Adaptive stochastic trajectory modelling in the chaotic advection regime

By J. G. E S L E R

Department of Mathematics, University College London,
Gower Street, London WC1E 6BT, UK
gavin@math.ucl.ac.uk

Motivated by the goal of improving and augmenting stochastic Lagrangian models of particle dispersion in turbulent flows, techniques from the theory of stochastic processes are applied to a model transport problem. The aim is to find an efficient and accurate method to calculate the total tracer transport between a source and a receptor when the flow between the two locations is weak, rendering direct stochastic Lagrangian simulation prohibitively expensive. Importance sampling methods that combine information from stochastic forward and back trajectory calculations are proposed. The unifying feature of the new methods is that they are developed using the observation that a perfect strategy should distribute trajectories in proportion to the product of the forward and adjoint solutions of the transport problem, a quantity here termed the 'density of trajectories' $D(\mathbf{x}, t)$.

Two such methods are applied to a 'hard' model problem, in which the prescribed kinematic flow is in the large Péclet number chaotic advection regime, and the transport problem requires simulation of a complex distribution of well-separated trajectories. The first, Milstein's measure transformation method, involves adding an artificial velocity to the trajectory equation and simultaneously correcting for the weighting given to each particle under the new flow. It is found that although a 'perfect' artificial velocity $\mathbf{v}^*$ exists, which is shown to distribute the trajectories according to D, small errors in numerical estimates of $\mathbf{v}^*$ cumulatively lead to difficulties with the method. A second method is Grassberger's 'go-with-the-winners' branching process, where trajectories found unlikely to contribute to the net transport (losers) are periodically removed, while those expected to contribute significantly (winners) are split. The main challenge of implementation, which is finding an algorithm to select the winners and losers, is solved by a choice that explicitly forces the distribution towards a numerical estimate of D generated from a previous back trajectory calculation. The result is a robust and easily implemented algorithm with typical variance up to three orders of magnitude lower than the direct approach.

1. Introduction

A common scenario in scalar transport problems in fluids is that a tracer is released at one location (the 'source' $\mathcal{S}$) and at some later time is measured at another (the 'receptor' $\mathcal{R}$). In many practical problems there is little interest in the tracer concentration at locations far from $\mathcal{R}$. For this type of problem a grid-based Eulerian transport model, designed to provide global solutions for the tracer concentration, is evidently inefficient. For certain flow set-ups, a Lagrangian model will offer greater efficiency, provided that sufficient numbers of trajectories are advected from source to receptor, i.e. the relative tracer concentration at the receptor is large in some sense. The present work addresses the opposite situation, in which the relative concentration at the receptor is not large, because the flow of particles from source to receptor is (in a certain sense) weak. The aim is to alleviate the difficulty associated with this situation by modifying and applying variance reduction / importance sampling methods developed for stochastic differential equations (e.g. Øksendal 2007; Kloeden & Platen 1992; Gardiner 2009), such as are widely used in mathematical finance and statistical physics, in order to redirect trajectories into the region of interest. Subsequently, we will refer to this process as adaptive trajectory modelling.

There are many possible practical applications. Consider, for example, an isolated release of a chemical that is highly toxic at low concentrations with the receptor an isolated centre of population. Alternatively, the source could be an erupting volcano and the receptor an airport or flight corridor (e.g. Devenish *et al.* 2011). Further, a common problem for meteorological and atmospheric chemistry aircraft measurement campaigns (e.g. Methven *et al.* 2006) is to quantify the influence of an upstream airmass (e.g. near a chemical source) upon another airmass downstream where measurements have been made. In addition to atmospheric dispersion modelling, adaptive trajectory modelling will have applications in ocean tracer studies (e.g. Proehl *et al.* 2005; Spivakovskaya *et al.* 2007), and the modelling of pollutants in ground water, i.e. flow through porous media (e.g. Dagan 1987; Zimmermann *et al.* 2001). Nor are there significant obstacles to extending our results to more sophisticated atmospheric models of turbulent dispersion (for example the so-called 'Lagrangian particle dispersion models' (LPDMs) e.g. Thomson 1987; Rodean 1996; Stohl *et al.* 2005, which are more sophisticated than the advection-diffusion set-up considered here). More broadly, the application of importance sampling detailed here may serve as a useful model problem for the development of 'particle filters' in data assimilation problems (Ades & van Leeuwen 2013), where many similar issues arise. For fluid dynamicists, the problem is of fundamental interest in that it brings to

the fore some interesting connections between Lagrangian and Eulerian descriptions of advection-diffusion problems (see e.g. Majda & Kramer 1999).

Here, two specific importance sampling techniques will be considered in the context of a model problem. These are the measure transformation method of Milstein (1995) (see also Kloeden & Platen 1992; Milstein & Tretyakov 2004), and the 'go-with-the-winners' (GWTW hereafter) branching process (Grassberger 1997, 2002). Both methods have recently (Haynes & Vanneste 2014) been used with some limited success in simple parallel flows, to obtain far-field concentrations in Taylor dispersion problems (Taylor 1953). Due to their relative simplicity, however, parallel flow problems obscure key issues as to how these methods might be applied effectively in situations of practical interest, such as large-scale atmospheric or oceanic transport.

The model problem to be investigated is chosen to highlight some of the key issues arising in practical atmosphere-ocean fluid transport problems. One important feature, perhaps unlikely to be encountered in the finance or molecular dynamics problems for which importance sampling methods have been developed, is 'chaotic advection', i.e. the exponential divergence of nearby trajectories of the underlying deterministic flow. It will be argued below that in the chaotic advection regime an important class of importance sampling techniques of the 'small noise'-type, which are based on results from large-deviation theory (e.g. Vanden-Eijnden & Weare 2012), are rendered impractical. Here the advection-diffusion equation is solved in a periodic channel bounded by sidewalls, with the prescribed flow in the channel consisting of a uniform current, on which is superimposed a linear combination of two waves propagating with different frequencies. Previously this flow has been used to explore chaotic advection in Rayleigh-Bernard convection (Weiss & Knobloch 1989) and in geophysical flows dominated by Rossby waves (Pierrehumbert 1991; Haynes *et al.* 2007). Other aspects of the model problem are also relevant to layer-wise two-dimensional (near isentropic) transport in the extratropical troposphere and stratosphere (e.g. Plumb 2002). For example, the Péclet number Pe is relatively high ($Pe = 5000$ in the numerical calculations below), and the flow exhibits a partial barrier to mixing (for details see Haynes *et al.* 2007). This partial mixing barrier, which is approximately co-located with the region of maximum flow speed, is representative of observed geophysical phenomena such as the edge of the stratospheric polar vortices, the extratropical tropopause, or the core of ocean currents such as the Gulf stream and Kuroshio, as well as in the Southern ocean (see e.g. Weiss & Provenzale 2008). The focus here will be on the (weak) transport across the barrier, which often needs to be quantified in practical applications.

A further reason for investigating the chosen model problem is that, provided the diffusivity exceeds a numerically determined threshold, a spectral method can be used to solve the advection-diffusion equation in the channel to very high accuracy (the 'PDE

solution' hereafter). It is to be emphasised that our model problem has been selected with the availability of this numerical PDE solution in mind. In a more general setting, a global PDE solution will be expensive and (typically) less accurate, and the advantages of the Lagrangian approach (e.g. flexibility, positivity, effortless computational parallelisation) will be more to the fore. Here, however, the PDE solution will act to benchmark the Lagrangian methods developed, and crucially can be used to deduce much about what an optimal importance sampling strategy should look like.

In section 2 the model problem is described, benchmark results from both forward and adjoint PDE solutions are presented, and a quantity termed the *density of trajectories* $D(\mathbf{x}, t)$ is introduced. It is argued that the aim of an optimal importance sampling strategy for $\mathcal{S} \to \mathcal{R}$ transport should be to distribute particles in proportion to D. In section 3 the stochastic representations of the forward and adjoint problems are introduced. Milstein's measure transformation method is then reviewed from an applied mathematics perspective. The choice of correcting velocity is discussed, and it is shown that under Milstein's optimal choice trajectories are indeed distributed in proportion to $D(\mathbf{x}, t)$. The resulting insight is used to define an effective scoring strategy for the GWTW algorithm. In section 4 practical 'adaptive' Lagrangian strategies to solve the model problem are developed and assessed using the PDE solution. A gain of several orders of magnitude in efficiency is shown to be possible. Section 5 gives the conclusions.

2. The model problem

2.1. *The forward problem*

The broad class of advection-diffusion problems that concern us here satisfy

$$(\partial_t + \mathbf{u} \cdot \nabla) c - \nabla \cdot (\boldsymbol{\kappa} \cdot \nabla c) = s(\mathbf{x}, t). \tag{2.1}$$

Here $c(\mathbf{x}, t)$ is a passive tracer mixing ratio, $\mathbf{u}(\mathbf{x}, t)$ is a known 'smooth' incompressible velocity field, $\boldsymbol{\kappa}(\mathbf{x}, t)$ is a symmetric diffusivity tensor, and $s(\mathbf{x}, t)$ a source term. The relevance of (2.1) as a model for large-scale transport in the atmosphere, at least in regions where the flow is dominated by stratification and rotation (i.e. away from the planetary boundary layer and regions of active convection), is well-established (see e.g. Haynes 2011, and refs. therein). In atmospheric applications the 'smoothness' property of $\mathbf{u}$ applies only to a suitably temporally or spatially averaged flow, with the effects of three-dimensional turbulent perturbations about this averaged flow incorporated into the eddy diffusivity $\boldsymbol{\kappa}$ (see e.g. Majda & Kramer 1999, for discussion of the applicability and limitations of the eddy diffusivity approach).

In the model problem to be considered, (2.1) will be solved in a domain $\mathcal{D}$, a channel that is periodic in the x-direction and is bounded by sidewalls in the y-direction, with dimensions $2\pi \times \pi$. No-flux boundary conditions are applied on the boundaries $\partial \mathcal{D}$ (i.e.

the sidewalls at $y = 0, \pi$)

$$\mathbf{n} \cdot \boldsymbol{\kappa} \cdot \nabla c = 0 \quad \text{on } \partial \mathcal{D}. \tag{2.2}$$

The flow velocity $\mathbf{u} = -\mathbf{k} \times \nabla \psi$ is specified by the streamfunction

$$\psi(\mathbf{x}, t) = -0.5y + \sin x \sin y + \epsilon \sin (x - ct) \sin 2y \tag{2.3}$$

with the amplitude and phase speed of the second wave taken below to be $\epsilon = 0.6$ and $c = 0.3$ respectively. A geophysical interpretation of the flow associated with (2.3) is that of a steady meandering jet supporting a travelling Rossby wave, hence (2.3) serves as a qualitative model of the flow experienced by fluid parcels on isentropic surfaces in (for example) the extratropical troposphere or stratosphere. At low values of ϵ, complete barriers to transport exist in the flow (2.3), most strikingly at the location of the jet core (region of maximum flow) which divides the channel into two (Haynes *et al.* 2007). At the present value $\epsilon = 0.6$ no formal barrier exists, nevertheless very little fluid is advected across the jet core, i.e. in practice there is a 'leaky' or partial transport barrier at the jet (see also Esler 2008). In the numerical simulations here the diffusivity $\boldsymbol{\kappa}$ is uniform and isotropic, i.e. $\boldsymbol{\kappa} = \kappa I$ where I is the identity matrix and $\kappa = 2 \times 10^{-4}$. With the given non-dimensionalisation κ can be identified with the inverse of the flow Péclet number $Pe = UL/K$, which is more generally defined in terms of the dimensional magnitudes U, L and K of the flow speed, flow length scale and diffusivity respectively. Here $Pe = 5000$. Several integrations have also been performed for other values of κ, including for spatially non-uniform configurations, for which the qualitative results described below have been reproduced.

The particular focus here will be on transport between a small isolated source region $\mathcal{S}$ active at $t = 0$ and a similar receptor region $\mathcal{R}$ at $t = T$. Hence the source in (2.1) will have the form $s(\mathbf{x}, t) = S(\mathbf{x})\delta(t)$. This choice is equivalent to setting $s = 0$ in (2.1) and solving the initial value problem with $c(\mathbf{x}, 0) = S(\mathbf{x})$. However the formalism of (2.1) is retained, as it allows the straightforward extension of all our results to sources with a more general time-dependence. Defining a general 'measurement' to be an integral quantity of the form

$$\mathcal{I} = \int_{-\infty}^{\infty} \int_{\mathcal{D}} c(\mathbf{x}, t) r(\mathbf{x}, t) \, \mathrm{d}\mathbf{x} \, \mathrm{d}t, \tag{2.4}$$

where the *receptor function* $r(\mathbf{x}, t)$ can in general be any integrable function, with the case of the isolated receptor at $t = T$ recovered by setting $r(\mathbf{x}, t) = R(\mathbf{x})\delta(t - T)$.

