

Resource Allocation for Lagrangian Tracking

BENJAMIN T. JONES, ANDREW SOLOW, AND RUBAO JI

Woods Hole Oceanographic Institution, Woods Hole, Massachusetts

(Manuscript received 18 June 2015, in final form 4 April 2016)

ABSTRACT

Accurate estimation of the transport probabilities among regions in the ocean provides valuable information for understanding plankton transport, the spread of pollutants, and the movement of water masses. Individual-based particle-tracking models simulate a large ensemble of Lagrangian particles and are a common method to estimate these transport probabilities. Simulating a large ensemble of Lagrangian particles is computationally expensive, and appropriately allocating resources can reduce the cost of this method. Two universal questions in the design of studies that use Lagrangian particle tracking are how many particles to release and how to distribute particle releases. A method is presented for tailoring the number and the release location of particles to most effectively achieve the objectives of a study. The method detailed here is a sequential analysis procedure that seeks to minimize the number of particles that are required to satisfy a predefined metric of result quality. The study assesses the result quality as the precision of the estimates for the elements of a transport matrix and also describes how the method may be extended for use with other metrics. Applying this methodology to both a theoretical system and a particle transport model of the Gulf of Maine results in more precise estimates of the transport probabilities with fewer particles than from uniformly or randomly distributing particle releases. The application of this method can help reduce the cost of and increase the robustness of results from studies that use Lagrangian particles.


1. Introduction

Particle transport has implications throughout oceanography. Phytoplankton and zooplankton that form the base of the marine food web cannot overcome ocean currents and are transported as small particles (Miller and Wheeler 2012). Higher trophic levels, including many invertebrates and fish, are transported as planktonic larvae (Pineda et al. 2007). Oil and other chemical pollutants often assemble into droplets that are transported as small particles (Lynch et al. 2015). Understanding the movement of these particles is critical to understanding marine ecosystems.

Our knowledge of particle transport may be represented as a connectivity matrix whose elements give the

probability of transport among discrete geographic regions (Cowen and Sponaugle 2009). One commonly used method to estimate connectivity matrices is to simulate many Lagrangian particles with an individual-based model (IBM) and to compute the ensemble average of the particle trajectories. IBMs simulate particle transport through Eulerian velocity fields that are produced by ocean circulation models. Because some computational overhead is required to produce the Eulerian velocity fields, IBMs operate most efficiently when simulating large batches of particles. Each particle responds to its local environment based on the attributes that have been prescribed to it, which may include buoyancy, swimming behavior, growth, or other relevant processes (Irisson et al. 2009). This feature allows IBMs to be configured for a variety of particle types and has resulted in their use across multiple disciplines of marine science (Lynch et al. 2015).

Accurate predictions with IBMs are dependent on correct specification of the input parameters. In addition to individual particle attributes that may be estimated from field and laboratory data, IBM studies universally require that the researcher choose how many particles to release and how to distribute particles among multiple

 Supplemental information related to this paper is available at the Journals Online website: <http://dx.doi.org/10.1175/JTECH-D-15-0115.1>.

Corresponding author address: Benjamin T. Jones, MS 33, Woods Hole Oceanographic Institution, 266 Woods Hole Road, Woods Hole, MA 02543.
E-mail: btjones@mit.edu

DOI: 10.1175/JTECH-D-15-0115.1

origin sites. The number and distribution of particle releases regulates the trade-off between computational time and result accuracy. Brickman and Smith (2002) present a discussion of the errors that may arise from releasing too few particles. The first type of error, which Brickman and Smith (2002) term U-I error, is that the number of particles is insufficient to capture the underlying statistics of the Eulerian velocity field. In the event of U-I error, an identically configured replicate trial will likely give different results. The second type of error, U-II error, is that the particle release distribution does not adequately sample a subarea of particular importance. When U-II errors occur, replicate trials with the same release locations will provide similar results, but the results do not accurately describe the properties of the region as a whole. Both Brickman and Smith (2002) and Simons et al. (2013) present methods to avoid these and similar errors. However, as we explain further in section 5, the methods presented by Brickman and Smith (2002) and Simons et al. (2013) require that the researcher first simulate extra particles, then retrospectively identify how many particles would have been required. IBM studies may simulate tens of millions of particles and consume vast computational resources (e.g., Watson et al. 2012; Jones et al. 2015), and so we seek an alternative method that reduces the required number of particles.

The second design issue, how to distribute particles across origin sites, is more difficult and has been less thoroughly explored in existing literature. One option is to uniformly distribute releases across origin sites (e.g., Watson et al. 2012; Jones et al. 2015). In the case of ecological studies, an alternative is to distribute particle releases based on known spawning distributions (Gallego and North 2009). However, knowledge of spawning distributions is often poor (Gallego and North 2009). As we will show, the choice of release distribution may have substantial implications for the number of particles that are required for statistical confidence, and the issue of optimizing this release distribution is not addressed by previously published methods. We propose a sequential method to optimize the particle release distribution across the origin sites.

We demonstrate our innovative method by estimating the elements of the connectivity matrix. We seek to answer the following questions: What is the minimum number of particles that are necessary to robustly estimate the transport probabilities? To minimize the required number of particles, how should particles be distributed across origin sites? Although our presentation is in the context of estimating the connection probabilities, the method may be applied to other objectives, such as parameterizing models of population

dynamics or assessing the contamination risk from pollutant spills. In addition to the description of our method here, we are also releasing a software package that implements it.

