Environmental and Ecological Statistics manuscript No.
(will be inserted by the editor)

Functional regression on remote sensing data in Oceanography

Nihan Acar-Denizli · Pedro Delicado · Gülay Başarır and Isabel Caballero

Received: date / Accepted: date

Abstract The aim of this study is to propose the use of a Functional Data Analysis (FDA) approach as an alternative to the classical statistical methods most commonly used in oceanography and water quality management. In particular we consider the prediction of Total Suspended Solids (TSS) based on Remote Sensing (RS) data. For this purpose several Functional Linear Regression Models (FLRMs) and classical non-functional regression models are applied to 10 years of RS data obtained from MEdium Resolution Imaging Spectrometer (MERIS) sensor to predict the TSS concentration in the coastal zone of the Guadalquivir estuary. The results of functional and classical approaches are compared in terms of their Mean Square Prediction Error (MSPE) values and the superiority of the functional models is established. A simulation

This investigation is partially supported by the Spanish Ministerio de Ciencia e Innovación and Fondo Europeo de Desarrollo Regional grant MTM2013-43992-R and by the project 2014-24 of Mimar Sinan Fine Arts University Coordinatorship of Scientific Research Projects.

N. Acar-Denizli
Department of Statistics, Mimar Sinan Güzel Sanatlar Üniversitesi, Istanbul, Turkey
Tel.: +90 535 4282031
E-mail: nihan.acar@msgsu.edu.tr

P. Delicado
Department of Statistics and Operational Research, Universitat Politècnica de Catalunya, Barcelona, Spain

G. Başarır
Department of Statistics, Mimar Sinan Güzel Sanatlar Üniversitesi, Istanbul, Turkey

I. Caballero
National Centers for Coastal Ocean Science, NOAA National Ocean Service, Silver Spring, USA

study has been designed in order to support these findings and to determine the best prediction model for the TSS parameter in more general contexts.

Keywords Functional linear regression models · Functional principal components · Functional partial least squares · Exponential regression models · Remote sensing data

1 Introduction

Satellite images recorded by remote sensors are a way to obtain information about the earth. They are widely used in environmental sciences to map land-use and to analyze crop production and water quality (Caballero et al, 2014b,a; Faivre and Fischer, 1997; Nezlin and DiGiacomo, 2005; Gitelson et al, 2015; Rawat and Kumar, 2015). Specifically, in oceanography and water quality management measuring ocean characteristics is a difficult and expensive task. It usually involves complex logistics and leads to expensive data measures. However, by using Remote Sensing (RS) data the future values of the relevant ocean characteristics such as Sea Surface Temperature (SST), Chlorophyll-a content (Chl-a) and Total Suspended Solids (TSS) can be predicted quickly and economically (Bernardello et al, 2016; Caballero et al, 2014a,b; Chen et al, 2015; Le et al, 2013).

Remote sensing data are composed of remote sensing reflectance (Rrs) values recorded at different wavelengths of a spectrum. The prediction models for ocean characteristics from RS data usually have low sample sizes (n, the number of in-situ observations) because of the difficulties in observing in-situ data and the moderate number of highly correlated predictors (p, the number of recorded wavelengths). In previous studies, the logarithm of TSS have been modeled mainly by linear regression on a single wavelength or a combination of different wavelengths as predictors (Caballero et al, 2014a,b; Nechad et al, 2010; Nezlin and DiGiacomo, 2005). Polynomial regression and regression based on dimension reduction techniques such as Principal Component Analysis (known as Empirical Orthogonal Functions, EOF, in environmental statistics) have also been used (Binding et al, 2003; Everson et al, 1997). However, collinearity and variable selection are some of the problems that may arise when dealing with classical regression models. Moreover, Hyperspectral Remote Sensing (HRS) data, which are measured on a large spectrum with high number of wavelengths, have recently been in use. This would imply regression models in which the number of predictors is larger than the number of observations ("large p small n" problem; see Hastie et al 2015, for instance). All the foregoing emphasizes the need for novel approaches instead of classical methods. One of these approaches is Functional Data Analysis (FDA), which enables us to work with continuous functions as predictor variables. General references for FDA are Ramsay and Silverman (2005), Horváth and Kokoszka (2012) and Kokoszka and Reimherr (2017). See also the recent review in Wang et al (2016).

Given the continuous structure of the spectrum, RS data can be treated as functional data measured over a range of wavelengths. Thus, for each pixel of the satellite image a function of Rrs values in the relevant spectrum is observed. This way TSS concentration can be predicted on Rrs functions by using Functional Linear Models (FLRMs). FLRMs are used to analyze linear relationships between variables when at least one of the related variables has a functional structure. FLRMs have a wide range of use in various fields such as chemometry, biomechanics and environmental sciences. Particularly, there are many studies that focus on modelling scalar responses on functional predictors (Aguilera et al, 2010; James, 2002; Cardot et al, 1999). A recent review on FLRMs for scalar responses can be found in Reiss et al (2017). For a more general review on functional regression models, see Morris (2015) or the general reference on FDA cited before.

FDA techniques have recently gained importance in the analysis of RS data. Cardot et al (2003) and Besse et al (2005) used Functional Linear Regression Models (FLRMs) to predict land use from remote sensing data obtained from the Vegetation sensor of the SPOT4 Satellite. Liu et al (2012) put forward a new rotation approach for functional factor analysis with an application on periodic remote sensing data. In oceanography, Gong et al (2015) used Functional Principal Components Analysis (FPCA) to model high-dimensional temperature curves and temperature surfaces of Lake Victoria. Nevertheless, there are still only a few studies that use the FDA approach to remote sensing satellite data in the field of oceanography (Lahet et al, 2001; Gong et al, 2015), even though there are many applications of multivariate analysis techniques in this field (Clarke et al, 2006; Caballero et al, 2014a).

It is worth mentioning the pioneering study by Lahet et al (2001) in which a functional data set is built representing reflectance as a function of wavelength. The authors use satellite data (observed reflectance at 11 wavelengths) and a classical spectrometric model to obtain smooth curves defined in the continuous range of the spectrum. Although Lahet et al (2001) do not at any point relate their work with the field of FDA, they propose a clustering method based on the maximum of these functional data.

The goal of our study is to propose the use of Functional Linear Regression Models for scalar responses to predict surface water characteristics based on reflectance functions. Specifically, we aim to use FLRMs as an alternative to classical regression models in order to predict the TSS concentration in the coastal zone of the Guadalquivir estuary on 10 years of RS data obtained from the MERIS sensor. MERIS is one of 10 sensors that in March, 2002, was deployed by the European Space Agency (ESA) on board the polar-orbiting Envisat-1 environmental research satellite.

The Guadalquivir estuary is of great significance from the ecologic, social and economic points of view (Ruiz et al, 2014). The river is the main source of freshwater inputs and nutrients to the estuary and the adjacent Gulf of Cádiz shelf, thereby regulating the high biological productivity of the basin (Navarro and Ruiz, 2006). The turbidity plume variability is important for the functioning of the estuary and the adjacent coastal region, where high turbid

levels frequently produce negative effects on water quality and clarity, which are the major determinants of the condition and productivity of an aquatic system as well as, the tractability of water for human consumption, recreation and manufacturing (Ruiz et al, 2014). Several prolonged turbid episodes originated hypoxia and inhibition of phytoplankton growth in the Guadalquivir coastal region (Navarro et al, 2012; Caballero et al, 2014b), generating alert in the policy-managers and stakeholders. Examination of the turbidity and suspended solid estimation is therefore required in order to assist to the challenging coastal management and water quality monitoring.

In the following section the structure of the in-situ and satellite data sets used in the analysis are described and detailed information on the matching process is provided. In Section 2.3, we explain the theoretical background of Functional Linear Regression Models that are used later to predict TSS concentration in the Region of Interest (ROI). The results of classical and functional approaches are summarized in Section 3. A simulation study designed to support findings and to compare the predictive performance of the models is set out in Section 4. The last Section summarizes our conclusions (based on both the real data application and the simulation study) about what the best prediction models are for tackling the problem. All the statistical computations and graphics in the paper has been done using the software R (R Core Team, 2017). Satellite images in Figures 2 and 8 has been created by Matlab (MATLAB, 2011).

2 Materials and Methods

2.1 Study Area and data sets

The data set is composed of two parts: the in-situ measurements of TSS concentrations collected from the sea surface and the satellite data recorded by MERIS between the years 2002 and 2011. The Region of Interest is determined by the coordinates $36° - 37°$ N latitude and $6° - 7°$ W longitude (see Figure 1).

2.1.1 In-Situ Data

The in-situ data consists of samples collected from the coastal region of the Gulf of Cádiz on the southwest coast of the Iberian Peninsula at the *Junta de Andalucía* station and by the cruises of *Reserva* and *Fluctuaciones* over different time periods. The sampling carried out by the *Junta de Andalucía* covers the period between April 2008 - May 2011, where the samples of *Reserva* and *Fluctuaciones* were collected between the periods July 2002 - September 2004 and May 2005 - May 2007, respectively. The coordinates of the *Junta de Andalucía* station was fixed at lattitude 36.78° N and longitude 6.37° W, where the coordinates of the *Reserva* and *Fluctuaciones* stations in the Region of Interest (ROI) were chosen according to the campaign planning (see Figure

