Advances in Water Resources 151 (2021) 103893

Contents lists available at ScienceDirect

Advances in Water Resources

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

A Computational Information Criterion for Particle-Tracking with Sparse or Noisy Data ☆

Nhat Thanh V. Tran [a,*], David A. Benson [b], Michael J. Schmidt [c], Stephen D. Pankavich [d]

[a] Department of Mathematics, University of California, Irvine, Irvine, CA, 92697, USA
[b] Hydrologic Science and Engineering Program, Department of Geology and Geological Engineering, Colorado School of Mines, Golden, CO, 80401, USA
[c] Center for Computing Research, Sandia National Laboratories, Albuquerque, NM 87185, USA
[d] Department of Applied Mathematics and Statistics, Colorado School of Mines, Golden, CO, 80401, USA

ARTICLE INFO

Keywords:
Computational Information Criterion
Lagrangian Modeling
Particle Methods
Diffusion-reaction Equation
Non-Gaussian Error Distribution

ABSTRACT

Traditional probabilistic methods for the simulation of advection-diffusion equations (ADEs) often overlook the entropic contribution of the discretization, e.g., the number of particles, within associated numerical methods. Many times, the gain in accuracy of a highly discretized numerical model is outweighed by its associated computational costs or the noise within the data. We address the question of how many particles are needed in a simulation to best approximate and estimate parameters in one-dimensional advective-diffusive transport. To do so, we use the well-known Akaike Information Criterion (AIC) and a recently-developed correction called the Computational Information Criterion (COMIC) to guide the model selection process. Random-walk and mass-transfer particle tracking methods are employed to solve the model equations at various levels of discretization. Numerical results demonstrate that the COMIC provides an optimal number of particles that can describe a more efficient model in terms of parameter estimation and model prediction compared to the model selected by the AIC even when the data is sparse or noisy, the sampling volume is not uniform throughout the physical domain, or the error distribution of the data is non-IID Gaussian.

1. Introduction

Numerical methods of all sorts are used to approximate the solutions of various model equations in hydrology. The independent variables in these models are discretized, and model coefficients are populated so as to faithfully reproduce some set of measured dependent variables (i.e., data). Of course, both the model solution and the measured data will contain errors; therefore, a perfect match of model to data is not necessarily desired. In fact, a perfect match will most often reduce the ability of the model to predict new sets of data, because the model is overfit to a single realization of noise (see the excellent overview by Konishi and Kitagawa (2008)). The bias between a model that is overfit and an underlying "true" model was classically addressed by Akaike (Akaike, 1974) by considering approximate measures of the entropy of the probability distributions associated with the likelihood that data arises from a specific candidate model. In short, these measures introduce an entropic penalty to maximum likelihood estimates of goodness of fit when the number of *adjustable* parameters or coefficients increases. In a similar fashion, the computational entropy of a model is a bias against its fitness and needs to be accounted for when using a model in a predictive mode.

For a few examples, consider first a particle-tracking model of contaminant transport. It is often visually pleasing to use a large number of particles in order to obtain a smooth (i.e., low concentration variance) histogram for comparison with data. However, if the noise in the data far exceeds the noise in the histogram, then the large number of particles is superfluous from a model prediction perspective. Indeed, the model smoothness is not implied by the data and should not be included in a prediction. In addition, if the model is used repetitively in a parameter estimation procedure, then these extra calculations may also become tangibly burdensome. Another example is stochastic Monte Carlo modeling of groundwater flow using an Eulerian (e.g., finite-difference) simulator. Oftentimes the spatial and temporal discretizations are thought to be a free modeling choice, and random realizations of hydraulic conductivity are generated, yielding models with arbitrary degrees of freedom. Subsequently, the models are weighted in relation to the goodness of fit (e.g., Beven and Binley (1992); Poeter and Anderson (2005); Ye et al. (2004)), but one must ask whether simpler models, in terms of

☆ This work was supported by the National Science Foundation under awards DMS-1614586 and DMS-1911145 and the US Army Research Office under Contract/Grant number W911NF-18-1-0338.

* Corresponding author.
E-mail addresses: nhattt@uci.edu (N.T.V. Tran), dbenson@mines.edu (D.A. Benson), mjschm@sandia.gov (M.J. Schmidt), pankavic@mines.edu (S.D. Pankavich).

https://doi.org/10.1016/j.advwatres.2021.103893
Received 10 September 2020; Received in revised form 18 January 2021; Accepted 16 March 2021
Available online 18 March 2021

both discretization and parameterization, should be elevated because of their computational simplicity.

Recently, Benson et al. (2020) showed that a computational entropy penalty, called the Computational Information Criterion or COMIC, can be easily used to address this issue for simple systems (i.e., those with constant discretization or particle numbers), but it has yet to be shown that those results can be generalized to more realistic modeling scenarios. Here, we show that measures of computational (entropy) penalty may be extended to less ideal cases. In the next section, we will review both the COMIC and the advection-dispersion (mixed hyperbolic and parabolic) partial differential equation (PDE) model of interest, and then discuss associated Lagrangian numerical methods. Sections 3 and 4 are devoted to understanding the effects of sparse and non-uniformly spaced data sets, respectively, on the information criterion and subsequent parameter estimation. Next, we reformulate the COMIC in Section 5 to account for a non-uniform discretization volume and demonstrate its equivalence (in terms of particle number selection) with the COMIC derived from a uniform sampling volume. In Section 6, we derive a more general version of the COMIC for non-Gaussian errors and non-uniform error variance and test it on noisy data sets. Conclusions are discussed and summarized in the final section.

