Advances in Water Resources 000 (2020) 103509

Contents lists available at ScienceDirect

Advances in Water Resources

journal homepage: www.elsevier.com/locate/advwatres

Entropy: (1) The former trouble with particle-tracking simulation, and (2) A measure of computational information penalty

David A. Benson[a,*], Stephen Pankavich[b], Michael J. Schmidt[b], Guillem Sole-Mari[c]

[a] Hydrologic Science and Engineering, Colorado School of Mines, Golden, CO 80401, USA
[b] Department of Applied Mathematics and Statistics, Colorado School of Mines, Golden, CO, 80401, USA
[c] Department of Civil and Environmental Engineering, Universitat Politècnica de Catalunya, Barcelona, Spain

ARTICLE INFO

Keywords:
Particle methods
Entropy
Mixing
Dilution index
Computational penalty
AIC

ABSTRACT

Traditional random-walk particle-tracking (PT) models of advection and dispersion do not track entropy, because particle masses remain constant. However, newer mass-transfer particle tracking (MTPT) models have the ability to do so because masses of all compounds may change along trajectories. Additionally, the probability mass functions (PMF) of these MTPT models may be compared to continuous solutions with probability density functions, when a consistent definition of entropy (or similarly, the dilution index) is constructed. This definition reveals that every discretized numerical model incurs a computational entropy. Similar to Akaike's (1974, 1992) entropic penalty for larger numbers of adjustable parameters, the computational complexity of a model (e.g., number of nodes or particles) adds to the entropy and, as such, must be penalized. Application of a new computational information criterion reveals that increased accuracy is not always justified relative to increased computational complexity. The MTPT method can use a particle-collision based kernel or an adaptive kernel derived from smoothed-particle hydrodynamics (SPH). The latter is more representative of a locally well-mixed system (i.e., one in which the dispersion tensor equally represents mixing and solute spreading), while the former better represents the separate processes of mixing versus spreading. We use computational means to demonstrate the fitness of each of these methods for simulating 1-D advective-dispersive transport with uniform coefficients.

1. Introduction

Entropy is a fundamental property possessed by any random variable or process, including a plume moving through natural media. From a thermodynamic viewpoint, an increase in entropy is an increase in mixing (dilution for a conservative solute). But entropy is also a measure of the information required to describe a system composed of random variables, so that the entropy of a computer simulation of a plume can be readily quantified. The entropy of a discrete random variable has been defined in a straightforward way by Shannon (1948). Unfortunately there is not a well-defined counterpart for continuous random variables. In fact, a commonly-used "definition" of the entropy of a continuous random variable (RV) can take on unphysical negative values. As such, the Kullback-Leibler divergence (or relative entropy), which is a measure of the *relative* difference between two continuous RVs, is often used instead (Kullback and Leibler, 1951; Kullback, 1968). Similarly, this quantity is also used to derive the entropic penalty of over-parameterization of models by Akaike (1974)—where it is assumed that any two random variables have similar measures (i.e., both discrete or continuous). When applied to numerical simulations, the Kullback-Leibler divergence assumes that different models use the same discretization length, which then cancels within the resulting expression. If differing discretizations are used, then the different entropies (information) associated with those discretizations must be accounted for.

This observation also applies to the Akaike information criterion, which is widely used to assess model fitness. Akaike's original information criterion (the AIC) simply penalizes different models for their number of adjustable parameters, which means that gains in accuracy between a model and a single realization of data may be counteracted by over-parameterization. In this paper, we show that the information penalty for highly discretized models means that minor gains in accuracy may be overwhelmed by losses due to computational complexity. In other words, a modeler knows intuitively that a good model is 1) accurate, 2) parsimonious, and 3) computationally efficient. The first point has been investigated thoroughly via convergence analysis, maximum likelihood estimation, etc. The second point was addressed by Akaike (1974) and later extensions of the AIC (Konishi and Kitagawa, 2008). The last point has lacked a theoretical foundation, so we address it herein. We also investigate, using the simplest setting possible (a 1-D diffusion problem), whether both particle-tracking and Eulerian finite-difference solutions display an optimal discretization where small

* Corresponding author.
E-mail address: dbenson@mines.edu (D.A. Benson).

https://doi.org/10.1016/j.advwatres.2020.103509
Received 22 April 2019; Received in revised form 10 January 2020; Accepted 11 January 2020
Available online xxx
0309-1708/

accuracy gains are unsupported by larger computational expenditures (i.e., information requirements).

We begin by reviewing the definitions of entropy and defining one that is consistent amongst discrete and continuous RVs. Our definition includes a discretization variable that is similar to the sampling volume of Kitanidis (1994a) for moving plumes. We then review the classical particle-tracking algorithm (e.g., Labolle et al., 1996), which does not directly track entropy because particles do not exchange mass. In this case the entropy can only be calculated after a continuous interpolation of concentration is performed, which is shown to have an effect on the entropy calculation. This is in contrast to newer particle-tracking algorithms that do exchange mass between particles during each timestep (Benson and Bolster, 2016; Sole-Mari et al., 2019). Because of the mass transfer, entropy automatically and continuously changes during a simulation. We investigate the entropy evolution represented by two types of inter-particle mass-transfer algorithms: 1) smoothed particle hydrodynamics (Gingold and Monaghan, 1977; Monaghan, 2012) and 2) a particle-collision-based algorithm (Benson and Bolster, 2016). The first method seeks to optimally solve a given deterministic PDE using particles, while the second implements a local physics-based set of equations for particle behavior. Typically, on the global scale, this method solves a stochastically perturbed equation of transport (Benson et al., 2019). Because of the subtle differences and similarities of the two methods, we investigate the growth of entropy simulated by both for a highly simplified problem: one-dimensional diffusion.

We wish to track entropy in the particle-tracking algorithms because it is a direct measure of mixing between dissimilar waters. Additionally, mixing is often the primary control on chemical reactions. In previous studies (Benson et al., 2017, 2019), we were forced to track the creation and destruction of chemical species (i.e., reaction rates) to compare numerical methods or upscaling techniques. With entropy consistently defined, the mixing performance of numerical and analytic techniques can be directly measured.

2. Mathematical background

The classical particle-tracking (PT) method is a way to eliminate numerical dispersion in the simulation of the advection-dispersion equation (ADE) given by

$$\frac{\partial c}{\partial t} = -\nabla \cdot (\mathbf{v}c) + \nabla \cdot (\mathbf{D}\nabla c). \tag{1}$$

Because the dispersion tensor may have spatial variability and resides inside a spatial derivative, one chooses specific values of the drift and pure diffusion terms in a numerical implementation of the associated Itô equation of particle motion $\Delta X = (\mathbf{v} + \nabla \cdot \mathbf{D})\Delta t + \mathbf{B}\sqrt{2\Delta t}\zeta$, where X is a particle position vector in d spatial dimensions, $\mathbf{v}(X)$ is a known velocity vector and $\mathbf{D}(X)$ is the local dispersion tensor at the position X at the beginning of the timestep, $\mathbf{BB^T} = \mathbf{D}$ is a Cholesky decomposition of the known diffusion tensor, and ζ is a d-dimensional vector of independent standard normal random variables (Kitanidis, 1994b; Labolle et al., 1996; Lichtner et al., 2002; Øksendal, 2003; Gardiner, 2004).

To approximate the solutions of Eq. (1), a large number of independent particles are transported according to the numerical Itô equation, and the histogram (or other interpolation) of these particles is used to recreate the function $c(x, t)$. If all N particles begin at the same location, then $c(x, t)$ is a density function and an approximation of the Green's function generated by Eq. (1). Because of the random dispersive motions of particles, the PT method accurately simulates the spread of a plume following the ADE. However, in its raw form (prior to creating the function $c(x, t)$), the PT method does not correctly simulate the mixing of dissimilar waters, or dilution of a conservative plume, because particles maintain constant mass.

Mixing and dilution can only be taken into account with post-processing of particle positions. Mixing and/or dilution are commonly measured by borrowing the definition of the entropy H_D of a discrete random variable X (see the seminal paper by Kitanidis (1994a) and recent extensions and applications by Chiogna et al. (2012); Chiogna and Rolle (2017); Sund et al. (2017)). Entropy is the expectation of the "information" contained within the probability mass function of that random variable. The information $I(p)$ is a non-negative function of an event's probability p that is defined as additive for independent events, i.e., $I(p_1) + I(p_2) = I(p_1 p_2)$. Because of this axiom, the functional form of information must be $I(p) \propto -\ln(p)$, so that the expected information is also strictly non-negative and defined by

$$H_D(X) = \mathbb{E}[I(P(X))] = -\sum_{i=1}^{N} p(x_i)\ln(p(x_i)), \tag{2}$$

for a discrete random variable (RV) with probability mass function $p(x)$ taking non-zero values at points $\{x_1, \ldots, x_N\}$.

By analogy, the expected information for a continuous RV is sometimes listed as

$$H_I(X) = -\int_{f(x)>0} f(x)\ln(f(x))dx \tag{3}$$

where the PDF of the continuous random variable X is $f(x)$ [L$^-$]. Of course $f(x)$ is not a probability, and thus the argument of ln () is not a dimensionless quantity. For these reasons Eq. (3) is not well defined on its own. In addition, because $f(x)$ may often be greater than unity, this usage for a continuous RV can violate the notion of entropy by assuming negative values; therefore, we use the subscript on H_I to represent "inconsistent" entropy. This definition is not without its utility; however, zero entropy means perfect order (zero mixing) and negative entropy has no physical meaning. In other words, this definition (3) for a continuous RV is only a loose analogy (see Appendix A). It does not follow from a Riemann-integral representation of Eq. (2), meaning

$$\int_{f(x)>0} f(x)\ln(f(x))dx \neq \lim_{\Delta x \to 0}\left[\sum_{i=1}^{N} f(x_i)\Delta x \ln(f(x_i)\Delta x)\right] \tag{4}$$

where $\{x_1, \ldots, x_N\}$ is a set of values at which $f(x_i) > 0$ for $i = 1, \ldots, N$, and the grid spacing $\Delta x = x_{i+1} - x_i$ is uniform for every $i = 1, \ldots, N-1$. In fact, the limit on the right side does not converge for any valid PDF. In practice, the evaluation of the entropy of some arbitrary continuous function $f(x)$ (like a plume moving through heterogeneous material) that does not have a convenient hand-integrable form, must impose a sampling interval ΔV. We use this new variable to conform with the usage in Kitanidis (1994a). With this finite sampling, an entropy H_C may be defined that is consistent with H_D in Eq. (2) by using the approximation that for small ΔV,

$$\mathbb{P}(x - \Delta V/2 < X < x + \Delta V/2) \approx f(x)\Delta V, \tag{5}$$

so that the argument of the logarithm in Eq. (3) is once again a dimensionless probability directly related to the sampling interval ΔV:

$$\begin{aligned} H_C(X) &= -\int_{f(x)>0} f(x)\ln(f(x)\Delta V)dx \\ &= -\ln(\Delta V) + H_I. \end{aligned} \tag{6}$$

Additionally, to construct a discrete approximation of the consistent entropy, we can merely approximate the integral in H_I so that

$$H_C(X) \approx -\ln(\Delta V) - \sum_{i=1}^{N} f(x_i)\Delta x \ln\left(f(x_i)\right). \tag{7}$$

Now we may identify this sampling volume ΔV as identical to the volume invoked by Kitanidis (1994a) to relate the discrete and continuous definitions of entropy, so that $H_D \approx H_C$. Most commonly, one would let $\Delta V = \Delta x$ in the sum of Eq. (7), but in estimation theory, this discretization may represent different things (Appendix A). Clearly, the choice of sampling interval ΔV both allows for a direct comparison of continuous to discrete processes and imposes some restrictions on how entropy is calculated, as we show later. Kitanidis (1994a) also defines the dilution index E as the product of the sampling volume and the exponential

of the entropy for discrete and continuous random variables. Using the consistent entropy provided by Eq. (7), this can be written as

$$\begin{aligned} E &= \Delta V e^{H_C} \\ &\approx \Delta V \exp\Big[-\ln(\Delta V) - \sum_{i=1}^{N} f(x_i)\Delta x \ln(f(x_i))\Big] \\ &\approx \exp\Big[-\sum_{i=1}^{N} f(x_i)\Delta x \ln(f(x_i))\Big]. \end{aligned} \tag{8}$$

As $\Delta x \to 0$, this uses the classical inconsistent definition of entropy for a continuous random variable, namely $E = \exp[-\int f(x)\ln(f(x))dx] = e^{H_I}$. For a discrete random variable, this becomes

$$E = \Delta V e^{H_D} = \Delta V \exp\left(-\sum_{i=1}^{N} p(x_i)\ln(p(x_i))\right). \tag{9}$$

Each definition (8) and (9) has the same units as X, i.e., a volume in the number of dimensions of random travel X, and has a reasonably well-defined physical meaning as the "size" of the volume occupied by either the ensemble of particles or the PDF $f(x)$ (Kitanidis, 1994a).

A real or simulated plume of conservative tracer is often idealized as a PDF of travel distance, i.e., the Green's function, when the spatial source is a normalized Dirac delta function $\delta(x)$. Without loss of generality, we will only consider plumes that have such a source function, so that we may use concentration as a PDF at any fixed time T, and thus $c(x,T) = f(x)$ in Eq. (7).

