Random forest-based understanding and predicting of the impacts of anthropogenic nutrient inputs on the water quality of a tropical lagoon

Xin Fang\(^1\,*\), Xiaoyan Li\(^1\), Yifei Zhang\(^1\), Yuan Zhao\(^3\), Jian Qian\(^1\), Chunling Hao\(^1\), Jiaqi Zhou\(^1\) and Yifan Wu\(^1\)

\(^1\) Second Institute of Oceanography, Ministry of Natural Resources, No. 36 BaochubeiLu Rd., Hangzhou, Zhejiang 310012, People’s Republic of China

\(^2\) School of Geography and Ocean Science, Nanjing University, No. 163 Xianlin Rd., Nanjing, Jiangsu 210093, People’s Republic of China

\(^3\) Ecological Environment Department of Hainan Province, No. 9 Meixian Rd., Haikou, Hainan 570203, People’s Republic of China

* Author to whom any correspondence should be addressed.

E-mail: fangx@sio.org.cn

Keywords: inorganic nitrogen, spatial distribution, anthropogenic activities, random forest, water quality prediction, management measures

Supplementary material for this article is available online

Abstract

Seawater quality degradation is caused by diverse, non-linearly interacting factors, knowledge of which is essential for understanding and predicting water quality trends. Currently, most water-quality research has been based on certain assumptions to employ linear approaches for solving simplified problems, such as numerical simulations or cumulative impact assessments. To improve the accuracy and ease of prediction, the random forest method has been increasingly employed as a good alternative to traditional prediction methods. In the present study, the random forest method was adopted to construct a model of the water quality response of Xincun Lagoon to anthropogenic nutrient inputs based on a limited amount of sample data, aiming to (a) identify the critical sources of nutrient inputs that affect the meeting of water quality objectives so as to minimize the socioeconomic impact on secondary stakeholders; and (b) predict the impact of a reduction of anthropogenic nutrient inputs on water quality improvement. It can be seen from the results that the intensity of stressors generated by different human activities presents an obvious non-linear superposition pattern, and the random forest method is one of the feasible solutions to this phenomenon; in addition, the impact on the lagoon ecosystem is not directly related to the intensity of the pressure source, for example, coastal aquaculture is more important than shallow sea cage aquaculture. Therefore, the method established in this paper can be used to identify the key pressure sources during the restoration of the lagoon environment, so as to achieve the unity of economy and effectiveness.

1. Introduction

As a result of both natural variability and anthropogenic activities (Halpern et al. 2008, Watson et al. 2018), most of the world’s ocean ecosystems are degrading (Halpern et al. 2012, Martin et al. 2016). This degradation is particularly pronounced in lagoons where there are strong land–sea interactions, fragile ecosystems, and significant impacts due to anthropogenic activities (Pérez-Ruzafa et al. 2013, Katsuki et al. 2019, Fang et al. 2020). Despite various protective measures or programs proposed by coastal countries for coastal zone areas since the last century, there have been no successful attempts to restore a degraded ecosystem to the historical baseline or the former ecosystem structure (Lotze et al. 2006). Coastal ecosystem restoration is a long and complex process, with various factors that contribute to ecosystem decline and have non-linear interactions with one another (Halpern and Fujita 2013). Although the cumulative impact assessment approach can be extended based on problem simplifications and assumptions (Holon et al. 2015, Mach et al. 2017, Lonsdale et al. 2020), it is still semiquantitative,
Understanding and predicting water quality trends is essential for the improvement of lagoon ecosystems not only because the public is sensitive to water quality deterioration, but also because water quality is highly correlated with the biological quality of water bodies (Crosa et al 2006). According to the *China Environment Status Bulletin of 2018*, the seawater quality of most of China’s bays is at a low-medium level and suffers from excessive levels of inorganic nitrogen and active phosphate relative to the water quality standards (Dou and Zhang 2019). The levels of these two water quality indicators directly affect the eutrophication index of water bodies. Therefore, the Ministry of Natural Resources of the People’s Republic of China proposed the Blue Bay Remediation Action Plan in 2016 to restore the ecological environmental quality of key bays and lagoons in China (Fang et al 2018), aiming to (a) improve the restoration rate of natural shorelines; (b) increase coastal wetland area; and (c) improve coastal seawater quality.

The first and second of these goals may be readily achieved through reasonable engineering measures, but there are some obstacles to achieving the third goal. The two main obstacles are as follows: (a) there is a conflict between water quality improvement goals and economic development goals. Bays and lagoons are hotspots for anthropogenic activities (such as tourism, aquaculture, ship transportation, chemical plant operation, and urban activities) (El Zetelli et al 2018). To improve water quality, the environmental management department will inevitably ban some development activities, which will directly affect the economic gain of some stakeholders, such as fishery farmers; (b) there is a conflict between water quality improvement goals and pollution source management measures. The Blue Bay Remediation Action Plan requires environmental managers to achieve a significant improvement trend in bay and lagoon water quality within two to three years. However, the response of water quality to changes in pollution sources is non-linear and delayed, and the time cost of a trial and error approach is too high. Therefore, it will be necessary to fully investigate the factors influencing seawater quality and construct a pollution source-seawater quality response model in order to develop reasonable pollution-source management measures consistent with the goals of water quality improvement.

