Estimating the accuracy of the Random Walk Simulation of Mass Transport Processes

Xuefei, Wu[1]

[1] Department of Engineering, Southwest University of Science and Technology

Mianyang, China

Dongfang, Liang[2]

[2] Department of Engineering, University of Cambridge

Cambridge, UK

Geliang Zhang[3*]

[3*] Southwest University of Finance and Economics

[*] E-mail: zgl@swufe.edu.cn

ABSTRACT

The mass transport processes always accompanies the flow phenomena and have attracted many researches. A lot of numerical methods have been developed to study them. These numerical methods can be classified into the Eulerian and the Lagrangian approaches. The Lagrangian approach has advantages in high stability and simplicity over the Eulerian approach, but suffers from heavy computational cost. In this paper, we are mainly concerned with the trade-offs between the accuracy and computational cost when applying the random walk method, which is a Lagrangian approach for examining the mass transport scenario. We introduce a linear model to assess the accuracy of the random walk method in several computational configurations. Studies on computational parameters, i.e. the size of time step and number of particles, are conducted with the focus on estimation of the longitudinal dispersion coefficient D_L in steady flows. The results show that the proposed linear model can satisfactorily explain the computational accuracy, both in sample and out-of-sample. Furthermore, we find a constant dimensionless parameter, which quantifies a generic relationship between the accuracy and the number of particles regardless of the flow and diffusion conditions. This dimensionless parameter is of theoretic value and offers guidelines for choosing the correct computational parameters to achieve the required numerical accuracy.

KEYWORDS: Random walk method; mass transport; error analysis; longitudinal dispersion coefficient; weak convergence;

statistical differential equation.

1. INTRODUCTION

Mass transport phenomena in fluid exist widely. Hence, extensive studies have been conducted by many researchers since the middle of the last century. Generally, the numerical methods for mass transport problems can be classified into either Eulerian or Lagrangian approaches. The Eulerian approach solves the mass transport equations on a control volume basis which is in a similar form as that for the flow field calculation. Consequently, Eulerian approach is grid-based and has gained its popularity on studying mass transport in finite computational domain (Benkhaldoun et al., 2007; Benson et al., 2017; Liang et al., 2010). The selection of the size of time step Δt in Eulerian approach often depends on the grid size Δx. For example, the size of the time step is restricted by the Courant-Friedrichs-Lewy condition for pure advection equation, i.e. Δt must be less than the time taken for the fluid with varying state to travel to adjacent grid points. For the pure diffusion equation, the explicit scheme such as the central differencing scheme requires $\Delta t \leq (\Delta x)^2/2D$, where D is the diffusion coefficient. The relative dominance of the advection or diffusion constraint can be assessed using the Peclet number Pe given by $Pe = (U \cdot \Delta x)/D$, where U is the local flow velocity.

On the other hand, the Lagrangian approach, a powerful method from individual particles perspective, has been widely applied as a counterpart of the Eulerian one. Although the flow field is traditionally solved through Eulerian approach, it has also been increasingly solved by Lagrangian methods. For example, the incompressible Navier-Stokes Equations have been solved by the smoothed particle hydrodynamics (SPH) method (Shao and Gotoh, 2005; Shao and Lo, 2003). The Lagrangian approach uses a large number of discrete massless particles to represent the pollutant cloud and tracks the pathway of each individual particle. The concentration, as well as other parameters such as the dispersion coefficient, can be obtained by studying the statistics of these particles' trajectories or their total ensemble. By definition, the Lagrangian approach is perfectly conservative and free from artificial diffusion near the steep concentration gradients. Besides, this mesh-free scheme limits its computation to the regions that the pollutant reaches, while the computation in Eulerian approach always needs to cover the entire flow domain regardless of the presence of the pollutant. Some more merits of Lagrangian approaches have been reported in recent years. (Zhang and Chen, 2007) found the Lagrangian approach performed better than the Eulerian one in the unsteady state condition. Saidi et al. (2014) concluded that Eulerian method cannot be applied to problems involving low concentration of particles while its Lagrangian counterpart can well detect the particles. Wu and Liang (2019) and Yang et al. (2019) compared the two algorithms through designed cases and found the Lagrangian approach achieved higher accuracy.

The advantages of the Lagrangian method listed above are accompanied by the high computational cost, which has been well recognized (Möbus et al., 2001; Neuman, 1993; Zhang and Chen, 2007). Furthermore, much less guidance has been reported concerning the selection of computational parameters in the Lagrangian approach than that in the Eulerian ones. As seen in the description of the method, the amount of the computation depends on two parameters: the number of particles N to present the pollutant cloud, and the size of the computational time step Δt. Therefore, the choices of N and Δt are crucial to the efficiency of the random walk method. According to many simulation cases in different aquatic environments, the more particles applied and the smaller the time step is, the more accurate and stable the simulation results will be, although these often lead to greater computational load. Hence, it is necessary to optimize the selection of the number of particles and the size of the time step. The random walk method studied in this paper is a typical representative of the Lagrangian approach. Unfortunately, there is little literature mentioning a detailed and optimized selection of these parameters for the random walk method applied in hydraulics and hydrodynamics.

Therefore, in this paper we empirically studied the impact of particles number and size of time step on the accuracy of longitudinal dispersion coefficient in the random walk method for steady flow, with the aim to minimize its computational cost and control the error of simulation for future studies. First, the theory of random walk method is introduced. Then an error model of N and Δt is presented, which is commonly used for Stochastic Differential Equations (SDE). In the current mathematical framework of the accuracy analysis in SDE approximations, the coefficients of N and Δt in the error model do not have analytical expressions and thus need to be estimated by case studies. The estimation of coefficients is usually difficult in practical situations, which requires careful experiment design in addition to huge computational cost. In this study, we have conducted the error analysis based on two types of steady flow for this model: Couette flow and open channel flow. In both types of steady flow, a dimensionless parameter of N is found to be a constant, shedding light on the possibility of the priori optimization of the simulation accuracy . For the Couette flow, it is found that the error model degrades into a simpler form, i.e. linear relationship with N by a constant dimensionless parameter, while Δt has little impact on the results as long as it does not exceed a threshold. In the end, we attempted to give some theoretically explanations for the model. For potential applications, findings of this paper can be used to accelerate the SDE-related simulations in water research.

