Elsevier Editorial System(tm) for Physics and Chemistry of the Earth
Manuscript Draft

Manuscript Number:

Title: Regionalization of precipitation for the Iberian Peninsula

Article Type: SI: EcoHCC14

Keywords: Precipitation, Cluster Analysis, Regionalization, Iberian Peninsula

Corresponding Author: Ms. Ana Parracho,

Corresponding Author's Institution:

First Author: Ana Parracho

Order of Authors: Ana Parracho; Paulo Melo-Gonçalves; Alfredo Rocha

Abstract: Temporal variability of precipitation over the Iberian Peninsula (IP) has high spatial gradients. Therefore, statistics of the temporal behavior of precipitation and derived quantities over the IP must be estimated taking into account these spatial gradients. Some statistics can be displayed over a map. However there are statistics, such as Probability Density Functions at each location of the IP, that are impossible to display in a map. Because of this, it is mandatory to reduce the number of degrees of freedom which, in this case, consists of a reduction of the time series representative of the IP domain. In this work, we present a spatial partition of the IP region into areas of similar precipitation. For that, daily E-OBS precipitation data for the years between 1986 and 2005 was used. The land-only high resolution data was obtained on a regular grid with 0.25° resolution in the IP domain. This data was subjected to a K-means Cluster Analysis in order to divide the IP into K regions. The clustering was performed using the squared Euclidean distance. Six clusters of IP grid points, defining 6 IP regions, were identified. The grid points in each region share the same time-varying behaviour which is different from region to region. The annual precipitation discriminates the following regions: (i) northwest Iberia (Spanish Galiza and Portuguese Minho); (ii) northwest Portugal (Beira Litoral); (iii) a large region ranging from the center to the western and southwestern shores of the Iberia; (iv) another large region extending from the center to the eastern and southeastern shores of the IP; (v) north (Asturias) and northeast Spain (Pyrenees); and, finally, (vi) a northeastern Iberia near France. The regions obtained for the four seasons of the year are similar. These results are consistent with the thermodynamic characteristics described in the available literature. Finally we emphasize that: (i) the methodology used here, based on Cluster Analysis, can be used to regionalize other areas of the world, and (ii) the identified regions of the IP can be used to represent the Iberian precipitation by six time series that can be subjected to further analysis, whose results can be presented in a concise manner.

Highlights

- Regionalization of precipitation over the Iberian Peninsula using k-means clusters
- Definition of six clusters of grid points, defining 6 precipitation regions
- Different precipitation probability density functions for the centroid of the clusters
- Precipitation indices trends at the centroid of the clusters

REGIONALIZATION OF PRECIPITATION FOR THE IBERIAN PENINSULA

A.C. Parracho[1], P. Melo-Gonçalves[1], A. Rocha[1]

[1]CESAM and Dept. Physics, University of Aveiro, Aveiro, Portugal

Abstract

Temporal variability of precipitation over the Iberian Peninsula (IP) has high spatial gradients. Therefore, statistics of the temporal behavior of precipitation and derived quantities over the IP must be estimated taking into account these spatial gradients. Some statistics can be displayed over a map. However there are statistics, such as Probability Density Functions at each location of the IP, that are impossible to display in a map. Because of this, it is mandatory to reduce the number of degrees of freedom which, in this case, consists of a reduction of the time series representative of the IP domain. In this work, we present a spatial partition of the IP region into areas of similar precipitation. For that, daily E-OBS precipitation data for the years between 1986 and 2005 was used. The land-only high resolution data was obtained on a regular grid with 0.25° resolution in the IP domain. This data was subjected to a K-means Cluster Analysis in order to divide the IP into K regions. The clustering was performed using the squared Euclidean distance. Six clusters of IP grid points, defining 6 IP regions, were identified. The grid points in each region share the same time-varying behaviour which is different from region to region. The annual precipitation discriminates the following regions: (i) northwest Iberia (Spanish Galiza and Portuguese Minho); (ii) northwest Portugal (Beira Litoral); (iii) a large region ranging from the center to the western and southwestern shores of the Iberia; (iv) another large region extending from the center to the eastern and southeastern shores of the IP; (v) north (Asturias) and northeast Spain (Pyrenees); and, finally, (vi) a northeastern Iberia near France. The regions obtained for the four seasons of the year are similar. These results are consistent with the thermodynamic characteristics described in the available literature. Finally we emphasize that: (i) the methodology used here, based on Cluster Analysis, can be used to regionalize other areas of the world, and (ii) the identified regions of the IP can be used to represent the Iberian precipitation by six time series that can be subjected to further analysis, whose results can be presented in a concise manner.

