MATHEMATICAL MODELING AND COMPUTING, Vol. 9, No. 2, pp. 311–317 (2022)

Stochastic machine learning modeling for the estimation of some uncertain parameters. Case study: Retardation factor in a radionuclide transport model

El Yamani M. A., Lazaar S.

Mathematics, Computer Sciences and Applications Team (ERMIA), University of AbdelMalek Essaadi, ENSA of Tangier, Morocco

(Received 13 October 2021; Accepted 7 February 2022)

In the present work, we define a stochastic model using machine learning techniques to generate random fields of some uncertain parameters. The proposed stochastic model is based on Bayesian inference and aims at reconstituting the parameters of interest and their credible intervals. The main goal of this work is to define a model that estimates the values of the uncertain parameters known only by their distribution probability functions and some observed spatial measurements. We note that this type of parameters may be associated with some mathematical models usually traduced by non-linear differential equations. In our case, we study the uncertainty of the retardation factor in a radionuclide transport model. To achieve a more realistic parameter estimation, Markov chain Monte Carlo (MCMC) algorithms are applied. We demonstrate that the obtained results confirm the feasibility of our proposed model and lead to a new understanding of contaminants' behavior.

Keywords: *Bayesian approach, machine learning, Metropolis–Hasting algorithm, partial differential equations, retardation factor, stochastic modeling.*

2010 MSC: 81T80, 62F15, 35A17, 62P12 **DOI:** 10.23939/mmc2022.02.311

1. Introduction

Because of industrial, agricultural and consumption-based lifestyles, the groundwater environment is being attacked by an increasing number of contaminants. As a result, it poses serious threats to the environment and human health. In order to create strategies for the protection of groundwater, the numerical simulation of transport contaminants is therefore essential to understand their behaviour in porous media. In the numerical study of complex phenomena, the modeling of transport contaminants is usually expressed by partial differential equations (PDE) that represent an important tool in understanding the behaviour of these phenomena in time and space [1–5]. It is due to the mathematical modeling of these phenomena through PDE targeting extremely precise forecasts. One of the critical part in the resolution of PDE is the estimation of the associated uncertain parameters that influence hugely the solution accuracy. As a result, several methods have been suggested in the literature for estimating uncertain parameters. Among the most popular methods, we find the Bayesian estimation [6–9]. In [6], authors propose the Bayesian estimation method to extract the geometrical and the electrical parameters used in CMOS pixelated nanoscale biosensor platforms. The same method has been developed in [8] to estimate the kinetic rate parameters of lactoferrin-mediated iron transport across blood–brain barriers from posterior probability density functions. As an alternative, many researchers proposed the Least Square Method [10,11] to estimate uncertain parameters. In [12], authors used the Maximum Likelihood method to estimate random parameters. This last proposed method was applied to experimental tumor data and the corresponding results were compared in terms of effectiveness to the results obtained when using Nonlinear Least Squares approach. The main conclusion is that the principle of Maximum Likelihood can give more reliable predictive results for individual tumor when combined with a Hidden Markov Model.

In [13], authors defined a conditional generator of parameter random fields coupling to an unconditional generator of parameters based on A. Mikhailov algorithm [14] (known as Palm process) and a kriged interpolator at observed localisations. This last work was an extension of an already published work [15–17] which deals with an unconditional random field generator. The objective of this present study is to apply a machine learning technique to predict the retardation factor related to radionuclide contaminants transport model in groundwater. This estimation will be achieved using Bayes' theorem coupled with the Metropolis–Hastings (MH) algorithm.

In section 2 of this paper, we describe briefly the transport of radionuclide contaminants equation. In section 3, we discuss in details the Bayesian inference for parameter estimation. Section 4 illustrates the application of the proposed methodology to simulate the retardation factor associated to the radionuclied transport model. We close this paper with a conclusion and some directive ideas for future work.

2. Radionuclide transport model

Several phenomena are likely to carry radionuclides from a deep geological repository into the biosphere, through the various containment barriers. Consequently, the migration of radionuclides in groundwater can be described by the advection and diffusion–dispertion equation [16] and may be written as:

$$\theta R_f \frac{\partial(c)}{\partial t} = -\nabla(qc) + \nabla[\theta(D_m + D_p)\nabla c] - \theta R_f \lambda c + S, \quad \Omega \times (0,T), \tag{1}$$

where $c = c(z,t)$ represents the concentration of the radionuclide per unit volume of fluid at location $z \in \Omega$ and time $t \in [0,T]$, R_f is the retardation factor which represents the sorption to the soil; θ is the porosity; q is the Darcy velocity; D_m is the molecular diffusion tensor; D_p is the dispersion tensor; λ is the radioactive decay constant; S is the source term and $T > 0$ is the final time of observation.

