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

LSCE

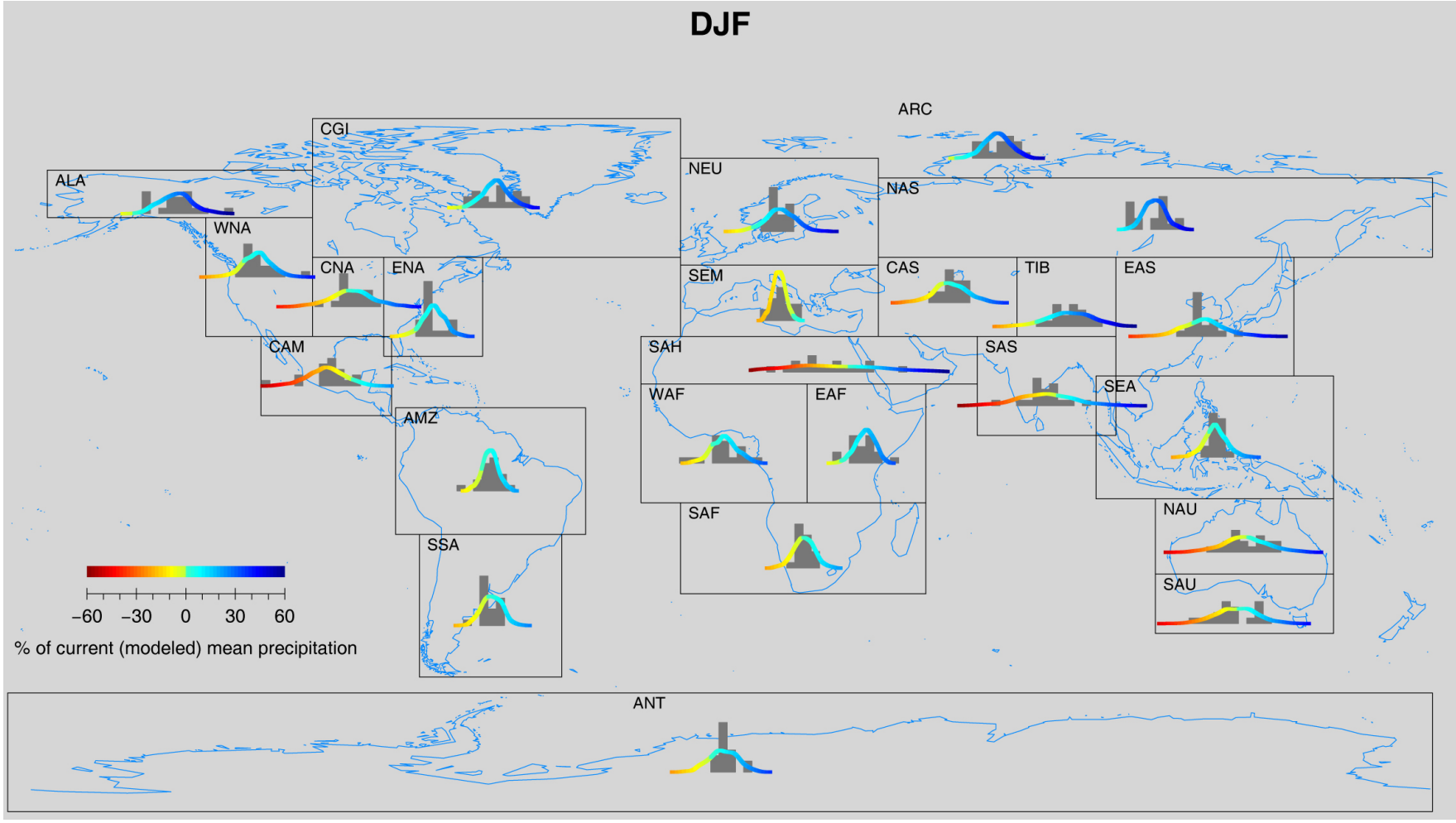
LABORATOIRE DES SCIENCES DU CLIMAT & DE L'ENVIRONNEMENT

philippe.naveau@lsce.ipsl.fr

Joint work with **Philomène Le Gall** (LSCE,IGE), Anne-Catherine Favre (IGE, Grenoble), Alexandre Tuel (Bern university)

Fundings: Xaida H2020, ANR Melody & T-REX, 80 Prime CNRS
arXiv:2111.00798 or hal-03409908v1

Classical regions found in the IPCC report



Motivation

Climatology

- How to identify regional clusters for multi-model heavy rainfall (CMIP) ?
- How to see changes in spatial rainfall structures due to anthrogenic forcings ?

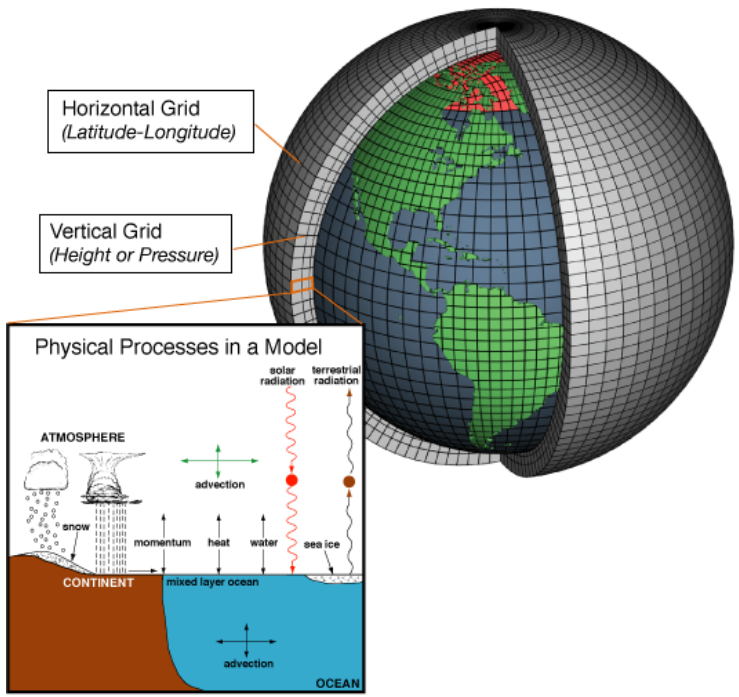
Statistics

- How to cluster spatially regions non-parametrically, without covariate and in compliance with EVT ?
- How to compare clusters under different RCP scenarios ?

Main tools

Climatology

- Global numerical physically based climate model (CMIP database)

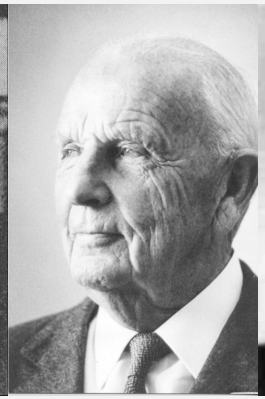


Main statistical tools

Univariate Extreme Value Theory

$$\lim_{n\to\infty} Pr(M_n \leq a_n x + b_n) = \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]_+^{-1/\xi}\right\}$$

Gumbel (1891-1966)



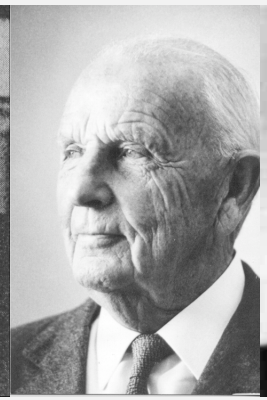
Weibull (1887-1979)

Fréchet (1878-1973)

Main statistical tools

Univariate Extreme Value Theory

$$\lim_{n\to\infty} Pr(M_n \leq a_n x + b_n) = \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]_+^{-1/\xi}\right\}$$