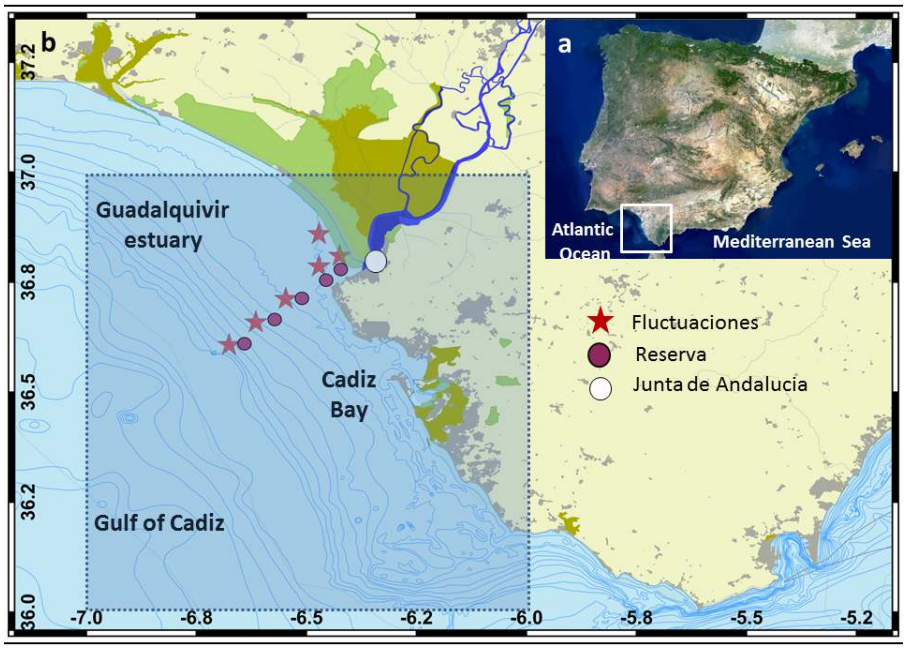
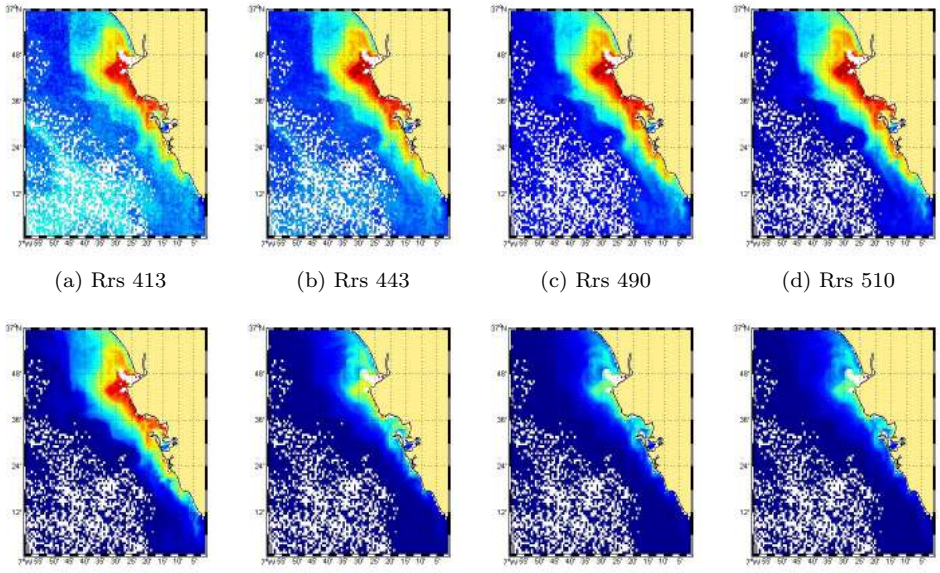


Fig. 1 The study area and ROI.
a) The study area. b) Map of the Guadalquivir estuary and the Gulf of Cádiz coastal area showing the ROI. Pink stars and circles indicate the *Fluctuaciones* and *Reserva* stations, respectively. The round white circle denotes the *Junta de Andalucía* station.

1). The surface samples taken for analysis were collected with a rosette sampler (5 m below the water surface) at a distance of between 1km and 25 km from the coast. The amount of TSS concentration in each sample is measured according to the protocols given in Caballero et al (2014a).

2.1.2 Satellite Data

The MERIS overpass time for central Europe is between 9:30 and 11:00 UTC, with a global coverage every 3 days. The MERIS sensor typically provides high coverage of the Gulf of Cádiz study area (overpass at GMT of approximately 10:30 am); however, cloudy and/or foggy conditions result in patchy spatial coverage and temporal gaps. The satellite data included within the Region Of Interest (ROI) was downloaded from the Ocean Colour Website (http://oceancolor.gsfc.nasa.gov) in hierarchical data format (hdf). SeaDAS image analysis software (SeaWifs Data Analysis System, version 6, http://seadas.gsfc.nasa.gov/) was used to convert data from hdf format to ascii format. The RS data set consists of Level-2 Remote Sensing Reflectance (Rrs) (sr^{-1}) values recorded at eight different wavelengths (413 nm, 443 nm, 490 nm, 510 nm, 560 nm, 620 nm, 665 nm, 681 nm) with 300 m full spatial



(a) Rrs 413 (b) Rrs 443 (c) Rrs 490 (d) Rrs 510

(e) Rrs 560 (f) Rrs 620 (g) Rrs 665 (h) Rrs 681

Fig. 2 Images recorded at each wavelength for the day 23-October-2009 at time 11:02:49.

resolution between the years 2002-2011. A Level-2 data product is the result of the sensor calibration and atmospheric correction, consisting of derived geophysical variables generated from the corresponding Level-1A product using the standard National Aeronautics and Space Administration (NASA) processing methodologies. Considering the resolution of the images, the data set consists of $370 \times 370 = 136900$ pixels images that are recorded for each wavelength in different time periods (Figure 2). The data set was passed through a quality control process corresponding to the L2 flags given in Caballero et al (2014a) in order to remove the suspicious and low-quality data points.

2.1.3 Data matching

The Rrs values recorded by the sensor are matched up with field measurements by considering the coordinate and the time at which the sample is collected. Data match-ups are performed by matching the in-situ observations with the satellite data within a square area of 5 by 5 pixels (1.5×1.5 km^2) whose central pixel contains the coordinates of the field measurement, which is a standard approach in remote sensing community. Then the satellite data are averaged over these 25 pixels (excluding those with low-quality data) and the results are matched to the in-situ observations. As a result of matching, a total of 71 observations are obtained. However, a careful consideration of scales is critical when comparing remotely-sensed data with in-situ observations, particularly because of the large spatio-temporal heterogeneity of estuarine and coastal wa-

ter properties influencing those measurements (Fettweis and Nechad, 2011). In this sense, the time difference between satellite overpasses and the collection time of in-situ samples is reduced by a filter of < 1.5 hours from acquisition, thus preventing temporal biases to further evaluate the results of each data set. As a result of this filtering process, the total number of observations decrease from 71 to 31. Furthermore, 6 observations were removed from the data set for various reasons: four observations for not being collected from the water surface; one observation due to the measurement error during filtering process, and one observation due to the missing values at some band values. Finally, analysis was conducted on 25 observations. If a wider time window of 4 or 5 hours is used, a major number of match-ups are obtained, although greater variability is encountered with the inconvenience of greater discrepancies between in-situ and RS observations.

2.2 Standard Regression Models

We refer as standard regression model to any linear regression, either simple (one explanatory variable) or multiple (many explanatory variables), where the response is the logarithm of the TSS ($Y = \log(\text{TSS})$) and the predictors are reflectance values recorded at a single wavelength or a combination of different wavelengths. Let X_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, p$, be the reflectance value at wavelength w_j at pixel i, and let Y_i be the logarithm of the TSS for the same pixel. Then the generic linear regression model is

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + \varepsilon_i,\ i = 1, \ldots, n, \tag{1}$$

where ε_i, $i = 1, \ldots, n$ are assumed to be independent zero mean random variables with a common variance σ^2. The corresponding model for TSS is

$$\text{TSS}_i = \exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + \varepsilon_i\right),\ i = 1, \ldots, n.$$

This last model for TSS is known as *exponential regression model* in the remote sensing literature. See, for instance, Caballero et al (2014a,b), Nechad et al (2010), or Nezlin and DiGiacomo (2005).

The regression model (1) is usually estimated by Ordinary Least Squares (OLS) solving the problem

$$\min_{\beta_0, \ldots, \beta_p} \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij}\right)^2.$$

Numerical instability can appear when the number of predictor $(p + 1)$ is large, compared with the number of observations n, or when the predictors

are highly correlated between them. In these circumstances it is advisable to use penalized versions of OLS estimation, solving instead the problem

$$\min_{\beta_0,\ldots,\beta_p} \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij} \right)^2 + \lambda \left(\sum_{j=1}^{p} |\beta_j|^d \right)^{1/d},$$

where $\lambda > 0$ is known as the penalization parameter. The most popular penalized OLS estimators are when $d = 1$ (LASSO, Least Absolute Shrinkage and Selection Operator) or $d = 2$ (Ridge regression). Hastie et al (2015) is a good reference for LASSO and Ridge regression.

An alternative way is to use elastic net estimator (Zou and Hastie, 2005) which is a compromise between Ridge and Lasso estimators. The elastic net objective function is

$$\min_{\beta_0,\ldots,\beta_p} \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij} \right)^2 + \lambda \left(\alpha \sum_{j=1}^{p} |\beta_j| + (1-\alpha) \sum_{j=1}^{p} \beta_j^2 \right).$$

Elastic net, which includes as special cases LASSO ($\alpha = 1$) and Ridge regression ($\alpha = 0$), can do variable selection and dealing with correlated predictors simultaneously.

2.3 Functional Linear Regression Models

A functional linear regression model with a scalar response is defined as

$$Y = \int_T \chi(t)\beta(t)dt + \epsilon, \tag{2}$$

where Y indicates the scalar response, ϵ is the random error term (with 0 mean and unknown variance σ^2), $\chi(t)$ and $\beta(t)$ define the (observable) functional predictor and the (unknown) functional parameter, respectively, which are defined as real functions on a continuous interval T. In order to estimate the unknown elements of this model (σ^2 and $\beta(t), t \in T$) we observe n pairs (χ_i, Y_i), $i = 1 \ldots, n$, following model (2) with independent errors $\epsilon_1, \ldots, \epsilon_n$.

The main additional difficulty with the functional linear regression model, as compared with the standard multiple regression model, is that tractable representations of functions χ and β are required. Several methods have been proposed to estimate model (2). In this study, we focus on three different approaches that are based on cubic B-spline basis expansion, functional principal components analysis, and functional partial least squares analysis, respectively.

2.3.1 B-spline Basis Approach

Both the predictor and the parameter estimate are functionals and can be approximately expanded in terms of cubic B-spline basis functions:

$$\chi_i(t) \approx \tilde{\chi}_i(t) = \sum_{k=1}^{K} c_{ik}\phi_k(t) = \mathbf{c}_i'\boldsymbol{\phi}(t),$$

$$\beta(t) \approx \tilde{\beta}(t) = \sum_{l=1}^{L} \hat{b}_l\theta_l(t) = \mathbf{b}'\boldsymbol{\theta}(t) = \boldsymbol{\theta}'(t)\mathbf{b},$$

where $\{\phi_1(t), \ldots, \phi_K(t)\}$ and $\{\theta_1(t), \ldots, \theta_L(t)\}$ are two (possible different) B-spline basis of functions defined on T, $\boldsymbol{\phi}(t) = (\phi_1(t), \ldots, \phi_K(t))'$, and $\boldsymbol{\theta}(t) = (\theta_1(t), \ldots, \theta_L(t))'$.

