An optimized machine learning approach to water pollution variation monitoring with time-series Landsat images

Yi Lin [a,b], Lang Li [a], Jie Yu [a,c,*], Yuan Hu [a], Tinghui Zhang [a], Zhanglin Ye [a], Awase Syed [d], Jonathan Li [d]

[a] College of Surveying and Geo-Informatics, Tongji University, Shanghai 200092, China
[b] Research Center of Remote Sensing & Spatial Information Technology, Shanghai 200092, China
[c] College of Environmental Science and Engineering, Tongji University, Shanghai 200092, China
[d] Department of Geography and Environmental Management University of Waterloo Waterloo, Ontario N2L 3G1, Canada

ABSTRACT

Non-point source (NPS) pollution has greatly threatened socio-economic development and human health due to water environment degradation. It is very important to quantitatively analyze spatio-temporal variation rules of NPS pollution sources surrounding drinking water source area (DWSA) and their impact on the water environment with time-series satellite images. In this paper, we study a systematic remote sensing monitoring method on DWSA of upper Huangpu River, Shanghai. Firstly, an optimized Extreme Learning Machine (ELM) classification algorithm, namely Mixed Kernel ELM with Particle Swarm Optimization (PSO-MK-ELM) was constructed. Based on the PSO-MK-ELM, four NPS pollution sources- farmland, building land, woodland, and water were identified accurately and efficiently. Then their corresponding spatiotemporal analysis was performed with 30 years (1989–2019) Landsat images. On the basis of NPS pollution source area and census data from 1989 to 2017, the principal pollutants discharged into DWSA were also calculated with the common Export Coefficient Model (ECM). Finally, the contributions of the spatial and temporal changes of NPS pollution sources on pollutant emissions were analyzed. The result indicates the PSO-MK-ELM has an advantage of efficiency and accuracy in NPS pollution source extraction and our results are expected to provide a scientific basis and data support for NPS pollution control and DWSA protection for better practices for environmental management in megacities worldwide.

1. Introduction

Among many natural resources, water resources is one of the key strategic and economic resources for ensuring sustainable economic and social growth (Li and Qian, 2018). The protection of the drinking water source area (DWSA) as an important water storage area is closely linked to the stability of civilization and the well-being of the people (Tang et al., 2019). In general, non-point source (NPS) pollution occurs from soil runoff, erosion, air deposition, flooding, intrusion or hydrological alteration (rainfall and melting snow), where it is difficult to trace pollution back to a particular source. This NPS pollution threatens water and environmental security as it is one of the major pollution types to DWSA (Chen et al., 2016). According to the Environmental Protection

Agency, NPS pollutants mainly originate from pollution sources including agricultural activities and urban living which finally converge into the water body via runoff effects (Ongley et al., 2010). The physical and anthropological variability due to land use/cover change (LUCC) and urbanization has significant impact on NPS pollutant emission (Ding and Liu, 2019; Duan et al., 2016). One of the main reasons for excessive emissions, is due to increasing population and their life style activities that have significant impact on LUCC. These variations in physical and anthropological activities due to rapid urbanization process result in more NPS pollutants being released (Zhang et al., 2017). Therefore, it is very important to continuously monitor changes in NPS pollution sources and to quantify their effect on DWSA.

The standard NPS emission investigation depends heavily on field

* Corresponding author at: College of Environmental Science and Engineering, Tongji University, Shanghai 200092, China. College of Surveying and Geo-Informatics, Tongji University, Shanghai 200092, China

E-mail addresses: linyi@tongji.edu.cn (Y. Lin), lilang@tongji.edu.cn (L. Li), 2011_jieyu@tongji.edu.cn (J. Yu), 1731975@tongji.edu.cn (Y. Hu), zhang_th@tongji.edu.cn (T. Zhang), 1434888@tongji.edu.cn (Z. Ye), awase008@gmail.com (A. Syed), junli@uwaterloo.ca (J. Li).

assessment and laboratory analysis, making it tedious and labor intensive (Shen et al., 2020). It further makes it difficult to continuously monitor large survey areas to predict pollutant emission. Duan et al. (2013) argues that these conventional methods only acquire sparsely sampled data making it difficult to evaluate water pollution quantitatively and with higher precision. With the recent technology advances, remote sensing technology serves to address these challenges by quickly acquiring image snapshots for identifying and monitoring pollution on large geographic surveys (Gómez et al., 2016). Since time-series remote sensing images can reveal spatial-temporal changes of land-use types including the NPS pollution sources, it can be used to estimate the NPS pollution emission and to analyze its impact on emission variations into a DWSA. Consequently, the remote sensing-based surveys are of spatio-temporal nature, to some extent, they can capture the impact of NPS pollutants on a DWSA.

The analysis of remote sensing images is dependent on two significant parameters, they being, LUCC over a temporal frame and its impact on NPS pollutant emission. Significant challenges arise by correlating NPS pollutant emission and LUCC with a long period: (1) to identify NPS pollution sources in time-series remote sensing images; (2) to quantify the NPS pollutant emission and analyze its variations with LUCC. The land-use types and area information of NPS pollution sources provide the basic information for NPS pollutant emission estimation and quantitatively analyzing the LUCC impacts on the variations of emission.

The first problem is usually solved by Machine Learning (ML) methods on chronological remote sensing images (Duro et al., 2012; Talukdar et al., 2020). Numerous researchers studied ML classifiers to extract ground information for monitoring and analyzing LUCC (Mahdianpari et al., 2018; Rudke et al., 2019; Tong et al., 2020). Generally, pixel-based classification can be divided into parametric and non-parametric methods (Phiri and Morgenroth, 2017). A widely used parametric method is the Maximum Likelihood, which acquires a desirable classification result when the data fulfills the normal distribution. However, the complex landscapes of an urban environment usually break the pre-assumption (Liu et al., 2011). On the contrary, the non-parametric classification methods like Support Vector Machine (SVM) and Artificial Neural Networks (ANNs) are not dependent on the data distribution (Mountrakis et al., 2011; Rokni et al., 2015). SVM is initially designed to deal with binary questions that maximize the distance between support vectors. It becomes popular for its capability to handle high-dimensional data with a few training samples (Ma et al., 2019). But it is relatively time-consuming for multi-class problems. ANNs maps the input into specific classes by assigning different weights on connected nodes and also the key is to initialize and adjust these weights to implicitly learn the rules between the input and the output (Maxwell et al., 2018). However, high training costs and intensive user-defined criteria are major obstacles to improve the accuracy. As an outstanding ANNs method, Extreme Learning Machine (ELM) can randomly assign the weights from input to the hidden layer which reduces the training time and only addresses the number of hidden layer neurons with fewer human interventions (Huang et al., 2015; Pal et al., 2013). However, classical ELM cannot guarantee high classification accuracy when dealing with highly nonlinear problems and thus it is natural to introduce a kernel function to deal with it. Further investigation is needed to simplify the model complexity and improve the NPS pollution source recognition accuracy. This can be achieved by the learning ability of the model and strong generalization kernel ELM method.

In addition to the aforementioned challenges, quantifying the NPS pollutant emission into the water needs to be addressed. The literature present two widely accepted models. The physical-based models represented by SWAT (Lai et al., 2020) strictly simulate physical processes by abundant parameters that limit their applications in data-lacking areas; on the contrary, the empirical models are exemplified with less input data for an appropriate pollutant emission estimate (Adu and Kumarasamy, 2018; Zhang et al., 2019). The Export Coefficient Model (ECM) is a widely used empirical model that calculates the total NPS pollutant emission as the sum of each NPS pollution source emission in accordance with its individual contribution (Ding et al., 2010; Johnes, 1996; Wu et al., 2015). However, the comprehensive study of the interrelationships between the long-term LUCC affected by the urbanization process and its effect on the NPS emission of DWSA in mega-cities is comparatively minimal. Furthermore, the well-researched NPS pollution is more concerned with sources linked to agricultural activities and the estimation of nitrogen and phosphorus emissions. (Duan et al., 2020; Wang et al., 2019).