The normalized concentration given by the classical PT method is represented as an interpolation of the N particles, namely

$$\begin{aligned} c_N(x,t) &= \frac{1}{m_{tot}} \sum_{i=1}^{N} \int_{\Omega} m_i \delta(z - X_i(t))\phi(x-z)dz \\ &= \frac{1}{m_{tot}} \sum_{i=1}^{N} m_i \phi(x - X_i(t)), \end{aligned} \tag{10}$$

where $c_N(x,t)$ [L^{-1}] is a reconstructed concentration function, m_{tot} is the total mass, Ω [L] is the physical domain, m_i is the mass of the i^{th} particle, $\delta(x - X_i(t))$ is a Dirac delta function centered at each particle location $X_i(t)$ for $i = 1, \ldots, N$, and $\phi(x)$ [L^{-1}] is a kernel function. The probability of a particle's whereabouts is simply $p(X_i) = m_i/m_{tot}$. For simplicity here, we will use constant $m_i = m = 1/N$, which means that each kernel must integrate to unity and $m_{tot} = 1$. In general, the kernel function is not known or specified in the PT method. A common choice uses simple binning of arbitrary size Δx, which is identified with a generalized kernel that depends not merely upon the distance between particle positions and binning grid points, but each separately. In particular, the binning kernel function $\phi(x, X_i(t))$ is defined by

$$\phi(x, X_i(t)) = \begin{cases} 1, & \text{if } x \in [x_\ell, x_{\ell+1}] \\ 0, & \text{else} \end{cases} \tag{11}$$

where $\ell = \text{ceil}\left(\frac{X_i(t) - x_1}{\Delta x}\right)$ is the binning gridpoint to the left of the particle position and ceil(x) is the "ceiling" function.

More recent methods recognize that each particle is a random sample with PDF that is the Green's function, so that the kernel associated with each particle should have the same shape as $c(x,t)$. This should be implemented as an iterative process, in which 1) a simple kernel is assumed in Eq. (10); 2) an estimated $\hat{c}(x,t)$ is constructed; 3) a new kernel is estimated $\hat{\phi}(x) \propto \frac{1}{h}\hat{c}(\frac{x}{h}, t)$ for some $h > 0$, which is then 4) re-used in Eq. (10) to re-estimate $\hat{c}(x,t)$ until closure is reached. The closest approximation of this procedure was given by Pedretti and Fernàndez-Garcia (2013), in which a specific functional form—typically Gaussian—is chosen for $\phi(x)$, and the "size" or bandwidth h of the kernel is a weighted average of a constant, global bandwidth and an adaptive bandwidth based on a single-pass estimation of $\hat{c}$. The weighting is a linear average of particle arrival time rank, pre-supposing that later arrivals are less dense. This method would be difficult to apply for multi-dimensional (spatial) pdfs, so more recent methods directly calculate local particle densities that are then used to estimate each particle's unique bandwidth (Sole-Mari and Fernàndez-Garcia, 2018). Because of the convolutional form in Eq. (10) it is easy to show that the interpolation adds the variance of the kernel to the variance of particle positions, so the bandwidth h of the kernel must be kept small to minimize numerical dispersion from the interpolation process. It is unclear how the "pre-choice" of kernel function changes estimates of the entropy, as we discuss in the following section.

3. Entropy calculation

A problem with previous PT methods is that they do not automatically track dilution. As particles move, they do so as Dirac delta functions (i.e., the kernel itself is a Dirac delta), and the entropy is based on:

$$c_N(x,t) = \frac{1}{m_{tot}} \sum_{i=1}^{N} m_i \delta(x - X_i(t)) = \sum_{i=1}^{N} \frac{1}{N}\delta(x - X_i(t)) \tag{12}$$

so that

$$H_D(X) = -\sum_{i=1}^{N} \frac{m_i}{m_{tot}} \ln\left(\frac{m_i}{m_{tot}}\right) = -\sum_{i=1}^{N} \frac{1}{N}\ln\left(\frac{1}{N}\right) = \ln(N). \tag{13}$$

Not only does the entropy depend on the number of particles, but it is also constant over all simulation times because m_i and N do not change (although particle-splitting will unnaturally increase entropy). This also reveals a key feature of particle-tracking algorithms: the use of more particles implies greater entropy (mixing). This effect was shown in the context of chemical reactions (Benson and Meerschaert, 2008) and measured via concentration autocovariance functions (Paster et al., 2014). On the other hand, if each m_i changes due to mass transfer between particles, then H_D will change. The question is: does it do so in a manner expected by physical principles?

For the particle simulations that follow, we assume a simple problem that is directly solvable: one-dimensional (1-D) diffusion from an initial condition $c(x,0) = \delta(x)$. The solution is Gaussian, with consistent entropy from finite sampling given by:

$$\begin{aligned} H_C(X) &= -\int \frac{e^{-x^2/4Dt}}{\sqrt{4\pi Dt}} \ln\left(\frac{e^{-x^2/4Dt}}{\sqrt{4\pi Dt}}\Delta V\right)dx \\ &= -\ln\left(\frac{\Delta V}{\sqrt{4\pi Dt}}\right) + \frac{1}{2} \\ &= -\ln(\Delta V) + \ln\sqrt{4\pi Dt} + \frac{1}{2} \end{aligned} \tag{14}$$

This reveals a few interesting points regarding entropy calculation. First, for any finite sampling volume, the initial condition has unphysical $H_C = -\infty$. The calculation only makes sense after some "setting time" $t > (\Delta V)^2/(4\pi e D) \approx 0.03(\Delta V)^2/D$. Second, for a reliable estimation of entropy, the sampling interval for a moving plume must remain constant, which means that the sampling volume must be constant in space. For instance, if an Eulerian model possesses finer grids in some areas, the plume will appear to have changing entropy if the Eulerian grid is used for entropy calculation. Third, the sampling interval must be held constant in time. Very often, PT results are sampled at increasingly larger intervals as a plume spreads out (in order to reduce sampling error, (see Chakraborty et al., 2009)). Clearly, if the sampling size $\Delta V \propto \sqrt{t}$, then the calculated entropy will remain erroneously constant over time. Fourth, there are two components of the entropy calculation: one given by the PDF, and one given by the act of sampling, or the amount of information used to *estimate* the probabilities (the term inside the logarithm). This implies that, all other things held equal, a finely discretized model has greater consistent entropy. Typically, a model's fitness is penalized by its excess information content, but that is only represented (currently) by adjustable parameters (e.g., Akaike, 1974; Hill and Tiedeman, 2007). The definition of consistent entropy H_C suggests that the number of nodes or total calculations in a model should also contribute

to the penalty. A simple example and a derivation of a computational information criterion for numerical models is explored in Section 6 and Appendix A.

Unfortunately, a general formula that relates entropy growth with the characteristics of the kernel $\phi(x)$ cannot be gained because

$$H(X) = -\int \sum_{i=1}^{N} m\phi(x-x_i)\ln\left(\Delta V m \sum_{i=1}^{N}\phi(x-x_i)\right)dx$$
$$= -\ln(\Delta V m) - m\int \sum_{i=1}^{N}\phi(x-x_i)\ln\left(\sum_{i=1}^{N}\phi(x-x_i)dx\right), \quad (15)$$

and the logarithm of the sum inside the last integral does not expand. As a result, we will rely on numerical applications of several different kernels in computing the consistent entropy of Eq. (7).

4. Mass-Transfer PT Method

A recent PT algorithm (Benson and Bolster, 2016) implements mass-transfer between particles coupled with random-walk particle-tracking (MTPT). The mass transfer between particle pairs is based on the conceptualization of mixing as a simple chemical reaction (see Benson and Meerschaert, 2008; Benson and Bolster, 2016; Benson et al., 2019b). Specifically, full mixing between two particles possessing potentially different masses (or moles) a and b of any species Z can be written as the irreversible pseudo-reaction $aZ + bZ \rightarrow \frac{a+b}{2}Z + \frac{a+b}{2}Z$. This full mixing only occurs between two particles based on their probability of co-location in a time step of size Δt, and the algorithm is applied to all potentially interacting particle pairs. The algorithm has been shown to act as a globally diffusive operator (Schmidt et al., 2018) if the local mixing is modeled as diffusive (i.e., particles move and/or collide by Brownian motion). This means that, even if particles are considered Dirac delta functions, their masses continually change, and so the total entropy H_D must also change. The diffusive nature of the mass transfer may be coupled with random walks to fully flesh out the local hydrodynamic dispersion tensor. So between diffusive mass transfer, random walks, and local advection, the mass experiences the Green's function of transport (which may be complex due to variable velocities, (e.g., Benson et al., 2019)). A key feature of this algorithm is that the number of particles encodes the degree of interparticle mixing, which is separate from, but related to, the spreading of a diffusing plume (Benson et al., 2019; Schmidt et al., 2018). Because fewer particles implies greater average separation, the mixing lags behind the spreading of particles to a greater degree as N is decreased (Paster et al., 2014). However, it remains to be shown that this effect is reflected in the entropy of a conservative plume.

To briefly review, the mass-transfer PT method calculates the probability of collision between particles. This probability becomes a weight of mass transfer (Benson and Bolster, 2016; Schmidt et al., 2018), with the understanding that co-located particles would be well-mixed. As a result, for the i^{th} particle, the mass of a given species m_i satisfies to first order

$$m_i(t+\Delta t) = m_i(t) + \sum_{j=1}^{N}\frac{1}{2}(m_j(t) - m_i(t))P_{ij} \quad (16)$$

for $i = 1, \ldots, N$. For local Fickian dispersion, each particle pair's collision probability is given by (Benson and Meerschaert, 2008):

$$P_{ij} = (\Delta s/(8\pi\eta D_{ij}\Delta t)^{d/2})\exp(-r^2/(8\eta D_{ij}\Delta t)), \quad (17)$$

where Δs is the particle support volume, D_{ij} is the average D between the i and j particles, r is the distance between the i and j particles, and $0 < \eta < 1$ is the fraction of the isotropic diffusion simulated by inter-particle mass transfer. The remainder $(1-\eta)$ is performed by random walks. Here we use the arithmetic average $D_{ij} = (D_i + D_j)/2$. It should be noted that the Δs does not actually change the calculation of mass transfer because the probabilities are normalized, namely

$$\sum_{j=1}^{N} P_{ij} = 1, \quad \text{for all} \quad i = 1, \ldots, N. \quad (18)$$

If P is constructed as a matrix, this amounts to row normalization, which does not guarantee columns summing to unity. In practice, an average of row and column normalization is used to construct a symmetric and probability (mass) preserving matrix in which $P_{ij} = P_{ji}$. The calculated probabilities are normalized to sum to unity because mass must either move to other particles (when $i \neq j$) or stay at the current particle (when $i = j$). When particle masses are not all the same and particles are close enough to exchange mass, then the masses must also change, and therefore the entropy $H_D = -\sum_{i=1}^{N} m_i \ln(m_i)$ must change.

As discussed in the Introduction, in the presence of dispersion gradients, particles undergoing random walks must be pseudo-advected by the true velocity plus the divergence of dispersion. In contrast, the probabilities in Eq. (16) should automatically adjust for these gradients because the probability of mass transfer is not given solely by D at the i^{th} particle. Transfer is automatically lower in the direction of lower D, as opposed to the random walk algorithm, which moves a particle with a magnitude given by the value of D at the particle (and hence moves it too far into regions of lower D). Therefore, while the mass transfer algorithm has been shown to be diffusive, it should also properly solve the ADE with its dispersion gradients. However, this effect has yet to be investigated, so we provide evidence via a case study of transport in shear flow in Appendix B.

Within Eqs. (16) - (18), it appears that the collision probabilities act as a kernel to redistribute mass. In other words, rather than create a new interpolated concentration function as a convolution of the particle masses, the collision probability directly re-distributes the particle masses via convolution. Because the convolution kernel is the collision probability, we will refer to Eq. (17) as the "collision kernel". Several researchers (Rahbaralam et al., 2015; Sole-Mari et al., 2017; Sole-Mari and Fernàndez-Garcia, 2018) have suggested that the kernel representing the mass transfer should actually be a function of total simulation time and/or particle number and local density (through the statistics of the particle distribution), and not merely the time interval over which the particle undergoes some small-scale motions. To summarize, these authors perform smoothing in order to most closely solve Eq. (1), i.e., the case in which mixing and dispersion are both equally modeled by the diffusion term. Another effect of this operation should be to most closely match the entropy of the (perfectly-mixed) analytic solution of the diffusion equation, so we investigate it here.

Recently, Sole-Mari et al. (2019) showed that MTPT can be generalized so that particles can use a Gaussian function (kernel) other than the particle/particle collision probability of Eq. (17) for the mass transfer. In doing so, the methodology can be made numerically equivalent to smoothed particle hydrodynamics (SPH) simulations. The choice of kernel has an effect on simulation accuracy (Sole-Mari et al., 2019), which we theorize also changes the entropy, or mixing, within the simulations. Specifically, for the mixing pseudo-reaction we study here, Sole-Mari et al. (2019) rewrite the mass transfer function (16) in the more general form

$$m_i(t+\Delta t) = m_i(t) + \sum_{j=1}^{N}\beta_{ij}(m_j(t) - m_i(t))P_{ij}, \quad (19)$$

where

$$\beta_{ij} = \frac{2\eta D_{ij}\Delta t}{h^2}, \quad (20)$$

and the expression for P_{ij} in Eq. (17) is also modified by the kernel bandwidth choice:

$$P_{ij} = (\Delta s/(2\pi h^2)^{d/2})\exp(-r^2/(2h^2)). \quad (21)$$

The kernel bandwidth h depends, at any time, on the global statistics of the particle distribution. For this reason, it is called an *adaptive* kernel

(Silverman, 1986). More specifically, we set it as the value that minimizes the asymptotic mean integrated squared error (AMISE) of a kernel density estimation. The following expression is valid for a density estimation with a Gaussian kernel and particles carrying identical masses (Silverman, 1986):

$$h_{\mathrm{DE}} = \left(\frac{d}{(2\sqrt{\pi})^d N \int (\nabla^2 f)^2 \mathrm{d}x} \right)^{1/(d+4)}, \tag{22}$$

where f is the (usually unknown) true distribution of solute mass. For the present diffusion benchmark problem, f is a zero-mean Gaussian with variance $2Dt$, so the density estimation kernel is Gaussian with (Sole-Mari et al., 2017)

$$h_{\mathrm{DE}} = 1.06 N^{-1/5} \sigma = 1.06 N^{-1/5} \sqrt{2Dt}. \tag{23}$$

This bandwidth can be used to interpolate the classical PT method, for example using a Gaussian kernel in Eq. (10).

