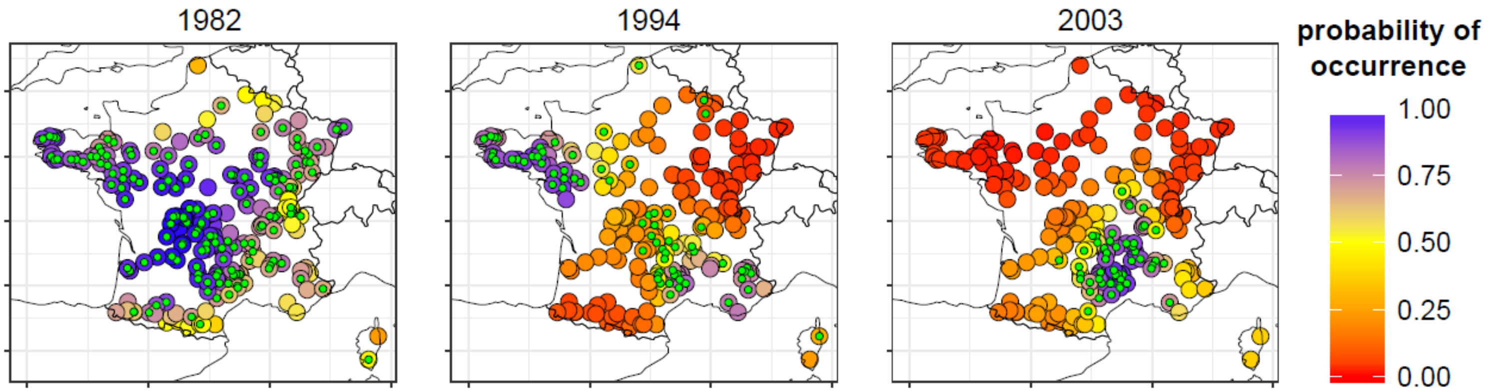
Estimation of hidden climate indices controlling flood occurrence

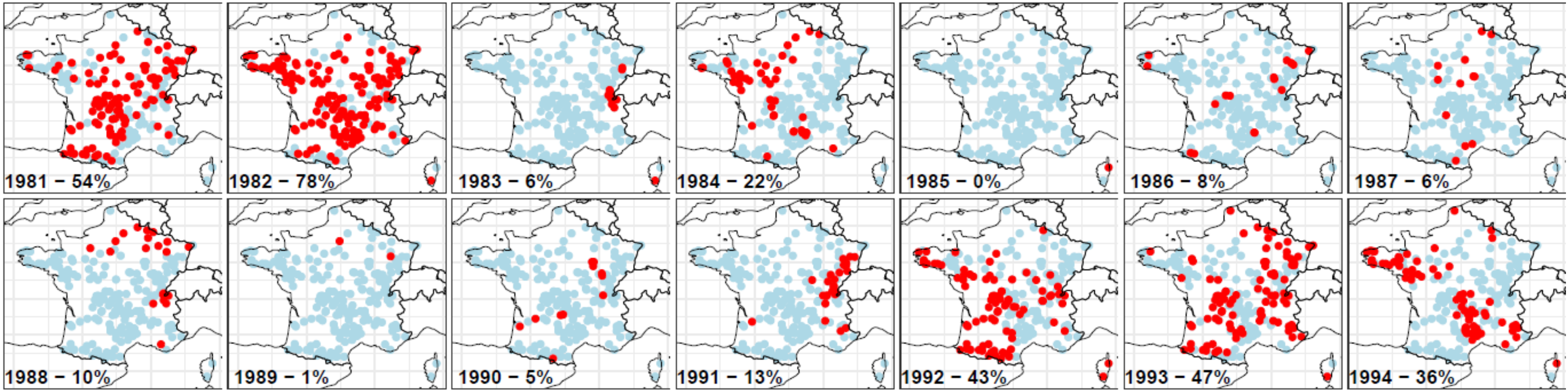
Benjamin Renard
Irstea Lyon, France



Motivation

Autumn flood occurrences in France

207 stations, 1969-2016
Floods exceeding a threshold such as Pr(occurrence) ≈ 0.2 at all sites



Clear clustering in space, maybe some clustering in time?
⇒ on any given year, %stations with occurrence tends to be either very low or very high, but is rarely close to its interannual average of 20%.

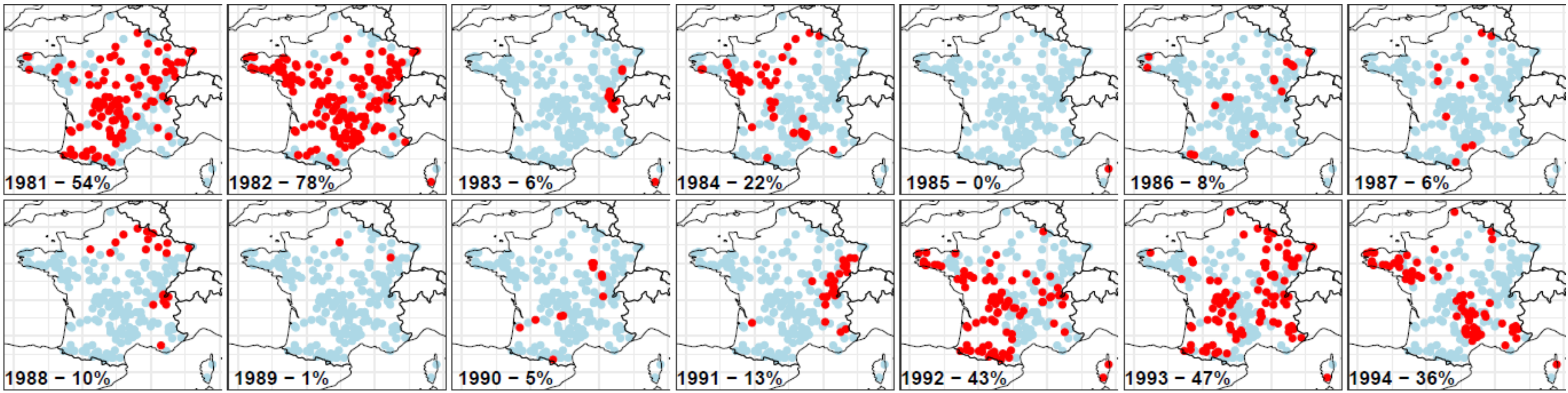
Consequences for risk management at this national scale: insurance companies, disaster response, etc.

Motivation

Autumn flood occurrences in France

207 stations, 1969-2016

Floods exceeding a threshold such as Pr(occurrence) ≈ 0.2 at all sites



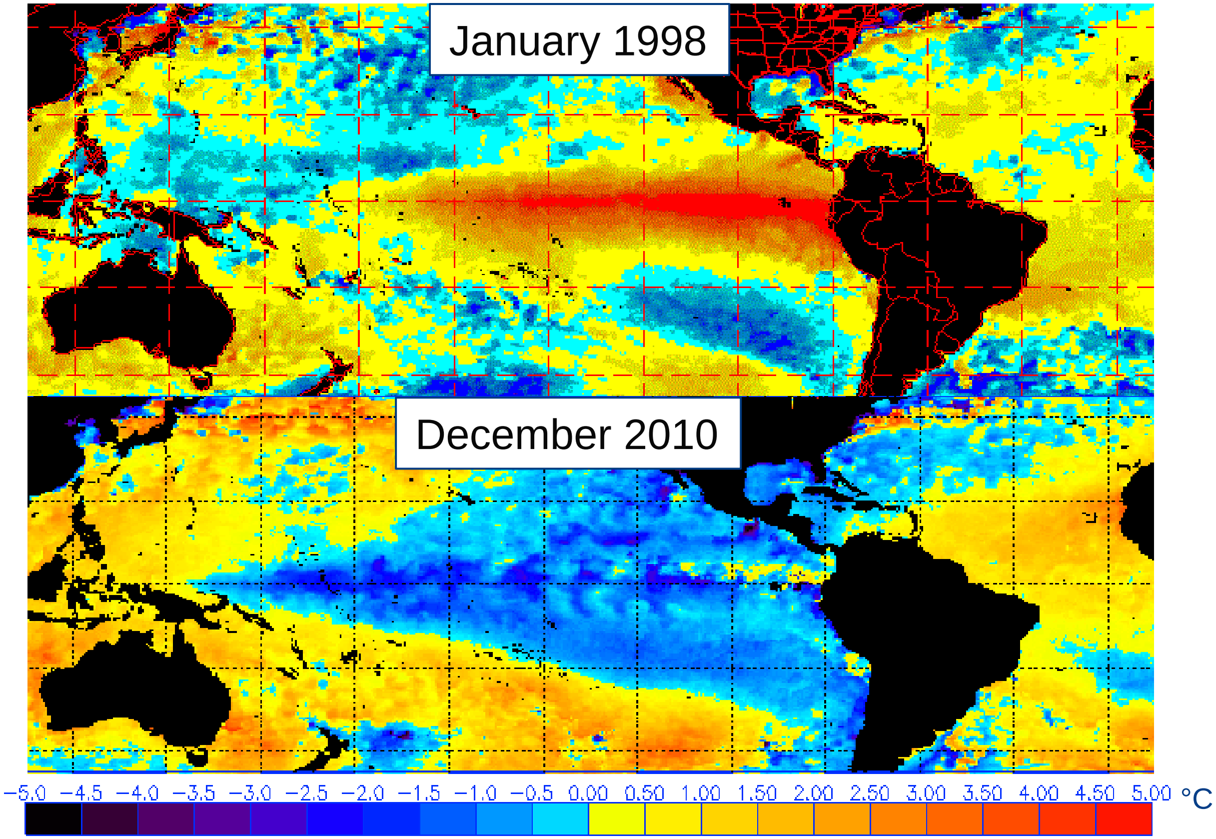
As hydrologists, we are interested in:

- Describing this space-time variability
- Understanding its origin (atmospheric / oceanic circulation)
- Making predictions such as Pr(occurrence | atmosphere-ocean)

  → *Seasonal forecasting, past reconstructions, future projections, etc.*

Introducing climate indices

Ex.: El Niño Southern Oscillation (ENSO)



Sea SurfaceTemperature (SST) anomalies

✓ **El Niño episode**

- Warming of the central and eastern tropical Pacific Ocean

- Worldwide impacts on oceanic and atmospheric circulation, and hence on hydrology

- e.g. in Southern US: increased precipitations

✓ **La Niña episode**

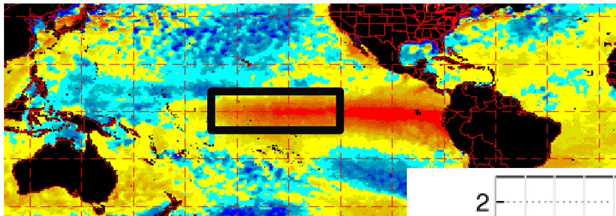
- Cooling of the central and eastern tropical Pacific Ocean

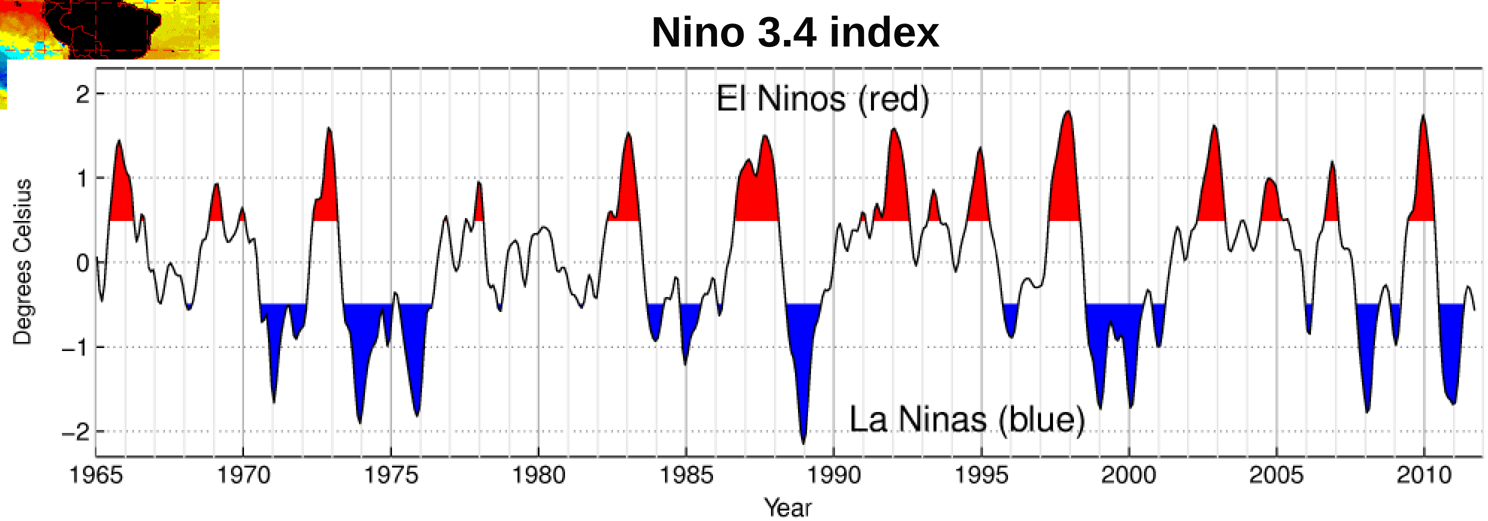
- e.g. in eastern Australia: increased precipitations

Some degree of seasonal predictability (a few months ahead)

Introducing climate indices

Ex.: El Niño Southern Oscillation (ENSO)





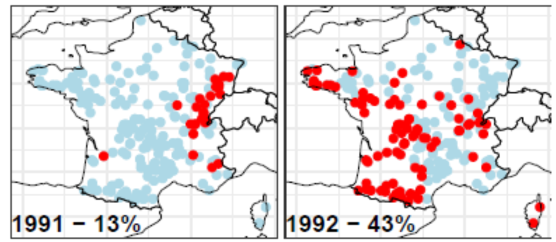
A climate index aims at summarizing a certain climate pattern using a single number.

Here, simple spatial average. Often PCA-like methods are used.

Using climate indices

Used as covariates in a regression model

Example for occurrence data



Occurrence 0/1 variable

Bernoulli distribution

$$\boxed{Y}(\boxed{t},\boxed{x}) \sim \boxed{B}(\boxed{\theta(t,x)})$$

Probability of occurrence, varying in time and space

time

space

Effect of climate index at site *x*

$$Logit\left(\theta(t,x)\right) = \boxed{\lambda_0(x)} + \boxed{\lambda_1(x)} * \boxed{I(t)}$$