Retardation factor. The long-term safety of storing nuclear waste in deep geologic repository mainly depends on the ability of natural and artificial barriers to slow down long-lived radionuclides from the possible migration to the biosphere. As a result, the retardation factor R_f appearing in equation (1) is successfully applied in studies of the migration of radionuclide components in groundwater. Therefore, it has a significant impact in the solution of the equation (1). For this purpose, a machine learning technique have been developed to estimate the retardation factor R_f based on the Bayesian inference. The retardation factor can be expressed as:

$$R_f = \left(1 + \frac{K_D\,\rho}{\eta}\right), \tag{2}$$

where K_D represents the distribution coefficient; ρ is the density of the soil and η is the porosity.

3. Bayesian inference for parameter estimation

The Bayesian inference is a statistical analysis based on using probability to represent all form of uncertainty. It is essentially based on two different types of information. First, the distribution of the observed data conditional on its parameters. This distribution is described by a likelihood function $L(\theta, Y) = p(Y|\theta)$, where θ represents the uncertain parameter to be estimated and Y represents the samples. The second one is the prior information about parameter value that express one's beliefs about this parameter before some evidence is taken into account. This information can be obtained from published papers, from past studies or from expert knowledge. Since the prior information of θ is fraught with uncertainty; it is modeled through a prior distribution with probability density function denoted $\pi(\theta)$. The Bayesian estimation of θ is to calculate the average of the so-called posterior distribution $\pi(\theta|Y)$ that represents the conditional probability of the parameters given the measurement data Y, resulting from the prior distribution $\pi(\theta)$ and the likelihood function $L(\theta, Y)$ according to Bayes' theorem. The posterior distribution can be expressed as:

$$\pi(\theta|Y) = \frac{p(Y|\theta) \times p(\theta)}{p(Y)}, \tag{3}$$

where $p(Y|\theta)$ denotes the likelihood function and $p(Y)$ is the marginal density of the data representing the evidence and it can be expressed as:

$$p(Y) = \int_{\theta} p(Y|\theta)p(\theta). \tag{4}$$

Since the evidence $p(Y)$ is known to be constant and difficult to be calculated, the equation (3) can be rewritten with a proportional expression as:

$$\pi(\theta|Y) \propto p(Y|\theta) \times p(\theta). \tag{5}$$

In addition, the likelihood function is expressed as $L(\theta, Y) = p(Y|\theta)$. Consequently:

$$\pi(\theta|Y) \propto L(\theta, Y) \times p(\theta). \tag{6}$$

We underline two big challenges to implement Bayesian approach: The first one is the specification of the prior probability distribution which expresses the subjective beliefs and the subjective uncertainty about the parameter; and the other one is the computing of the posterior distribution. In general, for a complex problem, the functional form of the posterior distribution is unknown and is difficult to be calculated. To overcome this difficulty, the Markov Chain Monte Carlo (MCMC) approaches can be used.

MCMC and the M–H algorithm. In Bayesian estimation, when the explicit computation of the posterior distribution of the parameter to be estimated is very complex, we use the Markov Chain Monte Carlo methods to provide the samples approximately distributed according to the law distribution. The Monte Carlo method relied to the generation of random number according to the proposal distribution. Moreover, the Markov Chain is a sequence of random numbers where each number is depending to the previous number in the sequence. It is used to generate a new value of the random variable from the proposal distribution with mean equal to the previous value of the random variable. According to equation (6), several values of the posterior distribution are generated. For that reason, the Metropolis–Hastings algorithm (M–H) is used to decide which proposed values of the estimated parameter θ^* are to be accepted or rejected. The implementation of the M–H algorithm requires a good specification of the proposal distribution $\pi(\cdot)$ and an initial state $\theta^{(0)}$. The M–H algorithm is described here after:

Algorithm 1 Metropolis Hastings.

Require:
The prior distribution $\pi(\theta)$.
The likelihood distribution $p(Y|\theta)$ of the measurement Y given the parameter value θ.
The number of sample N_s to be generated.