2. Model and Methods

To enable a complete analysis, we use an equation that does not require a numerical solution (i.e., an analytical solution is readily available). This restriction is not required. The simplest case of the one-dimensional, constant-coefficient advection-diffusion equation (ADE) is given by

$$\frac{\partial c}{\partial t} = -v\frac{\partial c}{\partial x} + D\frac{\partial^2 c}{\partial x^2}, \tag{1}$$

where $c(x,t)$ is the solution to the PDE, v is the velocity, and D is the diffusion coefficient. As we are interested in particle methods, we will restrict our attention to the initial condition $c(x,0) = \delta(x_0)$ so that $c(x,t)$ is a probability density function (PDF). In practice, the solution to the PDE is approximated by the numerical method $c_n(x,t)$, which is a function of a discretization parameter n, representing the number of particles in a particle method or nodes in a finite-difference approximation. The choice of n is arbitrary and we seek a systematic way to choose this modeling parameter.

The newly developed COMIC may be used to select models amongst these different levels of discretization (Benson et al., 2020). This criterion is a extension of Akaike's "an information criterion" (AIC) in which there is a penalization of the usage of more information (in the form of adjustable parameters) to describe the model. For completeness, we recall that the AIC is defined by

$$\mathrm{AIC} = -2\ln(\mathcal{L}(\hat{\theta})) + 2p \tag{2}$$

where $\mathcal{L}(\hat{\theta})$ is the likelihood function evaluated at the maximum likelihood estimate for the unknown parameters θ, and p is the number of parameters. For a computational model, the number of nodes or particles contributes to the entropy and must be accounted for when comparing model predictive fitness. In short, the (Kullback and Leibler, 1951; Kullback, 1968) relative entropy for a discretized model contributes an extra term to the AIC, and the COMIC takes the form

$$\mathrm{COMIC} = \mathrm{AIC} - \int c(x,t)\ln(\Delta V(x))\,dx \tag{3}$$

where $\Delta V(x)$ is the sampling volume, here given by the spacing between the particles or nodes. In the case that $\Delta V(x) = |\Omega|/n$ is constant where $|\Omega|$ represents the length of the spatial domain, then the COMIC will merely reduce to

$$\mathrm{COMIC} = \mathrm{AIC} + \ln(n) \tag{4}$$

up to a constant. It is this quantity that should be minimized in order to identify the best available predictive model Konishi and Kitagawa (2008); Benson et al. (2020).

For this study, we use two Lagrangian methods: random-walk particle-tracking (RWPT) and mass-transfer particle-tracking (MTPT). Both methods are described in detail elsewhere (Labolle et al., 1996; Salamon et al., 2006; Benson and Bolster, 2016; Schmidt et al., 2018), so we only summarize here. In its simplest form, the RWPT method places a number of particles n at the release position x_0, each with constant mass $1/n$. At every time step of duration Δt, each particle moves with mean $v\Delta t$ and random deviation $\sqrt{2D\Delta t}\xi$, where $\xi \sim \mathcal{N}(0,1)$ is a standard Normal random variable. At any desired time, bins of size Δx_i, centered at points x_i are constructed and the particle count n_i in each bin is converted to concentration by $c_n(x_i,t) = n_i/(n\Delta x_i)$.

Contrastingly, the MTPT method typically allows a portion of the dispersion to be performed by random walks as above, and the remaining portion is performed by mass transfer between particles (Benson et al., 2019, 2020; Sole-Mari et al., 2020; Herrera et al., 2009; Engdahl et al., 2017; Schmidt et al., 2019). The mass transfer between any and all particles is governed by the equation

$$m_i(t+\Delta t) = m_i(t) - \sum_{i=1}^{n}\big(m_i(t) - m_j(t)\big)W_{ij}, \tag{5}$$

where for each particle pair denoted i, j,

$$W_{ij} = \frac{(1/\sqrt{4\pi D\Delta t})\exp(-s_{ij}^2/4D\Delta t)}{\rho_{ij}} \tag{6}$$

is the normalized kernel that determines the weight of mass transfer between particles i and j (Benson and Bolster, 2016), ρ_{ij} is a normalizing constant that ensures conservation of mass and is typically taken to be the particle density (Sole-Mari et al., 2019; Schmidt et al., 2020), and s_{ij} is the distance between particles i and j. We note that the choice of the Gaussian kernel's bandwidth is a free parameter. In this case we choose it to be $\sqrt{2D\Delta t}$, making W a normalized version of the fundamental solution to the diffusion equation, and this is equivalent to choosing $\beta = 1$ within the convention of Sole-Mari et al. (2019). Because the particles continually change mass, they are typically pre-distributed throughout the domain with zero mass. One particle at the release point is given unit mass. For this study, when we employ an MTPT method all dispersion is modeled by mass transfer, i.e., the particles move only by their mean velocity with no random motion (in contrast to the RWPT simulations that only simulate dispersion via random-walk, with no mass transfer).

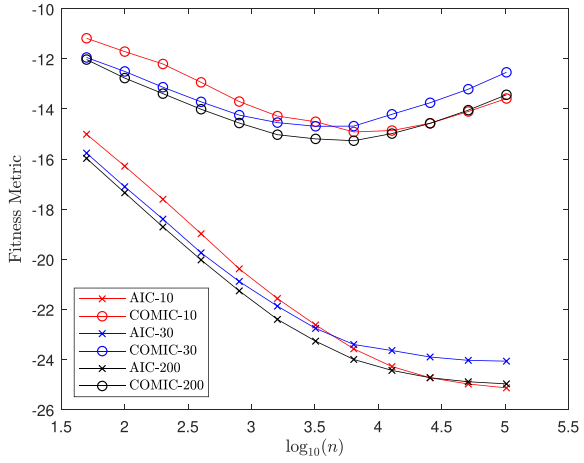
In subsequent sections we perform a variety of numerical simulations using the COMIC to select the optimal number of particles in a model and perform parameter estimation. All numerical simulations were conducted in MATLAB using a desktop computer with a 3.5 GHz Intel Core i7 processor and 16 GB of RAM. The code used to generate the results in this section is available at https://github.com/nhatthanhtran/entropy2020 (Tran, 2020b).