The stochastic Lagrangian methods described below will be verified against a numerical PDE solution of (2.1). The numerical method is a standard spectral method, exploiting the fact that, taking the Fourier transform of (2.1), the complex amplitudes of each Fourier mode of c satisfy an ordinary differential equation that is coupled only to its near neighbours in wavenumber space. High accuracy is possible once a resolution threshold
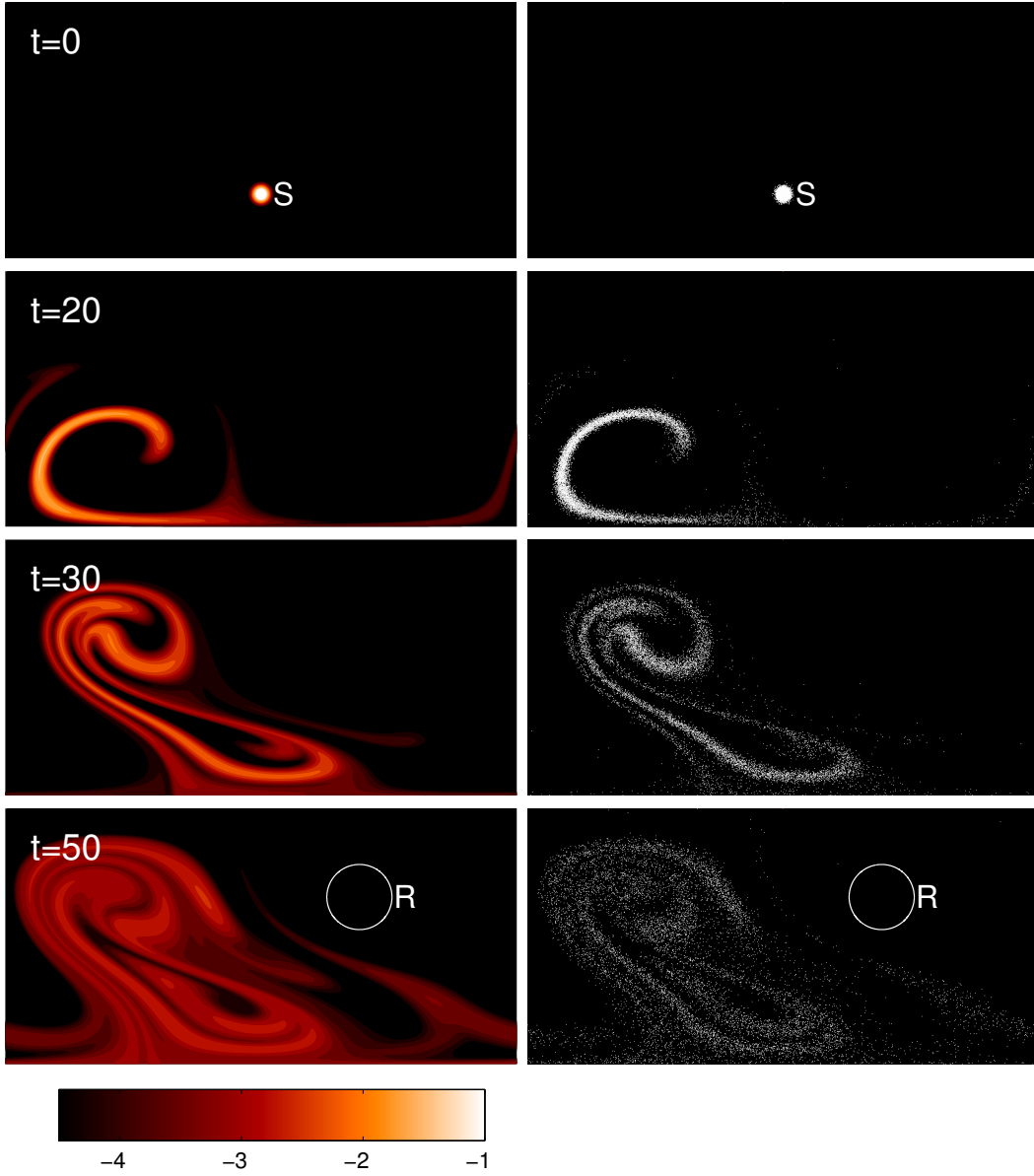
FIGURE 1. Left: Snapshots from the PDE simulation of the forward problem (2.1). The quantity contoured is $\log_{10} c(\mathbf{x}, t)$, for $t = 0, 20, 30, 50$, with contour interval 0.5 (see colour bar). Right: Scatterplot of 10^4 solutions $\mathbf{X}_t$ of the SDE (3.1) at the same times.

is crossed, because the spectral power of c drops off rapidly for wavenumbers $k \gtrsim \kappa^{-1/2}$. The integrations used below, mainly at a resolution of 512×256 wavenumbers, have been verified against a 1024×512 calculation with excellent agreement.

The specific initial condition $S(\mathbf{x}) = \exp\left(-|\mathbf{x} - \mathbf{x}_s|^2/2W_s^2\right)$, centred on $\mathbf{x}_s = (\pi, \ \pi/4)^\dagger$ and with horizontal scale $W_s = 1/20\sqrt{2}$, is applied at $t = 0$. Fig. 1 (left) shows snapshots of the developing PDE solution (note the logarithmic contour interval) at $t = 0, 20, 30$ and

$t = 50 = T$, where T is the receptor time, which corresponds to approximately five cycles of the periodic flow (2.3). The partial barrier to transport is evident in fig. 1, because the tracer is stirred and mixed across the lower left of the domain with little penetrating into the upper right. The receptor function, which is chosen to pick out a typical tracer filament (not visible in the $t = 50$ panel in fig. 1 as its concentration $c \lesssim 10^{-5}$), is taken to be $R(\mathbf{x}) = \exp\left(-|\mathbf{x} - \mathbf{x}_r|^2/2W_r^2\right)$ with $\mathbf{x}_r = (4.35, 2.05)^\dagger$ and $W_r = 1/5\sqrt{2}$ (see the circle in the $t = 50$ panel). Although concentrations in $\mathcal{R}$ are relatively low the quantity $\mathcal{I}$ can be calculated to high accuracy by the PDE method ($\mathcal{I} = 1.522583 \times 10^{-7}$ to 7 sig. figs.). The focus of all that follows is to develop efficient Lagrangian methods that can be used to estimate both $\mathcal{I}$ and the (weighted) distribution of tracer $c(\mathbf{x}, T)R(\mathbf{x})$ throughout the receptor region at the end of the simulation.

2.2. *The adjoint problem*

In the Lagrangian methods to be described below, a central role is taken by the solution, or approximate solution, of the adjoint of (2.1) (sometimes the 'retro-transport equation' e.g. Hourdin & Talagrand 2006), which provides an alternative means of obtaining $\mathcal{I}$. A brief review is warranted. The retro-transport equation is obtained by first expressing (2.1) using the operator formalism,

$$\mathcal{L}c = s, \tag{2.5}$$

where

$$\mathcal{L} \equiv \partial_t + (\mathbf{u} \cdot \nabla) - \nabla \cdot (\boldsymbol{\kappa} \cdot \nabla)$$

is the forward linear transport operator. Next it is necessary to define an inner product for integrable functions $f(\mathbf{x}, t)$, $g(\mathbf{x}, t)$ defined on $\mathcal{D} \times \mathbb{R}$

$$\langle f, g \rangle = \int_{-\infty}^{\infty} \int_{\mathcal{D}} f(\mathbf{x}, t)g(\mathbf{x}, t) \, \mathrm{d}\mathbf{x} \, \mathrm{d}t. \tag{2.6}$$

The inner product (2.6) can be used to define the adjoint $\mathcal{L}^\dagger$ of $\mathcal{L}$, according the usual definition $\langle f, \mathcal{L}^\dagger g \rangle = \langle \mathcal{L}f, g \rangle$ for all admissible f, g. It is a straightforward exercise in integration by parts to show that the adjoint transport operator satisfies

$$\mathcal{L}^\dagger \equiv -\partial_t - (\mathbf{u} \cdot \nabla) - \nabla \cdot (\boldsymbol{\kappa} \cdot \nabla).$$

In terms of the inner product, the integral quantity (2.4) is given by $\mathcal{I} = \langle c, r \rangle$. If we now define the *retro-transport* equation to be

$$\mathcal{L}^\dagger c^* = r, \tag{2.7}$$

then $\mathcal{I}$ can be manipulated as follows

$$\mathcal{I} = \langle c, r \rangle = \langle c, \mathcal{L}^\dagger c^* \rangle = \langle \mathcal{L}c, c^* \rangle = \langle s, c^* \rangle. \tag{2.8}$$

In other words, rather than solving (2.1) to obtain $\mathcal{I}$, an alternative is to solve (2.7) for

c^*, and then obtain $\mathcal{I}$ by evaluating the inner product with s. It is clear from the sign of the diffusion term in $\mathcal{L}^\dagger$ that for (2.7) to remain well-posed, it must be solved *backwards* in time. In the case of our instantaneous receptor at $t = T$ (2.7) takes initial conditions $c^*(\mathbf{x}, T) = R(\mathbf{x})$.

A formal interpretation of the adjoint solution $c^*(\mathbf{x}, t)$ is that it denotes sensitivity of $\mathcal{I}$ with respect to a small change in the source s at location $(\mathbf{x}, t)$, as can be expressed mathematically using the Frechet derivative $c^* = \delta\mathcal{I}/\delta s$. Alternatively, a more physical interpretation is that c^* is a measure the quantity of fluid at a given point that will subsequently arrive at the receptor at time $t = T$, weighted by $R(\mathbf{x})$. Fig. 2 (left) shows snapshots from the PDE solution of (2.7) at times $t = 50, 30, 20$ and 0. The transport barrier is again evident, as c^* remains low in the lower left of the domain. The accuracy of the solution can be checked directly by re-calculating $\mathcal{I} = \langle s, c^* \rangle$. Agreement with the forward calculation ($\mathcal{I} \approx 1.522583 \times 10^{-7}$) is obtained to 11 sig. figs., demonstrating the accuracy of the PDE algorithm.

2.3. *The density of trajectories*

A useful quantity is the product of the solutions of (2.1) and (2.7), i.e. the *density of trajectories* $D(\mathbf{x}, t) = c(\mathbf{x}, t) c^*(\mathbf{x}, t)$. Some intuition as to why D is significant for importance sampling follows if one interprets $c(\mathbf{x}, t)$ as (proportional to) the probability density of a trajectory originating in $\mathcal{S}$ at $t = 0$ arriving at $\mathbf{x}$ at time t. Then $c^*(\mathbf{x}, t)$ is the analogous quantity measuring the probability of the trajectory at $(\mathbf{x}, t)$ ending up in $\mathcal{R}$ at $t = T$. The product $D(\mathbf{x}, t)$ is therefore proportional to the probability of a trajectory passing through $(\mathbf{x}, t)$ as it travels from $\mathcal{S} \to \mathcal{R}$. Intuitively, an optimal Lagrangian algorithm for $\mathcal{S} \to \mathcal{R}$ transport should distribute trajectories in proportion to $D(\mathbf{x}, t)$, as will be demonstrated below.

Fig. 3 (left) shows snapshots of $D(\mathbf{x}, t)$ for our model problem. The $t = 20$ and $t = 30$ panels illustrate at a glance why we have described our problem as 'hard'. The greater part of the $D(\mathbf{x}, t)$ field is seen to occupy around ten or so small and disjoint regions of the domain. An effective adaptive trajectory algorithm must sample each of these disjoint regions efficiently without wasting trajectories elsewhere.

A property of $D(\mathbf{x}, t)$ (e.g. Hourdin & Talagrand 2006) is that, during times when source function $s(\mathbf{x}, t)$ and receptor function $r(\mathbf{x}, t)$ are zero, its domain-integral is con-
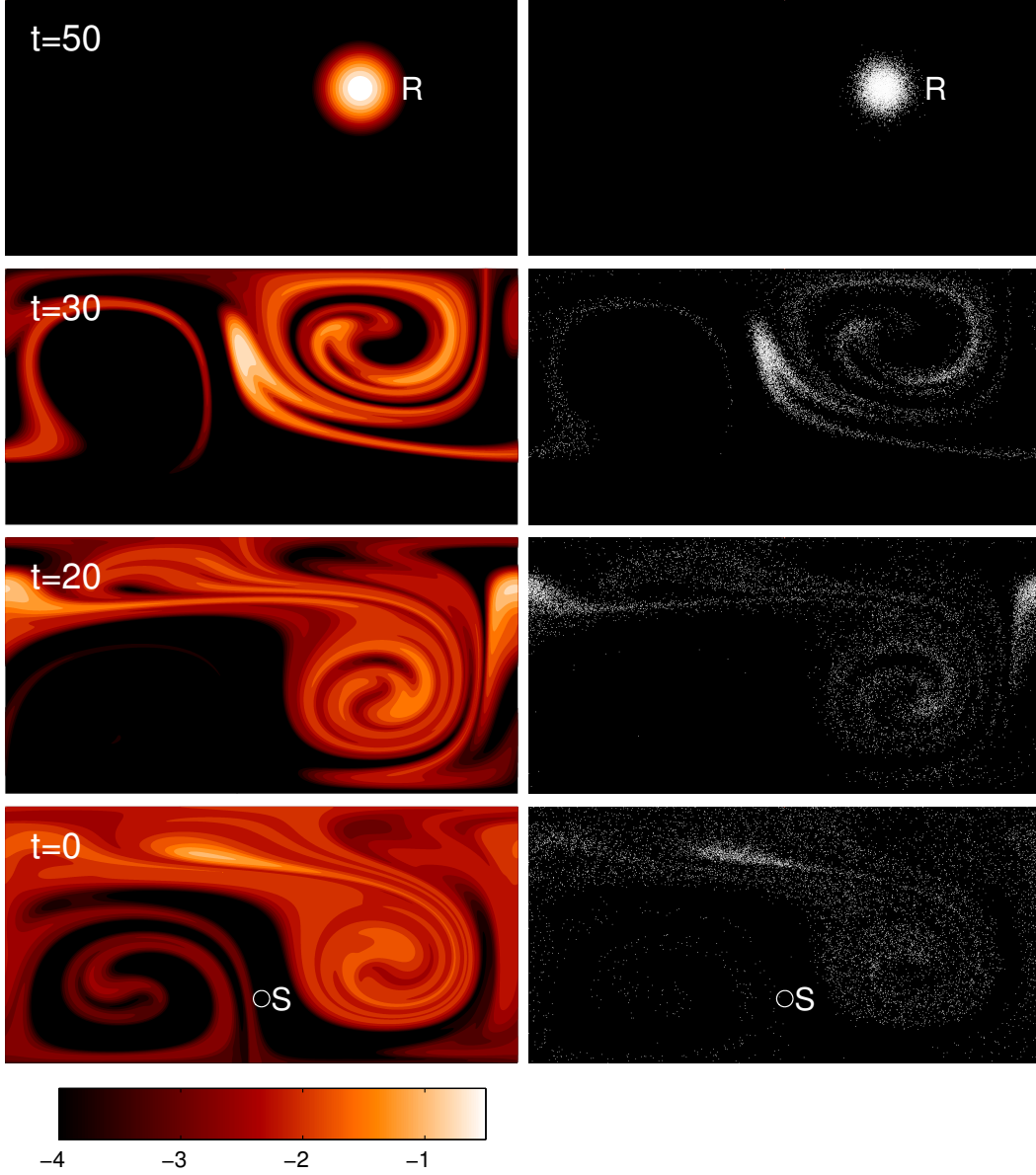
FIGURE 2. Left: Snapshots from the PDE simulation of the forward problem (2.1)). The quantity contoured is $\log_{10} c^*(\mathbf{x}, t)$, for $t = 50, 30, 20, 0$, and with contour interval 0.5. Right: Scatterplot of 10^4 solutions $\mathbf{X}_\tau$ of the SDE given by (3.1) again at $\tau = 0, 20, 30, 50$ (equivalently $t = 50, 30, 20, 0$).

