An improved source apportionment mixing model combined with a Bayesian approach for nonpoint source pollution load estimation

Zhongyue Yan, Jing Xu and Xiaohong Ruan

ABSTRACT

A nonpoint source (NPS) loads evaluation method based on the Bayesian source apportionment mixing model was established in this research. The model assumed that (1) the pollutant concentration from each source mixed with the others in the monitoring section during transport, (2) transport only considered first-order attenuation, (3) point source pollution had relatively stable emissions, and (4) the measurement error was random, unrelated, and consistent with a normal distribution (mean of 0). All unknown parameters in the model were taken as random variables, and their posterior distributions were derived by Markov chain Monte Carlo procedures based on historical data, literature, and empirical information. The outflow system of the Huaihe River was adopted as a case study to verify the feasibility of the model. Gelman–Rubin, automatic frequency control statistics, and the determination coefficient (R^2) verified the reliability. The results showed that the total loads of ammonia nitrogen (NH_4^+), chemical oxygen demand, total nitrogen, and total phosphorus from NPSs accounted for 16.35–27.58%, 18.78–25.69%, 21.68–29.71%, and 42.11–52.09%, respectively. The parameter sensitivity analysis showed that prior distribution of NPS concentration was the most sensitive one, which should be determined reasonably based on the empirical or historical information.

Key words | Bayesian approach, nonpoint source (NPS) pollution load, source apportionment mixing model (SAMM), uncertainty

Zhongyue Yan
Jing Xu
Xiaohong Ruan (corresponding author)
MOE Key Laboratory of Surficial Geochemistry,
School of Earth Sciences and Engineering,
Nanjing University,
No. 163 Xianlin Road, Qixia District,
Nanjing 210023, Jiangsu,
China
E-mail: *ruanxh@nju.edu.cn*

INTRODUCTION

Nonpoint source (NPS) pollution is often defined as diffuse pollution that occurs over a wide area and is associated with land use (Nikolaidis *et al.* 1998; Xu *et al.* 2013), and this form of pollution has had an important effect on water quality impairment worldwide (Chen *et al.* 2010; Zhuo *et al.* 2017). Traditional assessment methods for NPS pollution load estimation often demand a substantial amount of data consisting of basin data and river data, which are not always obtained during routine monitoring. Recently, an inverse modelling approach, which has been widely applied to environmental and hydrological problems (Woodbury & Ulrych 2000; Shen *et al.* 2006; Zou *et al.* 2007), has provided the possibility of estimating NPS loads based on limited data.

Currently, there are two simple and effective formulas used to estimate NPS loads with inverse models. The first one is the concentration formula (Liu *et al.* 2008; Gronewold *et al.* 2009; Shen & Zhao 2009, 2010), which estimates the NPS load via the concentration relationship between upstream and downstream regions. However, this formula lacks descriptions for the influences of point

doi: 10.2166/nh.2019.076

sources and discharge in the relationship. Therefore, the NPS load assessment results may be too high or too low. The second formula is the load formula (Chen *et al.* 2011, 2012; Zhao *et al.* 2014), which estimates the NPS load via load conservation and joins the point source and discharge descriptions. However, the requirements for upstream and downstream regions and point source flow monitoring are not easily met, and it is difficult to employ the formula for a wide range of applications. Thus, it would be beneficial to develop a robust method for load estimation by using incomplete routine monitoring data.

The source apportionment mixing model (SAMM) is often used to calculate the proportional contributions of pollution sources in a sample by using stable isotopes (Xue *et al.* 2012; Yang *et al.* 2013; Meghdadi & Javar 2017; Wang *et al.* 2017) or chemical elements (Massoudieh & Kayhanian 2013; Sharifi *et al.* 2014). The SAMM uses the characteristics of each source to identify and apportion the volumetric flow rate contributions of each source based on the mass conservation law. Based on this point of view, the SAMM can also be used to estimate the proportion of NPS pollution to recipient regions based on emission characteristics if the emission characteristics of the sources are not collinear. Furthermore, NPS loads can be estimated by using the contribution proportion and known point source load. This method can reduce the requirements for flow data and increase the applicability.

Due to the uncertainties of the model structure, parameters, and data (Lindenschmidt *et al.* 2007), it is obviously impossible to obtain the true values of the parameters. The Bayesian method, which treats all unknown variables as random variables, is usually applied to evaluate the parameters and uncertainties in an inverse model (Liu *et al.* 2008; Shen & Zhao 2010; Chen *et al.* 2011, 2012; Zhao *et al.* 2014). The Bayesian approach can make full use of empirical knowledge or historical data through prior distributions (Freni & Mannina 2010; Neuman *et al.* 2012); therefore, it can estimate parameters more accurately than traditional methods, especially when observation data are scarce. In addition, it addresses the uncertainty of the model parameters in a joint posterior distribution format (Massoudieh & Kayhanian 2013), which is convenient for analysing the risks of decision-making in water environment management (Malve & Qian 2006). Therefore, the Bayesian method is used to address the uncertainty of NPS loads in our model.

This study develops a new tool for estimating NPS loads from the point of view of source apportionment. This tool overcomes the disadvantages of needing more discharge monitoring requirements in traditional load formulas and can carry out NPS load assessments based on regular monitoring data to extensively support practical water quality management. The heavily polluted Huaihe River outflow system is invoked as a case study. The model results will help local decision makers quantify NPS loads and determine the key issue in water environment treatment.

METHODS AND MATERIALS

Flowchart of the Bayesian source apportionment mixing model

As shown in Figure 1, we simplify the model and turn it into a mathematical system by making a reasonable assumption of the physical conditions based on the conceptual model.

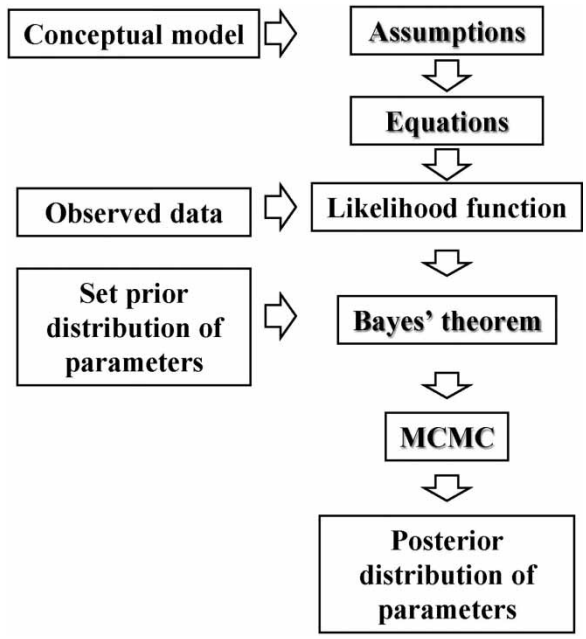

Figure 1 | The modelling flowchart.