2. THEORETICAL BACKGROUND

2.1. Random Walk Method

The random walk method originates from statistical physics which has been used to model movement in a wide variety of contexts: from time series of financial markets (Hamid et al., 2017; Mishra et al., 2015) to the dispersion in porous media (de Anna et al., 2013; Sole-Mari et al., 2017). In the simulation of mass transportation, a large number of particles are released to represent the pollutant cloud in the flow. The trajectories of each particle are tracked, then the streamwise variation of the ensemble of particles will reveal the dispersion rate, i.e. the longitudinal dispersion coefficient. The process is illustrated in Fig. 1, with takes the fully-developed turbent open channel flow as an example. An ensemble of particles are released at time t=0, and then move along the x-axis. The histogram of the particles reveals the concentration at different times after the release, as shown in the bottom subfigure.

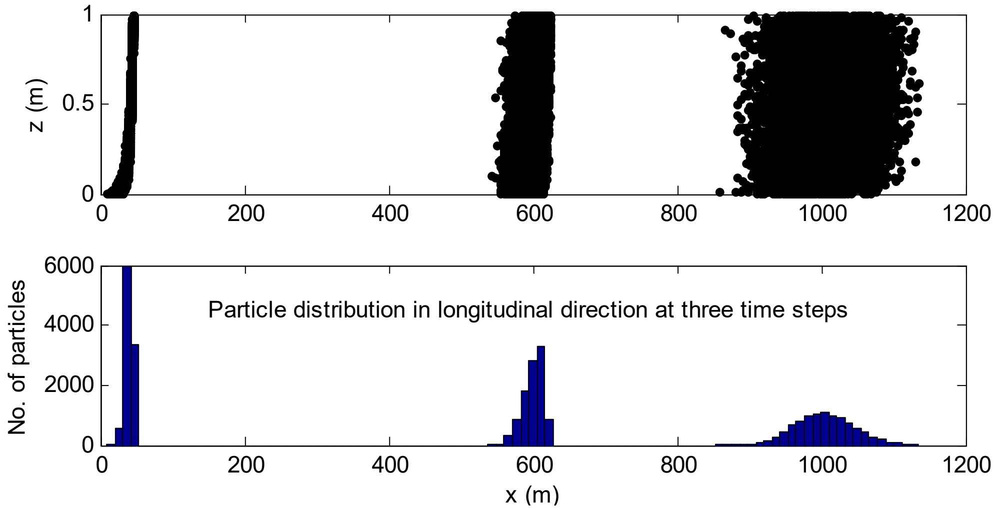


Figure. 1 Illustration of mass transportation simulation via the random walk method.

The displacement of each particle during each time step in the random walk method is described by Eq. 1. To simplify the derivation, we take the one-dimensional case here.

$$dx = a(x(t),t)dt + b(x(t),t)dW(t) \tag{1}$$

which consists of a deterministic component $a(x(t),t)dt$ and a random component $b(x(t),t)dW(t)$. W(t) here denotes a standard Wiener process, while x(t) denotes the x-axis position of each particle at time t. We introduce

$p(x,t)$ as the conditional probability density for $x(t)$, which subjects to the Fokker-Planck Equation as:

$$\frac{\partial p}{\partial t}+\frac{\partial}{\partial x}[a(x,t)p]=\frac{1}{2}\frac{\partial^2}{\partial x^2}[b(x,t)^2 p] \tag{2}$$

A more thorough analysis can be referred to the classic book in the field of statistics (Gardiner, 2004). The above analysis shows that the distribution of particles that move according to Eq. 1, which is the Ito stochastic differential equation (SDE), satisfies the Fokker-Planck equation, i.e. Eq. 2. Meanwhile, Eq.2 is similar in form to the mass transport equation (the probability density $\mathbf{p}$ is equivalent to the concentration $\mathbf{c}$), which is the foundation of the diffusion-advection phenomenon, shown as Eq. 3:

$$\frac{\partial c}{\partial t}+\frac{\partial}{\partial x}(u\cdot c)=\frac{\partial}{\partial x}(D\frac{\partial c}{\partial x}) \tag{3}$$

where $\mathbf{D}$ is the diffusion coefficient; u is the velocity of the flow; t is time. By now, we connected the Ito SDE with the mass transport equation, which is the bedrock of the random walk method.

In hydraulics, we use the longitudinal dispersion coefficient (denoted as D_L) to quantitatively analyze the mixing rate of pollutants in shear layers, which is a key parameter in water-quality modeling. As a measure of the spatially-averaged spreading rate of a tracer cloud, D_L could be determined by analyzing the statistics of the positions of a large number of particles. After a certain time has elapsed since the release of the particle ensemble, known as the Fickian limit, the standard deviation of particles' positions in the longitudinal direction increases linearly with time, so that D_L converges into a constant. By then D_L can be calculated through the time change of the longitudinal variance of the particle ensemble, as:

$$D_L=\frac{1}{2}\frac{\sigma_x(t_2)-\sigma_x(t_1)}{t_2-t_1} \tag{4}$$

In this paper, we conduct error analysis on the simulation of D_L , which is a typical example of weak convergence approximation for SDEs (Peter E. Kloeden, 2007). A strong convergence approximation for SDEs is needed when the trajectories of the ensemble of particles are taken into considerations, while a weak convergence approximation is applicable when only the distribution of the particles is concerned. The longitudinal dispersion coefficient can be calculated from the particle distribution and its development with time, rather than from the trajectories. Hence, the weak convergence analysis is adopted here.

2.2. Error Analysis

In the field of mathematics and finance, the statistical features of SDE have been well developed. First, the

SDE in Eq. 1 is discretized by a simple Euler discretization with time step Δt as Eq. 5:

$$x_{n+1} = x_n + a(x_n, t_n)\Delta t + b(x_n, t_n)\Delta W_n \tag{5}$$

In our case, D_L depends on the distribution of the particle ensemble, which is effectively related to the time-varying probability distribution $p(x,t)$, as described in Eq. 2 and Eq. 4. To calculate D_L, we want to compute the expectation of $f(x_T)$, i.e. $E[f(x_T)]$, where $f(x)$ is a scalar function with a uniform Lipschitz bound (Giles, 2008). To be specific in our context, $f(x_T)$ is the function that maps the ensemble of particles into D_L at instant T, where T indicates the time after the Fickian limit. To obtain $E[f(x_T)]$, the simplest estimation would be the mean of the discrete values $f(x_{T/\Delta t})$, from N independent path simulations, as shown in Eq. 6:

$$Y = N^{-1}\sum_{i=1}^{N} f(x_{T/\Delta t}^{(i)}) \tag{6}$$

In computational mathematics, it is well established that, provided that $a(x,t)$ and $b(x,t)$ satisfy certain conditions (Bally and Talay, 1996; Peter E. Kloeden, 2007; Talay and Tubaro, 1990), the expected mean square error (MSE) in the estimate Y is asymptotically of the form expressed in Eq. 7

$$MSE \approx C_1 N^{-1} + C_2 \Delta t^2 \tag{7}$$

which implies that the MSE comes with two sources: error due to the limited number of particles $\mathbf{N}$ and the size of time step Δt (Giles, 2015). Therefore, as long as C_1 and C_2 are established, the relationship between the choice of particles number N and the time step Δt will be known, given the required MSE of D_L. It is also worth mentioning that, Δt actually affects MSE in a much more complex way, which follows a polynomial function with a dominating quadratic component (BALLY and TALAY, 2009).