Gumbel (1891-1966) **Weibull (1887-1979)** **Fréchet (1878-1973)**

Multivariate

Let $\mathbf{M}_n := (M_{n,1}, \ldots, M_{nd})$ with $M_{nj} := \max(Y_{1j}, \ldots, Y_{nj})$

$$\mathbb{P}\left[\frac{\mathbf{M}_n - \mathbf{b}_n}{\mathbf{a}_n} \leq \mathbf{x}\right] = \mathbb{P}^n\left(\mathbf{Y} \leq \mathbf{a}_n\mathbf{x} + \mathbf{b}_n\right) \xrightarrow{\mathrm{d}} MGEV(\mathbf{x}), \qquad \text{as } n \to \infty.$$

see, e.g. Statistics of Extremes, Davison and Huser Annual Review of Statistics and Its Application 2015 2 :1, 203-235

Main issues

Climatology

- Global numerical models are approximations and do not capture all scales
- Heavy rainfall strongly vary in space and time and are heavy tailed

Main issues

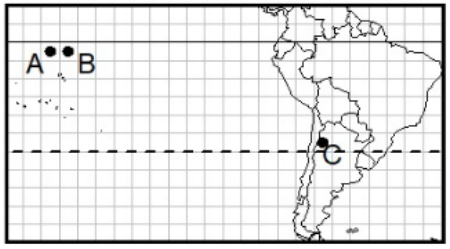
Climatology

- Global numerical models are approximations and do not capture all scales
- Heavy rainfall strongly vary in space and time and are heavy tailed

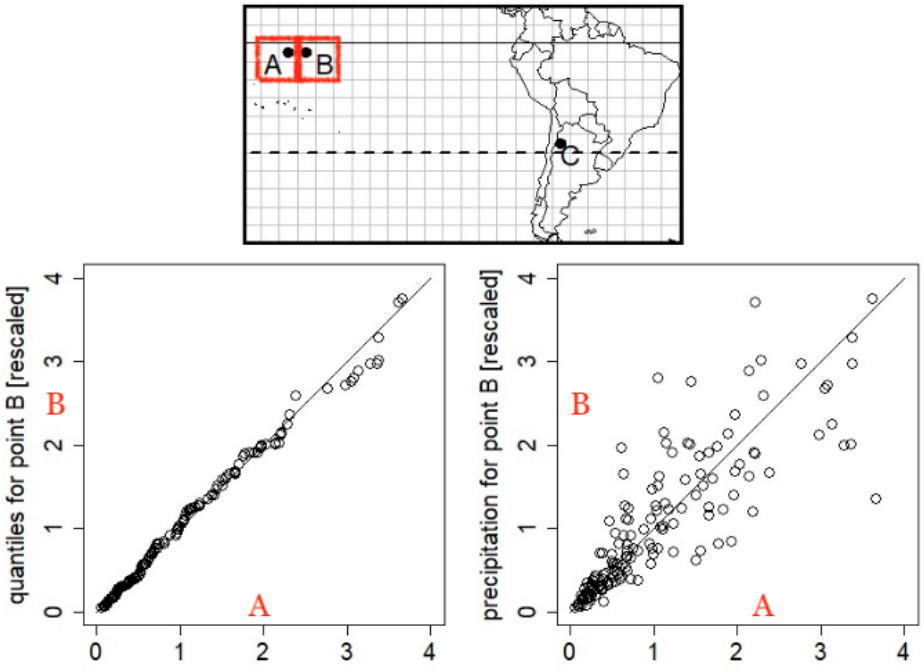
Statistics

- Pareto tail parameters are difficult to estimate in a non-stationary spatio-temporal context

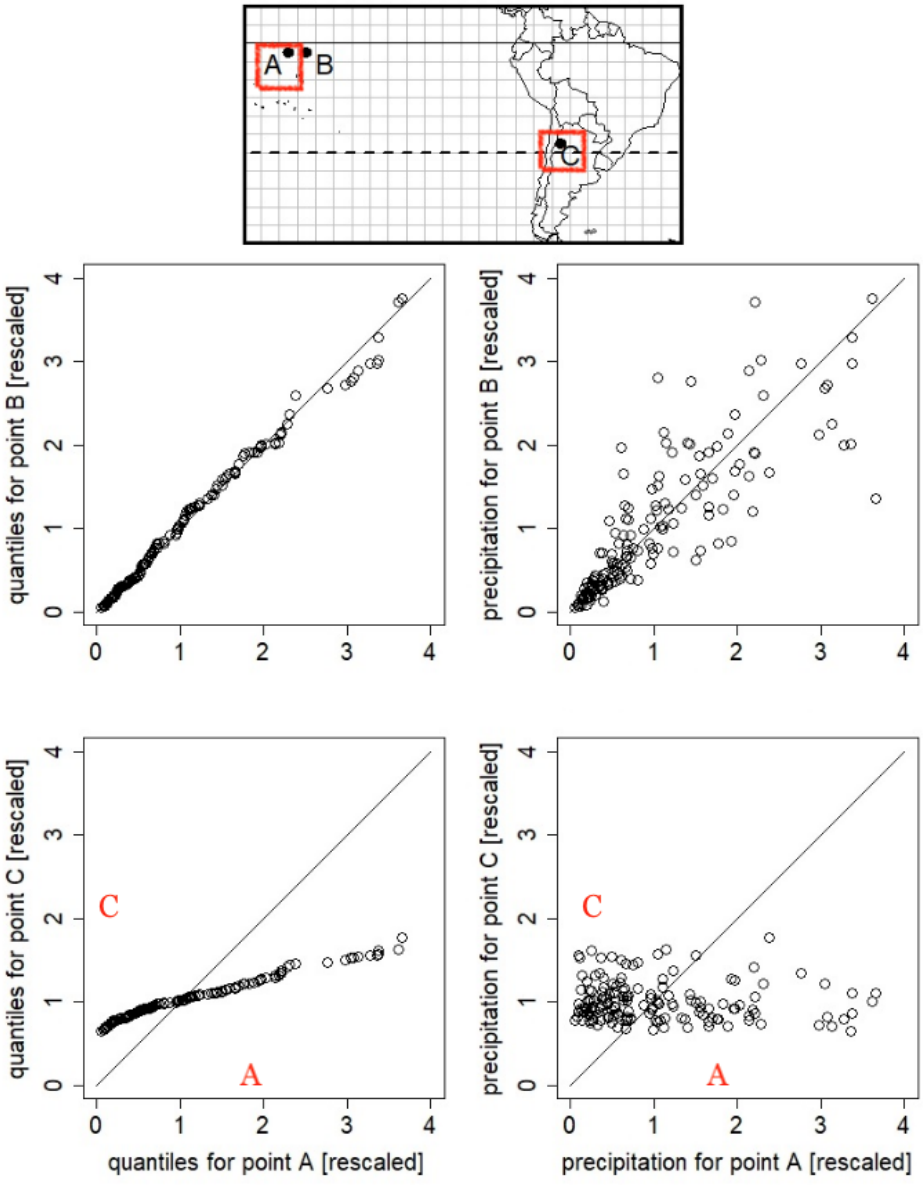
Example : zoom over south America



Example : zoom over south America



Example : zoom over south America



Regional regional analysis

- How to find grid point clusters that are “homogeneous", ie.

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

where Y_1 and Y_2 two positive continuous r.v., i.e. as $F_2(\lambda x) = F_1(x)$

see e.g., Le Gall et al. (2021). Improved Regional Frequency Analysis of rainfall data (Hal).
Carreau et al. (2017). Partitioning into hazard subregions for regional peaks-over- threshold modeling of heavy precipitation. WRR.
Hosking and Wallis (2005). Regional frequency analysis : an approach based on L-moments. Cambridge University Press.

See also Bobbia, Dombry and Varron (2021, proportional tail) or Daouiaa, Padoan and Stupfler (2021, pooling).

Regional regional analysis

- How to find grid point clusters that are “homogeneous", ie.