Replacing in (2) χ_i and β by their B-spline expansions, we obtain an approximated model

$$Y_i \approx \int_T \tilde{\chi}_i(t)\tilde{\beta}(t)dt + \epsilon_i = \mathbf{c}_i' \int_T \boldsymbol{\phi}(t)\boldsymbol{\theta}'(t)dt\,\mathbf{b} + \epsilon_i = \mathbf{c}_i'\mathbf{J}_{\phi\theta}\mathbf{b} + \epsilon_i.$$

In matrix notation we have $\mathbf{Y} = \mathbf{C}\mathbf{J}_{\phi\theta}\mathbf{b} + \boldsymbol{\epsilon}$, the usual expression of a multiple regression model with vector coefficient $\mathbf{b}$. Let $\hat{\mathbf{b}}$ be the ordinary least square estimator of $\mathbf{b}$ and define $\hat{\beta}(t) = \boldsymbol{\theta}'(t)\hat{\mathbf{b}}$. Then the fitted values of Y_i, $i = 1, \ldots, n$, are $\hat{Y}_i = \mathbf{c}_i'\mathbf{J}_{\phi\theta}\hat{\mathbf{b}} = \int_T \tilde{\chi}_i(t)\hat{\beta}(t)dt$.

In the procedure that we have just described, the smoothness of the estimated coefficient function $\hat{\beta}(t)$ is controlled only by the number L of elements in the corresponding B-spline basis. An alternative approach consists in adding a roughness penalty term to the sum of squared residuals:

$$\texttt{PENSSR}_\lambda(\tilde{\beta}) = \sum_{i=1}^{n} \left(Y_i - \int_T \tilde{\chi}_i(t)\tilde{\beta}(t)dt \right)^2 + \lambda \int_T \left(\tilde{\beta}''(t) \right)^2 dt.$$

See, for instance, Section 15.4 in Ramsay and Silverman (2005) or Marx and Eilers (1999) for an alternative approach leading to P-splines.

2.3.2 Functional Principal Components Regression

The main idea of Functional Principal Components Regression (FPCR) is to predict the scalar response Y from the principal component scores of functional data $\chi(t)$. This method is based on the Karhunen-Loève Theorem which states that a square integrable function $\chi(t)$, $t \in T$, with mean $\mu_\chi(t) = \mathbf{E}[\chi(t)]$ and the covariance function $c_\chi(t,s) = \mathbf{E}[(\chi(t) - \mu(t))(\chi(s) - \mu(s))]$, can be represented in terms of the eigenfunctions $\xi_1(t), \xi_2(t), \ldots$ of the covariance operator Γ_χ (the kernel function of which is the covariance function $c_\chi(t,s)$) as

$$\chi(t) = \mu_\chi + \sum_{j=1}^{\infty} f_j\xi_j(t),$$

where $f_j,\ j \geq 1$, are uncorrelated zero mean random variables (known as *scores*) satisfying $E[f_j^2] = \lambda_j$, where λ_j is the eigenvalue corresponding to the j-th eigenfunction of the covariance operator.

Consider a sample of independent identically distributed observations (χ_i, Y_i), $i = 1 \ldots, N$. Let $\bar{\chi} = (1/N) \sum_{i=1}^{N} \chi_i$ be the sample mean of functional predictors, and let

$$\hat{c}_\chi(t,s) = \frac{1}{N} \sum_{i=1}^{N} (\chi_i(t) - \bar{\chi}(t))(\chi_i(s) - \bar{\chi}(s)),\ s,t \in T,$$

be the empirical covariance function. According to the Karhunen-Loève Theorem, the functional observations χ_i can be approximated in terms of the first K eigenfunctions $\hat{\xi}_j(t)$ of the sample covariance operator $\hat{\Gamma}_\chi$ (the kernel of which is the empirical covariance function $\hat{c}_\chi(t,s)$) and the empirical functional principal component scores $\hat{f}_{ij} = \langle \chi_i - \bar{\chi}, \hat{\xi}_j \rangle$ for $i = 1, \ldots, N$ and $j = 1, \ldots, K$

$$\chi_i(t) \approx \tilde{\chi}_i(t) = \bar{\chi}_i + \sum_{j=1}^{K} \hat{f}_{ij} \hat{\xi}_j(t),\ i = 1, \ldots, N,\ j = 1, \ldots, K. \tag{3}$$

The empirical eigenfunctions satisfy the conditions $\int_T \hat{\xi}_j(t)\hat{\xi}_j(t)dt = 0$, for $i \neq j$ and $\int_T \hat{\xi}_j(t)\hat{\xi}_j(t)dt = 1$, so that they form an orthonormal basis in $L^2(T)$. Then the functional parameter of the FLRM (2) can be approximated by Equation (4) in terms of empirical functional components $\hat{\xi}_j$ related to the largest eigenvalues $\hat{\lambda}_j$ in a K dimensional space

$$\beta(t) \approx \tilde{\beta}(t) = \sum_{j=1}^{K} b_j \hat{\xi}_j(t). \tag{4}$$

Hence, the FLRM between the centered scalar response and the centered functional predictors $\tilde{\chi}_i(t) - \bar{\chi}(t)$ takes the form of

$$\begin{aligned} Y_i - \bar{Y} &\approx \int_T (\tilde{\chi}_i(t) - \bar{\chi}(t))\tilde{\beta}(t)dt + \epsilon_i = \int_T \left(\sum_{j=1}^{K} \hat{f}_{ij}\hat{\xi}_j(t) \right) \left(\sum_{h=1}^{K} b_h \hat{\xi}_h(t) \right) dt + \epsilon_i \\ &= \sum_{j=1}^{K} \sum_{h=1}^{K} \hat{f}_{ij} b_h \int_T \hat{\xi}_j(t)\hat{\xi}_h(t)dt + \epsilon_i = \sum_{j=1}^{K} \hat{f}_{ij} b_j + \epsilon_i. \end{aligned} \tag{5}$$

Assume that $\hat{b}_j$ refers to the ordinary least squares estimate of b_j in the Equation (5). Then the fitted values of $Y_i, i = 1, \ldots, N$ are computed as $\hat{Y}_i = \bar{Y} + \sum_{j=1}^{K} \hat{f}_{ij}\hat{b}_j$. Moreover, $\beta(t)$ is estimated by

$$\hat{\beta}(t) = \sum_{j=1}^{K} \hat{b}_j \hat{\xi}_j(t). \tag{6}$$

The number K of eigenfunctions can be chosen in different ways (see, for instance, Horváth and Kokoszka 2012, Section 3.3). When the main objective is to approximate χ_i by $\hat{\chi}_i$, as in Equation (3), the method based on CVP (cumulative percentage of total variance: $100 \times \sum_{j=1}^{K} \hat{\lambda}_j / \sum_{j=1}^{N} \hat{\lambda}_j$) give good results in practice: K is chosen in order that the CVP exceeds a desired level (85%, for instance). However, when the objective is fitting a functional regression model (as in our case) it could happen that the first K principal functions are not the most related with the response Y. In this context the selection of how many principal functions should be used (and which of them might be chosen) can be done by cross-validation or model selection criteria, such as Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC, also known as Schwarz Information Criterion, SIC).

Even if we have presented B-splines and FPCR separately, both methodologies can also be combined. For instance, FPCA for the predictor curves and a B-splines basis for the coefficient function (see, e.g., Goldsmith et al 2011, and references therein).

2.3.3 Functional Partial Least Squares Regression

In order to increase the predictive performance of FLRM, Preda and Saporta (2005) have proposed the Functional Partial Least Squares Regression (FPLSR) approach as an alternative to FPCR. This method is based on regressing the response on FPLS scores which are obtained from the maximization of the covariance between the functional predictor $\chi(t)$ and the scalar response Y. Since FPLS components are more related to the variability of the response, they are more relevant to predicting the outcome (Preda and Saporta, 2005; Aguilera et al, 2010; Febrero-Bande et al, 2015). Similarly to the case of FPCR, FPLS seeks for an orthonormal basis of functions $\{\phi_l\}_{l\geq 1}$ allowing predictors and the functional parameter to be expanded as $\chi(t) = \mu_\chi + \sum_{l=1}^{\infty} \upsilon_l \phi_l(t)$ and $\beta(t) = \sum_{l=1}^{\infty} c_l \phi_l(t)$, respectively. In practice, the FPLS components ϕ_l are obtained from an iterative algorithm (see, for instance, Delaigle et al 2012 or Febrero-Bande et al 2015 for detailed explanations) leading to the representations

$$\chi_i(t) \approx \tilde{\chi}_i(t) = \bar{\chi} + \sum_{l=1}^{L} \hat{\upsilon}_{il} \hat{\phi}_l(t),\ i = 1, \ldots, N, \qquad \beta(t) \approx \tilde{\beta}(t) \approx \sum_{l=1}^{L} c_l \hat{\phi}_l(t),$$

where c_l, $l = 1, \ldots, L$, indicate unknown constants, and the FPLS scores $\hat{\upsilon}_{il}$ are computed as $\hat{\upsilon}_{il} = \langle \chi_i - \bar{\chi}, \hat{\phi}_l \rangle$, $i = 1, \ldots, N$, $l = 1, \ldots, L$.

The FLRM model can be approximated by a finite linear combination of FPLS scores $\hat{\upsilon}_{il}$:

$$\begin{aligned} Y_i - \bar{Y} &\approx \int_T (\tilde{\chi}_i(t) - \bar{\chi}(t)) \tilde{\beta}(t) dt + \epsilon_i = \int_T (\tilde{\chi}_i(t) - \bar{\chi}(t)) \sum_{l=1}^{L} c_l \hat{\phi}_l(t) dt + \epsilon_i \\ &= \sum_{l=1}^{L} c_l \int_T (\tilde{\chi}_i(t) - \bar{\chi}(t)) \hat{\phi}_l(t) dt + \epsilon_i = \sum_{l=1}^{L} \hat{\upsilon}_{il} c_l + \epsilon_i. \end{aligned}$$

Let $\hat{c}_l$ be the ordinary least squares estimate of c_l. Then the predicted values of the responses are obtained from $\hat{Y}_i = \bar{Y}_i + \sum_{l=1}^{L} \hat{c}_l \hat{\upsilon}_{il}$. The $\beta(t)$ is estimated by

$$\hat{\beta}(t) = \sum_{l=1}^{L} \hat{c}_l \hat{\phi}_l(t). \tag{7}$$

2.4 Measuring the performance of the regression models

To compare the predictive performance of the different regression models, an adjusted version of Mean Square Prediction Error (aMSPE) based on Leave-One-Out Cross-Validation (LOOCV) is defined by

$$\text{aMSPE} = \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i^{(i)})^2}{\sum_{i=1}^{n} (Y_i - \bar{Y}^{(i)})^2}, \tag{8}$$

where $\hat{Y}_i^{(i)}$ is the prediction for observation i using the model fitted on the other $(n-1)$ observations, and $\bar{Y}^{(i)}$ indicates the mean computed after removing the i-th observation. Observe that the denominator of Equation (8) is n times the LOOCV estimation of the MSPE for the constant model, and it serves as a reference for other models. In our case we use always the logarithm of TSS values as the response values Y_i.

