Remote Sensing of Environment 163 (2015) 229–239

Contents lists available at ScienceDirect

Remote Sensing of Environment

journal homepage: www.elsevier.com/locate/rse

Retrieving the evolution of vertical profiles of Chlorophyll-a from satellite observations using Hidden Markov Models and Self-Organizing Topological Maps

A.A. Charantonis [a], F. Badran [b], S. Thiria [a]

[a] Laboratoire d'Océanographie et du Climat — Expérimentation et Approches Numériques, Université Pierre et Marie Curie, Tour 45, 5ème étage 4, place, Jussieu, 75005 Paris, France
[b] Laboratoire CEDRIC, Conservatoire National des Arts et Métiers, 292, rue Saint Martin, 75003 Paris, France

ARTICLE INFO

Article history:
Received 14 March 2013
Received in revised form 18 March 2015
Accepted 22 March 2015
Available online 21 April 2015

Keywords:
Inversion of satellite data
Evolution of vertical profiles of Chlorophyll-a
Hidden Markov Models
Self-Organizing Topological Maps

ABSTRACT

We present a statistical method, denoted PROFHMM, to infer the evolution of the vertical profiles of oceanic biogeophysical variables from sea-surface data. This method makes use of discrete Hidden Markov Models whose states are defined through Self-Organizing Topological Maps. The Self-Organizing Topological Maps are used to provide the states of the Hidden Markov Model, as well as improve its parameters. After introducing the general principles of PROFHMM, we present the results obtained in a case study in which the evolution of the vertical profiles of Chlorophyll-a was inverted from sea-surface data. We applied PROFHMM for the reconstruction of the evolution of the vertical distribution of Chlorophyll-a at BATS, by training it on the numerical outputs of the NEMO-PISCES model, and reproducing the evolution of this model by using a sequence satellite observations. We obtained a root mean square error of 0.0399 ng/l for the validation year 2008.

1. Introduction

The current density of satellite observations has allowed a quasi-continuous observation of the global ocean surface. The two-dimensional images provided by this coverage contain information on physical or biogeophysical variables but not on their vertical profiles (Dinnat, Boutin, Caudal, Etcheto, & Waldteufel, 2002; Feldman et al., 1989; Krishna Rao, Smith, & Koffler, 1972). Inverting the sea-surface data remotely sensed by satellite to obtain the evolution of the vertical profiles of biogeophysical variables usually requires a numerical modeling of their relations. Such models are, however, often faced with problems of non-linearity, complexity and incomplete knowledge of the mechanisms that govern these profiles.

In the present paper we attempt to infer the temporal evolution of vertical profiles of Chlorophyll-a from observed sea-surface satellite images. The biogeochemical activity of the oceans and the carbon cycle are two parts of a complex feedback system. A change in climate and an increase in the amount of available carbon can affect the oceanic primary production. In return, a change in the biogeochemical activity affects the climate, by modifying the albedo and carbon fixation rates, as well as the atmospheric and oceanic carbon concentrations (Hays, Richardson, & Robinson, 2005). It is therefore important to be able to determine the oceanic primary production, of which Chlorophyll-a is a proxy.

In recent years, many algorithms have been developed to infer the Chlorophyll-a concentration in ocean surface layers through satellite imaging (Brajard, Cédric, Cyril, & Thiria, 2006; Richardson, Risien, & Shillington, 2003). Studies using Self-Organizing Topological Maps have demonstrated that the vertical Chlorophyll-a distribution is related to various types of sea-surface data (Demarcq, Richardson, & Field, 2008; Richardson et al., 2002; Richardson et al., 2003; Silulwane, Richardson, Shillington, & Mitchell-Innes, 2001; Uitz, Claustre, Morel, & Hooker, 2006), and even been used to relate the evolution of the characteristic the local coastal upwelling and downwelling current patterns on the West Florida Shelf with the local wind forcing (Liu & Weisberg, 2005). Organizing and maintaining missions to perform in situ measurements of the vertical distribution of Chlorophyll-a can be cost-prohibitive, which can explain the spatial and temporal sparsity of such data. There are, however, large databases of extrapolated vertical profiles of Chlorophyll-a calculated by biogeochemical models such as the MERCATOR-VERT and NEMO-PISCES models (Gehlen et al., 2010; Gurvan & the NEMO team, 2012). It is generally accepted that these models are able to reproduce the dynamic processes that govern the evolution of the vertical profiles of Chlorophyll-a. There exist a large number of such databases, containing time-series spanning decades. Sea-surface measurements, on the other hand, are almost continuously accessible from satellite imagery.

These facts have lead us to investigate the possibility of inferring the evolution of the vertical profiles of Chlorophyll-a solely from sea-surface data. Here below we consider that the Chlorophyll-a vertical profiles correspond to a set of unobservable, so-called 'hidden states', and

E-mail address: anastase-alexandre.charantonis@locean-ipsl.upmc.fr (A.A. Charantonis).

http://dx.doi.org/10.1016/j.rse.2015.03.019
0034-4257/

that the multidimensional sea-surface data are "emitted" from these. Formulated this way, this problem is similar conceptually to the statistical modeling method known as Hidden Markov Models (HMM) (Juang, 2003). A HMM is fully defined by its states, its topology and its related probabilities.

We present a method that inverts sea-surface data in order to infer the vertical profiles of Chlorophyll-a by formalizing it as a classical HMM. Due to the high dimensions of the vectors involved in the reconstruction of the Chlorophyll-a profiles, the size of the database used to estimate the probabilities required to define the HMM is never sufficient. To circumvent this problem, and obtain a robust determination of the parameters of the HMM, we applied a Self-Organizing Topological Map (SOM) (Kohonen, 1990). SOMs are an established neuronal classification methods commonly used in geosciences (Liu & Weisberg, 2011) that produce topologically ordered classifications of data sets. They have been used in other cases to generate the topology of an HMM (Jaziri, Lebbah, Bennani, & Chenot, 2011). In PROFHMM, the topological aspect of the classifications provided is further used to improve the quality of the estimated HMM probabilities, given the otherwise insufficient available data.

The method we have developed, which is a combination of HMM and SOM, was named PROFHMM, for PROFile reconstruction through HMM. Taking as input a sequence of sea-surface variables, it inverts them and retrieves the most probable evolution of the hidden vertical distribution of Chlorophyll-a in the oceans.

PROFHMM is a very cost-efficient inversion method, since, once the training of the method is over, the computations needed are minimal and could be run on most personal computers.

After introducing the general principles of the method, we present the results obtained in a case study. The vertical profiles of Chlorophyll-a were reconstructed based on sea-surface data at the site of the Bermuda Atlantic Time Series (BATS) (32°N–64°W) of the JGOFS (Joint Global Ocean Flux Study) campaign (Doneya, Kleypasa, Sarmiento, & Falkowski, 2002), first using simulated data for both the vertical distribution profiles and the observation vectors, and then by using simulated data in conjunction with MODIS satellite data for the observation vectors obtained from NASA.

2. Data

The study of the oceanic primary production is linked with phytoplankton biomass and therefore with the Chlorophyll-a distribution. One cannot determine the vertical distribution of Chlorophyll-a without first understanding the parameters that influence the development of phytoplankton. It is generally accepted (Miller, 2003) that phytoplankton growth mainly depends on five variables: available shortwave radiation, available nutrients, herbivores and biology, water temperature, and water turbidity.

These variables cannot be easily monitored through a direct approach. Satellite imaging, however, can give us proxy information, which can be used in an empirical approach to determine the vertical profiles of Chlorophyll-a. Specifically in this study we used:

- sea-surface Chlorophyll-a concentration (SCHL)
- sea-surface temperature (SST)
- sea-surface height (SSH)
- downwelling shortwave radiation (SR)
- wind-speed (WS)

to infer the Chlorophyll-a vertical profiles, which are variables that more easily accessible than Chlorophyll-a.

There is currently a lack of in situ measurements of Chlorophyll-a with consistent revisit rates. The JGOFS Bermuda Atlantic Time Series (BATS), for example, makes in situ measurements of vertical profiles of Chlorophyll-a every month for three consecutive days, to an approximate depth of 200 m, which does not provide a good temporal sampling for inferring the dynamic processes that govern the development of phytoplankton in the area (Bates, Michaels, & Knap, 1996). This led us to use simulated Chlorophyll-a data, produced by the NEMO oceanic circulation model coupled to the PISCES biogeochemical model for testing the validity of our approach (Fig. 1).

