
Bertrand Saulquin, Ronan Fablet, Pierre Ailliot, Grégoire Mercier, David Doxaran, Antoine Mangin, and Odile Hembise Fanton d’Andon

Abstract—The spatial and temporal coverage of satellites provides data that are particularly well suited for the analysis and characterization of space–time-varying geophysical relationships. The latent-class models aim here to identify time-varying regimes within a dataset. This is of particular interest for geophysical processes driven by the seasonal variability. As a case example, we study the daily concentration of mineral suspended particulate matters estimated from satellite-derived datasets, in coastal waters adjacent to the French Gironde river mouth. We forecast this high-resolution dataset using environmental data (wave height, wind strength and direction, tides, and river outflow) and four latent-regime models: homogeneous and nonhomogeneous Markov-switching models, with and without an autoregressive term (i.e., the mineral suspended matter concentration observed the day before). Using a validation dataset, significant improvements are observed with the multiregime models compared to a classical multiregression and a state-of-the-art nonlinear model [support vector regression (SVR) model]. The best results are reported for a mixture of three regimes for the autoregressive model using nonhomogeneous transitions. With the autoregressive models, we obtain at day+1 for the mixture model forecasting performances of 93% of the explained variance, compared to 83% for a standard linear model and 85% using an SVR. These improvements are more important for the nonautoregressive models. These results stress the potential of the identification of geophysical regimes to improve the forecasting. We also show that nonhomogeneous transition probabilities and estimated autoregressive terms improve forecasting performances when observation data is lacking for short-time period of 1–15 days.

I. INTRODUCTION

The forecasting of a geophysical variable using statistical models is an alternative to model-based approaches which typically involve complex simulation and/or assimilation [1], [2]. For instance, coupled hydrodynamic and sediment transport models can be used to estimate the concentration of suspended particulate matters within the water column [3], while statistical approaches may use available satellite and model data to predict the same variable [4]. Many statistical approaches have been proposed and evaluated to forecast or infer a studied variable from predictors. Among them, linear multivariate regression [5] and nonlinear (polynomial) multivariate regression [6] are the most known. Numerous specific models dedicated to time-series analysis such as autoregressive moving average (ARMA) and autoregressive integrated moving average (ARIMA) models [7] have also been developed initially to address financial time series. These latest, which aim at studying the behavior of a time series without considering forcing factors, have also been applied to geophysical time series [8]. Nonlinear regressions, based on supervised learning strategies, such as neural networks [9] and support vector regressions (SVRs) [10] may provide relevant alternatives to estimate a variable from predictors. In the context of geophysical studies, they may nevertheless suffer from two major drawbacks. First, though relevant regression performances may be reported, these models may not be physically interpretative and may be very sensitive to the training dataset. Second, multiregime dynamics, often exhibited by geophysical processes driven by the seasonality [11], cannot be addressed by such nonlinear models, contrary to latent-regime models as demonstrated in our study.

We propose here to characterize time-varying relationships between a variable and its forcing parameters using latent-regime models, and hence optimize forecasting results. As an illustration, we address the concentration of inorganic suspended particle matters (SPIM), estimated from satellite data using a regional algorithm [12], [13], and observed in the
mouth of the Gironde estuary. In this area, sediments are mainly exported from the Gironde estuary [13], [14] and SPIM concentration clearly depends on the local physical forcing: swell, tide, wind, and river discharge. A minimum of energy has to be brought by waves and tides to re-suspend cohesive sediments accumulated at the bottom. Conversely, when sediments have been re-suspended in the water column by wave influence, their settling velocity depends on their size and density [15] and physico-chemical properties [16]. This example stresses that the relationships between the studied variable (SPIM) and the causing factors evolve in space and time and potentially requires advanced statistical methods to identify the underlying geophysical regimes.

From a methodological point of view, “latent regime regressions” also referred as “clusterwise regressions” [17], [18] are particularly appealing to identify such nonlinear and multi-regime patterns within a dataset. Each regime is associated with a linear regression and the overall nonlinear patterns are thus estimated as a combination of the different linear contributions. Regarding the temporal dynamics of these regimes, we here consider Markovian processes [19], which state the transitions in time between two regimes. The standard hidden Markov model (HMM) and nonhomogeneous hidden Markov model (NHHMM) are evaluated [19]. The inclusion of an autoregressive term (HMM-AR) and (NHHMM-AR) is also discussed. This aspect is motivated by the strong autocorrelation level depicted by geophysical time series [20]. When the observation of the previous day (referred as \( t-1 \)) is available, it is obvious, considering the strong natural autocorrelation of geophysical data that the forecast at time \( t \) should take into account the observation at time \( t-1 \). Conversely, for specific applications, or if the observations are not available during long periods (such as winter storms, or after a sensor failure), one may need to estimate the variable without using the observations of the previous days. We discuss here the choice between autoregressive or not autoregressive models for long lack of observation period using forecasting results from \( t+1 \) to \( t+15 \).