Although many studies have been conducted on seawater quality, the research has not been systematic and has many limitations (Wang et al 2019). For example, Zheng et al (2013) used a gray prediction method to predict seawater quality in a Rigs-to-Reefs area where discarded open-ocean oil platforms are transformed into reefs, but the article is a statistical analysis of historical water quality data, and does not consider the impact of changes in human activities on the water quality environment, nor does it consider the spatial changes of water quality; Grifoll et al (2010) employed a branch-decision scheme of decision-making theories to assess the risk of water quality degradation in ports, although the impact of different human activities on water quality has been considered, only the linear superposition is considered in the superposition analysis; and Palani et al (2008) employed an artificial neural network model to predict the salinity, water temperature, dissolved oxygen, and chlorophyll-a (Chl-a) concentration in the coastal waters of Singapore, this method considers the linear and non-linear relationship of different human activities, but also does not consider the spatial distribution of water quality indicators.

Halpern and Fujita (2013) stated that the assumption that the intensities of multiple anthropogenic stressors are linearly superimposed is problematic, as the observed stress at a given monitoring site may be less (a mitigating effect) or greater (a promoting effect) than the linear superposition result. As a result, when regulating anthropogenic activities (such as for bay remediation), there is a risk that the outcomes may be contrary to expectation. For example, termination of a mitigating factor may lead to an increase in the stress on the environment, or termination of a key factor may have a multiplier effect on reducing stress while minimizing the socio-economic impact of environmental remediation.

Just to avoid the aforementioned problems, based on a limited amount of sample data, the random forest method was adopted in the present study to construct a model of the response of the water quality in a tropical lagoon, taking the Xincun Lagoon as an example, to anthropogenic nutrient inputs.

Compared with other methods, random forest is a data-driven nonlinear modeling tool that can achieve satisfactory results even with incomplete data as long as effective data learning and sample training are used. Specifically, this method has the following advantages: (a) it can directly process high-dimensional data without the need for dimensionality reduction or feature selection; (b) it is completely data-driven and does not require prior knowledge to guide decisions; (c) it can identify the importance of features and the interaction between features; and (d) it allows for accurate classification of unbalanced data sets by means of the synthetic minority oversampling technique.

Therefore, this article intends to achieve the following two purposes: (a) identify the critical sources of nutrient inputs that affect meeting water quality objectives so as to minimize the socioeconomic impact on secondary stakeholders; and (b) predict the impact of a reduction of anthropogenic nutrient
inputs on water quality improvement and reduce the conflict between the two issues above.

2. Materials and methods

The methodology adopted in the present study was comprised of six steps: (a) using inorganic nitrogen as an example of nutrients, identifying the main anthropogenic activities that cause nutrient inputs, simplifying those activities as pollution discharge from point sources, determining the spatial distribution of the point sources, and estimating the pollution load from each point source; (b) in view of the data availability and the need for data format consistency, using ArcGIS 10.2 (Environmental Systems Research Institute, Inc.) to divide the study region into 6060 grid cells with a resolution of 0.000475° (latitude) × 0.000498° (longitude) in the CS_WGS_1984 coordinate system (the distance in the GCS2000 coordinate system was 52.973 m × 55.647 m) grid; (c) mapping the spatial diffusion of pollution from each point source (under an intensity of 1) in the study region using MIKE21 (Danish Hydraulic Institute); (d) averaging the monitored inorganic nitrogen concentration over the dry season and wet season, dividing the study region into 6060 grid cells, and using Kriging interpolation to estimate the mean inorganic nitrogen concentration in the grid cells not covered by the monitoring data; (e) constructing a model of the water quality response of Xincun Lagoon to anthropogenic inorganic nitrogen inputs based on the random forest method; and (f) determining the main factors influencing water quality and predicting the possible changes in water quality due to implementation of pollution source management measures.

2.1. Study area and field data

2.1.1. Research area

Xincun Lagoon (N18°22′–18°47′, E109°45′–110°08′) is located to the southeast of Hainan Island, the second largest island in China, and has an area of about 22.6 km². It is bordered to the south by Nanwan Monkey Island (currently the only island-type nature reserve for macaque monkeys in the world) and is a typical tropical coastal lagoon (Zhou et al. 2019). Xincun Lagoon is home to diverse ecosystems and includes a provincial seagrass reserve, with original and replanted mangroves. In the 1990s, the region was home to a large number of coral reefs, which have now disappeared due to changes in the ecological environment and destruction by anthropogenic activities (Fang et al. 2018). As shown in figure 1.