The simulation error is introduced to transform the system into an expressed form of the random variable required in the Bayesian method, which generates the likelihood function of the model by replacing the observation data. The prior distribution of parameters is defined by historical data or experience-based information. Then, the Bayes' theorem is used to express the posterior distribution of the parameters associated with the prior distribution and likelihood function, where the Markov chain Monte Carlo (MCMC) method is used to calculate the posterior distribution of parameters, which results in the final output of the NPS load evaluation.

Source apportionment mixing model

The SAMM is based on the following assumptions. (1) All sources that have a significant contribution to the receiving area have been identified, and the source categories include point sources, NPSs and upstream sources. (2) Pollutant transport in rivers only involves the first-order attenuation coefficient and no other reactions. (3) Point sources are characterized by constant emissions concentration and flow that are well known. The mass balance equation is expressed as follows:

$$\begin{cases} C_d = p_u C_u e^{(-kT)} + p_n C_n e^{\left(\frac{-kT}{2}\right)} + \sum_{i=1}^{m} p_i C_i e^{(-kT_i)} \\ p_u + p_n + \sum_{i=1}^{m} p_i = 1 \end{cases} \quad (1)$$

where C_d represents the concentration of the downstream receiver (mg/L); T and T_i represent the in-stream travel times (d), which are calculated from the mean velocity and flow-path length, respectively; p represents the volumetric flow rate contribution of each source to the recipient sample (dimensionless), where the sum is 1; k is the first-order decay rate of the pollution species (1/d); C_u, C_i, and C_n represent the pollution concentrations at upstream, point, and NPSs, respectively (mg/L); and m represents the number of point sources (dimensionless). Pollution from NPSs is assumed to be lost over half of the stream length on average (Alexander *et al.* 2006; Chen *et al.* 2011).

Then, the downstream concentration can be extended to many pollution forms as follows:

$$\begin{cases} C_{d,1} = p_u C_{u,1} e^{\left(\frac{-k_1 X}{v}\right)} + p_n C_{n,1} e^{\left(\frac{-k_1 X}{2v}\right)} + \sum_{i=1}^{m} p_i C_{i,1} e^{\left(\frac{-k_1 X_1}{v}\right)} \\ C_{d,2} = p_u C_{u,2} e^{\left(\frac{-k_2 X}{v}\right)} + p_n C_{n,2} e^{\left(\frac{-k_2 X}{2v}\right)} + \sum_{i=1}^{m} p_i C_{i,2} e^{\left(\frac{-k_2 X_1}{v}\right)} \\ C_{d,3} = p_u C_{u,3} e^{\left(\frac{-k_3 X}{v}\right)} + p_n C_{n,3} e^{\left(\frac{-k_3 X}{2v}\right)} + \sum_{i=1}^{m} p_i C_{i,3} e^{\left(\frac{-k_3 X_1}{v}\right)} \\ C_{d,4} = p_u C_{u,4} e^{\left(\frac{-k_4 X}{v}\right)} + p_n C_{n,4} e^{\left(\frac{-k_4 X}{2v}\right)} + \sum_{i=1}^{m} p_i C_{i,4} e^{\left(\frac{-k_4 X_1}{v}\right)} \\ \cdots \\ p_u + p_n + \sum_{i=1}^{m} p_i = 1. \end{cases} \quad (2)$$

When the point source (i) emission flow is known, the NPS pollution load can be expressed as follows:

$$L_{n,j} = \frac{p_n}{p_i} q_i C_{n,j} \quad (3)$$

$L_{n,j}$ represents the NPS pollution load of species j, q_i represents the point source discharge, $C_{n,j}$ represents the NPS concentration of species j, p_n represents the fractional concentration of pollution from the NPS, and p_i represents the fraction of pollution from point source i.

$$C_{d,j} = f(C_{u,j}, C_{i,j}, L_{n,j}, k_j, X, p_u, p_i, v, q_i) \quad (4)$$

The model in Equation (4) can be viewed as a forward model from p_n to C_j. The reverse model for Equation (5) is

$$L_{n,j} = f^{-1}(C_{d,j}, C_{u,j}, C_{i,j}, k_j, X, p_u, p_i, v, q_i) \quad (5)$$

Thus, the problem of load estimation has transformed into an inverse model with the purpose of fitting the set of C_n, p_u, p_n, and p_i. Then, the Bayesian approach is used to calibrate the unknown parameters.

Bayesian approach

Because the true values of unknown parameters will never be known exactly, the goal of this research is to infer their probability distributions from known information. Therefore, the Bayesian approach is introduced into the model, and all unknown parameters are treated as random

variables with a probability distribution (Reichert & Omlin 1997). Bayes' theorem is written as follows:

$$p(\theta|\mathrm{y}) = \frac{p(\theta)p(y|\theta)}{p(y)} \tag{6}$$

where θ represents the parameter that needs to be estimated; $p(\theta|y)$ represents the posterior probability of θ, which also represents the strength of our belief in θ when the data have been taken into account; $p(\theta)$ represents the prior distribution of θ, which is defined by historical or experience-based information; and $p(y|\theta)$ represents the likelihood function, which defines the probability that the data could be generated by the model with a parameter value of θ.

To define the likelihood function of the model and by assuming that the measurement errors are random, uncorrelated, and normally distributed, a normally distributed error ε_j is added to the SAMM, as shown in Equation (7), with a zero mean and a variance of σ^2. Therefore, the likelihood function for all observed values is given in Equation (8), where z represents the number of pollution species.

$$C_j = p_u C_{u,j} e^{\left(\frac{-k_j X}{v}\right)} + p_n C_{n,j} e^{\left(\frac{-k_j X}{2v}\right)} + \sum_{i=1}^{m} p_i C_{i,j} e^{\left(\frac{-k_j X_i}{v}\right)} + \varepsilon_j, \tag{7}$$

$$\prod_{j=1}^{z} \frac{1}{\sqrt{2\pi\sigma_j^2}} e^{-\frac{\left(C_j - p_u C_{u,j} e^{\left(\frac{-k_j X}{v}\right)} + p_n C_{n,j} e^{\left(\frac{-k_j X}{2v}\right)} + \sum_{i=1}^{m} p_i C_{i,j} e^{\left(\frac{-k_j X_i}{v}\right)}\right)^2}{2\sigma_j^2}} \tag{8}$$

Because the posterior distribution is a complex, high-dimension integral that is difficult to evaluate analytically, the MCMC was utilized to solve this issue, as it is commonly applied in Bayesian analyses (Lunn *et al.* 2000; Gelman 2006; Malve & Qian 2006).