Model parameter estimation is carried out from a dataset composed of 5862 time series of 1096 points in the mouth of the Gironde estuary in the \([3^\circ\text{W}–1^\circ\text{E}; 45–46.5^\circ\text{N}]\) area during the period 2007–2009. Validation is performed on the same area for using the data for the year 2010. We used EOFs to reduce the dimension of the space–time observations. This is a usual approach in spatio-temporal statistics [21], [22] although alternatives may be considered such as linear discriminant analysis [23], and we could also introduce a latent variable to describe the regime at each location and interact with the regimes at other locations. Nevertheless, such models are known to be very difficult to fit on the data and remain a research challenge for statisticians. We infer our mixture model using the expansion coefficients (ECs) of the first four modes of the EOF which explain 99% of the total variance. The variables used as predictors for the SPIM ECs are the wave height (WH) issued from a numerical model [24], the wind fields optimally interpolated from satellite observations [25], the tide coefficient [26], and the Gironde fresh water discharge (sum of the Garonne and Dordogne rivers contributions).

II. METHODS

A. Markov Switching Models

We address here the study of a two-dimensional scalar geophysical time series \( Y \). In an HMM [19], one states two different processes, the observed process \( Y \) and a hidden process \( Z \). The observed process (here the turbidity) is assumed to be temporally dependent of the hidden process. At a given time \( t \), the hidden variable \( Z_t \) is a discrete value which states the regime in play at time \( t \) characterized by a latent linear regression model with coefficient \( B_k \) between the variable \( Y_t \) and the predictor \( X_t \) [17]. The conditional likelihood of the observation \( Y_t \) given predictor \( X_t \) and regime \( Z_t \) is thus expressed as [17]

\[
P(Y_t|X_t, Z_t = k) \sim N(X_t B_k, \sigma_k)
\]

where \( N \) represents the multivariate Gaussian probability density function with mean \( X_t B_k \) and variance \( \sigma_k^2 \).

The hidden process \( Z_t \) is modeled as a first-order Markov chain characterized by its transition probability matrix between \( Z_{t-1} \) and \( Z_t \) [19]. In the simplest case (HMM), one assumes homogeneous transitions, i.e., time and context-independent transition matrix. The NHHMM allows the transition matrix between the hidden regimes to depend on a set of observed covariates \( S_t \). Hughes and Guttorp [28], [29] highlighted the added value of the NHHMM to characterize the links between the large-scale atmospheric measures and the small-scale spatially discontinuous precipitation field. In the NHHMM settings, the probability transition matrix is now time-dependent and conditioned by the covariates \( S_t \)

\[
P(Z_t = k | Z_{t-1} = l, S_t) = \frac{P(S_t | Z_{t-1} = l) \cdot P(Z_t = k | Z_{t-1} = l)}{\sum_{k,l} P(S_t | Z_{t-1} = l) \cdot P(Z_t = k | Z_{t-1} = l)}
\]

The nonhomogeneous transition is derived from the likelihood of the covariate \( S_t \) given transition from \( Z_{t-1} \) to \( Z_t \). We suppose that the probability density function of the covariates during this change of regime follows a normal distribution

\[
P(S_t | Z_{t-1} = l) = N(\mu_{i,k}, \Sigma_{i,k})
\]
wind, wave height (WH), tide coefficient and river discharge, and eventual lagged values of $Y_t$ (refers as autoregressive terms), and defines a general family of model which encompasses the most usual ones with regime switching. When no covariate is considered, i.e., $Z_t$ only depends $Z_{t-1}$, and $Y_t$ only depends on $(Y_{t-5}, \ldots, Y_{t-1})$ and $Z_{t-1}$, we retrieve the usual Markov switching autoregressive (MS-AR) model. If we further assume that $s = 0$ (without autoregressive component $Y_{t-1}$), then we obtain an HMM. When $Z_t$ does not dependent on $Z_{t-1}$, but only on $S_t$, it comprises the threshold autoregressive (TAR) model which is another important family of regime-switching models in the literature. HMMs, MS-AR, and TAR have been used in many fields of applications including geosciences [27].

A key property of the considered Markov switching models is the factorized expression of the joint likelihood of the observed and hidden processes. As an example, for an homogeneous Markov chain (HMM, [19]), characterized by its transition matrix $P(Z_t|Z_{t-1})$, its initial law $P(Z_0 = k|X_0, Y_0)$ and the conditional probability $P(Y_t|X_t, Z_t)$ (1), the conditional probability $P(Y_0^s, Z_0^s|X_0^s)$ at any time $s$ factorizes as

$$P(Y_0^s, Z_0^s|X_0^s) = \prod_{t=0}^{s} P(Y_t|X_t, Z_t) \cdot P(Z_t|Z_{t-1}) \cdot P(Z_0 = k|X_0, Y_0)$$

where $Y_0^s, X_0^s$, and $Z_0^s$, respectively) are Y values from $t = 0$ to $s$.

