Environ Monit Assess (2020) 192: 789
https://doi.org/10.1007/s10661-020-08731-2

Estimation of nitrogen and phosphorus concentrations from water quality surrogates using machine learning in the Tri An Reservoir, Vietnam

Nam-Thang Ha · Hao Quang Nguyen · Nguyen Cung Que Truong · Thi Luom Le · Van Nam Thai · Thanh Luu Pham

Received: 3 August 2020 / Accepted: 3 November 2020 / Published online: 26 November 2020

Abstract Surface water eutrophication due to excessive nutrients has become a major environmental problem around the world in the past few decades. Among these nutrients, nitrogen and phosphorus are two of the most important harmful cyanobacterial bloom (HCB) drivers. A reliable prediction of these parameters, therefore, is necessary for the management of rivers, lakes, and reservoirs. The aim of this study is to test the suitability of the powerful machine learning (ML) algorithm, random forest (RF), to provide information on water quality parameters for the Tri An Reservoir (TAR). Three species of nitrogen and phosphorus, including nitrite ($N\text{-}NO_2^-$), nitrate ($N\text{-}NO_3^-$), and phosphate ($P\text{-}PO_4^{3-}$), were empirically estimated using the field observation dataset (2009–2014) of six surrogates of total suspended solids (TSS), total dissolved solids (TDS), turbidity, electrical conductivity (EC), chemical oxygen demand (COD), and biochemical oxygen demand (BOD_5). Field data measurement showed that water quality in the TAR was eutrophic with an up-trend of $N\text{-}NO_3^-$ and $P\text{-}PO_4^{3-}$ during the study period. The RF regression model was reliable for $N\text{-}NO_2^-$, $N\text{-}NO_3^-$, and $P\text{-}PO_4^{3-}$ prediction with a high R^2 of 0.812–0.844 for the training phase (2009–2012) and 0.888–0.903 for the validation phase (2013–2014). The results of land use and land cover change (LUCC) revealed that deforestation and shifting

N.-T. Ha
Environmental Research Institute, School of Science, The University of Waikato, Hamilton 3216, New Zealand

N.-T. Ha
Faculty of Fisheries, The University of Agriculture and Forestry, Hue University, Hue 530000, Vietnam

H. Q. Nguyen
Graduate School of Systems and Information Engineering, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8572, Japan

N. C. Q. Truong
Center for Environmental Remote Sensing, Chiba University, Chiba 263-8522, Japan

T. Le
Dong Nai Technical Resources and Environment Center, Dong Khoi Street, Tan Hiep Ward, Bien Hoa City, Dong Nai Province 810000, Vietnam

V. N. Thai
Ho Chi Minh City University of Technology (HUTECH), 475A Dien Bien Phu Street, Binh Thanh District, Ho Chi Minh City 700000, Vietnam

T. L. Pham
Institute of Tropical Biology, Vietnam Academy of Science and Technology (VAST), 85 Tran Quoc Toan Street, District 3, Ho Chi Minh City 700000, Vietnam

T. L. Pham (✉)
Graduate University of Science and Technology, Vietnam Academy of Science and Technology, 18 Hoang Quoc Viet Street, Cau Giay district, Hanoi 100000, Vietnam
e-mail: thanhluupham@gmail.com

agriculture in the upper region of the basin were the major factors increasing nutrient loading in the TAR. Among the meteorological parameters, rainfall pattern was found to be one of the most influential factors in eutrophication, followed by average sunshine hour. Our results are expected to provide an advanced assessment tool for predicting nutrient loading and for giving an early warning of HCB in the TAR.

Keywords Tri An eutrophic reservoir · Water quality · Harmful cyanobacterial blooms · Random forest

Introduction

Water eutrophication has become a major environmental problem in developed and developing countries in the past few decades (Wang et al. 2019). In most developing countries, nutrient discharge to surface water will strongly increase in the following decades (van Puijenbroek et al. 2019), suggesting that water eutrophication will still remain a challenge. Harmful cyanobacterial blooms (HCB) are one of the most serious symptoms of water eutrophication that cause severe health issues and degraded water quality (Heisler et al. 2008; Reichwaldt and Ghadouani 2012; Dubey and Dutta 2020). These blooms can release toxins that enter the aquatic food chain and eventually put our health at risk (Morris 1999; Grattan et al. 2016). HCB have recently increased worldwide, and they represent a serious threat to drinking and recreational water resources (especially for countries that depend on surface water as a source of drinking water supply) and to the ecological and economic sustainability of ecosystems (Lu et al. 2019).

The major nutrient drivers of surface water eutrophication are nitrogen (N) and phosphorus (P) (Lewis et al. 2011). Traditionally, watershed nutrient management efforts to control eutrophication and HCB have focused on reducing P inputs (Schindler et al. 2008), as P limitation is common, and some HCB can fix atmospheric N_2 to satisfy their N requirements (Berrendero et al. 2016). However, N loading has increased dramatically in many watersheds because of the high-frequency application of fertilizer in agriculture, especially in developing countries, and the untreated wastewater discharge from households and industry. In many aquatic ecosystems, N loads increase faster than P loads. Therefore, N and P input constraints are likely needed for the long-term control of HCB in such systems (Lewis et al. 2011). In addition, a direct measurement of the forms of N and P is not practical, leading to the demanding of the prediction using other measurable water quality parameters as the surrogates.

Located in Dong Nai Province, the Tri An Reservoir (TAR) is a well-known man-made freshwater reservoir, built in 1984–1986, and provides a huge source of water for agriculture, industry, and public usage. However, in recent decades, the reservoir has fallen into the eutrophic category and suffered annual HCB as a consequence of nutrient enrichment (including N and P) from surrounding catchment areas (Nguyen et al. 2020; Pham et al. 2020a, b). Many previous works have focused on measuring water quality, analyzing the mechanism of HCB, and identifying harmful algae species (Truong et al. 2018; Trung et al. 2018; Pham et al. 2020a, b), but they have left a gap in the prediction of main nutrient loading in the TAR and other inland waters in Vietnam. In addition, the degradation of water quality and its relationship with deforestation and agricultural land conversion have not been studied to the same extent.

A multilinear solver has been used to predict the variation of biochemistry (i.e., nutrient species, chlorophyll-a (Chl-a) concentration) from a variety of data types based on linear or linear mixed models (Jones et al. 2001; Jones et al. 2004; Hollister et al. 2016). Even though these methods have demonstrated to be reliable, they have limitations such as independence, distribution assumption, and outlier sensitivity. In addition, the interaction between the group of nutrient species and other water quality parameters is complex in eutrophic inland lake and a non-linear relationship might exist (Qian et al. 2005). Therefore, the machine learning (ML) approach is exploited to do a prediction of nutrient concentration from different surrogates.

Different ML algorithms (i.e., random forest (RF), support vector machine (SVM), artificial neural network (ANN), Gaussian process, extremely randomized tree, adaptive boosting, gradient boosting, k-nearest-neighbors, extreme learning machine (ELM), M5 model tree) have been used to retrieve Chl-a (Park et al. 2015; Lou et al. 2016; Bui et al. 2017; Blix and Eltoft 2018; Keller et al. 2018; Li et al. 2018; Wang et al. 2018; Yi et al. 2018), turbidity, colored dissolved organic matter, diatom, dissolved organic carbon, and total suspended solids (TSS) (Keller et al. 2018; Ross et al. 2019). On the contrary, we observed only a very limited number of research papers applying ML algorithms for total phosphorous (TP) using SVM, ANN, radial basis function

neural network (RBFN), and an adaptive neuro-fuzzy inference system (ANFIS) approach (Chen and Liu 2015; García et al. 2019); nitrate ($N\text{-}NO_3^-$) using ANN (Jung et al. 2020) and TP; and dissolved phosphorous (DP) using ANN (Kim et al. 2012), TP, $N\text{-}NO_3^-$, and nitrogen and phosphorous forms using RF (Castrillo and García 2020; Shen et al. 2020) models. Of the published papers using the RF model, however, we found no or low coefficient of determination (R^2 reached less than 0.66 on average) reported for river/stream ecosystems with the concentration of Chl-a ranging from 17.4 to 92.9 μg/L. Other documents either conducted an estimation with/without a k-folds cross validation (CV) for a limited form of nutrient factors, applied for the river ecosystem, or mainly used the ANN model which is usually hard to extract the feature importance and control the model's hidden layers (Tu 1996; Mas and Flores 2008; Oyebode and Stretch 2019). The recent status leaves a gap for the identification of a reliable model for nutrient estimation in a HCB-suffering reservoir with a very high concentration of Chl-a (6–3400 μg/L) (Nguyen et al. 2020) like the TAR.

Herein, we consider the RF model (Breiman 2001), a powerful ML model with acknowledged merits in the literature (Hollister et al. 2016; Yajima and Derot 2018) and has the potential to be applied in water quality management (Belgiu and Drăguț 2016; Sihag et al. 2019). This model stands out from other tree structure–based models in that it maximizes the random selection of input data for the training and testing with the bootstrap sampling and bagging ensemble decision trees which could potentially improve the model performance (Breiman 2001; Yajima and Derot 2018). The RF algorithm does not require a normal distribution of the input data, is less sensitive to the outliers and less impacted by the data noise, and reduces the over-fitting in model prediction compared to linear and other ML models (Fawagreh et al. 2014; Parmar et al. 2019; Tyralis et al. 2019). The last advantage of the RF is the implementation of the model. It is easy to tune and control the RF hyper-parameters in a parallel computation using the popular Python library scikit-learn (Pedregosa et al. 2011). Therefore, the model is reliable and stable to do an experiment on accurate estimation of nutrient species from various surrogates.

In the present study, we evaluated the RF performance in retrieving $N\text{-}NO_2^-$, $N\text{-}NO_3^-$, and $P\text{-}PO_4^{3-}$ based on field data measurement from 2009 to 2014 in the TAR. Note that the physical relationships were documented between nutrient species (in the form of nitrogen and phosphorous compositions) and other water quality parameters. Six surrogates including TSS, TDS, turbidity, EC, COD, and BOD_5 were selected as inputs in our model because the nutrient species can be transferred with the particle matters (Marttila and Kløve 2009) in the particulate form of P and dissolved form of N (Wang et al. 2015), and showed a direct link to the suspended solid in the water. In addition, the relationship between nutrient forms and turbidity or TDS was well documented in the literature (Marttila and Kløve 2012; Xi et al. 2012; Wang et al. 2015; Paudel et al. 2019). Besides, nutrient species can exist in the form of ions in the water column which regulate the EC. Corwin et al. (2006) reported a strong correlation between the N form ($N\text{-}NO_3^-$) and EC (correlation coefficient range of 0.34–0.46) while Kim et al. (2007) determined a variation in correlation between P form ($P\text{-}PO_4^{3-}$) and EC in a nutrient removal system. On the other hand, the relationship between the parameters COD and BOD_5 and the nutrient species in various aquatic ecosystems has been reported (Davies-Colley et al. 1995; Carlsson et al. 1996; Zhang et al. 2011). COD identifies the oxygen demand of chemical reactions while BOD_5 indicates the oxygen demand of microorganisms, and both are essential for the transformation of nutrient forms in the water environment (Davies-Colley et al. 1995; Zhang et al. 2011). The response of nutrient parameters to meteorological factors and land use and land cover change (LUCC) was also examined from the correlation analysis as well. Our results contribute advanced approaches to understanding the current nutrient loading and to adopting a higher accuracy prediction of nutrient factors, supporting the foundation for mitigating HCB in inland lakes or reservoirs.