According to field surveys and data collection, there are numerous marine and terrestrial anthropogenic activities with frequent interactions around Xincun Lagoon. Anthropogenic activities at sea include shallow-sea cage aquaculture (typically by the Tanka people), operation of on-sea restaurants...
(reconstructed from Tanka fishing rafts), stake net fishing, and vessel fishing (the lagoon is home to one of China’s national central fishing ports with more than 1000 vessels of various sizes). Moreover, Xincun Lagoon has historically been a pearl cultivation base in the South China Sea, producing pearls weighing up to 6 g. Anthropogenic activities on land include coastal aquaculture (mainly farming grouper (*Epinephelus* spp.) and tilapia), urban activities (including those in parts of Xincun Town, Li’an Town, and Sancai Town), riverine inputs (from the Qugang River and Qugou River), farmland cultivation (mainly wax gourds and rice), animal husbandry (mainly cattle and sheep through free-range farming), and tourism (mainly a 1 day itinerary with Nanwan Monkey Island as the main attraction).

With the advancement of the China (HAINAN) Pilot Free Trade Zone, the central government and the Hainan provincial government have gradually imposed increasingly strict requirements for the ecological environmental quality in the zone, while the coastal lagoons (bays) represented by Xincun Lagoon are under high ecological environmental stress from both natural processes and anthropogenic activities. Given this context, the then State Oceanic Administration in 2016 (now the Ministry of Natural Resources) included Xincun Lagoon in the list of pilot areas for the first-round implementation of the Blue Bay Remediation Action Plan, aiming not only to develop suitable restoration measures to promote healthy development of the marine economy in the zone, but also to provide a reference for the restoration of similar lagoons (bays).

### 2.1.2. Data sources

To understand the ecological environmental parameters of Xincun Lagoon, field surveys were conducted in the region to assess water quality, Chl-a concentration, phytoplankton, zooplankton, benthos, and intertidal organisms in June 2017 (a summer wet season) and December 2017 (a winter dry season). Given that the main problem affecting Xincun Lagoon’s water quality is eutrophication, especially due to excessive inorganic nitrogen, inorganic nitrogen concentration was selected to represent nutrient inputs (Fang et al. 2020) and was evaluated and predicted under hypothetical water management scenarios in the present study. Details of the survey sites and inorganic nitrogen data are in appendix A (available online at stacks.iop.org/ERL/16/055003/mmedia).

Field surveys were conducted on the spatial distribution and intensity of anthropogenic activities that may contribute to nutrient inputs, including river inputs, raft aquaculture, stake net fishing, vessel fishing, coastal aquaculture, urban activities, farmland cultivation, and animal husbandry in Xincun Lagoon and the surrounding land areas (Fang et al. 2020). The annual discharge of inorganic nitrogen from each type of anthropogenic activity and its proportion of the total discharge from all types of anthropogenic activities were determined according to the Handbook of Emission and Discharge Coefficient of the First National Pollution Source Survey (State Council Leading Group for the First National Pollution Source Survey, 2008), as shown in Table 1. The spatial distribution of point sources and the amount of inorganic nitrogen discharged from them are detailed in appendix B. Other data such as water depth and the distribution of mangroves and seagrass beds were acquired from the local marine administrative department.

However, we did not consider the nitrogen input from atmospheric deposition, because compared with human activities, the amount of this part is very small and it is difficult to quantitatively evaluate.

### 2.2. Data analysis

The data were further processed to map the spatial distribution of inorganic nitrogen concentration, the response field of inorganic nitrogen to anthropogenic activities, and the spatial distribution of relevant parameters for subsequent analysis. Considering the requirements for the marine environmental management of Xincun Lagoon and the accuracy of the given data, the study region was divided into 6060 grid cells with a resolution of 0.000475° (latitude) × 0.000498° (longitude) in the GCS_WGS_1984 coordinate system, consisting of 205 level-1 grid cells, 909 level-2 grid cells, 864 level-3 grid cells, 1704 level-4 grid cells, and 2377 level-5 grid cells.

#### 2.2.1. Spatial concentration distribution of inorganic nitrogen of anthropogenic origin

The inorganic nitrogen monitoring data (appendix A) were subjected to Kriging interpolation in order to map the spatial distribution of inorganic nitrogen concentrations (appendix C). However, for the marine environmental department, regulating inorganic nitrogen by the concentration range according to the *National Seawater Quality Standard*
of China (GB 3097-1997) is a more appropriate practice. The spatial distribution of the ranges of inorganic nitrogen concentration (field-monitored data combined with Kriging-interpolated data) in the present study is displayed in figure 2. As shown in figure 2, the northern part of Xincun Lagoon and the secondary lagoons had seawater quality worse than Grade IV seawater quality (≥0.50 mg l\(^{-1}\)) as specified in GB 3097-1997, which may be attributed to the fact that relatively concentrated anthropogenic activities occur in these areas and thus a large quantity of inorganic nitrogen has been discharged.