### B. Estimation of the Model Parameters

The considered models involve two categories of parameters to be estimated: those of the observation model $\theta_0$, namely regression coefficient $B_k$ and standard deviation $\sigma_k$ for each regime and $\theta_s$ the parameters of the hidden Markov-switching process. For homogeneous models, $\theta_0$ is the transition matrix $P(Z_t|Z_{t-1})$, while for nonhomogeneous models, $\theta_s$ is the ($\mu_{k,l}, \sum_{k} I_{kl}$) parameters. Given observed $Y$ and $X$ series, we proceed to the estimation of model parameters according to a classical maximum likelihood (ML) criterion using an iterative expectation maximization (EM) procedure [30].

$$L(\theta) = \max \left( P\left( Y_0^T | X_0^T, S_0^T, \theta \right) \right)$$ (5)

where $T$ is the time-index of the last observation (i.e., all the series are observed) and $\theta = \{\theta_0, \theta_s\}$ is the set of parameters to be estimated. The EM procedure proceeds iteratively as follows: for a given initialization for the parameters, the procedure iterates estimation steps (E-step) of the posterior regime likelihood $P(Z_t = k|Y_0^T, X_0^T)$ of the regimes and the maximization step (M-step) to update the parameters given these posteriors [19]. The algorithm iterates until convergence between steps $n$ and $n+1$, i.e., $|L(\hat{\theta}(n)) - L(\hat{\theta}(n+1))| < 10^{-3}$. The posterior likelihoods are estimated in the E-step using the classical forward–backward recursions given series $X$ and $Y$ and current parameter estimate $\hat{\theta}(n)$ [19], [31], [32]. The M-step re-estimates the parameters $\theta(n+1)$. From the conditional dependencies involved in the considered models (cf. Fig. 1), it is often possible to break the optimization problem into several lower dimensional optimization problems which are much quicker to solve [32]. More precisely, it is possible to separate the observation model parameters $\theta_0$ and the hidden Markov-switching process $\theta_s$, in each regime, shown at the bottom of the page (7) and (8).

$$\hat{\theta}_s^{(n+1)} = \arg \max_{\theta_s} \left( \sum_{t} \log \left( P\left( Z_t = k | Z_{t-1} = l, S_t, \theta_s(n) \right) \right) \right) \times P\left( Z_t = k, Z_{t-1} = l | Y_0^T, S_0^T, \theta(n) \right)$$ (6)

### C. Forecasting Application

The considered multiregime regression models are applied to the short-term forecasting of series $Y$. More precisely, at a given time $t$, we aim at predicting variable $Y$ at time $t + dt$. We assume that prediction variables $X$ and covariates $S$, typically numerical simulations, are available up to time $t + dt$, whereas the variable $Y$ is only known up to time $t$. Thus, the forecast at time $t + dt$, denoted by $Y_{t+dt}$, is given by

$$\hat{Y}_{t+dt} = \arg \min_{B_k} \left( \sum_{t} \left( P\left( Z_t = k | Y_0^T, X_0^T, \theta(n) \right) (Y_t - B_k X_t)^2 \right) \right)$$ (7)

$$\hat{\sigma}_{k(n+1)} = \sum_{t} \left( P\left( Z_t = k | Y_0^T, X_0^T, \theta(n) \right) (Y_t - B_k X_t)^2 \right)$$ (8)
the conditional expectation of variable \(Y_{t+dt}\) given observations series up to time \(t\) and predictor series up to time \(t+dt\)

\[
\hat{Y}_{t+dt} = E(Y_{t+dt}|Y_0, X_{t+dt}^{t+dt}).
\]

(9)

For HMM, it resorts to

\[
\hat{Y}_{t+dt} = \sum_k P(Z_{t+dt} = k|Y_t^t, X_{0+dt}^t) \cdot X_{t+dt} B_k.
\]

(10a)

For NHHMM, it resorts to

\[
\hat{Y}_{t+dt} = \sum_k P(Z_{t+dt} = k|Y_t^t, X_{0+dt}^t) [\alpha \hat{Y}_{t+dt-1} + X_{t+dt} B_k]
\]

(11a)

and for the NHHMM-AR

\[
\hat{Y}_{t+dt} = \sum_k P(Z_{t+dt} = k|Y_t^t, X_{0+dt}^t, S_{0+dt}^t) \times [\alpha \hat{Y}_{t+dt-1} + X_{t+dt} B_k].
\]

(11b)

It might be noted that these predictions actually account for the uncertainties in the determination of the underlying regimes. Contrary to deterministic methods, confidence interval and uncertainties on \(\hat{Y}_{t+dt}\) can be derived which is a key issue for modeling considerations [33].

D. Model Performance Estimation

A key issue in practice, which has received lots of attention in the last few years, is the problem of model selection which aims at finding the “optimal” number of predictors and covariates [31]. Hereafter, we have chosen to use both the Bayes Information Criterion (BIC) and the explained variance (EVAR) as first guides. BIC index generally permits to select parsimonious models which fit the data well [34]. It is defined as

\[
\text{BIC} = -2 \log(L) + p \log(S)
\]

(12)

where \(L\) is the likelihood estimated, \(p\) is the number of parameters, and \(S\) is the number of observations. We also use the classical EVAR to characterize the model relevance

\[
\text{EVAR} = 1 - \var(\hat{Y}_{t+1} - Y_{t+1}) / \var(Y_{t+1}).
\]

(13)

BIC and EVAR are partially linked [34]. BIC tends to penalize complex models, whereas EVAR criterion only qualifies the result and may lead to the over-parameterization of a model that typically leads to errors when other datasets are tested using the same parameterization. Therefore, we consider both BIC and EVAR to assess the model performance.

