Spatio-temporal prediction for West African monsoon

Anestis Antoniadis[a], Céline Helbert[a], Clémentine Prieur[a,*], Laurence Viry[a]

[a] *Université de Grenoble, Laboratoire Jean Kuntzmann, BP 53, 38041 Grenoble cedex, France*

Abstract

In this paper, we propose a new approach for modeling and fitting high-dimensional response regression models in the setting of complex spatio-temporal dynamics. This study is motivated by investigating one of the major atmospheric phenomena which drives the rainfall regime in Western Africa : West African Monsoon. We are particularly interested in studying the influence of sea surface temperatures in the Gulf of Guinea on precipitation in Saharan and sub-Saharan.

Keywords: spatio-temporal modeling, filtering, multivariate penalized regression

1. Introduction

West African monsoon is the major atmospheric phenomenon which drives the rainfall regime in Western Africa. It is characterized by a strong spatio-temporal variability whose causes have not yet been determined in an unequivocal manner. However, there is a considerable body of evidence suggesting that spatio-temporal changes in sea surface temperatures in the Gulf of Guinea and changes in the Saharan and sub-Saharan albedo are major factors. One of the interest of physicists is to perform sensitivity analysis on West African monsoon (see Messager *et al.* (2004)). The main tool for simulating precipitation is a regional atmospheric model (MAR) whose performances were evaluated by

∗. Corresponding author. Email : clementine.prieur@imag.fr, phone : +33 (0)4 76 63 59 63, fax : +33 (0)4 76 63 12 63.

comparisons with precipitation data. Global sensitivity analysis of a model output consists in quantifying the respective importance of input factors over their entire range of values. Contrarily to deterministic approaches based on gradients, global analyses can be performed on nonlinear systems. Many techniques have been developed in this field (see Saltelli *et al.* (2000) for a review). Performing a global sensitivity analysis implies running the model a large number of times. However it can not be realized by running the MAR, as we work on large discretization grids in space and time, thus dealing with huge dimensions. A way for overcoming this issue is to fit a stochastic model which approximates the MAR by taking into consideration the spatio-temporal dynamic of the underlying physical phenomenon and with the ability to be run in a reasonable time. Statistical methods can be used to describe the behavior of a set of observations by focusing attention on the observations themselves rather than on the physical processes that produced them. One of those statistical methods is regression and in this paper we focus on the regression of precipitation on sea surface temperatures. As the numerical storage and processing of our model outputs (precipitation), as far as the statistical description of the data is concerned, require considerable computational resources, it will be run in a grid-computing environment (see Caron *et al.* (2006)). This grid deployment takes into account the scheduling of a huge number of computational requests and links with data-management between these requests, all of these as automatically as possible. It requires new developments which are not at the moment completely achieved. It explains why we fit our model with real data in this study : Reynolds climatological data on eighteen years (1983 to 2000) for sea surface temperatures and data collected by the french IRD (*Institut de Recherche pour le Développement*) during a period of eight years (1983 to 1990) for precipitation. The regression is achieved on the common period of observation (from 1983 to 1990). The poor quality of data over longer periods explain our restrictive choice (see Messager *et al.* (2004)). It

is clear that the study will be enhanced as soon as the grid deployment will be achieved, allowing a regression to be fitted on a longer period of eighteen years.

The paper is organized as follows. In Section 2 we give a brief description of the data. Our new approach for modeling both sea surface temperatures and precipitation is described in Section 3. Section 4 is devoted to the regression analysis. To conclude we mention in Section 5 some of the many interesting perspectives of our study.

2. Data description

This section is devoted to the description of our data sets, chosen in accordance with physicists. The data used for sea surface temperatures (SST) are the so-called Reynolds climatological data, generated by an optimal interpolation technique (Reynolds and Smith (1994)) which uses satellite and in-situ data. We obtain a value for SST at each of the 516 points of a spatial grid $\mathcal{G}$ located in the Gulf of Guinea. West African Monsson is a periodic phenomenon, active from May to September. We work with a time discretization : we have weekly data from March to November (to cover the active period of the physical phenomenon). For these data we have eighteen years of observations, from 1983 to 2000.

Precipitation data have been recorded by the Institut de Recherche pour le Développement (IRD) on a spatial grid $\mathcal{G}'$ of size 382 located in Western Africa with the greatest density of stations located between 5° N and 15° N (see Messager (2005)). We also have weekly data from March to November, but only from 1983 to 1990. After removing points on $\mathcal{G}'$ for which data were incomplete we work with 368 points.

Below we present a map focusing on the region of interest around the Gulf of Guinea (see left panel of Figure 2). We also show on the right panel of the same figure the 18 time-dependent curves of sea surface temperatures and the 8 time-dependent curves of precipitation at some fixed spatial point.

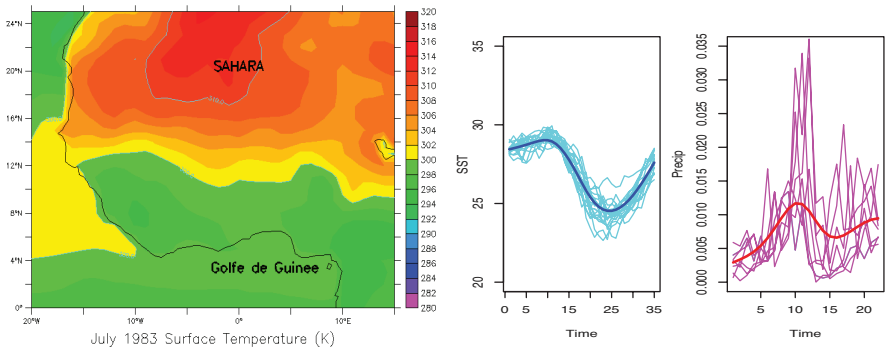

FIGURE 1: Left : Zone of interest for the study of West-African monsoon ; Right : Time-dependent curves for SST (left) *resp.* Precip (right) for each of the 18 (*resp.* 8) years of observations for some fixed spatial point $x \in \mathcal{G}$ (*resp.* $x' \in \mathcal{G}'$).

Both inputs (SST) and outputs (precipitation) depend on space and time. Spatial and time discretizations result in very high-dimensional data which are difficult to analyze with classical multivariate analysis. Functional data analysis (FDA) goes one big step further and seems the appropriate statistical tool to be used for analyzing our data for which time dynamics and spatial dynamics are a major component. Moreover an overarching theme in FDA is the necessity to achieve some form of dimension reduction of the infinite-dimensional data to finite and tractable dimensions and explains our choice to model inputs and outputs through spatio-temporal functional processes. For an introduction to the field of FDA, the two monographs by Ramsay and Silverman (2002, 2005) provide an accessible overview on foundations and applications, as well as a plethora of motivating examples.

3. Modeling inputs and outputs

The modeling for both inputs and outputs is described in this section. Our regression methodology to study the relationship between precipitation and the SST will be based

on such modeling. Our method is a new filtering approach based on Karuhnen-Loève decompositions and functional clustering. It allows reducing the dimensions involved in the data.

3.1. Functional modeling

Let $\mathcal{T}$ be a finite and closed interval of $\mathbb{R}$. We usually refer to $\mathcal{T}$ as time. The spatial regions of interest $\mathcal{R}$ and $\mathcal{R}'$ are both subsets of $\mathbb{R}^2$. In our applicative context, $\mathcal{T}$ is the annual time period from March to November. We model inputs (*resp.* outputs) on the spatial grid $\mathcal{G}$ (*resp.* $\mathcal{G}'$) described in Section 2. The phenomenon under study is a periodic phenomenon, with an active period from May to September, observed on N years (N is equal to 18 for SST and to 8 for precipitation). Let x be any point on the grid $\mathcal{G}$. Following Yao *et al.* (2005b), we consider that the ith observed time-dependent trajectory at point x corresponds to a sampled longitudinal curve viewed as realizations of random trajectories (X_i^x), $i = 1, \ldots, , N$, where X_i^x is assumed to belong to some Hilbert functional space $\mathbb{H} \subset \mathbb{L}^2(\mathcal{T})$. These X_i^x's are viewed as independent realizations of a stochastic process X^x with unknown smooth mean function $\mathbb{E}X^x(t) = \mu_{X^x}(t)$, and covariance function $\mathrm{Cov}\left(X^x(s), X^x(t)\right) = G_{X^x}(s,t)$. It is well known that under very mild conditions, there exists an orthogonal expansion of G_{X^x} (in the $\mathbb{L}^2$ sense) in terms of eigenfunctions $e_m(x,\cdot)$ with associated eigenvalues $\rho_m(x)$ (arranged in nonincreasing order), that is,