3. Sparsity of Data

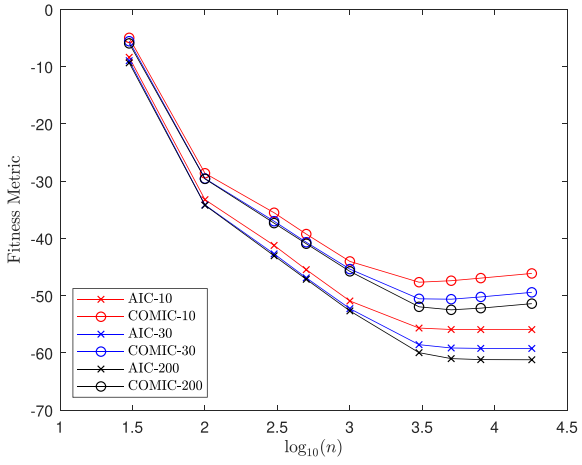
One problem that may arise when collecting data is a limited number of accessible locations. The sparsity of data could change the number of particles needed to fit a given data set (and subsequently predict others). We presume that this will also affect the optimal number of particles predicted by the COMIC. Thus, instead of assuming a large collection of data (e.g., ≥ 30 points), we will also consider relatively small data sets (≤ 10 points) within the domain, to assess the degree to which parameter estimates are affected by less data and to what extent the optimal number of particles changes in order to achieve more precise estimates. Because we are relying on smaller collections of data, modification to the AIC is necessary (Cavanaugh, 1997). This corrected fitness metric is often denoted as the AICc and defined by Hurvich and Tsai (1989)

$$\mathrm{AICc} = \mathrm{AIC} + \frac{2p^2+2p}{k-p-1}, \tag{7}$$

where k is the number of data points and p is the number of parameters in the model. This leads to the obvious modification of the COMIC with



(a) RWPT



(b) MTPT

Fig. 1. Fitness metrics for uniformly spaced data using $k = 10$, 30, and 200 data points: RWPT (top) and MTPT (bottom).

$\Delta V = |\Omega|/n$ as

$$\text{COMICc} = \text{AICc} + \ln(n). \tag{8}$$

Such a correction is relevant when we are attempting to compare different models with various numbers of data points and parameters. Simulations were conducted by setting the final time $T = 1$ with $D = 1$, $v = 0$ (i.e., no adjustable model parameters) and selecting $k = 10$, 30, or 200 "sample" data points $\hat{c}_1, ..., \hat{c}_k$ from exact values of the analytic solution of the ADE (1) on the interval $\Omega = [-5, 5]$ with uniform spacing. We calculate the AICc and COMICc with different particle numbers for both the random walk and mass transfer methods. For the former, the optimal number of particles as given by the COMICc appears to increase for the sparse ($k = 10$) data case to more than $n = 10^{4.3} \approx 20{,}000$ particles (Fig. 1a). On the other hand, the MTPT simulations appear relatively stable in terms of their COMIC fitness with an optimal number of particles around $n = 3{,}000$ (Fig. 1b).

We now use these results to fix the COMIC-optimal number of particles in simulations (20,000 for RWPT and 3,000 for MTPT) that now seek to estimate the parameters within the ADE. We choose data points from the analytic solution using $v = D = 1$ and use MATLAB's built-in `fminsearch` function to minimize the AIC with initial guesses for both coefficients of 0.5. Because the particles do not move in the MTPT algorithm, the solutions are the same if run multiple times (i.e., non-stochastic results). For 30 data points, the estimated values of D and v in the MTPT are within 10^{-6} of their true value 1. On the other hand, the RWPT method gives a different result for each realization because of the random walks, so we show a box/whisker plot for those results in Fig. 2.

For even fewer data points (say, $k = 6$), the random walk method requires significantly more particles to achieve a sufficient balance between goodness of fit and computational complexity. The qualitative behavior and shape of the AIC and COMIC for differing particle numbers are similar to Fig. 1, and hence are not shown. The mass transfer method displays similar behavior to other simulations - the COMIC optimal number of particles occurs at $n_{\text{COMIC}} \approx 3{,}000$ and the estimated values of D and v are both within 10^{-6} of their true value of 1. Hence, the MTPT is significantly more robust with respect to the size of the data. Finally, we briefly mention that simulations were performed for every chosen particle number in which parameter values were first estimated and then used to compute the fitness metric. Again, the convex shape of the AIC and COMIC for the RWPT method were nearly identical to Fig. 1, though the new optimal number of particles implied by the COMIC for the MTPT is $n_{\text{COMIC}} \approx 500$. One reason for this decrease in optimal particle number is that the MTPT estimates are very close to the exact solution, so that any noise in the parameter estimation dominates the overall error. Fig. 3 displays the fitness metrics from MTPT simulations with 30 data points. Using this new optimal particle number, we estimated the values of D and v to be within 10^{-5} of their true values, and these are similar to the parameter estimates obtained using 3,000 particles.

4. Non-uniformly Spaced Data

Another issue that may arise under less ideal data collection circumstances is that observations from a field site may not be constrained to a certain grid, meaning that the data may not exist at uniformly-spaced gridpoints. Therefore, we examine the effects of such spatial heterogeneity on the COMIC and elucidate how this will affect parameter estimation. To generate non-uniform data, we randomly select data points within the domain of interest. More specifically, we perform simulations for $k = 10$ and $k = 30$ data points, with $D = 1$, $v = 1$, $x_0 = 0$ and $T = 1$. Additionally, the numerical spacing $\Delta V = 10/n$ remains constant. Calculations of the COMIC for random data display similar results to the case of equally-spaced data, so we will use the optimal number of particles from the previous section, as well. For the RWPT method, we perform simulations with 5,000 particles for 30 data points and 20,000 particles for 10 data points. For the MTPT method both simulations use 3,000 particles. The initial guess for both parameters is 0.5 within all simulations. Due to the added randomness in the spacing of the data, we might anticipate that the parameter estimates would vary more than in the case of uniformly-spaced data; however, the results for these random walk parameter-estimation simulations are similar in both the magnitude and the variability of the estimates of both D and v (Fig. 2). For the MTPT method, a simulation with 30 data points yields estimates of D and v within 10^{-5} of their true values. Similarly, a simulation with 10 data points provides estimates of D and v within 10^{-4} of their true values. From these simulations, we conclude that the COMIC provides a useful and informative guide for the choice of particle number and parameter estimation even when the data is not uniformly spaced. Next, we will consider alterations to the COMIC in order to address non-constant sampling volumes.

