Utilization of a Bayesian probabilistic inferential framework for contamination source identification in river environment

Lijun Jing[1], Ruihui Chen[1], Xiaomei Bai[1], Fansheng Meng[2], Zhipeng Yao[3,a], Yanguo Teng[1] and Haiyang Chen[1,a]

[1] College of Water Sciences, Beijing Normal University, No 19, Xinjiekouwai Street, Beijing, 100875, China
[2] Chinese Research Academy of Environmental Sciences, No 8, Dayangfang, Chaoyang District, Beijing, 100012, China
[3] State Environmental Protection Key Laboratory of Quality Control in Environmental Monitoring, China National Environmental Monitoring Centre, Beijing, 100012, China

Abstract. In the environmental event of hazardous release into river, quick and accurate identification of the contamination source is important for emergence response. Generally, given a noisy and finite set of monitoring information, determining the source items (i.e. location, strength and release time) is an ill-posed inverse problem. In this study, a Markov chain Monte Carlo method combined with advection-dispersion equation (ADE) was proposed for the source identification of contamination event in river system based on a Bayesian probabilistic inferential framework. Case study with analytical solution for one-dimensional ADE showed that the proposed methodology was effective and the mean posterior errors for all source parameters were lower than 3%. Case simulation based on two-dimensional ADE with numerical solution obtained similar results and further demonstrated the utility of the proposed approach for source identification. We hope the study will provide a helpful guidance to develop approach for contamination event source identification to support environmental risk management of river system.

1 Introduction

Rivers are spatially open system and inevitably vulnerable to a variety of outer risks [1]. According to the China Statistical Yearbook, a total of 9339 accidents associated with water contamination took place in China from 1997 to 2008. Among them, one of the most serious threats is the contamination injection of toxic chemical compounds into river. For instance, in November 2005, approximately 100 tons of benzene were spilled into Songhuajiang River, China and caused severe social and ecological problems [2]. Once occurring, quick and accurate identification of the contamination sources is essential to manage the emergency response and mitigate environmental consequences in the river system [3]. However, identifying source terms in environments faces challenge since it is a problem to find what happened in the past with finite set of monitoring information, such as, a limited number of observations [4, 5].

In the past two decades, several deterministic and stochastic methods have been developed successively to facilitate the solution to the source identification problem for identifying the release locations of unknown sources, and estimate the release time and emission loads [2-11]. Among them, Bayesian approach has been applied widely for source identification in recent years as it has a number of distinctive attributes [10, 11]. The method converts contamination source identification into the reiterative computation of posterior probability distribution of the source parameters and can quantify random errors in the concentration data and their uncertainties [5]. For example, Yang et al. [10] constructed differential evolution algorithm based on Bayesian inference to identify multi-point sudden water pollution sources. Compared with the deterministic calculations, the Bayesian framework estimates the posterior distribution of historical contaminants in an expanded stochastic space with current measurement concentrations and thus restates the ill-posed inverse problem as a well-posed problem [8, 12]. Generally, the posterior distribution is composed of the prior distribution and the likelihood function. The former is the information related with an uncertain environmental parameter given the concentration history and the latter represents the probability of the data.

With rapid development of urbanization and quick increasement of population growth, the river environments are facing serious challenge on the contamination problems. Protecting ecological safety of river system has become a security issue because the demand for clean water is increasing in the world. Although contaminant source identification has been studied in river and lake pollution cases [3, 4], limited publications have been found for quickly identifying contamination event source in river system with a very finite set of monitoring information.

In this study, a Bayesian probabilistic inferential framework combined with advection-dispersion equation (ADE) was proposed for source identification of

[a] Corresponding author: superbg@163.com; yaozp@cnemc.cn

contamination event in river system given a set of limited concentration data. Bayesian inference, which can provide the probability density function (PDF) of the source parameters, has been used to formulate the problem of source event determination and the Markov Chain Monte Carlo (MCMC) stochastic sampling method with Metropolis algorithm was employed to generate the posterior distributions of source parameters. Summary statistics associated with each variable can be estimated once a series of MCMC samples have been obtained. Finally, two hypothetical cases were conducted to demonstrate the developed methodology.