Keywords

Precipitation, Cluster Analysis, Regionalization, Iberian Peninsula

Correspondence

Email: claudiabernardes@ua.pt

1. Introduction

Precipitation is a key variable in climate studies, with changes in its amount and spatial and temporal distributions having an important impact on both human activities (such as agriculture and drinking water resources), and natural hazards (such as droughts and floods). However, climate changes in precipitation are not yet well understood, due to their complexity and regional variability (López-Moreno et al., 2009).

This is true for the Iberian Peninsula (IP), where precipitation has a strong spatial gradient (with several small humid zones generally coinciding with mountainous headwaters (López-Moreno et al., 2008) amid larger dry areas) and a strong seasonal character (Garrido and Garcia, 1992, Serrano et al., 1998). Whereas for autumn, winter and spring, precipitation is mostly due baroclinic synoptic perturbations, moving eastward from Atlantic Ocean; for summer, the precipitation is mostly associated with convective storms due to ground heating, high moisture content, and upper instability (Serrano et al., 1998).

Therefore, it is important to monitor the trends in precipitation for this region, as well as study the trends in extreme precipitation events. However, some statistics (such as Probability Density Functions, which could be used in the study of extreme precipitation events) cannot be displayed over a map. In order to overcome this, a reduction of the number of time series representative of the IP domain is necessary.

Several attempts at regionalizing the precipitation over the Iberian Peninsula have been recorded, using for instance principal component analysis (Serrano et al., 1998; Rodriguez-Puebla et al., 1998; Rocha, 1999). However, in this work, a partition of the IP region into areas of similar precipitation is proposed, using cluster analysis. The data set used as well as the clustering method applied in this work are explained in the next section (Section 2), followed by a section presenting and discussing the results obtained (Section 3) and, finally, a section exposing the conclusions reached (Section 4).

2. Data set and Methods

2.1 E-OBS data

In this work, land-only daily observed precipitation data - E-OBS dataset (Haylock et al., 2008) from the EU-FP6 project ENSEMBLES was used. The dataset was handled with a

resolution of 0.25°, on a regular grid, over a region covering the IP, from roughly 36° to 44°N and from 10°W to 3°E. The period between 1986 and 2005 has been considered.

2.2 K-means cluster analysis

In order to partition the observations into a set number of clusters, k-means clustering was applied to the data. This data-partitioning algorithm assigns each domain grid point (*n* points in R^d) to one of the clusters defined by centroids, in which the number of clusters, k, is defined a priori. The k-means method was applied using the squared Euclidean distance measure (in which each centroid is the mean of the points in that cluster), and a heuristic k-means++ algorithm (Arthur and Vassilvitskii, 2007) for the cluster center initialization. According to these authors, the goal of this method is to find the k centers, C, that minimize:

$$\sum_{x \in X} min \|x - c\|_2^2 \quad (1)$$

Where x is a grid point and $c \in C$. And, inside the cluster, find y that minimizes:

$$\sum_{x \in X} min \|x - y\|_2^2 \quad (2)$$

So that:

$$y = \frac{1}{|X|} \sum_{x \in X} x \quad (3)$$

In short, the method initializes k random cluster centers and assigns each point in the domain to the nearest center. It then recomputes the optimum centers and repeats the assignment of each grid point. It repeats this procedure until the clustering does not change.

In this work, the number of clusters k=6 was determined after sensitivity tests were performed by running the test for several values of k (starting at k=3). It was decided that k=6 gives regions different enough from each other while maintaining consistent characteristics within the cluster. In addition, as this method uses initial random clusters, every single application yields slightly different results. In order to avoid this, the clustering is repeated using different initial cluster centroid positions. The number of times the clustering is repeated was defined by sensitivity tests, in which the number of replicates was increased until the final results from various runs of the program yielded the same (or very similar) results. In this case, the number of replicates was set to 7.