5. Non-Uniform Numerical Discretization Volume

In the original calculation of the COMIC, the spatial volume $\Delta V(x)$ along which the solution is calculated is assumed to be a constant, and this results in Eq. (4) representing the computational fitness metric. However, in performing certain simulations, this assumption may not hold as either the binning of RWPT particles — or positions of MTPT particles — may not be evenly distributed throughout the physical domain. In contrast to the previous section, this issue only arises due to

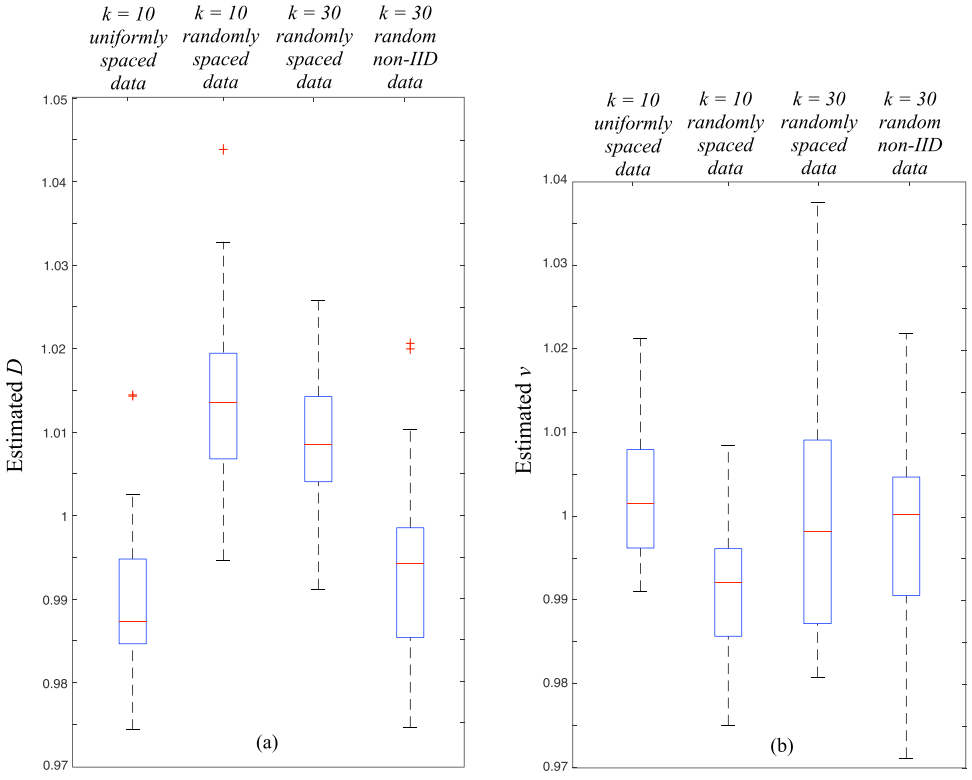


Fig. 2. Estimates of D and v using RWPT with different data spacings schemes, numbers, and distributional qualities.

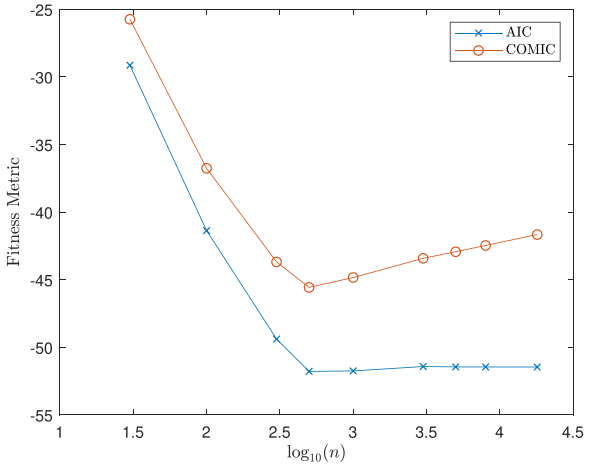


Fig. 3. Mass transfer fitness metrics determined by estimating the parameters D and v for various particle numbers.

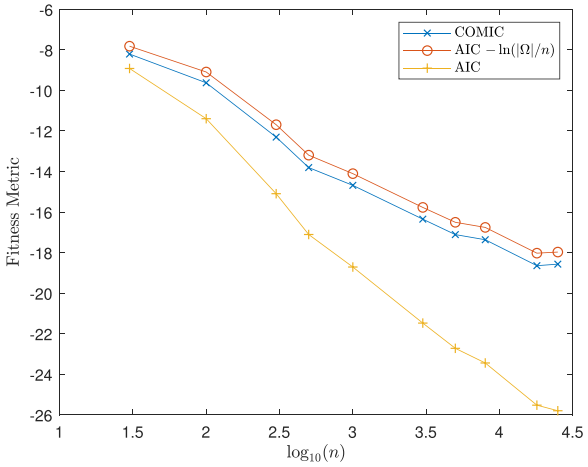


Fig. 4. Fitness metrics versus particle number in the MTPT method with spatially-varying sampling volume $\Delta V(x)$

the computational method rather than the collected data. In such a case, $\Delta V(x)$ will vary with x, and the calculation of the COMIC must instead use Eq. (3). One immediate drawback from this formulation is that the value of the true concentration $c(x,t)$ at any point is unknown. Still, the numerical approximation $c_n(x,t)$ can be used to estimate this function within the COMIC, which becomes

$$\text{COMIC} = \text{AIC} - \int c_n(x,t)\ln(\Delta V(x))\,dx. \tag{9}$$

In this direction, we perform MTPT simulations in which the particles are initially placed randomly within the domain, and the random spacing between particles will determine the sampling volume $\Delta V(x)$ at each of the n particle locations.

