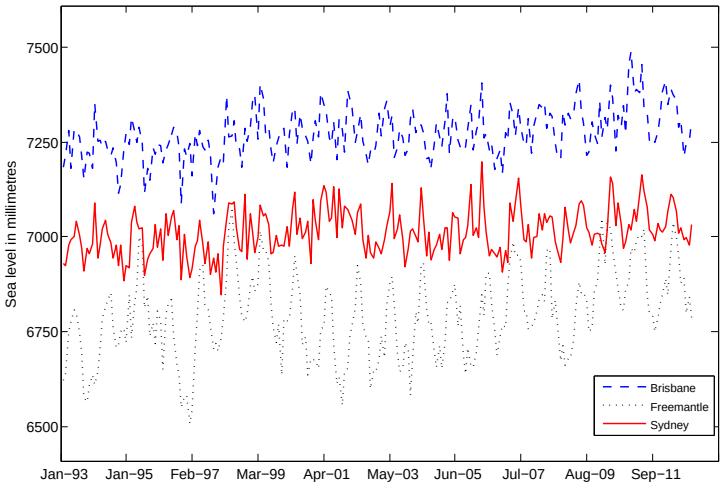
We consider time series of monthly sea levels collected at 39 different tide gauge stations around the coast of Australia. These are available from the web site of Permanent Service for Mean Sea Level (PSMSL) [15]. The PSMSL database contains monthly and annual mean sea level measurements from almost 200 national authorities from around the world, who are responsible for sea level monitoring in each country or region. In order to construct sea level times series, the PSMSL adjusts these measurements with reference to a common depth which is approximately 7000mm below mean sea level. Refer to PSMSL [15] for more details.

These 39 series for the period January 1993 to December 2012 were considered because they contained either no missing values or very few missing values. When there were missing values, each 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 web site 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, and the estimates obtained from fitting the GEV distribution to the annual maxima of each of the series. The graphs of all these series can be viewed at the PSMSL web site.

Table 3 Tide Gauge Sites and GEV Estimates

Coastal Direction	**Tide Gauge Site**	**GEV Estimates**		
		Location	**Scale**	**Shape**
NE	Bowen	7193.84	46.78	-0.19
NE	Brisbane	7357.39	49.14	-0.23
NW	Broome	7129.95	66.11	-0.62
SW	Bunbury	7038.35	62.33	-0.32
NE	Bundaberg	6802.40	41.75	0.19
S	Burnie	7105.04	37.17	-0.50
NE	Cairns	7122.41	45.52	-0.49
NE	Cape Ferguson	7334.36	49.72	-0.23
NW	Carnarvon	7082.80	86.56	-0.20
N	Darwin	7235.29	74.98	-0.36
SE	Eden	7130.83	44.06	-0.19
SW	Esperance	7199.57	57.24	-0.46
SW	Freemantle	6947.01	60.70	-0.34
SW	Geraldton	7163.99	72.73	-0.32
SE	Gold Coast Seaway	7027.16	61.74	-0.36
N	Haypoint	7056.43	50.35	-0.07
SW	Hillarys	7160.54	71.01	-0.43
S	Lorne	7212.14	43.15	-0.35
NE	Mackay	7135.77	52.48	-0.26
SE	Mooloolaba	7112.01	38.03	-0.32
NE	Mourilyan Harbour	7149.35	52.15	-0.33
SE	Newcastle	7134.05	50.58	-0.32
NW	Onslow	7056.77	82.58	-0.30
S	Port Adelaide	7092.52	138.68	-1.07
NE	Port Alma	7206.96	45.19	0.14
SE	Port Kembla	7188.27	42.47	-0.30
S	Port Lincoln	7116.12	54.54	-0.25
S	Port Pirie	7011.48	145.10	-1.00
S	Portland	7173.34	43.94	-0.39
NE	Rosslyn Bay	7245.28	46.52	-0.13
NE	Shute Harbour	7139.77	43.19	-0.17
S	Spring Bay	7228.43	40.46	-0.42
S	Stony Point	7146.90	44.52	-0.44
SE	Sydney	7097.65	42.90	-0.33
S	Thevenard	7168.06	61.06	-0.32
NE	Townsville	7075.55	48.42	-0.19
S	Victor Harbour	7102.70	63.34	-0.24
N	Weipa	7360.44	97.74	-0.33
N	Wyndham	7107.10	110.91	-0.16