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

The random function $X^x(t)$ where t denotes time and x location, may be decomposed into an orthogonal expansion

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

The above representation of a random function is known as the *Karhunen-Loève* expansion, although in the meteorological literature it is known as the Empirical Orthogonal Function (EOF) expansion. It can be shown that the truncated decomposition with N_x terms

$$X^{\text{trunc},x}(t) = \mu_{X^x}(t) + \sum_{m=1}^{N_x} \alpha_m(x) e_m(x,t),\ t \in \mathcal{T}. \tag{1}$$

minimises the mean integrated squared error $\mathbb{E}\left\{\int_{\mathcal{T}}[X^x(t) - X^{\text{trunc},x}(t)]^2 dt\right\}$. The spectral representation is optimal in the sense that this error is minimum compared to N_x terms of any orthogonal system (see, e.g. Cohen and Jones (1977)). In our case we take as N_x as the truncation needed at point x to explain more than 80 % of the variance.

In our analysis, for each spatial grid point in the Gulf of Guinea and each year of observation, sea surface temperature is measured during the active period on a temporal grid. A Karhunen-Loève decomposition is then performed at each location on the spatial grid (see e.g. Yao *et al.* (2005b)). In order to achieve an optimal (in the least-squares sense) representation of the observed process, the appropriate number of terms N_x depends on the location on the spatial grid. To simplify the analysis we will consider in the following that N_x is bounded above by a number M independent of x. As one can see from Figure 2 such an assumption with $M = 2$ (i.e. with a cumulative percentage of variance explained that is larger than 70%) seems perfectly valid for our data on sea surface temperatures.

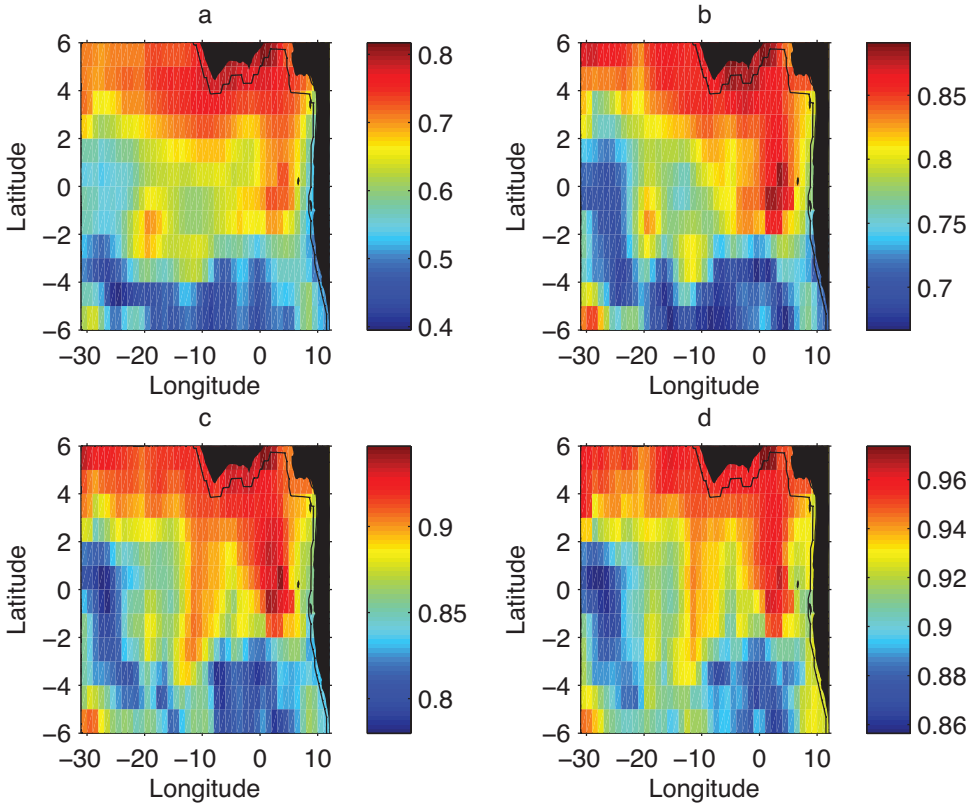

FIGURE 2: Percent cumulative variance for the SST on the map explained by reconstructing the SST using (a) one term (b) two terms (c) three terms and (d) four terms in the corresponding truncated *Karhunen-Loève* expansion

In our application, the number and shape of the eigenfunctions patterns over time are not known and the lack of stationarity makes them dependent on the spatial location. The estimation of these eigenfunctions at different spatial locations generates great amounts of high-dimensional data. It seems therefore reasonable to assume some kind of local stationarity by assuming that at least the resulting eigenfunctions are spatially piecewise constant.

Clustering algorithms become then crucial in reducing the dimensionality of such data. The choice of the clustering approach is described in Section 3.3. For the moment let us just assume that we know that there exist L_1 points $x_{0,1}, \ldots, x_{0,L_1}$ (with $L_1 \in \mathbb{N}^*$) on the spatial grid $\mathcal{G}$ partitioning $\mathcal{G} = \cup_{l=1}^{L_1} \mathcal{G}_l$ into L_1 subregions $\mathcal{G}_l$ that appear as a "natural" system of spatial coordinates that reflects the underlying internal and local stationary structures of the data. Given such a partition, for any x on $\mathcal{G}$, there exist $l \in \{1, \ldots, L_1\}$ and a specific point $x_{0,l}$ in $\mathcal{G}_l$ such that we can approximate $X^x(t)$ by

$$\widetilde{X}^x(t) = \mu_{X^x}(t) + \sum_{m=1}^{M} \widetilde{\alpha}_m(x) e_m(x_{0,l}, t), \ t \in \mathcal{T},$$

with $\widetilde{\alpha}_m(x) = \int_{\mathcal{T}} \widetilde{X}^x(t) e_m(x_{0,l}, t) dt$, for $m = 1, \ldots, M$.

The modeling for precipitation follows the same lines, leading to L_2 fixed grid points $y_{0,1}, \ldots, y_{0,L_2}$ (with $L_2 \in \mathbb{N}^*$) on the spatial grid $\mathcal{G}'$ partitioning $\mathcal{G}' = \cup_{l=1}^{L_2} \mathcal{G}'_l$ into L_2 subregions $\mathcal{G}'_l$. Then, given such a partition, for any y on $\mathcal{G}'$ there exist $l \in \{1, \ldots, L_2\}$ and a specific point $y_{0,l}$ in $\mathcal{G}'_l$ such that we can approximate $Y^y(t)$ by

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

with $\widetilde{\beta}_k(y) = \int_{\mathcal{T}} \widetilde{Y}^y(t) f_k(y_{0,l}, t) dt$ for $k = 1, \ldots, K$. The truncation number K is also assumed not to depend on $y \in \mathcal{G}'$ and will be chosen equal to 2 for our test case (see Figure 3).