In practice, C_1 and C_2 are usually unknown and have to be estimated case by case, which is undesirable. In this study, we collect evidence from different cases and try to establish some general guidelines for finding C_1 and C_2 that are independent of water environment configurations, such as flow field and diffusion coefficient.

3. CASE STUDIES

3.1. Couette flow

We first choose the dispersion phenomenon in Couette flow to verify the error analysis model, because this laminar shear flow gives the analytical solution for D_L (as Fisher first derived it in 1979), as shown in Eq. 8

$$D_{analytical} = \frac{U^2 H^2}{120 D_y} \tag{8}$$

In the Couette flow, water flow is assumed to go through two parallel plates of infinite extent, with the top plate moving at velocity U compared to the bottom one. The width between these two parallel plates is H, and D_y is the diffusion coefficient in the transverse direction.

3.1.1 Demonstration in a particular flow condition

Firstly, the random walk method has been applied on a verification case, whose parameters are presented in Table 1, with 100,000 particles and time step as 0.001 s. The simulated longitudinal dispersion coefficient is compared with the analytical solution, and the accuracy of this Lagrangian approach has been validated, as Fig. 2.

Table 1. Parameter values of a Couette flow case

Parameters	Values
H: flow width	1.0 m
U: velocity of the upper wall	1.0 m/s
D_y: diffusion coefficient in the transverse direction	0.001 m²/s

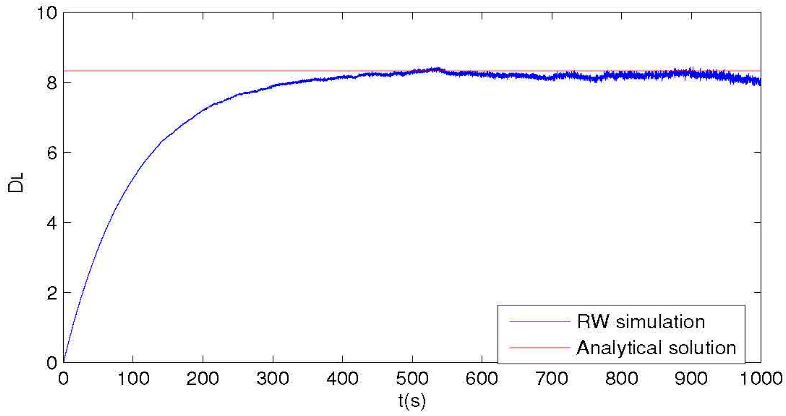


Figure 2. Evolution of the longitudinal dispersion coefficient D_L with time in Couette flow

Many simulations are carried out with different numbers of particles N and time steps Δt, as presented in Table 2. The results will be used to fit the parameters C_1 and C_2 in Eq. 7. Under each set of Δt and N, the random walk simulation was run for 20 times. In each run, the simulated D_L was compared with the analytical D_L, hence the error for each run was obtained. The MSE of D_L was calculated through the mean square error of the 20 simulations for one combination of the particles number N and time step Δt. The target parameters are the C_1 and C_2 in Eq. 7. Meanwhile, 30 sets of Δt and N (shown in Table 2) are available for fitting. This is a typical linear regression problem, and the target is to estimate two coefficients (C_1 and C_2) of the two variables (N, Δt). We choose the ordinary least squares (OLS) method to find C_1 and C_2. For the case in Table 1, we obtained $[C_1, C_2]$ = [125.8463, 1.5850]. And the fitted $R_{sq} = 0.9315$, which represents the fraction of the total sum of squares of MSE that the model explains. Please note that the linear model does not contain an intercept, thus the R_{sq} value should be interpreted as the fraction of total sum of squares of the error explained by the model, rather than the total variance explained by the model. The regression results are as shown in Fig. 3 and Fig. 4:

Table 2. Sets of time steps and particles numbers used to fit C_1 and C_2

Δt [s]: time step	0.05	0.05	0.05	0.05	0.05	0.075	0.075	0.075	0.075	0.075
N [1]: particles number	2000	1000	500	200	100	2000	1000	500	200	100
Δt [s]: time step	0.1	0.1	0.1	0.1	0.1	0.125	0.125	0.125	0.125	0.125
N [1]: particles number	2000	1000	500	200	100	2000	1000	500	200	100
Δt [s]: time step	0.15	0.15	0.15	0.15	0.15	0.2	0.2	0.2	0.2	0.2
N [1]: particles number	2000	1000	500	200	100	2000	1000	500	200	100

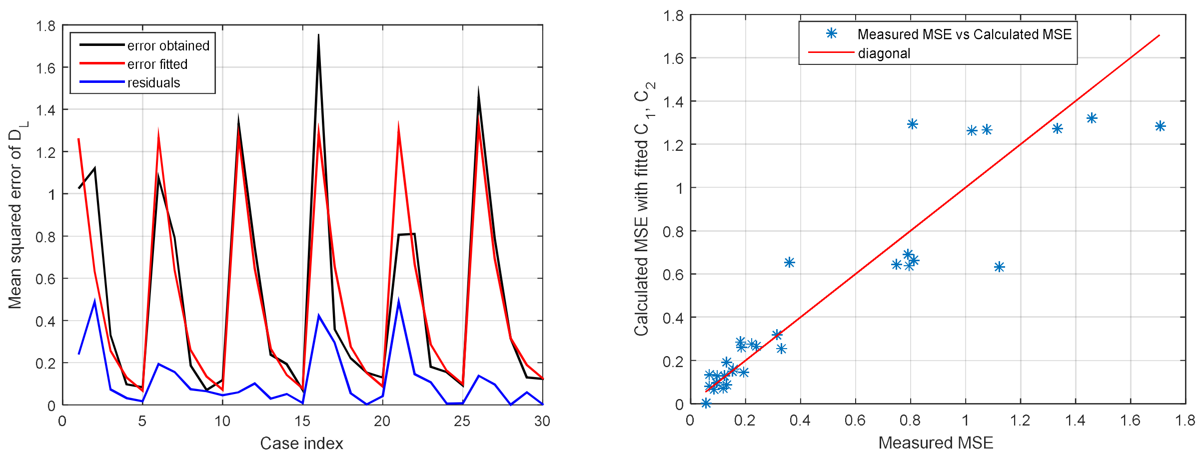


Figure 3. Left: Regression result of the fitting of C_1 and C_2; Right: Measured MSE vs Calculated MSE with fitted C_1 and C_2