The NPS pollutant emission estimated by the ECM usually relied on statistical data. Since the area information was evolved with time and space, it is difficult to monitor the LUCC information timely based on the field investigation. Consequently, we turned to a robust ELM method, which is expected to own both distinguishing learning ability and strong generalization ability in extracting the LUCC information on time-series images. Combining the extracted area LUCC information with other required input parameters (annual emission coefficient, removal rate, and the effluent coefficient), we could use the simple but effective ECM to estimate the NPS pollutant into the DWSA. In this paper, we used time-series Landsat images of DWSA of the upper Huangpu River in Shanghai to find how the LUCC from the 1980s to 2010s affects the NPS pollutant emission variations. First of all, we need to solve the problem of constructing an efficient and accurate extraction algorithm. Based on the optimized classification algorithm, using nine images from 1989 to 2019 to obtain the LUCC information and the spatiotemporal rules. Then ECM was utilized to estimate four typical pollutant emission from 1989 to 2017 in DWSA. Spatial-temporal variation analysis of NPS pollutant emission was finally conducted. Our contribution is that we provided a comprehensive framework from LUCC information extraction based on the time-series remote sensing images to quantitatively NPS pollution monitoring in a DWSA. Specially, we developed an efficient and reliable machine learning algorithm in LUCC information extraction, and this helped us to get rid of estimating the NPS pollutant emission based on the statistical data completely. Besides, it provides reliable data support for LUCC analysis and analyzes its impacts on NPS pollutant emission.

2. Study area and data sets

2.1. Study area

In this paper, DWSA of upper Huangpu River was taken as a case study. Fig. 1 showed the study area (red rectangle) and its zoom in. It is used in classification comparison and accuracy evaluation. Considering the NPS influence sphere, the quantitative NPS emission analysis area is selected in the second-order DWSA of the upper Huangpu River with a 3 km buffer zone (yellow curve). It includes about 754 km^2 water area and land area in total. Waters (blue curve) mainly includes Hengliaojing, Weigangjing, Xiatang, Taipu River, Lanlugang, Dianshan Lake (cyan curve), and Huangpu River. Among all the waters, Dianshan Lake is of utmost importance since it is the largest natural freshwater lake in Shanghai and the source of the Huangpu River.

2.2. Data

2.2.1. Time-series Landsat images

In this study, nine Landsat images with orbit number 118 / 39 were collected by TM, ETM+, and OLI from 1989 to 2019. The image parameters were illustrated in Table 1. To ensure the consistency of data analysis in the target area, the images were all acquired in summer. And six spectral bands with spatial resolution of 30 m were used in the experiment.

2.2.2. Statistical data

Annual reports were utilized in evaluating our method and calculating NPS pollutant emission. For accuracy assessment, area of
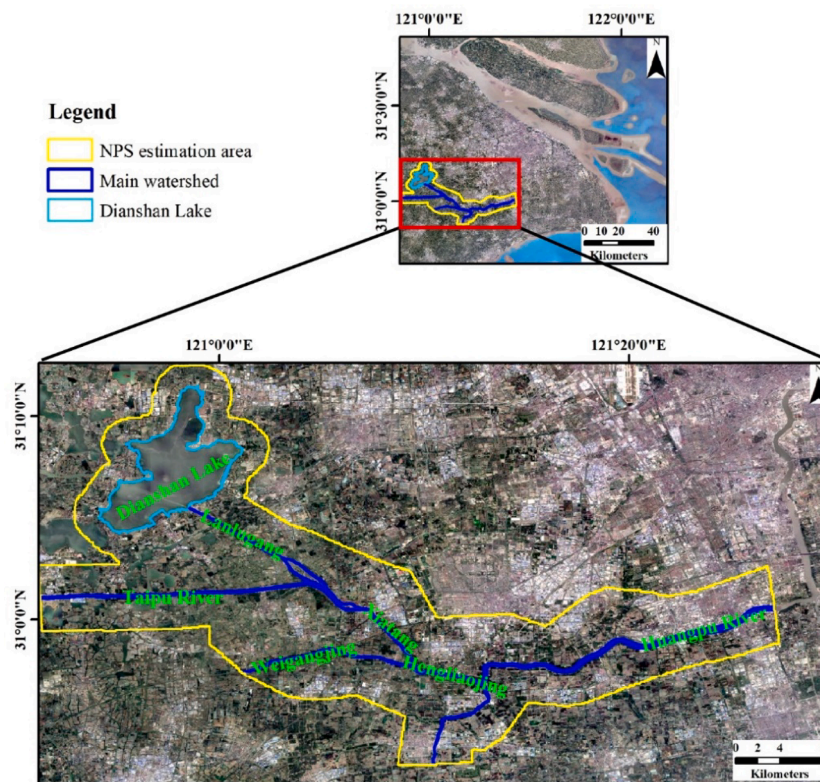
Fig. 1. Study area: DWSA of the upper Huangpu river. The yellow curve borders the second-order DWSA of the upper Huangpu River with a 3 km buffer area. The blue line delineates waters. The cyan closed curve is the boundary of Dianshan Lake. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Landsat images used in this study.

Data source	Sensor	Date (YYYY/MM/DD)
Landsat-5	TM	1989/06/05
Landsat-5	TM	1995/08/10
Landsat-7	ETM+	2000/09/18
Landsat-5	TM	2007/07/28
Landsat-5	TM	2009/07/17
Landsat-8	OLI	2013/08/29
Landsat-8	OLI	2015/08/03
Landsat-8	OLI	2017/08/24
Landsat-8	OLI	2019/07/29

Table 3
Population density of the water source nearby.

Year	Population density of water source nearby districts (people/km^2)				
	Qingpu	Songjiang	Minhang	Fengxian	Jinshan
1989	670	825	3410	757	929
1995	678	818	1467	757	975
2000	679	819	1759	733	905
2007	1160	1632	5113	1089	1139
2009	1217	1965	4894	1191	1179
2013	1787	2867	6830	1679	1331
2015	1804	2906	6845	1687	1362
2017	1799	2892	6836	1681	1367

administrative division, crop and intensive fish pond from 2015 annual report of Jinze Town, Qingpu District was gathered and displayed in Table 2. Population density statistical data used in NPS pollutant emission estimation was listed in Table 3. Pollutant emission coefficients and effluent coefficients of land-use types from the statistics department were listed in Table 4.

2.2.3. Reference dataset
The reference data of the 2015 Landsat-8 image of the whole DWSA of the upper Huangpu River was used to train four ML classifiers and

Table 2
Statistical data from 2015 annual report of Jinze Town, Qingpu District.

Type	Area (km^2)
Administrative region	108.42
Farmland	9.42
Intensive fish pond	9.86

assess classification accuracy. Table 5 listed 1841 samples used in this study.

3. Method

The overall method is to construct an intelligent ML algorithm on Landsat long time-series images of the upper Huangpu River, to analyze NPS pollution source spatial-temporal LUCC around the DWSA, to calculate the effluent pollutant emission and to find out the LUCC impact on pollutant emission changes. In detail, the Mixed Kernel ELM with Particle Swarm Optimization (PSO-MK-ELM) algorithm was utilized on nine Landsat imageries to extract ground object information followed by the analysis of LUCC in DWSA. Then, combining the estimated area from the PSO-MK-ELM with statistical reports, the pollutant emission of four major pollutants into the DWSA was calculated separately based on the ECM. Finally, the spatial-temporal NPS pollution source variation impact on pollutant emission variations was analyzed in detail. Fig. 2 showed the workflow of our method.

Table 4

NPS pollution sources emission and effluent coefficients.

