Coupling clustering algorithms and extreme value to theory to define spatial homogenous hydrological regions

Philippe Naveau[1,]; Philomène Le Gall[1,2]; Margaux Zafran[1], Anne-Catherine Favre[2]

[1] CNRS - Laboratoire des Sciences du Climat et de l'Environnement, Gif-sur-Yvette France

[2] IGE Institut des géosciences de l'Environnement, Grenoble, France

Note: this talk is part of **<IPS70: The statistics of extremes>** scheduled for **<Mon 15:00-16:30 Amsterdam (UTC+2)** organized by Anne Sabourin

Abstract:
Extreme precipitation often cause important floods and lead to important societal and economical damages. Rainfall extremes are subject to local orography features and their intensities can be highly variable. In this context, identifying climatically coherent regions is paramount to understand and analyze rainfall at the correct spatial scale. The main goal of this talk is to propose and study different dissimilarities and metrics that are tailored to capture both extremal behavior and hydrological features. We focus on blinding a margin-free distance adapted to extremes with the classical scale invariance constraint used in hydrological regional frequency analysis (RFA). In addition, we make the link between types of Kullback-Leibler divergences for extremes and unsupervised clustering techniques such as k-mediods and hclust approaches.

Both types of extremal dependences (asymptotic dependence and asymptotic independence) will be touched upon during this presentation.

A simulation data study will be detailed. One important aspect of this work is to be able to treat very large numerical climate outputs at the global scale. So, computationnal time and scalability will be key here.

Keywords:
Extreme Value Theory; Clustering; CMIP6 database; Kullback-Leibler divergences

1. Introduction:

The main motivation of this study is the analysis of heavy rainfall. Their distributions can highly variable in time, space and intensity. This leads practitioners to look after spatially coherent regions for which the distributional features are homogeneous, up to a multiplicative constant, and this reduces the uncertainties in the computation of high return levels (high quantiles). Historically, the class of statistical approaches that aim at partitioning a region of interest into homogeneous sub-regions is called regional frequency analysis (RFA) in hydrology. In the classical RFA approach proposed by Dalrymple (see e.g Dalrymple,1960) and developed by Hosking and his colleagues (see, e.g. Hosking and

Wallis, 1987), it is implicitly assumed that rainfall locations are (conditionally) independent and a cluster is defined as a scale invariant region, in particular the upper tail index that drives extremal behaviors is assumed to be constant within a cluster. Consequently, by gathering stations within a cluster, the sample size increases and the estimation of this common shape parameter describing the rainfall upper tail is imporved. Still, the question of how to find homogeneous clusters remains complex.

Various RFA techniques based on explanatory covariates (e.g., see Asadi, Engelke and Davi-son, 2018; Fawad et al., 2018, for recent work) were developed to identify homogeneous regions. Theses approaches consist in selecting covariates characterizing station, location or weather patterns to explain the spatial distribution of rainfall (e.g. see Hosking and Wallis, 2005; Burn, 1990; Evin et al., 2016). This variable selection step was used to predict the distribution at un-gauged sites. For instance, Carreau, Naveau and Neppel (2017) defined the scale parameter of a Generalized Pareto Distribution, the archetypical distribution for threshold exceedances, as a function of the weather station coordinates. However, picking relevant covariates requires data availability and subjectivity.

Other techniques entirely bypassed the selection of covariates by working directly with the moments used for homogeneity tests (Saf, 2009). For example, Le Gall et al. (2021) considered a ratio of Probability Weighted Moments (Greenwood et al., 1979). This ratio has the interesting feature of being invariant on a homogeneous region and it is fast, easy to infer, and simple asymptotic convergence properties. Still, the main drawback of this semi-parametric approach is that it is only focus on the marginal behaviors of rainfall data. The spatial dependence is completely ignored. Not accounting for the dependence leads to two issues. If two recording stations are strongly dependent, this strong link between the two stations reinforces the idea of grouping them in the same region. Accounting for the dependence in a RFA clustering algorithm should then improve its efficiency. The second issue is related to the statistical procedure to test the scale invariance homogeneity within a cluster. Assuming independence, while the data are dependent, leads to wrongly increase the rejection rate of homogeneity tests. To avoid two types of issues, various authors proposed clustering algorithms only based on the dependence structure. Recently, a parametric approach based on copula was introduced by Kim et al. (2019). They worked on cluster detection in mobility networks. Their proposal is to gather sites that are subject to intense traffic according to their covariates (e.g. geographical). The dependence strength within each cluster is then check by fitting a multivariate Gumbel copula. Non-parametric approaches based on exceedances were also proposed by Drees and Sabourin (2019);Janßen et al. (2020). Observations are projected onto the unit sphere. The dimension is then reduced by clustering using K-means algorithm or extremes (Janßen et al., 2020, Drees and Sabourin, 2019).

