Objective regional frequency analysis of extreme precipitation in Sicily, Italy

A. Forestieri*[1], F. Lo Conti[1], S. Blenkinsop[2], L.V. Noto[1], H. Fowler[2]

1 Università degli Studi di Palermo, Dipartimento di Ingegneria Civile, Ambientale, Aerospaziale, e dei Materiali, Palermo, Italy
2 Newcastle University, School of Civil Engineering and Geosciences, Newcastle, UK

*Corresponding author: angelo.forestieri@unipa.it

Abstract

The extreme events have large impacts on society and are likely to increase under climate change. For design and management decisions, particularly around hydraulic infrastructures, accurate estimates of precipitation magnitudes are needed at different durations. In this paper, the regional frequency analysis has been implemented and applied to precipitation data recorded in Sicily, Italy. Annual maximum series for rainfall durations of 1, 3, 6, 12 and 24 h provided by about 130 rain gauges were used. The Regional Frequency Analysis (RFA) has been used to identify the homogeneous regions using Principal Component Analysis (PCA) followed by a clustering analysis, through *k-means*, aimed to identify regional groups. Two regional probability distributions have been used in order to derive the Depth-Duration Frequency (DDF) curves: lognormal distribution with three parameters (GNO) and generalized extreme value distribution (GEV). The regional parameters of these distributions were estimated using the L-moment ratios approach while the relative bias and relative RMSE have been calculated using a simulation study of regional L-moment algorithm for the assessment of the accuracy.

1. Introduction

The extreme precipitations show intensification in many regions and this is of key importance to society as a result of the large impact through flooding (Trenberth et al., 2003). For design and management decisions, particularly around hydraulic infrastructures, accurate estimates of precipitation at different durations are needed. Frequently the historical available rainfall series are unsuitable for this estimation process because of the gaps present in the period of registration, and then the regional frequency analysis is necessary. *Regional rainfall frequency analysis* (RFA) plays an important role for several civil infrastructure design and non-structural problems involving natural hazards associated with extreme rainfall events.

In previous works, it has been demonstrated that RFA provides more reliable estimates of return periods for extreme rainfall even in the case when the dataset is not very large (Hosking and Wallis, 1988). RFA is also able to resolve the problem of the evaluation of precipitation extremes at ungauged sites within the same region.

The aim of this work consists in the design and the application of the RFA procedure for the area of Sicily, Italy, based on the selection of suitable procedures which take into account the data availability and the meteoclimatic features of the area. In previous works, related to the same area (Cannarozzo, et al. 1995; Lo Conti et al., 2007), the choice of the number and the extension of the homogeneous regions were made using hydrological criteria related principally on watersheds boundaries. In this work, we adopted an objective method to obtain the homogeneous regions that allow to consider several variables, linked to the extreme precipitation, in a robust mathematical framework.

2. Methods

Our analysis has been performed following the approach similar to that used by Jones et al. (2013). The first step of the methodology is the selection of the variables to include in the region identification. These variables directly influence the regional frequency analysis performances in terms of homogeneity. Variables adopted for this study are reported in Table 1.

Table 1: Variables used in rainfall region development.

Variable	Description
Z	Station elevation
AMR_d	Annual maximum rainfall for different durations (mm)
θ_d	Mean date of the events, represents a measure of the average time of occurrence of rainfall events.
$r_{m,d}$	Seasonality vector, provides a dimensionless measure of the spread of the data.
nDry	Number of days < 1mm
Rs/Rw	Ratio between summer rainfall (April - September) and winter rainfall (October - May)

The *Principal Component Analysis* (PCA), whose primary purpose is to reduce the number of variables obtaining some “latent” variables which result uncorrelated (Wilks, 2011), was performed on the initial dataset. It is particularly useful when one needs a data reduction procedure that makes no assumptions concerning an underlying causal structure that is responsible for co-variation in the data. The principal components can be defined as a linear combination of optimally-weighted observed variables. Usually, only the first few components obtained are retained and interpreted. The PCA analysis has been applied to the selection of variables estimated for each station.

The successive cluster analysis was performed using the reduced number of significant synthetic variables retrieved from the PCA to which the normalized latitude and longitude of stations have been added to support the grouping of continuous regions. Operationally, the cluster analysis was performed using the *k-means* clustering method (Hartigan and Wong, 1979). This method operates placing centroids of the initial clusters and reallocating group memberships on the basis of proximity to the cluster centroids. The algorithm is iterated until each data vector is closest to its group centroid and no further reallocations of memberships are made.

