Wetlands Ecol Manage
DOI 10.1007/s11273-017-9558-7

ORIGINAL PAPER

Modeling temporal patterns of methane effluxes using multiple regression and random forest in Poyang Lake, China

Lixiang Liu · Ming Xu · Renqiang Li

Received: 3 August 2016 / Accepted: 8 June 2017

Abstract In this study, we used two statistical models to predict daily CH_4 effluxes and compared the prediction accuracy of two models in Poyang Lake. Statistical models included linear model and Random forest model (RF) which can handle high dimensional non-linear relationships, categorical and continuous predictors, and highly collinear predictor variables. Seven climatic factors and water level data, together with the field CH_4 efflux at monthly intervals from 2011 to 2014 were used for model development and cross-validation. We found that the RF model provided the best prediction accuracy for daily CH_4 effluxes, whereas the linear model gave low prediction accuracy for CH_4 effluxes. The coefficient of determination was 0.93 and 0.63 for the "best" RF and linear models with the same climatic variables, respectively. The "best" linear model had the highest model-performance errors including the mean absolute error, root mean-square error, and the normalized root-mean-square error, followed by the "best" RF models. In addition, cross-validation results for the two "best" models also showed that the RF model was the best model for estimating CH_4 effluxes. We applied the optimum RF model to simulate daily CH_4 effluxes from 1 January 2011 to 31 December 2014, and then estimated the seasonal and annual CH_4 emissions in Poyang Lake. The mean CH_4 efflux in the summer was notably higher than that in the other seasons, with values of 0.097, 0.28, 0.11, and 0.045 mmol m^{-2} day^{-1} in the spring, in the summer, in the autumn, and in the winter over a 4-year period, respectively. The mean annual emission was 3.13 g m^{-2} $year^{-1}$, which was considerately lower than the mean global annual emission in lakes and that in the other subtropical lakes of the world. We found that the RF model may be used to estimate CH_4 effluxes and emissions in other lakes in the world.

Keywords Linear model · Random forest model · Model validation · Model selection

L. Liu · R. Li
Key Laboratory of Ecosystem Network Observation and Modeling, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China

M. Xu (✉)
Department of Ecology, Evolution and Natural Resources, Center for Remote Sensing and Spatial Analysis, Rutgers University, New Brunswick, NJ 08901, USA
e-mail: mingxu@crssa.rutgers.edu

Introduction

Lakes are an important source of atmospheric CH_4 emissions, which contribute to approximately 30% of all natural CH_4 emissions (IPCC 2013). Global lakes emit 8–71.6 Tg—CH_4/year into the atmosphere (Bastviken et al. 2004, 2011). However, these estimations remain largely uncertain due to the lack of field

Published online: 16 June 2017

CH_4 effluxes, large spatial and temporal variations, and the lack of continuous CH_4 effluxes based on long-term measurements (Yang et al. 2011; Ortiz-Llorente and Alvarez-Cobelas 2012; IPCC 2013; Selvam et al. 2014; Gondwe and Masamba 2014). More accurate estimations are necessary, since the accurate estimation of CH_4 emission from a lake is critical for extrapolating regional and global CH_4 budgets.

Currently, daily continuous measurements of CH_4 efflux for longer than one year using the eddy covariance (EC) method and the floating chamber method are rarely performed in lakes (Schubert et al. 2012; Xiao et al. 2014). Up to date, there were several studies using the EC method to measure CH_4 efflux in the lake, but these studies only reported with measurement durations shorter than one year (Edwards et al. 1994; Eugster et al. 2011; Schubert et al. 2012; Podgrajsek et al. 2014a, b; Sahlée et al. 2014; Xiao et al. 2014). Furthermore, the EC method used in the lake may also suffer from some limitations (Sahlée et al. 2014; Xiao et al. 2014), such as error propagation through the density corrections (Eugster et al. 2003; Miller et al. 2010; Lee and Massman 2011; Xiao et al. 2014), motion of the measuring platform due to unstable sediment below the measuring platform (Edson et al. 1998; Eugster et al. 2003; Blomquist et al. 2010; Norman et al. 2012), artificial density fluctuations from sensor self-heating of open-path sensors (Burba et al. 2008; Xiao et al. 2014) and expensive establishment and maintenance costs for EC systems (Eugster et al. 2011; Podgrajsek et al. 2014b). On the other hand, most available CH_4 effluxes were measured using the floating chamber method which usually calculated CH_4 efflux based on short-time measurements at weekly or monthly intervals (Xing et al. 2005, 2006; Rõõm et al. 2014). Thus, obtaining long-term CH_4 efflux measurements is relatively difficult in the lake.

Statistical models and process-based models have been developed to quantify temporal CH_4 effluxes at a fine scale, such as at hourly or daily levels (IPCC 2013; Zhu et al. 2013; Tan and Zhuang 2015). In previous studies, a few process-based models were developed to estimate CH_4 emissions in lakes (Kessler et al. 2012; Subin et al. 2012). Although process-based models are capable of modeling detailed physical and biological processes, they are more complicated and difficult in parameterizing the models as well as requiring too many inputs for running the models (Tang et al. 2010; Bridgham et al. 2013; Zhu et al. 2013; Tan and Zhuang 2015). In comparison, statistical models are simple and easy to run, and usually have good accuracy for specific lakes (Bastviken et al. 2004; Rasilo et al. 2015), which were usually based on short-term measurements of CH_4 effluxes to scale up to annual or longer temporal scales. Linear regression models and non-linear regression models have been used to simulate CH_4 efflux in various lakes in previous studies (Kettunen et al. 2000; Bastviken et al. 2004; Rõõm et al. 2014; Rasilo et al. 2015), but single linear and nonlinear regression models can not reflect the complex relationships between environmental factors and CH_4 effluxes including linear and nonlinear relationships (Schrier-Uijl et al. 2011; Rõõm et al. 2014). In particular, there was collinearity in a multiple linear model if the model included many environmental variables, especially climatic variables. Recently, RF models have been widely and successfully used in many fields, such as land-cover classification, soil science, and forest biomass (Peters et al. 2011; Dorigo et al. 2012; Taghizadeh-Mehrjardi et al. 2016; Sreenivas et al. 2016) because RF models can handle nonlinear relationships and highly collinear predictor variables as mentioned above (Prasad et al. 2006; Cutler et al. 2007). However, few studies have been reported to use the RF model to estimate CH_4 efflux in lakes or wetlands in general.

In this study, we measured daytime CH_4 efflux once every month using the floating chamber method from January 2011 to December 2014 in Poyang Lake. Then we built two types of statistical models between CH_4 efflux and climatic factors together with water level. Our objectives were to: (1) develop a "best" model for linear models and RF models, respectively; (2) compare the performance of the two optimum models and select the best model; (3) estimate and analyze the CH_4 efflux with the best model in Poyang Lake from 2011 to 2014.

Method

Study site

Poyang Lake lies in the south of China in Jiangxi province, the largest freshwater lake in China, and covers an area of 3283 km^2. The study site is located in the northern part of Poyang Lake in Xingzi county,

