Application of the Random Forest model for chlorophyll-*a* forecasts in fresh and brackish water bodies in Japan, using multivariate long-term databases

Hiroshi Yajima and Jonathan Derot

ABSTRACT

There is a growing world need for predicting algal blooms in lakes and reservoirs to better manage water quality. We applied the random forest model with a sliding window strategy, which is one of the machine learning algorithms, to forecast chlorophyll-*a* concentrations in the fresh water of the Urayama Reservoir and the saline water of Lake Shinji. Both water bodies are situated in Japan and have historical water records containing more than ten years of data. The Random Forest model allowed us to forecast trends in time series of chlorophyll-*a* in these two water bodies. In the case of the reservoir, we used the data separately from two sampling stations. We found that the best model parameters for the number of min-leaf, and with/without pre-selection of predictors, varied at different stations in the same reservoir. We also found that the best performance of lead-time and accuracy of the prediction varied between the two stations. In the case of the lake, we found the best combination of a min-leaf and pre-selection of predictors was different from that of the reservoir case. Finally, the most influential parameters for the random forest model in the two water bodies were identified as BOD, COD, pH, and TN/TP.

Key words | algal bloom, chlorophyll-*a*, lake and reservoir, RF model, water quality management

Hiroshi Yajima
Estuary Research Center,
Shimane University,
1060 Nishikawatsu-cho, Matsue, Shimane
690-8504,
Japan

Jonathan Derot (corresponding author)
Irstea, UR EABX,
50 avenue de Verdun, Cestas Cedex F-33612,
France
E-mail: *jonathan.derot@irstea.fr*

INTRODUCTION

For decades, in limnology and oceanography sciences, there has been a consensus regarding the usage of chlorophyll-*a* as a proxy for phytoplankton community biomass. It is also widely recognized that the variation within this first echelon of the food web in the aquatic environment can have a strong impact on the whole of an ecosystem. The intensification of anthropogenic pollution following the evolution of our society has had a direct influence on this phytoplankton biomass – mainly through a disruption in the nutrient flux (Howarth *et al.* 1995). These eutrophication phenomena are partly due to a decrease in the silicate flux and increased nitrogen inputs, which favor the replacement of diatoms by nitrogen-fixing cyanobacteria (Schindler 2006). It has been shown on many occasions that these disruptions are among major ecological issues in freshwater and coastal ecosystems (Smith *et al.* 2006). Furthermore, there is a direct link between inorganic nutrient pollution and the increased occurrence of toxic algae (Camargo & Alonso 2006). The increase of harmful algal blooms (HAB) is generally linked to eutrophication. This kind of phenomena could have a substantial negative impact on human health and the economy in the towns close to these water areas. Therefore, management of water quality in water bodies and the control of HAB are drawing considerable global attention.

Machine learning and data mining methods are numerical tools particularly well adapted in analyzing big datasets. In recent years, their utilization and application have been developed substantially. However, these methods of modeling

doi: 10.2166/hydro.2017.010

and prediction are not popular in studies of aquatic ecosystems compared with other science disciplines. The majority of studies that use machine learning algorithms in the field are performed with artificial neural networks (ANN) (Karunasingha *et al.* 2011). This type of model has been applied to a wide range of studies, such as forecasting the timing and duration of cyanobacteria blooms (Chan *et al.* 2007), and nutrient time series (Markus *et al.* 2010). In most cases, these ANN models have been customized (Ranković *et al.* 2012), or combined with genetic programming (GP) (Muttil & Chau 2006), to improve their performance. Less frequently, some modeling approaches have used different types of machine learning algorithms; such as, K-nearest neighbor (KNN) (Coopersmith *et al.* 2011), K-means (Chang *et al.* 2011), and hidden Markov model (HMM) (Rousseeuw *et al.* 2015).

The Random Forests (RF) model, which uses in part machine learning algorithms, was developed by Leo Breiman in 2001 (Breiman 2001). The model is commonly used in bioinformatics to perform genomic analysis, especially in cancer sciences (Touw *et al.* 2012), but is relatively unknown to environmental science (Kehoe *et al.* 2012). Recently, Harris & Graham (2017) used the RF model to predict cyanobacteria abundance in a eutrophic reservoir using a 14-year dataset. This model stands out from other tree structure-based models in that it uses random selection, which could potentially improve its performance (Breiman 2001). This RF model can also be used to create predictive models (Jiang *et al.* 2007), or forecast models (Lahouar & Slama 2015). The RF model has several advantages including no need for *a priori* determination of the initial assumptions, easy interpretation of outputs (Zhao & Zhang 2008), and the ability to select a small number of parameters – even when there is a large dataset available for inputs (Díaz-Uriarte & De Andres 2006). These favorable attributes seem to be suitable for the study of long-term multivariate databases in aquatic environments. However, the model also has the disadvantage of over-fitting problems (Breiman 2001), which is quite a common issue with ANN model applications (Tu 1996).

The main objective here was to understand the behavior of the RF model when applying it to a forecast model in a reservoir and a lake where long-term data exist. In this study, historical data recorded at Japan's Urayama Reservoir (manmade, with fresh water) and Lake Shinji (natural, with brackish water) was used. These contrasts enabled us to investigate the adaptability of the RF model in different hydrological and salinity conditions affecting algal blooms. In this study, the RF model was applied to forecast algal blooms detected in the form of a chlorophyll-*a* signal.

MATERIALS AND METHODS

Study sites and data

Urayama Reservoir

The Urayama Reservoir has been operated by the Japan Water Agency since April 1999, and is located on the outskirts of Tokyo's metropolitan area (35°57′08″N, 139°03′14″E). The reservoir provides hydro-electricity as well as supplying water for municipal potable supply and irrigation. The reservoir has a watershed of 51.6 km^2, surface water area of 1.2 km^2, total reservoir capacity of 58 × 10^6 m^3, and a maximum water depth of 49.3 m. It receives two inflows from the Urayama River and the Okubodani River, which have an average retention time of ten months (see Figure 1).

The dam's operational center has been conducting water quality sampling using a Van Dorn water sampler at the surface (0.5 m), middle (mid-depth), and bottom layers (1 m from bottom) on a monthly basis, at two points (U1 and U2) in the reservoir (see Figure 1). In an aquatic ecosystem, the greatest proportion of the phytoplankton biomass is mainly located in the euphotic zone. During preliminary tests with the RF model, it was noticed that the addition of sampling data performed at the middle and the bottom did not affect the outputs of the model significantly. Instead, the integration of these additional parameters remarkably increased the computational time in learning phases, and as a consequence, following tests focused on the surface water phytoplankton data. For the two stations of this reservoir, 26 parameters of both water quality and hydrological conditions (see Table 1) were used. Among these parameters, TN/TP and NH4/NOx were artificially created as there are suggestions that stoichiometry has an impact on HABs (Liu *et al.* 2011). It was confirmed that this addition

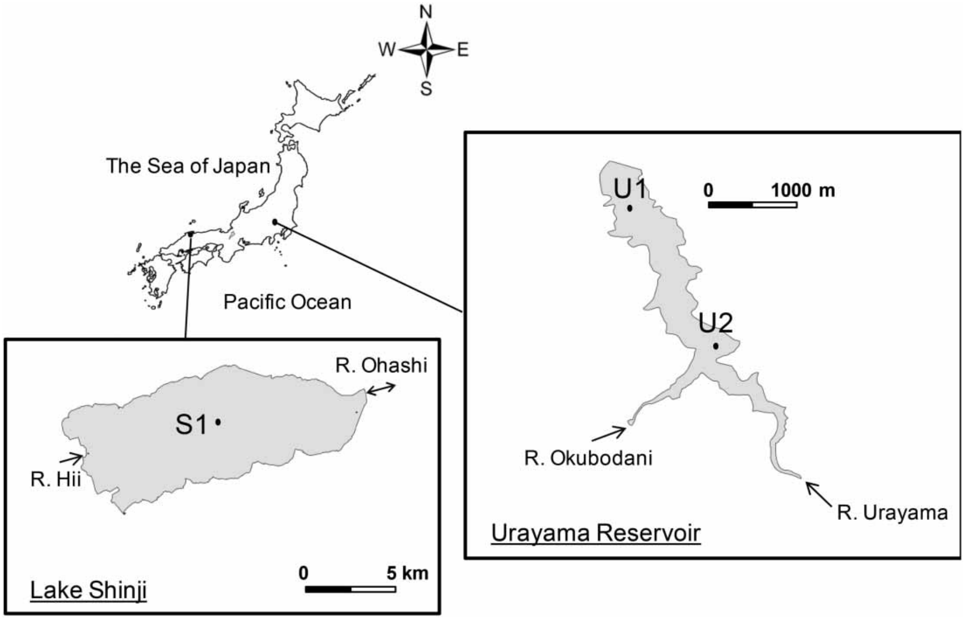

Figure 1 | Locations of sampling stations in the Urayama Reservoir and Lake Shinji.

improved the performance of the model; with little effect on the computational time.

