Journal of
Water & Climate Change

Assessing changes in intra- and inter-annual precipitation variability using advanced statistical inference techniques

Elsa Estevão Moreira [iD][a,b,*] and Vítor Augusto[b]

[a] Center for Mathematics and Applications (NOVA Math) and Department of Mathematics, NOVA FCT, Campus Caparica, Caparica 2829-516, Portugal
[b] Center for Mathematics and Applications (NOVA Math), NOVA FCT, Campus Caparica, Caparica 2829-516, Portugal
*Corresponding author. E-mail: efnm@fct.unl.pt

[iD] EEM, 0000-0003-0116-1917

ABSTRACT

Aiming to find statistically significant changes on the usual monthly weather conditions in Alentejo, Portugal, in the past 40 years, time series of precipitation from a grid of locations in Alentejo were studied using an unconventional approach. The time series were divided into 4 decades and disposed in contingency tables with two factors, Year and Month. Then, log-linear models were fitted to those tables. The model deviances for Year and Month represent the weight of each factor in explaining precipitation variability and Year:Month deviance is an estimate of the lack of independence between both factors, representing changes in intra-annual precipitation variability. The set of all the models deviances, per decade and location, were analysed using ANOVA techniques, first to compare the four decades, then to perform a crossed comparison between decades and between four pre-defined zones. Results indicate significant differences between the oldest and recent decades in terms of intra-annual precipitation variability, which could be interpreted as a trend towards softening the differences in precipitation between the wet months of the year. Furthermore, two homogeneous sub-regions could be defined in Alentejo, a large interior region and a smaller one close to the Atlantic Ocean.

Key words: ANOVA, backward elimination, climate change, Tukey multiple comparisons

HIGHLIGHTS

- Seeking for changes in the annual cycles and patterns of precipitation.
- Use of unusual methodology, namely log-linear models and analysis of variance.
- Findings suggest the differences in precipitation between the wet months are attenuating.
- Findings suggest Alentejo is not a homogeneous region regarding precipitation intra-annual variability.
- Two homogeneous regions could be defined in Alentejo.

1. INTRODUCTION

Portugal is located in Iberia Peninsula, the western tip of the Mediterranean basin, in the transition zone between the arid to semiarid climates of subtropical regions and the humid climates typical of northern Europe. Portugal has a marked precipitation seasonality: humid mild winters, and dry warm/hot summers, Lima *et al.* (2023). In particular, Alentejo is a water scarce region in southern Portugal, highly vulnerable to climate change. The frequent recurrence of severe and extreme droughts in Alentejo have been subject of many studies, e.g., Paulo *et al.* (2003), Paulo & Pereira (2006), Moreira *et al.* (2006), Moreira *et al.* (2012b) and Moreira *et al.* (2013).

Several studies for Iberia Peninsula indicate a strong evidence that climate change has produced a decrease in precipitation. For example, Senent-Aparicio *et al.* (2023) updates the trends in magnitude and seasonality of precipitation in Spain from 1951 to 2019 at different time scales, confirming the decreasing trend in annual precipitation in most of the Spanish territory, with particularly significance during March and June. In another study also for Spain, Ceballos *et al.* (2004), is confirmed a negative rainfall trend and an increase in intra-annual variability in rainfall. Regarding Portugal there are few studies with recent data that analyse the evolution of monthly precipitation. According to Lima *et al.* (2010), that studied 10 stations in Portugal, findings do not suggest overall important changes in precipitation distribution, over long time

spans ranging from 88 to 145 years up to 2007, only March exhibits some significant decreasing trends for some of the locations. However, studies seeking for changes in the annual distribution of the precipitation through months in Iberia Peninsula have a great importance. These kind of disturbances in the annual cycle have, naturally, great impact in crops which are adapted to the usual cadence of precipitation over the year. Moreover, citing (Wan *et al.* 2024), it is said that the ongoing intensification of the hydrological cycle due to global climate change alters intra-annual precipitation variability and changes in precipitation patterns lead to disparities in soil moisture.

With this work one aims detecting changes on the usual seasonal weather conditions in Alentejo, regarding precipitation and then in the future, assess the impact of those changes on vineyards in the Alentejo wine region, with the final aim of selecting which grapevine varieties can be better adapted to a changed climate. There are several studies assessing the effects of climate change on viticulture in Portugal. Citing a few, in Fraga *et al.* (2017) and Santos *et al.* (2020), the impacts that climate change can have on Portuguese and Mediterranean viticulture were addressed, where is reinforced the importance that changes in temperature, precipitation and other climatic indicators have on the wine earnings and its quality.

In particular, the present study sought to find changes in the annual cycles and patterns of precipitation in Alentejo in the past four decades that possibly could be attributed to climate change. The way to approaching this goal is an unconventional one, since uses statistical modelling and inference applied to time series. The typical methodology used in literature to achieve similar goals is the trend analysis. For instance, in Benetó & Khodayar (2023), were identified significant decreasing trends in precipitation during June and October and increasing trends during September and November, in the last seven decades, in the eastern Spain. Another study confirm precipitation decreasing trends in most of the Spanish territory, particularly significant during March and June (Senent-Aparicio *et al.* 2023).

Regarding the methodologies used, statistical models are a reasonable alternative to physical models for meteorologists and other scientific groups that aim to model climatic data. They often apply multivariate statistical techniques like the Canonical Correlation Analysis (CCA), robust multi-linear regression, Principal Component Analysis (PCA), Singular Spectrum Analysis (SSA) among others Wilks (2011), van den Dool (2006). Each methodology, independently of its complexity, has advantages and limitations. In climate studies, a methodology often used to study possible spatial patterns of variability and how they change with time is the Empirical Orthogonal Function (EOF) analysis that bases on PCA. For instance, in Zeleke *et al.* (2024), the Rotated Empirical Orthogonal Function (REOF) was used to detect the spatio-temporal variability of precipitation. In Portugal the spatial and temporal variability of precipitation and drought was assessed using R- and S-mode PCA (Martins *et al.* 2012). Results pointed to two distinct sub-regions in the country relative to both precipitation regimes and drought variability and no linear trend indicating drought aggravation or decrease was revealed (Martins *et al.* 2012). The log-linear modelling is a well-known methodology used to model categorical data with different purposes and have been used by the main author in the past to model climatic data. For instance, more complex log-linear models were used with the purpose of analyse and predict drought in Portugal (Moreira *et al.* 2012b, 2013, 2016, 2018). In these works, time series of Standardized Precipitation Index (SPI) and also Standardized Precipitation Evapotranspiration Index (SPEI) were used instead of raw precipitation time series, since the aim was to assess only drought. Moreover, Analysis of Variance (ANOVA)-like inference has been used by the main author to find significant differences between time series different periods regarding drought class transitions (Moreira *et al.* 2012b) and also to find homogeneous regions in Alentejo relative to drought class transitions, that is, to carry out a spatial analysis by using the latitude and longitude as factors in a two-way ANOVA (Moreira *et al.* 2013). With the same aim, a more usual technique when it comes to obtaining spatially homogeneous regions – cluster analysis – was applied to group log-linear models using Likelihood Ratio Test (LRT) *p*-values in Moreira *et al.* (2012a). The results of these last two studies differed, with the first one finding two different homogeneous regions and the second not. As a result, when comparing these two approaches, the ANOVA technique stood out for being more sensitive.

The data analysed in the present work consists in the amount of precipitation estimated monthly in the past 40 years for a set of locations representative of the region. These values were viewed has counts representing the number of precipitation mm^3 occurred in each month during periods of 10 years for each location and therefore they could be considered to form a contingency table with 2 categories or factors, Year and Month. To these contingency tables, saturated log-linear models were fitted and then, a temporal analysis was done by comparing the four decades using ANOVA applied to the deviances of the log-linear models regarding the two categories, Year and Month, as well as the interaction between both. A second temporal analysis was done by comparing the first decade with the most recent one, with aim of studied the relevance of each month considered in the models. Finally, a spatio-temporal analysis was performed through dividing Alentejo region in four zones and testing for significant differences using again ANOVA.

