Identification of a dynamical model for phytoplankton bloom based on high frequency measurements

Hafiz Ahmed, Rosane Ushirobira, Denis Efimov, Wilfrid Perruquetti

Identification of a dynamical model for phytoplankton bloom based on high frequency measurements

Hafiz Ahmed[*]

School of Mechanical, Aerospace & Automotive Engineering,
Coventry University,
CV1 5FB, United Kingdom.
E-mail: hafiz.h.ahmed@ieee.org
*Work done while working at Inria, Non-A team.

Rosane Ushirobira

Inria, Non-A team,
Villeneuve d'Ascq, France.
E-mail: rosane.ushirobira@inria.fr

Denis Efimov

Inria, Non-A team,
Villeneuve d'Ascq, France.
CRIStAL (UMR CNRS 9189),
Villeneuve-d'Ascq, France.
E-mail: denis.efimov@inria.fr

Wilfrid Perruquetti

Inria, Non-A team,
Villeneuve d'Ascq, France.
CRIStAL (UMR CNRS 9189),
Ecole Centrale de Lille,
Villeneuve-d'Ascq, France.
E-mail: wilfrid.perruquetti@inria.fr

Abstract High frequency measurements of various water characteristics and nutrients information of the Marel-Carnot sea monitoring station (Boulogne-sur-Mer, France) have been used to identify a physiological model for phytoplankton bloom through the fluorescence signal. An auto-regressive-moving-average with exogenous inputs (ARMAX) model is designed and tested based on the dataset. The model takes into account the effect of the measured water characteristics and nutrient level information. Through this study, it is demonstrated that the developed dynamical model can be used for estimating the fluorescence level (which characterizes the phytoplankton biomass) and for predicting the various states of phytoplankton bloom. Thus, the developed model can be used for monitoring phytoplankton biomass in the water which in turn might give information about unbalanced ecosystem or change in water quality.

Biographical notes: Hafiz Ahmed was born in Dhaka, Bangladesh. He received the M.Sc. degree in system, control, and information technology from Joseph Fourier University, Grenoble, France, in 2013. He received the Ph.D. degree in Automatic Control from the University of Lille, France in 2016. He was a Post-Doctoral Fellow with the Clemson University, Clemson, SC, USA. In May 2017, he joined North South University, Dhaka, Bangladesh as an Assistant Professor. Since September 2017, he is working as a Lecturer at the School of Mechanical, Aerospace & Automotive Engineering at Coventry University, United Kingdom. His research interests include modeling, identification, control and stability analysis of dynamical systems with application to environmental monitoring, systems biology, and renewable energy.

Rosane Ushirobira received the Ph.D. degree in mathematics from the University of Strasbourg, Strasbourg, France, in 1996. She was a Post-Doctoral Fellow with the University of Poitiers, Poitiers, France, and the Technische Universiteit Eindhoven, Eindhoven, The Netherlands. From 1999 to 2017, she was an Associate Professor with the University of Bourgogne, Dijon, France. She is an Inria researcher since 2017. In 2010, she defended her Habilitation thesis at Dijon. Her current research interests include Algebraic Lie theory, where she has published works on Lie algebras and superalgebras, the Weyl algebra, and their representations, automatic control theory, algebraic methods and applications in this area, especially in estimation problems.

Denis Efimov received the Ph.D. degree in Automatic Control from the Saint-Petersburg State Electrical Engineering University (Russia) in 2001, and the Dr.Sc. degree in Automatic control in 2006 from Institute for Problems of Mechanical Engineering RAS (Saint-Petersburg, Russia). From 2000 to 2009 he was research fellow of the Institute for Problems of Mechanical Engineering RAS, Control of Complex Systems Laboratory. From 2006 to 2011 he was working in the LSS (Supelec, France), the Montefiore Institute (University of Liege, Belgium) and the Automatic control group at IMS lab (University of Bordeaux I, France). Since 2011 he joined the Non-A team at Inria Lille center. He is a member of several IFAC TCs and a Senior member of IEEE. His main research interests include nonlinear oscillation analysis, observation and control, switched and nonlinear system stability.