2.2.2. Spatial distribution of inorganic nitrogen input from anthropogenic activities

Most studies assume that the response of seawater quality to anthropogenic activities of a given strength decays in a linear manner (Halpern et al 2008, Li et al 2015, Holon et al 2018) or an exponential manner (Parravicini et al 2012) with respect to the distance to the source of anthropogenic activities. This assumption may be reliable to some extent in open sea waters, but it is problematic in semi-enclosed small-scale lagoons. Therefore, it was assumed in the present study that each pollution source in appendix B (excluding on-sea restaurants, port ships, and tourism due to the relatively small proportion accounted for by these three types of anthropogenic activities in the total discharge of inorganic nitrogen and also due to the wastewater management measures in place) had an intensity of one unit (assign a value of 1), and such sources were subjected to MIKE21 Danish Hydraulic Institute (DHI) simulation (appendix D) to obtain the spatial distribution of inorganic nitrogen input from anthropogenic activities of unit intensity. Such spatial distribution is also known as a response field of inorganic nitrogen concentration to anthropogenic inputs, which was separately mapped for 54 different point sources of anthropogenic activities. The response field of inorganic nitrogen concentration to the inputs from the different types of anthropogenic activities are shown in appendix E.

2.2.3. Spatial distribution of relevant parameters

It is evident that in addition to anthropogenic inputs of inorganic nitrogen, changes in water depth, water exchange percentage, total nitrogen content in sediments, mangroves, seagrass beds, and open-ocean inorganic nitrogen concentration are also important factors contributing to changes in inorganic nitrogen content in seawater bodies. It has been suggested that deep water depth, especially in estuarine areas, can cause water stratification, which limits the flow of nutrients to the surface layer of semi-closed water bodies, resulting in nutrient accumulation over the long term (Ferreira et al 2011). The water exchange percentage in this study refers to the cumulative proportion of water exchanged in the study region over one month, which was simulated using a hydrologic model. The lower the water exchange percentage, the longer the water residence time, and the higher the accumulated concentration of inorganic nitrogen (Ferreira et al 2011). Sediments are an important source of dissolved inorganic nitrogen (DIN) in the water column (Wang et al 2016). In Mobile Bay, Alabama (USA), DIN released from sediments provides 36% of the nitrogen needed to maintain primary productivity (Cowan et al 1996).
A study on four bays in Guangxi, China found that while mangroves absorb inorganic nitrogen from the water column, mangrove litter also releases a considerable amount of inorganic nitrogen (He et al. 2014). Huang et al. investigated *Enhalus acoroides*, one of the dominant species in the study region, and found that the inorganic nitrogen content in the tissues of *Enhalus acoroides* was positively correlated with the inorganic nitrogen content in the water column, with inorganic nitrogen serving as the determining factor of the growth of this species (Huang et al. 2010). The present study considered the study region as a whole, while regarding open-ocean inorganic nitrogen as an external source of pollution. The spatial distribution of the above six parameters in the study region was mapped, as shown in appendix F.

### 2.3. Response model of water quality of Xincun Lagoon to inorganic nitrogen inputs

#### 2.3.1. Random forest method

The random forest is a machine learning algorithm proposed by Breiman in 2001 (Breiman 2001) that combines bagging and random feature selection to introduce additional diversity in the decision tree model. After an ensemble of decision trees is generated, the random forest approach uses voting or takes arithmetic means of all decision trees for prediction. This gives it a strong generalization capability, especially when dealing with big data. The random forest performs well for big data, can handle nonlinearities and interactions, does not require screening for features, can assess the importance of variables well, and is one of the best machine learning algorithms to date (Xiang et al. 2019). It has been applied to species invasion assessment (Fletcher et al. 2016, Bae et al. 2018), seagrass ecosystem prediction (Holon et al. 2018), and flood risk assessment (Lai et al. 2015).

#### 2.3.2. Model construction

The aforementioned 54 response fields were each multiplied by the inputs of inorganic nitrogen from each corresponding point source (appendix B), and the results were combined with the spatial distribution of the six parameters in section 2.2.3 to form a total of 60 predictor (independent) variables while using the concentration range of inorganic nitrogen in each grid cell as the response (dependent) variable. A random forest model was constructed based on the predictor and response variables using R 4.0.1 (R Core Development Team, 2020) with relevant tool packages (mainly including ggplot2, pROC Random Forest, and varSelRF) to identify the main factors that influenced the range of inorganic nitrogen concentration, and also to predict the range of inorganic nitrogen concentration in different hypothetical scenarios in which point source control measures would be taken.

The details of the R code for model construction are in appendix G. The backward variable termination method was adopted to select the relatively important factors influencing water quality. Next, the number of predictor variables randomly selected for splitting at each tree node (mtry) and the number of decision trees (ntree) were optimized, followed by fitting and validating the random forest model.