2 Methods

2.1 Bayesian formulation

By assuming the parameter vector m which indicates the leak characteristics: m (x_0, y_0, q_0, t_0), where {x_0, y_0} represents the spatial location of contamination source, q_0 is its strength, and t_0 is the release duration, the posterior probability for the m vector with the concentration C at measurement point and prior information I can be formulated as following:

$$P(m \mid C, I) = \frac{P(m \mid I)P(C \mid m, I)}{P(C \mid I)} \quad (1)$$

In the posterior distribution, the role of $P(C|I)$ is nothing more than a normalizing constant as C is the known data. Therefore, the posterior probability function $P(m|C,I)$ is proportional to the prior probability $P(m|I)$ and the likelihood function $P(C|m,I)$.

$$P(m \mid C, I) \propto P(m \mid I)P(C \mid m, I) \quad (2)$$

2.1.1 Assignment of the prior density function

$P(m|I)$ is the prior probability for the m vector and hypothetically satisfy the independent uniform distribution in environmental hydraulic model. Generally, this probability is set as a constant:

$$P(m \mid I) = \text{constant}, \ m \in \Re \quad (3)$$

Given a bounded estimation domain $\Re$, m (x_0, y_0, q_0, t_0) $\in \Re$, for example, the source location x_0 is assumed to be greater than zero but less than the length of river.

2.1.2 Assignment of the likelihood function

The likelihood function $P(C|m,I)$ was employed herein to represent the probability between an array of measurements and a certain set of observed concentrations. Commonly, the difference between the measured concentrations (C) and modelled values (R) largely arises from the measurement and model errors [11].

$$C_i = C_{true,i} + e_{meas,i} \quad (4)$$

$$R_i = C_{true,i} + e_{\mathrm{mod}el,i} \quad (5)$$

where, $C_{true,i}$, $e_{meas,i}$, $e_{model,i}$ are the true value of measurement concentration, the measurement noise and the model error at the ith observation point, respectively.

Generally, the noise of measurement error is generally assumed to be normally distributed with mean of zero and variance of σ_1^2 [13], and then can be defined as following:

$$P(C \mid C_{true}, I) \propto \exp\left(\sum_{i=1}^{n}\left[-\frac{(C_i - C_{true,i})^2}{2\sigma_1^2}\right]\right) \quad (6)$$

where n is the number of observations. Likewise, the difference $e_{model,i}$ between the modelled concentrations and true values is also assumed to be normally distributed with mean of zero and variance of σ_2^2:

$$P(C_{true} \mid m, I) \propto \exp\left(\sum_{i=1}^{n}\left[-\frac{(C_{true,i} - R_i(m))^2}{2\sigma_2^2}\right]\right) \quad (7)$$

Therefore, the likelihood function is then obtained by marginalizing the joint probability density function of C and C_{true} with respect to C_{true}:

$$P(C \mid m, I) = \int P(C \mid C_{true}, I)P(C_{true} \mid m, I)dC_{true} \propto \exp\left(\sum_{i=1}^{n}\left[-\frac{(C_i - R_i(m))^2}{2(\sigma_1^2 + \sigma_2^2)}\right]\right) \quad (8)$$

2.1.3 The posterior probability density function

Based on the prior distribution (Eq. 3) and likelihood function (Eq. 8), the relation of posterior probability density function $P(m|C,I)$ can be expressed as following:

$$P(m \mid C, I) \propto P(m \mid I)P(C \mid m, I) \propto I(m \in \Re)\exp\left(\sum_{i=1}^{n}\left[-\frac{(C_i - R_i(m))^2}{2(\sigma_1^2 + \sigma_2^2)}\right]\right) \quad (9)$$

2.2 Markov chain Monte Carlo

There is no doubting it will be difficult to analytically compute the posterior distribution with Eq. 9. In the majority of cases, we are interested in the highest density region of the posterior distribution. Therefore, the Markov chain Monte Carlo (MCMC) simulation is adopted in this study to estimate the posterior distribution of the parameter vector m by constructing the posterior density from which the mean, median, and other statistical analysis can be done.