In some cases, this allowed an augmentation of more than 30% for the correlation coefficient in forecasting chlorophyll-*a* concentrations. Furthermore, monthly water quality sampling and analysis included phytoplankton counting (water was collected from the surface by a bucket), which were grouped by taxa in the following manner: Cyanophyceae (CYANO); Bacillariophyceae (BACILLARIO); Chlorophyceae (CHLORO); Chrysophyceae (CHRYSO); Dinophyceae (DINO); Cryptophyceae (CRYPTO); and Euglenophyceae (EUGLENO) (http://mizukoku.nilim.go.jp/ksnkankyo/mizukokudam/system/download/H28D_Chousamanual_dam/H28D_06pura.pdf).

As a result, 33 parameters as inputs of the model for the reservoir were used. It is important to note that a bubble-plume artificial destratifier was installed in 2011 at the bottom of this reservoir; consequently the dynamics of this ecosystem have changed. Therefore, we only used the data between 1999 and 2010 to avoid bias during the learning phase of the model. Target signals of Chl.*a* for this study area can be seen in Figure 2(a), where the black line corresponds to station U1 and the blue line corresponds to station U2. It can be seen that there is some concentration discrepancy between the two stations, especially in 2006 and 2008. The sampling dates are summarized in Appendix 1. Recent data are available on the following website: www.water.go.jp/kanto/arakawa/data/suishitsu.html.

Lake Shinji

Lake Shinji is the seventh largest lake in Japan. It is located in the Shimane Prefecture in the western part of the main island (35°27′01″N, 132°56′58″E) (see Figure 1). It is connected to the Sea of Japan through River Ohashi, Lake Nakaumi, and the Sakai Channel. Seawater occasionally flows into the lake, which contributes to a variation in its salinity. The average salinities of the lake according to the data explained below are 3.6 and 4.0 psu at the surface and bottom, respectively. The variation ranges during the same period are 0.4–10.4 and 0.4–13.5 psu at the surface and bottom, respectively. The lake's surface area is 86.8 km^2 and the average water depth is 4.5 m. There are thirteen inflows (except River Ohashi) into the lake, and one in-out flows of River Ohashi, which makes a retention time of 60 days (Kamiya 1988). The lake is managed by the Ministry of Land, Infrastructure, Transport and Tourism, Japan (MLIT).

Table 1 | Parameters of predictors for the random forest model inputs

			Data availability	
No.	**Parameters**	**Names for predictors**	**Urayama Reservoir**	**Lake Shinji**
1	Air temperature	A. Temp	○	○
2	Water level	WL	○	
3	Inflow discharge	Inflow	○	
4	Outflow discharge	Outflow	○	
5	Transparency	Transp	○	○
6	Water depth	Depth	○	○
7	Water temperature	W. Temp	○	○
8	Turbidity	Turb	○	○
9	Dissolved oxygen	DO	○	○
10	pH	pH	○	○
11	Biochemical oxygen demand	BOD	○	
12	Chemical oxygen demand	COD	○	○
13	Chlorophyll-*a*	Chl.a	○	○
14	Pheophytin	Pheop	○	
15	Suspended solids	SS	○	○
16	E. coli	E. coli	○	○
17	Total nitrogen	TN	○	○
18	Nitrate	NO_3-N	○	○
19	Nitrite	NO_2-N	○	○
20	Ammonium	NH_4-N	○	○
21	Total phosphate	T-P	○	○
22	Phosphate	PO_4-P	○	
23	Dissolved total phosphorus	DTP	○	○
24	Dissolved inorganic phosphorus	DIP	○	○
25	Total nitrogen/total phosphorus	TN/TP	○	○
26	Ammonium/nitrate-nitrite	NH4/NO_x	○	
27	Cyanophyceae	CYANO	○	
28	Bacillariophyceae	BACILLARIO	○	
29	Chlorophyceae	CHLORO	○	
30	Chrysophyceae	CHRYSO	○	
31	Dinophyceae	DINO	○	
32	Cryptophyceae	CRYPTO	○	
33	Euglenophyceae	EUGLENO	○	
34	Dissolved inorganic nitrogen	DIN		○
35	Dissolved COD	DCOD		○
36	Total organic carbon	TOC		○
37	Chlorophyll-b	Chl.b		○
38	Chlorophyll-c	Chl.c		○
39	Chlorine	Cl^-		○

Note: In total, the Urayama Reservoir and Lake Shinji have 33 and 25 parameters, respectively. Water parameters were analyzed by the standard protocols based on Japanese Industrial Standard (JIS) (www.mlit.go.jp/river/shishin_guideline/kasen/suishitsu/houhou.html).

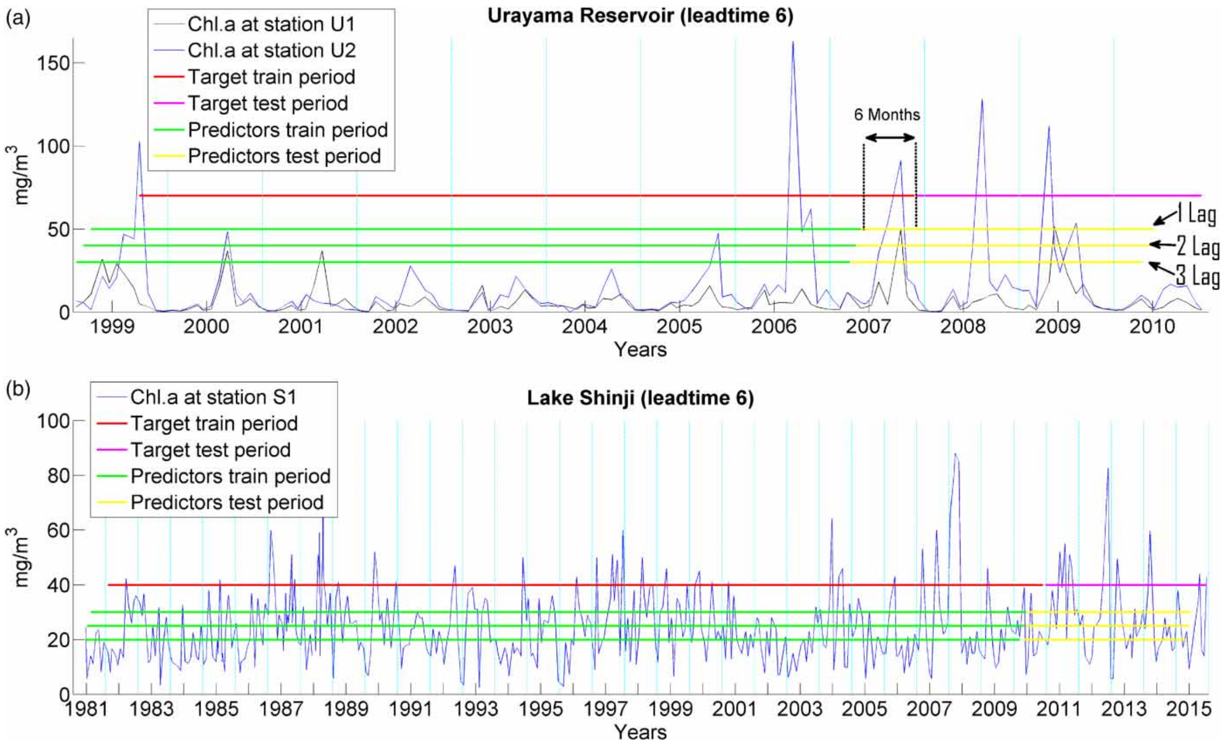

Figure 2 | Historical Chl.*a* data at two study sites, and example of lead-time equal to six months with three lag-time data periods. The top panel (a) corresponds to the Urayama Reservoir, and the bottom one (b) corresponds to Lake Shinji. Please refer to the online version of this paper to see this figure in colour: http://dx.doi:10.2166/hydro.2017.010.