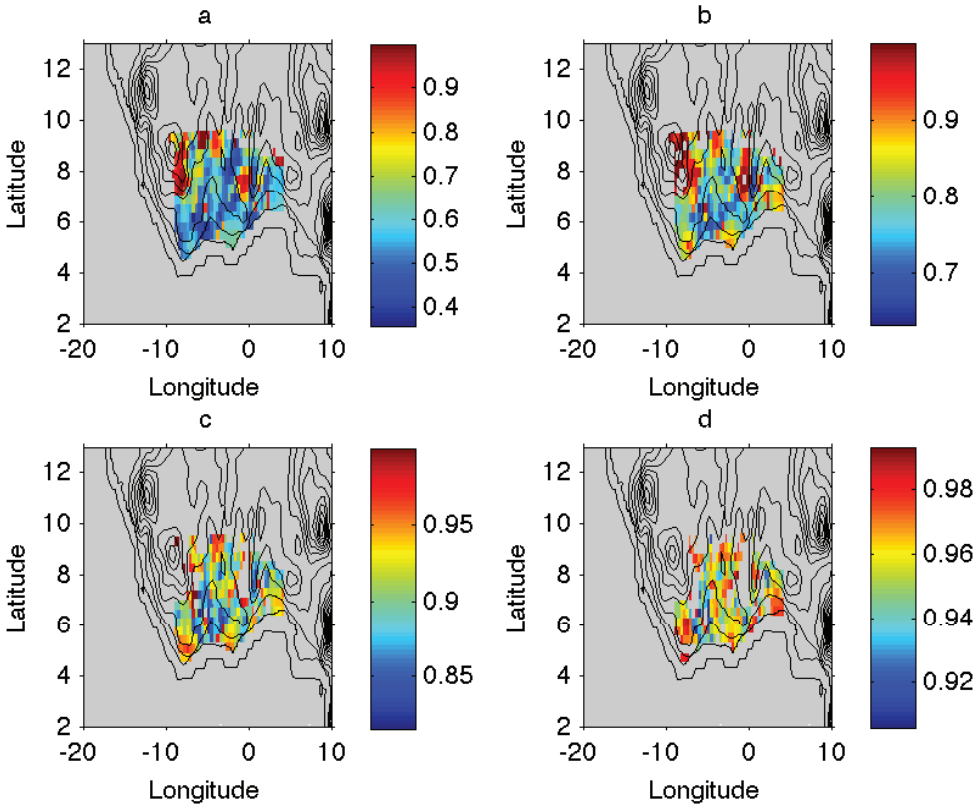

FIGURE 3: Percent cumulative variance for precipitation on the map explained by reconstructing the precipitation process using (a) one term (b) two terms (c) three terms and (d) four terms in the corresponding truncated *Karhunen-Loève* expansion.

In the following, if n_{SST} (*resp.* n_P) denotes the number of points on $\mathcal{G}$ (*resp.* $\mathcal{G}'$), we define the n_{SST}-dimensional (*resp.* the n_P-dimensional) vectors

$$\underline{\boldsymbol{\alpha}_m} = (\widetilde{\alpha}_m(x_1), \ldots, \widetilde{\alpha}_m(x_{n_{SST}}))^t, \quad m = 1, \ldots, M$$

and

$$\underline{\boldsymbol{\beta}_k} = \left(\widetilde{\beta}_k(y_1), \ldots, \widetilde{\beta}_k(y_{n_P})\right)^t, \quad k = 1, \ldots, K.$$

Note that in our application $n_{SST} = 516$ and $n_P = 368$.

3.2. Estimation procedure

We now describe our estimation procedure, following the main lines in Yao *et al.* (2005a). The methodology described below and used for our analysis has been implemented in Matlab and is freely available in the PACE (principal analysis by conditional expectation) package, downloadable from the internet (see Yao *et al.* (2010)).

We only deal here with SST since the procedure is the same for precipitation. Let x be any point on the spatial grid $\mathcal{G}$. Assume $x \in \mathcal{G}_l$ for some $l \in \{1, \ldots, L_1\}$. In a first step, we estimate the mean function $\mu_{X^x}(\cdot)$ based on the data from all individual curves. Mean and eigenfunctions are assumed to be smooth and we therefore use local linear smoothers (Fan and Gijbels, (1996)) for function and surface estimation, fitting local lines in one dimension and local planes in two dimensions by weighted least squares. The bandwidth b necessary for local smoothing is chosen by minimizing the cross-validation score given by $CV(b) = \sum_{i=1}^{N} \sum_{j=1}^{T} \{X_i^x(t_j) - \widehat{\mu}^{(-i)}(t_j; b)\}^2 / N$, where $t_1, \ldots, t_T$ is the time discretization of $\mathcal{T}$, N is the number of observed curves at x and $\widehat{\mu}^{(-i)}(t_j; b)$ is the estimation of $\mu_{X^x}(t_j)$ obtained without using the ith curve. To estimate the cross-covariance surface $G_X(s,t)$, $s, t \in \mathcal{T}$ we have used two-dimensional scatterplot smoothing. The raw cross-covariances $G_{X,i}(t_j, t_k) = \left(X_i^x(t_j) - \widehat{\mu}_{X^x}(t_j)\right)\left(X_i^x(t_k) - \widehat{\mu}_{X^x}(t_k)\right)$ are considered as input for the two-dimensional smoothing step. More precisely, the local linear surface smoother for the cross-covariance surface $G_X(s,t)$ is obtained as in Yao *et al.* (2005a)by minimizing :

$$\sum_{i=1}^{N} \sum_{1 \le j,k \le T} K_2\left(\frac{t_j - s}{h_1}, \frac{t_k - t}{h_2}\right) \left\{G_{X,i}(t_j, t_k) - f(\beta, (s,t), (t_j, t_k))\right\}^2$$

with respect to $\beta = (\beta_0, \beta_{1,1}, \beta_{1,2})$, leading to $\widehat{G}_X(s,t) = \widehat{\beta}_0(s,t)$, where $f(\beta,(s,t),(t_j,t_k)) = \beta_0+\beta_{1,1}(s-t_j)+\beta_{1,2}(t-t_k)$, K_2 is a given two-dimensional kernel, and where the bandwidths h_1 and h_2 are chosen again by cross-validation.

The estimates of eigenfunctions and eigenvalues correspond to the solutions $\widehat{e}_m(x_{0,l},\cdot)$ and $\widehat{\rho}_m$ of the following integral equations :

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

where the $\widehat{e}_m(x_{0,l},\cdot)$ are subject to $\int_{\mathcal{T}} \widehat{e}_m(x_{0,l},t)^2dt = 1$ and $\int_{\mathcal{T}} \widehat{e}_k(x_{0,l},t)\widehat{e}_m(x_{0,l},t)dt = 0$ for $m \neq k \leq M$. The eigenfunctions are estimated by discretizing the smoothed covariance, as described e.g. in Rice and Silverman (1991) or Capra and Müller (1997).

Finally, to complete the estimation procedure for SST, we have to estimate $\widetilde{\alpha}^i_m(x)$, for $i = 1,\ldots,N$ and $m = 1,\ldots,M$. We use the following projection estimates :

$$\sum_{j=2}^{T} X^x_i(t_j)\widehat{e}_m(x_{0,l},t_j)(t_j - t_{j-1}),$$

which are just numerical integration versions of $\widetilde{\alpha}^i_m(x) = \int_{\mathcal{T}} X^x_i(t)\widehat{e}_m(x_{0,l},t)dt$, for $m = 1,\ldots,M$. The estimation for each individual curve is needed in Section 4 for the selection procedure of the regression.

3.3. Functional clustering results

As mentioned previously, clustering algorithms are crucial in reducing the dimensionality of our data. The number and shape of the eigenfunctions patterns over time are not known. An ideal clustering method would provide a statistically significant set of clusters (and therefore of spatial regions) and curves derived from the data themselves without relying on a pre-specified number of clusters or set of known functional forms. Further, such a method should take into account the between time-point correlation inherent in time

series data. Some popular methods such as k-means clustering (see Hartigan and Wong (1978)), self-organizing maps (SOM) (see Kohonen (1997)) or hierarchical clustering (see Eisen *et al.* (1998)) do not satisfy this pre-requisite. One promising approach is to use a general multivariate Gaussian model to account for the correlation structure ; however, such a model ignores the time order of the eigenfunctions. The time factor is important in interpreting the clustering results of time series data. A curve-based clustering method called FCM was introduced in James and Sugar (2003) to cluster sparsely sampled time course genomic data. Similar approaches were proposed in Luan and Li (2003) to analyze time course gene expression data. In these methods, the mean gene expression profiles are modeled as linear combinations of spline bases. However, with different choices of bases or of the number of knots, one could get an array of quite different estimates of the underlying curves. Effective methods or guidance on how to select the basis or the number of knots are still lacking, which hinders the effective use of these methods in real applications. Here, we have used a data-driven clustering method, called smoothing spline clustering (SSC), that overcomes the aforementioned obstacles using a mixed-effect smoothing spline model and a rejection-controlled EM algorithm (see Ma *et al.* (2006)). A distinguishing feature of SSC is that it accurately estimates individual eigenvalue profiles and the mean eigenfunction profile within clusters simultaneously, making it extremely powerful for clustering time series data. Let us now present the way we fixed the number of clusters for our test case.