Essentially, the MCMC algorithm can be used to explore the state space of a random parameter using Markov chain mechanism and generate samples from the posterior distribution while spending most of the sampling steps in the highest density regions of the posterior state space [10, 14]. In other words, it generates a point series as a chain, and the distribution of these points follows the posterior probability density function $P(m|C,I)$. For a set of parameter vector m (x_0, y_0, q_0, t_0), the posterior probability is computed in a circular manner using the MCMC algorithm until the chain has converged on the target distribution. Generally, each new Markov chain m_k depends on the

previous part m_{k-1} and will be accepted if it improves the posterior probability density of the previous set [15]. Based on the MCMC results, a statistical analysis including histogram, mean and deviation can be obtained for each parameter.

2.3 Metropolis algorithm

Metropolis algorithm, a special case of Metropolis-Hastings [10, 16, 17], was adopted in this study. Its proposal distribution (p) satisfies symmetrical random sampling, $q(m^{(*)}|m^{(i)})=q(m^{(i)}|m^{(*)})$, and the acceptance probability (A) of Markov chain ranges from $m^{(i)}$ to $m^{(*)}$ is as follows.

$$A(m^{(i)}, m^{*}) = \min\left\{1, \frac{p(m^{*})q(m^{(i)} \mid m^{*})}{p(m^{(i)})q(m^{*} \mid m^{(i)})}\right\} = \min\left\{1, \frac{p(m^{*})}{p(m^{(i)})}\right\} \quad (10)$$

where $m^{(i)}$ is the sample generated in the ith iteration for parameter vector m.

The procedure flow of Metropolis algorithm is shown in Figure 1.

