Latent variable models for multi-variable space-time data

and applications in hydrology

Benjamin Renard[1] Mark Thyer[2] David McInerney[2]
Dmitri Kavetski[2] Michael Leonard[2] Seth Westra[2]

[1]INRAE, RiverLy Research Unit, Lyon, France

[2]School of Civil, Environmental and Mining Engineering, University of Adelaide, Australia

AppliBUGS 10 December 2021

Introduction

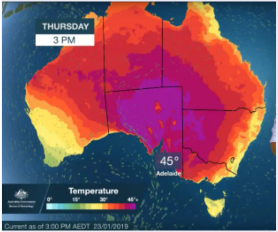
Resources and risk management for environmental systems

Often relies on the analysis of **several variables** measured at **many sites** and whose properties may **vary in time**.

Resources and risk management for environmental systems

Often relies on the analysis of **several variables** measured at **many sites** and whose properties may **vary in time**.

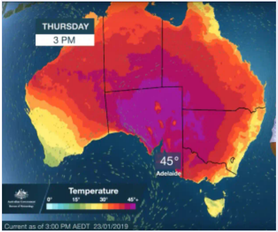
Example: Australian summers



Resources and risk management for environmental systems

Often relies on the analysis of **several variables** measured at **many sites** and whose properties may **vary in time**.

Example: Australian summers




The Sydney Morning Herald

INDEPENDENT. ALWAYS.

20 HOMES AND BUILDINGS DESTROYED IN SOUTH-WEST

THREE FIREFIGHTERS IN HOSPITAL WITH BURNS

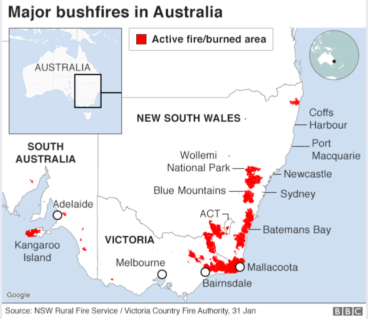
HUGE HEATWAVE TO HIT THE STATE TOMORROW

Fires close in on city

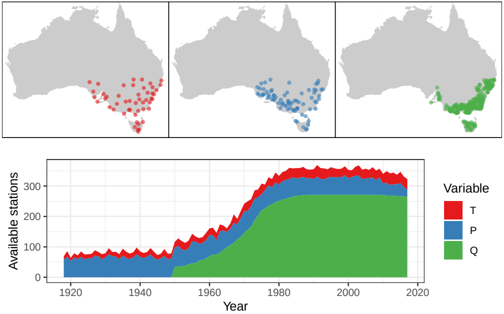
'Hazardous' air 28 days in the past two months

Australia records world's highest ever December temperature

The Premier has declared a state of emergency lasting seven days and warned people not travel near fire areas. FULL COVERAGE PAGES 2-5



Motivating dataset

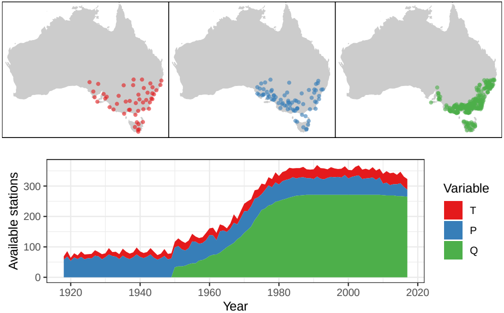


BoM's reference datasets for Temperature, Precipitation and Streamflow.

Variables of interest (DJF):

1. Number of heatwaves Tn
2. Heatwave intensities Tx
3. Dry-day duration Pd
4. Drought duration Qd

Motivating dataset



BoM's reference datasets for Temperature, Precipitation and Streamflow.

Variables of interest (DJF):

1. Number of heatwaves Tn
2. Heatwave intensities Tx
3. Dry-day duration Pd
4. Drought duration Qd

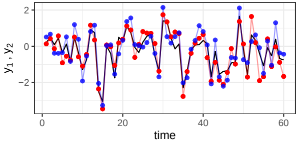
A model for Tn, Tx, Pd and Qd should handle...

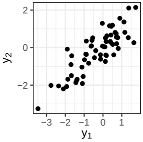
- Spatial dependence
- Inter-variable dependence
- Time variability and/or trend
- Data of different types, missing data

Latent variables as Hidden Climate Indices (HCI)

Variables affected by THE SAME climate index are dependant

$$\begin{cases} \color{red}{Y_{1,t} = \lambda_1 \tau_t + \varepsilon_{1,t}} \\ \color{blue}{Y_{2,t} = \lambda_2 \tau_t + \varepsilon_{2,t}} \\ \varepsilon_{1,t}, \varepsilon_{2,t} \overset{iid}{\sim} \mathcal{N}\left(0, \sigma^2\right) \end{cases}$$

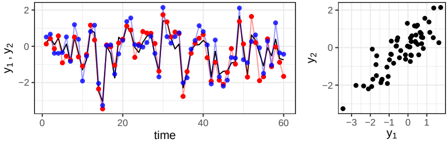




Latent variables as Hidden Climate Indices (HCI)

Variables affected by THE SAME climate index are dependant

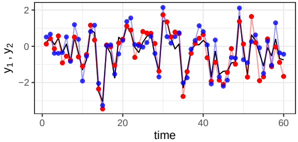
$$\begin{cases} Y_{1,t} = \lambda_1 \tau_t + \varepsilon_{1,t} \\ Y_{2,t} = \lambda_2 \tau_t + \varepsilon_{2,t} \\ \varepsilon_{1,t}, \varepsilon_{2,t} \overset{iid}{\sim} \mathcal{N}\left(0, \sigma^2\right) \end{cases}$$

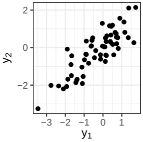


In practice, the time series τ_t is unknown (it is *hidden*).

Variables affected by THE SAME climate index are dependant

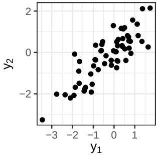
$$\begin{cases} Y_{1,t} = \lambda_1 \tau_t + \varepsilon_{1,t} \\ Y_{2,t} = \lambda_2 \tau_t + \varepsilon_{2,t} \\ \varepsilon_{1,t}, \varepsilon_{2,t} \overset{iid}{\sim} \mathcal{N}\left(0, \sigma^2\right) \end{cases}$$





In practice, the time series τ_t is unknown (it is *hidden*).