The idea of considering the amount of monthly precipitation as counts, i.e., categorical data, is new and is a different way of extracting information from precipitation time series. Also, when comparing with the previous works mentioned, herein the precipitation time series are modelled directly without transforming in drought indices, since the aim is different. Another novelty resides in using the resulting deviances from model fit, as observations of ANOVA and multiple comparison techniques, something that has never been done. The techniques are not new, but the approach presented has never been used for similar purposes.

2. MATERIALS AND METHODS

2.1. Data

The data used in this study consists in precipitation data-sets retrieved from European Centre for Medium-Range Weather Forecasts (ECMWF)[1] with 0.25 degrees of spatial resolution located over mainland Portugal. ERA5-ECMWF is a dataset available for public use and provides hourly, daily and monthly estimates of a large number of atmospheric, land and oceanic climate variables, like precipitation and temperature. The data covers the Earth on a 30 km grid and includes information from 1979 to 2019.

The data set to be studied is composed by 49 locations (grid points) in Alentejo region selected from the grid referred above, where the amount of monthly precipitation is expressed in cubic millimetres. Figure 1 show all grid points located in Portugal and the selected ones in Alentejo. Consequently, for each grid location there are 41 years with the amount of precipitation per month.

2.2. Methods

As said before in introduction, aiming a temporal analysis, the total period was divided in 4 sub-periods of 10 years in order to compare the decades of 1979–1988, 1990–1999, 2000–2009 and 2010–2019. Based on the graphical representation and some descriptive statistics of the data, the year of 1989 was chosen to not be included in the study since the distribution of precipitation over the months of that year, as well as the year total do not differ significantly from the previous year (1988) included in the first decade and also from the posterior year (1990) to be included in the second decade. Therefore, the non-inclusion of 1989 would not significantly affect the results of the study. That decision was also taken, since we wanted to further ahead, compare the earliest with latest decade available, so it was of our interest to consider the first year of data (1979).

To perform the analysis, the data for each location was disposed in contingency tables with two categories, the Year in rows and the Month in columns (Everitt 1977; Agresti 2002). In these tables, the amount of monthly precipitation was viewed as counts for the modelling purpose. Each mm^3 of precipitation counts one, so the total volume also represents de total number of precipitation mm^3 occurred in 9 of the 12 months of year, separated in periods of 10 years, as explained next. Then saturated log-linear models were fitted to those tables. Each contingency table has 10 rows, corresponding to the years of the decade in study, and only 9 columns, corresponding to the nine most important months of the year in terms of precipitation to viticulture in Alentejo region. June, July and August were not considered, since they are months when it hardly rains in Alentejo. The climate changes that have been taking place in Mediterranean regions are not affecting the summer months in terms of precipitation, as they remain very dry, but have become increasingly hotter. Vineyard crops in Alentejo are adapted to dry and hot summers, so in a study focussed on precipitation it makes sense for these months to be discarded as they have little influence on the production and quality of viticulture (Droulia & Charalampopoulos 2021; Wunderlich et al. 2023).

Moreover, their inclusion would result in contingency tables with many zeros, which must be avoided as it does not allow a good adjustment of the log-linear models, as well as, in their inference (Everitt 1977; Agresti 2002).

Table 1 presents an example of the contingency table for location 205, where each cell represents the number of mm^3 of precipitation occurred in month 1, 2, 3, 4, 5, 9, 10, 11, 12 of year $i = 1, \ldots, 10$, of the decade 1979–1988.

In a third analysis, a kind of spatio-temporal analysis was performed by dividing Alentejo region in four zones, which one with 12 locations, corresponding to the different coloured dots in Figure 1. Dots coloured in red form zone 1 (North), dots coloured in green form zone 2 (Center), dots coloured in yellow form zone 3 (West) and dots coloured in blue form zone 4 (East). Location corresponding to the only black dot was excluded from this third analysis, since leftover.

---

[1] More information in https://www.ecmwf.int/en/forecasts/dataset/ecmwf-reanalysis-v5.
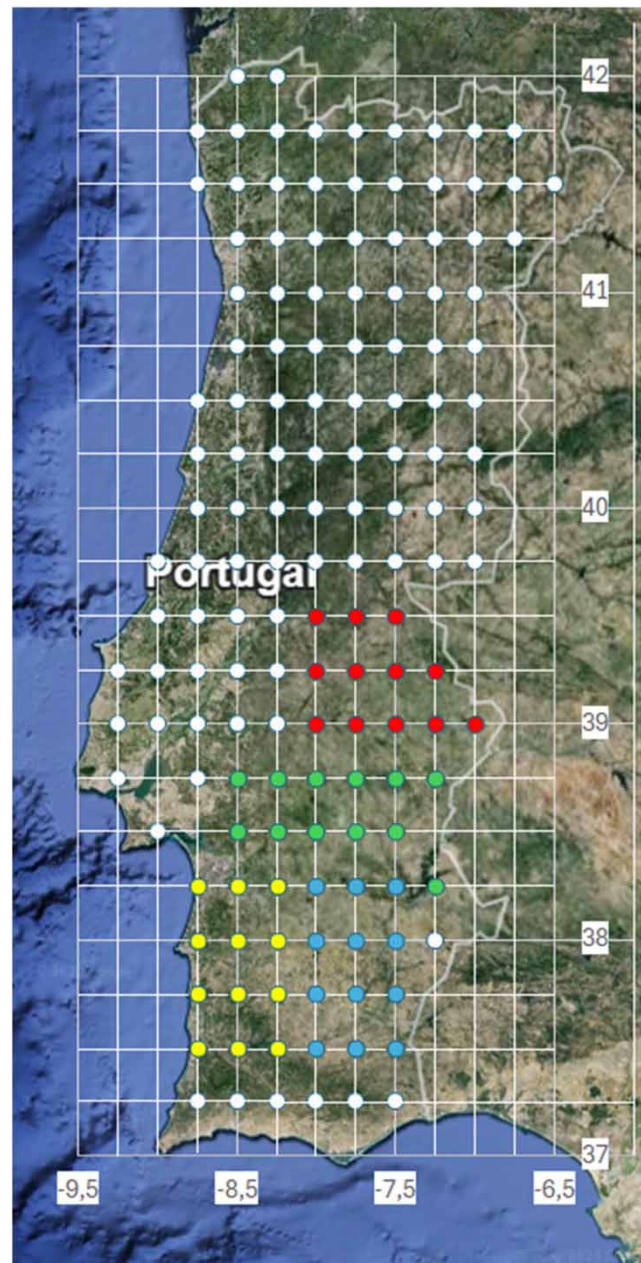
Figure 1 | Grid points located in Portugal where the selected grid points in Alentejo coloured. Dots coloured in red form zone 1 (north inland), dots coloured in green form zone 2 (center inland), dots coloured in yellow form zone 3 (south coast) and dots coloured in blue form zone 4 (south inland). Map copyright information: Google Data SIO, NOAA, U.S. Navy, NGA, GEBCO Landsat/Copernicus Inst. Geogr. Nacional, 2013.