The main difficulty related to the *k-means* method is that the number of clusters, *k*, must be predefined. There are different ways to define the number of clusters if they are not known a priori. Nevertheless, these methods may provide different results. Therefore, It is useful to try *k-means* with an initial range of values of clusters.

Once a set of physically plausible regions has been defined, it is necessary to assess their degree of homogeneity. Following the RFA approach proposed by Hosking and Wallis (2005), in this work three different tests were chosen: the discordancy measure *D* for each station, to examine possibly anomalous behaviour of individual stations, the *HW* (Hosking and Wallis, 1993) and the Anderson-Darling *AD* rank tests, (Stedinger et al., 1993) to assess homogeneity of the extreme rainfall regions obtained. Viglione et al. (2007) have compared the *HW* test and the *AD* test highlighting that, when the *L-skewness* coefficient for the region under analysis is lower than 0.23, the Hosking and Wallis heterogeneity measure HW_1 can be considered the preferred method; if *L-skewness* is greater than 0.23 the bootstrap Anderson-Darling test is preferable. In case of heterogeneity of a cluster, the possibility of combining it with another into a single region or that to subdivide it in two different clusters are evaluated.

The goodness of fit is evaluated for each homogenous region considering five parameter frequency distributions: generalized logistic (GLO), generalized extreme-value (GEV), lognormal (GNO), Pearson type III (PE3) and Generalized Pareto Distribution (GPA). After the choice of the frequency distributions, these were fitted to data from relative to sites in each homogeneous region. The distributions were fitted using the method of the L-moments (Hosking and Wallis, 2005) based on the dimensionless rescaled data *x'*. The quantiles relative to several return periods *T* were obtained setting *F(x')=1-1/T*. Finally, in order to evaluate the uncertainty and the reliability of the results obtained by statistical analysis, an assessment of the magnitude of uncertainty was carried out with a Monte Carlo simulation analysis.

3. Data

Sicily is the largest island in the Mediterranean Sea situated in the South of Italy with an area about 25.000 km^2.In order to apply the regional frequency analysis, this study used the extreme rainfall data published by the *Servizio Osservatorio delle Acque*. The rainfall data used in this research are the annual maxima rainfalls with durations equal to 1, 3, 6, 12 and 24 h and the daily time series. There is a total of 314 stations for sub-daily annual maxima data used for the analysis spanning the period 1928-2009 while for daily precipitation there are 382 stations spanning the period 1928-2009. However, some stations have many gaps due to non-operative periods and, for this reason, only those with at least 30 years of operation between the years 1972 and 2003 were selected, this being the period with maximum station operation. Thus, the observations for different durations comprise 124 stations, each with a minimum record length of 20 years.

4. Results

The Principal Components (PCs) evaluated separately for the different durations, were combined through an averaging operation because generally they showed similar score values among different durations. The residual percent variance method was used to select the principal components obtained from the analysis. It has been decided to retain the principal components which accounted for at least 5 % of the total variance in the input dataset then only the five principal components were selected.
Values of the score, for the first three principal components, highlight the responses of geography for the first component, while the second has shown a relation with the precipitation mean annual maxima and the greater value of the dry day; the third component is mainly related to the seasonality, with a particular influence in the east side where the events are short and intense. The five PCs obtained were used as input for the cluster analysis performed through the *k-means*.

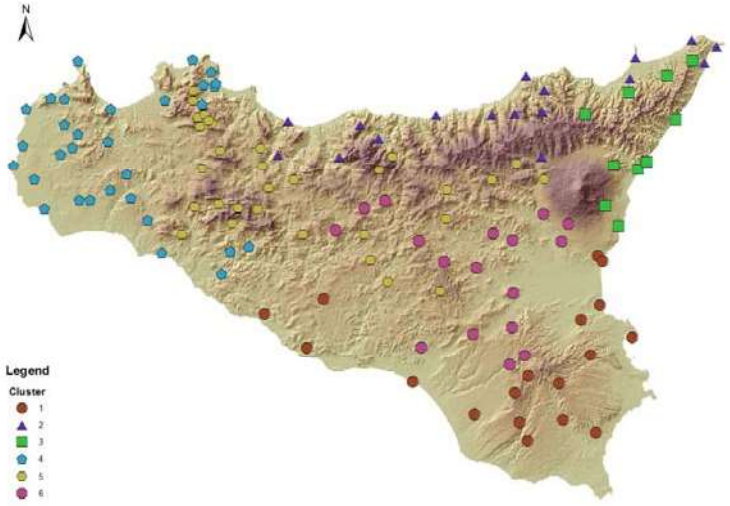

Fig 1: Homogeneous regions obtained with the K-means.