stant. The divergence theorem in two-dimensions can be used to show this as follows:

$$
\frac{d}{dt} \int_{\mathcal{D}} D(\mathbf{x}, t) \, d\mathbf{x} = \int_{\mathcal{D}} \left(c_t(\mathbf{x}, t) c^*(\mathbf{x}, t) + c(\mathbf{x}, t) c_t^*(\mathbf{x}, t) \right) d\mathbf{x} \tag{2.9}
$$

$$
= \int_{\mathcal{D}} \left(\left(-\mathbf{u} \cdot \nabla c + \nabla \cdot (\boldsymbol{\kappa} \cdot \nabla c) \right) c^* + c \left(-\mathbf{u} \cdot \nabla c^* - \nabla \cdot (\boldsymbol{\kappa} \cdot \nabla c^*) \right) \right) d\mathbf{x}
$$

$$
= \int_{\mathcal{D}} \nabla \cdot \left(\mathbf{u} c c^* + c^* \boldsymbol{\kappa} \cdot \nabla c - c \boldsymbol{\kappa} \cdot \nabla c^* \right) d\mathbf{x}
$$

$$
= \int_{\partial \mathcal{D}} \left(\mathbf{u} \cdot \mathbf{n} c c^* + c^* \left(\mathbf{n} \cdot \boldsymbol{\kappa} \cdot \nabla c \right) - c \left(\mathbf{n} \cdot \boldsymbol{\kappa} \cdot \nabla c^* \right) \right) ds = 0,
$$

where $\mathbf{n}$ is the unit normal to $\partial \mathcal{D}$. The final equality follows from no-normal flow at the boundary $\partial \mathcal{D}$, (i.e. $\mathbf{u} \cdot \mathbf{n} = 0$) and the no-flux boundary conditions (2.2) for c and c^* on $\partial \mathcal{D}$.

For our problem it follows that

$$\int_{\mathcal{D}} D(\mathbf{x}, t) \, \mathrm{d}\mathbf{x} = \mathcal{I}, \quad \text{at all times } t \in [0, T]. \tag{2.10}$$

3. Stochastic representations of the advection-diffusion problem

3.1. *Direct and adjoint stochastic representations*

The direct stochastic Lagrangian model corresponding to (2.1) consists of an Itô-type stochastic differential equation (SDE) for the particle position $\mathbf{X}_t$. The SDE in question is chosen in order that the probability density $p(\mathbf{x}, t)$ of the random variable $\mathbf{X}_t$ evolves in time according to a Fokker-Planck equation (sometimes forward Chapman-Kolmogorov equation) that is identical to (2.1). The method for deriving an SDE from a given Fokker-Planck PDE (or vice versa) is standard in stochastic calculus (e.g. section 4.3.4 of Gardiner 2009). In the case of (2.1), the SDE is

$$\mathrm{d}\mathbf{X}_t = \Big(\mathbf{u}(\mathbf{X}_t, t) + \nabla \cdot \boldsymbol{\kappa}(\mathbf{X}_t, t)\Big)\mathrm{d}t + (2\boldsymbol{\kappa})^{1/2}(\mathbf{X}_t, t) \cdot \mathrm{d}\mathbf{W}_t, \quad \mathbf{X}_0 \sim S(\mathbf{x})/S_{00}. \tag{3.1}$$

where $\mathrm{d}\mathbf{W}_t$ are the increments of a two-dimensional Brownian (Wiener) process, and the square root of a symmetric positive-definite tensor $\boldsymbol{\kappa}$ follows the standard definition. Notice that in (3.1) the probability density of the initial particle position $\mathbf{X}_0$ is specified to ensure $p(\mathbf{x}, 0) = S(\mathbf{x})/S_{00}$ (where the normalising constant S_{00} is the domain integral of $S(\mathbf{x})$). The consequence is that at subsequent times the respective distributions are in direct proportion, i.e. $p(\mathbf{x}, t) = c(\mathbf{x}, t)/S_{00}$.

Fig. 1 (right) shows scatterplots of $\mathbf{X}_t$ for an ensemble of $N = 10^4$ realisations of (3.1). Equation (3.1) is solved numerically using a suitable second-order Runge-Kutta time-stepping scheme for time-dependent SDEs (see Tocino & Ardanuy 2002, note that their scheme simplifies considerably when the diffusivity is constant and isotropic). Note that the Tocino-Ardanuy scheme is 'weak' in the sense that it convergences only in probability as step-size is reduced, as opposed to 'strong' in the sense of converging to individual sample paths of the stochastic process. It is weak convergence that is appropriate for the problem in hand. In comparison tests with the PDE solution, using the kernel density method described below, a time-step of $\delta t = 0.01$ was found to be adequate to ensure that the numerical error was significantly less than the statistical error when using $N = 10^6$ particles. It is immediately evident in Fig. 1 that the distribution of $\mathbf{X}_t$ follows that of $c(\mathbf{x}, t)$ as expected. It is equally apparent that exactly zero particles end up in the receptor region $\mathcal{R}$ (based the PDE solution, the probability of a particle arriving in $\mathcal{R}$ at $T = 50$ under (3.1) is about 2×10^{-5}). The integral $\mathcal{I}$ can be estimated from the
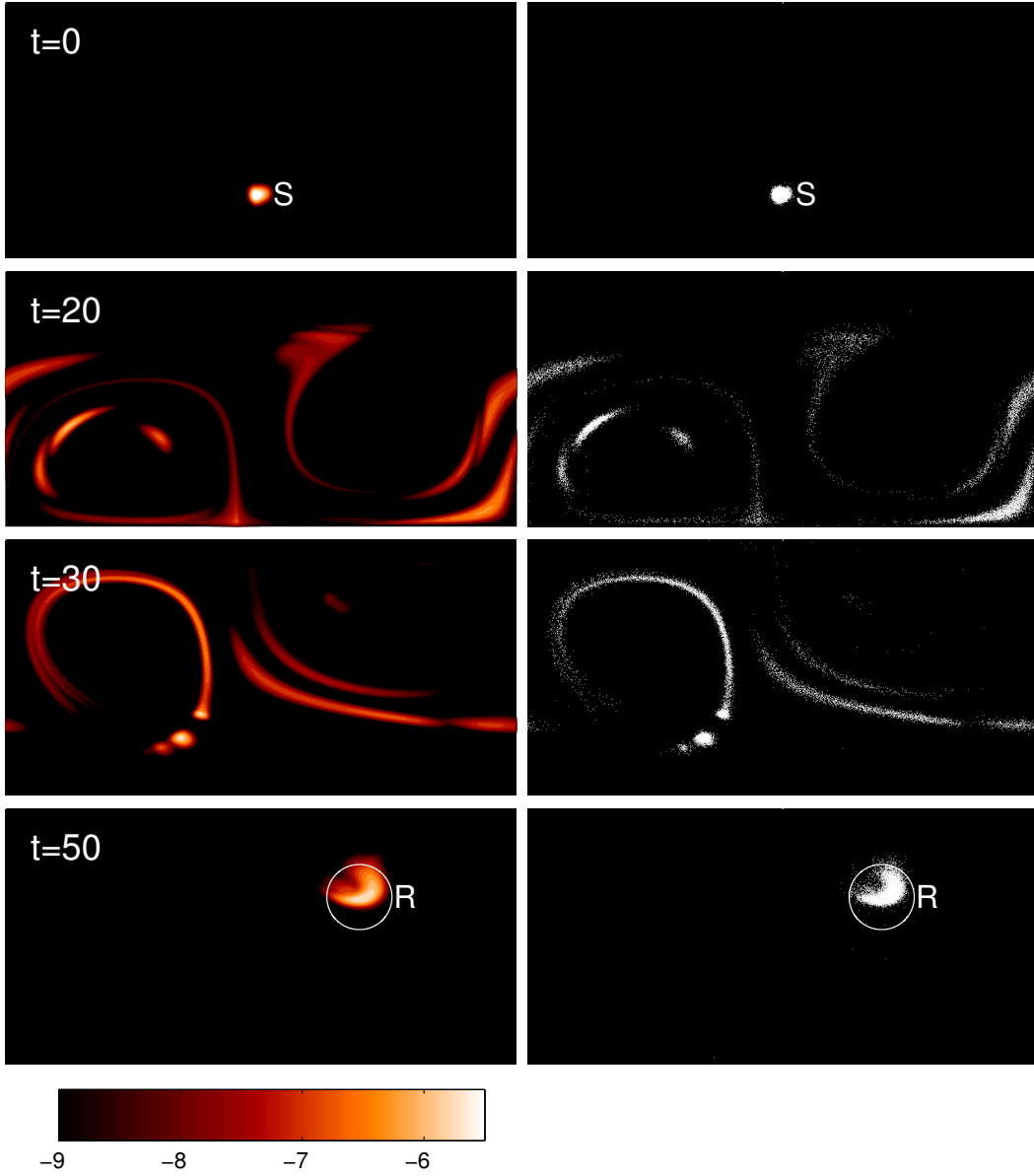
FIGURE 3. Left: Snapshots of the density of trajectories $D(\mathbf{x}, t)$ defined in §2.3 and calculated from the numerical solutions of forward (2.1) and retro (2.7) PDEs. The quantity contoured is $\log_{10}(D(\mathbf{x}, t))$, for $t = 0, 20, 30, 50$, and with contour interval 0.5. Right: Scatterplot of 10^4 solutions $\mathbf{Y}_t$ of the SDE experiment CHEAT also at $t = 0, 20, 30, 50$). CHEAT solves (3.6) with the correcting velocity $\mathbf{v}$ given by (3.16), as interpolated from the numerical solution of (2.7).

stochastic process using

$$\mathcal{I} = S_{00}\, \mathbb{E}(R(\mathbf{X}_T)) = S_{00} \int_{\mathcal{D}} R(\mathbf{x}) p(\mathbf{x}, T)\, \mathrm{d}\mathbf{x} = \langle c, r \rangle. \tag{3.2}$$

Fig. 1 makes clear, however, that efforts to calculate $\mathcal{I}$ by estimating the expectation in

(3.2) will require very large ensembles. Estimators based on (3.2) are described as 'high variance' estimators in the stochastic calculus literature, and efforts to improve them 'variance reduction methods' (e.g. Kloeden & Platen 1992).

The same approach can be taken with the adjoint 'retro-transport' equation (2.7). Its stochastic representation is

$$\mathrm{d}\bar{\mathbf{X}}_\tau = \left(-\mathbf{u}(\bar{\mathbf{X}}_\tau, \tau) + \nabla \cdot \boldsymbol{\kappa}(\bar{\mathbf{X}}_\tau, \tau)\right)\mathrm{d}\tau + (2\boldsymbol{\kappa})^{1/2}(\bar{\mathbf{X}}_\tau, \tau) \cdot \mathrm{d}\mathbf{W}_\tau, \quad \bar{\mathbf{X}}_0 \sim R(\mathbf{x})/R_{00}. \quad (3.3)$$

where $\tau = T - t$ runs backwards in time, and $\mathcal{I}$ is obtained from

$$\mathcal{I} = R_{00}\, \mathbb{E}(S(\bar{\mathbf{X}}_{\tau=T})).$$

Fig. 2 (right) shows scatterplots of solutions of (3.3). The distribution of $\bar{\mathbf{X}}_{\tau=T-t}$ by design matches $c^*(\mathbf{x}, t)$.