Sea surface temperatures :

We first performed the SSC clustering approach on the 516 estimated first eigenfunctions $t \to \widehat{e}_1(x, \cdot)$ obtained by the Karuhnen-Loève decomposition at each point x of the spatial grid $\mathcal{G}$. To determine a convenient number K of clusters several data- driven strategies can be defined, at least in the classical case. A first one amounts in inspecting basically the

within-cluster dissimilarity as a function of K. Many heuristics have been proposed trying to find a “kink" in the corresponding plot. A more formal argument has been proposed by Tibshirani *et al.* (2001) by comparing, using the gap statistic, the logarithm of the empirical within-cluster dissimilarity and the corresponding one for uniformly distributed data. An information theoretic point of view provided by Sugar and James (2003), considers the transformed distortion curve (K, d_K), where d_K denotes the minimum achievable distortion associated with fitting K centers to the data. The distortion d_K may be seen as a kind of average Mahalanobis distance between data and the set of cluster centers as a function of K. Jumps in the associated plot allow to select sensible values for K while the largest one may be the best choice for a mixture of multivariate distributions with common covariance. An asymptotic analysis (as the dimension goes to infinity) states that, when the number of clusters used is smaller than the true number, then the transformed distortion remains close to zero, before jumping suddenly and increasing linearly. This is the criteria we implemented to choose the number of clusters. We then plotted on Figure 4 below the jumps $d_K - d_{K-1}$ with respect to K.

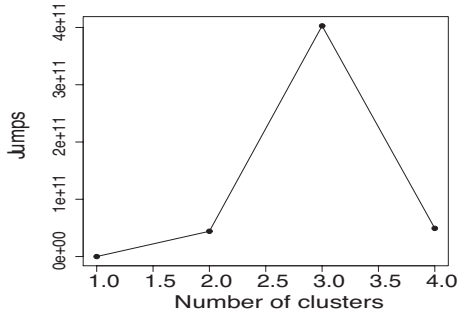

FIGURE 4: Jumps for the distortion measure $d_K - d_{K-1}$ with respect to the number K of clusters.

We see that $K = 3$ is a (at least local) maximum, hence we should fix $K = 3$. Projection

on the map for three (*resp.* two) clusters were drawn on the left panel (a) of Figure 5 (*resp.* right panel (b)).

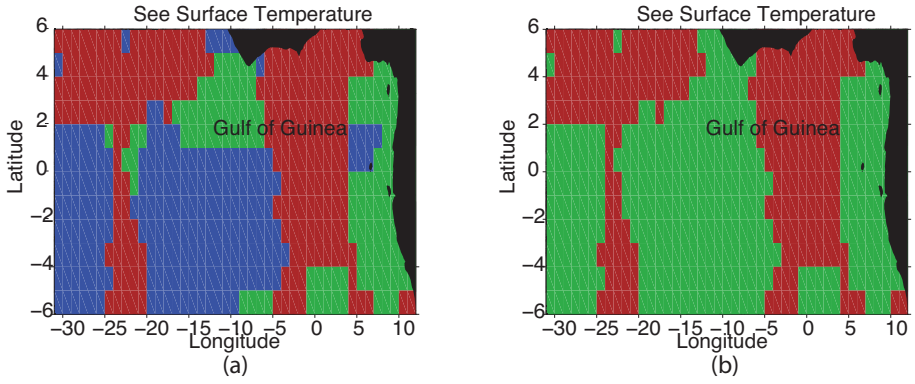

FIGURE 5: (a) Projection on the map for three clusters ; (b) for two clusters

Given the lack of observations, interpretation of the map with three clusters (see Figure 5 (a)) appeared difficult for physicists. On the map with two clusters (see Figure 5 (b)) we clearly see that a relevant factor is the distance from the coast. We thus prefer hereafter a choice of two clusters, which seems to be more robust.

The objective of the following figures is to see the overall trend of the eigenfunctions over time, uncovering spatial-specific variation patterns and extracting the dominant modes of variation. Considering two spatial clusters, in Figure 6 we collect for each cluster the estimated curves for the first eigenfunctions $t \to \widehat{e}_1(x, \cdot)$, $x \in \mathcal{G}$. The estimates of the mean functions in each cluster together with the upper and lower quartile curves in each cluster are given in Figure 7.

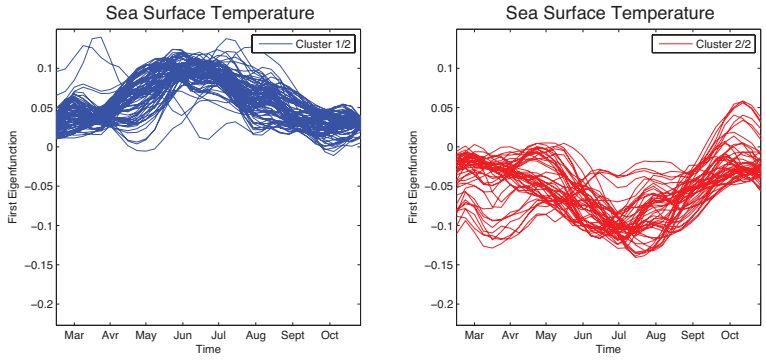

FIGURE 6: Estimated curves for the first eigenfunction by cluster

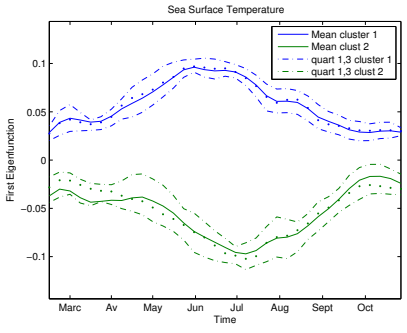

FIGURE 7: Mean curves on each cluster with their lower-upper quartile bands.

Let us now consider what happens for the second eigenfunctions $t \to e_2(x, \cdot)$, $x \in \mathcal{G}$. Figure 8 shows the estimated curves $t \to \widehat{e}_2(x, \cdot)$ on each of the two clusters obtained by applying the SSC procedure on the 18 curves $t \to \widehat{e}_1(x, \cdot)$.

The clustering structure seems also adapted for discriminating the second eigenfunctions, which supports the fact that a two clusters (denoted by $\mathcal{G}_l$, $l = 1, 2$) behavior of the

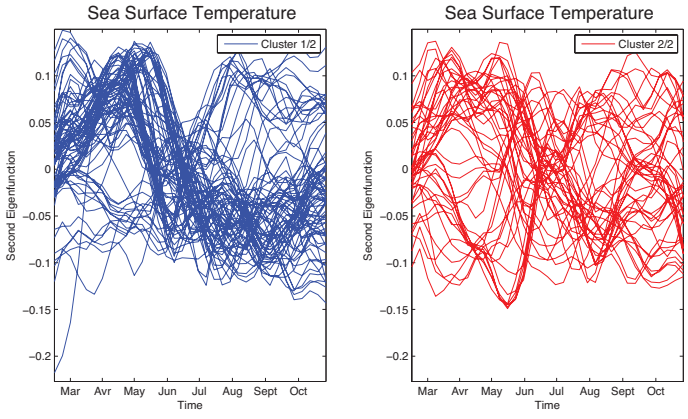

FIGURE 8: Estimated curves for the second eigenfunction by cluster

second eigenfunctions makes sense. It thus validates the assumption announced in Subsection 3.1 : there exist $x_{0,l} \in \mathcal{G}_l$, $l = 1, 2$ such that for any x on $\mathcal{G}_l$, we can approximate $X^x(t)$ by

$$\widetilde{X}^x(t) = \mu_{X^x}(t) + \sum_{m=1}^{2} \widetilde{\alpha}_m(x) e_m(x_{0,l}, t), \ t \in \mathcal{T},$$

with $\widetilde{\alpha}_m(x) = \int_{\mathcal{T}} \widetilde{X}^x(t) e_m(x_{0,l}, t) dt$, for $m = 1, 2$. It remains to choose the representative points $x_{0,1}$ and $x_{0,2}$ for each cluster. We considered the centroid for each cluster. These two points were not necessarily on the grid $\mathcal{G}$, thus for each cluster we chose the point on the grid which is the closest to the centroid.