Wilfrid Perruquetti was born in 1968 in Saint Gilles, France. In 1991, he received a M.Sc. in Automatic Control and graduated from "Institut Industriel du Nord" (French "Grande Ecole"). In 1994, he obtained a Ph.D. in Automatic Control, then joined the "Ecole Centrale de Lille" (French "Grande Ecole") as an Assistant Professor in 1995, where he is actually Full Professor (since 2003). He belongs to CRIStAL (CNRS, Ecole Centrale de Lille, Lille University) and to the Non-A project (Inria). He is Deputy Scientific Director of INS2I CNRS, permanent head of Non-A, Inria, and was from (2010-2014) scientific project manager at ANR, from (2007-2009) a scientific officer at the French Ministry of Education and Research (DGRI). He has published more than 300 journals, book chapters and conference papers.

1 Introduction

It is well-known that phytoplankton blooms have oscillatory explosive dynamics Evans and Parslow [1985]. The appearance of an atypical bloom process may serve as an indicator of the ecosystem balance loss signalizing that nutrient concentration has reached unacceptably high levels. Some phytoplankton species may produce toxin under certain conditions and their large emergence has noticeable dangerous implications on water quality and ecosystem health. For example, Algae blooms may damage higher organisms (such as zooplankton or shellfish) providing a vast economic implication Huppert et al. [2005]. Algae blooms are also responsible for the spread of some epidemics. In Bangladesh, for example, cholera outbreaks appear to originate from annual bloom events in the Bay of Bengal Colwell [1996]. From another side, the annual phytoplankton blooms are natural periodic events entrained by surrounding ecological and environmental conditions Fasham et al. [1990].

Phytoplankton cells are capable of integrating natural and human induced disturbances by changing their physiology. Thus it is important to prevent and detect phytoplankton blooms at an early stage and understand their physical and nutrient conditions of outbreak. Given the importance of phytoplankton blooms, many different modeling approaches have been developed for understanding their underlying dynamics. The existing literature is mainly devoted to the analysis and design of analytical simplified models Evans and Parslow [1985], Huppert et al. [2002, 2005] described by ordinary differential equations of the Lotka-Volterra type. However, data driven modeling approaches such as artificial neural networks or genetic algorithms have been applied to model and predict (mainly short term prediction) phytoplankton blooms Lee et al. [2003], Muttil and Lee [2005], Muttil and Chau [2006]. Recently in Rousseeuw et al. [2015], a numerical modeling approach has been applied for water quality monitoring based on the Hidden Markov Model framework. Such an advance became possible due to the implementation of a high frequency measurement system at Marel-Carnot station Hernandez-Farinas et al. [2014]. This station measures various water characteristics and nutrient information of the surrounding coastal area. Because of its high frequency nature, it generates a large amount of data signals. In order to create an effective water monitoring system, processing and analysis of this huge volume of data in an unmanned and systematic way is essential.

The goal of the present work is to identify a dynamical model for phytoplankton bloom using Marel-Carnot's data. The identification problem has been thoroughly analyzed and solved in the control engineering community (in applications to the technical systems). The occurrence of a large high frequency dataset for bio-systems opens the doors for this methodology application in this new area. The developed model obtain through system identification is described by differential equations using the theory presented in Walter and Pronzato [1997], Ljung [1998], Nelles [2001]), with the dynamics approximated on the given dataset of the Marel-Carnot's data[1].

The paper is organized as follows: Section 2 discusses the pre-processing of the Marel-Carnot's high frequency datasets to make them suitable for the model identification purpose. In Section 3, the model identification results and methods are given. The verification and analysis of the obtained model are summarized in the Section 4. Finally, Section 5 concludes the paper.

Table 1   Range and uncertainty in measured signals

Physico-chemical parameters	Range	Uncertainty
Temperature	−5 to +30°C	±0.1°C
Dissolved oxygen	0 to 20 mg/L	±0.2 mg/L
pH	6.8 to 8.5 UpH	±0.2 UpH
Turbidity	0 to 4000 NTU	±10%
Fluorescence	0 to 50 FFU	±10%
Nitrate	0.1 to 100 μmol/L	±5%
Phosphate	0.1 to 100 μmol/L	±5%
Silicate	0.1 to 100 μmol/L	±5%

2 Data Preparation

Marel-Carnot station is located in the harbor of Boulogne-sur-Mer which is in the north-east of France. The data is publicly available in the website of Ifremer[2]. In order to collect the data, several fixed buoys and ferry boxes were implanted in that area. We are considering the data between 2005 and 2009 for our model identification. The dataset has 14 different data signals. However, they were measured in variable frequency, *i.e.* the total amount of data points for each signal are not the same. The data can be classified into two categories based on the frequency of measurements:

1. 11 water characteristics (collected every 20 minutes),

2. 3 nutrient level information (collected every 12 hours).