Approach 1: explicitly modeling dependence

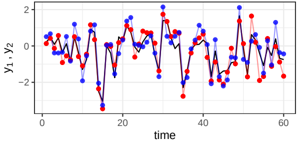


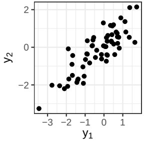
$$\begin{pmatrix} Y_{1,t} \\ Y_{2,t} \end{pmatrix} \overset{iid}{\sim} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right)$$

Latent variables as Hidden Climate Indices (HCI)

Variables affected by THE SAME climate index are dependant

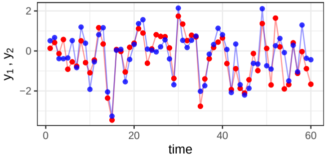
$$\begin{cases} \color{red}{Y_{1,t} = \lambda_1 \tau_t + \varepsilon_{1,t}} \\ \color{blue}{Y_{2,t} = \lambda_2 \tau_t + \varepsilon_{2,t}} \\ \varepsilon_{1,t}, \varepsilon_{2,t} \overset{iid}{\sim} \mathcal{N}\left(0, \sigma^2\right) \end{cases}$$





In practice, the time series τ_t is unknown (it is *hidden*).

Approach 2: uncovering the hidden climate index

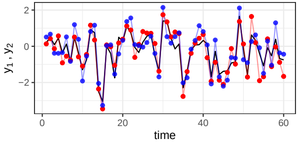


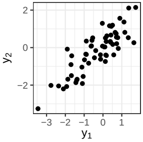
$$\begin{pmatrix} Y_{1,t} \\ Y_{2,t} \end{pmatrix} \overset{i}{\sim} \mathcal{N}\left(\begin{pmatrix} \color{orange}{\lambda_1 \tau_t} \\ \color{orange}{\lambda_2 \tau_t} \end{pmatrix}, \color{orange}{\sigma}^2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$$

Latent variables as Hidden Climate Indices (HCI)

Variables affected by THE SAME climate index are dependant

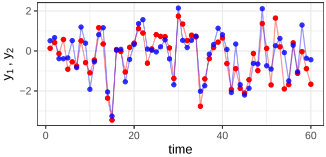
$$\begin{cases} Y_{1,t} = \lambda_1 \tau_t + \varepsilon_{1,t} \\ Y_{2,t} = \lambda_2 \tau_t + \varepsilon_{2,t} \\ \varepsilon_{1,t}, \varepsilon_{2,t} \overset{iid}{\sim} \mathcal{N}\left(0, \sigma^2\right) \end{cases}$$





In practice, the time series τ_t is unknown (it is *hidden*).

Approach 2: uncovering the hidden climate index



$$\begin{pmatrix} Y_{1,t} \\ Y_{2,t} \end{pmatrix} \overset{i}{\sim} \mathcal{N}\left(\begin{pmatrix} \lambda_1 \tau_t \\ \lambda_2 \tau_t \end{pmatrix}, \sigma^2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right)$$

Objective: any distribution, many sites, several variables, several HCIs

General modeling framework

Parent distributions of data

$$Y_v(\mathsf{s}, t) \sim \ \mathcal{D}_v \ \left(\boldsymbol{\theta}_v(\mathsf{s}, t) \right)$$

General modeling framework

Parent distributions of data

$$Y_v(\mathsf{s}, t) \sim \underbrace{\mathcal{D}_v}_{\text{any distribution, variable-specific}} (\,\boldsymbol{\theta}_v(\mathsf{s}, t)\,)$$

General modeling framework

Parent distributions of data

$$Y_v(\mathsf{s}, t) \sim \mathcal{D}_v \left(\underbrace{\boldsymbol{\theta}_v(\mathsf{s}, t)}_{\text{varies in space and time}} \right)$$

General modeling framework

Parent distributions of data

$$Y_v(\mathsf{s}, t) \sim \mathcal{D}_v \left(\boldsymbol{\theta}_v(\mathsf{s}, t) \right)$$

Parameters vary in space and time

$$\boldsymbol{\theta}_v(\mathsf{s}, t) = \lambda_{0,v}(\mathsf{s}) + \lambda_{1,v}(\mathsf{s})\tau_1(t)$$

General modeling framework

Parent distributions of data

$$Y_v(\mathsf{s}, t) \sim \mathcal{D}_v \left(\boldsymbol{\theta}_v(\mathsf{s}, t) \right)$$

Parameters vary in space and time

$$\boldsymbol{\theta}_v(\mathsf{s}, t) = \lambda_{0,v}(\mathsf{s}) + \underbrace{\lambda_{1,v}(\mathsf{s})\tau_1(t)}_{\text{SAME HCI affects all sites/variables}}$$

General modeling framework

Parent distributions of data

$$Y_v(\mathsf{s}, t) \sim \mathcal{D}_v \left(\boldsymbol{\theta}_v(\mathsf{s}, t) \right)$$

Parameters vary in space and time

$$\boldsymbol{\theta}_v(\mathsf{s}, t) = \lambda_{0,v}(\mathsf{s}) + \lambda_{1,v}(\mathsf{s})\tau_1(t) \underbrace{+ \cdots + \lambda_{K,v}(\mathsf{s})\tau_K(t)}_{\text{more HCIs}}$$

General modeling framework

Parent distributions of data

$$Y_v(\mathsf{s}, t) \sim \ \mathcal{D}_v \ \left(\boldsymbol{\theta}_v(\mathsf{s}, t) \right)$$

Parameters vary in space and time

$$\underbrace{g\left(\boldsymbol{\theta}_v(\mathsf{s}, t)\right)}_{\text{link function}} = \lambda_{0,v}(\mathsf{s}) + \lambda_{1,v}(\mathsf{s})\tau_1(t) \ + \cdots + \lambda_{K,v}(\mathsf{s})\tau_K(t)$$

General modeling framework

Parent distributions of data

$$Y_v(\mathsf{s}, t) \sim \mathcal{D}_v \left(\boldsymbol{\theta}_v(\mathsf{s}, t) \right)$$

Parameters vary in space and time

$$g\left(\boldsymbol{\theta}_v(\mathsf{s}, t)\right) = \lambda_{0,v}(\mathsf{s}) + \lambda_{1,v}(\mathsf{s})\tau_1(t) + \cdots + \lambda_{K,v}(\mathsf{s})\tau_K(t)$$

Temporal and spatial Gaussian processes for HCIs and their effects

$$\tau_k(t) \sim \mathcal{G}\left(\boldsymbol{\mu}_\tau, \Sigma_\tau\right);$$

General modeling framework

Parent distributions of data

$$Y_v(\mathsf{s}, t) \sim \mathcal{D}_v \left(\boldsymbol{\theta}_v(\mathsf{s}, t) \right)$$