III. DATA

A. Studied Variable

Nonalgal SPM concentrations (SPIM) are estimated using an analytical algorithm [12] defined as the difference between total SPM and phytoplankton biomass, the latter derived from the chlorophyll-a (chl-a, [36]). It incorporates mainly mineral SPM and smaller amounts of organic SPM not related to living phytoplankton. This method to derive nonalgal SPM from remote-sensing reflectance is based on the inversion of a simplified equation of radiative transfer, assuming that chlorophyll concentration is known. This merged dataset consists of fields of nonalgal surface SPM concentrations, derived from the Sea-viewing Wide Field-of-view Sensor (SeaWiFS), the Moderate Resolution Imaging Spectroradiometer (MODIS) and the Medium Resolution Imaging Spectrometer (MERIS) sensors, provided by the Ocean Color TAC (Thematic Application Facility) of MyOcean, and interpolated with a kriging method for the period 2007–2009 over the Gironde mouth river from 45°-1°E; 45–46.5°N [35]. Finally, 5682 continuous time series of 1096 days compose our initial dataset of mineral suspended particulate matters concentration. We first account for the space–time variability of the dataset, previously detrended and centered for each time series, using an EOF decomposition expressed here using the matrix form [21], [38]

\[
\text{Cov}(\text{SPIM}) = \text{UV}^T
\]

(14)

where \(U\) is a here 5682 × 5682 matrix containing the spatial modes (eigenvectors) of the covariance decomposition (ordered by percentage of EVAR). Associated with each spatial mode \(k\), its EC (also referred in the literature as principal component) is the time evolution of the \(k\)th mode

\[
\text{EC}_{\text{SPIM}}k = \text{SPIM}_{t} \ast U_k.
\]

(15)

Fig. 2 shows the four first spatial modes of the EOF decomposition. Fig. 3 depicts the four associated time series \(\text{EC}_{\text{SPIM}}k=1,4\). The first mode [Fig. 2(a)] comprises 85% of the total variance. It clearly addresses the seasonal cycle as shown in Fig. 3(a), where the switch between winter (high values of \(\text{EC}_{\text{SPIM}}1\) correspond here to high values of SPIM observed in winter) and summer periods is clearly visible. The variability around the seasonal mean is captured by the other modes [Figs. 2(c)–(e) and 3(c)–(e)]. Mode 2 refers to the inter-annual and the intra-seasonal variability in the shoreward gradient and represents 7% of the total variance. Mode 3 addresses some north–south gradients and represents 4% of the total variance and mode 4 is clearly influenced by the Gironde river [Fig. 2(d)], which brings sediments during water outflow, and represent 3% of the variance. By construction, EOF decomposition imposes the orthogonality [21] of the spatial modes (Fig. 2).

The reconstruction of the SPIM variable from the estimated ECs is performed as

\[
\text{SPIM}_t = \sum_k \text{EC}_{\text{SPIM}}k \ast U_k.
\]

(16)

The total EVAR using the four first modes is shown Fig. 4(b). On average, the EVAR represents 99% of the
Fig. 2. Spatial patterns of the EOF modes of the SPIM observed from satellite from 2007 to 2009 in the Gironde mouth river, presented as homogeneous correlation maps. From left to right and top to bottom, the first four EOF modes account, respectively, for 85%, 7%, 4%, and 3% of the total variance.

B. Predictors and Covariates

The predictors X are the variables used in the estimation of Y and Z any time (10a) and (10b), (11a) and (11b). We used here WH daily means of the Wave Watch 3 model (WW3; [24], [37]) provided by the IOWAAGA and PREVIMER programs, eastward and northward winds interpolated from QuickSCAT and ASCAT observations in conjunction with ECMWF forecasting [25], provided by Ifremer, tide index (SHOM, 2000) at Bordeaux and the flow measurement of river la Gironde. Similar to the SPIM data, all the data were log transformed. For the wind data which is signed, the transformed log variable was signed negatively a posteriori to the log transformation. The WH first mode of the EOF decomposition explained 98% of the total variance, 93% for the northward wind (WND1), and 96% for the eastward wind (WND2).

The choice of the predictors is performed here as follows. We first select as predictors the variable showing a significant correlation with the studied EC. Given these predictor datasets, we tested all the possible configurations and chose the predictors which provide the lower BIC and the greatest EVAR on the training dataset. Covariates are the normalized predictors used in the estimation of the EC but considered at \(t - 2\). In the same way, this time-lag has been estimated as the optimal lag using BIC and EVAR results on the training dataset.

