# Spatio-temporal metamodeling for West African monsoon

Clémentine Prieur
Université Joseph Fourier, EPI MOISE

AgroParisTech, 21/06/2011

# Multidisciplinarity

This work is part of a project involving:

- Physicists:
  - ⋆ Hubert GALLÉE (LGGE, Grenoble),
  - ⋆ Christophe MESSAGER (IFREMER, Brest),
  - ⋆ LTHE, Grenoble.
- Statisticians:
  - ⋆ Anestis ANTONIADIS,
  - ⋆ Céline HELBERT (EPI MOISE, Grenoble),
  - ⋆ Clémentine PRIEUR (EPI MOISE, Grenoble),
  - ⋆ Laurence VIRY (EPI MOISE, Grenoble).
- Computer specialists:
  - ⋆ Laurence VIRY (CIMENT, Grenoble),
  - ⋆ Eddy CARON (EPI GRAAL, Lyon).
  - ⋆ Benjamin DEPARDON (SysFera).

# Contents

# Context : African Monsoon Multidisciplinary Analysis

West African climate is driven by a monsoon phenomenon which is active from May to September.

The cumulative rainfall in West African, from the equatorial zone to the Sahelian one, is weak (500-600 mm).

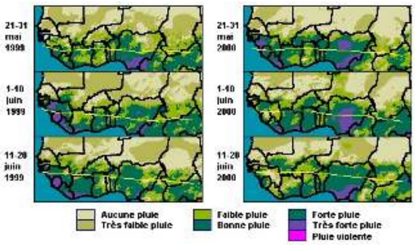
A small variation of the rainfall can have a dramatical impact on the agricultural activities, thus on the populations itself.

If this accumulation continues from one year to another, the consequences become sustainable ecosystems which tend to shift to type ecosystems Sahara.

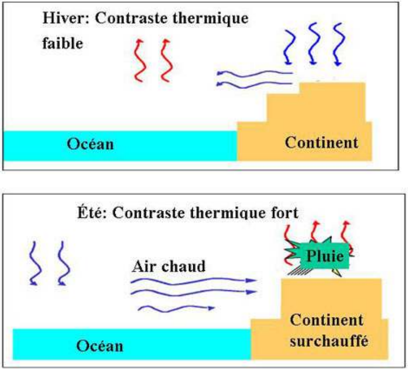
# West African monsoon

West African monsoon is related to

- the semi-annual deplacement of the InterTropical Zone of Convergence (ZCIT),
- the temperature gradient between the (sub-)Saharian zone and the equatorial atlantic cost in the gulf of Guinea.
- the dry trade winds from the north-east (particularly their most intense form *the harmattan*) are replaced by South-Western monsoon winds during summer.




# Temperature contrasts



# MAR - Regional Atmospheric Model

We have the following scheme

physical phenomena $\Rightarrow$ mathematical models $\Rightarrow$ simulation codes

- Underlying mathematical model (regional climate model)
  - ⋆ Atmosphere
    - hydrostatic primitive equations (Navier-Stokes-type)
    - parametrization of subgrid dynamic processes (turbulence, horizontal diffusion, digital filter)
    - hydrological cycle described by conservation equations: specific humidity, droplets and cloudy crystals, raindrops and snow particles
    - parametrization of cloud microphysical processes and atmospheric convection
  - ⋆ Surface
    - conservation of heat and soil water

# The MAR simulation code

huge computing power required, with a large number of runs

Note that

- a run: 15 days, $\sim$15H CPU,
- to simulate 1 year one needs to observe 2 years $\Rightarrow$ 48 (2x24) runs for 1 year, we simulate 17 years,
- 3 - 4 Go for inputs/outputs,
- these simulations are launched one after the other,
- the ending state of the simulation of one month is used as the initial state of the next month.

Local platforms: approximatively 15 parallel computer $\sim$ 2000 cores

$\Rightarrow$ we launch parallel simulations on 10 or even more nodes (according to the load of the platforms).

# DIET

DIET is a *middleware* able to find an appropriate server according to the information given in the client's request :

- problem to be solved, size of the data involved,
- the performance of the target platform (e.g. server load, available memory, communication performance),
- the local availability of data stored during previous computations.