2. A sequential Bayesian procedure

Consider a study system with n_o origins and n_d destinations. Let p_{ij} be the unknown probability that a particle released from origin i is at destination j at a specified time and let $\mathbf{P} = [P_{ij}]$ be the $n_o \times n_d$ matrix of these probabilities (Table 1). Our goal is to estimate \mathbf{P} to a specified precision using a minimal number of particles. Under the sequential Bayesian approach proposed here, the matrix \mathbf{P} is treated as a random variable. Throughout our description of this procedure, we follow the common statistics convention that random variables are indicated by uppercase letters (e.g., P_{ij}) and that realizations of these variables are indicated by lowercase letters (e.g., p_{ij}). As described in more detail below, at each step of the sequential procedure, the current value of an objective function measuring estimation precision is compared to a stopping criterion (Fig. 1). If the criterion is met, then the procedure terminates and each element of \mathbf{P} is estimated by its current expected value. If the criterion is not met, then the current distribution of \mathbf{P} is used to allocate a new batch of particles to the origins, these particles are released, the current distribution of \mathbf{P} is updated based on their observed destinations, and the procedure is repeated. In this section, we describe the basic statistical model, the stopping criterion, and the allocation rule.

a. Statistical model

Let $m_i^{(k)}$ be the number of particles through step k of the sequential procedure that has been released from origin i and let the random variable $X_{ij}^{(k)}$ be the number of these with destination j . Under the assumption that the destinations of different particles are independent and conditional on $\mathbf{p}_i = (p_{i1}, p_{i2}, \dots, p_{i,n_d})$, the vector $\mathbf{X}_i^{(k)} = (X_{i1}^{(k)}, X_{i2}^{(k)}, \dots, X_{i,n_d}^{(k)})$ has a multinomial distribution with $m_i^{(k)}$ trials and probability vector \mathbf{p}_i with the probability mass function given by Eq. (1). The probability mass function below describes the likelihood of observing any realization, $\mathbf{x}_i^{(k)}$, of the random variable $X_i^{(k)}$:

$$\text{pr}(\mathbf{x}_i^{(k)} | \mathbf{p}_i) \propto \prod_{j=1}^{n_d} p_{ij}^{x_{ij}^{(k)}}. \quad (1)$$

To implement the Bayesian approach, it is necessary to specify a prior distribution for the probability vector

TABLE 1. The parameters for our sequential analysis routine are collected and defined here. Following common statistics convention, random variables are indicated with capital letters and realizations of these variables are indicated with lowercase letters.

Symbol	Description
n_o	The total number of origin sites where particles are released.
n_d	The total number of destination sites where particles may arrive.
$\mathbf{P} = [P_{ij}]$	The connectivity matrix. Term p_{ij} is the unknown probability that a particle released from origin i will arrive at destination j . Term \mathbf{P} is the matrix of these probabilities, and the random variable P_i is the i th row of \mathbf{P} .
\mathbf{p}_i	A single realization of the random variable P_i .
$m_i^{(k)}$	The number of particles that have been released from origin i up to and including step k .
$X_{ij}^{(k)}$	A multinomially distributed random variable representing the number of particles released from origin i that arrive at destination j up to and including step k of the procedure. The vector $X_i^{(k)} = (X_{i1}^{(k)}, X_{i2}^{(k)}, \dots, X_{in_d}^{(k)})$.
$x_{ij}^{(k)}$	A single realization of the random variable $X_{ij}^{(k)}$ that gives the observed number of particles released from origin i and arriving at destination j up to and including step k of the procedure.
$\alpha_i^{(k)}$	The vector of parameters for the Dirichlet distribution for P_i at the end of step k . Term $\alpha_i^{(k)}$ is composed of $\alpha_{i1}^{(k)}, \alpha_{i2}^{(k)}, \dots, \alpha_{in_d}^{(k)}$.
$\mu_{ij}^{(k)}$	The mean of the Dirichlet distribution for P_i after step k .
$\sigma_{ij}^{(k)}$	The standard deviation of the Dirichlet distribution for P_i after step k .
$\text{CV}_{ij}^{(k)}$	The coefficient of variation of the Dirichlet distribution for P_i after step k .
$H^{(k)}$	The objective function used to determine when to terminate sampling and how to allocate particle releases.
δ	A threshold that determines when p_{ij} are too small to be relevant to the study goals.
π	A probability threshold that determines when p_{ij} are too small to be relevant to the study goals.
ε	The threshold value for $H^{(k)}$ that determines when sampling terminates.
b	The number of particles simulated in each batch.

P_i . A natural choice is the Dirichlet distribution with probability density function:

$$\text{pr}(\mathbf{p}_i) \propto \prod_{j=1}^{n_d} P_{ij}^{\alpha_{ij}-1} \quad (2)$$

with parameters $\alpha_{i1}, \alpha_{i2}, \dots, \alpha_{in_d}$. In the absence of prior information, it is again natural to take $\alpha_{ij} = 1$ for all i and j so that all possible values of \mathbf{P}_i are equally likely. It follows that the distribution of \mathbf{P}_i after step k is itself Dirichlet with updated parameters $\alpha_{ij}^{(k)} = 1 + x_{ij}^{(k)}$. This reflects the fact that the Dirichlet distribution is the conjugate prior distribution for multinomial data.

b. Stopping criterion

At step k , for each origin i , the current distribution of \mathbf{P}_i is Dirichlet with parameters $\alpha_{ij}^{(k)} = 1 + x_{ij}^{(k)}$, $j = 1, 2, \dots, n_d$. The decision whether to terminate the procedure and estimate p_{ij} by its current mean,

$$\mu_{ij}^{(k)} = \frac{\alpha_{ij}^{(k)}}{\sum_{j=1}^{n_d} \alpha_{ij}^{(k)}}, \quad (3)$$

or to release additional particles must be made on the basis of this distribution. One measure of the current uncertainty in P_{ij} is its coefficient of variation:

$$\text{CV}_{ij}^{(k)} = \frac{\sigma_{ij}^{(k)}}{\mu_{ij}^{(k)}}, \quad (4)$$

where

$$\sigma_{ij}^{(k)} = \sqrt{\frac{\alpha_{ij}^{(k)} \left(\sum_{j=1}^{n_d} \alpha_{ij}^{(k)} - \alpha_{ij}^{(k)} \right)}{\left(\sum_{j=1}^{n_d} \alpha_{ij}^{(k)} \right)^2 \left(\sum_{j=1}^{n_d} \alpha_{ij}^{(k)} + 1 \right)}} \quad (5)$$

is the current standard deviation of P_{ij} . We take as a measure of overall precision the objective function:

$$H^{(k)} = \max_{ij} [\text{CV}_{ij}^{(k)} : \text{pr}^{(k)}(P_{ij} > \delta) > \pi], \quad (6)$$

where $\text{pr}^{(k)}(P_{ij} > \delta)$ is the current probability that P_{ij} exceeds δ . Terms δ and π are small user-specified probabilities. The side condition is required because $\text{CV}_{ij}^{(k)}$ becomes excessively large if the current distribution of P_{ij} is concentrated near 0. The procedure terminates when $H^{(k)}$ first falls below a specified value ε . The choice of the constants δ and π is discussed below in section 6.

