

Article

Construction of Multi-Year Time-Series Profiles of Suspended Particulate Inorganic Matter Concentrations Using Machine Learning Approach

Pannimpullath R. Renosh ^{1,2,*} , Frédéric Jourdin ¹, Anastase A. Charantonis ³, Khalil Yala ², Aurélie Rivier ^{4,5}, Fouad Badran ², Sylvie Thiria ⁶, Nicolas Guillou ⁴ , Fabien Leckler ¹ , Francis Gohin ⁵ and Thierry Garlan ¹

¹ Service Hydrographique et Océanographique de la Marine (SHOM), 29228 Brest, France; frederic.jourdin@shom.fr (F.J.); fabien.leckler@shom.fr (F.L.); thierry.garlan@shom.fr (T.G.)

² Conservatoire National des Arts et Métiers (CNAM), 75003 Paris, France; yala-khalil@hotmail.fr (K.Y.); fouad.badran@cnam.fr (F.B.)

³ École Nationale Supérieure d'Informatique pour l'Industrie et l'Entreprise (ENSIIE), 91000 Évry, France; anastase.charantonis@ensiie.fr

⁴ Cerema, Direction Eau Mer et Fleuves, ER, Laboratoire de Génie Côtier et Environnement, Technopôle Brest-Iroise, 29280 Plouzané, France; aurelie.rivier@gmail.com (A.R.); nicolas.guillou@cerema.fr (N.G.)

⁵ IFREMER, Centre de Bretagne, Technopôle Brest-Iroise, 29280 Plouzané, France; Francis.Gohin@ifremer.fr

⁶ Laboratoire d'Océanographie et du Climat: Expérimentations et Approches Numériques (LOCEAN), 75005 Paris, France; sylvie.thiria@ocean-ipsl.upmc.fr

* Correspondence: pr.renosh@gmail.com

Received: 22 September 2017; Accepted: 8 December 2017; Published: 15 December 2017

Abstract: Hydro-sedimentary numerical models have been widely employed to derive suspended particulate matter (SPM) concentrations in coastal and estuarine waters. These hydro-sedimentary models are computationally and technically expensive in nature. Here we have used a computationally less-expensive, well-established methodology of self-organizing maps (SOMs) along with a hidden Markov model (HMM) to derive profiles of suspended particulate inorganic matter (SPIM). The concept of the proposed work is to benefit from all available data sets through the use of fusion methods and machine learning approaches that are able to process a growing amount of available data. This approach is applied to two different data sets entitled “Hidden” and “Observable”. The hidden data are composed of 15 months (27 September 2007 to 30 December 2008) of hourly SPIM profiles extracted from the Regional Ocean Modeling System (ROMS). The observable data include forcing parameter variables such as significant wave heights (H_s and H_{s50} (50 days)) from the Wavewatch 3-HOMERE database and barotropic currents (U_{bar} and V_{bar}) from the Iberian–Biscay–Irish (IBI) reanalysis data. These observable data integrate hourly surface samples from 1 February 2002 to 31 December 2012. The time-series profiles of the SPIM have been derived from four different stations in the English Channel by considering 15 months of output hidden data from the ROMS as a statistical representation of the ocean for ≈ 11 years. The derived SPIM profiles clearly show seasonal and tidal fluctuations in accordance with the parent numerical model output. The surface SPIM concentrations of the derived model have been validated with satellite remote sensing data. The time series of the modeled SPIM and satellite-derived SPIM show similar seasonal fluctuations. The ranges of concentrations for the four stations are also in good agreement with the corresponding satellite data. The high accuracy of the estimated 25 h average surface SPIM concentrations (normalized root-mean-square error—*NRMSE* of less than 16%) is the first step in demonstrating the robustness of the method.

Keywords: suspended particulate inorganic matter; self-organizing maps; Hidden Markov Model; machine learning; English Channel; ROMS

1. Introduction

An operational quantification of water turbidity is essential in many aspects of ocean and coastal management. For instance, in water quality monitoring, the European Marine Strategy Framework Directive (MSFD) requires operational measurements of long time series of turbidity levels (European Commission 2008, Annex III). The turbidity of the water column is also a key parameter in estimating the underwater visibility, as it is an optical measurement of water clarity. More specifically, marine operations involving divers or optical sensors may require the operational forecasts of the optical beam and diffuse attenuation coefficients of light traveling down the water column [1,2]. In this case, the decision-making may require parameter values, with explicit error specifications, in areas with possibly high temporal and spatial variability.

The quantitative analysis and forecast of turbidity profiles with its main optical properties and biochemical concentrations at any spatial and temporal scales are difficult tasks. There are coupled hydrodynamical and biophysical models of primary production and sediment transport for the forecast and quantification of turbidity. These models need dedicated research teams and thorough validation. The validation of these profiles at different depth levels is difficult. Surface validation is possible using ocean color satellites in cloud-free conditions at different spatiotemporal resolutions. In situ measurements of turbidity are furthermore scarce. However, more and more operational data and products are becoming available, in particular, those from satellite remote sensing and purely hydrodynamical models of the general circulation and wave parameters (e.g., <http://marine.copernicus.eu> for Europe and <http://data.shom.fr> for France).

The concept proposed here is to benefit from all of these available data sets through the use of fusion methods and machine learning approaches that are able to process this growing amount of data. For instance, in hydrology, recent works make use of machine learning methods such as Support Vector Regression, Random Forest Regression and Artificial Neural Networks. These works aim at improving spatial classification [3] and interpolation [4] of in situ data. When time is a factor, these data-driven approaches can also use analog Kalman filters [5] and analog Hidden Markov Models (HMMs) [6]. In optical remote sensing, machine learning is used to develop accurate algorithms for specific regional surface waters observed with hyperspectral [7] and multi-spectral [8] sensors. In the ocean color community, fusion methods and machine learning are emerging topics. Potential benefits of these methods include filling data gaps caused by cloud cover, inferring vertical distributions of biogeochemical variables observed at sea surface, emulate models of the upper ocean dynamics at high spatial resolution and developing relationships between environmental variables [9–12]. Further applications can be found in carbon dynamics [13–15], coastal turbidity [16,17], phytoplankton blooms [18–21], harmful algal blooms [22,23], and primary production [24].

In this field of research, the present work is the first attempt in recovering time series of in-depth turbidity profiles. For that, an HMM time-series analysis is combined with a previous classification of modeled turbidity profiles. This classification is performed by self-organizing maps (SOMs), which is an efficient means of interpreting similar patterns in complex multivariate data sets [25]. The purpose of the present investigation is to benefit from the hindcast output of a coupled model, previously validated in a research laboratory, and to combine it with the near-real-time (NRT) output of purely hydrodynamical models and satellite measurements, in order to produce an operational system of NRT forecasts (Figure 1). It should be noted that model hindcast results can be replaced by in situ measurements of the vertical profile when such a time series is available and long enough to be representative of turbidity range variations at the geographical point of study.

More precisely, the idea is to extend the application of hindcast results towards an operational and statistical system that performs error processing and error output. This concept is tested here on time series of suspended particulate inorganic matter (SPIM) vertical profiles at four locations in the English Channel (western Europe; Figure 2). The method consists of unsupervised classification and statistical analysis of a hindcast time series of SPIM profiles issued from the hydro-sedimentary numerical model set up by Guillou et al. in 2015 [26]. The analysis is then performed with the concomitant time series of

purely hydrodynamical model data and satellite-derived SPIM measurements. The statistical analysis allows for the determination of matrices of probabilities that can be further used in the processing of data for any other given period of time.

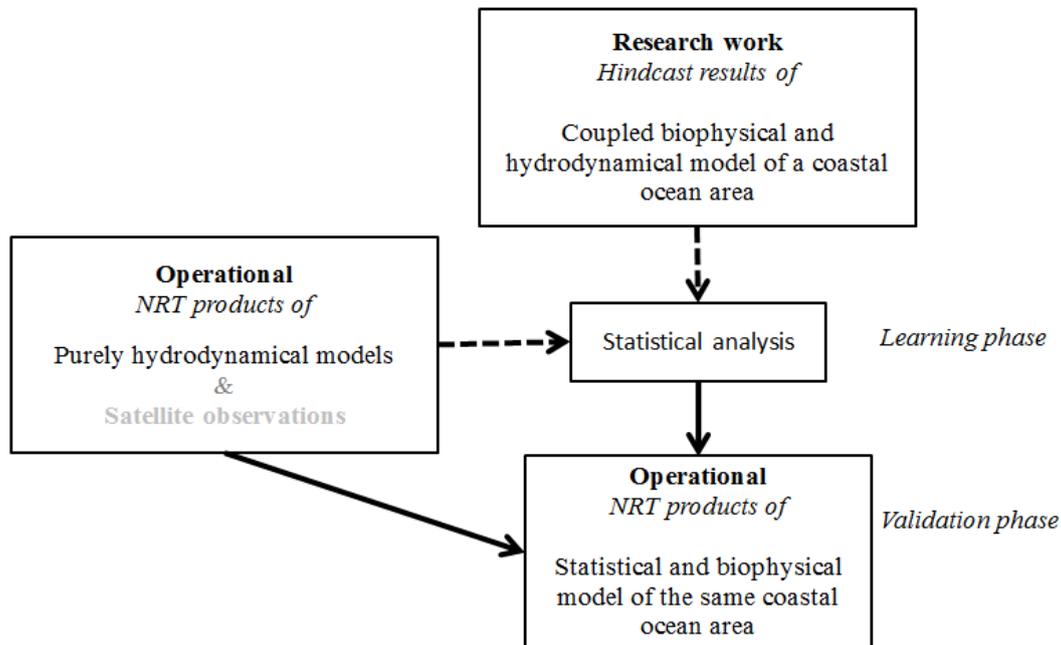


Figure 1. Schematic aim of the method. Plain arrows indicate data input of the operational statistical model. Dashed arrows indicate data input of the prior statistical analysis. Satellite measurements (in grey characters) can or can not be included in the system.