Finally, the results of this analysis are time series of the precipitation at the centroid, for the time period at study and a map in which each grid point is assigned to one of the 6 clusters. The results are presented in the next section, along with the tests used to mathematically evaluate them.

3. Results

3.1 Regionalization

The partition of the Iberian Peninsula into six clusters based on annual precipitation data resulted in the map show in Figure 1. This maps shows the delimitation of six precipitation regions: 1) a region in the Northwest of the IP; 2) a large region extending from the center to the western and southwestern shores; 3) a region in central Iberia; 4) a large region extending from the center to the eastern and southeastern shores; 5) a thin strip of land in the North of the IP (Asturias), extending to the northeast (Pyrenees); and 6) a small area in the northeastern Iberia, near France.

The results obtained for the seasonal precipitation are similar to those obtained for the annual precipitation, with only slight differences between the cold and warm seasons (Figure 2). Therefore, the regions remain unchanged throughout the annual cycle. Furthermore, these results are generally consistent with the thermodynamic characteristics of the Peninsula and are in agreement with its precipitation spatial distribution. The region of Atlantic component is represented by region 1 and (to a lesser extent) 2, the regions of Mediterranean component are represented by regions 4 and (to a lower degree) 6, the region of mountain influence is represented by region 5, and finally region 3 represents the area where the precipitation is under continental influence.

3.2 Validation using probability density functions and the K-S test

Although these results are encouraging, further investigation into their validity is necessary. Therefore, the Probability Density Functions (PDFs) for each centroid precipitation time series were estimated using the Kernel method (Silverman, 1986) with a normalized Kernel function. The density was evaluated at 100 equally spaced points between 1 and 30 mm of precipitation (since for precipitations over 30 mm, the probability density is very close to zero). The results are presented in Figures 3 and 4.

As shown in Figure 3, there is a difference between the PDFs for the annual precipitation at the centroid of each region. For instance, region 1 (under Atlantic influence) has, as expected, higher precipitation, with higher probability (than the other regions) for precipitation of over

about 8 mm. On the other hand, for region 4 (under Mediterranean influence) the precipitation is lower, with a higher probability of precipitation of less than 5 mm. The PDFs obtained for the seasonal precipitation (presented in Figure 4) show comparable results, with similar results from one season to another.

In order to test if the precipitation distributions at the centroids of the different regions are significantly different from each other, the Kolmogorov-Smirnov (K-S) test was performed. The K-S test (Wilks, 2006) establishes a null-hypothesis that the datasets belong to the same continuous distribution. This hypothesis is rejected if the discrepancy, D, is high enough.

$$D_S > \sqrt{-\frac{1}{2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\ln\left(\frac{\alpha}{2}\right)} \qquad (4)$$

In which n_1 and n_2 are the sizes of the data series, α is the significance level (5% in this case) and

$$D_S = max(|F1(x_1) - F2(x_2)|) \qquad (5)$$

Where $F1$ and $F2$ are the continuous cumulative distributions of the time series being tested. The results for this test, for both the annual and seasonal precipitation at the centroids, show that the six regions are statistically different from each other, at a 5% significance level.

In order to test how well the centroids represent the region they are assigned to, the precipitation distributions at the centroids were compared with the precipitation distributions of each grid point belonging to the region the centroid represents. The K-S test was performed and the percentage of grid points from which the centroid's precipitation is not significantly different is presented in Table 1.

From the results shown in the table, it can be concluded that the representativity of the centroids varies with the region and with the season of the year. For the annual precipitation, the best result was observed for region 1, in which close to 60% of grid points have precipitation distributions that are not different from the centroid's distribution. However, for regions 3 and 4, only a small percentage (approximately 7 and 2, respectively) of grid points are represented by the cluster. For region 4, the percentage remains relatively small throughout the four seasons of the year. The best results are observed for MAM, with higher percentages of grid points (most above around a third of grid points); and the worst results are for SON, with relatively low percentages across the entire domain.