intercept

climate index

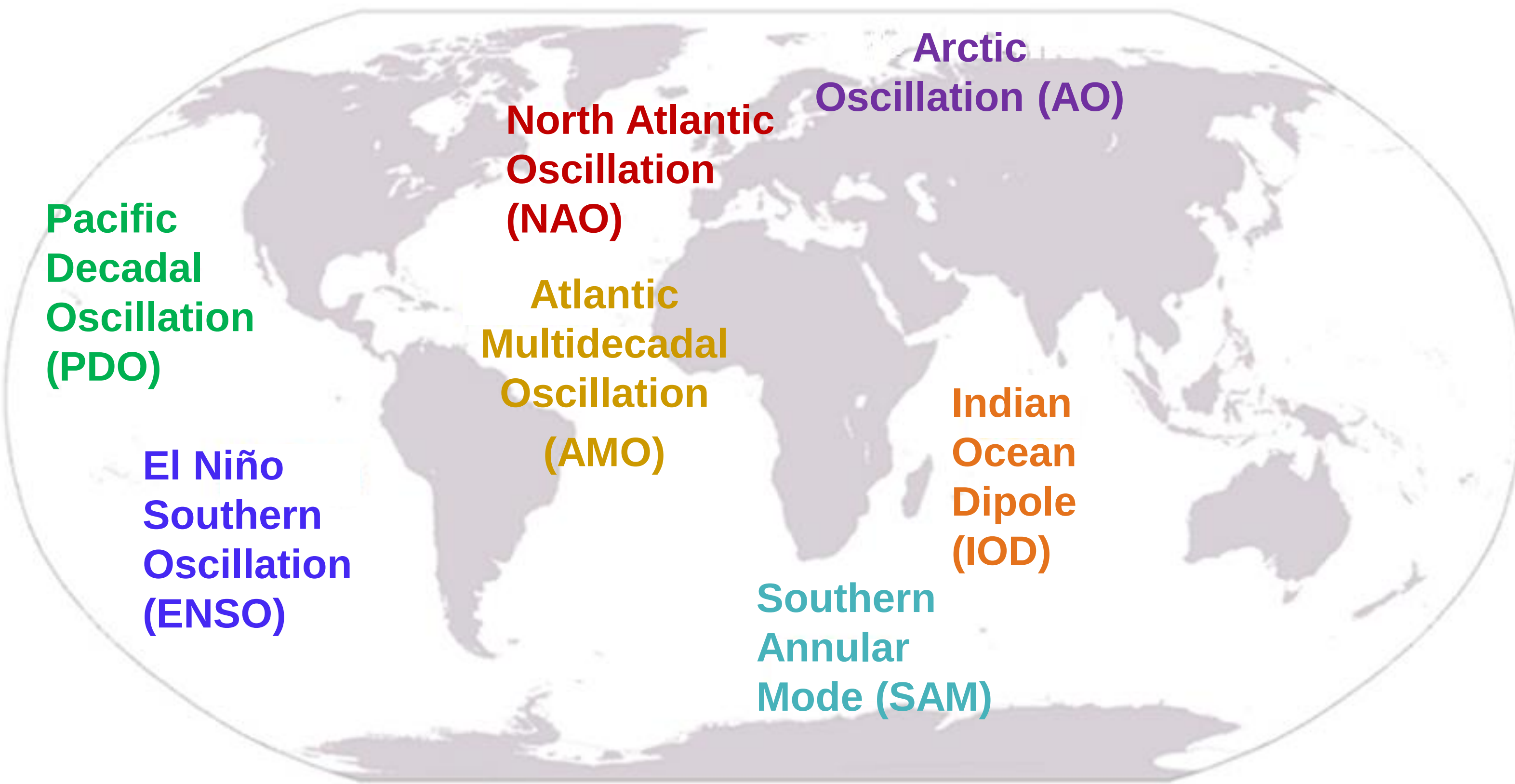
At one given site x: GLM

Other variable/distribution (e.g. annual maxima / GEV distribution): GAMLSS.

Multi-site: interest of hierarchical modeling for constraining spatial term $\lambda_1(x)$

Issues with climate indices

Many, many climate indices…



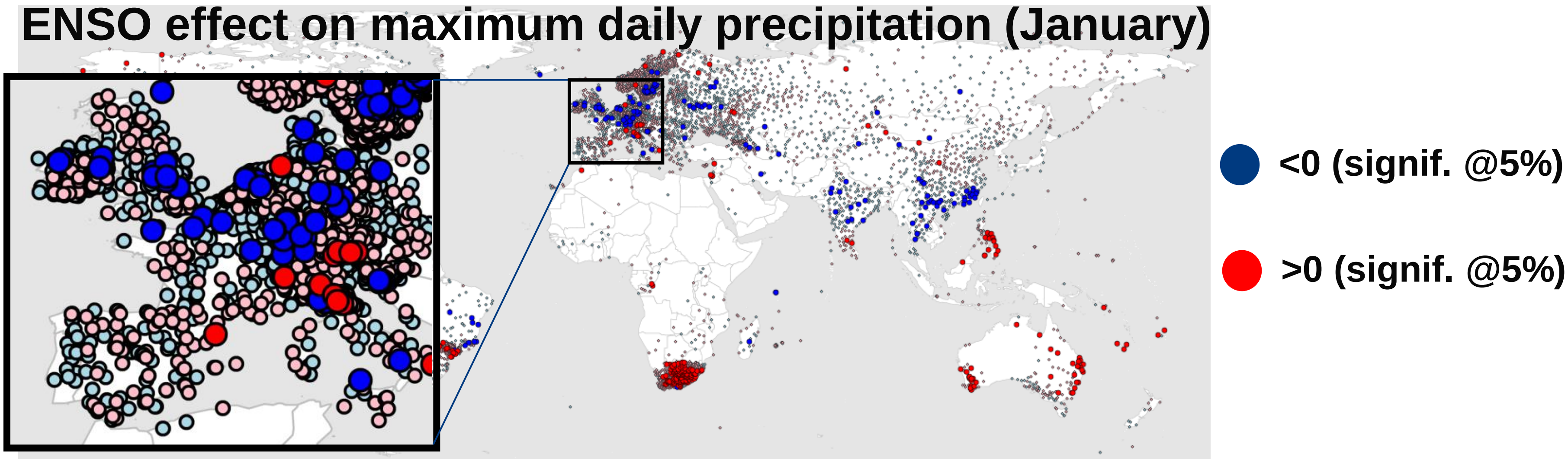
Nino 3.4
Nino 1.2
Nino 4
SOI
NAO.v1
NAO.v2
PDO
AO
PNA
AMO
IPO
MJO
MOI
WeMO
CAR
NTA
NAM
NP
AAO
SAM

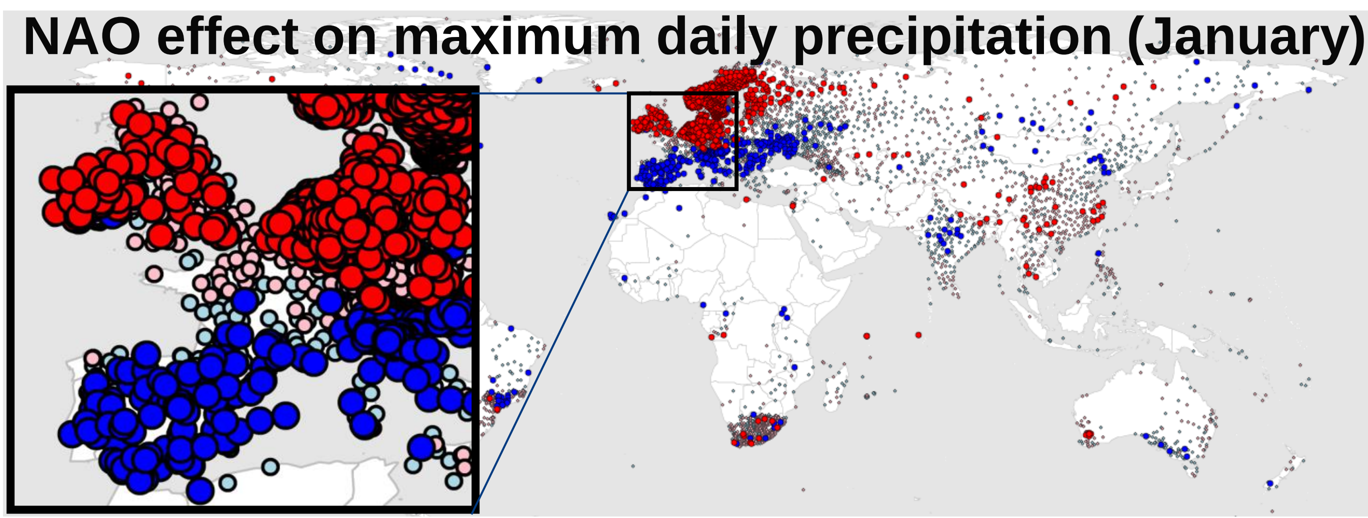
Taken from https://www2.ucar.edu/news/backgrounders/weather-maker-patterns-interactive-map

Issues with climate indices

They are sometimes poor predictors!

The (frustrating) case of France:

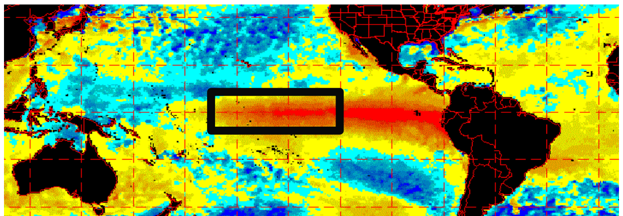




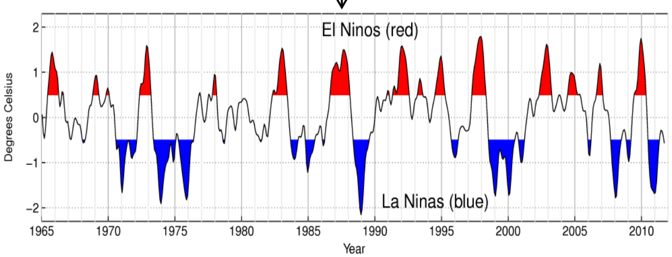
Hidden climate indices model

Look at the problem from the other way around

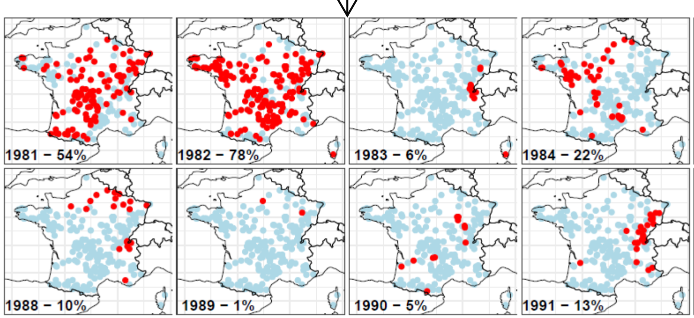
Instead of:



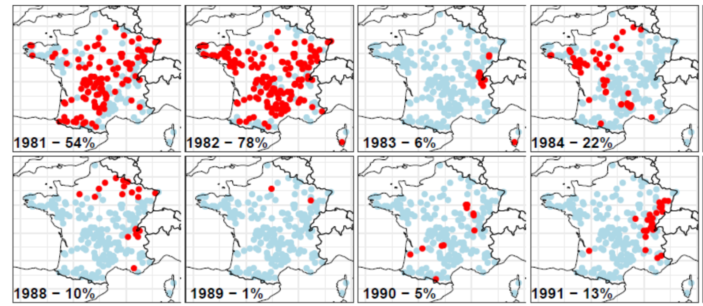
Define climate index



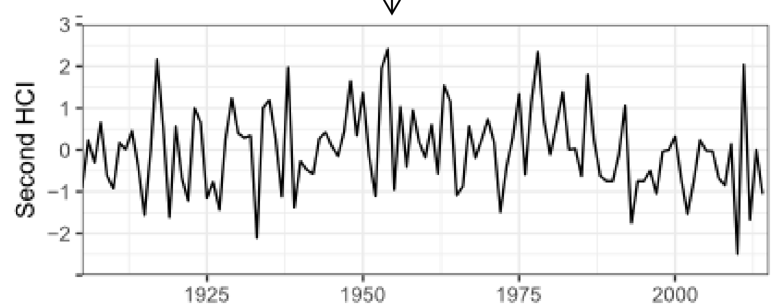
Covariate modeling



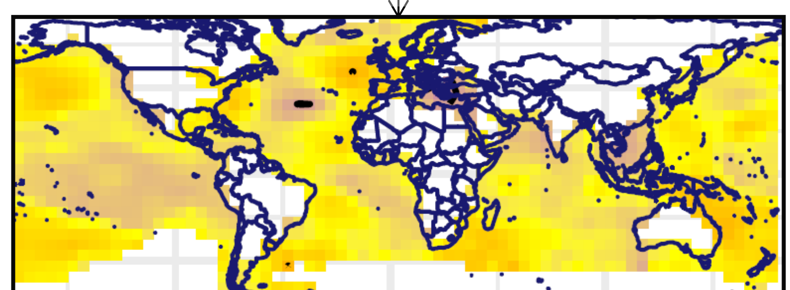
Let's try:



Look for "hidden" climate index



Look for associated climate pattern



Hidden climate indices model

Formulation

$$Y(t,x) \sim B(\theta(t,x))$$

Impact of 1st hidden climate index on each site

Kst component

$$Logit\left(\theta(t,x)\right) = \boxed{\boxed{\lambda_0(x)}} + \boxed{\boxed{\lambda_1(x)}} * \boxed{\tau_1(t)} + \ldots + \boxed{\boxed{\lambda_K(x)} * \tau_K(t)}$$

intercept

1st hidden climate index

$$\forall k = 0:K,\ \left(\lambda_k(x_1),\ldots,\lambda_k(x_n)\right) \sim MG\left(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)$$

$$\boldsymbol{\mu}_k = \boxed{\boldsymbol{Z}_k \boldsymbol{\beta}_k}$$

Spatial (e.g. elevation) effects

$$\left(\boldsymbol{\Sigma}_k\right)_{i,j} = \boxed{\gamma_1^2 \exp\left(-\left\|x_i - x_j\right\| / \gamma_2\right)}$$

Spatial dependence (“variogram”)

(conditional) independence of $\boldsymbol{Y} \mid \boldsymbol{\theta}$ (in both space and time)

Identifiability constraints:

$\forall k = 1:K,\ \left(\tau_k(t_1),\ldots,\tau_k(t_n)\right)$ has mean 0 and st. dev. 1

Hidden climate indices model

Estimation: stepwise strategy

$$Y(t,x) \sim B(\theta(t,x))$$

First estimate: $Logit\left(\theta(t,x)\right) = \lambda_0(x)$

Then: $Logit\left(\theta(t,x)\right) = \hat{\lambda}_0(x) + \lambda_1(x) * \tau_1(t)$

Then: $Logit\left(\theta(t,x)\right) = \hat{\lambda}_0(x) + \hat{\lambda}_1(x) * \hat{\tau}_1(t) + \lambda_2(x) * \tau_2(t)$

Etc.

Why a stepwise strategy?

Full model still not identifiable
Need additional constraint: uncorrelated $\left(\boldsymbol{\tau}_i, \boldsymbol{\tau}_j\right)$
Creates other problems…

Hidden climate indices model

Estimation: Bayesian / MCMC setup

Posterior distribution

$$p\left(\boldsymbol{\tau}_k, \boldsymbol{\lambda}_k, \boldsymbol{\beta}_k, \boldsymbol{\gamma}_k \mid \hat{\boldsymbol{\lambda}}_0, \hat{\boldsymbol{\lambda}}_{1:k-1}, \hat{\boldsymbol{\tau}}_{1:k-1}, \boldsymbol{y}, \boldsymbol{z}\right)$$

$$\propto \underbrace{p\left(\boldsymbol{y} \mid \boldsymbol{\tau}_k, \boldsymbol{\lambda}_k, \hat{\boldsymbol{\lambda}}_0, \hat{\boldsymbol{\lambda}}_{1:k-1}, \hat{\boldsymbol{\tau}}_{1:k-1}\right)}_{\text{likelihood}} \underbrace{p\left(\boldsymbol{\lambda}_k \mid \boldsymbol{\beta}_k, \boldsymbol{\gamma}_k, \boldsymbol{z}\right)}_{\text{hierarchical term}} \underbrace{p\left(\boldsymbol{\beta}_k, \boldsymbol{\gamma}_k, \boldsymbol{\tau}_k\right)}_{\text{prior}}$$

MCMC sampling

Customized block Metropolis sampler:

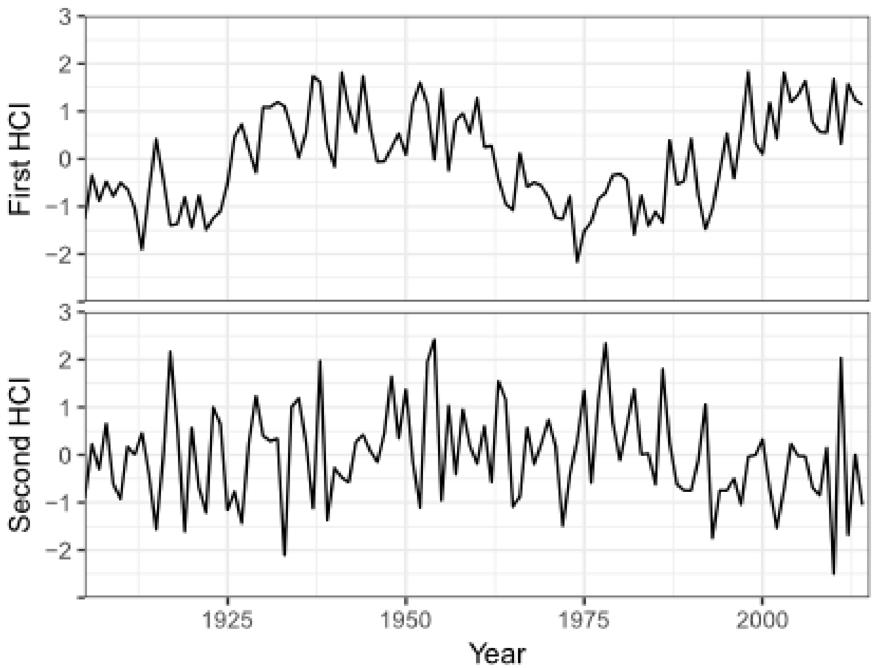
- Update parameter vector one-component-at-a-time
- Many simplifications in Metropolis ratio => quite fast!
- Adaptive: jump size adapted in order to keep acceptance rates within reasonable bounds
- Burn, thin, convergence check

