Non-parametric multimodel Regional Frequency Analysis applied to climate change detection and attribution

Philomène Le Gall, Anne-Catherine Favre, Philippe Naveau, Alexandre Tuel

Non-parametric multimodel Regional Frequency Analysis applied to climate change detection and attribution

Philomène Le Gall

Univ. Grenoble Alpes, CNRS, IRD, Grenoble INP, IGE, F-38000 Grenoble, France

E-mail: philomene.le-gall@univ-grenoble-alpes.fr

Anne-Catherine Favre

Univ. Grenoble Alpes, CNRS, IRD, Grenoble INP, IGE, F-38000 Grenoble, France

Philippe Naveau

Laboratoire des Sciences du Climat et de l'Environnement, ESTIMR, CNRS-CEA-UVSQ, Gif-sur-Yvette, France

Alexandre Tuel

Institute of Geography and Oeschger Centre for Climate Change Research, University of Bern, Switzerland

Summary. A recurrent question in climate risk analysis is determining how climate change will affect heavy precipitation patterns. Dividing the globe into homogeneous sub-regions should improve the modelling of heavy precipitation by inferring common regional distributional parameters. In addition, in the detection and attribution (D&A) field, biases due to model errors in global climate models (GCMs) should be considered to attribute the anthropogenic forcing effect. Within this D&A context, we propose an efficient clustering algorithm that, compared to classical regional frequency analysis (RFA) techniques, is covariate-free and accounts for dependence. It is based on a new non-parametric dissimilarity that combines both the RFA constraint and the pairwise dependence. We derive asymptotic properties of our dissimilarity estimator, and we interpret it for generalised extreme value distributed pairs.
As a D&A application, we cluster annual daily precipitation maxima of 16 GCMs from the coupled model intercomparison project. We combine the climatologically consistent subregions identified for all GCMs. This improves the spatial clusters coherence and outperforms methods either based on margins or on dependence. Finally, by comparing the natural forcings partition with the one with all forcings, we assess the impact of anthropogenic forcing on precipitation extreme patterns.

Keywords: CMIP; detection & attribution; extreme precipitation; F-madogram; Regional Frequency Analysis

1. Introduction

Since the early 19^{th} century, fossil fuels-based human activities have become one of the major forces of ecosystem and climate change, defining a new geological era, called *Anthropocene* (Crutzen, 2006) or *Capitalocene* (Malm and Hornborg, 2014; Campagne, 2017). The global warming caused by these activities induces important changes in

the climate system (IPCC, 2021). Working Group I of the IPCC, which assesses the physical science of climate change, summarises the latest advances in climate science to understand the climate system and assess climate change, by combining data from paleoclimate, observations and global circulation model (GCM) simulations. The latter are based on differential equations linked to the fundamental laws of physics, thermodynamics and chemistry. GCMs simulate the evolution of various climate variables on discretised tridimensional meshes with a typical horizontal resolution of 100 [km] or more. The coupled model intercomparison project (CMIP) (Meehl et al., 2000; Alexander and Arblaster, 2017) aims at comparing the performances of several dozen of GCMs developed by different research centres, e.g. see Table 1 in Appendix. As numerical experiments and approximations of the true climate system, these GCMs can produce different climate responses to different given inputs, e.g. emission scenarios. To reduce model errors and gain robustness in signal detection, GCMs are often analysed jointly. In particular, CMIP models have been used in the field of detection and attribution that aims at finding causual links between the climate response and known external forcings (see, e.g. Ribes et al., 2021; van Oldenborgh et al., 2021; Naveau et al., 2020). As a yardstick, the so-called "natural forcings" runs have not been influenced by human activities and were only driven by external forcings, *e.g.* solar variations, explosive volcanic eruptions like Mont Pinatubo in 1991 (see, e.g. Ammann and Naveau, 2010). Such a numerical setup can be viewed as a thought-experiment and it corresponds to a counterfactual world, but not to the observed one. In contrast, a factual world is produced by integrating all forcings, including rising greenhouse gazes, and factual runs aims at reproducing the observed climatology over the last century. Future periods, say 2071–2100, can also be explored with GCMs but future forcing and emission scenarios need to be chosen. For example, `RCP8.5` for CMIP5 (IPCC, 2013) and `SSP5-8.5` for CMIP6 (IPCC, 2021) will be analysed in this paper. In this context, a natural question is to wonder how the climate system will change under these scenarios.

Due to their large societal and economical impacts, a vast literature has be dedicated to answering this question for extreme events. In particular, heavy rainfall and heatwaves have received a particular attention, see chapters 10 and 11 in the Working Group I contribution of IPCC (2021) report. In this paper, we focus on annual maxima of daily precipitation from 1850 to 2100 provided by the factual (all forcings) and counterfactual (natural forcings only) models listed in Table 1 of the Appendix. Note that our main climatological goal is not to directly assess changes in heavy rainfall intensities and frequencies, but rather to detect how spatial patterns (clusters) of yearly maxima of daily precipitation could be modified by anthropogenic forcing.

To model yearly block maxima, one classical statistical approach is to impose a parametric generalised extreme value (GEV) distributions (see *e.g.* Coles et al., 2001; Davison et al., 2012). For example, each grid point of each individual CMIP model could be fitted with a spatial structure embedded within the GEV parameters (see, e.g. Kharin et al., 2013). However, the computational cost can be high (more than 200 years of precipitation data at thousands of grid points for 16 models), especially if the spatial dependence is included. Another aspect is the ease of interpretation. Well defined spatial patterns (clusters) in extreme precipitation are very useful for climatologists who can interpret them according to known physical phenomena (e.g., Pfahl et al., 2017;

Tandon et al., 2018; Dong et al., 2021). For example, the so-called regional frequency analysis (RFA) has been frequently used in hydrology, see Dalrymple (1960); Hosking and Wallis (2005), but it has been rarely implemented in a D&A context, especially within the CMIP repository. The main idea of RFA is to identity homogeneous regions with identical distributional features, up to normalising constants. More precisely two positive absolutely continuous random variables (r.v.) Y_1 and Y_2 are said to be homogeneous if there exists a positive constant λ such that

$$Y_2 \overset{\mathrm{d}}{=} \lambda Y_1,$$

where $\overset{\mathrm{d}}{=}$ denotes equality in distribution. This condition can be reformulated in terms of their cumulative distribution functions (cdf) $F_i(x) = \mathbb{P}(Y_i \leq x)$ with $i \in \{1, 2\}$ as

$$F_2(\lambda x) = F_1(x). \tag{1}$$

Hence, two climate model grid points are said to belong to the same homogeneous region if they satisfy (1). To visually understand this condition within the CMIP archive, three grid points, say A, B and C, from the CCSM4 counterfactual run are plotted in panel (a) of Figure 1. In panel (b), ranked annual precipitation maxima (rescaled by the empirical mean) of point A are compared to the ones from point B. Panel (d) provides the same information but between point A and point C. It appears that points A and B are likely to satisfy (1) and, consequently, could belong to the same homogeneous region. In contrast, the rescaled distribution at point A is much more heavy-tailed than at point C. This is not surprising because A and B are nearby and C far away from them. Still, panels (b) and (d) only rely on the marginal behaviours, and pairwise dependence information and/or covariates could help finding of homogeneous regions.

Various RFA techniques based on explanatory covariates (e.g., see Asadi et al., 2018; Fawad et al., 2018, for recent work) have been developed to identify homogeneous regions which rely on station location features and/or weather patterns to explain precipitation spatial distributions (see *e.g.* Burn, 1990; Hosking and Wallis, 2005; Evin et al., 2016). For example, Toreti et al. (2016) let scale parameters vary as a function of weather station locations. However, selecting relevant covariates is constrained by their availability, expert subjectivity and the scale of the problem. In particular, finding appropriate covariates for heavy rainfall patterns at the global scale is tedious. In addition, assessing the homogeneity of regions (Hosking and Wallis, 2005) relies on specific moments like skewness and kurtosis that are not necessary robust (based on the spatial independence assumption). Other techniques bypass the use of covariates by only working with the data at hand, here precipitation (Saf, 2009). For example, Le Gall et al. (2021) considered a ratio of probability weighted moments, see Greenwood et al. (1979) and applied a clustering algorithm on this ratio. More precisely, this ratio, denoted $\omega \in [0, 1]$, is mean and scale invariant, i.e. in compliance with (1), and it is a simple increasing function of ξ when rainfall extremes can be assumed to either follow a GEV or Pareto distribution with shape parameter ξ. To illustrate the spatial variability of CMIP rainfall tail index (i.e. of ω), panel (a) of Figure 2 displays the ratio ω at each grid point of a counterfactual `CCSM4` annual maxima run. Note that grid points A and B exhibit similar ω estimates, while grid point C differs (lighter tail).