This paper is organized in the following way. Sections 2 and 3 describe the study area, data set used and the statistical methods implemented for the analysis. Two experiments were performed: with and without satellite measurements in the statistical system. The results of these two experiments are displayed and discussed in Section 4, followed by the conclusions.

2. Study Area

The English Channel is a mega-tidal sea experiencing a tidal current of ≈ 1.0 m/s and a tidal range from 3 to 9 m [27–31]. This continental shelf sea also experiences large temporal variability in particle concentrations and particle-size class distributions [30,31]. In the English Channel, tidal currents and coastal waves primarily control particle resuspension and turbidity [30,32,33]. The vertical distributions of SPM in water columns of the English Channel largely depend upon two physical processes: re-suspension of bottom particles and advection of these particles by currents and waves. In the present investigation, we have selected four different stations (M1, M2, M3 and M4) from the highly dynamic coastal waters of the English Channel (Figure 2).

These four points correspond to the locations retained by Guillou et al. [26] for the local evaluation of numerical model predictions. These stations are, in particular, affected by different combinations of storms and tides [26]. Station M1, located near the Isle of Bréhat, Brittany coast, is strongly influenced by the spring/neap tidal cycle and is also exposed to swells coming from the west. Station M2, located in front of Barfleur, is inside a zone of strong tidal currents and moderately strong waves coming from the west. Station M3 is located east of the Isle of Wight in an area where the SPM concentration remains high compared to the other three stations throughout the year. This station is subject to tides and swells. The last station, M4, is located in the eastern part of the English Channel, where the tidal action is reduced in comparison to the other stations and the waves are weak to moderate [26,33,34].

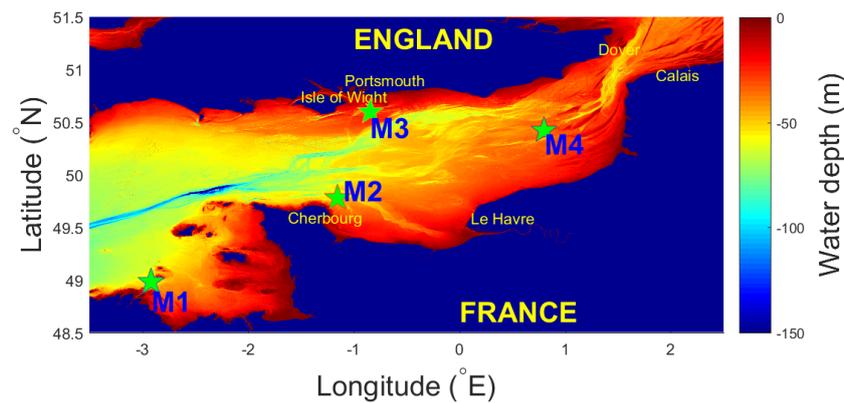


Figure 2. Study area map: four stations M1, M2, M3 and M4 are marked as green stars. The color bar indicates the bathymetric depth [35].

3. Data and Methodology

In HMM terminology, it is common to define two sets of data: observable data and hidden data (Figure 3). The hidden data are prediction targets. In our case, these are the vertical profiles of silts and sand concentrations extracted from a three-dimensional (3D) hydro-sedimentary numerical model. However, the PROFHMM approach by Charantonis et al. [19] retained in the present study does not use direct outputs of the oceanographic numerical model. It instead uses inferred statistical properties in terms of conditional probabilities between two consecutive time steps and between hidden and observable data. The observable data are the available and independent times series of observation values. These observations can be provided either by sensors or by numerical models and are also assumed to be noisy or partly erroneous. The method PROFHMM works on transition and emission probabilities between hidden data, which corresponds to the unknown parameters (Section 3.1), and observable classes, which correspond to forcing parameters (Section 3.2) [19]. For this purpose, we have classified these hidden and observable data using SOMs [36] (Section 3.3). At each time step, we extract the best matching unit. The transition and emission probabilities have been derived from the best matching units of hidden and observable data from time series analyses during a learning phase. From the sequence of concurrent observations, the “Viterbi Algorithm” [37] (Section 3.4) finds the most likely sequence of hidden states during the construction phase.

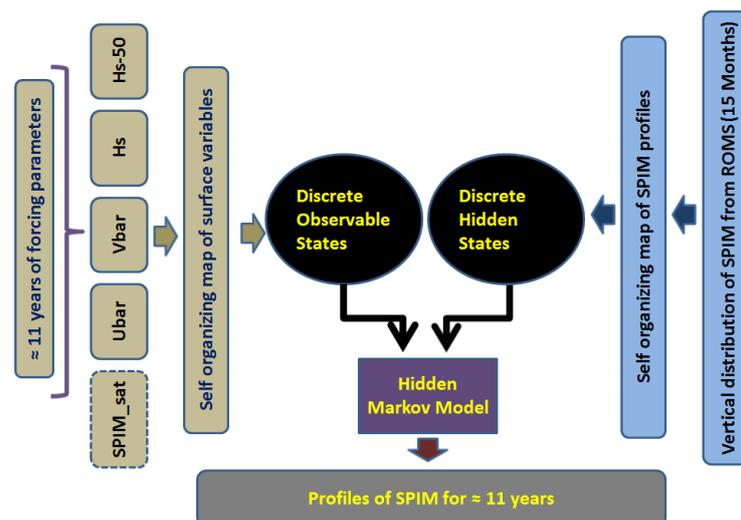


Figure 3. Schematic representation of the methodology. The observable variable shown in dotted outline is used only for the assimilation experiment (i.e., experiment 2).

3.1. Hidden Data

The word “hidden” comes from the HMM terminology. It can include different types of data sets, mainly model outputs. Here hidden data are selected as outputs of the 3D ROMS set up by Guillou et al. [26] in the English Channel. Each level corresponds to the SPIM concentration at a particular depth. Eleven depth levels are thus considered in sigma coordinates. The model assumes that the flow is turbulent over a rough bed. The particle assemblages are furthermore considered as spherically non-cohesive quartz components and the solid particles are not affected by the water-sediment mixture [26]. This 3D sedimentary ROMS model has been applied to the English Channel with spatial and temporal resolutions of 3 km and 30 s (with an output frequency of 1 h), respectively. The area selected for the model implementation extends from -3.3°W to 3°W and from 48.41°N to 51°N . This area is extended over a wider window between -4°W and 4.7°W and 48.40°N and 52.85°N to provide the boundary conditions. This numerical model derives particle concentrations for six grain-size classes of particle diameters ranging from 0 to 2 mm (Table 1).

Table 1. The particle grain-size classes of Regional Ocean Modeling System (ROMS) hydro-sedimentary model.

Sand Type	Diameter Range (μm)	Mean Diameter (μm)
Sand1	0–50	25
Sand2	50–100	75
Sand3	100–200	150
Sand4	200–500	350
Sand5	500–1000	750
Sand6	1000–2000	1500

We have selected the total SPIM concentrations for each depth level. These have been derived by the summation of all size classes of particles at a particular depth level.

$$SPIM_D = \sum_{i=1}^6 Sand(D, i) \quad (1)$$

where $SPIM_D$ is the total SPIM at depth D . These data begin from 27 September 2007 to 30 December 2008. This time span includes prominent winter storms observed as extreme event in that decade [38,39]. The components of tides, winds and waves are incorporated in this model. The interactions between waves and currents within and outside the bottom boundary layers are parameterized on the basis of Soulsby et al. [40]. Wave parameters are incorporated from regional Wavewatch III (WW3) simulations of the English Channel and the North Bay of Biscay using a spatial configuration based on a regular square grid with a mesh size of 2 mn entitled “NORGAS 2 mn”.

3.2. Observable Data

Here the word “observable” is the counterpart of “hidden” in the HMM terminology. It can deal with any kind of data, such as those issued from model predictions. The present study uses operational model outputs and satellite data as observable data. The satellite data (Section 3.5) is mainly used for validation. When it is used as observable data, this leads to an assimilation process.

3.2.1. Significant Wave Height (H_s)

The wave parameter used for the present study has been collected from the output of the WW3, version 4.09. The model has been run at IFREMER in the framework of “Marine Renewable Energies”. This is a spectral model of sea-state simulation that works on the basis of the conservation equation for wave action density. The scheme used for the configuration of the model is a propagation diagram

explicit for an unstructured grid [41]. The unstructured mesh grid (HOMERE) adapts the resolution at different scales from the coastal zone (refined mesh ≈ 200 m) to wide open ocean waters (≈ 10 km of mesh). We have selected the hourly time series of significant wave height (H_s) from the WW3 model output from the English Channel. The time series of $Hs50$ has been extracted from H_s , which is as the running average of H_s from the preceding 50 days and is arbitrarily defined by Rivier et al. in 2012 [34]:

$$Hs50_{(d)} = \frac{\sum_{i=0}^{49} Hs(d-i) \times (50-i)}{\sum_{i=0}^{49} (50-i)} \quad (2)$$

where $Hs(d-i)$ is relatively weighted $(50-i)$ for day $d-i$. The period of 50 days is reported by Rivier et al. [34] so as to give the best coefficient of determination in a linear regression between satellite data and the modeled output for 10 locations in the English Channel.