Data management is provided to allow persistent data to stay within the system for future re-use.

This feature avoids unnecessary communication when dependencies exist between different requests.

# DIET

Approach : a grid middleware approach with the Distributed Interactive Engineering Toolbox DIET
`http://graal.ens-lyon.fr/DIET/`

DIET is developped by the GRAAL (Algorithms and Scheduling for Distributed Heterogeneous Platforms) team of the LIP (Laboratoire de l'Informatique du Parallélisme) laboratory of the ENS Lyon.

Description : DIET consists of a set of elements that can be used together to build applications, using the classical Remote Procedure Call (RPC) paradigm.

# Context

Objective: performing a sensitivity analysis for our applicative case. What is the influence of input parameters (albedo, sea surface temperatures in the gulf of Guinea, . . . ) on the output (rainfall in West Africa)?

Our approach: global stochastic sensitivity analysis.

Main issues:

⋆ many-query setting,

⋆ running the MAR model is highly time consuming and requires huge storage capacities.

# Metamodel

To by-pass the problems due to MAR's complexity, we wish to fit a stochastic metamodel.

Main properties required for the surrogate model:

- taking into account the spatio-temporal dynamics,
- computing quickly the outputs.

$$\text{Sea Surface Temperature} \xrightarrow{\text{metamodel}} \text{Precipitation}$$

A first step is the spatio-temporal modeling of the inputs/outputs.

# Presentation of the day

Phase I: elaboration of an appropriate meta-model fitted on observations.

A. Antoniadis, C. Helbert, L. Viry, C. Prieur, Spatio-temporal prediction for West African monsoon. Submitted. `hal-00551303v1`

Key point: modeling and fitting high-dimensional response regression models in the setting of complex spatio-temporal dynamics.

# Modeling of inputs/outputs

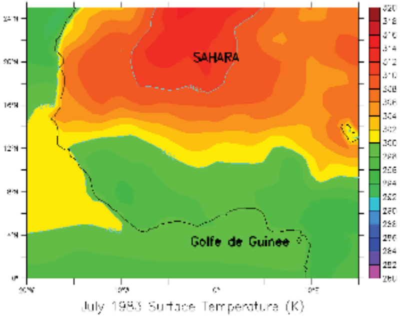
**Inputs: sea surface temperature**

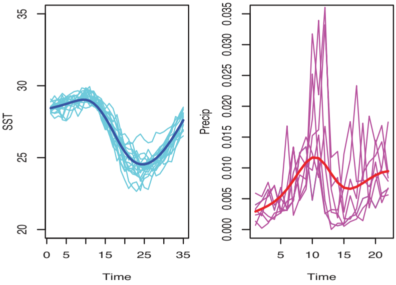
- Reynolds climatological data (satellite and in situ),
- weakly data on 18 years: 1983 to 2000,
- observations on a mesh $\mathcal{G}$ covering the area $\mathcal{R}$, $[5S : 5N] \times [30W : 10E]$, with a space step of $1^\circ$ latitude and longitude (516 points).

**Outputs: rainfall**

- weakly observations of IRD on 8 years: 1983 to 1990,
- obtained on a mesh $\mathcal{G}'$ covering the area $\mathcal{R}'$, $[10S : 30N] \times [20W : 20E]$ (368 points when removing incomplete data).

# Statistical framework



# Statistical framework

Data:

$\mathbf{X}^i := (X_i(x,t))_{x\in\mathcal{R},t\in\mathcal{T}}$ SST year $i = 1,\ldots,18$

$\mathbf{Y}^j := (Y_j(y,t))_{y\in\mathcal{R}',t\in\mathcal{T}}$ rainfall year $j = 1,\ldots,8$

Methodology:

Fix $x_0 \in \mathcal{G} \subset \mathcal{R}$.

Fix $y_0 \in \mathcal{G}' \subset \mathcal{R}'$.

West African monsoon is a periodic phenomenon, active period from May to September, observed on 8 years.

The input at point $x_0$ is a random trajectory $(X_i^{x_0})$, $i = 1,\ldots,18$.

The output at point $y_0$ is a random trajectory $(Y_j^{y_0})$, $j = 1,\ldots,8$.