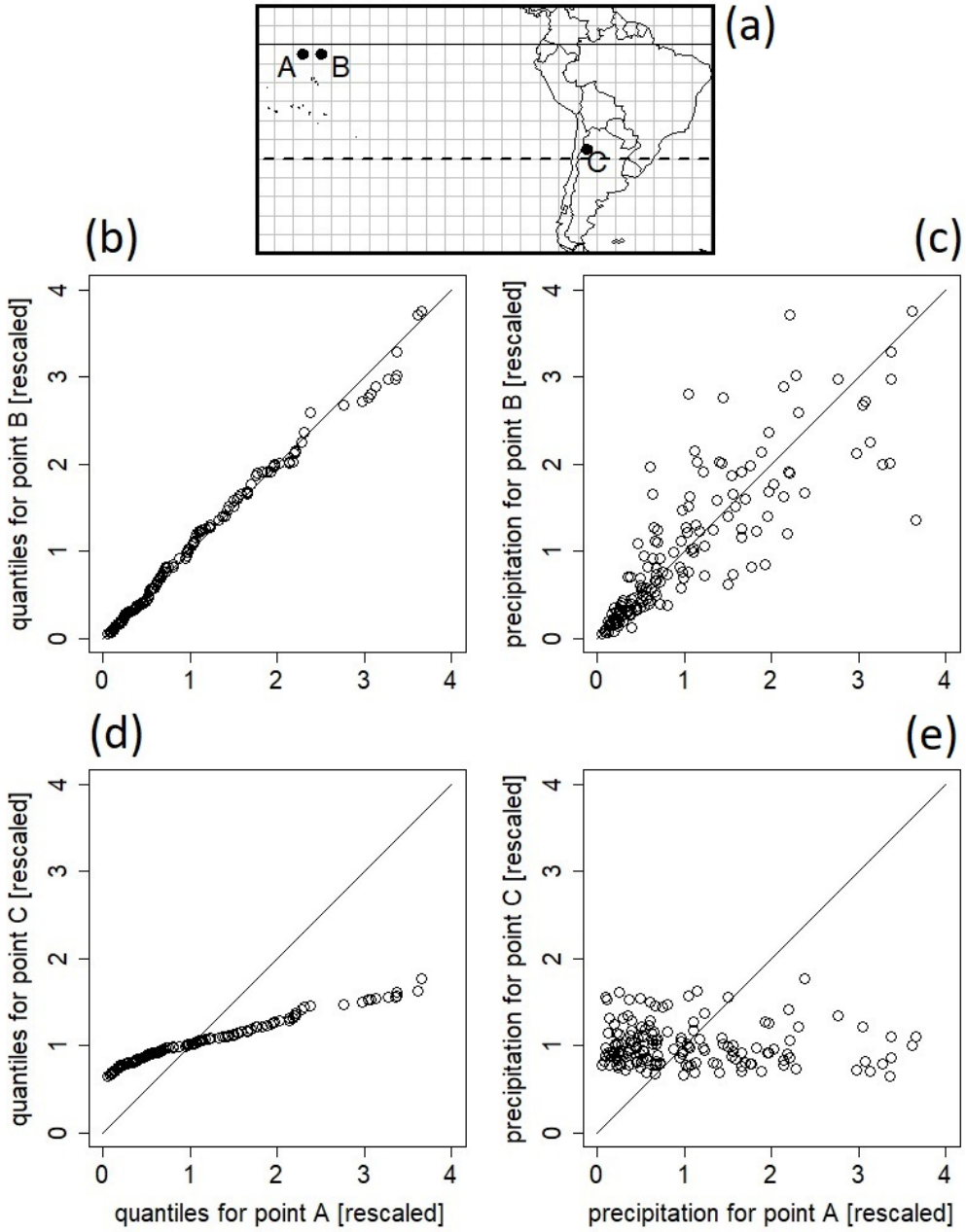

Fig. 1. Localisation (a), QQ-plots (b) and (d) and scatter plots (c) and (e) of annual precipitation maxima at three grid points A,B and C in the counterfactual run of the CCSM4 model (1850–2005). Panels (b) and (d) show the QQ-plots of rescaled precipitation for pairs (A,B) (b) and (A,C) (d). Panels (c) and (e) display the (rescaled) scatter plots for the same pairs.

All aforementioned RFA techniques has one major drawback. They rely on the assumption of pairwise independence or pairwise conditional independence (given the co-

variates). Note that Eq.(1) also constraints the marginal behaviour, but does not take into account of any information about the spatial dependence strength. Still, precipitation series at two nearby grid points are likely to be dependent. To illustrate this point, we can go back to Figure 2. Panels (b) and (e) display the scatter plots (rescaled by their means) between points A and B, and between points A and C, respectively. As expected from their local proximity, not only A and B have same similar marginals, but annual maxima of daily precipipation appears to be strongly correlated. This information coupled with constraint (1) should play an important role in improving RFA methods.

Modelling the dependence structure in clustering algorithms can be handled in different ways depending on the assumptions one is ready to make. Fully non-parametric or parametric approaches can be developed. Explanatory covariates can be included or difficult to find. For example, Kim et al. (2019) introduced a parametric approach based on copulas in the context of cluster detection in mobility networks. They grouped sites subject to intense traffic according to covariates (*e.g.* geographical), and checked the dependence strength within each cluster by fitting a multivariate Gumbel copula. Drees and Sabourin (2021) and Janßen et al. (2020) proposed approaches based on exceedances; after projecting observations onto the unit sphere, they reduced their dimension through K-means clustering (Janßen et al., 2020) and principal component analysis (Drees and Sabourin, 2021). Finally, Bernard et al. (2013) applied a non-parametric approach based on the F-madogram to weekly precipitation maxima. The so-called F-madogram (Cooley et al., 2006) is defined by

$$d = \frac{1}{2}\mathbb{E}\left|F_1\left(Y_1\right) - F_2\left(Y_2\right)\right|, \tag{2}$$

where Y_i is the continuous r.v. with cdf F_i. It is a distance which, by construction, is marginal-free because the r.v. $F_1\left(Y_1\right)$ and $F_2\left(Y_2\right)$ are both uniformly distributed on $[0, 1]$. Note that if Y_1 and Y_2 are equal in probability, the distance $d = 0$. Whenever the bivariate vector (Y_1, Y_2) follows a bivariate GEV distribution (see *e.g.* Gumbel, 1960; Tawn, 1988), this distance can be interpreted as linear transformation of the extremal coefficient (see *e.g.* Cooley et al., 2006; Naveau et al., 2009, and Section 2.2). Bernard et al. (2013), Bador et al. (2015) and later Saunders et al. (2021) computed this distance to build a pairwise dissimilarity matrix that was used as an input of a clustering algorithm. In these two former studies, a partitioning around medoids (`PAM`) algorithm (Kaufman and Rousseeuw, 1990) was applied whereas the latter used hierarchical clustering. But, the RFA requirement defined by (1) was not imposed, and so the marginal differences between Y_1 and Y_2 were not taken into account. To visualise this issue within the CMIP repository, it is simple to cluster a counterfactual `CCSM4` annual maxima run with the `PAM` algorithm† based on the distance d. The resulting map displayed in panel (b) of Figure 2 shows a few spatially coherent structures, but, overall is very patchy. In

†In all our CMIP analysis, `PAM` was applied separately to the southern and northern hemispheres. Global analysis (available upon request) were also made, but the climatological interpretation was not as clear as with the hemispheric scale. Also, different numbers of clusters were investigated and basic criteria like the silhouette coefficient were computed. No particular number could be clearly identified. But, in terms of interpretation, four clusters appear as a reasonable compromise between climate understanding, visual simplicity and statistical criteria.

addition, panel (a) related to the marginals behaviour appears to be unrelated to panel (b) that describes the spatial dependence. This was expected from the F-madogram distance, but it would make sense to cluster grid points that are both correlated but also the same type of marginal, see (1), the essence of the RFA.

To reach this goal, we propose the following work plan. In Section 2, we integrate the homogeneity condition (1) into a new definition of the F-madogram distance. The properties of this new dissimilarity, which we call RFA-madogram, is explained by analysing a special case: the logistic bivariate GEV model in Section 2.2. A non-parametic estimator of the RFA-madogram is proposed and its asymptotic consistency in law is detailed in Section 3. Concerning the CMIP database, we compute, in Section 4, a RFA-madogram dissimilarity matrix on annual maxima of daily precipitation for each CMIP models listed in Table 1, and then cluster them with the `PAM` algorithm. Finally, we propose a method to build a "central" partition that summarises the partitions obtained for each model and compare the spatial patterns obtained for counterfactual (1850–2005) and factual (2071–2100) experiments. Section 5 concludes the paper by providing a short discussion.

2. Joint modelling of dependence and homogeneity

2.1. RFA-madogram

To introduce homogeneity criteria, see Eq.(1), into distance defined in Eq.(2), we propose to define and study the following expectation

$$D(c, Y_1, Y_2) = \frac{1}{2}\mathbb{E}\left|F_2\left(cY_1\right) - F_1\left(\frac{Y_2}{c}\right)\right|, \tag{3}$$

where $c > 0$ is a normalising positive constant. The $D(c, Y_1, Y_2)$ is always non-negative and equal to zero for $c = \lambda$ when $Y_2 \overset{a.s}{=} \lambda Y_1$. The homogeneous regions are not defined *a priori*, so the existence of λ and its value are not known. We denote

$$c_{12}^* = \text{argmin}\{D(c, Y_1, Y_2) : c > 0\}.$$

Note that $D(c, Y_1, Y_2) = D\left(\frac{1}{c}, Y_2, Y_1\right)$, for all positive c. Therefore, $c_{12}^* = \frac{1}{c_{21}^*}$. The particular case of equality in distribution, $Y_1 \overset{d}{=} Y_2$, corresponds to the case where $c_{12}^* = c_{21}^* = 1$. An important feature of Eq.(3) is that, under the homogeneity condition of Eq.(1),

$$D(\lambda, Y_1, Y_2) = d(Y_1, Y_2),$$

where d is the classical F-madogram, see Eq.(2). To simplify notations, D or $D(c)$ will be a shortcut for $D(c, Y_1, Y_2)$.