Materials and methods

Study area

The Dong Nai River Basin (DNRB) (Fig. 1), the second largest catchment in Vietnam, is located in the country's main economic development region and accounts for 23% of Vietnam's gross domestic product (Asian Development Bank 2009). The DNRB originates from the Lang Biang Plateau at an elevation of 2000 m. It is approximately 437 km in length and has a total drainage area of 37,330 km^2. It flows through the TAR, which is

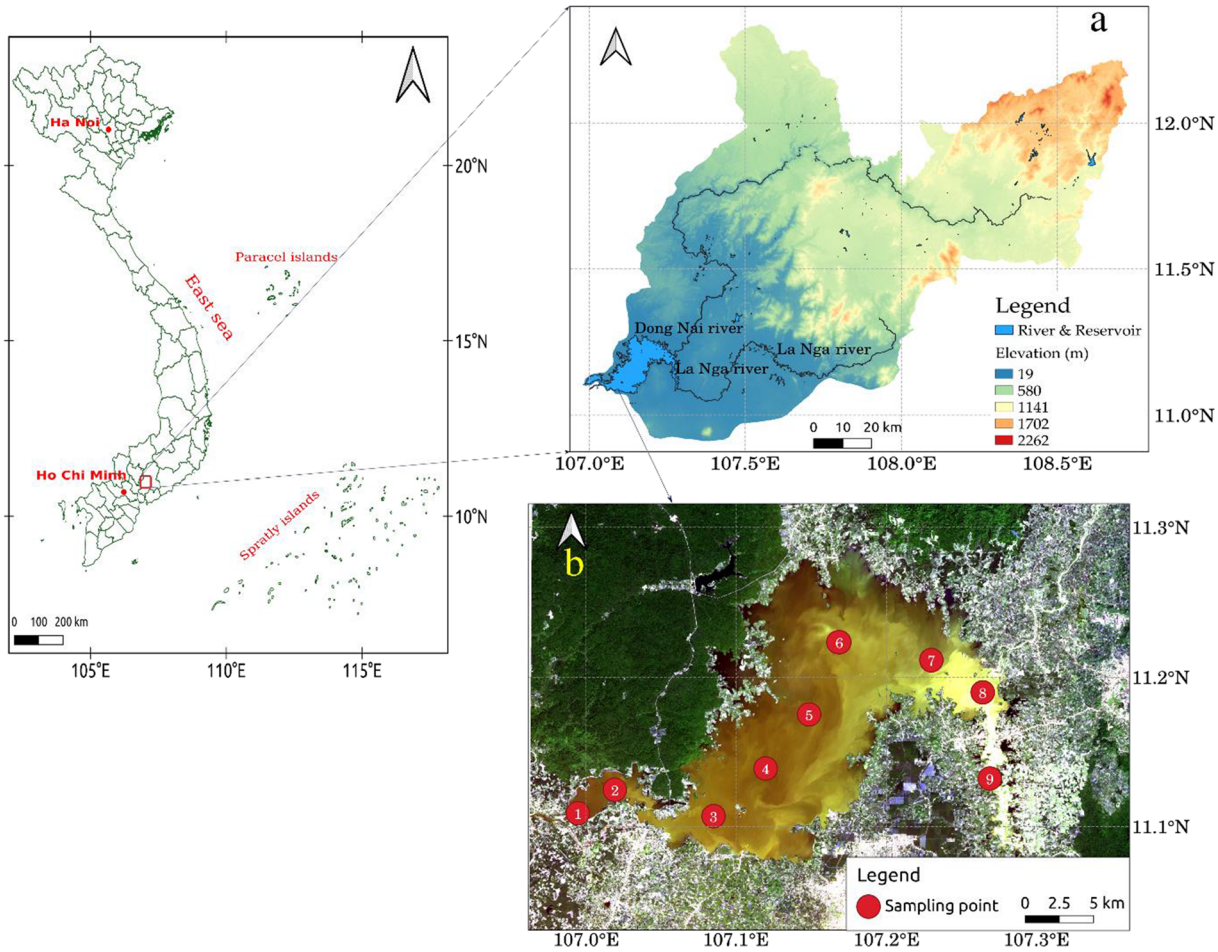

Fig. 1 Study site. **a** Tri An Reservoir watershed with two main rivers. **b** Tri An Reservoir with sampling locations

a major power dam in Southern Vietnam, and merges with the Be River and Saigon River (JICA 1996). The upstream river basin (upper part of the TAR), including the six provinces of Binh Thuan, Binh Phuoc, Dak Nong, Dong Nai, and Lam Dong, has been highly developed and primarily consists of agricultural and forest lands.

The TAR provides a freshwater resource for various sectors of agriculture, irrigation, fisheries, and hydropower operations. The reservoir's surface area is estimated at 320 km^2, and it has a maximum depth of 27 m and a water volume of 2.7 billion m^3. The topography and geolocation of the TAR make it an ideal place for loading nutrients from surrounding anthropogenic activities, leading to a high occurrence of HCB in recent years (Nguyen et al. 2020; Pham et al. 2020a, b). With the important role and the current status (degrading water quality) of the TAR, it is suggested to use novel approaches to improve the nutrient management as well as water quality that provide useful information for HCB mitigation in the reservoir.

Measurement and analysis of water quality parameters

Water quality data were collected bimonthly from nine monitoring stations in 2009–2014 (Fig. 1). Electrical conductivity (EC) and total dissolved solids (TDS) were measured in situ with a WTW multidetector. Turbidity was measured with a turbidimeter (Hach, 2100P, CO, USA). The chemical parameters were analyzed colorimetrically in triplicate with a spectrophotometer (Hach DR/2010) using the following American Public Health Association (Bridgewater et al. 2017) methods: nitrite 4500NO_2^-, nitrate 4500NO_3^- (B), and phosphate 4500PO_4^{3-} (B). Chemical oxygen demand (COD) and biochemical oxygen demand (BOD_5) were determined

by the consumption of oxygen and the difference in dissolved oxygen (DO) concentrations in the samples after 5 days, respectively, according to Bridgewater et al. (2017). To measure the TSS, approximately 300–400 mL of water samples was filtered into a pre-weighed glass fiber filter with a 0.45-μm pore size (Whatman, England) and dried completely at 95–105 °C. The TSS concentration was estimated gravimetrically.

RF performing for $N\text{-}NO_2^-$, $N\text{-}NO_3^-$, and $P\text{-}PO_4^{3-}$ retrieval

Water quality dataset

Due to the requirement of the long-term time series observation, we considered the six water quality parameters, which have relationship with the nutrient species (Table 1), collected bimonthly from 2009 to 2014, involving TSS, TDS, COD, BOD_5, EC, and turbidity as the input data for $N\text{-}NO_2^-$, $N\text{-}NO_3^-$, and $P\text{-}PO_4^{3-}$ prediction in the TAR.

We used a 4-year dataset (2009–2012) for training and a 2-year independent dataset (2013–2014) for validating the RF model. A total of 717 and 330 observations were used for the training and validating with the performance of RF in the periods of 2009–2012 and 2013–2014, respectively.

Introduction to random forest

Random forest (RF) (Breiman 2001) is a well-known algorithm, designed for both classification and regression tasks. RF is applied for a variety of problems, ranging from environment to water quality management which presented a consistent and reliable performance (Belgiu and Drăguț 2016; Nguyen et al. 2020; Zhang et al. 2019). Based on the classification and regression tree (CART) algorithm, RF uses 2/3 of the sample for the training and 1/3 of the sample (out-of-bag) for the testing, and then a majority voting is used to select the most desired decision from a large base of decision trees. The random selection of variable and bootstrapping aggregation (bagging) of the RF model support a reduction of data noise and variance during the training process. The most important parameters of RF algorithm include the number of decision trees, the maximum depth, the minimum sample of the leaf, and the minimum sample to split the branch of the decision trees. There are no default parameters of RF algorithm for the desired problems; instead, the parameter tuning is required to find the best combination among the parameters (Mohapatra et al. 2020).

Table 1 Pearson's correlation coefficients of $N\text{-}NO_2^-$, $N\text{-}NO_3^-$, $P\text{-}PO_4^{3-}$, and surrogates

	Turb.	EC	TSS	COD	BOD_5	TDS
$N\text{-}NO_2^-$	0.41	0.45	0.39	0.23	0.24	0.23
$N\text{-}NO_3^-$	0.50	0.17	0.37	0.16	0.22	0.16
$P\text{-}PO_4^{3-}$	0.65	0.28	0.67	0.04	0.22	0.38

Tuning the hyper-parameters for the RF model

We used a fivefold CV grid search in scikit-learn (Pedregosa et al. 2011) to tune the hyper-parameters of the RF model (Table 2). These hyper-parameters were maintained during the training and validation phases in 2009–2012 and 2013–2014.

Model performance and comparison in the prediction of $N\text{-}NO_2^-$, $N\text{-}NO_3^-$, and $P\text{-}PO_4^{3-}$

Multivariate linear and RF regression model performance

The multivariate linear regression (MLR) and the RF model were employed in the Python™ environment using the scikit-learn library (Pedregosa et al. 2011). Regarding the RF model, the ten-fold CV using the Shuffle split technique for data sampling (70% for the training and 30% for the testing) during the CV was applied to evaluate the performance of the RF model in the training (2009–2012) and validation (2013–2014) phases. Then, the best performance of the RF model for $N\text{-}NO_2^-$, $N\text{-}NO_3^-$, and $P\text{-}PO_4^{3-}$ prediction at each phase was extracted using a second fitting to the same dataset with 70% for the training and 30% for the testing. In addition, the function of feature importance in the RF model was used to measure the contribution of various water quality parameters to the prediction of selected parameters.

MLR and RF model comparison

We compare the skills of the MLR and RF models in $N\text{-}NO_2^-$, $N\text{-}NO_3^-$, and $P\text{-}PO_4^{3-}$ prediction using the standard metrics of R^2 and RMSE for both the training phase 2009–2012 and validation phase 2013–2014.

Table 2 Hyper-parameters of the RF model during the training and validation phases

For $N\text{-}NO_2^-$ retrieval		For $N\text{-}NO_3^-$ retrieval		For $P\text{-}PO_4^{3-}$ retrieval	
Hyper-parameters	Values	Hyper-parameters	Values	Hyper-parameters	Values
Max_depth	15	Max_depth	15	Max_depth	15
Bootstrap	True	Bootstrap	True	Bootstrap	True
N_estimator	100	N_estimator	100	N_estimator	100
Max_features	6	Max_features	6	Max_features	2
Min_sample_leaf	1	Min_sample_leaf	1	Min_sample_leaf	1
Min_sample_split	2	Min_sample_split	3	Min_sample_split	2

Auxiliary data collection

The meteorological parameters of average monthly rainfall, average wind speed, and average monthly sunshine hour collected from the Southern Regional Hydro-Meteorological Center, Vietnam (unpublished data), were used to elucidate the relationship with nutrient variables ($N\text{-}NO_2^-$, $N\text{-}NO_3^-$, and $P\text{-}PO_4^{3-}$).

The LUCC maps in 1973, 1994, and 2014 (modified from Truong et al. 2018) were used to evaluate the effect of shifting agriculture on water quality in the TAR.

Correlation analysis

The correlation analyses between water quality and the meteorological parameters (collected from nine sampling sites, as shown in Fig. 1) and between the water quality parameters and the amount of applied fertilizer (collected from stations 7 and 9 as shown in Fig. 1) were conducted to determine the effects of the meteorological parameters and the shifting agricultural activity on water quality in the TAR.

Evaluation criteria

We used the coefficient of determination (R^2) and root mean squared error (RMSE) to evaluate the performance of the MLR and RF models in the $N\text{-}NO_2^-$, $N\text{-}NO_3^-$, and $P\text{-}PO_4^{3-}$ retrieval. In addition, the Durbin-Watson test and autocorrelation function (ACF) plot were involved to check the autocorrelation from the time series data of $N\text{-}NO_2^-$, $N\text{-}NO_3^-$, and $P\text{-}PO_4^{3-}$. The Durbin-Watson test and ACF plot were employed in Python environment using statsmodels library (Seabold and Perktold 2010). A Durbin-Watson statistic d (Durbin-Watson Test 2008) is expected approximately 2 while the lag points are inside the 95% confidence interval. All data were presented as the mean ± standard deviation (SD). P values less than 0.05 were considered statistically significant.

The formulas of R^2, RMSE, and Durbin-Watson test are expressed as Eqs. (1), (2), and (3) respectively.

$$R^2 = 1 - \sum \frac{\left(x_i^{\text{measured}} - x_i^{\text{estimated}}\right)^2}{\left(x_i^{\text{measured}} - x_{\text{mean}}^{\text{measured}}\right)^2} \quad (1)$$

$$\text{RMSE} = \sqrt{\sum_{i=1}^{N} \frac{\left(x_i^{\text{measured}} - x_i^{\text{estimated}}\right)^2}{N}} \quad (2)$$

$$d = \sum_{t=2}^{T} \left((e_t - e_{t-1})^2 \right) / \sum_{t=1}^{T} e_t^2 \quad (3)$$

where x^{measured} is the measured values, $x^{\text{estimated}}$ is the values estimated from the model, T is the number of observations, and e_t is the residual of the model.