As highlighted above, the cluster analysis was performed using the variables obtained from the PCA and, additionally, the latitude and longitude normalized of each station to support the grouping of continuous regions. In this study, the possible range of values of *k* (i.e. number of clusters) was selected considering the minimum and maximum number of clusters used in previous work regarding the Sicily, i.e., from 3 to 8 (Cannarozzo et al., 1995; Lo Conti et al., 2007; Gabriele and Chiaravalloti 2013). Therefore, this range was hence examined evaluating the results obtained through the method of the Silhouette value (Rousseeuw, 1987) which provided the most robust and optimal solution that relative to *k* equal to 6.

At this point, the RFA tests described above are applied for the regions obtained. An analysis of the dependence of the L-moments values was achieved to confirm whether it was needed to consider a relation with the duration. The values of the regionally weighted L-moments were calculated for the different duration. The values of L-moments show clearly a behaviour of the L-moments of extreme rainfall for the duration of 1 hour different from that relative to the other durations (i.e. the latter results less variable). This behaviour could be due to the errors in the recorded data. Indeed, the uncertainty for the short duration is more evident because of the nature of the short extreme precipitations that could be due to particular meteorological conditions with different temporal and spatial extension compared with events of long duration. The distribution parameters are evaluated considering the mean value for each L-moment not including the values for the duration of 1h. When the test of homogeneity does not report a positive result for a region, the characteristics of sites that showed marked differences have been carefully examined, then these sites have been reassigned to other regions or deleted. In other cases, some clusters presented stations with discordant values although resulted homogeneous; in this case, stations were not deleted and they were used during the subsequent analysis.

On the basis of the homogeneous regions identified (Figure 1), the RFA was used to fit GNO and GEV distributions estimating the *x'* quantiles (Figure 2) with the L-moments method.

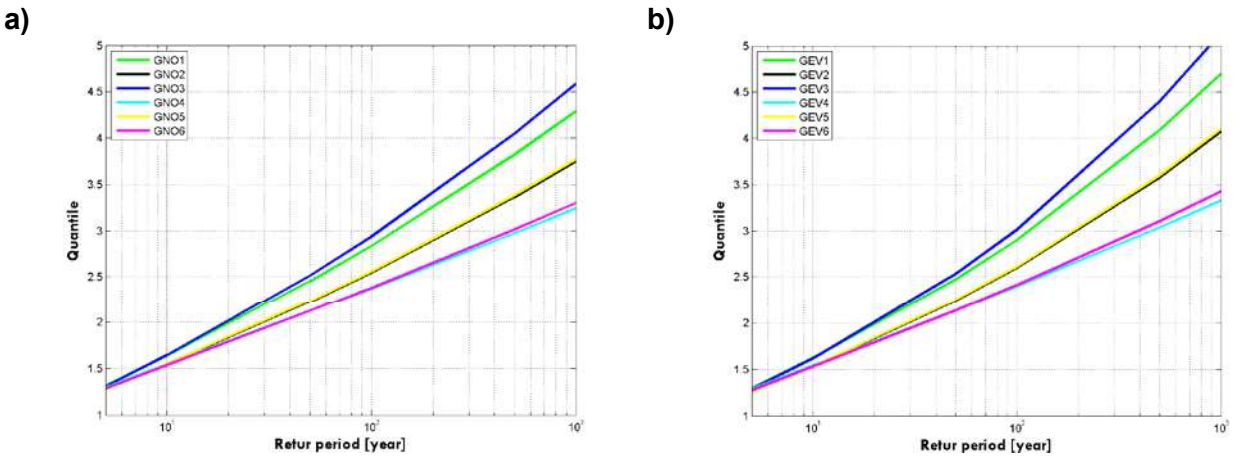

Fig 2: Growth curves for the different distributions, a) GNO and b) GEV.