Parameter sensitivity analysis

We investigated the parameter sensitivity of our model using one-parameter-at-a-time analysis (Fasham *et al.* 1990; Yoshie *et al.* 2007). Sensitivity index (S_p), the ratio of the fractional change in the model outputs to the fractional changes in the parameter value, was introduced to describe the influence of each parameter on the modelling outputs, which can be calculated by the following equation:

$$S_p = \frac{(E(p_{\mathrm{high}}) - E(p_{\mathrm{low}}))/E(p_b)}{(p_{\mathrm{high}} - p_{\mathrm{low}})/p_b} \tag{9}$$

where $E(p_b)$ is the model outputs in the baseline run (i.e., all parameters at their calibrated values p_b), and $E(p_{\mathrm{high}})$ and $E(p_{\mathrm{low}})$ are the model outputs for when the parameter is set to its high value p_{high} (for example, 10% higher than calibrated) and its low value p_{low} (for example, 10% lower than calibrated). Because our model is a Bayesian model where the parameter and outputs are distributions, so we use the mean of the distributions of the parameter and outputs to calculate the sensitivity index.

Case study description

The Huaihe River outflow is located in the lower Huaihe River Basin in northern Jiangsu Province and is one of the five flood protection channels for Hongze Lake, as shown in Figure 2(a). The upper inflow from Hongze Lake is controlled by the Erhe floodgate. The outflow towards the East China Sea is limited by the Haikou floodgate. The Erhe floodgate has always been closed, except in July and August of 2003, July and August of 2007, and April and May of 2010, where only a small amount of leakage occurred when the gate closed. The river is 163.5 km in length, with a basin area of 1,710 km^2. The discharge of the Huaihe River mainly derives from rainfall runoff in the watershed and urban sewage water when the upper floodgate is closed. In recent years, the Huaihe River outflow has been heavily polluted by both point and NPSs. It has been listed as a key object for the prevention and control of water pollution in the Huaihe River Basin. Identifying the source of river pollutants is essential for prevention and control plans and provides a scientific basis for prevention and control measures.

Ten sets of hydrological sites (H1) and two water-quality monitoring sites (S1 and S2) from 30 November to 4 December 2016, and 11 April to 18 April 2017, were collected. Moreover, six sewage outlets with monitoring data during the same period were supplied by the Jiangsu Province

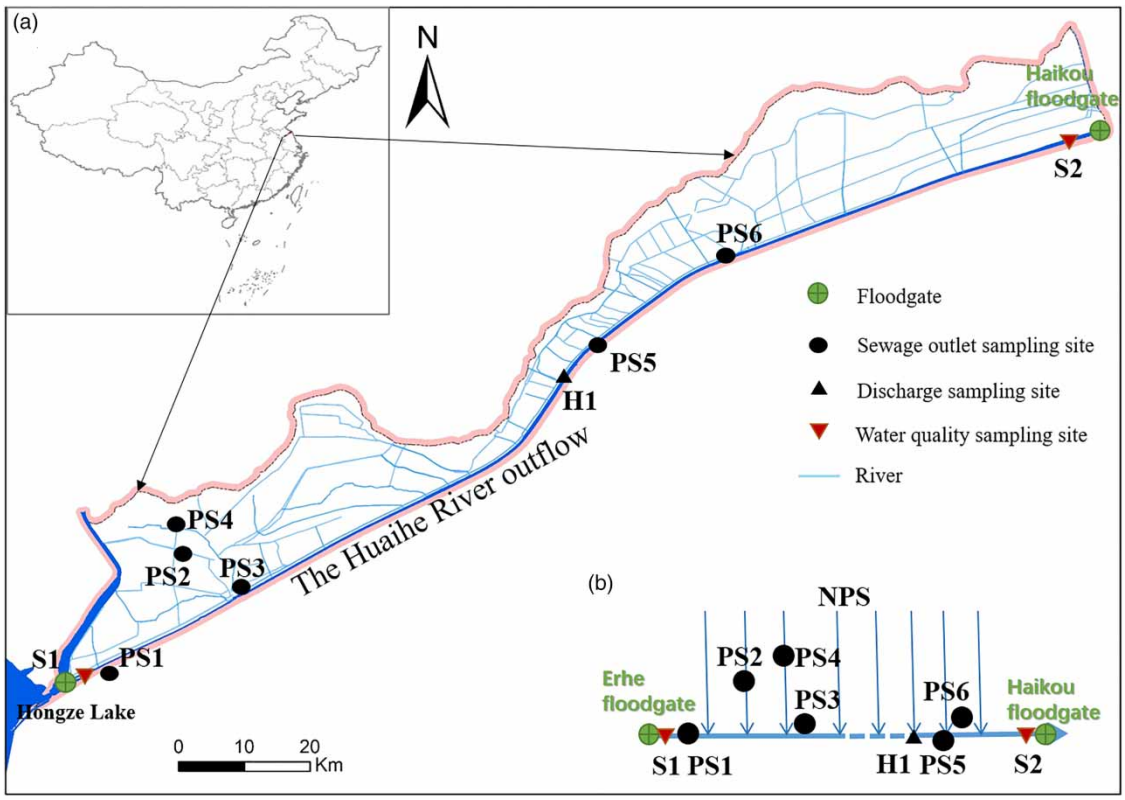

Figure 2 | Study area and sampling sites (a) and river generalizability of the study area (b).

Hydrology and Water Resources Investigation Bureau. Their spatial distributions are shown in Figure 2(a). Here, we generalized the river and point sources in the study area, as shown in Figure 2(b).

RESULTS AND DISCUSSION

Prior distribution determination and analysis

The observed data in Tables 1 and 2 for the variables considered in the model were initially analysed as references for the prior distribution of unknown parameters before the implementation of Bayesian statistics. As seen from Table 1, the water quality downstream (S2) is significantly worse than that upstream (S1), indicating that the interval flows into the river have a lower water quality than those upstream. The emission concentrations from point sources in Table 2 are far greater than those downstream (S2) in Table 1, indicating that the NPS inflow is cleaner than that of the point source. Therefore, the prior distribution of NPS pollutant concentrations is greater than that for upstream concentrations. From the above, the NPS pollutant concentrations were empirically determined as the normal distribution with the range of 0.09–8.14 mg/L for NH_4^+, 7.8–66.5 mg/L for chemical oxygen demand (COD),

Table 1 | The statistical characteristics of the sample data

Sample	Term	Median	Minimum	Maximum	S.D.
S1	NH_4^+ (mg/L)	0.10	0.09	0.54	0.14
S1	COD (mg/L)	11.05	7.80	18.00	3.17
S1	TN (mg/L)	2.61	2.09	3.00	0.32
S1	TP (mg/L)	0.08	0.02	0.28	0.08
S2	NH_4^+ (mg/L)	3.59	0.69	8.14	2.46
S2	COD (mg/L)	39.55	12.00	66.50	20.73
S2	TN (mg/L)	6.89	3.49	12.40	2.89
S2	TP (mg/L)	0.24	0.09	0.98	0.31
H1	Velocity (m/s)	0.75	0.21	1.15	0.38
H1	Discharge (m^3/s)	82.65	22.10	123.70	44.02

Table 2 | Point source information

Point source	Discharge (m^3/s)	NH_4^+ (mg/L)	COD (mg/L)	TN (mg/L)	TP (mg/L)
PS1	0.82	1.24	86.52	10.45	0.17
PS2	1.06	2.10	61.67	4.23	1.22
PS3	4.04	4.95	62.15	7.38	0.85
PS4	2.20	44.85	471.10	66.40	1.16
PS5	0.79	1.97	136.59	17.03	1.77
PS6	0.26	4.39	97.30	8.03	3.08
Weighted average[a]	9.17[b]	124.59	1557.01	204.51	9.62

[a]Weighted by the flow volume.
[b]Sum value.