IV. RESULTS

We summarize in Table I the prediction performance for the first four ECs of the SPIM issued from four models: HMM, NHHMM, HMM-AR, and NHHMM-AR. The number of considered modes for the mixture varies from 1 to 3. The one-mode models refer to a simple multivariate regression analysis. For each configuration we provide the BIC and EVAR_train on the training dataset (2007–2009) and EVAR_valid on the validation dataset (2010). Note that the selection of the predictors and
Fig. 3. EOF decomposition of the SPIM observed from satellite from 2007 to 2009 in the Gironde mouth river: from left to right and top to bottom, the ECs (EC_SPIM\textsubscript{1−4}) associated with the first four EOF modes depicted in Fig. 2, i.e., the time evolution of the spatial modes.

Fig. 4. (a) Initial SPIM variance. (b) Percentage of variance explained by the four first modes of the EOF decomposition of the suspended particulate matters.

resulting covariates is achieved as a prior step as described in Section II-D.

The first mode of the EOF decomposition explains 85% of the total variance. EC\_WH\textsubscript{1} and EC\_WND\textsubscript{2\textsubscript{1}} (the EC of the first EOF of the swell and westward winds, respectively) are identified as being the relevant predictors (cf. Section IV-B). This mode captures the mean seasonal variability of the SPIM, which is mainly driven by WH of the North Atlantic storms.
and at a second order by the eastward winds. For EC_SPIM1, when no autocorrelation term is used, the best fit is obtained for a 3-regime NHHMM model (BIC = 9873, EVAR_train = 90%, and EVAR_valid = 85%). When a first-order autocorrelation term is added, the 3-regime HMM-AR and NHHMM-AR models show the best results: BIC = 7997 (8000, respectively), EVAR_train = 98% and EVAR_valid = 97%. This stresses that when observations are available first-order autoregressive term (AR1) should be included to enhance the performances. The lag-1 autocorrelation observed value is 0.85 for EC_SPIM1, underlying the strong link between two successive observations. Compared to a single autoregressive model, i.e., without predictors and covariates and regime discretization (not shown), the gain value provided by X and S on EVAR_train is 98% and EVAR_valid is 97%. This stresses that when observations are available first-order autoregressive term (AR1) should be included to enhance the performances.

The second mode of the EOF decomposition of the SPIM variability explains 7% of the total variance. The selected predictors are the first mode of the eastward wind, the tide, and the river flow. The variability captured by EC_SPIM2 relates to the local eastward wind, which is not captured by the WH model, and the very coastal variability introduced by the tide and the river outflow. For the nonAR models the selected model was the three-regime NHHMM. It is interesting to note in this case that EVAR_valid increased from 50% to 73% between the HMM and the NHHMM, highlighting the contribution of the nonhomogeneous transition model.

The third mode of the EOF decomposition of the SPIM variability explains 4% of the total variance. It captures some inter-annual and intra-seasonal variability of the latitudinal gradient of the SPIM. The selected predictors are EC_WH1, EC_WND1 (northward), and the tide. Once again, three-regime NHHMM and HMM-AR provide the best results.

Regarding the fourth mode of the EOF decomposition of the SPIM variability, which accounts 3% of the total variance, EC_WH1, EC_WND2, the tide, and the river flow are selected as contributive predictors. We reconstruct 75% of EC_SPIM3 variance of the validation dataset using a three-regime NHHMM and 92% using the three-regime HMM-AR and NHHMM-AR.
The nuances of gray in the background highlight the temporal distribution of the regimes.

Fig. 5. Estimation of the EC_SPIM1 (in black) using EC_WH1, EC_WND2, and a single regression (green) and a 3-regime NHHMM (red). The nuances of gray in the background highlight the temporal distribution of the regimes.

Globally, we observe from Table III that three regimes are needed for all models to forecast optimally the EOF ECs at $t+1$. NHHMM outperforms HMM for forecasting results at $t+1$. The inclusion of an AR term clearly improves the results. NHHMM-AR and HMM-AR shows similar results at $t+1$. We will see (Section IV-B) that the added value of nonhomogeneous transitions for AR-model clearly appears for the long-term forecasting.

A. Example With the Estimation of EC_SPIM1

We report in Fig. 5 the temporal evolution of the three regimes of the NHHMM for EC_SPIM1 estimated at $t+1$. In Table II are shown the corresponding coefficients for each predictor and the intercept. The first regime (light gray, $Z_t = 1$), characterized by high-SPIM levels (intercept of 65), is referred as a “winter regime.” The “winter regime” also strongly relates to the WH (regression coefficient of 0.6). Dark gray periods ($Z_t = 3$) are identified as a “transition regime,” and medium gray ($Z_t = 2$) identified as the “summer regime.” From the “winter” to the “summer” regime, coefficient for WH decreases from 0.6 to 0.12. In summer, the energy brought by waves is “winter” to the “summer” regime, coefficient for WH decreases from 0.6 to 0.12. In summer, the energy brought by waves is

$$\hat{Y}_{t+dt} = \sum_{l=1}^{5} \sum_{i=1}^{2} \gamma_{l,i} X_{t+dt}^{l,i}$$

The probability of switching from regime 3 to 1 increases with WH and eastward winds normalized covariates.