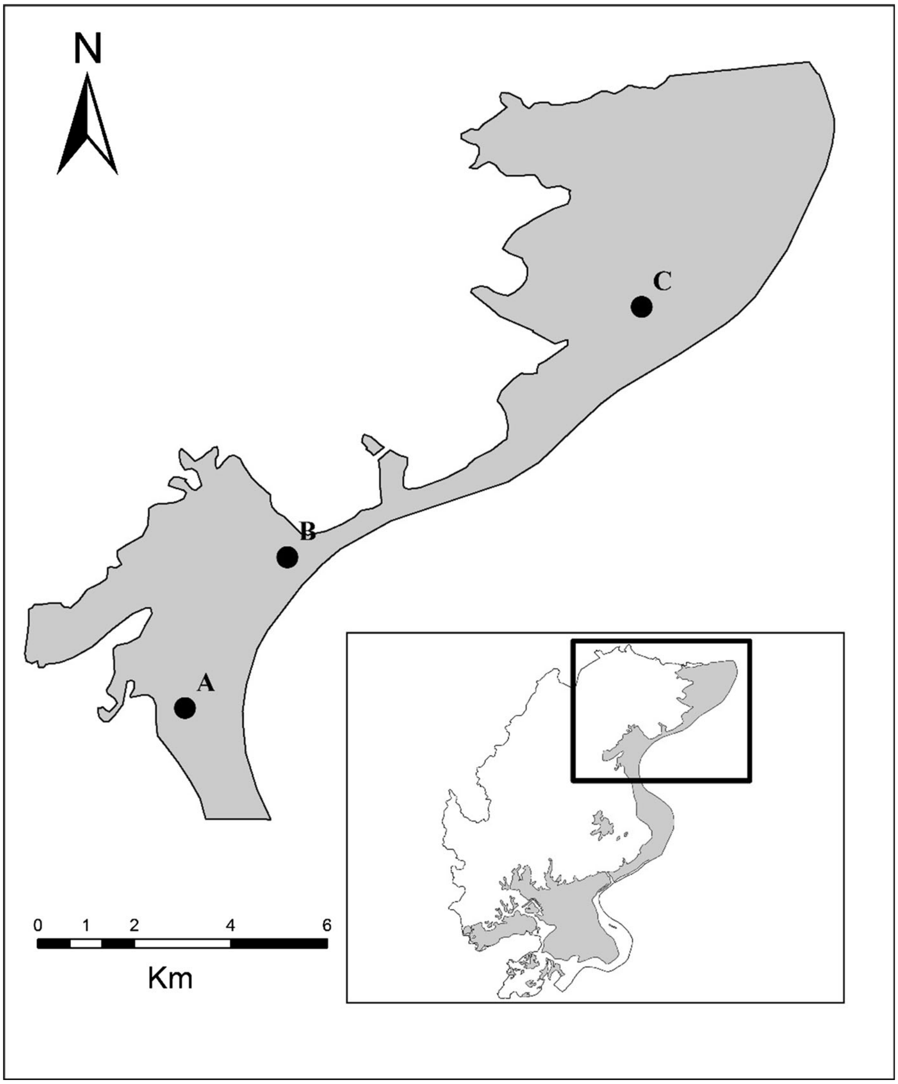

Fig. 1 Map of sampling sites in Poyang Lake

which is near the Poyang Lake laboratory for Wetland Ecosystem Research Station (operated by the Chinese Academy of Sciences) (Fig. 1). The region features a subtropical wet climate. Climatic records from the nearest weather station showed that the mean annual temperature was 18.1 °C, with a mean coldest (January) and warmest (July) monthly temperature of 5.2 and 29.8 °C from January 2011 to December 2014. The mean water level was 12.4 m during the 4-year period. The mean and maximum depths are 8 and 23 m in Poyang Lake, respectively. Vegetation in the lake is composed of macrophytes, including *Carex* sp. (*Carex cinerascens Kükenth.* and *Carexargyi Levl.et-Vant.*) and *Artemisia selengensis* in the hydrophyte zone, and the main submerged aquatic macrophytes, including *Ceratophyllum demersum*, *Potamogeton malaianus*, *Potamogeton crispus*, and *Hydrilla verticillata* (Wang et al. 2011).

Measurements of CH_4 efflux and climatic factors

Daytime CH_4 efflux was measured using a floating chamber and gas chromatography techniques at the air–water interface. The floating chamber was cylindrical, with an open surface of 0.031 m^2 and headspace volume of 6.28 L. Each sampling unit was made of an open-bottomed PVC chamber (20 cm in diameter and 100 cm in height) and equipped with styrofoam floats. A pressure relieve valve was installed on the top of the chamber to balance the pressure inside and outside of the chamber. A small fan remotely controlled by a wireless sensor was installed in each chamber to well mix the air inside the chamber during the sampling. The chamber was fully open at the bottom to account for the total gas emissions by ebullition and diffusive effluxes in previous studies (Bastviken et al. 2004; DelSontro et al. 2011; Schubert et al. 2012). More details on the chamber design, sampling method, and samples analysis are available in Liu et al. (2013).

Four chambers were simultaneously deployed from a small boat with the chambers about 10 m away from the boat to minimize the impact of the boat. Gas samples were taken every 20 min within a 1-h period. It should be noted that we took a gas sample (ambient concentration) immediately after the chamber was closed. The gas was extracted to a 12 ml evacuated glass vial by a 2 ml syringe needle with an air pump which enhanced the pressure in the vial to 3 bars. The gas samples were then transported immediately to a laboratory for CH_4 concentration analysis using gas chromatography equipped with a flame ionization detector (GC7890A, Agilent Technologies, Inc., Santa Clara, CA, USA) and a 2 m × 2 mm stainless steel column packed with 13XMS (80/100 mesh). We used silica gel and nitrogen (N_2) which ran at a flow rate of 30 mL min^{-1} as solid phase and the carrier gas. We calibrated the gas chromatograph (GC) for every four samples with a calibration gas of 2.03 ppm at 99.92% precision (China National Research Center for Certified Reference Materials, China). The oven and detector temperatures of the GC were set to 55 and 250 °C, respectively. CH_4 efflux was calculated using a linear regression model between CH_4 concentration versus time ($R^2 > 0.95$, $N = 4$).

In our previous study, we have explored the spatial variations of CH_4 efflux over the lake with 44 sampling locations (Liu et al. 2013). So we measured the CH_4 efflux in three sites to represent the average CH_4 efflux of the whole lake on the basis of our previous result in Poyang Lake. Samplings were carried out once every month from January 2011 to December 2014 at the three sites: Luoxingdeng, Mantianxing, and Huoyanshan. At each site, four chambers were placed approximately 10 m away from a small boat to minimize disturbance. The measurement was repeated from early morning to late afternoon, and we obtained a maximum of 6 measurements from each chamber every day. Additionally, we commenced three 24-h measurements to examine diel variations of CH_4 effluxes at the three sites from 24 to 25 July 2011 for summer, 5–6 September 2012 for fall, and 13–14 January 2013 for winter. These measurements were made every 2 h from 8:00 am to 8:00 am the next day local time, on each sampling day. Thus we collected a total amount of 12 efflux measurements for each chamber per 24 h. Finally we used the mean diurnal CH_4 efflux in our study.

We obtained the daily mean water level in the lake from the Xingzi Hydrological station. We used the atmospheric pressure, wind speed, and air temperature which were obtained from the Xingzi Meteorological station in our study. We also obtained the daily total (accumulated) solar radiation data from the Nanchang Radiation station from 1 January 2011 to 31 December 2014 in order to be consistent with other environmental factors data obtained in the lake. In addition, the daily total solar radiation was based on shortwave radiation.

Model development

We used eight environmental factors as our independent variables to build models for estimating the CH_4 efflux in Poyang Lake. These variables include average daytime air temperature (Tavg); maximum daytime air temperature (Tmax); minimum daytime air temperature (Tmin); mean daytime atmospheric pressure (Pa); average wind speed (wind); maximum daytime wind speed (windmax); average daily water level (WL); and daily total solar radiation (Rad). We chose these climatic factors due to data availability on a daily scale. In addition, previous studies have shown that single or multiple factors as mentioned above influenced CH_4 efflux dynamics in lakes (Bastviken et al. 2004; Rõõm et al. 2014; Rasilo et al. 2015; Tan and Zhuang 2015).

Linear model

Linear models were developed to estimate the linear relationships between daytime CH_4 effluxes and climatic variables. The CH_4 efflux was calculated as follows:

$$y = c + a_1 x_1 + a_2 x_2 + \ldots + a_n x_n \quad (1)$$

where y is CH_4 efflux (nmol m^{-2} s^{-1}); x_1, x_2, …, x_n are the predictor variables used to fit the model (i.e., environmental factors above mentioned); c is a constant parameter value; a_1, a_2, …, a_n are the linear regression coefficients. Specifically, we first built the full models with all the variables, and then used the backward stepwise selection to generate a model utilizing only significant variables.