2.09–12.4 mg/L for total nitrogen (TN), and 0.02–0.98 for total phosphorus (TP), in which the lowest values were taken from the upstream (S1) minimum and the highest values were taken from the downstream (S2) maximum.

To obtain the prior distribution of targeted parameters k_j in the model, a database of k_j for NH_4^+, COD, TN, and TP was compiled from the literature. Specifically, the prior distribution of k_j was determined as normal distribution (Liu *et al.* 2008; Chen *et al.* 2012), with a range of 0.005–0.70 d^{-1} for NH_4^+ (Alexander *et al.* 2000; Guo *et al.* 2008; Liu *et al.* 2008; Gao 2014; Zhang *et al.* 2015; Feng *et al.* 2017), 0.009–0.9 d^{-1} for COD (Guo *et al.* 2008; Gao 2014; Zhang *et al.* 2015), 0.005–0.5 d^{-1} for TN (Alexander *et al.* 2000; Feng *et al.* 2017), and 0.01–0.6 d^{-1} for TP (Zhang *et al.* 2015; Feng *et al.* 2017).

The in-stream travel times (T or Ti) are determined through the length between the points or NPS location with the downstream section divided by the mean velocity (Chen *et al.* 2011). The length was measured in Google Earth, and the velocity was defined as normal distribution (Liu *et al.* 2008) with mean value measured from H1 and a standard deviation of 0.01–0.02 m/s.

The prior distribution of the volumetric flow rate (p) was determined according to the point source flow and monitored flow data for H1, in which the NPS discharge was defined as normal distribution with a mean of 15 m^3/s and the interval value of 10–30 m^3/s by the measured discharge from points and H1. The prior distributions of the key parameters are summarized in Table 3.

Model calibration

The OpenBUGS software was used to develop and run this model. The observations of each pollutant from S2 were used to calibrate the model by adjusting the mean and variance of the prior distribution. The detailed data can be seen in Supplementary Material (Table S1) (available with the online version of this paper).

In order to be sure that the MCMC chain is a truly representative sample from the distribution, Gelman–Rubin Scale Reduction factor (SR) statistic (Gelman & Rubin 1992) and autocorrelation function (ACF) were used (Kruschke 2011). SR checks if the chains get stuck in unrepresentative regions of parameter space by the variability between MCMC chains

Table 3 | The prior distributions of key parameters and their value range

Symbol (unit)	Parameter	Variation range	Prior distribution
Cn1 (mg/L)	NPS concentration of NH_4^+	0.09–8.14	dnorm (2,1)
Cn2 (mg/L)	NPS concentration of COD	7.8–66.5	dnorm (25,0.04)
Cn3 (mg/L)	NPS concentration of TP	2.09–12.4	dnorm (4,1)
Cn4 (mg/L)	NPS concentration of TN	0.02–0.98	dnorm (0.2,1)
$k1$ (d^{-1})	First-order attenuation coefficient of NH_4^+	0.005–0.7	dnorm (0.09,100)
$k2$ (d^{-1})	First-order attenuation coefficient of COD	0.009–0.9	dnorm (0.15,100)
$k3$ (d^{-1})	First-order attenuation coefficient of TN	0.005–0.5	dnorm (0.10,100)
$k4$ (d^{-1})	First-order attenuation coefficient of TP	0.01–0.6	dnorm (0.09,100)
U.tau (s^2/m^2)	Reciprocal of velocity variance	400–10,000	dnorm (2500,0.01)
Qn (m^3/s)	NPS discharge	10–30	dnorm (15,0.01)

dnorm(α,β) denotes the norm distribution with the mean α and the reciprocal of variance β.

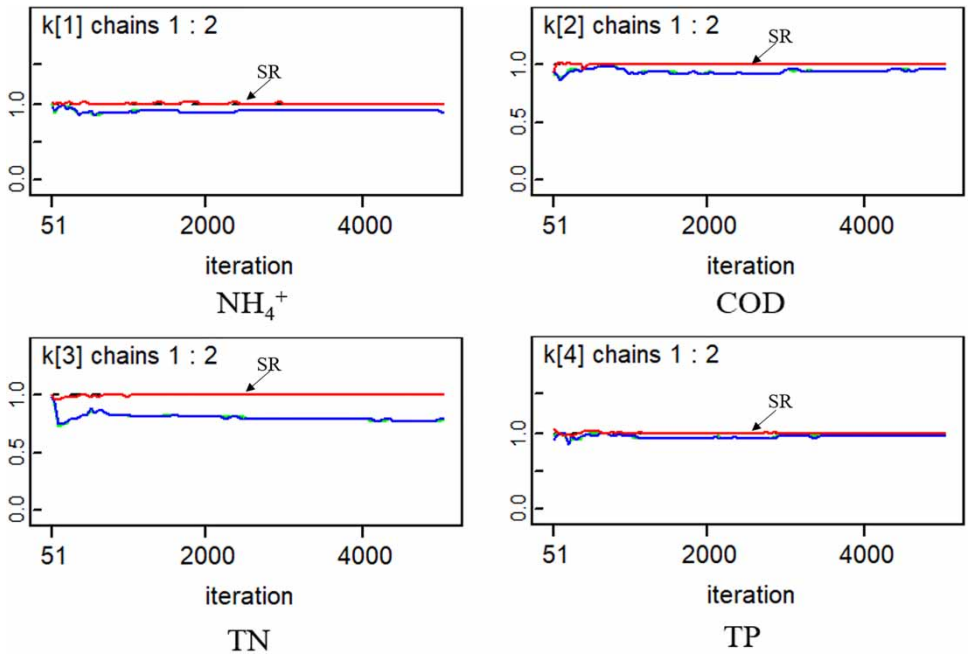

Figure 3 | SR statistic for two MCMC chains of k. The green lines represent the widths of the pooled runs, the blue lines represent the average widths within the individual runs, and the red lines represent their ratio SR (pooled/within). Please refer to the online version of this paper to see this figure in colour: http://dx.doi.org/10.2166/nh.2019.076.

relative to that within chains. If not, the SR should be around 1. ACF is to measure the correlation of the chain values with the values lag steps behind (or ahead). A highly auto-correlated chain is 'clumpy', which over-represents some values and under-represents other values. The way to mitigate the problem is to thin the chain until the ACF of lag (>0) steps close to zero. The determination coefficient (R^2) between the observed data and modelled values was calibrated, which should be close to 1.