Some regression models require the choice of tuning parameters (such as the number of components in PCR and PLSR models or the penalty parameter λ in LASSO, Ridge and Elastic net regression, and also α parameter in Elastic net). In such cases, an automatic choice mechanism (LOOCV, K-fold CV, GCV or BIC, for instance) within the LOOCV should be used, in order that the numerator of Equation (8) is n times an unbiased estimator of the MSPE for these regression models. In our examples we use LOOCV within LOOCV for non-functional models. For functional models, either GCV or BIC within LOOCV have been used (see Section 3.2 for more details).

3 Results

We present the results of fitting standard regression models and FLRMs to the 25 observations for which the time and space matching between TSS content (in-situ response) and Rrs values at 8 wavelengths (predictors from satellite) has been possible.

3.1 Results of Standard Regression Models

Several methods have been used to investigate the relationship between TSS concentration and Rrs values. Generally, Rrs values are proposed for use at the band which is most closely correlated with the response and for fitting

a regression model. In the literature, wavelength Rrs 665 has been found to be the band most closely with TSS concentration (Binding et al, 2003, 2005; Caballero and Navarro, 2016). For MERIS data, Nechad et al (2010) used the bands 665 nm and 681 nm to model TSS. In the study by Caballero and Navarro (2016), a simple regression model with band 665 nm was used to analyze the relationship between TSS and Rrs.

In our data set, the most correlated bands with the logarithm of the TSS content are found to be the wavelengths 665 nm ($r = 0.729, p < 0.001$) and 681 nm ($r = 0.734, p < 0.001$), while 510 nm was the least correlated wavelength ($r = 0.567, p = 0.003$). Therefore, two simple regression models were fitted between TSS values and Rrs values at the wavelengths 665 nm and 681 nm separately.

In-situ TSS measurements ranged between 3 and 327 mg/L, while Rrs values at the wavelength 681 nm ranged between 0 and 0.0275 sr^{-1}, and Rrs values at the wavelength 665 nm ranged between 0 and 0.028 sr^{-1}. As may be seen from the scatter plot of the observations in Figure 3, the dispersion of Rrs values are quite similar for both bands.

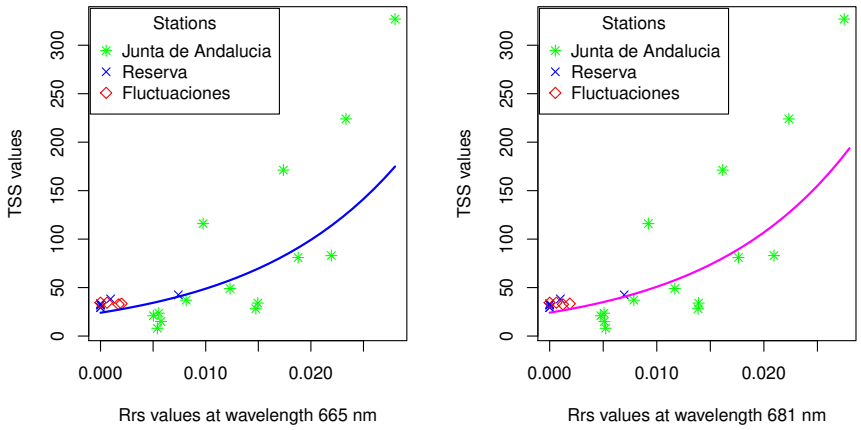


Fig. 3 The scatter plot between in-situ TSS and Rrs values at wavelengths 665 nm and 681 nm.
The blue curve indicates the relationship between TSS and Rrs 665 nm while the pink curve indicates the relationship between TSS and Rrs 681 nm.

The fitted simple regression model with band 665 nm is

$$\widehat{\text{TSS}} = \exp(3.18 + 70.83 * \text{Rrs } 665),$$

and the model with band 681 nm is

$$\widehat{\log(\text{TSS})} = 3.18 + 74.38 * \text{Rrs } 681.$$

The simple regression model with band 665 nm on the MERIS data of the same period was proposed in the study by Caballero et al (2014b). The MERIS data of this study consists in the last Ocean Color reprocessing by Mission required for improving the archived products (https://oceancolor.gsfc.nasa.gov/reprocessing/r2012.1/meris, December 2012). However, Caballero et al (2014b) used the first processing of the full resolution MERIS database. In this regard, the minor deviation between both data is explained for the different reprocessing history.

In addition to these univariate models, linear regression using all the band values, stepwise regression, LASSO estimation, Ridge regression and Elastic net regression are also applied to the data set, as they are implemented in the R library `glmnet` (Friedman et al 2010). In order to avoid correlation between the explanatory variables and to reduce dimension, regression models on Principal Components (PC) and on Partial Least Square Components (PLSC) are used as well, by the R library `pls` (Björn-Helge and Wehrens 2007).

All the models are compared in terms of the aMSPE criterion given in Equation (8). These values are shown in Table 1. Among the non-functional regression models, the best performance (in terms of aMSPE) corresponds to the Ridge and Elastic net regression models, followed by LASSO. Regression models based on PCR and PLSCR was exhibited a worst behavior than expected: aMSPE is larger than 1 in both cases.

Observing Equation (8) we realize that aMSPE values are the average of $n = 25$ values, namely

$$\frac{(Y_i - \hat{Y}_i^{(i)})^2}{(1/n)\sum_{i=1}^{n}(Y_i - \bar{Y}^{(i)})^2},\ i = 1, \ldots, n,$$

that have been obtained in the LOOCV process. In order to add variability information to the location information provided by aMSPE values, Figure 4 shows the box-plots of these $n = 25$ values for each regression model. These box-plots confirm that, among the non-functional models, Ridge, Elastic net and LASSO have the best performance, and that PCR and PLSCR have the worst. However the impression of linear model with all bands and stepwise regression is more positive compared to the simple regression models using only 665 or 681 nm bands, even though the aMSPE values are higher.

3.2 Results of Functional Regression Models

All the computations involving functional elements have been done using the R libraries `fda` (Ramsay et al 2017) and `fda.usc` (Febrero-Bande and Oviedo de la Fuente 2012). Cubic B-spline fitting with $K = 8$ basis is used to represent the data as functional objects. We have used $K = 8$ basis in order to interpolate the observed data (8 observations per function), given the low number of observed points at each function and the low noise level. The resulting curves evaluated in a fine grid are given in Figure 5.

Table 1 The Comparison of Standard versus Functional Regression Models

Model	aMSPE
Standard regression models	
Using only band 665 nm	0.51
Using only band 681 nm	0.50
Using all bands	0.59
Stepwise regression	0.54
LASSO	0.44
Ridge regression	**0.41**
Elastic net	**0.41**
PCR	1.35
PLSCR	1.08
Functional regression models	
FLRM with B-spline bases	0.44
FPCR	0.43
FPLSR	**0.39**

The outlier detection of the functional data set is carried out by a procedure based on weighting and bootstrap in which the number of bootstrap samples and the quantile to determine the cut off value obtained from Bootstrap sample are taken as 200 and 0.5, as suggested by default in Febrero-Bande and Oviedo de la Fuente (2012). Four different depth measures are used in computations: Fraiman-Muniz Depth (FMD), Modal Depth (MD), Random Tukey Depth (RTD) and Random Projection Depth (RPD). Although there are suspicious observations in the data set, none of them have been recognized as outliers.

The functional regression models used to model the logarithm of TSS content are FLRM based on the B-spline basis approach, FPCR and FPLSR. In FLRM, the number of basis function for representing the coefficient function ($\beta(t)$) has been chosen as $L = 8$, because this is the number of basis used for interpolating the observed points at each function. We control the smoothness of the estimated coefficient function with a roughness penalty approach. Specifically, within each of the 25 outer LOOCV runs, we use inner generalized cross-validation as it is implemented in the function `fregre.basis.cv` from library `fda.usc`, with the option `type.CV = GCV.S`.

BIC (or SIC) is used to identify the optimal number of components in FPCR and FPLSR. The CV criterion has not been considered since it requires more computing time (see for instance Febrero-Bande and Oviedo de la Fuente 2012). More precisely, we use functions `fregre.pc.cv` and `fregre.pls.cv` of library `fda.usc` (Febrero-Bande and Oviedo de la Fuente 2012) with the option `criteria = "SIC"`.

In FPCR, the fifth and the second components (with 4.3% and 6.8% explained variance, respectively) are chosen by BIC, while for the case of FPLSR the first two components are chosen. The chosen Functional Principal Components (FPCs) and Functional Partial Least Squares Components (FPLSCs) are displayed in Figure 6.