Ensure: Proposal distribution for sampling new values of the parameter;
Initialization : $\theta := \theta_0$;
for $k = 1 : N_s$
  Generate the next sample θ^* according to the proposal distribution;
  Compute the posterior probability of the new value θ^*;
  Compute the posterior probability of the previous value θ_{t-1};
  Compute the ratio :
  $r(\theta^*, \theta_{t-1}) = \frac{p(Y|\theta^*)\times\pi(\theta^*)}{p(Y|\theta_{t-1})\times\pi(\theta_{t-1})}$;
  The acceptance probability : $\alpha(\theta^*, \theta_{t-1}) = min[r(\theta^*, \theta_{t-1}); 1])$;
  if $\alpha = 1$ **then**
    $\theta_t := \theta^*$;
  else
    Draw u from Uniform(0,1);
    if $u < \alpha(\theta^*, \theta_{t-1})$ **then**
      $\theta_t := \theta^*$;
    else
      $\theta_t := \theta_{t-1}$;

To implement the M–H algorithm, the first step is to specify the likelihood function and the prior distribution. Let $\theta^{(t)}$ be the t^{th} sampling iteration. We start with an arbitrary initial parameter $\theta^{(0)}$, and we deduce the next value of the parameter θ^* according to the proposal distribution. In our synthetical case study, we use a lognormal distribution with a mean equal to the previous accepted parameter θ_{t-1} and variance of 6% of θ_{t-1} [7], the 6% of the variance is obtained from the synthetical observed values. After that, we compute the posterior distribution of the new value θ^* according to equation (6) and we compare it to the posterior probability of the previous value θ_{t-1}. If the acceptance probability α is equal to 1, then θ^* is accepted and we continue with the next value; if not, we generate a variable "u" with a uniform distribution $U_{[0,1]}$ and we compare it with the acceptance probability α; if "u" is less than the acceptance probability α, then θ^* is accepted; otherwise, we keep the old value θ_{t-1} and we restart sampling θ^*.

4. Numerical simulations and discussion

Geological statistical measurements have shown that the retardation factor R_f associated to the transport model of contaminants is distributed according to the lognormal law $\mathrm{Ln}(\mu, \sigma)$. For that, let consider a stochastic machine learning model for the prediction of R_f assumed to follow a lognormal distribution with unknown mean μ and unknown standard deviation σ. The estimation of the parameters (μ, σ) lead to the estimation of the uncertain parameter R_f. In order to apply the proposed methodology, we consider a synthetical prior function $\pi(R_f)$ follow a lognormal distribution with mean $\mu = 4$ and variance $\sigma^2 = 0.5$. Considering a likelihood function $p(X|R_f)$ assumed to be a lognormal distribution of mean $\mu = 3$ and variance $\sigma^2 = 1$. The synthetical observation values of the studied parameter R_f are $X_i = (33.71, 20.14, 7.16, 4.44, 18.68)$; they are generated randomly according to $p(X|R_f)$. The likelihood function can be expressed as follows:

$$p(X|R_f) = \frac{1}{x(\sigma_{R_f})\sqrt{2\pi}} \exp\left[-\frac{(\ln(x)-\mu)^2}{2(\sigma^2_{R_f})}\right]. \tag{7}$$

The details of the simulation using MH-algorithm is represented on table 1 and table 2. We note σ the standard deviation, A the acceptance rate, B the efficiency average, C the log marginal likelihood, D the Monte Carlo Stantard Error, M the median and $[a, b]$ the 95% credible interval.

Table 1. Machine learning data based on MCMC iteration started from five synthetical observation values.

MCMC iterations	Burn-in period	MCMC sample size	A	B	C
12500	2500	10000	0.4499	0.2347	−29.342521

Table 2. Comparative simulation of the estimated parameter using different types of prior distributions.

