Benefits of machine learning and sampling frequency on phytoplankton bloom forecasts in coastal areas

Jonathan Derot, Hiroshi Yajima, François G Schmitt

Benefits of machine learning and sampling frequency on phytoplankton bloom forecasts in coastal areas

Jonathan Derot[1,*], Hiroshi Yajima[1], François G. Schmitt[2].

(1) Estuary Research Center, Shimane University, 1060 Nishikawatsu-cho, Matsue, Shimane 690-8504, Japan

(2) CNRS, Univ. Lille, Univ. Littoral Cote d'Opale, UMR 8187, LOG, Laboratoire d'Océanologie et de Géosciences, F 62930 Wimereux, France

* Corresponding author. E-mail address: j.derot@soc.shimane-u.ac.jp

ORCID iDs

Derot Jonathan: https://orcid.org/0000-0001-6531-2732

Hiroshi Yajima: https://orcid.org/0000-0002-0361-1080

François G. Schmitt: http://orcid.org/0000-0001-6733-0598

Abstract

In aquatic ecosystems, anthropogenic activities disrupt nutrient fluxes, thereby promoting harmful algal blooms that could directly impact economies and human health. Within this framework, the forecasting of the proxy of chlorophyll a in coastal areas is the first step to managing these algal blooms. The primary goal was to analyze how phytoplankton bloom forecasts are impacted by different sampling frequencies, by using a machine learning model. The database used in this study was sourced from an automated system located in the English Channel. This device has a sampling frequency of 20 minutes. We considered 12 physicochemical parameters over a six-year period. Our forecast methodology is based on the random forest (RF) model and a sliding window strategy. The lag times for these sliding windows ranged from 12 hours to 3 months with four different sampling times until 1 day.

The results indicate that the optimal forecast was obtained for a 20 minutes time step, with an average R^2 of 0.62. Moreover, the highest values of fluorescence were predicted when the water temperature was approximately 11.8°C. Consequently, we demonstrated that the sampling frequency directly impacts the forecast performance of an RF model. Furthermore, this kind of model can recreate interactions that closely resemble biological processes. Our study suggests that the RF model can utilize the additional information contained in high-frequency datasets. The methodology presented here lays the foundation for the development of a numerical decision-making tool that could help mitigate the impact of these algal blooms.

Key words: Algal bloom, autonomous monitoring, English Channel, Random Forest model, individual conditional expectation plots, water quality management.

1. Introduction

The technological developments in recent decades, both numerical and material, help us to understand the complex processes present in aquatic ecosystems at a biological compartment level. In lacustrine environments, it has been noted that pairing of machine learning and high-frequency database inputs from automatic devices (or long-term sampling), engendered encouraging results in phytoplankton community forecasting (Yajima and Derot 2018; Thomas et al. 2018). For example, anthropogenic agricultural activities and the treatment of water and sewage, enrich nutrient levels in freshwater and coastal areas (Anderson et al. 2002; Smith et al. 2006; Roelke et al. 2010). These disruptive nutrient fluxes lead to eutrophication by promoting the development of toxic algae, causing harmful algal blooms (HAB) (Camargo and Alonso 2006; Schindler 2006). The size and intensity of these blooms has been increasing for over 20 years (Burkholder 2003; Glibert et al. 2005).

The occurrence of HABs has a negative socio-economic impact on drinking water, fisheries, agriculture, and tourism (Carmichael and Boyer 2016; Reynaud and Lanzanova 2017). Moreover, they are often associated with cyanobacteria proliferation (Backer et al. 2015). In the marine environment, the Prymnesiophyceae *Phaeocystis* is an organism that blooms in response to increased nutrient levels. This species generally impacts tourism because of the large quantities of foam that appears on beaches during these blooms (Veldhuis and Wassmann 2005). In some parts of the world, these algae generate losses in the aquaculture industry, which could potentially impact the economies of these countries (Chen et al. 2002). For many years, the problems relating to *Phaeocystis* were mostly confined to the English Channel (Lancelot et al. 1987; Lubac et al. 2008; Monchy et al. 2012; Danhiez et al. 2017). However, in recent years, this type of bloom has also been observed in other parts of the world, such as China and the Arabian Gulf (Lancelot et al. 2002; Schoemann et al. 2005).

In this context, the ability to forecast algal blooms is currently a major issue in ecology (Pennekamp et al. 2019). The development of HABs is often directly linked to nutrient pollution, also termed eutrophication (Heisler et al. 2008; Howarth et al. 2000; Lapointe et al. 2017). A numerical tool capable of understanding and forecasting HABs could help manage water quality, thereby enabling stakeholders to mitigate the impact of this toxic bloom. However, before creating this type of decision-making tool, it is imperative to focus on the prediction of more global biological processes, such as phytoplankton biomass. Classic hydro-ecological models work optimally for physical processes, but perform poorly when forecasts of the first echelon of the food web are involved (Shimoda and Arhonditsis 2016). This decrease in predictive performance can be explained by numerous complex interactions and nonlinear mechanisms between phytoplankton and environmental variables (Edwards et al. 2016). In addition, in open-ocean and coastal areas, there are strong currents and important phytoplankton migrations; therefore, it is increasingly complicated to forecast these biological processes using a machine learning model (Thomas et al. 2018).

Scientific literature contains a wide variety of numerical models concerning the prediction of phytoplankton biomass or phylum, including models such as the hydro-ecological (Bae and Seo 2018; Yajima and Choi 2013), autoregressive moving integrated moving average (ARIMA) (Chen et al. 2015), and random forest (RF) (Thomas et al. 2018; Yajima and Derot 2018; Shin et al. 2017; Kehoe et al. 2015; Rivero-Calle et al. 2015). There is also a wide variety of models based on neuronal networks: artificial neural networks (ANNs) (Shamshirband et al. 2019; G. Lee et al. 2016; S. Lee and Lee 2018); long short-term memory (LSTM) (Cho and Park 2019; Lee and Lee 2018; Cho et al. 2018); nonlinear autoregressive neural network (NAR) (Du et al. 2018), and deep belief network (DBN) (Zhang et al. 2016). Moreover, some of these models are coupled with a genetic algorithm (Lee et al. 2016) or wavelet-transform (Du et al. 2018; Shamshirband et al. 2019), or both (Recknagel et al. 2013). We can also find great diversity in the methods used to validate these models: coefficient of determination (R^2) (Du et al. 2018; Shamshirband et al. 2019; Lee et al. 2016; Lee and Lee 2018; Recknagel et al. 2013; Kehoe et al. 2015); root mean squared error (RMSE) (Du et al. 2018; Cho and Park 2019; Zhang et al. 2016; S. Lee and Lee 2018; Cho et al. 2018; Chen et al. 2015; Recknagel et al. 2013); mean absolute error (MAE) (Du et al. 2018; Shamshirband et al. 2019); mean squared error (MSE) (Lee et al. 2016; Rivero-Calle et al. 2015); mean relative error (MRE) (Zhang et al. 2016); absolute error peak (AEP) (Chen et al. 2015); and pseudo-R^2 (Thomas et al. 2018); area under curve (AUC) (Shin et al. 2017).

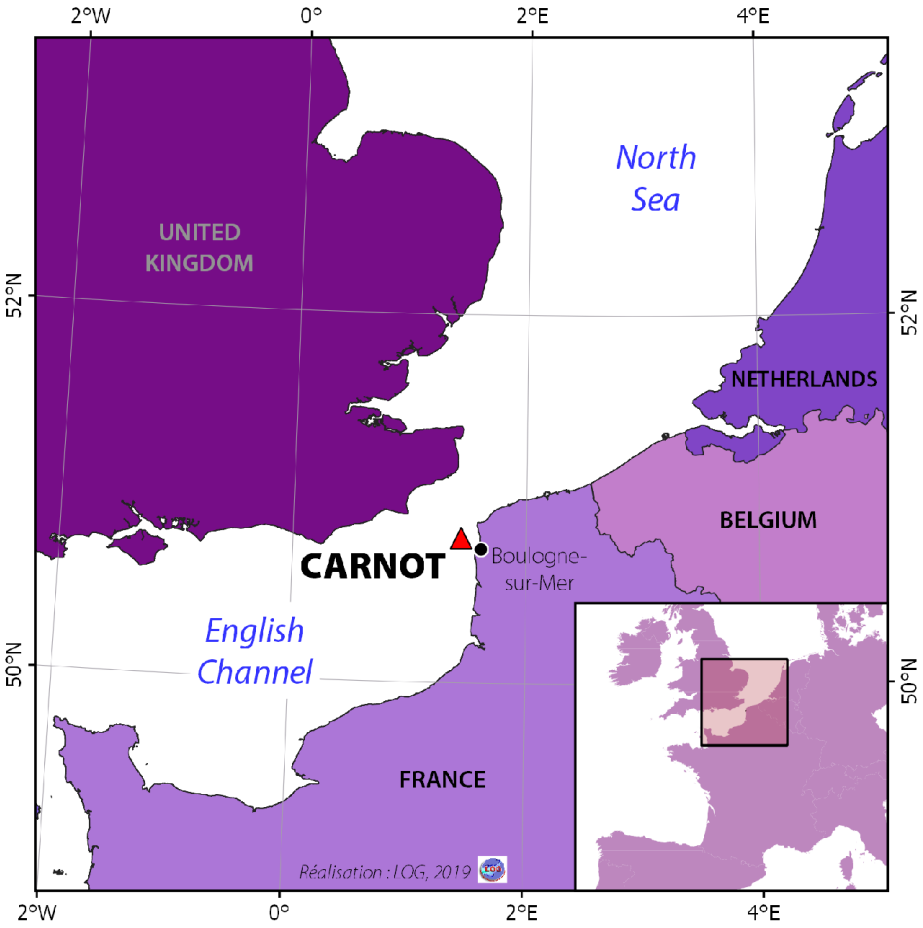
103 This highlights the lack of standardized protocol to forecast the parameters linked to
104 phytoplankton biomass in aquatic environments. Thus, it can be inferred that the use of this kind of
105 artificial intelligence based model in this branch of science is still in its infancy. However, the pairing
106 of high-frequency data with RF models seems to be an interesting alternative to forecast primary
107 production. One way to measure RF usefulness is the usage of the pseudo-R^2 coefficient that comes
108 from the cross-validation process, at an "out-of-bag" error level, and measures only the performance
109 of the learning phase (Breiman 2001). Although this coefficient has been used in several
110 environmental studies, some researchers are aware that pseudo-R^2 cannot be assessed as a true forecast
111 (Large et al. 2015; Teichert et al. 2016; Thomas et al. 2018). Therefore, in this study, we have chosen
112 to compare the raw data from an automated system to the output of the machine learning model, using
113 the coefficient of determination (R^2).
114 Automatic devices may have different sampling frequencies; for example, time steps of 10
115 minutes, 20 minutes or 4 hours can be found (Dur et al. 2007; Derot et al. 2016; Schmitt and Lefebvre
116 2016; Thomas et al. 2018). Currently, the impact of sampling frequency on the learning process is
117 poorly understood. Therefore, in this study, we have explored the capacity of the RF model to leverage
118 the supplementary information that is contained in the high sample frequency and water quality data.
119 Moreover, the predictive performance of this model was studied by varying the time steps used in our
120 database. Before creating a decision-making tool to help stakeholders with water quality management,
121 certain intermediate research stages are necessary. Furthermore, the coupling between machine
122 learning and hydrodynamic models exhibits encouraging results for the prediction of ecological
123 parameters directly linked to water quality (Cuttitta et al. 2018; Jia et al. 2018; Hanson et al. 2020).
124 Within this framework, the purpose of our study is to better understand the impact of sampling
125 frequency on algal bloom forecast capacity. Therefore, the results presented here could potentially
126 improve this type of coupled numerical model.

2. Material and methods

2.1. Automatic device and study area

The high-frequency dataset used in this study was producedby an automatic device called MAREL Carnot. MAREL is a French acronym for *Mesures Automatisées en Réseaux pour l'Environnement Littoral* (automated sampling network for coastal area). It belongs to a network of fixed platform networks along French coasts called COAST-HF (http://coast-hf.fr). The MAREL Carnot device used here is located in the eastern English Channel on the French coastal area. More specifically, this automatic system is situated at the exit of the Boulogne-sur-Mer harbor on the Carnot sea wall (50.7404 N; 1.5676 W) (Fig. 1). The Boulogne-sur-Mer harbor is the first fishing port in France. Consequently, it is subjected to significant anthropogenic pollution. Moreover, the English Channel has been affected by HAB generated by *Phaeocystis* for several decades (Lancelot et al. 1987; Lubac et al. 2008; Monchy et al. 2012; Danhiez et al. 2017). In the eastern part, this kind of bloom has been a recurrent event since the 1990s (Spilmont et al. 2009; Schmitt et al. 2011; Houliez et al. 2012; Grattepanche et al. 2011). This is one of the reasons for basing this kind of automated device in this area. It is pivotal to consider that the English Channel has very turbid waters and is subject to large tidal ranges. This has been tied to the fact that the bed of this sea is a continental shelf, with a maximal depth of 180 meters.