This is performed with varying particle numbers, and the resulting COMIC of Eq. (9) is computed for each simulation, once again using $k = 30$ data points generated from the analytic solution of the ADE. To account for the randomness in the initial particle spacing, we compute the ensemble average of 30 realizations (Fig. 4). From these simulations, two key observations become apparent:

1. The optimal number of MTPT particles given by the COMIC is 18,000, which is much larger than the 3,000 particles predicted by simulations with constant ΔV - see Fig. 1b. This can be explained by the the randomness of the particle spacing. For instance, Schmidt et al. (2018) show that the MTPT algorithm incurs increased error for randomly spaced, immobile particles because of mass-transfer "gaps" in areas of sparsely-distributed particles. The fitness of the method drops sharply compared to the uniformly-spaced scenario (Fig. 1), especially when the number of particles becomes small. This is clearly demonstrated by the AIC in Fig. 4, as

this curve possesses a large negative slope between 100 and 18,000 particles and only plateaus after 18,000 particles. At this point the particles are distributed tightly enough in the domain to maintain the diffusion process regardless of the variability in their spacing. It should be noted that partitioning a portion of the diffusion process to random walks alleviates the problem of persistent mass-transfer "gaps" (Benson et al., 2020).

2. The exact COMIC calculation, which is performed by approximating the integral in (9), and the uniformly-spaced COMIC, given by Eq. (4), have the same convex shape. For stationary particles, integrating a nonuniform $\Delta V(x)$ simply adds a constant value to computing the COMIC for constant ΔV. Hence, the curve is simply shifted vertically, without further influencing its shape or the position at which it attains its minimum value, and this property persists throughout multiple simulations. Both fitness metrics provide the identical optimal number of particles of 18,000. Thus, any calculation of the COMIC could use the Eq. (4) regardless of the sampling volume.

In addition to changes in the COMIC that arise from data collection or choice of numerical method, we will also consider differing statistical assumptions that affect the information criterion, and this is performed in the next section.

6. Non-IID/Non-Gaussian Error Processes

The original study of the COMIC (Benson et al., 2020) assumed that the residual between the (unknown) true concentration $c(x,t)$ and the data $\widehat{c}_i$ for $i = 1, ..., k$ at each location was an independent and identically distributed (IID) Normal random variable, so that

$$c(x_i, T) = \widehat{c}_i + \epsilon_i \tag{10}$$

for every $i = 1, ..., k$ where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ and the variance σ^2 must be estimated from the data. This led to a log-likelihood function (hence AIC) that used the average sum of squared errors for the variance estimate within the fitness metric. The assumption of IID Gaussian errors is not generally valid, and the properties of the error distribution are unknown in most cases. Here, we exploit the fact that the computational solution arises from a particle method in order to approximate the residuals. In this scenario, *Chakraborty et al* (Chakraborty et al., 2009) showed that (a) the concentration approximation generated by any particle method is proportional to a binomial random variable that is only asymptotically Normal (as $n \to \infty$, $\Delta x \to 0$, and $\sqrt{n}\Delta x \to \infty$ where Δx is the bin size of the method), and (b) individual concentration errors could be treated as independent with variance proportional to the concentration, i.e. $\sigma_i^2 = \frac{m_{\text{total}}}{n\Delta x} c_n(x_i, t)$, where m_{total} is the total mass. Indeed, this can be numerically verified using the RWPT and MTPT methods (Tran, 2020a). In Chakraborty et al. (2009), an alternative information criterion was also proposed for selecting a "best" model over all parameter choices with the desirable property that the chosen parameter estimate $\widehat{\theta}$ serves as a consistent estimator for the true parameter values θ. More specifically, this criterion arises from an optimal fitting procedure that serves to minimize the weighted mean square error function

$$\mathcal{E}(\theta) = \frac{1}{k} \sum_{i=1}^{k} w_i |\widehat{c}_i - c_n(x_i, T; \theta)|^2 \tag{11}$$

where θ is the vector of unknown model parameters, the minimization weights are

$$w_i = \frac{1}{m_{\text{total}} \widehat{c}_i}, \tag{12}$$

and the estimator $\widehat{\theta}$ is given by

$$\widehat{\theta} = \text{argmin}_{\theta \in \mathbb{R}^p} \ \mathcal{E}(\theta). \tag{13}$$

For the ADE problem described in previous sections, we merely have $\theta = [v, D]^T$. We note that in the case that errors are normally-distributed, minimizing (11) is equivalent to maximizing the log-likelihood function for a multivariate Gaussian distribution with $\widehat{\sigma}_i^2 = m_{\text{total}} \widehat{c}_i$ for $i = 1, \ldots, N$ (see Appendix).

As for the AIC, this information criterion does not account for the additional information incurred by taking large numbers of particles, and hence we augment it to create a new computational information criterion. Therefore, in this case we the define the COMIC by

$$\text{COMIC} = 2 \ln \left(\frac{1}{k} \sum_{i=1}^{k} \frac{1}{m_{\text{total}} \widehat{c}_i} |\widehat{c}_i - c_n(x_i, T)|^2 \right) + 2p - \int c_n(x, T) \ln(\Delta V(x)) \, dx. \tag{14}$$

where $\widehat{c}_i$ is the concentration data at location x_i and final time $t = T$. From the results of the previous section, it is beneficial to take ΔV constant, which reduces this formulation to

$$\text{COMIC} = 2 \ln \left(\frac{1}{k} \sum_{i=1}^{k} \frac{1}{m_{\text{total}} \widehat{c}_i} (\widehat{c}_i - c_n(x_i, T))^2 \right) + 2p - \ln(\Delta V). \tag{15}$$