It should be noted that the MCMC has a burn-in period due to the reduced influence of the initial value, and early simulation samples are often omitted by the Gelman–Rubin statistical diagnoses. In Figure 3, the Gelman–Rubin statistic for k chains shows that the model converges (SR close to 1) after several hundred iterations; therefore, the first 1,000 runs are discarded in our model to ensure that the chains converge. The chains were thinned to reduce their auto-correlation. The ACF for k chains with the thin step of 20 was shown in Figure 4, which shows that the chains are not 'clumpy' (ACF (lag > 0) close to 0).

The fit between the observed data and modelled values of the downstream section (S2) is calibrated by the deterministic coefficient, as shown in Figure 5. The results are acceptable when considering the complexities of pollutant transport and the measuring error. Therefore, the model can be used to assess NPS loads.

NPS pollution loads using the Bayesian source apportionment mixing model

The posterior distributions of k varied significantly among the different pollution species, as described in Figure 6. The statistics of the 10-day p_n are presented in Table 4, including the mean value, median value, standard deviation (S.D.), Monte Carlo error (MC error), and two confidence levels (2.5% and 97.5%). As shown in Table 3, the NPS contributions of p_n decrease as the water increases upstream when the Erhe floodgate opens, indicating that the model is reasonable.

The posterior distributions of the NPS loads are described in Figure 7, including the mean value and the 50% and 95% posterior confidence intervals. The mean NPS load of NH_4^+ was 20.47 g/s (0.259 g/m^3 d), which was lower than the mean value of 0.544 g/m^3 d in the study of the Hun-Taizi River system in China (Liu *et al.* 2008). The possible reason for this is that the load in the study by Liu includes point sources. The mean TN load is 39.81 g/s

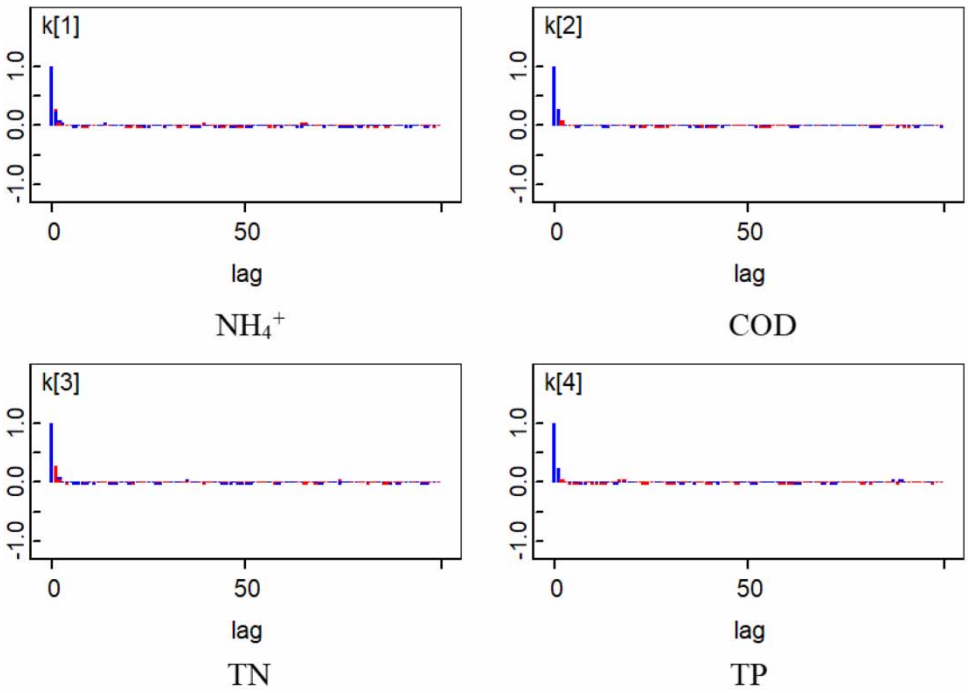

Figure 4 | The ACF for two MCMC chains of *k*.

(3.44×10^3 kg/d) and is in the range of $1.0 \times 10^3 - 36.4 \times 10^3$ kg/d in the study of the ChangLe river system in China (Chen *et al.* 2012). This result may be due to our evaluation period being the non-flood season, while the period in Chen's study was annual. The NPS load is often higher in flood season than the non-flood season (Shen & Zhao 2010; Chen *et al.* 2011).

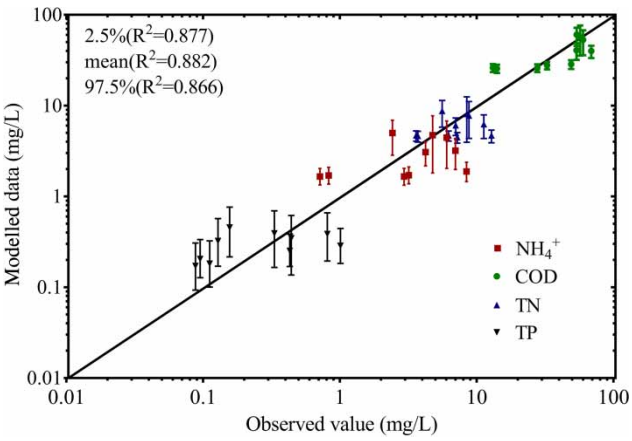

Figure 5 | The model fitting results for NH_4^+, COD, TN, and TP between observed data and the modelled mean and 2.5% and 97.5% confidence level values of S2.

The posterior distributions of the NPS loads proportional to the nonpoint and point source loads (PLns) in the river are shown in Figure 8. The total loads of NH_4^+, COD, TN, and TP from NPSs in the region accounted for approximately 16.35–27.58%, 18.78–25.69%, 21.68–29.71%, and 42.11–52.09%, respectively. It is clear that the point source load is the main source of pollution in this channel for COD, NH_4^+, and TN, but TP accounts for approximately half of the total pollution. The first national pollution source census (MOEP *et al.* 2010) shows that the proportions of COD, TN, and TP in water environments from agricultural NPS pollution are 43.7%, 57.2%, and 67%, respectively. Obviously, it is reasonable that the NPS pollution ratios in our study, where point sources are the main source of pollution, are lower than that in China.

This method can be applied on rivers with high velocity, where the algae biomass is relatively low and the uptake amount of ammonium by algae was small. However, in rivers with algae blooming and eutrophication, algae uptaking and the benthic flux should be considered to better describe the attenuation process of ammonium, and corresponding likelihood function by the Bayesian method should be developed to cover the coupled attenuation process.