144



145

146 **Fig. 1.** Location of the MAREL Carnot automatic device, in the eastern English Channel at the

147 Boulogne-sur-Mer port exit.

2.1.1. MAREL Carnot data

The MAREL Carnot sensors are attached to a floating system that nestles in a tube fixed to the sea wall. The data are constantly recorded at a depth of 1.5 meters below sea level. However, the measurement of the photosynthetically active radiation (P.A.R) parameter is an exception and for obvious reasons, the sensor is not installed in the tube, but on the top of the sea wall. Each parameter is recorded at a frequency of 20 minutes, except for the three nutrient parameters (nitrates, silicates, and phosphates) which are recorded with a periodicity of 12 hours (Dur et al. 2007; Derot et al. 2015; Zongo and Schmitt 2011; Huang and Schmitt 2014). Table 1 lists all the parameters that were used in our study. The data presented here can be obtained from the following sites: https://data.coriolis-cotier.org and the Seanoe site provided by Lefebvre et al. (2015).

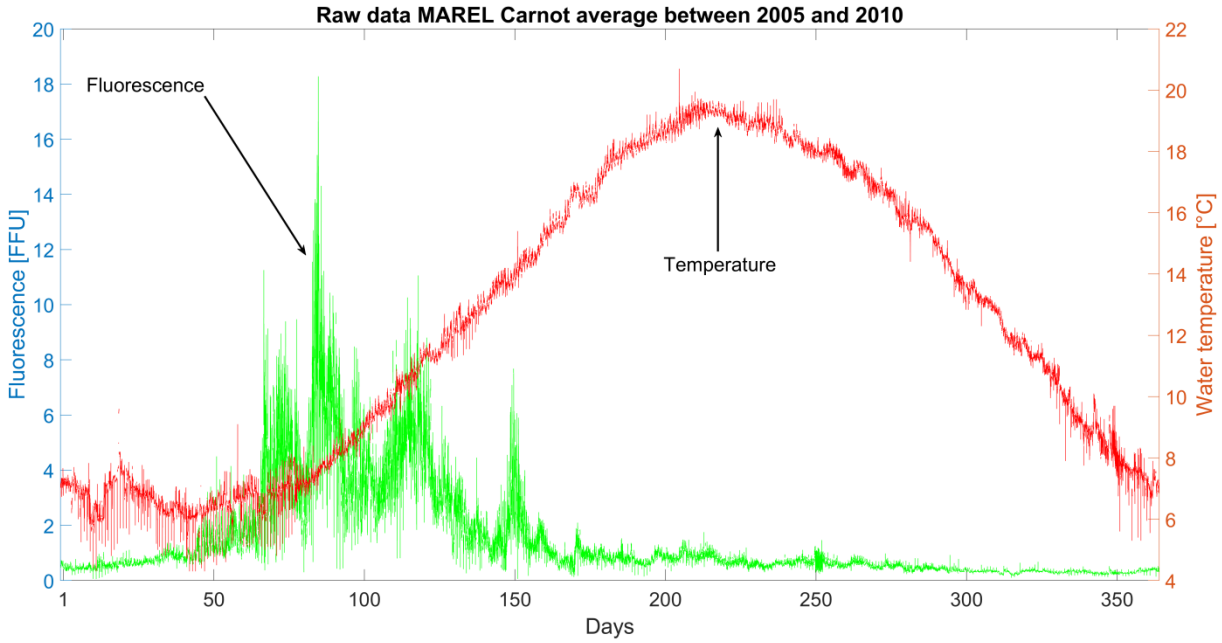
Not all the available parameters, recorded by the MAREL device, have been used to avoid the problems linked with collinearity. The term collinearity is used when the two predictors that are input into a machine learning model, display a strong correlation. In order to decrease the computation time and avoid creating an unstable model with a degraded predictive performance, it is better to avoid selecting predictors showing correlations (Kuhn and Johnson 2013). To this end, we discarded the readings for percentage of dissolved oxygen and salinity: they were too close to the concentration of dissolved oxygen and the conductivity readings, respectively. In each case, we selected the parameter with the most complete data set.

166 Time periods with missing data are an inherent problem for automatically generated datasets;
167 many factors can create these missing values: maintenance periods, internal system failures, and
168 vandalism (Dur et al. 2007; Derot et al. 2015). The percentage of missing data for each parameter is
169 listed in Table 1. The MAREL Carnot platform did not work for most of 2014. During the last four
170 years, the system experienced many problems; therefore, there is a high number of missing values for
171 many parameters. In order to avoid bias, our analysis only uses data between 2005 and 2010. In other
172 words, considering this 6-year period allows us to maintain consistency of the monitoring data used as
173 input to the RF model. In Fig. 2, we can see the averaged raw data for fluorescence and water
174 temperature for this time period of 6 years. Moreover, the type of machine learning used in this study
175 is able to manage the missing values (see Section 2.2.1 for further details). These missing values are
176 different for all the parameters. In some rare cases, some of these parameters could have more than
177 one month in raw missing values. In this context, we preferred to keep these missing values in the
178 inputs of the RF model, rather than using a numerical method to fill these gaps, which could create
179 greater bias.

180 **Table 1.** Physicochemical parameters (target signal and predictors) in the original MAREL Carnot
181 database, with number of readings, percentage of missing data between 2005 and 2010, sampling
182 frequency, and their associated units.

Parameters	Number of readings	Percentage of missing data	Sampling frequency	Units
Fluorescence	113530	13.60%	20 minutes	FFU
Water temperature	117176	10.83%	20 minutes	°C
P.A.R	62172	52.68%	20 minutes	µmol/s/m^3
Conductivity	114114	13.16%	20 minutes	mS/cm
Dissolved oxygen	111573	15.09%	20 minutes	mg/L
pH	97556	25.76%	20 minutes	pH
Turbidity	92992	29.23%	20 minutes	NTU
Wind direction	103304	21.38%	20 minutes	°
Wind speed	103716	21.07%	20 minutes	m/s
Nitrate	1982	45.70%	12 hours	µmol/L
Phosphate	1908	47.73%	12 hours	µmol/L
Silicate	1983	45.67%	12 hours	µmol/L

183



184

185 **Fig. 2**. Inter-annual mean (day-scale) of MAREL Carnot raw data for fluorescence and temperature

186 between 2005 and 2010.

2.2. Numerical analysis

2.2.1. Machine learning

There are many different types of machine learning models. The RF model is an evolution of the classification and regression tree (CART) model, which was created by the same scientist in 1984 (Breiman et al. 1984). Contrary to the CART model that only uses one tree structure; the RF mode is composed of a predetermined number of trees, hence the term “forest”. The input data of each tree comes from a random sub-sampling performed with a bootstrap technique, hence the term “random”. The first node of the tree is called the root node and split into two child nodes, and so forth until the terminal nodes, which contain the prediction of the model. By following this step, the RF model in regression mode will obtain an average between all the created trees. This stage is more generally referred to as “ensemble learning”. For the RF model, this ensemble method is based on a cross validation process via the out-of-bag (OOB) error. These OOB are mainly calculated from the mean squared error (MSE) in the form of a ratio, in order to give a weight to each predictor. It is important to note that these scores are ratio; therefore, they do not have units. The extraction of the OOB after the learning phase allows us to examine the relative importance of each predictor.

Recent studies have shown that the RF model is well adapted to forecasting changes in the phytoplankton community (Yajima and Derot 2018; Derot et al. 2020; Thomas et al. 2018). The tree structure combined with the bootstrap allows the RF model to effectively manage missing values in datasets, adapt to the study of nonlinear processes, and make no prior assumptions (Thomas et al. 2018; Breiman 2001). These properties coincide with the issues related to our long-term high-frequency sample database; the fluctuations in the phytoplankton abundance can be considered as a stochastic process (Derot et al. 2015), leading to the many gaps associated with sampling automation (see paragraph below). Within the framework of our study, the target signal is the phytoplankton biomass, measured by a proxy via fluorescence. The predictors are the remaining physicochemical parameters, as presented in Table 1. All our data are continuous; therefore, we used the RF model in regression mode.

213 Moreover, we used an individual conditional expectation (ICE) plot (Goldstein et al. 2015), to
214 identify if the interactions created during the learning phase are comparable with the real biological
215 mechanisms. These ICE plots are an improvement on the partial dependence plot (PDP) used several
216 times in previous scientific studies on phytoplankton and water environments (Friedman et al. 2001;
217 Cutler et al. 2007; Roubeix et al. 2016; Teichert et al. 2016; Derot et al. 2020). The PDP highlights the
218 marginal effect between a selected predictor and the target signal (Friedman 2001). In this way, it is
219 possible to observe the global relationship between these two variables. The ICE plots allow a much
220 more refined vision, accounting for the individual effect of the observations on the target. To
221 summarize, the PDP corresponds to the average of the ICE; however this average curve may
222 overshadow the complexity of the relationship created by the model during the learning phase
223 (Goldstein et al. 2015).
224 All our numerical analyses were conducted using the MATLAB software and its "*statistics*
225 *and machine learning*" toolbox. We used the "*TreeBagger*" function for the RF models and
226 "*plotPartialDependence*" for the ICE plots. Once the learning is completed, the function
227 "*TreeBagger*" creates a "fitted model object", which contains the model and all the related
228 information. By directly inserting this "object" in the function "*oobError*", it is possible to observe the
229 evolution of the out-of-bag error (Figs A1-A4). The figures showing the ranking of predictor
230 importance (Figs 5, A6 and A9-A11) are also from the same "object". We can extract these
231 permutation out-of-bag observations across each input, using the array
232 "*OOBPermutedVarDeltaError*". We used the function "*barh*" to visualize these ranking. In addition,
233 the "*rng*" function was set to 1, in order to ensure that the results of the random draw could be used for
234 reproducibility purposes. In our preliminary studies, we observed that 300 trees were sufficient to
235 ensure the stability of the learning phases (Figs. A1 to A4). Given this, we performed all the RF runs
236 in this study using this number of trees. The minimal number of observations per node was set to 5
237 (Derot et al. 2020).

2.2.2. Forecast methodology

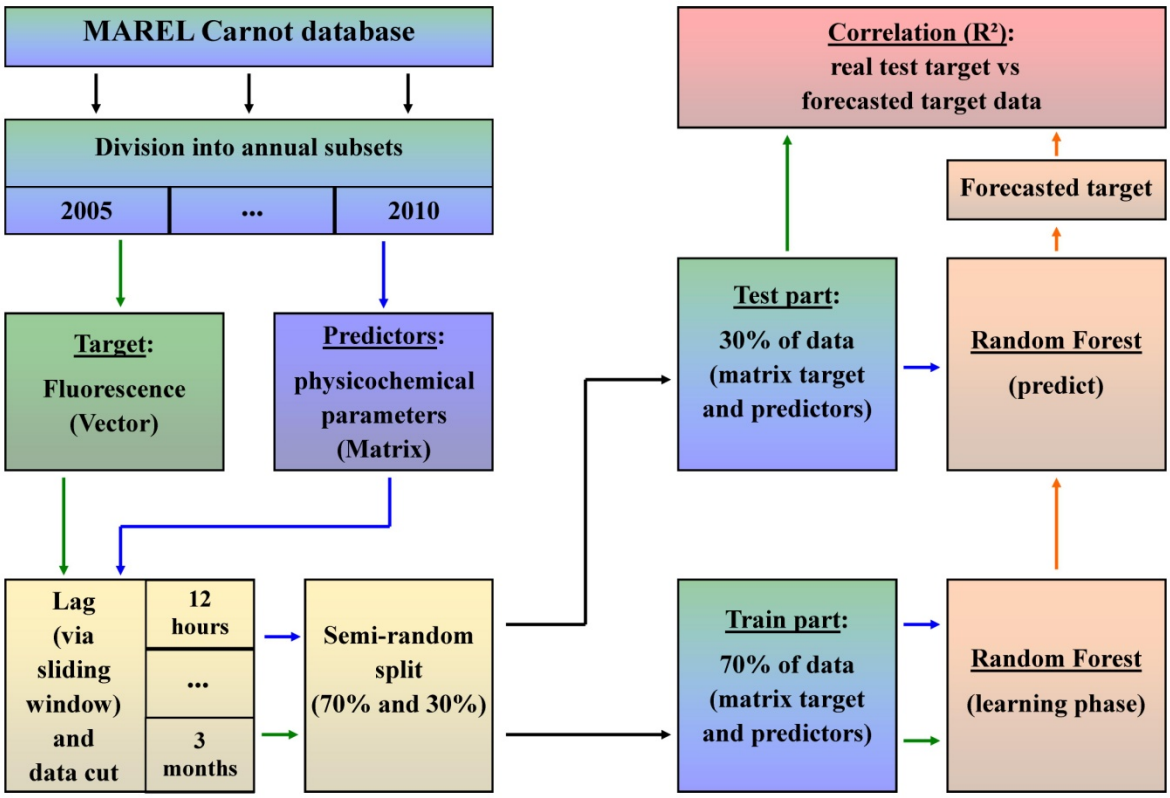
In order to understand the impact of the sampling frequency on the forecasts, we artificially created three databases with the following time steps: 1 hour, 12 hours, and 1 day from the original MAREL Carnot 20-minutes sample frequency database by performing a classical linear interpolation. These interpolations were conducted using the MATLAB function "*interp1*" on all parameters; the results are presented in Table 1. Subsequently, each of these datasets was split according to the year, from 2005 to 2010, inclusively. As mentioned above, the coupling between machine learning and hydrodynamic models could be a way to achieve decision-making tools (Jia et al. 2018; Hanson et al. 2020; Cuttitta et al. 2018). However, the calibration of the biogeochemistry solver linked to hydrodynamic models is fairly sensitive and directly impacts on the capacity to reproduce the dynamics of the phytoplankton (Shimoda and Arhonditsis 2016; Anderson 2005; Zhao et al. 2008). This is why the parameterization of this kind of solver is generally performed year by year (Yajima and Choi 2013). The division into annual subsets of our database was performed to meet this temporal limitation. In all cases fluorescence is always the target signal (green boxes in Fig. 3) and the other physicochemical parameters (Table 1) are the predictors (blue boxes in Fig. 3).