Kernel density methods (e.g. Silverman 1986; Wand & Jones 1994) can be used to estimate c and c^* from ensembles (e.g. $\mathbf{X}_t^{(i)}$, $i = 1, ..., N$) of solutions of (3.1) and (3.3) respectively. In terms of a symmetric kernel function $K(\mathbf{x})$ with unit integral and finite variance, suitable estimators are

$$\bar{c}(\mathbf{x}, t) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{\mathbf{x} - \mathbf{X}_t^{(i)}}{h}\right) + \text{'image terms'} \quad (3.4)$$

$$\bar{c}^*(\mathbf{x}, t) = \frac{1}{N^*h} \sum_{i=1}^{N^*} K\left(\frac{\mathbf{x} - \bar{\mathbf{X}}_{\tau=T-t}^{(i)}}{h}\right) + \text{'image terms'}.$$

Here h is a (small) kernel bandwidth, and 'image terms' refer to contributions from the images of trajectories, introduced to satisfy the boundary conditions. Formally, there are an infinite number of images generated by repeated reflection in the boundaries at $y = 0, \pi$, and located at periodic intervals of 2π in x. In practice, however, only the former need be considered and only for particles close to the boundaries. The accuracy of the estimate $\bar{c}^*$ of the adjoint solution will prove important in the construction of the Lagrangian estimators to be developed in sec. 4 below.

It is worth commenting that, if the kernel density estimates (3.4) are used to estimate $\mathcal{I}$ using the density of trajectories formula (2.10), the 'forward-reverse' representation of Milstein *et al.* (2004) is recovered as a straightforward consequence

$$\mathcal{I} \approx \int_\mathcal{D} \bar{c}(\mathbf{x}, t)\bar{c}^*(\mathbf{x}, t)\, \mathrm{d}\mathbf{x} = \frac{1}{NN^*h} \sum_{i=1}^{N} \sum_{j=1}^{N^*} K^{(2)}\left(\frac{\mathbf{X}_t^{(i)} - \bar{\mathbf{X}}_{\tau=T-t}^{(j)}}{h}\right) + \text{'image terms'} \quad (3.5)$$

Here $K^{(2)}$ is a new kernel generated from the self-convolution of the original kernel K (see e.g. pg. 49 of Silverman 1986). In many practical circumstances, particularly where source and receptor regions $\mathcal{S}$ and $\mathcal{R}$ are spatially and temporally localised, the forward-reverse estimator will have lower variance. It is particularly useful when the Green's function of the transport operator $\mathcal{L}$ (in the terminology of probability theory,

the transition density) is sought (corresponding to source $\mathcal{S}$ and receptor $\mathcal{R}$ being points in $\mathcal{D}$, for which case the estimator (3.2) has infinite variance). For non-localised $\mathcal{S}$ and $\mathcal{R}$ the estimator (3.5) is biased, and converges more slowly with ensemble sizes N and N^* compared with the unbiased estimator (3.2). However, because it can have a much lower starting variance, it can often be of practical value, see e.g. the ocean transport modelling study of Spivakovskaya *et al.* (2005).

3.2. *Importance sampling using Milstein's measure transformation method*

The measure transformation method of Milstein (1995) is based on Girsanov's theorem in stochastic calculus (e.g. Øksendal 2007, chap. 8). Usually, Girsanov's theorem is proved using results from measure theory, including the use of the Radon-Nikodym derivative, which may be unfamiliar to some applied mathematicians. Consequently, all of the results used here will be first proved using a relatively elementary, (and to the author's knowledge, novel), PDE-based approach. The starting point of Milstein's method is the extended stochastic process

$$d\mathbf{Y}_t = \Big(\mathbf{u}(\mathbf{Y}_t, t) + \mathbf{v}(\mathbf{Y}_t, t) + \nabla \cdot \boldsymbol{\kappa}(\mathbf{Y}_t, t)\Big)dt + (2\boldsymbol{\kappa})^{1/2}(\mathbf{Y}_t, t) \cdot d\mathbf{W}_t,$$
$$d\Theta_t = -2^{-1/2}\Theta_t \, \mathbf{v}(\mathbf{Y}_t, t) \cdot \boldsymbol{\kappa}^{-1/2}(\mathbf{Y}_t, t) \cdot d\mathbf{W}_t. \tag{3.6}$$

It is to be emphasised that it is the same three-dimensional Brownian increment $d\mathbf{W}_t$ in each equation of (3.6). The initial conditions of (3.6), which must be consistent with the initial condition $c(\mathbf{x}, 0) = S(\mathbf{x})$ of (2.1), are in general given by

$$\mathbf{Y}_0 \sim g(\mathbf{x})S(\mathbf{x})/G_{00}, \quad \Theta_0 = G_{00}g(\mathbf{Y_0})^{-1}, \quad \left(G_{00} = \int_{\mathcal{D}} g(\mathbf{x})S(\mathbf{x}) \, d\mathbf{x}\right). \tag{3.7}$$

In (3.6) $\mathbf{v}(\mathbf{x}, t)$ is a smooth but otherwise arbitrary vector field defined everywhere in $\mathcal{D}$ and similarly in (3.7) $g(\mathbf{x})$ is a smooth function in $\mathcal{D}$. Evidently, the particle trajectories $\mathbf{Y}_t$ in (3.6) differ from those of (3.1) in that they are advected by the corrected velocity field $\mathbf{u} + \mathbf{v}$, instead of by $\mathbf{u}$. There is also an additional equation for a stochastic variable Θ_t coupled to the trajectory equation. The key result obtained from Girsanov's theorem (following e.g. Kloeden & Platen 1992) is that, for any choice of $\mathbf{v}$ and g the process (3.6) satisfies

$$\mathcal{I} = \mathbb{E}\left(\Theta_T R(\mathbf{Y}_T)\right). \tag{3.8}$$

The result (3.8) will be established below using a PDE method. Note that the original direct estimate (3.2) is recovered under the choice $\mathbf{v} = 0$, $g = 1$. Note also that in order to frame our discussion in terms of 'correcting velocities', a form of the result has been given that is valid only when the diffusivity tensor $\boldsymbol{\kappa}$ is everywhere strictly positive definite, and thus invertible. In the more general situation in which one or more eigenvalues of $\boldsymbol{\kappa}$ is sometimes zero, care must be taken to restrict $\mathbf{v}$ to the direction of the eigenvector(s) of $\boldsymbol{\kappa}$ associated with its non-zero eigenvalue(s).

The aim of Milstein's method is to choose $\mathbf{v}$ and g judiciously in order that (3.8) is a 'lower variance' estimator of $\mathcal{I}$ compared with (3.2). The variable Θ_t can be understood as a trajectory 'weight', which is constant in the uncorrected process (3.1), but is here allowed to evolve. Under a particular choice of $\mathbf{v}$, the system (3.6) samples trajectories that, though (possibly) important for the problem of interest, may be highly improbable under the direct process (3.1). The trajectory weight Θ_t is designed to keep track of exactly how improbable the sampled trajectories are, essentially by integrating the relative probabilities of stochastic increments in the corrected versus the uncorrected process.

The result (3.8) can be established as follows. First, consider (3.6) as a *three*-dimensional vector-valued stochastic process, i.e. for $\mathbf{Z}_t = (Y_{1t} \; Y_{2t} \; \Theta_t)^\dagger$. The Fokker-Planck equation for the joint probability density $p(\mathbf{y}, \theta, t)$ for $\mathbf{Z}_t$ is then obtained as

$$p_t + \nabla \cdot ((\mathbf{u} + \mathbf{v})\,p) - \nabla \cdot (\boldsymbol{\kappa} \cdot \nabla p) + \nabla \cdot (\mathbf{v}\,(\theta p)_\theta) - \tfrac{1}{4}\left(\mathbf{v} \cdot \boldsymbol{\kappa}^{-1} \cdot \mathbf{v}\right)\left(\theta^2 p\right)_{\theta\,\theta} = 0, \quad (3.9)$$

where $\nabla \equiv \nabla_{\mathbf{y}}$ acts on the spatial variable $\mathbf{y}$ only. Next consider the integrated quantity

$$P(\mathbf{y}, t) = \int_0^\infty \theta p(\mathbf{y}, \theta, t)\,\mathrm{d}\theta. \qquad (3.10)$$

Multiplying (3.9) by θ and then integrating leads to (following some integration by parts)

$$\partial_t P + \nabla \cdot (\mathbf{u}P) - \nabla \cdot (\boldsymbol{\kappa} \cdot \nabla P) = 0. \qquad (3.11)$$

In other words (since here $\nabla \cdot \mathbf{u} = 0$) P satisfies our original advection-diffusion equation (2.1). Noting that the SDE initial condition (3.7) corresponds to $p(\mathbf{y}, \theta, 0) = \delta(\theta - G_{00}g(\mathbf{y})^{-1})g(\mathbf{y})S(\mathbf{x})/G_{00}$ it follows that the correct initial condition for (3.11), i.e. $P(\mathbf{y}, 0) = S(\mathbf{x})$, also holds. Consequently

$$P(\mathbf{y}, t) = c(\mathbf{y}, t)$$

at subsequent times t. The result (3.8) then follows from the definition of expectation

$$\mathbb{E}(\Theta_T R(\mathbf{Y}_T)) = \int_{\mathcal{D}} \int_0^\infty \theta R(\mathbf{y}) p(\mathbf{y}, \theta, T)\,\mathrm{d}\theta\,\mathrm{d}\mathbf{y} = \int_{\mathcal{D}} R(\mathbf{y}) P(\mathbf{y}, T)\,\mathrm{d}\mathbf{y} = \langle c, r \rangle = \mathcal{I}.$$
$$(3.12)$$

The fact that the PDE analogue of Girsanov's theorem / Milstein's method involves solving a PDE (3.9) in a higher dimensional space than the original, perhaps explains why the approach has not previously been exploited in the context of PDE methods for advection-diffusion problems. The great advantage of the stochastic representation is that the extra dimension can be accommodated at little extra cost.

The key question remains how to choose $\mathbf{v}$ in (3.6) and g in (3.7). This is a more delicate matter than might be supposed, as the following two examples will demonstrate.

3.3. *Milstein's method with a uniform correcting velocity*

Some intuition about the consequences of a poor choice of the correcting velocity $\mathbf{v}$ follows from setting $\mathbf{v} = V\mathbf{i}$, where V is a constant and $\mathbf{i}$ is the unit vector in the x-direction,
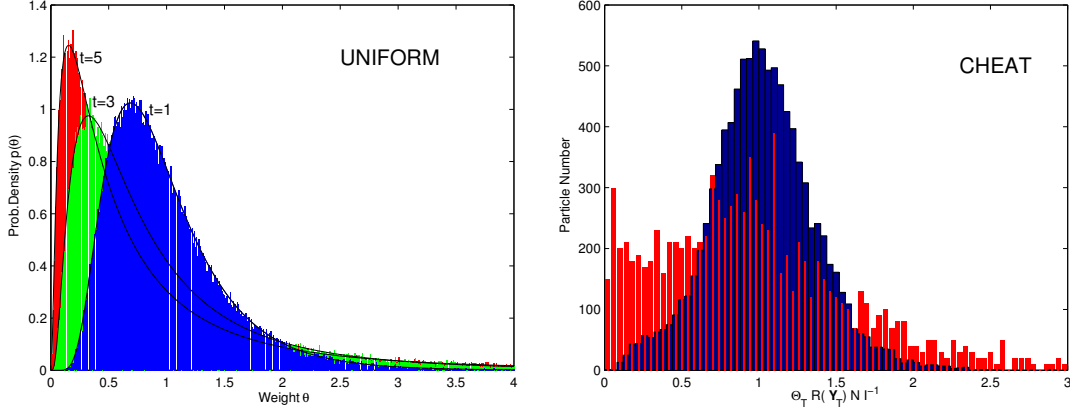
FIGURE 4. Left (UNIFORM): Normalised histograms of trajectory weight Θ_t in the uniform correcting velocity experiment with $V = 0.01$ at times $t = 1$ (blue),3 (green),5 (red), versus the corresponding analytic solutions (curves) given by (3.14). Right (CHEAT): Histograms of $\Theta_T R(\mathbf{Y}_t)/\mathcal{I}$ in the CHEAT experiments described in the text. Blue histogram: interpolation time $\Delta t = 0.1$, red histogram: $\Delta t = 0.2$.

together with $g(\mathbf{x}) = 1$. For the case of the constant and isotropic diffusivity problem under consideration, the trajectory weight equation (3.6) can be integrated explicitly to give

$$\Theta_t = \Theta_0 \exp\left(-\frac{V^2 t}{4\kappa} + \frac{V W_{1t}}{(2\kappa)^{1/2}} \right), \qquad (3.13)$$

where W_{1t} is the x-component of the vector-valued Wiener process $\mathbf{W}_t$. It follows that the random variable Θ_t/Θ_0 has probability density

$$\rho(\theta) = \frac{1}{(\pi t_*)^{1/2}\theta} \exp\left(-\frac{(\log\theta + \frac{t_*}{4})^2}{t_*} \right), \quad t_* = V^2 t/\kappa. \qquad (3.14)$$

The distribution $\rho(\theta)$ has a very long-tail for $t_* \gg 1$. It follows that after sufficient time, the estimator (3.8) will necessarily have very high variance, making it a poor estimator of $\mathcal{I}$. By way of illustration Fig. 4 (left) shows the evolution of the distribution (3.14) at $t_* = 0.5, 1.5, 2.5$. The results are compared to histograms of Θ_t in at corresponding times $t = 1, 3, 5$ in a numerical integration of (3.6), with $V = 0.01$ (recall that $\kappa = 2 \times 10^{-4}$). Even though the correcting velocity $V = 0.01$ is so small that it is difficult to distinguish by eye between corrected and uncorrected particle distributions (not shown), it is obvious that the variance of Θ_t is already alarmingly high after only 10% of the total integration time.