Pollution sources	Pollutant emission coefficients				Effluent coefficient
	TN ($kg \cdot hm^{-2} \cdot a^{-1}$)	TP ($kg \cdot hm^{-2} \cdot a^{-1}$)	COD ($kg \cdot people^{-1} \cdot a^{-1}$)	NH$_3$-N ($kg \cdot hm^{-2} \cdot a^{-1}$)	
Farmland	24.2	1.2	150	1.8	0.1
Woodland	4.8	0.2	80	0.2	0.15
Sewage	2.19	0.07	9.86	1.46	0.7
Aquaculture	8.9	4	252.4	4.7	0.90

* Data source of pollutant emission coefficients. The pollutant emission coefficients of farmland, woodland and aquaculture are from Handbook of The First National Pollution Source Survey Production and Emission Coefficients. And the data of Sewage is from Water Environment Improvement of Main Rivers into Lake Taihu Watersheds Technical Specification for Governance Planning.

** Data source of effluent coefficient. Effluent coefficient of four pollution sources are acquired from Water Environment Improvement of Main Rivers into Lake Taihu Watersheds Technical Specification for Governance Planning.

Table 5

Data set used in 2015 data classification.

Type of samples	Training data	Test data	Total number
Water	200	102	302
Building land	200	105	305
Woodland	200	106	306
Farmland	200	109	309
Bareland	200	104	304
Factories	200	115	315

3.1. Preprocessing

All the images used were preprocessed using radiometric correction, atmospheric correction, and geometric correction techniques. The radiometric correction and atmospheric correction were used to eliminate the absorption and attenuation effect during propagation and retrieve a surface reflectance associated with ground objects. And the geometric correction was to calibrate the location of pixels by using fifteen uniformly distributed ground control points.

3.2. Feature space construction

To enlarge differences of land-use types, we transformed original data into a suitable feature domain for machine-sensing. In this paper, we constructed a six-dimension space by spectral analysis and band selection. The spectral analysis results imply that the Normalized Difference Vegetation Index (NDVI), Normalized Difference Factory Index (NDFI), and the brightness component of Kanth-Thomas (K-T)

transformation are suitable features since they are good indicators for vegetation, factories and bareland. Besides, the band selection computed by the Optimum Index Factor (OIF) (Dwivedi and Rao, 1992) method manifests the optimal band combination is the near-infrared, red, and green band for TM / ETM + or near-infrared, red, and blue for OLI.

3.3. PSO-MK-ELM algorithm and accuracy assessment

3.3.1. The derivation of PSO-MK-ELM algorithm

The core of the PSO-MK-ELM algorithm is to improve the performance of classification by using a multi-kernel ELM method with parameters that are automatically determined by a bionic algorithm.

ELM is a generalized single layer feed-forward neural network (SLFN), which predicts the classification results by only one hidden layer in the original data space and avoids being influenced by weights and biases setting when the activation function is infinite differential (Huang et al., 2015). And the forward description of an ELM classification problem with t input with d nodes, l hidden nodes and m class labels can be written as:

$$\underset{t \times m}{\mathbf{Y}} = \underset{t \times l}{\mathbf{H}} \underset{l \times m}{\boldsymbol{\beta}} = \begin{pmatrix} \mathbf{h}_1(\mathbf{x}_1) & \mathbf{h}_2(\mathbf{x}_1) & \cdots & \mathbf{h}_l(\mathbf{x}_1) \\ \mathbf{h}_1(\mathbf{x}_2) & \mathbf{h}_2(\mathbf{x}_2) & \cdots & \mathbf{h}_l(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{h}_1(\mathbf{x}_t) & \mathbf{h}_2(\mathbf{x}_t) & \cdots & \mathbf{h}_l(\mathbf{x}_t) \end{pmatrix}_{t \times l} \begin{pmatrix} \boldsymbol{\beta}_1^T \\ \boldsymbol{\beta}_2^T \\ \vdots \\ \boldsymbol{\beta}_l^T \end{pmatrix}_{l \times m} = \begin{pmatrix} \mathbf{y}_1^T \\ \mathbf{y}_2^T \\ \vdots \\ \mathbf{y}_l^T \end{pmatrix}_{t \times m},$$

(1)

where **H** denotes the hidden layer output matrix to randomly map features in a nonlinear way, $\boldsymbol{\beta}$ refers to the weight vector linking the hidden
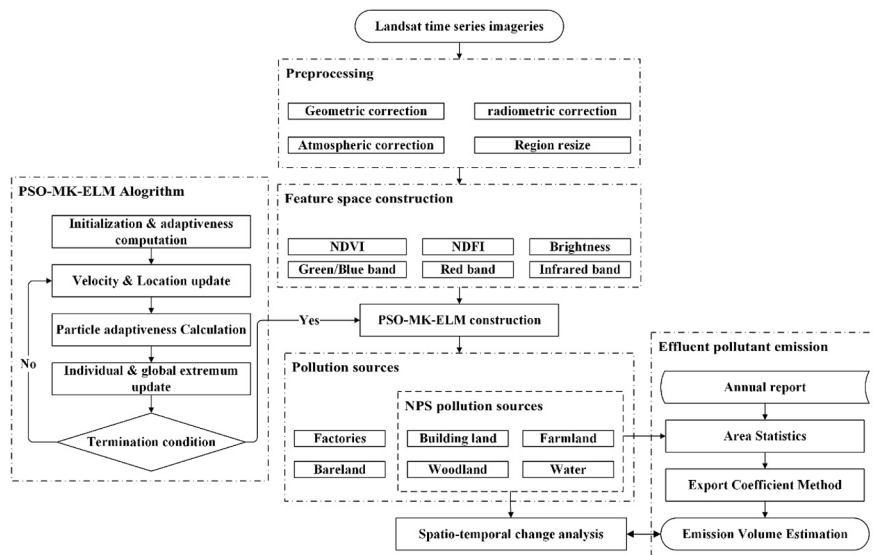

Fig. 2. Workflow of our method.

and output layer and $\mathbf{Y}$ is the output vector. In the light of regularization theory, β can be retrieved as:

$$\hat{\beta} = \left(\mathbf{H}^{\mathrm{T}}\mathbf{H} + \mathbf{I}/c\right)^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{Y}, \tag{2}$$

where $\mathbf{I}$ is an identity matrix and c is a regularization parameter.

Since the hidden layer output matrix maps the original data space into the ELM feature space nonlinearly, it can be substituted by a kernel matrix to realize the same purpose on the ground of the kernel theory. Then Equation (2) can be written as:

$$\hat{\beta} = \left(\mathbf{K}^{\mathrm{T}}\mathbf{K} + \mathbf{I}/c\right)^{-1}\mathbf{K}^{\mathrm{T}}\mathbf{Y}, \tag{3}$$

where the kernel matrix is represented as:

$$\mathbf{K} = \begin{pmatrix} K(\mathbf{x}_1, \mathbf{x}_1) & K(\mathbf{x}_1, \mathbf{x}_2) & \cdots & K(\mathbf{x}_1, \mathbf{x}_t) \\ K(\mathbf{x}_2, \mathbf{x}_1) & K(\mathbf{x}_2, \mathbf{x}_2) & \cdots & K(\mathbf{x}_1, \mathbf{x}_t) \\ \vdots & \vdots & \ddots & \vdots \\ K(\mathbf{x}_t, \mathbf{x}_1) & K(\mathbf{x}_t, \mathbf{x}_2) & \cdots & K(\mathbf{x}_t, \mathbf{x}_t) \end{pmatrix}_{t \times t}, \tag{4}$$

$K(\cdot, \cdot)$ stands for a kernel function. The Polynomial kernel function considers the contribution on the kernel value regardless of the spatial distance with a good generalization capability while the Gaussian kernel function includes only the influences of points near the center point with a high learning ability (Brailovsky et al., 1999; Liu et al., 2015). When linearly combining these two kernel functions, the mixed kernel function can be rewritten as:

$$K(\mathbf{x}, \mathbf{x}') = K_G\left(\mathbf{x}_G, \mathbf{x}_G'\right) + K_P\left(\mathbf{x}_P, \mathbf{x}_P'\right) \tag{5a}$$

$$K_G\left(\mathbf{x}_G, \mathbf{x}_G'\right) = exp\left(\frac{\|\mathbf{x}_G - \mathbf{x}_G'\|^2}{\sigma^2}\right) \tag{5b}$$

$$K_P\left(\mathbf{x}_P, \mathbf{x}_P'\right) = \left(\mathbf{x}_P \cdot \mathbf{x}_P' + a\right)^{\mathrm{b}} \tag{5c}$$

where K, K_G, K_P are the hybrid, the Gaussian, and the Polynomial kernel function, respectively. And $\mathbf{x}_G, \mathbf{x}_P$ are input vectors for the Gaussian and the Polynomial kernel function.