253 As in our previous phytoplankton forecast study (Yajima and Derot 2018), we used the sliding
254 window strategy to perform these forecasts (Herrera et al. 2010). To summarize, the sliding window is
255 a classical methodology for re-farming time series data when we a forecast analysis is to be performed
256 with machine learning models. A lag time is introduced between the target signal and the predictors.
257 For example, when the fluorescence is forecast with a lag time of one week, we removed the first
258 week of this target signal and the last week for all other physicochemical parameters (predictors). In
259 this way, a new input matrix is obtained, where the first value of the fluorescence corresponds to the
260 first value of day 8. The first values of all predictors are still the same. Therefore, there is always a
261 one-week lag between the target signal and the predictors. We used the following lag times for each of
262 our four databases and each year: no lag, 1 day, 3 days, 1 week, 2 weeks, 1 month, 2 months, 2 and a
263 half months and 3 months. Consequently, we performed nine RF runs for the1-year dataset. Therefore,
264 we made 45 RF runs for all the years in one database. In total, we performed 180 RF runs in this study
265 with our four datasets. It should be noted that by applying a sliding window of 2 weeks, we are forced
266 to remove some data at the beginning of the target signal vector and some data at the end of the
267 predictors' matrix (Yajima and Derot 2018). In order to perform our analyses using the same amounts
268 of data, we cut the same time period for each case, depending on the largest lag time, that is. 3 months
269 (yellow boxes in Fig. 3). We applied the same procedure for these 180 cases (Fig. 3).
270 First, the fluorescence vector was identified as the target signal (green boxes and arrows in
271 Fig. 3), and the predictor matrix contained the other physicochemical parameters (blue boxes and
272 arrows in Fig. 3). Second, we applied the sliding window with one definite lag time, and cut periods
273 depending on the 3 months lag (yellow boxes in Fig. 3). Third, we split our data into two parts, the
274 training part comprised 70% of the cut and lagged data and the remaining percentage was used for the
275 test part (yellow boxes in Fig. 3). This split was realized with a semi-random draw via the MATLAB
276 function "*cvpartition*". This function allowed the creation of two groups with similar intensity values.

Therefore, situations where in the training part contained all the high fluorescence values (bloom periods), leaving the test part with no bloom values, or vice versa were accounted for and skews in data were avoided. Fourth, we used the training part (both target and predictors) for the learning phase of the RF model (orange boxes in Fig. 3). Fifth, we used only the predictors from the test part to form a prediction or forecast of our target signal via the MATLAB function "*predict*". Finally, to control the quality of the predictions and forecasts; we performed a correlation between the predicted target signal and the real data from the test part (red box in Fig. 3). These R^2 coefficients were calculated with the coefficient of determination for each of our 180 cases as follows (Shamshirband et al. 2019; Lee et al. 2016; Lee and Lee 2018; Recknagel et al. 2013; Du et al. 2018; Kehoe et al. 2015):

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

where SS_{res} is the residual sum of squares, SS_{tot} is the total sum of squares, n is the number of observations, y_i is the observed data, $\hat{y}_i$ is the predicted data and $\bar{y}$ is the mean of the observed data.



292

293 **Fig. 3.** Conceptual diagram presenting the methodology used to measure the forecast quality,

294 considering all lag times and sampling frequencies.

3. Results and discussion

The primary purpose of this study is to demonstrate the capacity of a machine learning model to forecast phytoplankton blooms in coastal areas and to study the impact of the sampling frequency on the forecast performance of the RF model. For that purpose, we artificially reduced the time step and used different lag times with a sliding window strategy. First, we studied the evolution of the coefficients of determination, depending on several lag times and sampling frequencies. Second, we analyzed which predictors had the greatest influence on the learning phase. Subsequently, we compared the interactions created by the RF model with real biological mechanisms.

3.1. Sampling frequency and time-lag impacts

The evolution of the forecast performances and the dependence on lag times is depicted in Fig. 4 for the four datasets up to a time period of 3 months. This analysis was only performed on the test part. Each point on the four lines of Fig. 4 is derived from an inter-annual average for the years 2005-2010. For example, the second green point is calculated by averaging these 5 years with a lag time of 1 day. The time axis in Fig. 4 uses a log scale. Therefore, in order to avoid problems with the log of zero, we have put the results for no lag time in the annotations for this figure. The calculated averages of all lag times were included for each sampling frequency. Another point of view with the median instead of average is presented in Fig. A7 in the appendix and the range of error in shown Table A1. The green line in Fig. 4 represents the inter-annual average lag times depending on the sampling frequency of 20 minutes. Similarly, the blue, red, and magenta lines represent the coefficient of determination averages for the frequencies of 1 hour, 12 hours, and 1 day, respectively. It should be noted that we cannot apply a sliding window of 12 hours for a sampling frequency of 1 day. This is why the first point of the purple curve is missing in the Fig. 4. In the same way, the first point on the red curve, which corresponds to the 12 hours frequency, is equal to its coefficient without lag; it has been retained to maintain a visual coherence.

An example quantile-quantile plot from the test part is presented in the appendix (Fig. A5). The highest coefficients of determination were obtained for the 20 minutes and 1 hour frequencies. The accuracy of the RF model was evaluated for each frequency and lag time in the appendix in Fig. A8. Table A2 shows the error range linked to this Fig. A8. The average coefficient is slightly better for the green curve and R^2 with no lag time (see annotations). With respect to the frequencies of 12 hours and 1 day, they have the smallest averages. It is significant to note that all curves exhibit the same tendencies; R^2 generally starts to decrease after at two-month lag.

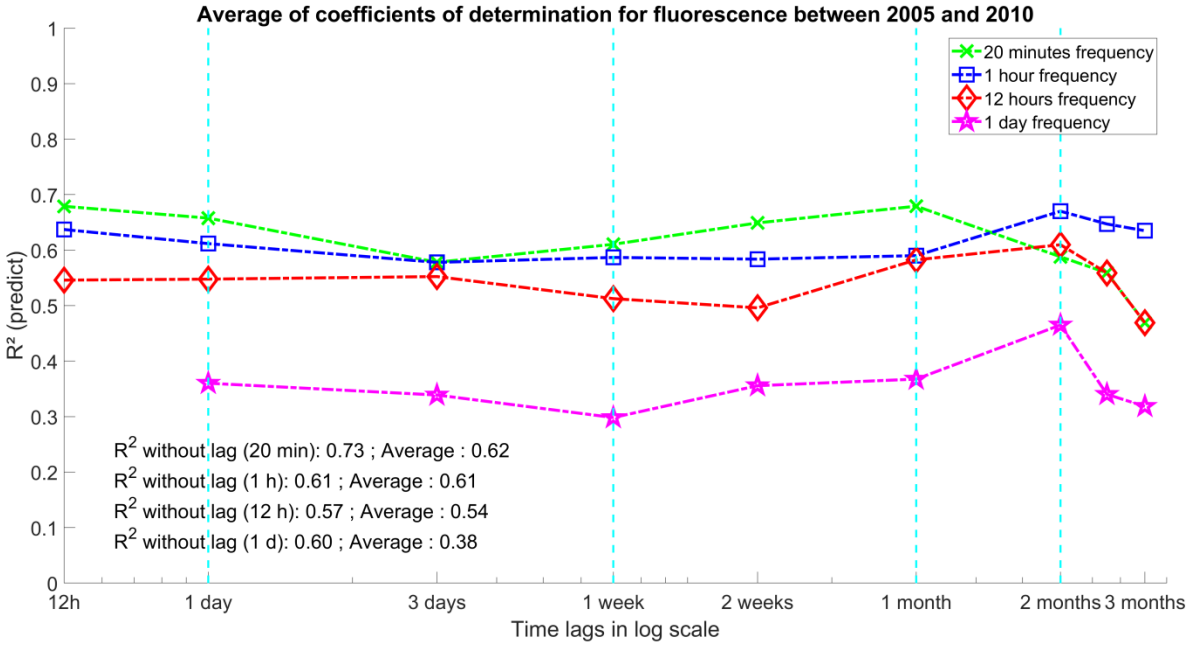


Fig. 4. Evolution of forecast performances depending on lag times and sampling frequencies from test part. The y-axis represents the inter-annual average coefficient of determination from the outputs of the RF models. The x-axis depicts the lag times from the sliding window on a logarithmic scale. The green, blue, red and magenta lines correspond to the sampling frequencies of 20 minutes, 1 hour, 12 hours, and 1 day, respectively. The annotations show the coefficient of determination with no lag time and the global averages of the coefficient of determination for each frequency. See Table A1 in the appendix for the range of error.

Our results indicate that the RF model has the ability to use the supplementary information, which is contained in the database from high-frequency sampling. As depicted in Fig. 4, the forecast capacities are generally better for sampling frequencies of 20 minutes and 1 hour than those of 12 hours and 1 day. It is also evident that the average R^2 coefficient is less than 0.4 for a sampling frequency of 1 day. In our previous works, we tested a close forecast strategy using another database from fresh water ecosystems (Yajima and Derot 2018). It was a long-term dataset over a period of 30 years with a bimonthly sampling frequency. With this lower time step, in various cases the coefficients of determination of the forecasted chlorophyll a were below a threshold of 0.5. In the current study with high-frequency database, it is observed that R^2 stays above this threshold until the 2 months lag, even when the sampling frequency is arterially decreased until 12 hours. Therefore, our findings demonstrate that the pairing of an RF model with a high-frequency dataset from an automatic system yields good forecast results on an annual scale. Water currents and phytoplankton migration have greater significance in open-ocean and coastal areas than in lakes. This complexity could make it difficult to obtain good forecast results for these types of ecosystems (Thomas et al. 2018).

Nevertheless, our study shows that this pairing strategy can also work in marine ecosystems. Consequently, in a water body where water quality management is a major societal issue, it is of pivotal importance to highlight the additional value provided by high sample frequency databases generated by automatic devices. The results of this study indicate that the forecast performance of the RF model increases with increasing sampling frequencies. In addition, it should be noted that although some other studies in similar fields have used the pseudo-R^2 to measure the performance of the RF model, the authors are aware that the coefficient does not assess the true forecast (Large et al. 2015; Teichert et al. 2016; Thomas et al. 2018). Thus, we split our dataset between a learning part and a test part (Fig. 3), and used the Pearson coefficient to measure the forecast performances. Consequently, the pairing between machine learning models and automatically generated high sample frequency databases could eventually lead to the creation of numerical decision-making models. Such a model could help stakeholders prevent HABs from hindering the economy as well as human health. In the next part of this section, we examine the influence of the predictors on the learning phase.

3.2. Physicochemical ranking

Once the learning phase has been completed, it is possible to extract the ranking predictor importance from the out-of-bag (OOB) permutated error. Thus, we can understand the relative impact that each predictor has on an RF model during the learning phase. As shown above, the original MAREL Carnot database with a frequency of 20 minutes provided the best forecast results. In order to examine the global ranking predictor importance, we performed an average of the 35 OOB errors for this frequency. The evaluated global ranking is presented in Fig. 5. It is observed that the nutrients appear to have had low impacts. However, it must be noted that owing to the device limitations, with the original time step of 20 minutes, these nutrients were actually recorded with a sampling frequency of 12 hours (Table 1). In regard to the other physicochemical parameters, water temperature had the most influence on these 35 learning phases, closely followed by the pH, and then the dissolved oxygen. Furthermore, the fourth most important predictor is the salinity measured via its proxy for conductivity. Apart from the nutrients, the photosynthetically active radiation (P.A.R) is the predictor with the least influence. In the appendix, this global ranking has also been evaluated for the other sampling frequencies: Fig. A9 for 1 hour, Fig. A10 for 12 hours and Fig. A11 for 1 day.