# Statistical modelling of the SST

$X_i^{x_0}$ is assumed to belong to some Hilbert functional space $\mathbb{H} \subset \mathbb{L}^2(\mathcal{T})$.

Assume $X^{x_0}$ is smooth, with

- unknown smooth mean function $\mathbb{E}X^{x_0}(t) = \mu_{X^{x_0}}(t)$,
- unknown smooth covariance function $\mathrm{Cov}\left(X^{x_0}(s), X^{x_0}(t)\right) = G_{X^{x_0}}(s,t)$.

We assume the existence of an orthogonal expansion of $G_{X^{x_0}}$:

$$G_{X^{x_0}}(s,t) = \sum_{m \geq 1} \rho_m(x_0) e_m(x_0, s) e_m(x_0, t), \; s, \; t \in \mathcal{T},$$

with eigenvalues $\rho_1(x_0) \geq \rho_2(x_0) \geq \ldots$

# Karuhnen-Loève decomposition

We perform a Karuhnen-Loève decomposition:

$$X^{x_0}(t) = \mu_{X^{x_0}}(t) + \sum_{m=1}^{\infty} \alpha_m(x_0) e_m(x_0, t), \; t \in \mathcal{T}.$$

For this fixed grid point, we use an appropriate truncation

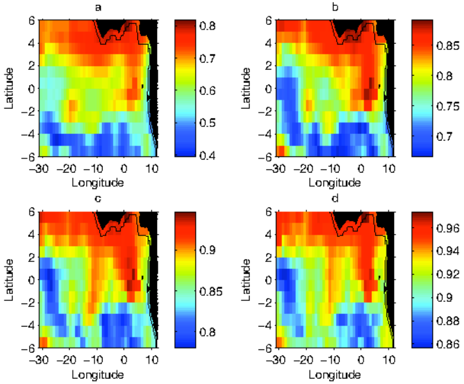
$$X^{KL,x_0}(t) = \mu_{X^{x_0}}(t) + \sum_{m=1}^{N_{x_0}} \alpha_m(x_0) e_m(x_0, t), \; t \in \mathcal{T}.$$

The truncation is chosen in order to explain more than 70% of the variance.

Assumption: the truncation $N_{x_0}$ does not depend on the spatial location that is $N_{x_0} = M$.

# For SST

First do a KL decomposition on each point $x$ of the mesh $\mathcal{G}$. The truncation criterion is based on the percentage of explained variance.




We can take $M = 2$.

# Model, stationarity by group

Estimation of time dependent eigenfunctions at different spatial locations generates great amounts of high-dimensional data: crucial need of clustering algorithms.

# Model, stationarity by group

Estimation of time dependent eigenfunctions at different spatial locations generates great amounts of high-dimensional data: crucial need of clustering algorithms.

Model:
We share $\mathcal{R}$ in $p$ subregions, $\mathcal{R}_1, \ldots, \mathcal{R}_p$. We choose $x_{0,1} \in \mathcal{R}_1, \ldots, x_{0,p} \in \mathcal{R}_p$.

If $j \in \{1, \ldots, p\}$, if $x \in \mathcal{R}_j$,

$$\tilde{X}^x(t) = \mu_{X^x}(t) + \sum_{m=1}^{M} \tilde{\alpha}_m(x) e_m(x_{0,j}, t),$$

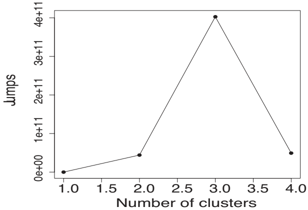
with $\tilde{\alpha}_m(x) = \int_{\mathcal{T}} \tilde{X}^x(t) e_m(x_{0,j}, t) dt$.

We need clustering algorithms for functional data, taking into account the between time-point correlation.

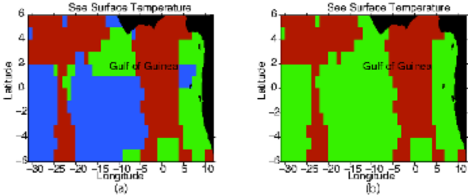
We need clustering algorithms for functional data, taking into account the between time-point correlation.

Data-driven clustering method by Ma *et al.* (2006): smoothing spline clustering (SSC)
mixed-effect smoothing spline + EM algorithm.