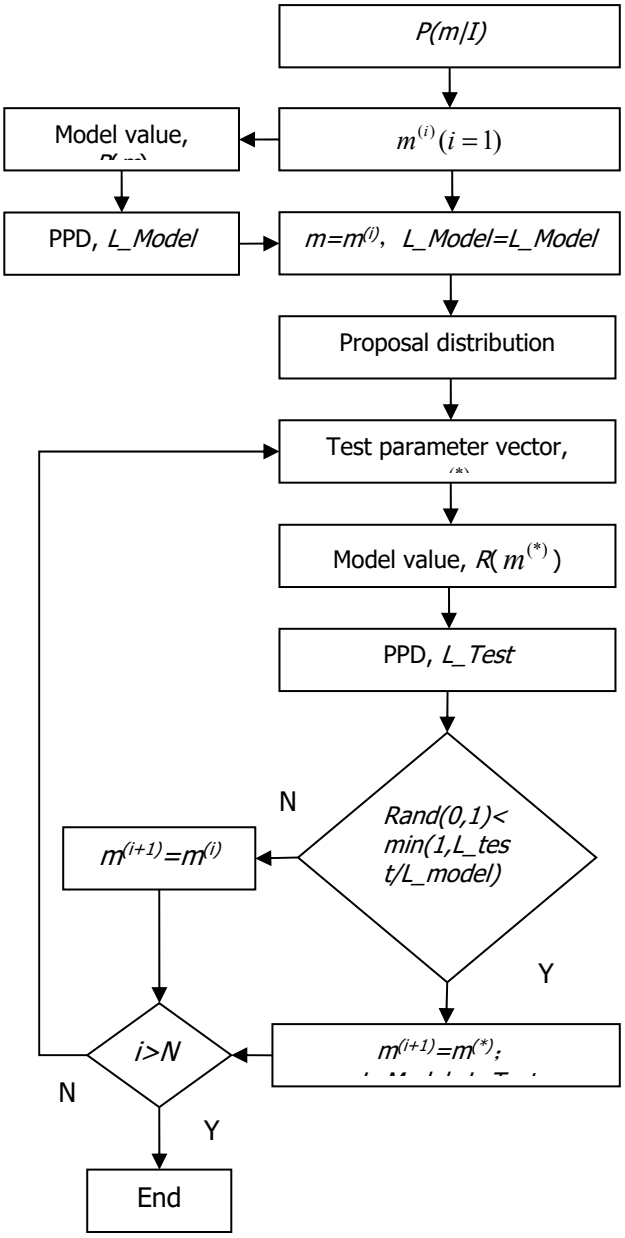

Figure 1. Flow of MCMC with Metropolis algorithm. PPD: posterior probability density; *m(i)* is the sample generated in the ith iteration for parameter vector m; *u* is a random number generated from the standard uniform distribution U(0, 1).

3 TWO CASES

In this study, two hypothetical cases were used to demonstrate the developed methodology for identifying contamination sources. Forward simulation of the spilled chemical compounds in a predetermined river was conducted to generate the concentration data set which was then regarded as the "measurements" from the observation points. Source identification was preformed based on the Bayesian probabilistic inferential framework combined with advection-dispersion equation, assuming that the location of spill location (x_0, y_0), the release duration (t_0) and the initial total leak mass (q_0) were unknown.

3.1 Case study Ⅰ: analytical solution for one-dimensional advection-dispersion equation

The one-dimensional transport processes of a mass concentration, C, in the fluid flows is generally governed by the advection-dispersion equation:

$$\begin{cases} \dfrac{\partial C}{\partial t} = E_x \dfrac{\partial^2 C}{\partial x^2} - u\dfrac{\partial C}{\partial x} - kC + \sum_{i=1}^{q} M_i \delta(x - x_i) \\ C(0,t) = 0, C(L,t) = 0 \ , t \in (0,T); C(x,0) = 0 \ \ , x \in (0,L) \end{cases} \tag{11}$$

where t is the forward time, s; u is the flow velocity, $m \cdot s^{-1}$; E_x represents the dispersion or diffusion tensors in x directions, $m^2 \cdot s^{-1}$; k describes the reaction coefficient, d^{-1}; q is the source number; M_i is the source strength of the ith source; δ is the Dirac delta function; x_i is the location of the ith source; L is the river length domain, m; T is the time domain, s. As descripted by Wu et al. [18], the advection-diffusion Eq. 11 can be solved using a brute-force approach.

In this scenario, a kind of contaminant (k=0.1 d^{-1}) with the total mass of 1.0×10^3 kg (q_0) was assumed to be released into a river at the point S and transported to the downstream river channel about 1.0×10^4 m. Two hours later (t_0), the contaminant plume was found and limited water samples (i.e. five) were quickly collected from five observation points (O_1, O_2, O_3, O_4 and O_5). Point S was defined as coordinate origin, and the distance from point S to point O_1 was 5.0×10^3 m (x_0). The distance among each observation point was 500m. The dispersion tensors in x directions was 0.6 $m^2 \cdot s^{-1}$; the mean velocity was about 0.36 $m \cdot s^{-1}$. Then, the source parameter vector of the contaminant event that should be identified is $m\ (x_0,\ q_0,\ t_0)$. Their actual values were 5.0×10^3 m, 1.0×10^3 kg, and 7.2×10^3 s, respectively.

The structural parameters (σ_1, σ_2 and proposed distribution step) were optimized using the iterative Expectation-Maximization method, a general iterative scheme for estimating the marginal PDF of a parameter from the joint PDF of the parameter and vector **m**. In specific, the optimization for the variances of measurement noise and model error was performed by using the likelihood portions of objective function. Meanwhile, the proposal distribution step was optimized using the prior PDF portion. In the presented study, the optimized values for σ_1, σ_2 and proposed distribution step were 0.01, 0.01 and percent 5 of prior density range, respectively.