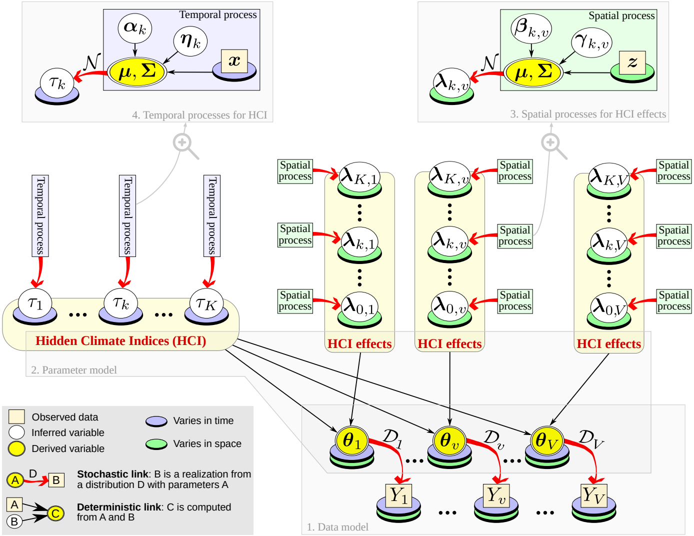
Parameters vary in space and time

$$g\left(\boldsymbol{\theta}_v(\mathsf{s}, t)\right) = \lambda_{0,v}(\mathsf{s}) + \lambda_{1,v}(\mathsf{s})\tau_1(t) + \cdots + \lambda_{K,v}(\mathsf{s})\tau_K(t)$$

Temporal and spatial Gaussian processes for HCIs and their effects

$$\tau_k(t) \sim \mathcal{G}\left(\boldsymbol{\mu}_\tau, \Sigma_\tau\right); \quad \lambda_{k,v}(\mathsf{s}) \sim \mathcal{G}\left(\boldsymbol{\mu}_\lambda, \Sigma_\lambda\right)$$

Schematics of an HCI model



Comments

Possible interpretations

- $g(\theta) = \lambda_0 + \sum \lambda_k \tau_k$ is similar to GLM... but with hidden covariates!

Comments

Possible interpretations

- $g(\theta) = \lambda_0 + \sum \lambda_k \tau_k$ is similar to GLM... but with hidden covariates!
- HCIs and their effects $\approx$ principal components and their loadings
- HCI model $\approx$ non-Gaussian probabilistic PCA
- see *Probabilistic Machine Learning* by Kevin P. Murphy (2022)

Inference

MCMC sampling from the posterior, Renard & Thyer (2019)

Inference

MCMC sampling from the posterior, Renard & Thyer (2019)

Difficulty 1: identifiability constraints

- HCIs should be orthonormal
- Stepwise inference: one component at a time
- Leads to 2 simpler constraints: each HCI has mean 0 and variance 1
- Possible solution: 'Givens Representation' of Pourzanjani et al. (2021)

Inference

MCMC sampling from the posterior, Renard & Thyer (2019)

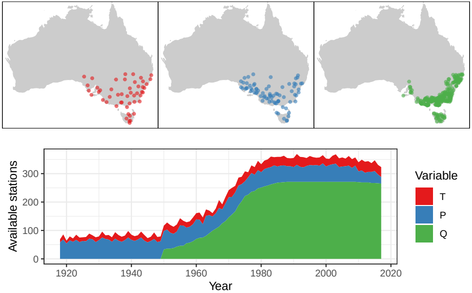
Difficulty 1: identifiability constraints

- HCIs should be orthonormal
- Stepwise inference: one component at a time
- Leads to 2 simpler constraints: each HCI has mean 0 and variance 1
- Possible solution: 'Givens Representation' of Pourzanjani et al. (2021)

DIfficulty 2: dimensionality

- Dimension of the posterior grows with both #sites and #time steps.
- No big deal for the likelihood (thank you conditional independence!)
- Bottleneck = covariance matrices in Gaussian hyperdistributions
- One solution: nearest-neighbor Gaussian process of Datta et al. (2016)

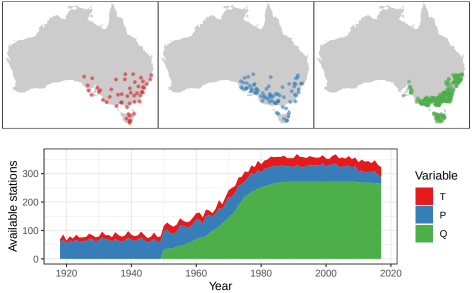
Data & Models



Number of heatwaves

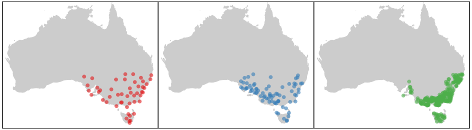
$$\begin{cases} Tn(\mathsf{s},t) \sim \mathcal{P}\left(\mu(\mathsf{s},t)\right) \\ log\left(\mu(\mathsf{s},t)\right) = \lambda_{Tn,0}(\mathsf{s}) + \sum_{k=1}^{3} \lambda_{Tn,k}(\mathsf{s})\tau_k(t) \end{cases}$$

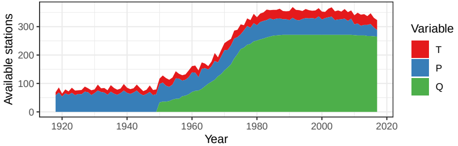
Data & Models



Heatwave intensities

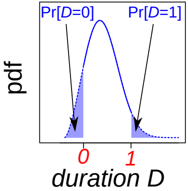
$$\begin{cases} Tx(\mathrm{s}, t) \sim \mathcal{GPD}\left(0, \sigma(\mathrm{s}, t), \xi(\mathrm{s})\right) \\ log\left(\sigma(\mathrm{s}, t)\right) = \lambda_{Tx,0}(\mathrm{s}) + \sum_{k=1}^{3} \lambda_{Tx,k}(\mathrm{s})\tau_k(t) \end{cases}$$

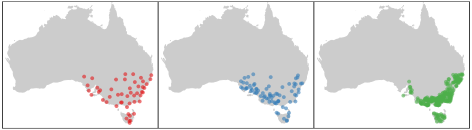


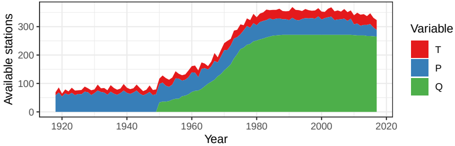


Drought duration

$$\begin{cases} Qd(\mathrm{s},t) \sim \mathcal{N}\left(\mu(\mathrm{s},t), \sigma(\mathrm{s})\right) \\ \mu(\mathrm{s},t) = \lambda_{Qd,0}(\mathrm{s}) + \sum_{k=1}^{3} \lambda_{Qd,k}(\mathrm{s})\tau_k(t) \end{cases}$$