In the case of MTPT, however, we do not have a variable density of particles with identical masses, but a constant density of particles with variable masses. As an approximation, we replace the number of particles N in Eq. (22) with the equivalent value for which the average particle density ρ would be equal in the two cases

$$\rho = N \int f^2 \mathrm{d}x, \tag{24}$$

which allows us to rewrite expression (22) as an approximation for MTPT:

$$h_{\mathrm{SPH}} = \left(\frac{d \int f^2 \mathrm{d}x}{(2\sqrt{\pi})^d \rho \int (\nabla^2 f)^2 \mathrm{d}x} \right)^{1/(d+4)}. \tag{25}$$

Once again, because of the simple benchmark problem studied herein, there is a very simple solution for the bandwidth, because the distribution f at any time is a Gaussian with variance $\sigma^2 = 2Dt$. Furthermore, if N particles are placed within an interval of length Ω with average spacing $\Omega/N = 1/\rho$ which doesn't change significantly during a simulation, then the bandwidth reduces to

$$h_{\mathrm{SPH}} = 0.82 \sigma^{4/5} \rho^{-1/5} \approx 0.82 (2Dt)^{2/5} (N/\Omega)^{-1/5}. \tag{26}$$

We have implemented the adaptive kernels as both the density interpolator ϕ of the classical random walk at any time (i.e., a Gaussian kernel with variance h_{DE}^2 in Eq. (10)) and also in the mass transfer coefficient (20) and the probability weighting function (21) with bandwidth h_{SPH} in the mass-transfer algorithm (19).

5. Results and discussion

All simulations use $D = 10^{-3}$ [$\mathrm{L^2T^{-1}}$] and are run for $t_{final} = 1000$ arbitrary time units. The spatial domain is arbitrary, but for the MTPT method, we randomly placed particles (with zero initial mass) uniformly on the interval [-5,5], which is approximately $\pm 3.5\sqrt{2Dt_{final}}$. Note that the units are arbitrary but must be internally consistent because of the scale-invariance of the solutions to the diffusion equation that follow (in 1-D) $c(x,t) = (2Dt)^{-1/2} c\left(x(2Dt)^{-1/2}, 1\right)$. As long as the same units of D (say, meters and seconds) are used for elapsed time and the units of the spatial domain, the solutions are universal. More extensive discussions of multi-scale invariance in 3-D are given by Schumer et al. (2003).

The MTPT method can represent a Dirac delta function initial condition by any number of particles. Here we place one particle at $x = 0$ with unit mass. To enable direct comparison of consistent entropy between all methods, we chose equivalent average particle spacing and sampling volume of $\Delta V = \Delta x = 10/N$. We investigate the calculation of entropy and dilution indices for 1) The PT method using bins of size Δx; 2) The PT method using constant-size Gaussian interpolation kernels; 3) The PT method using adaptive kernels with bandwidth given by Eq. (23); 4) The MTPT method using a collision probability kernel size of $\sqrt{4D\Delta t}$; and 5) The MTPT method using adaptive kernels with size given by Eq. (26). With the latter two mass-transfer scenarios, we also let the proportion of diffusion by mass transfer (versus random walks) vary and focus on the two cases of $\eta = 1$ and $\eta = 0.1$ to see the effect of the collision-based versus SPH-based kernel size.

5.1. PT Versus collision kernel MTPT

First, we simulated the classical PT algorithm with concentrations mapped both by binning and by Gaussian kernels with fixed variance

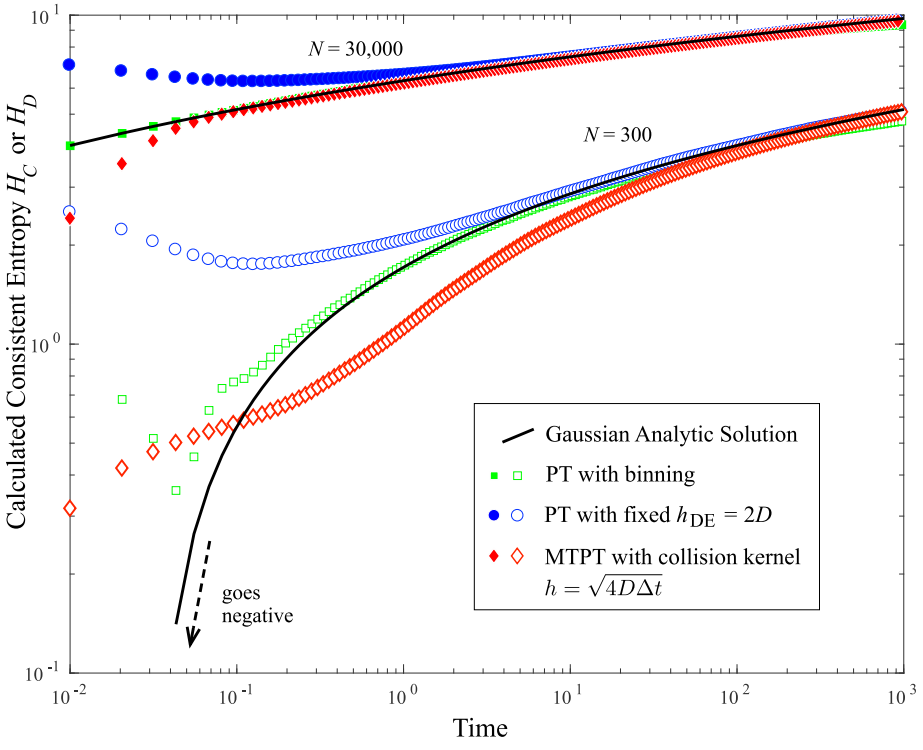

Fig. 1. Plot of calculated entropies H and H_C from single realizations of the 1-d random-walk diffusion problem.

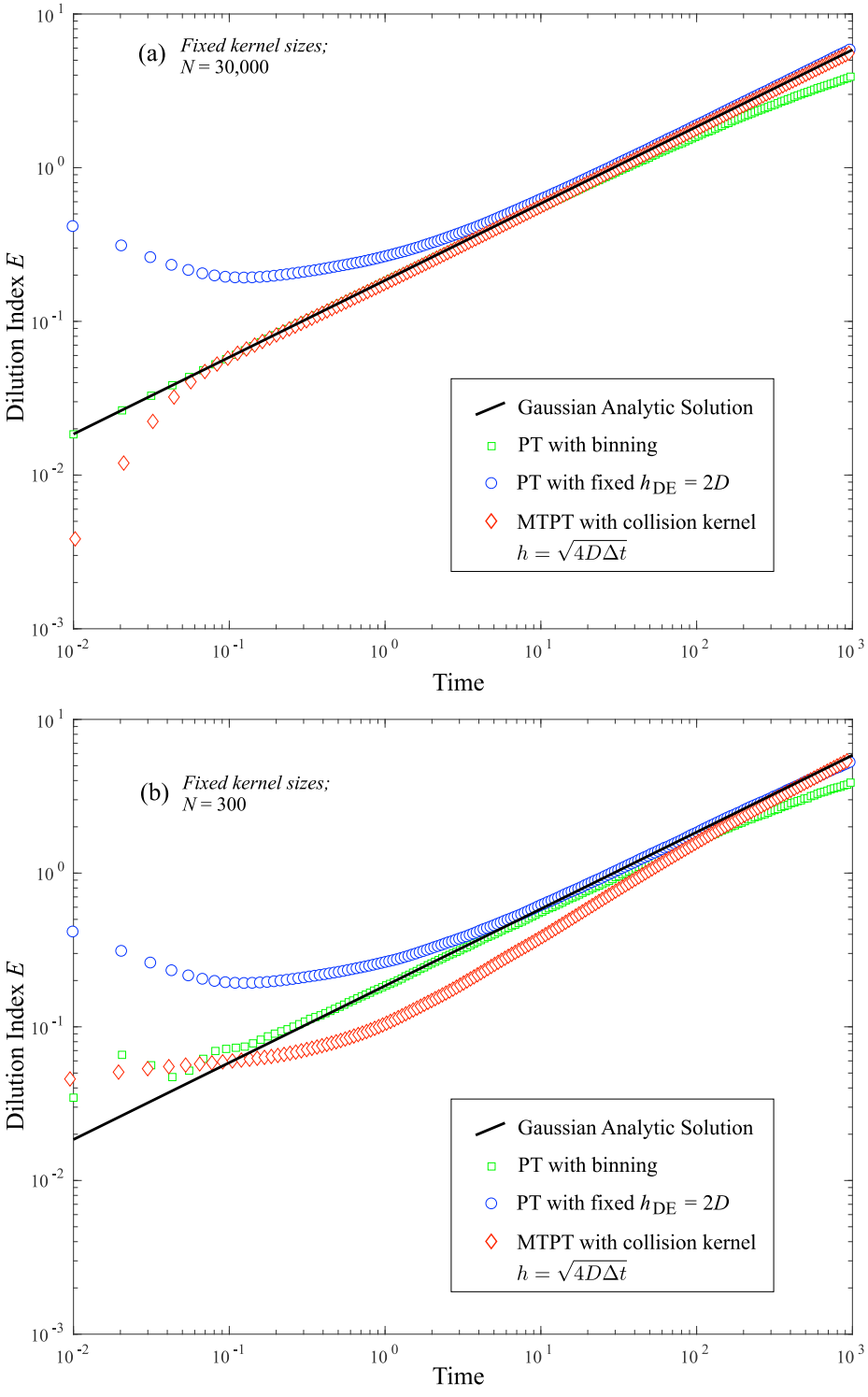

Fig. 2. Plots of calculated dilution indices E in the 1-d diffusion problem using interpolation of PT method and MTPT method for "fixed" collision kernels: (a) $N = 30,000$ and (b) $N = 300$.

$2D \times 1$ time unit. Because the simulations go from $t = 0.01$ to 1000, we chose a kernel size that is too big in the beginning and perhaps too small in the end (i.e., the kernel size is about 1/3 the spread of particles at $t = 10$). The calculated entropies from these simulations were compared to the analytic solution of Eq. (14) using $\Delta V = 10/N$ and the collision-kernel MTPT algorithm outlined in the previous Section 4. In these first MTPT simulations, we set the proportion of diffusion by mass transfer $\eta = 1$. In comparison to the other methods, the entropy from binned-PT concentrations matches the analytical solution very well at early times but significantly diverges later (Fig. 1). The difference between solutions is more obvious when looking at the dilution index E (Fig. 2). The fixed Gaussian-kernel interpolated concentrations over-estimate entropy and mixing at early time because a fixed kernel size is chosen that is typically larger than the actual diffusion distance for small times. The MTPT method underestimates entropy at early time relative to the analytic solution of Eq. (14) because the method, by design, does not perfectly mix concentrations. The random spacings impart regions where the particles are farther apart, and in these regions, the solutions are imperfectly mixed (i.e., imperfectly diffusive). As N gets larger, the solution is more

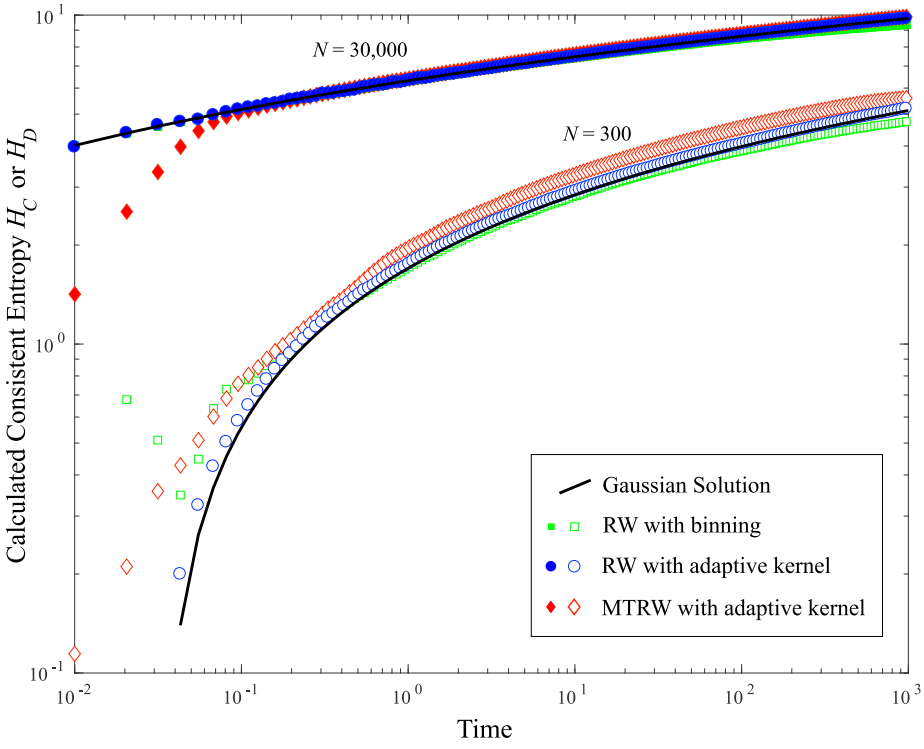

Fig. 3. Plot of calculated entropies H_C from ensemble averages of the 1-*d* random-walk diffusion problem using adaptive kernels for interpolation of simple random walks (blue circles) and for the mass-transfer particle-tracking algorithm (red diamonds) using $N = 30,000$ and $N = 300$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

perfectly-mixed and converges to the analytic diffusion kernel earlier (Figs. 1 and 2).