To make the above observations more precise a lower bound can be derived for the variance of the random variable $\Pi \equiv \Theta_T R(\mathbf{Y}_T)$. Exploiting the facts that $\mathbb{E}(\Pi) = \mathcal{I}$ (from 3.12), that $R(\mathbf{x})$ is bounded, and that the distribution of Θ_T is given by (3.14)

with $t_* = V^2 T/\kappa$, it is shown in Appendix A that

$$\mathrm{Var}(\Pi) \geqslant X(t_*, \mathcal{I}) S_{00} \mathcal{I} - \mathcal{I}^2, \tag{3.15}$$

where $X(t_*, \mathcal{I})$ is a function which, for fixed $\mathcal{I} > 0$, can be shown at large t_* to grow exponentially. For problems with fixed κ and T, as here, the inequality (3.15) allows an upper bound for V to be calculated, beyond which $\mathrm{Var}(\Pi)$ will certainly exceed the variance of the original estimator $\mathrm{Var}(S_{00} R(\mathbf{X}_T))$. For the problem at hand (with $T = 50$, $\kappa = 2 \times 10^{-4}$, $\mathcal{I} \approx 1.52 \times 10^{-7}$, $S_{00} \approx 7.85 \times 10^{-3}$) the bound obtained is rather low, $|V| < 0.03$. As above, a correcting velocity V of this magnitude has a very modest impact on the distribution of particles, making it clear that a naive application of Milstein's method with a uniform correcting velocity is worse than useless for our problem.

The broader significance of this result is that it gives an explicit demonstration that, unless $\mathbf{v}$ is chosen with great care, the variance of Θ_t will grow and overwhelm any possible reduction in variance obtained from directing more particles into the receptor region $\mathcal{R}$. In fact, as will be demonstrated next for the problem under consideration, great care is necessary in the choice of $\mathbf{v}$ to result in any variance reduction at all.

3.4. *Milstein's method with the optimal correcting velocity*

Kloeden & Platen (1992) comment that, in principle, a 'zero variance' estimator for $\mathcal{I}$, i.e. an optimal importance sampling strategy, can be constructed by setting $\mathbf{v} = \mathbf{v}^*$ where $\mathbf{v}^*$ is an ideal correcting velocity given by

$$\mathbf{v}^* = 2c^{*-1} \boldsymbol{\kappa} \cdot \nabla c^*. \tag{3.16}$$

Here c^* is the solution of the retro-transport (adjoint) equation (2.7). The standard proof for this is given in Appendix B, where it is shown that the corresponding optimal choice in (3.7) is $g(\mathbf{x}) = c^*(\mathbf{x}, 0)$. Of course, the expression (3.16) cannot be used directly to construct a purely Lagrangian estimator, because it requires knowledge of c^*, from which $\mathcal{I}$ can in any case be obtained directly using (2.8). Nevertheless, investigating solutions of the SDE system (3.6) with $\mathbf{v} = \mathbf{v}^*$ is instructive, both in providing intuition about how the 'perfect' importance sampling Lagrangian solution behaves, and as a test of the numerical implementation of (3.6).

Some results regarding the optimal importance sampling strategy follow from the observation that when $\mathbf{v} = \mathbf{v}^*$ and $g(\mathbf{x}) = c^*(\mathbf{x}, 0)$, the Fokker-Planck equation (3.9) has the exact solution

$$p(\mathbf{y}, \theta, t) = \delta\left(\theta - \mathcal{I} c^*(\mathbf{y}, t)^{-1}\right) c^*(\mathbf{y}, t) c(\mathbf{y}, t) \mathcal{I}^{-1}. \tag{3.17}$$

The result (3.17) can be verified by the straightforward, if tedious, process of insertion in (3.9), provided that care is taken over the properties of generalised functions, i.e. the

Dirac delta and its derivatives. That (3.17) is correctly normalised follows from (2.10). The solution (3.17) reveals the following:

(*a*) In the optimal importance sampling strategy, the distribution of particles in the domain is given by

$$\bar{p}(\mathbf{y}, t) = \int_0^\infty p(\mathbf{y}, \theta, t) \, \mathrm{d}\theta = D(\mathbf{y}, t) \, \mathcal{I}^{-1}. \tag{3.18}$$

In other words, Lagrangian particles are distributed according to the density of trajectories defined in section 2.3. (Note that this result can also be obtained, avoiding the use of generalised functions, by first integrating (3.9) with respect to θ to obtain an equation for $\bar{p}$.)

(*b*) In the optimal importance sampling strategy, the trajectory weights are everywhere inversely proportional to the local value of the adjoint solution, i.e. $\Theta_t \sim c^*(\mathbf{Y}_t, t)^{-1}$.

The results (*a*) and (*b*) show that the optimal strategy for Milstein's method, i.e. with $\mathbf{v} = \mathbf{v}^*$, is exactly consistent with the intuitive description of the importance of D given in §2.3. This intuition will be the basis of the purely Lagrangian strategies developed to estimate $\mathcal{I}$ in §4 below, including the means to choose 'winners' in the GWTW algorithm to be discussed next.

Equation (3.6) has been solved numerically using a correcting velocity $\mathbf{v} = \mathbf{v}^*$ obtained from numerical interpolation of the Eulerian solution of (2.7) described in §2.2 and shown in Fig. 2. The aim of this numerical experiment, named 'CHEAT' because the Eulerian solution that is being exploited can obviously be used to obtain $\mathcal{I}$ independently, is to check the results above and to gain some insight into their practical limitations. In the main experiment, cubic interpolation is used from the 1024×512 gridded Eulerian solution, with linear interpolation in time between outputs at intervals $\Delta t = 0.1$.

The particle positions for an ensemble of $N = 10^4$ realisations of CHEAT are shown in Fig. 3 (right). It is clear from Fig. 3 that the particles distribution in CHEAT closely follows D as predicted by (3.18). The trajectory weights, however, deviate from the prediction (*b*). Fig. 4 (right) shows a histogram (blue) of $\Theta_T R(\mathbf{Y}_T)/\mathcal{I}$ for CHEAT. If $\mathbf{v}^*$ were exact (and the numerics were perfect) then $\Theta_T R(\mathbf{Y}_T)/\mathcal{I}$ would be equal to unity. It is seen, however, to deviate substantially from this deterministic ideal. The main source of error is found to come from the linear interpolation in time between the calculated $\mathbf{v}^*$ fields which are stored at discrete time intervals $\Delta t = 0.1$. Evidence comes from a test in which Δt is increased to 0.2 (red histogram) which shows a large increase in the standard deviation of $\Theta_T R(\mathbf{Y}_T)/\mathcal{I}$.

The sensitivity of the CHEAT results to Δt is illustrative of a major practical difficulty with Milstein's method. Even relatively small errors in $\mathbf{v}^*$ cumulatively lead to large increases in the variance of the Lagrangian estimator. This finding influences our approach in section 4 below.

3.5. *Grassberger's 'go-with-the-winners' (GWTW) branching process*

An alternative to Milstein's method is the GWTW importance sampling strategy (Grassberger 1997, 2002). GWTW is a branching process which in its simplest form is applied at discrete intervals Δt in time. Trajectories are first assigned an initial position and weight Θ_0, according to (3.7), exactly as for Milstein's method. An ensemble of N realisations of the direct SDE (3.1) is then integrated forwards from $t = 0$, with each member of the ensemble having weight Θ_0. However in contrast to Milstein's method, in GWTW trajectory weights Θ_t are updated only at discrete times $t = t_j$ ($t_j = j\Delta t$), instead of continuously in time. Nevertheless the weight Θ_t performs exactly the same function for each algorithm.

At each branching time $t_j = j\,\Delta t$ ($j = 1, 2, 3...$), our implementation of the algorithm proceeds as follows:

(a) A scoring algorithm (see below) is used to assign each trajectory in the ensemble a score $S_{t_j}^{(i)}$ ($i = 1, ..., N$). The trajectories are then sorted into order of decreasing score.
(b) The trajectories with the lowest scores are then considered sequentially. The trajectory with the mth lowest score is designated as a 'loser' if its score is less than $1/3$ of the score of the trajectory with the mth highest score. The total number of losers found is denoted L.
(c) The weights of the L loser trajectories are then randomly, with probability one-half, either doubled $\Theta_{t_j} \to 2\Theta_{t_j}$ or set to zero. The $W (\approx L/2)$ trajectories with score zero are then removed from the calculation.
(d) The W trajectories with the highest scores, designated as the 'winners', are then split into two trajectories, each with half the weight of its parent, i.e. $\Theta_{t_j}/2$.

The steps *(a-d)* act to reduce the variance of the distribution of scores $\{S_{t_j}^{(i)},\ i = 1, ..., N\}$ at each branching time. This variance reduction occurs because any trajectory with a score less than around half of the ensemble mean score will be identified as a loser, and will either have its score doubled or will be removed completely. Any trajectory with a score much greater than 2-3 times the ensemble mean will likely be identified as a winner, and will be reproduced with each daughter trajectory having half its score. Provided that steps *(a-d)* are implemented sufficiently frequently in time, the net effect is to keep all scores within a factor of two or so of the ensemble mean. Crucially, this is achieved without either changing the number of trajectories, or by changing the overall expectation of the process.

The effectiveness of GWTW evidently depends upon the selection of an appropriate score S_{t_j} for each trajectory, together with the choice of $g(\mathbf{x})$ in the initial condition (3.7). There is a clear analogy between choosing S_{t_j} and choosing the correcting velocity $\mathbf{v}$ in Milstein's method. Just as with Milstein's method, we have found that a poor choice leads
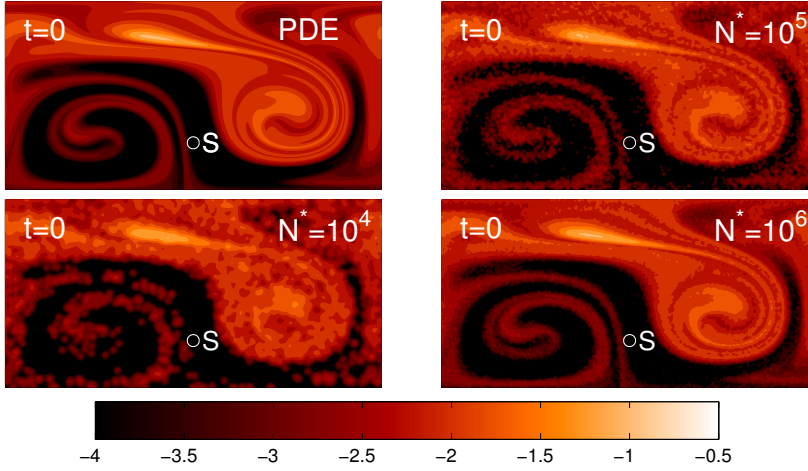
FIGURE 5. Solution $c^*(\mathbf{x}, 0)$ of the retro-transport equation (2.7) versus kernel density reconstructions $\bar{c}^*(\mathbf{x}, 0)$ from ensembles of solutions of (3.3) with $N = 10^4$, 10^5 and 10^6 trajectories respectively. The Gaussian kernel widths $h = 0.015$, $h = 0.02$ and $h = 0.04$ are used, which are close to the optimum value in each case.

to poor performance. Based on the above, the ideal choice can be guided by result (b) in §3.4, which gives the trajectory weights in the 'perfect' implementation of Milstein's method. If the *ideal* score at time t is taken to be the product of the trajectory weight and the local value of the adjoint solution, i.e.

$$S_t = \Theta_t \, c^*(\mathbf{X}_t, t), \tag{3.19}$$

then, based on the arguments given above, the GWTW algorithm will ensure that the distribution of values of S_t will remain within a factor of two or so of the ensemble mean $\langle S_t \rangle$. Hence the trajectory weights will satisfy $\frac{1}{2} \langle S_t \rangle c^*(\mathbf{X}_t, t)^{-1} \lesssim \Theta_t \lesssim 2 \langle S_t \rangle c^*(\mathbf{X}_t, t)^{-1}$, i.e. at any instant in time Θ_t is approximately inversely proportional to c^*, as it is in in the ideal implementation of Milstein's method. Of course, as with Milstein's method the exact field c^* is not available, but the result (3.19) will nevertheless guide our practical implementation of GWTW described in §4.2 below.

4. Adaptive Lagrangian estimators for $\mathcal{S} \to \mathcal{R}$ transport

Motivated by the above analysis, we proceed by considering forward estimators of $\mathcal{I}$ in which importance sampling is controlled by an estimate of the adjoint solution c^*, which is generated from a preliminary ensemble of integrations of the retro-transport SDE (3.3). Much depends on the quality of the estimate $\bar{c}^*$ of c^*, of which some illustrative examples generated using the kernel density method, are shown in Fig. 5.

Alternative approaches are possible. For example Vanden-Eijnden & Weare (2012) describe how results from large deviation theory can be used to obtain an estimate of $\mathbf{v}^*$

for Milstein's method that is asymptotically exact in the limit of small noise. To do this, it is necessary solve a variational problem for each trajectory at regular time intervals, with this problem being posed on the time interval $[t, T]$. Our efforts to use this method for our model problem have not been successful for reasons to be given below.