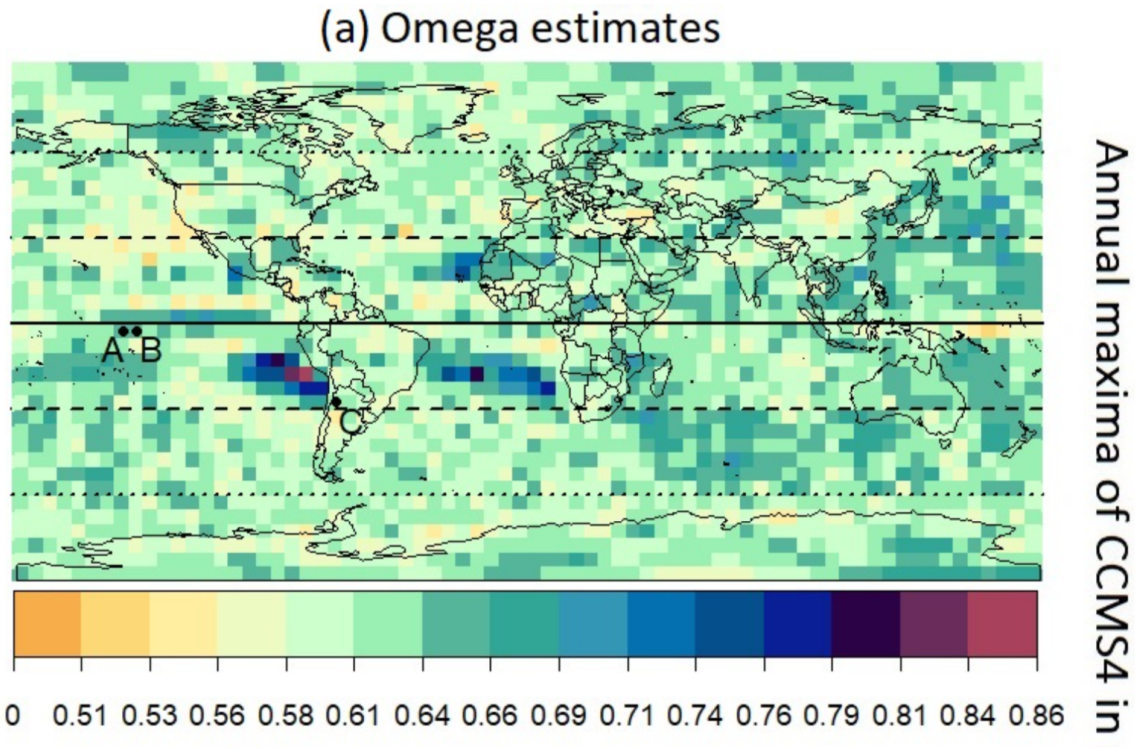
$$Y_2 \overset{\mathrm{d}}{=} \lambda Y_1,$$

where Y_1 and Y_2 two positive continuous r.v., i.e. as $F_2(\lambda x) = F_1(x)$

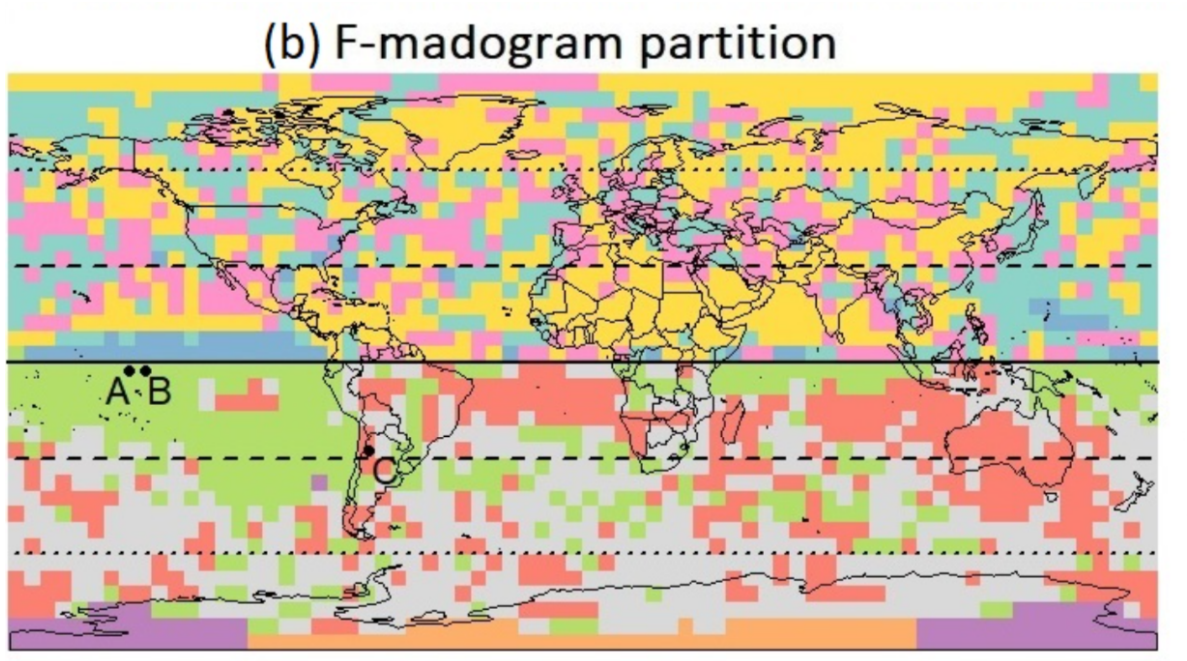
see e.g., Le Gall et al. (2021). Improved Regional Frequency Analysis of rainfall data (Hal).
Carreau et al. (2017). Partitioning into hazard subregions for regional peaks-over- threshold modeling of heavy precipitation. WRR.
Hosking and Wallis (2005). Regional frequency analysis : an approach based on L-moments. Cambridge University Press.

See also Bobbia, Dombry and Varron (2021, proportional tail) or Daouiaa, Padoan and Stupfler (2021, pooling).

Marginal behaviors : $\omega(GEV(\mu, \sigma, \xi) = \frac{3^\xi - 1}{2^\xi - 1} - 1$



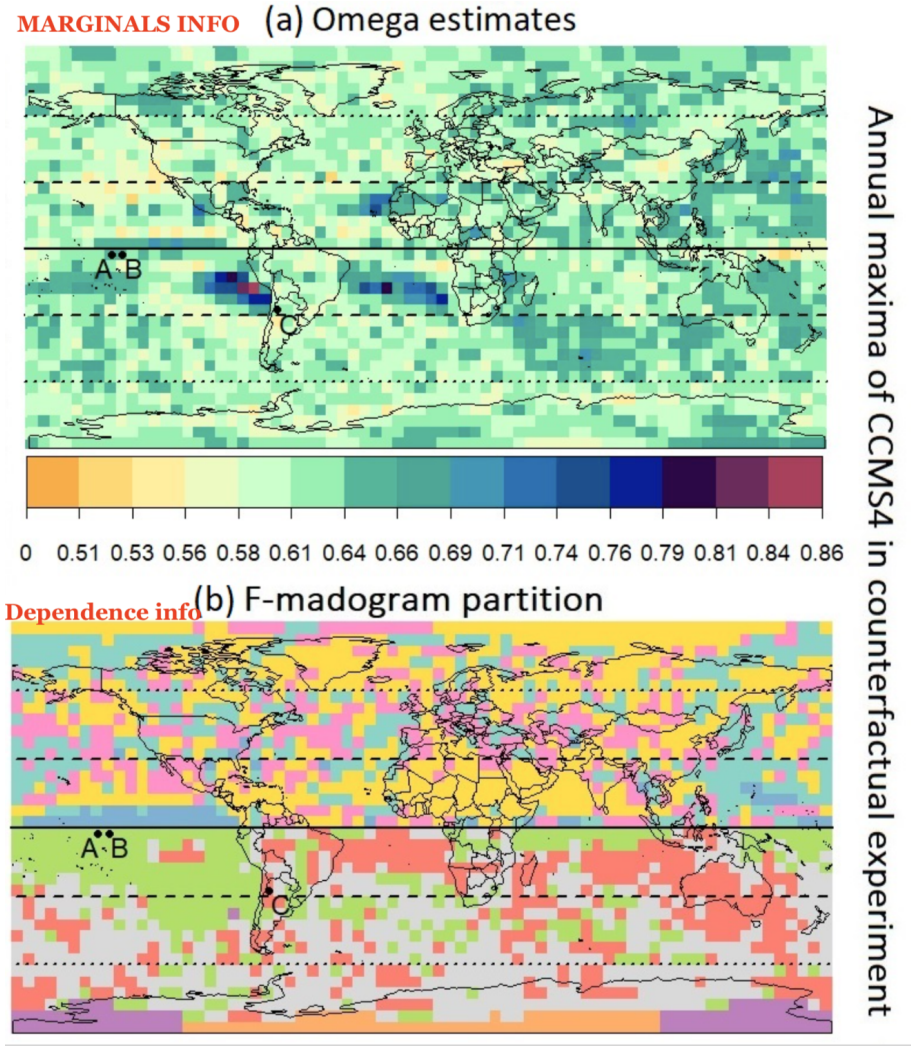
PAM clustering based on dependence $d = \frac{1}{2}\mathbb{E}\left|F_1\left(Y_1\right) - F_2\left(Y_2\right)\right|$