3.2.2. Barotropic Currents

The Iberian–Biscay–Irish (IBI) physical ocean reanalysis system is a regional reanalysis system at a spatial/horizontal resolution of $1/12^\circ$ and vertical resolution of 75 z – levels over the European North-East Atlantic. It is run by the company Mercator Ocean. The physical IBI reanalysis product was generated through the IBIRYS ocean reanalysis system for the 2002–2012 period [42]. The physical reanalysis products were generated using the physical model of the Nucleus for European Modeling of the Ocean (NEMO) [43]. This is freely available from the Copernicus Marine Environment Monitoring Service (CMEMS) website. These products (IBI_REANALYSIS_PHYS_005_002) consist of 3D monthly and daily mean fields of temperature, salinity, sea surface height, zonal, and meridional velocity components. Hourly means of surface fields such as the sea surface height, surface temperature, currents, and barotropic velocities, are also provided [44]. For the present study, we have extracted the hourly time series of barotropic velocities (northward $Vbar$ and eastward $Ubar$ components) from stations M1, M2, M3 and M4.

3.3. Self-Organizing Maps

PROFHMM uses SOMs in order to classify the available concurrent vertical distributions into the hidden states of the HMM and all observation vectors into the observable states. The SOMs are unsupervised classification algorithms that cluster data into discrete classes [36]. The discretization of the hidden and observable data has been done with 1100 states. By applying SOMs, we obtain a topological map containing states. Each state has a corresponding referent vector, whose value approximates the mean value of the data attributed to this class during the training. Two neighboring classes on the two-dimensional (2D) topological map will have referent vectors that are close in the data space. Thus, in PROFHMM, we have two maps, hidden and observable, respectively corresponding to the SOMs of the vertical SPIM and the SOMs of the input data, which are used to generate sequences of indexes of the hidden and observable states. By projecting sequences of data vectors on their corresponding topological maps, we can obtain a sequence of indexes corresponding to the class of each of the data vectors. The sequences obtained are used as inputs to train the HMM.

3.4. Hidden Markov Model

The Markov model used here is a stochastic model that assumes the first-order Markovian property, meaning that each consecutive state of the model depends solely on the previous state of the model, such as $P(X_t | X_1 X_2 \dots X_{t-1}) = P(X_t | X_{t-1})$, where P is the conditional probability and X_t is the state of the model at time t .

Expanding this principle, an 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 that has a statistical link with the hidden states. In order to model an HMM, a priori knowledge of the

transitions between the hidden states of the model and the probabilities of each observable state that have been emitted from a given hidden state must be found. The determination of these probabilities requires a data set containing concurrent sequences of hidden and observable states (i.e., the learning phase). If these a priori probabilities have been estimated, the Viterbi algorithm [37], which is an algorithm often associated with HMMs, can find the most likely sequence of unobserved states, given a sequence of concurrent observations. The Viterbi algorithm is a dynamic back-propagation algorithm whose objective is the inference of the most likely sequence of hidden states from a sequence of observable states. A normal computation of this probability would require the calculation of the probability of all possible sequences. The Viterbi algorithm circumvents this by calculating, in each time step, the maximum probability of reaching each possible state from the previous state and storing it for each state, and then iterating to the next step of the sequence. It is then able to retrieve the most likely way to reach each state at the end of the trajectory. The state with the maximum probability is selected, and then the algorithm backtracks, finding the preceding states.

3.5. Satellite Data

The satellite-derived non-algal SPM, also called the SPIM, is a multi-sensor product obtained by applying a semi-analytical algorithm to SeaWiFS, MODIS-Aqua, MERIS and VIIRS radiances. SPM was estimated following the semi-analytical algorithm proposed initially by Gohin et al. [45] and was modified and validated on in situ data collected in the coastal waters of the Bay of Biscay, the English Channel, and the western Mediterranean Sea [46]. This satellite product has been evaluated in detail by Jafar-Sidik et al. [47] through comparisons of a large set of in situ measurements of turbidity and SPM captured at the location of the Liverpool Bay mooring in the coastal waters of the Irish Sea. In this procedure, absorption and backscattering by phytoplankton were derived from preliminary estimations of Chlorophyll-a concentrations [48]. Then, SPM was obtained from radiance at 550 and 670 nm. Depending on the level of the retrieved SPM, the final SPM was chosen at 550 nm if both SPM concentrations (at 550 and 670 nm) were less than 4 mg/L. In the cases in which the SPM concentration (at 550 and 670 nm) was more than 4 mg/L, SPM(670) was chosen. Finally the satellite-derived SPM of each sensor available along the studied period (one or two sensors simultaneously) were interpolated by kriging [49]. Interpolation by kriging, widely used in the domain of spatial analysis, provides the best linear unbiased estimator given a space-time autocorrelation. Interpolations are carried out at each pixel on the 1 km grid of the images. Satellite data observed within 5 days before and 5 days after the day of interest were used to build the data sets used for the interpolation [49].

4. Results and Discussion

The present work focuses on two different experiments of constructing long-term SPIM profiles using the PROFHMM method. The first experiment (experiment 1) derives vertical profiles of SPIM using four modeled variables: H_s , H_{s50} , U_{bar} and V_{bar} (from NEMO and WW3 operational models). The main advantage of experiment 1 was that we could validate the derived surface SPIM with satellite data. The second experiment (experiment 2) focused on the assimilation of satellite data along with the operational model outputs. For this purpose, we integrated the satellite SPIM as an observable variable along with the four variables used in experiment 1. The present work mainly concentrates on experiment 1 because of the significant importance of the validation of the derived product. The results of the output from the learning phase of experiment 1 are explained in Section 4.1.

4.1. Output from Learning Phase

The vertical sections of the SPIM for 15 months between the original ROMS model output and PROFHMM have been compared. These sections show good agreement with each other for all of the four stations. Figures 4–7 show the vertical section of the SPIM between ROMS and PROFHMM at points M2 and M3. The derived SPIM exhibited seasonal and tidal fluctuations in their magnitude for all of the four stations.

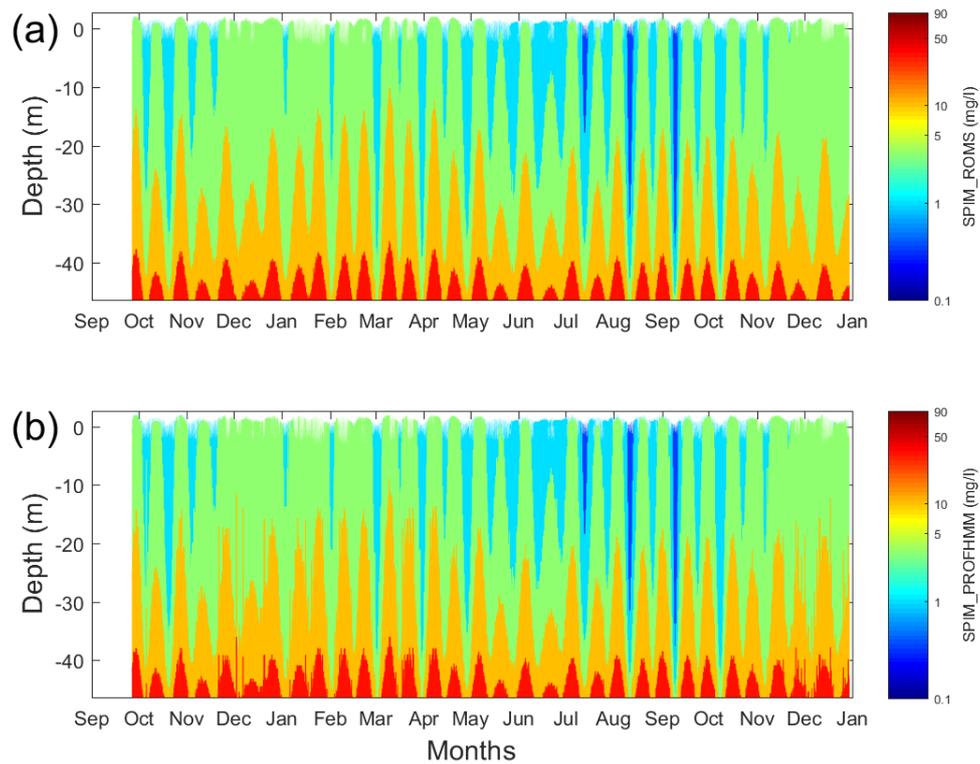


Figure 4. Learning phase of original suspended particulate inorganic matter (SPIM) section from Regional Ocean Modeling System (ROMS) data in (a), and corresponding PROFHMM-derived SPIM section in (b) from the point M2.