The restriction here to forward estimators is simply to focus the discussion. Indeed, because both forward and backwards trajectories are used, each of our approaches can be straightforwardly adapted to exploit the forward-reverse representation of Milstein *et al.* (2004) (see eqn. 3.5 above), which may lead to significant further improvements, particularly for very localised sources and receptors.

4.1. *Estimators based on Milstein's measure transformation method*

A purely Lagrangian forward estimator for $\mathcal{I}$ can be created using the the following four-stage algorithm:

(a) Generate an ensemble, of size N^*, of solutions of the retro-transport SDE (3.3).

(b) Use the kernel density method (3.4) to generate a gridded estimates $\bar{c}^*$ at time intervals Δt, of the retro-transport solution c^* of (2.7), using the back trajectory ensemble.

(c) Use the field $\bar{c}^*$ to generate a correcting velocity $\bar{\mathbf{v}}^*$ based on (3.16). This is multiplied by a smoothed step-function $F_\epsilon(\cdot)$ designed to smoothly reduce $\bar{\mathbf{v}}^*$ to near zero in regions where $\bar{c}^*$ is low, i.e.

$$\bar{\mathbf{v}}^* = 2F_\epsilon(\bar{c}^*)\,\bar{c}^{*-1}\boldsymbol{\kappa} \cdot \nabla\bar{c}^*, \tag{4.1}$$

Here $F_\epsilon(x)$ approaches unity for $x \gg \epsilon$ and decays monotonically and rapidly to zero for $x \ll \epsilon$ (below $F_\epsilon(x) = \exp\left(-8(\log x / \log \epsilon)^{12}\right)$ is used).

(d) Solve (3.6) using correcting velocity $\mathbf{v} = \bar{\mathbf{v}}^*$ from (4.1), using initial conditions (3.7), where the function $g(\mathbf{x}) = \epsilon + \bar{c}^*(\mathbf{x}, 0)$. This particular choice for $g(\mathbf{x})$ is consistent with trusting the estimate $\bar{c}^*$ to be a reasonable approximation for c^* only in regions where $\bar{c}^* \gtrsim \epsilon$.

The algorithm above requires the selection of a large number of numerical parameters (e.g. retro-ensemble size N^*, kernel size h, grid-spacing Δx, time interval Δt etc.), in addition to those associated with the SDE solutions themselves (e.g. timestep), and the choice of the function $F_\epsilon(\cdot)$. A complete numerical analysis and optimisation is therefore a formidable task, which we do not attempt here. Instead we focus on the key, and not immediately transparent, role of the parameter ϵ. The reason that a finite ϵ is required is that Milstein's method will fail whenever the trajectory passes through regions where the estimate $\bar{\mathbf{v}}^*$ differs significantly from the ideal $\mathbf{v}^*$. The failure can be explained by analogy with the 'uniform correcting velocity' problem of §3.3, which showed that the increase in $\mathrm{Var}(\Theta_t)$ due to a poorly directed correcting velocity can rapidly overwhelm any variance reduction due to redirecting trajectories into the region of interest.
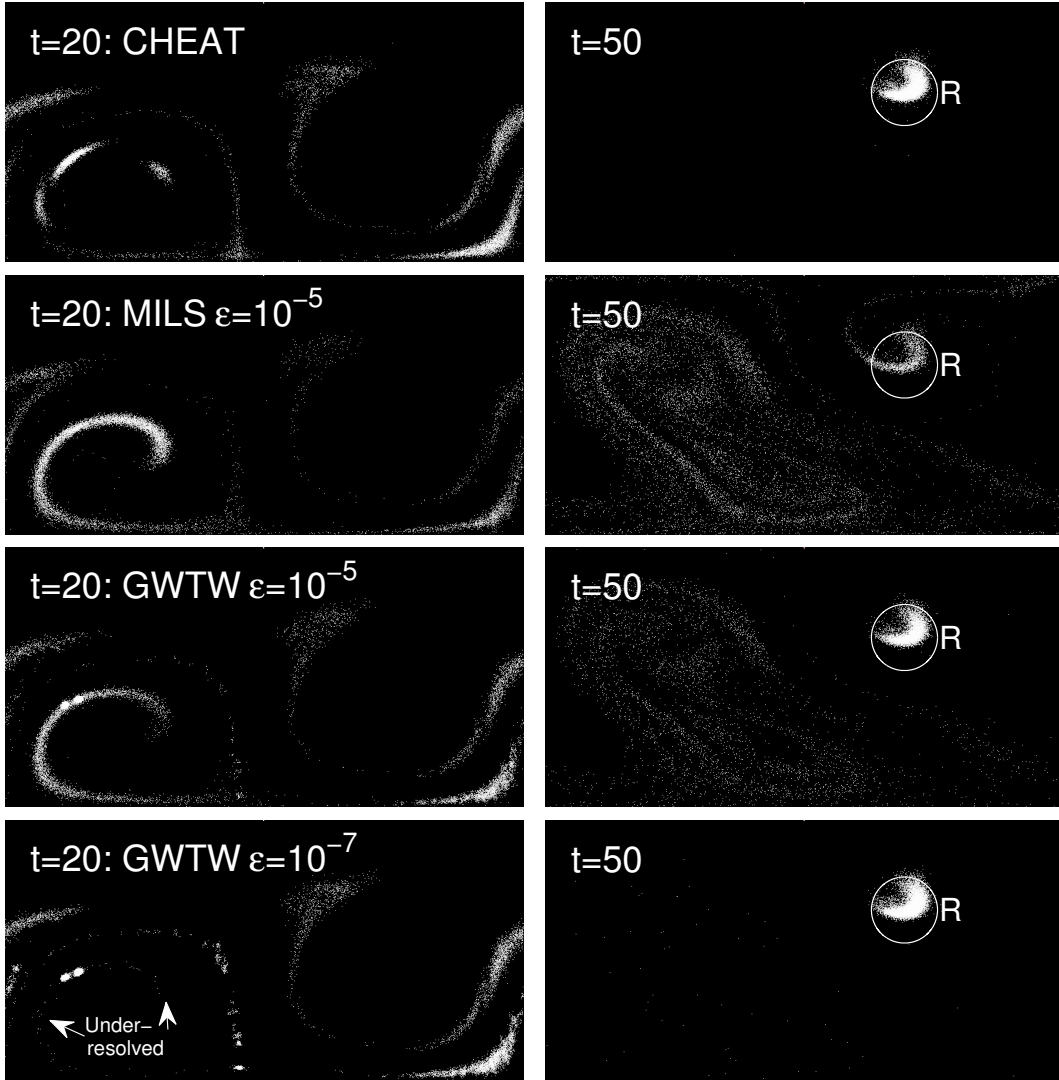
FIGURE 6. Trajectory distributions at $t = 20$ (left panels) and $t = 50$ (right panels) in (top row) the experiment CHEAT described in §3.4, in (second row) Milstein's method with $N^* = 10^5$ and $\epsilon = 10^{-5}$, in (third row) GWTW with $N^* = 10^5$ and $\epsilon = 10^{-5}$ and (bottom row) GWTW with $N^* = 10^5$ and $\epsilon = 10^{-7}$.

In our problem the estimate $\bar{\mathbf{v}}^*$ is poor where there are few adjoint trajectories contributing to $\bar{c}^*$, i.e. in all regions where $\bar{c}^*$ is small, and it turns out to be effective to set the correcting velocity $\mathbf{v}$ to be zero in these regions. It seems that such a fix will be necessary regardless of the method used to estimate $\bar{c}^*$, because the relative error in $\bar{c}^*$, which controls $\bar{\mathbf{v}}^*$, will be inevitably be higher where $\bar{c}^*$ is low. The smoothed step-function in (4.1) acts to suppress $\bar{\mathbf{v}}^*$ everywhere that $\bar{c}^* \lesssim \epsilon$. Broadly speaking, it
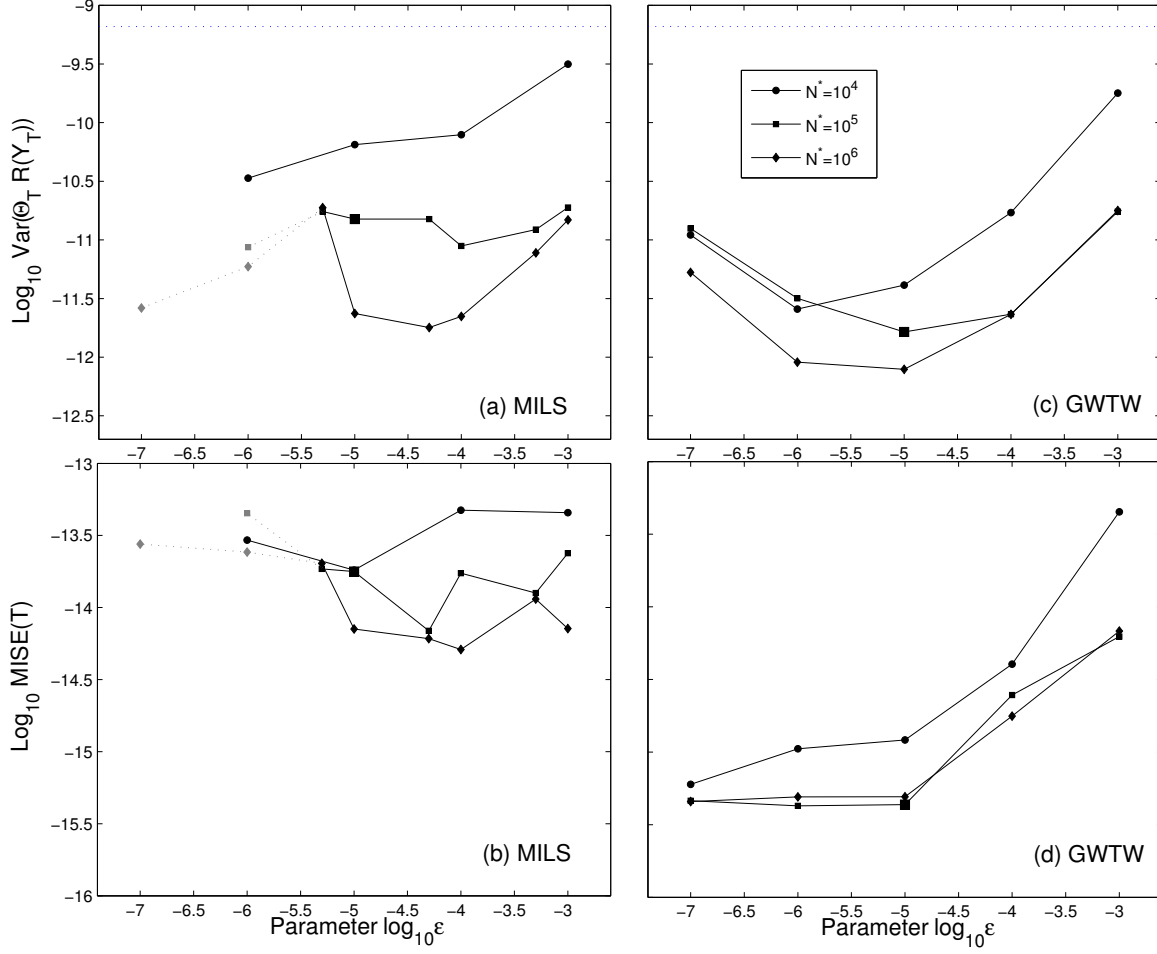
FIGURE 7. Upper panels (a,c): Calculated variance of the estimator $\Theta_T R(\mathbf{Y}_T)$, as a function of ϵ, for calculations using estimates of $\bar{c}^*$ based on $N^* = 10^4, 10^5$ and 10^6 trajectories. Points for which the calculated variance is inconsistent with the error in the sample mean are marked grey. (a) Milstein's method. (c) Go-with-the-winners. The variance of the direct estimator is indicated by the dotted line in each panel. Lower panels (b,d): As above, but showing MISE(T) (eqn. 4.2) for the same calculations.

turns out to be important to choose ϵ to be sufficiently large so that the estimate (4.1) is accurate throughout the region with $c^* \gtrsim \epsilon$. It will be seen below that the algorithm is not sensitive to the precise value of ϵ, with near optimal results obtained provided ϵ is within an order of magnitude or so of its optimal value.

Fig. 6 shows the trajectory positions at $t = 20$ and $t = 50$ in the CHEAT experiment (top row) and in Milstein's method with $N^* = 10^5$ and $\epsilon = 10^{-5}$. In the CHEAT experiment, trajectories are distributed almost exactly according to $D(\mathbf{x}, t)$, so CHEAT serves here as a reference which allows us to determine easily where the other methods

are over- and under-resolved in terms of their distribution of particles. It can be seen in the $t = 20$ panel that the Milstein's method integration with $\epsilon = 10^{-5}$ maintains some resolution in all relevant regions, but is over-resolved where $c(\mathbf{x}, t)$ is relatively large (compare Fig. 1) and somewhat under-resolved elsewhere. The distributions of trajectories for runs with other values of ϵ (not shown) are similar, but with fewer particles reaching $\mathcal{R}$ as ϵ is increased. It appears from Fig. 6 that the errors in Milstein's method do not primarily stem from a failure to simulate some of the $\mathcal{S} \to \mathcal{R}$ transport pathways, but from the excess variance in Θ_T introduced due to errors in $\mathbf{v}^*$, as discussed in section 3.3-3.4 above. Because $\text{Var}(\Theta_T)$ decreases as ϵ is increased, there is an optimal value of ϵ for which sufficient trajectories reach $\mathcal{R}$ with $\text{Var}(\Theta_T)$ not too large, and this optimal value depends on the details of the problem.