B. Forecasting of the SPIM on the 2010 Validation Dataset

We forecast SPIM fields from the reconstructed ECs (10a) and (10b), (11a) and (11b), (16). Fig. 7(a) and (b) compares the EVAR of the initial field (SPIM) using the three-regime NHHMM and NHHMM-AR models. On average, we were able to predict at $t+1$ 80% of the variance using the NHHMM [Fig. 7(a)] and 93% using the NHHMM-AR. The spatial distribution of the error is not homogeneous. Fig. 7(a) shows that EVAR_valid value is of 90% in the northern part with nevertheless poorer results in the south. Fig. 7(b) shows that the AR1 component of the model increases EVAR for the whole area.

We also consider the results of a standard multiregression analysis. If only one regime is considered NHHMM and HMM resort to a standard multivariate regression and NHHMM-AR and HMM-AR to a standard multivariate regression including an AR coefficient, the transition probability being equal to 1. Fig. 7(c) shows the obtained results with the standard multivariate regression and Fig. 7(d) the standard multivariate regression including an AR1. From Fig. 7(c) and 7(a), the gain in EVAR is in mean about 150% [from in mean 32% Fig. 7(c) to 80% Fig. 7(a)], while for the AR models, the gain is about from 11% [from in mean 83% Fig. 7(d) to 93% Fig. 7(b)].

Regarding the model forecasting performances, we report the short-term forecast results at different time steps using the 2010 validation dataset. Table III synthesizes the EVAR statistics using three regimes and the four tested models for the forecasting at $t+1$, $t+5$, and $t+15$.

The long-term forecasting results are globally better with the NHHMM-AR. At $t+15$ using the NHHMM, we are able to forecast 74% of the variance for 2010, compared to 40% for the HMM. In this case the time-varying regime transition probability $P(Z_t = k)\mid Z_{t-1} = l, S_{t+dt}$ helps in the estimation of $\hat{Y}_{t+dt}$ (covariates $S_{t+dt}^l$ are model outputs for which the short term predictions are assumed to be available). For autoregressive models, at $dt = 5$, we were able to forecast 82% of the 2010 SPIM variance with the NHHMM-AR compared to 77% with the NHHMM. In this case $Y_{t+dt}^l$ estimated using $X_{t+dt}^{l,i-1}$, $Y_{t}$, and the inhomogeneous transition properties, help to estimate $\hat{Y}_{t+dt}$. At $t+15$, NHHMM and NHHMM-AR show equivalent results underlying the maximal time-step for which the autoregressive term brings significant information.

An SVR model was also evaluated to evaluate the performances of a nonlinear model on the studied dataset. To perform the comparison, we train the SVR model (http://www.csie.ntu.edu.tw) for each EC using the same training dataset (2007–2009) and performed forecasting using the same validation dataset (2010). We used the setting as following: model epsilon-SVR ($s = 3$), linear or polynomial kernel ($t = 0$ or 1) and the same inputs (predictors, covariates) for
TABLE II
Estimated Regression Parameters for Each of the Three Regimes of the NHHMM and the HMM-AR for the First EOF EC of the SPIM: Regression Parameters Involve an Intercept and the Regression Coefficients of the Significant Forcing Parameters, i.e., the WH and the Eastward Wind Velocity

<table>
<thead>
<tr>
<th>Regime</th>
<th>EC_WH1</th>
<th>EC_WND2</th>
<th>Intercept</th>
<th>EC_WND1</th>
<th>Intercept</th>
<th>AR(1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NHHMM (1)</td>
<td>0.6037</td>
<td>-0.0632</td>
<td>65.0672</td>
<td>0.0910</td>
<td>-61.6442</td>
<td></td>
</tr>
<tr>
<td>NHHMM (2)</td>
<td>0.0006</td>
<td>0.0100</td>
<td>-24.6578</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NHHMM (3)</td>
<td>0.1210</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HMM-AR (1)</td>
<td>0.2383</td>
<td>-0.0033</td>
<td>4.4694</td>
<td>0.0050</td>
<td>-3.9531</td>
<td>0.86</td>
</tr>
<tr>
<td>HMM-AR (2)</td>
<td>-0.0050</td>
<td>0.0168</td>
<td></td>
<td></td>
<td></td>
<td>0.90</td>
</tr>
<tr>
<td>HMM-AR (3)</td>
<td>0.0354</td>
<td>0.0035</td>
<td>0.6079</td>
<td></td>
<td></td>
<td>0.90</td>
</tr>
</tbody>
</table>

Fig. 6. Nonhomogeneous transition probability between the "transition regime" (medium gray; Fig. 5) and the "winter regime" (light gray; Fig. 5) as a function of the normalized WH EC_WH and eastward wind EC_WND2.

V. DISCUSSION

We investigated the relevance of four regime-switching latent regression models, namely HMM, NHHMM, HMM-AR, and NHHMM-AR to characterize time-varying linear relationships between the high-resolution SPIM data (inorganic suspended matter concentration) and forcing conditions i.e., the WH, the northward and eastward winds, the tide and the river flow. SPIM data were issued from MODIS, SeaWiFS, and MERIS satellite data. As a case study, we considered a coastal area in the mouth of the Gironde estuary in the [3°W–1°E; 45°–46.5°N] area.