Random forest model

The random forest model is a decision-tree-based model that works by random sub-sampling of the given data set. The RF model performs recursive partitioning of data sets, and makes no assumptions regarding the distribution of the input data (Breiman 2001; Schroeder et al. 2010; Peters et al. 2011; Dorigo et al. 2012). RF models can capture non-linear relationships between the response variable (CH_4 efflux in our study) and predictor variables (climatic variables in our study), and can deal with correlated variables while producing a low generalization error (Cutler et al. 2007; Taalab et al. 2013; Zhang et al. 2014). We gradually added climatic factors in the models to compare the performance of different RF models.

Model validation

We used the cross-validation method to validate our two model types. We used randomly 75% of the data to fit models and the remaining 25% of the data to validate our models. The process was repeated 50 times to make sure that different subsamples were used exactly once as the validation data. In our study, we randomly divided the 51 CH_4 efflux and climatic data sets into training data (75%) and testing data (25%). These two approaches of CH_4 efflux estimation were fitted with training data and evaluated with testing data.

Statistical analysis

Linear and RF models were fitted using the R statistical software (R version 3.0.2; Ihaka and Gentleman 1996). Specially, we used these models based on the relationships between daily efflux and climatic factors to simulate daily CH_4 efflux over a 4-year period. We used the coefficients of determination to compare the performance of different models of the same type. In addition, three error statistics were used to distinguish the difference between the observed and predicted CH_4 efflux, including mean absolute error (MAE), root mean-square error (RMSE), and the normalized root-mean-square error (NRMSE), which were calculated to assess model accuracy in different types of models except for the coefficients of determination (Zhang et al. 2014). Then, we used analysis of variance (ANOVA) followed by Turkey post hoc tests and a paired T test to assess the seasonal and annual differences of predicted CH_4 efflux in the best model, respectively. All statistical analyses were performed with the SPSS 17.0 statistical software (SPSS, Inc., Chicago, USA). Graphs were made with the Sigma Plot 11.0 program (Systat Software, Inc., San Jose, CA, USA).

Result

Model development

For the linear model, all the variables together explained the highest variance in the observed CH_4 effluxes in the L1 model (Table 1). On the basis of all candidate variables, we dropped variables by the backward stepwise selection to generate the optimum linear model (L2 model). The coefficient of determination in the L2 model was largely unaffected compared to the L1 model when we dropped variables (Table 1). The input variables in the L2 model were the mean daily air temperature and water level. The observed versus predicted effluxes in the L2 model differed greatly from the ideal 1:1 line. In particular, the predicted variable in the L2 model even included negative values (the opposite direction of observed efflux) (Fig. 2). The mean difference between predicted and observed effluxes was 44% in the L2 model, but the relatively greatest single underestimation and overestimation reached up to 115–223 and 110–353%, respectively.

For the RF model, the RF8 model was the best model for simulating CH_4 effluxes in the RF models in our study. The R1 model was not improved by adding the maximum of daytime air temperature and

Table 1 Results of different model types for estimating daily CH_4 efflux

Model	Equation	R^2	MAE (nmol m^{-2} s^{-1})	RMSE (nmol m^{-2} s^{-1})	NRMSE (%)
Linear models					
L1	y = 0.32 × Tavg + 0.53 × WL + 0.44 × Tmin − 1.01 × Tmax−0.027 × Pa + 0.0011 × Rad + 1.63 × wind − 0.5 × windmax + 22.62	0.69	4.37	5.32	24.31
L2	y = 0.31 × Tavg + 0.64 × WL-7.18	0.63	2.87	3.62	16.53
Random forest models					
RF1	Temperature + water level	0.9272	1.06	1.59	7.28
RF2	Temperature + water level + pressure	0.9258	1.1	1.61	7.34
RF3	Temperature + water level + pressure + Tmax	0.9273	1.11	1.59	7.27
RF4	Temperature + water level + pressure + Tmin	0.9331	1	1.53	6.97
RF5	Temperature + water level + pressure + Tmax + Tmin	0.9314	1.04	1.56	7.14
RF6	Temperature + Water level + Pressure + Windmax + Tmax + Tmin	0.9366	0.97	1.49	6.79
RF7	Temperature + water level + pressure + wind + Windmax + Tmax + Tmin	0.9372	0.96	1.48	6.75
RF8	Temperature + water level + pressure + radiation + Wind + Windmax + Tmax + Tmin	0.9385	0.96	1.46	6.69

MAE mean absolute error, *RMSE* root mean square error, *NRMSE* the normalized root mean square error, *Tavg* average daytime air temperature, *WL* average daily water level, *Tmin* minimum daytime air temperature, *Tmax* maximum daytime air temperature, *Pa* mean daytime atmospheric pressure, *Rad* daily total solar radiation, *windmax* maximum daytime wind speed

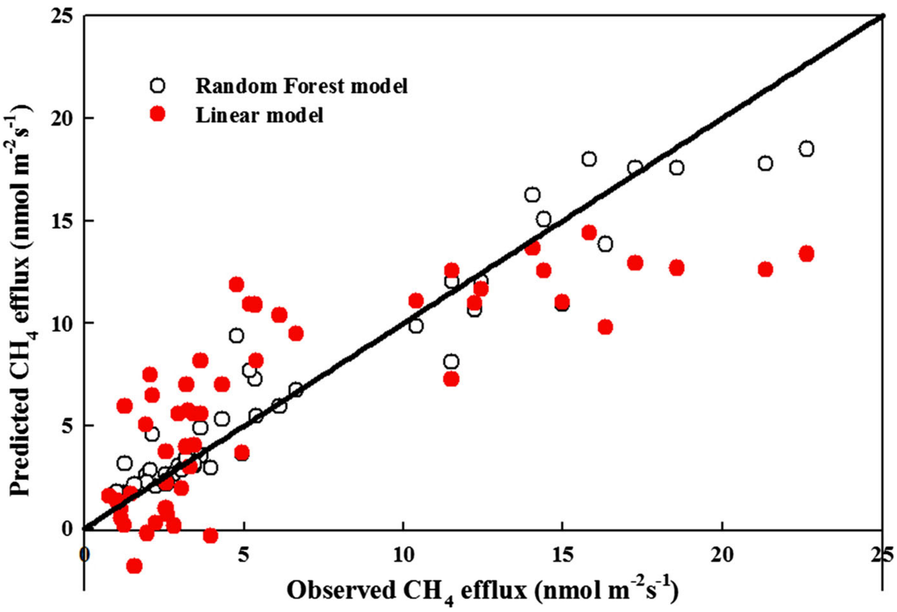

Fig. 2 Comparisons between the measured and the predicted CH_4 effluxes for two models. The *white*, *red* and *black circles* meant the measured against the predicted CH_4 effluxes for the RF1 and L2 models, respectively. The *black line* meant the ideal 1:1 regression line

atmospheric pressure to the model (RF2–RF3), but was improved significantly by adding the minimum of daytime air temperature (RF4 and RF5, respectively). The determination coefficient value of the RF4 model was highest in the RF1–RF5 models. In addition, the model was greatly improved by adding total solar radiation, the mean daytime wind speed, and the maximum of wind speed to the model simultaneously (RF6–RF8). The determination coefficient value of the R8 model was slightly higher than those in the RF7 model when we added total solar radiation and wind speed to the model, which showed the best performance of the model with the lowest errors (Table 1). We found that adding more variables, such as daily minimum and maximum air temperatures, daily atmospheric pressure, daily maximum and mean wind speeds, and the total solar radiation, barely improved the model's performance (Table 1). In order to compare the performance of two model types, we selected the RF1 model with the same variables as the L2 model. The observed versus predicted effluxes in the RF1 model were distributed closely around the 1:1 line. The difference between predicted and observed effluxes was, on average, 17% in the RF1 models (Fig. 2). The mean overestimation and underestimation of effluxes in the RF1 model were 33 and 14%, respectively.