The parameter mtry, which is the number of predictor variables randomly selected at each node in a
tree, was set at 8 to achieve the minimum out-of-bag (OOB) error rate, and the parameter ntree, which is the optimal number of decision trees, was determined to be 5000 to achieve a stable OOB error. With mtry = 8 and ntree = 5000, a random forest model was constructed using 70% of the data as training data, resulting in an OOB estimate of error rate of 0.85%. Next, the remaining 30% of data were used as testing data to validate the constructed model, which achieved a prediction accuracy of 99.95%, indicating that the model had high reliability (figure 3).

3. Results and discussion

3.1. Response results of the model construction

(a) A total of 24 point sources that strongly influenced the spatial distribution of inorganic nitrogen in the study region were selected from the 60 total point sources using the backward variable termination method, there are total nitrogen content in sediments (‘cjwzd’), urban activities (‘c1x23’), water depth (‘depth’), coastal aquaculture (‘gcd1’, ‘gcd10’, ‘gcd17’, ‘gcd19’, ‘gcd20’, ‘gcd21’, ‘gcd3’, ‘gcd9’), mangrove distribution (‘hs1’), farmland cultivation (‘nt44’), run-off input (‘quganghe’, ‘qugouhe’, ‘ruhaikou1’), water exchange percentage (‘sjhl’) and shallow-sea cage aquaculture. For MeanDecreaseAccuracy and MeanDecreaseGini, with the former standing for the mean decrease in accuracy calculated using the OOB error rate, and the latter standing for the mean decrease in Gini impurity. MeanDecreaseAccuracy was calculated by assigning a given predictor variable random but realistic values (the rest of the variables were left unchanged) and observing how the model prediction accuracy varied. The worse the model performed when a given predictor variable was randomized, the more important that variable was in predicting the response variable. The top five predictor variables in the order of decreasing importance were determined to be cjwzd (total nitrogen content in sediments), depth (water depth), c1x23 (point source 7 of urban activities), hsl (mangrove distribution), and sjhl (water exchange percentage). For MeanDecreaseGini, Gini impurity refers to the impurity of a node and is denoted by the Gini Index. The higher the Gini Index, the lower the node purity. Each time a particular predictor variable was used to split a node, the Gini Index values for the child nodes were calculated and compared to that of the original node. The changes in Gini Index were summed for each predictor variable and averaged across all trees. The mean change of Gini Index was a measure of variable importance. The top five variables in the order of decreasing importance were determined to be depth, hsl, cjwzd, sjhl, and c1x23.

(b) As shown in figure 4, predictor variable importance in the model was assessed by two metrics, MeanDecreaseAccuracy and MeanDecreaseGini, with the former standing for the mean decrease in accuracy calculated using the OOB error rate, and the latter standing for the mean decrease in Gini impurity. MeanDecreaseAccuracy was calculated by assigning a given predictor variable random but realistic values (the rest of the variables were left unchanged) and observing how the model prediction accuracy varied. The worse the model performed when a given predictor variable was randomized, the more important that variable was in predicting the response variable. The top five predictor variables in the order of decreasing importance were determined to be cjwzd (total nitrogen content in sediments), depth (water depth), c1x23 (point source 7 of urban activities), hsl (mangrove distribution), and sjhl (water exchange percentage). For MeanDecreaseGini, Gini impurity refers to the impurity of a node and is denoted by the Gini Index. The higher the Gini Index, the lower the node purity. Each time a particular predictor variable was used to split a node, the Gini Index values for the child nodes were calculated and compared to that of the original node. The changes in Gini Index were summed for each predictor variable and averaged across all trees. The mean change of Gini Index was a measure of variable importance. The top five variables in the order of decreasing importance were determined to be depth, hsl, cjwzd, sjhl, and c1x23.

3.2. Main factors affecting inorganic nitrogen concentration in the lagoon

The processes affecting inorganic nitrogen concentration in the water column of lagoons are complex, ranging from pollution-contributing processes, such as land use, population growth, pollution from point sources, pollution from agricultural non-point sources, and pollution from sediments) to pollution-mitigating processes, such as water exchange, mangrove absorption, and microbial transformation, all of which have complex, nonlinear interactions with one another. In view of data observability and accessibility, the present study included the response fields for 54 simplified point sources of pollution covering pollution-contributing processes, such as urban activities, run-off inputs, shallow-sea cage aquaculture, farmland cultivation, and coastal aquaculture, as 54 pollution-contributing factors, and also included water depth, open-ocean inorganic nitrogen concentration, total nitrogen content in sediments, water exchange percentage, mangrove distribution, and seagrass bed distribution as 6 relevant factors, totaling 60 predictor variables for further analysis.