In terms of spatial clustering of extreme rainfall and temperatures data, a series of articles have been based on a non parametric approach based on the F-madogram (Cooley, Naveau and Poncet, 2006). This metric is simply the expected L1-norm between two random variables that have been, marginally, transformed into uniform variables, see Padoan et al. (2014) for asymptotic consistency. As explicit expressions can be made between multivariate max-stable vectors and the F-madogram, this interpretable metric has been used in a clustering context by Bernard et al. (2013). These authors coupled the F-madogram with Partitioning Around Medoids (PAM) algorithm (Kaufman and Rousseeuw, 1990) to form homogeneous regions.

It was applied on various climatic variables such as temperature seasonal maxima (Bador et al., 2015) and precipitation annual maxima (Saunders, Stephenson and Karoly, 2020).

The main limit of this methodology is that, as all marginals were transformed to follow unit uniforms, all the information about the scale invariance requirement, a key feature of the RFA, was ignored. This leads us to adapt this approach to the RFA context.

2. **Methodology:**

Two methods will be presented.

2.1 Regular variation case

The first one is adapted to annual maxima (block of one year) and applied to yearly maxima of daily precipitation from 16 CMP6 global models. A new type of F-madogram is introduced and this dissimilarity incorporates a cost whenever the marginal invariance requirement is not satisfied. The asymptotic properties of the associated estimate are studied. Then, this dissimilarity is applied to classical clustering algorithms like PAM and means. This method is adapted to max-stable structures (regular variation case), see Le Gall et al. (2021).

But, it may inefficient for multivariate vectors that are asymptotically independent in the upper tails (hidden regular variation case).

2.1 Hidden regular variation case

We propose a dissimilarity taking into account on the one hand the proximity of the marginal laws of the excesses. The proximity of marginal distributions is evaluated using the non-parametric Kullback-Leibler estimator of excesses proposed by Naveau et al. (2014).

The bivariate dependence is estimated via the tail dependence coefficient studied by Ledford and Tawn (1996).

A new dissimilarity is then proposed and studied by making a convex combination of the non-parametric Kullback-Leibler and the tail dependence coefficient (see Zaffran and Naveau, 2021).

3. Result:

In this section, we focus on our main application: the analysis of yearly maxima of daily precipitation from 16 CMP6 global models.