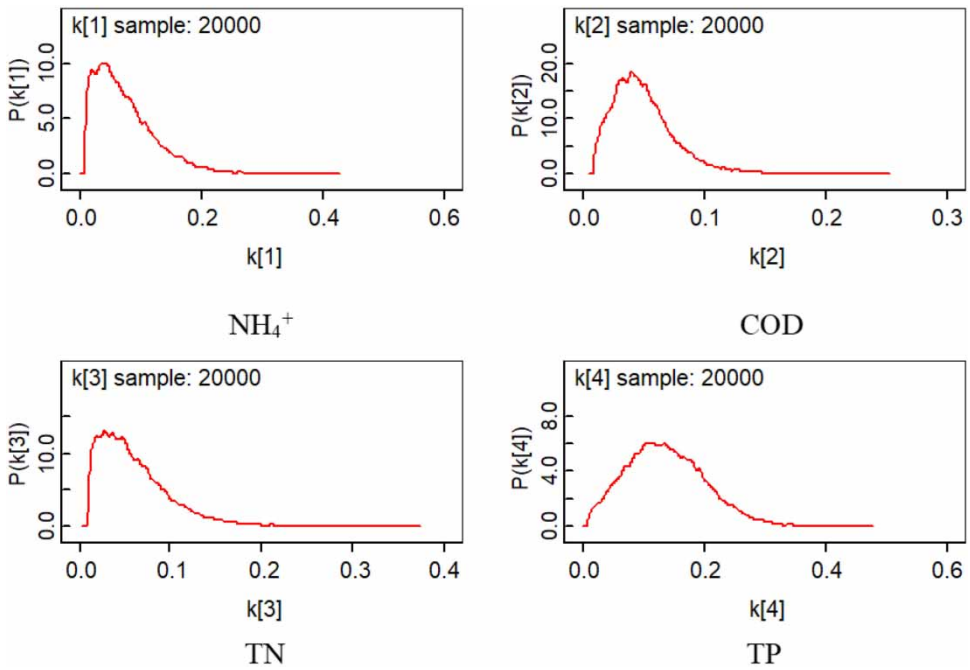

Figure 6 | The posterior distributions of *k*.

Table 4 | The posterior distribution of the NPS volumetric flow rate proportion P_n

$P_n[j]$	Mean	S.D.	MC error	2.50%	Median	97.50%
$P_n[1]$	0.5857	0.05101	5.21×10^{-4}	0.4989	0.5869	0.6679
$P_n[2]$	0.4725	0.06489	6.76×10^{-4}	0.3644	0.4727	0.5782
$P_n[3]$	0.4006	0.1036	0.001163	0.2122	0.4046	0.5731
$P_n[4]$	0.1911	0.05075	4.68×10^{-4}	0.1007	0.192	0.2759
$P_n[5]$	0.1525	0.043	4.12×10^{-4}	0.08456	0.1494	0.2339
$P_n[6]$	0.5949	0.04907	5.01×10^{-4}	0.5166	0.593	0.6794
$P_n[7]$	0.3339	0.09155	0.000952	0.1798	0.3318	0.4983
$P_n[8]$	0.1562	0.0441	4.46×10^{-4}	0.08713	0.1525	0.2409
$P_n[9]$	0.1493	0.04265	4.08×10^{-4}	0.0848	0.1452	0.2338
$P_n[10]$	0.153	0.0437	4.58×10^{-4}	0.08748	0.149	0.2413

$P_n[j]$ denotes the NPS volumetric flow rate proportion of the j_{th} monitored data set.

Parameter sensitivity analysis

To analyse the degree to which an input parameter affects the model outputs, the sensitivity index of model parameters for the model outputs PLns is calculated by Equation (9), and the results are given in Table S1. It is clear that the NPS concentration of pollution has the greatest impact on corresponding model output PLns, and the NPS discharge is the second sensitive parameter. The velocity and the first-order attenuation have low sensitivities. All the sensitivity indices are below one, which means that the model output shrinks the uncertainty of the parameters.

So, the sensitivity ranking of the input parameters sorted by the amount of influence in our model is Cn, Qn, Uh1, *k*, and U.tau. In practice, the parameters are not included in routine monitoring data, so experience or historical data should be used to define a reasonable prior distribution in

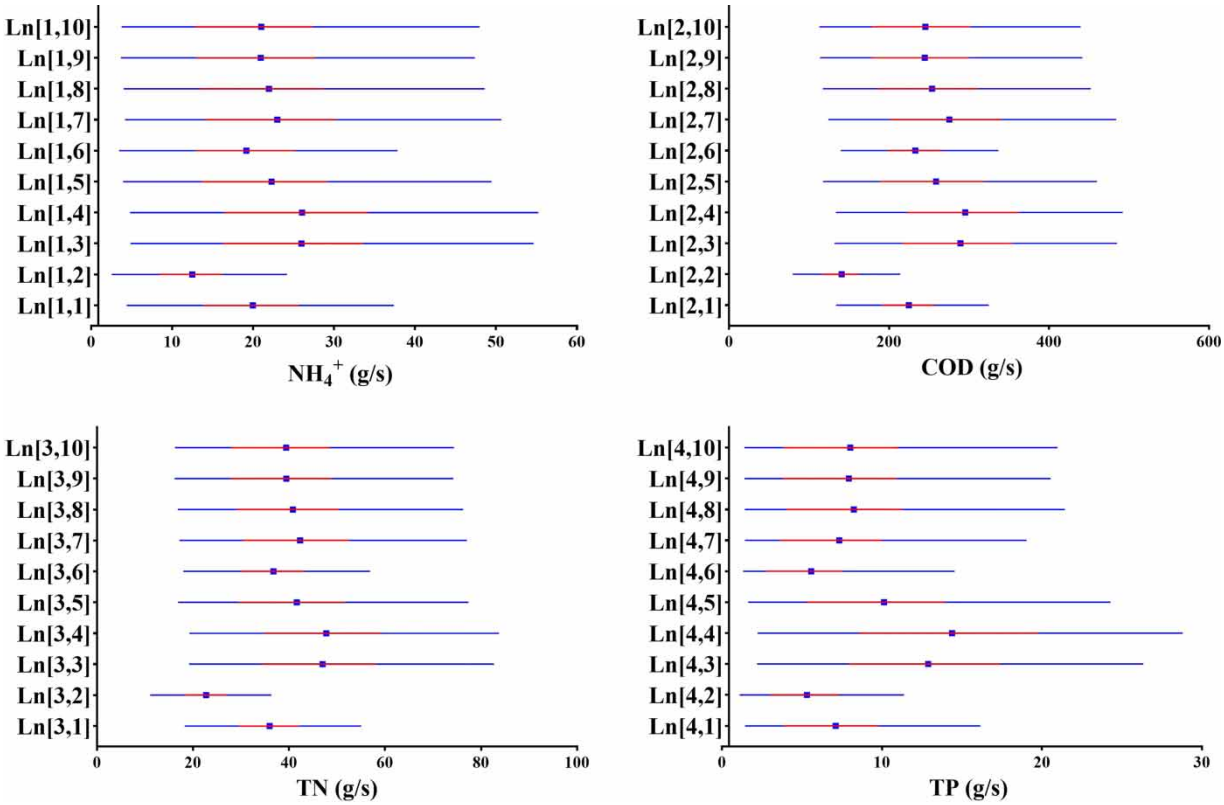

Figure 7 | The posterior distributions of the NPS loads of NH_4^+, COD, TN, and TP. $L_n[z,j]$ denotes the NPS load of the jth monitored data set for pollutant z. The solid diamonds represent the estimated posterior averages, the short red lines represent the 50% confidence intervals of the posterior distributions, and the long blue lines represent the 95% confidence intervals of the posterior distributions. Please refer to the online version of this paper to see this figure in colour: http://dx.doi.org/10.2166/nh.2019.076.