Water characteristics include 3 different kinds of dissolved oxygen (uncorrected, salinity corrected and saturated), water salinity level, air temperature, luminosity, water pH level, turbidity, fluorescence, water temperature and height of the sea level. The three nutrients are nitrate, phosphate and silicate. The amount of uncertainty in the measurements can be seen in table 1[3].

An important signal to be considered is the fluorescence signal. This signal characterizes the phytoplankton biomass. So, from this signal we can evaluate the amount of phytoplankton and its various states like in pre-blooming, blooming, post-blooming, no production of biomass and rare events. This signal recorded in between 2005 and 2009 with a 20 minutes measurement frequency can be seen in Fig. 1.

From Fig. 1, it can be seen that phytoplankton blooms are annual periodic events entrained by surrounding ecological and environmental conditions as mentioned in Colwell [1996]. It can also be noticed that the data points are missing very frequently deteriorating the quality of data. The data quality plays an important role in system identification. So, by using exactly this data, we cannot identify a very high quality model that captures most of the dynamics of the phytoplankton's biomass.

One of the ways to overcome this limitation of data (*i.e.* frequent missing data points) can be data averaging. Averaging is a very common technique in biological domain to reduce the effect of noise and missing data points. Moreover, it helps to reduce and simplify the computation and presentation of the data. After some mathematical manipulations, we found that 1-day averaging is able to captures almost the same dynamics of the original data but with scaled down amplitude. Moreover, it also smooths the original signal and reduces the noise as well. In Fig. 2, the 1-day averaged fluorescence signal can be seen.
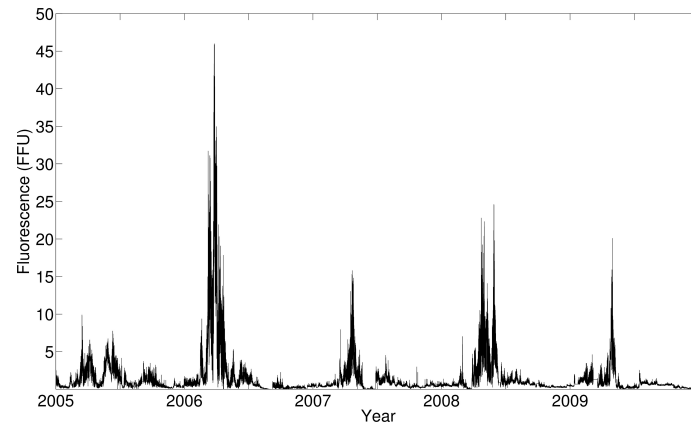
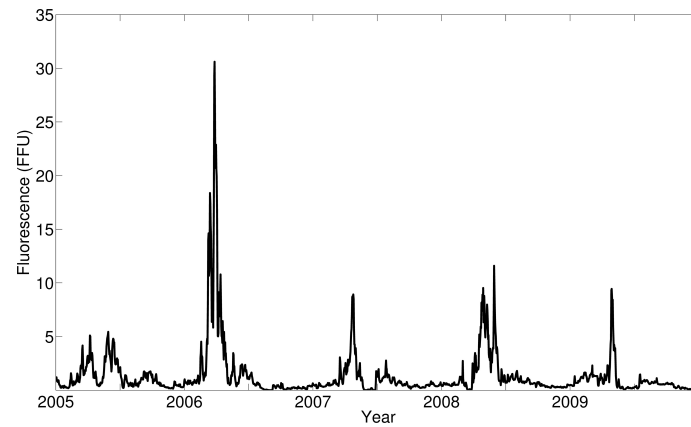**Figure 1** Fluorescence signal between 2005 and 2009