To quantify the influence of the quality of the estimate of $\bar{c}^*$ and the selection of ϵ, Fig. 7a shows calculated variances of the Lagrangian estimators, i.e. $\text{Var}(\Theta_T R(\mathbf{Y}_T))$. The variance of the direct estimator (i.e. $\text{Var}(S_{00} R(\mathbf{X}_T))$), calculated directly from the PDE solution, is marked by the dotted line. The results are plotted as a function of ϵ, and are given for the three kernel density reconstructions of $\bar{c}^*$, for $N^* = 10^4$, 10^5 and 10^6, as shown in Fig. 5. The variance in each case is calculated from a forward ensemble of size $N = 10^5$, using the standard statistical estimator. A difficulty is that the random variable $\Theta_T R(\mathbf{Y}_T)$ becomes long-tailed for calculations with $\epsilon = 10^{-6}$ or smaller, and in these case the values obtained for the variance are likely to be underestimates due to poor sampling of the tail. Points for which the error in the sample mean, compared to the expected value $\mathcal{I}$, is greater than the calculated value of $3(\text{Var}(\Theta_T R(\mathbf{Y}_T))/N)^{1/2}$ (i.e. three standard deviations), are highlighted as being inconsistent and are consequently marked in grey. If these points are ignored, the results show that between one and two orders of magnitude of variance reduction is possible, and confirm that performance improves with increasing N^* as is expected. The optimal value for ϵ is seen to be in the range $10^{-4} - 10^{-5}$.

An alternative measure of the accuracy of the calculations is given by the mean integrated square error (MISE) of the reconstructions of the tracer field within $\mathcal{R}$. Here MISE is defined to be

$$\text{MISE}(t) = \int_{\mathcal{D}} \left(c(\mathbf{x}, t) - \bar{c}(\mathbf{x}, t) \right)^2 R(\mathbf{x})^2 \, d\mathbf{x}, \tag{4.2}$$

where the reconstructions are obtained from the *weighted* kernel density formula (c.f. eqn. 3.4)

$$\bar{c}(\mathbf{x}, t) = \frac{1}{Nh} \sum_{i=1}^{N} \Theta_t^{(i)} K\left(\frac{\mathbf{x} - \mathbf{Y}_t^{(i)}}{h} \right) + \text{'image terms'}. \tag{4.3}$$

Contour plots of $c(\mathbf{x}, T)R(\mathbf{x})$ from the exact (PDE) solution, together with kernel density reconstructions $\bar{c}(\mathbf{x}, T)R(\mathbf{x})$ from the Milstein method (MILS) with $N = 10^5$, $N^* = 10^5$
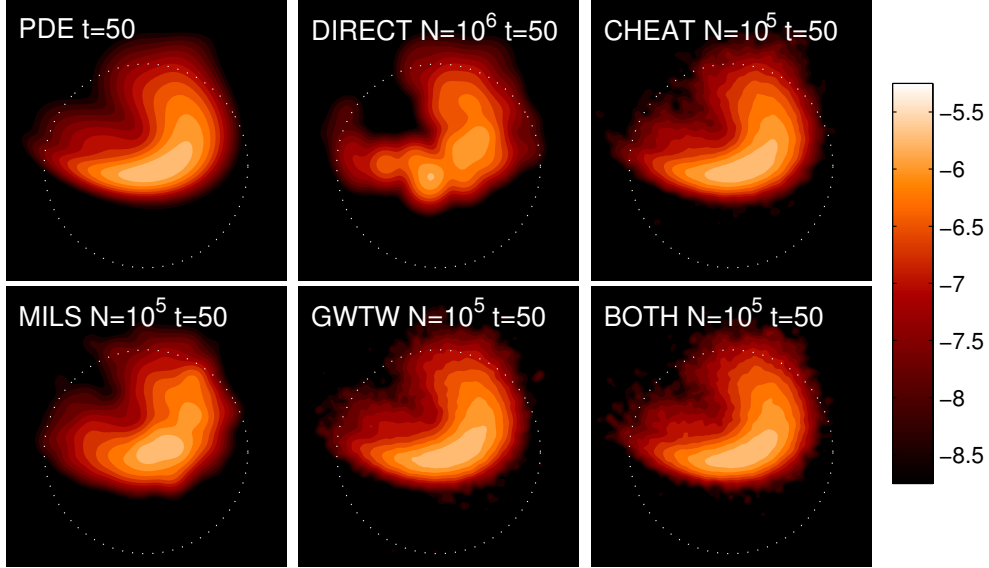
FIGURE 8. (Top left) Weighted PDE (Eulerian) solution $\log_{10} c(\mathbf{x}, T)R(\mathbf{x})$ in the vicinity of the receptor region $\mathcal{R}$. (Remaining panels) Kernel density estimates of $\log_{10} c(\mathbf{x}, T)R(\mathbf{x})$ from the stochastic Lagrangian integrations: DIRECT with $N = 10^6$, CHEAT with $N = 10^5$, MILSTEIN with $N = 10^5$ ($N^* = 10^5$, $\epsilon = 10^{-5}$), GWTW with $N = 10^5$ ($N^* = 10^5$, $\epsilon = 10^{-5}$) and finally an experiment combining both methods (BOTH) also with $N = 10^5$.

and $\epsilon = 1 \times 10^{-5}$, are shown in Fig. 8 (optimal kernel sizes are used). The results are clearly superior to the direct calculation with $N = 10^6$, for which only around 20 trajectories arrive in $\mathcal{R}$. Fig. 7b shows MISE(T) for the same set of simulations as in Fig. 7a, confirming that performance improves with increasing N^* and that $\epsilon = 10^{-4} - 10^{-5}$.

A further disadvantage of the implementation *(a-d)* of Milstein's method is the expense of calculating and storing the gridded correcting velocity $\bar{\mathbf{v}}^*$, as well as interpolating it during the forward calculations. As is clear from Fig. 4b, $\bar{\mathbf{v}}^*$ must be obtained and stored at a relatively high temporal resolution (here $\Delta t = 0.1$). The overall cost of this increases the expense of our forward calculations by around a factor of ten compared to the direct calculations, significantly reducing the gains made by variance reduction. Fortunately, better results are obtained with the GWTW algorithm, as will be seen next.

4.2. *Estimators based on 'go-with-the-winners' (GWTW)*

To create a purely Lagrangian estimator based on GWTW, steps *(a-b)* are followed as above, followed by:

(c 2) A forward run with GWTW scored according to

$$S_t = \Theta_t \left(\bar{c}^*(\mathbf{X}_t, t) + \epsilon \right). \tag{4.4}$$

In close analogy with Milstein's method, introducing an $\epsilon > 0$ acts effectively to switch off GWTW in regions where $\bar{c}^* \ll \epsilon$, in order that particles are not removed and added erroneously in regions where the estimate $\bar{c}^*$ of c^* is inaccurate.

The importance of choosing ϵ to be sufficiently large is seen in the lower two rows of Fig. 6, which shows the trajectory positions in GWTW runs with $N^* = 10^5$ and $\epsilon = 10^{-5}$ (third row) and $\epsilon = 10^{-7}$ (bottom row). For the $\epsilon = 10^{-7}$ GWTW run, from the $t = 20$ panel it is clear that some $\mathcal{S} \to \mathcal{R}$ transport pathways are significantly under-resolved, mainly in regions where $c(\mathbf{x}, t)$ is large (e.g. regions highlighted by arrows). The under-resolution is due to trajectories being removed too rapidly from regions where c^* is low, because in these regions there are no back trajectories, meaning that $\bar{c}^*$ is effectively zero and trajectory scores are too low. Elsewhere, there is evidence of erroneous trajectory clustering on the scale of the kernel size h, which is due to trajectories being reproduced close to local maxima in $\bar{c}^*$ associated with individual back trajectories. The consequence of the poor sampling of $\mathcal{S} \to \mathcal{R}$ transport pathways in the $\epsilon = 10^{-7}$ simulation is that, despite the spatial distribution of particles at $t = 50$ appearing to be reasonable, there is increased variance in statistics of the trajectory weights at $t = 50$, consistent with a relatively poor estimate of $\mathcal{I}$. Choosing $\epsilon = 10^{-5}$ improves the sampling of $\mathcal{S} \to \mathcal{R}$ pathways, but at the expense of over-resolving where c is large, and consequently a lower proportion of trajectories reaching $\mathcal{R}$ at $t = 50$. As with Milstein's method, there is evidently a trade-off between competing effects, and an optimal value of ϵ for the algorithm.

The performance of GWTW is quantified as a function of ϵ and N^* in Fig. 7c and d. It is clear that our implementation of GWTW performs substantially better than Milstein's method, with robust behaviour over a wide range of ϵ centred on the optimal value of around 10^{-5}. For runs with $N^* = 10^5$ the variance is reduced by 2.5 orders of magnitude, increasing to 3 orders of magnitude for $N^* = 10^6$. An even greater improvement over Milstein's method is evident for MISE(T), which is also apparent in the kernel density reconstructions of Fig. 8.

GWTW is also much cheaper to implement than Milstein's method. Interpolation of $\bar{c}^*$ and sorting of winners and losers happens only at longer intervals $\Delta t = 0.1$, whereas $\bar{v}^*$ in Milstein's method must be obtained by interpolation at every SDE time-step $\delta t = 0.01$ in order to integrate the trajectory weight equation. As a result GWTW adds only 10-20% to the cost of a forward run, making it remarkably cheap as well as robust. For our model problem, the computational cost of integrating $N = 10^5$ trajectories is comparable to that of using the kernel density method to make 500 constructions of $\bar{c}^*$ from the results, hence the overall cost of the algorithm (backwards run, construction of $\bar{c}^*$, forwards run) is around three times that of the direct method. In our model problem, fixing a level

of accuracy required for $\mathcal{I}$ in advance, the GWTW algorithm can obtain the required estimate at roughly 1% of the computational cost of using the direct method.

5. Conclusions

The focus of this work has been to address the following question: how is transport in an advection-diffusion problem, between a pre-selected source region $\mathcal{S}$ and receptor region $\mathcal{R}$, modelled most efficiently using stochastic Lagrangian methods? The main theoretical result, which underpins each of the methods subsequently formulated, and which we believe to be fundamental to the understanding of advection-diffusion problems, is that the ideal distribution of particles is proportional to the density of trajectories field $D(\mathbf{x}, t)$ defined by (2.9). The field D is the product of c, the solution of the forward transport problem $(\mathcal{S} \to \mathcal{R})$, and the adjoint solution c^* $(\mathcal{R} \to \mathcal{S})$. The corresponding ideal particle weights are inversely proportional to c^*.

Motivated by the above observations, a robust and flexible Lagrangian adaptive algorithm has been developed based on the 'go-with-the-winners' (GWTW) strategy (Grassberger 2002). The algorithm uses both forward and back trajectories, with the selection of winners and losers in the forward calculation controlled by an estimate of the adjoint solution, which is obtained from the preliminary backwards calculation. In our model problem, trajectories in the adaptive calculation have variance three orders of magnitude lower than those of the direct calculation, with the computational cost increased by just over a factor of two. There is also plenty of room for further optimisation, perhaps most promisingly because the method can be combined with the complementary forward-backwards representation of Milstein *et al.* (2004).

The GWTW algorithm is particularly suited to advection-diffusion in the chaotic advection regime, in which the underlying deterministic trajectories separate exponentially. Alternative 'small noise' approaches based on large-deviation theory (e.g. Vanden-Eijnden & Weare 2012) appear to be prohibitively expensive and complicated to implement in chaotic advection. In the method of Vanden-Eijnden & Weare (2012) a variational problem is solved for each trajectory, at regular time intervals, to obtain an correcting velocity $\mathbf{v}$ for use with Milstein's method. In the chaotic advection regime, the solutions to this variational problem are discontinuous in time and space across multiple interfaces which are not *a priori* predictable. This means that a solution for a particular trajectory cannot be found reliably using an iteration that begins with a previous solution.

The main weakness of the GWTW method described here, is that it is limited by the measurement ratio $\mathcal{I}/S_{00}$, i.e. the total tracer in the measurement region divided by the total released. Unlike the small noise methods discussed above, the GWTW method will fail when the measurement ratio is sufficiently low that there is little or no spatial overlap between forward and backwards trajectories at intermediate times in the calculation.

The Péclet number (Pe) dependence of the GWTW method is felt primarily through its influence on this measurement ratio. Further experiments with the simple model here indicate that GWTW will succeed over a wide range of Pe provided that the measurement region is chosen so that $\mathcal{I}/S_{00}$ is not too small (i.e. $\mathcal{I}/S_{00} \gtrsim 10^{-6} - 10^{-7}$ for the number of trajectories $N = 10^5 - 10^6$ used here). Note that, however, considering a sequence of flows with increasing Pe, the measurement region $\mathcal{R}$ will be required to be located increasingly close to the deterministic trajectories, to maintain the measurement ratio. Although measurement ratios much smaller than the value ($\approx 2 \times 10^{-5}$) considered here are unlikely to be of much practical interest in fluid problems, it is notable that values of $\mathcal{I}/S_{00}$ in the range $10^{-10} - 10^{-20}$ are typical in the molecular dynamics problems which have motivated the development of the small noise methods. For these problems, which are generally not in the chaotic regime, the small noise methods discussed above can be used.