c. Allocation rule

If the stopping criterion is not satisfied in step k , then step $k + 1$ begins by sequentially allocating each of a batch of b particles to an origin site. Consider allocating the first such particle under the assumption that, for each origin, the destination of this particle is known. For each origin, we would update the current distribution of \mathbf{P} to include this particle via Bayes's theorem, compute the value of the objective function $H^{(k)}$, and allocate the particle to the origin for which the value of $H^{(k)}$ is smallest. In practice, the destination of the particle

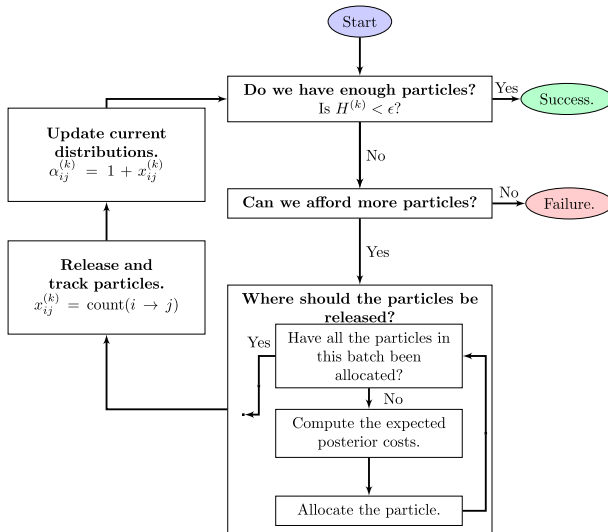


FIG. 1. The sequential analysis procedure is an iterative process. For each iteration, it first assesses whether enough particles have been simulated based on the stopping rule. If not and if additional particles are within the computational budget, then the particles are distributed according to the allocation rule. If at any time the stopping rule is satisfied or the budget is exhausted, the procedure is terminated with either a successful or failed result.

released at a particular origin is unknown until the entire batch has been allocated and the IBM has been run. For this reason, the particle is allocated to the origin with the smallest expected value of the stopping criterion, where this expected value is computed by integrating over the entire predictive distribution for the destination. For a single particle released from origin i , this predictive distribution is Dirichlet multinomial with one trial and parameters $\alpha_{ij}^{(k)}$, $j = 1, 2, \dots, n_d$.

A simulation approach to approximating the expected value of the stopping criterion for a single particle released from origin i proceeds as follows: Simulate a realization \mathbf{p}_i^* of \mathbf{P}_i from the Dirichlet distribution with parameters $\alpha_{ij}^{(k)}$, $j = 1, 2, \dots, n_d$. Simulate a destination from the multinomial distribution with one trial and probability vector \mathbf{p}_i^* . Update the current distribution of \mathbf{P}_i based on this simulated destination and compute the new value of the stopping criterion. Repeat the process many times and approximate the expected value of the stopping criterion by the average of its new values generated from these simulated destinations.

The same general approach is used to allocate the second particle except that destinations are simulated for both the first and second particles. However, in allocating the second particle, the origin of the first particle remains fixed at the origin selected as described above. The process is repeated for each particle in the batch. Because the origins of previously allocated particles are not reconsidered when allocating later

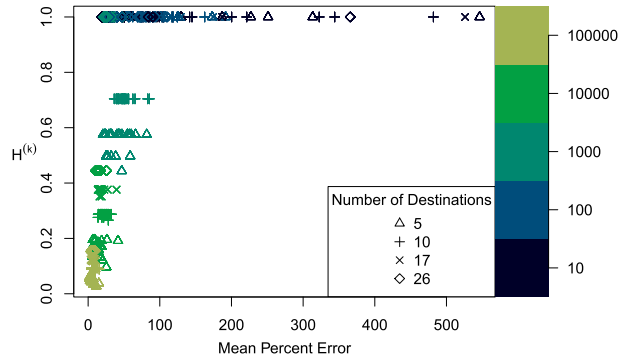


FIG. 2. The objective function (vertical axis) is plotted against the mean percent error in the estimated connectivity matrix (horizontal axis). Each data point was computed by randomly generating a matrix $\mathbf{x}^{(k)}$ from one of the artificially generated connectivity matrices. The color indicates the number of particles that were included in $\mathbf{x}^{(k)}$, and the plotting symbol indicates the number of destinations in the connectivity matrix.

particles, this procedure is not guaranteed to identify the optimal allocation of the batch of particles. Pseudocode to implement this allocation rule is provided in the [appendix](#).

3. Validation using artificial data

We validated our procedure using artificial data based on ecological networks (e.g., [Kininmonth et al. 2010](#); [Watson et al. 2011](#); [Jones et al. 2015](#)). For each replicate, we constructed a connectivity matrix and then drew multinomial samples from it that represent Lagrangian particles. Because we know the underlying connectivity matrix, this test ensures convergence to the correct solution.

Our objective function measures the precision of each p_{ij} , which may also be measured by the percent error in the estimated connectivity matrix when the true connectivity matrix is known. Because the connectivity matrix that was used to generate the artificial data is known, the artificial data may be used to assess the relationship between the objective function $H^{(k)}$ and the percent error. We randomly generated 25 connectivity matrices with each having $n_o = (4, 9, 16, \text{ and } 25)$ origins and $n_o + 1$ destinations. The first n_o destinations were the same as the origins, and between 0% and 10% of the particles returned to these origins. Destination n_d represented everywhere else. Each row of these connectivity matrices gives the probability vector for a multinomial distribution from which we took samples that represent Lagrangian particles. We treated these samples as a single run of a Lagrangian particle-tracking model and estimated the connectivity matrix, and then computed $H^{(k)}$ from this estimate. Term $H^{(k)}$ provides an upper bound on the percent error ([Fig. 2](#)), indicating