In light of these tests, one should proceed with caution when interpreting the results obtained using the cluster centroids, as they may not represent their entire region completely.

3.3 Application of regionalization to the study of precipitation indices trends

One of the possible applications of this regionalization is, for instance, the study of precipitation indices and their trends for each of the regions by using the centroid time series. In this work, several indices were computed for the six centroids, for the annual and seasonal precipitation, based on the indices defined by the CCI/CLIVAR/JCOMM Expert Team on Climate Change Detection and Indices (ETCCDI). Then, their trends were calculated, using the Theil-Sen regression (Theil, 1950 & Sen, 1968). The statistical significance of each trend was tested using the non-parametric Mann-Kendall Test (Mann, 1945 & Kendall, 1955), at a 5% significance level.

The indices showing more robust trends are the annual total precipitation (PRCPTOT), the maximum number of consecutive dry days (CDD) and the 90^{th} percentile of daily rainfall (Prec90p). These indices are presented in Table 2.

According to Table 2, the PRCPTOT shows a decreasing trend of more than 5 mm/year for the annual precipitation of region 1; an increase of nearly 7 mm/year in MAM, countered by a decrease of about 4.4 mm/year in JJA for the precipitation in region 1; an increase of 4.2 mm/year in MAM followed by a decrease of 4.6 in JJA for the precipitation in region 2; and, finally, a decrease in precipitation of 3.2 mm/year in SON for region 5. For this index, regions 3, 4 and 6 show no significant trends. In general, this index points to an annual decrease in precipitation over the entire IP, with stronger decrease during summer and a more significant increase during spring.

When it comes to the CDD, regions 1, 3, 4 and 5 show no significant trends, with several months having a zero days/year trend. On the other hand, region 2 shows a decrease in CDD of less than 1day/year for the annual precipitation, with slight increases in DJF, JJA and SON and DJF; on the other hand, region 4 has an increase of about half a day per year for DJF.

For Prec90p, regions 1, 2 and 4 show no significant trends. However, for MAM, region 2 shows an increase of about 0.3 mm/day/year, while region 6 shows a decrease of slightly smaller magnitude for Prec90p. Finally, region 6 shows an increase of close to 0.2 mm/day/year for JJA precipitation.

In general, however, trends appear to be slight, with a few inconsistencies between seasonal trends and annual trends. This may be due to the fact that the regionalization produces

different regions for annual and seasonal precipitations (see Figures 1 and 2); therefore the centroids are not collocated from one season to the other, or for the annual precipitation. As such, these results would need to be taken with caution and additional validation methods should be carried out.

4. Conclusions

This paper focused on the regionalization of the precipitation in the Iberian Peninsula, for use in climate studies. The regionalization was performed using k-means clustering on the observed precipitation data for the recent past: 1986 through 2005. This resulted in six regions with statistically different probability density functions for the precipitation at the centroids of the clusters. These centroids were then used to compute trends in precipitation indices, using the Theil-Sen regression. However, not all of the resulting trends were statistically significant, according to the Mann-Kendall Test. Three of the indices, with more statistically significant trends were chosen and shown for the six regions, for the annual and seasonal precipitation. Although this is one of the possible applications for this method, it can also be used in other studies and applied to other regions. Nevertheless, the results have to be interpreted with caution, as it was also shown that the centroids may not be representative of the entire region with which they are associated.

Acknowledgements

This study was supported by FEDER funds through the Programa Operacional Factores de Competitividade – COMPETE and by Portuguese national funds through FCT - Fundação para a Ciência e a Tecnologia, within the framework of the following projects: CLIPE PTDC/AAC-CLI/111733/2009; CLICURB EXCL/AAG-MAA/0383/2012.

References

Arthur, D., Vassilvitskii, S., 2007. K-means++: The advantages of careful seeding. Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics Philadelphia, PA, USA, 1027–1035.

Garrido, J., García, J.A., 1992. Periodic signals in Spanish monthly precipitation data. Theoretical and Applied Climatology, 45, 97–106.