Fig. 5 Monthly sea level series

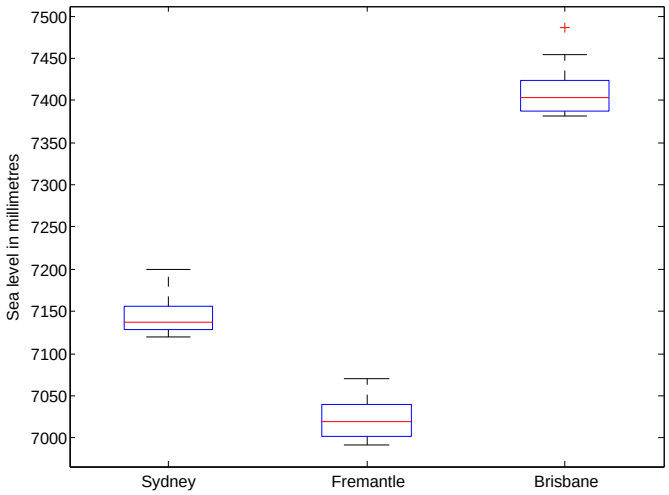


A report by the Australian Government's Department of Climate Change in 2009 [5] provides findings of the first national assessment of the risks of climate change on Australia's coastal areas. In particular, the report discusses the possible impact of rising sea levels on these coastal areas in the coming 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. Identifying areas with similar sea levels in this way 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.

Figure 5 shows three typical sea level series, namely, from tide gauges located near Fremantle, Sydney and Brisbane. Sydney is on the south east coast, Brisbane is on the central to north east coast while Fremantle is on the south west coast. The seasonal patterns in the series are apparent with the Brisbane series displaying higher sea levels, the Fremantle series displaying lower sea levels, with the Sydney series displaying sea levels between that of the other two sites. The Brisbane and Fremantle series each display a gentle slope. It should be note that the GEV distribution should be fitted to maxima that are stationary. In some cases, the series under consideration do display gentle slopes in the series of maxima. However, we proceeded with fitting the GEV distributions to the series of maxima without incorporating the slopes because none were steep enough to warrant incorporating a trend factor in the models. Furthermore, the series of maxima are too short (20 years) to identify whether any trend if it does exist, is linear or nonlinear. Figure 6 shows the values above the 95% percentile of each of these series, where the differences in the sea levels of these three series is clearly apparent.

We first applied the fuzzy c-means method to the GEV estimates to determine the appropriate number of clusters from the fuzzy silhouette coefficients. It was found that a 3 or 4 cluster solution appeared to be appropriate when m was set

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



to 2 or 1.8. Note that $1.5 < m < 2.5$ is an acceptable range for producing fuzzy clusters (Bezdek [2]). Table 4 shows the membership degrees for the 3-cluster fuzzy c-means solution for $m = 2$ and the equivalent hard cluster membership. A time series belongs to a particular hard cluster if its membership degrees is the highest for that cluster.

We can also graphically represent the membership degrees associated with each of the series by mean of a ternary plot which is a barycentric plot on three variables which sum to a constant (usually, this constant is represented as 1.0 or 100%). The plot graphically depicts the ratios of the three variables as positions in an equilateral triangle. In particular, in our case, Figure 7 depicts the ternary plot, a simple representation of the clusters based on membership degrees that summarises the partition structure reported in Table 4. In particular, the ternary plot clearly highlights the behaviour of the time series with fuzzy and non fuzzy membership.

From Table 4, it is clear that the two series in Cluster 2 have scale estimates that are very much larger that that those of all the other series, and shape estimates that are less than or approximately equal to -1. All the other series have shape estimates greater than -1. Hence, it appears that this cluster is based on the similarity of shape and scales estimates. The two series in Cluster 2 have almost crisp membership in this cluster and are represented by overlapping dots at the apex of the triangle in the ternary plot in Figure 7. Standardised location estimates (given in Column 4 of the table) were examined in order to search for patterns within the clusters associated with mean sea levels.