To verify the model's predictive ability for the MSE of D_L in Eq. 7, we use out-of-sample test for the model. This out-of-sample test is carried out as follows: we generate another set of Δt and N combinations (8 randomly chosen cases which are within the limit of the sets for fitting, as shown in Table. 3), and then run simulations to test the goodness of using the fitted C_1 and C_2 above in predicting the MSE of D_L. The results are shown in Fig. 4.

In this out-of-sample test, the R_{sq} value, is 0.8999, indicating that 89.99% of the total sum of squares of the MSE of the D_L can be captured using this set of $C_1$and C_2 values, which is also close to the R_{sq} value previously obtained from the parameter fitting procedure. Given the satisfactory predictability of the model in the out-of-sample test, which is comparable to the in-sample-fitting, the linear model introduced in Eq. (7) is verified to be a good model which can be used to estimate the MSE of D_L given different input of N and Δt. In addition, the OLS approach we have adopted for estimating the values of C_1 and C_2 is validated as an appropriate approach.

Table 3. Sets of time steps and particles numbers used for out-of-sample test

N [1]: number of particles	300	1500	800	250	600	750	1200	400
Δt [s]: time step	0.17	0.06	0.07	0.13	0.15	0.09	0.15	0.18

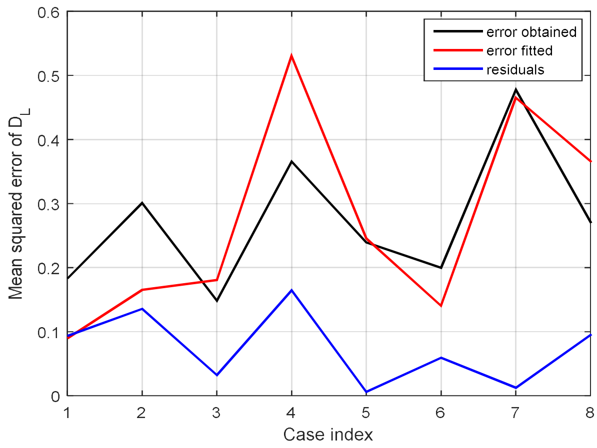


Figure 4. Regression result of the out-of-sample test with fitted $C_1$and C_2in Couette flow.

3.1.2 Nondimensionalization for Different Conditions

Besides the case we studied above, more flow conditions of Couette flow were tested. For each flow conditions, C_1 and C_2 were estimated, and the R_{sq} of them were calculated, as shown in the upper part of Table.4. We can see that among all the conditions, the R_{sq} values are quite high (above 0.9), the t-statistics of C_1 are significantly large, and p-values of C_1 are less than 0.1% significance level. Thus these tests provide further evidence to support the linear model established in the previous section.

Table 4. Different conditions in Couette flow: Top, estimation and its fitness of C_1 and C_2; Bottom, nondimensionalization of C_1 and C_2

D_y (m^2/s)	0.001	0.005	0.01	0.001	0.001	0.001	0.001

h (m)	1	1	1	0.1	0.2	1	1
U (m/s)	1	1	1	1	1	0.1	0.5
$D_{L,anl}$ (m^2/s)	8.333	1.667	0.833	0.083	0.333	0.083	2.083
Coef of C_1	125.846	5.691	1.439	0.014	0.259	0.013	7.865
95% CI of C_1	[109.997, 141.696]	[5.209, 6.174]	[1.277, 1.602]	[0.013, 0.016]	[0.236, 0.280]	[0.011, 0.014]	[6.875, 8.856]
t-statistic of C_1	15.562	23.126	17.400	21.166	22.784	15.562	15.562
p-value of C_1	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Coef of C_2	1.59E+00	6.83E-02	1.70E-02	-4.31E-06	1.88E-03	1.59E-04	9.91E-02
95% CI of C_2	[-2.387, 5.558]	[-0.053, 0.189]	[-0.024, 0.058]	[-3E-4, 3E-4]	[-0.004, 0.007]	[-2E-04, 6E-04]	[-0.149, 0.347]
t-statistic of C_2	0.782	1.107	0.819	-0.025	0.660	0.782	0.782
p-value of C_2	0.441	0.278	0.420	0.980	0.515	0.441	0.441
R_{sq}	0.932	0.968	0.944	0.960	0.966	0.932	0.932
Root mean square error	0.1856	0.0056	0.0019	1.56E-05	2.60E-04	1.85E-05	0.0116
Nondimensionalize							
$D_L/D_{L,anl}$ $C_{1,ndim}$	1.812	2.049	2.073	2.075	2.327	1.812	1.812
$\Delta t \cdot D_y/h^2$ $C_{2,ndim}$	2.28E+04	9.83E+02	2.45E+02	-6.21E-02	2.70E+01	2.28E+04	2.28E+04

The nondimensionalization step is to study the coefficient C_1 and C_2 in a dimensionless perspective. To nondimensionalize these two parameters, we divide D_L by $D_{L,anl}$ and divide Δt by D_y/h^2. We then replace the original D_L and Δt with these rescaled values, denoted as $D_{L,ndim}$, and Δt_{ndim} . Other methods of nondimensionalization have also been tried (as listed in table 4.1 in the Research Data uploaded alongside), but here we select the one that works best. Then we repeat the OLS fitting for the set of parameters C_1 and C_2 again, using these nondimensionalize values $D_{L,ndim}$ and Δt_{ndim}. The refitted C_1 and C_2 are donated as $C_{1,ndim}$ and $C_{2,ndim}$, respectively. The lower part of Table 4 shows the $C_{1,ndim}$ and $C_{2,ndim}$ values estimated using this nondimensionalizing approach. We can see that the dimensionless values of $C_{1,ndim}$ are close to 2 for all flow conditions. This indicates a theoretical explanation behind. We have tried to change 1/2 to 1/5 in Eq.4, found $C_{1,ndim}$ remains around 2, implying 1/2 in Eq.4 is not the reason for this number 2. Therefore, further research should be done to unravel the theory. $C_{1,ndim} = 2$ enables us to provide predictions for C_1 in different scenarios once we know the parameters for the flow environment, such as U, h and D_y, without the need to conduct simulations again.