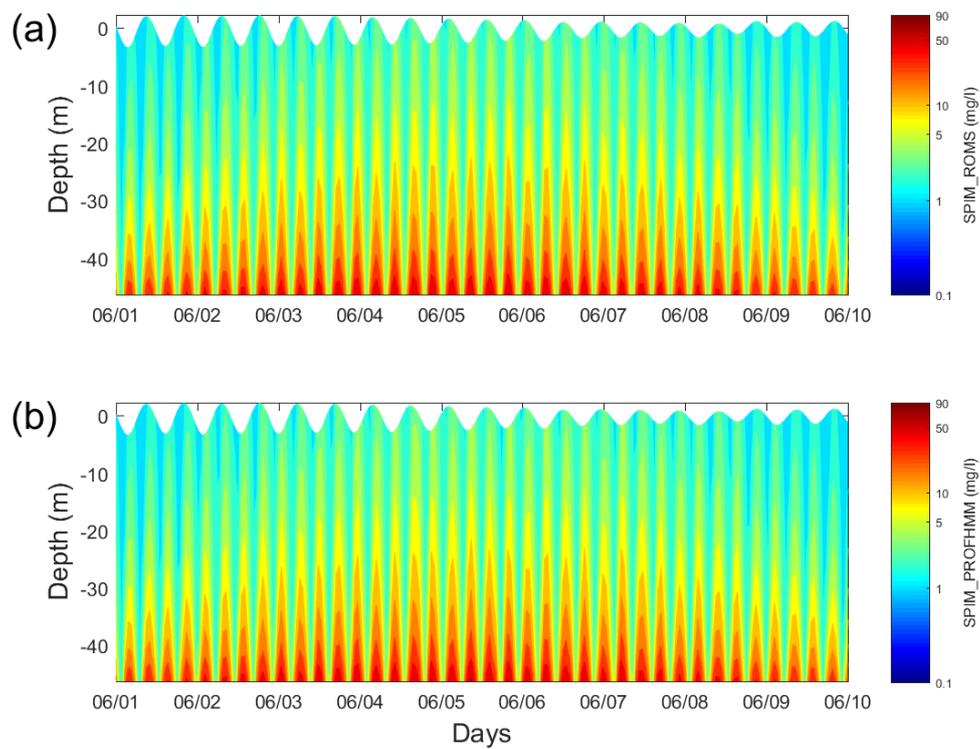


Figure 5. Learning phase of original suspended particulate inorganic matter (SPIM) section from Regional Ocean Modeling System (ROMS) data in (a), and corresponding PROFHMM-derived SPIM section in (b) from point M2 for 10-day spring tidal phase between 1 June 2008 and 10 June 2008.

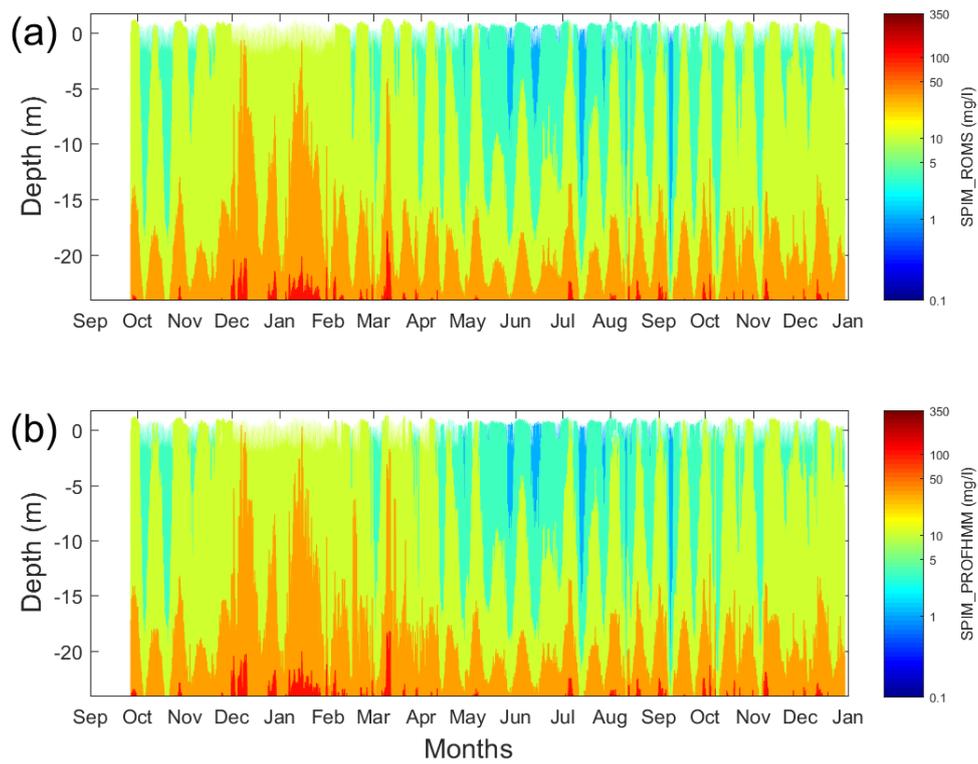


Figure 6. Learning phase of original suspended particulate inorganic matter (SPIM) section from Regional Ocean Modeling System (ROMS) data in (a), and corresponding PROFHMM-derived SPIM section in (b) from the point M3.

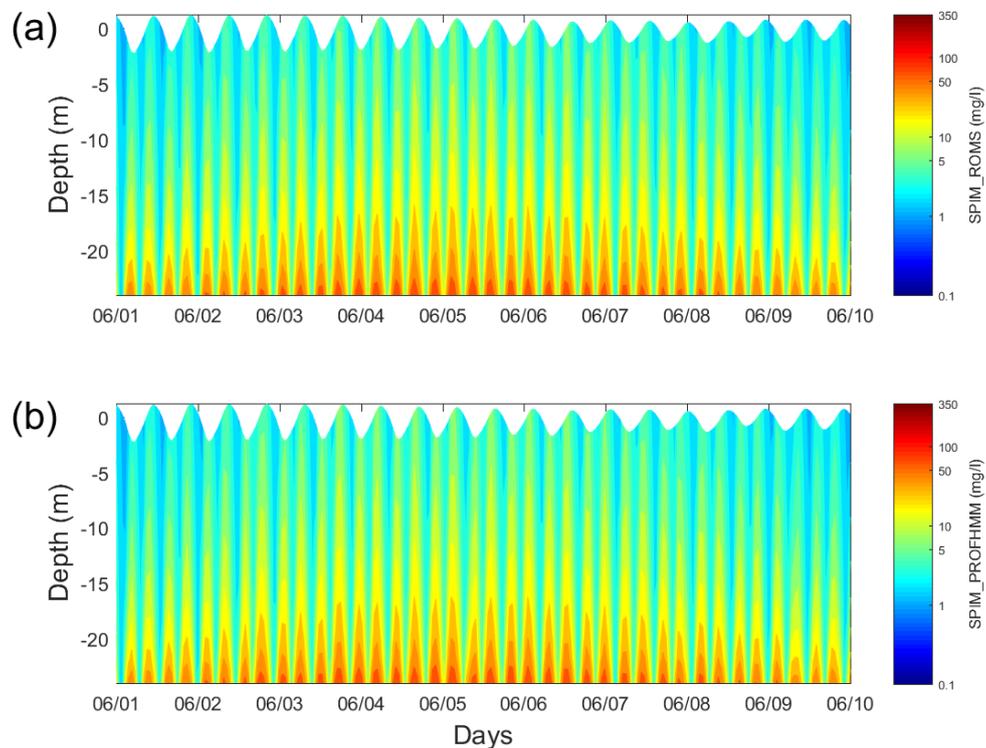


Figure 7. Learning phase of original suspended particulate inorganic matter (SPIM) section from Regional Ocean Modeling System (ROMS) data in (a), and corresponding PROFHMM-derived SPIM section in (b) from point M3 for 10-day spring tidal phase between 1 June 2008 and 10 June 2008.

The root-mean-square error (*RMSE*) and normalized root-mean-square error (*NRMSE*) between these two data sets (ROMS and PROFHMM) for the entire section of the learning phase have been performed using the following equations:

$$RMSE(mg/l) = \sqrt{\frac{\sum_{i=1}^n (SPIM_{ROMS_i} - SPIM_{PROFHMM_i})^2}{n}} \quad (3)$$

$$NRMSE(\%) = \frac{RMSE}{SPIM_{ROMS_{max}} - SPIM_{ROMS_{min}}} \quad (4)$$

The *RMSE* value of the SPIM between the ROMS and PROFHMM data for the M1 station was 1.43 mg/L and the *NRMSE* was 1.67%, which is small compared with the range of SPIM observed at the M1 station, ranging from 0.08 to 85.81 mg/L. In a similar way, the *RMSE* and *NRMSE* of SPIM (range from 0.33 to 83.37 mg/L) for the M2 station were 1.25 mg/L and 1.51%, respectively. The station M3 experienced high concentrations of SPIM compared with the other three stations because this station was more shallow and also more exposed to waves compared with the other three stations, and high particle resuspension occurred. The SPIM in this station showed a range of 0.55 to 339.48 mg/L with a *RMSE* of 5.10 mg/L and *NRMSE* of 1.50%. The *RMSE* and *NRMSE* of SPIM (range from 0.01 to 87.07 mg/L) derived for the M4 station were 1.68 mg/L and 1.93%, respectively. All four stations showed very low values of errors, which confirms that the methodology works well to quantify SPIM concentrations in the learning phase.