It is also important to note that neither the analytic solution nor the classical PT method represents the entropy of the initial condition correctly. The PT method, with all N particles placed at the origin, still has $H_D = \ln(N)$, while the entropy of the true Dirac delta initial condition is $H_D = -1\ln(1) = 0$. The analytic solution of Eq. (14) must use a calculation grid with finite Δx and sampling volume ΔV. In order for later-time entropies to match, this must be chosen as the same size as the bins for the PT method, i.e., $\Delta x = (x_{\max} - x_{\min})/N$, where the extents are chosen to almost surely see all particles in a simulation.

On the other hand, the MTPT method can represent the initial condition in many different ways, but here we simply placed one particle at the origin with unit mass, while the remaining $N - 1$ particles are placed randomly from the uniform distribution on $-5 < x < 5$ with zero mass. Because of this IC, the MTPT method can faithfully represent $H_D(t = 0) = 0$, and the effect of this deterministic, unmixed, IC stays with the simulations for a fair amount of time. At later time, both the fixed kernel PT and the MTPT methods converge to the analytic solution (Figs. 1, 2). At early times, however, the fixed kernel interpolator overestimates mixing when generating $c(x, t)$, not only with respect to the Gaussian solution, but also relative to the true initial condition with $H_D = 0$. Note also that the calculations of consistent entropy H_D depend strongly on N, but not the dilution index E which accounts for the different sampling (support) volumes.

5.2. Adaptive kernel versus collision kernel MTPT

We now turn to simulations using adaptive kernels, in which the particle-particle interaction probability has a time (and particle-number) varying kernel size by placing Eq. (25) into Eq. (21). This is predicated on the fact that a finite sampling of independent random variables is often used to create a histogram of those RVs. The idea is that a re-creation of the histogram should allow each sample to represent a larger domain than just its value, and a kernel should be assigned to spread each sample value. In the case of independent, mass-preserving random walks, the idea is clearly sound: for a delta-function initial condition, each particle is a sample with a PDF that is the Green's function, so that each particle's position could be viewed as a rescaled Green's function which is approximated by the histogram itself. The rescaling depends on the actual Green's function, which may vary in time and space, and the particle numbers. For independent particles undergoing Brownian motion, the Green's function is Gaussian with variance $2Dt$, and the kernel is shown to be Gaussian with zero mean and standard deviation given by Eq. (25). It is less clear whether this kernel should be used to represent the particle-particle interaction probability. First, the global statistics are not important to local mixing or reactions, i.e., a paucity of a reactant in one location is not informed by a wealth of reactant outside of the diffusion distance in one timestep. Second, the masses present on particles are anything but independent, as they depend strongly on their near-neighbors. Third, the kernels are designed to create a maximally smooth PDF based on random samples, but much research has shown that small-scale fluctuations are the most important driver of mixing and reaction rates. Thus, any kernel that smooths the local fluctuations is artificially increasing mixing and resulting reaction rates. However, much of this discussion is pure speculation, so we implement the kernel functions here as both interpolants of independent random walks and as weights in the mixing function.

For brevity and consistency with the previous results, we only show simulations with $N = 300$ and $N = 30,000$. Intermediate numbers track the same trends. For both particle numbers, the kernel-interpolated classical PT method has consistent entropy and dilution indices that match the diffusion equation analytic solution quite nicely (blue circles, Figs. 3 and 4). The kernels perform exactly as designed for optimally interpolating the PDF of independent, randomly-walking particles. For the lower particle number (300), the adaptive kernels in the MTPT algorithm match the analytic solution more closely than the collision kernel at early times (compare Figs. 1 and 3). The analytic solution assumes perfect mixing, i.e., local mixing and spreading are equal and characterized by the single coefficient D.

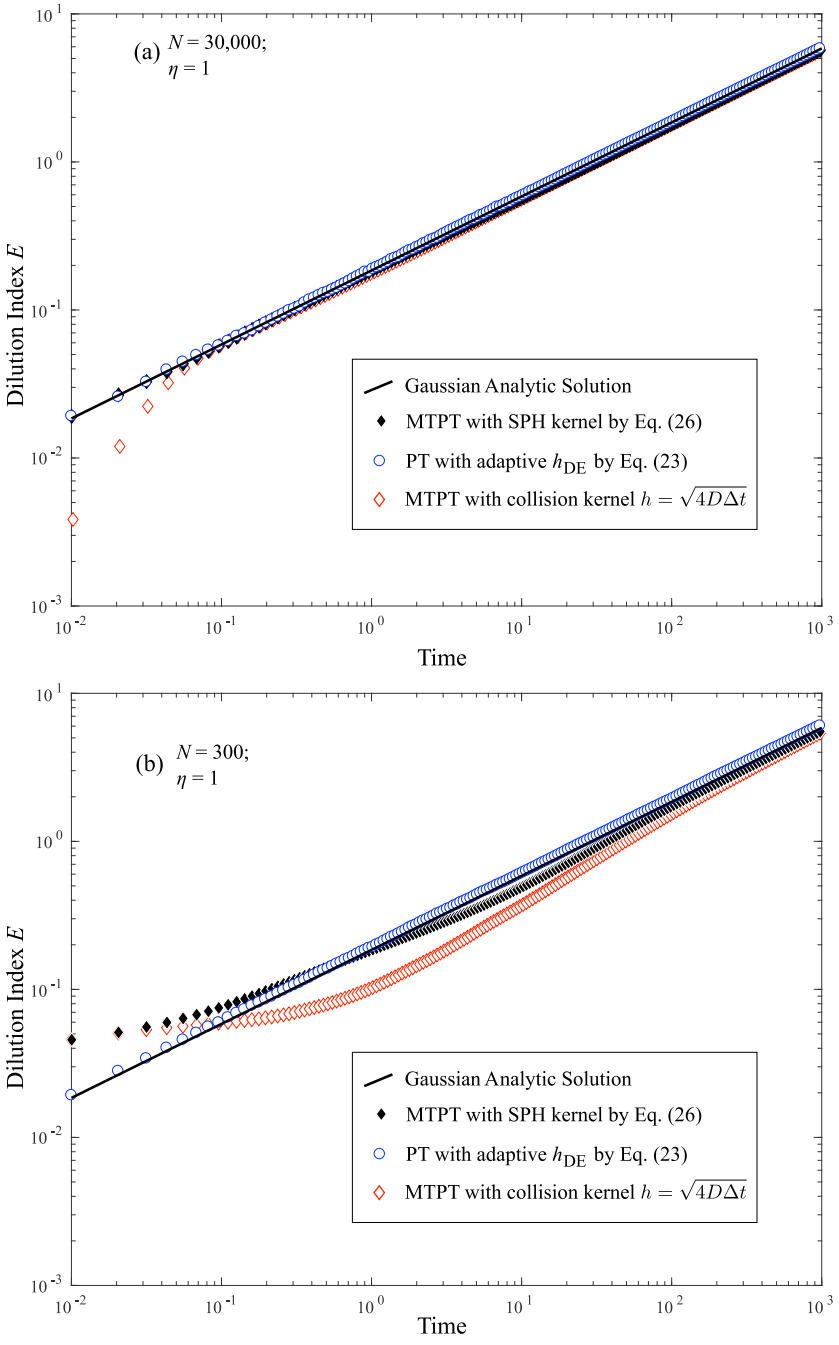

Fig. 4. Plots of calculated dilution indices E in the 1-D diffusion problem using adaptive kernels for interpolation of simple random walks (blue circles) and for the mass-transfer particle-tracking algorithm (red diamonds) for (a) $N = 30,000$ and (b) $N = 300$. MTPT with collision kernel results reproduced from Fig. 1 as grey diamonds for comparison. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

5.3. *Partitioning of local mixing and random walk spreading*

Recent studies (Benson et al., 2019; Schmidt et al., 2018) that employ the collision kernel for mass transfer have shown that mixing can be simulated as a smaller-scale (and smaller magnitude) process than solute spreading. This concept relies on the fact that upscaling by volume averaging and/or projection of 3-D concentrations to 2-D or 1-D replaces multi-valued concentrations with an average (e.g., Taylor, 1953). The spreading or warping of a concentration interface is a faster process than actual mass transfer across the interface. This is exemplified by miscible displacement of one fluid by another in laminar Poiseuille flow in a tube, where higher velocity in the center warps an initially sharp interface much faster than molecular diffusion actually mixes the fluids. When volume averaged to 1-D (Taylor, 1953), the spreading is given by a macro-dispersion coefficient that grows in time to an asymptotic value. Benson et al. (2019) showed that Taylor's macrodispersion can be performed by random walks, which causes particles to spread apart on average, while true mixing by molecular diffusion is performed by inter-particle mass transfer using the physics-based collision kernel. It is unclear whether using the adaptive SPH kernels as defined in Eq. (26) can achieve the same effect, given that the particle spreading is part of the evaluation of the kernel size for smaller-scale mixing. To investigate this effect, we chose a simple system in which the "macrodispersion" portion of D was the largest part (and also constant over time) by setting a constant mixing proportion $\eta = 0.1$ and re-ran the MTPT simulations for $N = 300$ and $N = 30,000$. Only the dilution indices are shown here, in Fig. 5. The differences between results for the collision kernel are small, while the adaptive kernel shows significantly decreased mixing.

This increased error for the adaptive kernel when $\eta \ll 1$ can be explained as follows. Expression (26) was obtained from (25) by assuming

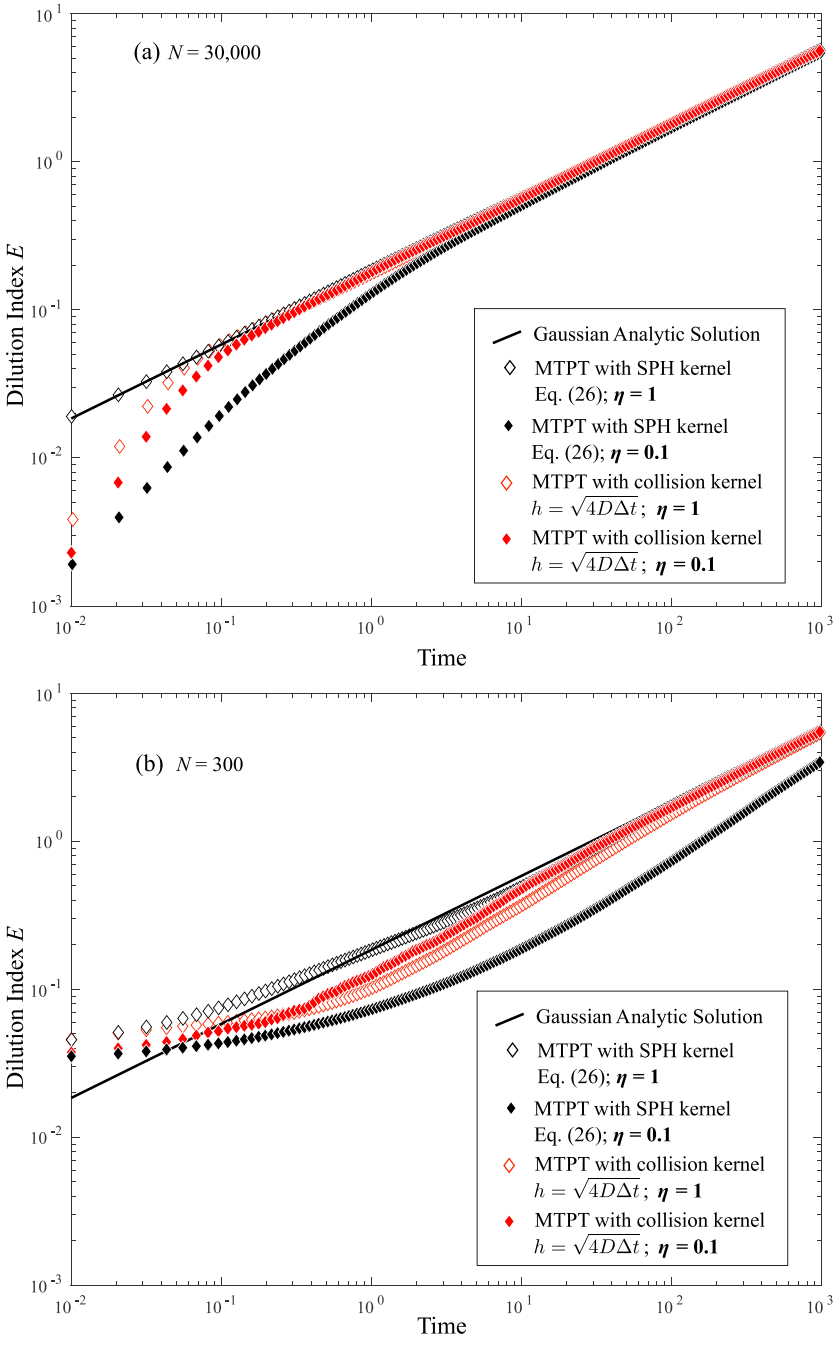

Fig. 5. Dilution indices for mixing/spreading proportions $\eta = 1$ and 0.1 for (a) $N = 30,000$ and (b) $N = 300$.

that the spatial distribution of the solute (f) is represented by a Gaussian function with variance $2Dt$. While this is approximately true for $\eta = 1$, the micro-scale variability generated when $\eta = 0.1$ (see Fig. 7a) suggests that f may not even be continuous and twice-differentiable to start with (which is a requisite for expression (25) to be valid). Nevertheless, if $\int(\nabla^2 f)^2 dx$ was to be numerically estimated each time step (such as in Sole-Mari and Fernàndez-Garcia (2018)), it would be much higher than for a Gaussian f with variance $2Dt$, because of the strong, small-scale concentration variations, suggesting that the truly optimal adaptive kernel obtained from Eq. (25) in this case would be much smaller than Eq. (26).