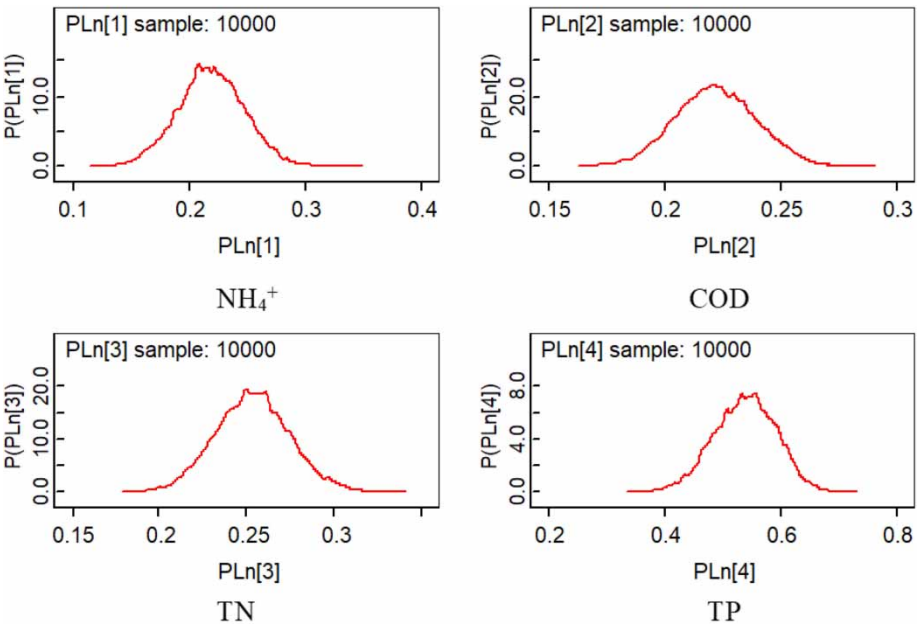

Figure 8 | The posterior distributions of NPS load proportions in pollution emissions PL_n.

order to get a precise modelling result. Besides, the more detailed the data, less uncertain is the result.

CONCLUSION

An improved inverse model based on the SAMM and coupled with the Bayesian approach for NPS load estimation was constructed. The model assumed that the pollutant concentrations in downstream samples were a mixture from various sources after attenuation, and the volumetric flow rate contributions were apportioned based on known information and pollution source emission characteristics. The model assessed NPS loads via a source analysis, which was carried out based on conventional and limited monitoring data without a sufficient number of discharge sites. The model not only considered the decrease in pollution transport but also addressed uncertainties in the estimation parameters, which are important for environmental risk management. This practical application shows that the model can serve as a simple and effective tool for researchers and managers to estimate NPS loads and uncertainties based on a small amount of water quality and hydrology data.

The Bayesian source apportionment mixing model (BSAMM) applied in this study roughly classified all pollution sources except for upstream and point source as non-point sources, which may include rural wastewater, agricultural wastewater, groundwater, and unmonitored industrial wastewater. If detailed pollution source data could be collected, more precise results regarding the source and volume of the pollution can be calculated.

Parameter sensitivity analysis shows the prior distribution of NPS concentration exert the greatest influence on modelling, and thereby sufficient empirical or historical information would facilitate a reasonable prior distribution and more precise result.

ACKNOWLEDGEMENT

This work was supported by the Major Science and Technology Program for Water Pollution Control and Treatment in China (No. 2017ZX07602-003 and No. 2014ZX07204-005).

REFERENCES

Alexander, R. B., Smith, R. A. & Schwarz, G. E. 2000 Effect of stream channel size on the delivery of nitrogen to the Gulf of Mexico. *Nature* **403** (6771), 758–761. doi:10.1038/35001562.

Alexander, R. B., Smith, R. A. & Schwarz, G. E. 2006 Comment on 'in-stream nitrogen attenuation: model-aggregation effects and implications for coastal nitrogen impacts'. *Environmental Science & Technology* **40** (7), 2485–2486. doi:10.1021/es052281m.

Chen, D., Dahlgren, R. A., Shen, Y. & Lu, J. 2012 A Bayesian approach for calculating variable total maximum daily loads and uncertainty assessment. *Science of the Total Environment* **430**, 59–67. doi:10.1016/j.scitotenv.2012.04.042.

Chen, D., Lu, J., Wang, H., Shen, Y. & Gong, D. 2011 Combined inverse modeling approach and load duration curve method for variable nitrogen total maximum daily load development in an agricultural watershed. *Environmental Science and Pollution Research* **18**, 1405–1413. doi:10.1007/s11356-011-0502-8.

Chen, D. J., Lu, J., Wang, H. L., Shen, Y. N. & Kimberley, M. O. 2010 Seasonal variations of nitrogen and phosphorus retention in an agricultural drainage river in East China. *Environmental Science and Pollution Research* **17** (2), 312–320. doi:10.1007/s11356-009-0246-x.

Fasham, M. J. R., Ducklow, H. W. & McKelvie, S. M. 1990 A nitrogen-based model of plankton dynamics in the oceanic mixed layer. *Journal of Marine Research* **48** (3), 591–639. doi:10.1357/002224090784984678.

Feng, S., Li, X. & Deng, J. 2017 Determination of comprehensive pollutants attenuation coefficients of the plain river networks in the upper reaches of Lake Taihu Basin. *Acta Scientiae Circumstantiae* **37** (3), 878–887. doi:10.13671/j.hjkxxb.2016.0125 (in Chinese).

Freni, G. & Mannina, G. 2010 Bayesian approach for uncertainty quantification in water quality modelling: the influence of prior distribution. *Journal of Hydrology* **392** (1–2), 31–39. doi:10.1016/j.jhydrol.2010.07.043.

Gao, Y. J. 2014 Research on comprehensive attenuation coefficient of pollutants in East Liaohe River. *Water Resources and Hydropower of Northeast China* **32** (1), 34–34 (in Chinese).

Gelman, A. 2006 Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis* **1** (3), 515–534.

Gelman, A. & Rubin, D. B. 1992 Inference from iterative simulation using multiple sequences. *Statistical Science* **7** (4), 457–472.

Gronewold, A. D., Qian, S. S., Wolpert, R. L. & Reckhow, K. H. 2009 Calibrating and validating bacterial water quality models: a Bayesian approach. *Water Research* **43** (10), 2688–2698. DOI: 10.1016/j.watres.2009.02.034.

Guo, R., Li, Y. B. & Fu, G. 2008 Controlling factors of degradation coefficient on organic pollutant in river. *Journal of Meteorology and Environment* **24** (1), 56–59 (in Chinese).