As shown by the results in section 3, with the exception of open-ocean inorganic nitrogen concentration, the relevant factors all had significant impacts on the spatial distribution of inorganic nitrogen concentration in the water column (figure 4). For pollution-contributing factors, the effects of 29 point sources of shallow-sea cage aquaculture—with the exception of seven point sources—were not as marked as expected, which was mainly attributed to the fact that such anthropogenic activities are mainly distributed at the entrance of the lagoon, where water depth conditions and water exchange percentages are optimal. On the contrary, a higher proportion (i.e. 8 of 11) of the point sources of coastal aquaculture exhibited marked effects, because coastal aquaculture is mainly distributed in the inner areas of the lagoon. Although anthropogenic discharge of inorganic nitrogen in these areas was obviously less than that of anthropogenic activities in other areas, the former areas had lower water exchange percentages, thereby increasing the impact of the discharged inorganic nitrogen on the local sea waters and eventually leading to the accumulation of inorganic nitrogen. The Qugang River and Qugou River both input high amounts of inorganic nitrogen to the water column of Xincun Lagoon. However, the two rivers have a large flow and therefore inorganic nitrogen is not limited to local areas of the lagoon but is likely to spread throughout the entire lagoon. In contrast, urban activities and farmland cultivation, which are both distributed on the north side of the lagoon
Figure 4. The top 20 predictor variables in the order of decreasing importance in the random forest model.

and adjacent to the mangrove forest zone, input a

dramatically lower amount of inorganic nitrogen,
thereby somewhat limiting their effects to local areas
of the lagoon.

3.3. Prediction under different regulatory scenarios
3.3.1. Termination of shallow-sea cage aquaculture
Surveys and data from the local fishery authority
indicate that shallow-sea cage aquaculture is mainly
concentrated at the entrance to Xincun Lagoon.
There are a total of 916 fish-farming households—
all of whom are unlicensed or have an expired
license—with a permanent population of about 3672
people, covering approximately 2.34 km²
and mainly
breeding cobia (*Rachycentron canadum*), pompano
(*Trachinotus ovatus*), and grouper (*Epinephelus* spp.).

Shallow-sea cage aquaculture has a significant
hindering effect on the hydrodynamic conditions in
the nearby sea waters, significantly reducing the flow
velocity in the aquaculture area and somewhat weak-
ening the water exchange capacity, which makes it
difficult for pollutants to diffuse out of the aquacul-
ture area and eventually results in local eutrophi-

cation (Boyd and Heasman 1998, Kuang et al 2019).

The local government plans to relocate all unlicensed
aquaculture activities in the lagoon to other places
for bay remediation while allowing some shallow-sea
cage aquaculture to continue on a much smaller scale
for leisure fisheries and sightseeing.

To predict the water quality of Xincun Lagoon
following the termination of all shallow-sea cage
aquaculture, the pollution discharge from each point
source of pollution in the shallow-sea cage aquacul-
ture area was set to 0 in the random forest model. It
was predicted that inorganic nitrogen concentrations
would meet the water quality standard for grade I sea-
water in all grid cells in the scenario above.

Shallow-sea cage aquaculture accounted for more
than half of the total discharge of inorganic nitro-
gen with discharge points located at the mouth of
the lagoon, affecting the entire lagoon area through
the movement of water. Therefore, a reduction in
the discharge from shallow-sea cage aquaculture
would significantly improve the water quality within
the lagoon. However, shallow-sea aquaculture is a

crucial means of livelihood for thousands of local
fishermen, thereby necessitating full consideration
of the potential socio-economic impact of regulat-
ory measures prior to eliminating shallow-sea cage
aquaculture.

3.3.2. Termination of coastal aquaculture
Coastal aquaculture can be divided into high-level
aquaculture and low-level aquaculture in mar-
ine administrative management, with the former
referring to aquaculture above the shoreline and the
latter below the shoreline. Surveys and remote sensing
images show that there are 22 coastal aquaculture
Spatial distribution of inorganic nitrogen levels in a hypothetical scenario in which all coastal aquaculture activities are prohibited.

facilities around Xincun Lagoon, covering an area of about 3.92 km² (most of them were constructed by local villagers without authorization). The main aquaculture species in this area are grouper (*Epinephelus* spp.), crimson snapper (*Lutjanus erythrops*), cobia (*Rachycentron canadum*), and shrimp. In autumn and spring, most of the breeding ponds are used to raise batches of fry (from current October to the following May).

Coastal aquaculture ponds are mainly reclaimed from wetlands. The construction process encroaches on ecosystems such as mangrove forests and salt marshes, which causes the coastline retreat and fragment (Berlanga–Robles *et al* 2011). At the same time, the indiscriminate discharge of aquaculture wastewater leads to an explosion of inorganic nitrogen levels in the surrounding waters, which may lead to biodiversity decline (Thomsen *et al* 2020).

The inorganic nitrogen discharge from each point source of pollution in the area of coastal aquaculture was set to 0 in the present study. The random forest model showed that the number of grid cells with grades I–V water quality were 769, 1439, 1083, 1937, and 832. The spatial distribution of these grid cells is shown in figure 5. Compared with the initial situation, the inorganic nitrogen concentration in the secondary lagoons was projected to reach the quality standard for grade I water after the elimination of coastal aquaculture, which further verified that the main source of inorganic nitrogen pollution in this area was coastal aquaculture.