It can be observed from Table 5 which shows the mean, maximum and minimum location values for each of the clusters, the mean location of Cluster 1 is greater than that of other clusters and it has the highest maximum and minimum location values. Furthermore, examination of the standardised location values re-

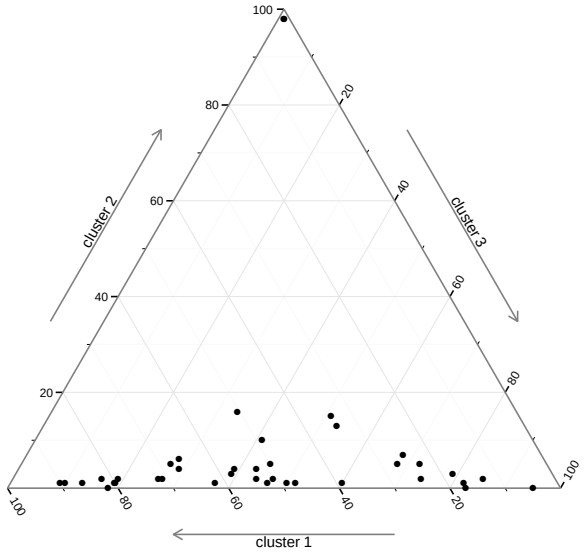
Table 4 3-Cluster Solution

Coastal Direction	Tide Gauge Station	GEV Estimates				Membership Degrees			Hard Cluster
		Location	Std Loc	Shape	Scale				
NE	Bowen	7193.84	0.55	46.78	-0.19	0.80	0.01	0.19	1
NE	Brisbane	7357.39	2.13	49.14	-0.23	0.66	0.06	0.28	1
NW	Broome	7129.95	-0.07	66.11	-0.62	***0.49***	0.10	***0.41***	1
S	Burnie	7105.04	-0.31	37.17	-0.50	***0.53***	0.04	***0.43***	1
NE	Cairns	7122.41	-0.14	45.52	-0.49	***0.58***	0.03	***0.39***	1
NE	Cape Ferguson	7334.36	1.91	49.72	-0.23	0.68	0.05	0.27	1
N	Darwin	7235.29	0.95	74.98	-0.36	0.67	0.04	0.29	1
SW	Esperance	7199.57	0.61	57.24	-0.46	0.79	0.02	0.19	1
SW	Geraldton	7163.99	0.26	72.73	-0.32	***0.54***	0.02	***0.44***	1
SW	Hillarys	7160.54	0.23	71.01	-0.43	***0.57***	0.04	***0.39***	1
S	Lorne	7212.14	0.73	43.15	-0.35	0.90	0.01	0.10	1
NE	Mackay	7135.77	-0.01	52.48	-0.26	***0.49***	0.01	***0.50***	1
SE	Mooloolaba	7112.01	-0.24	38.03	-0.32	***0.51***	0.02	***0.47***	1
NE	Mourilyan Harbour	7149.35	0.12	52.15	-0.33	0.81	0.00	0.18	1
SE	Newcastle	7134.05	-0.03	50.58	-0.32	0.62	0.01	0.37	1
NE	Port Alma	7206.96	0.68	45.19	0.14	***0.50***	0.05	***0.45***	1
SE	Port Kembla	7188.27	0.50	42.47	-0.30	0.90	0.01	0.09	1
S	Portland	7173.34	0.35	43.94	-0.39	0.86	0.01	0.13	1
NE	Rosslyn Bay	7245.28	1.05	46.52	-0.13	0.71	0.02	0.26	1
NE	Shute Harbour	7139.77	0.03	43.19	-0.17	***0.52***	0.01	***0.46***	1
S	Spring Bay	7228.43	0.88	40.46	-0.42	0.82	0.02	0.16	1
S	Stony Point	7146.90	0.09	44.52	-0.44	0.71	0.02	0.27	1
S	Thevenard	7168.06	0.30	61.06	-0.32	0.81	0.01	0.19	1
N	Weipa	7360.44	2.16	97.74	-0.33	***0.51***	0.16	***0.34***	1
S	Port Adelaide	7092.52	-0.43	138.68	-1.07	0.01	0.98	0.01	2
S	Port Pirie	7011.48	-1.22	145.10	-1.00	0.01	0.98	0.01	2
SW	Bunbury	7038.35	-0.96	62.33	-0.32	0.13	0.02	0.85	3
NE	Bundaberg	6802.40	-3.24	41.75	0.19	***0.34***	0.13	***0.53***	3
NW	Carnarvon	7082.80	-0.53	86.56	-0.20	0.27	0.05	0.68	3
SE	Eden	7130.83	-0.06	44.06	-0.19	***0.47***	0.01	***0.51***	3
SW	Freemantle	6947.01	-1.84	60.70	-0.34	0.25	0.07	0.68	3
SE	Gold Coast Seaway	7027.16	-1.07	61.74	-0.36	0.18	0.03	0.79	3
N	Haypoint	7056.43	-0.78	50.35	-0.07	0.24	0.02	0.73	3
NW	Onslow	7056.77	-0.78	82.58	-0.30	0.23	0.05	0.72	3
S	Port Lincoln	7116.12	-0.20	54.54	-0.25	0.17	0.00	0.82	3
SE	Sydney	7097.65	-0.38	42.90	-0.33	0.39	0.01	0.60	3
NE	Townsville	7075.55	-0.60	48.42	-0.19	0.17	0.01	0.82	3
S	Victor Harbour	7102.70	-0.33	63.34	-0.24	0.05	0.00	0.95	3
N	Wyndham	7107.10	-0.29	110.91	-0.16	***0.34***	0.15	***0.51***	3