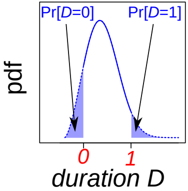




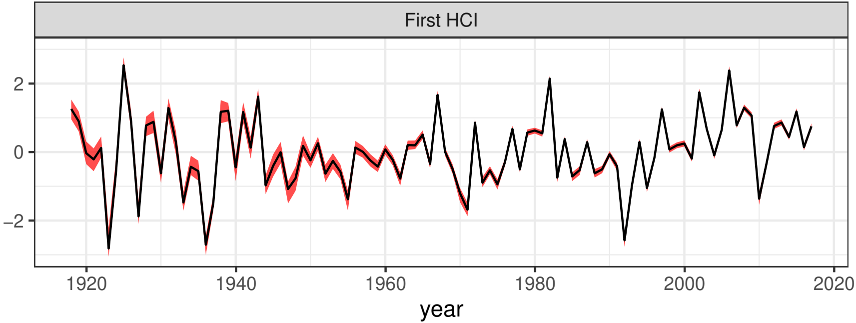


Dry-day duration

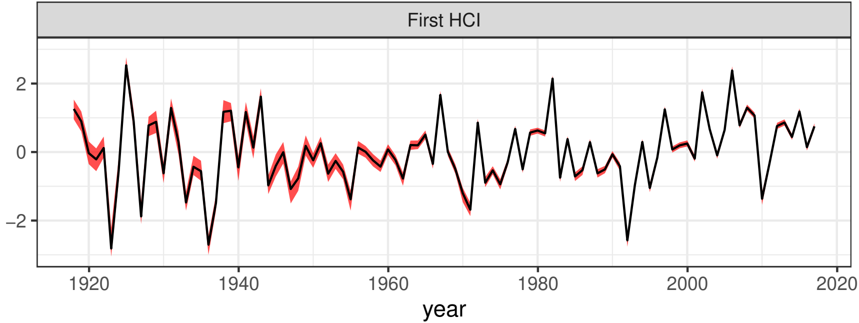
$$\begin{cases} Pd(s,t) \sim \mathcal{N}\left(\mu(s,t), \sigma(s)\right) \\ \mu(s,t) = \lambda_{Pd,0}(s) + \sum_{k=1}^{3} \lambda_{Pd,k}(s)\tau_k(t) \end{cases}$$



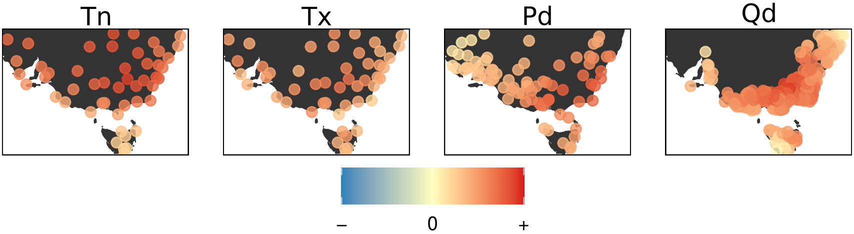
First HCI and its effects



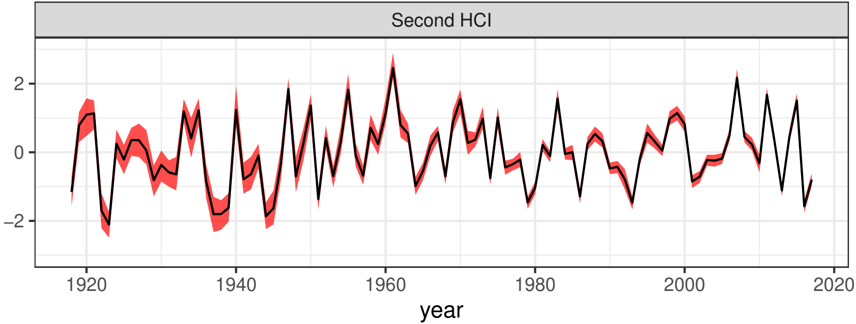
First HCI and its effects



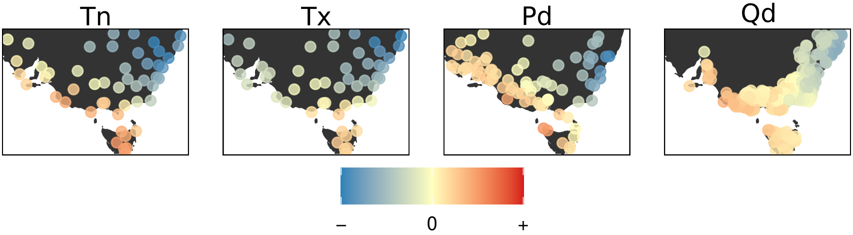
Effect on...



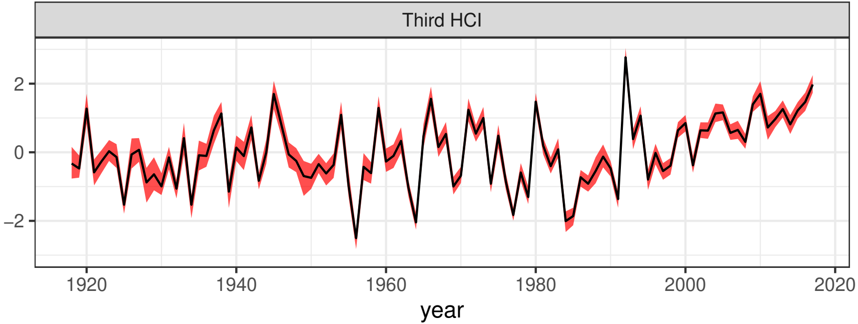
Second HCI and its effects



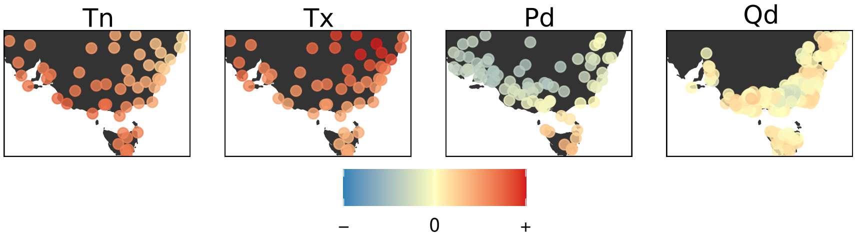
Effect on...