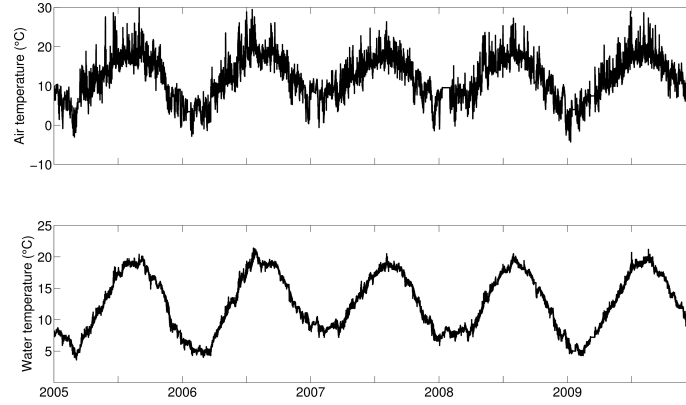**Figure 2** 1-day averaged fluorescence signal between 2005 and 2009

Figure 3   Air and water temperature
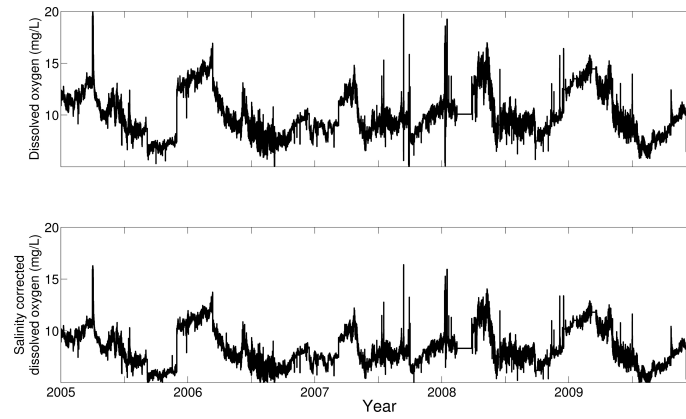

Figure 4   Uncorrected and salinity corrected dissolved oxygen

Besides fluorescence, we also have 13 more signals. For example, we have two different temperature signals like air and water temperature. In general, the temperatures of water and air are of periodic nature. The temperatures rise in the summer and go down in the winter. Moreover, the water temperature of the sea varies less than the air above the water. The temperatures are presented in Fig. 3.

Dissolved oxygen is a very important criterion for water quality measurement. It is well-known that many chemical and biological reactions in ground water and surface water depend directly or indirectly on the amount of available oxygen. Dissolved oxygen is necessary in aquatic systems for the survival and growth of many aquatic organisms and used as an indicator of the health of surface water bodies Wilde et al. [1998].

Many factors influence the dissolved oxygen measurements, and salinity is one of them. The on-board sensor implanted in the ferry box calculated the dissolved oxygen without considering the salinity. This sensor-calculated oxygen is named "uncorrected oxygen". Later it was corrected by incorporating the effect of salinity. This dissolved oxygen is labelled "salinity corrected oxygen". The uncorrected and salinity corrected dissolved oxygen are given in Fig. 4.
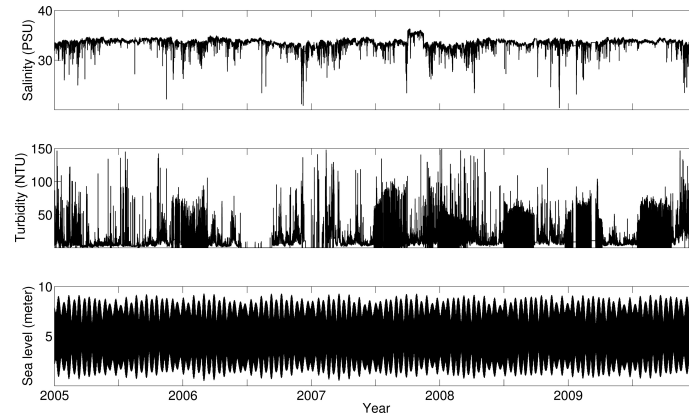
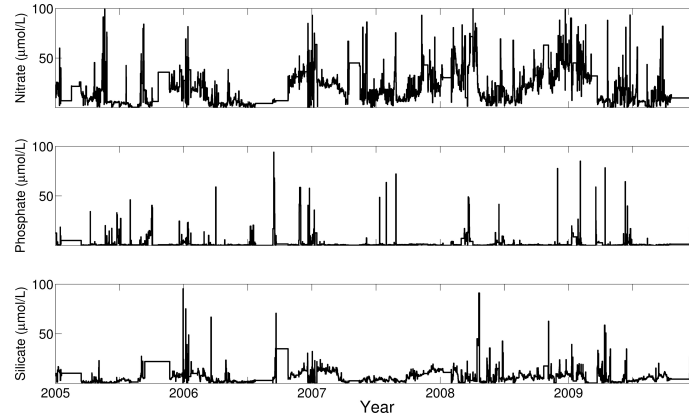**Figure 5** Water salinity, turbidity and height of the sea level