Haylock, M.R., Hofstra N., Klein Tank, A.M.G., Klok, E.J., Jones, P.D., New, M., 2008. A European daily high-resolution gridded dataset of surface temperature and precipitation. Journal of Geophysical Research (Atmospheres), 113, D20119, doi:10.1029/2008JD10201.

Kendall, M.G., 1955. Rank Correlation Methods. Griffin, London.

López-Moreno, J.I., García-Ruiz, J.M., Beniston, M., 2008. Environmental Change and water management in the Pyrenees. Facts and future perspectives for Mediterranean mountains. Global and Planetary Change, 66, 300–312.

López-Moreno, J.I., Vicente-Serrano, S.M., Angulo-Martínez, M., Beguerías, S., Kenawy, A., 2009. Trends in daily precipitation on the northeastern Iberian Peninsula, 1955–2006. International Journal of Climatology, 30, 1026–1041, doi: 10.1002/joc.1945.

Mann, H.B., 1945. Nonparametric tests against trend. Econometrica, 13, 245–259.

Rocha, A., 1999, Low-Frequency Variability Of Seasonal Rainfall Over The Iberian Peninsula And ENSO. International Journal of Climatology, 19, 889–901

Rodriguez-Puebla, C., Encinas, A.H., Nieto, S., Garmendia J., 1998. Spatial And Temporal Patterns Of Annual Precipitation Variability Over The Iberian Peninsula. International Journal Of Climatology, 18, 299-316.

Sen, P.K., 1968. Estimates of the regression coefficient based on Kendall's tau. Journal of the American Statistical Association, 63, 1379-1389.

Serrano, A., García, J.A., Mateus, V.L., Cancillo, M.L., Garrido, J., 1998. Monthly Modes of Variation of Precipitation over the Iberian Peninsula. Journal Climate, 12, 2894-2919

Silverman, B.W., 1986. Density Estimation for Statistics and Data Analysis. Chapman and Hall: London, 175 pp.

Theil, H., 1950. A rank-invariant method of linear and polynomial regression analysis, I. Proc. Kon. Ned. Akad. v. Wetensch.A53, 386-392.

Wilks, D. S., 2006. Statistical Methods In Atmospheric Sciences. Elsevier, 649 pp.

List of figures

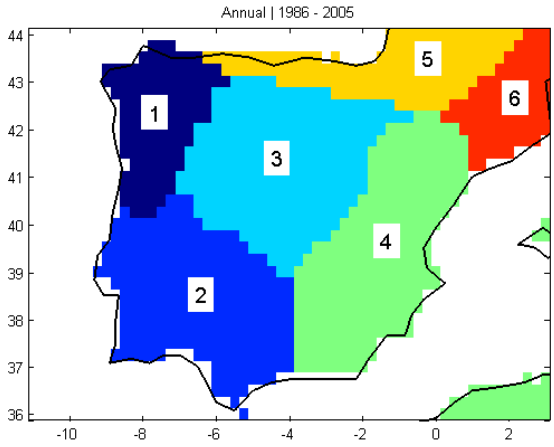

FIGURE 1: PRECIPITATION REGIONS DEFINED BY THE K-MEANS CLUSTER ANALYSIS FOR THE ANNUAL DATA FROM 1986 TO 2005.