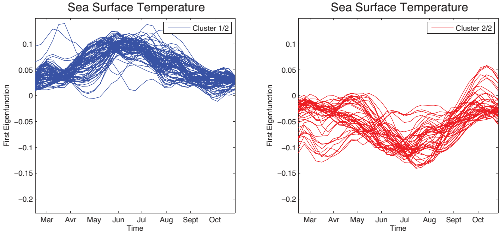
For the choice of the number of clusters, we consider the transformed distortion curve $(K, d_K)$, where $d_K$ denotes the minimum achievable distortion associated with fitting $K$ centers to the data.



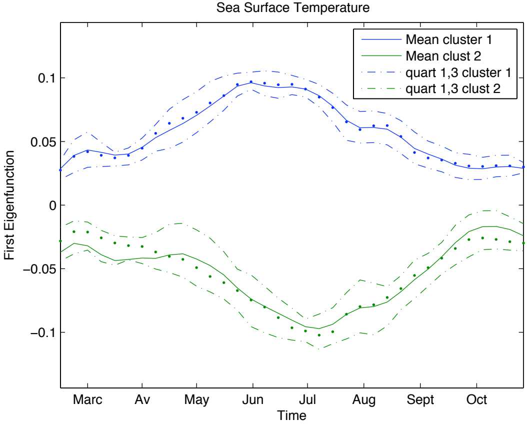

Jumps for the distortion measure $d_K - d_{K-1}$
with respect to the number $K$ of clusters




(a) Projection on the map for three clusters; (b) for two clusters

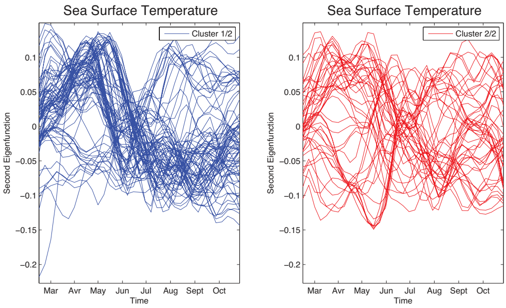



Estimated curves for the first eigenfunction by cluster



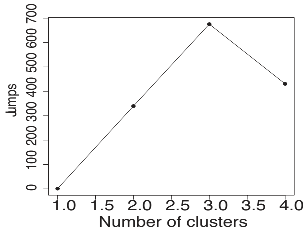

Mean curves on each cluster with their lower-upper quartile bands

# What happens for the second eigenfunctions

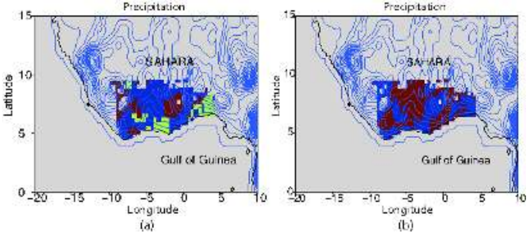



Estimated curves for the second eigenfunction by cluster

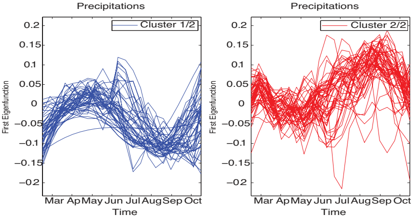
# Following the same idea for the rainfall



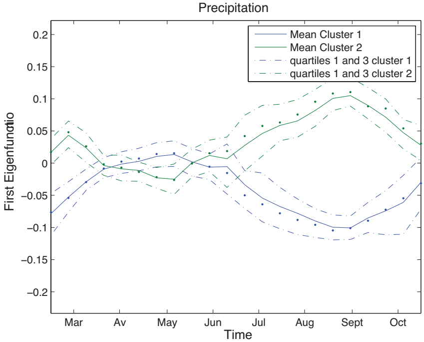

Jumps for the distortion measure $d_K - d_{K-1}$ with respect to the number $K$ of clusters