4.2. Output from Out-of-Learning Phase

The construction of multi-year SPIM profiles has been executed with two different types of experiments.

4.2.1. Experiment 1

In this experiment, almost 11 years of time series data of the four variables (*Hs*, *Hs50*, *Ubar* and *Vbar*) were used as observables, and 15 months of SPIM profiles from the ROMS hydro-sedimentary model were considered as hidden data [34] in order to derive ≈ 11 years of SPIM profiles.

Time Series of Surface SPIM

We have pointed out that the methodology works well in all of the four stations in the learning phase, relying on these four observable variables. The profiles of SPIM have been derived through PROFHMM for ≈ 11 years with a sampling interval of 1 h. The validation of SPIM at each depth is out of the scope of this article because of the unavailability of sedimentary validation data for the 11-year period considered for the present study. For this reason, we have selected the surface satellite *SPIM_sat* for the validation of our statistical model. The hourly *SPIM_sat* values have been derived by the linear interpolation of daily data. As it is smooth over a semi-diurnal tidal period, we have computed 25 h of running average time series of surface PROFHMM data for the comparison with *SPIM_sat* data. As a consequence, the amplitude range of turbidity fluctuations having periods below 25 h (particularly from semi-diurnal tides) cannot be quantitatively validated here.

The time series of surface SPIM derived through PROFHMM and *SPIM_sat* showed good temporal agreement in terms of magnitude order with each other for the four stations (Figure 8), except for station M4 in particular, which showed low SPIM concentrations. The point M1 followed a slight overestimation in *SPIM_PROFHMM* compared with *SPIM_sat* during the summer (Figure 8a). At station M4, low values of surface SPIM (< 0.1) were observed in the ROMS model. This discrepancy may have been due to the inability of the satellite to measure very low concentrations of surface SPIM [46]. The time series of *SPIM_PROFHMM* and *SPIM_sat* showed very similar seasonal fluctuations throughout the 11 year period. We observed very high concentrations of SPIM during the winter and

comparatively low SPIM during the spring and summer. This was observed throughout the time series for the four stations (Figure 8). The periods of turbid events are well explained/predicted by PROFHMM using only the significant wave height and barotropic currents. It can furthermore be seen that there is a better fit between *SPIM_PROFHMM* and *SPIM_sat* during the time of learning phase (the red colored rectangle in each panel of Figure 8). An increase in the length of hidden data (more than 15 months) would lead to a better fit between *SPIM_PROFHMM* and *SPIM_sat* all the way along the time series.

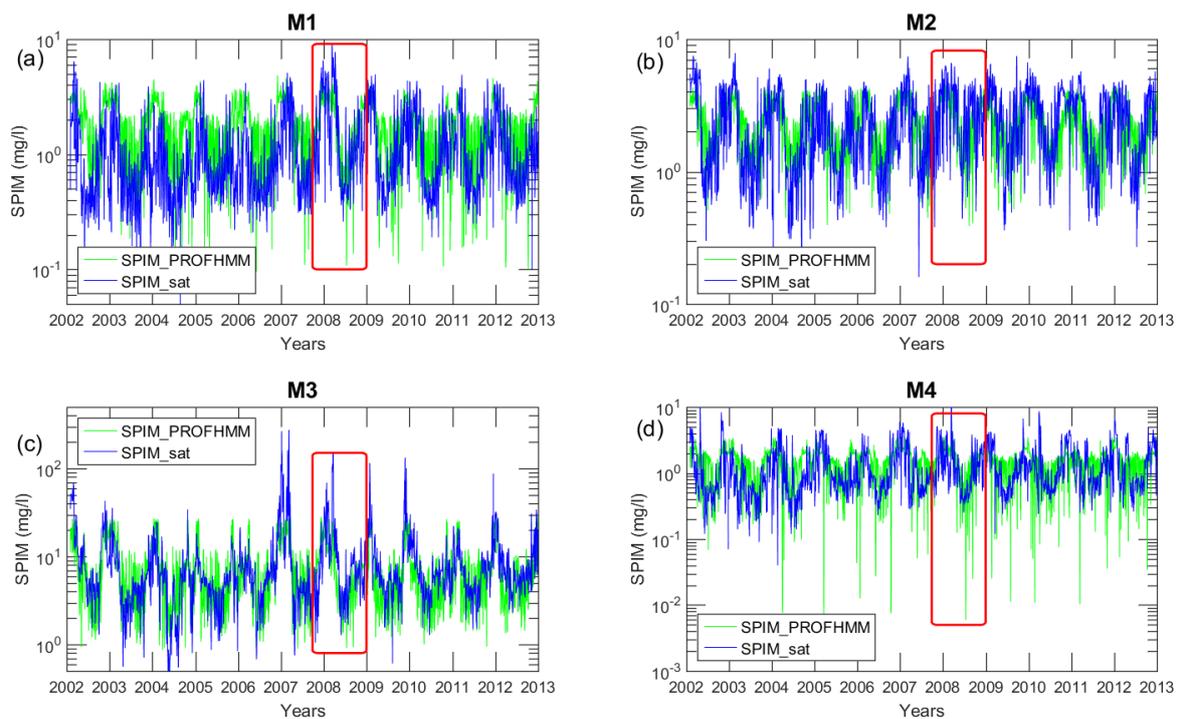


Figure 8. Time series of 25 h running average surface suspended particulate inorganic matter (SPIM) derived through PROFHMM method along with satellite SPIM at station M1 in (a), at M2 in (b), at M3 in (c), and at M4 in (d). The breadth of the red colored rectangle indicates the time span of the learning phase.

The discrepancy between surface *SPIM_PROFHMM* and *SPIM_sat* is noticeable at point M3 in particular. The point M3 is located in the turbid area near the Isle of Wight, which exhibits strong near-surface SPIM with yearly averaged values of 10–15 mg/L [26]. The winter storm effects have been derived throughout the prediction by the PROFHMM model, but there are some discrepancies in terms of intensity of these turbid events between surface *SPIM_PROFHMM* and *SPIM_sat*, in particular during winter for the years 2007, 2008, 2009 and 2010 when the satellite observed high concentrations of SPIM.

These discrepancies could have several causes:

- The satellite algorithm overestimates the SPIM concentration when finer particles dominate the suspension during storms [33].
- The kriging interpolation of satellite images becomes imprecise during storms because of the high cloud coverage [45].
- The assumptions in the ROMS model (non-cohesive particles, settling velocity, critical shear stress, etc.) are partially valid [33].

Error Distribution of Surface SPIM

The error distribution between *SPIM_PROFHMM* and *SPIM_sat* has been derived using a normalized histogram (Figure 9).

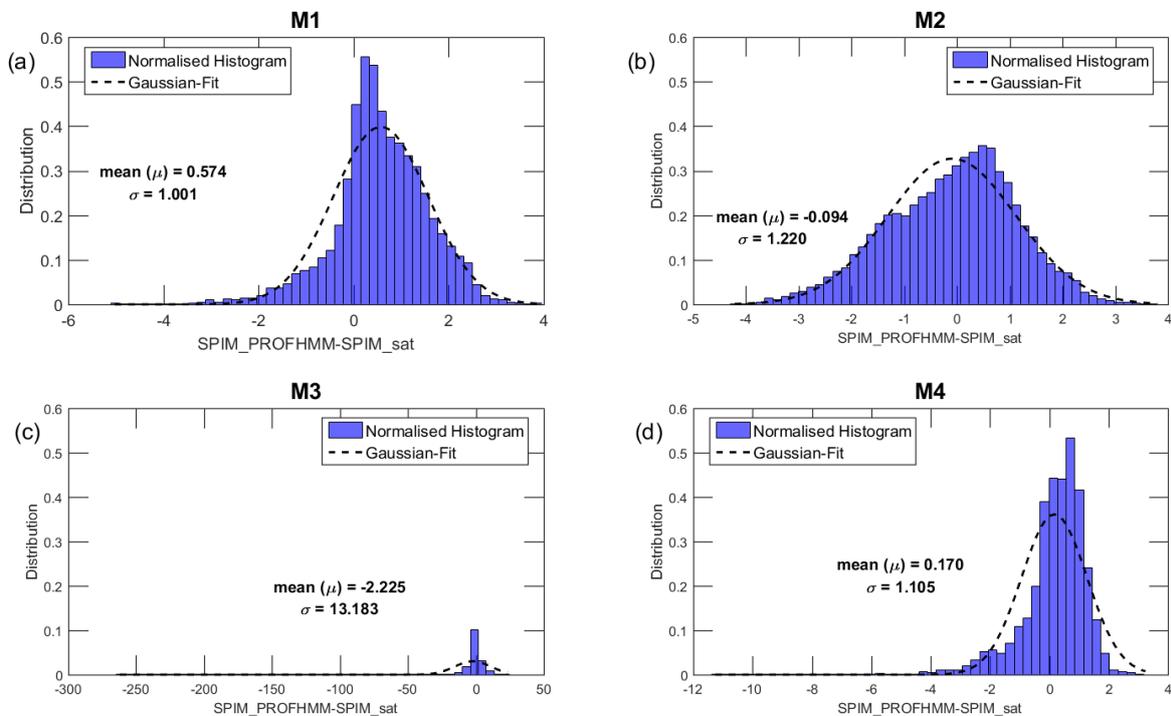


Figure 9. The error distribution of suspended particulate inorganic matter (SPIM) between PROFHMM and satellite SPIM along with the Gaussian fit for (a) M1, (b) M2, (c) M3, and (d) M4 stations.