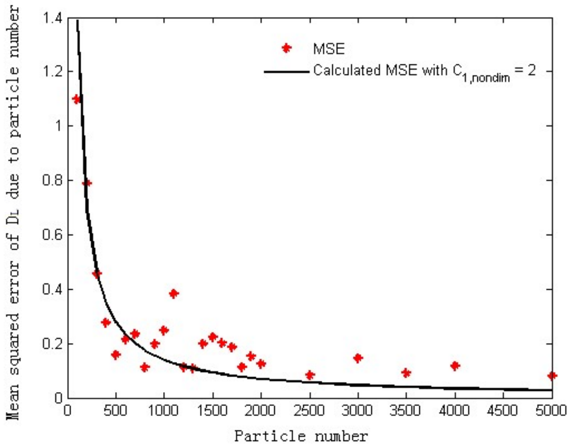
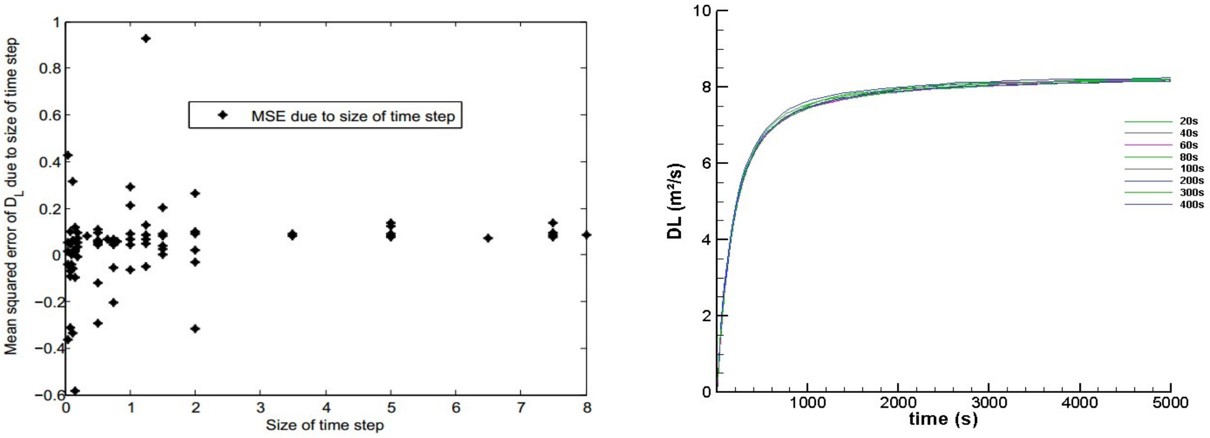


Figure. 5 Predicted MSE compared with the measured MSE in Couette flow(the dots represent the measured MSE, while the solid line represents the predicted MSE by the error model with dimensionless parameter C1 = 2. The predicted MSE via the dimensionless parameter explained 89.9% of the total sum of squares of the measured MSE.)

To prove that $C_{1,nondim} = 2$, we choose another 25 sets of N and $\mathbf{\Delta t}$ (as listed in Table.4.2 in the Research Data uploaded alongside). The values of N and Δt are selected because we want to minimize the MSE from Δt (abbreviated as $MSE_{\Delta t}$), thus MSE due to N (abbreviated as MSE_{N}) is predominant. Each combination of N and Δt is used in the random walk model for the simulation of the flow case in Table.1 for 20 times, then the MSE can be obtained for this set of N and Δt. Meanwhile, with $C_{1,nondim} = 2$, and comparatively small $MSE_{\Delta t}$, the MSE predicted by the error model can be calculated as $MSE_{ndim} \approx MSE_N = D_{L,ana}^2 \cdot 2 \cdot N^{-1}$. The measured MSE are then compared with the predicted MSE, as shown in Fig.5. The black line which represents the predicted MSE can fit the measured MSE well, illustrated by red dots. With $C_{1,nondim} = 2$, the predicted MSE explained 89.9% of the total sum of squares of the measured MSE. Therefore, the inference that $C_{1,nondim} = 2$ for Couette flow is tenable.

However, we are unable to find a unified dimensionless value for C_2 in the current stage. It can be seen from Table 4 (and Table 4.1 in the Research Data uploaded alongside), that the estimation of C_2, which implies the influence of Δt, has much lower confidence. One possible reason for such inaccuracy may be that the weights that two parts carry are not balanced. Therefore, we tried more sets of N and Δt (86 sets in total), with a very large number of N fixed at 50000, i.e. using 50000 particles to reduce the part of MSE due to the number of particles.

Assuming $MSE_N \approx 2N^{-1}$ stands true, the time-step-part of error $MSE_{\Delta t}$ can then be obtained by subtracting MSE_N from the total MSE. We then plotted the relationship between $MSE_{\Delta t}$ and size of time step as Fig.6(a). Compared with that of the particles numbers as shown in Fig.5, $MSE_{\Delta t}$ are scattered randomly around the zero line, with no evident regularity. Such a distribution implies that the accuracy of the numerical simulation of longitudinal dispersion coefficient D_L in Couette flow may be independent of the size of the time step, i.e. $C_2 \approx 0$. This conclusion is confirmed by Fig.6(b), which shows that all the development lines of D_L converge together for Δt from 20s to as large as 400s. In conclusion, the choice of the size of the time step in the given steady flow actually has very little impact on the simulation of the D_L. This evidence is also in line with the large p-values of C_2 in Table. 4, suggesting $C_2 \approx 0$.



(a) (b)

Figure. 6 (a). Mean squared error of D_L due to the size of the time step with the number of particles is fixed as 50000; (b). D_L development with time calculated by different sizes of the time step (Couette flow)

3.2. Open Channel Flow

The above conclusions might be a special case regarding the Couette flow. Hence, we also did simulations on another steady flow condition, e.g. open channel flow with logarithmic velocity distribution. Interestingly, we reached the same conclusion: $C_{1,ndim} \approx 2$, and D_L is irrelevant with the size of time step in a certain range when velocity is small enough. However, when the velocity is moderately large, it can be noticed that a larger D_L leads to a larger MSE. (The analytical solution for the longitudinal dispersion coefficient $D_{L,anl}$ of logarithmic flow, as well as the flow conditions, can be found in the Research Data uploaded alongside.)