veal that most are greater than zero. The series in this cluster could be generally associated with higher sea levels.

For Cluster 3, the minimum location value is lower than than of the other clusters and the mean location in very similar to that of Cluster 2. Furthermore the standardised location values in Cluster 3 are all negative, leading to the observation that the series in this cluster could generally be associated with lower sea levels. While the series in Cluster 2 have almost crisp membership degrees (0.98 in each case), the standardised location estimates are also negative and smaller than any of the negative standardised location estimates in Cluster 1, indicating that these series could be associated with lower sea levels as well. However, many series in Clusters 1 and 3 have substantial fuzzy membership degrees (identified in bold italics in the membership degrees column in Table 4) in both clusters

Fig. 7 Ternary Plot Depicting Cluster Membership



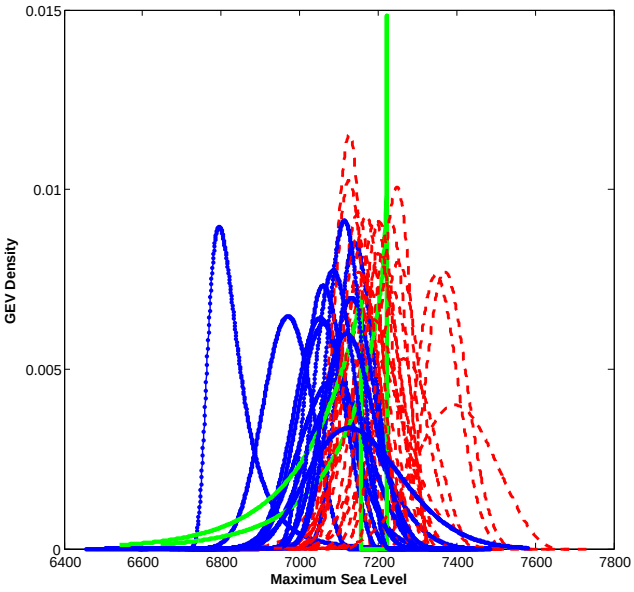
indicating that over some periods of time, they have higher sea levels (positive standardised location estimates) while over other periods of time, they have lower sea levels (negative standardised location estimates) but it is not always clear which of the GEV estimates or which combination of estimates contribute most to this fuzziness. However, internally optimized weights from the weighted fuzzy c-means cluster solution, for the location, scale and shape estimates were 0.097, 0.755 and 0.148, respectively, indicating that the scale estimate compared to the location and the shape estimates contribute most to cluster separation. Those series in Clusters 1 and 3 with substantial fuzziness are depicted away from the corners at the base of the triangle in the ternary plot.