Model comparison

The optimum model performance for two model types

We used the RF1 model to compare with the "best" linear model with the same variables. We found that adding more variables, such as daily minimum and maximum air temperatures, daily atmospheric pressure, daily maximum and mean wind speeds, and total solar radiation, barely improved the model's performance (Table 1).We compared the "best" model for two model types (L2 and RF1) by evaluating their performance using prediction accuracy that included MAE, RMSE and NRMSE except for the coefficients of determination. We found that the RF1 model was the best model for predicting CH_4 effluxes of the two model types. It had the overall highest coefficient of determination and the lowest model errors. The coefficient of determination, MAE, RMSE and NRMSE in the RF1 model were 0.93, 1.06 nmol m^{-2} s^{-1}, 1.59 nmol m^{-2} s^{-1}, and 7.28%, respectively (Table 1). For the L2 model, the coefficient of determination, MAE, RMSE and NRMSE values were 0.63, 2.87 nmol m^{-2} s^{-1}, 3.62 nmol m^{-2} s^{-1}, and 16.53%, respectively.

Model validation

In addition, we also used cross-validation to compare the two models (L2 and RF1) by evaluating their performance with our field data. We found that the RF1 model was the best model among the two models as we mentioned above. The RF1 model had the highest coefficient of determination and lowest model errors in 50 repeated times (Fig. 3). Specifically, the mean validated coefficient of determination, MAE, RMSE and NRMSE in the RF1 model were 0.86 (range 0.74–0.93), 1.51 nmol m^{-2} s^{-1} (range 1.19–.03), 2.38 nmol m^{-2} s^{-1} (range 1.69–3.26) and 10.89% (range 7.77–14.91%), respectively. For the L2 model, the mean validated coefficient of determination, MAE, RMSE and NRMSE were 0.68 (range 0.56–0.74), 2.86 nmol m^{-2} s^{-1} (range 2.62–3.21), 3.59 nmol m^{-2} s^{-1} (range 3.27–4.24) and 16.46% (range 14.94–19.37%), respectively.

Model application

We estimated the daily CH_4 efflux, seasonal and annual CH_4 emissions of the Poyang Lake from 2011 to 2014 based on our best model, the RF1 model including the mean daily air temperature and water level. We found that rapid fluctuation of daily CH_4 effluxes often occurred in January, July, around May and September when the mean daily air temperature or the average daily water level sharply changed at the daily scale (Fig. 4). At the seasonal scale, we found that the mean CH_4 emission in the summer was significantly higher than those in other seasons (Fig. 5), with mean values of 0.097, 0.28, 0.11, and 0.045 mmol m^{-2} day^{-1} in the spring, in the summer, in the autumn and in the winter, respectively. In our study, the mean annual CH_4 emission was 3.13 g m^{-2} $year^{-1}$ with annual emissions of 2.63, 3.54, 3.19, and 3.17 g m^{-2} $year^{-1}$ in 2011, in 2012, in 2013 and in 2014, respectively (Fig. 6). There was a significant difference in the annual CH_4 emission between 2012 and 2011 ($p < 0.05$), whereas the

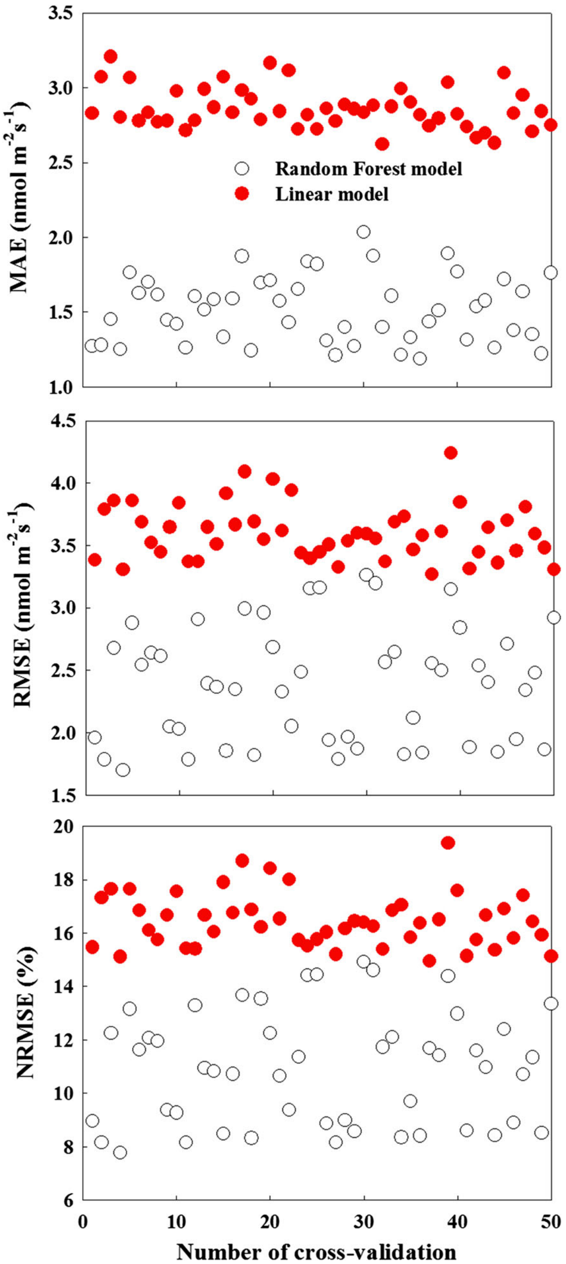

Fig. 3 Cross-validation statistics of two models for daily CH_4 effluxes. The *white*, *red* and *black circles* meant the errors in 50 cross-validations for the RF1 and L2 models, respectively. The *different letters* in the figure panels meant the mean absolute error (**a**), root mean square error (**b**), and the normalized root mean square error (**c**) in 50 cross-validations for two models

differences were not significant for the annual CH_4 emission among the other years ($p > 0.05$). The mean annual CH_4 emission in 2014 was 20.73% higher than that in 2011.

Discussion

Model performance

In our study, we selected the RF1 model for predicting long-term CH_4 effluxes in Poyang Lake. We found that the RF1 model had the highest coefficient of determination and the lowest model-performance errors (Table 1), yet the coefficient of determination and errors for linear model showed far poorer model performance. Our result showed that the RMSE in the RF1 model with the value of 1.59 nmol m^{-2} s^{-1} was less than half the standard deviation of the observed data with the value of 2.98 nmol m^{-2} s^{-1}, which has been considered to be appropriate for model evaluation (Singh et al. 2005). It should be noted that it was difficult to compare our results with other CH_4 efflux estimations using statistical models, because most studies were conducted in wetlands and few studies in lakes. Dengel et al. (2013) found that the coefficient of determination for an artificial neural network model, which gap-filled EC CH_4 effluxes, ranged from 0.37 to 0.94 in six high latitude wetlands. An earlier study also found that the artificial neural network model, which can predict the underlying nonlinear relationship between independent variables and dependent variables, showed better performance for simulating monthly CH_4 emission from 1990 to 2009 in wetlands in northern high latitudes (Zhu et al. 2013), where the RSME was approximately 8 times than that in the RF1 model in our study. On the other hand, a number of studies have supported the fact that the RF model has strong predictive potential for other research topics (Whitcomb et al. 2009; Peters et al. 2011; Vincenzi et al. 2011; Zhang et al. 2014; Shi et al. 2015). In particular, previous studies have confirmed that the RF model outperformed other statistical models, including linear model, artificial neural networks, nonlinear model, such as the exponential model in simulating annual precipitation in Mainland China (Prasad et al. 2006; Taalab et al. 2013; Zhu et al. 2013; Hengl et al. 2015; Shi et al. 2015). However, there were still some prediction errors and some scatters around the 1:1 line in the RF1 model (Table 1; Fig. 2). High effluxes were under-predicted while low effluxes were over-predicted in the RF model in our study. Although most studies showed that the RF model can provide accurate predictions without over-fitting the data (Prasad et al. 2006; Peters et al. 2011; Taalab et al. 2013; Zhang