see, e.g. Saunders et al.(2020). A regionalisation approach for rainfall based on extremal dependence, Extremes.

Bador et al. (2015). Spatial clustering of summer temperature maxima from the CNRM-CM5 climate model ensembles & E-OBS over Europe. Weather and climate extremes.

Decoupling the pb into : “marginals" versus "dependence"




Our new dissimilarity

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

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

New Dissimilarity

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

Recall RFA constraint : $\exists\lambda, \quad F_2(\lambda x) = F_1(x)$

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

New Dissimilarity

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

Recall RFA constraint : $\exists\lambda, \quad F_2(\lambda x) = F_1(x)$

Optimization step

$$c^* = \text{argmin}\{D(c) : c > 0\}.$$

New Dissimilarity

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

Inference

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

Convergence under smoothness copula condition

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

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

with $a_c(u) = F_2(cF_1^{\leftarrow}(u))$ and $\mathrm{Cov}(\mathbb{D}(\boldsymbol{u}),\mathbb{D}(\boldsymbol{v})) = C(\boldsymbol{u}\wedge\boldsymbol{v}) - C(\boldsymbol{u})C(\boldsymbol{v})$

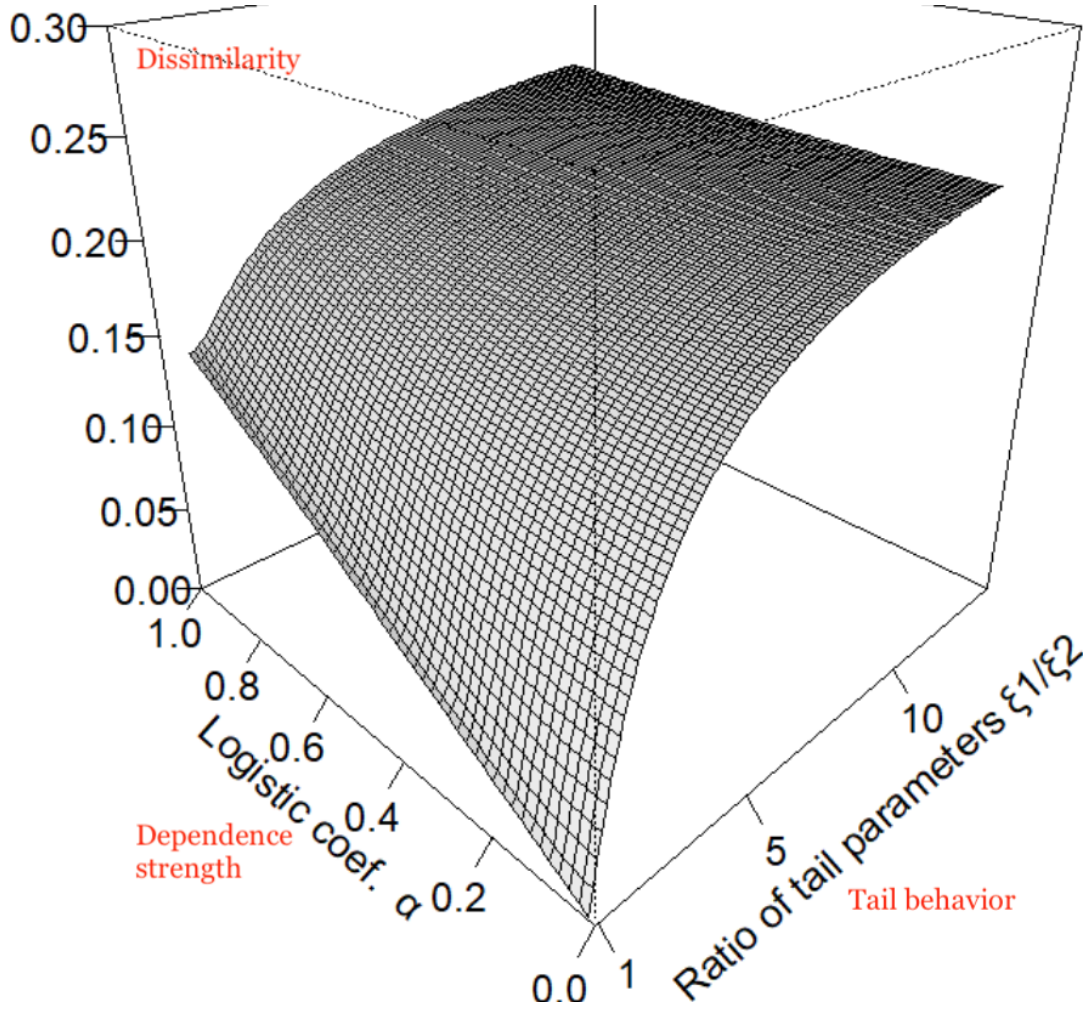
see, e.g., Multivariate nonparametric estimation of the Pickands dependence function using Bernstein polynomials. Marcon, Padoan, PN, Muliere, Segers. Journal of Statistical Planning and Inference, 2017, 183, 1-17

An example (bi-logistic max-stable model)

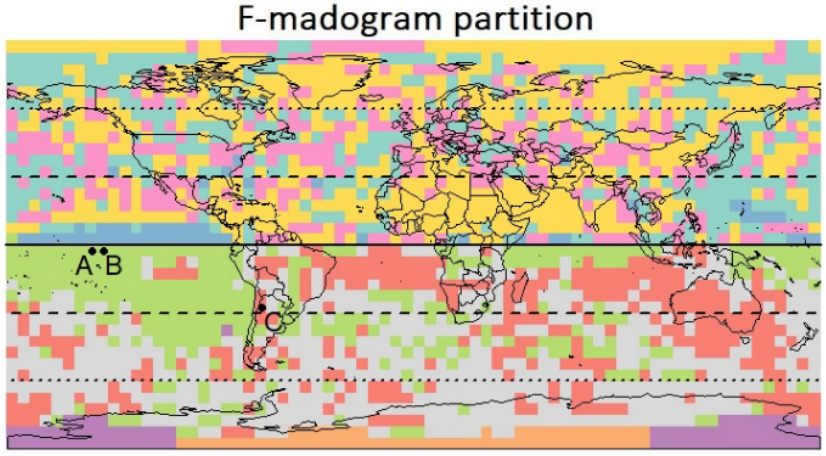
$$F(y_1, y_2) = \exp\left\{-V\left[\frac{-1}{\log F_1(x)}, \frac{-1}{\log F_2(y)}\right]\right\} \text{ \& } V(x, y) = \left(x^{-\frac{1}{\alpha}} + y^{-\frac{1}{\alpha}}\right)^{\alpha}$$