Based upon the optimization results, the metropolis algorithm was conducted to run for 10 000 propositions of set of parameters. The iteration curves of source parameters showed that the Markov chain of all modeled parameters became convergence after 2000 iterations. Further, in order to verify the MCMC results, the posterior distribution was calculated directly, and the marginal distribution of each source parameter was evaluated by numerically integrating the full posterior distribution over the remaining source parameters, shown in Figure 2. The solid vertical line represents the true parameter value, and the dashed line is the mean of the MCMC samples. Shaded regions represent 95% highest probability density intervals based on the MCMC samples.

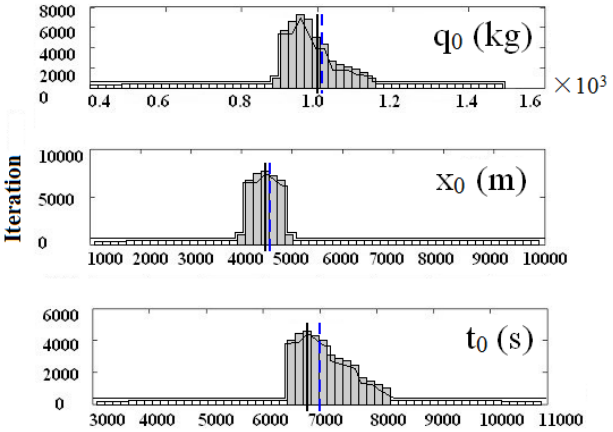

Figure 2. Posterior histogram of model parameters generated from MCMC samples with analytical solution for one-dimensional advection-dispersion equation.

It can be seen the stair-step appearance of the marginal distributions for q_0, x_0, t_0 was concentrated near to the initial assumed value. The posterior histogram of total leak mass q_0 had maximum probability near to 1.0×10^3kg, which was very well resolved. The posterior probability density of source location x_0 was concentrated in the range of 4.8×10^3 and 5.3×10^3 m (actual value= 5.0×10^3 m). With regard to the spilled time t_0, its maximum posterior probability was distributed at the range from 6.5×10^3 to 8.0×10^3 s, which included the actual value 7.2×10^3 s.

To determine the robust of the Bayesian methodology, Monte Carlo simulation was utilized to generate twenty data sets for random model parameters with errors of 10%. Results showed that the method of Bayesian computation for source event identification was steady and robust. The average errors of modeled parameters (q_0, x_0, t_0) for the twenty simulations were 2.9%, 2.5% and 2.2%, respectively. Relatively, the source strength had somewhat higher variation ranged from 0.5% to 8.5%, followed by source location x_0 (2.5%, 5.9%) and spilled time duration t_0 (2.2%, 4.7%).

3.2 Case study Ⅱ: numerical solution for two-dimensional advection-dispersion equation

The two-dimensional transport processes of a mass concentration, C, in the fluid flows are governed by the advection-dispersion equation:

$$\frac{\partial C}{\partial t} + u_x \frac{\partial C}{\partial x} + u_y \frac{\partial C}{\partial y} = D_x \frac{\partial^2 C}{\partial x^2} + D_y \frac{\partial^2 C}{\partial y^2} + KC + \sum_{i=1}^{q} s_i \delta(x - x_i) \tag{12}$$

Where t is the forward time, s; u_x and u_y are the depth-integrated velocity components in x and y directions, respectively, $m \cdot s^{-1}$; D_x and D_y represent the dispersion or diffusion tensors in x and y directions, respectively, $m^2 \cdot s^{-1}$; K describes the reaction coefficient, d^{-1}; q is the source number; s_i is

the spilled rate of the *i*th source; x_i is the location of the *i*th source.

The above equation can be solved with finite element method by which partial differential equations can be solved approximately [19]. In the present research, ANSYS software (v10.0) which offers a comprehensive field of engineering simulation was employed to compute the modelled concentration.