5.4. Distributional entropy

As noted earlier, particle simulations display greater entropy with an increasing number of particles (e.g., Figs. 1 and 3). In a similar way that the consistent entropy is related to classically defined inconsistent entropy for a continuous RV by adding the sampling portion: $H_C = -\ln(\Delta V) + H_I$, the portion of the entropy of a discrete RV can be partitioned into particle number and underlying "structure" of the PMF: $H_{\text{PMF}} = \ln(\Omega/N) + H_D$. Using this adjustment, the amount of mixing (given by the rate of convergence to the Gaussian) between simulations with different particle numbers can be compared (Fig. 6). Here, we ran MTPT simulations using the collision kernel with particle numbers in the set {100, 300, 1000, 3000, 10000, 30000}. For smaller N, the ensemble average of up to 20 realizations is used because of differences between individual runs. Quite clearly, the smaller particle numbers experience delayed convergence to the well-mixed Gaussian. This is a feature of the MT algorithm that is usually reflected in reduced reaction rates. But a simple measurement of the reduced entropy creation *rate* with smaller particle numbers is a sufficient demonstration of suppressed mixing.

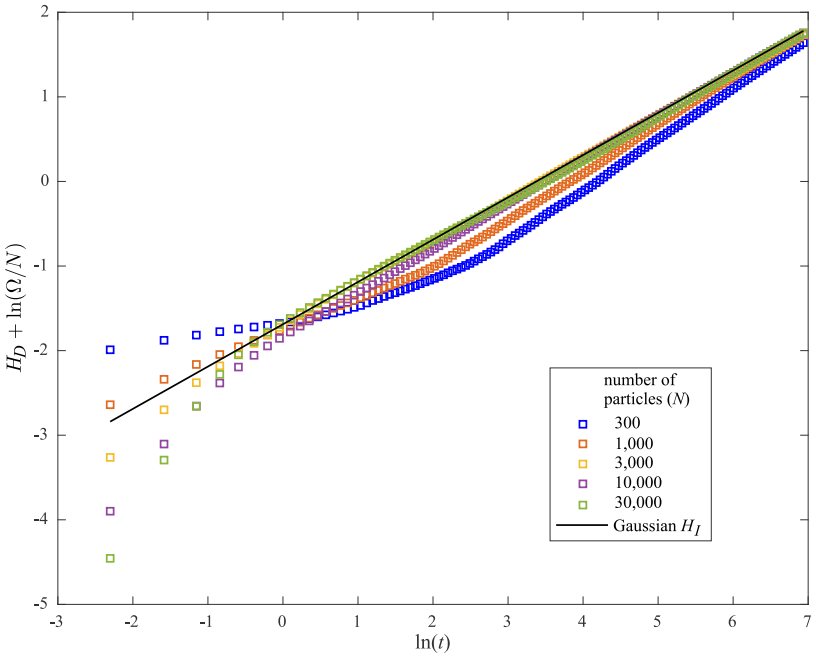

Fig. 6. Plots of relative, or PMF, entropy $H_{\text{PMF}} = H_D + \ln(\Omega/N)$ growth over time for different particle numbers diffusing under the MT algorithm. Also plotted is the $H_I(t)$ for a Gaussian diffusion (i.e., Eq. (14) using $\Delta V = 1$).

It is also instructive to inspect the plots of the calculated PMFs and PDFs from the $\eta = 0.1$ simulations (Fig. 7). The collision kernel MTPT method is notable because the degree of mixing and the shape of the plume are somewhat independent. Random walks may place particles with different masses in arbitrarily close proximity, and some time must elapse before local mixing equilibrates those masses (e.g., Fig. 7a). The result is the mass (or concentration) at any single position in 1-D space possesses substantial variability. This feature—concentration fluctuations at any point in space—has been exploited to perform accurate upscaling of transport and reaction in heterogeneous velocity fields (Benson et al., 2019; Cirpka and Kitanidis, 2000a, 2000b; Dentz et al., 2000; Dentz and Carrera, 2007). On the other hand, the fixed kernel interpolation of classic PT methods replaces this concentration variance at every location with concentration variability in space (see blue circles in Fig. 7b).

6. Computational entropy penalty

Numerical models provide discrete estimates of dependent variables that may be continuous functions of time and space. Often, the functions are non-negative and can be normalized to unit area so that they are PDFs. Therefore, the underlying "true" PDF has a certain entropy, and the sampling, or computational, procedure used to approximate these functions adds some artificial entropy because of the information required by the discretization. One desirable trait of a model is a parsimonious representation of the true physical process, i.e. fewer model parameters are preferred. At the same time, a more straightforward and accurate computational process is also preferred. Considerable attention has been paid to parsimonious (few parameter) models, but less attention has been paid to model computational requirements. Eq. (7) shows that, if a true PDF can be estimated via very few sampling points or nodes, there is less additional entropy incurred in the calculation. That is to say, if two models (with the same parametric parsimony) yield equivalent estimates of the underlying "true" dependent variable, then the model that estimates the PDF with the coarsest sampling, or least computationally intensive structure, is preferred from an entropic standpoint. Augmenting the Kullback-Leibler (inconsistent) representation of model entropy with the consistent entropy (Appendix A) allows us to compare the discrete PMFs obtained from computational approximation with the underlying PDF, and ultimately results in the COMputational Information Criterion (COMIC) as a natural extension of Akaike's information criterion (Akaike, 1974; 1992). To emphasize the influence of computational entropy, we illustrate two examples here by estimating a true diffusion given by a Gaussian with variance $2Dt$ by several numerical calculations with zero adjustable parameters (i.e., D is a known parameter).

6.1. Finite-difference example

For simplicity, we set $\Delta V = \Delta x = \Omega/\mathcal{N}$ for a fixed domain Ω and $\mathcal{N}$ nodes, and then compared the numerical estimation of the Green's function of the 1-D diffusion equation given by implicit finite-difference (FD) models with different discretizations $\Delta x \in \{0.4, 0.12, 0.04, 0.012, 0.004, 0.0012, 0.0004\}$. Other numerical parameters were held constant, including $\Omega = [-6, 6]$, $D = 10^{-3}$, and $\Delta t = 0.05$. We use all of the data from each model to calculate the mean SSE (i.e., the mean SSE is independent of subsampling), so here $n = \mathcal{N}$. Clearly a smaller Δx provides a better estimate of the analytic solution of a Gaussian with variance $2Dt$, but at what cost? Do 100 nodes suffice? A million? Because there are no adjustable parameters, the AIC, which is given in terms of the log-likelihood function $\text{AIC} = 2\ln(\text{SSE}/\mathcal{N})$, is a decreasing function of the number of nodes $\mathcal{N}$ (Fig. 8a). If, however, one factors in the penalty of $\ln(\Delta V)$, there is an optimal tradeoff of accuracy and computational entropy at $\mathcal{N} \approx 3000$ at almost every time step (Fig. 8b). Fewer nodes are not sufficiently accurate, and more nodes are superfluous for this particular problem, as shown by plotting the relative fitness criteria (AIC versus COMIC) for each discretization at some time (Fig. 8c).

Four important points regarding the COMIC immediately arise:

1. A model is typically sampled at a finite and fixed number of data measurement locations. We also sampled the many FD models and analytic solution at 15 randomly chosen "measurement points" common to all simulations and found nearly identical (albeit more noisy) results. However, we have not yet investigated the effect of additional sample noise on discerning the optimal discretization.
2. The AIC was derived with the assumption that the number of sample points and computational burden of models is identical and do not contribute to the relative AIC. Often, the common factors are elim-

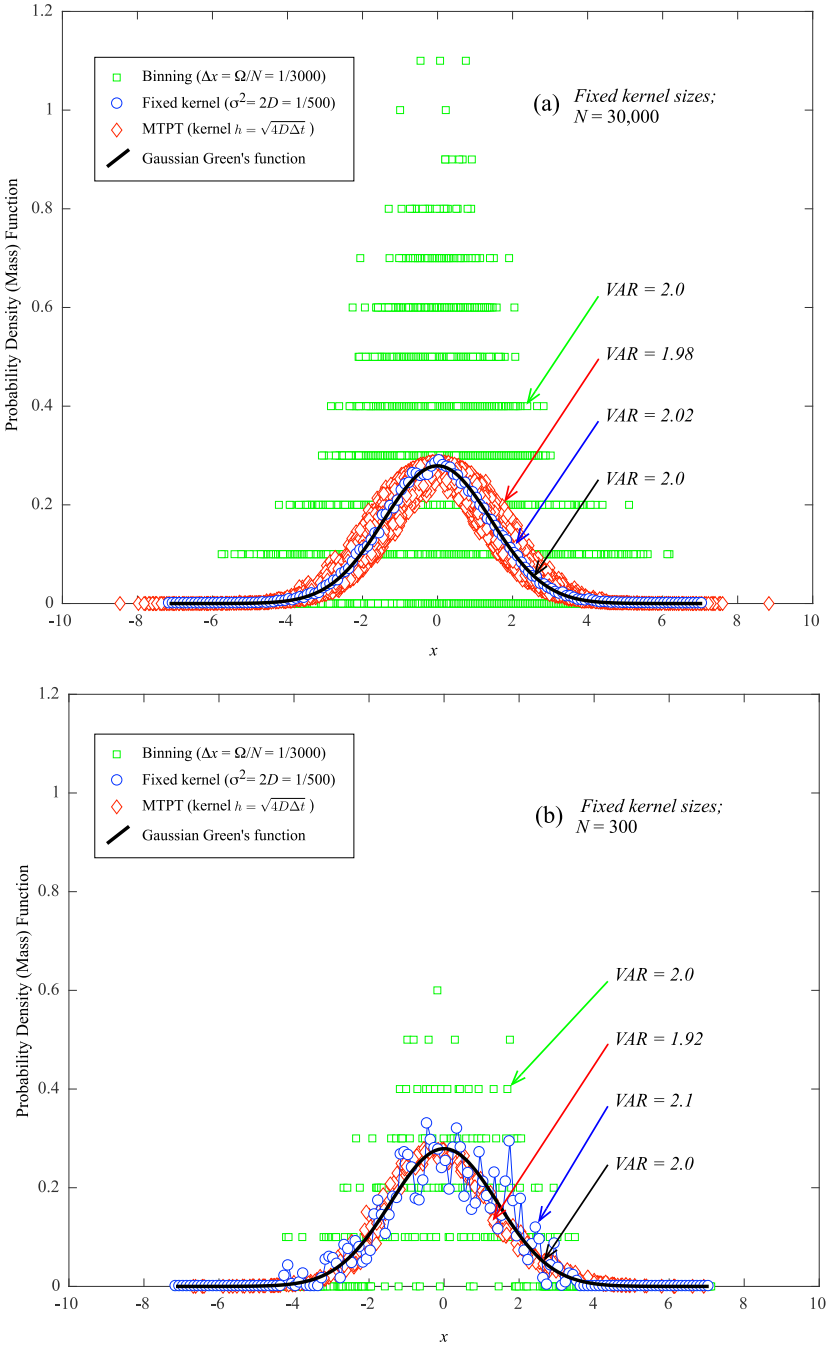

Fig. 7. Plot of calculated PMFs and PDFs (and their variances) in the 1-*d* diffusion problem using "fixed" kernels for (a) $N = 30,000$ and (b) $N = 300$.

inated from the AIC, and some arbitrary constants are also added, with no effect on *relative* AIC. In considering the COMIC, however, the choice of likelihood function and inclusion of constants may change the optimal model, so care in the choice of AIC is required.

3. The numerical solutions at some final time T are actually conditional densities of the joint densities $c(x, t)$, so that increased number of timesteps should also increase computational entropy (i.e., Δt contributes to the multidimensional ΔV, see Appendix A). Here we held the time step size constant for all FD models, so that the temporal sampling $t = j\Delta t$ has no effect on the *relative* entropy.
4. We used a constant spatial discretization $\Delta V = \Delta x$ to simplify the comparative Kullback-Leibler measures. Some models use variably-spaced grids, so the resulting computational entropy is more complicated than we investigate here.