Optimal parameter selection in Equation (3) is a key issue since parameters can affect the classification accuracy. Fortunately, it can be solved by using the PSO algorithm. The idea of the PSO algorithm is to initialize a set of particles and each particle parameterized by velocity, position and fitness attribute suggests a potential optimal solution in space and search for the ideal parameter by iterating in D-dimension space (Zhang et al., 2015). The algorithm is not only easy to operate and fast to compute, but also advantageous in many application fields (Tamiminia et al., 2017; Wang et al., 2018). In particular, it is suitable to deal with multivariable and nonlinear problems. Consequently, we used the PSO algorithm to search for optimal solutions of Gaussian kernel function, Polynomial kernel function and regularization parameters.

3.3.2. Accuracy assessment

Accuracy assessment in this paper was conducted in both global and local aspects. The first assessment was set to evaluate the PSO-MK-ELM algorithm in the whole range of the DWSA. The second accuracy assessment measured the algorithm in subset zones, evaluating the accuracy with statistical data. And the reference data of the former accuracy assessment consisted of 1841 samples from the ground survey over the DWSA and from the corresponding high-resolution images, and 200 training samples of each land-use type were randomly selected and the remaining samples were used for the test. The details of samples were listed in Table 5. And the latter reference data was statistical data from the 2015 annual report of Jinze Town, Qingpu District.

The first accuracy assessment compared our PSO-MK-ELM algorithm with other widely used algorithms in terms of total accuracy, Kappa coefficient, and computational time consumption. We constructed a confusion matrix based on the reference data for the calculation of the first two indexes. Then the total accuracy and Kappa coefficient could be

used to assess overall classification accuracy among different classification algorithms and the consistency of the predicted class with the actual one. And the computational time was recorded during the classification since it reflects the classification efficiency directly. The second accuracy assessment was designed to evaluate the PSO-MK-ELM classification results with its accordingly statistical data. The purpose of the assessment was to find out the difference between the statistical area and the extracted area. Consequently, the absolute and the relative error were chosen as the indicators.

3.4. NPS pollutant effluent emission calculation

ECM is a classical method that utilizes empirical parameters to reckon the NPS pollutant emission. It originated from the LUCC and lake eutrophication assessment of North America in the 1970s, popularized for computational simplification in nitrogen and phosphorus loss prediction in 1996, started to be highly concerned in China since 2000 (Duan et al., 2020). ECM is simple in formula with a few parameters and it is easy to operate. Besides, it can output a reliable emission even in the shortage of in-situ monitoring data. Considering the advantages of ECM, we utilized it to estimate the NPS pollutant emission. ECM is formulated as:

$$L = \sum_{i=1}^{n} A_i E_i (1 - T_i) f_i, \tag{6}$$

where L represents the total pollutant emission from different pollution sources per year in the unit of ton or kilogram. n stands for the number of NPS pollution sources; A_i means an area or a quantity of the i-th pollution source. Area information of NPS pollution sources was acquired and counted via the PSO-MK-ELM classification results. The urban population is reckoned by the multiplication of building land area and the population density. E_i refers to the annual emission coefficient of the i-th pollution source. T_i expresses the removal rate of the i-th pollution source, which usually refers to the capability to get rid of the domestic sewage. And we chose 60% removal rate in the following estimation. f_i is the i-th effluent coefficient that was determined according to a Technical Specification released by the Department of Ecology and Environment of Jiangsu Province.

3.5. Interrelation between LUCC information and NPS emission of DWSA

Area variations of land-use types affect the contribution of NPS pollutant emission that discharged into DWSA. Based on the PSO-MK-ELM algorithm, we acquired the LUCC information of the 30-year range. And it was used as an input parameter of ECM when used in the NPS pollutant emission estimation discharged into DWSA. Then the LUCC information utilized for spatiotemporal analysis could reveal how land-use type changes affected the emission variations.

4. Results and discussion

4.1. LUCC results

We input the blue, red, and near-infrared band into the Gaussian kernel function when the rest NDVI, the first component of K-T transform, and NDFI into the Polynomial kernel function in the experiments. To compare the classification performance of different ML algorithms, Support Vector Machine (SVM), Maximum Likelihood, Classical ELM, and PSO-MK-ELM were used on the reference data. The classification results were shown in Fig. 3. And the classification accuracy evaluation was listed in Table 6.

The PSO-MK-ELM performed better than other three methods when comprehensively evaluated the accuracy and time cost of four algorithms. The total accuracy of four classifiers exceeds 0.84, which indicates those ML algorithms are capable of separating ground objects in
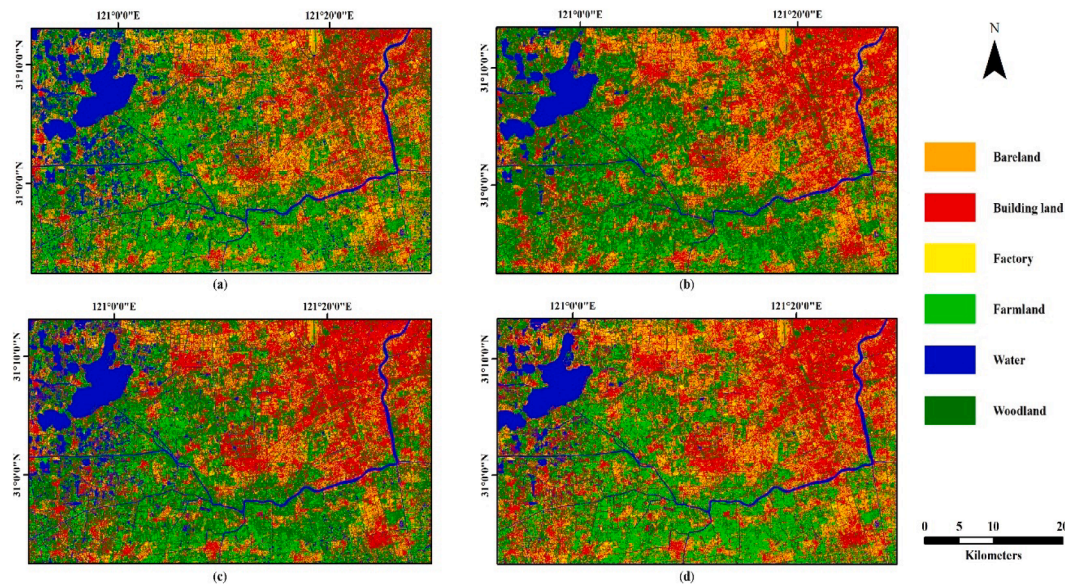
Fig. 3. Classification results of reference data. (a) SVM (b) Maximum Likelihood (c) Classical ELM (d) PSO-MK-ELM.

Table 6
Classification accuracy evaluation statistics.