Kruschke, J. K. 2011 *Doing Bayesian Data Analysis: A Tutorial with R and BUGS*. Elsevier Academic Press, San Diego, CA, USA, pp. 623–624.

Lindenschmidt, K. E., Fleischbein, K. & Baborowski, M. 2007 Structural uncertainty in a river water quality modelling system. *Ecological Modelling* **204** (3), 289–300. doi:10.1016/j.ecolmodel.2007.01.004.

Liu, Y., Yang, P., Hu, C. & Guo, H. 2008 Water quality modeling for load reduction under uncertainty: a Bayesian approach. *Water Research* **42** (13), 3305–3314. doi:10.1016/J.Watres.2008.04.007.

Lunn, D. J., Thomas, A., Best, N. & Spiegelhalter, D. 2000 WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing* **10** (4), 325–337.

Malve, O. & Qian, S. S. 2006 Estimating nutrients and chlorophyll a relationships in Finnish lakes. *Environmental Science and Technology* **40** (24), 7848–7853. doi:10.1021/es061359b.

Massoudieh, A. & Kayhanian, M. 2013 Bayesian chemical mass balance method for surface water contaminant source apportionment. *Journal of Environmental Engineering* **139** (2), 250–260. doi:10.1061/(ASCE)EE.1943-7870.0000645.

Meghdadi, A. & Javar, N. 2017 Quantification of spatial and seasonal variations in the proportional contribution of nitrate sources using a multi-isotope approach and Bayesian isotope mixing model. *Environmental Pollution* **235**, 207–222. doi:10.1016/j.envpol.2017.12.078.

Ministry of Environment of the People's Republic of China (MOEP), National Bureau of Statistics of the People's Republic of China (NBOS), Ministry of Agriculture of the People's Republic of China (MOA) 2010 *The First National Pollution Source Census*. Available from: http://www.stats.gov.cn/tjsj/tjgb/qttjgb/qgqttjgb/201002/t20100211_30641.html (in Chinese).

Neuman, S. P., Xue, L., Ye, M. & Lu, D. 2012 Bayesian analysis of data-worth considering model and parameter uncertainties. *Advances in Water Resources* **36**, 75–85. doi:10.1016/j.envpol.2017.12.078.

Nikolaidis, N. P., Heng, H. H., Semagin, R. & Clausen, J. C. 1998 Non-linear response of a mixed land use watershed to nitrogen loading. *Agriculture Ecosystems & Environment* **67** (2), 251–265. doi:10.1016/S0167-8809(97)00123-0.

Reichert, P. & Omlin, M. 1997 On the usefulness of overparameterized ecological models. *Ecological Modelling* **95** (2–3), 289–299. doi:10.1016/S0304-3800(96)00043-9.

Sharifi, S., Haghshenas, M. M., Deksissa, T., Green, P., Hare, W. & Massoudieh, A. 2014 Storm water pollution source identification in Washington, DC, using Bayesian chemical mass balance modeling. *Journal of Environmental Engineering* **140** (3), 04013015. doi:10.1061/(ASCE)EE.1943-7870.0000809.

Shen, J., Jia, J.-J. & Sisson, M. A. 2006 Inverse estimation of nonpoint sources of fecal coliform for establishing allowable load for Wye River, Maryland. *Water Research* **40** (18), 3333–3342. doi:10.1016/j.watres.2006.07.035.

Shen, J. & Zhao, Y. 2009 A Bayesian approach for estimating bacterial nonpoint source loading in an estuary with limited observations. *Journal of Environmental Science and Health (A)* **44** (14), 1574–1584. doi:10.1080/10934520903263553.

Shen, J. & Zhao, Y. 2010 Combined Bayesian statistics and load duration curve method for bacteria nonpoint source loading estimation. *Water Research* **44** (1), 77–84. doi:10.1016/j.watres.2009.09.002.

Wang, Y. L., Liu, X. Y., Song, W., Yang, W., Han, Bin, Dou, X. Y., Zhao, X. D., Song, Z. L., Liu, C. Q. & Bai, Z. P. 2017 Source appointment of nitrogen in $PM_{2.5}$ based on bulk δ^{15}n signatures and a Bayesian isotope mixing model. *Tellus B: Chemical and Physical Meteorology* **69** (1), 1299672. doi:10.1080/16000889.2017.1299672.

Woodbury, A. D. & Ulrych, T. J. 2000 A full-Bayesian approach to the groundwater inverse problem for steady state flow. *Water Resources Research* **36** (8), 2081–2093. doi:10.1029/2000WR900086.

Xu, K., Wang, Y., Su, H., Yang, J., Li, L. & Liu, C. 2013 Effect of land-use changes on nonpoint source pollution in the Xizhi River watershed, Guangdong, China. *Hydrological Processes* **27** (18), 2557–2566. doi:10.1002/hyp.9368.

Xue, D., Baets, B. D., Cleemput, O. V., Hennessy, C., Berglund, M. & Boeckx, P. 2012 Use of a Bayesian isotope mixing model to estimate proportional contributions of multiple nitrate sources in surface water. *Environmental Pollution* **161**, 43–49. doi:10.1016/j.envpol.2011.09.033.

Yang, L., Han, J., Xue, J., Zeng, L., Shi, J., Wu, L. & Jiang, Y. 2013 Nitrate source apportionment in a subtropical watershed using Bayesian model. *Science of the Total Environment* **463–464**, 340–347. doi:10.1016/j.scitotenv.2013.06.021.

Yoshie, N., Yamanaka, Y., Rose, K. A., Eslinger, D. L., Ware, D. M. & Kishi, M. J. 2007 Parameter sensitivity study of the NEMURO lower trophic level marine ecosystem model. *Ecological Modelling* **202** (1–2), 26–37. doi:10.1016/j.ecolmodel.2006.07.043.

Zhang, Y. L., Shen, J., Shi, S. J., Han, L. Q. & Yao, Z. P. 2015 Dynamic change of the river pollutions' composite degradation coefficient in Huaihe River's tributary. *Environmental Monitoring in China* **31** (2), 64–67 (in Chinese).

Zhao, Y., Sharma, A., Sivakumar, B., Marshall, L, Wang, P. & Jiang, J. 2014 A Bayesian method for multi-pollution source water quality model and seasonal water quality management in river segments. *Environmental Modelling & Software* **57**, 216–226. doi:10.1016/J.Envsoft.2014.03.005.

Zhuo, D., Liu, L., Yu, H. & Yuan, C. 2017 A national assessment of the effect of intensive agro-land use practices on nonpoint source pollution using emission scenarios and geo-spatial data. *Environmental Science & Pollution Research* **25** (2), 1–23. doi:10.1007/s11356-017-0118-8.

Zou, R., Lung, W. S. & Wu, J. 2007 An adaptive neural network embedded genetic algorithm approach for inverse water quality modeling. *Water Resources Research* **43**, 2539–2545. doi:10.1029/2006WR005158.

First received 28 May 2018; accepted in revised form 28 January 2019. Available online 26 February 2019