Results

Water quality variation in the TAR in the 2009–2014 period

The annual mean and standard deviation values of the water quality variables in the 2009–2014 period are shown in Fig. 2. The concentrations of TSS and TDS did not vary much over the sampling period at a range of 5–14 mg/L and 34–41.2 mg/L, respectively. The concentrations of EC and $N\text{-}NO_2^-$ did not show a trend over the sampling period. Conversely, the concentrations of

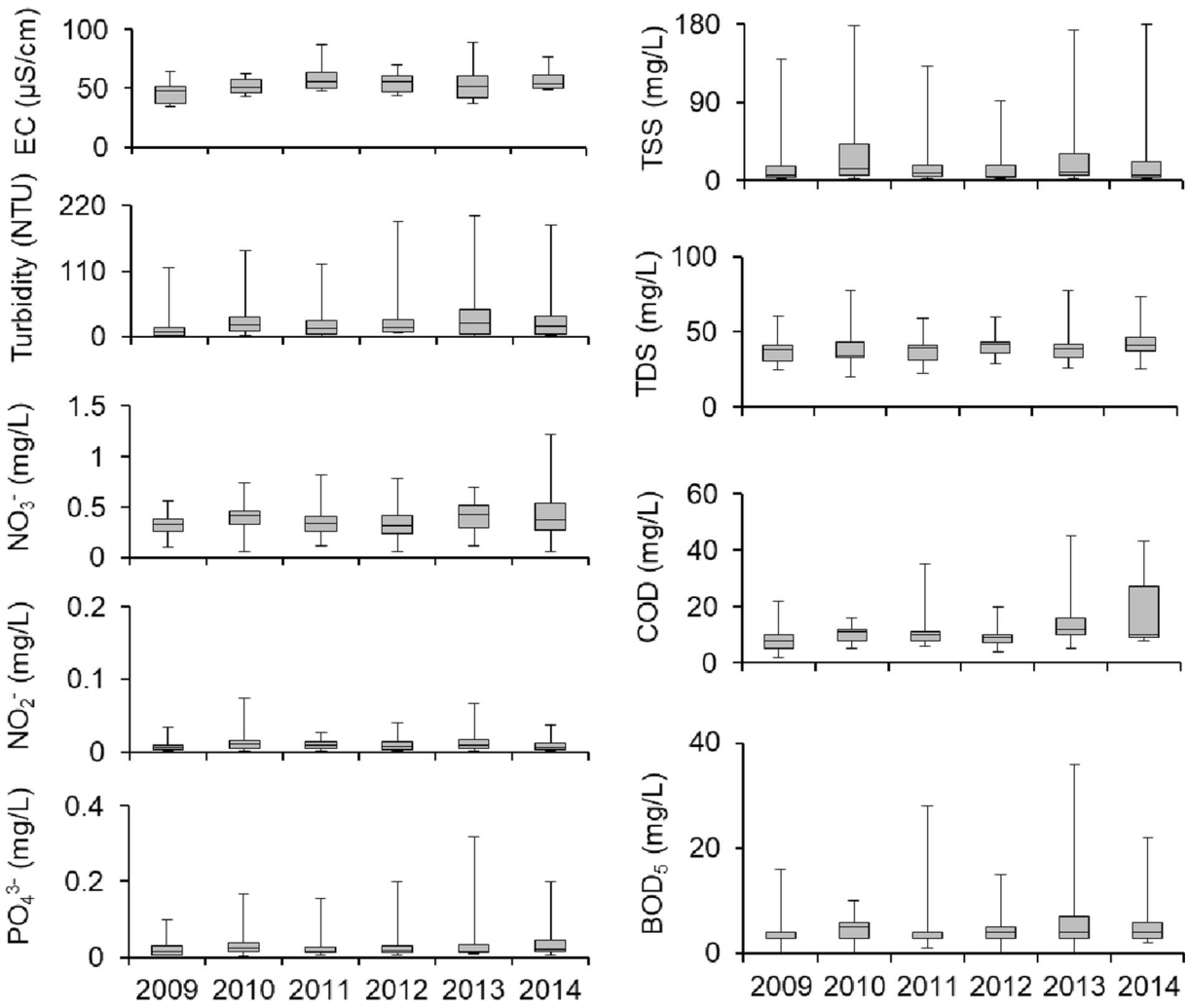

Fig. 2 Annual water quality variables during the 2009–2014 period in the TAR

$N\text{-}NO_3^-$, $P\text{-}PO_4^{3-}$, and turbidity increased during the study period. The mean $N\text{-}NO_3^-$ concentration went from 0.3 mg/L (range of 0.1–0.56 mg/L) in 2009 to 0.37 (range 0.06–1.22 mg/L) in 2014, the mean $P\text{-}PO_4^{3-}$ concentration went from 0.01 mg/L (range of 0.005–0.099 mg/L) in 2009 to 0.02 mg/L (range of 0.005–0.2 mg/L) in 2014, and the mean turbidity concentration went from 8.5 NTU (range of 1.0–122 NTU) in 2009 to 13 NTU (range of 2.0–166.0 NTU) in 2014. The concentrations of BOD_5 and COD also increased during the study period. The mean BOD_5 concentration increased from 3.0 mg/L (range of 2.0–6.0 mg/L) in 2009 to 4 mg/L (range of 2.0–32.0 mg/L) in 2014, and the mean COD concentration increased from 8.0 mg/L (range of 2.0–22.0 mg/L) in 2009 to 10 mg/L (range of 8.0–43.0 mg/L) in 2014.

Table 3 Ten-fold CV for $N\text{-}NO_2^-$, $N\text{-}NO_3^-$, and $P\text{-}PO_4^{3-}$ retrieval in the training phase (2009–2012)

	$N\text{-}NO_2^-$			$N\text{-}NO_3^-$			$P\text{-}PO_4^{3-}$		
	R^2	RMSE	Mean	R^2	RMSE	Mean	R^2	RMSE	Mean
RF model	0.737	0.0039	0.0098	0.806	0.084	0.308	0.744	0.01	0.022

Table 4 RF and MLR model performance for $N\text{-}NO_2^-$, $N\text{-}NO_3^-$, and $P\text{-}PO_4^{3-}$ retrieval in the training phase (2009–2012)

	$N\text{-}NO_2^-$		$N\text{-}NO_3^-$		$P\text{-}PO_4^{3-}$	
	R^2	RMSE	R^2	RMSE	R^2	RMSE
RF	0.812	0.003	0.844	0.076	0.817	0.008
MLR	0.457	0.005	0.411	0.148	0.629	0.012

$N\text{-}NO_2^-$, $N\text{-}NO_3^-$, and $P\text{-}PO_4^{3-}$ retrieval

Performance of the RF and the MLR model during the training phase

The RF model was trained for $N\text{-}NO_2^-$, $N\text{-}NO_3^-$, and $P\text{-}PO_4^{3-}$ prediction using collected data from 2009 to 2012. The ten-fold CV presented a good R^2 value for all the retrieved parameters (Table 3). RF performed well in $N\text{-}NO_3^-$ retrieval with an R^2 of 0.806, whereas

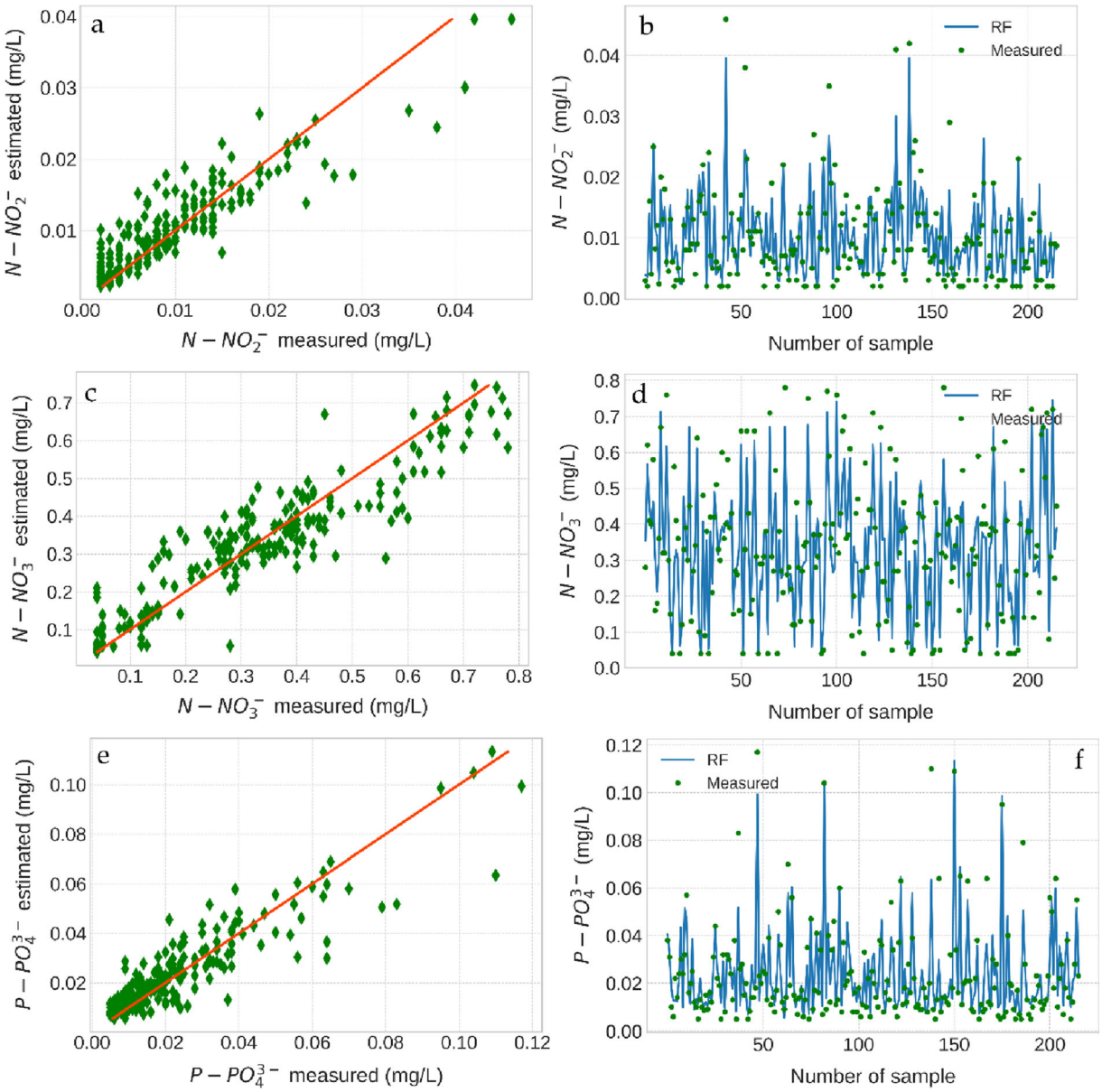

Fig. 3 Best performance from RF models of $N\text{-}NO_2^-$ (**a**), $N\text{-}NO_3^-$ (**c**), and $P\text{-}PO_4^{3-}$ (**e**) retrieval from water parameters in 2009–2012, and observed and predicted (**b**, **d**, **f**) values

Table 5 Durbin-Watson test for $N\text{-}NO_2^-$, $N\text{-}NO_3^-$, and $P\text{-}PO_4^{3-}$ parameters in the training phase (2009–2012)

	$N\text{-}NO_2^-$	$N\text{-}NO_3^-$	$P\text{-}PO_4^{3-}$
d statistic	1.76	2.09	2.06

lower values were observed for $N\text{-}NO_2^-$ and $P\text{-}PO_4^{3-}$ (R^2 was 0.737 and 0.744, respectively). The RMSEs were significantly lower than the mean values of $N\text{-}NO_2^-$ (0.0098), $N\text{-}NO_3^-$ (0.308), and $P\text{-}PO_4^{3-}$ (0.022) (Table 3).