Type of samples	Total Accuracy	Kappa Coefficient	Time consuming (s)
SVM	0.9155	0.9033	6.2420
Maximum Likelihood	0.8419	0.8228	5.1868
Classical ELM	0.8619	0.8373	1.4192
PSO-MK-ELM	0.9267	0.9026	2.7332

the research area to a satisfying accuracy. But our method and SVM can both distinguish the surface information with 5% accuracy improvement than the other two methods. Furthermore, our method is 1.12% more accurate than SVM and only consumes about 43.787% computational time of SVM to achieve the same classification accuracy level. Indeed, the kappa coefficient of the PSO-MK-ELM is slightly lower than the SVM, but higher total accuracy and less computation time still validate its effectiveness.

4.2. Accuracy validation

4.2.1. Comparison between LUCC extraction and statistical data

The area of administrative region, farmland, and intensive fish pond listed in Table 7 were extracted from Fig. 3 (d) based on the urban division and the urban river vector range (Fig. 4). And the values were compared with the data in Fig. 2. The extracted area (105.27 km^2) of Jinze Town administrative region is nearly consistent with the statistical one (108.42 km^2) with −3.15 km^2 absolute error and 2.91% relative error. Similarly, farmland and intensive fish pond have absolute errors of 0.94 km^2 and −0.79 km^2 with relative errors of 9.98% and 8.01% accordingly. It demonstrates that the proposed method can accurately distinguish three types of land with the highest absolute error of no more

Table 7
Area statistics from the PSO-MK-ELM classifier.

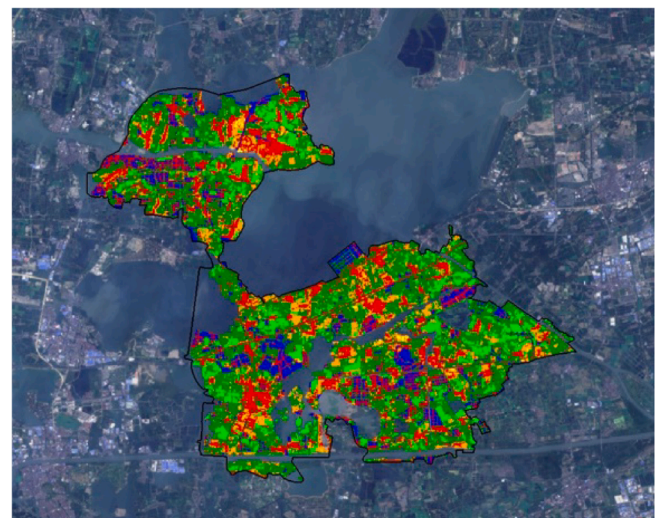| Type	Estimated area (km^2)	Absolute error (km^2)	Relative error (%)
Administrative region	105.27	−3.15	2.91
Farmland	10.36	0.94	9.98
Intensive fish pond	10.65	−0.79	8.01

Fig. 4. Ground object information of Jinze Town.

than 3.2 km^2. Indeed, the relative errors of farmland and fish pond are higher than that of the administrative region. The result is probably due to floating plants (water hyacinth, cyanobacteria) which are prone to be misclassified into the grains in summer. Because the administrative region is less likely to be affected by seasonal changes and therefore shows a good performance in the relative error. It is then logical to attribute discrepancies between the extracted area and the statistical area from the annual book to seasonal differences in data acquisition. In general, our method can fulfill the accuracy requirements for a further quantitative study.

4.2.2. Comparison of detail retention capabilities

To compare the identification ability of different classifiers from the perspective of detail preservation, a fish pond near the bank was extracted (Fig. 5). When compared with the SVM classification result, both the classical ELM and our method show a better recognition ability to identify vegetation covering the fish pond. What's more, our method is more accurate to distinguish woodland from other ground objects such as building land.
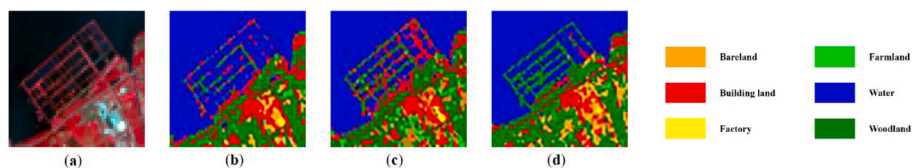
Fig. 5. Classification results comparison of different methods: (a) False color (band 4, 3, 2), (b) SVM, (c) Classical ELM, (d) PSO-MK-ELM.

4.3. LUCC spatiotemporal analysis

In this part, nine images (1989–2019) of the second-order DWSA with a 3 km buffer zone were processed by the PSO-MK-ELM algorithm. The classification results were displayed in Fig. 6 and the extracted area data were listed in Table 8. To illustrate the data in Table 8 vividly for spatiotemporal analysis, we drew a line chart to explain the variation tendency (Fig. 7).

In the past 30 years (1989–2019), six land-use types were recognized and extracted. According to area statistics and line graph drawing after the PSO-MK-ELM classification, the dynamic spatio-temporal change laws of LUCC can be analyzed.

- The area of building land was in general slowly increasing except for the decreasing period from 1989 to 1995 and from 2013 to 2015. As a kind of NPS pollution source, building land in residential areas could exert an adverse impact on the water quality improvement of DWSA considering the sewage and domestic waste emission. From 1989 to 2019, the area increased only 2.7988 km^2 per year on average. This gradual increase indicates an outstanding balance of urban development and environmental protection under urbanization. As for the decreasing period, it was likely to be caused by demolition and reconstruction of the process of urbanization.

- The trend of farmland variation showed a negative correlation with woodland. Farmland area extended constantly and reached at the peak of 307.5759 km^2 in 2000 as more food was required after more individuals settling down into the city under the urbanization while woodland decreased for the livelihood and development of the increasing population. After that, the extension of farmland decreased and the decreasing trend was consistent with the gradual increase of woodland area until 2009. Since other land types were not detected anomalies during the same span, it can attribute to the implementation of the Returning Land from Farmland to Forest Policy. After 2009, although farmland started to increase while woodland decreased, the variation rate is milder comparing with the previous periods.

- The area of water almost varied little in the past 30 years in terms of its amplitude. It decreased about 11.5587 km^2 in total that indicates the water body within the DWSA remained relatively stable and thus can infer that the aquaculture area did not vary significantly. And the constant area of the water body should be attributed to the protection laws of DWSA.

- The area of factories started to see a gradual increase since the year 2000, while by the year 2019, it saw a gradual decline. In the years 1989 and 1995, the area of factories was recorded as 0 km^2 which mirrors the DWSA characteristics of low-level industrialization, agriculture-oriented status, and early stage of urbanization within that period. From 2000 to 2019, the area could vary no more than 5.7360 km^2 annually. And a reduction has been noted from the years 2017 to 2019 due to industrial structure adjustments.

