Physics and Chemistry of the Earth 94 (2016) 22–28

Contents lists available at ScienceDirect

Physics and Chemistry of the Earth

journal homepage: www.elsevier.com/locate/pce

Regionalization of Europe based on a *K*-Means Cluster Analysis of the climate change of temperatures and precipitation

M.J. Carvalho*, P. Melo-Gonçalves, J.C. Teixeira, A. Rocha

Department of Physics & CESAM, University of Aveiro, Aveiro, Portugal

ARTICLE INFO

Article history:
Received 2 March 2015
Received in revised form
23 December 2015
Accepted 18 May 2016
Available online 24 May 2016

Keywords:
Climate change
Surface temperatures
K-Means Clustering
Precipitation
Europe

ABSTRACT

In order to study climate change on a regional scale using Earth System Models, it is useful to partition the spatial domain into regions according to their climate changes. The aim of this work is to divide the European domain into regions of similar projected climate changes using a simulation of daily total precipitation, minimum and maximum temperatures for the recent-past (1986–2005) and long-term future (2081–2100) provided by the Coupled Model Intercomparison Project (CMIP5). The difference between the long-term future and recent-past daily climatologies of these three variables is determined. Aiming to objectively identify the grid points with coherent climate changes, a *K*-Mean Cluster Analysis is applied to these differences. This method is performed for each variable independently (univariate version) and for the aggregation of the three variables (multivariate version). A mathematical approach to determine the optimal number of clusters is pursued. However, due to the method characteristics, a sensitivity test to the number of clusters is performed by analysing the consistency of the results. This is a novel method, allowing for the determination of regions based on the climate change of multiple variables. Results from the univariate application of this method are in accordance with results found in the literature, showing overall similar regions of changes. The regions obtained for the multivariate version are mainly defined by latitude over European land, with some features of land-sea interaction. Furthermore, all regions have statistically different distributions of at least one of the variables, providing confidence to the regions obtained.

1. Introduction

Climate change studies are usually carried out either globally or regionally. Either way, they usually focus on areas with different climate characteristics and large variability. In Europe, temporal variability of daily surface climate variables (such as minimum and maximum temperatures and precipitation) has high spatial gradients. Therefore, statistics of the temporal behaviour of a particular variable or its derived quantities over the target domain 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 grid point of the domain, that are impossible to be displayed 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 domain. This, together with the large amount of data, adds up to the need to define regions to be analysed using either grid point as a representation of the region or the average behaviour of the grid points.

In an attempt to divide the overland areas of the globe into a manageable number of regions, each with simple shape and representing a different climatic regime, several authors (such as Sillmann et al. (2013, 2014)) have followed the approach of Giorgi and Francisco (2000). When studying the uncertainty in regional climate change prediction using ensemble simulations from a coupled Atmosphere-Ocean Global Climate Model (AOGCM), they proposed a division of the domain, creating rectangle-like overland areas, admitting, however, that this was a subjective approach to the issue. On a regional scale, the simple use of geographic markers has been extensively used in order to define regions. For example, in their study of European heat waves in present-day and future climates, Lau and Nath (2014) simply divided the domain into three regions: Russia, eastern Europe and western Europe. Much like this, when studying record high maximum and low minimum temperatures, Meehl et al. (2009) used the 100°W meridian to divide the

* Corresponding author.
E-mail address: mariajcarvalho@ua.pt (M.J. Carvalho).

http://dx.doi.org/10.1016/j.pce.2016.05.001
1474-7065/

United States of America into eastern and western USA. Fischer and Schär (2009) simply use the Iberian Peninsula, Scandinavia and France as key regions when studying the PRUDENCE (Christensen and Christensen (2007)) regional climate model scenarios for temperature and the driving processes in temperature extremes while Wójcik (2015) divides Poland taking into account some orographic characteristic in order to study the reliability of CMIP5 simulations in reproducing atmospheric circulation. An upgrade to this methodology is the approach of Fischer et al. (2014) who segregate grid points based on their altitude, in order to study projected changes in precipitation intensity and frequency in Switzerland. Locally, changes in precipitation in the Iberian Peninsula have been studied by Gonzalez-Hidalgo et al. (2010) following the regions defined by hydrological basins, which is relates directly to water management.

Due to the subjectivity of the regionalization methodologies described above, several authors have pursued a more objective approach. For example, Richman and Lamb (1985, 1987) used Principal Component Analysis (PCA) in order to study the spatial distribution of three- and seven-day rainfall events for the USA and Canada, respectively. In a later work, White et al. (1991) applied Rotation Principal Analysis to monthly precipitation in Pennsylvania using different rotation algorithms in order to assess the sensitivity of the regionalization result to rotation. They found that the resulting regions varied widely with the use of different rotation algorithms.