Figure 6 Nitrate, phosphate and silicate signals

Some other important water characteristics are turbidity, salinity and sea level. Turbidity is a measure of water clarity which can affect the color of water. A high turbidity increases the water temperature which in turn reduces the concentration of dissolved oxygen. Excessive algal growth is one of the reasons of high turbidity Wang [1974]. Salinity has an effect on the density of the sea water (high salinity increases the density). These signals can be seen in Fig. 5.

Nutrients (like nitrate, silicate and phosphate) play an important role in the production of the phytoplankton biomass. So, data about these three nutrients can give us very valuable information regarding various states of the phytoplankton biomass production. These three nutrient signals can be seen in Fig. 6.

Like fluorescence signal, 1-day averaging was also applied to all other signals to overcome the data limitation (*i.e.* frequent missing data points, variable frequency sampling *etc.*).
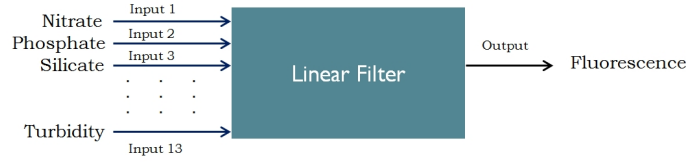
Figure 7   Structure of the dynamical model under consideration

3   Model Identification

Our goal is to identify a dynamical model that will be able to predict the phytoplankton bloom considering surrounding ecological and environmental conditions. In model/system identification, selecting the inputs and outputs of the model is very important. In order to model the phytoplankton blooms, the fluorescence signal, which characterizes the biomass of the phytoplankton has been chosen as the output of our model, and other 13 signals (3 nutrients and 10 water characteristics) have been considered as inputs of our model.

Moreover, since the dynamical relation between the inputs and output are not very well known in our case, we have taken an engineering approach based on black-box model architecture. One of the reasons of using this approach is our previous work experience in biological domain. We have previously successfully applied black-box approach in modeling the valve opening activity of a population of oysters where the relation of the inputs (like sun position, moon position and water tide level) with the output (*i.e.* valve opening activity) was also not well known Ahmed et al. [2017, 2014].

The graphical representation of the model structure considering fluorescence as output and other signals (dissolved oxygen, luminosity, salinity, pH, turbidity, nitrate, phosphate, silicate, sea level, water temperature, air temperature) as inputs can be seen in Fig. 7.

The generic structure of the dynamical model for phytoplankton bloom can be represented as follows:

$$y_k = F(y_{k-1}, \ldots, y_{k-\nu}; x_{1,k}, \ldots, x_{1,k-\mu}; \ldots;$$
$$x_{13,k}, \ldots, x_{13,k-\mu}; \theta) + \epsilon_k \tag{1}$$

where y is the output (*i.e.* fluorescence), $x_1, \ldots, x_{13}$ denote the inputs, $\nu \geq 0$ is the number of past output values taken into account to determine the fluorescence level at the next time instant k (*i.e.* in this model $k-1$ is the current time instant and y_{k-1} is the current fluorescence level state), $\mu \geq 0$ is the number of past input values used in the decision on the value of y_k (the model (1) assumes that the instant values of $x_{1,k}, \ldots, x_{13,k}$ can be used for this decision), $\epsilon_k \in \mathbb{R}$ is the disturbance representing additional uncertain influencing forces and approximation errors (it is assumed that this term is sufficiently small if the model has been well identified); $\theta \in \mathbb{R}^q, q > 0$ is the vector of constant parameters of the model (1) and $F : \mathbb{R}_+^{\nu+13\mu+13+q} \to \mathbb{R}_+$ is a function defining the model structure.