- The area of bareland fluctuateed along with the changes of other LUCC types. From 1989 to 2007, bareland increased in a relatively violent way when reaching the peak value of 71.3493 km^2 mainly caused by the sharply varied farmland and woodland. From 2007 to 2009, the increasing woodland and building land area covered the decreasing farmland under the influence of the growing population
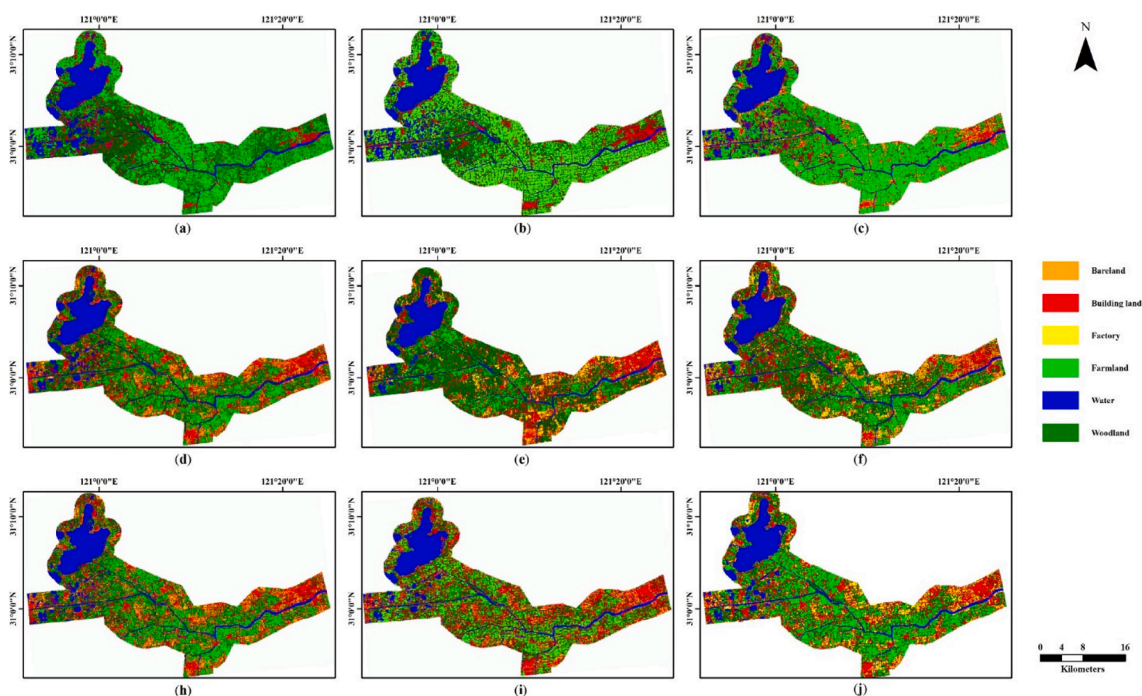

Fig. 6. PSO-MK-ELM classification results of the second-order DWSA with a 3 km buffer zone: (a) 1989, (b) 1995, (c) 2000, (d) 2007, (e) 2009, (f) 2013, (g) 2015, (h) 2017, (i) 2019.

Table 8
Area statistics of the second-order DWSA.

Year	Building land (km^2)	Farmland (km^2)	Woodland (km^2)	Bareland (km^2)	Factories (km^2)	Water (km^2)
1989	104.8509	228.8844	299.9592	0	0	120.5865
1995	89.4474	295.9209	235.8972	8.3061	0	124.7094
2000	125.9631	307.5759	119.0673	31.1976	49.6674	120.8097
2007	151.9974	110.5587	236.8539	71.3493	63.1296	120.3921
2009	156.0888	58.9734	312.7491	31.3416	73.6209	121.5072
2013	167.8977	93.9276	263.0754	31.5621	85.8429	111.6792
2015	151.9974	110.5587	236.8539	53.3493	90.1296	111.3921
2017	169.4304	144.6678	193.0518	38.0628	98.2944	110.4777
2019	188.8155	174.8196	177.6096	33.2847	70.0614	109.0278
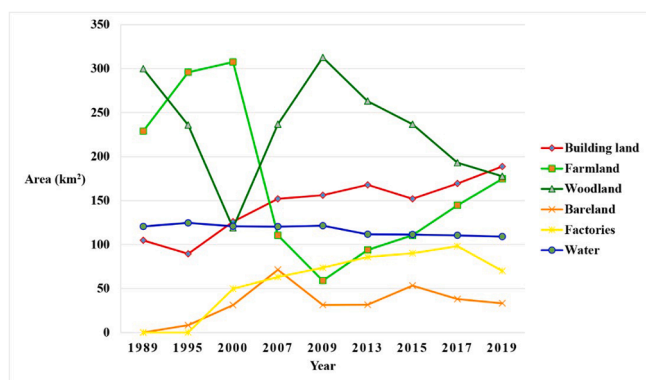
Fig. 7. Line chart of LUCC statistical data from 1989 to 2019.

and the environmental conservation policy, which caused the descending trend as the whole. From 2009 to 2019, bareland area changed a little in general. And the fluctuation around the year 2015 was caused by the demolition and reconstruction of Shanghai.

4.4. NPS pollution source variation analysis

NPS pollution sources in our study refers to woodland, farmland, building land and water, which are related to the agricultural activities and urban life. In this part, we focused on the NPS pollution source variations especially the spatial-temporal changes resulted from the urbanization process. According to Fig. 7, the spatiotemporal evolution of NPS pollution sources could be divided into three stages:

The first stage (1989–2000) depicted an urbanization period with an agriculture-dependent production. Urgent needs for improving living conditions resulted in demolition and reconstruction, which was reflected by building land trend to decrease first and then it started to increase. With the upgraded houses, more people were attracted to settle in this area which required more farmland to cultivate and more houses to live in. And woodland area therefore shrunk.

The second stage (2000–2009) displayed an accelerated urbanization phase with some measures of environmental protection measures. Building land was still expanding in order to deal with population growth pressure at a quicker rate. As for farmland, it was converted into woodland that could help protect the water source ecological environment after the implementation of Returning Land from Farming to Forest Policy nationally. And it caused a significant conversion between woodland and farmland. Besides, it also changed the mode of production of the citizens nearby. That is to say, some people were making a living from farmland to factories that was presented in the increasing area of factories and decreasing farmland.

The third stage (2009–2019) delineated a highly urbanized level. Building land and farmland were still expanding but at a mild growing pace under the urbanization effect. And the descending area of woodland mainly became farmland to fulfill the basic need for food.

It is worth noting that water was the only NPS pollution source

whose area could keep stable. And the relative stability of water area was determined by the nature of DWSA.

4.5. ECM for NPS pollutant emission estimation

According to the statistical information, the effluent pollutant emissions of NPS pollution sources were displayed in Table 9. In the shortage of detailed census data of 2019, the emission estimation utilized eight images from 1989 to 2017. From Table 9, it could be seen that:

- TN. From 1989 to 2017, the TN emission increased from 597.342 t to 1651.668 t with an average increase of 37.6545 t annually. And the corresponding emission ranged from 8.18% to 13.18%. The emission of TN achieved the peak and trough in 1995 and 2015 separately.
- TP. The TP emission was 60.603 t in 1989 and only added 29.885 t after 28 years, which means that the annual average increasing amount was 1.0673 t. The corresponding rate of emission contribution varied in the range of 0.72% and 1.05%. The emission of TP reached a maximum of 90.488 t in 2017 and a minimum of 57.385 t in 1995.
- COD. The emission of COD in 1989 was 5350.433 t and added up to 9775.928 t in 2017 with a 1.83 times increment over 28 years. The COD ranked first in terms of the absolute pollutant emission and the per year contribution ratio which indicates it as the most influential pollutant within the research area.
- NH$_3$-N. The initial emission of NH$_3$-N equaled 338.538 t in 1989 and it came to 1013.117 t in 2017. The ratio of the annual increment was 24.0921 t. And it varied from 4.23% to 8.38% in the sampling years. And the highest and the lowest emission load was the 1044.539 t of 2013 and 230.464 t of 1995.

4.6. NPS pollutant emission analysis

Typical pollutant emission from 1989 to 2017 were displayed in Fig. 8.

In general, four pollutants including TN, TP, COD and NH$_3$-N discharged into the second-order DWSA of the upper Huangpu River were affected by agriculture, forestry, aquaculture and resident life. According to the annual emission contribution, the pollutant emission ranked from top to bottom is COD (77.88%-86.54%), TN (8.18%-13.18%), NH$_3$-N (4.23%-8.38%), and TP (0.72%-1.05%). Conformed with the changes of pollution source areas, the in-water emission was then varied. Under the pressure of urbanization, domestic sewage emission increased and the COD emission was subsequently increased. So did the NH$_3$-N emission. And the TN emission was influenced by farmland and building land, and the latter ascending tend outweighed the descending former one that caused the increased TN emission in total. Considering the shrunk farmland and stable aquaculture, the TP emission in general decreased.