Using this criterion, we perform simulations of the random walk and mass transfer methods to compute the value of $2 \ln(\mathcal{E})$ and the COMIC. The formulation and implementation of these methods is analogous to that of the previous section with $k = 30$ randomly-spaced data points. From Fig. 5, the optimal number of particles for the MTPT method

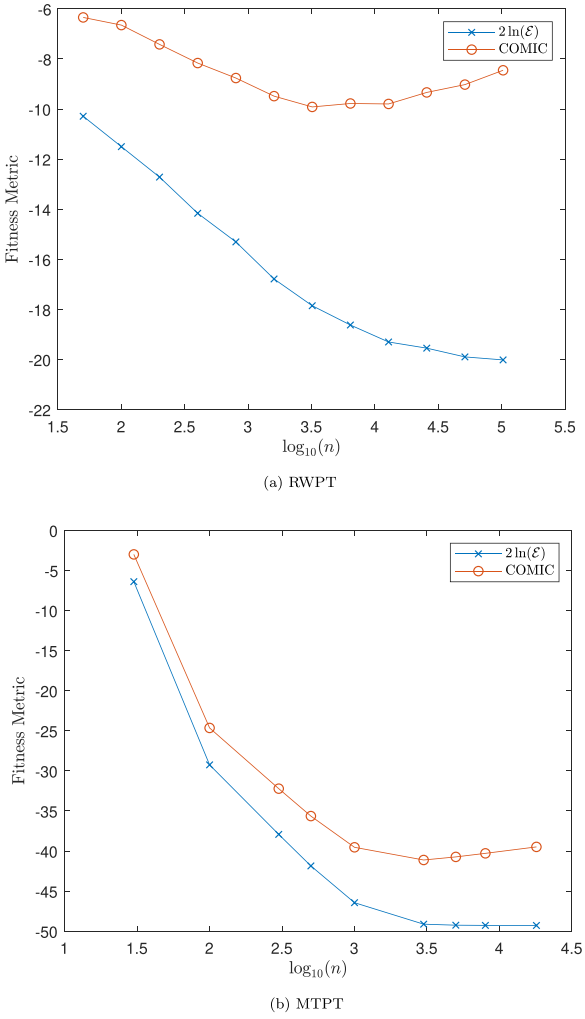
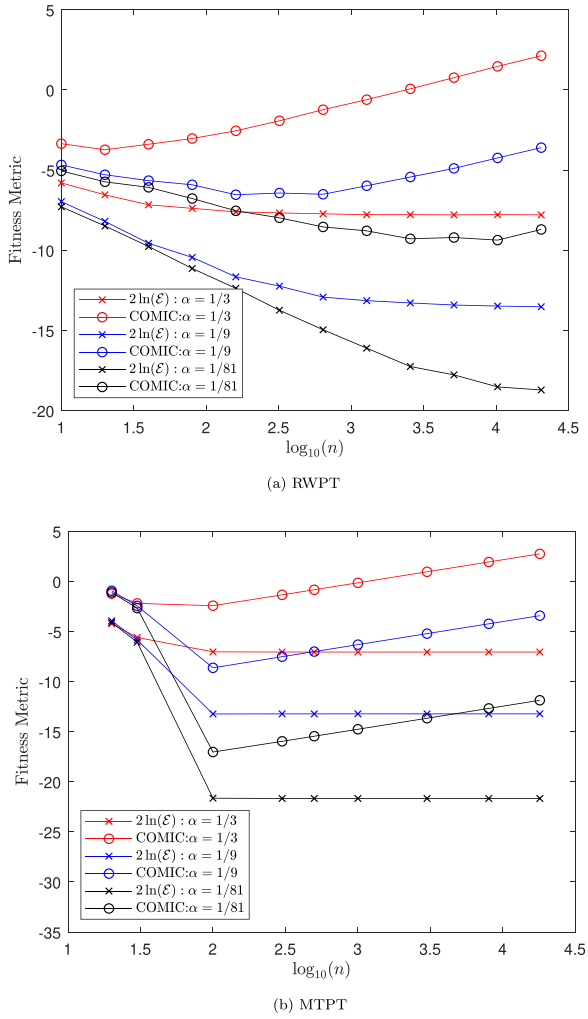


Fig. 5. Fitness metrics for non-Gaussian error processes - RWPT (top) and MTPT (bottom). Here, the COMIC is computed from Eq. (15), and these curves can be directly compared with those of Fig. 1 with $k = 30$.



(a) RWPT

(b) MTPT

Fig. 6. Fitness metrics for Non-Gaussian Error Processes with noisy data - RWPT (top) and MTPT (bottom).

is 3,000, while for the RWPT method, it ranges between 3,200 and 6,400 particles. We take the midpoint of these and assume the optimal number of particles is about 5,000. Then, this predicted value is essentially the same as that stemming from IID Gaussian error simulations (Benson et al., 2020). Using the optimal number of particles to perform parameter estimation provides MTPT estimates within 10^{-4} of the true values of $D = 1$ and $v = 1$. Because the data is random, simulation outcomes may vary, but multiple runs with different data display similar qualitative behavior. Estimated values of D and v using the RWPT method are also similar - see Fig. 2. The maximal absolute error in the estimates of D and v are about 2.5% and 3%, respectively, which are comparable to the parameter estimation performed for IID Gaussian error simulations. Therefore, the COMIC demonstrates consistency among different error assumptions and estimators. In other words, optimal model discretization does not strongly depend on the weights in the sum of squared errors.

Lastly, we perform similar simulations with noisy data, i.e. data that differs from the analytic solution due to measurement error modeled by independent random concentration noise at each spatial location. Fig. 6 shows the fitness metrics arising from simulated data that is normally distributed with mean equal to the exact concentration from the analytic solution of eq. (1) and standard deviation given by $\alpha = 1/3,\ 1/9,$ or $1/81$ of the exact concentration so that $\mathrm{Var}(\hat{c}_i) = \alpha^2 c(x_i, T)^2$. We note that any negative sample points are discarded, in order to guarantee non-negativity of the concentration data set. Of course, the exact value of α can be tuned to adjust for the noise in the data.