2.2.1. Log-linear models

The observed frequencies in these contingency tables represented by n_{ij}, $i = 1, \ldots, 10$ and $j = 1, \ldots, 9$, consists of the number of precipitation (mm^3) occurred in month $j = 1, \ldots, 9$ of year $i = 1, \ldots 10$. To each of the contingency tables mentioned before, a saturated log-linear model with two categories or factors (the Year and the Month)

$$\log(m_{i,j}) = \lambda + \lambda_i^Y + \lambda_j^M + \lambda_{i,j}^{YM}, \quad i = 1, \ldots, 10, \; j = 1, \ldots, 9, \tag{1}$$

Table 1 | Contingency table for location 205 (Évora district), for the decade 1979–1988

Year/Month	1	2	3	4	5	9	10	11	12
1979	10	33	19	178	50	13	35	100	264
1980	151	71	151	12	40	34	132	32	100
1981	55	19	133	70	18	96	56	137	151
1982	89	108	52	89	2	26	190	98	90
1983	68	71	45	26	19	10	151	25	41
1984	0	15	36	15	24	7	135	114	63
1985	40	63	100	42	1	48	204	171	45
1986	18	107	17	57	36	52	19	41	36
1987	77	95	28	148	50	33	37	35	53
1988	111	78	16	64	18	3	44	52	235

was fitted, where $m_{i,j}$ represent the expected frequencies and the parameters λ, λ_i^Y, λ_j^M and $\lambda_{i,j}^{YM}$ are, respectively, the grand mean, the effect of the Year, the effect of the Month and the interaction between both (Everitt 1977; Agresti 2002). Regarding the model parameters, the following conditions must be verified:

$$\sum_{i=1}^{10} \lambda_i^Y = 0, \quad \sum_{j=1}^{9} \lambda_j^M = 0, \quad \sum_{i=1}^{10}\sum_{j=1}^{9} \lambda_{ij}^{YM} = 0,$$

as a result, model (1) has one parameter λ, 9 independent parameters λ_i^Y, 8 independent parameters λ_j^M and $8 \times 9 = 72$ independent parameters $\lambda_{i,j}^{YM}$, which leads to a total of 90 independent parameters, equal to the number of observed frequencies, that is, the model fits perfectly to the data, being then called a saturated model.

For this model it is assumed that the observed frequencies are independent and identically distributed (i.i.d.) random variables with Poisson distribution, therefore the model parameters are estimated using maximum likelihood method.

Regarding the main effects Year and Month, the hypothesis to be tested are

$$H_{0Y}: \lambda_1^Y = \cdots = \lambda_{10}^Y = 0 \quad \text{vs} \quad H_{1Y}: \exists i \in \{1, \ldots, 10\}: \lambda_i^Y \neq 0,$$
$$H_{0M}: \lambda_1^M = \cdots = \lambda_9^M = 0 \quad \text{vs} \quad H_{1M}: \exists j \in \{1, \ldots, 9\}: \lambda_j^M \neq 0.$$

Regarding the interaction, the hypothesis of statistical independence between Year and Month are

$$H_{0YM}: \forall i \in \{1, \ldots, 10\}, \forall j \in \{1, \ldots, 9\}, \ \lambda_{i,j}^{YM} = 0$$

vs

$$H_{1YM}: \exists i \in \{1, \ldots, 10\}, \exists j \in \{1, \ldots, 9\}: \lambda_{i,j}^{YM} \neq 0.$$

If the null hypothesis H_{0YM} is rejected, the model (1) holds, meaning that factors Year and Month are not independent.

When testing hypothesis H_{0Y}, H_{0M} and H_{0YM}, the model goodness of fit is also being tested. To do that, the chi-squared statistic

$$G^2 = 2 \sum_{i=1}^{10}\sum_{j=1}^{9} n_{i,j} \log\left(\frac{n_{i,j}}{\hat{m}_{i,j}}\right) \tag{2}$$

called residual deviance, is used, with $\hat{m}_{i,j} = \exp\left(\hat{\lambda} + \hat{\lambda}_i^Y + \hat{\lambda}_j^M + \hat{\lambda}_{i,j}^{YM}\right)$, where $\hat{\lambda}$, $\hat{\lambda}_i^Y$, $\hat{\lambda}_j^M$, $\hat{\lambda}_{i,j}^{YM}$ are the parameters of maximum likelihood estimates. When the null hypothesis holds, G^2 statistic has asymptotically central chi-square distribution, with a

number of degrees of freedom equal to the difference between the number of parameters in the saturated model and number of parameters in a particular unsaturated model. If the null hypothesis H_{0YM} is not rejected, G^2 equals zero and has zero degrees of freedom.

Taking the saturated model as a starting point, backward elimination method (Agresti 2002), can be used to obtain a simplest model. This method consists of, step by step, removing the set of parameters that are not statistically significant to model the observations, then creating a sub-model with fewer parameters, without significant loss of information. So, considering model (1) with 90 parameters, in a first step, the null hypothesis that the set of parameters representing the interaction between the level j of factor M with all levels of factor Y are all equal to zero

$$H_0^{(j)}: \lambda_{1j}^{YM} = \ldots = \lambda_{10j}^{YM} = 0 \quad \text{vs} \quad H_1^{(j)}: \exists i, k \in \{1, \ldots, 10\}: \lambda_{ij}^{YM} \neq \lambda_{kj}^{YM}$$

is tested. Testing these hypotheses is equivalent to test if the interaction parameters corresponding to the Month j can be eliminated from the model. In order to do that, G^2 is computed for the model without these 10 parameters, which has chi-square distribution with 10 degrees of freedom (df), the difference between the number of parameters in the saturated model and in the simplest model. $H_0^{(j)}$ is rejected if G^2 exceeds the chi-square quantile with 10 df and significant level α. Not rejecting $H_0^{(j)}$ indicates that those parameters can be eliminated from the model, since these are not significant. The same procedure is applied for all months $j = 1, \ldots, 9$ for the YM interactions. An equivalent test could be performed for the parameters representing each level of factors Y and M. At the end, a simplest model is obtained, containing only the parameters that are significant to explain the observations.

2.2.2. Analysis of variance

Analysis of Variance (commonly named as ANOVA) is a statistical technique used to test if there are significant differences between several levels, named treatments, of the same factor, in case of one-way ANOVA (Montegomery 2013). For each treatment there is a sample with several observations named replicates. To perform ANOVA, it is assumed that the samples are i.i.d. with normal distribution and the same variance σ^2. However, even if the data does not verify these assumptions, ANOVA can still be applied, namely when the samples sizes are large and the number of observations is the same in all treatments (balanced ANOVA).

ANOVA with one factor is the simplest linear model in analysis of variance, but more than one factor may exist, in particular the case with two factors (two-way ANOVA) is well known. In two-way ANOVA, there are two different factors that may influence the response variable, each one with a certain number of levels. For each combination of the two-factor's levels there is a sample with the same size. In this situation, using hypotheses testing, it is studied the influence of each factor on the response variable, as well as if there is interaction between both factors (Montegomery 2013).

Lets consider the case with two factors: factor A with r levels, factor B with s levels and n replicates (sample's size n). Using the notation used in Montegomery (2013), the ANOVA linear model with fixed effects and interaction is written as

$$Y_{i,j,k} = \mu + \tau_i + \beta_j + (\tau\beta)_{i,j} + e_{i,j,k}, \tag{3}$$

for $i = 1, \ldots, r, j = 1, \ldots, s$ and $k = 1, \ldots, n$, where μ is the overall mean, τ_i is the effect of the i-th level of the row factor A, with $\sum_{i=1}^{r} \tau_i = 0$, β_j is the effect of the j-th level of column factor B, with $\sum_{j=1}^{s} \beta_i = 0$, $(\tau\beta)_{i,j}$ is the two-factor interaction effect, with $\sum_{i=1}^{r} (\tau\beta)_{i,j} = 0$ and $\sum_{j=1}^{s} (\tau\beta)_{i,j} = 0, e_{i,j,k} \sim N(0, \sigma^2)$ i.i.d. are the random error components, and $Y_{i,j,k}$ is the k-th observed response when factor A is at the i-th level and factor B is at the j-th level.