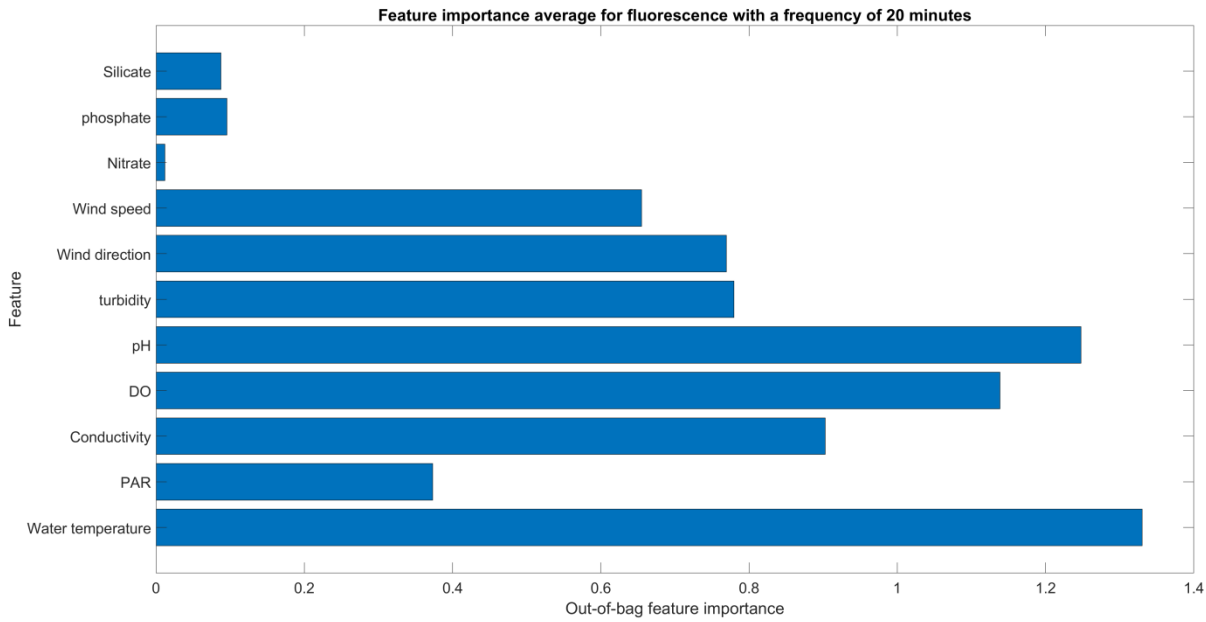


Fig. 5. Ranking of predictor importance based on the average of the out-of-bag (OOB) error, from the 35 runs performed with a time step of 20 minutes. Legend of abbreviations: P.A.R for photosynthetically active radiation and DO for dissolved oxygen.

382 Water temperature, salinity, nutrients, dissolved oxygen, and pH are the water quality
383 indicators that are used in the water framework directive because of their direct impact on biological
384 processes (Best et al. 2007; Millero 2016). Among the physicochemical parameters, we temperature
385 and pH were observed to be the most important predictors, but the impact of the other predictors, such
386 as dissolved oxygen, turbidity, and salinity, are non-negligible. Furthermore, the sampling frequency
387 and lag time can also strongly impact the predictor ranking, as depicted in Fig. A6, where the out-of-
388 bag importance of the temperature is very low. Despite its relative importance for a frequency of 20
389 minutes, the temperature alone is not sufficient to predict the chlorophyll a correctly. For this
390 frequency, we obtained an R^2 without lag that was equal to 0.31, when only the temperature was used
391 as a predictor and other physicochemical parameters were remove. The values of OOB that were
392 extracted from the 35 learning phases with at frequency of 20 minutes were consistent with the water
393 quality indicators, except for the nutrients (Fig 5). Nevertheless, if only the OOB from the databases
394 with a frequency over 12 hours were considered (Fig. A6); the nutrients crucially impacted the
395 learning phase of the RF models. This leads us to believe that the low impact of nutrients shown in this
396 high-frequency database is an artifact caused by recording system limitations.
397 Consequently, in light of the significant influence of the temperature on the learning phases of
398 the RF models, it considered to have been systematically account for when a parameter directly linked
399 to the primary production is predicted with machine learning based on tree structure. Furthermore, it is
400 important to harmonize all sampling frequencies from automatic devices in order to prevent this type
401 of bias. In this context, when designing or upgrading an automatic station, similar to a MAREL buoy,
402 having several types of sensors, we suggest installing sensors having the highest possible common
403 sampling frequency. This could increase the biological prediction linked to the phytoplankton biomass
404 via a machine learning model. Next, we will study the parallels between the interactions created by the
405 RF model and real biological mechanisms.

3.3. Learning phase interactions

Machine learning models are often considered “black boxes” because we cannot understand the interaction between the predictor that the model creates during its learning phase. However, it is possible to transform the RF models into “gray boxes” with the partial dependence plot (PDP) and the individual conditional expectation (ICE) plots. In the previous section, it was seen that the water temperature, conductivity, dissolved oxygen, and pH had the most influence on the average in our learning phases. Therefore, we extracted one ICE plot for each of these 4 predictors with a time step of 20 minutes (Fig.6). The red lines are the PDP, and all the gray lines and the blue points have been derived from the ICE method. In reference to the water temperature, we can see that the RF model predicts high fluorescence values mainly around 11.8°C; there is also a slight increment over 17.0°C. The other three predictors exhibit a common pattern. That is to say, the predicted values of fluorescence are low until the predictors reach a certain threshold. The results of this study exhibit high predicted fluorescence values for a pH over 8.25, conductivity over 47 mS/cm, and concentration in dissolved oxygen over 9 mg/L. It is also important to note that high fluorescence values are mainly predicted for very high conductivities and pH values. However there is a large spread in the high fluorescence values for the dissolved oxygen from the middle range to the higher value.

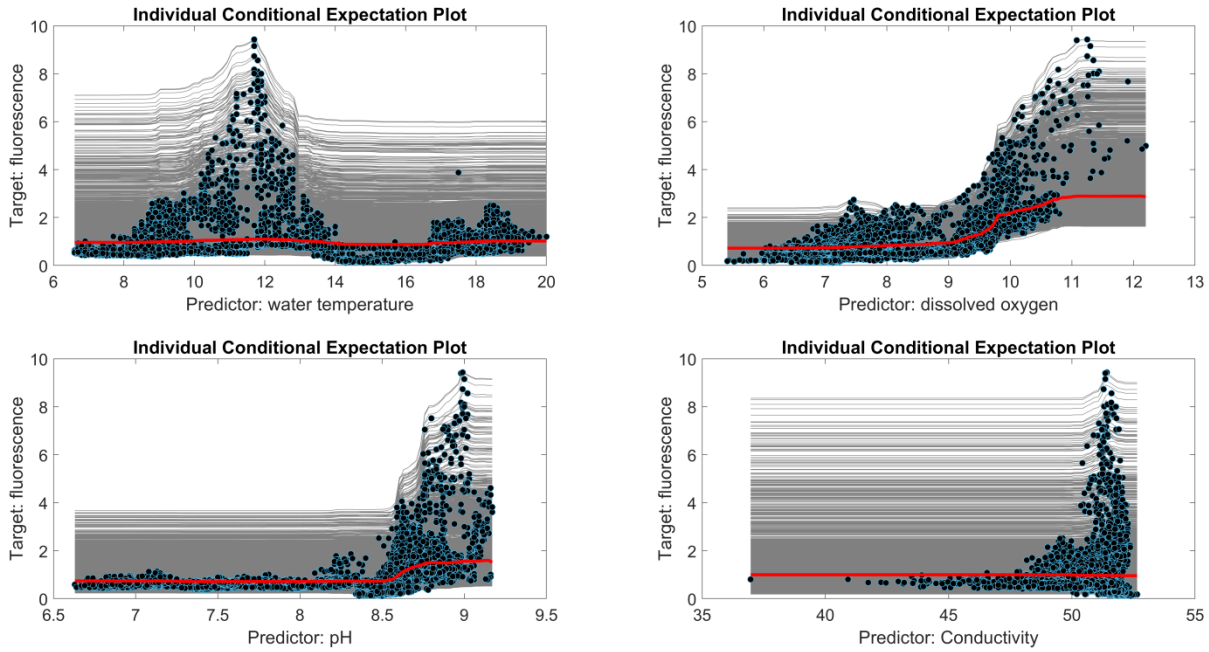


Fig. 6. Individual conditional expectation (ICE) plots for the four most influent predictors with a time step of 20 minutes. The red lines denote the partial dependence plot (PDP); the gray lines and the blue represent from the ICE analyses. Upper left: water temperature; upper right: dissolved oxygen; lower left: pH; lower right: conductivity.

Although the interactions created during the learning phase of an RF model are difficult to comprehend, it may be possible to obtain some links between these interactions and the biological processes that occur in the ecosystems being studied. In Fig. 6 (upper left), It can be observed that the highest fluorescence data are predicted for temperatures of approximately 11.8°C. This result is consistent with those in previous literature (Schoemann et al. 2005; Jahnke 1989). The MAREL Carnot device is located in the English Channel; in this area, the main issue is the harmful algal bloom (HAB) linked to *Phaeocystis globosa*. For this type of phytoplankton, regardless of whether the light conditions are limited, the optimal growth rate appears to be between 10°C and 14°C (Jahnke 1989; Schoemann et al. 2005). Therefore, our ICE analysis of water temperature illustrates that even without prior assumptions, an RF model can account for some real biological processes.

4. Conclusion

In this study we found that the average coefficient of determination, which is the index of the quality of the forecast, decreases when the sampling frequency increases. The coefficient for the 20-minute time step was 0.24 larger than that for the 1-day time step. From our analyses, we observed that the nutrients had a limited impact on the learning phase with the highest sampling frequency. In regard to the water temperature, the averaged OOB error reached 13, while that for the phosphate concentration was only approximately 0.1. Creation of the ICE plot for the water temperature allowed us to illustrate that the RF model predicted the highest fluorescence values of approximately 11.8°C. Consequently, the results suggest that RF models can use the additional information contained in high-frequency databases. It is supposed that the apparent low influence of the nutrients was a bias due to the difference in sampling frequencies. Moreover, although the RF model has no prior assumptions, it was able to create some interactions closely resembling the biological processes present in our study area.

The decrease in the sampling frequency is not the only factor impacting forecast capacity. It should be kept in mind that different time steps between the input parameters can introduce biases into the learning process of an RF model. Therefore, it is imperative to have harmonized sampling frequencies in datasets from automated devices. Several studies in the environmental science literature have claimed that the pairing between high-frequency or long-term datasets with an RF model could overcome the limitations of conventional models (linear, generalized linear model …) (Kehoe et al. 2015; Thomas et al. 2018; Rivero-Calle et al. 2015). The results of our study seem to confirm this hypothesis.

461 Furthermore, some of the automatic systems are equipped with flow cytometers, enabling
462 differentiation between the phytoplankton groups responsible for the HAB (Thomas et al. 2018).
463 Therefore, pairing these types of datasets with machine learning models could aid in the creation of
464 numerical decision-making tools that can help stakeholders with water quality management. In the
465 long run, this kind of tool could have a benefit the economy and human health. Within this framework,
466 we are currently exploring the possibilities of applying this type of pairing on an inter-annual scale, in
467 order to increase the lengths of the forecasted periods.

468 **Acknowledgements**

469 This research was funded by a grant from the Japan Society for the Promotion of Science (JSPS).
470 Derot J. benefited of Postdoctoral Fellowship for Research in Japan. The MAREL Carnot system
471 belongs to a fixed platform network along French coasts called COAST-HF (http://coast-hf.fr). The
472 data presented here can be obtained in the following sites: https://data.coriolis-cotier.org and also the
473 Seanoe site given in Lefebvre et al. (2015).