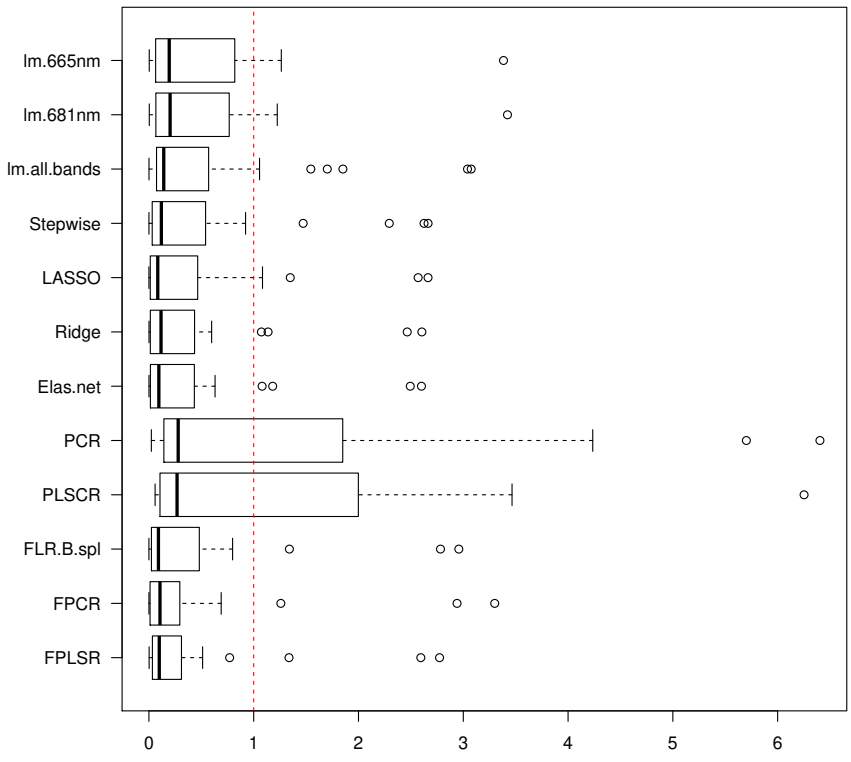


Fig. 4 Comparison of regression models. For each model, the box-plot corresponds to the $n = 25$ values of $(Y_i - \hat{Y}_i^{(i)})^2/\{(1/n)\sum_{i=1}^{n}(Y_i - \bar{Y}^{(i)})^2\}$ whose average is the corresponding aMSPE value in Table 1. The vertical red dashed line at 1 serves as a reference.

As may be seen in the left panel of Figure 6, the second FPC gives higher positive weight to the band values around 550 nm, while it gives negative weights to the lowest and mainly to the highest band values. The fifth FPC is harder to interpret: it has an oscillatory behavior giving positive weight to low, medium and high band values, and negative to the rest. For the case of FPLSR (the panel on the right in Figure 6), the first FPLSC is a stable function (it could be interpreted as an average of all bands, with higher weights for higher bands), while the second FPLSC is similar to the second FPC, except for a sign change. The parameter estimates of models FPCR and FPLSR ($\hat{b}_j$ and $\hat{c}_l$ in Equations 6 and 7) are given in Table 2.

The aMSPE values computed by LOOCV are found to be lower for the FPCR and FPLSR approaches compared to the B-spline basis approach, and

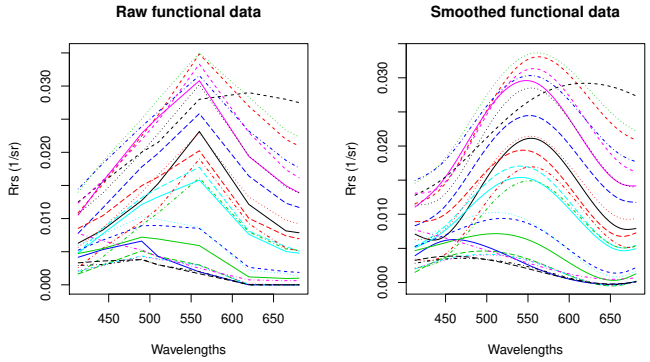


Fig. 5 Raw functional data and curves smoothed by cubic B-splines.

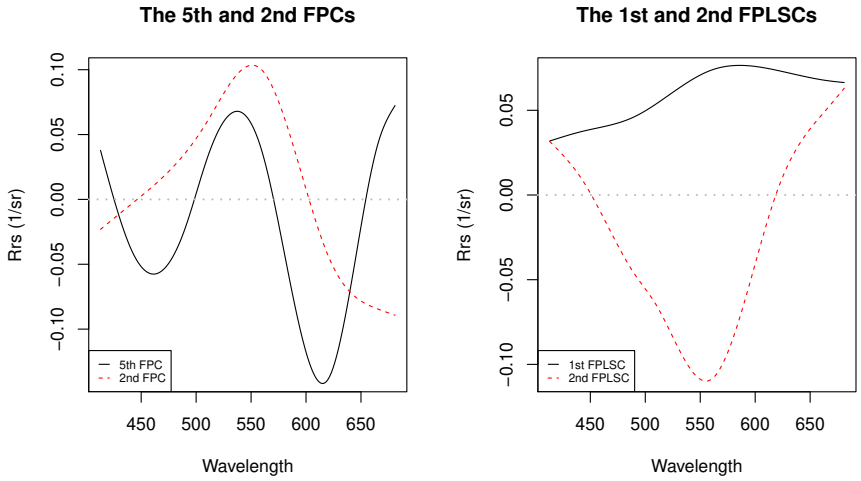


Fig. 6 Components from FPCA (left) and FPLS (right) that have been selected by BIC to be included in FPCR and FPLSR Models.

this one (the worst among functional regression models) coincides with Ridge regression (the best result for standard regression models). This is a very positive result for functional regression models when compared with the standard ones.

The functional parameter estimations of FPCR and FPLSR models (computed as in Equations 6 and 7) are shown in Figure 7. The functional parameter estimate of the FPLSR model in the right panel of Figure 7 indicates that the difference between extreme (whether high or low) and medium bands is an effective way to predict the response. The functional parameter estimate

Table 2 The coefficient estimates of functional components

Parameter	Estimate	Standard Error	p value
FPCR			
Intercept	3.76	0.10	< 0.001
FPC 5	-24.05	3.92	< 0.001
FPC 2	-8.34	3.12	0.013
FPLSR			
Intercept	3.76	0.10	< 0.001
FPLSC 1	0.12	0.005	< 0.001
FPLSC 2	0.20	0.001	< 0.001

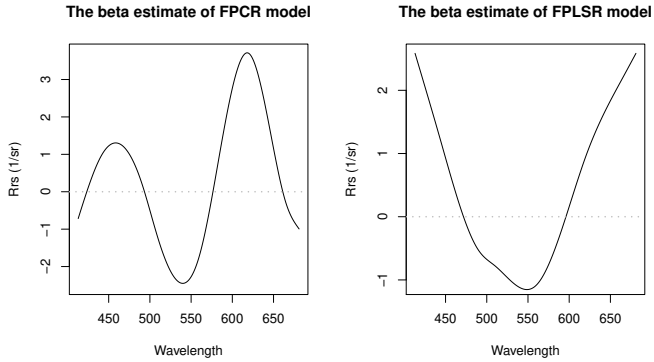


Fig. 7 The functional parameter estimations of FPCR and FPLSR Models.

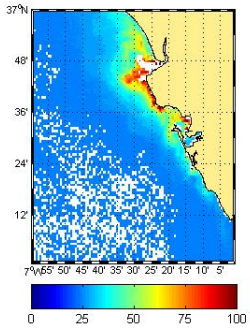
of the FPCR model in the left panel has an interpretation quite coincident with the previous one, with the following differences: the very extreme bands have lower weight (close to 0) and the relevant medium bands are now more concentrated.

Regarding Table 1, the FPCR and FPLSR models give better predictions than the standard models. The FPCR model with two components (chosen by BIC) has the lowest aMSPE value (thus predicting the logarithm of TSS better than other models) closely followed by FPLSR.

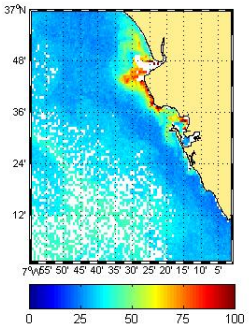
As an example, the predicted values of the logarithm of TSS concentrations on the day 23-October-2009 obtained from four different models are shown in Figure 8.

4 Simulation study

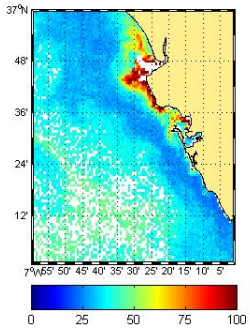
A simulation study has been designed in order to compare the predictive performance of the proposed regression models and to support findings. For the simulation study, the area of interest is that described in Section 2: $36° - 37°$ N latitude and $6° - 7°$ W longitude (see Figure 1).



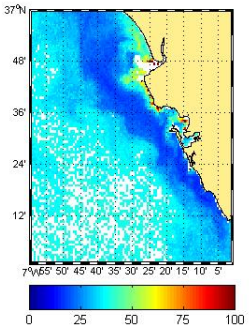
(a) Exponential Regression with 665 nm



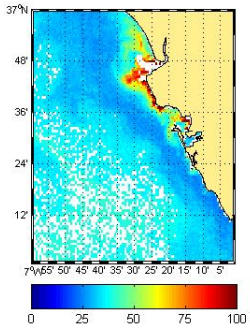
(b) Elastic Net Regression with all the bands



(c) FLRM with Basis Expansion



(d) FPCR with 2 components chosen by BIC



(e) FPLSR with 2 components chosen by BIC

Fig. 8 Predicted Images for TSS concentrations (in mg/L) on the day 23-October-2009 at time 11:02:49. Pixels that have been removed by the control quality after L2 flagging are represented in white.

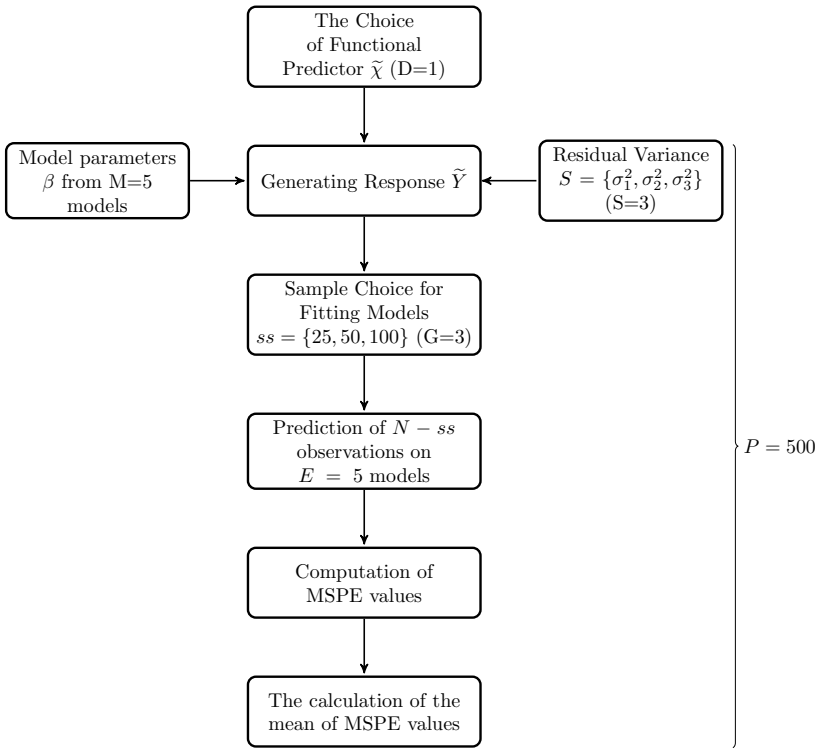


Fig. 9 Design of the simulation study.