To better visualize the RF's performance in the training phase, we extracted the best performance of the RF model in the $N\text{-}NO_2^-$, $N\text{-}NO_3^-$, and $P\text{-}PO_4^{3-}$ estimation, as shown in Fig. 3. The scatter plots presented a good prediction of the RF model for selected parameters, with R^2 values ranging from 0.812 to 0.844 (Table 4). In addition, RF performed well in this training phase when the predicted values were close to the observed values. The Durbin-Watson statistic values range from 1.76 to 2.09 (Table 5) while almost the lag points are inside the 95% confidence interval (Fig. 4a–c), indicating that the retrieval models from the time series data of $N\text{-}NO_2^-$, $N\text{-}NO_3^-$, and $P\text{-}PO_4^{3-}$ are not affected by the autocorrelation during the training phase 2009–2012. The data analysis also determines a significant outperforming of the RF to the MLR model (Table 4) with higher values of R^2 and an improvement of RMSE for $N\text{-}NO_2^-$ (40%), $N\text{-}NO_3^-$ (55.8%), and $P\text{-}PO_4^{3-}$ (33.3%).

Performance of the RF and the MLR model during the validation phase

To validate the predictable capability of the RF model in the $N\text{-}NO_2^-$, $N\text{-}NO_3^-$, and $P\text{-}PO_4^{3-}$ retrieval, we tested the performance of RF using an independent dataset in 2013–2014. The results indicated a promising performance of RF with a high R^2 value of the ten-fold CV for $N\text{-}NO_2^-$ (0.823), $N\text{-}NO_3^-$ (0.814), and $P\text{-}PO_4^{3-}$ (0.830). The accuracy of the validation results was high when comparing the RMSEs with the mean values of the selected parameters. The RMSEs presented significantly lower values than the mean values of $N\text{-}NO_2^-$ (0.0128), $N\text{-}NO_3^-$ (0.325), and $P\text{-}PO_4^{3-}$ (0.023) (Table 6).

The best performances of RF in the $N\text{-}NO_2^-$, $N\text{-}NO_3^-$, and $P\text{-}PO_4^{3-}$ estimation were also extracted to provide an overview of the model's prediction in the validation phase, as shown in Fig. 5. The validation results showed a very good prediction of RF in $N\text{-}NO_2^-$, $N\text{-}NO_3^-$, and $P\text{-}PO_4^{3-}$ in the TAR, with a high R^2 of 0.890–0.903 (Table 7). The predicted values closely reached or overlapped with the observed values. Similar to the training phase (2009–2012), we observed no autocorrelation from the retrieval of the time series of $N\text{-}NO_2^-$, $N\text{-}NO_3^-$, and $P\text{-}PO_4^{3-}$ with the Durbin-Watson statistic values ranging from 1.87 to 1.94 (Table 8) and the 95% confidence interval containing the lag points (Fig. 4d–f). For the validation dataset, the MLR model performed an acceptable R^2 prediction (Table 7) for $N\text{-}NO_2^-$ (0.638) and $P\text{-}PO_4^{3-}$ (0.746), and however a very low value for $N\text{-}NO_3^-$ (0.219). Despite a better performance of the MLR model for $N\text{-}NO_2^-$ and $P\text{-}PO_4^{3-}$ retrieval using dataset 2013–2014, the skill metrics determined an outperforming of RF model in this case (Table 7) with an improvement of RMSE for $N\text{-}NO_2^-$ (50%), $N\text{-}NO_3^-$ (62.1%), and $P\text{-}PO_4^{3-}$ (44.4%).

Feature importance in the $N\text{-}NO_2^-$, $N\text{-}NO_3^-$, and $P\text{-}PO_4^{3-}$ retrieval

The function of feature importance in the RF model provided an overview of the most contributing factors to the successful prediction of $N\text{-}NO_2^-$, $N\text{-}NO_3^-$, and $P\text{-}PO_4^{3-}$ in both the training and validation phases (Fig. 6). Given the importance value of over 0.1, EC, TDS, and turbidity were the main contributors for $N\text{-}NO_2^-$ and $N\text{-}NO_3^-$, and EC, TDS, TSS, and turbidity for the $P\text{-}PO_4^{3-}$ estimation in the training phase.

For the validation phase, the most important parameters were EC, TDS, TSS, and turbidity for $N\text{-}NO_2^-$ and $P\text{-}PO_4^{3-}$ estimation while EC, COD, TDS, and turbidity are the main contributors for $N\text{-}NO_3^-$ estimation (Fig. 6).

Table 6 Ten-fold CV for $N\text{-}NO_2^-$, $N\text{-}NO_3^-$, and $P\text{-}PO_4^{3-}$ retrieval in the validation phase (2013–2014)

	$N\text{-}NO_2^-$			$N\text{-}NO_3^-$			$P\text{-}PO_4^{3-}$		
	R^2	RMSE	Mean	R^2	RMSE	Mean	R^2	RMSE	Mean
RF model	0.823	0.0048	0.0128	0.814	0.074	0.325	0.830	0.010	0.023

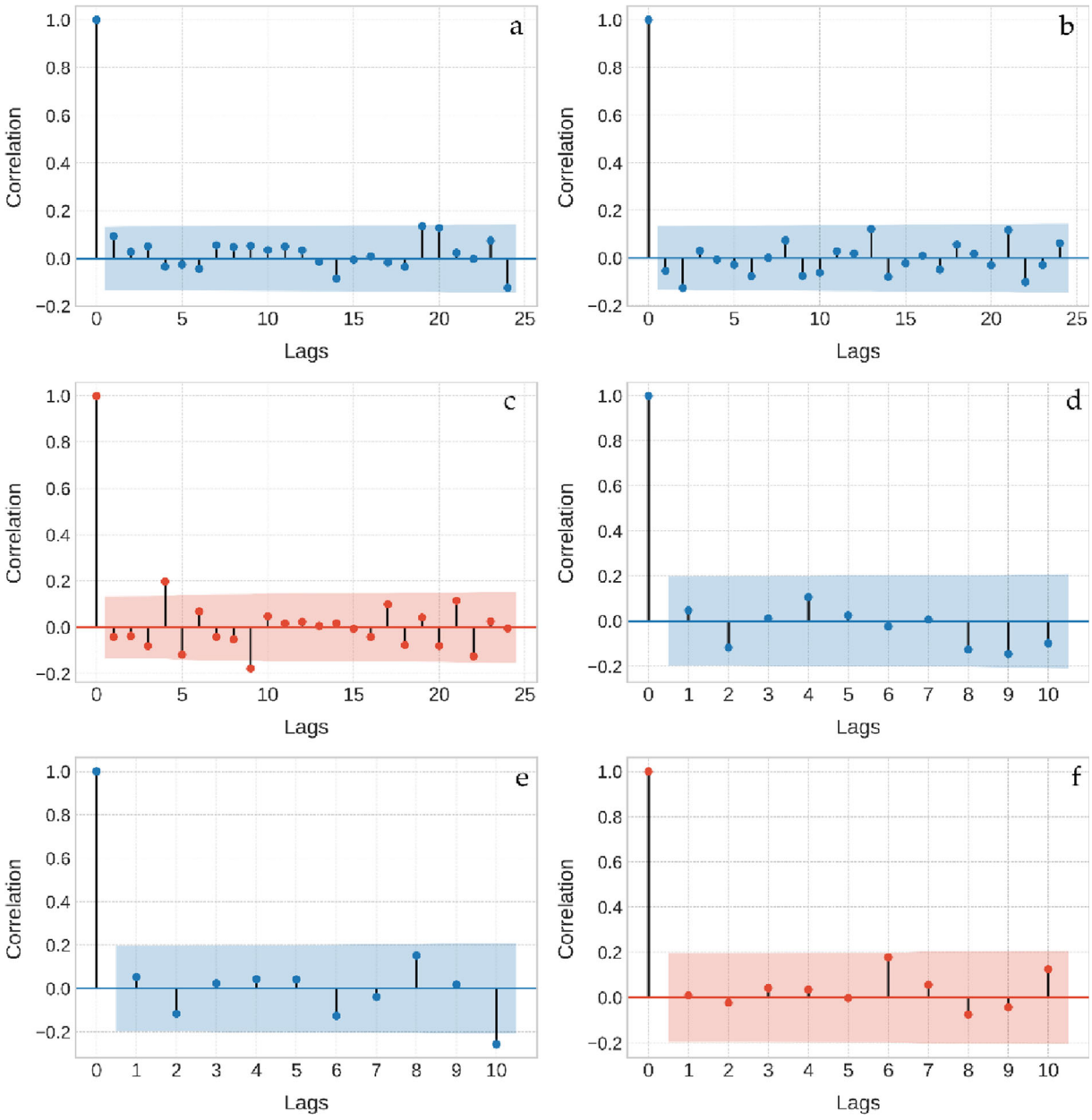

Fig. 4 The autocorrelation plots for $N\text{-}NO_2^-$ (**a**), $N\text{-}NO_3^-$ (**b**), and $P\text{-}PO_4^{3-}$ (**c**) retrieval during the training phase 2009–2012, and the validation phase 2013–2014 (**d**–**f**)

Governing factors controlling eutrophication in the TAR

Land use/cover change effects

The LUCC significantly changed during the monitoring period (1973–2014), with a high proportion of LUCC converted to agricultural land (Figs. 7 and 8) in Lam Dong Province. Forest coverage was decreased to 45% (2014) from 73% (1994), whereas agricultural land area increased from 3458 km^2 (24% coverage in 1994) to 6565 km^2 (51% coverage in 2014) (Truong et al. 2018). Among planting crops, coffee-planted areas rapidly increased during the 1990–2014 period (Fig. 8), raising a

Table 7 RF and MLR model performance for $N\text{-}NO_2^-$, $N\text{-}NO_3^-$, and $P\text{-}PO_4^{3-}$ retrieval in the validation phase (2013–2014)

	$N\text{-}NO_2^-$		$N\text{-}NO_3^-$		$P\text{-}PO_4^{3-}$	
	R^2	RMSE	R^2	RMSE	R^2	RMSE
RF	0.890	0.004	0.888	0.059	0.903	0.005
MLR	0.638	0.008	0.219	0.156	0.746	0.009

significant concern about deforestation (Meyfroidt et al. 2013) and nutrient loading in the TAR.

Lam Dong Province covers more than 1.5 million ha of coffee-planted areas with a total annual output of about 400,000 tons (Lam Dong Department of Statistic 2018). It is the country's second largest coffee-growing area. The coffee yield was strongly affected by the organic matter content: total nitrogen (N), potassium (K), and phosphorus (P) (Tiemann et al. 2018). However, farmers generally apply unbalanced proportions of chemical fertilizers (higher than recommended), posing threats to the sustainability of the environment and increasing the contamination of offsite water resources (Byrareddy et al. 2019; PVFCCo 2016; Tiemann et al. 2018).

Nutrients arriving from the reservoir catchment area, through runoff during rainfall events, can stimulate phytoplankton growth, and combined with a seasonal increase in water temperature, it would facilitate *Dolichospermum lemmermannii* proliferation (Bresciani et al. 2018). Owing to the high demand for fertilizers (Tiemann et al. 2018), the increase in coffee areas was hypothesized as the main factor for eutrophication in the lower region of the TAR basin. We tested this hypothesis by analyzing the correlation between the applied amount of fertilizer (N-P-K) and the monitored concentration of nutrient factors ($N\text{-}NO_2^-$, $N\text{-}NO_3^-$, $P\text{-}PO_4^{3-}$) using a ratio of 480-240-350 kg/ha/year of N-P-K (Table S1, S2, Supplementary material) in the rainy season (PVFCCo 2016). The source points were located at the outlets of Dong Nai and La Nga Rivers (stations 7 and 9, Fig. 1).

Our results determined a positive correlation between the applied fertilizer and the concentration of nutrient factors in the TAR. Accordingly, $N\text{-}NO_2^-$, $N\text{-}NO_3^-$, and $P\text{-}PO_4^{3-}$ presented a clear relationship with the applied N and P fertilizers at the outlets of Dong Nai and La Nga Rivers (R^2 = 0.40–0.46) (Fig. 9).

Table 8 Durbin-Watson test for $N\text{-}NO_2^-$, $N\text{-}NO_3^-$, and $P\text{-}PO_4^{3-}$ parameters in the validation phase (2013–2014)

	$N\text{-}NO_2^-$	$N\text{-}NO_3^-$	$P\text{-}PO_4^{3-}$
d statistic	1.88	1.87	1.94

Effect of meteorological factors

The effects of meteorological factors on the eutrophication phenomenon in the TAR were explored using the precipitation character, average sunshine hour, wind speed, and their relationships with nutrient concentrations on a monthly scale (Fig. 10). The results indicated a strong effect of the rainfall on eutrophication in the TAR, followed by the average sunshine hour. High correlations were found only between $P\text{-}PO_4^{3-}$ and meteorological parameters (precipitation and average sunshine hour). Conversely, low correlations were observed between wind speed and the nutrient parameters.