The one-way ANOVA model only includes the overall mean, the single factor effect with r levels and the random error components for the r levels and n replicates.

For these kind of models, the hypotheses to test are about the equality of factor A levels

$$H_0: \tau_1 = \cdots = \tau_r = 0 \quad \text{vs} \quad H_1: \exists i \in \{1, \ldots, r\}: \tau_i \neq 0,$$

about the equality of factor B levels

$$H_0: \beta_1 = \cdots = \beta_s = 0 \quad \text{vs} \quad H_1: \exists j \in \{1, \ldots, s\}: \beta_j \neq 0$$

and about if factor A interacts with factor B

$$H_0 : \forall i \in \{1, \ldots, r\} \; \forall j \in \{1, \ldots, s\}, \; (\tau\beta)_{i,j} = 0 \quad \text{vs} \quad H_1 : \exists i \in \{1, \ldots, r\} \; \exists j \in \{1, \ldots, s\} : (\tau\beta)_{i,j} \neq 0$$

Testing these hypotheses involves building an ANOVA table (Table 2), to perform a F test, where the sum of squares for factors A, B and interaction, as well as, for the error are presented jointly with the corresponding degrees of freedom. The way of computing those sum of squares is not exposed here for simplicity but can found in Montegomery (2013).

The effects of factor A, B and interaction AB are significant, if the values of statistics F_0^A, F_0^B and F_0^{AB}, respectively, do not exceed the $1 - \alpha$ quantile of the F distribution with the corresponding degrees of freedom. Usually statistical software like R, have implemented the computation of this table and present the test p-values.

When obtaining significant effects, i.e., very low p-values, one can use multiple comparison methods, like Tukey test, to find which pairs of factor levels are significantly different (Montegomery 2013).

3. RESULTS AND DISCUSSION

Saturated log-linear models were fitted to the 49 contingency tables, mentioned in Section 2.1, using R software and the code described in Dunn & Smyth (2018). All models are composed by $9 \times 10 = 90$ independent parameters, which is equal to the number of cells in the contingency tables. From the R output, an ANOVA type table as in the example presented in Table 3 is obtained.

For log-linear models, these tables are equivalent to the ANOVA table presented in Section 2.2.2, although presenting different information. There, the residual deviance and the deviance (the difference) can be found, obtained after adding the factors Year, Month and Year:Month interaction, to the model with just one parameter (the grant mean – NULL), as well as, the corresponding degrees of freedom and p-values from the Chi-square test for the models obtained. The deviance value for the Year is a measure of the weight of factor Year in explaining precipitation variability, the deviance value for the Month is a measure of the weight of factor Month in explaining precipitation variability and the deviance value for Year:Month is an estimate of the interaction between the Year and Month, representing the lack of independence between both factors.

As expected, the results obtained for the fitted models indicate that both Year, Month factors, and Year:Month interaction are highly significant for all locations and decades, that is, all parameters referring to the Year, Month and interaction are extremely relevant in the model to explain the variability of precipitation. In particular, for Year:Month interaction higher values are obtained (2,681.63 for the case in Table 3), revealing the lack of independence between the factors Year and Month. This indicates that over the years there have been significant changes in the distribution of precipitation throughout the months of the year, otherwise the deviance for the interaction should be close to zero. The greater the value relative to the interaction deviation, the greater its importance in explaining precipitation. Given the highly significant values obtained for the Year, Month and their interaction, it was decided to carry out an in-depth analysis of the model deviances separately.

3.1. Temporal analysis 1

Aiming to evaluate the differences between the four decades per location, initially a graphical comparison of the deviances for Year:Month interaction (value 2,681.63 in case of Table 3 example) was performed. With these values the graph presented in Figure 2 was drawn, where the x-axis represents the location code ordered as shown in Figure 1.

Table 2 | ANOVA table for a two-factor model with interaction

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F statistic
A	SS_A	$r-1$	$MS_A = \frac{SS_A}{r-1}$	$F_0^A = \frac{MS_A}{MS_E}$
B	SS_B	$s-1$	$MS_B = \frac{SS_B}{s-1}$	$F_0^B = \frac{MS_B}{MS_E}$
Interaction	SS_{AB}	$(r-1)(s-1)$	$MS_{AB} = \frac{SS_{AB}}{(r-1)(s-1)}$	$F_0^{AB} = \frac{MS_{AB}}{MS_E}$
Error	SS_E	$rs(n-1)$	$MS_E = \frac{SS_E}{rs(n-1)}$	
Total	SS_T	$rsn-1$		

Table 3 | ANOVA type table for the decade 1979–1988, location 205 (Evora district)

Source	Df	Deviance	Residual Df	Residual Deviance	*p*-value
NULL			89	3,574.1	
Year	9	248.82	80	3,325.3	$< 2.2e - 16$
Month	8	643.69	72	2,681.6	$< 2.2e - 16$
Year:Month	72	2,681.63	0	0.0	$< 2.2e - 16$
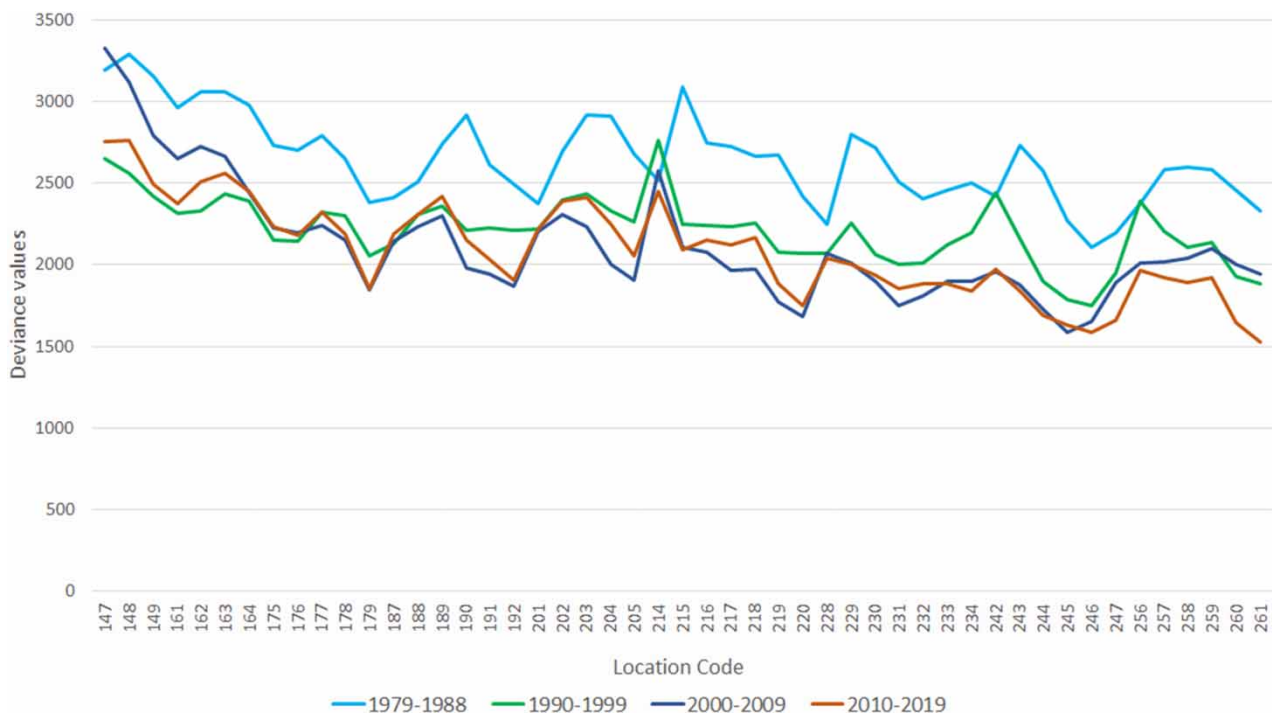
Figure 2 | Deviance values for each location and decade regarding Year:Month interaction.