Third HCI and its effects

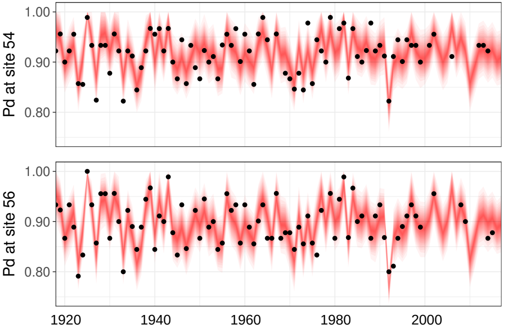


Effect on...



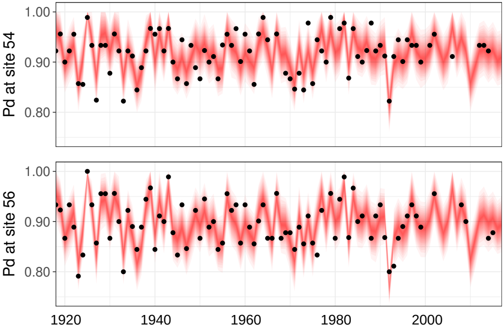
Probabilistic predictions

Time-varying distributions

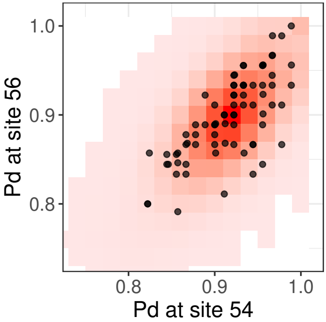


Probabilistic predictions

Time-varying distributions

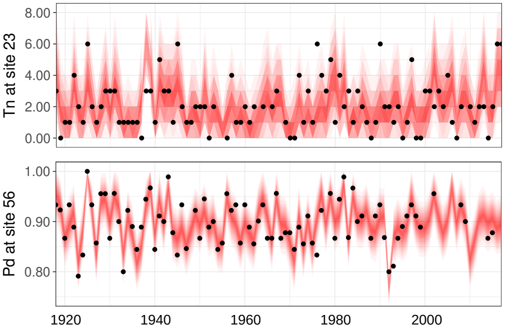


Joint bivariate distribution

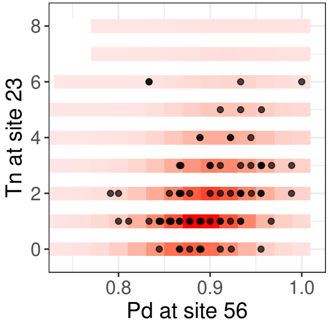


Probabilistic predictions

Time-varying distributions

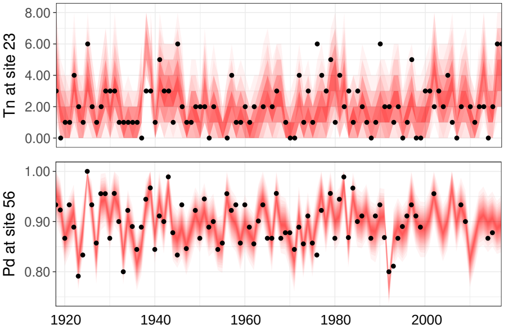


Joint bivariate distribution

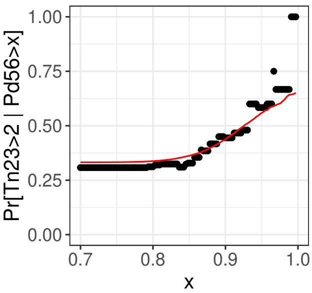


Probabilistic predictions

Time-varying distributions

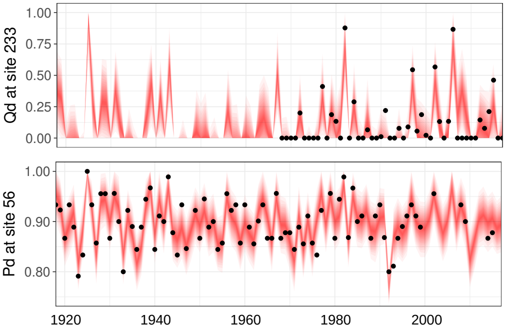


Conditional probabilities



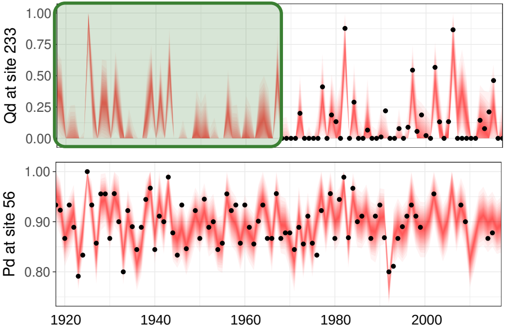
Probabilistic predictions

Time-varying distributions



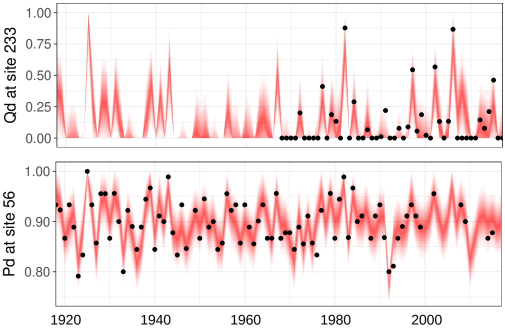
Probabilistic predictions

Time-varying distributions

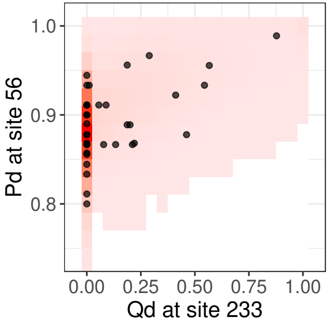


Probabilistic predictions

Time-varying distributions

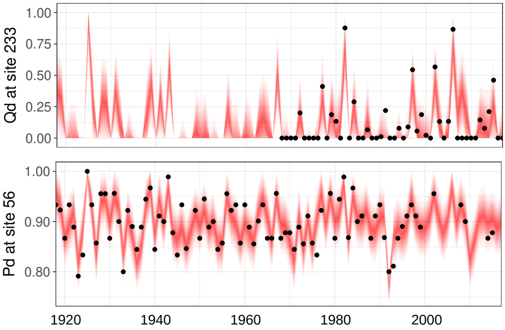


Joint bivariate distribution

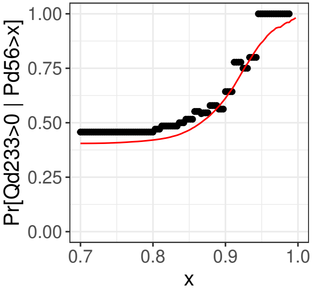


Probabilistic predictions

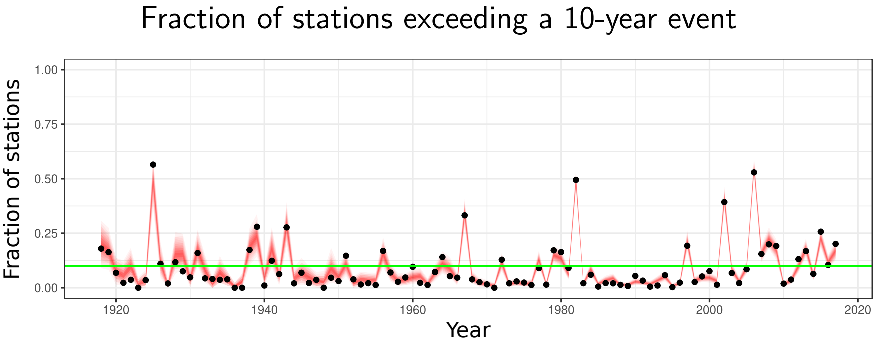
Time-varying distributions



Conditional probabilities



Probabilistic predictions



$\Longrightarrow$ Consequences in terms of risk management

Conclusion

Summary: the HCI modeling framework

- A general probabilistic model for multi-variable space-time data
- Based on hidden climate indices extracted from the target data
- Flexible: can handle a wide range of hydro-meteorological datasets

Conclusion

Summary: the HCI modeling framework

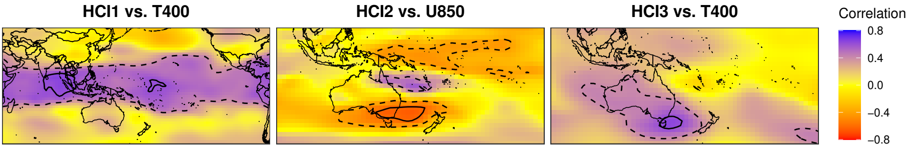
- A general probabilistic model for multi-variable space-time data
- Based on hidden climate indices extracted from the target data
- Flexible: can handle a wide range of hydro-meteorological datasets

Other noticeable results [not shown here]

- Probabilistic predictions are reliable, including in cross-validation