Precipitation :

The procedure adopted for analyzing precipitation is similar. The number K of clusters was chosen by considering the plot $(K, d_K - d_{K-1})$ which attains a local maximum at $K = 3$ (see Figure 9). Projection on the map for three (*resp.* two) clusters were drawn on the left panel (a) of Figure 10 (*resp.* right panel (b)). Once more, the interpretation of the map

with three clusters (see Figure 10 (a)) appeared difficult for physicists. However on the map with two clusters (see Figure **??** (b)) we see that a relevant factor is the topography. We thus prefer hereafter a choice of two clusters, which seems to be more robust.

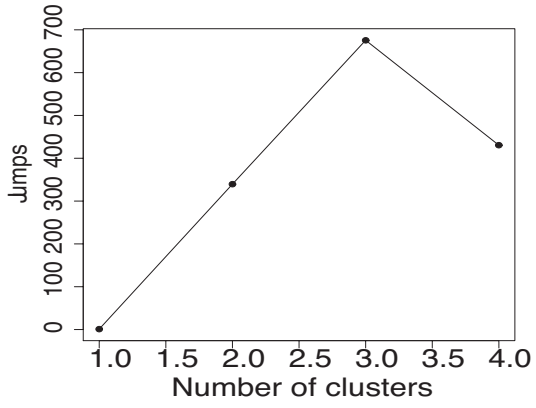

FIGURE 9: Jumps for the distortion measure $d_K - d_{K-1}$ with respect to the number K of clusters.

Considering two spatial clusters, in Figure 11 we collect for each cluster the estimated curves $t \to \widehat{f}_1(y,t)$, $y \in \mathcal{G}'$. The estimates of the mean functions in each cluster together with the upper and lower quartile curves in each cluster are given in Figure 12.

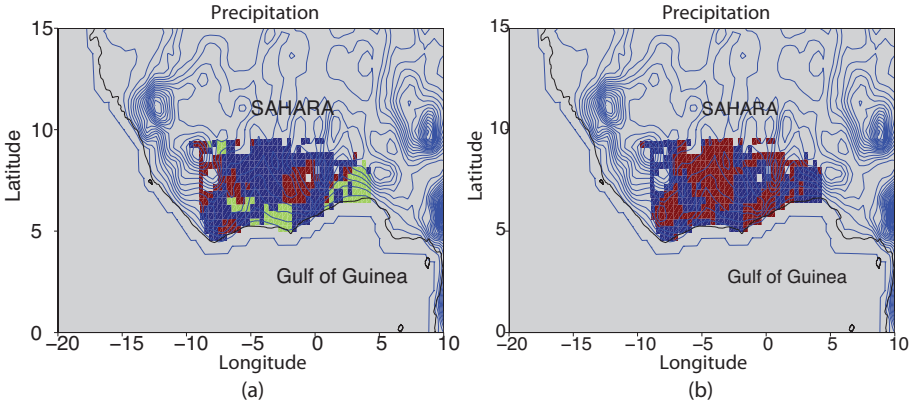

FIGURE 10: (a) Projection on the map for three clusters ; (b) for two clusters

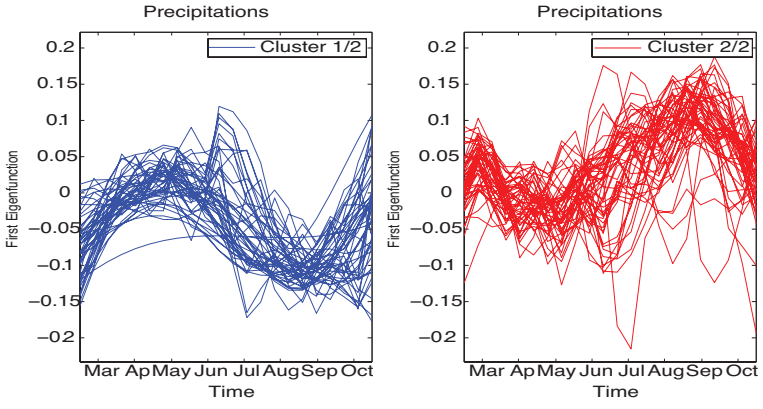

FIGURE 11: Precipitation data : estimated curves by cluster for the first eigenfunction.

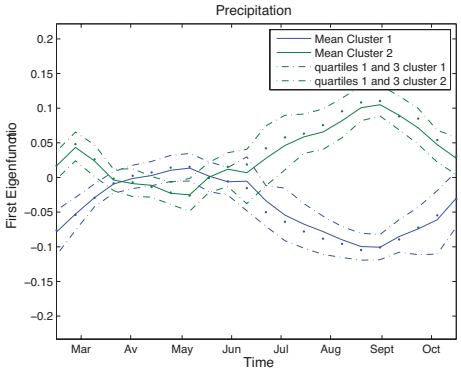

FIGURE 12: Mean curves on each cluster with their lower-upper quartile bands.

Let us now consider what happens for the second eigenfunctions $t \to f_2(y, \cdot)$, $y \in \mathcal{G}'$. Figure 13 shows the estimated curves $t \to \widehat{f}_2(y, \cdot)$ on each of the two clusters obtained by applying the SSC procedure on the 8 curves $t \to \widehat{f}_1(y, \cdot)$.

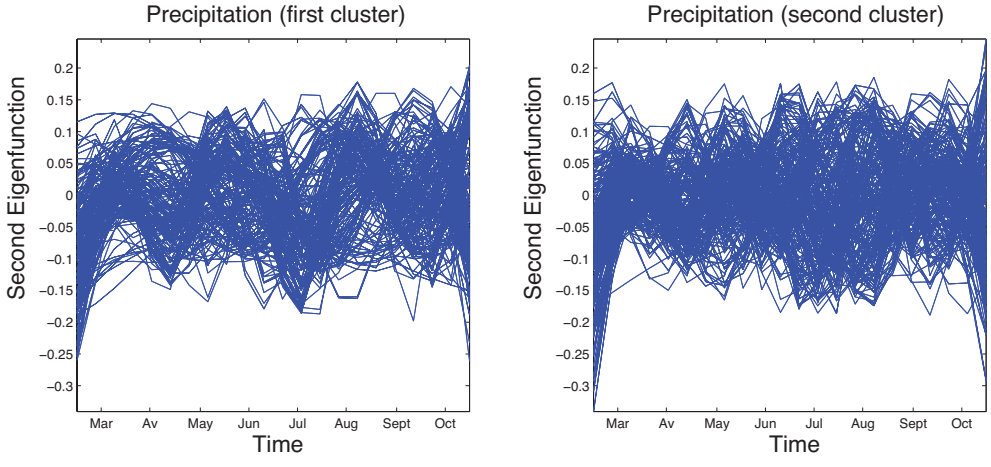

FIGURE 13: Precipitation data : estimated curves by cluster for the second eigenfunction.

No further clustering structure appears in the estimated curves which supports the fact that using the two clusters (denoted by $\mathcal{G}'_l$, $l = 1, 2$) obtained by SSC on the first eigenfunctions for discriminating the clusters makes sense. It thus validates again the assumption announced in Subsection 3.1 : there exist $y_{0,l} \in \mathcal{G}'_l$, $l = 1, 2$ such that for any $y \in \mathcal{G}'_l$, we can approximate $Y^y(t)$ by

$$\widetilde{Y}^y(t) = \mu_{Y^y}(t) + \sum_{k=1}^{2} \widetilde{\beta}_k(y) f_k(y_{0,l}, t),\ t \in \mathcal{T}$$

with for $k = 1, 2$, $\widetilde{\beta}_k(y) = \int_{\mathcal{T}} \widetilde{Y}^y(t) f_k(y_{0,l}, t) dt$. The difference is that, contrarily to what happens fro SST, we do not define $t \to f_2(y_{0,l}, t)$ as the second eigenfunction obtained at point $y_{0,l}$ but as the mean curve $t \to f_2(t)$ of all curves $t \to f_2(y, t)$, $y \in \mathcal{G}'$. Thus it does not depend on space. It remains to choose the representative points $y_{0,1}$ and $y_{0,2}$ for each cluster. We considered the centroid for each cluster. These two points are not necessarily on the grid $\mathcal{G}'$, thus for each cluster we chose the point on the grid that is the closest to the corresponding centroid.