The identification procedure for the above mentioned model consists of selecting an appropriate structure for the function of $F(.)$ with posterior calculation of the corresponding vector of parameters θ. Out of many possible choices for $F(.)$, we have chosen ARMAX model structure. The ARMAX (auto-regressive-moving-average with exogenous inputs) model is one of the most popular model structures used for identification in various fields of science Ljung [1998], Nelles [2001], Walter and Pronzato [1997]. Its advantages include

linearity with respect to all signals and parameters, different effective methods for calculation of the parameters, simplicity of stability analysis of the obtained model, robustness and sensitivity with respect to perturbations Ljung [1998]. Therefore, the function $F(.)$ in equation (1) is selected in the following form:

$$F(.) = \sum_{i=1}^{\nu} \theta_i^y y_{k-i} + \sum_{j=0}^{\mu} \theta_j^{x_1} x_{1,k-j} + \cdots + \sum_{j=0}^{\mu} \theta_j^{x_{13}} x_{13,k-j}$$

with $\theta = \left[\theta_1^y, \ldots, \theta_\nu^y, \theta_1^{x_1}, \cdots, \theta_\mu^{x_1}, \ldots, \theta_1^{x_{13}}, \cdots, \theta_\mu^{x_{13}} \right]^T$.

Using least-square method Ljung [1998], Nelles [2001], Walter and Pronzato [1997], the estimate $\hat{\theta}$ of the parameters θ have been calculated for Marel-Carnot high frequency dataset as

$$\hat{\theta} = \left(\Phi^T \Phi \right)^{-1} \Phi^T \Upsilon,$$

where $\Upsilon = \left[y_{\nu+1}, \ldots, y_L \right]^T$ and the jth row of the matrix Φ is

$$\Phi_j = \left[y_{j+\nu-1}, \ldots, y_j, x_{1,j+\nu}, \ldots, x_{1,j+\nu-\mu}, \ldots, x_{13,j+\nu}, \ldots, x_{13,j+\nu-\mu} \right]$$

for all $1 \leq j \leq L - \nu$, where L is the length of the training dataset.

4   Model Verification and Analysis

Two performance costs J_ϵ and J_p have been calculated to evaluate the obtained accuracy of the designed model. The cost J_ϵ estimates the average amplitude of ϵ_k (the approximation errors in the model):

$$J_\epsilon = \sqrt{\frac{\sum_{j=1}^{L-\nu} \epsilon_j^2}{L - \nu}}$$

where $\epsilon_j = y_{j+\nu} - \hat{y}_{j+\nu}, \hat{y}_{j+\nu} = \Phi_j^T \hat{\theta}, 1 \leq j \leq L - \nu$.

The cost J_p evaluates the prediction quality of the model on an infinite time interval:

$$J_p = \sqrt{\frac{\sum_{j=1}^{L-\nu} e_j^2}{L - \nu}}$$

where $e_j = y_{j+\nu} - \tilde{y}_{j+\nu}, 1 \leq j \leq L - \nu$ and

$$\tilde{y}_k = y_k, \quad 1 \leq k \leq \nu$$

$$\tilde{y}_{\nu+j} = \sum_{i=1}^{\nu} \theta_i^y \tilde{y}_{\nu+j-i} + \sum_{z=0}^{\mu} \theta_i^{x_1} x_{1,\nu+j-z} + \cdots$$
$$+ \sum_{z=0}^{\mu} \theta_i^{x_{13}} x_{13,\nu+j-z}, \quad 1 \leq j \leq L - \nu$$
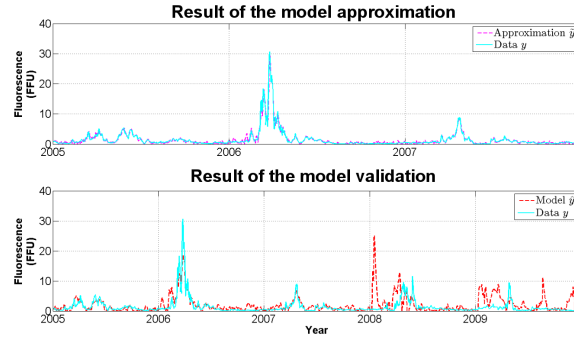
Figure 8   The result of the model numerical verification (training dataset is of 3 years)