- HCIs $\neq$ standard indices such as NINO, SAM, IOD, etc.
- Replacing HCIs with standard indices underestimates dependence

Work in progress

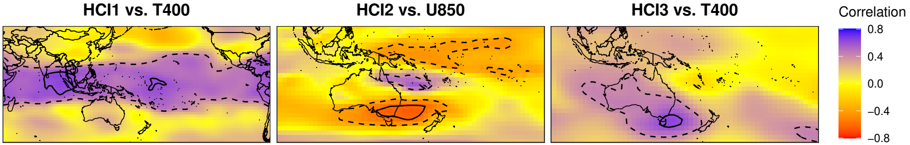
Predicting HCIs from large-scale climate variables?



$\Longrightarrow$ Downscaling device for past reconstructions, seasonal forecasting or future projections

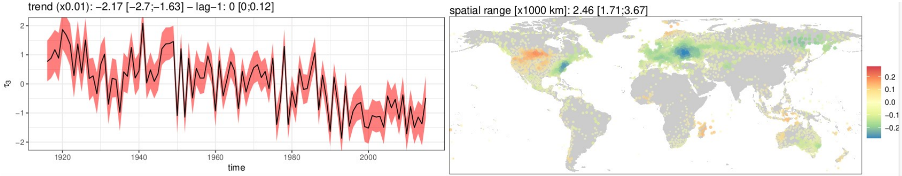
Work in progress

Predicting HCIs from large-scale climate variables?



$\Longrightarrow$ Downscaling device for past reconstructions, seasonal forecasting or future projections

Application to global floods and extreme precipitation



Thank you!

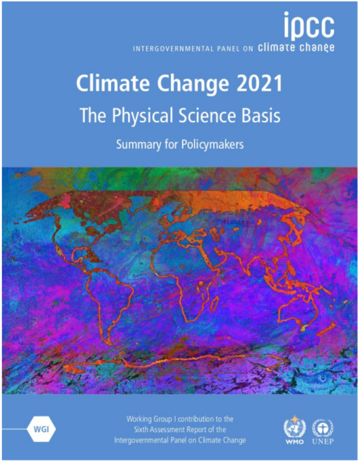
Renard & Thyer (**2019**). Revealing Hidden Climate Indices from the Occurrence of Hydrologic Extremes. *Water Resources Research*.

Renard et al. (**2021?**). A Hidden Climate Indices Modeling Framework for Multi-Variable Space-Time Data. *Submitted to Water Resources Research*.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 835496

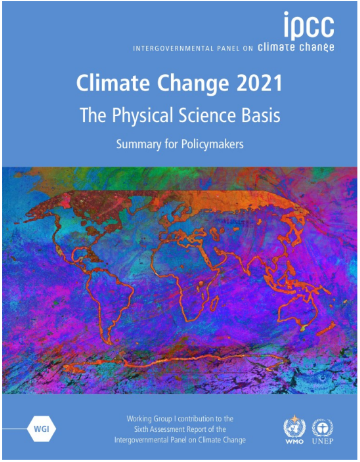
https://globxblog.inrae.fr/

https://github.com/STooDs-tools



" The frequency and intensity of heavy precipitation events have increased since the 1950s over most land area for which observational data are sufficient for trend analysis (high confidence), and human-induced climate change is likely the main driver"

Application 2: context

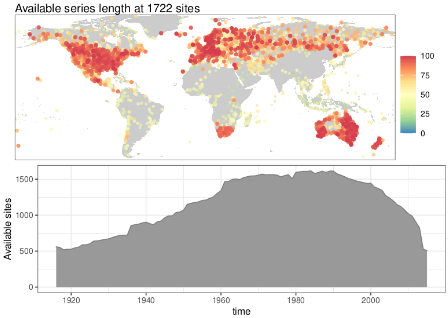


"*The frequency and intensity of heavy precipitation events have increased since the 1950s over most land area for which observational data are sufficient for trend analysis (high confidence), and human-induced climate change is likely the main driver*"

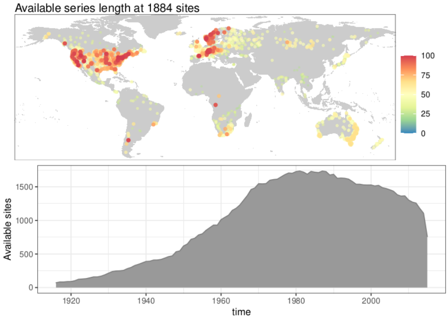
"*Confidence about peak flow trends over past decades on the global scale is low, but there are regions experiencing increases [. . .] and regions experiencing decreases [. . .]*"