The mean and standard deviations of the error derived for each station show mean values near to zero, except at the station M3 (Table 2). The standard deviation of the error for M1 is (± 1.00), for M2 is (± 1.22), for M3 is (± 13.18) and for M4 is (± 1.11). Because of the small number of but large outliers in the distribution at M3, a large value of the standard deviation was obtained at this point. The median absolute deviations showed a better estimation of error (Table 2), as they were less affected by few outliers. These showed values of around 1 or 2 mg/L.

Table 2. The error estimated for surface suspended particulate inorganic matter (SPIM) between PROFHMM and satellite for four stations.

<i>SPIM_PROFHMM-SPIM_sat</i>	M1	M2	M3	M4
<i>SPIM_sat</i> range (mg/L)	0.05–9.26	0.13–7.85	0.38–274	0.04–14.76
Mean/Bias (mg/L)	0.57	−0.09	−2.23	0.17
Standard deviation (mg/L)	1.00	1.22	13.18	1.11
Median (mg/L)	0.54	0.03	−0.76	0.36
Median absolute deviation (mg/L)	0.56	0.83	2.21	0.53
Root-mean-square error (mg/L)	1.15	1.22	13.37	1.12
Normalized root-mean-square error (%)	12.53	15.85	4.88	7.60

The *RMSE* and *NRMSE* between *SPIM_PROFHMM* and *SPIM_sat* have been derived for the four stations. The *RMSE* values were found to be low (Table 2) compared with the range of SPIM observed in each station. The normalized *RMSE* (using satellite min and max values) also showed a percentage less than 16% for all of the four stations.

4.2.2. Experiment 2

In this experiment, we incorporated *SPIM_sat* as a variable along with the variables of experiment 1. The time series of SPIM profiles were derived for ≈ 11 years. The vertical sections of the time series showed more or less similar results to those of experiment 1. In particular, in both experiments, high frequency turbidity fluctuations (from semi-diurnal tides) displayed ranges that were qualitatively consistent with ranges of the ROMS and with the tidal current (Figures 10 and 11). Moreover, the introduction of smooth *SPIM_sat* data (without semi-diurnal tidal signal) did not seem to weaken the resulting tidal component in the final SPIM signal. The error derived for the learning phase of experiment 2 showed higher values at M1, M2 and M3 and slightly lower values for station M4 compared with experiment 1 (Table 3). This discrepancy at station M4 may have been due to the addition of inconsistent smooth/non-semi-diurnal tides (*SPIM_sat*) into the system as an observable parameter, while the other four observable variables were incorporated with this semi-diurnal tidal signal. At station M4, the surface SPIM values were sometimes relatively low; the satellite data likely failed to retrieve these concentrations accurately.

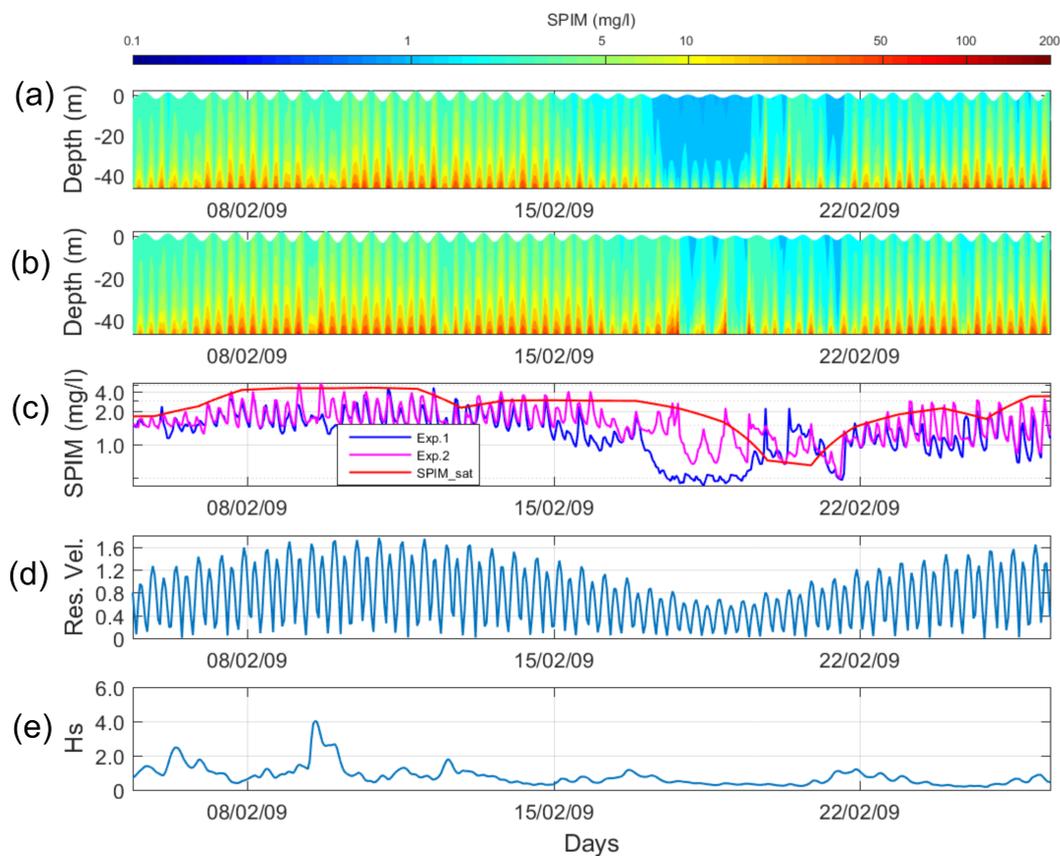


Figure 10. Section of 21 days of suspended particulate inorganic matter (SPIM) from out-of-learning phase for Exp. 1 in (a) and Exp. 2 in (b), surface SPIM (Exp. 1, Exp. 2 and *SPIM_sat*) in (c), resultant barotropic velocity in (d), and significant wave height (H_s) in (e) at station M2.

The assimilation of timely smooth satellite data within experiment 2 may have weakened the resulting SPIM variabilities compared with variabilities obtained with experiment 1. Figures 10 and 11 give an example for the cases of both tidal cycle (spring-neap) and storm events. The resultant velocity (*Res.Vel.*) is derived from the barotropic currents using the following equation:

$$Res.Vel. = \sqrt{Ubar^2 + Vbar^2} \quad (5)$$

Figure 10 is a representation of the tidal signal (spring-neap) from point M2, and Figure 11 is the storm event from point M3. In Figure 10, it is observed that during neap tides, experiment 1 produced clear tidal information for the surface SPIM compared with that of experiment 2. In Figure 10c, while comparing the two curves between days 15 February and 22 February, the tidal signal in experiment 2 seemed to be perturbed by the introduction of smoother satellite information. The addition of smooth (without semi-diurnal tidal signal) satellite SPIM as an observable variable in experiment 2 produced SPIM that always tried to fit the satellite data, keeping semi-diurnal tidal information. However, this trade-off seemed to perturb some tidal periods (higher periods seen around 18 February 2009). The introduction of smooth *SPIM_sat* as an observable parameter may have confused the system while deriving the SPIM concentration. In Figure 11e, we can see higher values of *Hs* than are typically considered as a wave event. During this event, experiment 2 led to a weaker tidal signal compared to experiment 1 (Figure 11a).