are the estimates of the fluorescence signal y_k generated independently by the designed model in the presence of the same inputs signals and L is the length of the validation dataset. Example of the obtained estimates $\hat{y}$ and $\tilde{y}$ (for $\nu = \mu = 10$, *i.e.* considering 10 days of memory) can be seen in Fig. 8, 9 and 10 respectively for different training and validation dataset. In Fig. 8, data from the first three years have been used for training. Similarly, in Fig. 9, data from the first four years have been used for training. Finally, the entire dataset has been used for training in Fig. 10. The entire dataset has been used for validation in all the cases.

From the model numerical verification results, we can conclude that the model demonstrates a sufficiently good accuracy of representation of the dynamic behavior of fluorescence of phytoplankton using the high frequency data measurements. However, the accuracy depends on the length of the training dataset. From Fig. 8 and 9, it is clear that the accuracy of our model decreases as we reduce the length of the training dataset. This is partly because of the frequently missing data points. In system identification, the quality of the identified model depends heavily on the quality of the dataset. Frequently missing data points degrades the quality of the dataset which in turns affects the quality of the identified model.

Another important point is that the behaviour of the fluoresence signal is significantly different from year to year. While most of the inputs were very periodic, the fluorescence signal didn't follow the same way. So, a model trained on the earlier years may perform less satisfactorily in the later years during the validation phase. This limitation impacts directly the accuracy of the prediction. This problem can be overcome by using data of longer length i.e., $10 - 15$ years or more. Other factors may also had an impact on the algal growth. For example, unmeasured parameters like zooplankton, direction of wind, wind velocity *etc.* may have played an important role on the growth of the phytoplankton. It was reported in the literature Fitch and Moore [2007], Moser et al. [2012] that wind direction has an impact on the growth. Considering all these factors may improve the prediction accuracy in the future.

5   Conclusion

This paper presents the first development of a dynamical model for phytoplankton bloom from system identification point of view. The developed model takes into account the effect
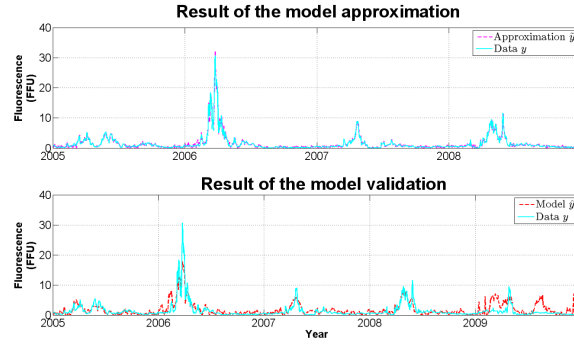
Figure 9   The result of the model numerical verification (training dataset is of 4 years)
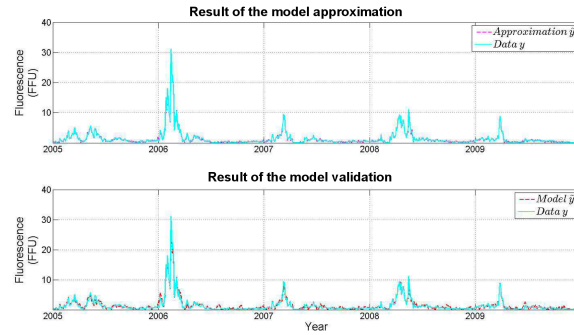

Figure 10   The result of the model numerical verification (training dataset is of 5 years)

of the various water characteristics such as dissolved oxygen, water salinity, pH, sea level, air and water temperature and luminosity along with information about three different nutrients (namely nitrate, silicate and phosphate). A generic "black-box" modeling approach has been used and an ARMAX structure of the model has been selected. The obtained model has been able to predict successfully the fluorescence level (which characterizes phytoplankton biomass) with a good prediction error.