Prior Distribution	μ	σ	D	M	$[a, b]$
Lognormal	2.935173	0.4087142	0.009229	2.928029	[2.147885; 5.738067]
Uniform	4.124898	0.1172293	0.004267	4.089218	[4.00323; 4.425667]

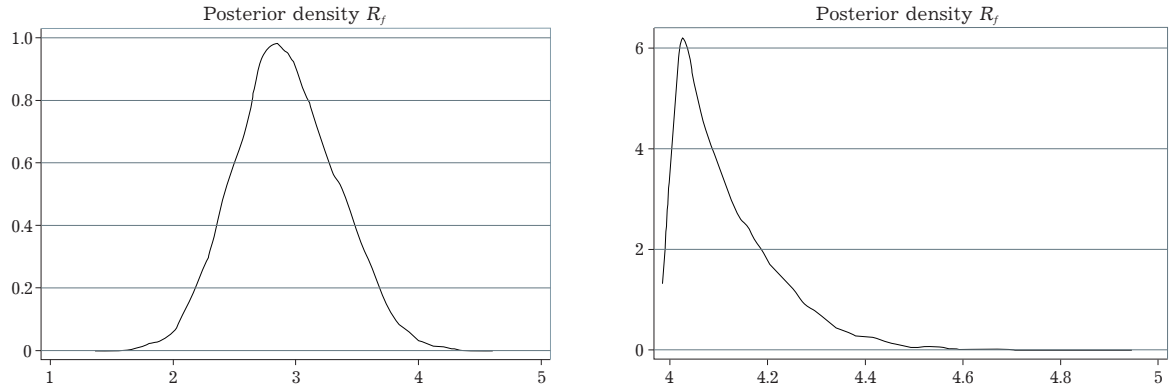

Fig. 1. Posteriori distribution of the estimated parameter R_f obtained using the Bayesian simulation with Lognormal prior distribution (left figure) and Uniform prior distribution (right figure).

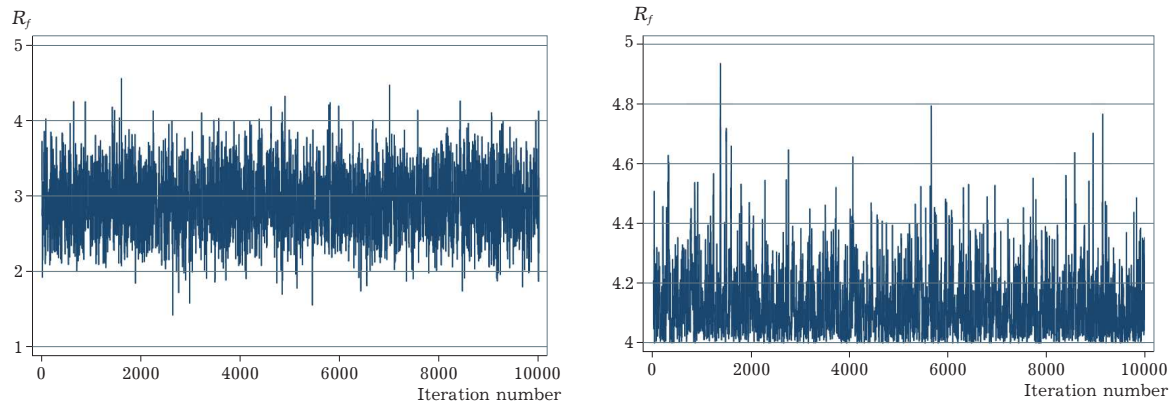

Fig. 2. Representation of the estimated parameter R_f obtained using the Bayesian simulation with Lognormal prior distribution (left figure) and Uniform prior distribution (right figure).

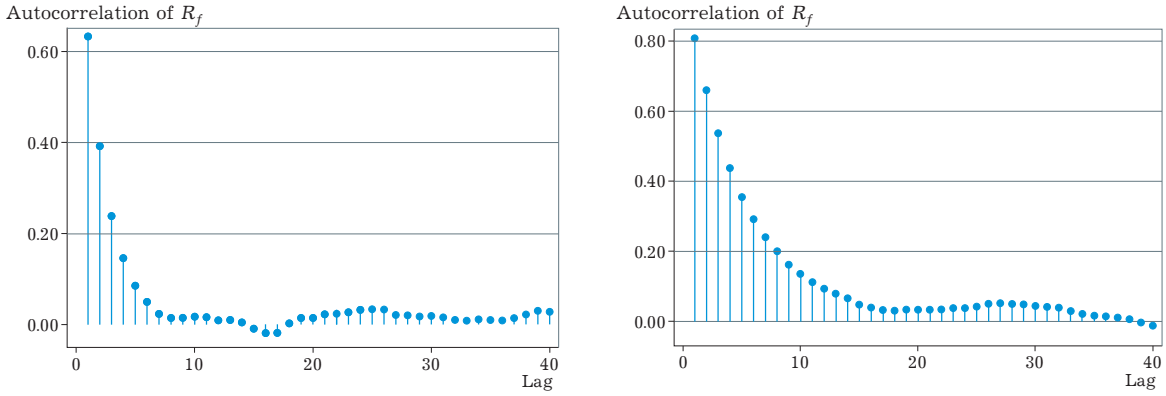

Fig. 3. Autocorrelation functions of the generater parameter simulated using Bayesian inference with Lognormal prior distribution (left figure) and Uniform prior distribution (right figure).