Discussion

Application of the RF model in nutrient retrieval

In this study, the performance of the RF model in the $N\text{-}NO_2^-$, $N\text{-}NO_3^-$, and $P\text{-}PO_4^{3-}$ retrieval was tested using measured data for the training (2009–2012) and validation (2013–2014) phases. The RF was stable and reliable during the ten-fold CV and outperformed the MLR model. This result indicates that RF was well trained using 4 years of observation data with a large variation in water quality parameters. The model was capable of catching the data trend, managing the variation of the predicted parameters, and therefore presenting a reliable estimation of the $N\text{-}NO_2^-$, $N\text{-}NO_3^-$, and $P\text{-}PO_4^{3-}$ concentrations when validating with independent measured data in 2013–2014.

For all retrieval parameters, the hyper-parameters were almost similar to low max_depth (15), number of trees (100), and min_sample_leaf (1) and different from max_feature (range of 2–6) and min_sample_split (range of 2–3). The low values of the hyper-parameters supported a faster run and a simpler tree structure, thus reducing overestimation when retrieving the selected parameters.

Recently, ML has been applied to nutrient retrieval in inland waters. Good performances were obtained using long-term data observation with the SVM (R^2 = 0.9, water quality dataset 2006–2014) (García et al. 2019),

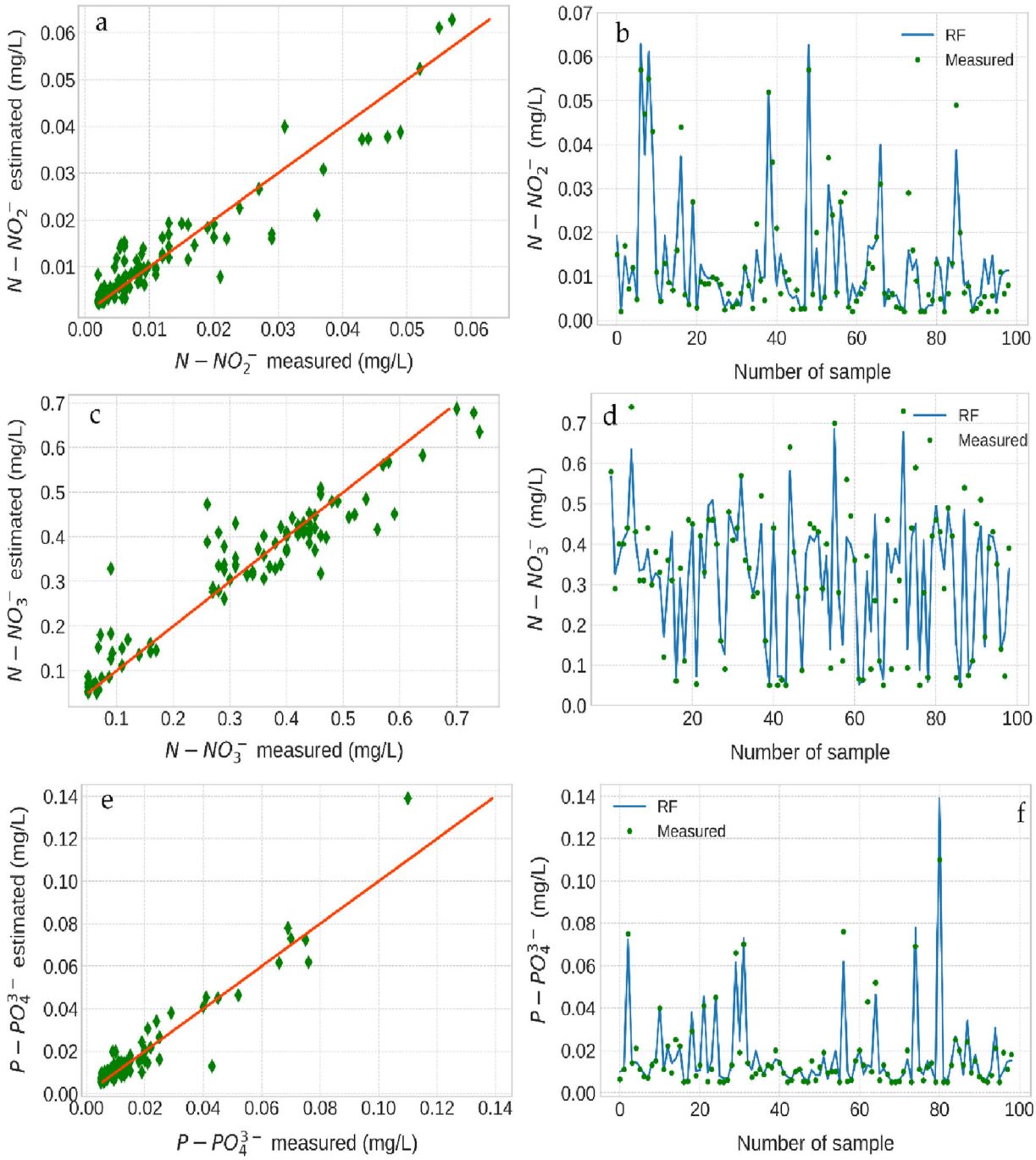

Fig. 5 Best performance from RF model of N-NO_2^- (**a**), N-NO_3^- (**c**), and P-PO_4^{3-} (**e**) retrieval from water parameters in 2013–2014, and observed and predicted (**b**, **d**, **f**) values

adaptive neuro-fuzzy inference system ($R^2 = 0.86$, water quality dataset 1993–2013) (Chen and Liu 2015), ANN (feed-forward network, $R^2 = 0.99$, hydro-meteorological dataset 1994–2000) (Kim et al. 2012), RF (no R^2 reported, water quality dataset 2009–2012) (Castrillo and García 2020), and RF (Pearson's coefficients reached approximately 0.66 on average, environmental dataset 1994–2018) (Shen et al. 2020) models. Despite a

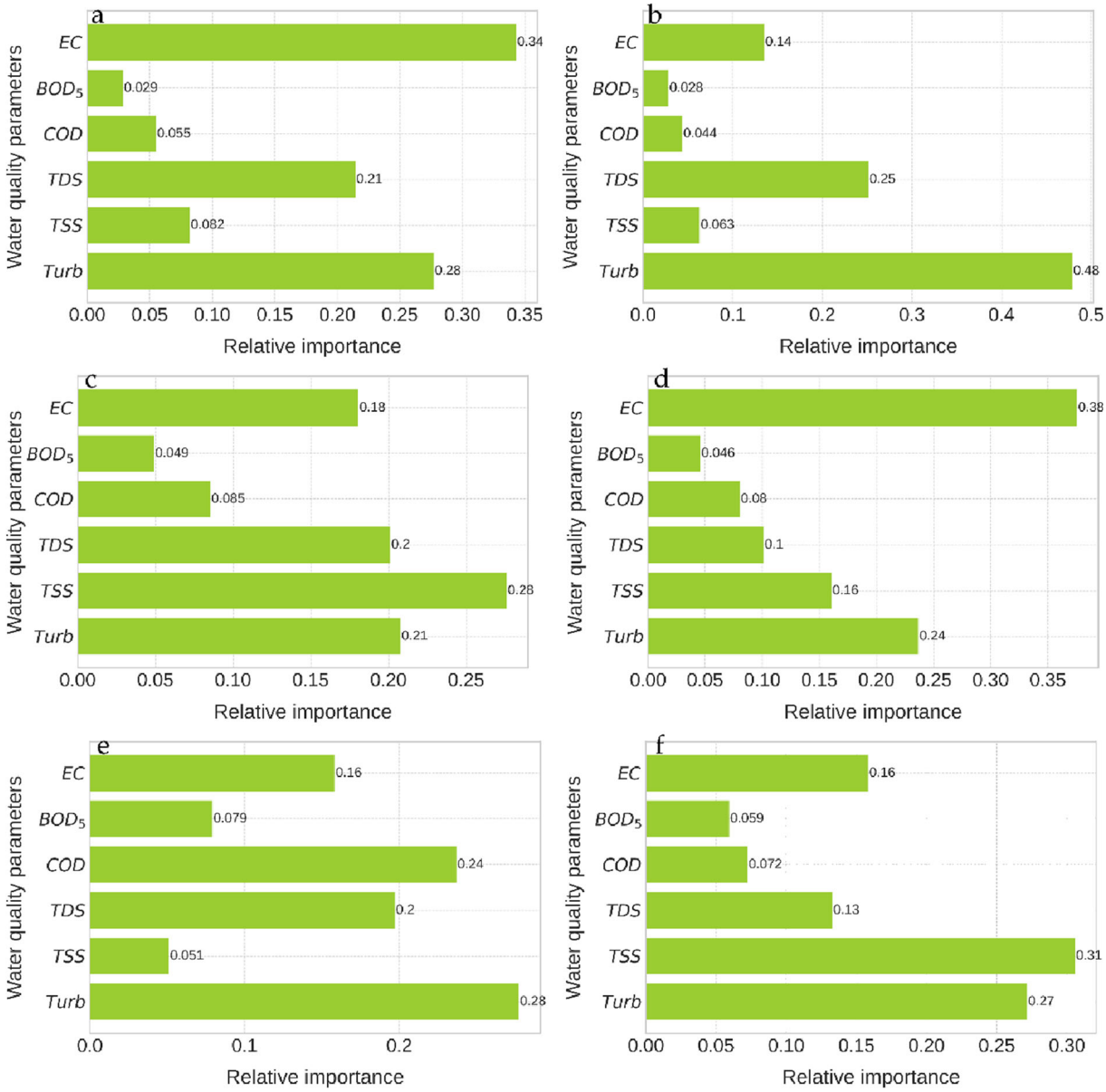

Fig. 6 The importance of input features for N-NO_2^- (**a**), N-NO_3^- (**b**), and P-PO_4^{3-} (**c**) retrieval during the training phase 2009–2012, and the validation phase 2013–2014 (**d**, **e**, **f**)

similar approach of using the RF model, it is necessary to note the differences between Castrillo and García (2020) and Shen et al. (2020) and our case study, including the water ecosystem applied, the input parameters, the wide range of Chl-a concentration, and the performance of the RF model. Here, we challenged the RF model using a unique set of hyper-parameters to train and validate the performance in a 6-year time series data (2009–2014) for a HCB-suffering reservoir with a very high and wide range of Chl-a concentration (6–3400 μg/L). The ten-fold CV for each of the training and validation phases identified a stable and reliable estimation of N-NO_2^-, N-NO_3^-, and P-PO_4^{3-} (R^2 = 0.737–0.806 in 2009–2012; R^2 = 0.814–0.830 in 2013–2014) using the surrogates of EC, TDS, TSS, COD, BOD_5, and turbidity. The application of the RF model in our study therefore is unique and applicable to various regions, indicates a sensitive performance for various