Table 5 Means, Maxima and Minima of the 3-Cluster Solution

	Mean	**Maximum**	**Minimum**
Cluster 1	7191.80	7360.44	7105.04
Cluster 2	7052.00	7092.52	7011.48
Cluster 3	7049.30	7130.83	6802.40

Figure 8 shows the GEV density function for each of the series of annual maxima under consideration. The broken-line red density curves towards the right-hand side of the x-axis are those for the series of annual maxima from Cluster 1 (mostly higher sea levels), while the light green density curves near the centre are those for the series annual maxima from Cluster 2, and the solid blue density curves towards the left-hand side are those for the series of annual maxima from

Fig. 8 GEV Density Functions: 3 Clusters



Cluster 3 (lower sea levels). Both the separation of Clusters 1 and 3 as well as fuzzy nature of some the series in these clusters are apparent from the positioning of the density curves on the x-axis. It is also clear that the grouping of the two series in Cluster 2 is strongly influenced by the shape estimates but in terms of location they are closer to the series in Cluster 3 which are associated with the lower sea levels.

4.1 Validation of the Cluster Solution

As a means of validating this fuzzy cluster solution, we applied the k-means method with 3 clusters to the GEV estimates and found the crisp clusters had a 97% agreement with the hard clusters identified from the fuzzy c-means analysis. We also applied the other fuzzy clustering methods to the GEV estimates and obtained relatively good coherence between the cluster solutions. In particular, there was a 85% agreement between the fuzzy c-means and weighted fuzzy c-means methods, and a 97% agreement between the fuzzy c-means method and each of the fuzzy c-medoids and weighted fuzzy c-medoids methods. For all methods the two series with the very large scale estimates and with the shape estimates that were less than or equal to -1 grouped together in one cluster. There were very few discrepancies amongst these fuzzy methods in cluster membership of the other two clusters as indicated by the percentages of agreement given above. This compatibility in cluster solutions provides further validation of the obtained fuzzy c-means 3-cluster solution.

4.2 A 4-Cluster Solution

The fuzzy silhouette coefficients based on the fuzzy c-means method also indicated that a 4-cluster solution might be an option. We applied the fuzzy c-means method to the GEV estimates and obtained the cluster solution shown in Table 6, with the clusters means, maxima and minima given in Table 7.