Hence for both precipitation and sea surface temperatures, we obtain a decomposition where the basis functions are functions depending on time only and whose coefficients are spatially indexed and time independent. The relative mean squared error (MSE^{rel}) for the reconstruction of sea surface temperatures and precipitation is estimated by leave-one-out cross-validation (see eq. (2) below for the definition of MSE^{rel}). The panels in Figure 14 below display this relative mean squared error for SST reconstruction (left) and precipitation (right) on the appropriate map.

Leave-one-out cross-validation relative mean squared error estimation for SST at each point $x \in \mathcal{G}$ is defined by

$$MSE^{rel,SST}(x) = \frac{1}{22} \sum_{j=1}^{T} \frac{\frac{1}{18} \sum_{k=1}^{18} \left(\widetilde{X}^{(-k),x}(t_j) - X_k^x(t_j) \right)^2}{\frac{1}{18} \sum_{k=1}^{18} \left(X_k^x(t_j) \right)^2}, \tag{2}$$

where $\widetilde{X}^{(-k),x}(t_j)$ is the estimation of $X^x(t_j)$ obtained without using the kth curve $X_k^x(\cdot)$. The procedure for the estimation of the relative mean squared error for precipitation is similar.

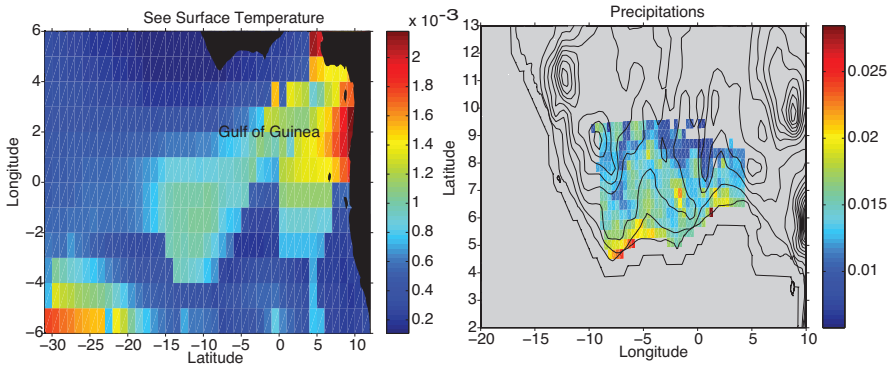

FIGURE 14: (Relative Mean Squared Error for the reconstruction of SST (left) and of Precipitation (right)

4. Multivariate regression model, a double penalized approach

This section concerns the regression approach we have adopted for prediction (see Subsection 4.1). We also discuss in this Section the selection procedure of the tuning parameters for our application (see Subsection 4.2).

4.1. Regression procedure

As mentioned in the introduction, we intend to use a novel method recently developed by Peng *et al.* (2010) in integrated genomic studies which we describe below for the sake of completeness. The method uses an ℓ_1-norm penalty on a least squares procedure to control the overall sparsity of the coefficient matrix in a multivariate linear regression model. In addition, it also imposes a *group* sparse penalty, which in essence is the same as the *group*

lasso penalty proposed by Bakin (1999), Antoniadis and Fan (2001) and Obozinski *et al.* (2008). This penalty puts a constraint on the ℓ_2 norm of regression coefficients for each predictor, which thus controls the total number of predictors entering the model, and consequently facilitates the detection of important predictors.

More precisely, consider a multivariate regression problem with q response variables $Y_1, \ldots, Y_q$ and p prediction variables $X_1, \ldots, X_p$:

$$Y_j = \sum_{i=1}^{p} X_i B_{ij} + \epsilon_j, \quad j = 1, \ldots, q, \tag{3}$$

where the error terms $\epsilon_1, \ldots, \epsilon_q$ have a joint distribution with mean 0 and covariance Σ. In the above, we assume without any loss of generality that all the response and prediction variables are standardized to have zero mean and thus there is no intercept term in equation (3). Our primary goal is to identify non-zero entries in the $p \times q$ regression coefficient matrix $B = (B_{ij})$ based on n i.i.d. samples from the above model which is exactly the problem addressed by Peng *et al.* (2010). Under normality assumptions, B_{ij} can be interpreted as proportional to the conditional correlation $\mathrm{Cor}(Y_j, X_i | X_{-(i)})$, where $X_{-(i)} := \{X_i' : 1 \leq i' \neq i \leq p\}$. In the following, we use $\mathbf{Y}_j = (Y_j^1, Y_j^2, \ldots, Y_j^n)^T$ and $\mathbf{X}_i = (X_i^1, X_i^2 \ldots, X_i^n)^T$ to denote respectively the sample of the jth response variable and that of the ith prediction variable. We also use $\mathbf{Y} = (\mathbf{Y}_1 : \cdots : \mathbf{Y}_q)$ to denote the $n \times q$ response matrix, and use $\mathbf{X} = (\mathbf{X}_1 : \cdots : \mathbf{X}_p)$ for the $n \times p$ prediction matrix. We shall focus on the cases where both q and p are larger than the sample size n. For example, in the applied study of West African Monsoon discussed later, we regress $\left(\underline{\boldsymbol{\alpha}_1}, \underline{\boldsymbol{\alpha}_2}\right)$ on $\left(\underline{\boldsymbol{\beta}_1}, \underline{\boldsymbol{\beta}_2}\right)$.Hence, for this application, the sample size is 8, while the number of spatial components are respectively $p = 2 \times n_{SST}$ and $q = 2 \times n_P$. In the application $n_{SST} = 516$ and $n_P = 368$. When $q > n$, whatever the value of p is, the ordinary least square (OLS) solution is not unique, and regularization becomes indispensable. The choice of suitable regularization depends heavily on the type of data structure we envision. Recently, ℓ_1-norm based sparsity constraints such as lasso (Tibshirani

(1996)) have been widely used under such high-dimension-low-sample-size settings. In our application, we will impose an ℓ_1-norm penalty on the coefficient matrix B to control the overall sparsity of the multivariate regression model but in addition, we put constraints on the total number of predictors entering the model which is essentially the `remMap` idea. This is achieved by treating the coefficients corresponding to the same predictor (one row of B in model (3) as a group, and then penalizing their ℓ_2 norm. A predictor will not be selected into the model if the corresponding ℓ_2-norm is shrunken to 0. Thus this penalty facilitates the identification of master predictors which affect (relatively) many response variables. Specifically, for model (3), we wiil use the following criterion

$$\ell_{(\lambda,\mu)}(\mathbf{Y}, B) = \frac{1}{2}\|\mathbf{Y} - \mathbf{X}B\|_F^2 + \lambda \sum_{j=1}^{p} \|\mathbf{C}_j \cdot B_j\|_1 + \mu \sum_{j=1}^{p} \|\mathbf{C}_j \cdot B_j\|_2, \tag{4}$$

where $\mathbf{C}$ is a $p \times q$ 0-1 matrix indicating the coefficients of B on which penalization is imposed. In the above equation $\mathbf{C}_j$ and B_j are the jth rows of $\mathbf{C}$ and B respectively, while $\|\cdot\|_F$ denotes the Frobenius norm of matrices, $\|\cdot\|_1$ and $\|\cdot\|_2$ are respectively the ℓ_1 and ℓ_2 norms of vectors and "$\cdot$" stands for the Hadamard product (entry-wise multiplication). The selection matrix $\mathbf{C}$ is pre-specified based on prior knowledge : if we know in advance that predictor X_i affects response Y_j , then the corresponding regression coefficient B_{ij} will not be penalized and we set $C_{ij} = 0$. When there is no such prior information, $\mathbf{C}$ can be simply set to a constant matrix $C_{ij} = 1$. Finally, an estimate of the coefficient matrix B is $\widehat{B}_{\lambda,\mu} := \text{argmin}_B \ell_{(\lambda,\mu)}(\mathbf{Y}, B)$.

In the above criterion function, the ℓ_1 penalty induces the overall sparsity of the coefficient matrix B. The ℓ_2 penalty on the row vectors $\mathbf{C}_j \cdot B_j$ induces row sparsity of the product matrix $\mathbf{C} \cdot B$. As a result, some rows are shrunken to be entirely zero. Consequently, predictors which affect relatively more response variables are more likely to be selected into the model. We will refer to the proposed estimator $\widehat{B}_{\lambda,\mu}$ as the remMap (REgularized Mul-

tivariate regression for identifying MAster Predictors) estimator in connexion with the `remMap` theory and R-package developed by Peng *et al.* (2010) for regularized multivariate Regression for identifying master predictors in integrative genomics studies of breast cancer.