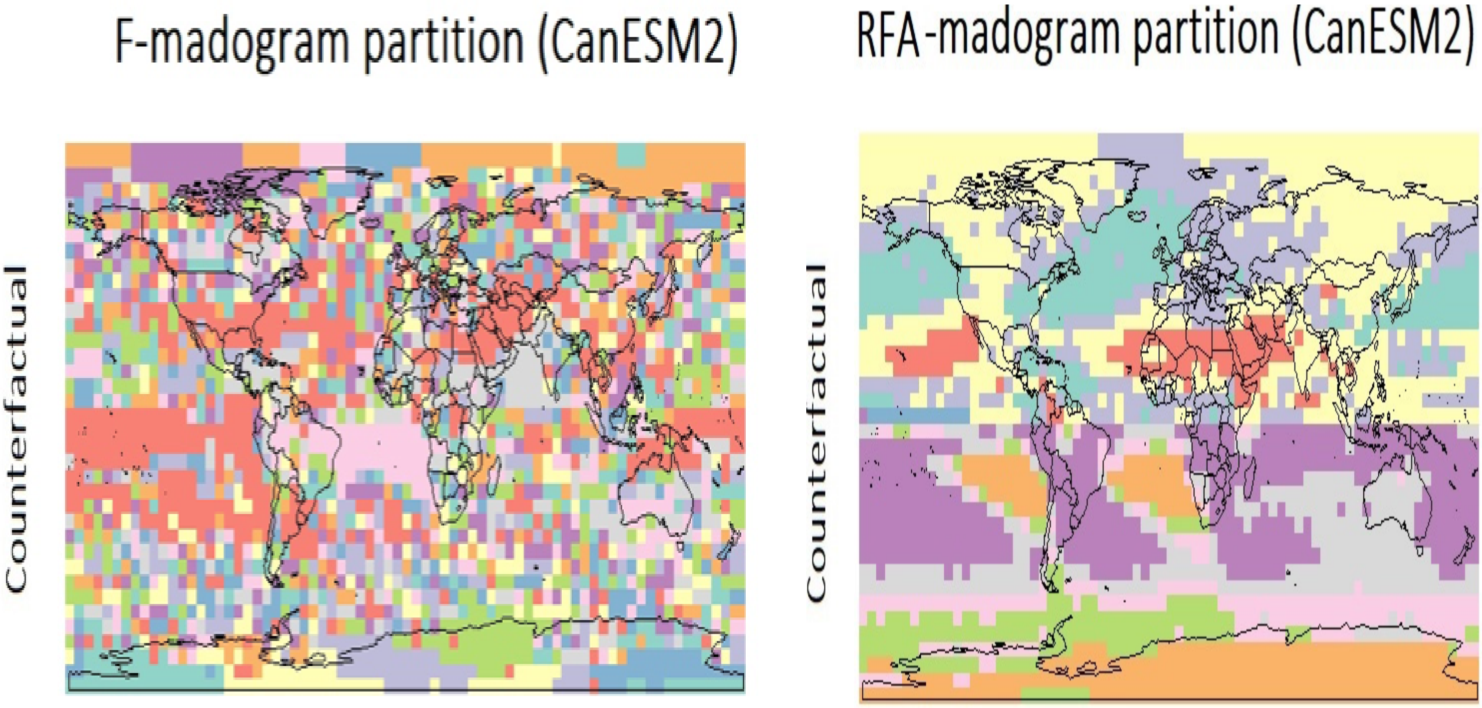
Data description

Global climate model outputs like any numerical simulations correspond to an approximation of the true system under study, here the climate system.

In the realm of Detection and Attribution (D\&A), either in a transient setup or in the context of extreme event attribution (EEA), numerous review studies , see e.g. Naveau et al. (2020), listed different sources of variability, uncertainties and errors. In particular, these reviews highlighted that model error in numerical experiments like the Coupled Model Intercomparison Project (CMIP) can be large and has to be taken into account in any D\&A statistical analysis}. This research field aims at answering questions related to {\it relative} changes between two worlds.In D\&A with transient runs, the two worlds correspond to global coupled climate runs with all forcings (ALL) and with only natural forcings (NAT), respectively.

To find relative changes, it is important to identify homogenous regions in the factual and counterfactual worlds.

For example, the figure below compares the 10 clusters obtained from the counterfactual extreme rainfall of the Canadian model. The left panel, by only taking into account the dependence, provides a “noisy” clustering. In contrast, the right panel has been obtained from our new F-madogram that penalizes a marginal behavior (scale invariance). The obtained partionning provides then spatially and climatologically coherent subregions.


F-madogram partition (CanESM2)
RFA-madogram partition (CanESM2)
Counterfactual
Counterfactual

The analysis for each of the 16 CMIP factual and counterfactual worlds is available upon request and will be shown. If time allowed, an algorithm will presented to combine all these clsuters (to gain in robustness).

2. Discussion and Conclusion:

In this talk, the main goal is to present fast, easy-to-implement and tailored for extremes clustering algorithms. A key feature is to impose scale invraince (RFA) within each cluster, in particular the upper tail index in each cluster should be identical.
The main application is the study of heavy rainfall, either recorded by weatehr stations (not shown here) or from global climate models under different scenarios.

Bador, M., Naveau, P., Gilleland, E., Castella`, 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(20):7929–7937.

Boucefiane, A. and Meddi, M. (2019). Regional growth curves and extreme precipitation events estimation in the steppe area of northwestern Algeria. Atmo´sfera, 32(4):287–303.