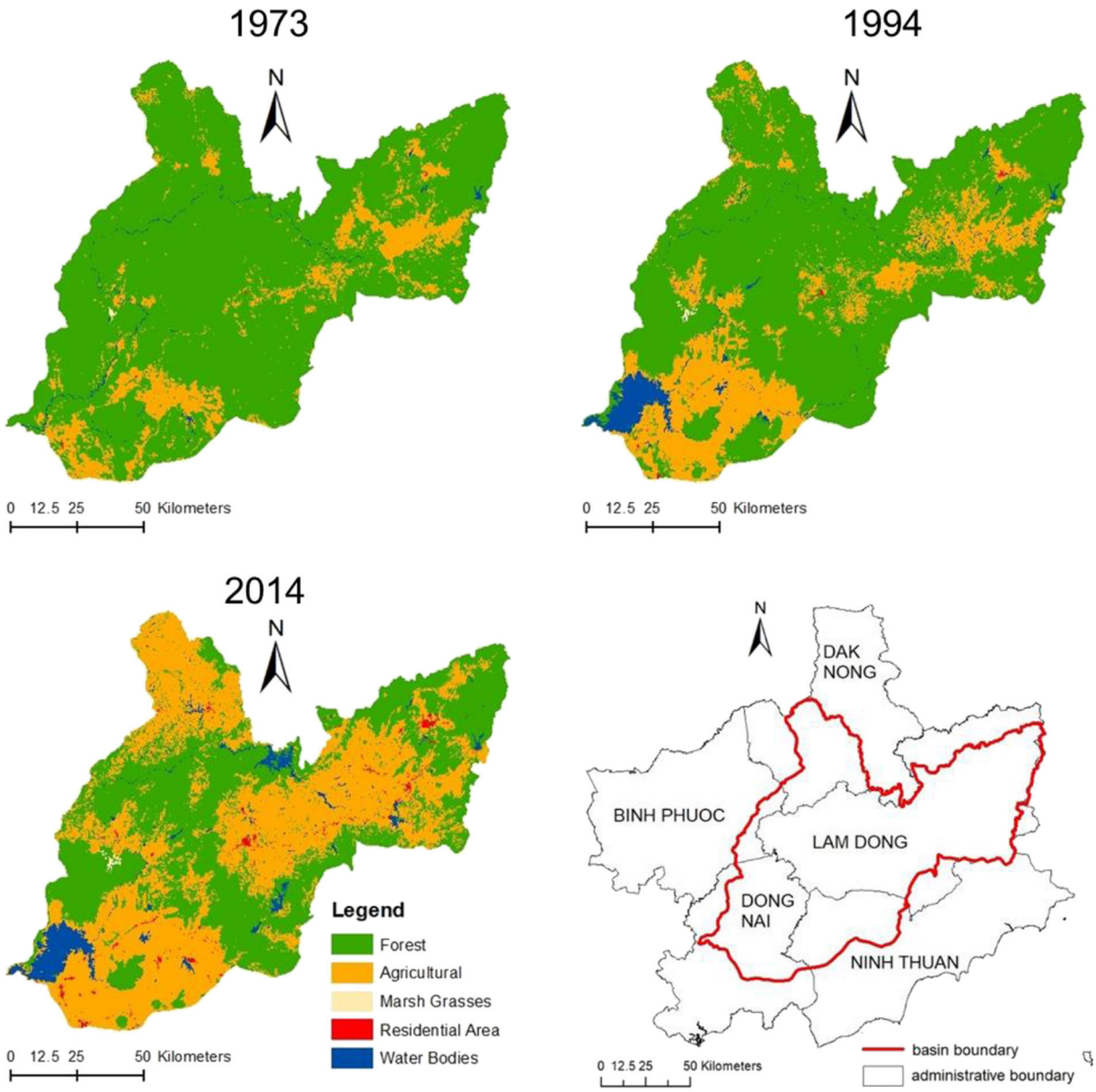

Fig. 7 Land use/cover change during 1973–2014 period and administrative boundaries of the upstream of Dong Nai River Basin (adapted from Truong et al. 2018)

chemical forms of N and P, diversifies the selection of ML algorithms for nutrient factor retrieval, and provides a good foundation for further assessment of nutrient loading in inland lakes or reservoirs.

Note that the importance of input features is different from that of the retrieved parameters and observed periods with a larger contribution of the EC, TDS, TSS, COD, and turbidity parameters to the retrieval of N-NO_2^-, N-NO_3^-, and P-PO_4^{3-}. Figure 6 indicates a slight difference in feature importance of training (2009–2012) and validation (2013–2014); however, the most contributing factor was the same with the EC for N-NO_2^-, turbidity for N-NO_3^-, and TSS for P-PO_4^{3-} in the selected phases. N-NO_2^- and N-NO_3^- levels in the water have exhibited significant positive relationship with EC (Calvi et al. 2018). Turbidity is the measure of relative clarity of the water, while TSS can be used as the indicator of sediment in the reservoir, which usually

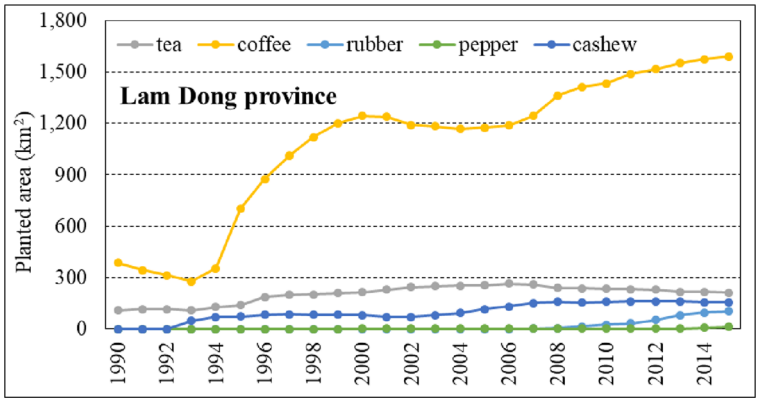

Fig. 8 Change in planted area industrial crops during the 1973–2014 period in Lam Dong Province

consists of silt, fine sand, and algae (Herschy 2012). *Turbidity* levels are positively related to *TSS (*Daphne et al. 2011). In the TAR, high turbidity and TSS concentration were primarily caused by rainfall runoff (Pham et al. 2020a, b). Our results agree with previous studies that high turbidity and TSS levels led to an increase in nutrients of the water (Nguyen et al. 2020; Pham et al. 2020a, b).

The application of the RF model to nutrient species retrieval, however, has some limitations. Despite an attempt to tune the hyper-parameters, the performance of RF (R^2 values) using a ten-fold CV in the training phase was not as high as expected. This could lead to the demand for a more effective optimizing tool for RF and for ML models to employ long-term observation data. In addition, the complex water environment in and the anthropogenic activities surrounding the TAR resulted in a large variation of water quality parameters, more complex interactions among the parameters, and different contributions from various parameters to the retrieval of N-NO_2^-, N-NO_3^-, and P-PO_4^{3-}. Our results determined this variation to be an explanatory factor in the performance of the RF model in the 2009–2012 period.

Effects of LUCC and meteorological factors on nutrient variation in the TAR

Clearly, water eutrophication in the TAR was affected by several factors, including LUCC and climatic effects. Our analysis showed evidence of nutrient concentration increasing in proportion to the amount of applied fertilizer in agricultural land in the upstream regions. Brito et al. (2019) demonstrated that an application of 100 kg N, P/ha/year on the watershed may resulted in loading of 2.5–2.8 kg N and 1.5 kg P/year to the reservoir. In addition, soil erosion, deforestation, and agricultural land conversion were identified as the driving factors of degrading water quality in the lower section of the basin. Shang (2019) reported the same trend, in which fertilizer application in agricultural lands was a major source of nitrogen (N-NO_3^- and NH_4-N) export

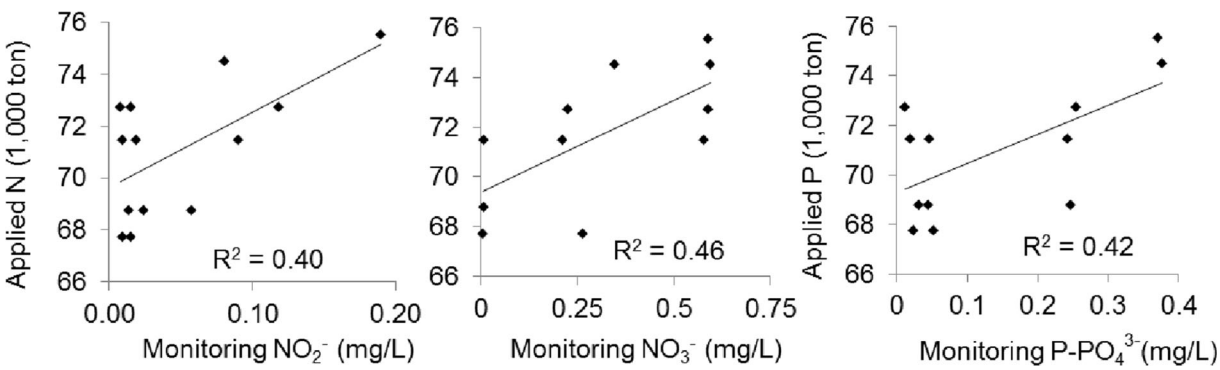

Fig. 9 Correlation between the applied amount of fertilizer (N-P-K) and the monitored concentration of nutrients (N-NO_2^-, N-NO_3^-, and P-PO_4^{3-})

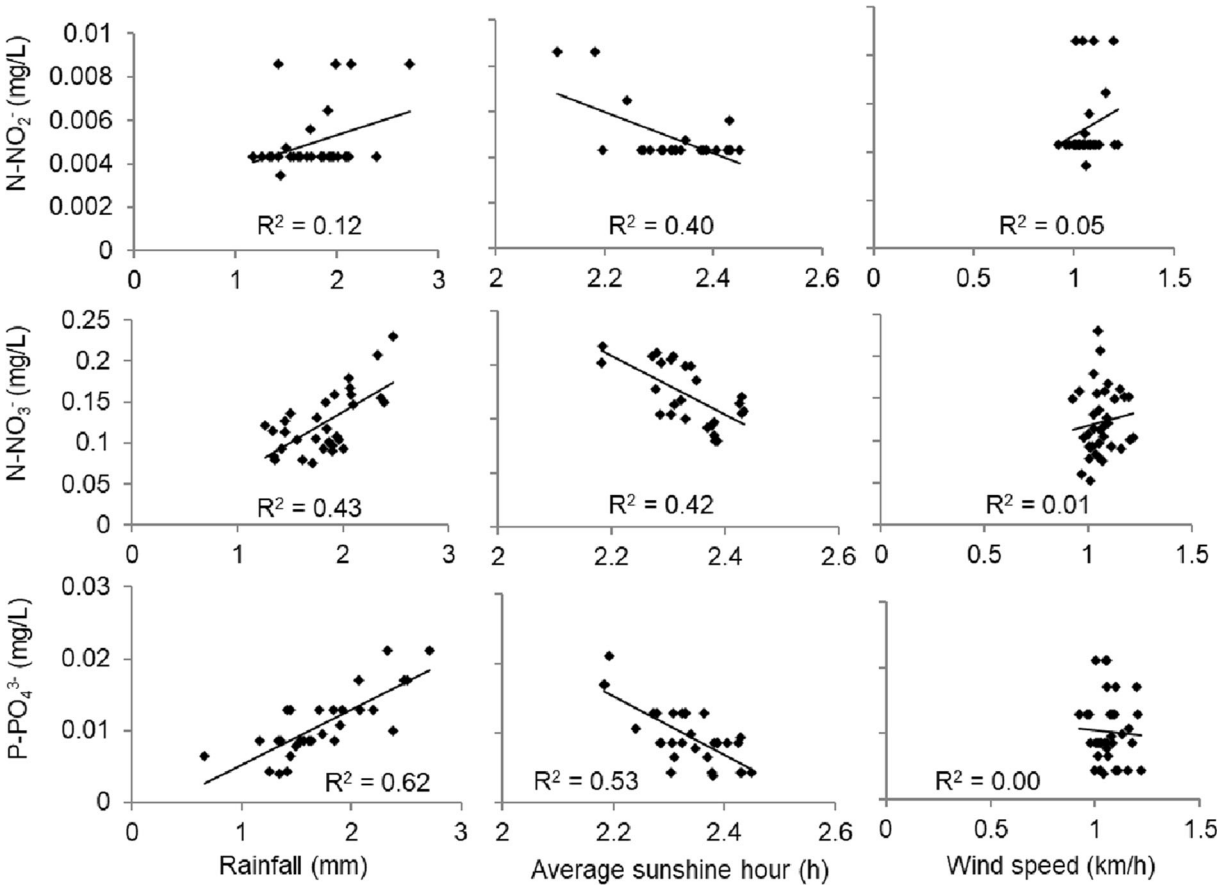

Fig. 10 Correlation between nutrient concentrations and meteorological factors

and concluded that watersheds with more agricultural land had larger nitrogen loads. In other basins, TN and TP loading increased correspondingly with the conversion from forest to agricultural land (Bi et al. 2018; Tong and Chen 2002).

In the TAR, rainfall, which will occur more frequently in the following decades (Park et al. 2011; Wang et al. 2012), is the most important climatic factor in the loading increment of nutrient species (R^2 = 0.43–0.62, according to the upper trend of $N\text{-}NO_3^-$ and $P\text{-}PO_4^{3-}$).