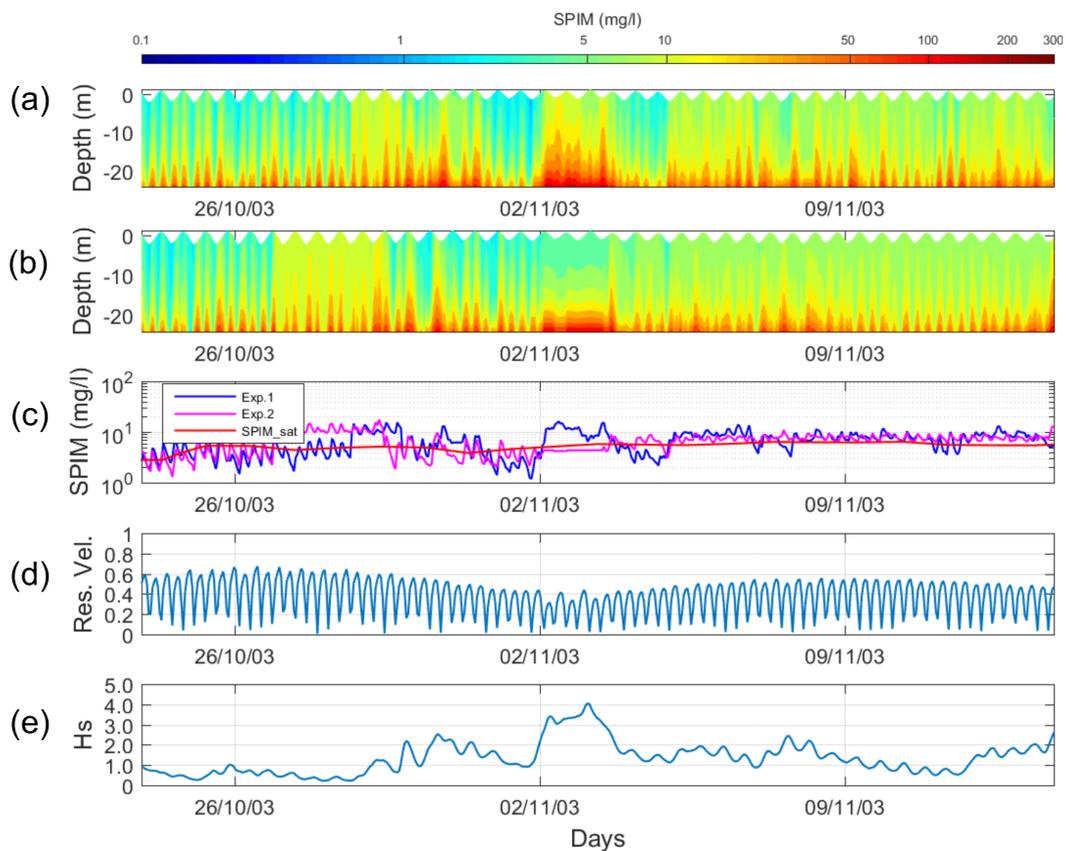


Figure 11. Section of 21 days suspended particulate inorganic matter (SPIM) from out-of-learning phase for Exp. 1 in (a) and Exp. 2 in (b), surface SPIM (Exp. 1, Exp. 2 and *SPIM_sat*) in (c), resultant barotropic velocity in (d) and significant wave height (*Hs*) in (e) at station M3.

Table 3. The error estimated for the entire section (15 months temporally and 11 depth levels) of learning phase for four stations.

Experiment	Different Error Estimators	M1	M2	M3	M4
Exp. 1	Root-mean-square error (mg/L)	1.43	1.25	5.10	1.68
Exp. 1	Normalized root-mean-square error (%)	1.67	1.51	1.50	1.93
Exp. 2	Root-mean-square error (mg/L)	1.92	1.47	5.75	1.39
Exp. 2	Normalized root-mean-square error (%)	2.24	1.77	1.70	1.60

5. Conclusions

In the English Channel, we demonstrated the feasibility of deriving vertical SPIM profiles over a long time period (≈ 11 years) while only using three local and average hydrodynamic parameters (H_s , U_{bar} and V_{bar}). In these highly dynamical waters, the SPIM-profile time evolution follows statistical properties that are given by a previous analysis of a shorter time series (≈ 1 year) hindcast coming from a dedicated hydro-sedimentary deterministic model. By nature, a priori statistics are sensitive to the range and magnitude of processes sampled by the hindcast simulation. For this reason, the choice of this hindcast must be representative of the final extension period of the statistical simulation. For instance, here we used hidden data from 27 September 2007 to 30 December 2008. This period included a winter storm observed as the most prominent extreme event in the last decade [38,39]. The validation of the resulting and averaged (25 h) SPIM concentrations with satellite non algal particles (NAP) concentrations measured at the ocean surface led to *NRMSE* of less than 16%, which was a first step in demonstrating the robustness of this methodology. Further validation tests with in situ measurements of the whole vertical profile are required. When operational forecasts of H_s , U_{bar} and V_{bar} are available, this statistical method would then be able to deliver turbidity forecasts. Furthermore, the method is able to easily “assimilate” satellite data. Here, a first experiment used NAP concentrations from L4 satellite products that do not contain the semi-diurnal tidal signal. A further experiment with the assimilation of NAP measured by future geostationary satellites (such as the MTG/FCI—Meteosat Third Generation/Flexible Combined Imager, [50]) is desired. The processing would then benefit from L4 products having a high temporal (≈ 1 h) resolution. Data collected over nights and cloudy areas are then needed ([51] with DINEOF processing for completion over cloudy areas). This data-driven approach is general in the sense that it can analyze a hindcast simulation from any other validated hydro-sedimentary model, having different resolutions, coordinates, configurations, assumptions (laws), particle type/shape, and so forth. Such a possibility should be further tested, for instance, with an experiment in a different area where turbid river plumes are a major process, a case which has not been tested here.

Acknowledgments: The authors would like to thank Direction générale de l’armement (DGA) of France for the financial support.

Author Contributions: S.T., F.B., A.A.C. and K.Y. conceived the general method. P.R.R., F.J., and A.A.C. designed the experiment. P.R.R. performed the experiment and data analysis. P.R.R. and F.J. wrote the paper. N.G. and A.R. provided the ROMS hydro-sedimentary data and also contributed to improving the paper quality. F.G. provided the satellite SPIM data. F.L. provided WW3/HOMERE data. T.G. contributed to the improvement of the article by personal communication.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

SPM	Suspended particulate matter
SOMs	Self-organizing maps
HMM	Hidden Markov model
SPIM	Suspended particulate inorganic matter
ROMS	Regional Ocean Modeling System
<i>NRMSE</i>	Normalized root-mean-square error
WW3	Wavewatch III
<i>RMSE</i>	Root-mean-square error
NAP	Non-algal particles

References

1. Zaneveld, J.R.V.; Pegau, W.S. Robust underwater visibility parameter. *Opt. Express* **2003**, *11*, 2997–3009.
2. Lee, Z.; Shang, S.; Hu, C.; Du, K.; Weidemann, A.; Hou, W.; Lin, J.; Lin, G. Secchi disk depth: A new theory and mechanistic model for underwater visibility. *Remote Sens. Environ.* **2015**, *169*, 139–149.
3. Sahoo, S.; Jha, M.K. Pattern recognition in lithology classification: Modeling using neural networks, self-organizing maps and genetic algorithms. *Hydrogeol. J.* **2017**, *25*, 311–330.
4. Diesing, M.; Kroger, S.; Parker, R.; Jenkins, C.; Mason, C.; Weston, K. Predicting the standing stock of organic carbon in surface sediments of the North–West European continental shelf. *Biogeochemistry* **2017**, *135*, 183–200.
5. Lguensat, R.; Viet, P.H.; Sun, M.; Chen, G.; Fenglin, T.; Chapron, B.; Fablet, R. Data-Driven Interpolation of Sea Level Anomalies Using Analog Data Assimilation. Technical Report. 2017. Available online: <https://hal.archives-ouvertes.fr/hal-01609851> (accessed on 22 September 2017).
6. Fablet, R.; Viet, P.H.; Lguensat, R. Data-driven Models for the Spatio-Temporal Interpolation of satellite-derived SST Fields. *IEEE Trans. Comput. Imaging* **2017**, *3*, 647–657.
7. Wang, X.; Zhang, F.; Ding, J. Evaluation of water quality based on a machine learning algorithm and water quality index for the Ebinur Lake Watershed, China. *Sci. Rep.* **2017**, *7*, 12858.
8. Wattelez, G.; Dupouy, C.; Lefevre, J.; Ouillon, S.; Fernandez, J.M.; Juillot, F. Application of the Support Vector Regression Method for Turbidity Assessment with MODIS on a Shallow Coral Reef Lagoon (Voh-Kone-Pouembout, New Caledonia). *Water* **2017**, *9*, 737.
9. Le Traon, P.Y.; Antoine, D.; Bentamy, A.; Bonekamp, H.; Breivik, L.A.; Chapron, B.; Corlett, G.; Dibarboure, G.; DiGiacomo, P.; Donlon, C.; et al. Use of satellite observations for operational oceanography: Recent achievements and future prospects. *J. Oper. Oceanogr.* **2015**, *8* (Suppl. S1), s12–s27.
10. She, J.; Allen, I.; Buch, E.; Crise, A.; Johannessen, J.A.; Le Traon, P.-Y.; Lips, U.; Nolan, G.; Pinardi, N.; Reißmann, J.H.; et al. Developing European operational oceanography for Blue Growth, climate change adaptation and mitigation, and ecosystem-based management. *Ocean Sci.* **2016**, *12*, 953–976.
11. Ban, Y.; Marullo, S.; Eklundh, L. European Remote Sensing: Progress, challenges, and opportunities. *Int. J. Remote Sens.* **2017**, *38*, 1759–1764.
12. McCarthy, M.J.; Colna, K.E.; El-Mezayen, M.M.; Laureano-Rosario, A.E.; Méndez-Lázaro, P.; Otis, D.B.; Toro-Farmer, G.; Vega-Rodriguez, M.; Muller-Karger, F.E. Satellite Remote Sensing for Coastal Management: A Review of Successful Applications. *Environ. Manag.* **2017**, *60*, 323–339.
13. Sasse, T.P.; McNeil, B.I.; Abramowitz, G. A novel method for diagnosing seasonal to inter-annual surface ocean carbon dynamics from bottle data using neural networks. *Biogeosciences* **2013**, *10*, 4319.
14. Parard, G.; Charantonis, A.A.; Rutgeron, A. Remote sensing algorithm for sea surface CO₂ in the Baltic Sea. *Biogeosci. Discuss.* **2014**, *11*, 12255–12294.
15. Jang, E.; Im, J.; Park, G.H.; Park, Y.G. Estimation of fugacity of carbon dioxide in the East Sea using in situ measurements and Geostationary Ocean Color Imager satellite data. *Remote Sens.* **2017**, *9*, 821.
16. Saulquin, B.; Fablet, R.; Ailliot, P.; Mercier, G.; Doxaran, D.; Mangin, A.; d’Andon, O.H.F. Characterization of time-varying regimes in remote sensing time series: Application to the forecasting of satellite-derived suspended matter concentrations. *IEEE J. STARS* **2015**, *8*, 406–417.
17. Chang, N.B.; Bai, K.; Chen, C.F. Integrating multisensor satellite data merging and image reconstruction in support of machine learning for better water quality management. *J. Environ. Manag.* **2017**, *201*, 227–240.
18. Blondeau-Patissier, D.; Gower, J.F.; Dekker, A.G.; Phinn, S.R.; Brando, V.E. A review of ocean color remote sensing methods and statistical techniques for the detection, mapping and analysis of phytoplankton blooms in coastal and open oceans. *Prog. Oceanogr.* **2014**, *123*, 123–144.
19. Charantonis, A.A.; Badran, F.; Thiria, S. Retrieving the evolution of vertical profiles of chlorophyll-a from satellite observations, by using hidden Markov models and self-organizing maps. *Remote Sens. Environ.* **2015**, *163*, 229–239.
20. Krasnopolsky, V.; Nadiga, S.; Mehra, A.; Bayler, E.; Behringer, D. Neural networks technique for filling gaps in satellite measurements: Application to ocean color observations. *Comput. Intell. Neurosci.* **2016**, *2016*, 29.
21. Kwon, Y.S.; Jang, E.; Im, J.; Baek, S.H.; Park, Y.; Cho, K.H. Developing data-driven models for quantifying *Cochlodinium polykrikoides* using the Geostationary Ocean Color Imager (GOCI). *Int. J. Remote Sens.* **2018**, *39*, 68–83.