The data we wanted to classify as hidden states were the NEMO-PISCES output data vectors at BATS, which contained the average vertical Chlorophyll-a distribution at 17 depth levels (from 5 to 217 m) and the temperature distribution at nine of these depth levels. The inclusion of the temperature in the hidden Chlorophyll-a data vectors permits a better representation of the physical conditions that constrain the Chlorophyll-a development. These vertical distribution profiles (of dimension $26 = 17 + 9$) were five-day averages of the model, spanning the period from 1992 to 2008 and located in a $2° \times 2°$ square centered on BATS. In order to have additional vectors from which to infer the possible states of vertical distribution of Chlorophyll-a, the vectors at the neighboring grid points of the model were also taken into account. This gave us 9 grid points, with 73 five-day averaged profiles per year, during 17 years for a total of 11,169 profiles for the determination of the hidden states. This is done under the assumption that all the Chlorophyll-a profiles in a $6° \times 6°$ area are similar and can be used for the determination of the SOMs. Doing so, we increase by a factor of nine the size of the training data set. We will refer to these vectors as hidden vertical distribution (HVD) vectors. A single such vector, taken at time t, is referred to as $\boldsymbol{x}^t_{hid}$.

In addition to the vertical profiles, the NEMO-PISCES model provides the associated observable surface variable values, such as SSH, SST, SCHL, WS and SR. These values constitute the components of the model surface (MS) vectors.

We also obtained MODIS observations from 24 June 2002 up to 14 June 2011. These observations consist in SST and SCHL values, which are averaged over the valid pixels of the studied zone, and over the 5-day periods corresponding to those of the model outputs. Empty segments in the temporal series were estimated through linear interpolation. So we run a more realistic experiment by using two-dimensional remote-sensing (RS) vectors containing the SST and SCHL provided by MODIS.

Consecutive time sequences of the vectors presented are used to generate time-series, denoted S_{hid}, for the sequences of HVD vectors, and S_{obs}, for the sequences of MS or RS vectors. A single vector, taken at time t, is denoted $\boldsymbol{x}^t_{hid}$ ($\in R^{26}$) for the HVD vectors, as $\boldsymbol{x}^t_{obs}$ ($\in R^2$ *or* R^5) for an observable vector, or more specifically $\boldsymbol{x}^t_{MS}$ or $\boldsymbol{x}^t_{RS}$, for a MS or RS vector.

Fig. 1. The location of BATS.

3. Method

In this section we discuss the statistical models known as Hidden Markov Models and Self-Organizing Topological Maps, and their combination in our method. The flowchart linking the different components of the PROFHMM method is shown in Fig. 2.

In Section 3.1 we first introduce the HMM methodology, along with a brief description of the SOM that is necessary in order to understand the HMM architecture. The SOM and their role in PROFHMM are further detailed in Section 3.2.

3.1. Hidden Markov Models

A Markov model is a state-based stochastic model that assumes the first-order Markovian property, meaning that the probability of the observed system being in a given state depends solely on the previous state in which was the model, that is: $P(X_t|X_1X_2 \ldots X_{t-1}) = P(X_t|X_{t-1})$, where X_t is the state of the model at time *t*.

Expanding this principle, a Hidden Markov Model (HMM) is a stochastic model with two sequences: one sequence of hidden states that follows the first-order Markovian property, and one sequence of observable states which have a statistical link with the hidden states. The Viterbi algorithm (Viterbi, 1967) then finds the most likely sequence of hidden states, given a sequence of concurrent observations. There exist alternatives for reconstructing sequences (Hagenauer & Hoeher, 1989; Viterbi, 1998), but we focus on the Viterbi algorithm in this paper. PROFHMM could however, be applied with a different reconstructing algorithm.

The HMM are algorithms which allow us to infer the most likely sequence of some discrete, hidden states, given a series of concurrent observations. To do so, we have to discretize all the available HVD vectors into a set of finite states, in such a way that each state corresponds to a referent vector of the vertical distribution of Chlorophyll-a. The discretization needs to be very fine in order to obtain the most accurate reconstruction possible. It should therefore permit a partition of the set of HVD vectors into subsets, each one having a very small standard deviation.

The discrete, hidden states need to be connected among themselves through a probability matrix. The probabilities in this matrix correspond to a statistical learning of the dynamic processes governing the temporal transitions between the hidden states. These are referred to as transition probabilities.

The observations need to be consistent in nature and need to be linked, through a probability density function or matrix, with the hidden states. This density function, or the probability matrix elements, corresponds to the existing links between the observations and the hidden states and these elements are referred to as emission probabilities.

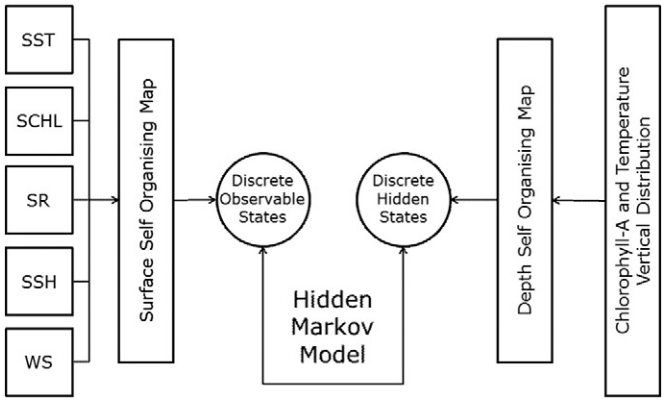

Fig. 2. Flowchart showing the links between the different modules of PROFHMM and the observable and hidden vectors.

Therefore, when inverting sea-surface data to retrieve the vertical profiles of Chlorophyll-a by using HMM, we are faced with three major problems: the determination of the hidden states, their transition probabilities and the emission probabilities. The continuous, multidimensional nature of variables included in the sea-surface observations, in conjunction with the fine discretization of the hidden states, imposes a number of constraints on the determination of the conditional probability density functions. However, the calculation of the emission probability becomes easier when the observations are clustered into observable states.

To solve these problems we propose the use of the Self-Organizing Topological Maps (SOMs) in order to discretize both the sea-surface observations and the hidden vertical Chlorophyll-a distribution data sets.

SOMs are unsupervised classification algorithms that cluster data into discrete classes. These classes are arranged on a map in such a way that classes that are close on the map represent situations that are close in the original data space. The general concepts of SOMs are further detailed in Section 3.2. In the present study, the SOM classification is applied twice, once to the satellite sea-surface data and once to the vertical profiles of Chlorophyll-a connected to these images. This generates two topological map classifications. The classes generated on the map containing the vertical profiles correspond to the hidden states of the HMM, whereas the classes of the map containing the sea-surface observations correspond to the observable states. The choice of performing separate classifications of the observable and hidden data sets allows the HMM to represent the dynamics of the unobservable states independently from the observations.

However, when discretizing the observation space into states, only a limited region of the observation space may correspond to a particular hidden vertical distribution state. This, in turn, makes for less reliable probability density functions associated with these states which are less present during the training phase. This could be resolved by limiting the number of states, giving each state enough data to properly estimate its probability density functions, but this would severely decrease the detail of the inverted profiles, since the reconstruction of the vertical distributions depends on the hidden states.

To overcome this problem in our model, we make the assumption that each hidden state generates observation vectors according to a Gaussian probability density mixture. Two such distinct mixtures of Gaussians, issued from two hidden states close in terms of profiles of HVD vectors, need to determine proximate regions in the observation space. This allows us to improve the estimation of the probabilistic properties of one state by using the properties of its "neighboring" states. That implies that our method requires that its states are topologically arranged.

3.2. Self-Organizing Topological Maps

Self-Organizing Topological Maps (SOMs) are clustering methods based on neural networks. They provide a clustering of a learning data set into a reduced number of subsets, called classes, which share some statistical characteristics.

Each class is represented by its referent vector $r(i)$ which approaches the mean value of the elements belonging to it. The topological aspect of the maps can be justified by considering the map as an undirected graph on a two-dimensional lattice whose vertices are the N classes. This graph structure permits the definition of a discrete distance $d(C(i), C(j))$ between two classes C(i) and C(j), defined as the length of the shortest path between C(i) and C(j) on the map. The nature of the SOM training algorithm forces a topological ordering upon the map and, therefore, any neighboring classes C(i) and C(j) on the map ($d(C(i), C(j)) = 1$) have referent vectors $r(i)$ and $r(j)$ that are close in the Euclidean sense in the data space.

Let us consider a vector *x* that is of the same dimensions and nature as the data used to generate the topological map; we can find the index of the class to which it is classified by choosing: index $= (\|x - r(i)\|)$,

therefore assigning it to the class whose referent is closest to it in the Euclidean sense (Fig. 3). A classified vector x will be represented by its class index, C(index).

In PROFHMM, we train two SOMs, the first one containing the observations, denoted $sMap_{obs}$, and the second one containing the distributions of the hidden states, denoted $sMap_{hid}$. The number of classes in $sMap_{obs}$ and $sMap_{hid}$ correspond to the number of states of the HMM, N_{obs} and N_{hid}. Their respective classes and referent vectors will be denoted C_{hid} and C_{obs}, r_{hid} and r_{obs}. We used the algorithms provided by the MATLAB somtoolbox, specifically the functions som_make, som_batchtrain, som_bmus, in order to train our maps and classify our data. The determination of many of the SOM parameters is automatically calculated by this toolbox, by applying the default parameters.