In this scenario, the total spilled mass was 5.3 tons (q_0) and the contaminant plume was found one hour (t_0) later since the mass-release. Ten time histories of "measured" concentration were generated with ANSYS software (v10.0) at intervals of 5 minutes as the assumed observations. The dispersion tensors in *x*, *y* directions was 0.6 and 0.01 $m^2 \cdot s^{-1}$, respectively, and the mean velocity in x, y directions was about 0.36 and 0.002 $m \cdot s^{-1}$, respectively. Then, the contaminant event source parameters to identify were x_0, y_0, q_0 and t_0. Their actual values were 500m, 35m, 5.3×10^3kg and 3.6×10^3s, respectively.

Likewise, the Expectation-Maximization approach was employed to optimize the structural parameters. With the optimization values ($\sigma_1 = \sigma_2 = 0.01$; percent 5 of prior density range for the proposal distribution step), the metropolis algorithm was conducted to run for 10 000 propositions of the set of parameters. Markov chain of all target parameters became converged after 800 iterations. Figure 3 shows the posterior histogram of model parameters generated from MCMC samples. Similarly, the solid vertical line and dashed line represent the true parameter value and the mean of MCMC samples, respectively. It can be seen that the stair-step appearance of the marginal x_0, y_0, q_0, t_0 distribution was concentrated near to the initial assumed value.

Although the posterior histogram showed that the Bayesian methodology combined with the numerical solution for two-dimensional advection-dispersion equation could yield a good approximation for the source terms, including the total discharge mass q_0, the source location (x_0, y_0) and the contamination elapsed time t_0, it was important to further assess the identification errors related to the contamination event source. It can be seen from Figure 3 that the chain was quickly converged after 800 iterations. Therefore, the first 800 "burn-in" samples were discarded conservatively, and the remaining 9200 samples were used to generate the posterior statistical analysis. The results demonstrated that the Bayesian methodology generated an honest identification of the contamination event in that the source items were accurately predicted according to the available context parameters, such as the data set of 'measured concentrations'. The posterior histogram of source location (x_0, y_0) had maximum probability near to (484m, 33.9m) with mean error of (3.14%, 3.17%) and the mean posterior PDF for source strength q_0 concentrated between 5.2×10^3kg and 5.5×10^3kg with mean of 5.3×10^3kg. Regarding the spilled time t_0, its maximum posterior probability distributed at the range from 3.5×10^3 s to 3.9×10^3 s, which included the actual value 3.6×10^3 s. These results suggested that the agreement between the MCMC results and the posterior distribution was good.

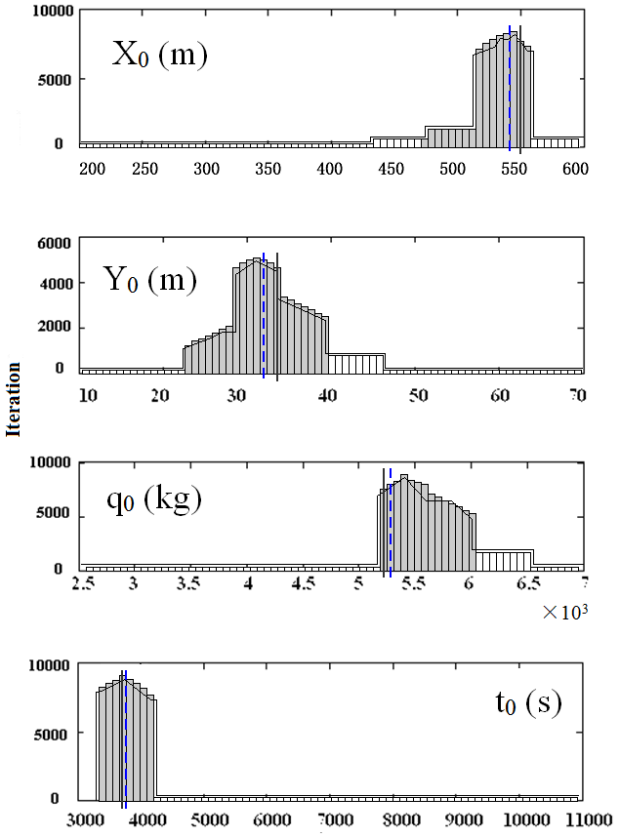

Figure 3. Posterior histogram of model parameters generated from MCMC samples with numerical solution for two-dimensional advection-dispersion equation.