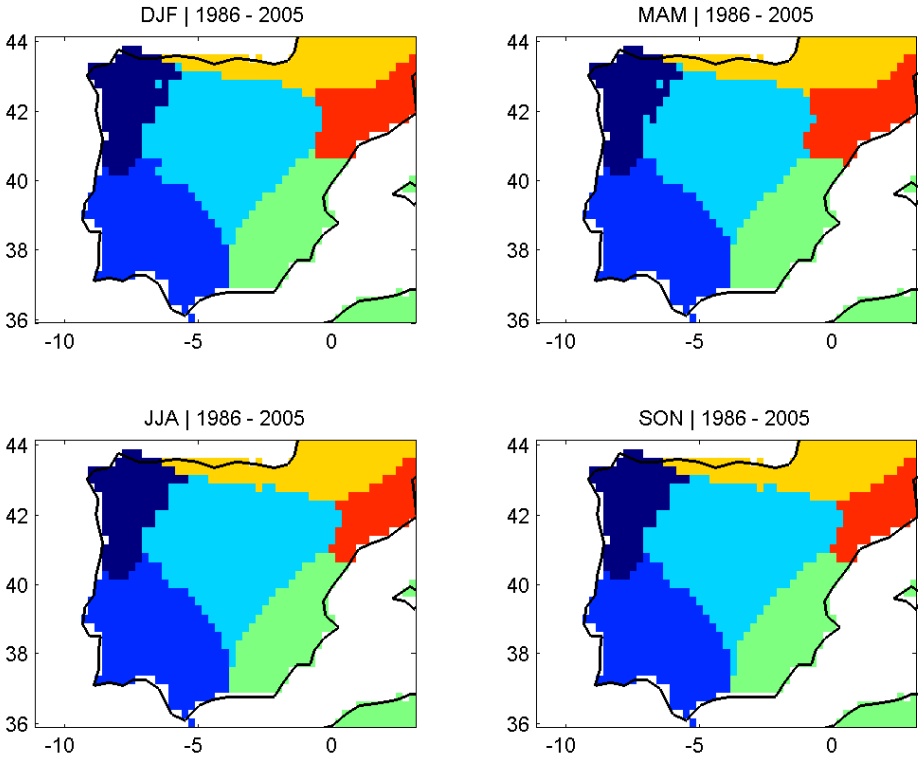

FIGURE 2: PRECIPITATION REGIONS DEFINED BY THE K-MEANS CLUSTER ANALYSIS FOR THE SEASONAL DATA FROM 1986 TO 2005.

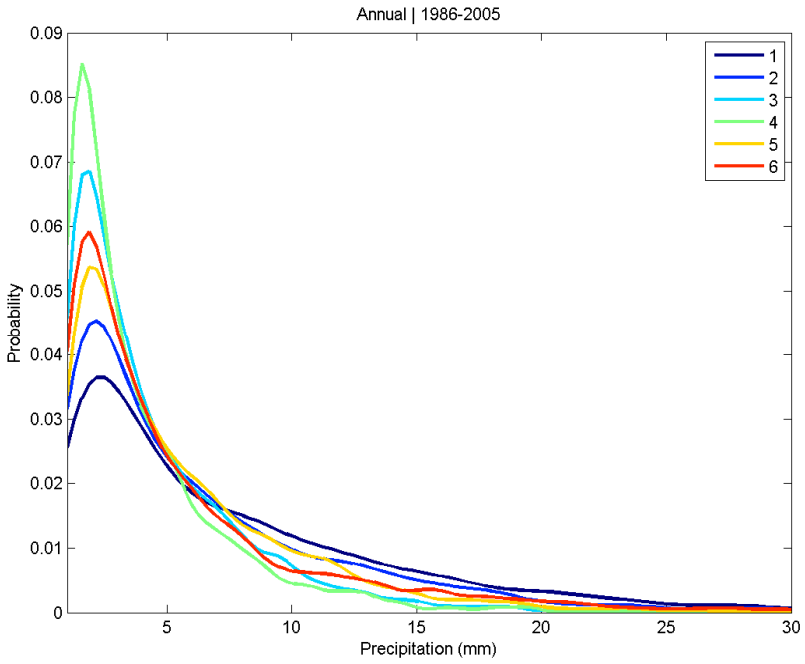

FIGURE 3: PROBABILITY DENSITY FUNCTIONS (PDFS) FOR EACH CENTROID (1-6) DEFINED BY THE K-MEANS CLUSTER ANALYSIS OF THE ANNUAL PRECIPITATION DATA FROM 1986 TO 2005.