The key point from a RFA point of view is that, if Eq.(1) is satisfied, D behaves as the classical F-madogram distance. Note that D is not a true distance, but a dissimilarity. The triangle inequality is satisfied under homogeneity condition but may not be valid in general. Still, D captures information about the extremal dependence like the F-madogram, and, in addition, it encapsulates marginal information concerning the

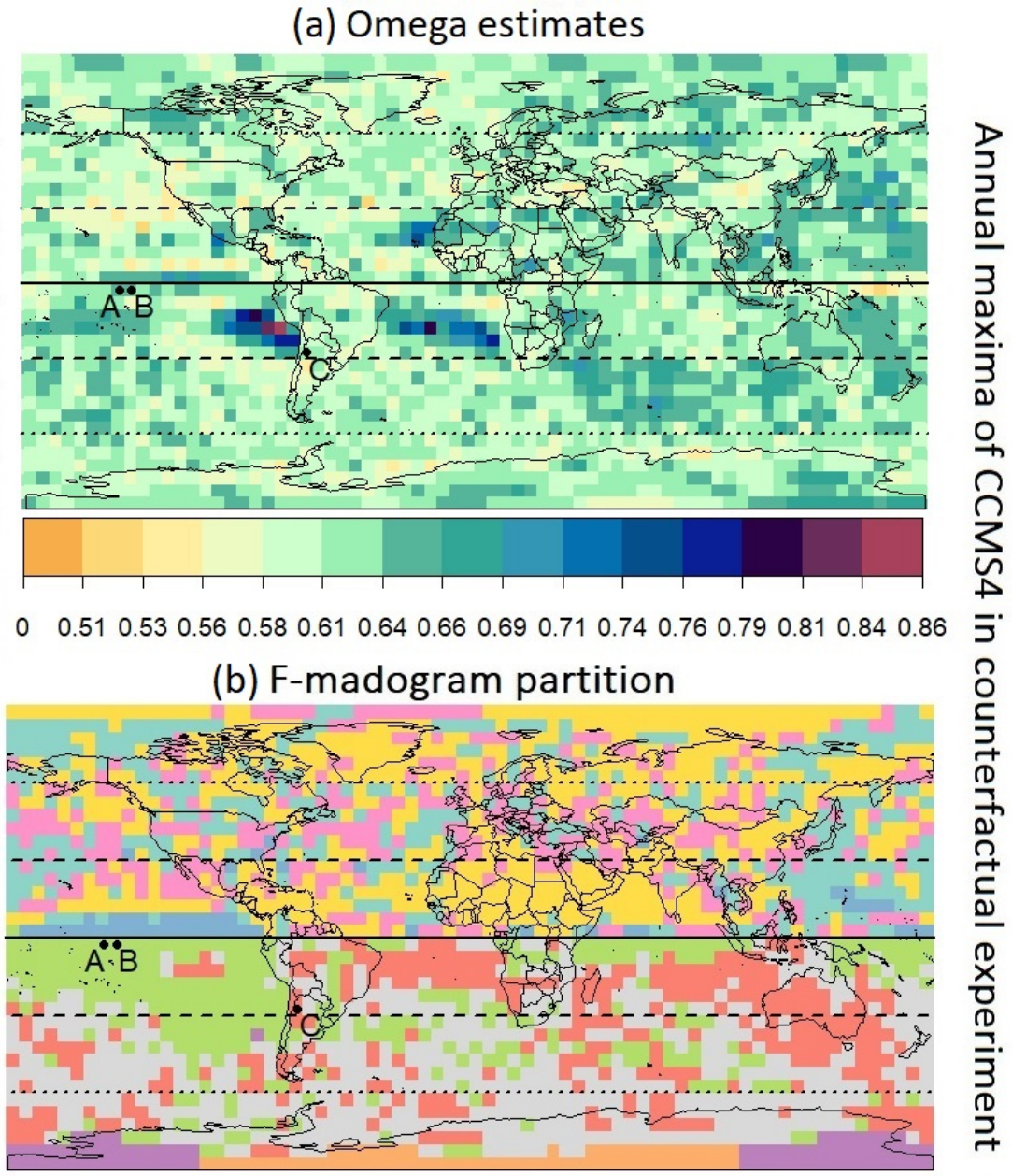

Fig. 2. Two summaries of the structure of the precipitation annual maxima of counterfactual (1850–2005) `CCSM4` model. (a) Pointwise ω ratio (Le Gall et al., 2021). High values of $\hat{\omega}$ correspond to heavy tailed distributions. (b) Results of `PAM` clustering with the F-madogram distance (Bernard et al., 2013), with four clusters for each hemisphere separately. Each color corresponds to a cluster.

departure from Eq.(1). More precisely, one can show (see Appendix A for the proof) that

$$2\left|d-D\right| \leq \mathbb{E}\left[\Delta(c, Y_1)\right] + \mathbb{E}\left[\Delta(c, Y_2/c)\right], \tag{4}$$

where the function $\Delta(c,x) = \left|F_2\left(cx\right) - F_1\left(x\right)\right|$ measures the difference between the rescaled cdfs.

To deepen our understanding of D, we comment on the special case of a bivariate-GEV distributions.

2.2. RFA-madogram for bivariate GEVs

In this section, we suppose that the bivariate vector (Y_1, Y_2) follows a max-stable distribution (Coles et al., 2001; Fougères, 2004; Guillou et al., 2014) with dependence function $V(.,.)$

$$\mathbb{P}(Y_1 \leq x; Y_2 \leq y) = \exp\left[-V\left\{\frac{-1}{\log F_1(x)}, \frac{-1}{\log F_2(y)}\right\}\right],$$

where F_i corresponds to a GEV marginal cdf. If $F_i(x) = \exp\left\{-\left(\frac{x}{\sigma_i}\right)^{-1/\xi_i}\right\}$ with $\xi_1 = \xi_2 = \xi$, then the equality $Y_2 \overset{d}{=} \frac{\sigma_2}{\sigma_1} Y_1$ holds and we are in the homogeneity case.The shape parameter ξ describes the common upper-tail behaviour. The larger ξ is, the heavier the upper-tail of the distribution. Although complex, Eq. (8) in Appendix C, summarises how $D(c)$ can be expressed in function of $V(.,.)$ and the marginal parameters.

To simplify the dependence strength interpretation, it is common to focus on the extremal coefficient defined as the scalar θ_{12} such that

$$\mathbb{P}(Y_1 \leq u, Y_2 \leq u) = \left\{\mathbb{P}(Y_1 \leq u)\mathbb{P}(Y_2 \leq u)\right\}^{\frac{\theta_{12}}{2}}.$$

It is equal to $\theta_{12} = V(1,1)$. If Y_1 and Y_2 are independent, then $\theta_{12} = 2$, while if they are fully dependent, then $\theta_{12} = 1$. Appendix D provides the mathematical details to link the extremal coefficient with $D(c)$. It allows to find an optimal value for rescaling parameter c_{12}^*. For example, it is possible to show that $c_{12}^* = \frac{\sigma_2}{\sigma_1} = \lambda$. for the logistic GEV model,

$$V(x,y) = \left(x^{-\frac{1}{\alpha}} + y^{-\frac{1}{\alpha}}\right)^{\alpha}, \text{ with } \alpha > 0. \tag{5}$$

In particular, the value of the dissimilarity $D(c_{12}^*)$ can be plotted as a function of the logistic coefficient α and of the ratio ξ_1/ξ_2. From Figure 3, one can see that the full dependence case corresponds to $\alpha \approx 0$, and the independence case to $\alpha = 1$. In addition, the ratio ξ_1/ξ_2 varies between 1 (homogeneity case) and 10, i.e. cases with $\xi_1 = 0.1$ and $\xi_2 = 0.01$. The dissimilarity is small when both the dependence is strong and the marginals are homogeneous (leftmost corner). Large dissimilarities correspond to the opposite cases, a near independence and/or strong heterogeneity in the shape parameters (rightmost corner). Note also that, as the homogeneity and the dependence strength decrease jointly, dissimilarity increases (concavity of the surface). These features correspond to our goal that, given the same dependence strength, the price to pay is high

when the RFA condition (1) does not hold. In other words, our aim to cluster grid points that are jointly strongly dependent and in compliance with (1) seems, at least conceptually, to have been reached. The remaining question is to know if this strategy works in practice with the CMIP archive. To answer this, we need to first check that a non-parametric estimator can be developed and its asymptotic properties can be well understood.

3. RFA-madogram Inference

Given $\mathcal{X} \subset \mathbb{R}^n$ and $n \in \mathbb{N}$, let $\ell^\infty(\mathcal{X})$ denote the spaces of bounded real-valued functions on $\mathcal{X}$. For $f\colon \mathcal{X} \to \mathbb{R}$, let $\|f\|_\infty = \sup_{\boldsymbol{x}\in\mathcal{X}} |f(\boldsymbol{x})|$. The arrows "$\xrightarrow{\text{a.s.}}$", "$\Rightarrow$", and "$\rightsquigarrow$" denote almost sure convergence, convergence in distribution of random vectors (see Vaart, 1998, Ch. 2) and weak convergence of functions in $\ell^\infty(\mathcal{X})$ (see Vaart, 1998, Ch. 18–19), respectively. Let $L^2(\mathcal{X})$ denote the Hilbert space of square-integrable functions $f\colon \mathcal{X} \to \mathbb{R}$, with $\mathcal{X}$ equipped with n-dimensional Lebesgue measure; the L^2-norm is denoted by $\|f\|_2 = \left\{\int_{\mathcal{X}} f^2(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\right\}^{1/2}$.