3.3. The PROFHMM method

In PROFHMM, we use the correspondence between the hidden vertical Chlorophyll-a distribution profiles, $\boldsymbol{x}^t_{hid}$, and their class indices, $C_{hid}(i^t)$, provided by the SOM, $sMap_{hid}$, for translating the time-series of vectors, $S_{hid} = \{\boldsymbol{x}^1_{hid}, \ldots, \boldsymbol{x}^T_{hid}\}$, into the corresponding times-series of class indices, $SI_{hid} = \{C_{hid}(i^1), \ldots, C_{hid}(i^T)\}$. Similarly, we translate the time-series of observations, $S_{obs} = \{\boldsymbol{x}^1_{obs}, \ldots, \boldsymbol{x}^T_{obs}\}$, into the time-series of class indices, $SI_{obs} = \{C_{obs}(i^1), \ldots, C_{obs}(i^T)\}$.

We consider two phases: training and retrieval.

During the training phase, a number of concurrent couples of sequences of indices, SI_{hid} and SI_{obs}, whose length in consecutive time-steps is denoted as L_{seq}, are used in order to estimate the elements of the Transitions matrix, $Tr_{B\text{-}W}$, and the Emissions matrix, $Em_{B\text{-}W}$. We have then at our disposal a set of sequences, $A = \{seq_i, i \in 1 \ldots N_{seq}\}$, where N_{seq} is the number of sequences. In the present paper this has not been important, since we work with a single training sequence, yet it is included because of its importance when such continuous sequences do not exist. These probabilities are estimated by using the Baum–Weltch algorithm (Baum et al., 1970), which is a particular case of a generalized expectation–maximization algorithm that takes as input all the concurrent sequences of SI_{obs} and SI_{hid} and outputs the most likely matrices to have generated them through a hidden Markov process.

$Tr_{B\text{-}W}$ contains the transition probabilities of the hidden states

$$tr_{i,j} = P\left(C_{hid}\left(i^t\right) = i \,\middle|\, C_{hid}\left(i^{t-1}\right) = j\right) \tag{1}$$

where

$$\sum_{i=1}^{N_{hid}} tr_{i,j} = 1. \tag{2}$$

$Tr_{B\text{-}W}$ corresponds, in a physical sense, to the underlying dynamics that govern the hidden states, containing the probabilities $tr_{i,j}$ of going from a hidden state j to the hidden state i at time t.

$Em_{B\text{-}W}$ contains the a posteriori probabilities of each observed state to have been emitted by a hidden state,

$$e_{i,j} = P\left(C_{obs}\left(i^t\right) \,\middle|\, C_{hid}\left(j^t\right)\right) \tag{3}$$

Fig. 3. The classification of a HDV through the Self-Organizing Topological Map.

where

$$\sum_{i=1}^{N_{hid}} e_{i,j} = 1. \tag{4}$$

In a physical sense, $Em_{B\text{-}W}$ corresponds to the link existing between the observed quantities and the dynamics of the unobserved quantities, $e_{i,j}$, presenting the probability of having a given observed state i, given the concurrent hidden state j at time t.

Another probability matrix that needs to be calculated is the initial probability matrix Π, whose components, π_i, represent the average revisit rate of each hidden state given an infinite sequence. This matrix is used during the retrieval phase in order to optimize the starting point of the reconstruction of the most likely sequence.

For the retrieval phase, HMMs use the Viterbi algorithm, which is a well known dynamic programming algorithm, for inferring the most likely sequence of indices, SI_{hid}, given the previously estimated parameters $Tr_{B\text{-}W}$, $Em_{B\text{-}W}$ and Π of the HMM and a sequence of observation indices, SI_{obs}.

3.4. Optimizing the estimation of probabilities

The method presented up to now can present some inconsistencies, if applied without taking into account the specificities of the problem, namely how the states with low frequency in the database hampers the estimation of the HMM probabilities. Therefore, once we have created the topological maps and have acquired the states of the HMM, we need to focus on some problems inherent in the hidden state reconstructions.

The Viterbi algorithm may present problems when performing a reconstruction based on probabilities estimated with a training data set that omits transitions that exist in reality. A balance needs to be found between the amount of discretization provided by the SOMs, which affects the detail of the reconstruction, and the correctness of the state selected for the reconstruction. Bigger map sizes contain more detail, but generate less reliable HMM parameters, while smaller ones have very reliable parameters but tend to reconstruct seasonal averages. The size of the maps, which define the values of N_{obs} and N_{hid}, are therefore optimized in order to get the best results for the HMM. This optimization is an iterative process of training both maps with different sizes, then running PROFHMM and performing a cross-validation using a separate validation data set (Kohavi, 1995).

Yet, even with an optimization, there will be some situations and transitions that are seldom encountered in the training data and result in null probabilities in the probability matrices, $Em_{B\text{-}W}$ and $Tr_{B\text{-}W}$ that we estimated in the first pass of the Baum–Weltch algorithm.

Due to this usual lack of sufficient data in the concerned domains, $Em_{B\text{-}W}$ and $Tr_{B\text{-}W}$ need to be adjusted. This is done by taking into account the properties of the SOM. A major characteristic of the present method is to use the topological order in order to improve the accuracy of the estimated probability matrices. SOMs allow us to modify the probabilities by allowing each state to communicate via a diminutive probability with each of its neighboring states.

This is done by considering the neighborhood matrices, NM_{obs} and NM_{hid}, of dimensions (N_{obs}, N_{obs}) and (N_{hid}, N_{hid}), where

$$NM_{SOM}(i,j) = \begin{cases} 1, & \text{if } d(C_{SOM}(i), C_{SOM}(j)) < 2 \\ 0, & \text{else} \end{cases} \tag{5}$$

with d(i,j) being the discrete distance on the map, with SOM representing alternatively either $sMap_{obs}$ or $sMap_{hid}$.

Taking into account the neighborhood states for calculating the final transition probabilities consists in increasing the probability of reaching a class j from a class i by an amount proportional to the sum of the previously calculated probabilities of reaching the neighbor classes of class j on either $sMap_{obs}$ or $sMap_{hid}$.

We therefore modified our algorithms in order to take into account multiple concurrent time sequences of indices during the training. For each of those training sequences, we apply the Baum–Weltch algorithm and get two estimated initial probability matrices, $Tr_{B\text{-}W(seq)}$, $Em_{B\text{-}W(seq)}$.

In order to favor the data observed during training, we add a weighting term, w_c, to the initial probabilities and we further multiply it by the square root of the total length of each training sequence used in the initial Baum–Weltch algorithm pass, noted L_{seq}, since this length is a measure of confidence in the correctness of the estimated parameters. This weighting term is selected at the same time with the sizes of the SOMs, through the same iterative process. The matrices obtained are increased by adding a constant to avoid null probabilities.

The final Em and Tr matrices, Em_{final} and Tr_{final}, are computed by applying for $1 \le i \le N_{obs}$ and for $1 \le j \le N_{hid}$, using all the sequences of the training data set A:

$$Em_{final}(i,j) = \sum_{N_{seq}} \left(w_c * \sqrt{L_{seq}} * Em_{B-W(\text{seq})}(i,j) + \sum_{k=1}^{N_{hid}} \left(NM_{hid}(j,k) * Em_{B-W(\text{seq})}(i,k) \right) \right) + 1, \quad (6)$$

Which is normalized to fit the constraint where

$$\sum_{i=1}^{N_{hid}} e_{i,j} = 1 \quad (7)$$

and for $1 \le i, j \le N_{hid}$, using all the sequences of the training data set A:

$$Tr_{final}(i,j) = \sum_{N_{seq}} \left(w_c * \sqrt{L_{seq}} * Tr_{B-W(\text{seq})}(i,j) + \sum_{k=1}^{N_{hid}} \left(NM_{hid}(i,k) * Tr_{B-W(\text{seq})}(i,k) \right) \right) + 1 \quad (8)$$

which is normalized to fit the constraint

$$\sum_{i=1}^{N_{hid}} tr_{i,j} = 1. \quad (9)$$

In the applications shown in Section 4, there is only one training sequence from which we obtained the transition and emission probabilities and the inclusion of the sequence length in the formula is not of importance, yet it can be an important factor for other studies. Each observation increases the emission probability of the most likely hidden state and the probability of its neighbors. In PROFHMM, we assume that when we classify a sea-surface observation on $sMap_{obs}$, it is classified correctly. However, if that is not the case, it would have been classified in one of the neighboring classes (Fig. 4). This is taken into account by using the Em_{final} and Tr_{final} matrices.