MLIT has a long-term monthly water quality data at the surface (1.0 m) and the bottom (1.0 m from bottom) of S1 (see Figure 1). This study used the data collected between 1981 and 2015. The sampling dates are summarized in Appendix 1. Some parameters of these historical data do not have complete time series. As to the inorganic nitrogen, the missing values for NH4-N, NO2-N and NO3-N are not necessarily present at the same time. Therefore, we obtained 40.1% of missing values of the NH4/NOx ratio. Consequently, we decided not to use this ratio as a predictor for this lake. It should also be noted that, contrary to the previous site, there is no information about phytoplankton species. Finally, 25 parameters for inputs of the model (see Table 1) were used. In Figure 2(b), the target signal of Chl.*a* (see blue curve) between 1981 and 2015 used for this study area can be seen. Moreover, the data between June and November from 1985 to 1988 has two datasets, except August 1988, which has three datasets (see Appendix 1). These data were simply averaged during the same month, which allowed a regular time step of one month to be kept. Recent data are available on the following website: www.cgr.mlit.go.jp/izumokasen/shiryokan/jokyo/index.html.

RF and CART models

The RF model is based on a tree structure proposed by Breiman (2001) as an evolution of the Classification And Regression Tree (CART) model created in 1984 by the same author (Breiman *et al.* 1984). The algorithm performs a movement from the root to the terminal's nodes, which contain the predictions. This movement is done iteratively where each child node is subdivided into branches (see Appendix 2 for additional information, equations, and diagrams about these two models). In the case of the CART model, only one decision tree is built, whereas a predetermined number of decision trees are used for the RF model. In other words, the RF model is made up of a multiplicity of CART models (see Figure A2.1 in Appendix 2). Hence the term *forest* is used. In order to increase the robustness of this model, it uses a statistical inference technique, called a *Bootstrap*. Each of the trees which make up a particular forest is built up from random sub-sampling of datasets. Therefore, if several runs with the same database are performed without saving all the tree diagrams, a small difference between outputs of this RF model is found. This bootstrap method is based on a random draw with replacement. Hence the term *Random* is used in the name of this model. The final prediction of the model output is determined by an ensemble of methods among all results from each tree making up the forest. This can be called the 'majority vote'.

When applying the RF model to a prediction or a forecast, it requires a longer period for the learning phase than the prediction or forecast period to have accurate results. In Figure 2(a) and 2(b), the separation between the learning (see red and green lines) and forecast period (magenta and yellow lines) can be seen. Considering the data availability, the forecast periods of the Urayama Reservoir and Lake Shinji were set for two and five years, respectively. In this study, the construction of the RF models was performed using the function *TreeBagger* from Matlab software.

One of the techniques to avoid over-fitting problems in the RF model application is to limit the minimum leaf size (min-leaf). This parameter determines the minimum number of observations that is used to create each child node; smaller values of the min-leaf need a deeper learning process. The technical documentation of this Matlab function recommends using a minimal number of five observations per child node (min-leaf = 5) by default.

Forecast method

In order to test the performance of the RF model to forecast Chl.*a* in the study areas, a sliding window strategy (Herrera *et al.* 2010) was implemented. The behavior of the model through a variation of two parameters (lead-time and lag-time) was also studied. The lead-time corresponds to shift-time of the sliding window (see Figure A3.1 in Appendix 3). That is to say, the shift-time that presents itself between the date of forecasting point and the date of predictors that are used in parallel as inputs to the RF model. In this study, the lead-time between one and six months was varied. In Figure 2(a) and 2(b), the red line represents the learning periods of the target signal (Chl.*a*), and the three green lines on each panel represent the learning periods used with predictors (lag time = 1, 2, and 3, respectively). The time difference between the right end of the red line and the right end of the first green line on each panel (lag time = 1) corresponds to a lead-time of six months in this case (see the black annotation on Figure 2(a)).

Using multiple lag-times for predictions may allow better results of Chl.*a* prediction to be achieved. In this study, the parameter from one to three months was tested. All predictors at the lead-time in cases where the offset was equal to 1 were used. In the case where the offset was equal to 2 or 3, all predictors with offsets of one and two months before the lead-time were added, respectively. Hence, the increase in lag-time increased the predictor number inputs of the model. For example, in the case of the Urayama Reservoir with a lag-time equal to 3, the total number of input parameters expanded from 33 to 99. These 3-different lag-times are represented in green for the learning periods and in yellow for the prediction periods in Figure 2(a) and 2(b).

In the process of the RF model application, an average of out-of-bag (OOB) error permutation performed on all trees of the forest was calculated; which is a parameter to show the relative importance of each predictor used in the input of the model. In addition to varying the min-leaf (MF), tests with an automatic pre-selection of predictors in the process of the learning were also performed. In these cases, only the ten most influential predictors with the most influence were selected by a pre-run for the final run.

Protocol of testing strategy

As a preliminary study, sensitivity tests to evaluate the suitable min-leaf number for our cases were performed. Then the RF model was applied to each station (U1, U2, and S1). First, the min-leaf and lag-time variation was tested with four different cases: (1) the default settings proposed in the RF model (the min-leaf is equal to 5 and no pre-selection for inputs. Consequently, all predictors were used; (2) the min-leaf is equal to 12 and no pre-selection for inputs; (3) the min-leaf is equal to 5 and automatic pre-selection for inputs; (4) the min-leaf is equal to 12 and automatic pre-selection for inputs. Then raw Chl.*a* data at each station was compared with the best result among three lag-times for four test cases (min-leaf = 5; min-leaf = 12; min-leaf = 5 and pre-selection; min-leaf = 12 and pre-selection). Depending on the different model settings, this gave a better understanding of the RF model prediction.

CART model application was also explored for our study, and showed the RF model outperforming the CART model. Consequently, only the results for the RF model are described (see Appendix 4 for the performance test between the RF and CART models).

RESULTS AND DISCUSSION

In this study we applied the RF model to forecast the general pattern of Chl.*a* in two different water bodies in Japan. First, we have summarized the results of sensitivity tests of the RF model according to the optimal tree number and the minimal number of the predictor (min-leaf parameter). Second, we have shown the performances of the forecast for both the Urayama Reservoir and Lake Shinji.

Finding the optimal tree number and leaf size

We calculated the mean-squared error (MSE) from the OOB errors provided by the *TreeBagger* toolbox for five different min-leaf: 5, 10, 12, 20, and 50, and between 1 and 1,000 decision trees for U2 at the Urayama Reservoir. The minimum number of trees required, which is necessary to achieve the best results, depends on the nature and the quality of the input dataset. Using a small number of trees allowed us to minimize the computing time without affecting the final results of this test. In our case, the MSE was stable for more than 200 trees (see Figure 3, and Figures A5.1–A.5.5 in Appendix 5). Consequently, we decided to use 200 trees in the rest of our study for three sampling stations. We also see that after the stabilization phase of these MSE, the min-leaf values 5 and 12 seemed to have lower MSE than the others. Therefore, we used these two values for our application. It should be noted that the results achieved for other two sampling points were consistent with those presented in Figure 3 (see Figures A5.6–A5.15 in Appendix 5).