An example (bi-logistic max-stable model)

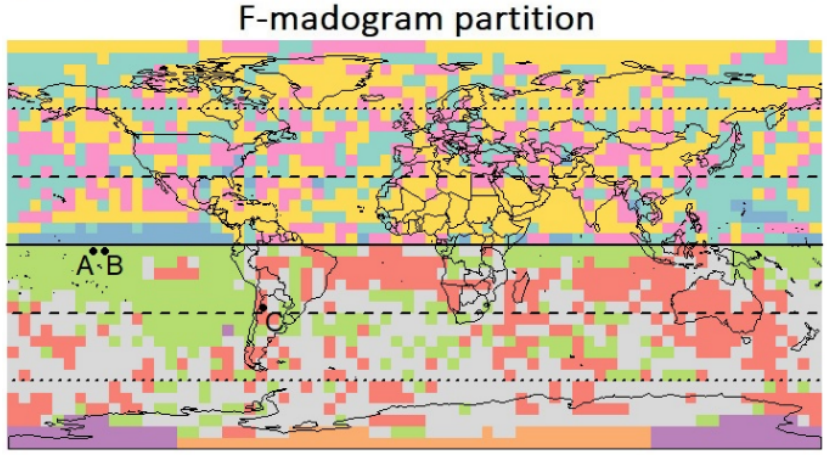
$$F(y_1, y_2) = \exp\left\{-V\left[\frac{-1}{\log F_1(x)}, \frac{-1}{\log F_2(y)}\right]\right\} \text{ \& } V(x,y) = \left(x^{-\frac{1}{\alpha}} + y^{-\frac{1}{\alpha}}\right)^{\alpha}$$



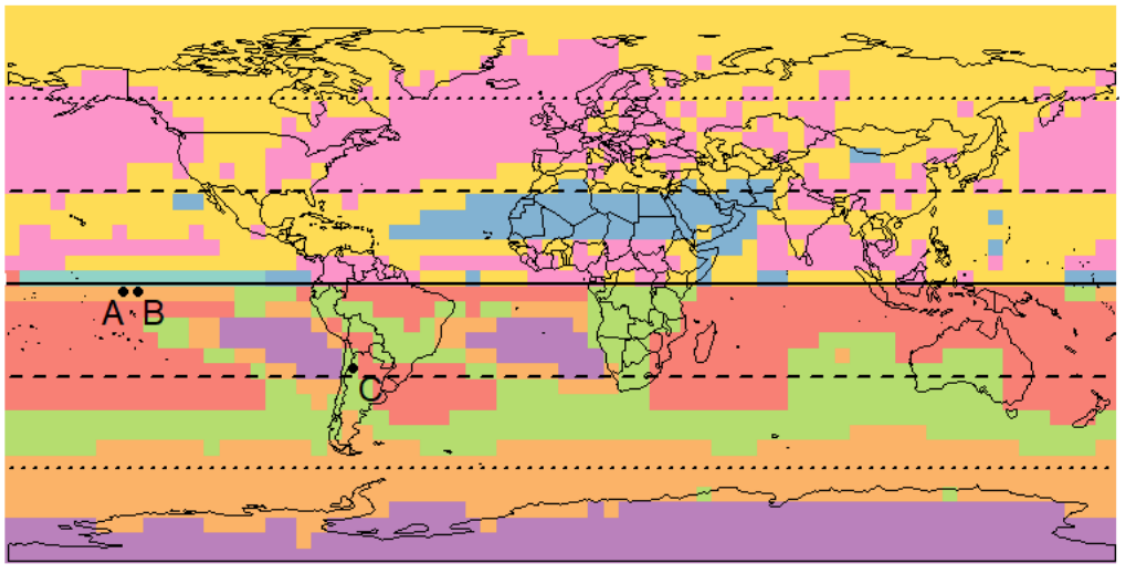
PAM clustering based on dependence : $d = \frac{1}{2}\mathbb{E}\left|F_1\left(Y_1\right) - F_2\left(Y_2\right)\right|$



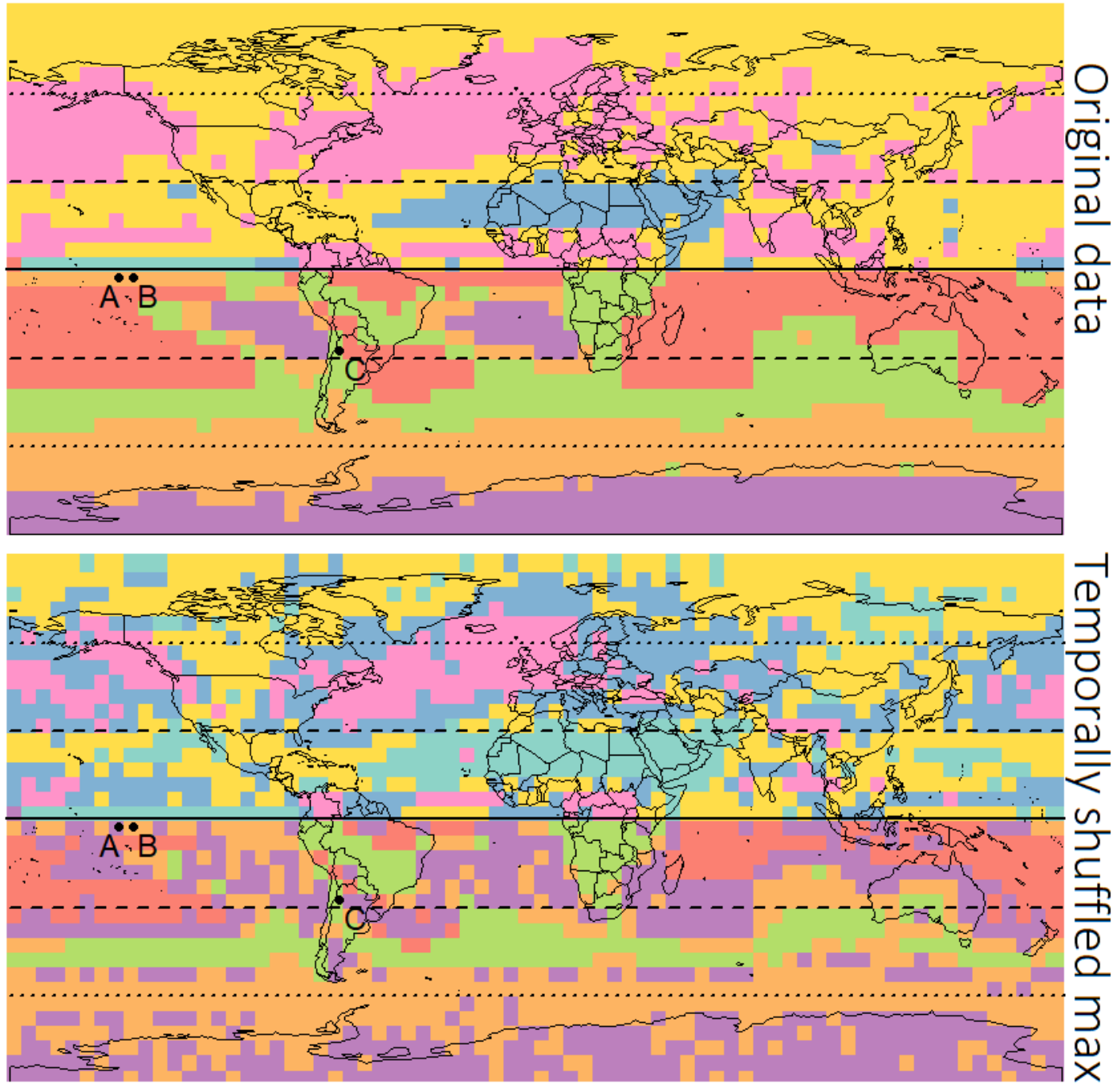
PAM clustering based on dependence : $d = \frac{1}{2}\mathbb{E}\left|F_1\left(Y_1\right) - F_2\left(Y_2\right)\right|$