As can be seen in Figure 2, the deviance for Year:Month interaction is noticeably higher in the decade starting in 1979 than the others, for all locations with two exceptions. Aiming to statistically confirm this observation, a One-Way ANOVA and Tukey multiple comparison methods were applied to these deviance values, using R. Therefore, the deviance values from models fitted to each decade and location were considered as random samples in a one-way ANOVA, where the only factor is the decade with four levels and each location is a repetition. Given that the assumptions for applying ANOVA may not be fulfilled, as said in Section 2.2.2, a balanced data situation must exist, that is, the number of observations per treatment (sample size) must be the same. This is the case, since the same number of locations per decade is considered. A balanced data situation allows still having a robust ANOVA for departures from assumptions as normality, independence between observations and non-similar variances (Scheffé 1999; Ito 2002).

The results from one-way ANOVA and Tukey methods can be seen in Tables 4 and 5. The results in Table 4 show highly significant differences between the four decades. Table 5, where all pairs of decades are compared, show that, in fact, there are significant differences between the first decade and the other three, because the first three *p*-values approximately zero.

These results mean that the year changing influenced more the distribution of precipitation across the months in the first decade than in the most recent three decades, that is, the intra-annual variability was higher in the 1980s than in the recent past. These results may sound a little strange, since the opposite could be expected. However, this may also be due to less differences in precipitation between the months in recent decades. In other words, the differences in precipitation between

Table 4 | One-way ANOVA table applied to the deviance values per decade regarding Year:Month interaction

Source	Df	Sum Sq	Mean Sq	F value	p-value
Decade	3	7,010,896	2,336,965	28.74	2.29E-14
Residuals	132	10,734,736	81,324		

Table 5 | Tukey multiple comparison table applied to the deviance values per decade regarding Year:Month interaction

decade pair	diff	lwr	upr	p-value
2-1	−478.633	−658.604	−298.663	0.000
3-1	−531.063	−711.033	−351.092	0.000
4-1	−552.325	−732.295	−372.354	0.000
3-2	−52.429	−232.400	127.541	0.873
4-2	−73.691	−253.662	106.279	0.711
4-3	−21.262	−201.233	158.709	0.990

months (intra-annual variability) may be smoothing out, making it more homogeneous from September to May. This conclusion could be in line with current common sense that differences between seasons are attenuating.

A similar analysis was performed with the deviances corresponding to the Year and Month factors, which represent the isolated contributions of the factors Year and Month for modelling the precipitation. These values are obtained from the third column of Table 3 and two graphs shown in Figures 3 and 4 were produced.

From the analysis of graph in Figure 3 (deviances for factor Year), it can be see that the decade with highest deviation is, in general, the fourth (2010–2019), followed closely by the second (1990–1999) and then by the first and third decade, which are also close to each other, but with lower deviances. This indicate that the Year changing is more important for explaining the variability of precipitation in fourth and second than in first and third decades, thus a kind alternate behaviour may be present. As a result, differences in annual precipitation between the years (inter-annual variability) are higher for fourth and second decades than for the first and third decades. To confirm statistically this behaviour, once more an one-Way ANOVA followed by a Tukey multiple comparison was applied to those deviances. Results can be found in Tables 6 and 7.
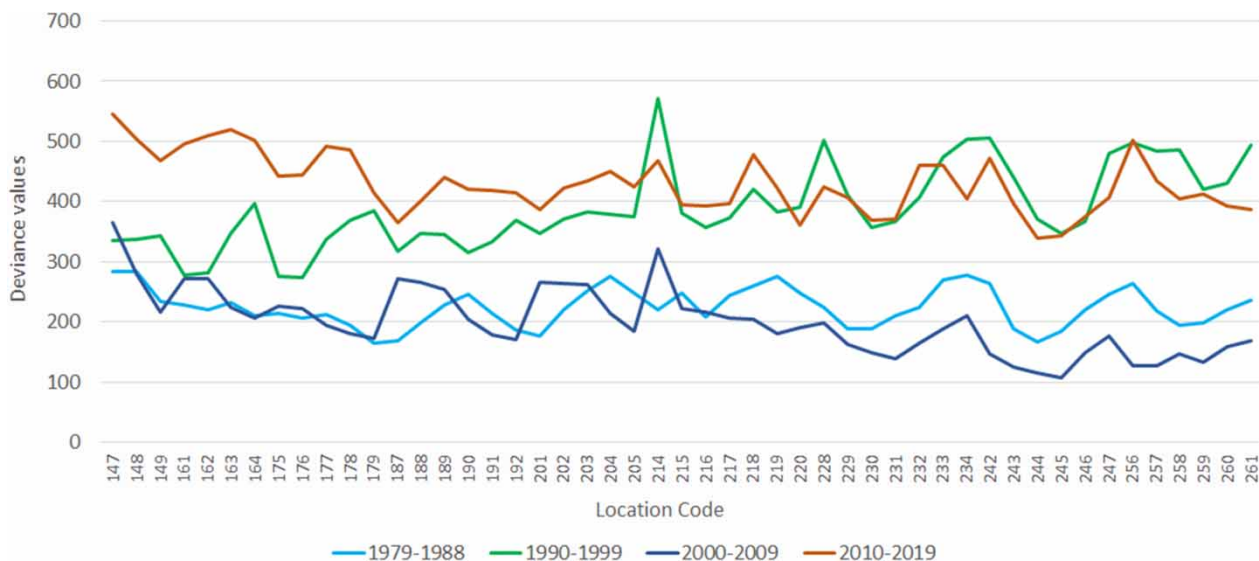

Figure 3 | Deviation values for each decade regarding factor Year.

Table 6 | One-way ANOVA table applied to the deviance values per decade regarding Year factor

Source	Df	Sum Sq	Mean Sq	F value	p-value
Decade	3	1,187,473	395,824	150.7	2E-16
Residuals	132	346,629	2,626		

Table 7 | Tukey multiple comparison table applied to the deviance values per decade regarding factor Year

decade pair	diff	lwr	upr	p-value
2-1	144.472	112.132	176.812	0.000
3-1	−7.057	−39.397	25.283	0.941
4-1	210.316	177.976	242.656	0.000
3-2	−151.529	−183.869	−119.190	0.000
4-2	65.844	33.504	98.184	0.000
4-3	217.373	185.033	249.713	0.000

The one-way ANOVA (Table 6) shows that there are significant differences between the decades and Tukey test (Table 7) shows significant differences for all pairs of decades except between third and first (p-value =0.941). These results contradict partially the possible alternate behaviour observed, since second and fourth are considered statistically different and should be similar in case of alternate behaviour. Looking to Figure 3, it can see that for the first half set of locations approximately, in fact second and fourth have very different deviances.