Fig.7 shows the results by these varying Δt from 0.1s to 0.5s, with the same particles number 50,000. As can be seen, despite the differences in the sizes of the time step, there is some small difference among the simulated

D_L. The simulation of D_L for open channel flow is dependent on Δt, although the impact is relatively small in this range.

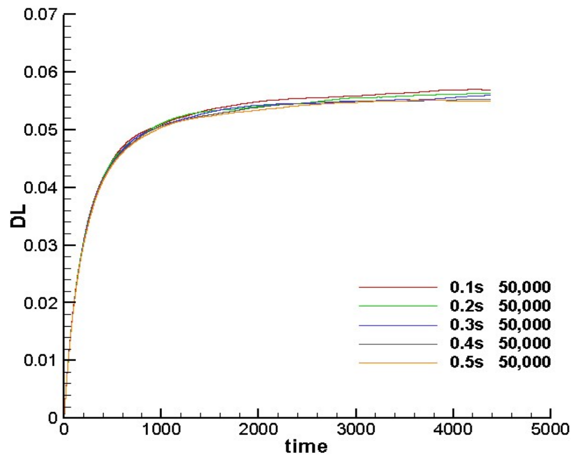


Figure. 7 D_L development with time calculated by different sizes of the time step in open channel flow

The impact of N and Δt is displayed in details in Table. 5. The same ordinary least squares (OLS) method is utilized to find the values of C_1 and C_2. We also extend the calculation to other flow conditions, as listed in Table.5, and nondimensionalization was done by the same method as illustrated in the above section. Again, $C_{1,ndim}$ converges to 2 after being nondimensionalized, despite the changing of flow conditions. The confidence of such a regression model is quite high, as proved by the high value of the R_{sq} in all configurations. However, impact of the time step is now more obvious in the open channel flow than that in the Couette flow. The 2nd and the 3rd columns show that a significant non-zero C_2 can be estimated from simulations when the flow velocity is larger than 0.05 m/s, suggesting a positive impact of time step on the simulated error. Although the estimated coefficient C_2 seems to be small, the modeled relationship between it and the measured MSE is strong and cannot be omitted, as indicated by the large t-statistics and the p-values associated with it. These results are consistent with Eq. 7.

Table 5. Different conditions in open channel flow: Top, estimation and its fitness of C_1, C_2;

Bottom, nondimensionalization of C_1, C_2

$u_* \ (m/s)$	0.01	0.05	0.1	0.01	0.01
$d \ (m)$	1	1	1	2	5
$D_{z,avg} \ (m^2/s)$	6.83E-04	3.42E-03	6.83E-03	1.37E-03	3.42E-03
$D_{L,anl} \ (m^2/s)$	0.0586	0.293	0.586	0.117	0.293

Coefficient C_1		6.90E-03	1.70E-01	8.01E-01	2.52E-02	1.67E-01
95% CI of C_1		[6.1E-03, 7.8E-03[	[1.57E-01, 1.83E-01]	[7.28E-01, 8.74E-01]	[2.24E-02, 2.80E-02]	[1.50E-01, 1.84E-01]
t-statistic of C_1		16.110	24.989	21.436	17.638	19.256
p-value of C_1		0.000	0.000	0.000	0.000	0.000
C_2		1.00E-04	1.35E-02	1.30E-01	7.00E-04	2.40E-03
95% CI of C_2		[-1.54E-04, 2.67E-04]	[1.02E-02, 1.69E-02]	[1.11E-01, 1.48E-01]	[-3.42E-05, 1.37E-03]	[-1.91E-03, 6.63E-03]
t-statistic of C_2		0.524	7.948	13.835	1.864	1.083
p-value of C_2		0.604	0.000	0.000	0.073	0.288
R_{sq}		0.9346	0.9798	0.9813	0.9491	0.9545
Root mean square error		9.84E-06	1.56E-04	8.57E-04	3.28E-05	1.99E-04
NonDimensionalize						
$D_L/D_{L,anl}$	$C_{1,ndim}$	2.0121	1.9796	2.3318	1.8329	1.95
$\Delta t \cdot D_y/h^2$	$C_{2,ndim}$	3.515E+04	1.35E+04	8.08E+03	4.16E+05	1.47E+06

The MSE_N for different particles numbers are plotted in Fig.8. The red dots are the measured MSE from 20 runs of the model, while the black line shows the predictions made with $C_{1,ndim} = 2$. It is clear that the larger number of particles leads to the smaller MSE_N. Besides, the predicted black line can well fit the measured data with R_{sq} = **96.2%**.

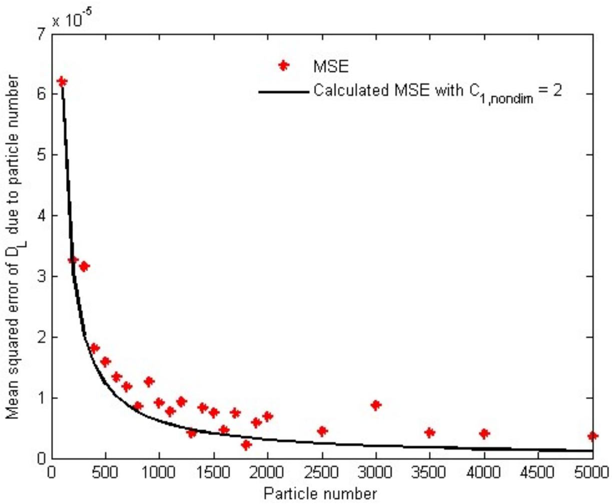


Figure. 8 Predicted MSE compared with the measured MSE in open channel flow (the dots represent the measured MSE,

while the solid line represents the predicted MSE by the error model with dimensionless parameter C1 = 2. The predicted MSE via the dimensionless parameter explained 96.2% of the total sum of squares of the measured MSE .)

4. DISCUSSIONS

a) Error due to particles number

The random walk scheme, which tries to use a relatively smaller number of samples to approximate the probability distribution, belongs to the class of Monte Carlo methods. Monte Carlo methods are popular methods in many fields, and the error analysis of the Monte Carlo methods has been well studied. It is known that, as Mackay (2003) claimed, Monte Carlo methods are usually used to solve two problems:

(1) To generate samples $\{x^{(r)}\}_{r=1}^{R}$ from a probability distribution $P(x)$.

(2) To estimate expectations of functions $\phi(x)$ under this distribution.