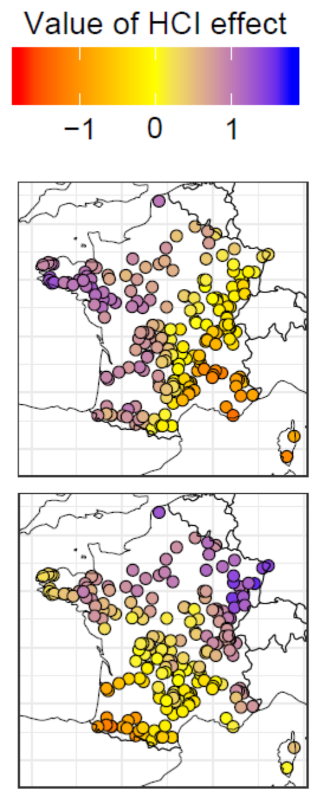
Synthetic case study

Data generation

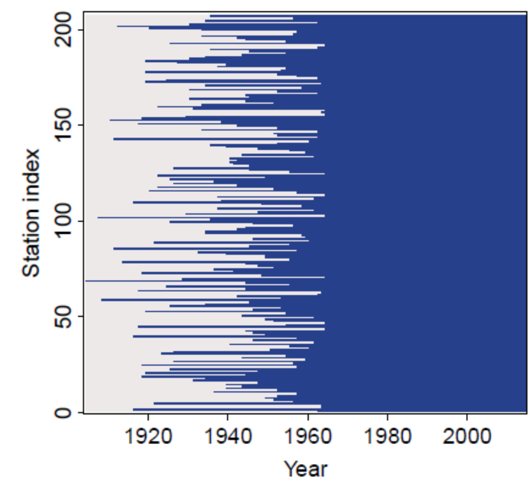
207 sites, 1905-2014

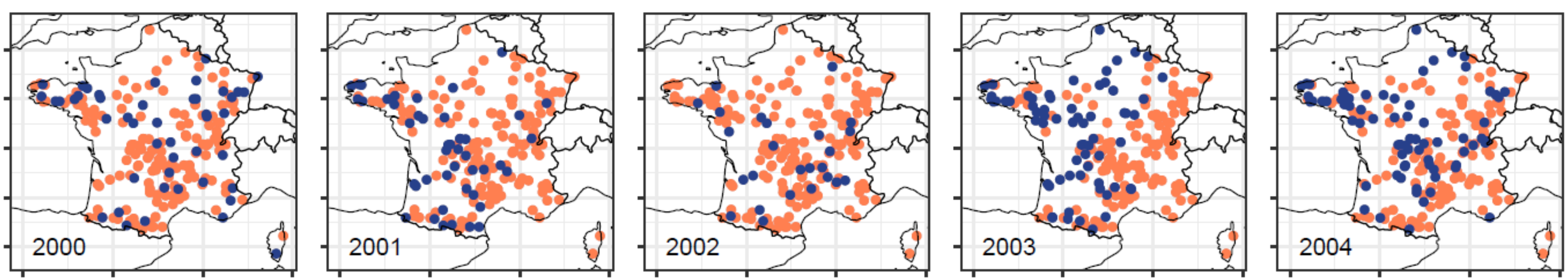
$$\lambda_0 \equiv Logit(0.2)$$





Missing data mask:





Synthetic case study

Estimated model

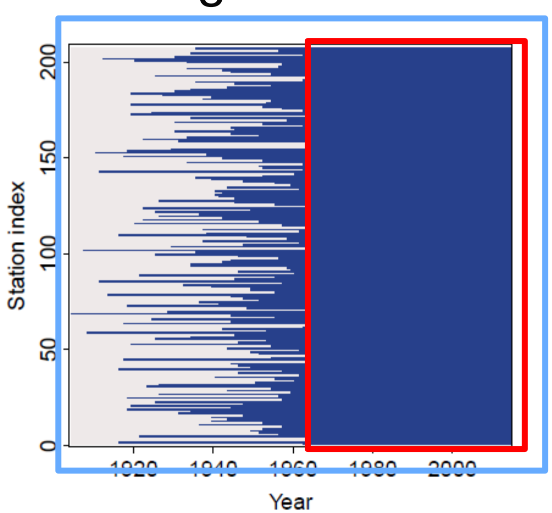
3-component model

$$Logit\big(\theta(t,x)\big) = \lambda_0(x) + \lambda_1(x) * \tau_1(t) + \lambda_2(x) * \tau_2(t) + \lambda_3(x) * \tau_3(t)$$

$$\big(\lambda_0(x_1),\ldots,\lambda_0(x_n)\big) \sim MG\big(\mu_0 \boldsymbol{1}_n, \sigma_0^2 \boldsymbol{I}_n\big)$$

$$\big(\lambda_k(x_1),\ldots,\lambda_k(x_n)\big) \sim MG\big(\mu_k \boldsymbol{1}_n, \boldsymbol{\Sigma}_k\big)$$

Missing data mask:

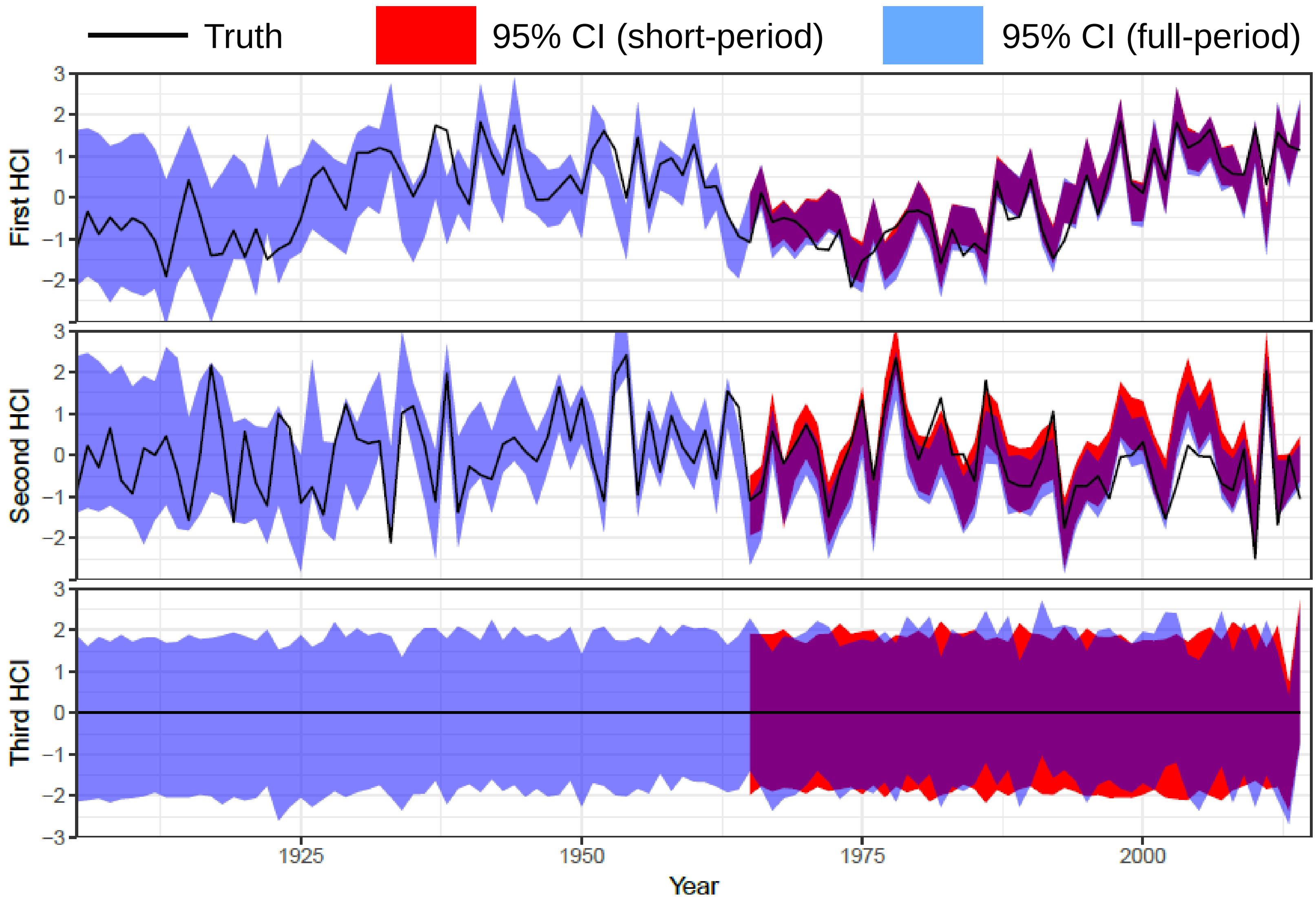


Short-period estimation

Full-period estimation

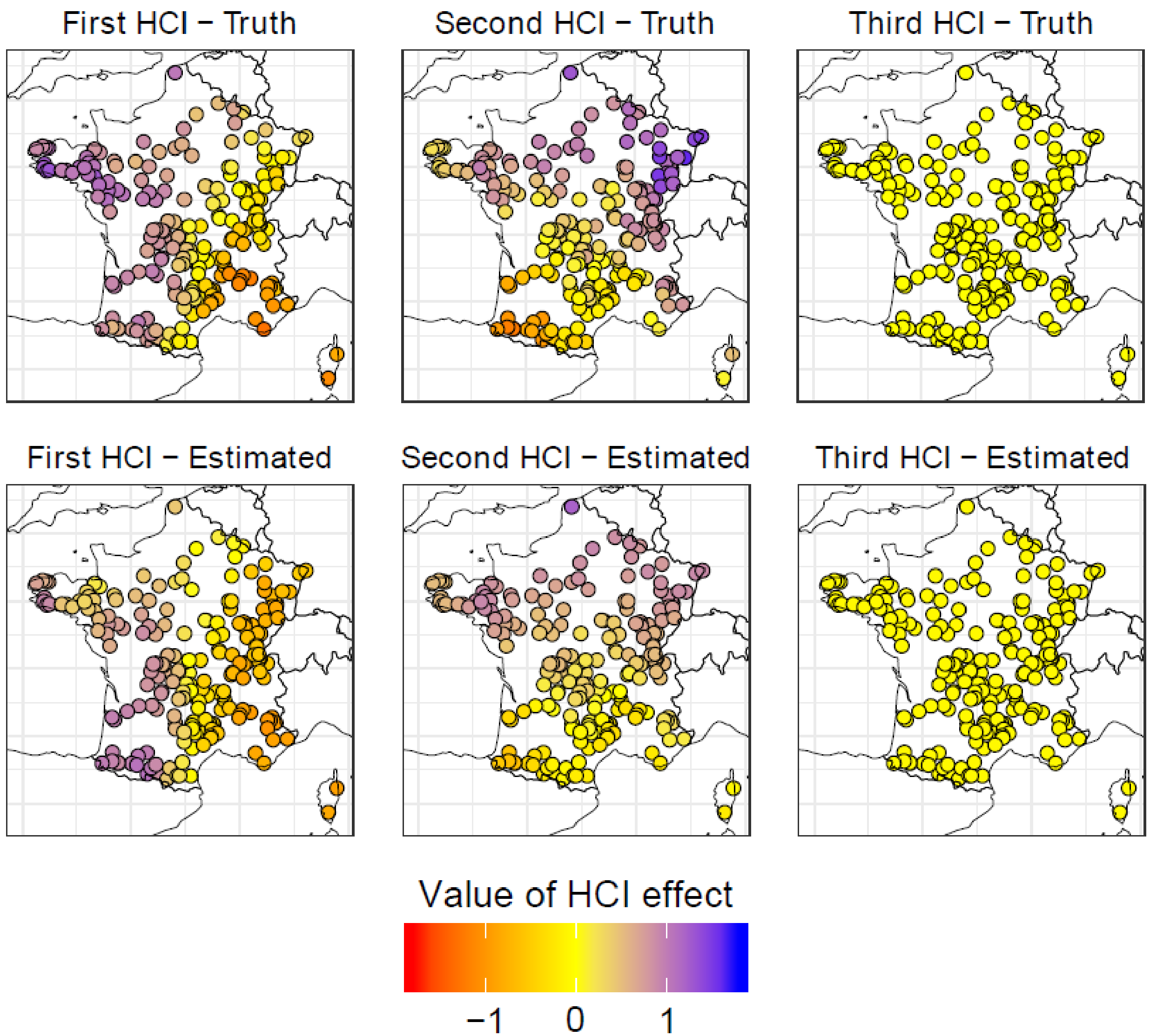
Synthetic case study

Ability to recover hidden climate indices



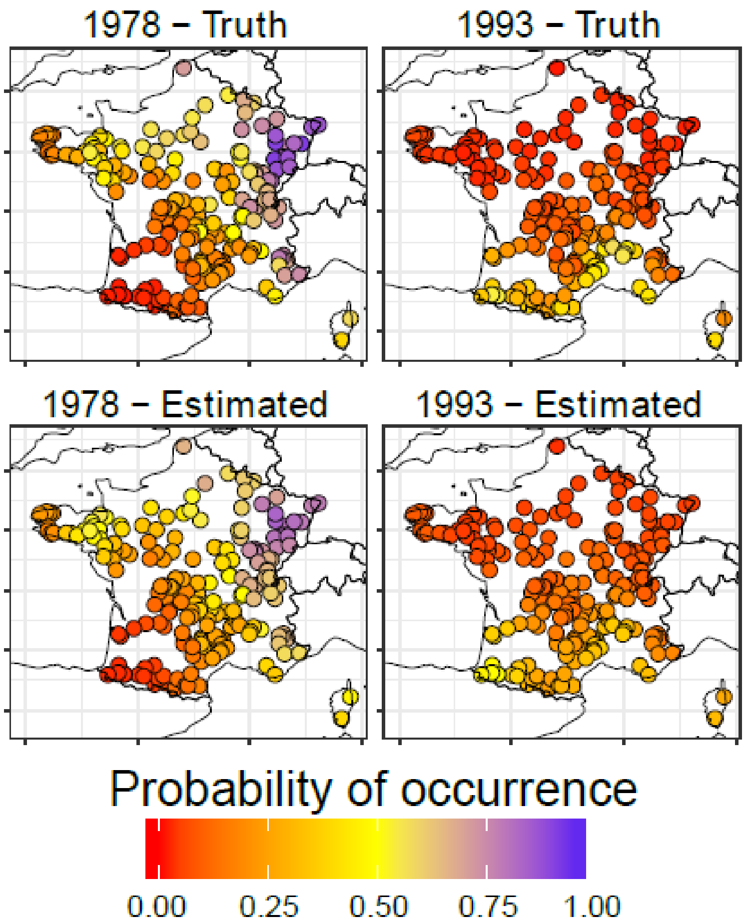
Synthetic case study

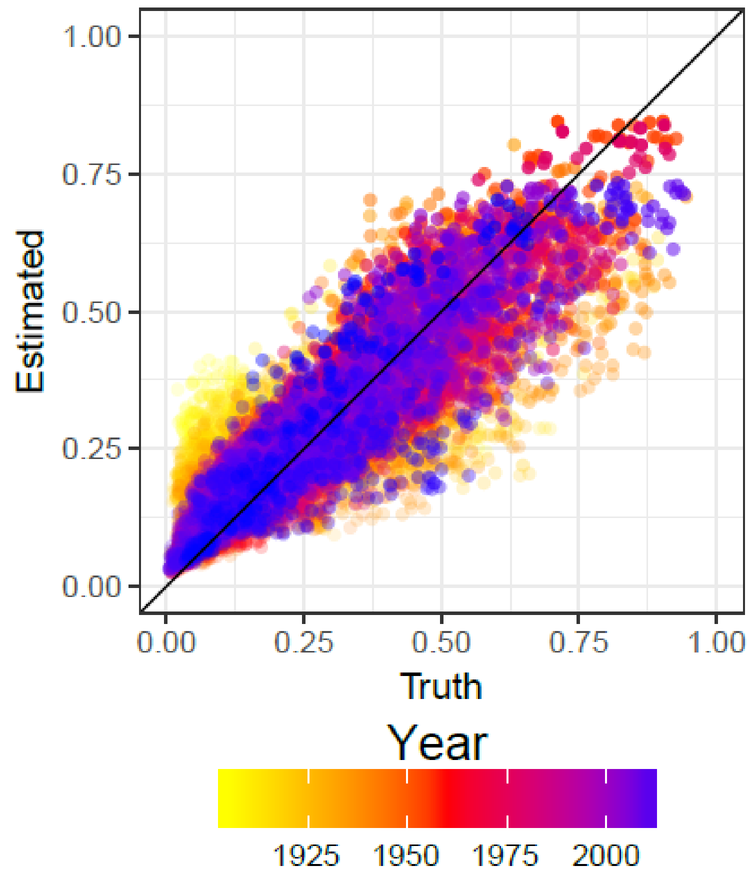
Ability to estimate HCI effects



Synthetic case study

Ability to estimate occurrence probabilities