4.2. Implementation and results

In this subsection, we describe the different steps for the implementation of the remMAP procedure on our application. A first step is to fit both parameters λ and μ. These parameters are adjusted by v-fold cross-validation. The prediction error obtained by 4-fold cross-validation is drawn on Figure 15 below. We note that there does not exist a unique minimum. For $\lambda = 1$ and $\mu = 4$, the error seems to reach a value close to the minimum.

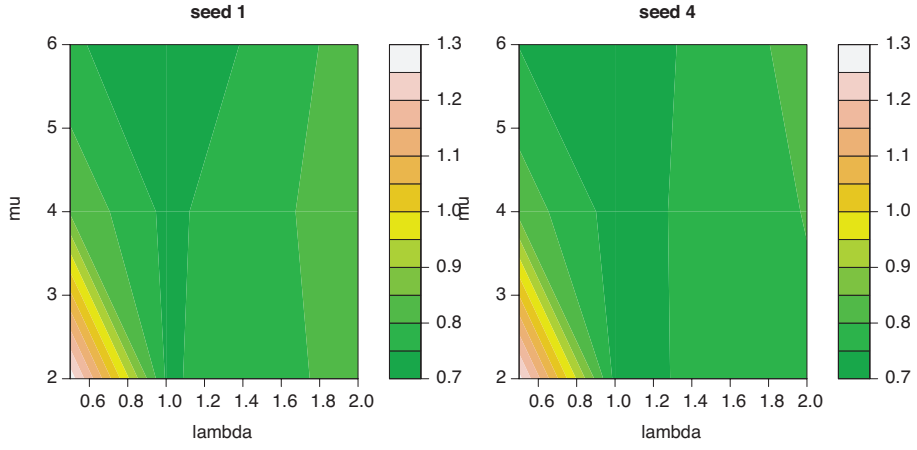

FIGURE 15: Cross validation for the choice of λ and μ (with two different germs)

On Figure 16 we note that the regression coefficients matrix B, estimated using $\lambda = 1$ and $\mu = 4$ for the penalties, is sparse. This is a consequence of using the remMAP methodology.

FIGURE 16: Regression coefficients matrix B estimated with $\lambda = 1$ and $\mu = 4$

It seems quite interesting to display on a map the spatial points on the grid $\mathcal{G}$ corresponding to the nonzero rows of the matrix B (left) and the spatial points on $\mathcal{G}'$ influenced by the nonzero rows of B (right) (see Figure 17 below). As one may see, the two regions seem complementary and cover quite well the region of interest for precipitation.

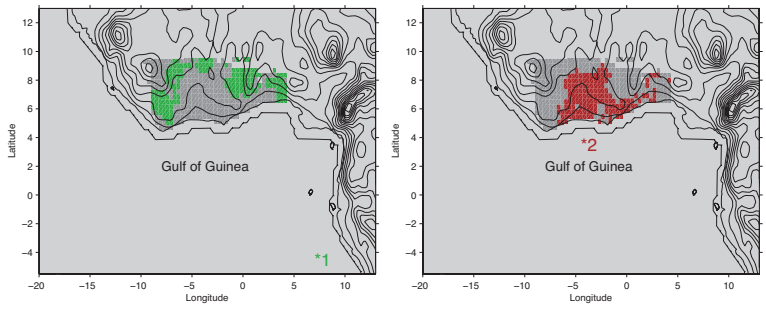

FIGURE 17: Spatial location for the average responses indicated by the retained coefficients for both predictors (points 1 and 2 on the map).

Using the results of the regression and given the retained regression coefficients we the proceed to the reconstruction of Precipitation on the grid $\mathcal{G}'$. Define first

$$\left(\underline{\boldsymbol{\beta}_{1,reg}}, \underline{\boldsymbol{\beta}_{2,reg}}\right)^T = \left(\underline{\boldsymbol{\alpha}_1}, \underline{\boldsymbol{\alpha}_2}\right) \widehat{B}\,.$$

Then for $l = 1, 2$, for $y \in \mathcal{G}'_l$, let

$$Y^{y,reg}(t) = \widehat{\mu}_{Y^y}(t) + \underline{\boldsymbol{\beta}_{1,reg}}\, \widehat{f}_1(y_{0,l}, t) + \underline{\boldsymbol{\beta}_{1,reg}}\, \widehat{f}_2(t)\,.$$

The relative mean squared error (RMSE) estimated by leave-one-out cross-validation (see (5) below) is displayed on the map (see Figure 18). Notice however some points of high RMSE which are close to the coast. We have also plot the annual and weekly boxplots for the relative MSE (see Figure 19). The relative error is between 0.3 and 0.4 which is not so bad if we consider we did not have many observations to conduct the study. As one can see on the right panel of the figure this error is not constant over time, with bad reconstructions for some weeks.

$$MSE^{Precip,reg}(y) = \frac{1}{22}\sum_{j=1}^{22} \frac{\frac{1}{8}\sum_{k=1}^{8}\left(Y^{(-k),y,reg}(t_j) - \widehat{\widetilde{Y}_k^y}(t_j)\right)^2}{\frac{1}{8}\sum_{k=1}^{8}\left(\widehat{\widetilde{Y}_k^y}(t_j)\right)^2}, \tag{5}$$

with $\widehat{\widetilde{Y}_k^y}(t_j) = \widehat{\mu}_{Y^y}(t_j) + \widehat{\widetilde{\beta}_1^k}(y)\widehat{f}_1(y_{0,l}, t_j) + \widehat{\widetilde{\beta}_2^k}(y)\widehat{f}_2(t_j)$.

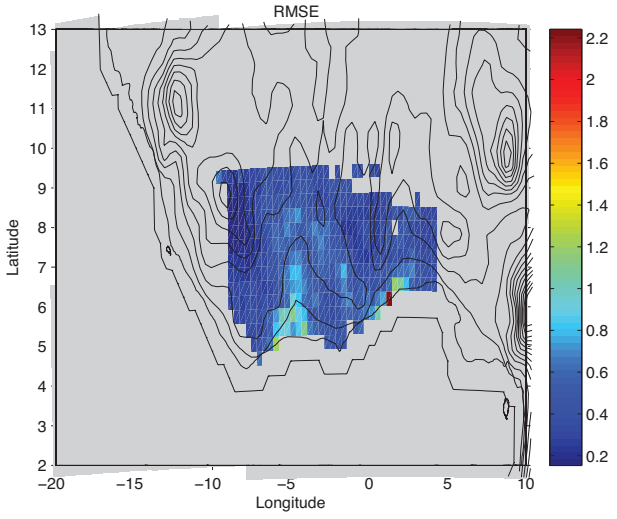

FIGURE 18: Relative MSE for the reconstructed precipitation by regression on the map.

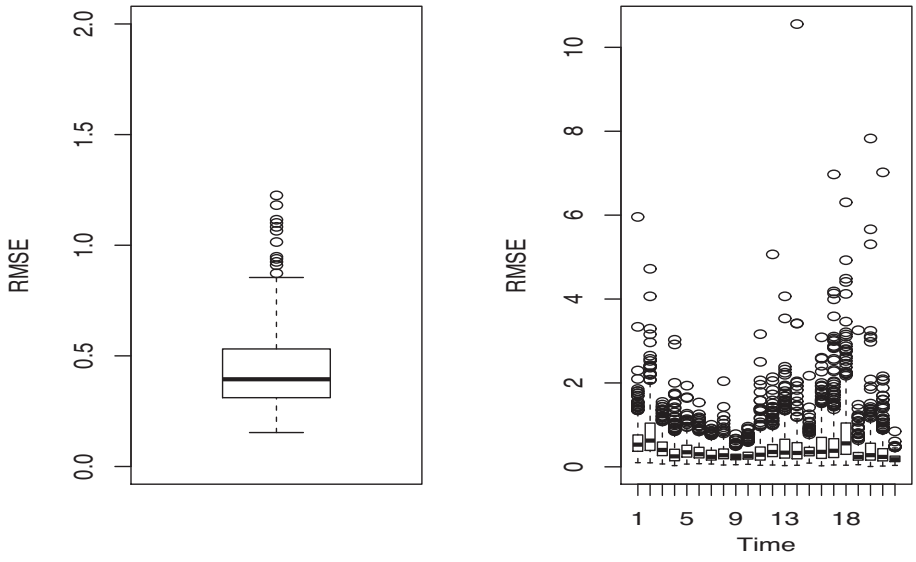

FIGURE 19: Boxplots of the relative MSE per year (left) and per week (right).