The above modifications permit the Viterbi algorithm to circumvent the problems of impossible transitions or emissions due to insufficient data in the training sequences that resulted in null probabilities in the estimated parameters.

4. Results

As noted previously, the method was tested with two configurations, once by using model forcing and outputs in order to train the HMM and reconstruct the HVD vectors, and once when using satellite observations in order to achieve the same reconstruction.

In both cases, the $sMap_{hid}$ was trained by taking into account all available profiles surrounding BATS, corresponding to 11,169 profiles (Section 2).

As noted before, we used the SOM referent vectors in order to reconstruct the evolution of the vertical distribution of Chlorophyll-a. Thus even if PROFHMM finds the "best" state given by SOM for the reconstruction, we still slightly diverge from reality. Indeed we used a vector quantization of the possible HVDs: only a discrete number of possible HVDs are available instead of the necessary continuous values required for an exact reconstruction of the hidden vectors. This imposes a limit to the quality of our reconstructions. We note as "optimum reconstruction" the best possible reconstruction we could obtain if the HMM gave us the "real" sequence of states of the $sMap_{hid}$. This optimum reconstruction therefore can be represented by a series of indices, SI_{opt}, we would have obtained by attributing to each HVD of our data base their "best" state on $sMap_{hid}$. The performance of PROFHMM is therefore bound by the reconstruction of this series of indices. We have therefore included the percentage of corresponding indices between SI_{opt} and the reconstructed series of indices, SI_{rec}, as a measure of performance of PROFHMM. It must be noted that even when the algorithm does not return exactly the same index, the reconstruction can still be valid if the two states whose indices we compared are neighbors on the $sMap_{hid}$.

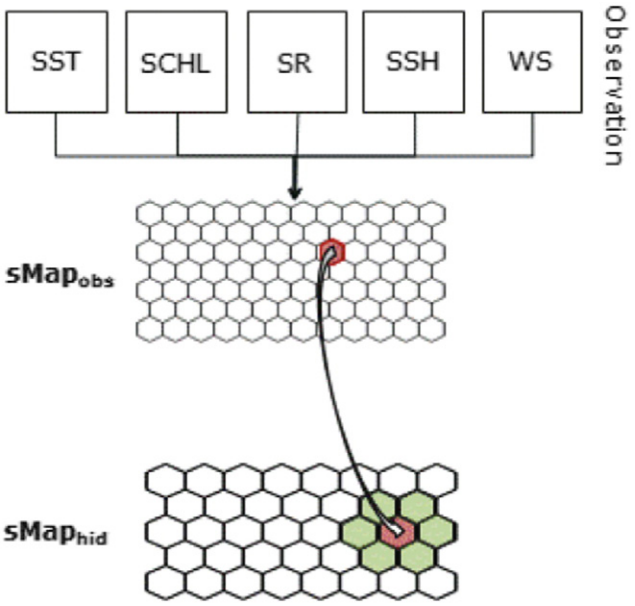

Fig. 4. The emission probability of each class $C_{obs}(i)$ of the $sMap_{obs}$ from a class $C_{hid}(j)$ of $sMap_{hid}$ takes into account the probability of being emitted by a class $C_{hid}(k)$ neighboring $C_{hid}(j)$.

The 10 first time steps of each reconstruction by PROFHMM tend to have poor performances, since the reconstruction is greatly based on the observations, and the dynamic processes of the hidden states have not been long enough to drive the system to the most likely state. The performances shown below, which include these 10 first time steps, would be improved if we ignore these time steps.

The PROFHMM reconstructions obtained also provided the evolution of the vertical profiles of temperature. The results in reconstruction of temperature are slightly worse (they present a mean RMSE of 0.276 °C in the case of the reconstruction based on satellite observations), but, as with the results in Chlorophyll-a, they still follow the general shape and intensity of the evolution of the temperature profiles. These slightly worse results were to be expected since the temperature profiles were only included to "guide" the evolution of the Chlorophyll-a profiles, improving the transition probabilities, yet were undersampled in the training, containing only nine levels of temperature when compared to the seventeen levels of Chlorophyll-a. In addition to that, the HMM and SOM parameters of the method were optimized based on the results obtained on the Chlorophyll-a reconstruction. For all these reasons, we chose to only focus on the results regarding the reconstruction of the evolution of the vertical distribution of Chlorophyll-a.

4.1. Reconstruction from model surface vectors

We chose 2-D hexagonal SOMs. In the NEMO-PISCES application, all the available HVD and MS vectors have been used during the learning phase of the SOM. The optimal architecture, after the cross-validation

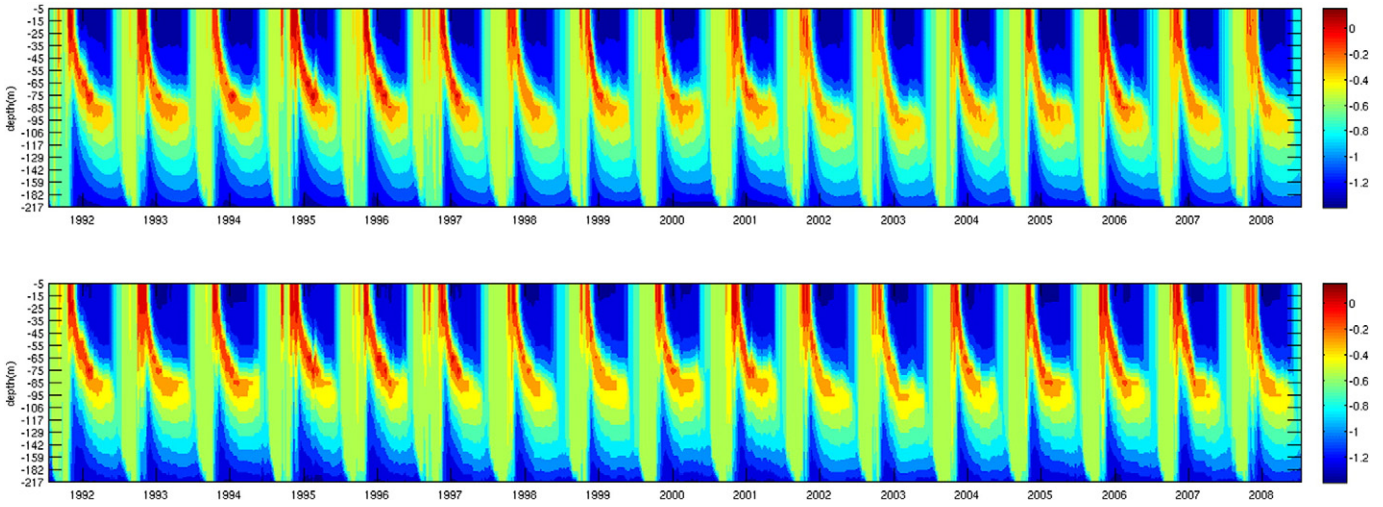

Fig. 5. The Chlorophyll-a values of the NEMO-PISCES model (top) and those given by the PROFHMM inversion (bottom), at BATS, for the period 1992 to 2008. The *X* axis represents time, the *Y* axis represents depth, and the color bar is in $\log_{10}$ [ng/l]. The last three years are validation years.

process, was found to have 294 (21×14) states both for the $sMap_{hid}$ and the $sMap_{obs}$.

For the estimation of the HMM parameters, we only took 14 years (1992–2005) for the training, each including 73 five-day-mean steps. Therefore we only have a unique training sequence, $L_{seq} = 1022$ time steps. We kept three years (2006–2008), or 219 five-day-mean steps, to validate our approach.

We can see, from Fig. 5, that the PROFHMM reconstruction respects the form and general intensity of the Chlorophyll-a profile evolution, throughout the period 1992–2008. At first glance there is no apparent difference between the reconstructions of the training (1992–2005) and the validation years (2005–2008).