In this section, given a sample of bivariate observations, say $(\mathbf{Y}_1, \ldots, \mathbf{Y}_n)^t$, we focus on the asymptotic properties of two RFA-madogram estimators. Two cases can be studied: when the marginal distributions, F_1 and F_2, are known or when we need to use their empirical estimator, say $\widehat{F}_1$ and $\widehat{F}_2$. In both cases, the copula function of the bivariate vector $(Y_1, Y_2)^t$, say $C(u_1, u_2)$, that captures the dependence structure needs to be inferred. To derive our asymptotic results, we adapt the main ingredients of Theorem 2.4 from Marcon et al. (2017) to our settings, see Appendix B for details. With the notation

$$a_c(u) = F_2\left\{cF_1^{\leftarrow}(u)\right\},$$

we can write

$$D(c) = \frac{1}{2}\mathbb{E}\left|a_c\left(U_1\right) - a_c^{\leftarrow}\left(U_2\right)\right|,$$

where the bivariate vector $\mathbf{U} = (U_1, U_2)^t$ follows the copula $C(\mathbf{u})$. This leads us to the estimators

$$D_n(c) = \frac{1}{n}\sum_{i=1}^{n} D_c\left(\mathbf{U}_i\right), \text{ with } \mathbf{U}_i = (F_1(Y_{1,i}), F_2(Y_{2,i}))^t \text{ and } D_c\left(\mathbf{U}_i\right) = \left|a_c\left(U_{1,i}\right) - a_c^{\leftarrow}\left(U_{2,i}\right)\right|.$$

If F_1 and F_2 are unknown and are replaced by their empirical estimators, we have, with $\hat{a}_c(u) = \hat{F}_2\left\{c\widehat{F}_1^{\leftarrow}(u)\right\}$,

$$\widehat{D}_n(c) = \frac{1}{n}\sum_{i=1}^{n} \hat{D}_c\left(\widehat{\mathbf{U}}_i\right), \text{ with } \widehat{\mathbf{U}}_i = \left\{\widehat{F}_1(Y_{1,i}), \widehat{F}_2(Y_{2,i})\right\}^t \text{ and } \widehat{D}_c\left(\widehat{\mathbf{U}}_i\right) = \left|\hat{a}_c\left(\widehat{U}_{1,i}\right) - \hat{a}_c^{\leftarrow}\left(\widehat{U}_{2,i}\right)\right|.$$

In practice, $\widehat{D}_n(c)$ is directly computed from the expression

$$\widehat{D}_n(c) = \frac{1}{n}\sum_{i=1}^{n}\left|\widehat{F}_2(cY_{1,i}) - \widehat{F}_1(Y_{2,i}/c)\right|.$$

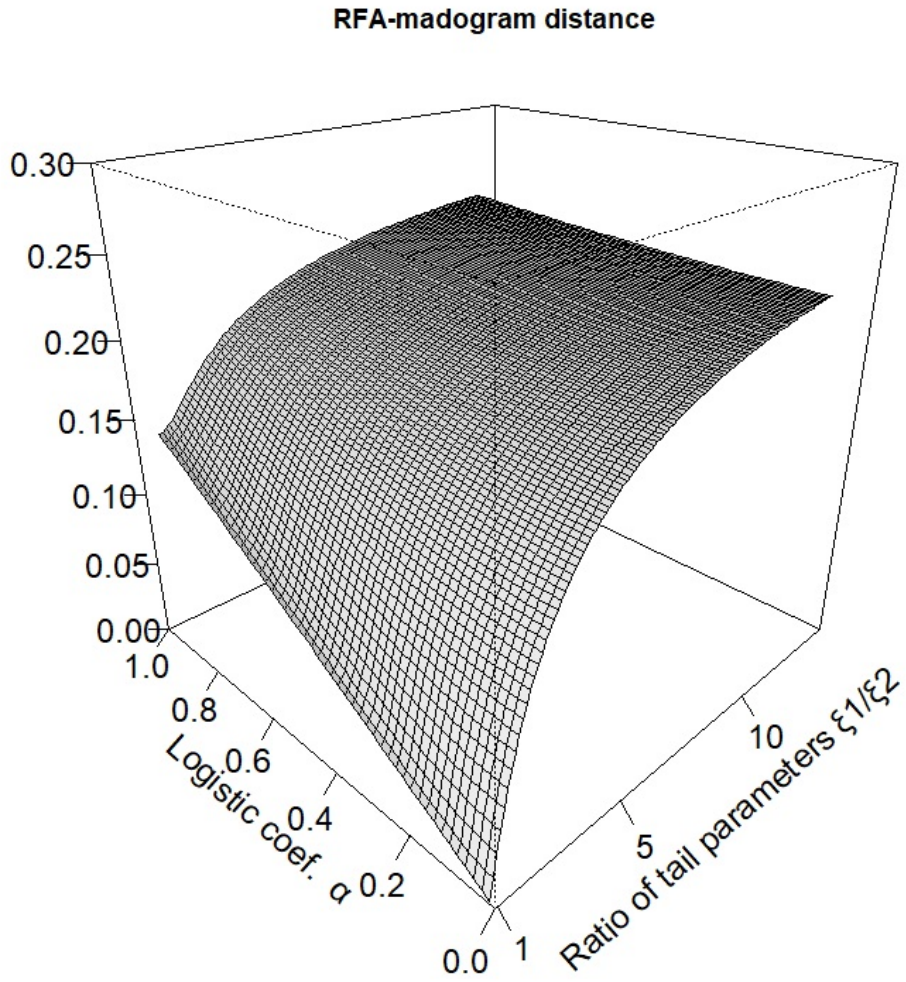

Fig. 3. Distance (z-axis) D defined in Eq.(3) in the logistic bivariate GEV model example. The normalising coefficient is chosen as the optimal one, c^*. The x and y-axis indicate the dependency coefficient α in the logistic dependence, see Eq.(5) and the ratio of tail parameters $i.e.$ the homogeneity of the two r.v. A ratio equal to one corresponds to the homogeneous case. A ratio equal to 10 can be illustrated by the realistic case of $\xi_1 = 0.1, \xi_2 = 0.01$.

Still, the definition of $\widehat{D}_n(c)$ with $\widehat{\mathbf{U}}_i$ facilitates the derivation of theoretical results by leveraging existing properties of the empirical copula

$$C_n(\mathbf{u}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(\mathbf{U}_i \leq \mathbf{u}) \text{ and by writing } D_n(c) = \int_{[0,1]^2} D_c\left(\mathbf{U}\right) dC_n\left(\mathbf{u}\right).$$

In particular, the following classical smoothness condition on copula C is needed, see Example 5.3 in Segers (2012) for details.

Condition (S)

For every $i \in \{1, 2\}$, the partial derivative of C with respect to u_i exists and is continuous on the set $\{\boldsymbol{u} \in [0, 1]^2 : 0 < u_i < 1\}$.

PROPOSITION 3.1. *Let $(\mathbf{Y}_1, \ldots, \mathbf{Y}_n)^t$ be n independent and identically distributed random vectors whose common distribution has continuous margins and a copula function C that satisfies condition (S).*

Let $\mathbb{D}$ be a C-Brownian bridge, that is, a zero-mean Gaussian process on $[0, 1]^2$ with continuous sample paths and with covariance function given by

$$\mathbb{C}ov(\mathbb{D}(\boldsymbol{u}), \mathbb{D}(\mathbf{v})) = C(\boldsymbol{u} \wedge \mathbf{v}) - C(\boldsymbol{u})\, C(\mathbf{v}), \qquad \boldsymbol{u}, \mathbf{v} \in [0, 1]^2. \tag{6}$$

Here $\boldsymbol{u} \wedge \mathbf{v}$ denotes the vector of componentwise minima. We define the Gaussian process $\widehat{\mathbb{D}}$ on $[0, 1]^2$ by

$$\widehat{\mathbb{D}}(\boldsymbol{u}) = \mathbb{D}(\boldsymbol{u}) - \frac{\partial C}{\partial u_1}\, \mathbb{D}(u_1, 1) - \frac{\partial C}{\partial u_2}\, \mathbb{D}(1, u_2)$$

Then we can write that

a) We have $\|D_n(c) - D(c)\|_\infty \to 0$ almost surely as $n \to \infty$. Moreover, as $n \to \infty$,

$$\sqrt{n}\left\{D_n(c) - D(c)\right\} \rightsquigarrow$$
$$\frac{\{1 + D(c)\}^2}{2}\left[\int_0^1 \left\{\mathbb{D}(a_c^{\leftarrow}(x), 1) - \mathbb{D}(a_c^{\leftarrow}(x), a_c(x))\right\} dx + \int_0^1 \left\{\mathbb{D}(1, a_c(x)) - \mathbb{D}(a_c^{\leftarrow}(x), a_c(x))\right\} dx\right]$$

b) We have $\|\widehat{D}_n(c) - D(c)\|_\infty \to 0$ almost surely as $n \to \infty$, and as $n \to \infty$,

$$\sqrt{n}\left\{\widehat{D}_n(c) - D(c)\right\} \rightsquigarrow \left[-\left\{1 + D(c)\right\}^2 \int_0^1 \widehat{\mathbb{D}}\left\{a_c^{\leftarrow}(x), a_c(x)\right\}\, \mathrm{d}x\right]_{c>0}.$$

4. Analysis of CMIP precipitation for 16 models under two experiments

We now apply the RFA-madogram to the problem of partitioning annual precipitation maxima from 16 CMIP GCMs (see Table 1 in Appendix) into homogeneous regions. For each hemisphere of a given GCM run, we estimate the dissimilarity matrix $D(c^*)$ (Eq.(3)) between each pair of grid points. To cluster from a dissimilarity matrix, the `PAM` clustering algorithm is implemented as it is fast, adapted to max-stable distributions (Bernard et al., 2013), and it does not require the triangle inequality (Schubert and Rousseeuw, 2021). The counterfactual (1850–2005) and factual (2071–2100) runs

are analysed separately and later compared to identify possible differences.
With 16 partitions in four clusters for each 16 counterfactual (factual) hemispheric runs, GCM in-between-model error becomes an issue in terms of interpretation. We therefore summarise them in one "central" partitions, which we obtain in two steps. First, partitions for each counterfactual hemispheric runs are relabelled so as to minimise the pairwise difference between two partitions by taking each (alternatively) as reference score. As an example with five grid points, the partitions {1 1 1 2 2 3} and {3 3 3 1 1 2} are equal up to the permutation $(1, 3, 2)$. Then, we compute the probability of each grid point to belong to each of the clusters, and associate the corresponding grid point to the cluster of highest probability. For instance, grid point B is assigned to cluster 1 for 6 models out of 16, to cluster 2 for 9 models and to cluster 3 for only one model. In the so-called central partition, B is then assigned to cluster 2 with probability 9/16. Partitions for the factual experiment are relabelled in order to minimise the difference with the counterfactual central partition.