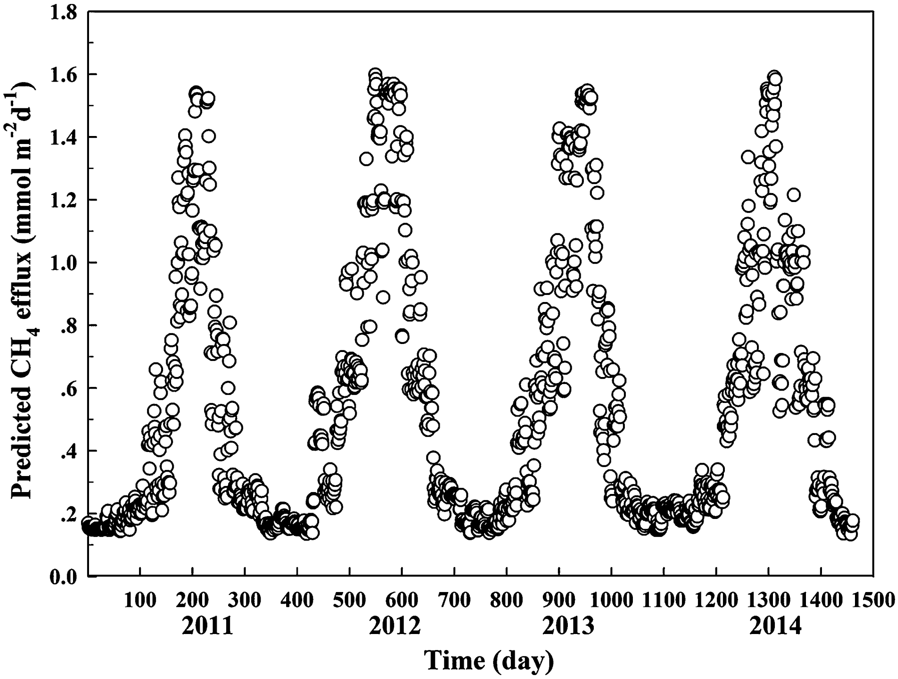

Fig. 4 Predicted daily CH_4 efflux variations in the optimum Random forest model over a 4-year period. The *x*-axis data meant time sequence from the 1 January 2011 to 31 December 2014

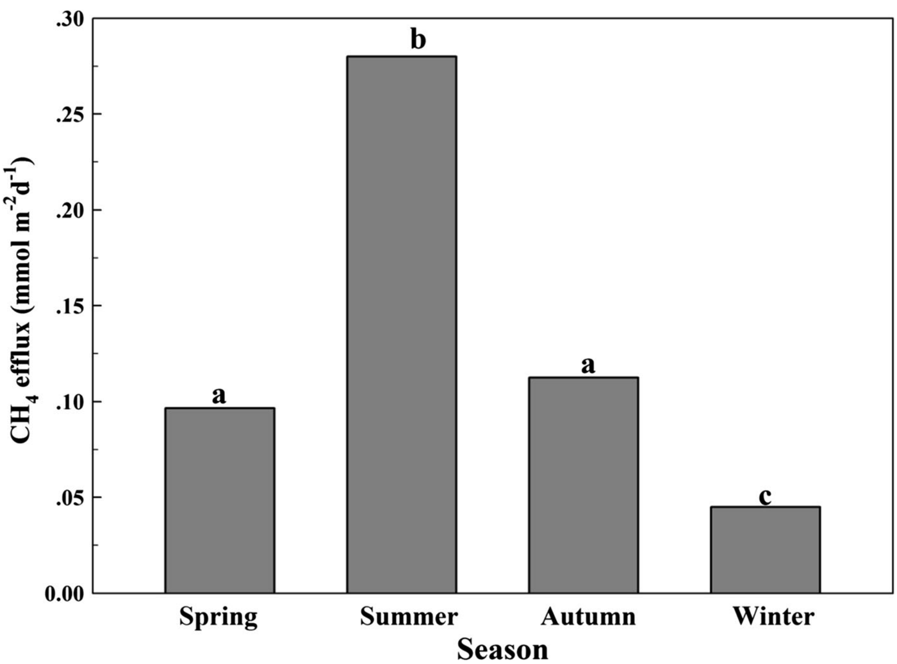

Fig. 5 Mean seasonal CH_4 effluxes in the RF1 model over a 4-year period. *Different letters* at the top of the *bar*s meant significant differences among the four seasons

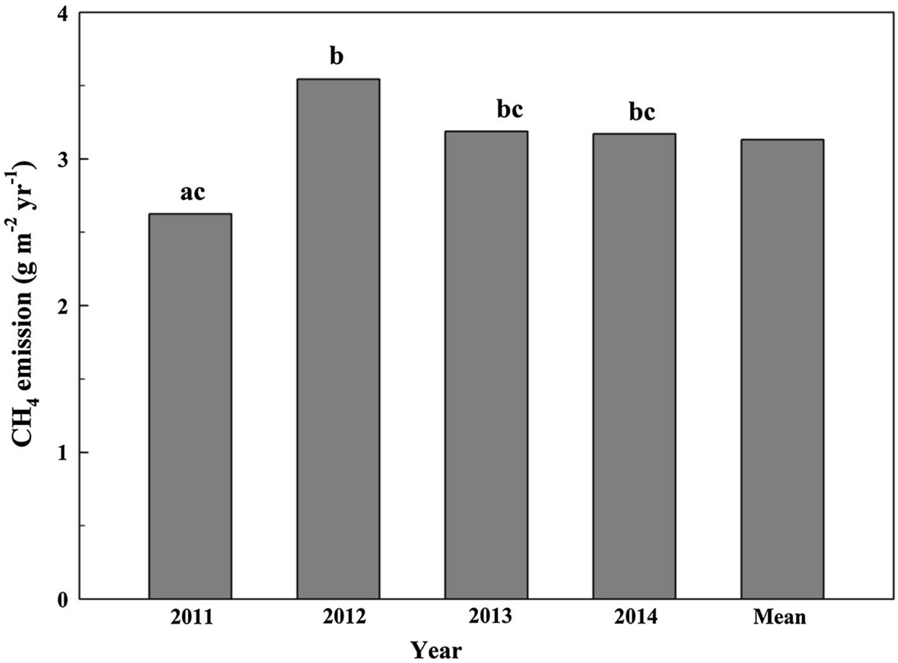

Fig. 6 Annual CH_4 emissions in the RF1 model during a 4-year period. *Different letters* at the top of the bars meant significant differences among the four years

et al. 2014; Shi et al. 2015), some studies indicated that the RF model over-fitted for soil properties mapping and Ruditapes philippinarum yield in previous studies (Statnikov et al. 2008; Vincenzi et al. 2011; Hengl et al. 2015). This may be due to the occurrence of noisy datasets which resulted in the smoothing effect of regression in the gap-filling processes (Statnikov et al. 2008; Hengl et al. 2015). However, in order to avoid the over-fitting impact, we used the cross-validation with independent datasets. Our results confirmed that the RF model is consistently outperforming the regression models in modeling CH_4 emissions in Poyang Lake. In the current study, we found that the RF model is also superior to other models in estimating CH_4 efflux in Poyang Lake in southern China.

Comparison with previous CH_4 efflux

Cumulative annual CH_4 efflux using the RF model in Poyang Lake in our study was relatively low in comparison with other lakes based on measurements in the world. The mean cumulative annual CH_4 efflux in Poyang Lake was 3.13 gCH_4 m^{-2} $year^{-1}$, which was significantly lower than the mean annual CH_4 effluxes of 3.61 gCH_4 m^{-2} $year^{-1}$ reported by Bastviken et al. (2011) and 64.30 gCH_4 m^{-2} $year^{-1}$ reported by Ortiz-Llorente and Alvarez-Cobelas (2012) in global lakes. In particular, CH_4 emissions in our study also were much lower than many previous reports from similar regions. For example, Xing et al. (2005, 2006) found mean annual emissions of 8.5 gCH_4 m^{-2} $year^{-1}$ in Donghu Lake and 7.7 gCH_4 m^{-2} $year^{-1}$ in Biandantang Lake in China. The lower CH_4 emissions in our study may be attributed to the low concentration of carbon substrates in the water and sediments in Poyang Lake. The dissolved organic carbon (DOC) concentration in Poyang Lake was merely 3.3 mgL^{-1} which was much lower than that of 5.8 mgL^{-1} in Biandantang Lake and 7.4 mgL^{-1} in Donghu Lake, two subtropical lakes in China (Xing et al. 2005, 2006). Additionally, we found that the average organic carbon content in the sediments in Poyang Lake was 0.89% which was much lower than that of 30.76% averaged over 5 temperate lakes that gave an average CH_4 emission rate of 34.2 gCH_4 m^{-2} $year^{-1}$ in the Netherlands (Schrier-Uijl et al. 2011).