References

Anderson, T. R. (2005). Plankton functional type modelling: running before we can walk? Journal of Plankton Research, 27(11), 1073-1081. http://dx.doi.org/10.1093/plankt/fbi076

Anderson, D. M., Glibert, P. M., & Burkholder, J. M. (2002). Harmful algal blooms and eutrophication: nutrient sources, composition, and consequences. *Estuaries, 25*(4), 704-726. http://dx.doi.org/10.1007/BF02804901

Backer, L., Manassaram-Baptiste, D., LePrell, R., & Bolton, B. (2015). Cyanobacteria and algae blooms: review of health and environmental data from the harmful algal bloom-related illness surveillance system (HABISS) 2007–2011. *Toxins, 7*(4), 1048-1064. http://dx.doi.org/10.3390/toxins7041048

Bae, S., & Seo, D. (2018). Analysis and modeling of algal blooms in the Nakdong River, Korea. Ecological modelling, 372, 53-63. http://dx.doi.org/10.1016/j.ecolmodel.2018.01.019

Best, M., Wither, A., & Coates, S. (2007). Dissolved oxygen as a physico-chemical supporting element in the Water Framework Directive. *Marine pollution bulletin, 55*(1-6), 53-64. http://dx.doi.org/10.1016/j.marpolbul.2006.08.037

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). Classification and regression trees. Wadsworth Int. Group, 37(15), 237-251. http://dx.doi.org/10.1201/9781315139470

Burkholder, J. M. (2003). Cyanobacteria. *Encyclopedia of Environmental Microbiology*.

Camargo, J. A., & Alonso, Á. (2006). Ecological and toxicological effects of inorganic nitrogen pollution in aquatic ecosystems: a global assessment. *Environment international, 32*(6), 831-849. http://dx.doi.org/10.1016/j.envint.2006.05.002

Carmichael, W. W., & Boyer, G. L. (2016). Health impacts from cyanobacteria harmful algae blooms: Implications for the North American Great Lakes. *Harmful algae, 54*, 194-212. http://dx.doi.org/10.1016/j.hal.2016.02.002

Chen, Q., Guan, T., Yun, L., Li, R., & Recknagel, F. (2015). Online forecasting chlorophyll a concentrations by an auto-regressive integrated moving average model: Feasibilities and potentials. Harmful algae, 43, 58-65. http://dx.doi.org/10.1016/j.hal.2015.01.002

Chen, Y.-Q., Wang, N., Zhang, P., Zhou, H., & Qu, L.-H. (2002). Molecular evidence identifies bloom-forming Phaeocystis (Prymnesiophyta) from coastal waters of southeast China as Phaeocystis globosa. *Biochemical Systematics and Ecology, 30*(1), 15-22. http://dx.doi.org/10.1016/S0305-1978(01)00054-0

Cho, H., Choi, U., & Park, H. (2018). Deep learning application to time-series prediction of daily chlorophyll-a concentration. WIT Trans. Ecol. Environ, 215, 157-163. http://dx.doi.org/10.2495/EID180141

Cho, H., & Park, H. Merged-LSTM and multistep prediction of daily chlorophyll-a concentration for algal bloom forecast. In IOP Conference Series: Earth and Environmental Science, 2019 (Vol. 351, pp. 012020, Vol. 1): IOP Publishing http://dx.doi.org/10.1088/1755-1315/351/1/012020

Cutler, D. R., Edwards Jr, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., et al. (2007). Random forests for classification in ecology. *Ecology, 88*(11), 2783-2792. http://dx.doi.org/10.1890/07-0539.1

Cuttitta, A., Torri, M., Zarrad, R., Zgozi, S., Jarboui, O., Quinci, E. M., et al. (2018). Linking surface hydrodynamics to planktonic ecosystem: the case study of the ichthyoplanktonic assemblages in the Central Mediterranean Sea. Hydrobiologia, 821(1), 191-214. http://dx.doi.org/10.1007/s10750-017-3483-x

Danhiez, F., Vantrepotte, V., Cauvin, A., Lebourg, E., & Loisel, H. (2017). Optical properties of chromophoric dissolved organic matter during a phytoplankton bloom. Implication for DOC estimates from CDOM absorption. *Limnology and Oceanography, 62*(4), 1409-1425. http://dx.doi.org/10.1002/lno.10507

Derot, J., Jamoneau, A., Teichert, N., Rosebery, J., Morin, S., & Laplace-Treyture, C. (2020). Response of phytoplankton traits to environmental variables in French lakes: New perspectives for bioindication. *Ecological indicators, 108*, 105659. http://dx.doi.org/10.1016/j.ecolind.2019.105659

Derot, J., Schmitt, F. G., Gentilhomme, V., & Morin, P. (2016). Correlation between long-term marine temperature time series from the eastern and western English Channel: Scaling analysis using empirical mode decomposition. *Comptes Rendus Géoscience, 348*(5), 343-349. http://dx.doi.org/10.1016/j.crte.2015.12.001

Derot, J., Schmitt, F. G., Gentilhomme, V., & Zongo, S. B. (2015). Long-term high frequency phytoplankton dynamics, recorded from a coastal water autonomous measurement system in the eastern English Channel. *Continental Shelf Research, 109*, 210-221. http://dx.doi.org/10.1016/j.csr.2015.09.015

Du, Z., Qin, M., Zhang, F., & Liu, R. (2018). Multistep-ahead forecasting of chlorophyll a using a wavelet nonlinear autoregressive network. Knowledge-Based Systems, 160, 61-70. http://dx.doi.org/10.1016/j.knosys.2018.06.015

Dur, G., Schmitt, F. G., & Souissi, S. (2007). Analysis of high frequency temperature time series in the Seine estuary from the Marel autonomous monitoring buoy. *Hydrobiologia, 588*(1), 59-68. http://dx.doi.org/10.1007/s10750-007-0652-3

Edwards, K. F., Thomas, M. K., Klausmeier, C. A., & Litchman, E. (2016). Phytoplankton growth and the interaction of light and temperature: A synthesis at the species and community level. *Limnology and Oceanography, 61*(4), 1232-1244. http://dx.doi.org/10.1002/lno.10282

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1, Vol. 10): Springer series in statistics New York, NY, USA:.

Glibert, P. M., Anderson, D. M., Gentien, P., Granéli, E., & Sellner, K. G. (2005). The global, complex phenomena of harmful algal blooms. http://dx.doi.org/10.5670/oceanog.2005.49

Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics, 24*(1), 44-65. http://dx.doi.org/10.1080/10618600.2014.907095

Grattepanche, J.-D., Breton, E., Brylinski, J.-M., Lecuyer, E., & Christaki, U. (2011). Succession of primary producers and micrograzers in a coastal ecosystem dominated by Phaeocystis globosa blooms. Journal of Plankton Research, 33(1), 37-50. http://dx.doi.org/10.1093/plankt/fbq097

Hanson, P. C., Stillman, A. B., Jia, X., Karpatne, A., Dugan, H. A., Carey, C. C., et al. (2020). Predicting lake surface water phosphorus dynamics using process-guided machine learning. Ecological modelling, 430, 109136. http://dx.doi.org/10.1016/j.ecolmodel.2020.109136

Heisler, J., Glibert, P. M., Burkholder, J. M., Anderson, D. M., Cochlan, W., Dennison, W. C., et al. (2008). Eutrophication and harmful algal blooms: a scientific consensus. Harmful algae, 8(1), 3-13. http://dx.doi.org/10.1016/j.hal.2008.08.006

Houliez, E., Lizon, F., Thyssen, M., Artigas, L. F., & Schmitt, F. G. (2012). Spectral fluorometric characterization of Haptophyte dynamics using the FluoroProbe: an application in the eastern English Channel for monitoring Phaeocystis globosa. Journal of Plankton Research, 34(2), 136-151. http://dx.doi.org/10.1093/plankt/fbr091

Herrera, M., Torgo, L., Izquierdo, J., & Pérez-García, R. (2010). Predictive models for forecasting hourly urban water demand. *Journal of Hydrology, 387*(1-2), 141-150. http://dx.doi.org/10.1016/j.jhydrol.2010.04.005

Howarth, R. W., Anderson, D., Cloern, J. E., Elfring, C., Hopkinson, C. S., Lapointe, B., et al. (2000). Nutrient pollution of coastal rivers, bays, and seas. Issues in ecology(7), 1-16.

Huang, Y., & Schmitt, F. G. (2014). Time dependent intrinsic correlation analysis of temperature and dissolved oxygen time series using empirical mode decomposition. *Journal of Marine Systems, 130*, 90-100. http://dx.doi.org/10.1016/j.jmarsys.2013.06.007

Jahnke, J. (1989). The light and temperature dependence of growth rate and elemental composition of Phaeocystis globosa Scherffel and P. pouchetii (Har.) Lagerh. in batch cultures. *Netherlands Journal of Sea Research, 23*(1), 15-21. http://dx.doi.org/10.1016/0077-7579(89)90038-0

Jia, X., Karpatne, A., Willard, J., Steinbach, M., Read, J., Hanson, P. C., et al. (2018). Physics guided recurrent neural networks for modeling dynamical systems: Application to monitoring water temperature and quality in lakes. arXiv preprint arXiv:1810.02880.

Kehoe, M. J., Chun, K. P., & Baulch, H. M. (2015). Who smells? Forecasting taste and odor in a drinking water reservoir. Environmental science & technology, 49(18), 10984-10992. http://dx.doi.org/10.1021/acs.est.5b00979

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling* (Vol. 26): Springer. http://dx.doi.org/10.1007/978-1-4614-6849-3

Lancelot, C., Billen, G., Sournia, A., Weisse, T., Colijn, F., Veldhuis, M. J., et al. (1987). Phaeocystis blooms and nutrient enrichment in the continental coastal zones of the North Sea. *Ambio*(1).

Lancelot, C., Rousseau, V., Schoemann, V., & Becquevort, S. (2002). On the ecological role of the different life forms of Phaeocystis. *LIFEHAB: Life history of microalgal species causing harmful blooms. European Commission publication no: EUR, 20361*, 71-75.

Lapointe, B. E., Herren, L. W., & Paule, A. L. (2017). Septic systems contribute to nutrient pollution and harmful algal blooms in the St. Lucie Estuary, Southeast Florida, USA. Harmful algae, 70, 1-22. http://dx.doi.org/10.1016/j.hal.2017.09.005

631 Large, S. I., Fay, G., Friedland, K. D., & Link, J. S. (2015). Quantifying patterns of change in marine
632 ecosystem response to multiple pressures. *PloS one, 10*(3), e0119922.
633 http://dx.doi.org/10.1371/journal.pone.0119922
634
635 Lee, G., Bae, J., Lee, S., Jang, M., & Park, H. (2016). Monthly chlorophyll-a prediction using neuro-
636 genetic algorithm for water quality management in Lakes. Desalination and Water Treatment,
637 57(55), 26783-26791. http://dx.doi.org/10.1080/19443994.2016.1190107
638
639 Lee, S., & Lee, D. (2018). Improved prediction of harmful algal blooms in four Major South Korea's
640 Rivers using deep learning models. International journal of environmental research and public
641 health, 15(7), 1322. http://dx.doi.org/10.3390/ijerph15071322
642
643 Lefebvre, A. (2015). MAREL Carnot data and metadata from Coriolis Data Centre. SEANOE.
644
645 Lubac, B., Loisel, H., Guiselin, N., Astoreca, R., Artigas, L. F., & Mériaux, X. (2008). Hyperspectral
646 and multispectral ocean color inversions to detect Phaeocystis globosa blooms in coastal
647 waters. *Journal of Geophysical Research: Oceans, 113*(C6).
648 http://dx.doi.org/10.1029/2007JC004451
649
650 Millero, F. J. (2016). *Chemical oceanography*: CRC press. http://dx.doi.org/10.1201/b14753
651
652 Monchy, S., Grattepanche, J.-D., Breton, E., Meloni, D., Sanciu, G., Chabé, M., et al. (2012).
653 Microplanktonic community structure in a coastal system relative to a Phaeocystis bloom
654 inferred from morphological and tag pyrosequencing methods. *PloS one, 7*(6), e39924.
655 http://dx.doi.org/10.1371/journal.pone.0039924
656
657 Pennekamp, F., Iles, A. C., Garland, J., Brennan, G., Brose, U., Gaedke, U., et al. (2019). The intrinsic
658 predictability of ecological time series and its potential to guide forecasting. *Ecological*
659 *Monographs*, e01359. http://dx.doi.org/10.1002/ecm.1359
660
661 Recknagel, F., Ostrovsky, I., Cao, H., Zohary, T., & Zhang, X. (2013). Ecological relationships,
662 thresholds and time-lags determining phytoplankton community dynamics of Lake Kinneret,
663 Israel elucidated by evolutionary computation and wavelets. Ecological modelling, 255, 70-
664 86. http://dx.doi.org/10.1016/j.ecolmodel.2013.02.006
665
666 Reynaud, A., & Lanzanova, D. (2017). A global meta-analysis of the value of ecosystem services
667 provided by lakes. *Ecological Economics, 137*, 184-194.
668 http://dx.doi.org/10.1016/j.ecolecon.2017.03.001
669
670 Rivero-Calle, S., Gnanadesikan, A., Del Castillo, C. E., Balch, W. M., & Guikema, S. D. (2015).
671 Multidecadal increase in North Atlantic coccolithophores and the potential role of rising CO2.
672 Science, 350(6267), 1533-1537. http://dx.doi.org/10.1126/science.aaa8026
673
674 Roelke, D. L., Grover, J. P., Brooks, B. W., Glass, J., Buzan, D., Southard, G. M., et al. (2010). A
675 decade of fish-killing Prymnesium parvum blooms in Texas: roles of inflow and salinity.
676 *Journal of plankton research, 33*(2), 243-253. http://dx.doi.org/10.1093/plankt/fbq079
677
678 Roubeix, V., Danis, P.-A., Feret, T., & Baudoin, J.-M. (2016). Identification of ecological thresholds
679 from variations in phytoplankton communities among lakes: contribution to the definition of
680 environmental standards. *Environmental monitoring and assessment, 188*(4), 246.
681 http://dx.doi.org/10.1007/s10661-016-5238-y