Urayama Reservoir forecast

Station 1 (U1) forecast

In the case of station U1, the correlation coefficients of R^2 were generally improved when we used a min-leaf of 12,

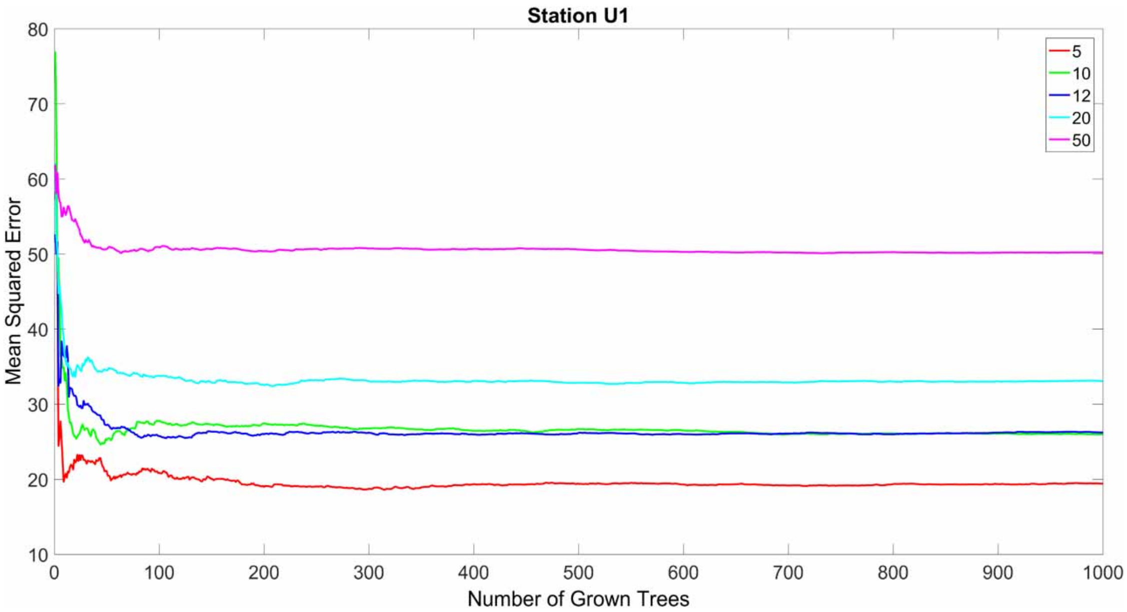

Figure 3 | Sensitivity tests of the RF model at station U1 in the Urayama Reservoir depending on the minimal number of the predictor (min-leaf parameter) used for the separation between each node. In the legend, each min-leaf corresponds to a different color: red for 5, green for 10, blue for 12, cyan for 20, and purple for 50. Please refer to the online version of this paper to see this figure in colour: http://dx.doi:10.2166/hydro.2017.010.

with and without pre-selection (see Figure 4(a)–4(d)). In these two cases, when min-leaf was equal to 12, the R^2 variation corresponding to lead-time showed a 'v' shape. That is to say, there was a strong decline of R^2 values when lead-time was two or three months, and the highest R^2 values were performed for a lag of 1: $R^2 = 0.41$ for a lead-time of one month and without pre-selection, and $R^2 = 0.36$ for a lead-time of five months and with pre-selection (see the red line in Figure 4(b) and 4(d).

Raw Chl.*a* data at station U1 was plotted in parallel with the best results for four different previous cases in Figure 4 (min-leaf = 5; min-leaf = 12; min-leaf = 5 and pre-selection; min-leaf = 12 and pre-selection) (see Figure 5). In this figure, we used the same color code as in Figures 7 and 9. The general variations of Chl.*a* were forecasted in four cases. Although the first outbreak of bloom at the fifth month was practically undetected, the second bloom around the tenth month was predicted. The most severe bloom, around the seventeenth month, was also underestimated in each case. Regarding the two last blooms (months 28 and 34), all the outputs of the model were coherent, despite the overestimation found around the twenty-fifth month. The best correlation coefficient ($R^2 = 0.41$) of the forecast among the four different settings was the solid blue line (min-leaf of 12 without pre-selection).

Station 2 (U2) forecast

The different tendencies for the variations of R^2 were found at station U2 compared with station U1 (see Figures 4 and 6). In the case of station U2, the correlation coefficients of R^2 were not improved by using a min-leaf of 12 instead of 5 (see Figure 6(a)–6(d)). Moreover, the highest R^2s were mainly observed for a lead-time of three months. The lowest values were generally seen around lead-times of two and five months. In these cases, the R^2 variations follow a pattern of a 'w' shape, and we see the highest value with a min-leaf 5 and with pre-selection ($R^2 = 0.61$) (see Figure 6(c)). Even though these two stations were located in the same reservoir, we found that the best lead-times differed and the best result was obtained with a lag-time of one month. We also found overestimations for several blooms (months eight and 17), and an underestimation around the twenty-fifth month at stations U1 and U2 in Urayama Reservoir (Figures 5 and 7, respectively).

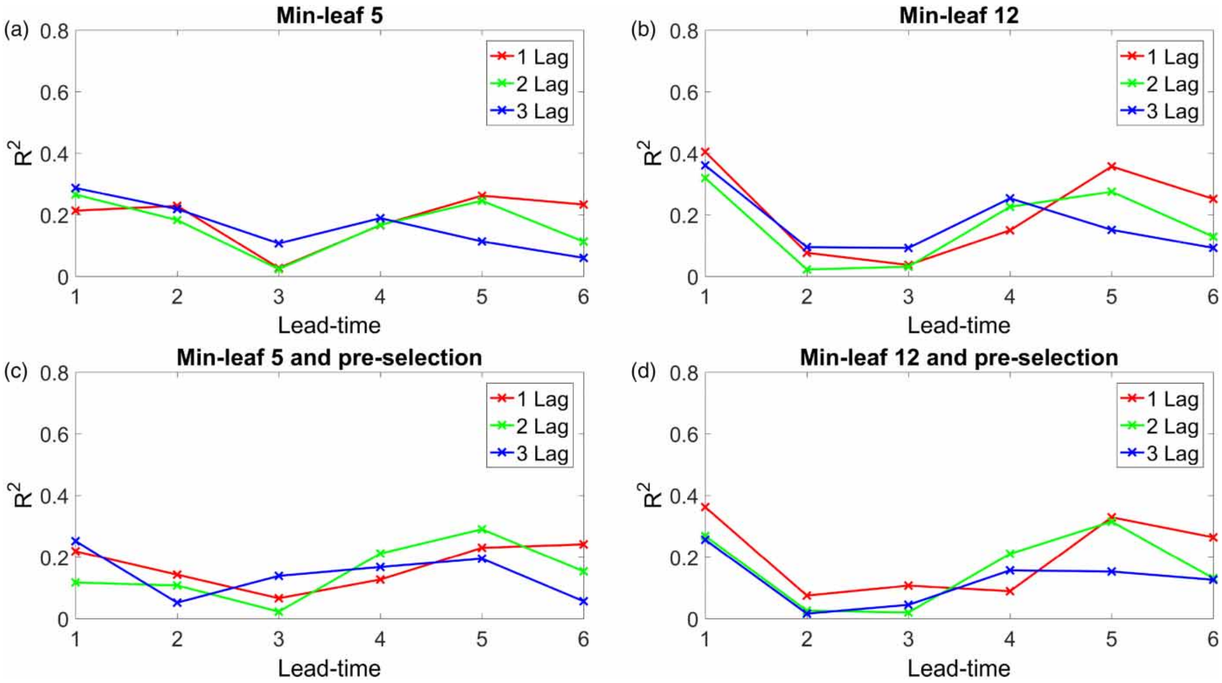

Figure 4 | Evolution of R^2 coefficient depending on lead-time and lag-time for station U1 in the Urayama Reservoir. Each panel corresponds to a different setup: min-leaf equal to 5 or 12 and with/without pre-selection.