Specifically, the LUCC variation and its relations to NPS pollutant emission into the water could be obviously seen as:

Table 9
NPS effluent pollutant emission and its contribution ratio.

| Year	NPS effluent pollutant emission (t) and its annual emission percentage (%)							
	TN	Percent	TP	Percent	COD	Percent	NH$_3$-N	Percent
1989	597.342	9.41	60.603	0.96	5350.433	84.30	338.538	5.33
1995	446.006	8.18	57.385	1.05	4719.270	86.54	230.464	4.23
2000	557.867	9.33	59.625	1.00	5050.838	84.50	309.058	5.17
2007	1084.687	11.69	75.566	0.81	7437.045	80.14	683.256	7.36
2009	1133.942	11.74	77.346	0.80	7725.411	80.00	719.947	7.46
2013	1623.355	13.02	89.818	0.72	9711.919	77.88	1044.539	8.38
2015	1494.197	12.83	85.646	0.74	9108.787	78.21	957.261	8.22
2017	1651.668	13.18	90.488	0.72	9775.928	78.01	1013.117	8.09
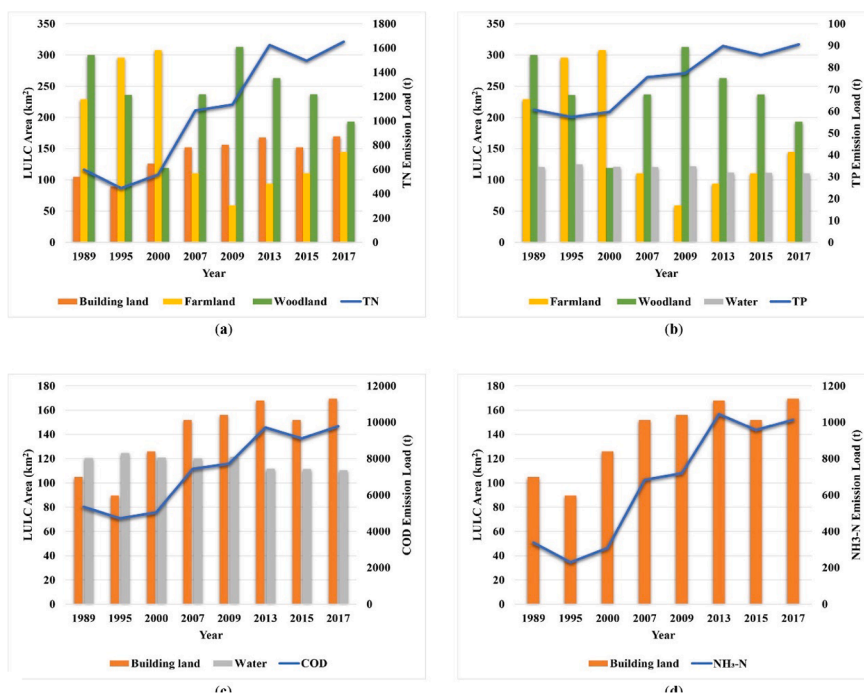
Fig. 8. Four NPS pollutant emission and its major sources. (a) TN, (b) TP, (c) COD, and (d) NH$_3$-N.

- TN. The emission of TN was mainly derived from farmland, woodland and building land. While farmland was decreasing, building land area expanded faster and thus caused a gradual increase of TN. From the period 1989 to 1995 and the period 2013 to 2015, TN emission descended because of the increasing farmland. And then from 2009 to 2015, farmland remained relatively stable, TN emission changed milder than in the previous years.

- TP. Considering the major contributions of TP were from water, farmland and woodland, TP emission was therefore going down as the result of slowing changes of fish pond farming and significant down towards changes of farmland.

- COD. Water and building land were both chief LUCC types contributing to COD emission with closely related to the aquaculture and urban life. With the continuously increasing areas of building land, COD emission increased accordingly.

- NH$_3$-N. As the principal output of the domestic sewage discharged from building land, the NH$_3$-N emission was directly influenced by the gradual increase in building land as the result of a growing population.

5. Conclusions

In this paper, we proposed a comprehensive method to monitor long-term water pollution variation in a DWSA by using Landsat images. An optimized machine learning algorithm (PSO-MK-ELM) was constructed to acquire LUCC information in 30 years for studying the NPS pollution source variation rules. Following the spatial-temporal variation analysis, four pollutants released into the DWSA were estimated to quantify their contribution to water quality. Finally, spatio-temporal variations of NPS pollutant emission were analyzed under the impact of LUCC.

The contributions in this study are stated as follows.

(1) The classification results validate that the PSO-MK-ELM algorithm is highly accurate and efficient to extract NPS pollution sources in DWSA.

(2) The results of long term spatial-temporal LUCC monitoring show that the distribution variations of NPS pollution sources should be attributed to urbanization process and policy of environmental conservation. Specifically, building land variation is mainly influenced by the demolition and reconstruction needs in different urbanization stages. The opposite change trend of farmland and woodland is mainly caused by Returning Land from Farming to Forest Policy. And the relatively stable area of water is the result of laws of DWSA conservation.

(3) The emission reveals that COD is the dominant NPS pollutant from the late 1980s to the late 2010s in the study area while TP contributes the least to the DWSA at the same time. And the variation of NPS pollutant emission is conformed with the

distribution change of NPS pollution sources. It verifies again that the urbanization process and environmental protection policies are the forces to influence LUCC in DWSA and on the change of NPS pollutant emission.

Our study proves that the intelligent remote sensing monitoring method using time-series satellite images can provide a scientific data analysis basis for water resources protection and pollution source control. In addition, it is possible to extend the research to other water source protection areas worldwide for the sustainable development.

CRediT authorship contribution statement

Yi Lin: Conceptualization, Methodology, Resources, Writing - original draft, Writing - review & editing, Supervision, Funding acquisition. **Lang Li:** Software, Validation, Formal analysis, Writing - original draft, Writing - review & editing. **Jie Yu:** Conceptualization, Methodology, Writing - review & editing, Project administration. **Yuan Hu:** Visualization, Writing - original draft. **Tinghui Zhang:** Data curation, Validation. **Zhanglin Ye:** Investigation, Software. **Awase Syed:** Writing - review & editing. **Jonathan Li:** Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

References

Adu, J.T., Kumarasamy, M.V., 2018. Assessing non-point source pollution models: A review. Polish J. Environ. Stud. https://doi.org/10.15244/pjoes/76497.

Brailovsky, V.L., Barzilay, O., Shahave, R., 1999. On global, local, mixed and neighborhood kernels for support vector machines. Pattern Recognit. Lett. https://doi.org/10.1016/S0167-8655(99)00086-0.

Chen, L., Wang, G., Zhong, Y., Shen, Z., 2016. Evaluating the impacts of soil data on hydrological and nonpoint source pollution prediction. Sci. Total Environ. 563–564, 19–28. https://doi.org/10.1016/j.scitotenv.2016.04.107.

Ding, X., Liu, L., 2019. Long-term effects of anthropogenic factors on nonpoint source pollution in the upper reaches of the Yangtze river. Sustain. 11 https://doi.org/10.3390/su11082246.

Ding, X., Shen, Z., Hong, Q., Yang, Z., Wu, X., Liu, R., 2010. Development and test of the Export Coefficient Model in the Upper Reach of the Yangtze River. J. Hydrol. 383, 233–244. https://doi.org/10.1016/j.jhydrol.2009.12.039.

Duan, M., Du, X., Peng, W., Jiang, C., Zhang, S., Ding, Y., 2020. Quantitative assessment of background pollutants using a modified method in data-poor regions. Environ. Monit. Assess. 192 https://doi.org/10.1007/s10661-020-8122-8.