Carreau, J., Naveau, P., and Neppel, L. (2017). Partitioning into hazard subregions for regional peaks-over- threshold modeling of heavy precipitation. Water Resources Research, 53(5):4407–4426.

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

DALRYMPLE, T. (1960). Flood-frequency analyses, manual of hydrology: Part 3 Technical Report, USGPO,.

Davison, A. C. and Huser, R. (2015). Statistics of extremes. Annual Review of Statistics and its Application, 2:203–235.

de Fondeville, R. and Davison, A. C. (2018). High-dimensional peaks-over-threshold inference. Biometrika, 105(3):575–592.

DREES, H. and SABOURIN, A. (2019). Principal component analysis for multivariate extremes.arXiv preprintarXiv:1906.11043.

Evin, G., Favre, A.-C., and Hingray, B. (2018). Stochastic generation of multi-site daily precipitation focusing on extreme events. Hydrology and Earth System Sciences, 22(1):655–672.

Ferreira, A. and de Haan, L. (2015). On the block maxima method in extreme value theory: PWM estimators. The Annals of Statistics, 43(1):276–298.

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(5):1049–1054.

Hosking, J. and Wallis, J. (1993). Some statistics useful in regional frequency analysis. Water resources research, 29(2):271–281.

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

JANSSEN, A., WAN, P. et al. (2020).k-means clustering of extremes.Electronic Journal of Statistics141211–1233.

Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data Clustering: A Review. ACM Comput. Surv., 31(3):264–323.

KATZ, R. W., PARLANGE, M. B. and NAVEAU, P. (2002). Statistics of extremes in hydrology.Advances in WaterResources251287–1304.

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)6899–120.

Jalbert, J., Favre, A.-C., B´elisle, C., and Angers, J.-F. (2017). A spatiotemporal model for extreme precipitation simulated by a climate model, with an application to assessing changes in return levels over North America. Journal of the Royal Statistical Society: Series C (Applied Statistics), 66(5):941–962.

Kaufman, L. and Rousseeuw, P. J. (1990). Finding groups in data: an introduction to cluster analysis, volume 344. John Wiley & Sons.

Le Gall, Naveau P, Favre A.C, Tuel A (2021). Non-parametric Regional Frequency Analysis (personal communication)

Li, M., Li, X., and Ao, T. (2019). Comparative Study of Regional Frequency Analysis and Traditional At-Site Hydrological Frequency Analysis. Water, 11(3):486.

Malekinezhad, H. and Zare-Garizi, A. (2014). Regional frequency analysis of daily rainfall extremes using L-moments approach. Atmo´sfera, 27(4):411 – 427.

NAVEAU, P., GUILLOU, A., COOLEY, D. and DIEBOLT, J. (2009). Modelling pairwise dependence of maximain space.Biometrika 961–17.

Naveau, P. et al. (2014). “A non-parametric entropy-based approach to detect changes inclimate extremes”. In :Journal of the Royal Statistical Society : Series B (StatisticalMethodology)76.5, p. 861-884.doi:10.1111/rssb.12058.

Naveau, P., A. Hannart, and A. Ribes, 2020: Statistical methods for extreme event attribution in climate science. *Annual Review of Statistics and Its Application*, **7 (1)**, 89–110,

Naveau, P., A. Ribes, F. W. Zwiers, A. Hannart, A. Tuel, and P. Yiou, 2018: Revising return periods for record events in a climate event attribution context. *Journal of Climate*, **31**, 3411–3422.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics, 20:53–65.

Saunders, K. (2018). An investigation of Australian rainfall using extreme value theory. PhD thesis, Melbourne University.

SAUNDERS, K., STEPHENSON, A. and KAROLY, D. (2020). A regionalisation approach for rainfall based onextremal dependence.Extremes1–26.

TAWN, J. A. (1988). Bivariate extreme value theory: models and estimation.Biometrika75397–41

Viglione, A., Laio, F., and Claps, P. (2007). A comparison of homogeneity tests for regional frequency analysis. Water Resources Research, 43(3). W03428, doi:10.1029/2006WR005095.

Zaffran, M. et P. Naveau (2021). “Spatial clustering of rainfall extremes by couplingKullback-Leibler divergence and tail coefficient”. In :rapport interne du LSCE (àsoumettre prochainement).