682 Schindler, D. W. (2006). Recent advances in the understanding and management of eutrophication.
683 *Limnology and Oceanography, 51*(1part2), 356-363.
684 http://dx.doi.org/10.4319/lo.2006.51.1_part_2.0356
685
686 Schmitt, F. G., Landry, Y., Revillion, M., Bordé, C., Gentilhomme, V., & Herbert, V. (2011). Blooms
687 de Phaeocystis sur la Côte d'Opale: investigations historiques, in Du naturalisme à l'écologie,
688 edité par FG Schmitt.
689
690 Schmitt, F.G., & Lefebvre A. (2016). *Mesures à haute résolution dans l'environnement marin côtie.*
691 Paris: CNRS Editions.
692
693 Schoemann, V., Becquevort, S., Stefels, J., Rousseau, V., & Lancelot, C. (2005). Phaeocystis blooms
694 in the global ocean and their controlling mechanisms: a review. *Journal of Sea Research,*
695 *53*(1-2), 43-66. http://dx.doi.org/10.1016/j.seares.2004.01.008
696
697 Shimoda, Y., & Arhonditsis, G. B. (2016). Phytoplankton functional type modelling: running before
698 we can walk? A critical evaluation of the current state of knowledge. Ecological modelling,
699 320, 29-43. http://dx.doi.org/10.1016/j.ecolmodel.2015.08.029
700
701 Shamshirband, S., Jafari Nodoushan, E., Adolf, J. E., Abdul Manaf, A., Mosavi, A., & Chau, K.-w.
702 (2019). Ensemble models with uncertainty analysis for multi-day ahead forecasting of
703 chlorophyll a concentration in coastal waters. Engineering Applications of Computational
704 Fluid Mechanics, 13(1), 91-101. http://dx.doi.org/10.1016/j.ecolmodel.2015.08.029
705
706 Shin, J., Yoon, S., & Cha, Y. (2017). Prediction of cyanobacteria blooms in the lower Han River
707 (South Korea) using ensemble learning algorithms. Desalination and Water Treatment, 84, 31-
708 39. http://dx.doi.org/10.5004/dwt.2017.20986
709
710 Smith, V. H., Joye, S. B., & Howarth, R. W. (2006). Eutrophication of freshwater and marine
711 ecosystems. *Limnology and Oceanography, 51*(1part2), 351-355.
712 http://dx.doi.org/10.4319/lo.2006.51.1_part_2.0351
713
714 Spilmont, N., Denis, L., Artigas, L. F., Caloin, F., Courcot, L., Créach, A., et al. (2009). Impact of the
715 Phaeocystis globosa spring bloom on the intertidal benthic compartment in the eastern English
716 Channel: A synthesis. Marine pollution bulletin, 58(1), 55-63.
717 http://dx.doi.org/10.1016/j.marpolbul.2008.09.007
718
719 Teichert, N., Borja, A., Chust, G., Uriarte, A., & Lepage, M. (2016). Restoring fish ecological quality
720 in estuaries: implication of interactive and cumulative effects among anthropogenic stressors.
721 *Science of the Total Environment, 542*, 383-393.
722 http://dx.doi.org/10.1016/j.scitotenv.2015.10.068
723
724 Thomas, M. K., Fontana, S., Reyes, M., Kehoe, M., & Pomati, F. (2018). The predictability of a lake
725 phytoplankton community, over time-scales of hours to years. *Ecology letters, 21*(5), 619-628.
726 http://dx.doi.org/10.1111/ele.12927
727
728 Veldhuis, M. J., & Wassmann, P. (2005). Bloom dynamics and biological control of a high biomass
729 HAB species in European coastal waters: a Phaeocystis case study. *Harmful algae, 4*(5), 805-
730 809. http://dx.doi.org/10.1016/j.hal.2004.12.004
731
732 Yajima, H., & Choi, J. (2013). Changes in phytoplankton biomass due to diversion of an inflow into
733 the Urayama Reservoir. Ecological engineering, 58, 180-191.
734 http://dx.doi.org/10.1016/j.ecoleng.2013.06.030

Yajima, H., & Derot, J. (2018). Application of the Random Forest model for chlorophyll-a forecasts in fresh and brackish water bodies in Japan, using multivariate long-term databases. *Journal of Hydroinformatics, 20*(1), 206-220. http://dx.doi.org/10.2166/hydro.2017.010

Zhang, F., Wang, Y., Cao, M., Sun, X., Du, Z., Liu, R., et al. (2016). Deep-learning-based approach for prediction of algal blooms. Sustainability, 8(10), 1060. http://dx.doi.org/10.3390/su8101060

Zhao, J., Ramin, M., Cheng, V., & Arhonditsis, G. B. (2008). Competition patterns among phytoplankton functional groups: How useful are the complex mathematical models? acta oecologica, 33(3), 324-344. http://dx.doi.org/10.1016/j.actao.2008.01.007

Zongo, S., & Schmitt, F. G. (2011). Scaling properties of pH fluctuations in coastal waters of the English Channel: pH as a turbulent active scalar. *Nonlinear Processes in Geophysics, 18*(6), 829-839. http://dx.doi.org/10.5194/npg-18-829-2011

Appendix

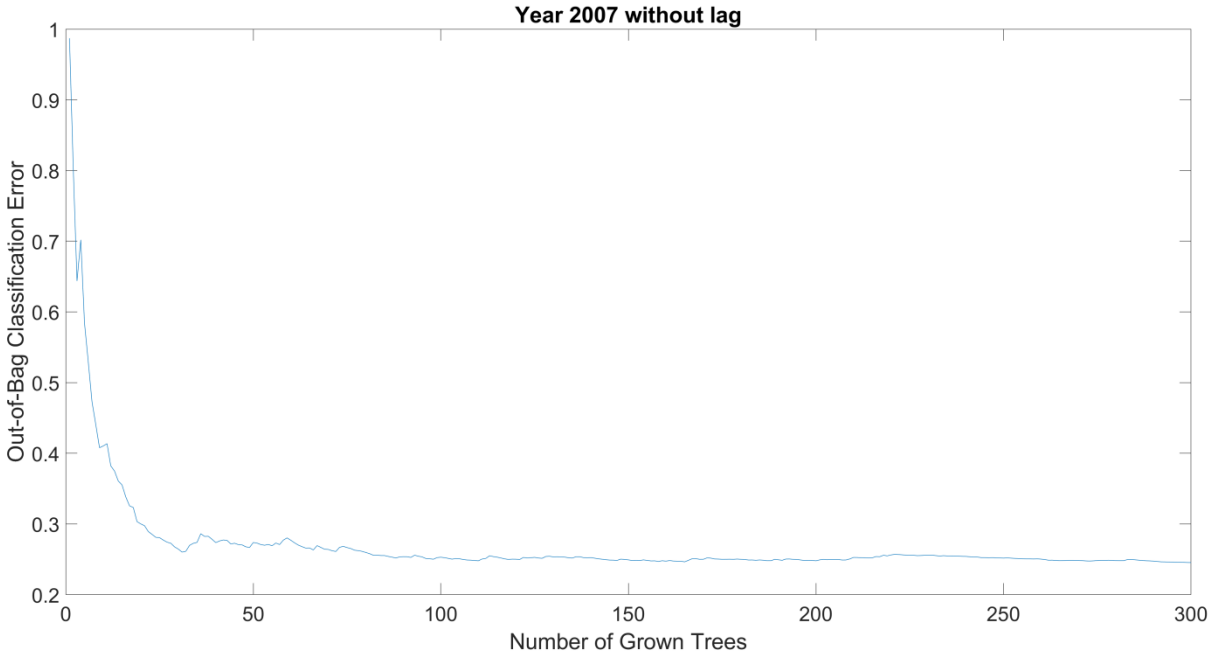


Fig. A1. Evolution of the out-of-bag error for the year 2007 without lag time.

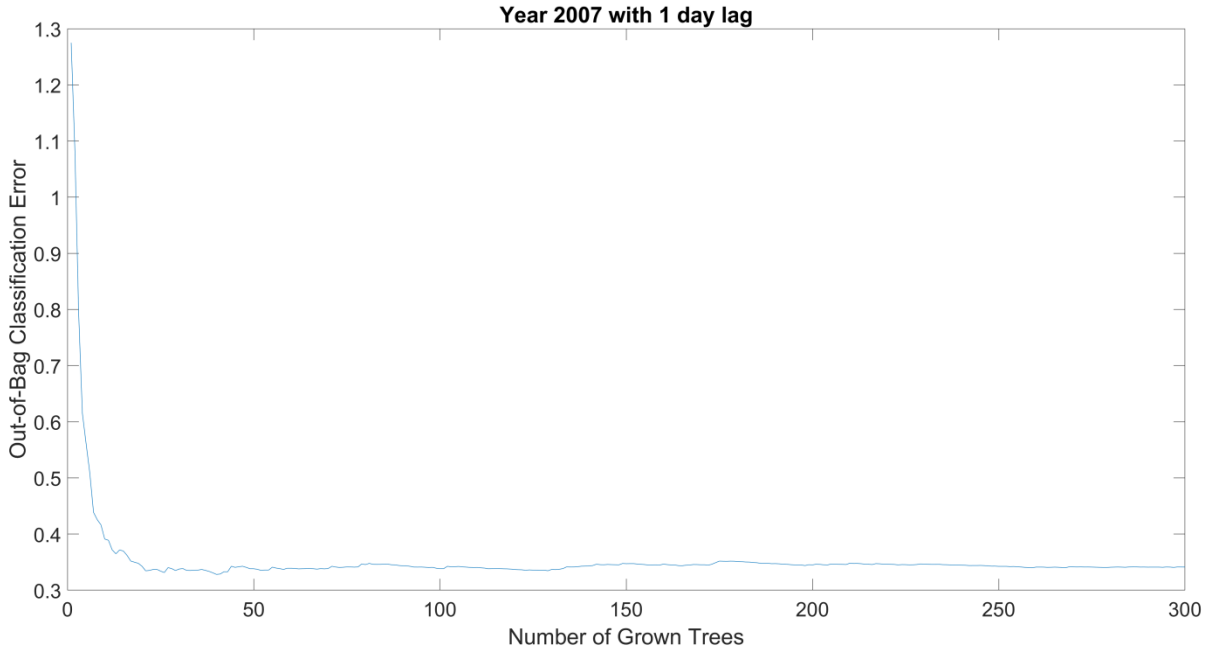
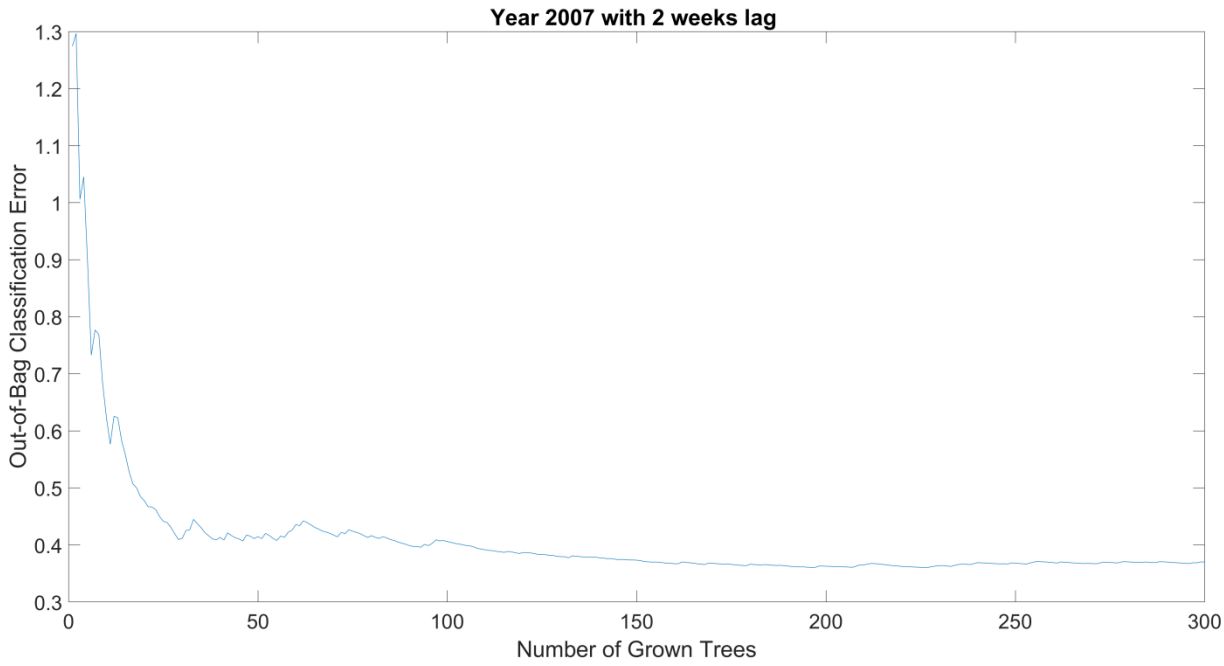


Fig. A2. Evolution of the out-of-bag error for the year 2007 with lag time of 1 day.