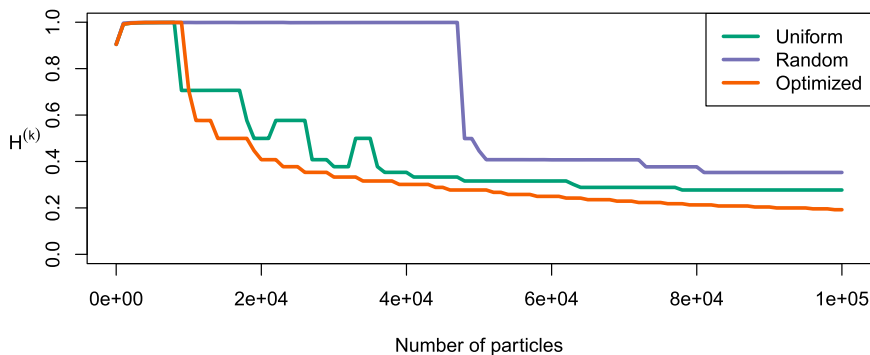


FIG. 3. Ten sequential simulations were run using nine node artificially generated connectivity matrices. The results of all 10 simulations were similar, and so only one of them is plotted here. The number of particles included for the estimate for $H^{(k)}$ is depicted on the horizontal axis, and the particle allocation scheme is given by the color of the line.

that it is a valuable error metric. The value of $H^{(k)}$ is inversely related to the number of particles that followed each possible pathway. When few particles have been simulated relative to the number of destinations, $H^{(k)}$ is large, indicating that these few particles may not provide a good estimate for the connectivity matrix. However, as the number of particles increases, both $H^{(k)}$ and the percent error decrease, and so small $H^{(k)}$ correctly indicates that the percent error is small. Fewer particles are required for connectivity matrices with fewer destinations because having fewer destinations results in larger transport probabilities under our connectivity matrix-generating scheme. Although the expected value of the posterior percent error could have been used instead of the CV-based objective function, the CV has the practical benefit of an analytic solution and accurately indicates when the percent error is small.

We also tested that the allocation rule results in faster convergence of $H^{(k)}$ than either uniformly or randomly distributing particle releases. The uniform distribution represents the null case where particles are released throughout the domain, and the random distribution mimics particle releases based upon criteria that do not correlate well with the flow patterns (e.g., species distributions). Our method consistently outperformed both alternatives in 10 simulations, and the simulations revealed interesting aspects of $H^{(k)}$ (Fig. 3). The objective function initially reacts only to the missing connections that have the largest CV, and $H^{(k)}$ reduces to $\sqrt{m_i^{(k)}(m_i^{(k)} + 2)^{-1}}$ for these connections. Therefore, the objective function initially increases asymptotically toward 1 until these missing connections are identified and then subsequently decreases. Because our allocation rule assumes that the objective function monotonically decreases as more particles are simulated, this property of the objective function is problematic. The threshold

number of particles required to satisfy $P(p_{ij} \geq \delta) < \pi$ may be computed by solving the relation $\pi = F(\delta, 1, n_i)$, where $F(\delta, 1, m_i^{(k)})$ is the cumulative distribution function of the beta distribution with shape parameters 1 and $m_i^{(k)}$ evaluated at δ , and we recommend that users release this number of particles from each origin in the first batch. Once the missing connections are identified, the allocation rule outperforms the alternatives, and $H^{(k)}$ decreases in proportion to the square root of the number of particles. Term $H^{(k)}$ may also increase when p_{ij} are approximately equal to δ . In this scenario, connections alternate between satisfying and not satisfying $P(p_{ij} \geq \delta) \geq \pi$, and rapid changes in the value of $H^{(k)}$ occur as shown by the uniform allocation scheme in Fig. 3. However, these changes are transient features, and so the allocation rule performs well in spite of them. In all trials, the random distribution resulted in poor convergence of the objective function, suggesting that allocation schemes based on spawning distributions should be avoided when the objective is to precisely estimate the connectivity matrix.

Overall, our method performs well on artificial networks that represent ecological networks. It converges to the correct solution, and converges more quickly than either null distribution of particle releases.

4. Validation using a realistic tracking simulation

We further validated our method using a simulation of the Gulf of Maine as a representative IBM study. Our simulation is based upon that of Huret et al. (2007). For brevity, we describe only where our study differs from the original. We used a particle-tracking model to simulate cod larval dispersal during January 1995. We forced the particle-tracking model with hourly output from the Finite Volume Coastal Ocean Model (FVCOM; Chen et al. 2003). FVCOM was configured

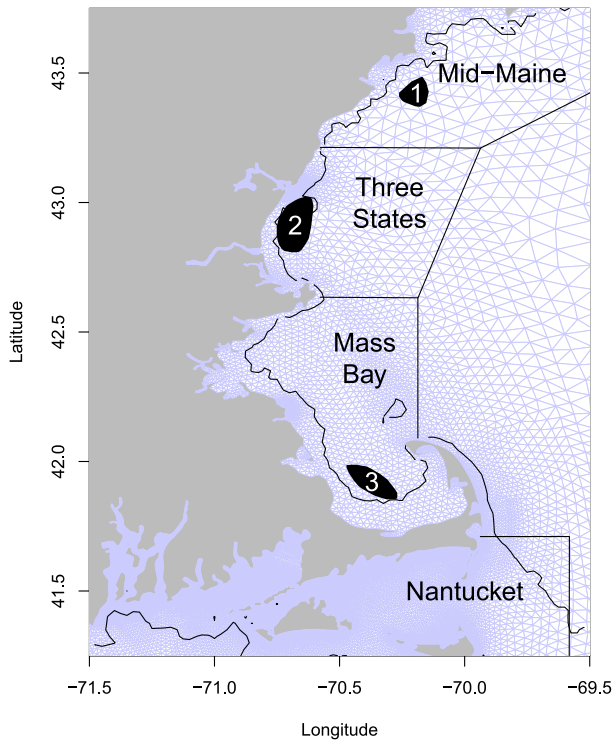


FIG. 4. The study regions are depicted here. The numbered sites are the particle release locations. The straight boundary lines indicate the destination regions, and the black line nearshore indicates the 30-m isobath that was used to determine suitable habitat. The blue background mesh is the FVCOM mesh.

using the third generation of the Gulf of Maine mesh, which contains 48 451 nodes and 90 415 elements that smoothly transition from 200-m resolution at the coastline to 15-km resolution in the central Gulf of Maine and extends from Maryland to Nova Scotia, Canada (Fig. 4).