In future works, more sophisticated model structures have to be analyzed. This paper did not include any biological hypothesis in the modeling process. This is under consideration for future investigation. This work approximated the dynamics of the model from the dataset. In our research ahead, an alternative for this approximation may include the development of ordinary differential equation based dynamical model and comparing the result with the dataset for validation purpose. Regarding inputs of the model, the vectors of inputs can be extended by including wind velocity and direction of the wind, since they may play an important role in the accumulation of biomass for the phytoplankton. Finally, some improvements can be made by considering the impact of correlation among the various inputs on parameter estimates.

Acknowledgement

References

Hafiz Ahmed, Rosane Ushirobira, Denis Efimov, Damien Tran, and Jean-Charles Massabuau. Dynamical model identification of population of oysters for water quality monitoring. In *Control Conference (ECC), 2014 European*, pages 152–157, June 2014. doi: 10.1109/ECC.2014.6862479.

Hafiz Ahmed, Rosane Ushirobira, Denis Efimov, Damien Tran, Mohamedou Sow, Pierre Ciret, and Jean-Charles Massabuau. Monitoring biological rhythms through the dynamic model identification of an oyster population. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 47(6):939–949, 2017.

Rita R Colwell. Global climate and infectious disease: the cholera paradigm*. *Science*, 274(5295):2025–2031, 1996.

Geoffrey T Evans and John S Parslow. A model of annual plankton cycles. *Biological oceanography*, 3(3):327–347, 1985.

MJR Fasham, HW Ducklow, and SM McKelvie. A nitrogen-based model of plankton dynamics in the oceanic mixed layer. *Journal of Marine Research*, 48(3):591–639, 1990.

Dillon T Fitch and J Keith Moore. Wind speed influence on phytoplankton bloom dynamics in the southern ocean marginal ice zone. *Journal of Geophysical Research: Oceans (1978–2012)*, 112(C8), 2007.

Tania Hernandez-Farinas, Dominique Soudant, Laurent Barille, Catherine Belin, Alain Lefebvre, and Cedric Bacher. Temporal changes in the phytoplankton community along the french coast of the eastern english channel and the southern bight of the north sea. *ICES Journal of Marine Science: Journal du Conseil*, 71(4):821–833, 2014.

Amit Huppert, Bernd Blasius, and Lewi Stone. A model of phytoplankton blooms. *The American Naturalist*, 159(2):156–171, 2002.

Amit Huppert, Bernd Blasius, Ronen Olinky, and Lewi Stone. A model for seasonal phytoplankton blooms. *Journal of Theoretical Biology*, 236(3):276–290, 2005.

Joseph HW Lee, Yan Huang, Mike Dickman, and AW Jayawardena. Neural network modelling of coastal algal blooms. *Ecological Modelling*, 159(2):179–201, 2003.

Lennart Ljung. *System identification*. Springer, 1998.

GAO Moser, AM Ciotti, MFC Giannini, RT Tonini, and J Harari. Changes in phytoplankton composition in response to tides, wind-induced mixing conditions, and freshwater outflows in an urbanised estuarine complex. *Brazilian Journal of Biology*, 72(1):97–111, 2012.

Nitin Muttil and Kwok-wing Chau. Neural network and genetic programming for modelling coastal algal blooms. *International Journal of Environment and Pollution*, 28(3): 223–238, 2006.

Nitin Muttil and Joseph HW Lee. Genetic programming for analysis and real-time prediction of coastal algal blooms. *Ecological modelling*, 189(3):363–376, 2005.

Oliver Nelles. *Nonlinear system identification: from classical approaches to neural networks and fuzzy models*. Springer, 2001.

K. Rousseeuw, E. Caillault, A. Lefebvre, and D. Hamad. Hybrid hidden markov model for marine environment monitoring. *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, 1(8):204–213, 2015.

Eric Walter and Luc Pronzato. *Identification of parametric models*. Springer Verlag New-York, 1997.

Wun-Cheng Wang. *Effect of turbidity on algal growth*. Illinois State Water Survey, 1974.

FD Wilde, DB Radtke, J Gibs, and RT Iwatsubo. National field manual for the collection of water-quality data-field measurements: Us geological survey techniques of water-resources investigations, book 9, chap. *A6 (variously paginated)*, 1998.

Note

[1] http://www.ifremer.fr/difMarelCarnot/
[2] see 1
[3] http://conference-marelcarnot2014.fr/Donnees_MAREL.html