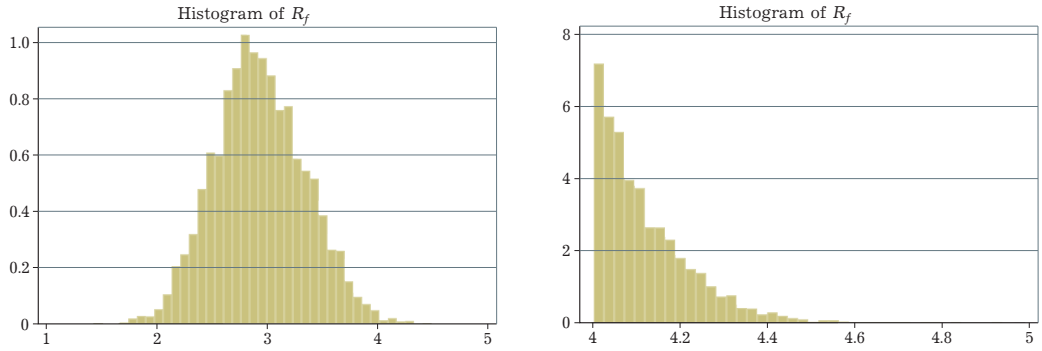

Fig. 4. Histrogram of the generated parameter using Metropolis Hastings algorithm with Lognormal prior distribution (left figure) and Uniform prior distribution (right figure).

The Monte Carlo standard error (D) obtained from the Bayesien model is very low, and demonstrates that the number of samples is sufficient for reaching a high numerical precision of the estimator.

5. Conclusion

In this study, we presented a machine learning technique for analyzing the uncertainty of the retardation factor in a radionuclide transport model. This uncertainty was estimated using the Bayesian inference implemented with Metropolis Hasting algorithm. The density function of the estimated uncertain parameter (Fig. 4) is similar to the proposal distribution. The autocorrelation function dies-off quickly

which explains the huge heterogeneity of the retardation factor. Among the advantages of the proposed method, we can mention the ability to draw strong conclusions with prior knowledge about what we are measuring and with small data sets. Moreover, when the prior distribution is uniform, the Bayesian inference becomes the maximum probability method, because this last does not take into account the prior distribution, which means that the Bayesian inference is more advanced than the maximum probability. A future study will aim at comparing the results of this model with other machine learning models based on neural networks.

[1] Ndaïrou F., Area I., Nieto J. J., Torres D. F. M. Mathematical modeling of COVID-19 transmission dynamics with a case study of Wuhan. Chaos, Solitons and Fractals. **135**, 109846 (2020).

[2] Ma Y., Xiao X., Yu W., Shang W., Tan P., Wu Z., Ni M. Mathematical modeling and numerical analysis of the discharge process of an alkaline zinc-cobalt battery. Journal of Energy Storage. **30**, 101432 (2020).

[3] Kareem W. A., Izawa S., Klein M., Fukunishi Y. A hyperbolic partial differential equation model for filtering turbulent flows. Computers & Fluids. **190**, 156–167 (2019).

[4] Locatelli L., Binning P. J., Sanchez-Vila X., Søndergaard G. L., Rosenberg L., Bjerg P. L. A simple contaminant fate and transport modelling tool for management and risk assessment of groundwater pollution from contaminated sites. Journal of Contaminant Hydrology. **221**, 35–49 (2019).

[5] Majee S., Shit G. C. Modeling and simulation of blood flow with magnetic nanoparticles as carrier for targeted drug delivery in the stenosed artery. European Journal of Mechanics – B/Fluids. **83**, 42–57 (2020).