Particles that represent cod larvae were released from three spawning sites along the coast of New England (Fig. 4) throughout January 1995. The spawning grounds were taken from the map published in Huret et al. (2007). Particle release locations within each spawning region were randomly selected in time and space from a uniform distribution. Particle destinations were computed from the position of the particle at 60 days age.

Our first test validated the use of a multinomial distribution. The multinomial distribution assumes independence between particles, which may not be appropriate if particles are released closely in space and time. We released 1000 particles from each spawning ground and estimated the connectivity matrix. We repeated this process 100 times with different release locations and timing and obtained 100 estimates for each element of the connectivity matrix. We assumed that the mean of these 100 trials represents the expected

outcome, and tested this assumption using the variance test from Brickman and Smith (2002). The variance of the mean leveled off after 40–60 trials, indicating that our use of 100 trials is sufficient (supplemental Fig. S1). We computed the χ^2 statistic for each element, $\sum_{k=1}^{100} (p_{ij} - \hat{p}_{ij}^{(k)})^2 p_{ij}^{-1}$, where $\hat{p}_{ij}^{(k)}$ is the estimate of p_{ij} from the k th trial and p_{ij} is the mean of these estimates across all 100 trials. The observed distributions of the χ^2 statistics did not differ from those that would result from multinomial sampling (Fig. 5).

Our second test evaluated the allocation rule. We sequentially released batches of 500 particles whose distribution was determined by our allocation rule, either by a uniform distribution or by a randomly chosen distribution, until our computational budget of 50 000 particles was exhausted. During these tests, we set $\varepsilon = 0.1$, $\delta = 0.005$, and $\pi = 0.05$. In three repetitions, our methodology consistently increased the convergence rate of $H^{(k)}$ (Fig. 6). Only the optimized distribution scheme satisfied the stopping criterion within the budget by reaching the threshold value of 0.1. Upon exhausting the budget, $H^{(k)} = 0.11 \pm 0.0039$ (mean plus/minus standard deviation) for uniformly distributed particles and $H^{(k)} = 0.28 \pm 0.030$ for the random distribution. The optimized distribution satisfied the stopping criterion after simulating $26\,666 \pm 3253$ particles. At the point where the optimized distribution satisfied the stopping criterion, $H^{(k)} = 0.14 \pm 0.008$ for the uniform distribution and $H^{(k)} = 0.40 \pm 0.017$ for the random distribution.

5. Alternative methods

Choosing the number of Lagrangian particles is a fundamental component of IBM studies, and previous publications have described alternative methods to address this issue. Brickman and Smith (2002) proposed the variance test as a method to identify the presence of both U-I and U-II errors. To apply the variance test, researchers first generate a set of release locations that evenly distributes b particles throughout a single origin site. The researchers then perform t replicate simulations using this release distribution. Variability among the trials emerges due to a stochastic component in the particle velocities, and this variability is quantified with the test statistic $V^{(k)}$. To compute $V^{(k)}$, the researchers draw a random sample of k trials from the t trials available. Term $V^{(k)}$ is the mean variance in a sample of size k divided by k . Term $V^{(k)}$ decays with increasing k , and the researchers may be confident that their results are not subject to U-I error when the $V^{(k)}$ versus k curve levels off. To protect against U-II error, they suggest modifying the variance test to use increasing b instead of increasing k .

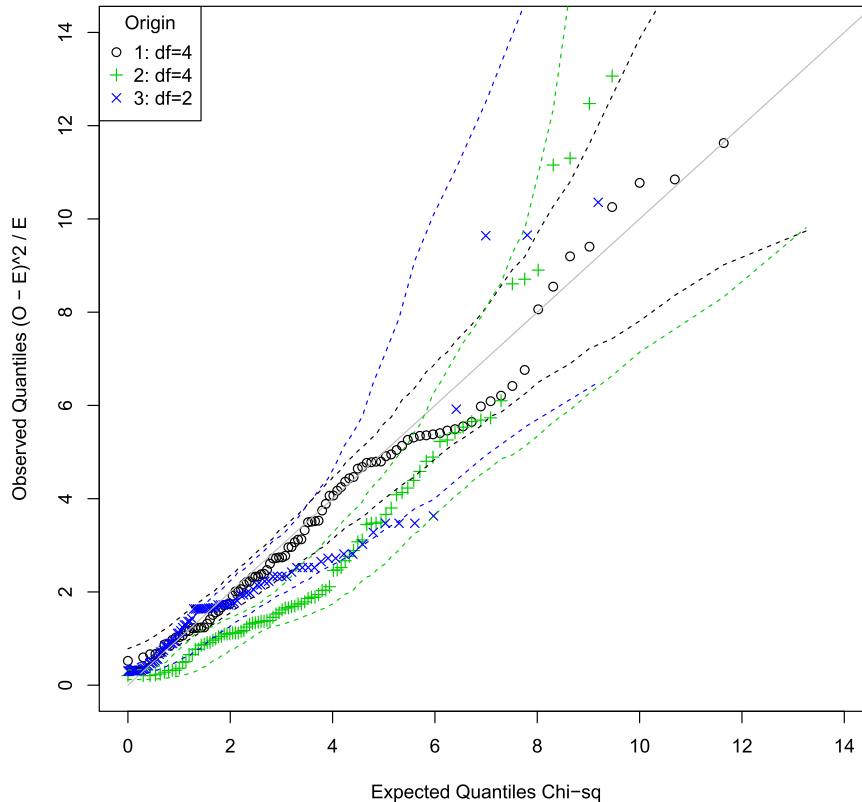


FIG. 5. The expected quantiles from a χ^2 distribution are plotted against the observed quantiles of the χ^2 statistic from many particle-tracking simulations. The dashed lines indicate a 95% confidence interval, and the solid line indicates a one-to-one relationship. For origins 1 and 2, we observed five possible destinations, and so there are 4 degrees of freedom in the χ^2 distribution. For origin 3, particles only went to three destinations due to strongly directional southern flow, and so there are only 2 degrees of freedom.