4.1.1. Reconstruction of the year 2005

The year 2005 was the last year in the training data set. In Fig. 6, which is a zoom on the year 2005, the reconstruction follows the form and intensity of the NEMO-PISCES Chlorophyll-a values. The transition between states with different concentrations of Chlorophyll-a from March (at the level of the third tick on the X axis of Fig. 6, from the surface to fifty-five meters below surface the optimal reconstruction in the

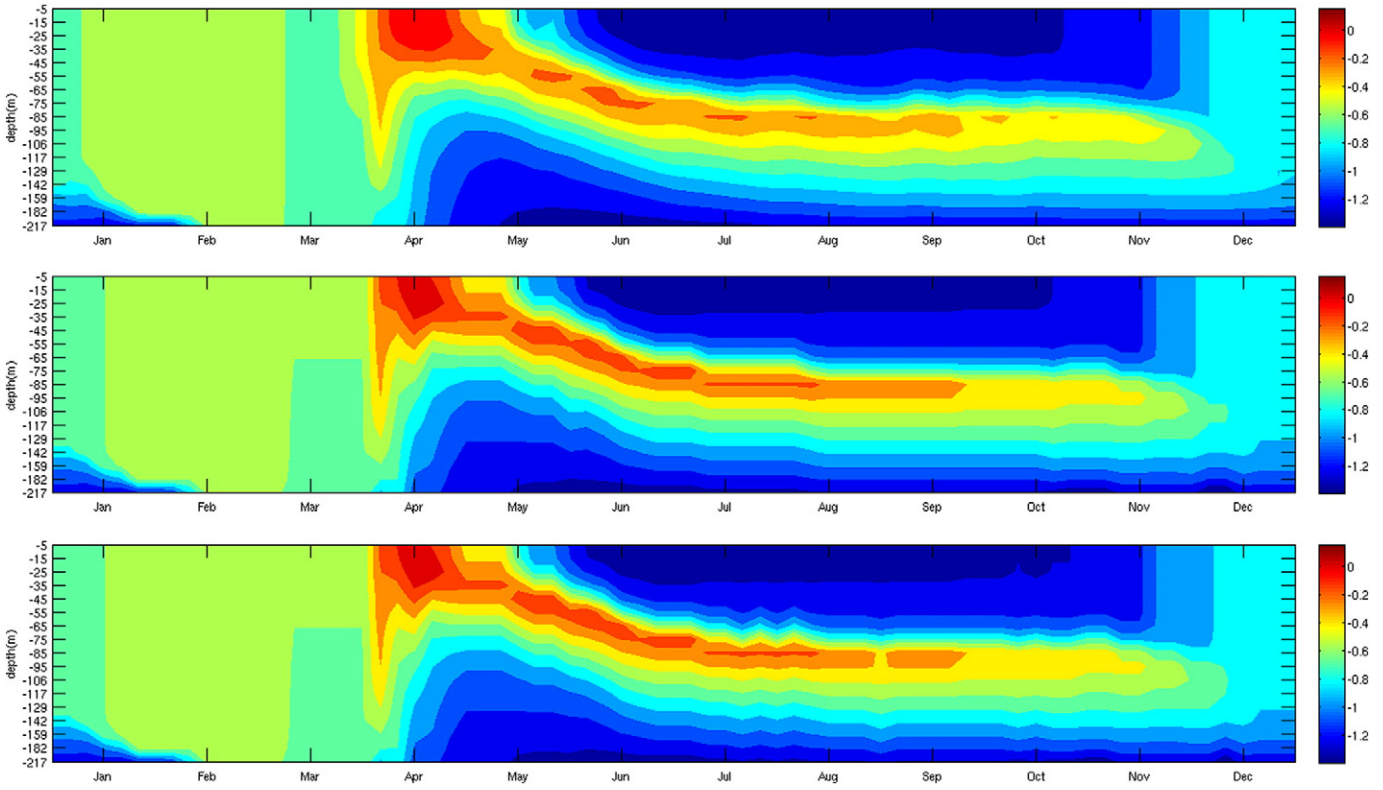

Fig. 6. The Chlorophyll-a values of the NEMO-PISCES model (top), the result of the PROFHMM inversion (middle), and the optimum reconstruction of the NEMO-PISCES model (bottom), at BATS for the year 2005. Each tick on the X axis represents the middle of the corresponding month, the Y axis represents depth, and the color bar is in log10 [ng/l].

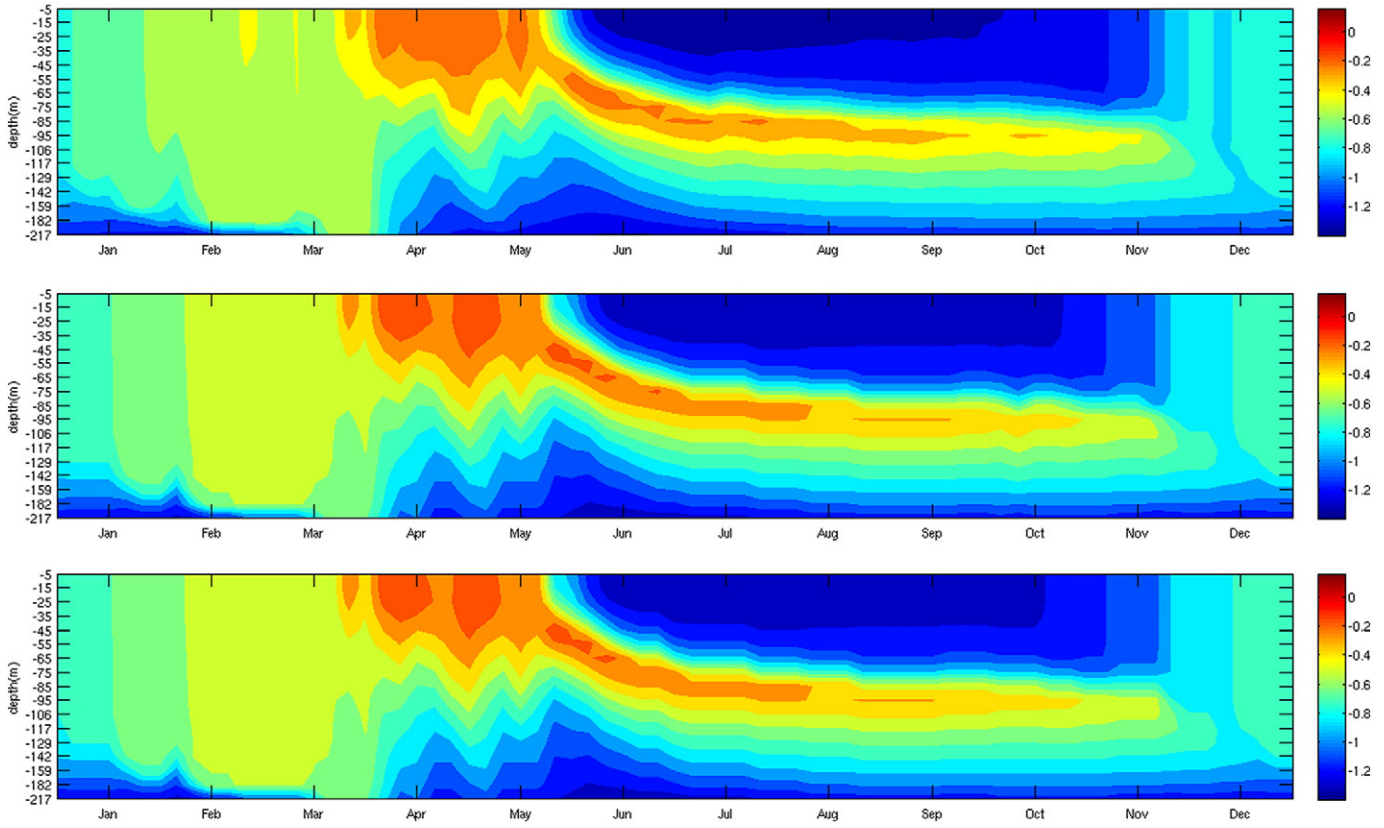

Fig. 7. The Chlorophyll-a values of the NEMO-PISCES model (top), the result of the PROFHMM inversion (middle), and the optimum reconstruction of the NEMO-PISCES model (bottom), at BATS for the year 2008. Each tick on the X axis represents the middle of the corresponding month, the Y axis represents depth, and the color bar is in log10 [ng/l].

second panel and the PROFHMM reconstruction in the third panel both indicate a higher concentration than the outputs of the NEMO model in the first panel) to the beginning of April (at the level of the fourth tick on the X axis of Fig. 6, from the surface to thirty-five meters below surface the optimal reconstruction in the second panel and the PROFHMM reconstruction in the third panel both indicate a higher concentration than the outputs of the NEMO model in the first panel) is less accurate than in the rest of the reconstruction. This seems to be due to the fact that the model situations during this interval (top panel) are poorly represented when projected on $sMap_{hid}$ (bottom panel), and therefore this is a problem of discretization as mentioned in the beginning of Section 4, and not a problem of erroneous reconstruction by the Viterbi algorithm, since PROFHMM retrieves the same states with the "optimum reconstruction" (bottom panel). We may assume that, given a finer discretization and a longer training sequence, this drawback would disappear.