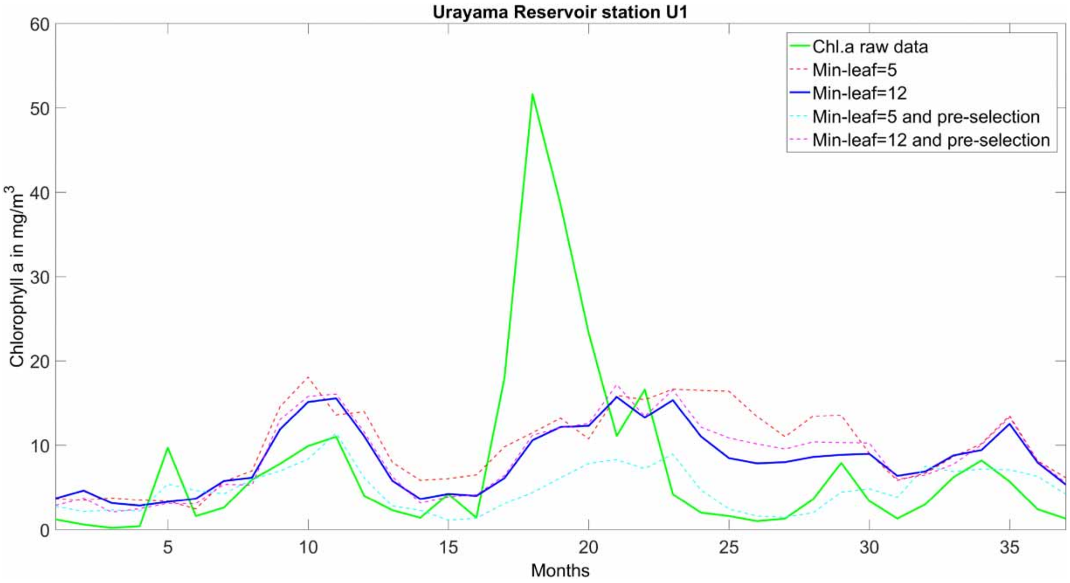

Figure 5 | Comparison between the raw Chl.*a* data at station U1 in the Urayama Reservoir and the best result of each setup used in Figure 4. Color code: the raw Chl.*a* data in green; the min-leaf is equal to 5 in red (see Figure 4(a)); the min-leaf is equal to 12 in blue (see Figure 4(b)); the min-leaf is equal to 5 with pre-selection in cyan (see Figure 4(c)); the min-leaf is equal to 12 with pre-selection in magenta (see Figure 4(d)). Please refer to the online version of this paper to see this figure in colour: http://dx.doi:10.2166/hydro.2017.010.

Lake Shinji station (S1) forecast

For Lake Shinji, the best results were obtained with a lag of 1 and 2 (see Figure 8). In Figure 8(a)–8(d), the highest R^2 coefficients were located for a lead-time of one month. The strongest correlation ($R^2 = 0.52$) was found with the following settings: min-leaf is equal to 12 with pre-selection for a lag-time of two (see Figure 8(d)). The correlation

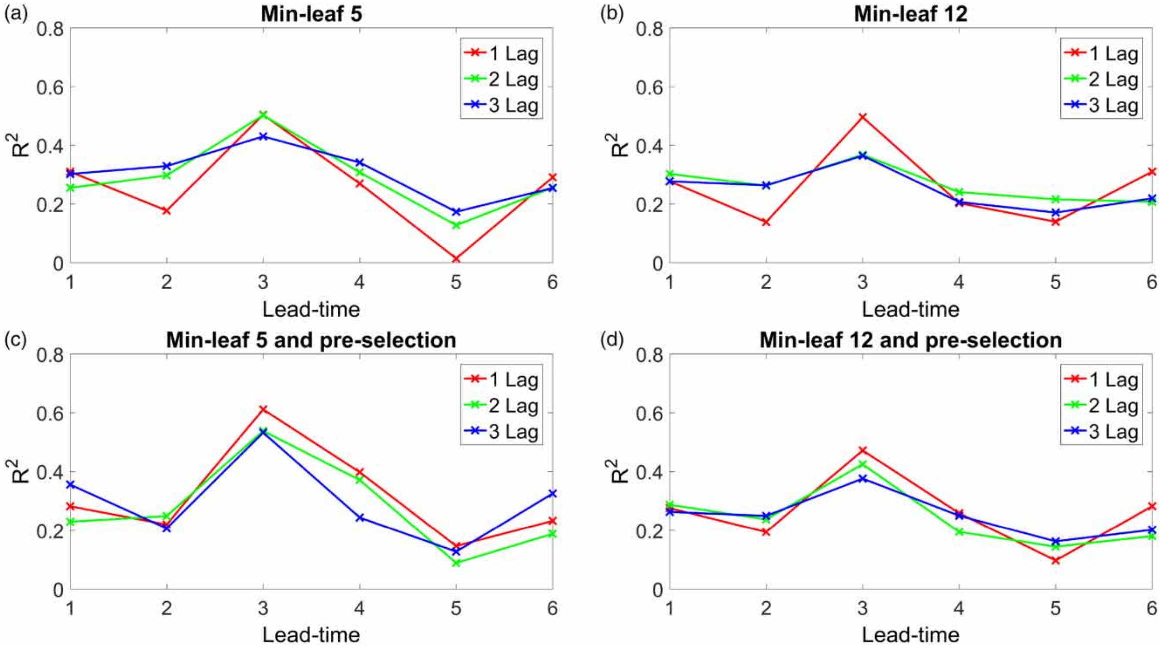

Figure 6 | Evolution of R^2 coefficient depending on lead-time and lag-time for station U2 in the Urayama Reservoir. Each panel corresponds to a different setup: min-leaf equal to 5 or 12 and with/without pre-selection.

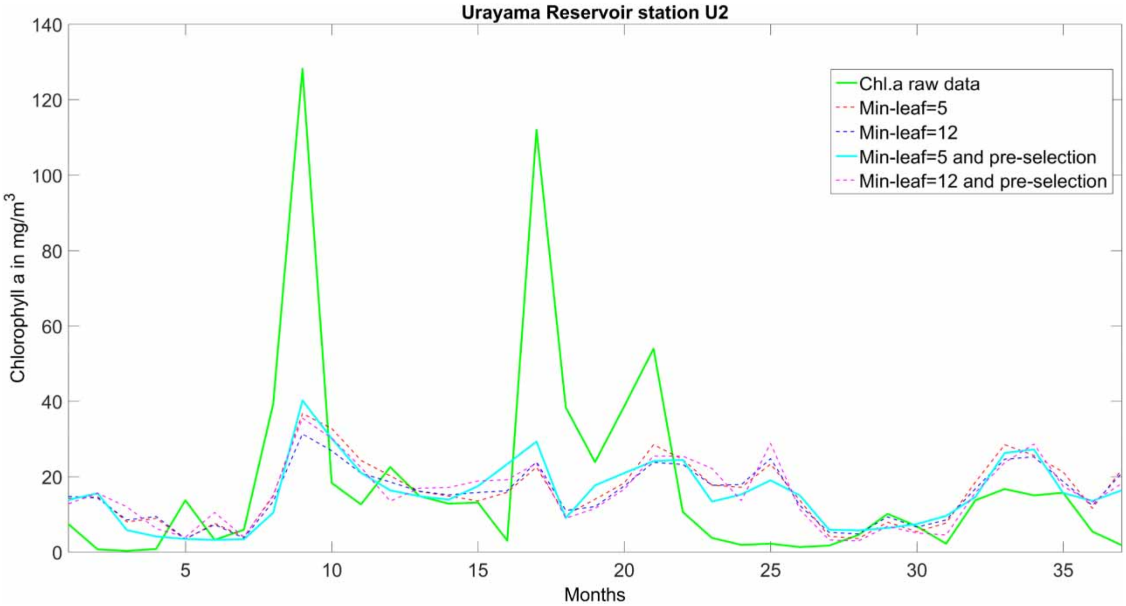

Figure 7 | Comparison between the raw Chl.*a* data of for station U2 in the Urayama Reservoir and the best forecast results (lead time of three months) of each setup we used in Figure 6. Color code is the same as in Figure 5.