In addition, CH_4 effluxes demonstrated distinct seasonal variations in Poyang Lake, which was consistent with other studies (Casper et al. 2000; Huttunen et al. 2003; Duan et al. 2005; Xing et al. 2005, 2006; Palma-Silva et al. 2013). The CH_4 effluxes showed a sharp increase in the summer and fell to lower values for the rest of the year, which was in agreement with previous studies (Xing et al. 2005, 2006; Duan et al. 2005; Ortiz-Llorente and Alvarez-Cobelas 2012). The CH_4 effluxes ranged from 1.00 to 22.65 nmol m^2 s^{-1}, which fell within the reported range of CH_4 effluxes from lakes in other regions (−1388.89 to 4340.28 nmol m^2 s^{-1}) (Ortiz-Llorente and Alvarez-Cobelas 2012). This may be attributed to high temperature and higher substrate availability during the summer. Previous studies have found that warmer temperatures increased CH_4 effluxes in the lake (Bastviken et al. 2008; Marinho et al. 2009; Palma-Silva et al. 2013; Rõõm et al. 2014). On the other hand, earlier studies have found that the amount of labile organic matter, including allochthonous inputs of terrestrial organic matter during the summer flooding and autochthonous production within-lake by phytoplankton and benthic algae in the summer, was very high (Crump et al. 2003; Xing et al. 2005, 2006; Bade et al. 2007). In particular, previous studies have also shown that fresh organic carbon from dead algae stimulates CH_4 emission in lakes (Huttunen et al. 2002; Xing et al. 2005) because the degradation of dead alga and algal exudates, such as the methylated compounds, are the precursors for CH_4 production (Ferrón et al. 2012; Xiao et al. 2015).

Variable importance

In the current study, we found that temperature and water level were critical indicators by assessing the importance of various predictor factors on CH_4 efflux in lakes (Fig. 7). The assessment of variable importance in the RF8 model showed that the mean daytime air temperature, the minimum daytime air temperature, the maximum daytime air temperature and the mean daily water level contributed the most to modeling CH_4 efflux, followed by total solar radiation, the maximum wind speed and the mean wind speed. Previous studies have reported a strong positive relationship between CH_4 effluxes and sediment temperature (Xing et al. 2005; Duan et al. 2005; Wik et al. 2014). This could be due to the fact that high temperatures stimulate the growth of methanogenic communities resulting in methane production (Nozhevnikova et al. 2007; Rooney-Varga et al. 2007; Duc

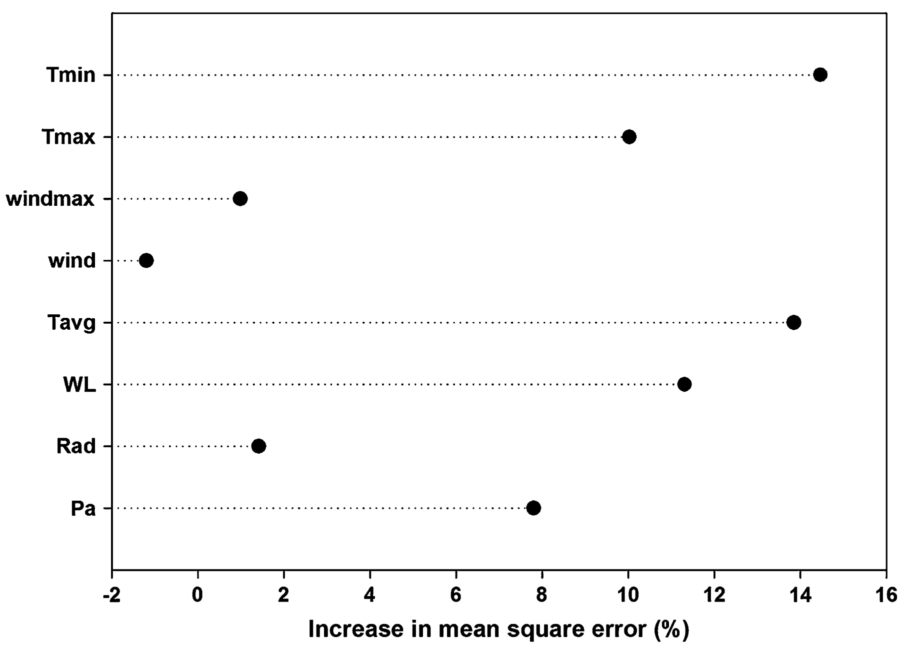

Fig. 7 Relative importance of predictor variables for CH_4 effluxes in the RF1 model

et al. 2010). However, few studies demonstrated that the minimum daytime air temperature and the maximum daytime air temperature influenced CH_4 effluxes on the basis of temporal measurements in lakes. This may be explained by the low variation of sediment temperature in lakes. In addition, earlier studies also found that more CH_4 was emitted at high water levels and less CH_4 was released at low water levels in freshwater systems (Bergström et al. 2007; Chen et al. 2011; Yang et al. 2012, 2013; Xiao et al. 2013), which was due to heavy rain and the resulting runoff leading to large amounts of external nutrients and CH_4-rich inputs (Chen et al. 2007; Hu et al. 2007). Moreover, most studies found that wind speed had a major influence on diel CH_4 efflux (Wanninkhof 1992; Palma-Silva et al. 2013; Xiao et al. 2013) because high wind speed mechanically induced turbulence that physically influenced the efficiency of gas transfer at the air–water interface and brought CH_4-rich water from the bottom to the surface in lakes, but few study reported a significant relationship between CH_4 efflux and wind speed at longer scales.

Conclusion

In the present study we attempted to simulate CH_4 efflux on the basis of monthly filed data using several statistical models over a 4-year period in the sediment of Poyang Lake. This paper used linear and random forest models to predict temporal variations of CH_4 efflux from 2011 to 2014 in Poyang Lake. We first built several models for two model types, respectively. Then we compared these models and developed an optimum model for two model types through the coefficient of determination and the significant variables in the model, respectively. We also compared the two optimum models and selected the optimum model by the coefficient of determination, three well-known error statistics and cross-validation results. Results showed that the random forest model was the best model to predict daily CH_4 efflux which the coefficient of determination was the highest and error statistics was the lowest. We also found that air temperature and water level was the most important variables for simulating CH_4 efflux in lakes. We used the best model to predict daily CH_4 efflux from 2011 to 2014 in Poyang Lake. The distinct seasonal variations of predicted CH_4 effluxes in our study were in accordance with other studies. Our study suggested that random forest model can be used to estimate long-term CH_4 effluxes and emissions in other lake in the world. The finding may have important implications in managing CH_4 emissions in lakes by considering long-term continuous variations of CH_4 efflux.

Acknowledgements This research was supported by the National Basic Research Program of China (973 Program, 2012CB417103), Assessment and Valuation of Ecosystem Services in Qinghai province, China (No.2013-N-556). We gratefully acknowledge the Poyang Lake Laboratory Wetland Ecosystem Research, CAS for permission to access the study site and assistance with field work.