For example, Figure 4 shows the central partitions in four clusters by hemisphere. Intense colours correspond to points that belong to the same cluster in most, if not all, model partitions. Beginning with the counterfactual experiment, we first note that the clusters are very coherent spatially, in stark contrast to marginal- (ω) and dependence-based (F-madogram) partitions (Figure 2), even though no geographical covariates were used in the clustering.
The Northern Hemisphere is dominated by two clusters (pink and yellow), with two others (blue and turquoise) with limited spatial extent. The distribution is more even in the Southern Hemisphere, and also more zonally symmetric.
These partitions, driven both by homogeneity and dependence, are generally consistent with precipitation climatology. In the Northern Hemisphere, the pink cluster extends over the storm track regions of the North Atlantic and Pacific Oceans, and over the Inter-Tropical Convergence Zone (ITCZ) around 10°N. The blue cluster covers the dry subtropics above the Sahara, Southwest Asia and southwest of North America. The turquoise cluster is located in the dry zone above the cold Pacific tongue, while the yellow cluster includes most regions with semi-arid and continental climates. Still, it also includes monsoon-dominated regions (*e.g.*, India) and the dry Arctic.
In the Southern Hemisphere, arid regions in Antarctica and in the dry descent regions at the eastern edge of the subtropical anticyclones are grouped together in the purple cluster, while the red cluster covers much of the wet tropics. The orange and green clusters correspond to the Southern Hemisphere storm track.

Most of the clusters appear to be quite robust across GCMs. Notable exceptions are the ITCZ regions in the Northern Hemisphere, and the equatorial Pacific and the eastern Indian Ocean west of Australia in the Southern Hemisphere. This lack of robustness may be due to the choice of cluster number. In any case, some differences are expected across GCMs, as they differ in their representation of storm tracks, monsoons or ITCZ location and dynamics.
At first order, it appears that homogeneity of the distributions plays the dominant role, with arid or wet regions grouped together in both hemispheres. Still, the clustering is by design not only based on marginal distributions but also on dependence strength. To measure the importance of dependence in the spatial structure, we apply our clustering

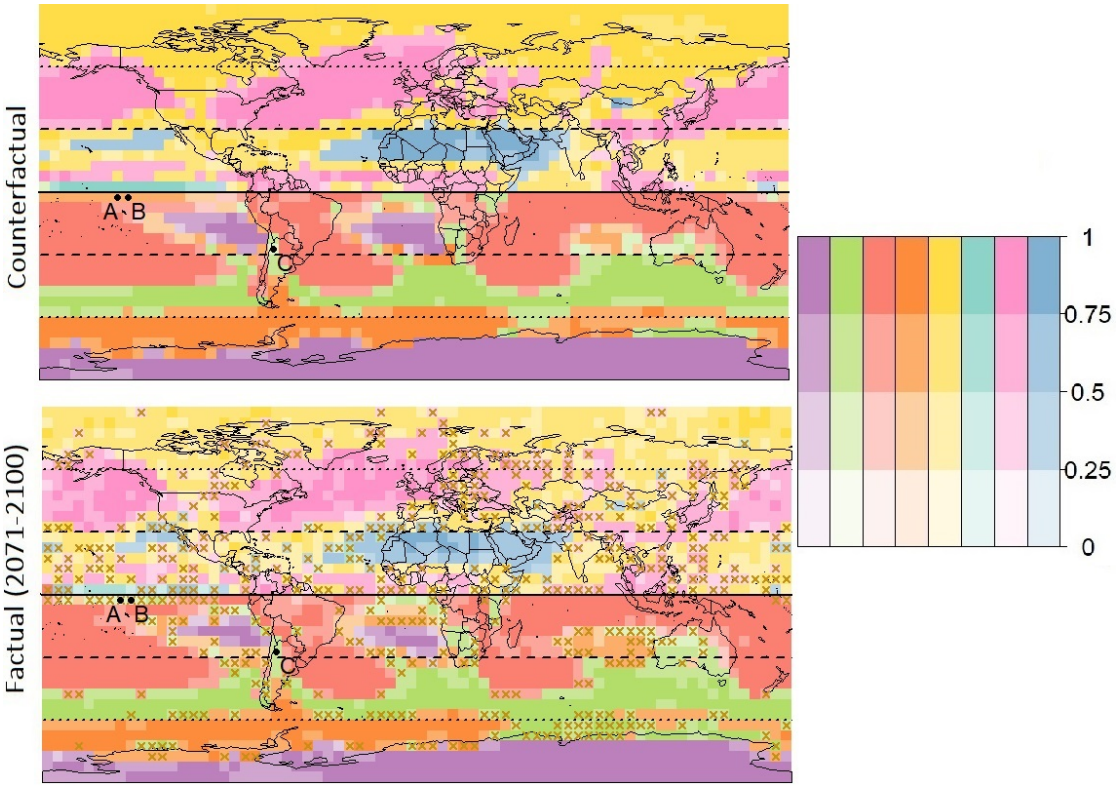

Fig. 4. Central partitions of CMIP models (with four clusters for each hemisphere), for (top) the counterfactual experiment (1850–2005) and (bottom) the factual experiment (2071–2100). Each colour corresponds to a cluster, with the shade indicating the probability of belonging to that cluster. In the bottom map, brown crosses indicate points where the most likely cluster is different between the counterfactual and the factual experiments.

algorithm to temporally shuffled annual maxima at each grid point. This removes any spatial dependence between variables while preserving their marginal distributions. The results of Figure 5 for the `CCSM4` model show a much less spatially coherent partition for the shuffled data. The dependence thus plays an important role in the coherence of the partition. This role can be further quantified by computing the relative difference between RFA-madogram on shuffled and non-shuffled data (with respect to the medoids). For about 2/3 of the grid points, the RFA-madogram takes lower values on the non-shuffled data, in particular near the medoids.

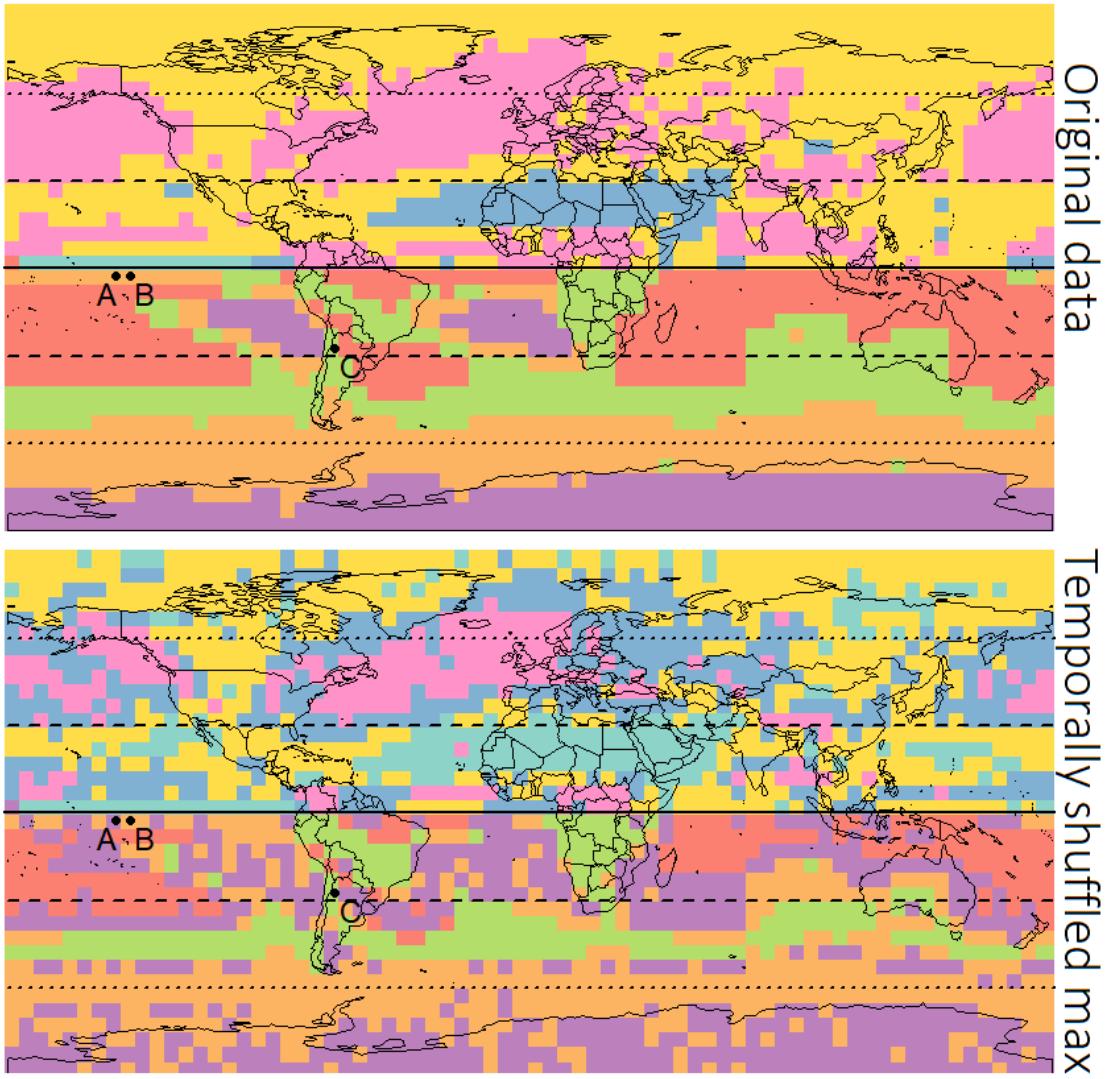

Fig. 5. Partition of `CCSM4` model in the counterfactual experiment based on the RFA-madogram dissimilarity $D(c^*)$ and `PAM` algorithm, for (top) original data, and (bottom) data randomly shuffled in time at each grid point. The clustering algorithm is applied to each hemisphere independently.