There are several noteworthy features shown in a single realization of the simulations (different simulations show similar results):

1. For the RWPT simulations, the optimal number of particles is strongly determined by the standard deviation of the independent noise held within the data. In short, the noisier the data, the more noise can be contained within the model solution (i.e., fewer particles should be used).
2. For the MTPT solution, the optimal value of the COMIC occurs at $n = 100$ particles, which is far fewer than the optimal number of particles when $\alpha = 0$. In contrast to the RWPT solutions, the number of particles is not strongly affected by the magnitude of noise in the data. This can be explained by the formulation of the COMIC, in which the dominant term in the calculation is the AIC, until n becomes sufficiently large. The numerical simulation is converging to the exact solution of the PDE, i.e. a normal distribution, and as the number of particles nears 100, the approximate solution closely agrees with the exact solution. However, assuming substantial noise in the data, the exact solution is far from the expected normal distribution, which means that the AIC will remain large even as the computational approximation converges. This can be seen by the graph of $2\ln(\mathcal{E})$ within Fig. 6, which plateaus for simulations with more than 100 particles.
3. If a given data set is particularly noisy or fails to faithfully represent the underlying solution of the PDE, one would actually do best to use fewer particles in a simulation. When significant noise exists in the data, a low-noise, high-n solution is not expected to be a better predictor because it is apt to be over-fit to peculiarities in a single noisy data set.

7. Conclusions

We have investigated the use of the COMIC to select parsimonious and robust computational models for simple advective-diffusive transport in a variety of realistic, data-driven scenarios, including noisy, sparse, and spatially-heterogeneous data sets, non-uniform sampling volumes, and non-IID errors. In the case of non-uniform sampling volumes, we have shown that the calculation of the COMIC can be further simplified using the average spacing between particles throughout the domain, and this does not influence the particle number selected by the algorithm. The results of our simulations demonstrate that under any of these conditions, the COMIC is a flexible criterion that allows the user to select an appropriate number of particles in a simulation so as to guarantee the use of minimal computational information to construct a descriptive model. In particular, we find that the use of large particles numbers is often superfluous in these simulations and needlessly increases the complexity of a model. This highlights the importance of selecting a suitably efficient computational model with minimal information content to best make predictions based on a single given data set. In particular, we find the following general rules regarding the optimal discretization for RWPT and MTPT simulations and/or parameter estimation of advection and diffusion in 1-D:

1. For “perfect” (no noise) data that is uniformly spaced, the number of data does not strongly effect the optimal number of particles. In general, RWPT (with binning) required $\approx 20,000$ particles, MTPT required $\approx 3,000$.
2. For either solution technique, the parameters can be estimated quite accurately. The optimal number of particles in the MTPT simulations drops to ≈ 500 when estimating the velocity and diffusion parameters.
3. Non-uniformly spaced data does not change the ability to estimate parameters.

4. Non-uniformly spaced particles in the MTPT technique significantly increases the optimal number of particles because of increased error in the simulations.
5. Non-uniform numerical discretization adds a constant to the COMIC but does not change the shape of the COMIC versus discretization curve, hence the minimum is not shifted. This means that the computational entropy penalty can be estimated by $\ln(n)$.
6. The classical MLE fitness metric of average squared error, which comes from an assumption of IID Normal errors, is a reasonably good estimator for errors in which concentration variance is proportional to concentration. However, as the noise in data increases, the optimal number of RWPT particles decreases sharply. In short, the error of the solution is directly tied to particle numbers, and hyper-accurate solutions are not representative of the noisy data (and may overfit the errors). The accuracy of the MTPT method is not tied to particle numbers in the same way (once a minimum particle spacing is achieved), and so the optimal particle number remains ≈ 100 for all noise levels.

Declaration of Competing Interest

None.

CRediT authorship contribution statement

Nhat Thanh V. Tran: Software, Investigation, Writing - original draft, Visualization. **David A. Benson:** Conceptualization, Supervision, Writing - review & editing. **Michael J. Schmidt:** Software, Writing - review & editing. **Stephen D. Pankavich:** Conceptualization, Supervision, Writing - review & editing.

Acknowledgments

Sandia National Laboratories is a multi-mission laboratory managed and operated by the National Technology and Engineering Solutions of Sandia, L.L.C., a wholly owned subsidiary of Honeywell International, Inc., for the DOE's National Nuclear Security Administration under contract DE-NA0003525. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

Appendix A. MLEs for non-IID Multivariate Gaussian Processes

We first recall the process of obtaining maximum-likelihood parameter estimates $\widehat{\theta}$ under the assumption that the errors between model and observations are independent, zero-mean Gaussians. In this case the likelihood function is given by

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

where k is the number of observation points, $\Sigma(\theta)$ is a diagonal (due to independence) covariance matrix of errors, and y is a vector of residuals satisfying $y_i = \widehat{c}_i - c(x_i, T)$ for $i = 1, ..., k$. Now, if the errors are further assumed to be identically-distributed, then Σ is a constant multiple of the identity and depends only upon a single variance parameter. In this case, $\Sigma = \sigma^2 \mathbb{I}$ where σ^2 is the assumed variance of the error at each spatial data point x_i for $i = 1, ..., k$. The log-likelihood function then becomes

$$\ln(L) = -\frac{k}{2}\ln(2\pi) - \frac{k}{2}\ln\sigma^2 - \frac{k}{2\sigma^2}\frac{\mathrm{SSE}}{k} \tag{A.2}$$

where $\mathrm{SSE} = \mathrm{y} \cdot \mathrm{y} = |\mathrm{y}|^2$ represents the sum of squared errors. Maximizing this function using standard tools provides an estimator of the observation variance, namely $\widehat{\sigma}^2 = \mathrm{SSE}/\mathrm{k}$. Removing any constant terms that do not change from one model to another, the corresponding log-likelihood evaluated at the MLE is

$$\ln(\widehat{L}) = -\ln\left(\frac{\mathrm{SSE}}{k}\right). \tag{A.3}$$

Alternatively, if one does not assume that the errors are identically-distributed, then Σ is not a constant multiple of the identity, but merely diagonal, so that