References

Bade DL, Carpenter SR, Cole J, Pace M, Kritzberg E, Van de Bogert MC, Cory RM, McKnight DM (2007) Sources and fates of dissolved organic carbon in lakes as determined by whole-lake carbon isotope additions. Biogeochemistry 84:115–129

Bastviken D, Cole J, Pace M, Tranvik L (2004) Methane emissions from lakes: dependence of lake characteristics, two regional assessments, and a global estimate. Glob Biogeochem Cycle 18:1–12

Bastviken D, Cole J, Pace M, Van de Bogert MC (2008) Fates of methane from different lake habitats: connecting whole-lake budgets and CH_4 emissions. J Geophys Res 113:G02024

Bastviken D, Tranvik LJ, Downing JA, Crill PM, Enrich-Prast A (2011) Freshwater methane emissions offset the continental carbon sink. Science 331:50

Bergström I, Mäkelä S, Kankaala P, Kortelainen P (2007) Methane efflux from littoral vegetation stands of southern boreal lakes: an upscaled regional estimate. Atmos Environ 41:339–351

Blomquist BW, Huebert BJ, Fairall CW, Faloona IC (2010) Determining the sea–air flux of dimethylsulfide by eddy correlation using mass spectrometry. Atmos Meas Technol 3:1–20

Breiman L (2001) Random forests. Mach Learn 45:5–32

Bridgham S, Cadillo-Quiroz H, Keller JK, Zhuang QL (2013) Methane emissions from wetlands: biogeochemical, microbial, and modeling perspectives from local to global scales. Glob Change Biol 19:1325–1346

Burba GG, McDermitt DK, Grelle A, Anderson DJ, Xu L (2008) Addressing the influence of instrument surface heat exchange on the measurements of CO_2 flux from open-path gas analyzers. Global Change Biol 14:1854–1876

Casper P, Maberly SC, Hall GH, Finlay BJ (2000) Fluxes of methane and carbon dioxide from a small productive lake to the atmosphere. Biogeochemistry 49:1–19

Chen YG, Bai XH, Li XH, Hu ZX, Liu WL (2007) Primary study of the methane flux on the water-air interface of eight lakes in winter, China. J Lake Sci 19:11–17

Chen H, Yuan XZ, Chen ZL, Wu YY, Liu XS, Zhu D, Wu N, Zhu QA, Peng CH, Li WZ (2011) Methane emissions from the surface of the Three Gorges Reservoir. J Geophys Res 116:1–5

Crump BC, Kling GW, Bahr M, Hobbie JE (2003) Bacterioplankton community shifts in an Arctic Lake correlate with seasonal changes in organic matter source. Appl Environ Microbiol 69:2253–2268

Cutler DR, Edwards TC, Beard KH, Cutler A, Hess KT, Gibson J, Lawler JJ (2007) Random forests for classification in ecology. Ecology 88:2783–2792

DelSontro T, Kunz MJ, Kempter T, Wuest A, Wehrli B, Senn DB (2011) Spatial heterogeneity of methane ebullition in a large tropical reservoir. Environ Sci Technol 45:9866–9873

Dengel S, Zona D, Sachs T, Aurela M, Jammet M, Parmentier FJW, Oechel W, Vesala T (2013) Testing the applicability of neural networks as a gap-filling method using CH_4 flux data from high latitude wetlands. Biogeosciences 10:8185–8200

Dorigo T, Lucieer A, Podobnikar T, Čarni A (2012) Mapping invasive Fallopia japonica by combined spectral, spatial, and temporal analysis of digital orthophotos. Int J Appl Earth Obs 19:185–195

Duan XN, Wang XK, Mu YJ, Ouyang ZY (2005) Seasonal and diurnal variations in methane emissions from Wuliangsu Lake in arid regions of China. Atmos Environ 39:4479–4487

Duc NT, Crill P, Bastviken D (2010) Implications of temperature and sediment characteristics on methane formation and oxidation in lake sediments. Biogeochemistry 100:185–196

Edson JB, Hinton AA, Prada KE, Hare JE, Fairall CW (1998) Direct covariance flux estimates from mobile platforms at sea. J Atmos Oceanic Technol 15:547–562

Edwards GC, Neumann HH, den Hartog G, Thurtell GW, Kidd GE (1994) Eddy correlation measurements of methane fluxes using a tunable diode laser at the Kinosheo Lake tower site during the Northern Wetlands Study (NOWES). J Geophys Res 99:1511–1517

Eugster W, Kling G, Jonas T, McFadden JP, Wüest A, MacIntyre S, Chapin FS III (2003) CO_2 exchange between air and water in an Arctic Alaskan and midlatitude Swiss lake: importance of convective mixing. J Geophys Res 108(D12):4362

Eugster W, Delsontro T, Sobek S (2011) Eddy covariance flux measurements confirm extreme CH_4 emissions from a Swiss hydropower reservoir and resolve their short-term variability. Biogeosci Discuss 8:5019–5055

Ferrón S, Ho DT, Johnson ZI, Huntley ME (2012) Air–water fluxes of N_2O and CH_4 during microalgae (Staurosira sp.) cultivation in an open raceway pond. Environ Sci Technol 46:10842–10848

Gondwe MJ, Masamba WRL (2014) Spatial and temporal dynamics of diffusive methane emissions in the Okavango Delta, northern Botswana, Africa. Wetlands Ecol Manage 22:63–78

Hengl T, Heuvelink GBM, Kempen B, Leenaars JGB, Walsh MG, Shepherd KD, Sila A, MacMillan RA, de Jesus JM, Tamene L, Tondoh JE (2015) Mapping soil properties of Africa at 250 m resolution: random forests significantly improve current predictions. PLoS ONE 10:e0125814

Hu Q, Feng S, Guo H, Chen G, Jiang T (2007) Interactions of the Yangtze river flow and hydrologic processes of the Poyang Lake, China. J Hydrol 347:90–100

Huttunen JT, Väisänen TS, Hellsten SK, Heikkinen M, Nykänen H, Jungner H, Niskanen A, Virtanen MO, Lindqvist OV, Nenonen OS, Martikainen PJ (2002) Fluxes of CH_4, CO_2, and N_2O in hydroelectric reservoirs Lokka and Porttipahta in the northern boreal zone in Finland. Glob Biogeochem Cycle 16:1–17

Huttunen JT, Alm J, Liikanen A, Juutinen S, Larmola T, Hammar T, Silvola L, Martikainen PJ (2003) Fluxes of methane, carbon dioxide and nitrous oxide in boreal lakes and potential anthropogenic effects on the aquatic greenhouse gas emissions. Chemosphere 52:609–621

Ihaka R, Gentleman R (1996) R: a language for data analysis and graphics. J Comput Graph Stat 5:299–314

IPCC (2013) Climate Change 2013: The Physical Science Basis: Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. In: Qin D, Plattner GK, Tignor M, Allen SK, Boschung J, Nauels A, Xia Y, Bex V, Midgley PM (eds) Stocker TF. Cambridge University Press, Cambridge, p 507

Kessler MA, Plug LJ, Anthony KMW (2012) Simulating the decadal- to millennial-scale dynamics of morphology and sequestered carbon mobilization of two thermokarst lakes in NW Alaska. J Geophys Res 117:G00M06

Kettunen A, Kaitala V, Alm J, Silvola J, Nykanen H, Martikainen PJ (2000) Predicting variations in methane emissions from boreal peatlands through regression models. Boreal Environ Res 5:115–132

Lee XH, Massman WJ (2011) A perspective on thirty years of the Webb, Pearman, and Leuning density correction. Boundary-Layer Meteorol 139:37–59

Liu LX, Xu M, Lin M, Zhang X (2013) Spatial variability of greenhouse gas effluxes and their controlling factors in the Poyang Lake in China. Pol J Environ Stud 22:749–758

Marinho CC, Palma-Silva C, Albertoni EF, Trindade CR, Esteves FA (2009) Seasonal dynamics of methane in the water column of two subtropical lakes differing in trophic status. Braz J Biol 69:631–637

Miller SD, Marandino C, Saltzman ES (2010) Ship-based measurement of air-sea CO_2 exchange by eddy covariance. J Geophys Res 115:D02304

Norman M, Rutgersson A, Sørensen LL, Sahlée E (2012) Methods for estimating air-sea fluxes of CO_2 using high-frequency measurements. Boundary-Layer Meteorol 144:379–400

Nozhevnikova AN, Nekrasova V, Ammann A, Zehnder AJB, Wehrli B, Holliger C (2007) Influence of temperature and high acetate concentrations on methanogenensis in lake sediment slurries. FEMS Microbiol Ecol 62:336–344

Ortiz-Llorente MJ, Alvarez-Cobelas M (2012) Comparison of biogenic methane emissions from unmanaged estuaries, lakes, oceans, rivers and wetlands. Atmos Environ 59:328–337

Palma-Silva C, Marinho CC, Albertoni EF, Giacomini LB, Barros MPF, Furlanetto LM, Trindade CRT, de Assis Esteves F (2013) Methane emissions in two small shallow neotropical lakes: the role of temperature and trophic level. Atmos Environ 81:373–379