Table 6 4-Cluster Solution

Coastal	Tide Gauge	GEV Estimates				Membership Degrees				Hard
Direction	Station	Location	Std Loc	Shape	Scale					Cluster
SW	Bunbury	7038.35	-0.96	62.33	-0.32	0.89	0.08	0.01	0.02	1
NE	Bundaberg	6802.40	-3.24	41.75	0.19	0.41	0.29	0.09	0.21	1
NW	Carnarvon	7082.80	-0.53	86.56	-0.20	0.64	0.20	0.03	0.13	1
SW	Freemantle	6947.01	-1.84	60.70	-0.34	0.64	0.22	0.04	0.10	1
SE	Gold Coast Seaway	7027.16	-1.07	61.74	-0.36	0.79	0.15	0.01	0.05	1
N	Haypoint	7056.43	-0.78	50.35	-0.07	0.52	0.31	0.02	0.16	1
NW	Onslow	7056.77	-0.78	82.58	-0.30	0.75	0.15	0.02	0.08	1
NE	Townsville	7075.55	-0.60	48.42	-0.19	***0.48***	***0.40***	0.01	0.11	1
S	Victor Harbour	7102.70	-0.33	63.34	-0.24	0.66	0.26	0.01	0.07	1
N	Wyndham	7107.10	-0.29	110.91	-0.16	0.43	0.25	0.10	0.23	1
NW	Broome	7129.95	-0.07	66.11	-0.62	***0.30***	***0.41***	0.07	0.21	2
S	Burnie	7105.04	-0.31	37.17	-0.50	0.23	0.58	0.03	0.17	2
NE	Cairns	7122.41	-0.14	45.52	-0.49	0.19	0.64	0.02	0.15	2
SE	Eden	7130.83	-0.06	44.06	-0.19	0.19	0.62	0.01	0.19	2
SW	Esperance	7199.57	0.61	57.24	-0.46	0.14	0.49	0.02	0.36	2
SW	Geraldton	7163.99	0.26	72.73	-0.32	0.29	0.41	0.02	0.28	2
SW	Hillarys	7160.54	0.23	71.01	-0.43	0.27	0.45	0.03	0.25	2
S	Lorne	7212.14	0.73	43.15	-0.35	0.07	***0.41***	0.01	***0.51***	2
NE	Mackay	7135.77	-0.01	52.48	-0.26	0.09	0.83	0.00	0.07	2
SE	Mooloolaba	7112.01	-0.24	38.03	-0.32	0.15	0.73	0.01	0.11	2
NE	Mourilyan Harbour	7149.35	0.12	52.15	-0.33	0.01	0.97	0.00	0.01	2
SE	Newcastle	7134.05	-0.03	50.58	-0.32	0.00	0.99	0.00	0.00	2
SE	Port Kembla	7188.27	0.50	42.47	-0.30	0.07	0.56	0.01	0.36	2
S	Port Lincoln	7116.12	-0.20	54.54	-0.25	0.23	0.68	0.01	0.08	2
S	Portland	7173.34	0.35	43.94	-0.39	0.08	0.73	0.01	0.19	2
NE	Shute Harbour	7139.77	0.03	43.19	-0.17	0.18	0.58	0.01	0.23	2
S	Stony Point	7146.90	0.09	44.52	-0.44	0.11	0.74	0.01	0.14	2
SE	Sydney	7097.65	-0.38	42.90	-0.33	0.18	0.73	0.01	0.08	2
S	Thevenard	7168.06	0.30	61.06	-0.32	0.13	0.62	0.01	0.24	2
S	Port Adelaide	7092.52	-0.43	138.68	-1.07	0.01	0.01	0.97	0.01	3
S	Port Pirie	7011.48	-1.22	145.10	-1.00	0.01	0.01	0.98	0.01	3
S	Spring Bay	7228.43	0.88	40.46	-0.42	0.10	0.39	0.02	0.49	4
NE	Bowen	7193.84	0.55	46.78	-0.19	0.08	0.33	0.01	0.59	4
NE	Brisbane	7357.39	2.13	49.14	-0.23	0.10	0.20	0.03	0.68	4
NE	Cape Ferguson	7334.36	1.91	49.72	-0.23	0.08	0.17	0.02	0.73	4
N	Darwin	7235.29	0.95	74.98	-0.36	0.16	0.28	0.03	0.54	4
NE	Port Alma	7206.96	0.68	45.19	0.14	0.23	0.30	0.03	0.44	4
NE	Rosslyn Bay	7245.28	1.05	46.52	-0.13	0.06	0.15	0.01	0.78	4
N	Weipa	7360.44	2.16	97.74	-0.33	0.20	0.25	0.10	0.45	4

We observe from Table 7 that Cluster 3 consists of the two series with the dominant scale and shapes values that also grouped together in the 3-cluster solution. As before, from the standardised locations of the series in this cluster it would appear they are associated with lower sea levels.

Table 7 Means, Maxima and Minima of the 4-Cluster Solution

	Mean	**Maximum**	**Minimum**
Cluster 1	7029.63	7107.10	6802.40
Cluster 2	7143.91	7212.14	7097.65
Cluster 3	7052.00	7092.52	7011.48
Cluster 4	7270.25	7360.44	7193.84

The standardised locations of the series in Cluster 1 are all negative with 8 of the 10 being less than -0.55 implying that the series in this cluster could generally be associated with lower sea levels. Some of these series have reasonable fuzzy membership degrees in Cluster 2 which consists of series with both positive standardised locations no more than 0.61 and negative standardised locations no less than -0.38. The standardised locations of the series in Cluster 4 are all positive being no less than 0.55 and these these series could generally be associated with higher sea levels.

In Cluster 2, some of the series with negative standardised locations have reasonable fuzzy membership in Cluster 1, while some of the series with positive standardised locations have reasonable fuzzy membership in Cluster 4. The series in Cluster 2 could mostly be associated with sea levels that are neither very high nor very lower.