4 CONCLUSIONS

During the emergency response of accidental incidents in river system, it is important to quickly determine the contamination source terms using limited measured information for protecting the water resources and minimizing the social losses. This study proposed a Bayesian probabilistic inferential framework to identify the contaminant event sources in river by combining with advection-dispersion equation given a finite monitoring data. Regarding both measurement and model errors in the probability density function, the method was coupled with a fast stochastic sampling algorithm to promote the convergence efficiency for quickly determining the source term parameters. Results of the two case simulations showed that the proposed methodology was effective and robust for source event identification on chemical spill in river environment. The posterior mean errors for all source parameters were lower than 5%, which demonstrated the utility of the approach for inverse source identification.

Acknowledgments

This study was financially supported by Major Science and Technology Program for Water Pollution Control and Treatment of China

(2017ZX07302002) and Beijing Natural Science Foundation (8172030).

References

1. A. Azizullah, M.N. Khattak, P. Richter, D. Hader, Environ. Int. **37**, 479-497 (2011)
2. Z. Li, X.Z. Mao, T.S. Li, S.Y. Zhang, Adv. Water Resour. **88**,68-79 (2016)
3. A. Ghane, M. Mazaheri, J.M.V. Samani, J. Environ. Manage. **180**,164-171 (2016)
4. W.P. Cheng, Y.F. Jia, Adv. Water Resour. 33, 397-410 (2010)
5. J.B. Wang, J.S. Zhao, X.H. Lei, H. Wang, Environ. Pollut. **241**, 759-774 (2018)
6. J. Atmadja, A.C. Bagtzoglou, Water Resour. Res. **37**, 2113-2125 (2001)
7. A.M. Michalak, P.K. Kitanidis, J. Hydraul Res. **42**, 9-18 (2004)
8. J.B. Wang, N. Zabaras, Int. J. Heat Mass Tran. **49**, 939-950 (2006)
9. T. Xu, J.J. Gomez-Hernandez, Water Resour. Res. **52**, 6587-6595 (2016)
10. H. Yang, D. Shao, B. Liu, J. Huang, X. Ye, Stoch. Environ. Res. Risk Assess. **30**, 507-522 (2016)
11. A. Keats, E. Yee, F.S. Lien, Atmos. Environ. **41**, 465-479 (2007)
12. A. Hazart, J.F. Giovannelli, S. Dubost, L. Chatellier, Signal Process. **96**, 346-361 (2014)
13. C.A. Kastner, A. Braumann, P.L. Man, S. Mosbach, G.P. Brownbridge, J. Akroyd, C. Himawan, Chem. Eng. Sci. **89**, 244-257 (2013)
14. S.P. Brooks, J. R. Stat. Soc. **47**, 69-100 (2010)
15. E. Yee, F.S. Lien, A. Keats, R.D.J. Amours, Wind Eng. Ind. Aerod. **96**, 1805-1816 (2008)
16. N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, E. Teller, J. Phys. Chem. Solids **21**, 1087-1091 (1953)
17. W.K. Hastings, Biometrika. **57**, 97-109 (1970)
18. Z.K. Wu, H.M. Fan, X.R. Chen, J. Hydrodyn. **23**, 121-125 (2008)
19. Y.F. Jia, S.S. Wang, Y.C. Xu, Int. J. Eng. Sci. **3**, 57-71 (2002)