Suppose a set of N samples $\{x^{(i)}\}_{i=1}^{N}$ is generated from a probability distribution $P(x)$, we can then obtain an estimator $\hat{\Phi}$ by using the following equation:

$$\hat{\Phi}=\frac{1}{N}\sum_{i=1}^{N}\phi(x^{(i)}) \tag{9}$$

This estimator $\hat{\Phi}$ is an unbiased estimator of the expectation of ϕ. Furthermore, the variance of $\hat{\Phi}$ will decrease as $\frac{\sigma^2}{N}$, where $\sigma^2$is the variance of ϕ.

In our case, given a known time-discretization, Euler discretization, and all other initial setups of a simulation, we are actually trying to use N particles to track their trajectories $\{x^{(i)}\}_{i=1}^{N}$, and $\phi(x^{(i)})$ here is the function for longitudinal dispersion coefficient. $\hat{\Phi}$ is the estimator for $\phi(x^{(i)})$ (i.e. D_L). Therefore, the left side of Eq. (9) is the simulated D_L. The variance of the simulated D_L, hence, will decrease as $\frac{\sigma^2}{N}$. σ^2 here, is the variance of ϕ. Since $\hat{\Phi}$ is an unbiased estimator, the variance and the Mean square error (MSE) are equivalent. Thus, the MSE due to the limited number of particles, denoted as MSE_N, will decrease as $\frac{\sigma^2}{N}$, i.e. $MSE_N \approx \frac{\sigma^2}{N}$.

Since the initial conditions and other settings of simulations are fixed, the target probability distribution $P(x)$, although unknown, is a determined distribution. Thus, its variance σ^2 is constant. Therefore, $MSE_N \approx \frac{\sigma^2}{N}$ can be expressed as $MSE_N \approx C_1 N^{-1}$, where C_1 is a constant to be determined in different cases and is dependent on the variances of the $P(x)$.

b) Size of the time step

The calculation of the longitudinal dispersion coefficient, which concerns the ensemble of the particles cloud rather than the exact trajectory of each particle, belongs to weak convergence approximation of SDE (Bally and Talay, 1996). It has been proved that the Euler scheme, one of the simplest discretization schemes and the one we used in the paper, converges with a weak order of 1, i.e. the mean square error because the time step is proportional to the square of Δt, as indicated in Eq.7. This becomes noticeable in open channel flow case studies with a large flow speed.

However, it is found in our simulations that, the size of the time step has little influence on the calculation of the longitudinal dispersion coefficient D_L in Couette flows. We may comprehend this by drawing an analogy between the mass transportation in flows and the motions of a group of antelopes. The distribution of this group of antelopes will stay all the same no matter they take a large or small step to jump, as long as the velocity of their motion and the total travel time is the same.

Now we try to explain such degradation of Eq.7 from a statistical perspective. Without loss of generality, the one-dimensional scenario is taken as an example. The position of a particle is calculated as:

$$x = x_0 + u\Delta t + \frac{\partial D_x}{\partial x}\Delta t + \sqrt{2D_x\Delta t}R \tag{10}$$

We now consider two sizes of the time step, Δt_{large}, Δt_{small}, and $\Delta t_l = N\Delta t_s$. For simplicity's sake, D_x is also assumed to be constant, thus $\frac{\partial D_x}{\partial x} \equiv 0$. As the velocity u is also a constant in a steady flow, the position of the particle calculated by Δt_l and Δt_s are given as below respectively:

$$x_l = x_0 + u\Delta t_l + \sqrt{2D_x\Delta t_l}R_l \tag{11a}$$

$$x_s = x_0 + u\Delta t_s + \sqrt{2D_x\Delta t_s}R_s \tag{11b}$$

Taking $\Delta t_l = N\Delta t_s$ into Eq.11, the final position of the particle after the same time length will be

$$x_{l,final} = x_0 + u \cdot N\Delta t_s + \sqrt{2D_x N\Delta t_s}\,R_l \tag{12a}$$

$$x_{s,final} = x_0 + N \cdot u\Delta t_s + \sqrt{2D_x\Delta t_s}\,(R_{s1} + R_{s2} + \cdots + R_{sN}) \tag{12b}$$

where R_l as well as R_{s1}, R_{s2},…, R_{sN} are all random numbers following a normal distribution with zero average and unit variance.

What we care about here, as stated in weak convergence, is not the trajectory or the exact position of the particle, but the collective distribution of all the particles. Specifically, it is the longitudinal dispersion coefficient D_L that we want to predict. As calculated by Eq. 4, D_L is dependent on the variance of particles ensemble. From the knowledge of statistics, it is known that

$$\mathrm{Var}(\sqrt{N\Delta t_s}R_l) = N\Delta t_s \mathrm{Var}(R_l) \tag{13a}$$

$$\mathrm{Var}(\sqrt{\Delta t_s}(R_{s1} + R_{s2} + \cdots + R_{sN})) = \Delta t_s N \mathrm{Var}(R_{s1}) \tag{13b}$$

Substitute Eq.12 and Eq.13 to Eq.4, it can be proved that D_L will stay the same despite the choice of the size of the time step, provided that the initial condition is all the same. This derivation ignores the diffusion effect in the simulations. Only in some certain and simple cases, such as in a linear flow-velocity profile, the calculation process of diffusion coefficient may absorb the diffusion effect and thus seem to be independent of the time step.

5. CONCLUSIONS

It is well known that the choice of the number of particles N, and the size of time step Δt, is of vital importance in the implementation of random walk methods. The choice of these two parameters in random walk methods relies on the balance between the accuracy (measured as MSE) and the computational cost. In this paper, we present an empirical study on the reliance of the MSE of the longitudinal dispersion coefficient D_L on these two parameters in a quantitative approach. The following conclusions are made for steady flows:

(1) After nondimensionalization, i.e. normalizing D_L by the analytical value $D_{L,anl}$, the value of $C_{1,ndim}$ converges to a constant of 2 regardless of the flow conditions.

(2) For the Couette flow, the accuracy of the random walk simulations for the longitudinal dispersion coefficient seems to be independent of Δt. However, when the velocity profile is highly nonlinear, extremely large Δt values will decrease the accuracy of the random walk simulation.