In addition to having been used to identify Weather Types (or Weather Regimes) by Santos et al. (2005), Cluster Analysis has also been used to define climatic regions. DeGaetano and Shulman (1990) applied a flexible clustering technique to the first three principal components of several climatological variables in order to identify regions of coherent plant hardiness. Much like this, Fovell and Fovell (1993) used *K*-Means Cluster Analysis in order to identify climatic zones of the Conterminous United States based on both temperature and precipitation.

Despite being objective, these studies regionalize the domain using observed (or modelled) data for a given period and therefore obtain regions of coherent climate. In a changing climate, the main interest becomes knowing the regions of coherent changes, instead of the definition of regions with the same climate characteristics, since they can change in time. The main goal of this work was to identify regions with consistent climate changes in precipitation and surface minimum and maximum temperature seasonal cycles in Europe. This is a novel method, allowing for the determination of regions based on the climate change of multiple variables.

2. Method and data

The data used was provided by the Coupled Model Intercomparison Project Phase 5 (CMIP5), simulated by the MPI-ESM-LR model with the r1i1p1 initialisation, with a horizontal resolution of 1.9° horizontal resolution (Giorgetta et al., 2013). As a representation of the recent-past climate, the last 20 years (1986–2005) of the historical experiment which runs from 1850 to 2005 were used. The future climate used was simulated by using the 8.5 Representative Concentration Pathway (RCP8.5) which stabilises radiative forcing at 8.5 $W \cdot m^{-2}$ in the year 2100 without exceeding that value (Riahi et al., 2011). From the future climate experiment, which runs from 2006 to 2100, only the long-term future (from 2081 to 2100) was used in this work, since changes for this period relative to the recent-past are expected to be greater than those found for both the near-term (2016–2035) and mid-term (2046–2065) periods. The variables used were the daily minimum and maximum near-surface air temperatures as well as total daily precipitation, which includes both liquid and solid phases from all types of clouds (both large scale and convective). The simulations are available for the entire globe. However, this study focused on a domain containing only Europe: 25 °N–70 °N, 45 °W–65 °E. This domain is presented in Fig. 1 where the model grid points corresponding to the aforementioned resolution are also plotted.

Taking each of the two climates – recent-past and long-term future – daily climatologies of each of the variables were determined, for each of the domain grid points, using a 15 day-running window as a low frequency filter. Taking each grid point, the difference between the recent-past and long-term future climatologies were determined, creating a measure of the changes in the seasonal cycle.

The challenge was then identifying grid points where the changes in the seasonal cycle of the variables are similar. To the difference fields, the *K*-Means Cluster Analysis was applied. This is a non-hierarchical clustering method which starts by computing the centroids for each cluster and then calculates the distances between the current data vector and each of the centroids, assigning the vector to the cluster whose centroid is closest to it. Since this is a dynamic method, meaning that vectors can change cluster after being assigned to it, this process is repeated until all vectors are assigned a cluster and their members are closest to the centroid than to the mean of other clusters (Wilks, 2011). The mathematical condition for the cluster C_k and the k centroids μ_k can be expressed as Equation (1).

$$\textit{Minimize} \quad \sum_{k=1}^{K} \sum_{x_n \in C_k} \|x_n - \mu_k\|^2 \quad \textit{with respect to } C_k,\ \mu_k \tag{1}$$

It is important to note that, unlike when applied to determine weather types, the *K*-Means Clustering is done in space and not time, resulting in each grid point (instead of a time step) being assigned to a cluster.

The described methodology was applied to the climatology differences of each of the three variables independently (univariate version) and using the daily climatology differences of a synthetic joint variable composed by concatenating the temporal-varying spatial fields of the three variables (multivariate version). Since the goal is to determine one set of regions on which to base the analysis, the univariate version is only used to, once more, analyse the consistency of the multivariate results and ultimately validate them, since this is a novel statistical approach.

A major drawback of applying a clustering technique is the need to choose a priori the number of clusters (and consequently, in this case, regions). Therefore, and in order to determine the optimal number of clusters, the Gap Statistics was used as described by Pham et al. (2005). As discussed in their paper, this statistics uses the distortion of the cluster and is determined as follows Equation (2):

$$f(k) = \begin{cases} 1 & \text{if } \mathrm{K} = 1 \\ \dfrac{S_K}{\alpha_K S_{K-1}} & \text{if } S_{K-1} \neq 0,\ \forall K > 1 \\ 1 & \text{if } S_{K-1} = 0,\ \forall K > 1 \end{cases} \tag{2}$$

with α_K defined as in Equation (3):

$$\alpha(K) = \begin{cases} 1 - \dfrac{3}{4N_d} & \text{if } K = 2 \text{ and } N_d > 1 \\ \alpha_{K-1} + \dfrac{1 - \alpha K - 1}{6} & \text{if } K > 2 \text{ and } N_d > 1 \end{cases} \tag{3}$$

where S_k is the sum of the cluster distortion when the number of