Finally, from the analysis of the graph presenting the deviances for factor Month (Figure 4), it can see again that the first decade has higher deviances compared with the other three, as in Figure 2, which indicates that the contribution factor
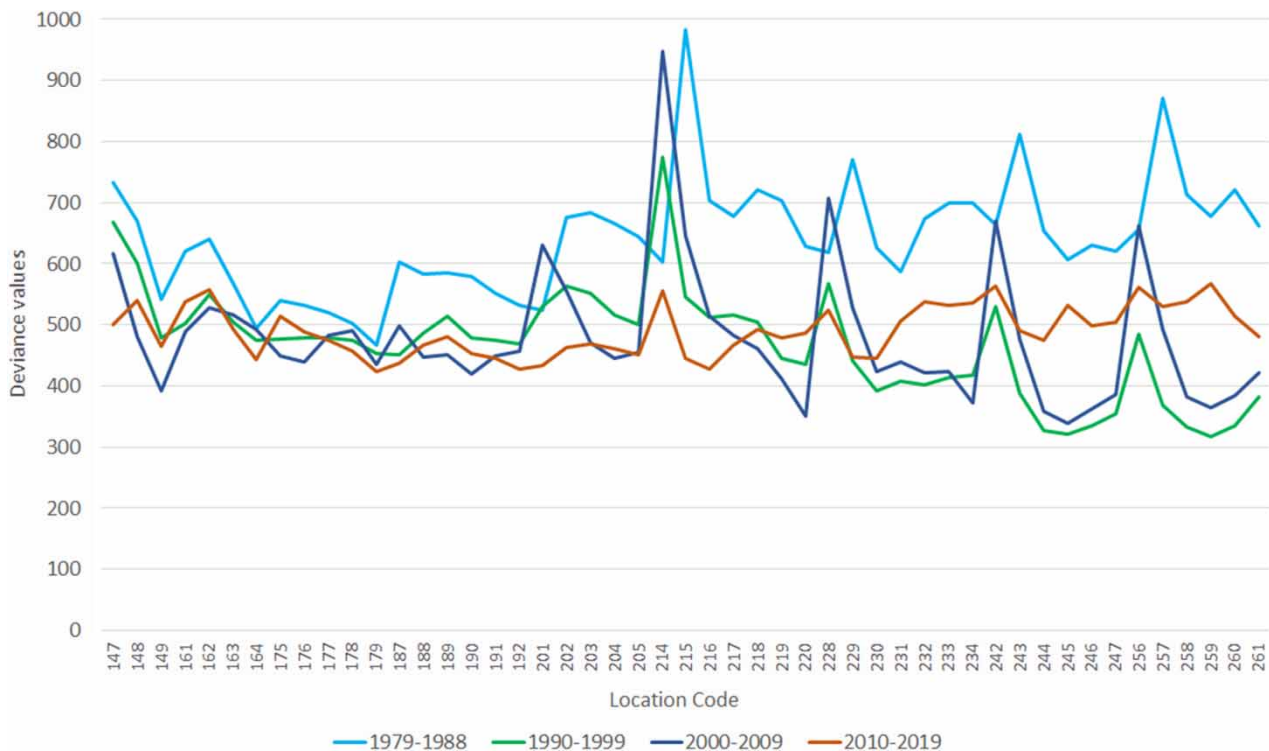
Figure 4 | Deviation values for each decade regarding factor Month.

Month to explaining precipitation variability higher in the first decade. Thus, there was more differences in precipitation between those 9 months of the year in first decade. Moreover, when comparing the deviances for factor Month with those for factor Year, the first ones are generally greater, therefore factor Month has more weight in explaining the precipitation variability occurred in each of the four decades, as a result has also more weight in the Year:Month interaction.

As previously, an one-Way ANOVA and Tukey test were performed on the deviance values per decade regarding factor Month and results showed in Tables 8 and 9, are similar to those in Tables 4 and 5 for Year:Month interaction.

In addition to what has already been said, these results seem to point a certain trend towards a decrease of the intra-annual variability of precipitation, that is, the differences in precipitation between months may be fading over in the 3 recent decades, in line with the conclusion drawn from the Year:Month interaction analysis. Furthermore, decades with less inter-annual variability seem to alternate with decades with more inter-annual variability. However, the decades with the higher inter-annual variability are the second and the fourth, which may be interpreted as decades having some years with high levels of precipitation and other with very low levels.

3.2. Temporal analysis 2

In a different analysis, the first (1979–1988) and fourth (2010–2019) decades were compared, in order to evaluate the importance that each of the nine months has in the saturated model fitted to each location. These decades were chosen to see if there were large differences in the relevance of each month between the most recent and oldest decade.

To perform that analysis, the backward elimination method mentioned in Section 2.2.1 was applied by eliminating from the models, a set at a time, the parameters relative to each month. For instance, the parameters relative to February were all considered zero, then the sub-model was refitted and tested for goodness of fit. The procedure was repeated for other months considered in the models, but all the sub-models were rejected by the chi-square test, meaning that the parameters relative to each Month are all relevant. The corresponding deviances were retained and compared between months and also between the first and last decade.

Figure 5 shows the deviance values of the models without the parameters relative to each month from January to December, in both decades. When comparing the same month between the two decades, the deviances are, in general, higher in the older decade than in the most recent one, which is in line with the results obtained in last section. Therefore, the contribution to explain precipitation variability for each particular month is higher in the elder years. The month with biggest differences between decades is December, followed March, October and January, that is, with higher changes in precipitation from first to fourth decade. This indicate a substantial decrease of the importance of December in explaining variability and less substantial decrease for March, October and January. Moreover, December is also the month with highest importance in

Table 8 | One-way ANOVA table applied to the deviance values per decade regarding factor Month

Source	Df	Sum Sq	Mean Sq	F	p-value
Decade	3	619,408	206,469	26.29	2.13E-13
Residuals	132	1,036,627	7,853		

Table 9 | Tukey multiple comparison table applied to the deviance values per decade regarding factor Month

decade pair	diff	lwr	upr	p-value
2-1	−152.806	−208.732	−96.879	0.000
3-1	−154.794	−210.720	−98.867	0.000
4-1	−159.645	−215.571	−103.719	0.000
3-2	−1.988	−57.914	53.939	1.000
4-2	−6.839	−62.766	49.087	0.989
4-3	−4.851	−60.778	51.075	0.996

Figure 5 | Deviances corresponding to each Month contribution to the models in the first (1979–1988) and last decade (2010–2019).

explaining variability in both decades, mainly in the first decade and February the one with less. On the other hand, April is the month with closest deviances in both decades, which relates with fewer changes in precipitation for this month.

3.3. Spatio-temporal analysis

In two previous analyses, the aim was focused in comparing the four decades and to see how they vary with the locations without considering their spatial coordinates. In this section, aiming a spatio-temporal analysis, Alentejo region was divided in four zones as seen in Figure 1. That division came out with the aim of assess the influence of latitude and longitude, and consequently also the ocean proximity (western zone) in opposition to inland (eastern zone). Therefore, the following zones were defined: zone 1 - north inland, zone 2 - central inland, zone 3 - south coast and zone 4 - south inland.

In a previous work Moreira *et al.* (2013), a different division of the Alentejo region was also made into four zones and as a result of applying a two-way ANOVA, it was possible to define two zones in terms of drought behaviour, one near the ocean and the other inland. From another previous work with the same aim (Moreira *et al.* 2012a), using cluster analysis of log-linear models, the opposite conclusion was achieved, i.e., the entire Alentejo could be considered a homogenous region. Therefore, when comparing these two approaches, the ANOVA technique stood out as more sensitive for determining differences between zones and should also be used in this case. The downside of ANOVA, unlike other techniques as clustering, is that different sub-regions must be defined in advance according to our beliefs, in order to test if there are significant differences between the sub-regions (zones). Moreover, since Alentejo is a small region, it does not made sense to define more than four sub-regions.

At first, the spatial maps in Figure 6 were created, where it can be seen the weight of deviances concerning Year:Month interaction per location inside each zone in the four decades. The same maps were created for the Year and Month factors, not being presented so as not to overload the article with figures. The aim was to find significant differences between the zones over decades. In order to fulfill this objective, a two-way ANOVA with interaction was performed considering the factors zone and decade, each one with four levels. To obtain a balanced ANOVA, each zone should include equal number of locations (12).