Model calibration was carried out using 2007-to-2009 datasets, whereas 2010 dataset was used as an independent validation dataset of 1-to-15-day forecasting performances.

An optimal number of three regimes were identified to capture the different geophysical dynamics and optimize forecasting performances. Autoregressive and nonhomogeneous models showed better performances. For the 2010 validation dataset, 1-day NHHMM forecasts explained 80% of the variance, whereas an NHHMM-AR model explained 93% of the variance. The natural high autocorrelation level observed in geophysical time-series makes the observation of the previous day an important predictor to consider. The comparison to other models clearly stresses the relevance of the proposed latent-class models, whereas the EVAR of 1-day forecast for a standard multivariate linear regression was of 32% (83%, respectively) without (respectively with) an first-order autoregressive term, the nonlinear SVR model reached, respectively, 40% and 80% of EVAR. The gain of 100% between the NHHMM and the SVR model (16% between the NHHMM-AR and the SVR including an AR1 term, respectively) pointed out the relevance of the multiregime approaches. The SVR model failed here in retrieving regime shifts.

As illustrated for the first SPIM EOF component (Fig. 5), the proposed multiregime setting identified three different relationships between the observed turbidity, the WH and the wind. We did not drive the model to account for seasonal regimes but these regimes exhibited seasonally discriminated patterns, with two leading factors: the mean SPIM level (intercept) and the eastward WH. The later was interpreted as a feature of the minimum of energy to be brought by the eastward swell to re-suspend the sediments. This is regarded as a key characteristic of the latent-regime model compared to other nonlinear regression models, such as neural networks [39] or SVR [10], which can hardly be interpreted in general.

Regarding long-term forecast performance, at t + 15 best results obtained were of 74% of EVAR for the NHHMM and 75% for the NHHMM-AR. For short period, typically from 1 to 15 days, when the observed Y is not available, NHHMM-AR provided the best results. In this case, the available predictors X^{t+dt}, covariates S^{t+dt}, and the estimated Y^{t+dt} help in the estimation of Y^{t+dt}. At t + 15, NHHMM and NHHMM-AR showed similar results. Hence, a 15-day period could be regarded as the maximal time interval, beyond which one may only consider covariates. It may also be noted that, in case of
sensor failure and/or long missing data periods (e.g., series of successive storms in the case-study region), though no satellite observations might be available, one could still reach relevant SPIM prediction accounting in average for about 75% of the variance.

In the future, we will address the forecasting of the chl-a using satellite-derived observations such as the photosynthetic available radiation, the temperature, the suspended particulate matters (as index of available nutrients), and light attenuation [40]. In this more complicated case, second-order relationships between the variable and its predictors have to be evaluated, the chl-a dynamic being not anymore a passive result of the forcing conditions, as expected with the SPIM, but having its proper characteristics depending on each phytoplankton specie. Extensions of the considered latent regime setting to other inverse problems in satellite sensing data analysis are also under investigation, such as latent regime inversion procedures for satellite-derived chl-a concentration to account for different

---

**Fig. 7.** EVAR for the 2010 validation dataset reconstructed using (a) the 3-regime NHHMM and (b) NHHMM-AR, compared to a standard multivariate regression (c) without AR and (d) including an AR1.

**TABLE III**

<table>
<thead>
<tr>
<th>dt (days)</th>
<th>HMM</th>
<th>HMM-AR</th>
<th>NHMM</th>
<th>NHHMM-AR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>73</td>
<td>93</td>
<td>80</td>
<td>93</td>
</tr>
<tr>
<td>5</td>
<td>63</td>
<td>80</td>
<td>77</td>
<td>82</td>
</tr>
<tr>
<td>15</td>
<td>40</td>
<td>70</td>
<td>74</td>
<td>75</td>
</tr>
</tbody>
</table>

EVAR (13) for the forecast at t + 1, t + 5, and t + 15 of the 2010 validation dataset. For each model, three latent-regimes are used.
water types (turbid or not turbid) and/or the presence of specific phytoplankton species.

ACKNOWLEDGMENT

The authors would like to thank A. Sottolichio from the Université of Bordeaux 1, for the provision of the in situ Gironde flow measurement and the MCGS (Marine Collaborative Ground Segment; http://www.mccgs.fr) project which aim at making the most of ESA Sentinels satellites potential for users driven services based on high level products. MCGS addresses the need of the European Space Agency to build up data processing centers in conjunction with the Copernicus Program for the provision of services to local and national, public and private European institutions or entities involved in marine activities.

REFERENCES

[34] H. S. Bhat and N. Kumar, On the Derivation of the Bayesian Information Criterion. School of Natural Sciences, Univ. California, 2010.