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \cdots & \sigma_k^2 \end{bmatrix}.$$

Hence, the error distribution at each spatial data point x_i possesses a different variance σ_i^2. In this case, the MLE can be generated using similar methods as above, and the resulting values of $\widehat{\sigma}_i^2$ are given by the multivariate sum of squared errors

$$\widehat{\sigma}_i^2 = \frac{1}{N}\sum_{\ell=1}^{N} |(y_\ell)_i|^2$$

where $y_1, ..., y_N$ is a collection of residual vectors with each y_ℓ representing a single sample from all spatial points, namely

$$(y_\ell)_i = \widehat{c}_i - c(x_i, T)$$

for the ℓth sample with $i = 1, ..., k$. Of course, if only one sample is collected at each spatial gridpoint so that $N = 1$ and y represents the single data vector, then the variance estimator reduces to $\widehat{\sigma}_i^2 = y_i^2$. Unfortunately, this does not provide a sharp estimate, as the standard deviation of the error distribution is merely equal to the data value at every spatial point. Therefore, to provide a realistic estimate of these values, we would require multiple concentration measurements at each spatial data point and at a fixed time T. In the absence of this data or a suitable approximation, the variance of the error distribution cannot be accurately determined. Because of this difficulty, *Chakraborty et al* (Chakraborty et al., 2009) proposed an alternative minimization criterion, which uses a weighted mean square error as described within Section 6, and this is derived only under the assumption that the error is approximately Gaussian. This criterion enables us to generate a consistent parameter estimate when only a single concentration sample is available at each spatial location.

References

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Autom. Control 19 (6), 716–723.
Benson, D.A., Bolster, D., 2016. Arbitrarily complex chemical reactions on particles. Water Resources Research 52 (11), 9190–9200. https://doi.org/10.1002/2016WR019368.
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. Advances in Water Resources 123, 40–53. https://doi.org/10.1016/j.advwatres.2018.11.001.
Benson, D., Pankavich, S., Schmidt, M., Sole-Mari, G., 2020. Entropy: (1) the former trouble with particle-tracking simulation and (2) a measure of computational information penalty. Advances in Water Resources 137, 103509. https://doi.org/10.1016/j.advwatres.2020.103509.
Beven, K., Binley, A., 1992. The future of distributed models: Model calibration and uncertainty prediction. Hydrological Processes 6 (3), 279–298. https://doi.org/10.1002/hyp.3360060305.
Cavanaugh, J.E., 1997. Unifying the derivations for the akaike and corrected akaike information criteria. Statistics & Probability Letters 33 (2), 201–208. https://doi.org/10.1016/S0167-7152(96)00128-9.
Chakraborty, P., Meerschaert, M., Lim, C., 2009. Parameter estimation for fractional transport: A particle-tracking approach. Water Resources Research 45 (10), W10415.
Engdahl, N.B., Benson, D.A., Bolster, D., 2017. Lagrangian simulation of mixing and reactions in complex geochemical systems. Water Resources Research 53 (4), 3513–3522. https://doi.org/10.1002/2017WR020362.
Herrera, P.A., Massabó, M., Beckie, R.D., 2009. A meshless method to simulate solute transport in heterogeneous porous media. Advances in Water Resources 32 (3), 413–429. https://doi.org/10.1016/j.advwatres.2008.12.005.
Hurvich, C.M., Tsai, C.-L., 1989. Regression and time series model selection in small samples. Biometrika 76 (2), 297–307. https://doi.org/10.1093/biomet/76.2.297.
Konishi, S., Kitagawa, G., 2008. Information Criteria and Statistical Modeling. 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.

Poeter, E., Anderson, D., 2005. Multimodel ranking and inference in ground water modeling. Groundwater 43 (4), 597–605. https://doi.org/10.1111/j.1745-6584.2005.0061.x.

Salamon, P., Fernàndez-Garcia, D., Gómez-Hernández, J.J., 2006. A review and numerical assessment of the random walk particle tracking method. Journal of Contaminant Hydrology 87 (3–4), 277–305. https://doi.org/10.1016/j.jconhyd.2006.05.005.

Schmidt, M.J., Engdahl, N.B., Pankavich, S.D., Bolster, D., 2020. A mass-transfer particle-tracking method for simulating transport with discontinuous diffusion coefficients. Advances in Water Resources 140, 103577. https://doi.org/10.1016/j.advwatres.2020.103577.

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

Schmidt, M.J., Pankavich, S.D., Navarre-Sitchler, A., Benson, D.A., 2019. A Lagrangian method for reactive transport with solid/aqueous chemical phase interaction. Journal of Computational Physics: X 100021. https://doi.org/10.1016/j.jcpx.2019.100021.

Sole-Mari, G., Fernández-Garcia, D., Sanchez-Vila, X., Bolster, D., 2020. Lagrangian modeling of mixing-limited reactive transport in porous media: Multirate interaction by exchange with the mean. Water Resources Research 56 (8). https://doi.org/10.1029/2019WR026993. e2019WR026993. E2019WR02699310.1029/2019WR026993

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. Advances in Water Resources 126, 108–115. https://doi.org/10.1016/j.advwatres.2019.02.009.

Tran, N., 2020. Entropic Criteria for Computational Models of Advection-Diffusion Equations. Colorado School of Mines, 1500 Illinois Ave.; Golden, C0 80401.

Tran, N. T., 2020b. Entropy2020. 10.5281/zenodo.4018153

Ye, M., Neuman, S.P., Meyer, P.D., 2004. Maximum likelihood bayesian averaging of spatial variability models in unsaturated fractured tuff. Water Resources Research 40 (5). https://doi.org/10.1029/2003WR002557.