Global datasets for hydrologic extremes

Precipitation: Hadex 2+3



Streamflow: GSIM



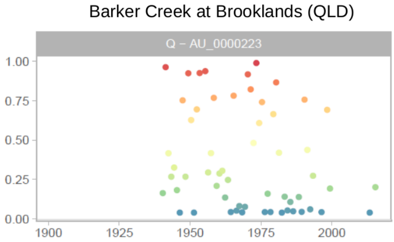
Objectives

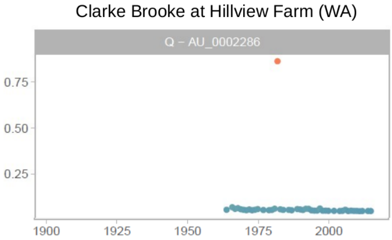
- Look for trends, low-frequency variability and teleconnections for both P and Q extremes
- Analyze long 100-year period (vs. a typical 50-year)
- Attempt at predicting extreme P/Q from large-scale climate

Analyzed variables

Non-exceedance probability ($\Leftrightarrow$ return period) of the largest event of the season

Example: Maximum streamflow in December-January-February for 2 Australian stations

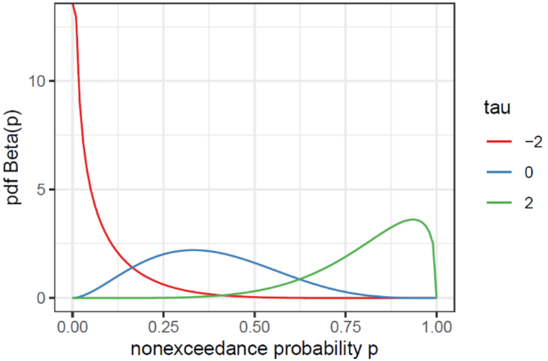




Model

Beta distribution reparameterized in terms of mean μ and precision γ

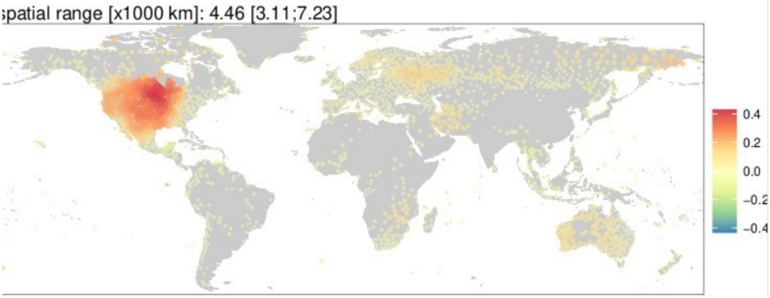
$$\begin{cases} Y_v(s,t) \sim Beta_v\left(\mu_v(s,t), \gamma_v(s)\right) \\ logit\left(\mu_v(s,t)\right) = \lambda_{v,0}(s) + \lambda_{v,1}(s)\tau_1(t) + \ldots + \lambda_{v,K}(s)\tau_K(t) \end{cases}$$



Model

Beta distribution reparameterized in terms of mean μ and precision γ

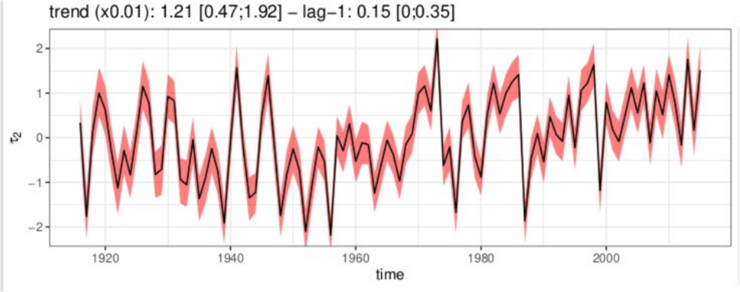
$$\begin{cases} Y_v(\mathrm{s}, t) \sim Beta_v\left(\mu_v(\mathrm{s}, t), \gamma_v(\mathrm{s})\right) \\ logit\left(\mu_v(\mathrm{s}, t)\right) = \lambda_{v,0}(\mathrm{s}) + \lambda_{v,1}(\mathrm{s})\tau_1(t) + \ldots + \lambda_{v,K}(\mathrm{s})\tau_K(t) \\ \boldsymbol{\lambda}_k \sim NNGP\left(\boldsymbol{\beta}, \mathrm{V}\right); V_{i,j} = \nu_0^2 exp(-d_{i,j}/\nu_1) \end{cases}$$