Bertrand Saulquin is pursuing the Ph.D. degree in signal processing at the Telecom Bretagne School, Brest, France, in 2014. He is a Research Engineer with the ACRI-ST, Sophia-Antipolis, France. He has been working on satellite data analysis and geostatistics for the past 7 years in the frame of several ESA and EU projects (MARCOAST; Aquamar, ECOOP, MyOcean I & II, and OSS2015). His research interests include multivariate statistical analysis and data mining for the analysis of geophysical processes such as the inversion of water reflectances and the forecasting of the turbidity in complex coastal waters.
Ronan Fablet received the Graduate degree from the Ecole Nationale Supérieure de l’Aéronautique et de l’Espace (SUPAERO), Toulouse, France, in 1997, and the Ph.D. degree in signal processing and telecommunications from the University of Rennes, Rennes, France, in 2001.

In 2002, he was an INRIA Postdoctoral Fellow with Brown University, Providence, RI, USA. From 2003 to 2007, he held a Full-Time Research Position in the field of signal and image processing applied to fisheries science with Ifremer Brest, Plouzané, France. In 2008, he joined the Signal and Communications Department, Télécom Bretagne, Brest, France, as an Associate Professor, and has been holding a Professor position since 2012.

Grégoire Mercier was born in France in the 1971. He received the degree in engineering from the Institut National des Télécommunications, Evry, France in 1993, and the Ph.D. degree from the University of Rennes, Rennes, France, in 1999.

Since 1999, he has been with the École Nationale Supérieure des Télécommunications de Bretagne, Brest, France, where he is currently an Associate Professor with the Image and Information Processing Department. His research interests include remote-sensing image compression and segmentation, especially in hyperspectral and synthetic aperture radar.

David Doxaran received the Ph.D. degree in ocean colour remote sensing and numerical transport modelling of sedimentary fluxes in the Gironde estuary from the Université Bordeaux 1, Talence, France, in 2002.

He joined the Laboratoire d’Océanographie de Villefranche as permanent CNRS Researcher at the end of 2007. His research project combines the use of field bio-optical measurements and ocean color satellite data with sediment transport modeling to provide new estimates of suspended particles fluxes exported by rivers into the coastal ocean, track and understand the dynamics of terrestrial particles in river plumes.

Antoine Mangin received the Ph.D. degree in physics from the Ingénieur de l’École Nationale Supérieure des Arts et Métiers, in 1990.

Since 2000, he has been a Scientific Manager with the ACRI-ST Company, Sophia-Antipolis, France. He has been working with a number of research projects mixing physical oceanography and ocean color. He is a Director of French Scientific Cluster for Ocean Colour (GIS COOC) and member of national and international working group for ocean radiometry. His research interests include multisensors data merging (mainly European upcoming Sentinel-2 and Sentinel-3).

Grégoire Mercier was born in France in the 1971. He received the degree in engineering from the Institut National des Télécommunications, Evry, France in 1993, and the Ph.D. degree from the University of Rennes, Rennes, France, in 1999.

Since 1999, he has been with the École Nationale Supérieure des Télécommunications de Bretagne, Brest, France, where he is currently an Associate Professor with the Image and Information Processing Department. His research interests include remote-sensing image compression and segmentation, especially in hyperspectral and synthetic aperture radar.

David Doxaran received the Ph.D. degree in ocean colour remote sensing and numerical transport modelling of sedimentary fluxes in the Gironde estuary from the Université Bordeaux 1, Talence, France, in 2002.

He joined the Laboratoire d’Océanographie de Villefranche as permanent CNRS Researcher at the end of 2007. His research project combines the use of field bio-optical measurements and ocean color satellite data with sediment transport modeling to provide new estimates of suspended particles fluxes exported by rivers into the coastal ocean, track and understand the dynamics of terrestrial particles in river plumes.

Antoine Mangin received the Ph.D. degree in physics from the Ingénieur de l’École Nationale Supérieure des Arts et Métiers, in 1990.

Since 2000, he has been a Scientific Manager with the ACRI-ST Company, Sophia-Antipolis, France. He has been working with a number of research projects mixing physical oceanography and ocean color. He is a Director of French Scientific Cluster for Ocean Colour (GIS COOC) and member of national and international working group for ocean radiometry. His research interests include multisensors data merging (mainly European upcoming Sentinel-2 and Sentinel-3).

Odile Hembise Fanton d’Andon received the Ph.D. degree in fluid mechanics in 1997.

He is the General Manager with the ACRI-ST, Sophia-Antipolis, France, a French SME. She has set up and led a number of European research projects bridging ocean color products to marine and coastal applications. Among others, she has been leading the GlobColour project for ESA, as well as ODESA software development, evolution and dissemination for the exploitation of MERIS mission. Currently, she is the Coordinator of the FP7 project OSS2015 and actively involved in the preparation of upcoming ESA/Copernicus Sentinel-3 mission.

Pierre Ailliot received the Ph.D. degree in mathematics and applications from the University of Rennes, Rennes, France, in 2004.

Since 2007, he has been an Associate Professor with the Mathematics Department, University of Brest, Brest, France. His research interests include environmental statistics with a focus on stochastic weather generators, regime switching models, and extreme value analysis.