3.4. Application prospects of the random forest method

Holon *et al* (2018) found that most anthropogenic activities exhibit complex nonlinear effects on ecosystem degradation, and the nonlinear trend can be well predicted using the random forest method. While the present study mainly investigated the impact of anthropogenic nutrient inputs on water quality rather than on ecosystems, a similar conclusion was reached, that is, anthropogenic activities exhibited obvious nonlinear effects on inorganic nitrogen concentration in the water column. Such nonlinear effects cannot be modeled using mechanical methods such as numerical simulation.

Through the establishment of the random forest model in this paper, it is possible to accurately identify the key factors affecting the water quality of the lagoon, rather than just blindly selecting the human activities with the largest amount of nutrient input, in fact, which are generally the main source of income for local residents. In the subsequent restoration of the lagoon’s ecological environment,
this situation should be fully considered, and the marine environment should be improved while minimizing the impact on local residents and achieving relative sustainability.

3.5. Limitations
As mentioned earlier, random forest models have many advantages with good prediction ability (Holon et al. 2018). However, the model has relatively high requirements for the data used (including data coverage and accuracy), as was observed in the present study. In particular, the data coverage determined the prediction scope while the data accuracy determined the prediction accuracy. Therefore, in order to improve the scope and accuracy of predictions long-term observation and investigation in the study region is necessary.

Another limitation for random forest models is that feature selection with random forest algorithms depends on setting an arbitrary cutoff, which may cause a few feature variables of high importance to be filtered out. In addition, random forest models fail to fully consider the impact of the correlation of feature variables on the model prediction accuracy. To extract as much useful feature information or patterns from high-dimensional data as possible, multiple rounds of computation can be performed to obtain average values. Alternatively, random forest algorithms could be optimized.

4. Conclusions
This present study employed a random forest model to explore the main factors contributing to the increase of inorganic nitrogen concentration in the water column of a lagoon and predict how the lagoon water quality would vary in response to a reduction of anthropogenic nutrient inputs.

The modeling results showed that anthropogenic activities that create a large amount of inorganic nitrogen inputs do not necessarily cause large changes in the lagoon water quality. In some cases, more attention should be paid to anthropogenic activities that are not a large contributor of inorganic nitrogen but have obvious effects on local waters, such as urban activities and farmland cultivation on the north side of the lagoon. Such anthropogenic activities are inconspicuous, but considering them is essential for ecological restoration because management of these activities has a low socioeconomic impact while having a multiplier effect on remediation.

Data availability statement
All data that support the findings of this study are included within the article (and any supplementary files).

Funding
This work was supported by Scientific Research Fund of the Second Institute of Oceanography, MNR, Grant No. JG1917.

Conflict of interest
The authors report no declarations of interest.