Finally, on Figure 20 below we plotted for some fixed points in $\mathcal{G}'$ the curve reconstructed by regression (continuous line) for precipitation, the one obtained by the filtering modeling of Section 3 (circles) and the observations themselves (dots). As one can see the regression prediction curve somehow smooths the observations in quite a natural way and the methodology seems promising for pursuing via this model a sensitivity analysis, but this is beyond the scope of the present work.

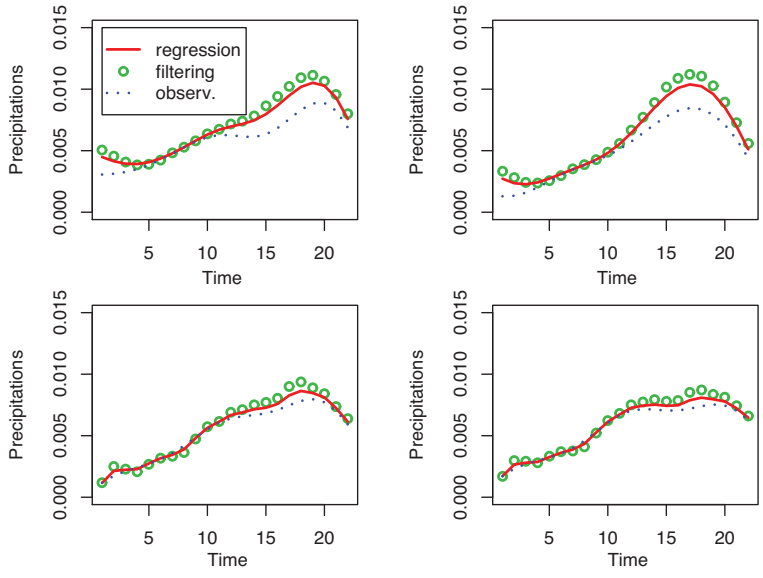

FIGURE 20: For 4 spatial points selected in the domain $\mathcal{G}'$ a display of the reconstructed precipitation curve (red), the reconstruction curve with truncated Karhunen-Loève decomposition (circles) and the observed precipitation (dots).

5. Conclusion and perspectives

Motivated in investigating the West African Monsoon, we present a new approach for modeling and fitting high-dimensional response regression models in the setting of complex spatio-temporal dynamics. We were particularly interested in developing an appropriate regression based methodology for studying the influence of sea surface temperatures in the Gulf of Guinea on precipitation in Saharan and sub-Saharan. However, one central issue in the analysis of such data consists in taking into account the spatio-temporal dependence of the observations. For most of the applications that we are aware of the spatio-temporal dynamics are usually modeled as time function-valued (spatially stationary) processes al-

lowing the development of efficient prediction procedures based on appropriate principal component like decompositions and regression. In practice, however, many observed spatial functional time series cannot be modeled accurately as stationary. To handle spatial variation in a natural way we have segmented the space into regions of similar spatial behavior using in the process an efficient clustering technique that clusters the times series into groups that may be considered as stationary so that in each group more or less standard regression prediction procedures can be applied. Furthermore to avoid regression models that are far too complex for prediction, and inspired by similar approaches used in modern genomic data analysis we have used an appropriate regularization method that has proven to be quite efficient for the data we have analyzed. However, a major lack in this study is that it was implemented with only 8 years of observations. As explained in the introduction, our developments are to be considered as a first step and must be tested on larger data sets. Concerning our application, we indeed wait for computing tools under development to run the MAR on years 1983 to 2000. An idea for the future is also to find a relevant way to perturb initial maps of SST and then to run MAR on these new inputs. From this new sample, of greater size, we should obtain finer results for fitting our model. Our model will be used further to perform sensitivity analysis which is a major issue in the study of West African monsoon.

Acknowledgments

This work has been partially supported by French National Research Agency (ANR) through COSINUS program (project COSTA-BRAVA n° ANR-09-COSI-015) and the IAP Research Network P6/03 of the Belgian State (Belgian Science Policy).

References

[1] Antoniadis, A. and Fan, J. (2001). Regularization of wavelets approximations (with discussion). *J. Amer. Statist. Assoc.*, **96**(455), 939–963.

[2] Bakin, S. (1999). *Adaptive regression and model selection in data mining probems*. Ph.D. thesis, Australian National University, Cambera, Australia.

[3] Capra, W. B. and Müller, H. G. (1997). An accelerated-time model for response curves. *Journal of the American Statistical Association*, (92), 72–83.

[4] Caron, E., Chouhan, P. K., and Dail, H. (2006). Godiet : A deployment tool for distributed middleware on grid'5000. In *EXPGRID workshop. Experimental Grid Testbeds for the Assessment of Large-Scale Distributed Apllications and Tools.*, Paris. France. HPDC-15, IEEE.

[5] Cohen, A. M. and Jones, D. E. (1977). A technique for the solution of eigenvalue problems. *J. Inst. Math. Appl.*, **20**(1), 1–7.

[6] Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, **95**, 14863–14868.

[7] Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Application*. Chapman and Hall, London.

[8] Hartigan, J. A. and Wong, M. A. (1978). A k-means clustering algorithm. *App. Statist.*, **28**, 100–108.

[9] James, G. and Sugar, C. (2003). Clustering for sparsely sampled functional data. *Journal of the American Statistical Association*, **98**, 397–408.

[10] Kohonen, T. (1997). *Self-Organizing Maps*. Springer, New York.

[11] Luan, Y. and Li, H. (2003). Clustering of time-course gene expression data using a mixed-effects model with b-spline. *Bioinformatics*, **19**, 474–482.

[12] 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.

[13] Messager, C. (2005). *Couplage des composantes continentale et atmosphérique du cycle de l'eau aux échelles régionale et climatique. Application à l'Afrique de l'Ouest.* Ph.D. thesis, University Joseph Fourier, Grenoble, France.

[14] 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.

[15] Obozinski, G., Wainwright, M. J., and Jordan, M. I. (2008). Union support recovery in high-dimensional multivariate regression. Technical Report (to appear in Annals of Statistics) 761, Dept. of Statistics. University of California at Berkeley.

[16] 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. *Annals of Applied Statistics*, **4**(1), 53–77.

[17] Ramsay, J. O. and Silverman, B. W. (2002). *Applied Functional Data Analysis : Methods and Case Studies*. Springer-Verlag.

[18] Ramsay, J. O. and Silverman, B. W. (2005). *Functional data analysis*. Springer-Verlag, second edition edition.

[19] Reynolds, R. W. and Smith, M. T. (1994). Improved global sea surface temperature analysis using optimal interpolation. *Journal of Climate*, **7**, 929–948.

[20] Rice, J. A. and Silverman, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *J. Roy. Statist. Soc. Ser. B*, **53**(1), 233–243.

[21] Saltelli, A., Chan, K. P. S., and Scott, E. M. (2000). *Sensitivity Analysis*. John Wiley & Sons, New York.

[22] Sugar, C. A. and James, G. M. (2003). Finding the number of clusters in a dataset : an information-theoretic approach. *Journal of the American Statistical Association*, **98**(463), 750–763.

[23] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, **58**, 267–288.

[24] Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *J. Roy. Statist. Soc. Ser. B*, **63**(2), 411–423.

[25] Yao, F., Müller, H.-G., and Wang, J.-L. (2005a). Functional data analysis for sparse longitudinal data. *J. Amer. Statist. Assoc.*, **100**(470), 577–590.

[26] Yao, F., Müller, H.-G., and Wang, J.-L. (2005b). Functional linear regression analysis for longitudinal data. *Ann. Statist.*, **33**(6), 2873–2903.

[27] Yao, F., Liu, B., Müller, H.-G., and Wang, J.-L. (2010). *PACE 2.7*. University of California at Davis, http ://anson.ucdavis.edu/-mueller/data/programs.html.