We now turn to the comparison of the central partitions between the factual and

counterfactual experiments. The overall partition structure is very similar in both experiments (Figure 4). The clusters are better defined in the counterfactual experiment (*i.e.* cluster probabilities closer to 1) because the sample size is much larger than for the factual experiment (155 *versus* 30 years). Globally, differences between the two central partitions are not significant compared to variability of model partitions compared to the central partition for either the factual or the counterfactual experiment (not shown). Hence, we cannot conclude to more spatial pattern variability in the factual world.
The most likely cluster changes for a number of grid points, however, as indicated by crosses on the bottom panel of Figure 4. In the Northern Hemisphere, the pink (humid) and blue (arid) clusters expand slightly Northwards. More specifically, the probability of a given grid point to belong to the pink cluster generally increases at high latitudes, while the probability to belong to the blue cluster increases around the 25°N latitude. In the Southern Hemisphere the green cluster (humid) also expands Southwards.
While the resolution of our analysis is rather low (5°), these differences are consistent with the expected polewards shift of major climate zones under climate change, particulary the arid subtropics and the storm track regions of both hemispheres (Scheff and Frierson, 2012).

5. Conclusion

When considering multivariate data, extreme value theory can be difficult to handle. Reducing the dimensionality of extreme precipitation data set is then a challenging task. Our main goal in this work was to show that a simple and fast clustering approach based on an interpretable dissimilarity could highlight climatologically coherent regions.

The proposed approach coupled the main RFA idea, *i.e.* a normalising factor, with the dependence structure via the F-madogram. The introduced dissimilarity has links with extreme value theory via the extremal coefficient and tail parameters. The RFA-madogram neither requires estimating any marginal parameters nor dependence parameters. It is fully data-driven and bypasses the need of selecting relevant covariates or dependence structure.

Our analysis of annual maxima of daily precipitation from each CMIP model provides more spatially coherent hemispheric regions than some other non-parametric methods focusing on only one aspect (either homogeneity or dependence). Another contribution of this work is the handling of multi-partitions as our selected CMIP set has 16 GCM runs. Our combining approach enables us to compare one multi-model partition of the factual (all forcings) world with another multi-model partition of counterfactual (natural forcings) world. It appears that spatial variability between all models for the factual (resp. counterfactual) experiment appears to be significantly higher than between the two factual and counterfactual experiments.

In this work, we focus on the spatial structure of annual maxima precipitation in CMIP models, and on the forcing impact. We did not directly study the changes in rainfall distributions and frequencies. One interesting perspective would be to model precipitation intensities and dependence structure within each cluster. This could be useful for the D&A community. Another aspect is that the statistical approach developed therein is easy-to-implement and flexible, e.g. it can be used on non-gridded products.

For example, it could be applied to large weather networks, reanalysis (ERA 5) and radar products. Such datasets have finer spatial resolution scales than GCMs, and the dependence structure could be stronger, and consequently the analysis of heavy rainfall spatial patterns at fine spatial scales improved.

Acknowledgement

Within the CDP-Trajectories framework, this work is supported by the French National Research Agency in the framework of the "Investissements d'avenir" program (ANR-15-IDEX-02).

Part of this work was supported by the DAMOCLES-COST-ACTION on compound events, the French national program (FRAISE-LEFE/INSU and 80 PRIME CNRS-INSU), and the European H2020 XAIDA (Grant agreement ID: 101003469). The authors also acknowledge the support of the French Agence Nationale de la Recherche (ANR) under reference ANR-20-CE40-0025-01 (T-REX project), and the ANR-Melody.

A. Proof of Eq. (4)

We can write that

$$\begin{aligned} 2D(c) &= \mathbb{E}\left|F_2\left(cY_1\right) - F_1\left(Y_1\right) + F_1\left(Y_1\right) - F_2\left(Y_2\right) + F_2\left(Y_2\right) - F_1\left(Y_2/c\right)\right| \\ &\leq \mathbb{E}\left|F_2\left(cY_1\right) - F_1\left(Y_1\right)\right| + \mathbb{E}\left|F_1\left(Y_1\right) - F_2\left(Y_2\right)\right| + \mathbb{E}\left|F_2\left(Y_2\right) - F_1\left(Y_2/c\right)\right|, \\ &\leq 2d + \mathbb{E}\left[\Delta(c, Y_1)\right] + \mathbb{E}\left[\Delta(c, Y_2/c)\right] \end{aligned}$$

In the same way, we can write that

$$\begin{aligned} 2d &= \mathbb{E}\left|F_1\left(Y_1\right) - F_2\left(cY_1\right) + F_2\left(cY_1\right) - F_1\left(Y_2/c\right) + F_1\left(Y_2/c\right) - F_2\left(Y_2\right)\right|, \\ &\leq 2D(c) + \mathbb{E}\left[\Delta(c, Y_1)\right] + \mathbb{E}\left[\Delta(c, Y_2/c)\right]. \end{aligned}$$

It follows that the inequality expressed in Eq.(4) is valid. □

B. Proof of Proposition 3.1

Let $a(u)$ be any continuous non-decreasing function from $[0,1]$ to $[0,1]$ and denote its inverse by $a^{\leftarrow}(u)$. The map

$$\phi : \ell^{\infty}([0,1]^2) \to \ell^{\infty}([0,1]) : f \mapsto \phi(f)$$

defined by

$$(\phi(f))(a) = \frac{1}{2}\left(\int_0^1 f\left(a^{\leftarrow}(u), 1\right) du + \int_0^1 f\left(1, a(u)\right) du\right) - \int_0^1 f\left(a^{\leftarrow}(u), a(u)\right) du$$

is linear and bounded, and therefore continuous. To continue, we need the following lemma.

LEMMA B.1. *For any cumulative distribution function H on $[0,1]^2$ and for any non-decreasing function $a(.)$ on $[0,1]$, the function*

$$\delta\left(\mathbf{u}\right) = \frac{1}{2}\left|a(u_1) - a^{\leftarrow}(u_2)\right|$$

satisfies

$$\int_{[0,1]^2} \delta\left(\mathbf{u}\right) dH\left(\mathbf{u}\right) = (\phi(H))(a). \tag{7}$$

Table 1. List of 16 CMIP models considered, with institutions, belonging countries and native horizontal resolution (longitude by latitude in degree). AOR (UoT): Atmosphere and Ocean Research Institute (The University of Tokyo); CSIRO: Commonwealth Scientific and Industrial Research Organisation; DOE: Department of Energy; JAMSTEC: Japan Agency for Marine-Earth Science and Technology; NIES: National Institute for Environmental Studies; NSF: National Science Foundation. Most models come from the CMIP phase 5, those coming from phase 6 are indicated by *. In this paper, models are regridded to a resolution of $5°$ x $5°$.

Model	Institute	Country	Resolution
CanESM2	Canadian Centre for Climate Modelling and Analysis	Canada	2.8° x 2.8°
CanESM5*			2.8° x 2.8°
CCSM4	National Center for Atmospheric Research (NCAR)	USA	1.3° x 0.9°
CESM1-CAM5	NSF, DOE and NCAR	USA	1.3° x 0.9°
CNRM-CM5	Centre National de Recherches Météorologiques	France	1.4° x 1.4°
CNRM-CM6-1*			1.4° x 1.4°
ACCESS1-3	CSIRO and Bureau of Meteorology	Australia	1.9° x 1.3°
CSIRO-Mk3-6-0			1.9° x 1.9°
IPSL-CM5A-LR	Institut Pierre Simon Laplace	France	3.8° x 1.9°
IPSL-CM5A-MR			2.5° x 1.3°
IPSL-CM6A-LR*			2.5° x 1.3°
MIROC-ESM	JAMSTEC, AOR (UoT), NIES	Japan	2.8° x 2.8°
MIROC-ESM-CHEM			2.8° x 2.8°
MRI-CGCM3	Meteorological Research Institute	Japan	1.1° x 1.1°
MRI-ESM2-0*			1.1° x 1.1°
NorESM1-M	Norwegian Climate Centre	Norway	2.5° x 1.9°

Proof of Lemma B.1: Note that

$$\delta(\mathbf{u}) = \max\left(a(u_1), a^{\leftarrow}(u_2)\right) - \frac{1}{2}\left(a(u_1) + a^{\leftarrow}(u_2)\right).$$

For any $\mathbf{u} \in [0,1]^2$, we have

$$\max\left(a(u_1), a^{\leftarrow}(u_2)\right) = 1 - \int_0^1 \mathcal{I}\left(u_1 \leq a^{\leftarrow}(u), u_2 \leq a(u)\right) du$$

and

$$\frac{1}{2}\left(a(u_1) + a^{\leftarrow}(u_2)\right) = 1 - \frac{1}{2}\left(\int_0^1 \mathcal{I}\left(u_1 \leq a^{\leftarrow}(u)\right) du + \int_0^1 \mathcal{I}\left(u_2 \leq a(u)\right) du.\right)$$