References
Bae M J, Murphy C A and García-Berihuete E 2018 Temperature and hydrologic alteration predict the spread of invasive Largemouth Bass (Micropterus salmoides) Sci. Total Environ. 639 58–66
Ban N C, Aldina H M and Ardron J A 2010 Cumulative impact mapping: advances, relevance and limitations to marine management and conservation, using Canada’s Pacific waters as a case study Mar. Policy 34 876–86
Berlanga-Robles C A, Ruiz-Luna A, Bocco G and Velerdy Z 2011 Spatial analysis of the impact of shrimp culture on the coastal wetlands on the northern coast of Sinaloa, Mexico Ocean Coast. Manage. 54 535–43
Boyd A J and Heasman K G 1998 Shellfish mariculture in the Benguela system: water flow patterns within a mussel farm in Saldanha Bay, South Africa J. Shellfish Res. 17 25–32 (https://ia802807.us.archive.org/20/items/charthouse_38131_shellfishmaricultureinthebengu1981/shellfishmaricultureinthebengu1981.pdf)
Breiman L 2001 Random forests Mach. Learn. 45 5–32
Cowan J L W, Pennock J R and Boynton W R 1996 Seasonal and interannual patterns of sediment-water nutrient and oxygen fluxes in Mobile Bay, Alabama (USA): regulating factors and ecological significance Mar. Ecol. Prog. Ser. 141 229–45
Crosa G, Froehrich J, Nikolayenko V, Stefani F, Galli P and Calamari D 2006 Spatial and seasonal variations in the water quality of the Amu Darya River (Central Asia) Water Res. 40 2237–45
Dou Q and Zhang Z Y 2019 Analysis on environmental pollution in China’s coastal ecosystem J. Resour. Ecol. 10 424–31
El Zrelli R, Rabaioui I, Ben Alaya M, Daighbouj N, Castet S, Besson P, Michiel S, Bejaoui N and Courjault-Rade P 2018 Seawater quality assessment and identification of pollution sources along the central coastal area of Gabes Gulf (SE Tunisia): evidence of industrial impact and implications for marine environment protection Mar. Pollut. Bull. 127 445–52
Fang X, Hao C L, Cheng C G, Ni W, Wang J C and Zhang Y F 2018 Pressure analysis and spatial quantitative assessment of impact of human activities on lagoon ecosystem: a case study in Xincun and Li’an Lagoons, Hainan, China Appl. Ecol. Environ. Res. 16 6253–66
Fang X, Li X, Xiang Y, Hao C, Zhao Y and Zhang Y 2020 Cumulative impact of anthropogenic nutrient inputs on lagoon ecosystems—a case study of Xincun Lagoon, Hainan, China Reg. Stud. Mar. Sci. 35 101213
Ferreira J G et al 2011 Overview of eutrophication indicators to assess environmental status within the European marine strategy framework directive Estuar. Coast. Shelf. Sci. 93 117–31
Fletcher D H, Gillingham P K, Britton J R, Blanchet S and Gozlan R E 2016 Predicting global invasion risks: a management tool to prevent future introductions Sci. Rep. 6 26316
Grifoll M, Jordà G, Borja Á and Espino M 2010 A new risk assessment method for water quality degradation in harbour domains, using hydrodynamic models Mar. Pollut. Bull. 60 69–78
Halpern B S et al 2012 An index to assess the health and benefits of the global ocean Nature 488 615–20
Halpern B S and Fujita R 2013 Assumptions, challenges, and future directions in cumulative impact analysis Ecosphere 4 1–11
Halpern B S, Walbridge S, Selkoe K A, Kappel C V, Micheli F, D’Ago R C, Bruno J F, Casey K S, Ebert C and Fox H E 2008 A global map of human impact on marine ecosystems Science 319 948–52
He B M, Wei M X, Fan H Q, Pan L H and Cao Q X 2014 Spatio-temporal change of inorganic nitrogen content and the evaluation of eutrophication in the surface seawaters of mangrove areas in Guangxi Bays J. Appl. Oceanogr. 33 140–8
Holm F, Marre G, Pargiacini V, Mouquet N, Boekel T, Descamp P, Tribot A S, Boissery P and Deter J 2018 A predictive model based on multiple coastal anthropogenic pressures explains the degradation status of a marine ecosystem: implications for management and conservation Biol. Conserv. 222 125–35
Holm F, Mouquet N, Boissery P, Bouchchoua M, Delaruelle G, Tribot A S and Deter J 2015 Fine-scale cartography of human impacts along French Mediterranean Coasts: a relevant map for the management of marine ecosystems PLoS One 10 e0135473
Huang D J, Huang X P and Huang Z G 2010 Spatio-temporal variation of TN & TP contents in Enhalus acoroides and responses to nutrient load in Xincun Bay, Hainan Mar. Environ. Sci. 29 40–3
Katsuki K, Seto K, Tsujimoto A, Takata H and Sonoda Y 2019 Impact of anthropogenic eutrophication, Lake Mokoto, Japan Ecol. Indic. 19 253–63
Kryvonenko A, Vore A, Vassallo P, Micheli F, Montefalcone M, Morri C, Paoli C, Albertelli G, Fabiano M and Bianchi C N 2012 Understanding relationships between conflicting human uses and coastal ecosystems status: a geospatial modeling approach Ecol. Indic. 19 17–33
Kryvonenko A, Vore A, Vassallo P, Micheli F, Montefalcone M, Morri C, Paoli C, Albertelli G, Fabiano M and Bianchi C N 2012 Understanding relationships between conflicting human uses and coastal ecosystems status: a geospatial modeling approach Ecol. Indic. 19 253–63
Krzemien A, Herbeck L S and Jennerjahn T C 2020 The end of resilience: surpassed nitrogen thresholds in coastal waters led to severe seagrass loss after decades of exposure to aquaculture effluents Mar. Environ. Res. 160 104986
Wang Y L, Yuan H M, Song J M, Li X G, Li N, Qu B X, Kang X M, Wang Q D, Xing J W and Liang X M 2016 Benthic exchange of dissolved inorganic nitrogen and impact of relevant environmental factors at sediment-water interface in Jiaozhou Bay Mar. Sci. 41 104–14
Wang Z, Bu C, Li H and Wei W 2019 Seawater environmental Kuznets curve: evidence from seawater quality in China’s coastal waters J. Cleaner Prod. 219 925–35
Watson J E M, Venter O, Lee J, Jones K R, Robinson J G, Possingham H P and Allan J R 2018 Protect the last of the wild Nature 563 27–30
Xiang T, Dong X H, Guo C and Li Z J 2019 The impacts of climate change on the distribution of Pelteobagrus fulvidraco by randomized forest model Acta Hydrob. Sin. 43 1262–70
Zheng Y, Zheng X, Gao Z and Zhang Y 2013 Prediction of seawater quality in Rigs-to-Reefs area based on grey systems theory Procedia Environ. Sci. 18 236–42
Zhou L, Yang Y, Wang Z, Jia J, Mao L, Li Z, Fang X and Gao S 2019 Investigating ENSO and WPWP modulated typhoon variability in the South China Sea during the mid–late Holocene using sedimentological evidence from southeastern Hainan Island, China Mar. Geol. 416 105987