Simons et al. (2013) propose an alternative method to test the related question, how many particles are required to ensure that a simulation closely approximates a reference solution? The first step in their method is to compute a single large trial with b particles and compute a reference solution. Because this solution is computed from the largest number of particles available, they assume that it provides the best representation of the underlying flow and they seek to replicate it with a reduced number of particles. They begin by drawing a random subset of s particles from the pool of b particles, and compute a sample solution from this subset. They then compare the sample solution to the reference solution by computing the fraction of unexplained variance (FUV) between the solutions as $FUV^{(s)} = 1 - r^2$, where r is the linear correlation coefficient between the solutions. Repeating this process many times for multiple values of s , they obtain a curve that plots $FUV^{(s)}$ against s . Finally, they threshold this curve when $FUV^{(s)}$ is sufficiently small to identify an appropriate value of s .

Although our procedure, the variance test, and the FUV method all address similar questions, our method is structured differently from the others in order to reduce the required number of particles. Both the variance test and the FUV method begin by simulating a large pool of trials or particles, and then they subsample from this pool to estimate the variability in the results. For the variance test, t must be greater than k to subsample and compute $V^{(k)}$. For the FUV method, b must be greater than s to estimate $FUV^{(s)}$. In contrast, our method alternates between simulating particles and assessing convergence and then terminates as soon as convergence is achieved. However, this design choice prohibits subsampling from a larger pool to estimate the variability in the results, and instead we estimate the variability from the properties of the posterior distribution for each p_{ij} . Each of the three proposed methods has merits in addressing issues related to the number of required particles for IBM studies, but each method differs slightly in how each does it.

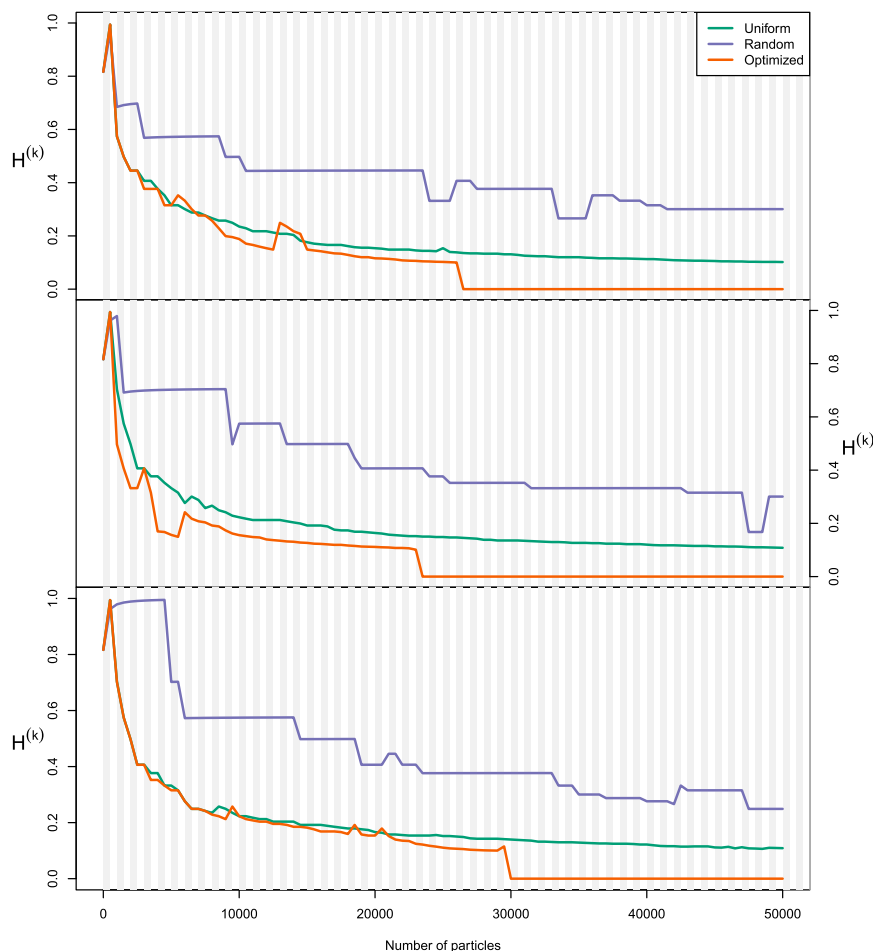


FIG. 6. Particles were released uniformly, randomly, and using the allocation rule three times in a particle-tracking model for the Gulf of Maine. Particles were simulated in batches of 500, which are indicated by the shaded regions, and a total budget of 50 000 particles was permitted. The colored lines display the decrease in value for the objective function during each simulation and under each particle release scheme.

6. Discussion

We provide a flexible and reliable method to match particle release counts and distributions to the specific objectives of a particular study. The method avoids both the U-I and U-II errors discussed in [Brickman and Smith \(2002\)](#). U-I errors occur when replicate simulations would result in substantially different results. Our method avoids this error by evaluating a stopping criterion and continuing the simulation until variability in the results is sufficiently small. U-II errors occur when the release distribution skips over subregions of particular importance. Whereas [Brickman and Smith \(2002\)](#) evenly distribute particle releases throughout each origin and reuse the same release locations for each trial, we draw a new set of release locations from a uniform distribution for each step. This procedure avoid U-II

errors altogether, because a large number of randomly drawn points will represent the underlying structure of each origin. Although we draw the release locations within each origin from a uniform distribution in our examples, egg production models or finescale field data may be used to generate these distributions when such information is available ([Gallego and North 2009](#)). Our method also addresses how to distribute releases among multiple origins in order to minimize the number of particles required to achieve statistical confidence, which has not been done by prior studies.

Although our method assumes that b particles are simulated in each batch, choosing b is dependent on the specific IBM being used. IBMs may be operated in on-line mode and load the Eulerian velocity fields directly from a hydrodynamic model, or in offline mode and read the velocity fields from archived output of a

hydrodynamic model. In either case, there is a computational cost to operating the hydrodynamic model or reading the circulation fields. This cost is incurred every time a batch of particles is simulated, but it is largely independent of the number of particles being simulated in each batch. A trade-off emerges where small b allows our method to most effectively allocate particles among origins and terminate most quickly, but large b increases the efficiency of the IBM and reduces the cost per particle. Choosing an optimal value of b may reduce the computational cost required to achieve convergence, but the choice of b does not influence when our method deems that convergence has been achieved. The computational overhead of loading the velocity fields is specific to each IBM and hydrodynamic model configuration, and so we recommend that researchers choose b such that their IBM operates with a reasonable level of efficiency.