6.2. Mass-transfer particle-tracking examples

Regarding this last point stated above for finite-difference models, the main thrust of this paper is the entropy of particle methods. The particles are typically randomly spread in space, so that a constant ΔV is not possible. However, using the inconsistent entropy isolates the correspondence of the N particles to an underlying PMF (e.g., Fig. 6). In the case of perfectly-mixed Fickian diffusion, this enables a direct comparison of the fitness of the particle methods to simulating diffusion, and the correction term $-\ln(\Omega/N)$ is the entropy associated with computation. We use this correction, in analogy with the FD results above, to assess the entropic fitness of MTPT methods and test several intuitive hypotheses. First, prior research has shown that fewer particles in the collision kernel MTPT method represent poorer mixing (hence poor fitness when modeling perfectly-mixed Fickian diffusion). In the absence of mixing

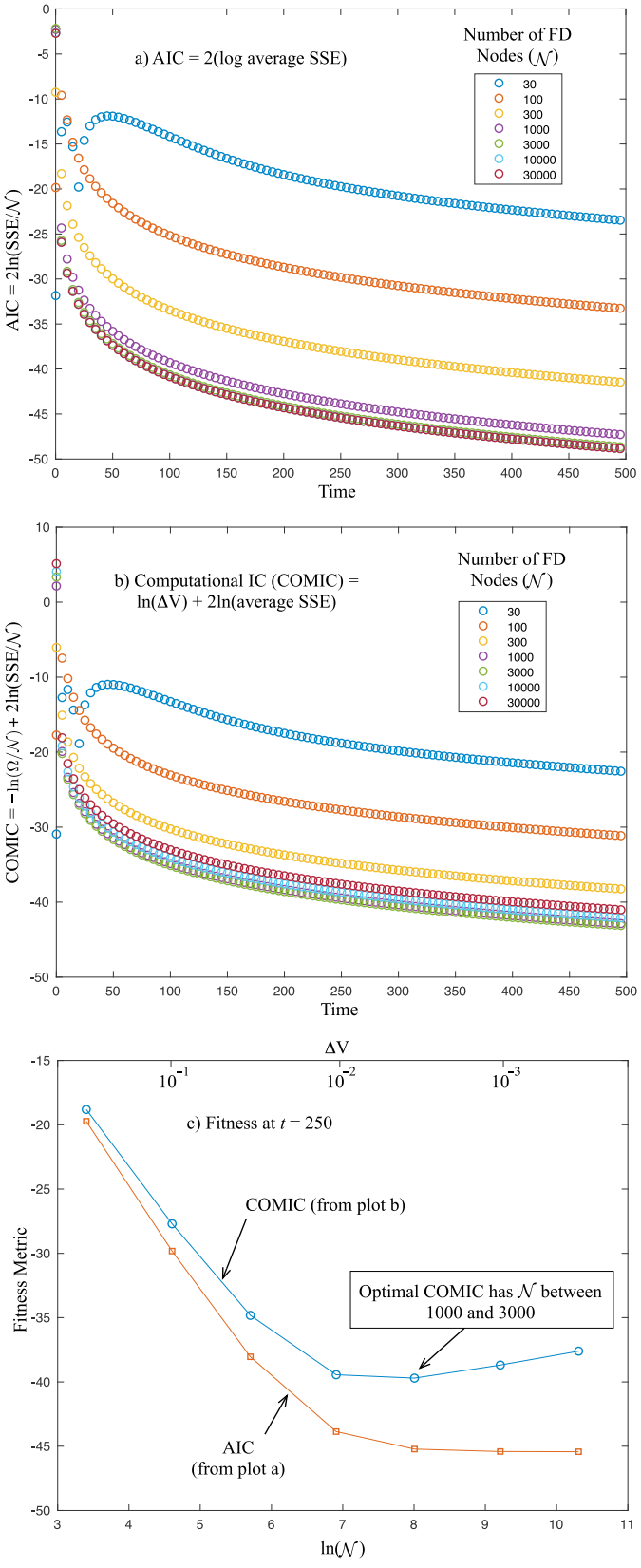

Fig. 8. Plots of relative model fitness measures for FD model: (a) log-likelihood function ln(SSE/$\mathcal{N}$); (b) computational information criteria COMIC=− ln(ΔV) + ln(SSE/$\mathcal{N}$); and, (c) both measures versus discretization at a single time t = 250.

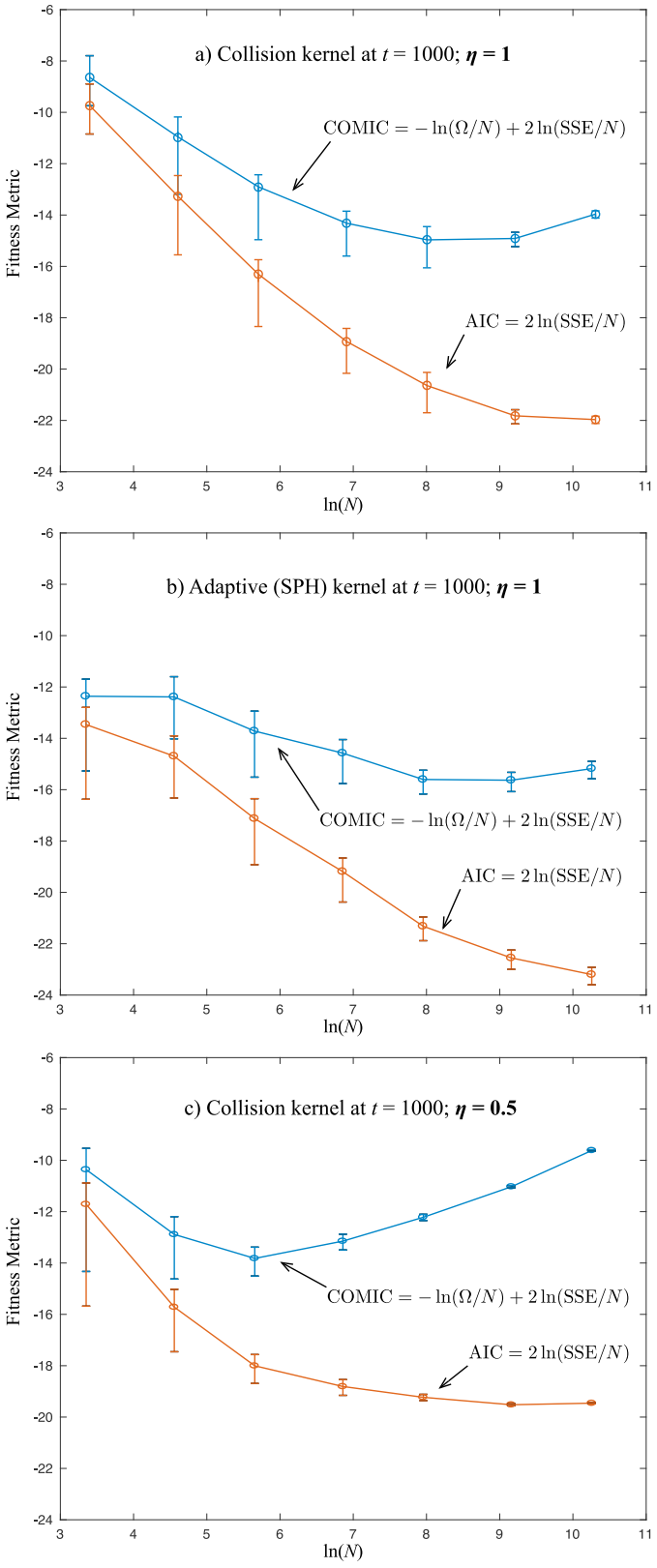

Fig. 9. Plots of ensemble statistics of relative model fitness measures for three MTPT models of Fickian diffusion at t = 1000: a) Using the collision kernel with all diffusion by mass transfer (η = 1); b) adaptive SPH kernel using Eq. (26) and full diffusion by mass transfer (η = 1); (b) collision kernel and half diffusion by mass transfer and half by random walks (η = 0.5). Error bars are ± one standard deviation in ensemble results.

by random walks (i.e., $\eta = 1$), we hypothesize that adding more particles will yield an improved average SSE, but that the overall model entropic fitness (measured by a smallest COMIC) reaches a maximum at some point. Indeed, a statistically significant minimum is found between $N = 1000$ and $N = 10,000$ particles, with an estimated minimum at $\approx$ 3000 particles (Fig. 9a).

On the other hand, the adaptive SPH kernel is constructed to best match Fickian diffusion by everywhere adjusting for particle density and number. Therefore, we hypothesize that the model entropic fitness will be relatively stable across a broad range of particle numbers. This is also found to be true in simulations (Fig. 9b), and COMIC fitness only suffers in a significant way for $N < 100$. Finally, in contrast to the collision kernel for $\eta = 1$ (shown in Fig. 9a), we hypothesize that splitting the diffusion between mass transfer and random walks will improve (for this example) the fitness of smaller particle number simulations by eliminating persistent "mixing gaps" where large random distances between particles prevents convergence to a well-mixed Gaussian. However, at some point, the model SSE will not improve with the addition of more particles because the "noise" of concentrations around the Gaussian will be saturated (see, e.g., Fig 7 a). Fig. 9c reveals exactly this behavior in the COMIC: adding random walks decreases the optimal number of particles to $\approx$ 300.

To summarize the MTPT entropic fitness for simulating Fickian diffusion: 1) for the SPH kernel, small particle numbers are sufficient and equally fit (by design); 2) similarly to the FD method, the collision kernel has a minimum COMIC around 3000 particles; and 3) with the collision kernel, partitioning diffusion by mass transfer and random walks promoted mixing and fitness for smaller particle numbers ($\approx$ 300) and clearly shows the superfluous nature of large particle numbers for simulating Fickian diffusion.

7. Conclusions

Classical PT methods do not track entropy until a concentration function is mapped from particle positions. The choice of bins or kernels for this mapping cannot be arbitrary, as the choice directly changes the calculated entropy, or degree of mixing, of a moving plume. The newer mass-transfer method directly simulates entropy without any such mapping (because particle masses continually change), and does so with several beneficial features. First, the zero-entropy initial condition, and its effect on the early portions of a simulation, are accurately tracked. Second, the particle number is an integral part of the mixing rate of a plume. Higher particle numbers simulate more complete mixing at earlier times, as shown by the convergence of entropy to that of a Gaussian. The MTPT method can use physically-based particle collision probabilities for the mixing kernel, or adaptive kernels dictated by the SPH algorithm. These adaptive kernels more closely match the analytic Gaussian solution's entropy when solving the diffusion equation in one pass (i.e., all mass transfer given by the diffusion coefficient). However, when the diffusion/dispersion is split between local inter-particle mixing and spreading by random walks, the adaptive-kernel entropies change substantially and do not match the Gaussian solution for small particle numbers. The collision kernel does not generate the same effect. We suggest that the adaptive SPH kernels only be used to solve locally well-mixed problems (i.e., where the dispersion tensor represents both mixing and dispersion equally), whereas the collision kernel may partition mixing and spreading as the physics of the problem dictate (Benson et al., 2019).

The fact that discrete (or discretized) approximations to real, continuous functions carry a sampling (or computational) entropy means that metrics which compare different simulations based on information content must be penalized by that computational information. For this purpose, we define a computational information criterion (COMIC) based on Akaike's AIC that includes this penalty. We show how a finite-difference solution of the 1-d diffusion equation has a well-defined optimal solution of about 3000 nodes in terms of combined accuracy and computational requirements. When the MTPT is used to simulate Fickian diffusion, these simulations show that the collision kernel also has a minimum COMIC around 3000 particles, but the adaptive SPH kernel, by design, is fit over a large range of particle numbers. Adding some diffusion by random walks makes the collision kernel a better fit for smaller particle numbers ($N \approx 300$), and shows that simulations of Fickian diffusion for large number of particles is computationally superfluous. We anticipate that this new entropy-based fitness metric may discount some overly computationally-intensive models that previously have been deemed optimal in terms of data fit alone.

Declaration of Competing Interest

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

CRediT authorship contribution statement

David A. Benson: Conceptualization, Methodology, Software, Writing - original draft, Writing - review & editing. **Stephen Pankavich:** Formal analysis, Writing - original draft, Writing - review & editing. **Michael J. Schmidt:** Methodology, Software, Writing - review & editing. **Guillem Sole-Mari:** Methodology, Software, Writing - original draft, Writing - review & editing.

Acknowledgements

We thank the editor, reviewers Daniele Pedretti and Olaf Cirpka, and one anonymous reviewer for extremely helpful comments. This material is based upon work supported by, or in part by, the US Army Research Office under Contract/Grant number W911NF-18-1-0338. The authors were also supported by the National Science Foundation under awards EAR-1417145, DMS-1614586, DMS-1911145, EAR-1351625, EAR-1417264, EAR-1446236, CBET-1705770, and DMS-1911145. The first author thanks the students in his class "GEGN 581: Analytic Hydrology" for inspiring this work. Two matlab codes for generating all results in this paper (one finite-difference and one particle-tracking) are held in the public repository https://github.com/dbenson5225/Particle_Entropy.

Appendix A. Computational information criterion and maximum likelihood estimators

The ultimate goal of this section is to derive an extension of *Akaike's* "an information criterion" (AIC) (Akaike, 1974) that establishes an objective function to be optimized in order to select a model and a minimal number of parameters that best fits a given set of data. Such an extension must incorporate the results of Section 2, which introduced a consistent notion of entropy that allows one to compare the relative entropies of a continuous PDF with that of a discrete PMF approximation. Of course, this discussion will first require some background knowledge of the Kullback-Leibler divergence, the basic formulation of the AIC, and maximum likelihood estimators, each of which we provide below.