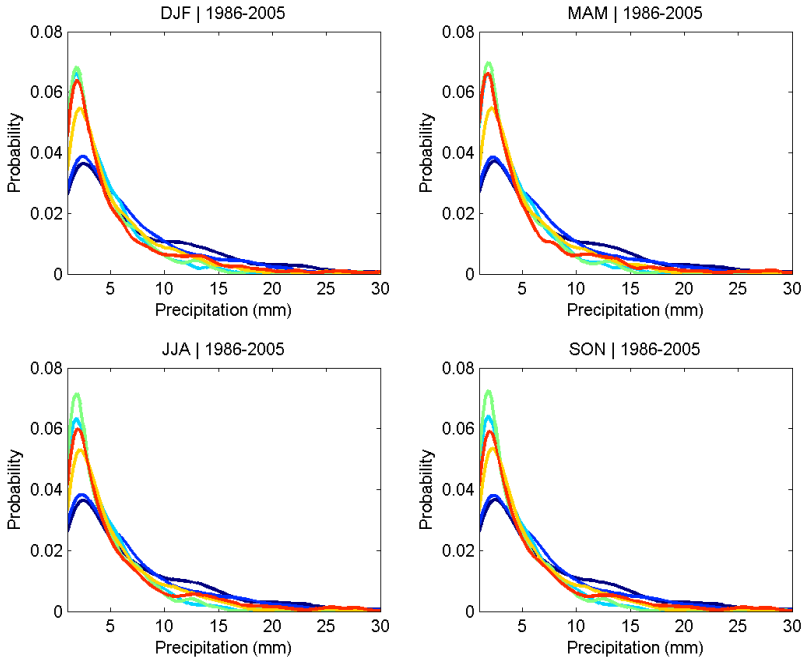

FIGURE 4: AS IN FIGURE 3, BUT FOR SEASONAL DATA.

TABLE 1: PERCENTAGE OF GRID POINTS FOR WHICH THE PRECIPITATION DISTRIBUTION IS NOT SIGNIFICANTLY DIFFERENT FROM THE CENTROID'S DISTRIBUTION, FOR EACH REGION AND SEASON OF THE YEAR.

Region	Annual	DJF	MAM	JJA	SON
1	58.1197	35.8974	63.2479	41.0256	29.0598
2	38.3333	13.7500	84.1667	47.5000	15.8333
3	7.0313	7.7160	39.1975	18.2099	4.0123
4	1.7182	9.6220	10.3093	7.5601	2.7491
5	29.6552	20.6250	31.2500	6.2500	11.8750
6	26.0870	32.0755	35.8491	36.7925	14.1509

TABLE 2: TRENDS OF PRCPTOT, CDD AND PREC90P INDICES, FOR ANNUAL AND SEASONAL PRECIPITATIONS, BETWEEN 1986 AND 2005. STATISTICALLY SIGNIFICANT TRENDS , AT THE 5% LEVEL, ARE UNDERLINED.

Index	Region	Annual	DJF	MAM	JJA	SON
PRCPTOT (mm/year)	1	-5.3121	-4.1058	6.7950	-4.4478	2.0201
	2	-1.2503	-2.5514	4.2162	-4.6150	4.0068
	3	-0.9648	0.1574	1.7450	-0.9946	-0.1148
	4	-0.6935	0.4861	0.8250	-0.3421	-0.6112
	5	-5.2653	0.0304	-1.4657	1.8884	-3.2002
	6	-4.3402	1.7139	-1.6113	2.5263	-0.4339
CDD (N_{days}/year)	1	-0.2250	0.0000	0.0000	0.1270	-0.1000
	2	-0.7735	0.1667	0.0000	1.0000	0.0955
	3	0.0000	0.1484	0.0000	0.3333	0.0871
	4	-0.0651	0.4495	0.0000	-0.0313	0.0000
	5	0.0646	0.0000	0.0000	-0.1082	0.1082
	6	-0.1667	-0.2330	0.1250	0.0000	0.0000
Prec90p (mm/day/year)	1	0.0213	0.1347	0.2757	0.0131	0.0668
	2	-0.0014	-0.1717	0.2812	-0.1383	0.2629
	3	-0.0043	0.1174	0.0264	0.0207	0.0616
	4	0.0293	-0.0318	0.0953	-0.0411	-0.0082
	5	0.0074	-0.0059	-0.0381	0.0548	-0.1072
	6	-0.0635	0.0711	-0.2422	0.1906	0.0417