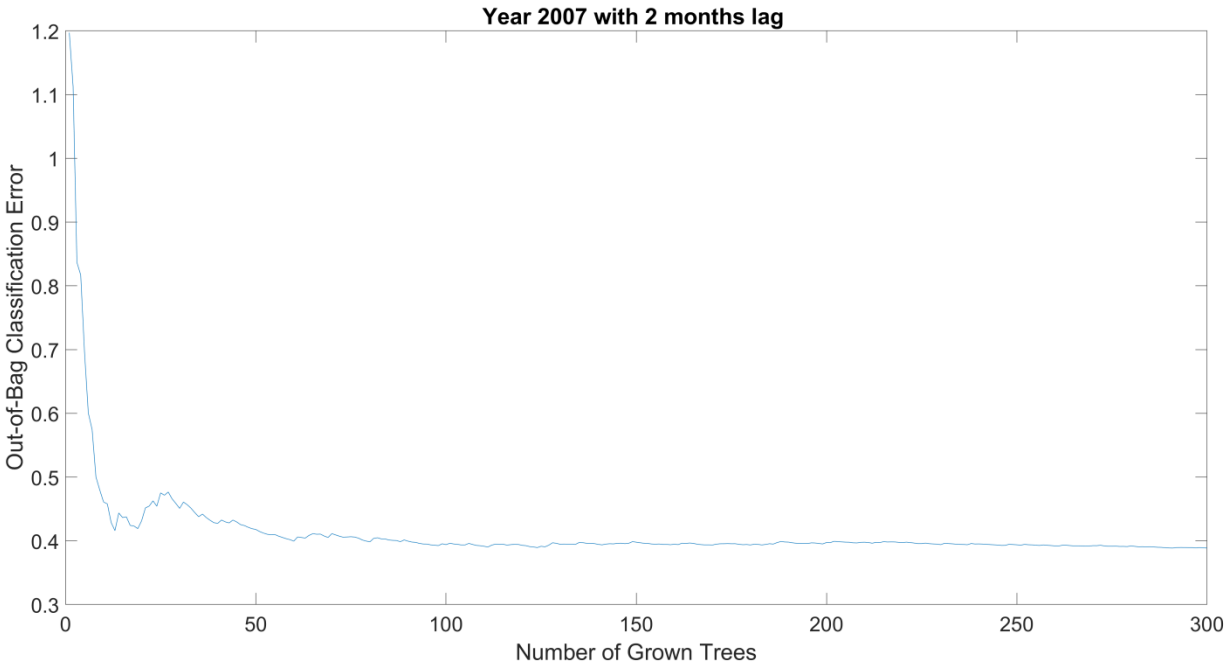


757

758 **Fig. A3.** Evolution of the out-of-bag error for the year 2007 with lag time of 2 weeks.

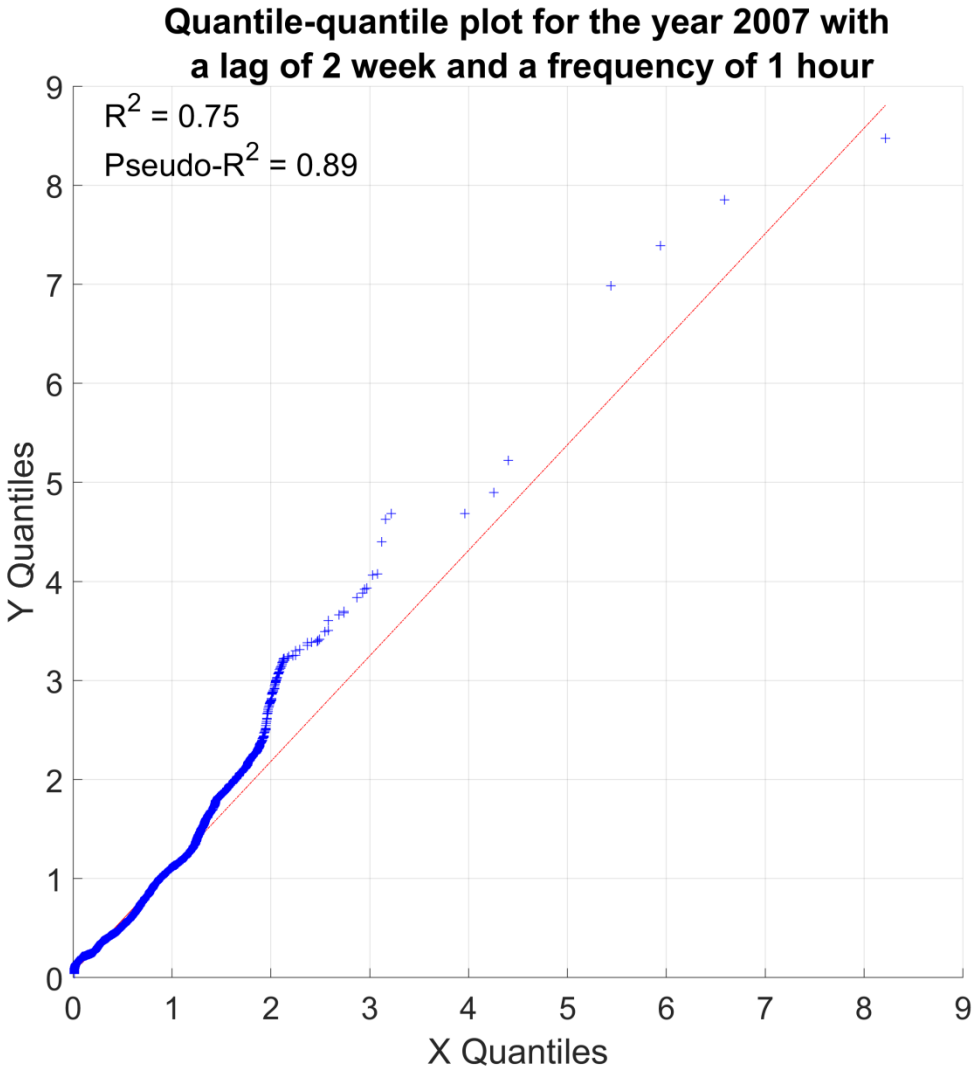
759

760



761

762 **Fig. A4.** Evolution of the out-of-bag error for the year 2007 with lag time of 2 months.

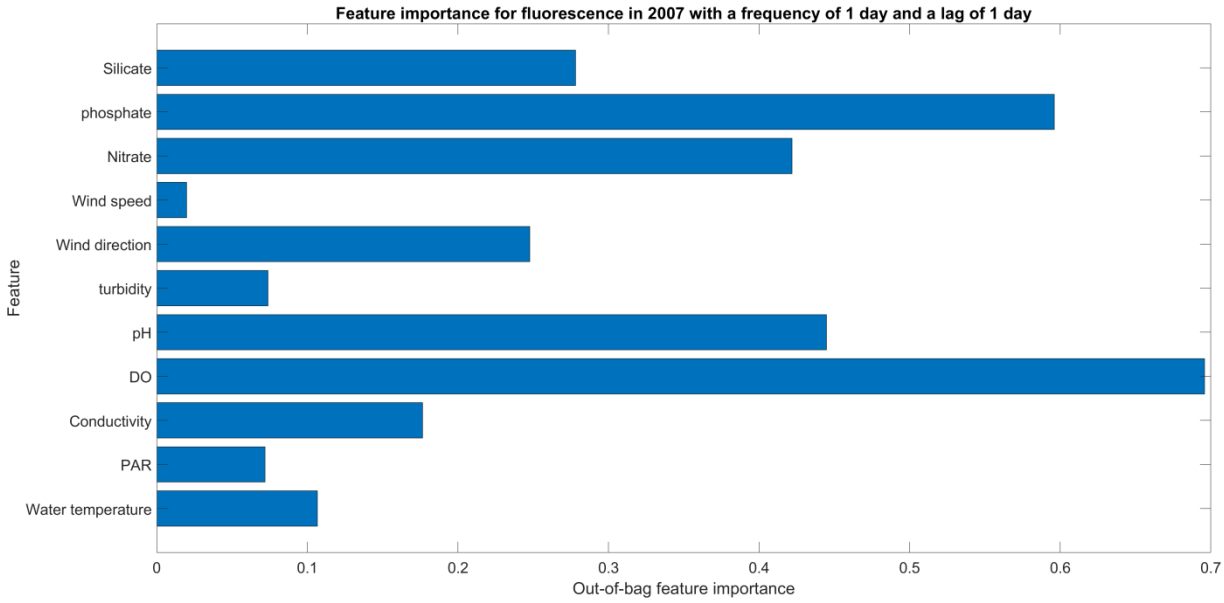


763

764 **Fig. A5.** Quantile-quantile plot from the test part; for the year 2007 with a frequency of 1 hour and a

765 lag time of 2 weeks.

766



767

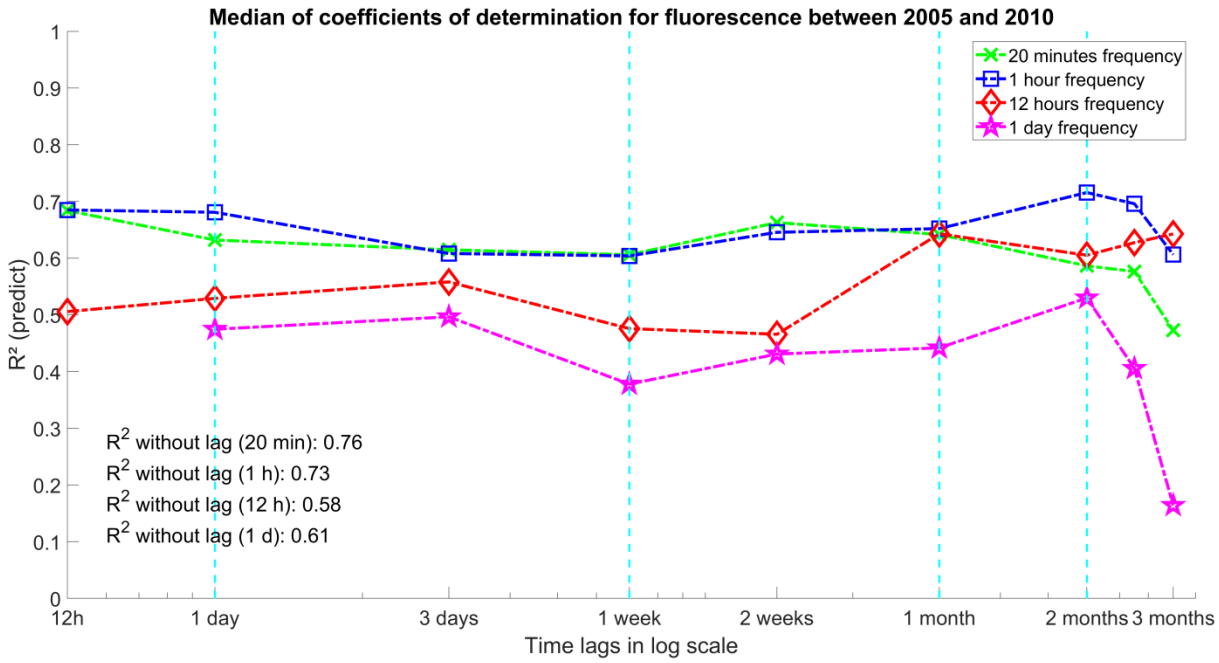
768 **Fig. A6.** Ranking of predictor importance from the out-of-bag (OOB) error, for the year 2007 this a

769 lag time of 1 day and a sampling frequency of 1 day.