coefficients R^2 were generally improved when we used a min-leaf of 12 instead of 5. We can see this output of model forecast (lead time of one month) in Figure 9 via the purple solid line. Overall, the raw Chl.*a* data dynamics were well forecasted by the RF model, with the exception of the underestimations of blooms around the twenty-third and fortieth month. It can be noticed that the best R^2 of 0.52 was slightly higher than the Urayama Reservoir's case

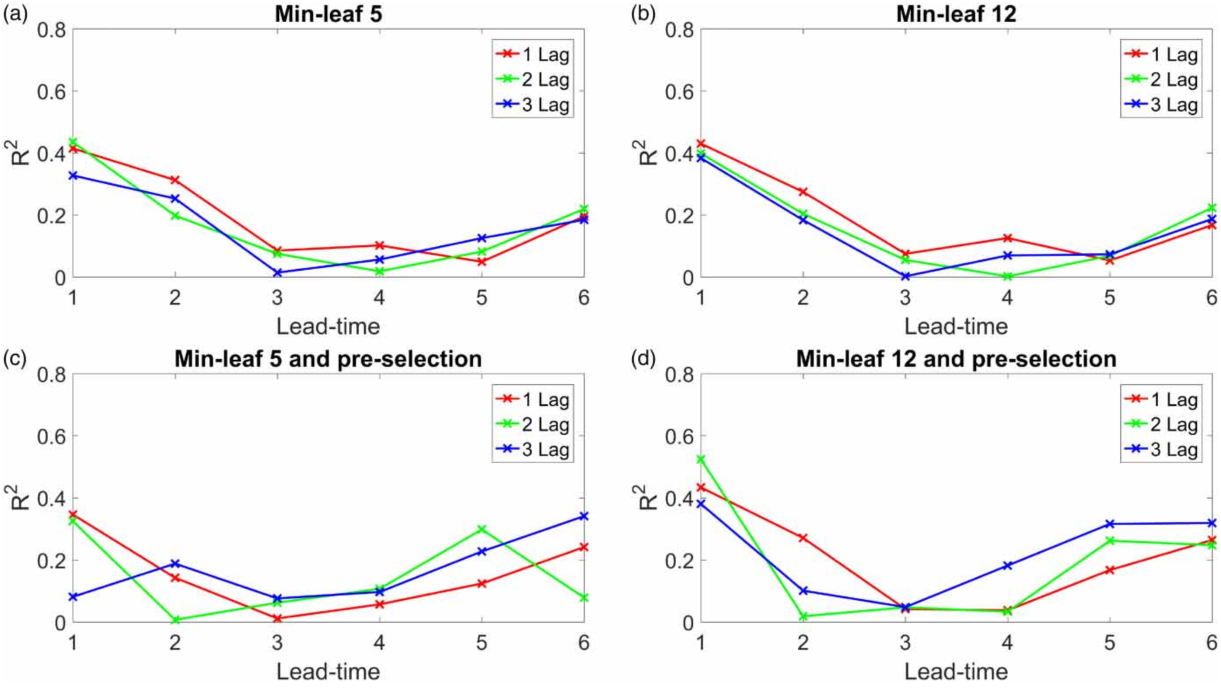

Figure 8 | Evolution of R^2 coefficient depending on lead-time and lag-time in Shinji Lake. Each panel corresponds to a different setup: min-leaf equal to 5 or 12 with/without pre-selection.

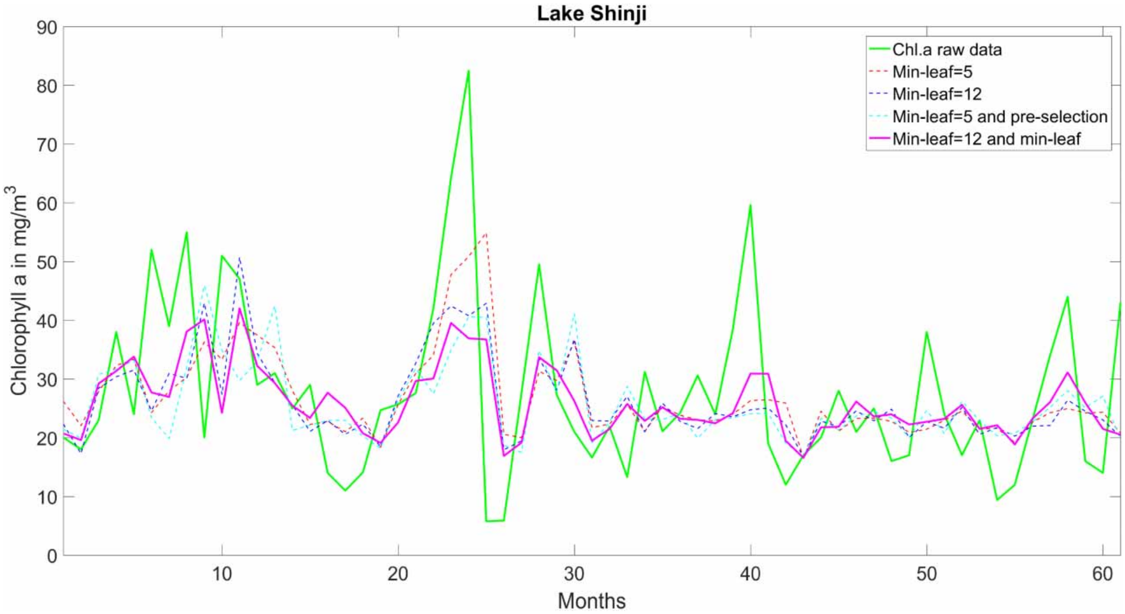

Figure 9 | Comparison between the raw data of Chl.*a* in Lake Shinji and the best forecast results (lead-time of one month) of each setup we used in Figure 8. The color code is same as in Figures 5 and 7. Please refer to the online version of this paper to see this figure in colour: http://dx.doi:10.2166/hydro.2017.010.

($R^2 = 0.41$). It may also be noted that the increase in a min-leaf does not always increase R^2 coefficients (see Figure 8 (a)–8(d)). This tendency was the same as for station U2 but different from station U1.

Lake Shinji has much longer datasets than the Urayama Reservoir. In order to examine the impact of the data length on the forecast capability of the RF model, we reduced the original dataset of 30 years for the learning period to 80, 60, and 40% for the case of min-leaf = 12 and with pre-selection. We can see that the correlation coefficient decreased as the dataset for the learning period reduced (see Figure 10). This shows that the dataset volume is one of the key factors that affects the model performance.

Importance of parameters

The ranking of predictor importance for each station was summarized based on an average of OOB permutated error after 100 runs (see Table 2). The RF model uses a bootstrap approach, and we obtained a slightly different predictor importance for each run. To minimize the random effect linked to bootstrap, we used average values. For stations U1 and U2 at the Urayama Reservoir, the following four parameters were commonly important for the forecast of Chl.*a*: BOD (0.67, 0.71), COD (0.63, 0.65), pH (0.43, 0.31) and TN/TP (0.30, 0.26). For S1 at Lake Shinji, the two parameters of COD (0.43) and pH (0.23) showed important roles as found for the Urayama Reservoir, even though Turb. (0.51) showed a higher value than COD.