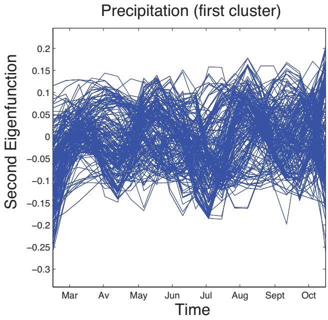

(a) Projection on the map for three clusters; (b) for two clusters

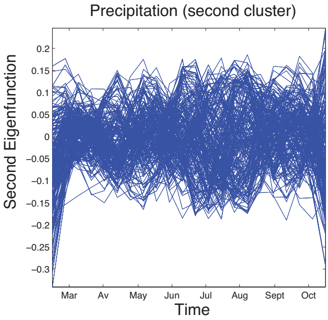



Estimated curves by cluster for the first eigenfunction




Mean curves on each cluster with their lower-upper quartile bands




Estimated curves by cluster for the second eigenfunction

# Conclusions for rainfall

We share $\mathcal{R}'$ in $q$ subregions, $\mathcal{R}'_1, \ldots, \mathcal{R}'_q$. For any $l \in \{1, \ldots, q\}$, we choose $y_{0,l} \in \mathcal{R}'_l$.

If $l \in \{1, \ldots, q\}$, if $y \in \mathcal{R}'_l$,

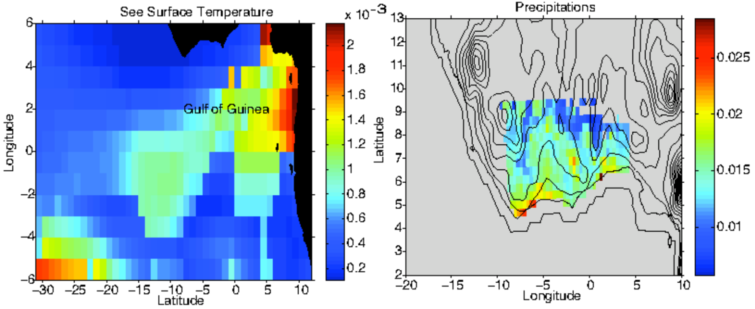
$$\tilde{Y}^y(t) = \mu_{Y^y}(t) + \sum_{k=1}^{K} \tilde{\beta}_k(y) f_k(y_{0,l}, t),$$

with $\tilde{\beta}_k(y) = \int_{\mathcal{T}} \tilde{Y}^y(t) f_k(y_{0,l}, t) dt$.

We do not define $t \longrightarrow f_2(y_{0,l}, t)$ as the second eigenfunction at point $y_{0,l}$ but as the mean curve $t \longrightarrow f_2(t)$ of all curves $t \longrightarrow f_2(y, t)$, $y \in \mathcal{G}'$.

We take $K = 2$ and $q = 2$.

# Relative Mean Squared Error




Relative Mean Squared Error for the reconstruction of SST (left) and of Precipitation (right)

# Estimation procedure: SST

Let $x \in \mathcal{G}$. Then, $\exists j \in \{1, \ldots, p\}$ such that $x \in \mathcal{G}_j$.

Tools for estimating $\mu_{X^x}(\cdot)$ and $G_X(s,t)$:
local linear smoothing + cross-validation
(see Fan & Gijbels 1996, Yao *et al.* 2005).

Tools for estimating eigenfunctions and eigenvalues: one solves the eigenequations

$$\int_{\mathcal{T}} \widehat{G}_X(s,t)\hat{e}_m(x_{0,j},s)ds = \hat{\rho}_m \hat{e}_m(x_{0,j},t),$$

with $\int_{\mathcal{T}} \hat{e}_m(x_{0,j},t)^2 dt = 1$ and $\int_{\mathcal{T}} \hat{e}_k(x_{0,j},t)\hat{e}_m(x_{0,j},t)dt = 0$ for $m < k$.
Eigenfunctions estimated by discretizing smoothed covariance.

# Estimation procedure: SST

Tools for estimating $\tilde{\alpha}^i_m(x)$, for $m = 1, \ldots, M$ and each year $i = 1, \ldots, 18$: we use projection estimates

$$\sum_{k=2}^{T} X_i^x(t_k)\hat{e}_m(x_{0,j}, t_k)(t_k - t_{k-1})\,.$$

The estimation for each individual curve is needed for the selection procedure of the regression.