Duan, W., He, B., Nover, D., Yang, G., Chen, W., Meng, H., Zou, S., Liu, C., 2016. Water quality assessment and pollution source identification of the eastern poyang lake basin using multivariate statistical methods. Sustain. 8 https://doi.org/10.3390/su8020133.

Duan, W., Takara, K., He, B., Luo, P., Nover, D., Yamashiki, Y., 2013. Spatial and temporal trends in estimates of nutrient and suspended sediment loads in the Ishikari River, Japan, 1985 to 2010. Sci. Total Environ. 461–462, 499–508. https://doi.org/10.1016/j.scitotenv.2013.05.022.

Duro, D.C., Franklin, S.E., Dubé, M.G., 2012. A comparison of pixel-based and object-based image analysis with selected machine learning algorithms for the classification of agricultural landscapes using SPOT-5 HRG imagery. Remote Sens. Environ. 118, 259–272. https://doi.org/10.1016/j.rse.2011.11.020.

Dwivedi, R.S., Rao, B.R.M., 1992. The selection of the best possible Landsat TM band combination for delineating salt-affected soils. Int. J. Remote Sens. 13, 2051–2058. https://doi.org/10.1080/01431169208904252.

Gómez, C., White, J.C., Wulder, M.A., 2016. Optical remotely sensed time series data for land cover classification: A review. Photogramm. Remote Sens. ISPRS J. 10.1016/j.isprsjprs.2016.03.008.

Huang, Gao, Huang, Guang Bin, Song, S., You, K., 2015. Trends in extreme learning machines: A review. Neural Networks, 10.1016/j.neunet.2014.10.001.

Johnes, P.J., 1996. Evaluation and management of the impact of land use change on the nitrogen and phosphorus load delivered to surface waters: The export coefficient modelling approach. J. Hydrol. 183, 323–349, 10.1016/0022-1694(95)02951-6.

Lai, G., Luo, J., Li, Q., Qiu, L., Pan, R., Zeng, X., Zhang, L., Yi, F., 2020. Modification and validation of the SWAT model based on multi-plant growth mode, a case study of the Meijiang River Basin. China. J. Hydrol. 585, 10.1016/j.jhydrol.2020.124778.

Li, P., Qian, H., 2018. Water resources research to support a sustainable China. J. Water Resour. Dev. Int. 10.1080/07900627.2018.1452723.

Liu, K., Shi, W., Zhang, H., 2011. A fuzzy topology-based maximum likelihood classification. ISPRS J. Photogramm. Remote Sens. 66, 103–114, 10.1016/j.isprsjprs.2010.09.007.

Liu, X., Wang, L., Huang, G.-B., Zhang, J., Yin, J., 2015. Multiple kernel extreme learning machine. Neurocomputing, 10.1016/j.neucom.2013.09.072.

Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., Johnson, B.A., 2019. Deep learning in remote sensing applications: A meta-analysis and review. Photogramm. Remote Sens. ISPRS J. 10.1016/j.isprsjprs.2019.04.015.

Mahdianpari, M., Salehi, B., Rezaee, M., Mohammadimanesh, F., Zhang, Y., 2018. Very deep convolutional neural networks for complex land cover mapping using multispectral remote sensing imagery. Remote Sens. 10, 10.3390/rs10071119.

Maxwell, A.E., Warner, T.A., Fang, F., 2018. Implementation of machine-learning classification in remote sensing: An applied review. J. Remote Sens. Int. 10.1080/01431161.2018.1433343.

Mountrakis, G., Im, J., Ogole, C., 2011. Support vector machines in remote sensing: A review. Photogramm. Remote Sens. ISPRS J. 10.1016/j.isprsjprs.2010.11.001.

Ongley, E.D., Xiaolan, Z., Tao, Y., 2010. Current status of agricultural and rural non-point source Pollution assessment in China. Pollut. Environ. 10.1016/j.envpol.2009.10.047.

Pal, M., Maxwell, A.E., Warner, T.A., 2013. Kernel-based extreme learning machine for remote-sensing image classification. Remote Sens. Lett. 4, 853–862, 10.1080/2150704X.2013.805279.

Phiri, D., Morgenroth, J., 2017. Developments in Landsat land cover classification methods: A review. Remote Sens, 10.3390/rs9090967.

Rokni, K., Ahmad, A., Solaimani, K., Hazini, S., 2015. A new approach for surface water change detection: Integration of pixel level image fusion and image classification techniques. J. Appl. Earth Obs. Geoinf. Int. 10.1016/j.jag.2014.08.014.

Rudke, A.P., Fujita, T., de Almeida, D.S., Eiras, M.M., Xavier, A.C.F., Rafee, S.A.A., Santos, E.B., de Morais, M.V.B., Martins, L.D., de Souza, R.V.A., Souza, R.A.F., Hallak, R., de Freitas, E.D., Uvo, C.B., Martins, J.A., 2019. Land cover data of Upper Parana River Basin, South America, at high spatial resolution. J. Appl. Earth Obs. Geoinf. Int. 10.1016/j.jag.2019.101926.

Shen, W., Zhang, L., Li, S., Zhuang, Y., Liu, H., Pan, J., 2020. A framework for evaluating county-level non-point source pollution: Joint use of monitoring and model assessment. Sci. Total Environ. 722, 10.1016/j.scitotenv.2020.137956.

Talukdar, S., Singha, P., Mahato, S., Shahfahad, Pal, S., Liou, Y.A., Rahman, A., 2020. Land-use land-cover classification by machine learning classifiers for satellite observations-A review. Remote Sens. https://doi.org/10.3390/rs12071135.

Tamiminia, H., Homayouni, S., McNairn, H., Safari, A., 2017. A particle swarm optimized kernel-based clustering method for crop mapping from multi-temporal polarimetric L-band SAR observations. J. Appl. Earth Obs. Geoinf. Int. 10.1016/j.jag.2017.02.010.

Tang, X., Zhai, A., Ding, X., Zhu, Q., 2019. Safety Guarantee System of Drinking Water Source in Three Gorges Reservoir Area and its Application in Huangjuedu Drinking Water Source Area. Sustain. 11, 10.3390/su11247074.

Tong, X.Y., Xia, G.S., Lu, Q., Shen, H., Li, S., You, S., Zhang, L., 2020. Land-cover classification with high-resolution remote sensing images using transferable deep models. Remote Sens. Environ. 237, 10.1016/j.rse.2019.111322.

Wang, D., Tan, D., Liu, L., 2018. Particle swarm optimization algorithm: an overview. Soft Comput, 10.1007/s00500-016-2474-6.

Wang, G., Li, J., Sun, W., Xue, B., A, Y., Liu, T., 2019. Non-point source pollution risks in a drinking water protection zone based on remote sensing data embedded within a nutrient budget model. Water Res. 157, 238–246. https://doi.org/10.1016/j.watres.2019.03.070.

Wu, L., Gao, J.E., Ma, X.Y., Li, D., 2015. Application of modified export coefficient method on the load estimation of non-point source nitrogen and phosphorus pollution of soil and water loss in semiarid regions. Environ. Sci. Pollut. Res. 22, 10647–10660, 10.1007/s11356-015-4242-z.

Zhang, L., Gao, S., Wei, B., Li, Y., Li, H., Wang, L., Ye, B., 2017. Effects of Urbanization on rural drinkingwater quality in Beijing. China. Sustain. 9, 10.3390/su9040461.

Zhang, L., Wang, Z., Chai, J., Fu, Y., Wei, C., Wang, Y., 2019. Temporal and spatial changes of non-point source N and p and its decoupling from agricultural development in water source area of middle route of the south-to-north water diversion project. Sustain. 11, 10.3390/su11030895.

Zhang, Y., Wang, S., Ji, G., 2015. A Comprehensive Survey on Particle Swarm Optimization Algorithm and Its Applications. Probl. Eng. Math. 10.1155/2015/931256.