Our method also assumes that the multinomial distribution is an appropriate model for the particle destinations, which implies that the trajectories are independent. Multiple releases that are closely located in time and space may result in correlation among trajectories. However, randomly chosen release locations within an origin, releases separated by at least the velocity decorrelation scale, or tracking durations longer than the Lagrangian decorrelation time will likely avoid this concern. Each particle may only contribute to one destination, which excludes settlement criteria based on the proportion of time that a particle spends within a destination region (e.g., Huret et al. 2007). An alternative is to assign each particle a probability of settling during each time step and then remove it from further consideration after settlement (e.g., Tian et al. 2009b).

Our examples focus on a single objective function and stopping rule that reflect our objectives from applying this procedure. Because the CV responds to the uncertainty in each p_{ij} relative to the value of that p_{ij} , it is appropriate for use when the estimates for p_{ij} are multiplied together and errors would be multiplicative (e.g., in a matrix projection population model). Likewise, ignoring very small p_{ij} was chosen to reflect that very low connectivity rates among subpopulations may not substantially impact population demographics (Hastings 1993; Lowe and Allendorf 2010). Choosing the parameters δ and π for this objective function is study specific, but here we present some examples for consideration. In ecology, only a few migrants per generation are necessary to maintain genetic homogeneity, and many fish spawn millions of eggs each year (Slatkin 1987). Therefore, studies examining genetic connectivity must quantify even rare connections, and $\delta = 10^{-6}$ may be appropriate. However, a more frequent exchange of individuals is required for connectivity to influence

population dynamics, and so studies examining population demographics may set $\delta = 10^{-2}$ (Hastings 1993; Lowe and Allendorf 2010). The second parameter, π , is analogous to the significance level in frequentist statistical tests, and so we suggest $\pi = 0.05$ as a default value. However, these are merely default suggestions, and researchers may alter them based upon the goals of individual studies.

More broadly, users may replace Eq. (6) with an appropriate representation of what is important in their system. The objective function must take the parameter vectors α as an argument and return a single scalar value that quantifies the quality of α . For example, ecological studies that include population connectivity as one component of a population model may quantify the variability of the results differently. Realized population connectivity patterns include spawning distributions and postsettlement survival (Watson et al. 2010). Researchers seeking to estimate these patterns may develop a population model that includes these processes, evaluate the population model using many credible values for the connectivity matrix \mathbf{P} and seek to minimize the variance in the evaluations. Either the output of a particle-tracking model or the objective function must include all processes relevant to the study, including, for example, survival and growth of larvae and loss of particles to the model boundaries. The allocation rule relies on two assumptions that any choice of objective function must satisfy. First, the objective function must decrease as the quality of the estimated connectivity matrix increases. Second, releasing more particles from an origin must reduce the contribution of that origin to the value of the objective function. We suggest that practitioners test these assumptions when using a new objective function. The software package associated with this publication includes methods for performing this test. Our allocation rule is a greedy heuristic that provides an improved, but suboptimal, particle distribution. In the future, we hope to provide theory that bounds the difference between the output of our allocation rule and the optimal solution.

Particle counts in particle-tracking studies vary widely from a few thousand (e.g., Huret et al. 2007; Tian et al. 2009a) to tens of millions (e.g., Watson et al. 2012; Jones et al. 2015). Field research that relies on parentage, tagging, or drifter data may be limited to only a few hundred sample points (Almany et al. 2007; Planes et al. 2009). The appropriate number of particles is dependent on the study goals, and readers and authors must take care to avoid drawing conclusions beyond those that can be justified by the number of particles. Our method provides a robust and quantitative way to determine the count and distribution of particle releases, which can help

researchers obtain more precise estimates of transport probabilities with reduced costs, draw appropriate conclusions from tracking experiments, and thus lead to better understanding of marine ecosystems.

7. Code availability

An online interface to our method is available (<http://btjones.scripts.mit.edu/index.fcgi/research/sequential-analysis-method>). Source code and instructions for installing and accessing our method are available (<https://github.com/btjones16/sequential-analysis-software>). The source code repository includes R and C++ libraries, together with a Simplified Wrapper and Interface Generator (SWIG) interface file that allows access to the C++ library from Python, Octave, and other scripting languages (Beazley 1996).

Acknowledgments. We thank Katie Samuelson, the Larval Ecology reading group at WHOI, Editor Carlos Lozano, and two anonymous reviewers for their helpful feedback regarding this manuscript, and Changsheng Chen for providing the FVCOM output. This research was supported by the Department of Defense (DoD) through the National Defense Science and Engineering Graduate Fellowship (NDSEG) program and the National Science Foundation through Grant OCE-1459133 and Grant OCE-1031256.

APPENDIX

Pseudocode Implementation of Allocation Rule

This appendix provides pseudocode for a naïve implementation of the allocation rule. We recommend that readers refer to the software packages referenced in the main text for more computationally efficient implementations.

```

function computeExpectedPosteriorCost
( $\alpha_i^{(k)}$ ,  $n$ )
  while estimate for  $H^{(k+1)}$  not converged do
     $p_i^*$   $\leftarrow$  draw from a Dirichlet distributions
    with parameters  $\alpha_i^{(k)}$ 
     $d \leftarrow$  draw  $n$  particles from a multinomial
    distribution with parameters  $p_i^*$ 
     $\alpha_i^{(k+1)} \leftarrow \alpha_i^{(k)} + d$ 
    estimate  $H^{(k+1)}$ 
  end while
  return estimated  $H^{(k+1)}$ 
end function
 $b \leftarrow$  number of particles to allocate
 $m \leftarrow (0, 0, \dots, 0)$  (release distribution)

```

```

for 1, 2, ...,  $b$  do
  for  $i$  in origins do
     $H_i^{(k+1)} \leftarrow$  computeExpectedPosteriorCost
    ( $\alpha_i^{(k)}$ ,  $m_i + 1$ )
  end for
   $i \leftarrow \arg \min(H_i^{(k+1)})$ 
   $m_i = m_i + 1$ 
end for