In the scientific literature, we found several studies which are in-line with the results obtained in Table 2. The COD, which strongly influenced the RF model in our two types of aquatic ecosystems, is commonly used for highlighting the organic contamination and the phenomena of eutrophication. A study based on a long-term monitoring (nine years) of this parameter in Tokyo Bay (Kawabe & Kawabe 1997) demonstrated that, first, amplitude variations of COD concentration are associated with the nitrogen and phosphorus concentrations, and second, the seasonal variation of COD (spring bloom and winter period) are related to the solar radiation and water temperature with a lag of one month. Another study, which used a machine learning model (ANN) to perform predictions of BOD in a Turkish river, showed that the COD was the most influential parameter for this model (Dogan *et al.* 2009). As regards pH, a study showed that it could influence the diatom growth rate (Chenl & Durbin 1994). Another author suggested that

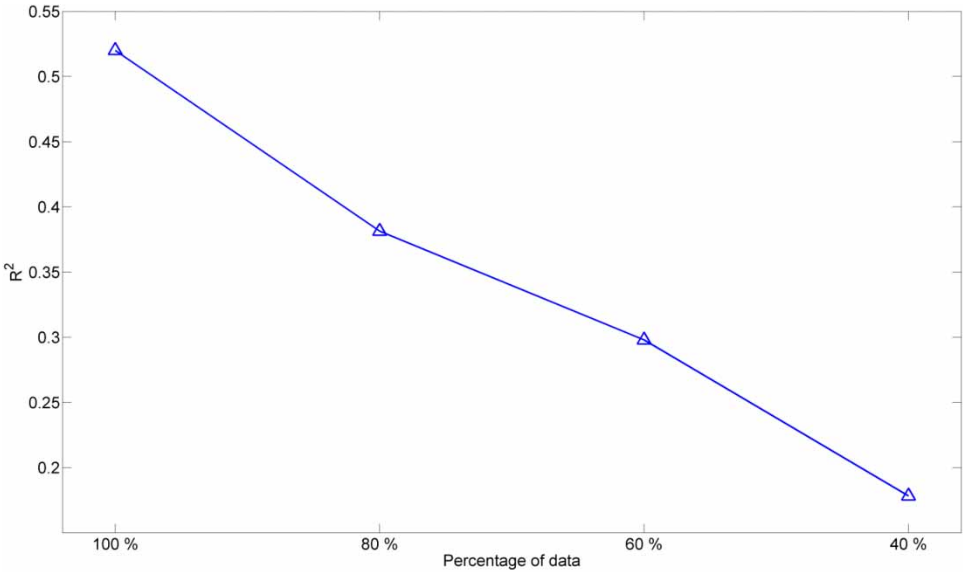

Figure 10 | Correlation coefficient depending on the percentage of the learning period for Lake Shinji. The dataset for the learning period of 30 years was reduced by up to 40% in the case of min-leaf = 12, and pre-selection.

the pH could also have an impact on the species succession of phytoplankton (Hansen 2002). In the case of predicting cyanobacteria abundance in a eutrophic reservoir using the RF model (Harris & Graham 2017) compared with the Urayama Reservoir case, pH and TN/TP were not important (they did not have BOD and COD as parameters). Also, temperature was one of the top three important parameters, it was not in the Urayama Reservoir's case. These differences may be due to a different constitution of phytoplankton, a different limiting factor for phytoplankton: nitrogen or phosphorus, or hydrological conditions. More importantly, the different dataset of input parameters as predictors may have affected more than these. However, it is difficult to clarify the reason because of the nature of the RF model.

Table 2 | Ranking of predictor importance for each station based on an average of out-of-bag (OOB) permutated error after 100 runs

Ranking	Station U1	Station U2	Station S1
1	BOD (0.67)	BOD (0.71)	Turb. (0.51)
2	COD (0.63)	COD (0.65)	COD (0.43)
3	pH (0.43)	TP (0.36)	SS (0.41)
4	DIP (0.35)	CHLORO (0.33)	Chl.c (0.36)
5	BACILLARIO (0.34)	pH (0.31)	Chl.b (0.29)
6	TN/TP (0.30)	TN/TP (0.26)	TN (0.26)
7	Pheop. (0.29)	SS (0.24)	pH (0.23)
8	WL (0.28)	W. Temp. (0.19)	TP (0.22)
9	SS (0.27)	WL (0.16)	NO3 (0.19)
10	CHLORO (0.25)	DIP (0.13)	Transp. (0.13)

Values in parentheses indicate average OOB after 100 runs. Detailed figures are shown in Appendix 6

We also calculated the correlation matrix between parameters for each sampling station to compare OOB rankings (see Appendix 7). In the case of the Urayama Reservoir, important parameters with correlation coefficients $R^2 > |0.75|$ between Chl.*a* were BOD (0.87, 0.91) and COD (0.80, 0.92) for both U1 and U2. They ranked first and second by OOB ranking (see Table 2). In the case of Lake Shinji, $R^2 > |0.75|$ were TN (0.96), COD (0.93), SS (0.93) and TP (0.82), which were ranked second, third, sixth and eighth by OOB ranking, respectively (Table 2). While the turbidity was ranked first, the correlation coefficient with Chl.*a* was 0.31. Therefore, a strong correlation with Chl.*a* was not necessarily the most important parameter for the RF model. This is one of the advantages of machine learning models, which are able to detect conditional linkages between some parameters that cannot be found directly with basic statistics. In addition, the RF model is used for forecasting rather than prediction, which may explain some of the differences observed between the OOB ranking and the correlation matrix.

Forecast capability of RF model

Usually RF models are used to perform predictions (Jiang *et al.* 2007; Harris & Graham 2017). That is to say, learning periods are the same for the target signal and the predictors. In our case, we separated the datasets into two stages of the learning period and test period. We also introduced a lead-time for forecasting and lag-time for predictors (see Figure 2 and Appendix 3). This allowed us to test the forecast capability of this model. The best correlation coefficients results were 0.41, 0.61, and 0.52 for U1, U2, and S1, respectively. We could not forecast Chl.*a* values accurately. The main reason for this may be due to the data availability on a monthly basis. The hydrological and water quality conditions can significantly change in a month and the abundance of each phytoplankton taxa can vary substantially as well. Moreover, non-uniform horizontal distribution, or even diurnal vertical distribution for the phytoplankton existence, could affect the results. These issues can be solved if applying output datasets calculated by a detailed ecosystem simulation model, as was examined for the Urayama Reservoir, using a 3D hydrodynamic and ecosystem model (Yajima & Choi 2013). In the case of forecasting, as long as we can predict the evolution of phytoplankton, and know the algal bloom in advance, this model will be useful for water quality management.

CONCLUSIONS

We applied the RF model to both natural and artificial ecosystems, and we could forecast the general trend of the Chl.*a* signal. When Chl.*a* became relatively high, the model tended to make overestimates or underestimates. This kind of bloom can be interpreted as an extreme event, or it can be interpreted as a stochastic process at the level of their probability density. Consequently, it was difficult to make an accurate modeling for these events with a non-regular monthly basis data.

In the process of applying the RF model to the three stations, with two different types of water bodies, we obtained some important information. For the choice of tree number, it was enough to use 200 trees per forest in the model. As to the min-leaf number, we examined 5, 10, 12, 20 and 50 leafs, and a min-leaf of 12 generally performed well. Moreover, we tested a selection of input parameters for predictors. In general, the pre-selection outperformed without pre-selection application. This allowed us to save some computation time for the analysis. As to the lead-time for the forecast, we obtained the best results at one month lead-time for stations U1 and S1, and a three month lead-time for station U2, even though it is difficult to understand the reason for this.