A1. Kullback-Leibler divergence

We begin with a review of the Kullback-Leibler divergence and its extension to the inconsistent entropy for continuous RVs. Following Kullback (1968), we first consider the likelihood of two competing hypotheses h_1 and h_2 given some knowledge of the probability of an event x, and note that Bayes' Theorem provides a representation for the conditional probability of each hypothesis given x, namely:

$$P(h_i|x) = \frac{P(x|h_i)P(h_i)}{P(x|h_1)P(h_1) + P(x|h_2)P(h_2)} \tag{27}$$

for $i = 1, 2$. Next, this statement can be generalized to continuous RVs so that if h_1 and h_2 now represent the events that a random variable

X comes from a distribution represented by the PDFs $f_1(x)$ and $f_2(x)$, respectively, then the conditional probability of each hypothesis given that $X = x$ is now

$$P(h_i|x) = \frac{f_i(x)P(h_i)}{f_1(x)P(h_1) + f_2(x)P(h_2)} \tag{28}$$

for $i = 1, 2$. Taking logarithms and rearranging this expression then yields

$$\ln\left(\frac{f_1(x)}{f_2(x)}\right) = \ln\left(\frac{P(h_1|x)}{P(h_2|x)}\right) - \ln\left(\frac{P(h_1)}{P(h_2)}\right). \tag{29}$$

The right side of this equality is a measure of the difference between the logarithm of the odds in favor of hypothesis 1 (versus hypothesis 2) after the observation of $X = x$ relative to before this observation. In other words, this difference represents exactly the information contained within the observation that $X = x$, and the left side of the equality, often referred to as the log-likelihood ratio, is the information in favor of h_1 (and against h_2). To obtain the mean value of this, we merely integrate over all possible observations and against the PDF $f_1(x)$, which gives

$$I(f_1, f_2) = \int f_1(x) \ln\left(\frac{f_1(x)}{f_2(x)}\right) dx. \tag{30}$$

This quantity is defined to be the Kullback-Leibler divergence (KLD), and represents the entropy of f_1 relative to f_2. Notice that this expression can also be separated into two integrals so that

$$I(f_1, f_2) = \int f_1(x) \ln f_1(x) dx - \int f_1(x) \ln f_2(x) dx, \tag{31}$$

and the former of these two integrals is directly related to the inconsistent entropy of Eq. (3). Next, we will use the KL divergence to establish the AIC.

A2. Akaike information criterion (AIC)

The AIC was originally established to select a model and associated parameter values that are a best predictor of potential future data based on some set of given data. It is well-known that adding parameters will reduce data/model misfit for a single set of observations, but the added parameters will often cause worse fits for newly collected data (Konishi and Kitagawa, 2008). In particular, consider a variety of different models defined by distinct parameter vectors θ and corresponding PDFs $h(y|\theta)$ arising from data values $y_1, \ldots, y_n$, along with a single vector of "true" parameter values θ_0 with PDF $g(y) = h(y|\theta_0)$. The problem of interest is how to optimally select both a number of model parameters k and their associated values θ to best approximate θ_0 given that we have incomplete knowledge of the latter quantity. In fact, the information provided to make this decision arises only from the given data, which is merely a collection of n independent sample values, each representing a realization of a random variable Y with PDF $g(y)$. Ultimately, the AIC yields an approximate criterion for the selection of parameters, which entails minimizing the quantity

$$-2\sum_{i=1}^{n} \ln h(y_i|\hat{\theta}) + 2k \tag{32}$$

over the number of parameters k, where $\hat{\theta}$ is the maximum likelihood estimate for θ. Furthermore, this process corresponds to minimizing the underlying entropy among such models.

In the context of computing concentrations as in previous sections, we could consider a function $c(x, t)$ for which we have a coupled set of observed data, say $\{(x_i, c_i) : i = 1, \ldots, n\}$, which represents values of the concentration measured at differing spatial points and at a fixed time $t = T$. Here, the function c can be a solution to a PDE (e.g., Eq. (1)) and may depend upon some parameters θ, for instance, D in Eq. (1). Of course, the parameter values are unknown and must be inferred from the given data, which may contain some noise due to errors in measurement. We can then define y_i to be the corresponding error between c and the measured data c_i for every $i = 1, \ldots, n$ and consider the associated PDF for each of these errors, denoted $h(y|\theta)$. Additionally, θ_0 represents the "true" parameter values so that the underlying PDF can be represented by $g(y) = h(y|\theta_0)$. The AIC will then provide a criterion to select the parameter values (and number of parameters) that is a best predictor of any future concentration data, thereby selecting a specific model.

The selection criterion is based on the entropy maximization principle, which states that the optimal model is obtained by maximizing (over the given data, on which θ depends) the expected value of the log-likelihood function, namely

$$S(g, h(\cdot|\theta)) = \int g(y) \ln(h(y|\theta)) dy. \tag{33}$$

This quantity is not a well-defined (i.e., strictly positive) counterpart to entropy, as discussed in the main text. Thus, it is typically implemented in a relative sense among models using the Kullback-Leibler divergence of g and h, given by

$$I(g, h(\cdot|\theta)) = \int g(y) \ln\left(\frac{g(y)}{h(y|\theta)}\right) dy = S(g, g) - S(g, h(\cdot|\theta)). \tag{34}$$

As noted in the previous section, this can be interpreted as a measurement of the relative similarity between the probability distributions g and h. As Akaike (1974) notes, maximizing the expected log-likelihood above is equivalent to minimizing $I(g, h(\cdot|\theta))$ over the given data. Of course, since θ_0 is unknown and $g(y) = h(y|\theta_0)$ depends upon knowledge of the "true" parameter values, we cannot directly compute $I(g, h(\cdot|\theta))$. Instead, this quantity must be suitably approximated. Following Akaike (1974, 1992), when the number of data points n is sufficiently large, an approximation of the KLD using the Fisher information matrix can be utilized, and classical estimation techniques imply

$$I(g, h(\cdot|\theta)) \approx \left(\sum_{i=1}^{n} \ln g(y_i) - \sum_{i=1}^{n} \ln h(y_i|\hat{\theta})\right) + k, \tag{35}$$

where k is the number of estimated parameters within θ, and $\hat{\theta}$ is the maximum-likelihood estimate for θ. Here, k appears in order to correct for the bias introduced by approximating the "true" parameter values with their corresponding maximum-likelihood estimates. Finally, since the sum involving g is constant for any choice of model parameters, it can be omitted in computing the minimization. Therefore, the AIC may be defined (with a scaling factor of two, as in Akaike (1974)) by

$$\text{AIC} = -2\ln(\text{maximum likelihood}) + 2k, \tag{36}$$

or in the notation described herein

$$\text{AIC}(\hat{\theta}) = -2\sum_{i=1}^{n} \ln h(y_i|\hat{\theta}) + 2k. \tag{37}$$

It is this quantity that one wishes to minimize (over k, where $\hat{\theta}$ may depend upon k) in order to select the best model approximation to g, and this is the basis of our departure. Prior to formulating an extension of the AIC, we provide a brief review involving maximum likelihood estimators.

A3. Maximum likelihood estimators (MLEs)

Though we have not mentioned the process of obtaining the maximum-likelihood estimates $\hat{\theta}$ described above, useful discussions of MLEs for models with unknown structure are provided by Hill and Tiedeman (2007) and Brockwell and Davis (2016). As an example, consider the scenario in which the errors between model and observations are independent, zero-mean Gaussians. In this case the likelihood function is given by

$$L(y; \theta) = \left[(2\pi)^n |\Sigma(\theta)|\right]^{-1/2} \exp\left(-\frac{1}{2} y^T \Sigma(\theta)^{-1} y\right), \tag{38}$$

where n is the number of observation points, $\Sigma(\theta)$ is a covariance matrix of errors that depends upon some unknown parameter vector θ, and y is a vector of residuals satisfying $y_i = c_i - c(x_i, T)$ for $i = 1, ., n$. Recall that

c_i is the measured concentration and $c(x_i, T)$ represents the concentration at the spatial data point x_i and time T given by the PDE solution. Therefore, the log-likelihood function is

$$\ln(L) = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma| - \frac{1}{2}y^T\Sigma^{-1}y. \tag{39}$$

In practice, the observation errors are often assumed to be independent, and Σ is diagonal. Furthermore, the variance of each observation is often unknown or estimated during the model regression (although numerous approximations can be applied - see Chakraborty et al. (2009) for assumed concentration errors in particle tracking), so it is assumed that Σ depends only upon a single variance parameter, denoted by σ^2, and thus satisfies $\Sigma = \sigma^2\mathbb{I}$. The last term in Eq. (39) is more conveniently given in terms of the sum of squared errors $\text{SSE} = y \cdot y = |y|^2$ so that

$$\ln(L) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{n}{2\sigma^2}\frac{\text{SSE}}{n}. \tag{40}$$

Because this function should be maximized, one takes the derivative with respect to σ^2 and sets it to zero to compute the value of σ^2 at which the maximum occurs. This provides an estimator of the observation variance, namely $\hat{\sigma}^2 = \text{SSE}/n$, so that the corresponding maximum is

$$\ln(L) = -\frac{n}{2}\left(1 + \ln(2\pi) + \ln\left(\frac{\text{SSE}}{n}\right)\right). \tag{41}$$

Because the number of observations is usually fixed, the $\frac{n}{2}$ term is canceled from all terms (as maximizing $\ln(L)$ also maximizes $\frac{2}{n}\ln(L)$). Hence, the MLE is $\hat{\sigma}^2 = \text{SSE}/n$ and the quantity $-\ln(\text{SSE}/n)$ provides a relative estimate for the value of the log-likelihood function evaluated at the MLE.

A4. Computational information criterion

Returning to the formulation of the AIC for a model of concentration, we wish to alter the original derivation so that one may compare a variety of discrete computational models to the true, continuous model using their relative entropy to evaluate their similarity. In such a case, we would like to use the KLD to measure the relative entropy of the error distribution between the observed data and the solution of the PDE model, but we must also correct this criterion for the fact that our discrete approximations to the PDE solution change with the resolution of the chosen numerical method, which is often described by a single parameter N. For instance, this parameter can represent the number of particles N in a stochastic particle method or identify the spatial grid size, $\Delta x = \frac{\Omega}{N}$, in a finite difference method. Ultimately, we wish to establish a criterion to select the best of these approximate models depending upon the value of N.

As before, letting (x_i, c_i) for $i = 1, \dots, n$ represent the given pairs of concentration data and $c(x, t)$ denote the true solution to the PDE model, we can describe the statistical model incorporating measurement error by

$$c_i = c(x_i, T) + \epsilon$$

for $i = 1, \dots, n$, where T is a known measurement time and $\epsilon \sim h(y|\theta)$ is a random variable with distribution h that encodes each of the associated random errors. Because we often do not possess an analytic solution for c, it is necessary to approximate the PDE solution at time $t = T$ with a number of suitable numerical models, the solutions of which we denote by $c_N(x)$, with the model of interest changing with the value of N. Hence, discrete approximation error, in addition to measurement error from the data, must be incorporated into the model selection criterion. Because the KLD can account for the latter quantity, we can establish a computational information criterion merely by correcting the AIC by the difference in entropy between the PDE solution and the numerical approximation. In this way if $H_{rel}(f_1, f_2)$ represents the relative entropy between f_1 and f_2, then the new information criterion can be expressed as

$$\text{COMIC}(\hat{\theta}, N) = H_{rel}(c, c_i) + H_{rel}(c_N, c).$$

The first relative entropy here is given by the AIC, while the second is merely the difference between the discrete entropy of the approximate solution and the inconsistent entropy of the PDE solution, or $H_D - H_I$, where these quantities are given by Eqs. (2) and (3), respectively. In Section 2, a sampling volume was introduced to relate these terms using the consistent entropy of Eq. (6), and we found (see Fig. 6 with $\Delta V = \frac{\Omega}{N}$)

$$H_D - H_I \approx -\ln(\Delta V).$$

Using this, we can finally define an adjusted criterion to the AIC, which we name COMIC or the COMputational Information Criteria, given by

$$\text{COMIC}(\hat{\theta}; \Delta V) = -2\sum_{i=1}^{n}\ln h(y_i|\hat{\theta}) + 2k - \ln(\Delta V). \tag{42}$$

From an information theory perspective, this computational penalization can also be seen as a limitation on the information content needed to represent the approximate solution $c_N(x)$. Assuming the selected numerical method converges to the underlying PDE solution as $N \to \infty$, the values of $c(x, t)$ can be computed to an arbitrarily large degree of precision by merely choosing N sufficiently large in the computational model. However, in doing so, one must continue to store increasingly large amounts of information to gain smaller and smaller levels of accuracy. Thus, a trade-off results between the desired gain in precision and the stored information content, and the COMIC provides an efficient criterion for penalizing such considerations to select a parsimonious and computationally efficient (low information content) model.

In order to focus on the computational implications of this adjustment to the model selection criterion, we consider the case in which the errors between model and observations are Gaussian with variance σ^2, as in the example illustrated within the previous section, and assume that no other parameters (e.g., D) require estimation. In this case, the log-likelihood function evaluated at the maximum-likelihood estimate is proportional to the log of the average sum of squared errors (SSE) given by Eq. (41). Upon removing constants, the form of the COMIC becomes

$$\text{COMIC}(\Delta V) = 2\ln\left(\frac{\text{SSE}}{n}\right) - \ln(\Delta V), \tag{43}$$

where

$$\text{SSE} = \sum_{i=1}^{n}(c_i - c_N(x_i))^2. \tag{44}$$

For numerical models with equivalent SSE, their measure of distributional entropy is the same, but their computational entropy would be $-\ln(\Delta V)$, so that the model fitness should be adjusted by this difference in information content.