4.1.2. Reconstruction of the year 2008

The year 2008 (Fig. 7) is a validation year for the PROFHMM reconstructions. The reconstruction still follows the form and intensity of the NEMO-PISCES values. As with the reconstruction of the year 2005, there are two regions presenting errors: one during late February (the values in the yellow −0.4 log10 [ng/l] range in the first twenty five meters in the top panel are green −0.6 log10 [ng/l] in the lower two panels) and one during late May (the values in the deep orange −0.1 log10 [ng/l] range in the first twenty five meters in the top panel are in light orange −0.3 log10 [ng/l] in the lower two panels). The significant error in May can also be seen in Fig. 8, where most of

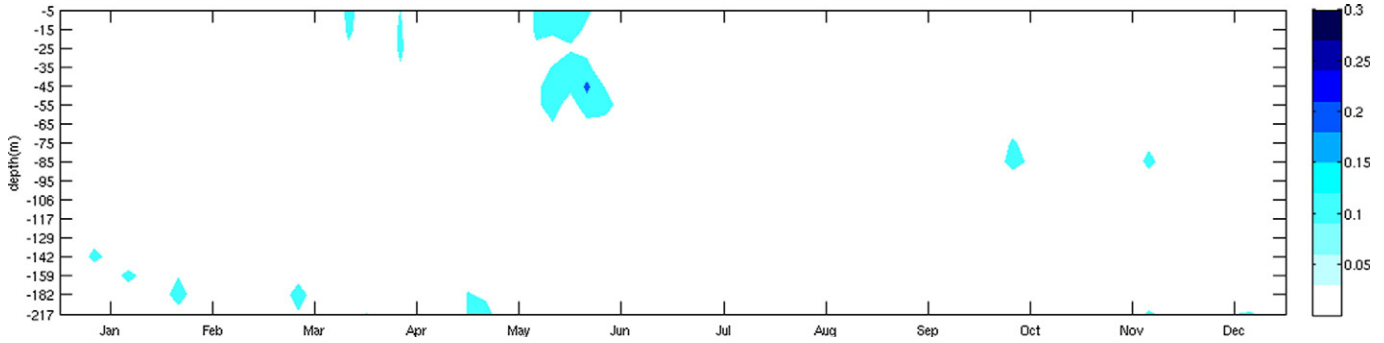

Fig. 8. The absolute error computed in Chlorophyll-a values between the NEMO-PISCES model and the result of the PROFHMM inversion, for the year 2008. Each tick on the X axis represents the middle of the corresponding month, the Y axis represents depth, and the color bar is in log10 [ng/l].

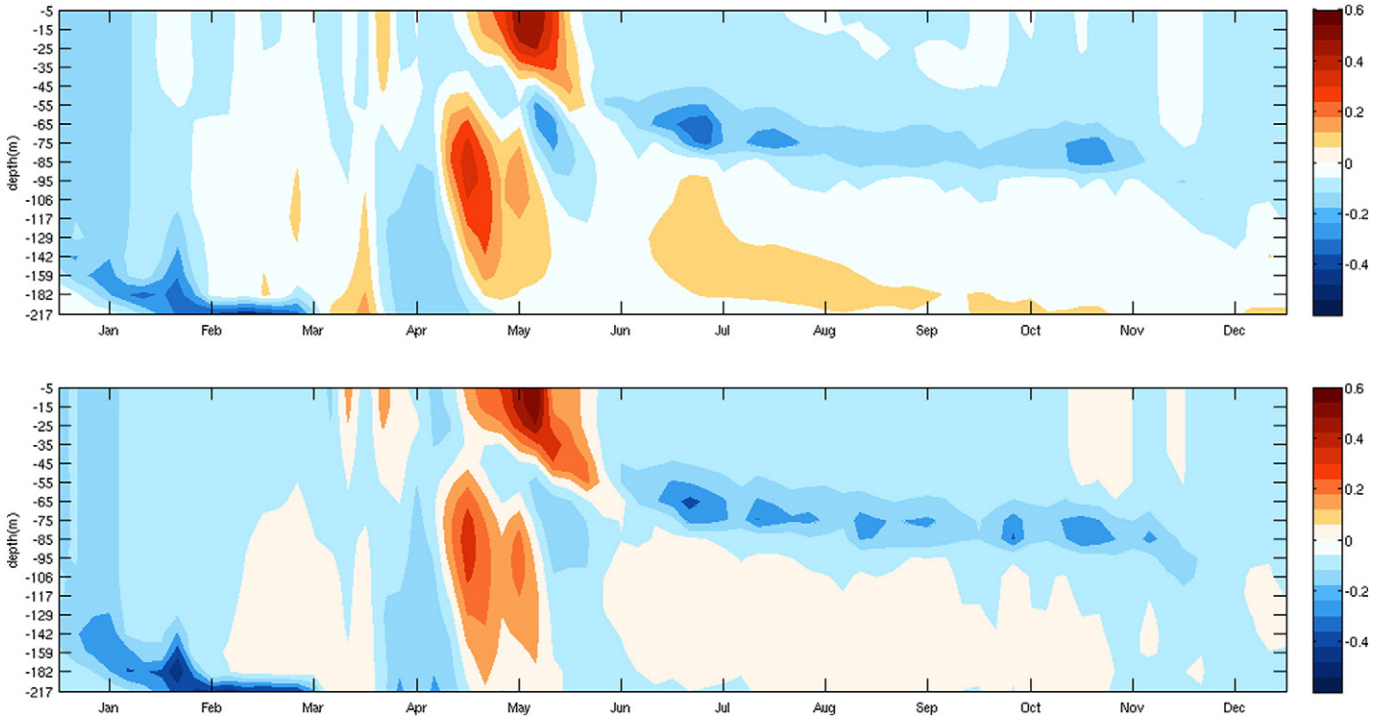

Fig. 9. The difference in Chlorophyll-a values between the average year of the NEMO-PISCES model for the period 1992–2008 and the year 2008 of the NEMO-PISCES model (top), and the difference in Chlorophyll-a values between the average year of the NEMO-PISCES model for the period 1992–2008 and the result of the PROFHMM reconstruction for the test year 2008 (bottom). Each tick on the X axis represents the middle of the corresponding month, the Y axis represents depth, and the color bar is in log10 [ng/l].

the absolute errors (in blue) are less than 0.02 $\log_{10}$ [ng/l] and none of the errors exceeds 0.2 $\log_{10}$ [ng/l]. However, the error in late February is not seen. This is due to the log scale used and the color scale.

In order to ensure that we are not repeatedly reconstructing the mean year, we calculated the climatology of the NEMO-PISCES model for the period 1992–2008 and subtracted it from both the NEMO-PISCES Chlorophyll-a data and from the PROFHMM MS reconstruction. The climatology was computed by averaging over the 17 available years, the vertical distributions of Chlorophyll-a of each of the 73 five-day steps that constitute each year. As seen in Fig. 9, for the year 2008, PROFHMM manages to reconstruct the Chlorophyll-a concentration anomaly in the water column in May and the lingering decrease in Chlorophyll-a concentration in below the thermocline from June to November.

4.2. Reconstruction from remote-sensing vectors

When applying PROFHMM to MODIS satellite data, we again used the NEMO-PISCES simulated data to represent the vertical profiles of Chlorophyll-a. The available data provided by the NEMO-PISCES

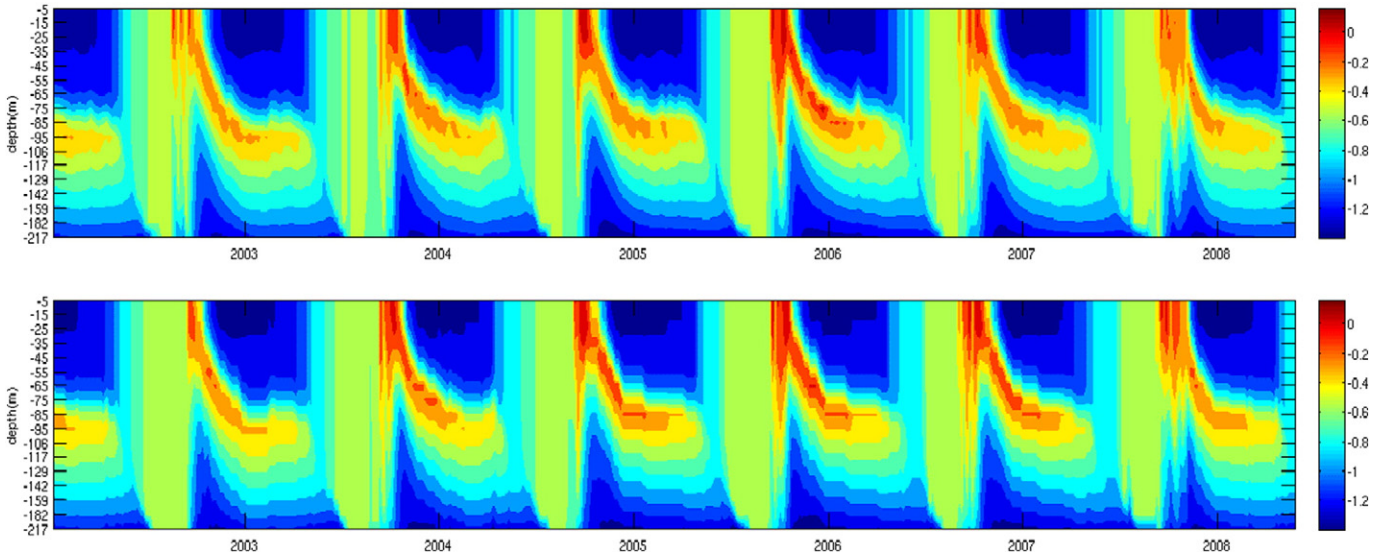

Fig. 10. The Chlorophyll-a values of the NEMO-PISCES model (top) and those given by the PROFHMM inversion based on MODIS satellite data (bottom), at BATS, for the period 2002 to 2008. The *X* axis represents time, the *Y* axis represents depth, and the color bar is in $\log_{10}$ [ng/l]. The last year (2008) is a validation year.