Subtracting both expressions and integrating over H implies

$$\begin{aligned}\int_{[0,1]^2} \delta(\mathbf{u})\, dH(\mathbf{u}) =& \frac{1}{2}\left(\int_{[0,1]^2}\int_0^1 \mathcal{I}\left(u_1 \leq a^{\leftarrow}(u)\right) du dH(u_1,u_2)\right.\\ &\left.+ \int_{[0,1]^2}\int_0^1 \mathcal{I}\left(u_2 \leq a(u)\right) du dH(u_1,u2)\right)\\ &- \int_{[0,1]^2}\int_0^1 \mathcal{I}\left(a^{\leftarrow}(u_1) \leq u, a(u_2) \leq u\right) du dH(u_1,u_2).\end{aligned}$$

The stated lemma can be deduced by applying Fubini's theorem on the three double integrals. □

By Lemma B.1, we obtain for $a_c(u) = F_2(cF_1^{\leftarrow}(u))$

$$D_n(a_c) = (\phi(C_n))(a_c) \text{ and } D(a_c) = (\phi(C))(a_c).$$

this leads to

$$||D_n(a_c) - D(a_c)||_\infty \leq 2||C_n - C||_\infty.$$

Classical results about empirical copulas gives uniform strong consistency, see Segers .... Similar arguments can be used for $\widehat{D}_n(\hat{a}_c)$. Now, we can consider the empirical process

$$\mathbb{D}_n = \sqrt{n}(C_n - C), \qquad \widehat{\mathbb{D}}_n = \sqrt{n}(\widehat{C}_n - C).$$

and we can write

$$\sqrt{n}\big(D_n(a_c) - D(a_c)\big) = (\phi(\mathbb{D}_n))(a_c) \text{ and } \sqrt{n}\big(\widehat{D}_n(\hat{a}_c) - D(\hat{a}_c)\big) = (\phi(\widehat{\mathbb{D}}_n))(\hat{a}_c).$$

We recall now that in the space $\ell^\infty([0,1]^d)$ equipped with the supremum norm, $\mathbb{D}_n \rightsquigarrow \mathbb{D}$, as $n \to \infty$, where $\mathbb{D}$ is a C-Brownian bridge, and, as condition (S) holds, then $\widehat{\mathbb{D}}_n \rightsquigarrow \widehat{\mathbb{D}}$, as $n \to \infty$, where $\widehat{\mathbb{D}}$ is the Gaussian process defined in (3.1), see Segers (2012) for details. In addition, $\hat{a}_c$ converges in probability to a_c. The continuous mapping theorem then implies, as $n \to \infty$,

$$\sqrt{n}\big(D_n(a_c) - D(a_c)\big) = \phi(\mathbb{D}_n) \rightsquigarrow \phi(\mathbb{D}), \quad \sqrt{n}\big(\widehat{D}_n(\hat{a}_c) - D(\hat{a}_c)\big) = (\phi(\widehat{\mathbb{D}}_n)) \rightsquigarrow \phi(\widehat{\mathbb{D}}),$$

in $\ell^\infty([0,1])$. From the continuity of its sample paths and by the form of the covariance function (6), the Gaussian process $\widehat{\mathbb{D}}$ satisfies

$$\mathbb{P}\{\forall\, u \in [0,1] : \widehat{\mathbb{D}}(u,1) = \widehat{\mathbb{D}}(1,u) = 0\} = 1.$$

Ths provides all the elements to conclude the proposition. □.

C. Expression of $D(c)$ in the bivariate GEV case

As $|a-b| = 2\max(a,b) - a - b$, we have

$$2D(c) = 2\mathbb{E}\left[\max\left(F_2\left(cY_1\right), F_1\left(Y_2/c\right)\right)\right] - \mathbb{E}\left[F_2\left(cY_1\right)\right] - \mathbb{E}\left[F_1\left(Y_2/c\right)\right]$$

To deal with each term, we recall that the quantile function of $F(x;\xi,\sigma) = \exp\left[-\left(\frac{x}{\sigma}\right)^{-1/\xi}\right]$ is

$$F^{-1}(u;\sigma,\xi) = \sigma\left(-\log u\right)^{-\xi} = \sigma z^{\xi}, \text{ with } z = -1/\log(u),$$

This implies that

$$Y_i \overset{\mathrm{d}}{=} \sigma_i Z_i^{\xi_i},$$

where Z_i follows an unit Fréchet. If follows that, with $a_{12} = \left(\frac{c\sigma_1}{\sigma_2}\right)^{-1/\xi_2}$,

$$F_2\left(cY_1\right) \overset{\mathrm{d}}{=} \exp\left[-\left(\frac{cY_1}{\sigma_2}\right)^{-1/\xi_2}\right] \overset{\mathrm{d}}{=} \exp\left(-a_{12}Z_1^{-\xi_1/\xi_2}\right),$$

then

$$F_2\left(cY_1\right) \overset{\mathrm{d}}{=} \exp\left(-a_{12}W_1\right) \text{ with } W_1 = Z_1^{-\xi_1/\xi_2}.$$

In the same way, with $a_{21} = \left(\frac{\sigma_2}{c\sigma_1}\right)^{-1/\xi_1}$,

$$F_1\left(Y_2/c\right) \overset{\mathrm{d}}{=} \exp\left[-\left(\frac{Y_2}{c\sigma_1}\right)^{-1/\xi_1}\right] \overset{\mathrm{d}}{=} \exp\left(-a_{21}Z_2^{-\xi_2/\xi_1}\right)$$

then

$$F_1\left(Y_2/c\right) \overset{\mathrm{d}}{=} \exp\left(-a_{21}W_2\right) \text{ with } W_2 = Z_2^{-\xi_2/\xi_1}.$$

By noticing that W_i follows a Weibull distribution with $\mathbb{P}(W_1 > w) = \exp(-w^{-\xi_2/\xi_1})$, the expectation $\mathbb{E}[F_2\left(cY_1\right)]$ can be linked as the Laplace transform of a Weibull r.v.

$$\mathbb{E}\left[F_2\left(cY_1\right)\right] = \mathbb{E}\left[\exp\left(-a_{12}W_1\right)\right] \text{ and } \mathbb{E}\left[F_1\left(Y_2/c\right)\right] = \mathbb{E}\left[\exp\left(-a_{21}W_2\right)\right].$$

For the bivariate structure, we can write that, for any $u \in (0,1)$,

$$\begin{aligned}\mathbb{P}\left[\max\left(F_2\left(cY_1\right),F_1\left(Y_2/c\right)\right)\leq u\right] &= \mathbb{P}\left[\max\left(\exp\left(-a_{12}Z_1^{-\xi_1/\xi_2}\right),\exp\left(-a_{21}Z_2^{-\xi_2/\xi_1}\right)\right)\leq u\right],\\ &= \mathbb{P}\left[Z_1\leq\left(\frac{-a_{12}}{\log u}\right)^{\xi_2/\xi_1},\ Z_2\leq\left(\frac{-a_{21}}{\log u}\right)^{\xi_1/\xi_2}\right],\\ &= \exp\left\{-V\left[\left(\frac{-a_{12}}{\log u}\right)^{\xi_2/\xi_1},\left(\frac{-a_{21}}{\log u}\right)^{\xi_1/\xi_2}\right]\right\}.\end{aligned}$$

Since the r.v. $\max\left(F_2\left(cY_1\right),F_1\left(Y_2/c\right)\right)\leq u$ is positive, in the general setup, we have

$$\begin{aligned}D &= \int_0^1\left(1-\exp\left\{-V\left[\left(\frac{a_{12}}{-\log u}\right)^{\xi_2/\xi_1},\left(\frac{a_{21}}{-\log u}\right)^{\xi_1/\xi_2}\right]\right\}\right)du\\ &\quad-\frac{1}{2}\mathbb{E}\left[\exp\left(-a_{12}W_1\right)\right]-\frac{1}{2}\mathbb{E}\left[\exp\left(-a_{21}W_2\right)\right],\end{aligned}\tag{8}$$

where W_i follows a Weibull distribution with $\mathbb{P}(W_1>w)=\exp(-w^{\xi_1/\xi_2})$. Note that

$$\left(a_{12}\right)^{\frac{\xi_2}{\xi_1}}=\frac{1}{a_{21}}$$

Conversely, $\left(a_{21}\right)^{\frac{\xi_1}{\xi_2}}=\dfrac{1}{a_{12}}$

D. Homogeneous case

In the special case where $\xi_1=\xi_2=\xi$, we denote $\theta_c=V\left(a_{12},a_{21}\right)$, where $a_{12}=\left(\dfrac{c\sigma_1}{\sigma_2}\right)^{-1/\xi}=1/a_{21}$. Then, we have

$$\begin{aligned}\mathbb{P}\left[\max\left(F_2\left(cY_1\right),F_1\left(Y_2/c\right)\right)\leq u\right] &= \exp\left\{V\left[\left(\frac{\sigma_2}{c\sigma_1}\right)^{-1/\xi},\left(\frac{c\sigma_1}{\sigma_2}\right)^{-1/\xi}\right]\log u\right\},\\ &= u^{V(a_{12},a_{21})}.\end{aligned}$$