# A multivariate regression approach

Define

$$\underline{\alpha_m} = (\widetilde{\alpha}_m(x_1), \ldots, \widetilde{\alpha}_m(x_{\#\mathcal{G}})) \quad m = 1, 2$$

and

$$\underline{\beta_k} = \left(\widetilde{\beta}_k(y_1), \ldots, \widetilde{\beta}_k(y_{\#\mathcal{G}'})\right) \quad k = 1, 2\,.$$

Context: the sample size (8) is much smaller than the spatial components ($\#\mathcal{G} = 516$, $\#\mathcal{G}' = 368$).

# Penalization approach

$$Y_j = \sum_{i=1}^{2\,\#\mathcal{G}} X_i B_{ij} + \epsilon_j, \quad j = 1, \ldots, 2\,\#\mathcal{G}',$$

where the error terms $\epsilon_1, \ldots, \epsilon_{2\,\#\mathcal{G}'}$ have a joint distribution with mean 0 and covariance $\Sigma$.

$$\left(Y_1, \ldots, Y_{2\,\#\mathcal{G}'}\right) = \left(\underline{\beta_1}, \underline{\beta_2}\right),$$

$$\left(X_1, \ldots, X_{2\,\#\mathcal{G}}\right) = \left(\underline{\alpha_1}, \underline{\alpha_2}\right).$$

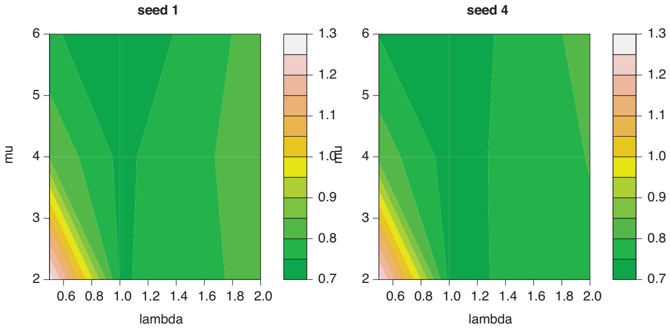
# Penalization approach

$$\ell_{(\lambda,\mu)}(\beta, B) = \frac{1}{2}\|\beta - \alpha B\|_F^2 + \lambda \sum_{i=1}^{2\,\#\mathcal{G}} \|\mathbf{C}_i \cdot B_i\|_1 + \mu \sum_{i=1}^{2\,\#\mathcal{G}} \|\mathbf{C}_i \cdot B_i\|_2,$$

where $\mathbf{C}$ is a $2\,\#\mathcal{G} \times 2\,\#\mathcal{G}'$ 0-1 matrix indicating the coefficients of $B$ on which penalization is imposed.

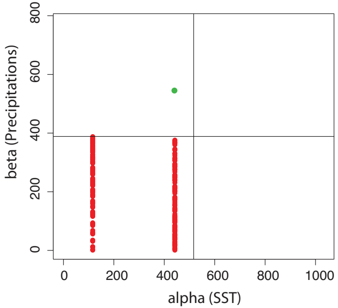
Finally, an estimate of the coefficient matrix $B$ is $\widehat{B}_{\lambda,\mu} := \mathbf{argmin}_B \ell_{(\lambda,\mu)}(\mathbf{Y}, B)$ (see Peng *et al.*, 2010).

# Implementation on our test-case



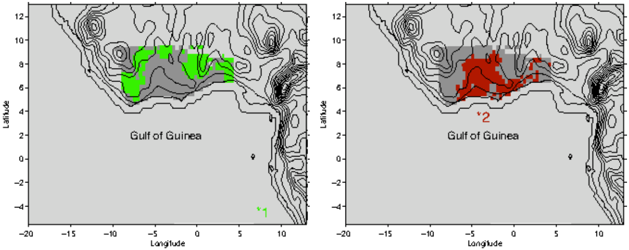

Cross validation for the choice of
$\lambda$ and $\mu$ (with two different germs)