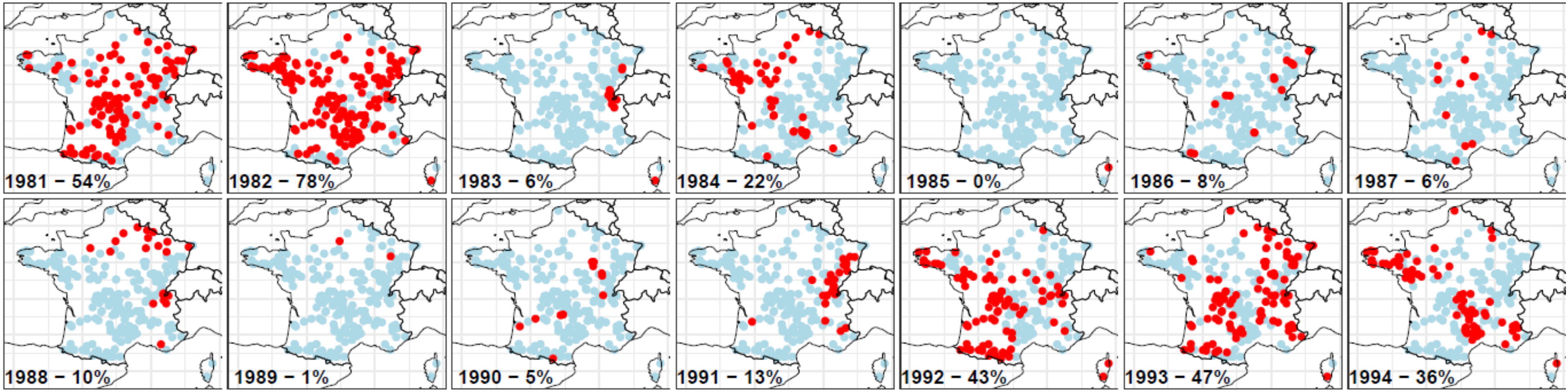


Autumn flood occurrences in France

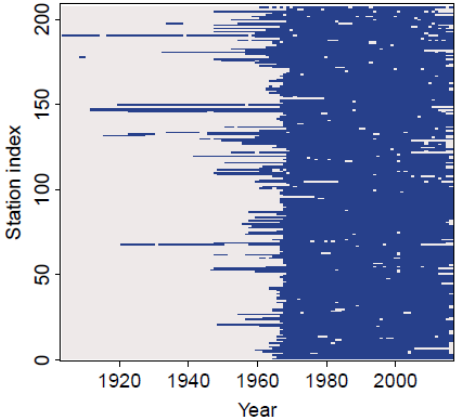
Data

207 stations, 1905-2014

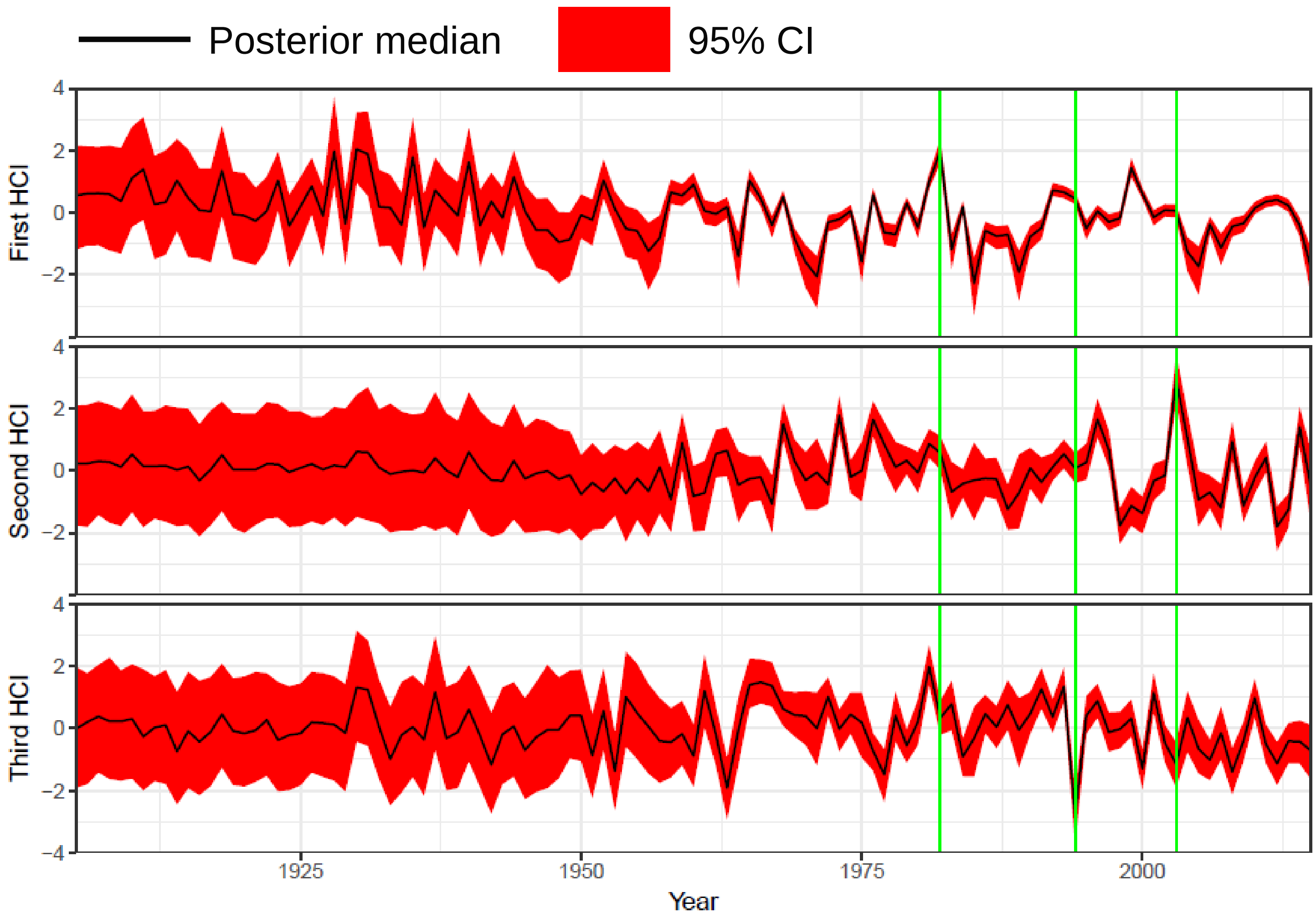
Floods exceeding a threshold such as Pr(occurrence) ≈ 0.2 at all sites



Missing data



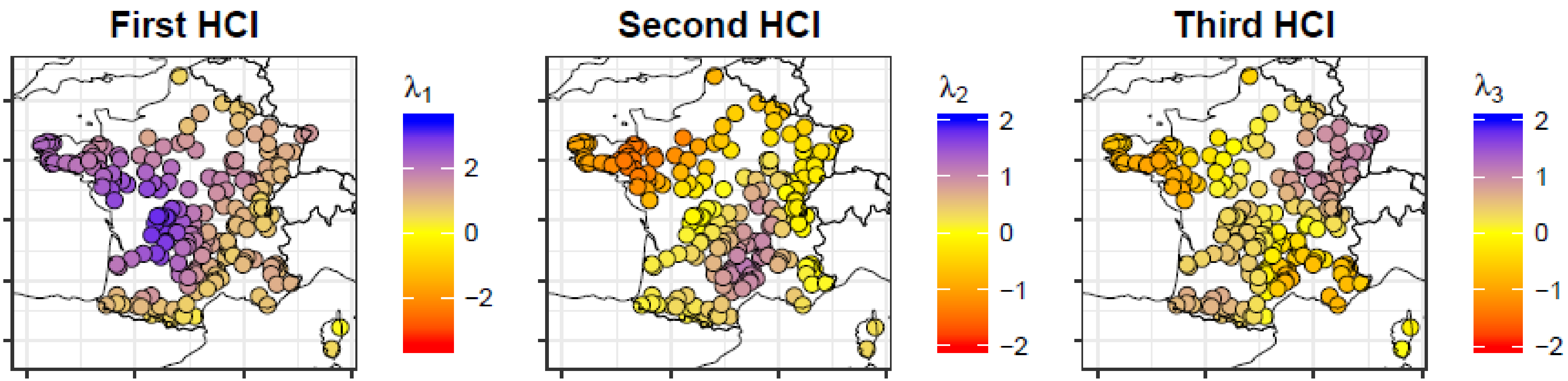
Autumn flood occurrences in France
Hidden climate indices