Finally, an assessment of the accuracy has been performed evaluating the BIAS and RMSE of quantiles estimated with fitted distributions with reference to those obtained from a Monte Carlo procedure reproducing the variability of data. Values of the relative BIAS resulted near zero or slightly negative for lower return periods. In particular, the GNO distribution reported a very low value. The RMSE values for the GNO and GEV distributions are very similar each other with an exception regarding the cluster 3, in particular in the extreme upper tail where $F \geq 0.99$ ($T \geq 100$ years). The cluster 3 has shown, for high return periods, a significant increase of the values of BIAS and RMSE, probably due to the small number of stations inside the region. The cluster 6, despite being identified as "possibly heterogeneous", does not show RMSE values greater than the other clusters.

Since the values of the L-moment for the duration of 1h were not considered for the evaluation of the parameters of the distributions, the relative BIAS and RMSE for the GEV and GNO for the duration of 1h was evaluated comparing with the regional value achieved considering the real heterogeneity through the L-CV values for 1h and the regional growth curve previously obtained.

A comparison of BIAS and RMSE for the regional curve and for 1h duration are negligible, the greater difference is in the cluster 6 for both distributions. Finally, the regional growth curve obtained for durations lower than 3h shows acceptable performances.

5. Conclusions

This article presents a new set of regions for the RFA of precipitation in Sicily, obtained considering the matching between the at site characteristics with those relative to the extreme rainfall such as magnitude, the timing of events, seasonality and distribution between summer and winter events. The methodology adopted allows for a better and more robust regional frequency analysis, compared with traditional approaches. In previous works, the number of regions obtained with hydrological characteristic varied between three and eight. Through this new approach, homogeneous regions have been obtained with a cluster analysis, using new variables achieved by PCA. The best number of regions is equal to six, then similar to the number provided by previous works but with important spatial differences.

Quantiles have been obtained fitting two extreme different distributions, the GEV and the GNO. The precipitations for 1h duration showed a different behaviour due to particular kind of precipitation, usually convective precipitation with limited spatial extension. The GNO distribution has given result better than GEV, although both distributions have three parameters.

References

Cannarozzo, M, D'Asaro, F., and Ferro, V. (1995), "Regional Rainfall and Flood Frequency Analysis for Sicily Using the Two Component Extreme Value Distribution." *Hydrological sciences journal* 40(1): 19–42.

Lo Conti, F., Noto, L., La Loggia. G., and Cannarozzo. M. (2007), "Regional Frequency Analysis of Extreme Precipitation in Sicily, Italy." *Variability in space and time of extreme rainfalls, floods and droughts*.

Gabriele, S. and Chiaravalloti, F. (2013), "Using the Meteorological Information for the Regional Rainfall Frequency Analysis: An Application to Sicily." *Water resources management* 27(6): 1721–35.

Hartigan, J. A. and Wong M. A. (1979), "Algorithm AS 136: A K-Means Clustering Algorithm." *Applied statistics*: 100–108.

Hosking, J. R. M. and Wallis, J. R. (1988), "The Effect of Intersite Dependence on Regional Flood Frequency Analysis." *Water Resources Research* 24(4): 588–600.
http://dx.doi.org/10.1029/WR024i004p00588.

Hosking, J. R. M. and Wallis, J. R. (1993), "Some Statistics Useful in Regional Frequency Analysis." *Water Resources Research* 29(2): 271–81.
http://dx.doi.org/10.1029/92WR01980.

Hosking, J. R. M. and Wallis, J. R. (2005), *Regional Frequency Analysis: An Approach Based on L-Moments*. Cambridge University Press.

Jones, M. R., Fowler, H. J., Kilsby, C. G., and Blenkinsop, S. (2013), "An Assessment of Changes in Seasonal and Annual Extreme Rainfall in the UK between 1961 and 2009." *International Journal of Climatology* 33(5): 1178–94. http://dx.doi.org/10.1002/joc.3503.

Rousseeuw, P. J. (1987), "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis." *Journal of Computational and Applied Mathematics* 20(0): 53–65. http://www.sciencedirect.com/science/article/pii/0377042787901257.

Stedinger, J. R., Vogel, R. M., and Foufoula-Georgiou, E. (1993), “Frequency Analysis of Extreme Events.”

Trenberth, K. E., Dai, A., Rasmussen, R. M., and Parsons, D. B. (2003), “The Changing Character of Precipitation.” *Bulletin of the American Meteorological Society* 84(9): 1205–17.

Viglione, A., Laio, F., and Claps, P. (2007), “A Comparison of Homogeneity Tests for Regional Frequency Analysis.” *Water Resources Research* 43(3).

Wilks, D. S. (2011), 100 *Statistical Methods in the Atmospheric Sciences*. Academic press.