Fig. 11 Intensive blooms of cyanobacteria during the rainy season in the Tri An Reservoir

High rainfall may result in strong soil erosion, strengthen the velocity of freshwater flow, increase nutrient loading into the lower basin (Li et al. 2015), and benefit HCB (Mu et al. 2019; Reichwaldt and Ghadouani 2012). Intense HCB are commonly observed from June to October in the TAR (Fig. 11). Geographically, the topography of the TAR may lead to a high volume of surface runoff into the reservoir, especially during the 6 months of the rainy season (May to October). Alongside the long monthly sunshine hour in the dry season (R^2 = 0.40–0.53), the TAR may suffer further anthropogenic pollutant loads, which may bring HCB to a higher frequency in the future. Therefore, we suggest an integrated investigation of hydrological, meteorological, and nutrient loading in the TAR catchment area to better understand the mechanism of large HCB.

Conclusions

To our knowledge, this research is the first to predict nutrient factors in the TAR using the ML approach. Our results indicated that RF is a sensitivity predictor of N-NO_2^-, N-NO_3^-, and P-PO_4^{3-} retrieval for both the training (R^2 ten-fold CV = 0.737–0.806) and validation (R^2 ten-fold CV = 0.814–0.830) phases. Using RF is rational in our case with small predictions of RMSEs. During the training and validation phases, EC, TSS, and turbidity were among the most important contributors to the estimation of N-NO_2^-, N-NO_3^-, and P-PO_4^{3-} in the TAR.

The conversion of a large-scale forest to agricultural land in the 1973–2014 period led to an increase in fertilizer application and resulted in the degradation of water quality in the TAR. Our results determined a positive correlation (R^2 = 0.40–0.46) between applied fertilizers (N and P in ton/ha/year) and monitoring nutrients (N-NO_2^-, N-NO_3^-, P-PO_4^{3-}).

Eutrophication in the TAR, with an annual average concentration of N-NO_3^- and P-PO_4^{3-} of 0.3 mg/L and 0.01 mg/L, respectively, led to intensive HCB from June to October. Variations in recent rainfall patterns have significantly increased nutrient loads from the watershed into the reservoir. Based on our results, strict management strategies are recommended to improve the efficiency of fertilizer use in agriculture.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10661-020-08731-2.

Funding This study was mainly granted by the Vietnam Academy of Science and Technology (VAST) under grant number "KHCBSS.02/19-21", and in part by the International Foundation for Science (IFS) under grant number "I-2-A-6054-1".

References

Asian Development Bank (ADB). (2009). *Water: vital for Viet Nam's future*. Vietnam: Ha Noi.

Belgiu, M., & Drăguţ, L. (2016). Random forest in remote sensing: a review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing, 114*, 24–31. https://doi.org/10.1016/j.isprsjprs.2016.01.011.

Berrendero, E., Valiente, E. F., Perona, E., Gómez, C. L., Loza, V., Muñoz-Martín, M. Á., & Mateo, P. (2016). Nitrogen fixation in a non-heterocystous cyanobacterial mat from a mountain river. *Scientific Reports, 6*(1), 30920. https://doi.org/10.1038/srep30920.

Bi, W., Weng, B., Yuan, Z., Ye, M., Zhang, C., Zhao, Y., Yan, D., & Xu, T. (2018). Evolution characteristics of surface water quality due to climate change and LUCC under scenario simulations: a case study in the Luanhe River Basin. *International Journal of Environmental Research and Public Health, 15*(8), 1724. https://doi.org/10.3390/ijerph15081724.

Blix, K., & Eltoft, T. (2018). Machine learning automatic model selection algorithm for oceanic chlorophyll-a content retrieval. *Remote Sensing, 10*(5), 775. https://doi.org/10.3390/rs10050775.

Breiman, L. (2001). Random forest. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324.

Bresciani, M., Cazzaniga, I., Austoni, M., Sforzi, T., Buzzi, F., Morabito, G., & Giardino, C. (2018). Mapping phytoplankton blooms in deep subalpine lakes from Sentinel-2A and Landsat-8. *Hydrobiologia, 824*(1), 197–214. https://doi.org/10.1007/s10750-017-3462-2.

Bridgewater, L. L., Baird, R. B., Eaton, A. D., Rice, E. W., & American Public Health Association, American Water Works Association, & Water Environment Federation (Eds.). (2017). *Standard methods for the examination of water and wasterwater* (23rd ed.). Washington, DC: American Public Health Association.

Brito, D., Neves, R., Branco, M., Prazeres, Â., Rodrigues, S., Gonçalves, M., & Ramos, T. (2019). Assessing water and nutrient long-term dynamics and loads in the Enxoé temporary river basin (Southeast Portugal). *Water, 11*(2), 354. https://doi.org/10.3390/w11020354.

Bui, M.-H., Pham, T.-L., & Dao, T.-S. (2017). Prediction of cyanobacterial blooms in the Dau Tieng Reservoir using an artificial neural network. *Marine and Freshwater Research, 68*(11), 2070. https://doi.org/10.1071/MF16327.

Byrareddy, V., Kouadio, L., Mushtaq, S., & Stone, R. (2019). Sustainable production of robusta coffee under a changing climate: a 10-year monitoring of fertilizer management in

coffee farms in Vietnam and Indonesia. *Agronomy, 9*(9), 499. https://doi.org/10.3390/agronomy9090499.

Calvi, C., Dapeña, C., Martinez, D. E., & Quiroz Londoño, O. M. (2018). Relationship between electrical conductivity, 18O of water and NO3 content in different streamflow stages. *Environmental Earth Sciences, 77*(6), 248. https://doi.org/10.1007/s12665-018-7427-1.

Carlsson, H., Aspegren, H., & Hilmer, A. (1996). Interactions between wastewater quality and phosphorus release in the anaerobic reactor of the EBPR process. *Water Research, 30*(6), 1517–1527. https://doi.org/10.1016/0043-1354(95)00333-9.

Castrillo, M., & García, Á. L. (2020). Estimation of high frequency nutrient concentrations from water quality surrogates using machine learning methods. *Water Research, 172*, 115490. https://doi.org/10.1016/j.watres.2020.115490.

Chen, W.-B., & Liu, W.-C. (2015). Water quality modeling in reservoirs using multivariate linear regression and two neural network models. *Advances in Artificial Neural Systems, 2015*, 1–12. https://doi.org/10.1155/2015/521721.

Corwin, D. L., Lesch, S. M., Oster, J. D., & Kaffka, S. R. (2006). Monitoring management-induced spatio–temporal changes in soil quality through soil sampling directed by apparent electrical conductivity. *Geoderma, 131*(3), 369–387. https://doi.org/10.1016/j.geoderma.2005.03.014.

Daphne, L., Djati Utomo, H., & Kenneth, L. (2011). Correlation between turbidity and total suspended solids in Singapore rivers. *Journal of Water Sustainability, 1*, 313–322.

Davies-Colley, R. J., Hickey, C. W., & Quinn, J. M. (1995). Organic matter, nutrients, and optical characteristics of sewage lagoon effluents. *New Zealand Journal of Marine and Freshwater Research, 29*(2), 235–250. https://doi.org/10.1080/00288330.1995.9516657.

Dubey, D., & Dutta, V. (2020). Nutrient enrichment in lake ecosystem and its effects on algae and macrophytes. In V. Shukla & N. Kumar (Eds.), *Environmental concerns and sustainable development* (pp. 81–126). Singapore: Springer Singapore. https://doi.org/10.1007/978-981-13-6358-0_5.

Durbin–Watson Test. (2008). *In The concise encyclopedia of statistics (pp. 173–175)*. New York: Springer New York. https://doi.org/10.1007/978-0-387-32833-1_122.

Fawagreh, K., Gaber, M. M., & Elyan, E. (2014). Random forests: from early developments to recent advancements. *Systems Science & Control Engineering, 2*(1), 602–609. https://doi.org/10.1080/21642583.2014.956265.

García, N. P. J., García-Gonzalo, E., Alonso Fernández, J. R., & Díaz Muñiz, C. (2019). Water eutrophication assessment relied on various machine learning techniques: a case study in the Englishmen Lake (Northern Spain). *Ecological Modelling, 404*, 91–102. https://doi.org/10.1016/j.ecolmodel.2019.03.009.

Grattan, L. M., Holobaugh, S., & Morris, J. G. (2016). Harmful algal blooms and public health. *Harmful Algae, 57*, 2–8. https://doi.org/10.1016/j.hal.2016.05.003.

Heisler, J., Glibert, P. M., Burkholder, J. M., Anderson, D. M., Cochlan, W., Dennison, W. C., Dortch, Q., Gobler, C. J., Heil, C. A., Humphries, E., Lewitus, A., Magnien, R., Marshall, H. G., Sellner, K., Stockwell, D. A., Stoecker, D. K., & Suddleson, M. (2008). Eutrophication and harmful algal blooms: a scientific consensus. *Harmful Algae, 8*(1), 3–13. https://doi.org/10.1016/j.hal.2008.08.006.

Herschy, R. W. (2012). Lake sediments. In L. Bengtsson, R. W. Herschy, & R. W. Fairbridge (Eds.), *Encyclopedia of lakes and reservoirs*. Dordrecht: Encyclopedia of Earth Sciences Series. Springer. https://doi.org/10.1007/978-1-4020-4410-6_26.

Hollister, J. W., Milstead, W. B., & Kreakie, B. J. (2016). Modeling lake trophic state: a random forest approach. *Ecosphere, 7*(3), e01321. https://doi.org/10.1002/ecs2.1321.

JICA. (n.d.) (1996). The master plan study on Dong Nai River and surrounding basins water resources development: final report: Vol. 4. Appendix II: Topography and geology, appendix III: Meteorology and hydrology. Japan International Cooperation Agency: Nippon Koei Co., Ltd. https://openjicareport.jica.go.jp/617/617/617_123_11309523.html. Accessed 20 Sep 2017.

Jones, K. B., Neale, A. C., Nash, M. S., Van Remortel, R. D., Wickham, J. D., Riitters, K. H., & O'Neill, R. V. (2001). Predicting nutrient and sediment loadings to streams from landscape metrics: a multiple watershed study from the United States Mid-Atlantic region. *Landscape Ecology, 16*(4), 301–312. https://doi.org/10.1023/A:1011175013278.

Jones, J. R., Knowlton, M. F., Obrecht, D. V., & Cook, E. A. (2004). Importance of landscape variables and morphology on nutrients in Missouri reservoirs. *Canadian Journal of Fisheries and Aquatic Sciences*. https://doi.org/10.1139/f04-088.

Jung, K., Bae, D.-H., Um, M.-J., Kim, S., Jeon, S., & Park, D. (2020). Evaluation of nitrate load estimations using neural networks and canonical correlation analysis with k-fold cross-validation. *Sustainability, 12*(1), 400. https://doi.org/10.3390/su12010400.

Keller, S., Maier, P., Riese, F., Norra, S., Holbach, A., Börsig, N., Wilhelms, A., Moldaenke, C., Zaake, A., & Hinz, S. (2018). Hyperspectral data and machine learning for estimating CDOM, chlorophyll a, diatoms, green algae and turbidity. *International Journal of Environmental Research and Public Health, 15*(9), 1881. https://doi.org/10.3390/ijerph15091881.

Kim, K.-S., Yoo, J.-S., Kim, S., Lee, H. J., Ahn, K.-H., & Kim, I. S. (2007). Relationship between the electric conductivity and phosphorus concentration variations in an enhanced biological nutrient removal process. *Water Science and Technology: A Journal of the International Association on Water Pollution Research, 55*(1–2), 203–208. https://doi.org/10.2166/wst.2007.053.

Kim, R. J., Loucks, D. P., & Stedinger, J. R. (2012). Artificial neural network models of watershed nutrient loading. *Water Resources Management, 26*(10), 2781–2797. https://doi.org/10.1007/s11269-012-0045-x.

Lam Dong Department of Statistic (2018). *Lam Dong statistical yearbook 2018*. Statistical Publishing House (Vietnam). Accessed 2 Jan 2020.

Lewis, W. M., Wurtsbaugh, W. A., & Paerl, H. W. (2011). Rationale for control of anthropogenic nitrogen and phosphorus to reduce eutrophication of inland waters. *Environmental Science & Technology, 45*(24), 10300–10305. https://doi.org/10.1021/es202401p.

Li, X., Huang, T., Ma, W., Sun, X., & Zhang, H. (2015). Effects of rainfall patterns on water quality in a stratified reservoir subject to eutrophication: implications for management. *Science of The Total Environment, 521–522*, 27–36. https://doi.org/10.1016/j.scitotenv.2015.03.062.