Therefore, a relatively large time step Δt can be applied to minimize the computational expenses without compromising the accuracy for the numerical simulation of D_L in steady flows. However, the time step must be limited for the correct treatment of boundary conditions and source terms. Furthermore, given that the nondimensionalized value $C_{1,ndim} \approx 2$ is valid for any flow conditions, the absolute value of C_1 can be calculated as $2D_{L,anl}^2$. Once the analytical solution or empirical value of the dispersion coefficient $D_{L,anl}^2$ is known, the relationship between the accuracy of predicted D_L value and the computational parameters can be expressed by:

$$MSE \approx c_1 N^{-1} \tag{14}$$

This degraded error model provides guidance on the choice of the number of the particles needed to achieve the desired accuracy. On the other hand, it can be used to estimate the computational uncertainty MSE_N for a given value of N in different flow scenarios.

We are currently extending this research to unsteady flows, and the related details will be described in another paper. Both the dispersion coefficient and other parameters will be examined by our error model for the random walk simulations.

ACKNOWLEDGMENTS

We are grateful to the financial support by the National Natural Science Foundation of China (51809219), the Open Research Fund Program of State key Laboratory of Hydroscience and Engineering (sklhse-2019-B-02), the Royal Academy of Engineering UK-China Urban Flooding Research Impact Programme (Grant No. UUFRIP\100051) and the Ministry of Education and State Administration of Foreign Experts Affairs 111 Project (Grant No. B17015).

REFERENCES

Bally, V., Talay, D., 1996. The Law of the Euler Scheme for Stochastic Differential Equations: II. Convergence Rate of the Density. Monte Carlo Methods Appl. https://doi.org/10.1515/mcma.1996.2.2.93

BALLY, V., TALAY, D., 2009. The Law of the Euler Scheme for Stochastic Differential Equations: II. Convergence Rate of the Density. Monte Carlo Methods Appl. 2. https://doi.org/10.1515/mcma.1996.2.2.93

Benkhaldoun, F., Elmahi, I., Seaïd, M., 2007. Well-balanced finite volume schemes for pollutant transport on unstructured meshes. J. Comput. Phys. 226, 180–203.

Benson, D.A., Aquino, T., Bolster, D., Engdahl, N., Henri, C. V., Fernàndez-Garcia, D., 2017. A comparison of Eulerian and Lagrangian transport and non-linear reaction algorithms. Adv. Water Resour. 99, 15–37. https://doi.org/10.1016/J.ADVWATRES.2016.11.003

de Anna, P., Le Borgne, T., Dentz, M., Tartakovsky, A.M., Bolster, D., Davy, P., 2013. Flow Intermittency, Dispersion, and Correlated Continuous Time Random Walks in Porous Media. Phys. Rev. Lett. 110, 184502. https://doi.org/10.1103/PhysRevLett.110.184502

Gardiner, 2004. Handbook of stochastic methods. Adv. Math. (N. Y). 55, 101. https://doi.org/10.1016/0001-8708(85)90015-5

Giles, M.B., 2015. Multilevel Monte Carlo methods. Acta Numer. 24, 259–328. https://doi.org/10.1017/S096249291500001X

Giles, M.B., 2008. Multilevel Monte Carlo Path Simulation. Oper. Res. 56, 607–617. https://doi.org/10.1287/opre.1070.0496

Hamid, K., Suleman, M.T., Ali Shah, S.Z., Imdad Akash, R.S., 2017. Testing the Weak Form of Efficient Market

Hypothesis: Empirical Evidence from Asia-Pacific Markets, SSRN. https://doi.org/10.2139/ssrn.2912908

Liang, D., Wang, X., Falconer, R.A., Bockelmann-Evans, B.N., 2010. Solving the depth-integrated solute transport equation with a TVD-MacCormack scheme. Environ. Model. Softw. 25, 1619–1629. https://doi.org/10.1016/j.envsoft.2010.06.008

Mishra, A., Mishra, V., Smyth, R., 2015. The Random-Walk Hypothesis on the Indian Stock Market. Emerg. Mark. Financ. Trade 51, 879–892. https://doi.org/10.1080/1540496X.2015.1061380

Möbus, H., Gerlinger, P., Brüggemann, D., 2001. Comparison of Eulerian and Lagrangian Monte Carlo PDF methods for turbulent diffusion flames. Combust. Flame 124, 519–534. https://doi.org/10.1016/S0010-2180(00)00207-8

Neuman, S., 1993. Eulerian-Lagrangian theory of transport in space-time nonstationary velocity fields: Exact nonlocal …. Water Resour. Res 29, 633–645.

Peter E. Kloeden, E.P., 2007. Numerical solution of stochastic differential equations. Stochastics Stoch. Reports 47, 121–126. https://doi.org/10.1080/17442509408833885

Saidi, M.S., Rismanian, M., Monjezi, M., Zendehbad, M., Fatehiboroujeni, S., 2014. Comparison between Lagrangian and Eulerian approaches in predicting motion of micron-sized particles in laminar flows. Atmos. Environ. 89, 199–206. https://doi.org/10.1016/J.ATMOSENV.2014.01.069

Shao, S., Gotoh, H., 2005. Turbulence particle models for tracking free surfaces Turbulence particle models for tracking free surfaces Modèles particulaires turbulents pour suivre les surfaces libres. J. Hydraul. Res. J. J. Hydraul. Res. 43, 22–1686. https://doi.org/10.1080/00221680509500122

Shao, S., Lo, E.Y.M., 2003. Incompressible SPH method for simulating Newtonian and non-Newtonian flows with a free surface. Adv. Water Resour. 26, 787–800. https://doi.org/10.1016/S0309-1708(03)00030-7

Sole-Mari, G., Fernàndez-Garcia, D., Rodríguez-Escales, P., Sanchez-Vila, X., 2017. A KDE-Based Random Walk Method for Modeling Reactive Transport With Complex Kinetics in Porous Media. Water Resour. Res. 53, 9019–9039. https://doi.org/10.1002/2017WR021064

Talay, D., Tubaro, L., 1990. Expansion of the global error for numerical schemes solving stochastic differential equations. Stoch. Anal. Appl. 8, 483–509. https://doi.org/10.1080/07362999008809220

Wu, X. fei, Liang, D., 2019. Study of pollutant transport in depth-averaged flows using random walk method. J. Hydrodyn. 31, 303–316. https://doi.org/10.1007/s42241-018-0105-7

Yang, F and Liang, D and Wu, X and Xiao, Y., 2019. On the application of the depth-averaged random walk method to solute transport simulations. Journal of hydroinformatics. IWA Pub.

Zhang, Z., Chen, Q., 2007. Comparison of the Eulerian and Lagrangian methods for predicting particle transport in enclosed spaces. Atmos. Environ. 41, 5236–5248. https://doi.org/10.1016/j.atmosenv.2006.05.086