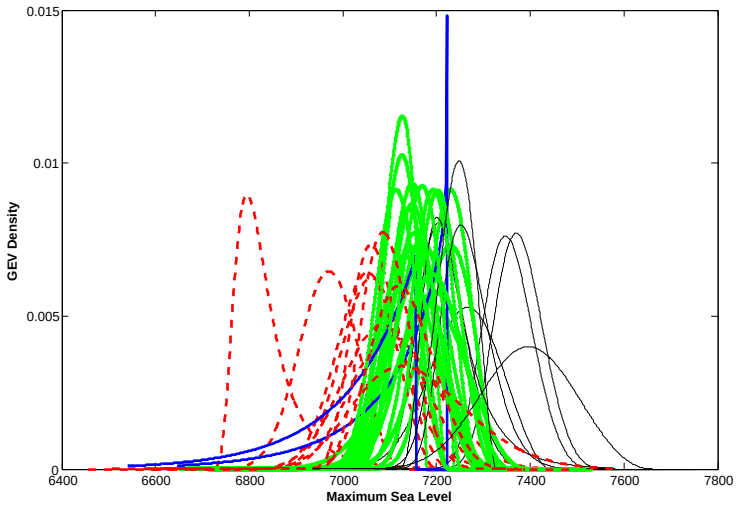
Hence, it appears that the 4-cluster solution gives a somewhat clearer separation of series with high and low sea levels than the 3-cluster solution does. This is also apparent from the cluster means, maxima and minima in Table 7. This separation is also apparent in Figure 8 which depicts the GEV density curves associated with the four clusters with different colours and patterns. Namely, GEV fits to series associated with low sea levels are depicted by broken red curves while those associated with between low and high sea levels are depicted by solid light green curves; those associated with high sea levels are depicted by thin solid black curves, while GEV fits to series dominated by the shape estimates are depicted as thick solid blue curves. As well, the fuzzy membership of some the series across the clusters is also apparent.

We also obtained the crisp 4-cluster solution using the k-means method and found there was a 95% agreement with the equivalent hard clusters from this fuzzy c-means 4-cluster solution.

4.3 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. 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 4-cluster solution obtained from the fuzzy c-means method. The results are presented in Table 8 and they confirm and complement our previous interpretation, that is, (1) the first cluster corresponds to localities having low sea levels that could be no more than 7296 millimetres in periods of 100 years, (2) the second cluster corresponds to localities having not very high or not very low sea levels, (3) the third cluster corresponds to localities having low

Fig. 9 GEV Density Functions: 4 Clusters



sea levels that could be no more than 7188 millimetres in periods of 100 years and (4) the fourth cluster corresponds to sites having the high sea levels greater than 7440 millimetres in periods of 25 or more years. Compared to the maximum and mean locations of the clusters in Table 8, these returns appear to be realistic.

In particular, the difference between the 25 year returns and the corresponding location means based on the GEV fit to the 1993 to 2012 series of annual maxima, which correspond to projected sea level rises for series in 2037 in Clusters 1, 2, 3 and 4 are 0.209, 0.138, 0.131 and 0.169 metres, respectively. These projections fall mostly 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 (CSIRO) for 2030, relative to 1990 (refer to Table 2.1 in [5]).

Table 8 Cluster Location Means and Maxima and 25, 50 and 100 Year Mean Returns Levels

	Location		**Returns**		
	max	**mean**	**25 years**	**50 years**	**100 years**
Cluster 1	7107	7030	7238	7267	7296
Cluster 2	7212	7147	7284	7294	7303
Cluster 3	7093	7052	7183	7186	7188
Cluster 4	7360	7270	7440	7461	7482

Remark 3

Hourly and daily sea level time series are available from other web sites. However, there are limitations with using such data mainly because most of these time series contain missing values for several months in each of the years that measurements were taken. Hence, the reason we used monthly rather than daily sea level time series in our application.

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 obtained. 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, viz., sea-level time series, all three GEV estimates are capable of contributing to cluster separation. From 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, which in this case of the application to groups of similar sea-level time series, may 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 analyzed with the proposed procedure.

The future directions that we will be embarking on in analysing real time series extremes is, (1) extending GEV fitting to extremes with trend, and (2) examining the fuzzy behaviour of the series by incorporating their spatial features as an added source of information.