Peters J, Van Coillie F, Westra T, De Wulf R (2011) Synergy of very high resolution optical and radar data for object-based olive grove mapping. Int J Geogr Inf Sci 25:971–989

Podgrajsek E, Sahlée E, Rutgersson A (2014a) Diurnal cycle of lake methane flux. J Geophys Res Biogeosci 119:236–248

Podgrajsek E, Sahlée E, Bastviken D, Holst J, Lindroth A, Tranvik L, Rutgersson A (2014b) Comparison of floating chamber and eddy covariance measurements of lake greenhouse gas fluxes. Biogeosciences 11:4225–4233

Prasad AM, Iverson LR, Liaw A (2006) Newer classification and regression tree techniques: bagging and random forests for ecological prediction. Ecosystems 9:181–199

Rasilo T, Prairie YT, del Giorgio PA (2015) Large-scale patterns in summer diffusive CH_4 fluxes across boreal lakes, and contribution to diffusive C emissions. Glob Change Biol 21:1124–1139

Rõõm EI, Nõges P, Feldmann T, Tuvikene L, Kisand A, Teearu H, Nõges T (2014) Years are not brothers: two-year comparison of greenhouse gas fluxes in large shallow Lake Võrtsjärv, Estonia. J Hydrol 519:1594–1606

Rooney-Varga JN, Giewat MW, Duddleston KN, Chanton JP, Hines ME (2007) Links between archaeal community structure, vegetation type and methanogenic pathway in Alaskan peatlands. FEMS Microbiol Ecol 60:240–251

Sahlée E, Rutgersson A, Podgrajsek E, Bergström H (2014) Influence from surrounding land on the turbulence measurements above a lake. Boundary-Layer Meteorol 150:235–258

Schrier-Uijl AP, Veraart AJ, Leffelaar PA, Berendse F, Veenendaal EM (2011) Release of CO_2 and CH_4 from lakes and drainage ditches in temperate wetlands. Biogeochemistry 102:265–279

Schroeder R, Rawlins MA, McDonald KC, Podest E, Zimmermann R, Kueppers M (2010) Satellite microwave remote sensing of North Eurasian inundation dynamics: development of coarse-resolution products and comparison with high-resolution synthetic aperture radar data. Environ Res Lett 5:015003

Schubert CJ, Diem T, Eugster W (2012) Methane emissions from a small wind shielded lake determined by eddy covariance, flux chambers, anchored funnels and boundary model calculations: a comparison. Environ Sci Technol 46:4515–4522

Selvam BP, Natchimuthu S, Arunachalam L, Bastviken D (2014) Methane and carbon dioxide emissions from inland waters in India-implications for large scale greenhouse gas balances. Glob Change Biol 20:3397–3407

Shi YL, Song L, Xia Z, Lin YR, Myneni RB, Choi S, Wang L, Ni XL, Lao CL, Yang FK (2015) Mapping annual precipitation across Mainland China in the period 2001–2010 from TRMM3B43 product using spatial downscaling approach. Remote Sens 7:5849–5878

Singh J, Knapp HV, Arnold JG, Demissie M (2005) Hydrologic modeling of the Iroquois River watershed using HSPF and SWAT. J Am Water Resour Assoc 41:343–360

Sreenivas K, Dadhwal VK, Kumar S, Sri Harsha G, Mitran T, Sujatha G, Janaki Rama Suresh G, Fyzee MA, Ravisankar T (2016) Digital mapping of soil organic and inorganic carbon status in India. Geoderma 269:160–173

Statnikov A, Wang L, Aliferis C (2008) A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. BMC Bioinformatics 9:319

Subin ZM, Riley WJ, Mironov D (2012) An improved lake model for climate simulations: model structure, evaluation, and sensitivity analyses in CESM1. J Adv Model Earth Syst 4:M02001

Taalab KP, Corstanje R, Creamer R, Whelan MJ (2013) Modelling soil bulk density at the landscape scale and its contributions to C stock uncertainty. Biogeosciences 10:4691–4704

Taghizadeh-Mehrjardi R, Nabiollahi K, Kerry R (2016) Digital mapping of soil organic carbon at multiple depths using different data mining techniques in Baneh region, Iran. Geoderma 266:98–110

Tan ZL, Zhuang QL (2015) Arctic lakes are continuous methane sources to the atmosphere under warming conditions. Environ Res Lett 10:054016

Tang J, Zhuang Q, Shannon RD, White JR (2010) Quantifying wetland methane emissions with process-based models of different complexities. Biogeosciences 7:3817–3837

Vincenzi S, Zucchetta M, Franzoi P, Pellizzato M, Pranovi F, De Leo GA, Torricelli P (2011) Application of a Random Forest algorithm to predict spatial distribution of the potential yield of Ruditapes philippinarum in the Venice lagoon, Italy. Ecol Model 222:1471–1478

Wang YY, Yu XB, Li WH, Xu J, Chen YW, Fan N (2011) Potential influence of water level changes on energy flows in a lake food web. Chin Sci Bull 56:2794–2802

Wanninkhof R (1992) Relationship between wind speed and gas exchange over the ocean. J Geophys Res 97:7373–7382

Whitcomb J, Moghaddam M, McDonald K, Kellndorfer J, Podest E (2009) Mapping vegetated wetlands of Alaska using L-band radar satellite imagery. Can J Remote Sens 35:54–72

Wik M, Thornton BF, Bastviken D, Maclntyre S, Varner RK, Crill PM (2014) Energy input is primary controller of methane bubbling in subarctic lakes. Geophys Res Lett 41:555–560

Xiao SB, Wang YC, Liu DF, Yang ZJ, Lei D, Zhang C (2013) Diel and seasonal variation of methane and carbon dioxide fluxes at Site Guojiaba, the Three Gorges Reservoir. J Environ Sci 25:2065–2071

Xiao W, Liu SD, Li HC, Xiao QT, Wang W, Hu ZH, Hu C, Gao YQ, Shen J, Zhao XY, Zhang M, Lee XH (2014) A flux-gradient system for simultaneous measurement of the CH_4, CO_2, and H_2O fluxes at a lake–air interface. Environ Sci Technol 48:14490–14498

Xiao SB, Liu WG, Yang H, Liu D, Wang Y, Peng F (2015) Extreme methane bubbling emissions from a subtropical shallow eutrophic pond. Austin Biom and Biostat 1:1–6

Xing YP, Xie P, Yang H, Ni LY, Wang YS, Rong KW (2005) Methane and carbon dioxide fluxes from a shallow hypereutrophic subtropical lake in China. Atmos Environ 39:5532–5540

Xing YP, Xie P, Yang H, Wu AP, Ni LY (2006) The change of gaseous carbon fluxes following the switch of dominant producers from macrophytes to algae in a shallow subtropical lake of China. Atmos Environ 40:8034–8043

Yang H, Xie P, Ni LY, Flower RJ (2011) Underestimation of CH_4 emission from freshwater lakes in China. Environ Sci Technol 45:4203–4204

Yang L, Lu F, Wang XK, Duan XN, Song WZ, Sun BF, Chen S, Zhang QQ, Hou PQ, Zheng FX, Zhang Y, Zhou XP, Zhou YJ, Ouyang ZY (2012) Surface methane emissions from different land use types during various water levels in three major drawdown areas of the Three Gorges Reservoir. J Geophys Res 117:1–11

Yang L, Lu F, Wang XK, Duan XN, Song WZ, Sun BF, Zhang QQ, Zhou YJ (2013) Spatial and seasonal variability of diffusive methane emissions from the Three Gorges Reservoir. J Geophys Res 118:471–481

Zhang J, Huang S, Hogg EH, Lieffers V, Qin Y, He F (2014) Estimating spatial variation in Alberta forest biomass from a combination of forest inventory and remote sensing data. Biogeosciences 11:2793–2808

Zhu XD, Zhuang QL, Qin ZC, Glagolev MH, Song LL (2013) Estimating wetland methane emissions from the northern high latitudes from 1990 to 2009 using artificial neural networks. Glob Biogeochem Cycle 27:1–13