Appendix B. Effect of $\nabla \cdot \mathbf{D}$ on the mass-transfer algorithm

We illustrate the effect of spatially-variable $\mathbf{D}$ in simple 2-d shear flow, borrowing the parabolic velocity profile $v_y = 0$ and $v_x = -y^2 - by$ of Hagen-Poiseuille flow. The domain used here is $0 < x < 400$; $0 < y < 1$, with concentrations initially zero everywhere except for a strip $90 < x < 110$ with concentration 1/20, i.e., initial mass=1. The x-domain is periodic, so particles that exit at $x = 400$ are re-introduced at $x = 0$. We show a scenario with heterogeneous and anisotropic diffusion $\mathbf{D} = \begin{bmatrix} \alpha_L v_x & 0 \\ 0 & \alpha_T v_x \end{bmatrix}$, with longitudinal and transverse dispersivities $\alpha_L = 10^{-2}$; $\alpha_T = 10^{-3}$. Dispersive transport was simulated for $t = 500$ with timestep size $\Delta t = 1$ either solely by mass transfer or solely by random walks. Because the mass transfer algorithm can move mass among all particles in the domain, a total of 20,000 particles were placed in the

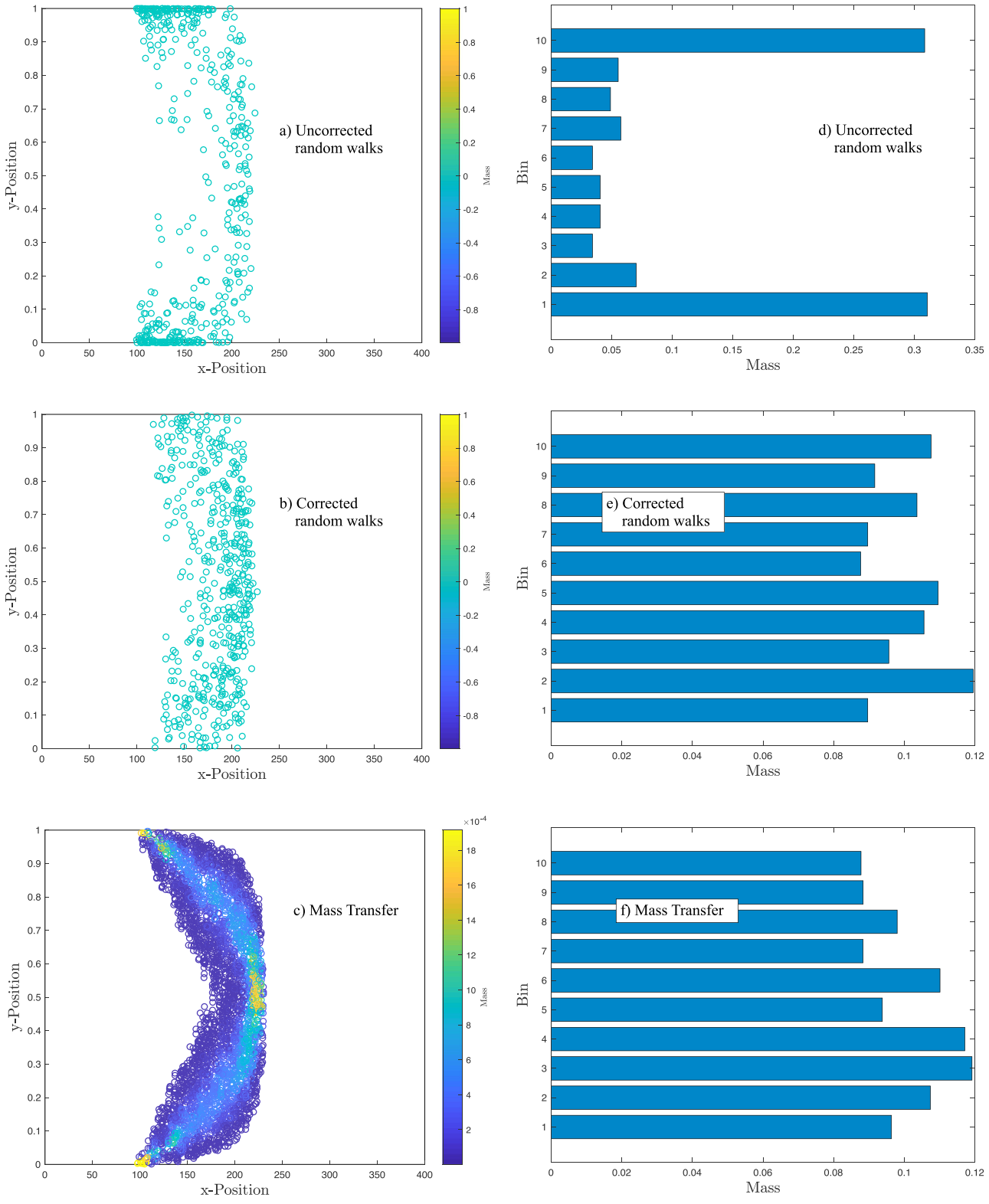

Fig. 10. a-c) Particle positions and masses in shear flow simulations. For clarity, only those particles with mass $> 10^{-6}$ are shown. d-f) Histograms on binned masses versus lateral y-position.

400×1 domain, with an average of 100 particles in the initial non-zero concentration strip. This gives plenty of "clean" particles on either side of the strip.

Pure random walks without the drift correction term migrate all particles, including those with mass, to the lower **D** regions (Fig. 10a). The drift correction eliminates the lateral bias (Fig. 10b and e). The mass transfer algorithm has no apparent bias or need for $\nabla \cdot \mathbf{D}$ correction (Fig. 10c and f). As an aside, the mass-transfer method quite clearly shows the regions of greatest, and least, shear and mixing (Fig. 10c).

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.advwatres.2020.103509.

References

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Autom. Control AC-19 (6), 716–723.

Akaike, H., 1992. Information theory and an extension of the maximum likelihood principle. Springer Series in Statistics, Perspectives in Statistics, pp. 610–624.

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.

Benson, D.A., Bolster, D., 2016. Arbitrarily complex chemical reactions on particles. Water Resour Res 52 (11), 9190–9200. https://doi.org/10.1002/2016WR019368.

Benson, D.A., Meerschaert, M.M., 2008. Simulation of chemical reaction via particle tracking: diffusion-limited versus thermodynamic rate-limited regimes. Water Resour. Res. 44, W12201. https://doi.org/10.1029/2008WR007111.

Benson, D.A., Pankavich, S., Bolster, D., 2019. On the separate treatment of mixing and spreading by the reactive-particle-tracking algorithm: an example of accurate upscaling of reactive Poiseuille flow. Adv. Water Resour. 123, 40–53. https://doi.org/10.1016/j.advwatres.2018.11.001.

Benson, D.A., Schmidt, M.J., Bolster, D., Harmon, C., Engdahl, N.B., 2019. Aging and mixing as pseudo-chemical-reactions between, and on, particles: perspectives on particle interaction and multi-modal ages in hillslopes and streams. Adv. Water Resour. 103386. https://doi.org/10.1016/j.advwatres.2019.103386.

Brockwell, P.J., Davis, R.A., 2016. Introduction to Time Series and Forecasting. Springer Texts in Statistics, third Springer.

Chakraborty, P., Meerschaert, M.M., Lim, C.Y., 2009. Parameter estimation for fractional transport: Aparticle-tracking approach. Water Resources Research 45 (10), W10415. https://doi.org/10.1029/2008WR007577. https://www.agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2008WR007577.

Chiogna, G., Hochstetler, D.L., Bellin, A., Kitanidis, P.K., Rolle, M., 2012. Mixing, entropy and reactive solute transport. Geophysical Research Letters 39 (20). https://doi.org/10.1029/2012GL053295. https://www.agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2012GL053295.

Chiogna, G., Rolle, M., 2017. Entropy-based critical reaction time for mixing-controlled reactive transport. Water Resources Research 53 (8), 7488–7498. https://doi.org/10.1002/2017WR020522. https://www.agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/2017WR020522.

Cirpka, O.A., Kitanidis, P.K., 2000a. An advective-dispersive stream tube approach for the transfer of conservative-tracer data to reactive transport. Water Resources Research 36 (5), 1209–1220. https://doi.org/10.1029/1999WR900355. https://www.agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/1999WR900355.

Cirpka, O.A., Kitanidis, P.K., 2000b. Characterization of mixing and dilution in heterogeneous aquifers by means of local temporal moments. Water Resources Research 36 (5), 1221–1236. https://doi.org/10.1029/1999WR900354. https://www.agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/1999WR900354.

Dentz, M., Carrera, J., 2007. Mixing and spreading in stratified flow. Phys. Fluid. 19, 17107.

Dentz, M., Kinzelbach, H., Attinger, S., Kinzelbach, W., 2000. Temporal behavior of a solute cloud in a heterogeneous porous medium: 1. point-like injection. Water Resources Research 36 (12), 3591–3604. https://doi.org/10.1029/2000WR900162. https://www.agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2000WR900162.

Gardiner, C.W., 2004. Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences, 4, 3rd Ed Springer, Berlin.

Gingold, R.A., Monaghan, J.J., 1977. Smoothed particle hydrodynamics: theory and application to non-spherical stars. Mon Not R Astron Soc 181 (3), 375–389. https://doi.org/10.1093/mnras/181.3.375.

Hill, M.C., Tiedeman, C.R., 2007. Effective Groundwater Model Calibration: with Analysis of Data, Sensitivities, Predictions, and Uncertainty. John Wiley & Sons.

Kitanidis, P.K., 1994. The concept of the Dilution Index. Water Resources Research 30 (7), 2011–2026. https://doi.org/10.1029/94WR00762. https://www.agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/94WR00762.

Kitanidis, P.K., 1994. Particle-tracking equations for the solution of the advection-dispersion equation with variable coefficients. Water Resources Research 30 (11), 3225–3227. https://doi.org/10.1029/94WR01880. https://www.agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/94WR01880.

Konishi, S., Kitagawa, G., 2008. Information Criteria and Statistical Modeling. Springer Series in Statistics. Springer, New York, NY.

Kullback, S., 1968. Information Theory and Statistics. Dover Publications.

Kullback, S., Leibler, R.A., 1951. On information and sufficiency. Ann. Math. Statist. 22 (1), 79–86. https://doi.org/10.1214/aoms/1177729694.

Labolle, E.M., Fogg, G.E., Tompson, A.F.B., 1996. Random-walk simulation of transport in heterogeneous porous media: local mass-conservation problem and implementation methods. Water Resour. Res. 32 (3), 583–593.

Lichtner, P.C., Kelkar, S., Robinson, B., 2002. New form of dispersion tensor for axisymmetric porous media with implementation in particle tracking. Water Resources Research 38 (8). https://doi.org/10.1029/2000WR000100. 21–1–21–16.

Monaghan, J., 2012. Smoothed particle hydrodynamics and its diverse applications. Annu. Rev. Fluid Mech. 44 (1), 323–346. https://doi.org/10.1146/annurev-fluid-120710-101220.

Øksendal, B., 2003. Stochastic differential equations. Stochastic Differential Equations. Universitext. Springer, Berlin, Heidelberg.

Paster, A., Bolster, D., Benson, D.A., 2014. Connecting the dots: semi-analytical and random walk numerical solutions of the diffusion–reaction equation with stochastic initial conditions. J. Comput. Phys. 263, 91–112. https://doi.org/10.1016/j.jcp.2014.01.020.

Pedretti, D., Fernàndez-Garcia, D., 2013. An automatic locally-adaptive method to estimate heavily-tailed breakthrough curves from particle distributions. Adv Water Resour. 59, 52–65. https://doi.org/10.1016/j.advwatres.2013.05.006.

Rahbaralam, M., Fernàndez-Garcia, D., Sanchez-Vila, X., 2015. Do we really need a large number of particles to simulate bimolecular reactive transport with random walk methods? a kernel density estimation approach. J. Comput. Phys. 303, 95–104. https://doi.org/10.1016/j.jcp.2015.09.030.

Schmidt, M.J., Pankavich, S.D., Benson, D.A., 2018. On the accuracy of simulating mixing by random-walk particle-based mass-transfer algorithms. Adv. Water Resour. https://doi.org/10.1016/j.advwatres.2018.05.003.

Schumer, R., Benson, D.A., Meerschaert, M.M., Baeumer, B., 2003. Multiscaling fractional advection-dispersion equations and their solutions. Water Resour. Res. 39, 1022.

Shannon, C.E., 1948. A mathematical theory of communication. Bell Syst. Techn. J. 27, 379–423.

Silverman, B.W., 1986. Density Estimation for Statistics and Data Analysis. Chapman and Hall, London.

Sole-Mari, G., Fernàndez-Garcia, D., 2018. Lagrangian modeling of reactive transport in heterogeneous porous media with an automatic locally adaptive particle support volume. Water Resources Research 54 (10), 8309–8331. https://doi.org/10.1029/2018WR023033. https://www.agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2018WR023033.

Sole-Mari, G., Fernández-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 Resources Research 53 (11), 9019–9039. https://doi.org/10.1002/2017WR021064. https://www.agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/2017WR021064.

Sole-Mari, G., Schmidt, M.J., Pankavich, S.D., Benson, D.A., 2019. Numerical equivalence between SPH and probabilistic mass transfer methods for Lagrangian simulation of dispersion. Adv. Water Resour. 126, 108–115. https://doi.org/10.1016/j.advwatres.2019.02.009.

Sund, N.L., Porta, G.M., Bolster, D., 2017. Upscaling of dilution and mixing using a trajectory based spatial markov random walk model in a periodic flow domain. Adv Water Resour 103, 76–85. https://doi.org/10.1016/j.advwatres.2017.02.018.

Taylor, G., 1953. Dispersion of soluble matter in solvent flowing slowly through a tube. Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences 219 (1137), 186–203. https://doi.org/10.1098/rspa.1953.0139. http://www.rspa.royalsocietypublishing.org/content/219/1137/186.full.pdf.