The simulation study consists of four main steps (Figure 9 shows the simulation scheme). The first step is to choose the day and hour to be simulated. This is done by choosing the satellite image with the maximum number of full pixels in the last wavelength 681 nm, since the amount of Rrs in this band is usually lower compared to other bands. So the image chosen as the functional predictor corresponds to the day 11-July-2009 and hour 10:31:31; it has 9843 pixels with valid values for wavelength 681 nm. The Rrs values at the eight wavelengths for this day has been merged and the functional data object has been created. A total of 66 observations were removed due to the missing values at the other wavelengths. Finally, 9777 functional observations were left (N=9777).

In the second step the TSS state of the sea is simulated. In order to do that, the scalar response vector $\widetilde{Y}$ is generated according to the estimated models in Section 3, using as predictors either Rrs values or Rrs functions for each pixel of the chosen image. Five models ($M = 5$) are used in generating the response: a

simple linear regression with 665 nm wavelength; Elastic net regression; FLRM with the B-spline Basis approach; FPCR with all components, and FPCR with the two components chosen by the BIC criterion. FPLSR could not be used in the generation of the response, since this method requires knowledge of the covariance between the response and the predictor. Here, the covariance can not be computed because the response is unknown.

The response vector generated when using the simple linear regression model (with band 665 nm as predictor) is denoted by $\widetilde{Y}_{lm}$. The Elastic net regression model estimated in Section 3.1 has all non-null coefficients. The response vector generated with this model is denoted by $\widetilde{Y}_{elas}$. The general form of simulated FLRMs can be expressed as follows:

$$\widetilde{Y}_i = \int \tilde{\chi}_i^*(t)\tilde{\beta}(t)dt + \epsilon_i. \tag{9}$$

Here, $\tilde{\chi}_i^*(t)$ denotes the functional predictor which is composed of Rrs values at 8 different wavelengths, and $\tilde{\beta}(t)$ is the functional parameter estimate that is taken respectively from the functional models used in the application: FLRM with B-spline basis expansion, FPCR with all components and FPCR with the components chosen by BIC. The responses of the functional linear models are denoted respectively by $\widetilde{Y}_{\text{Bs}}$, $\widetilde{Y}_{\text{FPC 8}}$ and $\widetilde{Y}_{\text{FPC 2}}$.

The response $\widetilde{Y}_{Bs}$ for FLRM with B-spline basis approach is generated from

$$\widetilde{Y}_{Bs} = \int \tilde{\chi}^*(t)\tilde{\beta}(t)dt + \epsilon_i = \mathbf{c}_i^*\mathbf{J}_{\phi\theta}\mathbf{b}^* + \epsilon_i,$$

where $\mathbf{J}_{\phi\theta}$ is the matrix computed from the inner product of basis functions $\boldsymbol{\phi}(t) = [\phi_1(t), .., \phi_8(t)]'$ and $\boldsymbol{\theta}(t) = [\theta_1(t), ..., \theta_7(t)]'$ that are used to extend $\tilde{\chi}_i^*(t) = \sum_{k=1}^{8} c_{ik}^*\phi_k(t)$ and $\tilde{\beta}(t) = \sum_{l=1}^{7} b_l^*\theta_l(k)$, respectively. Here, $\mathbf{c}_i^* = [c_{i1}^*, ..., c_{i8}^*]'$ and $\mathbf{b}^* = [b_1^*, ..., b_7^*]'$ indicate the computed coefficient vectors to expand the functional predictor and the parameter function in the corresponding basis functions.

The expressions $\mathbf{b}^*$ and $\mathbf{J}_{\phi\theta}$ in this model are replaced by the related parameters of the FLRM in the Section 3.2, whereas $\mathbf{c}_i^*$ is obtained from the smoothing of the chosen functional predictor $\tilde{\chi}_i^*(t)$ on the 8 functions forming a B-spline basis.

In the case of FPCR, Model (9) is approximated by,

$$\widetilde{Y}_{FPC} = \bar{Y}_i + \mathbf{F}_{pc}^*\mathbf{b}^* + \epsilon_i,$$

where $\bar{Y}_i$ is the mean of the real response vector, $\mathbf{F}_{pc}^*$ denotes the score matrix which is computed from $\mathbf{F}_{pc}^* = \langle \chi, \hat{\xi} \rangle$, and $\mathbf{b}^*$ is the parameter vector obtained from the expansion $\beta_{FPC}(t) = \sum_{j\in J} b_j^*\hat{\xi}_j$ used in the application. J indicates the set of components used in the expansion of the data, which is equal to $\{1, \ldots, 8\}$ for FPCR with all the components, and is equal to $\{5, 2\}$ for FPCR with the BIC method.

The error term ϵ_i is generated from the normal distribution with zero mean and the variance equal to the raw residual variance of the model from

which the functional parameter is taken. The raw residual variances of the models in Section 3 are quite high. Hence, three different values ($S = 3$) of residual variance are used in the simulation: the raw residual variances of the related model ($\sigma_1^2 = \hat{\sigma}^2$), one fifth of the model residual variance($\sigma_2^2 = \hat{\sigma}^2/5$), one tenth of the model residual variance ($\sigma_3^2 = \hat{\sigma}^2/10$). For each value of the residual variance and each type of model, the response is generated five hundred ($P = 500$) times.

The third step consists on simulating the in-situ data collection campaigns. Three ($G = 3$) different sample sizes $ss = \{25, 50, 100\}$ are considered. ss out of $N = 9777$ pixels are randomly chosen from the image. Therefore, the simulation has $M \times S \times G$, i.e. 45, possible scenarios. Following the same procedures as in Section 3, and using the pixels in the sample of size ss, for each scenario a total of $E = 5$ models are estimated: a simple linear regression with band 665 nm; Elastic net regression (parameters λ and α chosen by LOOCV); the B-spline basis expansion; the FPCR BIC, and the FPLSR BIC models.

The last step consists in using the estimated models for predicting the responses $\tilde{Y}_i$ for the other $(N - ss)$ pixels. The predictive performance of the models are measured in terms of the their MSPE values.

Finally steps 2, 3 and 4 are repeated $P = 500$ times. So the mean of MSPE values are taken over the $P = 500$ simulations. The results of the simulation study are given in Table 3. The best results for each scenario are marked in bold. Figure 10 offers an alternative way to look at the these data: for each estimating model i in a particular scenario, the ratio

$$\frac{\min(\text{mean MSPE}_i)}{\text{mean MSPE}_i}$$

is represented with a color code. The models for which this ratio is closer to one (lighter colors) have better performances.

In general, the best predictions are obtained from the models used for generating responses. This is clearer for larger sample sizes and/or smaller residual variances. As seen from the first block of Table 3, in the case of low sample size and high residual variance (σ_1^2), all the response types have been predicted better by Elastic net. However, when sample sizes increases or residual variance decreases, the Elastic net is beaten by other estimators, even for data generated according to the Elastic net model.

The mean MSPE values of the models FPCR BIC and FPLSR BIC are usually very similar. It turns out that with the increment of the sample size the predictive performance of functional linear models improves and get closer to each other. The responses based on functional models ($\widetilde{Y}_{Bs}$,$\widetilde{Y}_{FPC\ 8}$, $\widetilde{Y}_{FPC\ 2}$) are predicted better by functional regression models.

Although the response generated from simple linear regression model ($\widetilde{Y}_{lm}$) is predicted better by the simple linear regression model, one may observe that the predictions given by Elastic net and the functional linear models are only slightly worst, particularly in the case of low residual variance and high sample size (last block of Table 3; see also Figure 10).

Table 3 Simulation results. Mean of MSPE values. Each row corresponds to one of the 45 scenarios (combinations of generating model, sample size and residual variance). The columns correspond to the 5 estimating models.