No trend, autocorrelation or low-frequency variability in estimated HCI's

Autumn flood occurrences in France

HCI effects



First HCI: always positive, but large only in oceanic part of France

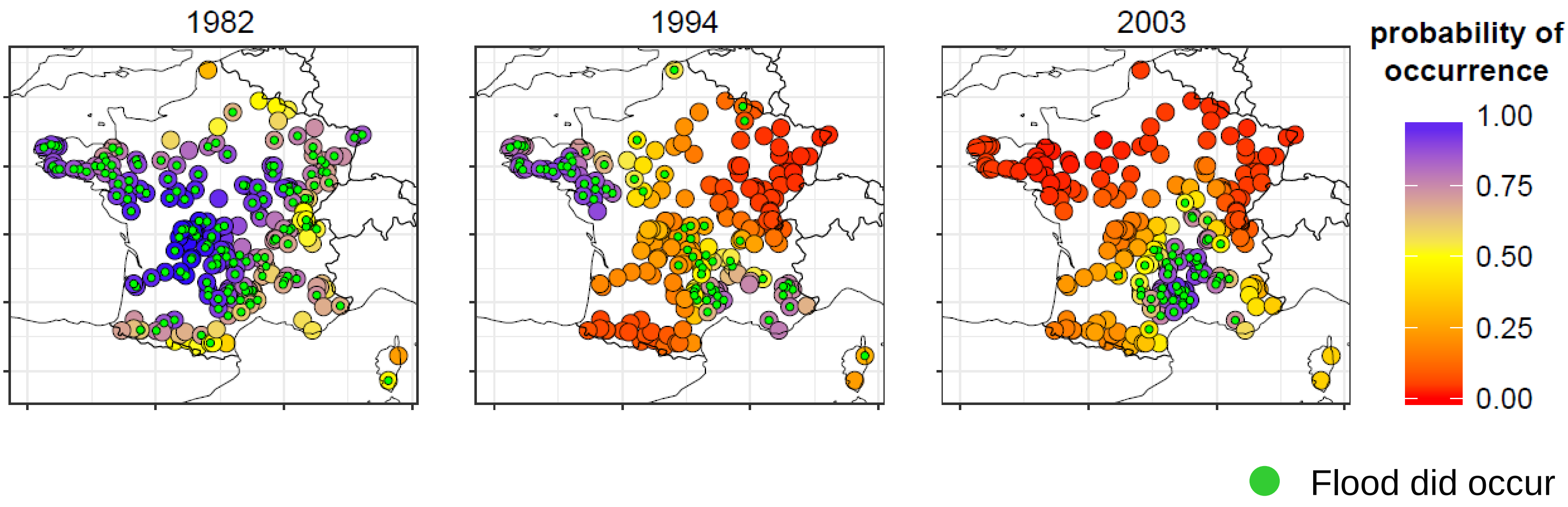
Second HCI: opposition Britany – Cevennes

Third HCI: effect gets smaller…

Overall, first HCI has by far the largest effects

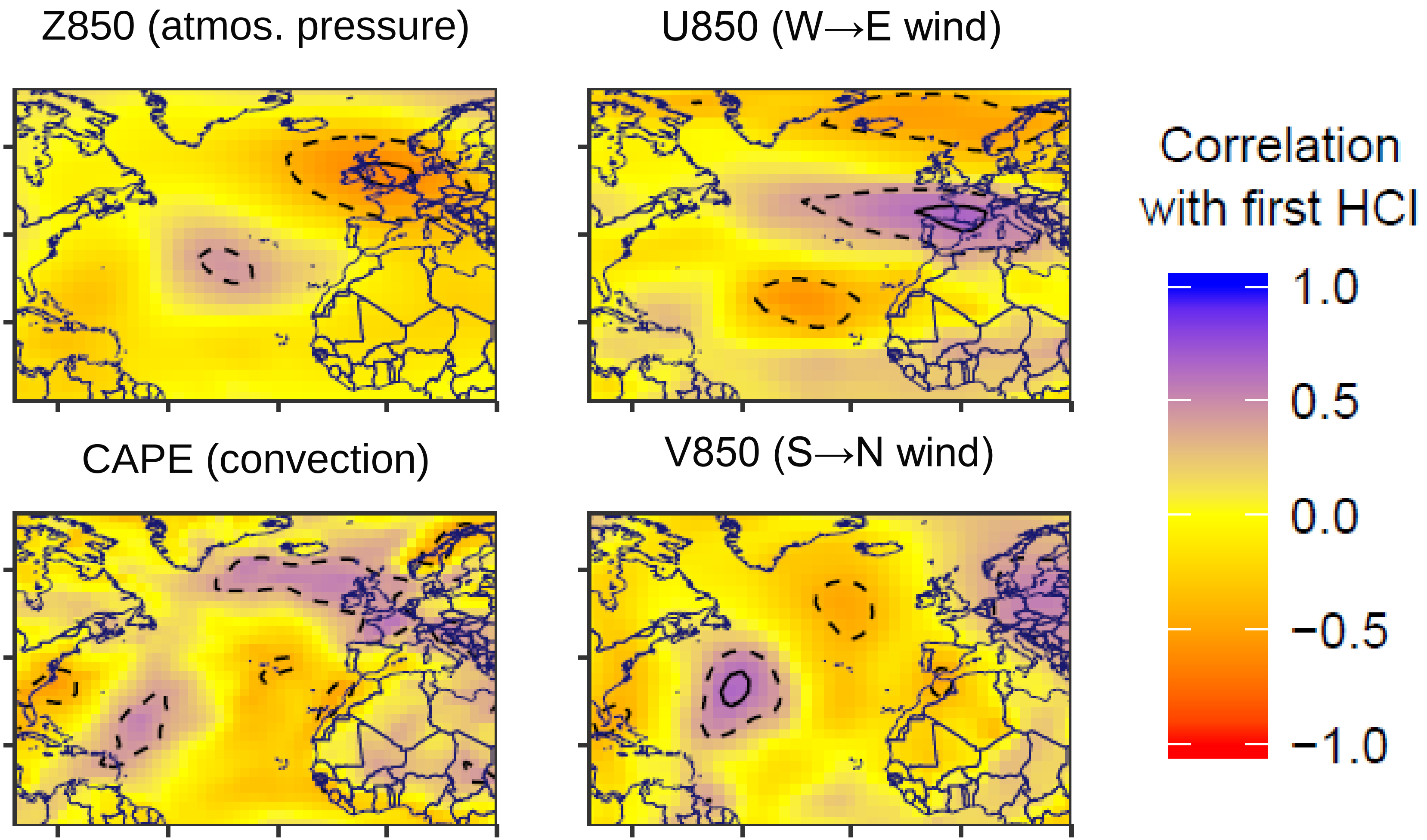
Autumn flood occurrences in France

Probabilities of occurrence



Autumn flood occurrences in France

Climate patterns associated with 1st HCI



Conclusions

A Bayesian hierarchical model to describe the space-time variability of (flood) occurrence data

Based on the identification of hidden climate indices

Useful in cases where standard climate indices have poor predictive capability

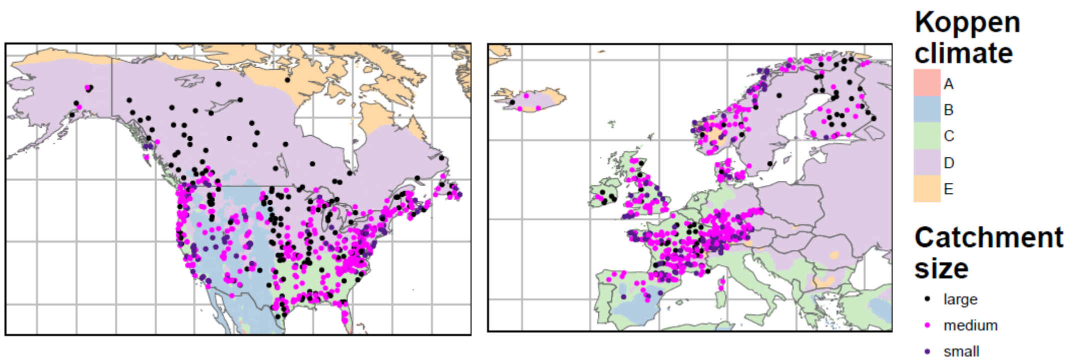
Synthetic case study shows that extracting hidden climate indices from occurrence data alone is feasible

Real-life case study suggests that hidden climate indices are linked with specific climate patterns, giving hope for predictictability

Future work

Prediction, cross-validation

- So far, link with large-scale climate only exploratory (correlation maps)
- Develop the method to predict occurrences from large-scale climate data (as opposed to standard climate indices)
- Implement cross-validation experiments



Apply to an even larger spatial scale… but computational issues

- Number of sites is the computational bottleneck
- Virtually intractable with thousands of sites
- Numerical tricks are possible (approximation of the Metropolis ratio)

Future work

Clarify links with other existing approaches

- "Continuous" version of a Hidden Markov model
- Principal component analysis
  - *The rationale is quite similar, adaptation to occurrence data*
  - *If Gaussian rather than Bernoulli distribution, back to PCA? (cf. Tipping and Bishop 1999)*
  - *Spatial hyperdistribution makes a difference though*

Tipping, M. E., and C. M. Bishop (1999), Probabilistic Principal Component Analysis, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 61(3), 611-622.

Modelling intensities rather than occurrences

- Replace occurrences by seasonal maxima, Bernoulli by GEV
- 3 parameters, but see GLM / GAMLSS
- Are intensities predictable from large-scale climate? (my guess: not much)
- Link with spatial extremes theory (cf. Reich and Shaby 2012)

Reich, B. J., and B. A. Shaby (2012), A Hierarchical Max-Stable Spatial Model for Extreme Precipitation, Ann Appl Stat, 6(4), 1430-1451.