PAM clustering based on $D(c) = \frac{1}{2}\mathbb{E}\left|F_2\left(cY_1\right) - F_1\left(\frac{Y_2}{c}\right)\right|$



Impact of the spatial dependence

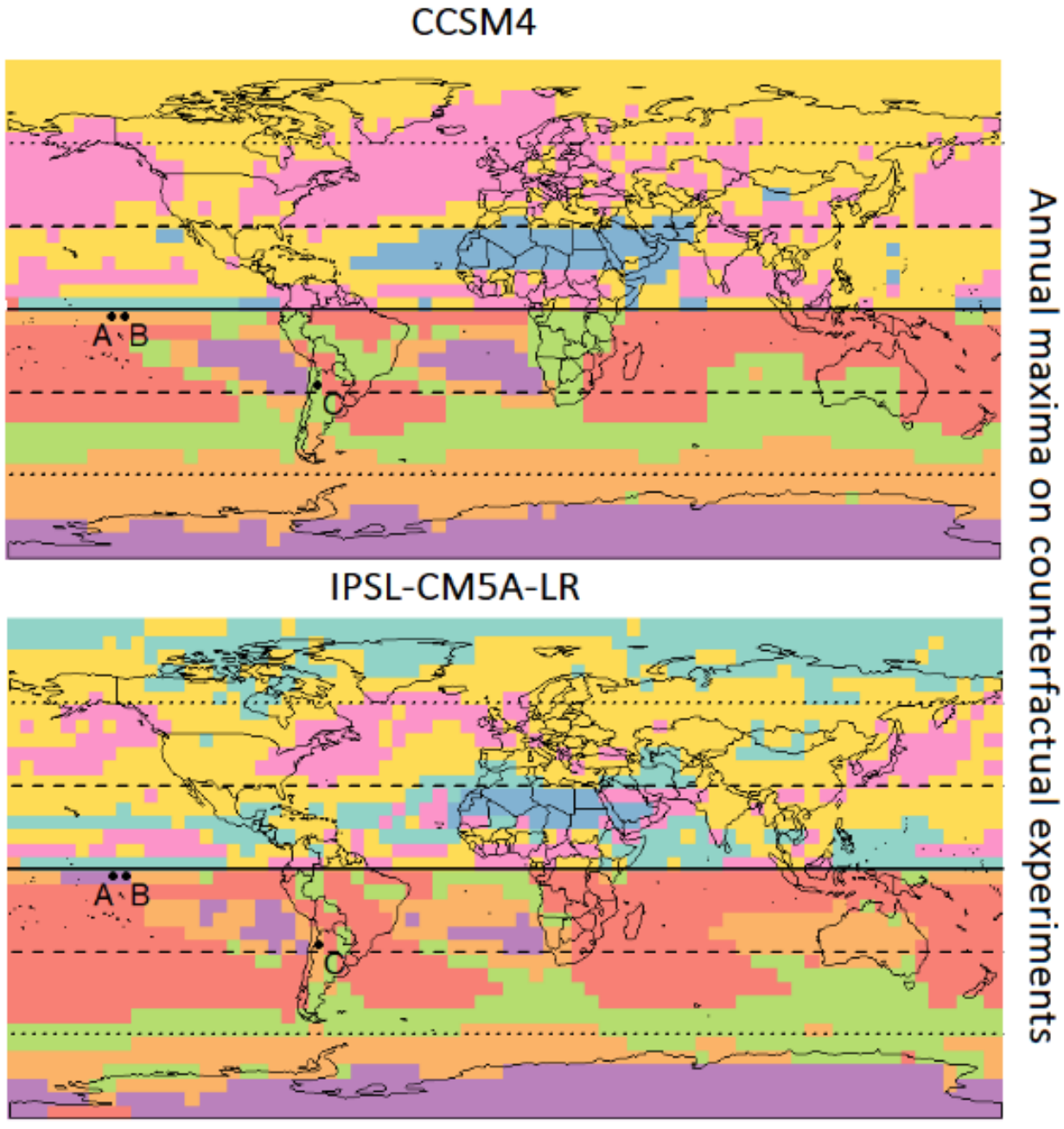


Climate multi-model error

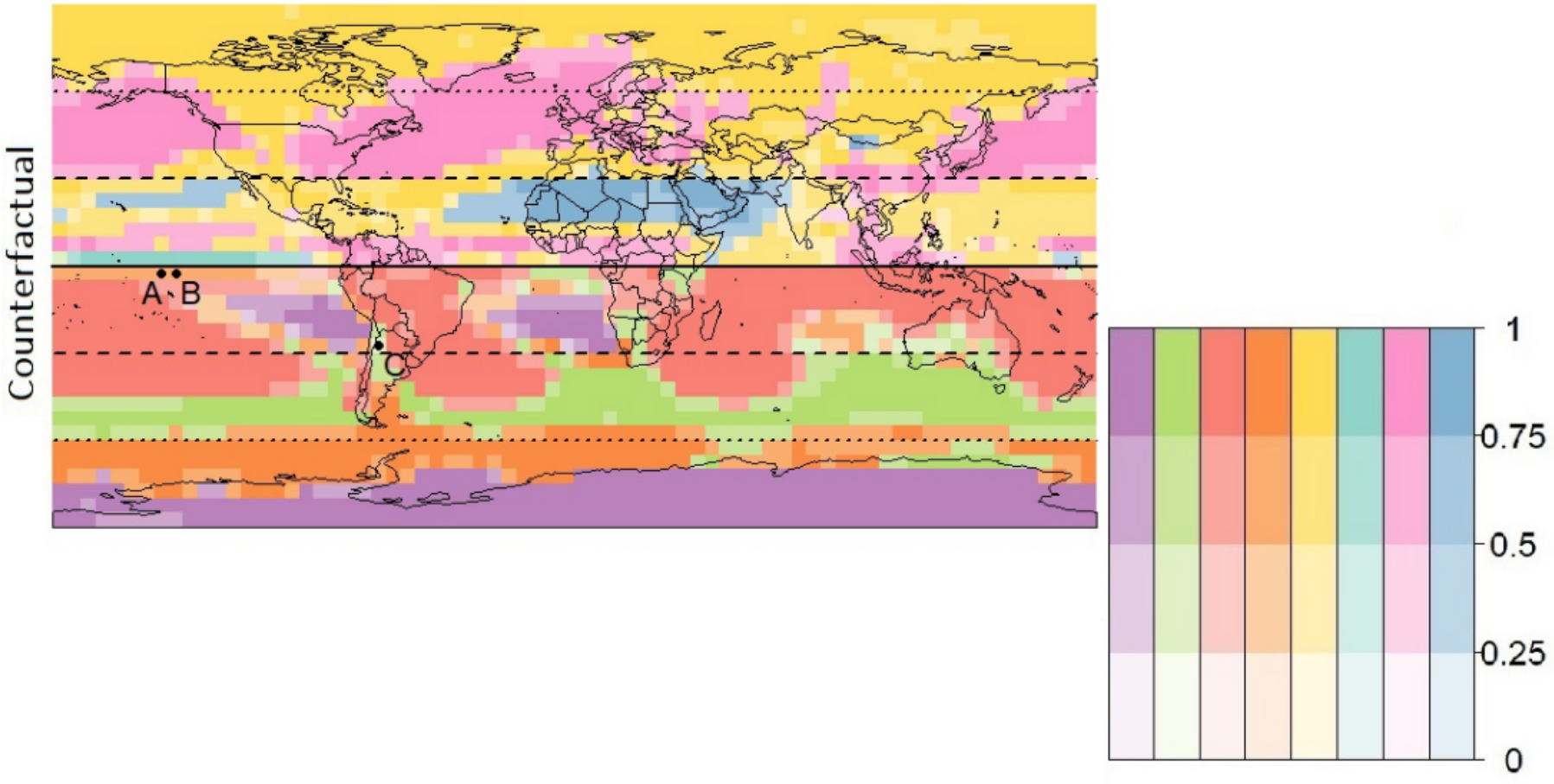
16 climate models

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

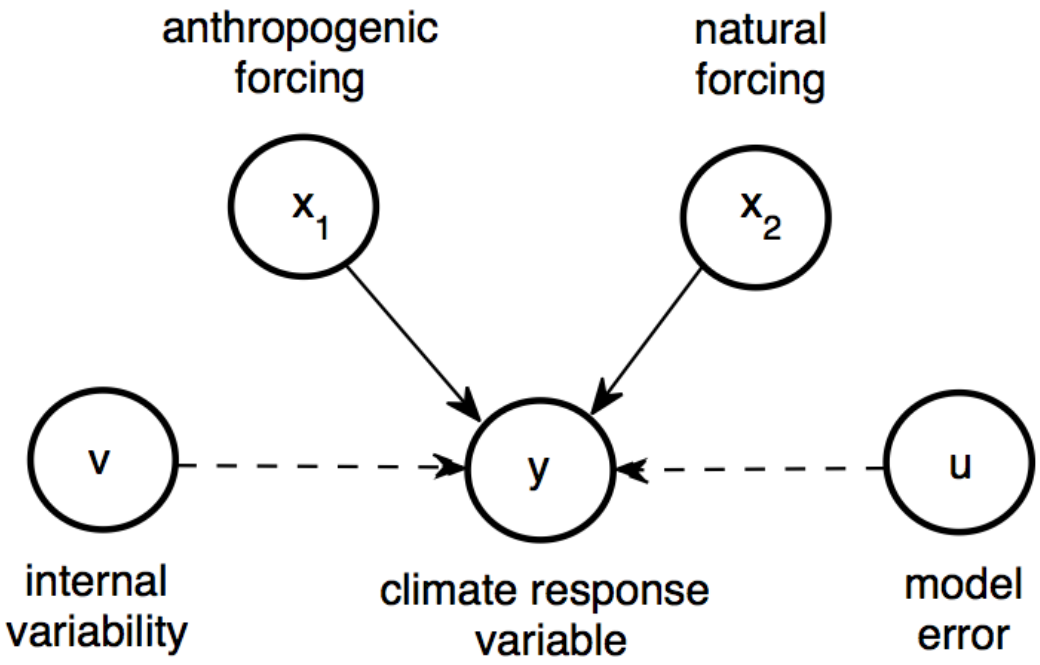
Two out of 16 climate models



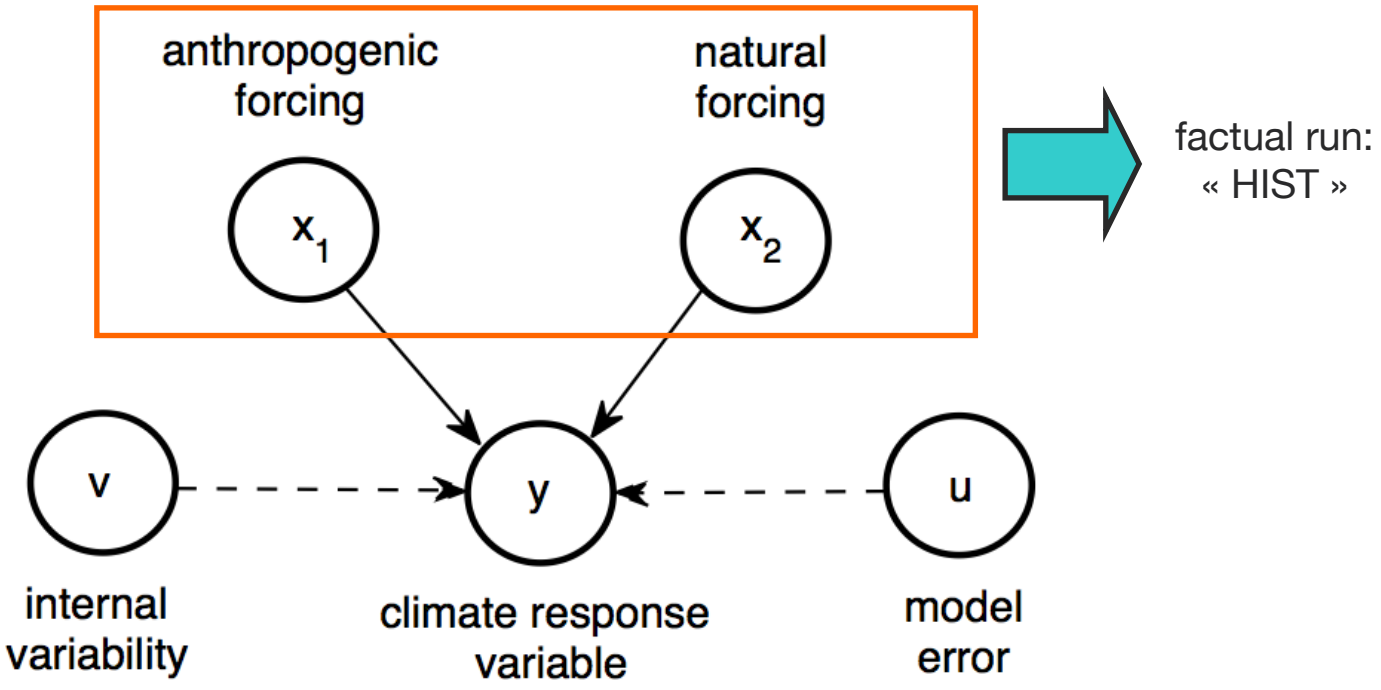
Merging all 16 climate models



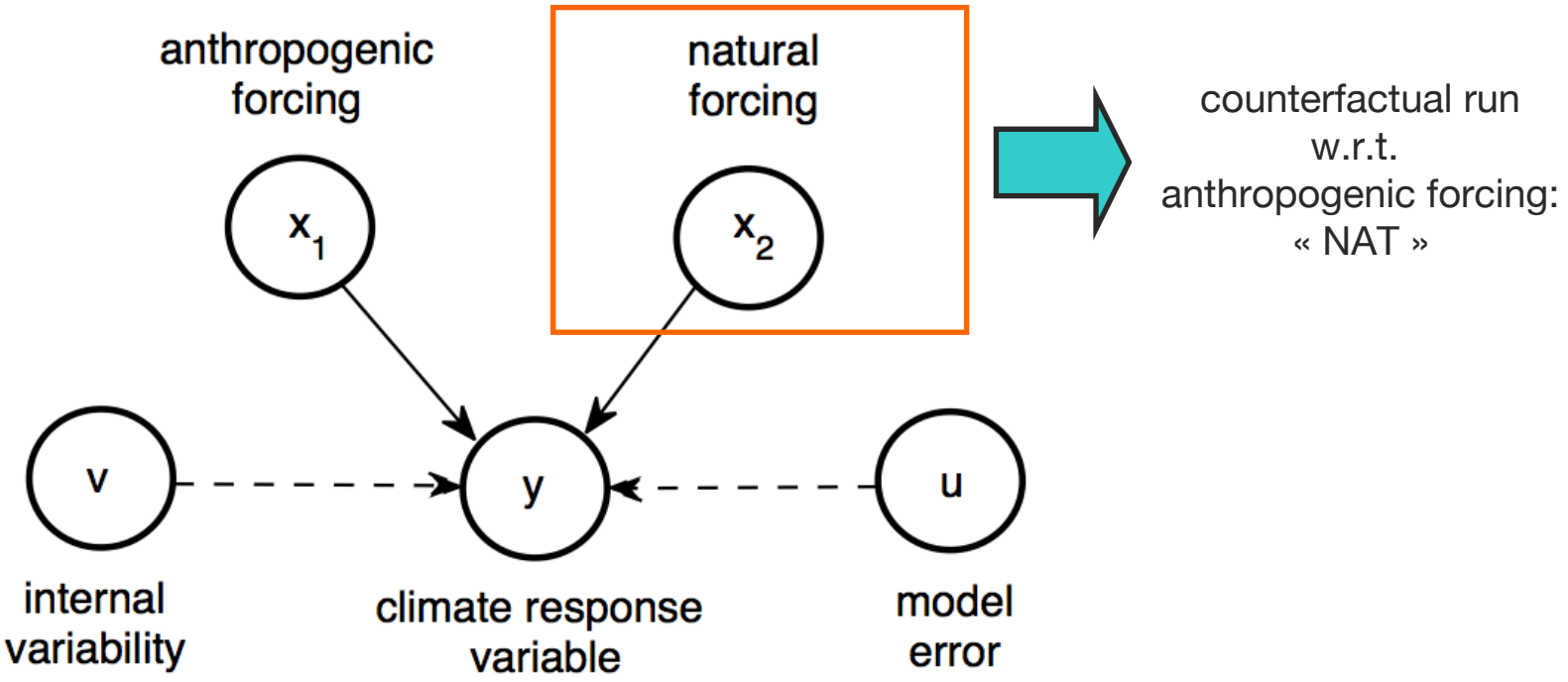
Climate change



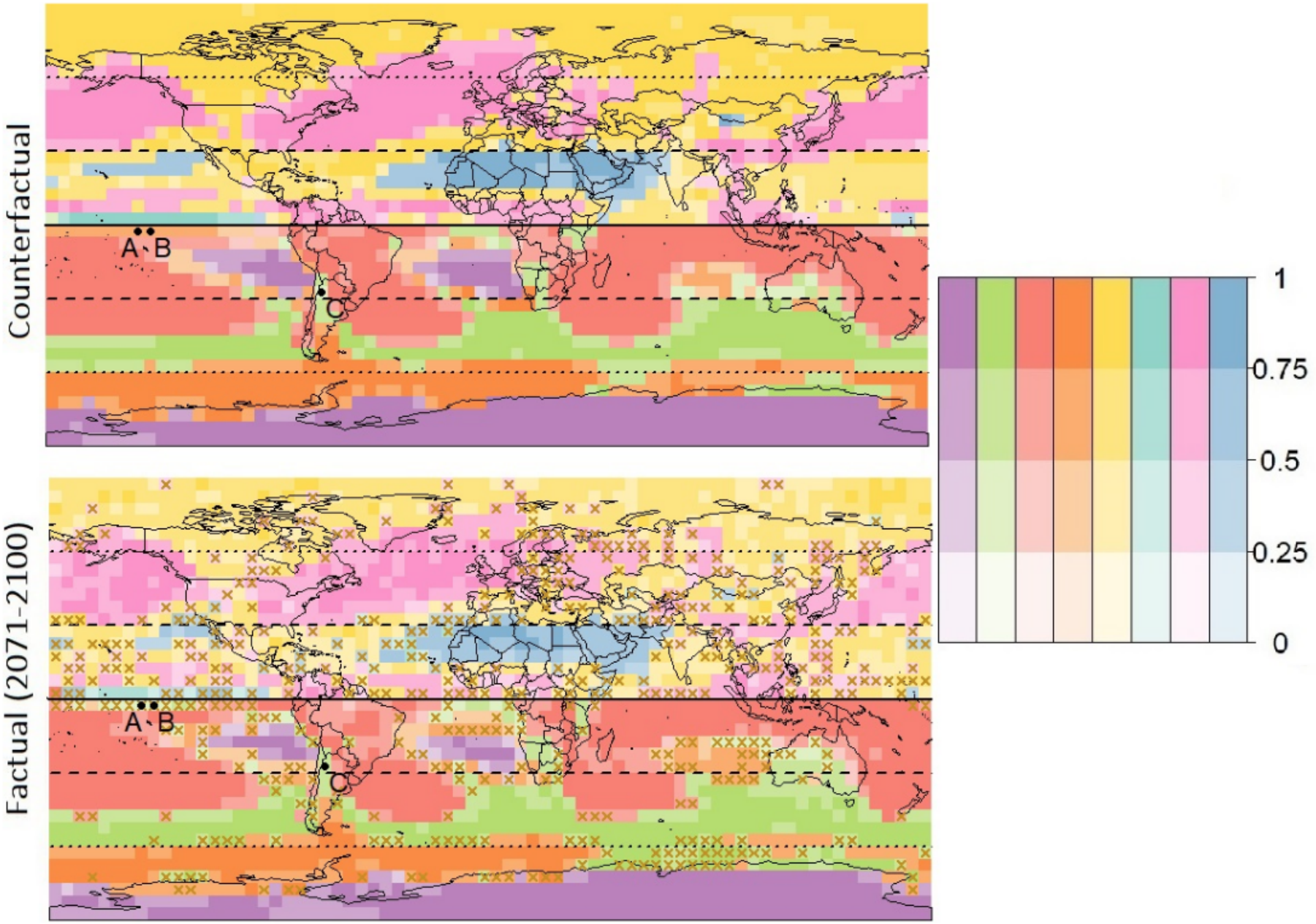
Factual world



Counterfactual world



Comparing the merging of all 16 climate models under climate change



Conclusions

Main messages

- Fast and simple dissimilarity in compliance with Extreme Value Theory and the RFA constraint
- Provide coherent spatial clusters over a large dataset of global climate models
- Today climate produce less stable clusters compared to a counterfactual world without anthropogenic forcings
- Global models have a crude spatial resolution scales, so a small move by the cluster can have a large impact

Conclusions

Main messages

- Fast and simple dissimilarity in compliance with Extreme Value Theory and the RFA constraint