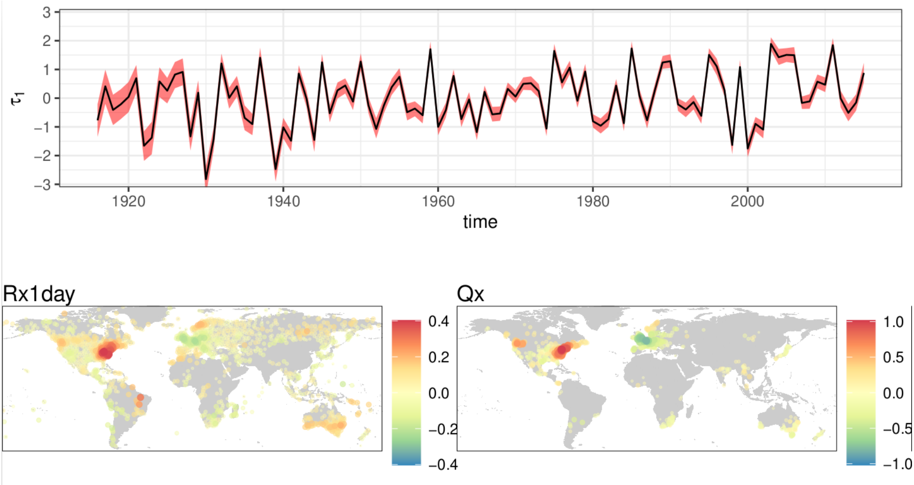
Model

Beta distribution reparameterized in terms of mean μ and precision γ

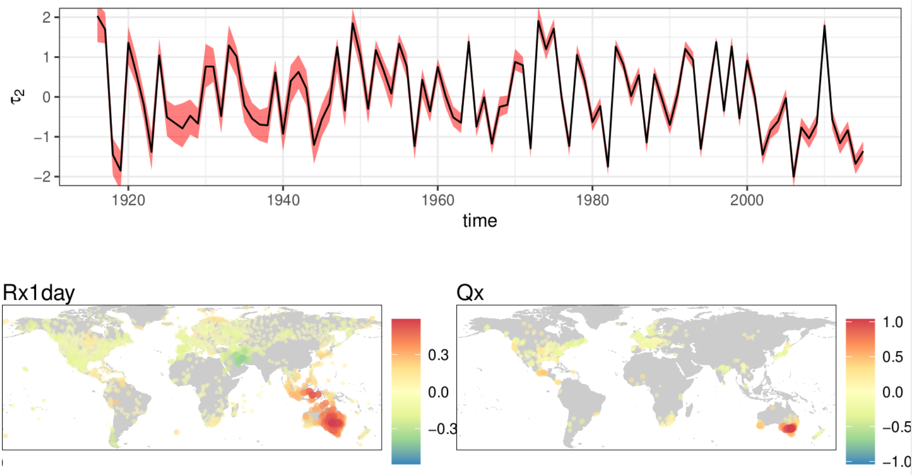
$$\begin{cases} Y_v(\mathrm{s},t) \sim Beta_v\left(\mu_v(\mathrm{s},t), \gamma_v(\mathrm{s})\right) \\ logit\left(\mu_v(\mathrm{s},t)\right) = \lambda_{v,0}(\mathrm{s}) + \lambda_{v,1}(\mathrm{s})\tau_1(t) + \ldots + \lambda_{v,K}(\mathrm{s})\tau_K(t) \\ \boldsymbol{\lambda}_k \sim NNGP\left(\boldsymbol{\beta}, \mathrm{V}\right); V_{i,j} = \nu_0^2 exp(-d_{i,j}/\nu_1) \\ \boldsymbol{\tau}_k \sim NNGP\left(\mathrm{m}, \mathrm{W}\right); W_{i,j} = u_0^2 exp(-d_{i,j}/u_1); m_t = \eta_0 + \eta_1 t \end{cases}$$



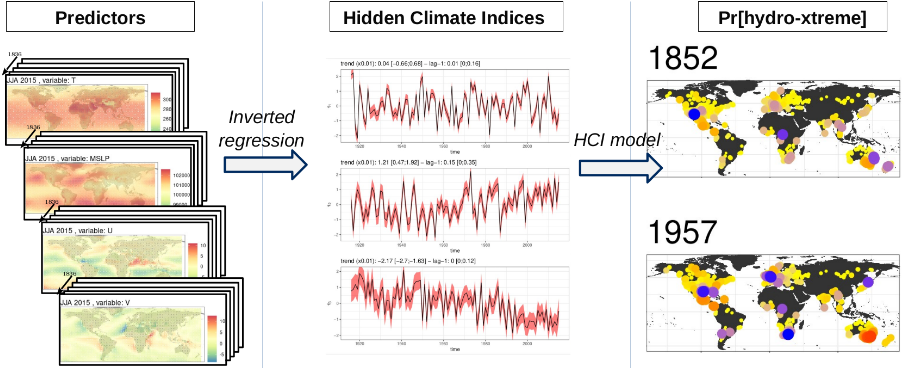
Step 1: identify components



Step 1: identify components



Step 2: past reconstructions from 1836



Thank you!

Renard & Thyer (**2019**). Revealing Hidden Climate Indices from the Occurrence of Hydrologic Extremes. *Water Resources Research*.

Renard et al. (**2021?**). A Hidden Climate Indices Modeling Framework for Multi-Variable Space-Time Data. *Submitted to Water Resources Research*.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 835496

https://globxblog.inrae.fr/

https://github.com/STooDs-tools

Downscaling approach

Method: inverted regression

Step 1: $w(s,t)$: climate field at time t and location s

$\widehat{\tau}_k(t)$: estimated HCI's (from previous analysis)

Goal: estimate $\psi_k(s)$'s in:

$$w(s,t) = \psi_0(s) + \psi_1(s)\widehat{\tau}_1(t) + \ldots + \psi_K(s)\widehat{\tau}_K(t) + \varepsilon(s,t)$$

Step 2: $w(s,t^*)$: climate field at time t^* and location s

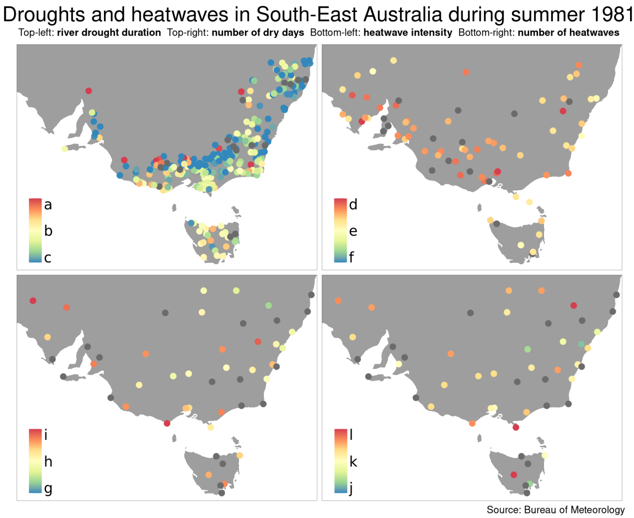
$\widehat{\psi}_k(s)$: estimated from previous step

Goal: estimate $\tau_k(t^*)$'s in:

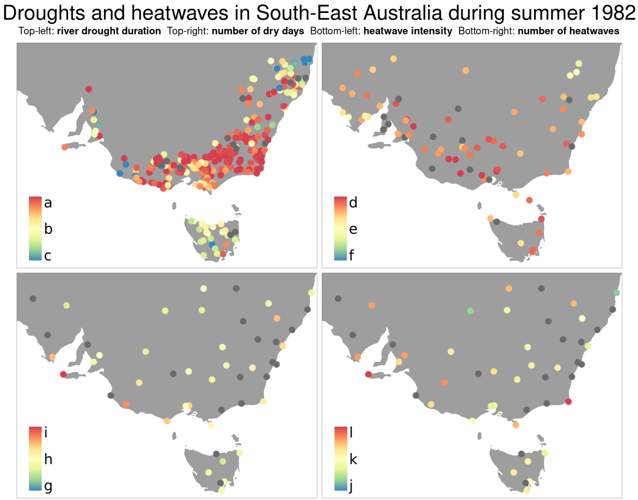
$$w(s,t^*) = \psi_0(s) + \widehat{\psi}_1(s)\tau_1(t^*) + \ldots + \widehat{\psi}_K(s)\tau_K(t^*) + \varepsilon(s,t^*)$$

Alternatives: LASSO, RIDGE and other form of penalised regression, but first attempts inconclusive

Motivating dataset: a few years



Motivating dataset: a few years



Source: Bureau of Meteorology

Motivating dataset: a few years