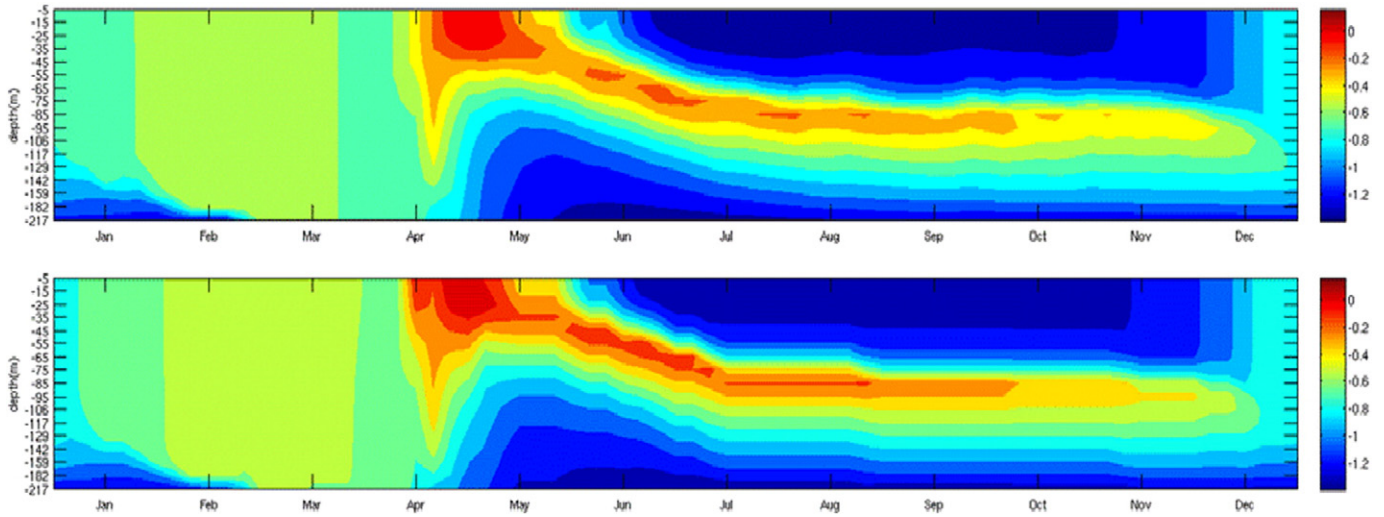

Fig. 11. The Chlorophyll-a values of the NEMO-PISCES model (top), the result of the PROFHMM inversion based on MODIS data (bottom), at BATS for the year 2005. The *X* axis represents time, the *Y* axis represents depth, and the color bar is in $\log_{10}$ [ng/l].

model and the MODIS sea-surface temperature and Chlorophyll-a data were concurrent only through the last months of the year 2002 and the entire 2003–2008 period, which represents roughly six years of data.

We kept the year 2008 as a validation set, and used the rest of the data for the training of both the $sMap_{obs}$ and the HMM, corresponding to 402 five-day steps. This led us to decrease N_{obs} to 80 classes (arranged in an 8×10 matrix), due to the limited amount of vectors to train the map. This limited amount of vectors could provide a less discretized space and the omission of certain important states in our model. As we used again the NEMO-PISCES data to perform our reconstructions, we kept the same $sMap_{hid}$ as the one used in our first experiment. The degradation of the results due to the use of satellite images is minimal, as shown in Fig. 10.

4.2.1. Reconstruction of the year 2005

The year 2005 was in the training data for both experiments (MS and RS). In Fig. 11, which is a zoom of the MODIS-based reconstruction of the year 2005, the reconstruction again follows the form and intensity of the NEMO-PISCES Chlorophyll-a values. The reconstruction has slightly more pronounced high values.

4.2.2. Reconstruction of the year 2008

The year 2008 is, as stated above, a validation year for both PROFHMM reconstructions. Although we only used six years of data for the training of the HMM, and 402 RS vectors for the training of the $sMap_{obs}$, the reconstruction still follows the form and intensity of the NEMO-PISCES values. Again, the reconstruction of the test year presents slightly higher values, as shown in Fig. 12.

4.3. Comparisons

In order to have some quantifiable results we computed the RMS between the PROFHMM reconstruction of Chlorophyll-a and the NEMO-PISCES model values, as well as the RMS between the optimum reconstruction and the NEMO-PISCES model values.

We also located, in each case, the 10% of the minimum and 10% maximum values of Chlorophyll-a in the NEMO-PISCES model, and

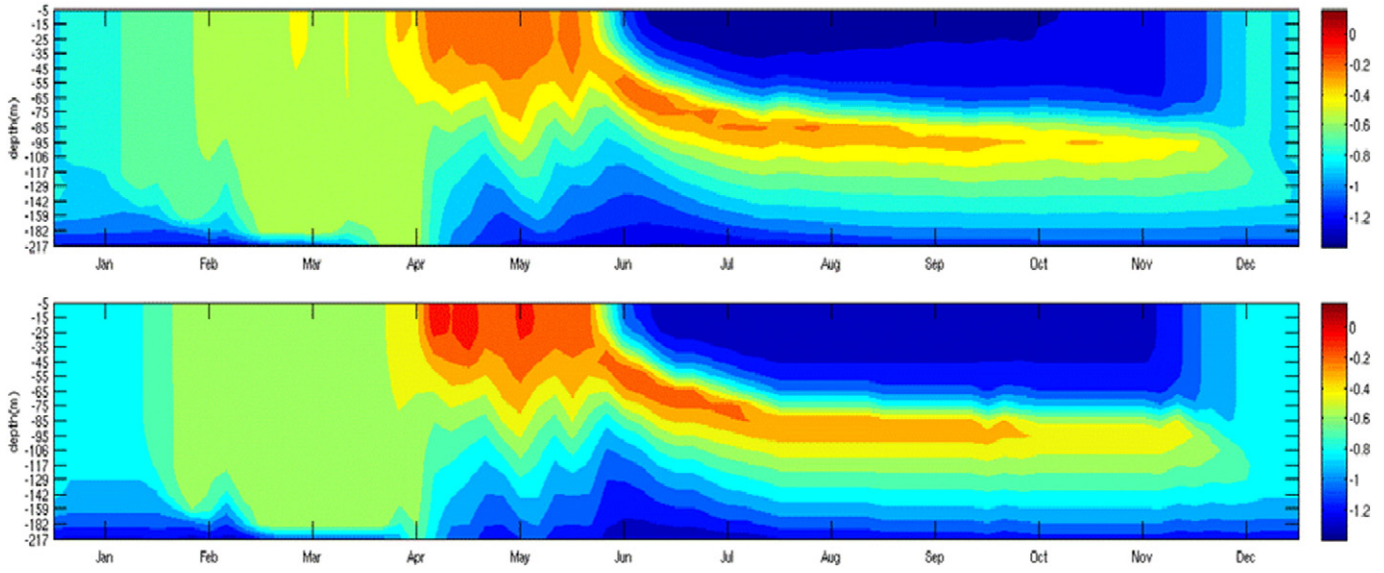

Fig. 12. The Chlorophyll-a values of the NEMO-PISCES model (top), the result of the PROFHMM inversion based on MODIS satellite data (bottom), at BATS for the year 2008. The color bar is in $\log_{10}$ [ng/l].

Table 1
PROFHMM RMS performances in [ng/l].

	MS data performances				RS data performances			
	RMS	Min	Max	%	RMS	Min	Max	%
1992–2008 INVERTED	0.0455	0.0034	0.1192	88.07%	–	–	–	–
1992–2008 OPTIMUM	0.0423	0.0032	0.1158		–	–	–	
2005 INVERTED	0.0411	0.0072	0.0422	93.15%	0.0499	0.0067	0.0514	84.93%
2005 OPTIMUM	0.0400	0.0069	0.0411		0.0400	0.0069	0.0411	
2008 INVERTED	0.0303	0.0076	0.0310	91.78%	0.0399	0.0096	0.0408	75.34%
2008 OPTIMUM	0.0302	0.0076	0.0309		0.0302	0.0076	0.0309	
2002–2008 INVERTED	0.0453	0.0097	0.0466	84.63%	0.0590	0.0217	0.0856	75.74%
2002–2008 OPTIMUM	0.0406	0.0077	0.0418		0.0406	0.0077	0.0418	

calculated the RMS between them and their co-localized points in the PROFHMM and optimum reconstructions. This allows us to affirm that the extreme values are also well reconstructed by PROFHMM.