- Provide coherent spatial clusters over a large dataset of global climate models
- Today climate produce less stable clusters compared to a counterfactual world without anthropogenic forcings
- Global models have a crude spatial resolution scales, so a small move by the cluster can have a large impact

Future climatological work

- How to regions will change with regional climate models (finer scale)
- How to integrate rainfall observations in the analysis

Future statistical work

- Dealing with the hidden regular variation case
- How to downscale models or upscale observations to find the best optimal scale

Future work on difference sources of error and/or uncertainty in D&A

- Natural climate internal variability
- Natural forcing variations
- Model uncertainty from approximating the true climate system with numerical experiments
- Observational uncertainties due to instrumental errors, homogenization problems and mismatches between data sources
- Sampling uncertainty in space and time
- Statistical modeling error by assuming a specific statistical model, e.g., assuming a generalized extreme value distribution for independent block maxima.
- Inferential uncertainties

PAM (Partioning Around Medoid) by Kaufman and Rousseeuw

Thus, we explain here in details the PAM procedure. The user provides:

- a number of clusters k ;
- a matrix containing all the pairwise dissimilarities, $D \in \mathbb{R}_+^{n,n}$.

The element $D_{i,j}$, for $(i,j) \in [\![1,n]\!]^2$, represents the dissimilarity between point x_i and point x_j.

The aim of the algorithm is to find k medoids solution of:

$$\underset{(m_i, i=1,\dots,k) \subset [\![1,n]\!]}{\operatorname{argmin}} \sum_{j=1}^{n} \min_{m \in (m_i, i=1,\dots,k)} D_{j,m} \tag{2.4}$$

To understand this equation more intuitively, one can see it like this: let's assume we have k medoids, among the n points of the data-set, at our disposal. Each non-medoid point of the data-set is associated to (the cluster of) its closest medoid, in the sense of the dissimilarity matrix D. This is the information contained in $\min_{m \in (m_i, i=1,\dots,k)} D_{j,m}$. This gives us a partition of the data-set. Then, we can define the global cost of any partition, which is simply the sum of the dissimilarities (in the sense of D) between each point and its closest medoid. We see that the partition and its global cost is only determined by the choice of the k medoids. Hence, to minimize the global cost, we are looking for the k points among the n points of the data-set that will achieve the minimal coast possible.