770

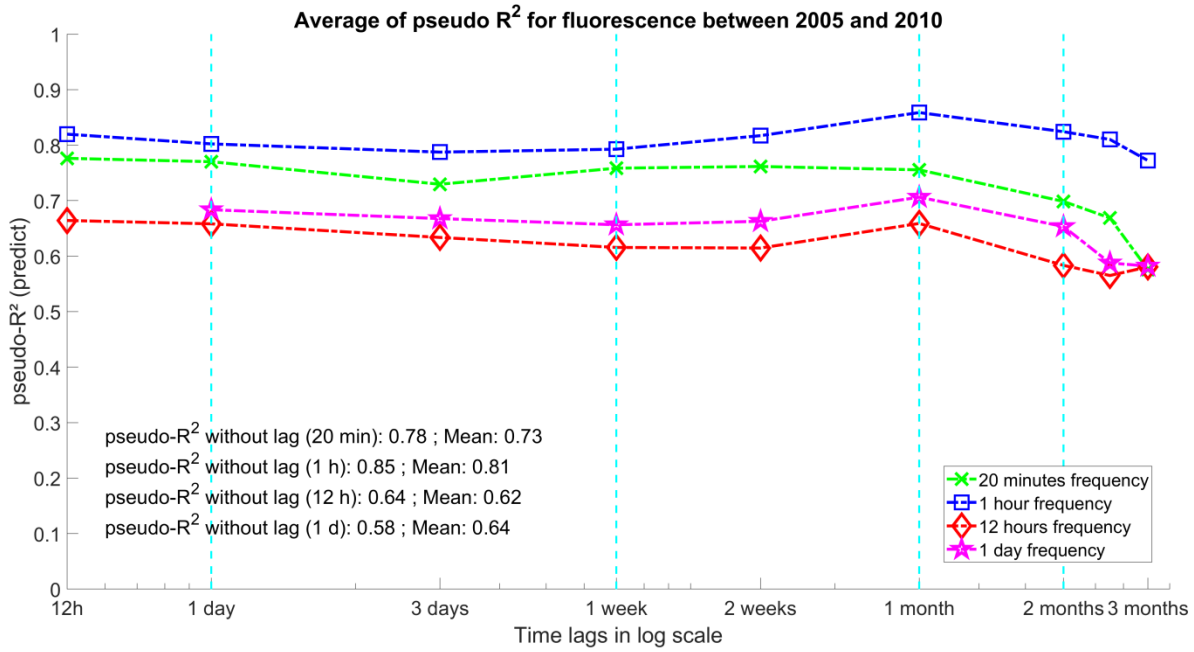
	Frequency 20 minutes	**Frequency 1 hours**	**Frequency 12 hours**	**Frequency 1 day**
Lag = 0	0.14	0.23	0.13	0.07
Lag = 12 hours	0.05	0.15	0.19	∅
Lag = 1 day	0.07	0.17	0.18	0.26
Lag = 3 days	0.13	0.16	0.18	0.26
Lag = 1 week	0.10	0.15	0.20	0.21
Lag = 2 weeks	0.09	0.20	0.21	0.23
Lag = 1 month	0.06	0.16	0.18	0.21
Lag = 2 months	0.12	0.09	0.11	0.21
Lag = 2.5 month	0.15	0.08	0.20	0.19
Lag = 3 months	0.15	0.09	0.30	0.23

771 **Table A1.** Error ranges linked to Fig. 4 calculated via the standard deviation.



772

773 **Fig. A7.** Evolution of forecast performances depending on lag times and sampling frequencies from
774 test part. The y-axis represents the inter-annual median from the outputs of the RF models. The x-axis
775 denotes the lag times from the sliding window on a logarithmic scale. The green, blue, red and
776 magenta lines correspond to the sampling frequencies of 20 minutes, 1 hour, 12 hours, and 1 day,
777 respectively. The annotations display the median of coefficient of determination with no lag time for
778 each frequency.



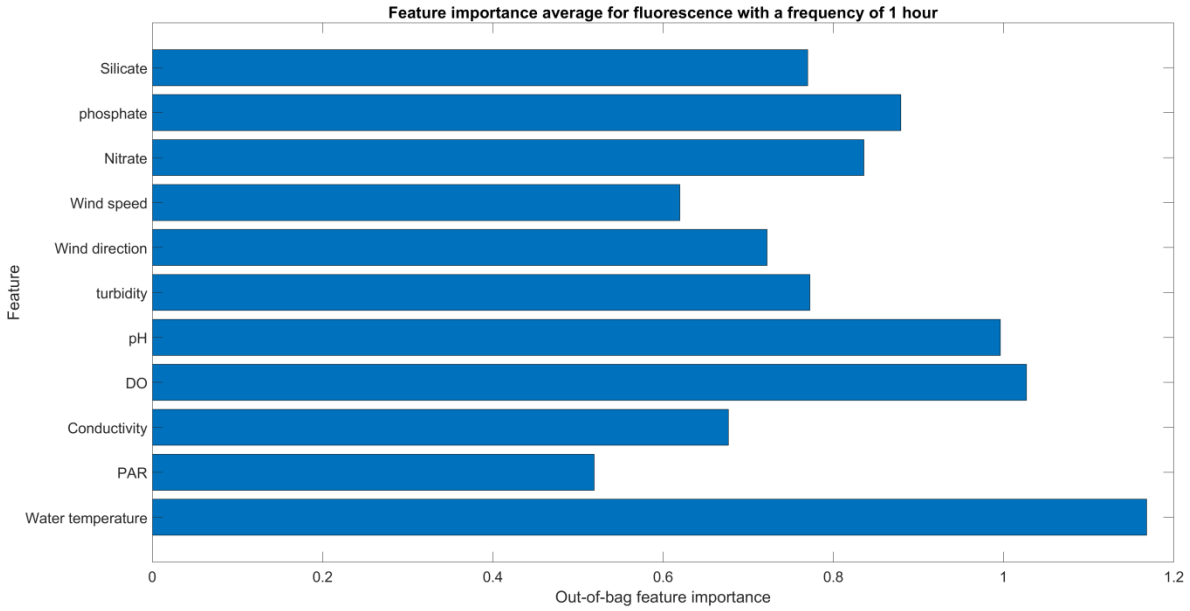
779

780 **Fig. A8.** Evolution of forecast OOB error depending on lag times and sampling frequencies from test
781 part. The y-axis represents the inter-annual mean extracted after the learning part. The x-axis denotes
782 the lag times from the sliding window on a logarithmic scale. The green, blue, red and magenta lines
783 correspond to the sampling frequencies of 20 minutes, 1 hour, 12 hours, and 1 day, respectively. The
784 annotations display the mean of OOB with no lag time for each frequency.

	Frequency 20 minutes	Frequency 1hous	Frequency 12 hours	Frequency 1 day
Lag = 0	0.18	0.11	0.16	0.18
Lag = 12 hours	0.09	0.12	0.14	ø
Lag = 1 day	0.09	0.15	0.16	0.10
Lag = 3 days	0.09	0.15	0.17	0.10
Lag = 1 week	0.06	0.16	0.18	0.07
Lag = 2 weeks	0.08	0.13	0.18	0.08
Lag = 1 month	0.06	0.07	0.19	0.11
Lag = 2 months	0.14	0.12	0.13	0.07
Lag = 2.5 month	0.17	0.13	0.15	0.09
Lag = 3 months	0.15	0.17	0.10	0.05

785 **Table A2.** Error ranges linked to Fig. A8 calculated via the standard deviation.

786



787

788 **Fig. A9.** Ranking of predictor importance based on the average of the out-of-bag (OOB) error, from
789 the 35 runs performed with a time step of 1 hour. Legend of abbreviations: P.A.R for
790 photosynthetically active radiation and DO for dissolved oxygen.

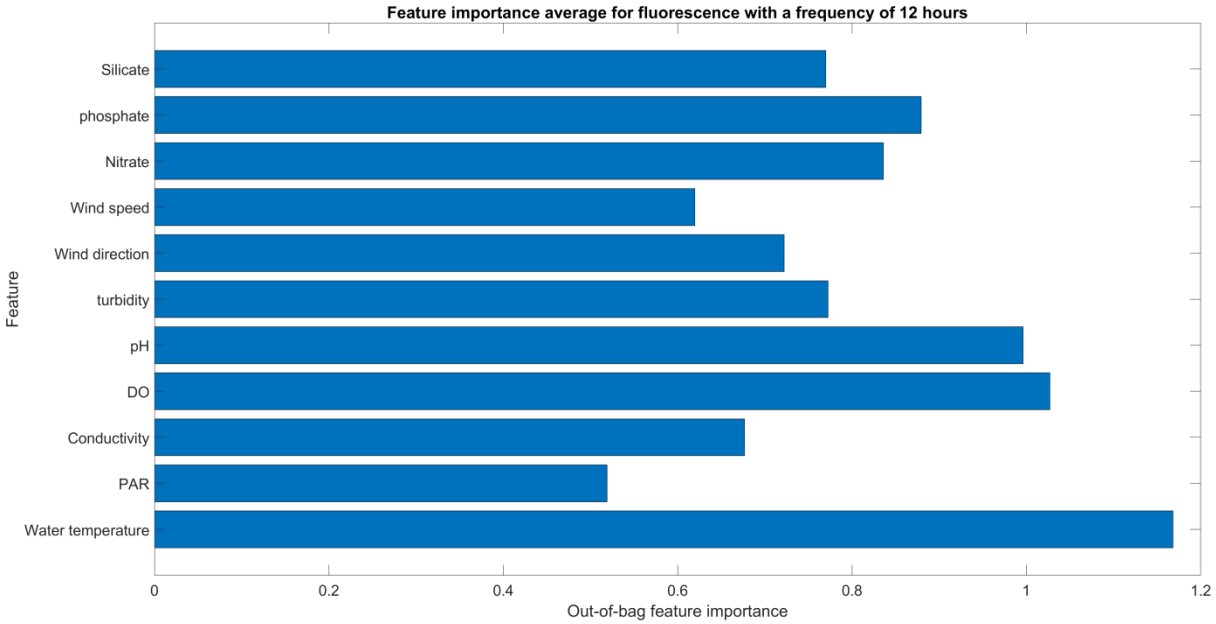


Fig. A10. Ranking of predictor importance based on the average of the out-of-bag (OOB) error, from the 35 runs performed with a time step of 12 hours. Legend of abbreviations: P.A.R for photosynthetically active radiation and DO for dissolved oxygen.

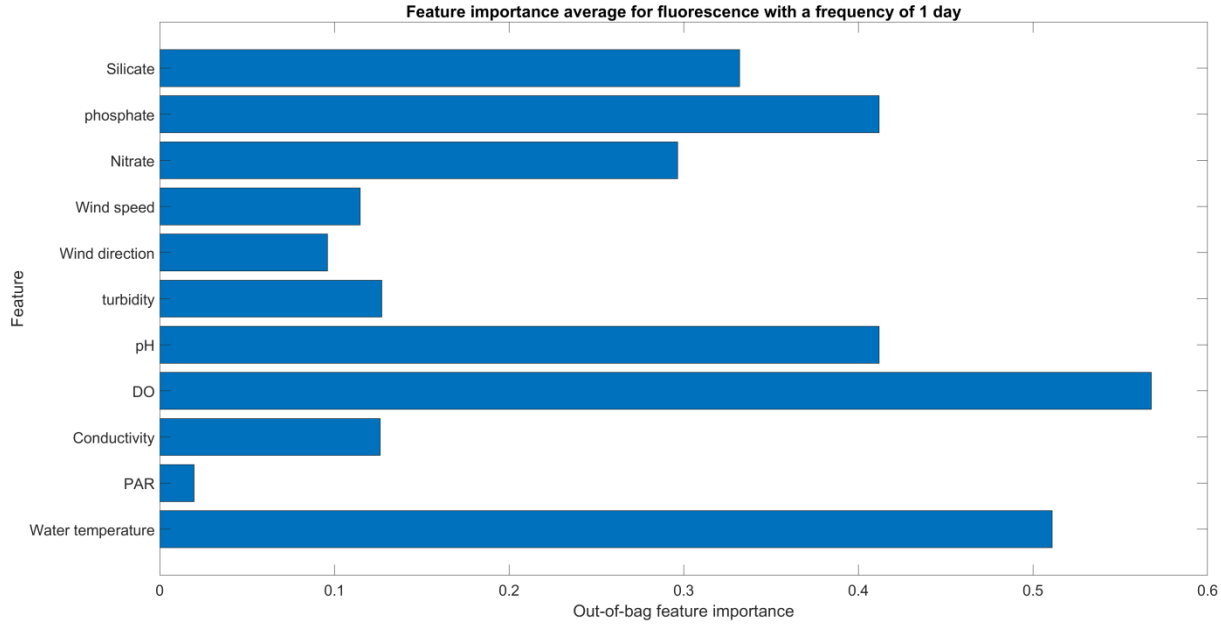


Fig. A11. Ranking of predictor importance based on the average of the out-of-bag (OOB) error, from the 35 runs performed with a time step of 1 day. Legend of abbreviations: P.A.R for photosynthetically active radiation and DO for dissolved oxygen.