[6] Stadlbauer B., Cossettini A., Morales J. A., Pasterk E. D., Scarbolo P., Taghizadeh L., Heitzinger C., Selmi L. Bayesian estimation of physical and geometrical parameters for nanocapacitor array biosensors. Journal of Computational Physics. **397**, 108874 (2019).

[7] Choi W., Kikumoto H., Choudhary R., Ooka R. Bayesian inference for thermal response test parameter estimation and uncertainty assessment. Applied Energy. **209**, 306–321 (2018).

[8] Khan A. I., Liu J., Dutta P. Bayesian inference for parameter estimation in lactoferrin-mediated iron transport across blood-brain barrier. Biochimica et Biophysica Acta (BBA) – General Subjects. **1864** (3), 129459 (2020).

[9] Fang T., Mackillop W., Jiang W., Hildesheim A., Wacholder S., Chen B. E. A Bayesian method for risk window estimation with application to HPV vaccine trial. Computational Statistics & Data Analysis. **112**, 53–62 (2017).

[10] Ji Y., Jiang X., Wan L. Hierarchical least squares parameter estimation algorithm for two-input Hammerstein finite impulse response systems. Journal of the Franklin Institute. **357** (8), 5019–5032 (2020).

[11] Li M., Liu X. The least squares based iterative algorithms for parameter estimation of a bilinear system with autoregressive noise using the data filtering technique. Signal Processing. **147**, 23–34 (2018).

[12] Patmanidis S., Chignola R., Charalampidis A. C., Papavassilopoulos G. P. A comparison between Nonlinear Least Squares and Maximum Likelihood estimation for the prediction of tumor growth on experimental data of human and rat origin. Biomedical Signal Processing and Control. **54**, 101639 (2019).

[13] El Yamani M. A., Lazaar S. Conditional Assessment of Uncertain Parameters Using Palm Probabilistic Approach and Kriging Interpolation. AI2SD 2019: Advanced Intelligent Systems for Sustainable Development (AI2SD'2019). 27–33 (2020).

[14] Mikhailov G. A. Monte Carlo methods for solving problems with stochastic parameters. Russian Journal of Numerical Analysis and Mathematical Modelling. **2** (2), 137–157 (1987).

[15] Lazaar S. Random fields for uncertain parameters related to a transport model – Monte Carlo method. Seminar at "Service de Métrologie Nucléaire". University Libre of Brussels, Belgium (2000).

[16] Díaz M. J. M., Lazaar S., Ortegón G. F. On the numerical simulation of uncertain parameters in a radionuclide transport model. Comptes Rendus Mathematique. **345** (7), 415–420 (2007).

[17] Lazaar S. A numerical simulation of some uncertain parameters related to a radionuclide transport model. Seminar at Departamento de ecuaciones diferenciales y análisis numérico of Seville University, Spain (2010).

Стохастичне моделювання машинного навчання для оцінки деяких невизначених параметрів. Приклад: коефіцієнт уповільнення в моделі поширення радіонуклідів

Ель-Ямані М. А., Лазар С.

Група математики, комп'ютерних наук та додатків (ERMIA), Університет АБДЕЛМАЛЕКУ ЕССААДІ, ENSA з Танжеру, Марокко

У цій роботі визначено стохастичну модель із застосуванням методів машинного навчання для створення випадкових полів з деякими невизначеними параметрами. Запропонована стохастична модель заснована на баєсовому висновуванні і спрямована на відновлення шуканих параметрів та їхніх достовірних інтервалів. Основною метою даної роботи є визначення моделі, яка б оцінювала значення невизначених параметрів, відомих лише за їхніми функціями ймовірності розподілу та деякими просторовими спостережуваними вимірюваннями. Зауважимо, що цей тип параметрів може бути пов'язаний з деякими математичними моделями, які зазвичай описуються за допомогою нелінійних диференціальних рівнянь. У нашому випадку вивчається невизначеність коефіцієнта уповільнення в моделі поширення радіонуклідів. Для досягнення більш реалістичної оцінки параметрів застосовуються алгоритми Монте-Карло марковських ланцюгів (MCMC). Продемонстровано, що отримані результати підтверджують доцільність визначення запропонованої нами моделі та призводять до нового розуміння поведінки забруднювачів.

Ключові слова: *байєсовий підхід, машинне навчання, алгоритм Метрополіса–Гастінгса, рівняння з частинними похідними, коефіцієнт уповільнення, стохастичне моделювання.*