From the ranking of predictor importance analysis, a strong correlation with Chl.*a* was not necessarily the most important parameter for the RF model forecast in the case of Lake Shinji. This suggested that the most important predictors did not necessarily have a strong statistical correlation with a target parameter. In addition, the size of the database had a great impact on the forecast performance in Lake Shinji and a smaller dataset decreased the performance. Although we only studied two different types of ecosystems, the number of predictors used in the inputs of the model, as well as the size of the datasets, may have affected the model's performance rather than the type of the ecosystem.

In our future work, we will use long-term high frequency time series in order to evaluate the effect of database contents for the forecast capability of the RF model. Moreover, we need to explore a new technique for forecasting, as shown in a recent study, which used a cubist model in the prediction of high-intensity-cyanobacteria-bloom that excelled the RF model (Harris & Graham 2017). In our next study, we would also like to include the most representative phytoplankton taxa independently as predictors, instead of grouping, which might be helpful in considering phytoplankton competition and coexistence. Furthermore, other predictors that have a connection with the specific characteristics of each taxon, in the form of morpho-functional characteristics (presence of flagella, cell shape, colonial or solitary species, and so on), may be included. In this way we would add a layer of information which could greatly improve the learning performed by the RF model. In this case, we will not only have quantitative but also qualitative data, which can be handled easily by the RF model.

ACKNOWLEDGEMENTS

We would like to thank the Arakawa River Dam joint operational center, the Japan Water Agency; the Izumo River Office, the Ministry of Land, Infrastructure, Transport and Tourism, Japan, for providing the data used in this study. We are also grateful to Prof. J. Brooke (The University of Adelaide) for his helpful comments on this manuscript. www.englisheditor.webs.com is thanked for its English proofing.

REFERENCES

Breiman, L. 2001 Random forests. *Mach. Learn.* **45** (1), 5–32.

Breiman, L., Friedman, J., Stone, C. J. & Olshen, R. A. 1984 *Classification and Regression Trees.* Wadsworth International Group, Belmont, California.

Camargo, J. A. & Alonso, Á. 2006 Ecological and toxicological effects of inorganic nitrogen pollution in aquatic ecosystems: a global assessment. *Environ. Int.* **32** (6), 831–849.

Chan, W. S., Recknagel, F., Cao, H. & Park, H. D. 2007 Elucidation and short-term forecasting of microcystin concentrations in Lake Suwa (Japan) by means of artificial neural networks and evolutionary algorithms. *Water Res.* **41** (10), 2247–2255.

Chang, K., Gao, J. L., Wu, W. Y. & Yuan, Y. X. 2011 Water quality comprehensive evaluation method for large water distribution network based on clustering analysis. *J. Hydroinform.* **13** (3), 390–400.

Chenl, C. Y. & Durbin, E. G. 1994 Effects of pH on the growth and carbon uptake of marine phytoplankton. *Mar. Ecol. Prog. Ser.* **109** (83–94), 83–94.

Coopersmith, E. J., Minsker, B. & Montagna, P. 2011 Understanding and forecasting hypoxia using machine learning algorithms. *J. Hydroinform.* **13** (1), 64–80.

Díaz-Uriarte, R. & De Andres, S. A. 2006 Gene selection and classification of microarray data using random forest. *BMC Bioinform.* **7** (1), 1.

Dogan, E., Sengorur, B. & Koklu, R. 2009 Modeling biological oxygen demand of the Melen River in Turkey using an artificial neural network technique. *J. Environ. Manage.* **90** (2), 1229–1235.

Hansen, P. J. 2002 Effect of high pH on the growth and survival of marine phytoplankton: implications for species succession. *Aquat. Microb. Ecol.* **28** (3), 279–288.

Harris, T. D. & Graham, J. L. 2017 Predicting cyanobacterial abundance, microcystin, and geosmin in a eutrophic drinking-water reservoir using a 14-year dataset. *Lake Reservoir Manage.* **1** (17), 1040–2381.

Herrera, M., Torgo, L., Izquierdo, J. & Pérez-García, R. 2010 Predictive models for forecasting hourly urban water demand. *J. Hydrol.* **387** (1), 141–150.

Howarth, R., Jensen, H. S., Marino, R. & Postma, H. 1995 Transport to and processing of P in near-shore and oceanic waters. In: *Phosphorus in the Globale Environment: Transfers, Cycles, and Management* (H. Tiessen, ed.). Wiley, New York, pp. 323–345.

Jiang, P., Wu, H., Wang, W., Ma, W., Sun, X. & Lu, Z. 2007 Mipred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. *Nucleic Acids Res.* **35** (suppl 2), W339–W344.

Kamiya, H. 1988 Calculation of back flow into Lake Shinji and Lake Nakaumi. *Report of the Shimane Prefectural Institute of Public Health and Environmental Science* **30**, 94–95 (in Japanese).

Karunasingha, D. S., Jayawardena, A. & Li, W. 2011 Evolutionary product unit based neural networks for hydrological time series analysis. *J. Hydroinform.* **13** (4), 825–841.

Kawabe, M. & Kawabe, M. 1997 Factors determining chemical oxygen demand in Tokyo Bay. *J. Oceanogr.* **53** (5), 443–453.

Kehoe, M., O'Brien, K., Grinham, A., Rissik, D., Ahern, K. & Maxwell, P. 2012 Random forest algorithm yields accurate quantitative prediction models of benthic light at intertidal sites affected by toxic Lyngbya majuscula blooms. *Harmful Algae* **19**, 46–52.

Lahouar, A. & Slama, J. B. H. 2015 Day-ahead load forecast using random forest and expert input selection. *Energy Convers. Manage.* **103**, 1040–1051.

Liu, X., Lu, X. & Chen, Y. 2011 The effects of temperature and nutrient ratios on microcystis blooms in Lake Taihu, China: an 11-year investigation. *Harmful Algae* **10** (3), 337–343.

Markus, M., Hejazi, M. I., Bajcsy, P., Giustolisi, O. & Savic, D. A. 2010 Prediction of weekly nitrate-N fluctuations in a small agricultural watershed in Illinois. *J. Hydroinform.* **12** (3), 251–261.

Muttil, N. & Chau, K. W. 2006 Neural network and genetic programming for modelling coastal algal blooms. *Int. J. Environ. Pollut.* **28** (3–4), 223–238.

Ranković, V., Radulović, J., Radojević, I., Ostojić, A. & Čomić, L. 2012 Prediction of dissolved oxygen in reservoirs using adaptive network-based fuzzy inference system. *J. Hydroinform.* **14** (1), 167–179.

Rousseeuw, K., Caillault, E. P., Lefebvre, A. & Hamad, D. 2015 Hybrid hidden Markov model for marine environment monitoring. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **8** (1), 204–213.

Schindler, D. W. 2006 Recent advances in the understanding and management of eutrophication. *Limnol. Oceanogr.* **51** (1), 356–363.

Smith, V. H., Joye, S. B. & Howarth, R. W. 2006 Eutrophication of freshwater and marine ecosystems. *Limnol. Oceanogr.* **51** (1part2), 351–355.

Touw, W. G., Bayjanov, J. R., Overmars, L., Backus, L., Boekhorst, l., Wels, M. & van Hijum, S. A. 2012 Data mining in the life sciences with random forest: a walk in the park or lost in the jungle? *Brief. Bioinform.* **14** (3), 315–326.

Tu, J. V. 1996 Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *J. Clin. Epidemiol.* 49 (11), 1225–1231.

Yajima, H. & Choi, J. 2013 Changes in phytoplankton biomass due to diversion of an inflow into the Urayama Reservoir. *Ecol. Eng.* **58**, 180–191.

Zhao, Y. & Zhang, Y. 2008 Comparison of decision tree methods for finding active objects. *Adv. Space Res.* **41** (12), 1955–1959.

First received 10 January 2017; accepted in revised form 6 September 2017. Available online 7 November 2017