References

1. Alonso, A.M., De Zea Bermudez, P., Scotto, M.G., Comparing generalized Pareto models fitted to extreme observations: an application to the largest temperatures in Spain, Stoch Environ Res Risk Assess., 28, 1221–1233 (2014).
2. Bezdek J.C., Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, NewYork (2001).
3. Coles, S., An introduction to statistical modeling of extreme values Springer-Verlag: London (2001).
4. Campello R.J.G.B. and Hruschka E.R., A fuzzy extension of the silhouette width criterion for cluster analysis, Fuzzy Sets and Systems,157, 2858 –2875 (2006)
5. Department of Climate Change, Commonwealth of Australia, www.climatechange.gov.au , ISBN: 978-1-921298-71-4 (2009).
6. D'Urso, P., Fuzzy clustering, in Handbook of Cluster Analysis, Chapman and Hall, C. Hennig, M. Meila, F. Murtagh, R. Rocci (eds.), in press (2015).
7. Everitt, B.S., Landau, S. and Leese, M., Cluster analysis (4th ed.) Arnold Press: London (2001).
8. Heiser, W.J., Groenen, P.J.F., Cluster differences scaling with a within-clusters loss component and a fuzzy successive approximation strategy to avoid local minima, Psychometrika, 62, 63 – 83 (1997).
9. Hwang, H., De Sarbo, W.S. and Takane, Y., Fuzzy Clusterwise Generalized Structured Component Analysis. Psychometrika , 72, 181–198 (2007).
10. Krishnapuram,R., Joshi, A., Nasraoui, O., Yi, L., Low-complexity fuzzy relational clustering algorithms for web mining, IEEE Transactions on Fuzzy Systems, 9 4, 595–607 (2001).
11. Maharaj, E.A., Alonso A.M. and D'Urso, P., Clustering Seasonal Time Series Using Extreme Value Analysis: An Application to Spansih Temperature Time Series Preprint submitted to Communications in Statistics: Case Studies and Data Analysis (2015).
12. Maharaj, E.A., D'Urso, P., Galagedera, D. U. A., Wavelets-based fuzzy clustering of time series, Journal of Classification,27 2, 231–275 (2010).

13. McBratney, A.B. and Moore, A.W., Application of fuzzy sets to climatic classification, Agricultural and Forest Meteorology, 35, 165 –185 (1985).
14. Méndez, F. J., Menéndez, M., Luceño, A. and Losada, I. J, Analysing monthly extreme sea levels with a time-dependent GEV model, J. Atmos. Ocean. Technol., 24, 894 – 911 (2007).
15. Permanent Service for Mean Sea Level (PSMSL) "Tide Gauge Data", Retrieved 24 Nov 2014 from http://www.psmsl.org/data/obtaining/. Simon J. Holgate, Andrew Matthews, Philip L. Woodworth, Lesley J. Rickards, Mark E. Tamisiea, Elizabeth Bradshaw, Peter R. Foden, Kathleen M. Gordon, Svetlana Jevrejeva, and Jeff Pugh (2013) New Data Systems and Products at the Permanent Service for Mean Sea Level, Journal of Coastal Research, 29, 3, 93 – 504. doi:10.2112/JCOASTRES-D-12-00175.1 (2014).
16. Reiss, R-D. and Thomas, M., Statistical Analysis of Extreme Values with Applications to Insurance, Finance, Hydrology and Other Fields 3rd Edition Birkhauser, Basel, Boston, Berlin (2000).
17. Safadi, T. and Peña, D., Bayesian analysis of dynamic factor models: an application to air pollution and mortality in Sao Paulo, Brazil, Environmetrics, 19, 582–601 (2008).
18. Scotto, M.G., Barbosa, S.M. and Alonso, A.M., Clustering Time Series of Sea Levels: Extreme Value Approach, J. Waterway, Port, Coastal, Ocean Eng., 136, 2793–2804 (2010).
19. Scotto, M.G., Barbosa, S.M. and Alonso, A.M., Extreme value and cluster analysis of European daily temperature series,Journal of Applied Statistics, 38 12, 215–225 (2011).
20. Tsimplis, M. N., and Blackman, D. L., Extreme sea-level distribution and return periods in the Aegean and Ionian seas, Estuarine Coastal Shelf Sci, 44, 79–89 (1997).
21. Unnikrishnan, A. S., Sundar, D., and Blackman, D. L., Analysis of extreme sea level along the east coast of India, J. Geophys. Res., 190, C06023 (2004).