The two-way ANOVA was performed using R, considering as observations the deviance values per decade and zone relative to the Year:Month interaction, then to the Year and finally to the Month factor. Results can be found in Tables 10, 11 and 12, respectively.
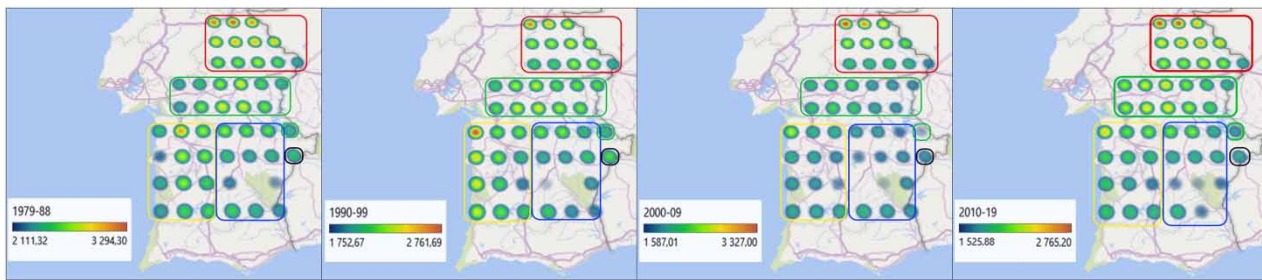
Figure 6 | Deviance weight for Year:Month interaction per location inside zone 1 (outlined in red), zone 2 (outlined in green), zone 3 (outlined in yellow) and zone 4 (outlined in blue) per decade.

Table 10 | Two-way ANOVA table applied to the deviance values per decade regarding Year:Month interaction

Source	Df	Sum Sq	Mean Sq	F value	p-value
Zone	3	6,360,742	2,120,247	42.991	2E-16
Decade	3	9,493,574	3,164,525	64.165	2E-16
Zone:Decade	9	866,273	96,253	1.952	0.0476
Residuals	176	8,680,089	49,319		

Table 11 | Two-way ANOVA table applied to the deviance values per decade regarding the factor Year

Source	Df	Sum Sq	Mean Sq	F value	p-value
Zone	3	6,857	2,286	1.27	0.286
Decade	3	1,926,723	642,241	356.86	2E-16
Zone:decade	9	194,483	21,609	12.01	1.1E-14
Residuals	176	316,748	1,800		

Table 12 | Two-way ANOVA table applied to the deviance values per decade regarding the factor Month

Source	Df	Sum Sq	Mean Sq	F value	p-value
Zone	3	142,106	47,369	8.055	0.000046454
Decade	3	934,469	311,490	52.971	2E-16
Zone:Decade	9	295,778	32,864	5.589	0.000000885
Residuals	176	1,034,956	5,880		

Concerning the Year:Month interaction (Table 10), results show that highly significant differences exist (very low p-values) between the zones indicating that Alentejo is not a homogeneous region. Results also show highly significant differences between decades, confirming the results obtained in the previous temporal analysis (Section 3.1). The interaction between factor zone and decade is also significant at 5% level (p-value = 0.0476), meaning that there were changes in the spatial distribution of the precipitation across the months over the four decades.

Paying attention to the contribution of factor Year (Table 11), results show that the differences between zones are not significant at 5% significance (p-value =0.286), but differences between decades are highly significant, again in line with the results obtained in the previous temporal analysis (Section 3.1). The interaction between factor zone and decade is also

Table 13 | Tukey multiple comparison table applied to the deviance values per decade regarding factor Month

zone pair	diff	lwr	upr	*p*-value
2-1	−7.286458	−65.3669577	50.79404	0.9880860
3-1	51.619583	−6.4609160	109.70008	0.1008012
4-1	−19.825000	−77.9054993	38.25550	0.8127592
3-2	58.906042	0.8255423	116.98654	0.0454341
4-2	−12.538542	−70.6190410	45.54196	0.9438237
4-3	−71.444583	−129.5250827	−13.36408	0.0090138

highly significant, indicating that the inter-annual variability of precipitation went through some spatial changes over the four decades.

Finaly, looking to the contribution of factor Month (Table 12), results show that the differences between zones are highly significant, which leads us to conclude again the heterogeneity of Alentejo region now regarding intra-annual variability. For this case, the Tukey test was performed to determine which zones are different (Table 13) and results show that zones 1, 2 and 4 are not different at 1% significance level, indicating that these zones may be clustered regarding the Month factor. These results lead us to conclude that two homogenous regions can be defined in Alentejo regarding intra-annual precipitation variability, one formed by zones 1, 2 and 4 (inland) and the other formed only by zone 3 close to Atlantic Ocean. This discover partially agree with the results in Moreira *et al*. (2013), where also two homogeneous regions were found, although regarding drought behaviour. The predefined regions in both works are not quite the same, but both point to the influence of ocean proximity regarding intra-annual precipitation, as well as to drought behaviour.

As expected for this case, differences between the decades are also highly significant, as well as the interaction zone: decade, that is, intra-annual variability has been changed spatially over decades, confirming again the results from temporal analysis 1.

The spatio-temporal analysis presented in this section intercepts the findings of temporal analysis 1, as it confirms the results obtained, but has the addition of providing information on the regional spatial homogeneity of the findings, as well as on the interaction between time and space.

4. CONCLUSIONS

In this research study, saturated log-linear models with two categories, Year and Month, were fitted to contingency tables accounting the number of mm^3 of precipitation estimated for each month during periods of 10 years (four decades) in Alentejo region. The model's, deviances for the factor Year, Month and Year:Month interaction were then used as observations to be analysed for significant differences between decades using a balanced ANOVA and Tukey multiple comparisons methods.

At first the analysis aimed just to be temporal through comparing the four decades. Then four zones were predefined in Alentejo aiming a spatio-temporal analysis, which was performed applying a two-way balanced ANOVA, considering decade and zone as factors each with 4 levels.

Interpreting the results presented a challenge, yet it enabled us to glean some valuable findings. First, significantly differences in precipitation distribution across the months over the years were found between the first decade and the three more recent ones. These findings suggest that differences in precipitation between months (intra-annual variability) may be smoothing out, since they are less accentuated in recent decades. Although these changes may be due to global climate change, there is no way to prove it. In fact, other factors such as teleconnections, solar cycles, anthropogenic activities, may be influencing it.

Regarding the effect of the year changing, i.e., inter-annual variability within a decade, results indicate that the differences in precipitation distribution between the years of the second and fourth decades are higher than between the years of the first and third decades, showing an alternating behaviour, where a decade with less precipitation variability between the years is followed by a decade with more precipitation variability between the years, although with partial statistical significance. Furthermore, it also stands out that the source of variability coming from the factor Month has more weight in explaining overall precipitation variability than the source coming from the factor Year.

With a different objective, the first and last decades were compared in order to evaluate the importance of each month in explaining precipitation variability. The first and immediate conclusion was that, for both decades, no month could be discard from the model. All the months considered (September to May) are very important to explain precipitation variability. However, February seems to be the one with lower importance and December with higher, in both decades. The month with the biggest decrease between the first and fourth decades is also December, followed by March, October and January, meaning a substantial decrease of the importance of December in explaining variability and less substantial decrease for March, October and January. On the other hand, April is the month with closest deviances in both decades, which relates to fewer changes in precipitation for this month.

To close findings, the spatio-temporal analysis allowed to conclude that Alentejo is not a homogeneous region regarding precipitation intra-annual variability, since distribution of precipitation across months has undergone changes over decades differently, in different sub-regions of Alentejo. Furthermore, two homogeneous regions could be defined in Alentejo regarding intra-annual precipitation variability, a large interior region and a smaller one close to the Atlantic Ocean.