22. Gokaraju, B.; Durbha, S.S.; King, R.L.; Younan, N.H. A machine learning based spatio-temporal data mining approach for detection of harmful algal blooms in the Gulf of Mexico. *IEEE J. STARS* **2011**, *4*, 710–720.
23. Vilas, L.G.; Spyrakos, E.; Palenzuela, J.M.T.; Pazos, Y. Support Vector Machine-based method for predicting *Pseudo-nitzschia* spp. blooms in coastal waters (Galician rias, NW Spain). *Prog. Oceanogr.* **2014**, *124*, 66–77.
24. Li, Z.; Cassar, N. Satellite estimates of net community production based on O₂/Ar observations and comparison to other estimates. *Glob. Biogeochem. Cycle* **2016**, *30*, 735–752.
25. Pearce, J.; Commodore, A.; Neelon, B.; Boaz, R.; Bozigar, M.; Wilson, S.; Svendsen, E. A novel approach for characterizing neighborhood-level trends in particulate matter using concentration and size fraction distributions: A case study in Charleston, SC. *Air Qual. Atmos. Health* **2017**, 1–12, doi: 10.1007/s11869-017-0503-y.
26. Guillou, N.; Rivier, A.; Gohin, F.; Chapalain, G. Modeling Near-Surface Suspended Sediment Concentration in the English Channel. *J. Mar. Sci. Eng.* **2015**, *3*, 193–215.
27. Desprez, M. Physical and biological impact of marine aggregate extraction along the French coast of the Eastern English Channel: short- and long-term post-dredging restoration. *ICES J. Mar. Sci.* **2000**, *57*, 1428–1438.
28. Seuront, L.; Schmitt, F.G. Multiscaling statistical procedures for the exploration of biophysical couplings in intermittent turbulence. Part II. Applications. *Deep Sea Res. II* **2005**, *52*, 1325–1343.
29. Korotenko, K.A.; Sentchev, A.V.; Schmitt, F.G. Effect of variable winds on current structure and Reynolds stresses in a tidal flow: Analysis of experimental data in the eastern English Channel. *Ocean Sci.* **2012**, *8*, 1025–1040.
30. Renosh, P.R.; Schmitt, F.G.; Loisel, H.; Sentchev, A.; Mériaux, X. High frequency variability of particle size distribution and its dependency on turbulence over the sea bottom during re-suspension processes. *Cont. Shelf Res.* **2014**, *77*, 51–60.
31. Renosh, P.R.; Schmitt, F.G.; Loisel, H. Intermittent particle dynamics in marine coastal waters. *Nonlinear Process. Geophys.* **2015**, *22*, 633–643.
32. Velegrakis, A.F.; Michel, D.; Collins, M.B.; Lafite, R.; Oikonomou, E.K.; Dupont, J.P.; Bishop, C. Sources, sinks and resuspension of suspended particulate matter in the eastern English Channel. *Cont. Shelf Res.* **1999**, *19*, 1933–1957.
33. Guillou, N.; Rivier, A.; Chapalain, G.; Gohin, F. The impact of tides and waves on near-surface suspended sediment concentrations in the English Channel. *Oceanologia* **2017**, *59*, 28–36.
34. Rivier, A.; Gohin, F.; Bryère, P.; Petus, C.; Guillou, N.; Chapalain, G. Observed vs. predicted variability in non-algal suspended particulate matter concentration in the English Channel in relation to tides and waves. *Geo-Mar. Lett.* **2012**, *32*, 139–151.
35. Biscara, L.; Schmitt, T.; Corréard, S.; Creach, R. Modèles numériques de bathymétrie pour la prévision Hydrodynamique du dispositif vigilance vagues-submersions. In Proceedings of the Actes des XIIIèmes Journées Nationales Génie Côtier—Génie Civil, Dunkerque, France, 2–4 July 2014; Volume 155, pp. 2–4.
36. Kohonen, T. The self-organizing map. *Proc. IEEE* **1990**, *78*, 1464–1480.
37. Viterbi, A. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans. Inf. Theory* **1967**, *13*, 260–269.
38. Haigh, I.; Nicholls, R.; Wells, N. Rising sea levels in the English Channel 1900 to 2100. *Proc. Inst. Civil Eng. Mar. Eng.* **2010**, *164*, 81–92.
39. Wadey, M.P.; Nicholls, R.J.; Haigh, I. Understanding a coastal flood event: The 10th March 2008 storm surge event in the Solent, UK. *Nat. Hazards* **2013**, *67*, 829–854.
40. Soulsby, R.L.; Hamm, L.; Klopman, G.; Myrhaug, D.; Simons, R.R.; Thomas, G.P. Wave-current interaction within and outside the bottom boundary layer. *Coast. Eng.* **1993**, *21*, 41–69.
41. Roland, A.; Cucco, A.; Ferrarin, C.; Hsu, T.W.; Liau, J.M.; Ou, S.H.; Umgiesser, G.; Zanke, U. On the development and verification of 2-D coupled wave-current model on unstructured meshes. *J. Mar. Syst.* **2009**, *78*, S244–S254.
42. Levier, B.; Benkiran, M.; Reffray, G.; Sotillo, M. *IBIRYS: A Regional High Resolution Reanalysis (Physical and Biogeochemical) over the European North East Shelf*; EGU: München, Germany, 2014.
43. Madec, G. *NEMO Ocean General Circulation Model Reference Manual*; Internal Report; LODYC/IPSL: Paris, France, 2008.
44. Sotillo, M.G.; Cailleau, S.; Lorente, P.; Levier, B.; Aznar, R.; Reffray, G.; AmoBaladrón, A.; Chanut, J.; Benkiran, M.; Alvarez-Fanjul, E. The MyOcean IBI Ocean Forecast and Reanalysis Systems: Operational products and roadmap to the future Copernicus Service. *J. Oper. Oceanogr.* **2008**, *8*, 63–79.

45. Gohin, F.; Loyer, S.; Lunven, M.; Labry, C.; Froidefond, J.M.; Delmas, D.; Huret, M.; Herbland, A. Satellite-derived parameters for biological modelling in coastal waters: Illustration over the eastern continental shelf of the Bay of Biscay. *Remote Sens. Environ.* **2005**, *95*, 29–46.
46. Gohin, F. Annual cycles of chlorophyll-a, non-algal suspended particulate matter, and turbidity observed from space and in-situ in coastal waters. *Ocean Sci.* **2011**, *7*, 705–732.
47. Jafar-Sidik, M.; Gohin, F.; Bowers, D.; Howarth, J.; Hull, T. The relationship between Suspended Particulate Matter and Turbidity at a mooring station in a coastal environment: Consequences for satellite-derived products. *Oceanologia* **2017**, *59*, 365–378.
48. Gohin, F.; Druon, J.N.; Lampert, L. A five channel chlorophyll concentration algorithm applied to SeaWiFS data processed by SeaDAS in coastal waters. *Int. J. Remote Sens.* **2002**, *23*, 1639–1661.
49. Saulquin, B.; Gohin, F.; Garrello, R. Regional Objective Analysis for Merging High-Resolution MERIS, MODIS/Aqua, and SeaWiFS Chlorophyll-a Data from 1998 to 2008 on the European Atlantic Shelf. *IEEE Trans. Geosci. Remote Sens.* **2011**, *49*, 143–154.
50. Ruddick, K.; Neukermans, G.; Vanhellefont, Q.; Jolivet, D. Challenges and opportunities for geostationary ocean colour remote sensing of regional seas: A review of recent results. *Remote Sens. Environ.* **2014**, *146*, 63–76.
51. Liu, X.; Wang, M. Analysis of ocean diurnal variations from the Korean Geostationary Ocean Color Imager measurements using the DINEOF method. *Estuar. Coast. Shelf Sci.* **2016**, *180*, 230–241.



© 2017 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).