Sample size 25						
Type of Response	Residual Variance	Lin. Reg. on 665 nm	Elastic Net Reg.	FLRM Basis	FPCR BIC	FPLSR BIC
$\widetilde{Y}_{exp}$	σ_1^2	1.10	**1.04**	1.33	1.40	1.27
	σ_2^2	**0.32**	0.56	0.40	0.47	0.43
	σ_3^2	**0.10**	0.20	0.14	0.19	0.17
$\widetilde{Y}_{elas}$	σ_1^2	1.25	**1.04**	1.26	1.30	1.26
	σ_2^2	1.50	0.47	**0.36**	0.40	**0.36**
	σ_3^2	1.56	0.20	0.15	**0.13**	**0.13**
$\widetilde{Y}_{Bs}$	σ_1^2	1.27	**1.02**	1.21	1.28	1.29
	σ_2^2	1.57	0.50	**0.22**	0.29	0.23
	σ_3^2	1.60	0.24	**0.06**	0.08	0.07
$\widetilde{Y}_{FPC\ 8}$	σ_1^2	1.31	**1.02**	1.21	1.51	1.31
	σ_2^2	1.52	0.33	0.49	**0.24**	0.25
	σ_3^2	1.53	0.20	0.42	**0.06**	**0.06**
$\widetilde{Y}_{FPC\ 2}$	σ_1^2	1.40	**1.05**	1.22	1.28	1.29
	σ_2^2	1.96	0.58	0.29	0.30	**0.28**
	σ_3^2	2.04	0.35	0.11	0.09	**0.08**
Sample size 50						
Type of Response	Residual Variance	Lin. Reg. on 665 nm	Elastic Net Reg.	FLRM Basis	FPCR BIC	FPLSR BIC
$\widetilde{Y}_{exp}$	σ_1^2	**0.96**	1.00	1.02	1.02	0.99
	σ_2^2	**0.28**	0.40	0.32	0.35	0.34
	σ_3^2	**0.09**	0.12	0.10	0.11	0.11
$\widetilde{Y}_{elas}$	σ_1^2	1.07	**1.00**	**1.00**	1.02	1.02
	σ_2^2	1.11	0.32	0.26	**0.25**	**0.25**
	σ_3^2	1.22	0.11	0.10	**0.08**	**0.08**
$\widetilde{Y}_{Bs}$	σ_1^2	1.09	0.98	**0.96**	0.98	0.99
	σ_2^2	1.15	0.29	**0.17**	0.19	**0.17**
	σ_3^2	1.16	0.12	**0.05**	**0.05**	**0.05**
$\widetilde{Y}_{FPC\ 8}$	σ_1^2	1.07	0.92	**0.89**	0.94	0.92
	σ_2^2	1.11	0.20	0.34	**0.12**	0.13
	σ_3^2	1.14	0.13	0.28	**0.03**	**0.03**
$\widetilde{Y}_{FPC\ 2}$	σ_1^2	1.15	0.99	**0.97**	0.98	1.00
	σ_2^2	1.36	0.32	0.18	0.18	**0.17**
	σ_3^2	1.37	0.18	0.06	**0.05**	**0.05**
Sample size 100						
Type of Response	Residual Variance	Lin. Reg. on 665 nm	Elastic Net Reg.	FLRM Basis	FPCR BIC	FPLSR BIC
$\widetilde{Y}_{exp}$	σ_1^2	**0.92**	0.99	0.95	0.95	0.94
	σ_2^2	**0.27**	0.33	0.28	0.30	0.30
	σ_3^2	**0.08**	0.10	0.09	0.09	0.09
$\widetilde{Y}_{elas}$	σ_1^2	1.01	0.97	**0.91**	0.94	0.93
	σ_2^2	0.96	0.24	0.23	**0.21**	0.22
	σ_3^2	0.95	0.08	0.08	**0.06**	**0.06**
$\widetilde{Y}_{Bs}$	σ_1^2	1.01	0.95	**0.86**	0.89	0.88
	σ_2^2	0.97	0.21	**0.15**	0.16	**0.15**
	σ_3^2	0.97	0.08	**0.04**	**0.04**	0.05
$\widetilde{Y}_{FPC\ 8}$	σ_1^2	1.01	0.85	**0.81**	0.82	**0.81**
	σ_2^2	0.99	0.17	0.28	**0.09**	**0.09**
	σ_3^2	0.99	0.10	0.23	**0.02**	**0.02**
$\widetilde{Y}_{FPC\ 2}$	σ_1^2	1.04	0.95	**0.86**	0.88	0.90
	σ_2^2	1.09	0.23	**0.15**	**0.15**	**0.15**
	σ_3^2	1.10	0.12	0.05	**0.04**	**0.04**

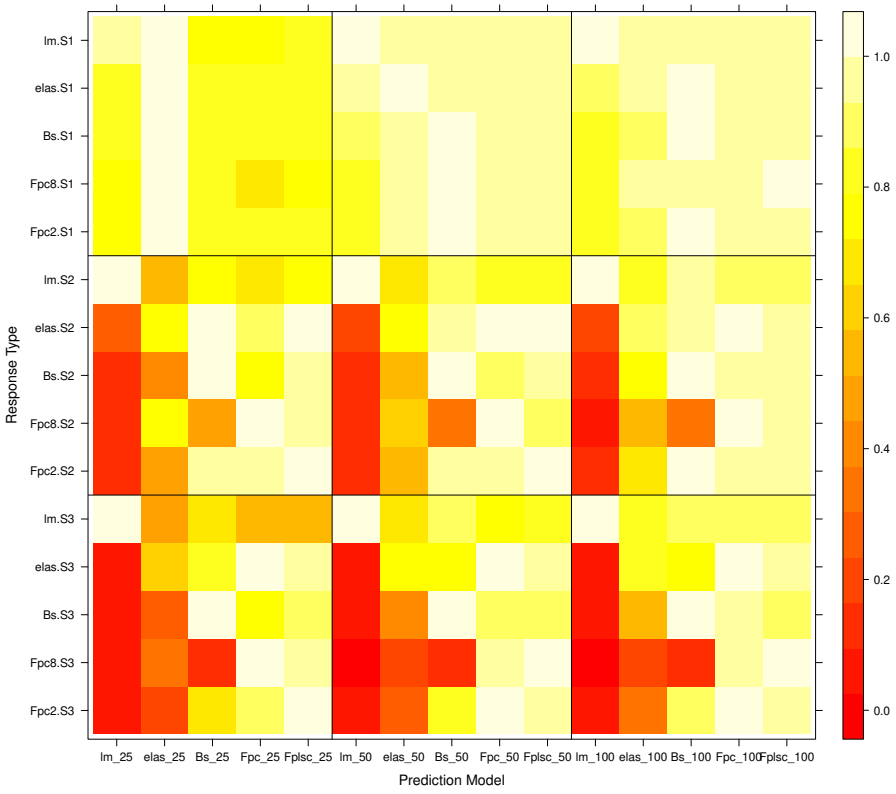


Fig. 10 Performances of the estimation models as the sample size increases and the response type changes

In general, the mean MSPE values are found to be higher for the linear regression than for the other models, and higher for Elastic net than for functional models (this is clearer for smaller residual variances). Considering all types of response and the residual variance, FLRMs based on dimension reduction methods (FPCR BIC and FPLSR BIC) give the best predictions among all the models.

In summary, when the sample size is small and the residual variance is high, the results obtained from all the models are comparable, with certain advantage for Elastic net. However, for large samples with low residual variance it is recommended to use functional models (specially those based on dimension reduction methods) instead of non functional approaches.

Table 3 and Figure 10 summarize the average simulation results, but they do not inform about the variability of MSPE values among the $P = 500$

simulated data sets. In order to gain an insight into this variability, Figure 11 shows the box plots of the $P = 500$ MSPE values (in logarithms) for the 5 estimating models in the 15 scenarios corresponding to sample size $ss = 25$. These graphics support what we have said before, when analyzing Table 3 and Figure 10 (and the same is true for the box plots corresponding to sample sizes $ss = 50$ or 100, not included here): the method that corresponds to the model the data is simulated from performs best; for the highest residual variance, Elastic net gives the best results (lowest and most concentrated MSPE values); for lower residual variances, functional regression models are preferred (lower and more concentrated MSPE values) to Elastic net (with the exception of data generated from model FPC 8), whereas simple linear model performs well only when responses are generated with a simple linear model.

There are several ways to extend the simulation study (to be considered in further studies). First, it is possible to use the coordinates of the chosen pixels in a determined route in analogy with the routes followed by the cruises *Reserva* and *Fluctuaciones*. Secondly, a greater number of satellite images can be used when generating the response. Moreover, matching can be conducted on the basis of the exact pixels instead of considering 2×2 box area, as in the study by Acar-Denizli et al (2017).

5 Conclusion

In this study, the use of functional linear models is proposed as an alternative to classical regression models to make predictions from RS data in oceanography and water quality management. In this sense, various statistical models are applied to predicting TSS content in the coastal region of Guadalquivir estuary and a comparison is made between the performance of these models. The results indicate that the functional models predict the TSS parameter better than other classical approaches with lower prediction errors. This finding is supported by the simulation study. According to the results of the simulation, as the sample size increases and the error variance decreases, the predictive performance of FLRMs generally gets better than that of the linear regression models. In concordance with the results obtained in Table 1, FLRMs based on dimension reduction methods yield better predictions than other models in simulation. In particular, in the case of large sample size and low residual variance, the use of functional prediction models is highly recommended.

Two of the limitations of this study are the low number of in-situ observations and the few available wavelengths. Unfortunately, because of the sensors that were used over that period, it is not possible to increase the number of wavelengths. Recently, new sensors that record data in a wider spectrum have been employed. Moreover, new in-situ samples are currently being collected. We hope that in the near future the matching between both sources of information can be successfully carried out. In fact, the use of functional models on hyperspectral data has recently been gaining importance. Ferraty et al (2017) predict some environmental related parameters from hyperspec-

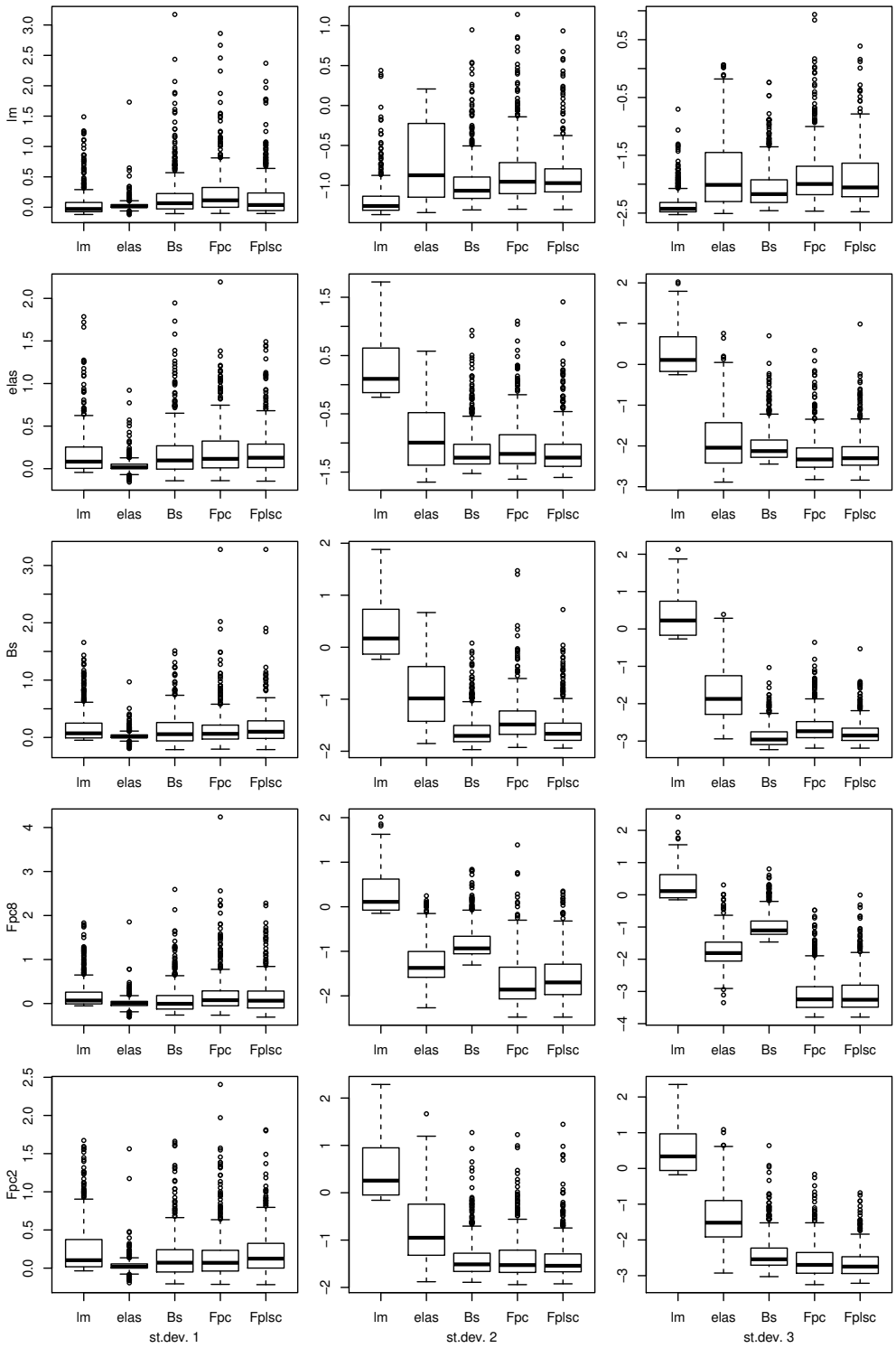


Fig. 11 Simulation results for sample size $ss = 25$: Box plots of the $P = 500$ MSPE values (in logarithms) for the 5 generating models (rows), the 3 different standard deviations (columns) and the 5 estimating models (each corresponding to a box plot at the panels).