```

REFERENCES

- Almany, G. R., M. L. Berumen, S. R. Thorrold, S. Planes, and G. P. Jones, 2007: Local replenishment of coral reef fish populations in a marine reserve. *Science*, **316**, 742–744, doi:10.1126/science.1140597.
- Beazley, D. M., 1996: SWIG: An easy to use tool for integrating scripting languages with C and C++. *Proceedings of the Fourth Conference on USENIX Tcl/Tk Workshop, 1996*, Vol. 4, USENIX Association, 15–15. [Available online at <http://dl.acm.org/citation.cfm?id=1267498.1267513>.]
- Brickman, D., and P. C. Smith, 2002: Lagrangian stochastic modeling in coastal oceanography. *J. Atmos. Oceanic Technol.*, **19**, 83–99, doi:10.1175/1520-0426(2002)019<0083:LSMICO>2.0.CO;2.
- Chen, C., H. Liu, and R. C. Beardsley, 2003: An unstructured grid, finite-volume, three-dimensional, primitive equations ocean model: Application to coastal ocean and estuaries. *J. Atmos. Oceanic Technol.*, **20**, 159–186, doi:10.1175/1520-0426(2003)020<0159:AUGFVT>2.0.CO;2.
- Cowen, R. K., and S. Sponaugle, 2009: Larval dispersal and marine population connectivity. *Annu. Rev. Mar. Sci.*, **1**, 443–466, doi:10.1146/annurev.marine.010908.163757.
- Gallego, A., and E. W. North, 2009: Initial conditions: Spawning locations. *Manual of Recommended Practices for Modelling Physical Biological Interactions during Fish Early Life*, E. W. North, A. Gallego, and P. Petitgas, Eds., ICES Cooperative Research Report, Vol. 295, International Council for Exploration of the Sea, 20–23.
- Hastings, A., 1993: Complex interactions between dispersal and dynamics: Lessons from coupled logistic equations. *Ecology*, **74**, 1362–1372, doi:10.2307/1940066.
- Huret, M., J. A. Runge, C. Chen, G. Cowles, Q. Xu, and J. M. Pringle, 2007: Dispersal modeling of fish early life stages: Sensitivity with application to Atlantic cod in the western Gulf of Maine. *Mar. Ecol.: Prog. Ser.*, **347**, 261–274, doi:10.3354/meps06983.
- Irisson, J.-O., J. M. Leis, C. B. Paris, and H. I. Browman, 2009: Behavior and settlement. *Manual of Recommended Practices for Modelling Physical Biological Interactions during Fish Early Life*, E. W. North, A. Gallego, and P. Petitgas, Eds., ICES Cooperative Research Report, Vol. 295, International Council for Exploration of the Sea, 42–59.
- Jones, B. T., J. Gyory, E. K. Grey, M. Bartlein, D. S. Ko, R. W. Nero, and C. M. Taylor, 2015: Transport of blue crab larvae in the northern Gulf of Mexico during the Deepwater Horizon oil spill. *Mar. Ecol.: Prog. Ser.*, **527**, 143–156, doi:10.3354/meps11238.
- Kinmonth, S., G. Death, and H. Possingham, 2010: Graph theoretic topology of the Great but small Barrier Reef world. *Theor. Ecol.*, **3**, 75–88, doi:10.1007/s12080-009-0055-3.

- Lowe, W. H., and F. W. Allendorf, 2010: What can genetics tell us about population connectivity? *Mol. Ecol.*, **19**, 3038–3051, doi:10.1111/j.1365-294X.2010.04688.x.
- Lynch, D. R., D. A. Greenberg, A. Bilgili, D. J. J. McGillicuddy, J. P. Manning, and A. L. Aretxabaleta, 2015: *Particles in the Coastal Ocean: Theory and Applications*. Cambridge University Press, 560 pp.
- Miller, C. B., and P. A. Wheeler, 2012: *Biological Oceanography*. 2nd ed. Wiley-Blackwell, 464 pp.
- Pineda, J., J. A. Hare, and S. Sponaugle, 2007: Larval transport and dispersal in the coastal ocean and consequences for population connectivity. *Oceanography*, **20** (3), 22–39, doi:10.5670/oceanog.2007.27.
- Planes, S., G. P. Jones, and S. R. Thorrold, 2009: Larval dispersal connects fish populations in a network of marine protected areas. *Proc. Natl. Acad. Sci. USA*, **106**, 5693–5697, doi:10.1073/pnas.0808007106.
- Simons, R. D., D. A. Siegel, and K. S. Brown, 2013: Model sensitivity and robustness in the estimation of larval transport: A study of particle tracking parameters. *J. Mar. Syst.*, **119 & 120**, 19–29, doi:10.1016/j.jmarsys.2013.03.004.
- Slatkin, M., 1987: Gene flow and the geographic structure of natural populations. *Science*, **236**, 787–792, doi:10.1126/science.3576198.
- Tian, R. C., and Coauthors, 2009a: Dispersal and settlement of sea scallop larvae spawned in the fishery closed areas on Georges Bank. *ICES J. Mar. Sci.*, **66**, 2155–2164, doi:10.1093/icesjms/fsp175.
- , and Coauthors, 2009b: Modeling the connectivity between sea scallop populations in the Middle Atlantic Bight and over Georges Bank. *Mar. Ecol.: Prog. Ser.*, **380**, 147–160, doi:10.3354/meps07916.
- Watson, J. R., S. Mitarai, D. A. Siegel, J. E. Caselle, C. Dong, and J. C. McWilliams, 2010: Realized and potential larval connectivity in the Southern California Bight. *Mar. Ecol.: Prog. Ser.*, **401**, 31–48, doi:10.3354/meps08376.
- , D. A. Siegel, B. E. Kendall, S. Mitarai, A. Rassweiler, and S. D. Gaines, 2011: Identifying critical regions in small-world marine metapopulations. *Proc. Natl. Acad. Sci. USA*, **108**, E907–E913, doi:10.1073/pnas.1111461108.
- , B. E. Kendall, D. A. Siegel, and S. Mitarai, 2012: Changing seascapes, stochastic connectivity, and marine metapopulation dynamics. *Amer. Nat.*, **180**, 99–112, doi:10.1086/665992.