The statistical techniques chosen to perform temporal and spatio-temporal analysis are well known. However, from our point of view, although unusual, the approach used is new and proved to be useful in obtaining some findings about the possible effects of climate change in the Alentejo region.

In terms of future work, we intend to deepen the analysis and try to find other types of changes in annual precipitation cycles that were not possible to achieve with the present approach; namely, whether there is a change in annual and seasonal precipitation cycles. We would also like to carry out a more in-depth analysis of the locations, namely those close to the Alqueva Dam (the largest artificial lake in Europe, located in Alentejo) to find some evidence of its influence on the climate of these locations. We also intend to carry out a similar study on the maximum and minimum temperatures recorded above and below critical values, to find significant changes in the distribution of temperatures across the months over the past 40 years.

ACKNOWLEDGEMENTS

DATA AVAILABILITY STATEMENT

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

Agresti, A. (2002) *An Introduction to Categorical Data Analysis*. Hoboken, New Jersey: John Wiley & Sons.

Benetó, P. & Khodayar, S. (2023) On the need for improved knowledge on the regional-to-local precipitation variability in eastern Spain under climate change, *Atmospheric Research*, **290**, 1–17. https://doi.org/10.1016/j.atmosres.2023.106795.

Ceballos, A., Martínez-Fernández, J. & Luengo-Ugidos, M. Á. (2004) Analysis of rainfall trends and dry periods on a pluviometric gradient representative of Mediterranean climate in the Duero Basin, Spain, *Journal of Arid Environments*, **58** (2), 215–233. https://doi.org/10.1016/j.jaridenv.2003.07.002.

Droulia, F. & Charalampopoulos, I. (2021) Future climate change impacts on European viticulture: a review on recent scientific advances, *Atmosphere*, **12** (4), 495. https://doi.org/10.3390/atmos12040495.

Dunn, P. K. & Smyth, G. K. (2018) *Generalized Linear Model With Examples in R*. Springer: New York.

Everitt, B. S. (1977) *The Analysis of Contingency Tables*. London: Chapman & Hall.

Fraga, H., Atauri, I., Malheiro, A., Moutinho-Pereira, J. & Santos, J. (2017) Viticulture in Portugal: a review of recent trends and climate change projections, *OENO One*, **51** (2), 61–69. https://doi.org/10.20870/oeno-one.2017.51.2.1621.

Ito, K. (2002) 7 Robusteness of ANOVA and MANOVA test procedures. In: Krishnaiah, P. R. (ed.) *Handbook of Statistics*, vol. 1. pp. 199–236.

Lima, M. I. P., Carvalho, S. C. P., Lima, J. L. M. P. & Coelho, M. F. E. S. (2010) Trends in precipitation: analysis of long annual and monthly time series from mainland Portugal, *Advances in Geosciences*, **25**, 155–160. https://doi.org/10.5194/adgeo-25-155-2010.

Lima, D. C. A., Bento, V. A., Lemos, G., Nogueira, M. & Soares, P. M. M. (2023) A multi-variable constrained ensemble of regional climate projections under multi-scenarios for Portugal - Part I: an overview of impacts on means and extremes, *Climate Services*, **30**, 100377. https://doi.org/10.1016/j.cliser.2023.100351.

Martins, D. S., Raziei, T., Paulo, A. A. & Pereira, L. S. (2012) Spatial and temporal variability of precipitation and drought in Portugal, *Natural Hazards and Earth System Sciences*, **12**, 1493–1501.

Montegomery, D. C. (2013) *Design and Analysis of Experiments*. John Wiley & Sons, Singapore.

Moreira, E., Paulo, A., Pereira, L. & Mexia, J. (2006) Analysis of SPI drought class transitions using log-linear models, *Journal of Hydrology*, **331**, 349–359. https://doi.org/10.1016/j.jhydrol.2006.05.022.

Moreira, E., Mexia, J. T. & Pereira, L. S. (2012a) Clustering of loglinear models using LRT *p*-values to assess homogeneous regions relative to drought class transitions, *Journal of Statistical Computation and Simulation*, **82** (2), 293–308. https://doi.org/10.1080/00949655.2011.640680.

Moreira, E., Mexia, J. T. & Pereira, L. S. (2012b) Are drought occurrence and severity aggravating? A study on SPI drought class transitions using log-linear models and ANOVA-like inference, *Hydrology and Earth System Sciences*, **16**, 3011–3028.https://doi.org/10.5194/hess-16-3011-2012.

Moreira, E., Mexia, J. & Pereira, L. (2013) Assessing homogeneous regions relative to drought class transitions using an ANOVA-like inference. Application to Alentejo, Portugal, *Stochastic Environmental Research and Risk Assessment*, **27** (1), 183–193. https://doi.org/10.1007/s00477-012-0575-z.

Moreira, E., Pires, C. & Pereira, L. (2016) SPI drought class prediction driven by the Northern Atlantic Oscillation index using log-linear modelling, *Water*, **43** (8), 1–18. https://doi.org/10.3390/w8020043.

Moreira, E., Russo, A. & Trigo, R. (2018) Monthly prediction of drought classes using log-linear models under the influence of NAO for early-warning of drought and water management, *Water*, **65** (10), 1–24. https://doi.org/10.3390/w10010065.

Paulo, A. A. & Pereira, L. S. (2006) Drought concepts and characterization. Comparing drought indices applied at local and regional scales, *Water International*, **31**, 37–49. https://doi.org/10.1080/02508060608691913.

Paulo, A. A., Pereira, L. S. & Matias, P. (2003) Analysis of local and regional droughts in southern Portugal using the theory of runs and the Standardized Precipitation Index. In: Rossi, G., Cancelliere, A., Pereira, L. S., Oweis, T., Shatanawi, M., & Zairi, A. (eds.) *Tools for Drought Mitigation in Mediterranean Regions*. Dordrecht: Kluwer, pp. 55–78.

Santos, J. A., Fraga, H., Malheiro, A. C., Moutinho-Pereira, J., Dinis, L. T., Correia, C., Moriondo, M., Leolini, L., Dibari, C., Costafreda-Aumedes, S. & Kartschall, T. (2020) A review of the potential climate change impacts and adaptation options for European viticulture, *Applied Sciences*, **10**, 3092. https://doi.org/10.3390/app10093092.

Scheffé, H. (1999) *The Analysis of Variance*. John Wiley & Sons, Hoboken, NJ.

Senent-Aparicio, J., López-Ballesteros, A., Jimeno-Sáez, P. & Pérez-Sánchez, J. (2023) Recent precipitation trends in Peninsular Spain and implications for water infrastructure design, *Journal of Hydrology*, **45**, 101308. https://doi.org/10.1016/j.ejrh.2022.101308.

van den Dool, H. (2006) *Empirical Methods in Short-Term Climate Prediction*. Oxford University Press, Oxford, UK.

Wan, Q., Li, L., Liu, B., Xie, M. & Zhang, Z. (2024) Altered intra-annual precipitation patterns affect the N-limitation status of soil microorganisms in a semiarid alpine grassland, *Ecological Indicators*, **158**, 111457. https://doi.org/10.1016/j.ecolind.2023.111457.

Wilks, D. S. (2011) *Statistical Methods in the Atmospheric Sciences*, 3rd edn. International Geophysics, Academic Press, Oxford, UK.

Wunderlich, R. F., Lin, Y. & Ansari, A. (2023) Regional climate change effects on the viticulture in Portugal, *Environments*, **10** (1), 5. https://doi.org/10.3390/environments10010005.

Zeleke, T., Lukwasa, A., Beketie, K. & Ayal, D. (2024) Analysis of spatio-temporal precipitation and temperature variability and trend over Sudd-Wetland, Republic of South Sudan, *Climate Services*, **34**, 100451. https://doi.org/10.1016/j.cliser.2024.100451.