tral remote sensing data by using nonparametric FLRMs. It is foreseen that in the future FLRMs will gain more importance in the analysis of sensor data in oceanography and water quality management as well as providing more efficient prediction models.

There is a need to further assess the potential of FDA on the validation and calibration of local or regional empirical models, which is the common practice in coastal and estuarine waters. Therefore, this study is a first approximation to applying this methodology in a significant spot for the challenging coastal water management, and can be further extended to other regions, satellites and geophysical parameter in order to test its suitability.

References

Acar-Denizli N, Delicado P, Başarır G, Caballero I (2017) Functional linear regression models for scalar responses on remote sensing data: an application to oceanography. In: Functional Statistics and Related Fields, Springer, pp 15–21

Aguilera AM, Escabias M, Preda C, Saporta G (2010) Using basis expansions for estimating functional PLS regression: Applications with chemometric data. Chemometrics and Intelligent Laboratory Systems 104(2):289–305

Bernardello R, Serrano E, Coma R, Ribes M, Bahamon N (2016) A comparison of remote-sensing sst and in situ seawater temperature in near-shore habitats in the western mediterranean sea. Marine Ecology Progress Series 559:21–34

Besse PC, Cardot H, Faivre R, Goulard M (2005) Statistical modelling of functional data. Applied Stochastic Models in Business and Industry 21(2):165–173

Binding C, Bowers D, Mitchelson-Jacob E (2003) An algorithm for the retrieval of suspended sediment concentrations in the irish sea from seawifs ocean colour satellite imagery. International Journal of Remote Sensing 24(19):3791–3806

Binding C, Bowers D, Mitchelson-Jacob E (2005) Estimating suspended sediment concentrations from ocean colour measurements in moderately turbid waters; the impact of variable particle scattering properties. Remote sensing of Environment 94(3):373–383

Björn-Helge M, Wehrens R (2007) The pls package: Principal Component and Partial Least Squares Regression in R. Journal of Statistical Software, Articles 18(2):1–23, DOI 10.18637/jss.v018.i02, URL https://www.jstatsoft.org/v018/i02

Caballero I, Navarro G (2016) Análisis multisensor para el estudio de los patrones de turbidez en el estuario del guadalquivir. Revista de teledetección: Revista de la Asociación Española de Teledetección 46:1–17

Caballero I, Morris E, Prieto L, Navarro G (2014a) The influence of the guadalquivir river on the spatio-temporal variability of suspended solids and chlorophyll in the eastern gulf of cadiz. Mediterranean Marine Science 15(4):721–738

Caballero I, Morris EP, Ruiz J, Navarro G (2014b) Assessment of suspended solids in the guadalquivir estuary using new deimos-1 medium spatial resolution imagery. Remote Sensing of Environment 146:148–158

Cardot H, Ferraty F, Sarda P (1999) Functional linear model. Statistics & Probability Letters 45(1):11–22

Cardot H, Faivre R, Goulard M (2003) Functional approaches for predicting land use with the temporal evolution of coarse resolution remote sensing data. Journal of Applied Statistics 30(10):1185–1199

Chen X, Han X, Feng L (2015) Towards a practical remote-sensing model of suspended sediment concentrations in turbid waters using meris measurements. International Journal of Remote Sensing 36(15):3875–3889

Clarke E, Speirs D, Heath M, Wood S, Gurney W, Holmes S (2006) Calibrating remotely sensed chlorophyll-a data by using penalized regression splines. Journal of the Royal Statistical Society: Series C (Applied Statistics) 55(3):331–353

Delaigle A, Hall P, et al (2012) Methodology and theory for partial least squares applied to functional data. The Annals of Statistics 40(1):322–352

Everson R, Cornillon P, Sirovich L, Webber A (1997) An empirical eigenfunction analysis of sea surface temperatures in the western north atlantic. Journal of Physical Oceanography 27(3):468–479

Faivre R, Fischer A (1997) Predicting crop reflectances using satellite data observing mixed pixels. Journal of Agricultural, Biological, and Environmental Statistics 2(1):87–107

Febrero-Bande M, Oviedo de la Fuente M (2012) Statistical computing in functional data analysis: The R package fda.usc. Journal of Statistical Software 51(4):1–28, URL http://www.jstatsoft.org/v51/i04/

Febrero-Bande M, Galeano P, González-Manteiga W (2015) Functional principal component regression and functional partial least-squares regression: An overview and a comparative study. International Statistical Review 0(0):1–23, DOI 10.1111/insr.12116, URL http://dx.doi.org/10.1111/insr.12116

Ferraty F, Zullo A, Fauvel M (2017) Nonparametric regression on contaminated functional predictor with application to hyperspectral data. Econometrics and Statistics (in press)

Fettweis MP, Nechad B (2011) Evaluation of in situ and remote sensing sampling methods for spm concentrations, belgian continental shelf (southern north sea). Ocean Dynamics 61(2-3):157–171

Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software 33(1):1–22, URL http://www.jstatsoft.org/v33/i01/

Gitelson AA, Peng Y, Arkebauer TJ, Suyker AE (2015) Productivity, absorbed photosynthetically active radiation, and light use efficiency in crops: Implications for remote sensing of crop primary production. Journal of plant physiology 177:100–109

Goldsmith J, Bobb J, Crainiceanu C, Caffo B, Reich D (2011) Penalized functional regression. Journal of Computational and Graphical Statistics 20:830–851

Gong M, Miller C, Scott E (2015) Functional pca for remotely sensed lake surface water temperature data. Procedia Environmental Sciences 26:127–130

Hastie T, Tibshirani R, Wainwright M (2015) Statistical learning with sparsity. CRC press

Horváth L, Kokoszka P (2012) Inference for functional data with applications, vol 200. Springer Science & Business Media

James GM (2002) Generalized linear models with functional predictors. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64(3):411–432

Kokoszka P, Reimherr M (2017) Introduction to Functional Data Analysis. CRC Press

Lahet F, Ouillon S, Forget P (2001) Colour classification of coastal waters of the Ebro river plume from spectral reflectances. International journal of remote sensing 22(9):1639–1664

Le C, Hu C, Cannizzaro J, English D, Muller-Karger F, Lee Z (2013) Evaluation of chlorophyll-a remote sensing algorithms for an optically complex estuary. Remote Sensing of Environment 129:75–89

Liu C, Ray S, Hooker G, Friedl M (2012) Functional factor analysis for periodic remote sensing data. The Annals of Applied Statistics 6(2):601–624

Marx BD, Eilers PHC (1999) Generalized linear regression on sampled signals and curves: A p-spline approach. Technometrics 41(1):1–13

MATLAB (2011) version 7.10.0 (R2010a). The MathWorks Inc., Natick, Massachusetts

Morris JS (2015) Functional regression. Annual Review of Statistics and Its Application 2:321–359

Navarro G, Ruiz J (2006) Spatial and temporal variability of phytoplankton in the gulf of cádiz through remote sensing images. Deep Sea Research Part II: topical studies in oceanography 53(11):1241–1260

Navarro G, Huertas IE, Costas E, Flecha S, Díez-Minguito M, Caballero I, López-Rodas V, Prieto L, Ruiz J (2012) Use of a real-time remote monitoring network (rtrm) to characterize the guadalquivir estuary (spain). Sensors 12(2):1398–1421

Nechad B, Ruddick K, Park Y (2010) Calibration and validation of a generic multisensor algorithm for mapping of total suspended matter in turbid waters. Remote Sensing of Environment 114(4):854–866

Nezlin NP, DiGiacomo PM (2005) Satellite ocean color observations of stormwater runoff plumes along the San Pedro Shelf (southern California) during 1997–2003. Continental Shelf Research 25(14):1692–1711

Preda C, Saporta G (2005) PLS regression on a stochastic process. Computational Statistics & Data Analysis 48(1):149–158

R Core Team (2017) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL https://www.R-project.org

Ramsay J, Silverman B (2005) Functional Data Analysis. Springer, USA

Ramsay JO, Wickham H, Graves S, Hooker G (2017) fda: Functional Data Analysis. R package version 2.4.7. URL http://CRAN.R-project.org/package=fda

Rawat J, Kumar M (2015) Monitoring land use/cover change using remote sensing and gis techniques: A case study of hawalbagh block, district almora, uttarakhand, india. The Egyptian Journal of Remote Sensing and Space Science 18(1):77–84

Reiss PT, Goldsmith J, Shang HL, Ogden RT (2017) Methods for scalar-on-function regression. International Statistical Review 85(2):228–249

Ruiz J, Polo MJ, Díez-Minguito M, Navarro G, Morris EP, Huertas E, Caballero I, Contreras E, Losada MA (2014) The guadalquivir estuary: a hot spot for environmental and human conflicts. In: Environmental Management and Governance, Springer, pp 199–232

Wang JL, Chiou JM, Mueller HG (2016) Functional data analysis. The Annual Review of Statistics and Its Application 3(2):257–295

Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society Series B 67:301–320