We can write

$$D=\int_0^1 1-u^{\theta_c}du-\frac{1}{2}\mathbb{E}\left[\exp\left(-a_{12}W_1\right)\right]-\frac{1}{2}\mathbb{E}\left[\exp\left(-a_{21}W_2\right)\right]\tag{9}$$

where $W_i, i=1,2$ has cdf equal to $\exp(-x)$.
Hence,

$$D=\frac{\theta_c}{\theta_c+1}-\frac{1}{2(1+a_{12})}-\frac{1}{2(1+a_{21})}.$$

To minimise D as a function of c, we study the variations of $r : x \longmapsto \frac{V\left(x, \frac{1}{x}\right)}{1+V\left(x, \frac{1}{x}\right)} - \frac{1}{2(1+x)} - \frac{x}{2(1+x)}$. We suppose that V **is differentiable.** If the previous function r admits a minimum, its derivative cancels in some c_0. The r' cancels if and only if the derivative of $x \longmapsto \frac{V\left(x, \frac{1}{x}\right)}{1+V\left(x, \frac{1}{x}\right)}$ cancels, if and only if there exists x s.t. $\frac{\partial V}{\partial x}\left(x, \frac{1}{x}\right) = \frac{1}{x^2}\frac{\partial V}{\partial y}\left(x, \frac{1}{x}\right)$. In the special case where the **dependence is logistic** *i.e.*

$$V(x,y) = \left(\frac{1}{x^{1/\alpha}} + \frac{1}{y^{1/\alpha}}\right)^{\alpha},$$

we have $\frac{\partial V}{\partial x}\left(x, \frac{1}{x}\right) = \frac{\partial V}{\partial y}\left(x, \frac{1}{x}\right)$, for all positive x. Therefore, if r admits a minimum, it is for $x = \pm 1$. Eventually, for logistic dependence, D is minimal for

$$c = \frac{\sigma_2}{\sigma_1}.$$

References

Alexander, L. V. and Arblaster, J. M. (2017) Historical and projected trends in temperature and precipitation extremes in Australia in observations and CMIP5. *Weather and Climate Extremes*, **15**, 34–56.

Ammann, C. M. and Naveau, P. (2010) Statistical volcanic forcing scenario generator for climate simulations. *Journal of Geophysical Research: Atmospheres*, **115**.

Asadi, P., Engelke, S. and Davison, A. C. (2018) Optimal regionalization of extreme value distributions for flood estimation. *Journal of Hydrology*, **556**, 182–193.

Bador, M., Naveau, P., Gilleland, E., Castellà, M. and Arivelo, T. (2015) Spatial clustering of summer temperature maxima from the CNRM-CM5 climate model ensembles & E-OBS over Europe. *Weather and climate extremes*, **9**, 17–24.

Bernard, E., Naveau, P., Vrac, M. and Mestre, O. (2013) Clustering of maxima: Spatial dependencies among heavy rainfall in France. *Journal of Climate*, **26**, 7929–7937.

Burn, D. H. (1990) Evaluation of regional flood frequency analysis with a region of influence approach. *Water Resources Research*, **26**, 2257–2265.

Campagne, A. (2017) *Le capitalocène: aux racines historiques du dérèglement climatique*. Éditions Divergences.

Coles, S., Bawa, J., Trenner, L. and Dorazio, P. (2001) *An introduction to statistical modeling of extreme values*, vol. 208. Springer.

Cooley, D., Naveau, P. and Poncet, P. (2006) Variograms for spatial max-stable random fields. In *Dependence in probability and statistics*, 373–390. Springer.

Crutzen, P. J. (2006) The "anthropocene". In *Earth System Science in the Anthropocene*, 13–18. Springer.

Dalrymple, T. (1960) Flood-frequency analyses, manual of hydrology: Part 3. *Tech. rep.*, USGPO,.

Davison, A. C., Padoan, S. A. and Ribatet, M. (2012) Statistical modeling of spatial extremes. *Statistical science*, **27**, 161–186.

Dong, S., Sun, Y., Li, C., Zhang, X., Min, S.-K. and Kim, Y.-H. (2021) Attribution of extreme precipitation with updated observations and cmip6 simulations. *Journal of Climate*, **34**, 871 – 881.

Drees, H. and Sabourin, A. (2021) Principal component analysis for multivariate extremes. *Electronic Journal of Statistics*, **15**, 908–943.

Evin, G., Blanchet, J., Paquet, E., Garavaglia, F. and Penot, D. (2016) A regional model for extreme rainfall based on weather patterns subsampling. *Journal of Hydrology*, **541**, 1185–1198.

Fawad, M., Ahmad, I., Nadeem, F. A., Yan, T. and Abbas, A. (2018) Estimation of wind speed using regional frequency analysis based on linear-moments. *International Journal of Climatology*, **38**, 4431–4444.

Fougères, A.-L. (2004) Multivariate extremes. *Monographs on Statistics and Applied Probability*, **99**, 373–388.

Greenwood, J. A., Landwehr, J. M., Matalas, N. C. and Wallis, J. R. (1979) Probability weighted moments: definition and relation to parameters of several distributions expressable in inverse form. *Water resources research*, **15**, 1049–1054.

Guillou, A., Naveau, P. and Schorgen, A. (2014) Madogram and asymptotic independence among maxima. *REVSTAT-Statistical Journal*, **12**.

Gumbel, E. J. (1960) Distributions des valeurs extremes en plusieurs dimensions. *Publications de l'Institut de statistique de l'Université de Paris*, **9**, 171–173.

Hosking, J. R. M. and Wallis, J. R. (2005) *Regional frequency analysis: an approach based on L-moments*. Cambridge University Press.

IPCC (2013) *Summary for Policymakers*, book section SPM, 1–30. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press. URL: `www.climatechange2013.org`.

IPCC (2021) *Climate Change 2021: The Physical Science Basis*. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press (In Press). URL: `https://www.ipcc.ch/report/ar6/wg1/downloads/report/IPCC_AR6_WGI_Full_Report.pdf`.

Janßen, A., Wan, P. et al. (2020) k-means clustering of extremes. *Electronic Journal of Statistics*, **14**, 1211–1233.

Kaufman, L. and Rousseeuw, P. (1990) *Finding groups in data: an introduction to cluster analysis*. Wiley Series in Probability and Statistics, Wiley.

Kharin, V. V., Zwiers, F., Zhang, X. and Wehner, M. (2013) Changes in temperature and precipitation extremes in the cmip5 ensemble. *Climatic change*, **119**, 345–357.

Kim, H., Duan, R., Kim, S., Lee, J. and Ma, G.-Q. (2019) Spatial cluster detection in mobility networks: a copula approach. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **68**, 99–120.

Le Gall, P., Favre, A.-C., Naveau, P. and Prieur, C. (2021) Improved regional frequency analysis of rainfall data. (Submitted).

Malm, A. and Hornborg, A. (2014) The geology of mankind? A critique of the Anthropocene narrative. *The Anthropocene Review*, **1**, 62–69.

Marcon, G., Padoan, S., Naveau, P., Muliere, P. and Segers, J. (2017) Multivariate nonparametric estimation of the pickands dependence function using bernstein polynomials. *Journal of Statistical Planning and Inference*, **183**, 1–17.

Meehl, G. A., Boer, G. J., Covey, C., Latif, M. and Stouffer, R. J. (2000) The coupled model intercomparison project (CMIP). *Bulletin of the American Meteorological Society*, **81**, 313–318.

Naveau, P., Guillou, A., Cooley, D. and Diebolt, J. (2009) Modelling pairwise dependence of maxima in space. *Biometrika*, **96**, 1–17.

Naveau, P., Hannart, A. and Ribes, A. (2020) Statistical methods for extreme event attribution in climate science. *Annual Review of Statistics and Its Application*, **7**, 89–110.

van Oldenborgh, G. J., van der Wiel, K., Kew, S., Philip, S., Otto, F., Vautard, R., King, A., Lott, F., Arrighi, J., Singh, R. and van Aalst, M. (2021) Pathways and pitfalls in extreme event attribution. *Climatic Change*, **166**, 13.

Pfahl, S., O'Gorman, P. A. and Fischer, E. M. (2017) Understanding the regional pattern of projected future changes in extreme precipitation. *Nature Climate Change*, **7**, 423–427.

Ribes, A., Qasmi, S. and Gillett, N. P. (2021) Making climate projections conditional on historical observations. *Science Advances*, **7**.

Saf, B. (2009) Regional flood frequency analysis using L-moments for the West Mediterranean region of Turkey. *Water Resources Management*, **23**, 531–551.

Saunders, K., Stephenson, A. and Karoly, D. (2021) A regionalisation approach for rainfall based on extremal dependence. *Extremes*, **24**, 215–240.

Scheff, J. and Frierson, D. M. W. (2012) Robust future precipitation declines in cmip5 largely reflect the poleward expansion of model subtropical dry zones. *Geophysical Research Letters*, **39**. URL: `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2012GL052910`.

Schubert, E. and Rousseeuw, P. J. (2021) Fast and eager k-medoids clustering: O (k) runtime improvement of the PAM, CLARA, and CLARANS algorithms. *Information Systems*, **101**, 101804.

Segers, J. (2012) Asymptotics of empirical copula processes under non-restrictive smoothness assumptions. *Bernoulli*, **18**, 764–782.

Tandon, N. F., Zhang, X. and Sobel, A. H. (2018) Understanding the dynamics of future changes in extreme precipitation intensity. *Geophysical Research Letters*, **45**, 2870–2878.

Tawn, J. A. (1988) Bivariate extreme value theory: models and estimation. *Biometrika*, **75**, 397–415.

Toreti, A., Giannakaki, P. and Martius, O. (2016) Precipitation extremes in the mediterranean region and associated upper-level synoptic-scale flow structures. *Climate dynamics*, **47**, 1925–1941.

Vaart, A. W. v. d. (1998) *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.