These RMSs are shown in Table 1, along with the percentage of retrieved indices (SI_{ret}) corresponding to the optimum indices (SI_{opt}).

The term inverted corresponds to the PROFHMM reconstruction, while the term optimum corresponds to the "optimum reconstruction". The years above these terms correspond to the period being reconstructed. MS DATA performances correspond to the reconstructions done with sea-surface observations taken from the model outputs, while RS DATA performances correspond to the reconstruction done with the help of MODIS satellite images.

In Table 1, the RS performances for the year 2008 as seen in the difference OPTIMUM–INVERTED is less than the same difference computed over the complete period (2002–2008). This indicates that the performances on the training were poorer for this experiment, but that they would greatly improve given a longer training data set. The MIN and MAX columns correspond to the RMS calculated over the points between the bottom 10% and top 10% of the Chlorophyll-a values in the NEMO-PISCES outputs, and their respective reconstructions through PROFHMM.

5. Conclusion

In the present paper we have introduced PROFHMM, which is an inversion method based on SOM and HMM. PROFHMM is able to reconstruct the temporal evolution of hidden profiles of biogeochemical variables from observable data at the top layer of the profile. We applied this method for the reconstruction of the Chlorophyll-a vertical profiles at BATS, using model outputs and satellite data as sea-surface observations. The method was also applied at the HOT (Hawaii Ocean Time-series) location of the JGOFS campaign and, for which we obtained similar performances (not shown in the article).

Upon developing PROFHMM, we realized that it could be modified to a more general, statistical, non-linear method that could be applied to the reconstruction of other oceanic variables.

We intend to expand the method in order to reconstruct the spatial evolution of the temperature field on the transect of the ARAMIS campaign. However, PROFHMM is general enough to be applicable to a multitude of other problems in geophysics, for which it is possible to learn, in a statistical way, the dynamics of a geophysical model, while observing a sub-dimension of the variables.

PROFHMM is very efficient in terms of calculations. Its cost efficiency could allow its integration into reanalysis or forecasting models for improving assimilation by producing accurate first guesses of the model evolution.

Long-term perspectives of PROFHMM include taking into account the amount of incomplete satellite observations in the inversions, and expanding the principle to Bayesian fields in order to reconstruct space–time evolutions of geophysical or biogeochemical variables.

Acknowledgments

The research presented in this paper was financed by the Centre National de l'Etude Spatial (CNES, French National Center for Space Studies), and the Délégation Gouvernementale pour l'Armement (DGA, French Military Research Delegation), both of which we thank for their support. We also thank Cyril Moulin and Laurent Bopp, of the Laboratoire des Sciences du Climat et l'Environnement (LSCE; Climate and Environmental Sciences Laboratory), for their help with NEMO-PISCES, and Michel Crépon of the Laboratoire Océanographique et du Climat — Expérimentation et Approches Numériques (LOCEAN; Oceanographic and Climate Laboratory — Experimentation and Numerical Approaches) for his input on the method.

References

Bates, Nicholas R., Michaels, Anthony F., & Knap, Anthony H. (1996). Seasonal and interannual variability of oceanic carbon dioxide species at the US JGOFS Bermuda Atlantic Time-series Study (BATS) site. *Deep Sea Research Part II: Topical Studies in Oceanography*, *43*(2), 347–383.

Baum, L. E., Petrie, T., Soules, G., & Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, *41*, 164–171.

Brajard, J., Cédric, J., Cyril, M., & Thiria, S. (2006). *Use of a neuro-variational inversion for retrieving oceanic and atmospheric constituents from satellite ocean color sensor: Application to absorbing aerosols neural networks. Earth Sciences and Environmental Applications of Computational Intelligence*, *19(2)*, Amsterdam: Elsevier.

Demarcq, H., Richardson, A. J., & Field, J. G. (2008). Generalised model of primary production in the southern Benguela upwelling system. *Marine Ecology Progress Series*, *354*, 59.

Dinnat, E., Boutin, J., Caudal, G., Etcheto, J., & Waldteufel, P. (2002). Influence of sea surface emissivity model parameters in L-band for the estimation of salinity. *International Journal of Remote Sensing*, *23*, 5117–5122.

Doneya, S. C., Kleypasa, J. A., Sarmiento, J. L., & Falkowski, P. G. (2002). The US JGOFS synthesis and modeling project — An introduction. *Deep-Sea Research Part II*, *49*, 1–20.

Feldman, G. C., Kuring, N. A., Ng, C., Esaias, W. E., McClain, C. R., Elrod, J. A., et al. (1989). Ocean color: Availability of the global data set. *Eos*, *70*, 634–641.

Gehlen, M., Moussaoui, A., Perruche, C., Dombrowsky, E., Aumont, O., Brasseur, P., et al. (2010). Integration of biogeochemistry and ecology to Mercator ocean systems: Recent advances and future developments of the Green Mercator initiative. *MyOcean Science Days 1–2*.

Gurvan, M., & the NEMO team (2012). *NEMO ocean engine — Version 3.4*. Note du Pôle de Modélisation de l'Institut Pierre-Simon Laplace, 27. Paris: Institut Pierre-Simon Laplace (ISSN 1288-1619).

Hagenauer, J., & Hoeher, P. (1989). A Viterbi algorithm with soft-decision outputs and its applications. *Proc. IEEE GLOBECOM Conference, Dallas, Texas, USA, November 1989* (pp. 47.11–47.17).

Hays, G. C., Richardson, A. J., & Robinson, C. (2005). Climate change and marine plankton. *Trends in Ecology & Evolution*, *20*(6), 337–344.

Jaziri, R., Lebbah, M., Bennani, Y., & Chenot, J. -H. (2011). SOS-HMM: Self-organizing structure of Hidden Markov Model, artificial neural networks and machine learning — ICANN 2011. *Lecture Notes in Computer Science*, *6792*, 87–94.

Juang, B. -H. (2003). *Hidden Markov Models. Encyclopedia of telecommunications*. Wiley Online Library (onlinelibrary.wiley.com).

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, *2(12)*. (pp. 1137–1143). San Mateo, California: Morgan Kaufmann.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, *78*(9).

Krishna Rao, P., Smith, W. L., & Koffler, R. (1972). Global sea-surface temperature distribution determined from an environmental satellite. *Monthly Weather Review*, *100*(1), 10–14.

Liu, Y., & Weisberg, R. (2005). Patterns of ocean current variability on the West Florida Shelf using the self-organizing map. *Journal of Geophysical Research*0148-0227, *110*, C06003. http://dx.doi.org/10.1029/2004JC002786.

Liu, Y., & Weisberg, R. H. (2011). *A review of self-organizing map applications in meteorology and oceanography*. INTECH Open Access Publisher.

Miller, C. B. (2003). *Biological oceanography0632055367.*

Richardson, A. J., Pfaff, M. C., Field, J. G., Silulwane, N. F., & Shillington, F. A. (2002). Identifying characteristic chlorophyll a profiles in the coastal domain using an artificial neural network. *Journal of Plankton Research*, *24*, 1289–1303.

Richardson, A. J., Risien, C., & Shillington, F. A. (2003). Using self-organizing maps to identify patterns in satellite imagery. *Progress in Oceanography*, *59*, 223–239.

Silulwane, N. F., Richardson, A. J., Shillington, F. A., & Mitchell-Innes, B. A. (2001). Identification and classification of vertical chlorophyll patterns in the Benguela upwelling system and Angola–Benguela Front using an artificial neural network. *South African Journal of Marine Science*, *23*, 37–51.

Uitz, J., Claustre, H., Morel, A., & Hooker, S. B. (2006). Vertical distribution of phytoplankton communities in open ocean: An assessment based on surface chlorophyll. *Journal of Geophysical Research*, *111*, C08005. http://dx.doi.org/10.1029/2005JC003207.

Viterbi, A. J. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, *13*(2), 260–269. http://dx.doi.org/10.1109/TIT.1967.

Viterbi, A. J. (1998). An intuitive justification and a simplified implementation of a MAP decoder for convolutional codes. *IEEE Journal on Selected Areas in Communications*, *16*(2), 260–264.

Web References

SOM TOOLBOX. http://www.cis.hut.fi/projects/somtoolbox/package/docs2/somtoolbox.html

NASA OceanColor. website http://oceancolor.gsfc.nasa.gov/

LOCEAN (2002–2009). *Altimétrie sur un rail atlantique et mesures in situ.* ARAMIS http://aramis.locean-ipsl.upmc.fr.