Li, X., Sha, J., & Wang, Z.-L. (2018). Application of feature selection and regression models for chlorophyll-a prediction in a shallow lake. *Environmental Science and Pollution Research, 25*(20), 19488–19498. https://doi.org/10.1007/s11356-018-2147-3.

Lou, I., Xie, Z., Ung, W. K., & Mok, K. M. (2016). Freshwater algal bloom prediction by extreme learning machine in Macau storage reservoirs. *Neural Computing and Applications, 27*(1), 19–26. https://doi.org/10.1007/s00521-013-1538-0.

Lu, J., Zhu, B., Struewing, I., Xu, N., & Duan, S. (2019). Nitrogen–phosphorus-associated metabolic activities during the development of a cyanobacterial bloom revealed by metatranscriptomics. *Scientific Reports, 9*(1), 2480. https://doi.org/10.1038/s41598-019-38481-2.

Marttila, H., & Kløve, B. (2009). Retention of Sediment and Nutrient Loads with Peak Runoff Control. *Journal of Irrigation and Drainage Engineering, 135*(2), 210–216.

Marttila, H., & Kløve, B. (2012). Use of turbidity measurements to estimate suspended solids and nutrient loads from peatland forestry drainage. *Journal of Irrigation and Drainage Engineering, 138*(12), 1088–1096. https://doi.org/10.1061/(ASCE)IR.1943-4774.0000509.

Mas, J. F., & Flores, J. J. (2008). The application of artificial neural networks to the analysis of remotely sensed data. *International Journal of Remote Sensing, 29*(3), 617–663. https://doi.org/10.1080/01431160701352154.

Meyfroidt, P., Vu, T. P., & Hoang, V. A. (2013). Trajectories of deforestation, coffee expansion and displacement of shifting cultivation in the Central Highlands of Vietnam. *Global Environmental Change, 23*(5), 1187–1198. https://doi.org/10.1016/j.gloenvcha.2013.04.005.

Mohapatra, N., Shreya, K., & Chinmay, A. (2020). Optimization of the random forest algorithm. In S. Borah, V. Emilia Balas, & Z. Polkowski (Eds.), *Advances in data science and management* (Vol. 37, pp. 201–208). Singapore: Springer Singapore. https://doi.org/10.1007/978-981-15-0978-0_19.

Morris, J. G. (1999). An emerging public health problem with possible links to human stress on the environment. *Annual Review of Energy and the Environment, 24*(1), 367–390. https://doi.org/10.1146/annurev.energy.24.1.367.

Mu, M., Wu, C., Li, Y., Lyu, H., Fang, S., Yan, X., Liu, G., Zheng, Z., Du, C., & Bi, S. (2019). Long-term observation of cyanobacteria blooms using multi-source satellite images: a case study on a cloudy and rainy lake. *Environmental Science and Pollution Research*. https://doi.org/10.1007/s11356-019-04522-6.

Nguyen, H.-Q., Ha, N.-T., & Pham, T.-L. (2020). Inland harmful cyanobacterial bloom prediction in the eutrophic Tri An Reservoir using satellite band ratio and machine learning approaches. *Environmental Science and Pollution Research, 27*, 9135–9151. https://doi.org/10.1007/s11356-019-07519-3.

Oyebode, O., & Stretch, D. (2019). Neural network modeling of hydrological systems: a review of implementation techniques. *Natural Resource Modeling, 32*(1), e12189. https://doi.org/10.1111/nrm.12189.

Park, J.-H., Inam, E., Abdullah, M. H., Agustiyani, D., Duan, L., Hoang, T. T., Kim, K.-W., Kim, S. D., Nguyen, M. H., Pekthong, T., Sao, V., Sarjiya, A., Savathvong, S., Sthiannopkao, S., Keith Syers, J., & Wirojanagud, W. (2011). Implications of rainfall variability for seasonality and climate-induced risks concerning surface water quality in East Asia. *Journal of Hydrology, 400*(3–4), 323–332. https://doi.org/10.1016/j.jhydrol.2011.01.050.

Park, Y., Cho, K. H., Park, J., Cha, S. M., & Kim, J. H. (2015). Development of early-warning protocol for predicting chlorophyll-a concentration using machine learning models in freshwater and estuarine reservoirs, Korea. *Science of The Total Environment, 502*, 31–41. https://doi.org/10.1016/j.scitotenv.2014.09.005.

Parmar, A., Katariya, R., & Patel, V. (2019). A review on random forest: an ensemble classifier. In J. Hemanth, X. Fernando, P. Lafata, & Z. Baig (Eds.), *International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018* (pp. 758–763). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-03146-6_86.

Paudel, B., Montagna, P. A., & Adams, L. (2019). The relationship between suspended solids and nutrients with variable hydrologic flow regimes. *Regional Studies in Marine Science, 29*, 100657. https://doi.org/10.1016/j.rsma.2019.100657.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, & Cournapeau, D. (2011). scikit-learn: machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Pham, T.-L., Tran, T. H. Y., Hoang, N. S., Ngo, X. Q., & Tran, T. T. (2020a). Co-occurrence of microcystin- and geosmin-producing cyanobacteria in the Tri An Reservoir, a drinking-water supply in Vietnam. *Fundamental and Applied Limnology/Archiv für Hydrobiologie, 193*(4), 299–311. https://doi.org/10.1127/fal/2020/1296.

Pham, T.-L., Tran, T. H. Y., Shimizu, K., Li, Q., & Utsumi, M. (2020b). Toxic cyanobacteria and microcystin dynamics in a tropical reservoir: assessing the influence of environmental variables. *Environmental Science and Pollution Research*. https://doi.org/10.1007/s11356-020-10826-9.

PVFCCo (2016). Polyhalite application improves coffee (*Coffea robusta*) yield and quality in Vietnam. International Potash Institute, e-ifc, No. 47, December 2016, pp. 12–19. https://www.ipipotash.org/uploads/udocs/e-ifc-47-dec2016-coffee-vietnam.pdf. Accessed 1 Oct 2020

Qian, S. S., Reckhow, K. H., Zhai, J., & McMahon, G. (2005). Nonlinear regression modeling of nutrient loads in streams: a Bayesian approach. *Water Resources Research, 41*(7). https://doi.org/10.1029/2005WR003986.

Reichwaldt, E. S., & Ghadouani, A. (2012). Effects of rainfall patterns on toxic cyanobacterial blooms in a changing climate: between simplistic scenarios and complex dynamics. *Water Research, 46*(5), 1372–1393. https://doi.org/10.1016/j.watres.2011.11.052.

Ross, M. R. V., Topp, S. N., Appling, A. P., Yang, X., Kuhn, C., Butman, D., Simard, M. & Pavelsky, T.M. (2019). AquaSat: a data set to enable remote sensing of water quality for inland waters. *Water Resources Research*, 2019WR024883. https://doi.org/10.1029/2019WR024883

Schindler, D. W., Hecky, R. E., Findlay, D. L., Stainton, M. P., Parker, B. R., Paterson, M. J., Beaty, K. G., Lyng, M., & Kasian, S. E. M. (2008). Eutrophication of lakes cannot be controlled by reducing nitrogen input: results of a 37-year

whole-ecosystem experiment. *Proceedings of the National Academy of Sciences, 105*(32), 11254–11258.

Seabold, S., & Perktold, J. (2010). Statsmodels: econometric and statistical modeling with Python. In *9th Python in Science Conference*.

Shang, L. (2019). Climate change and land use/cover change impacts on watershed hydrology, nutrient dynamics – a case study in Missisquoi River watershed (Graduate College Dissertations and Theses). Vermont. Retrieved from https://scholarworks.uvm.edu/graddis/1016. Accessed 20 Jan 2020.

Shen, L. Q., Amatulli, G., Sethi, T., Raymond, P., & Domisch, S. (2020). Estimating nitrogen and phosphorus concentrations in streams and rivers, within a machine learning framework. *Scientific Data, 7*(1), 161. https://doi.org/10.1038/s41597-020-0478-7.

Sihag, P., Mohsenzadeh Karimi, S., & Angelaki, A. (2019). Random forest, M5P and regression analysis to estimate the field unsaturated hydraulic conductivity. *Applied Water Science, 9*(5), 129. https://doi.org/10.1007/s13201-019-1007-8.

Tiemann, T., Maung Aye, T., Duc Dung, N., Minh Tien, T., Fisher, M., Nalin de Paulo, E., & Oberthür, T. (2018). Crop nutrition for Vietnamese robusta coffee. *Better Crops with Plant Food, 102*(3), 20–23. https://doi.org/10.24047/BC102320.

Tong, S. T. Y., & Chen, W. (2002). Modeling the relationship between land use and surface water quality. *Journal of Environmental Management, 66*(4), 377–393. https://doi.org/10.1006/jema.2002.0593.

Trung, B., Dao, T.-S., Faassen, E., & Lürling, M. (2018). Cyanobacterial blooms and microcystins in Southern Vietnam. *Toxins, 10*(11), 471. https://doi.org/10.3390/toxins10110471.

Truong, N., Nguyen, H., & Kondoh, A. (2018). Land use and land cover changes and their effect on the flow regime in the upstream Dong Nai River Basin, Vietnam. *Water, 10*(9), 1206. https://doi.org/10.3390/w10091206.

Tu, J. V. (1996). Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of Clinical Epidemiology, 49*(11), 1225–1231. https://doi.org/10.1016/S0895-4356(96)00002-9.

Tyralis, H., Papacharalampous, G., & Langousis, A. (2019). A brief review of random forests for water scientists and practitioners and their recent history in water resources. *Water, 11*(5), 910. https://doi.org/10.3390/w11050910.

van Puijenbroek, P. J. T. M., Beusen, A. H. W., & Bouwman, A. F. (2019). Global nitrogen and phosphorus in urban waste water based on the shared socio-economic pathways. *Journal of Environmental Management, 231*, 446–456. https://doi.org/10.1016/j.jenvman.2018.10.048.

Wang, S., Qian, X., Han, B.-P., Luo, L.-C., & Hamilton, D. P. (2012). Effects of local climate and hydrological conditions on the thermal regime of a reservoir at Tropic of Cancer, in southern China. *Water Research, 46*(8), 2591–2604. https://doi.org/10.1016/j.watres.2012.02.014.

Wang, X., Liu, Z., Miao, J., & Zuo, N. (2015). Relationship between nutrient pollutants and suspended sediments in upper reaches of Yangtze River. *Water Science and Engineering, 8*(2), 121–126. https://doi.org/10.1016/j.wse.2015.04.003.

Wang, X., Gong, Z., & Pu, R. (2018). Estimation of chlorophyll a content in inland turbidity waters using WorldView-2 imagery: a case study of the Guanting Reservoir, Beijing, China. *Environmental Monitoring and Assessment, 190*(10), 620. https://doi.org/10.1007/s10661-018-6978-7.

Wang, X., Daigger, G., de Vries, W., Kroeze, C., Yang, M., Ren, N.-Q., Liu, J., & Butler, D. (2019). Impact hotspots of reduced nutrient discharge shift across the globe with population and dietary changes. *Nature Communications, 10*(1), 2627. https://doi.org/10.1038/s41467-019-10445-0.

Xi, B.-D., Zhang, Y.-L., & Xu, Q.-J. (2012). Possibility of total dissolved solid as one of nutrient baselines in inner Mongolia-Xinjiang plateau. *Huan Jing Ke Xue= Huanjing Kexue, 33*(10), 3308–3313.

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. https://doi.org/10.2166/hydro.2017.010.

Yi, H.-S., Lee, B., Park, S., Kwak, K.-C., & An, K.-G. (2018). Short-term algal bloom prediction in Juksan weir using M5P model-tree and extreme learning machine. *Environmental Engineering Research*. https://doi.org/10.4491/eer.2018.245.

Zhang, H., Cui, B., Hong, J., & Zhang, K. (2011). Synergism of natural and constructed wetlands in Beijing, China. *Ecological Engineering, 37*(2), 128–138. https://doi.org/10.1016/j.ecoleng.2010.08.001.

Zhang, L., Huettmann, F., Zhang, X., Liu, S., Sun, P., Yu, Z., & Mi, C. (2019). The use of classification and regression algorithms using the random forests method with presence-only data to model species' distribution. *MethodsX, 6*, 2281–2292. https://doi.org/10.1016/j.mex.2019.09.035.