# Implementation on our test-case



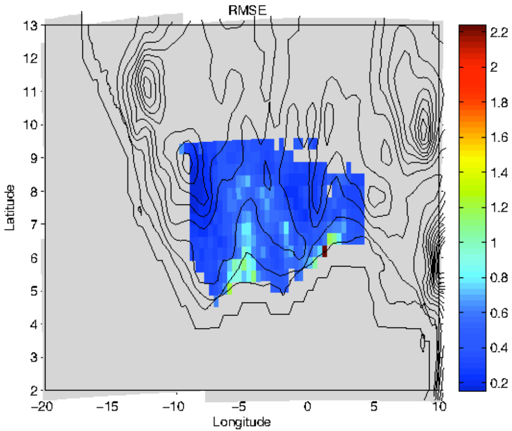

Regression coefficients matrix $B$
estimated with $\lambda = 1$ and $\mu = 4$

# Implementation on our test-case



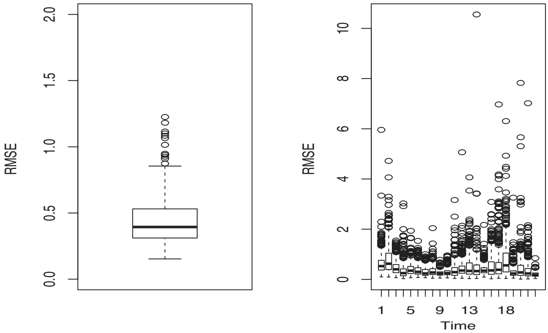

Spatial location for the average responses indicated by the retained coefficients for both predictors (points 1 and 2 on the map).

# Relative MSE



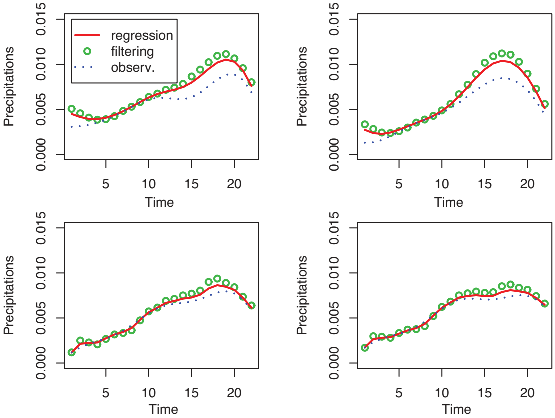

Relative MSE for the reconstructed precipitation by regression on the map.

# Annual and weekly relative MSE




Boxplots of the relative MSE per year ( left) and per week (right)

# Comparison




Reconstructed precipitation curve via regression (red), via the truncated Karhunen-Loève decomposition (circles) and observed precipitation (dots).

# Perspectives and conclusions

⋆ Our spatio-temporal modeling for the inputs (resp. the outputs) seems relevant: small relative error on the grid.

⋆ DIET middleware used to transparently execute MAR workflows on the Ciment grid. Soon available.

Perspectives:

- Working with the code outputs or/and also getting more years of observations, phase in progress.
- Consider also the influence of albedo, . . .
- Provide an even more transparent access to the grid through DIETWebboard.

- Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Application.* London: Chapman and Hall.
- Hörmann, S. and Kokoszka, P. (2010). Weakly dependent functional data. *The Annals of Statistics* **38**, 3, 1845-1884.
- Ma, P, Castillo-Davis, C. I., Zhong, W. and Liu, J. S. (2006). A data-driven clustering method for time course gene expression data. *Nucleic Acids Research* **34**, 4, 1261-1269.
- Messager, C., Gallée, H. and Brasseur O. (2004). Precipitation sensitivity to regional SST in a regional climate simulation during the West African monsoon for two dry years. *Climate Dynamics* **22**, 249-266.
- Peng, J., Zhu, J., Bergamaschi, A., Han, W., Noh D.- Y., Pollack, J. R. and Wang, P. (2010). Regularized Multivariate Regression for Identifying Master Predictors with Application to Integrative Genomics Study of Breast Cancer. `http://arxiv.org/abs/0812.3671`
- Yao, F., Müller H.-G. and Wang J.-L. (2005). Functional Data Analysis for Sparse Longitudinal Data. *Journal of the American Statistical Association* **100**, 470, 577-590.
- Caron, E. and Desprez, F. (2006). DIET: A scalable toolbox to build network enabled servers on the grid. *International Journal of High Performance Computing Applications* 20(3): 335-352.