Will the GWTW algorithm be straightforward to implement in state-of-the-art operational atmospheric Lagrangian trajectory models, such as FLEXPART (Stohl *et al.* 2005) or NAME (Jones *et al.* 2007)? Some practical and technical issues must be overcome, such as the increase from two to three dimensions, and the fact that diffusivities in the atmosphere are highly anisotropic, with dispersion in the horizontal primarily driven by a combination of vertical diffusivity and vertical shear in the mean wind (c.f. shear dispersion, Taylor 1953). Further, turbulent diffusivities are spatially inhomogeneous, most notably at the top of the atmospheric boundary layer where there is a rapid decrease. Finally, Poisson (jump) processes are likely to be required in regions of convection (Forster *et al.* 2007). Notwithstanding these potential difficulties, there is no obstacle in principle to the suitable adaptation of the principles outlined in this work to the operational scenario.

Appendix A. Lower bound on the variance of Milstein's estimator with uniform drift

In §3.3, the estimator $\Pi = \Theta_T R(\mathbf{Y}_T)$ for $\mathcal{I}$, given by (3.8), was discussed for the case with a uniform correcting velocity $\mathbf{v} = V\mathbf{i}$. Here we obtain the lower bound (3.15) on its variance.

We will work under the assumption that $R(\mathbf{x})$ is non-negative and bounded above by a constant C, ($C = 1$ in our example). The fact that $\Theta_0 = S_{00}$ (because $g(\mathbf{x}) = 1$) can then be used to write the inequality

$$\mathbb{P}(x < \Pi < \infty) \leqslant \mathbb{P}(x/CS_{00} < \Theta_T/\Theta_0 < \infty) \tag{A 1}$$

If the (unknown) probability density of Π is introduced as $\pi(x)$, (A 1) can be expressed as

$$P_\pi(x) \leqslant P(x/CS_{00}) \qquad (A\,2)$$

where the cumulative densities P_π and P are defined to be

$$P_\pi(x) = \int_x^\infty \pi(q)\,\mathrm{d}q, \qquad P(x) = \int_x^\infty \rho(q)\,\mathrm{d}q,$$

and $\rho(x)$ is the density of the random variable Θ_T/Θ_0 given by (3.14).

Next, note that the variance of Π can be written, using integration by parts, as

$$\mathrm{Var}(\Pi) = \int_0^\infty x^2 \pi(x)\,\mathrm{d}x - \mathcal{I}^2 = 2\int_0^\infty T(x)\,\mathrm{d}x - \mathcal{I}^2, \quad \text{where} \quad T(x) = \int_x^\infty P_\pi(q)\,\mathrm{d}q,$$

where the fact that $\mathbb{E}(\Pi) = \mathcal{I}$ has been used. Using the inequality (A 2) we can write

$$T(x) = \mathcal{I} - \int_0^x P_\pi(q)\,\mathrm{d}q \geqslant \mathcal{I} - S_{00}C \int_0^{x/S_{00}C} P(q)\,\mathrm{d}q.$$

Next define $X = X(t_*, \mathcal{R})$, where $\mathcal{R} = \mathcal{I}/S_{00}C$ satisfies $0 < \mathcal{R} < 1$, to be the root of the equation

$$\int_0^X P(q)\,\mathrm{d}q = \frac{\mathcal{R}}{2}. \qquad (A\,3)$$

After integrating, a nonlinear equation satisfied by X is found

$$\mathrm{erfc}\left(\frac{t_* - 4\log X}{4\sqrt{t_*}}\right) + X\,\mathrm{erfc}\left(\frac{t_* + 4\log X}{4\sqrt{t_*}}\right) = \mathcal{R} \qquad (A\,4)$$

where $\mathrm{erfc}(\cdot)$ denotes the complementary error function. The quantity X has been defined in order that the inequality

$$T(x) \geqslant \mathcal{I}/2, \quad \text{for} \ \ 0 \leqslant x \leqslant S_{00}CX,$$

is satisfied. As a consequence, because $T(x)$ is non-negative everywhere,

$$\mathrm{Var}(\Pi) = 2\int_0^\infty T(x)\,\mathrm{d}x - \mathcal{I}^2 \geqslant 2\int_0^{XS_{00}C} T(x)\,\mathrm{d}x - \mathcal{I}^2 \geqslant \mathcal{I}^2\,(X/\mathcal{R} - 1)$$

which is (3.15) as required.

For the bound (3.15) to be useful it is necessary to understand the behaviour of $X(t_*, \mathcal{R})$ as $t_* \to \infty$ for fixed $\mathcal{R}$. Note that for $t_* \ll 1$ $X \approx \mathcal{R}/2$ and the bound is negative (i.e. useless). At large t_*, the asymptotic behaviour of (A 4) can be seen to be $X \sim \exp(t_*/4)$, hence the bound will always become useful as t_* (or V) increases. For the results discussed in §3.3 numerical solutions of (A 4) have been obtained using a standard root finding method.

Appendix B. The zero variance property of the 'ideal' estimator

The result that setting $\mathbf{v} = \mathbf{v}^*$ in Milstein's method leads to a zero variance estimator is a standard one in stochastic calculus (see e.g. exercise 16.2.1 of Kloeden & Platen

1992) but is included here as a useful alternative perspective on the results of §3.4. To show the result it suffices to apply the multivariate version of Itô's lemma, which states that, for a smooth function $f(\mathbf{z}, t)$ and a vector-valued random process $\mathbf{Z}_t$,

$$\mathrm{d}f(\mathbf{Z}_t, t) = (\partial_t f)\mathrm{d}t + (\nabla_\mathbf{z} f) \cdot \mathrm{d}\mathbf{Z}_t + \tfrac{1}{2}\mathrm{d}\mathbf{Z}_t \cdot \nabla_\mathbf{z}\nabla_\mathbf{z} f \cdot \mathrm{d}\mathbf{Z}_t.$$

The lemma is applied to the stochastic variable $\Theta_t c^*(\mathbf{Y}_t, t)$, with the identification $\mathbf{Z}_t^\dagger = (\mathbf{Y}_t^\dagger\ \Theta_t)$ as in §3.2. Then,

$$\begin{aligned}
\mathrm{d}\left(\Theta_t c^*(\mathbf{Y}_t, t)\right) = {} & \Theta_t(\partial_t c^*)\mathrm{d}t + c^*\mathrm{d}\Theta_t + \Theta_t \nabla c^* \cdot \mathrm{d}\mathbf{Y}_t \\
& + \tfrac{1}{2}\Theta_t\left(\mathrm{d}\mathbf{Y}_t \cdot \nabla\nabla c^* \cdot \mathrm{d}\mathbf{Y}_t\right) + \mathrm{d}\Theta_t\left(\mathrm{d}\mathbf{Y}_t \cdot \nabla c^*\right). \quad \text{(B 1)}
\end{aligned}$$

Next, using (3.6) to substitute for $\mathrm{d}\mathbf{Y}_t$ and $\mathrm{d}\Theta_t$, and discarding terms of $o(\mathrm{d}t)$, after some working it can be seen that

$$\mathrm{d}\left(\Theta_t c^*(\mathbf{Y}_t, t)\right) = \Theta_t\left(\nabla c^* - \tfrac{1}{2}c^*\mathbf{v} \cdot \boldsymbol{\kappa}^{-1}\right) \cdot (2\boldsymbol{\kappa})^{1/2} \cdot \mathrm{d}\mathbf{W}_t - \Theta_t \mathcal{L}^\dagger c^*\, \mathrm{d}t \qquad \text{(B 2)}$$

$$= 0 \quad \text{on setting } \mathbf{v} = \mathbf{v}^*. \qquad \text{(B 3)}$$

Integrating in time, $\Theta_T c^*(\mathbf{Y}_T, T) = \Theta_0 c^*(\mathbf{Y}_0, 0) = \text{constant}$ provided we make the choice $\Theta_0 = G_{00} c^*(Y_0, 0)^{-1}$, or equivalently that $g(\mathbf{x}) = c^*(\mathbf{x}, 0)$ as anticipated in §3.4. Since $\Theta_T c^*(\mathbf{Y}_T, T) = \Theta_T R(\mathbf{Y}_T)$ we have the desired result that the latter is a deterministic (i.e. zero variance) estimator of $\mathcal{I}$.

REFERENCES

ADES, M. & VAN LEEUWEN, P. J. 2013 An exploration of the equivalent weights particle filter. *Q. J. Roy. Meteorol. Soc.* **139**, 820–840.

DAGAN, G. 1987 Theory of solute transport by groundwater. *Ann. Rev. Fluid Mech.* **19** (1), 183–213.

DEVENISH, B. J., THOMSON, D. J., MARENCO, F., LEADBETTER, S. J., RICKETTS, H. & DACRE, H. F. 2011 A study of the arrival over the United Kingdom in April 2010 of the Eyjafjallajokull ash cloud using ground-based lidar and numerical simulations. *Atmos. Environ.* **48**, 152–164.

ESLER, J. G. 2008 Robust and leaky transport barriers in unstable baroclinic flows. *Phys. Fluids* **20**, 116602.

FORSTER, C., STOHL, A. & SEIBERT, P. 2007 Parameterization of convective transport in a Lagrangian particle dispersion model and its evaluation. *J. Appl. Meteorology* **46**, 403–422.

GARDINER, C. W. 2009 *Stochastic Methods: A Handbook for the Natural and Social Sciences*, 4th edn. Berlin: Springer, 447pp.

GRASSBERGER, P. 1997 Prune-enriched Rosenbluth method: simulations of θ polymers of chain length up to 1000000. *Phys. Rev. E* **56**, 3682–3693.

GRASSBERGER, P. 2002 Go with the winners: a general Monte-Carlo strategy. *Comp. Phys. Comm.* **147**, 64–70.

HAYNES, P. H. 2011 Transport and mixing of atmospheric pollutants. In *Environmental Haz-*

ards: The Fluid Mechanics and Geophysics of Extreme Events (ed. H. K. Moffatt & E. F. Shuckburgh). World-Scientific.

HAYNES, P. H., POET, D. & SHUCKBURGH, E. F. 2007 Transport and mixing in dynamically consistent flows. *J. Atmos. Sci.* **64**, 3640–3651.

HAYNES, P. H. & VANNESTE, J. 2014 Dispersion in the large-deviation regime. Part I: shear flows and periodic flows . *J. Fluid Mech.* **745**, 321–350.

HOURDIN, F. & TALAGRAND, O. 2006 Eulerian backtracking of atmospheric tracers. I: Adjoint derivation and parametrization of subgrid-scale transport. *Q. J. R. Meteorol. Soc.* **132**, 567–583.

JONES, A. R., THOMSON, D. J., HORT, M. & DEVENISH, B. 2007 The U.K. Met Office's next-generation atmospheric dispersion model, NAME III. In *Air Pollution Modeling and its Application XVII (Proceedings of the 27th NATO/CCMS International Technical Meeting on Air Pollution Modelling and its Application)* (ed. C. Borrego & A.-L. Norman), pp. pp. 580–589. Berlin: Springer.

KLOEDEN, P. E. & PLATEN, E. 1992 *Numerical Solution of Stochastic Differential Equations*, 4th edn. Berlin: Springer-Verlag, 636pp.

MAJDA, A. J. & KRAMER, P. R. 1999 Simplified models for turbulent diffusion: theory, numerical modelling and physical phenomena. *Physics Reports* **314**, 237.

METHVEN, J., ARNOLD, S. R., STOHL, A., EVANS, M. J., AVERY, M., LAW, K., LEWIS, A. C., MONKS, P. S., PARRISH, D. D., REEVES, C., SCHLAGER, H., ATLAS, E., BLAKE, D. R., COE, H., CROSIER, J., FLOCKE, F. M., HOLLOWAY, J. S., HOPKINS, J. R., McQUAID, J., PURVIS, R., RAPPENGLUCK, B., SINGH, H. B., WATSON, N. M., WHALLEY, L. K. & WILLIAMS, P. I. 2006 Establishing Lagrangian connections between observations within air masses crossing the Atlantic during the ICARTT experiment. *J. Geophys. Res.* **111**, D23S62.

MILSTEIN, G. N. 1995 *Numerical solution of stochastic differential equations*. Kluwer, Dordrecht, 177pp.

MILSTEIN, G. N., SCHOENMAKERS, J. G. M. & SPOKOINY, V. 2004 Transition density estimation for stochastic differential equations via forward-reverse representations. *Bernoulli* **10** (2), 281–312.

MILSTEIN, G. N. & TRETYAKOV, M. V. 2004 *Stochastic Numerics for Mathematical Physics*. Springer, 594 pp.

ØKSENDAL, B. 2007 *Stochastic Differential Equations: An introduction with applications*, 6th edn. Berlin: Springer-Verlag, 369pp.

PIERREHUMBERT, R. T. 1991 Chaotic mixing of tracer and vorticity by modulated travelling Rossby waves. *Geophys. Astrophys. Fluid Dyn.* **58**, 285–319.

PLUMB, R. A. 2002 Stratospheric transport. *J. Met. Soc. Japan* **80**, 793–809.

PROEHL, J. A., LYNCH, D. R., McGILLICUDDY JR., D. J. & LEDWELL, J. R. 2005 Modeling turbulent dispersion on the north flank of Georges bank using Lagrangian particle methods. *Cont. Shelf Res.* **25**, 875 –900.

RODEAN, H. C. 1996 *Stochastic Lagrangian Models of Turbulent Diffusion*. Chicago: American Meteorological Society, 84pp.

SILVERMAN, B. W. 1986 *Density estimation for statistics and data analysis*. London, UK: Chapman and Hall, 176pp.

Spivakovskaya, D., Heemink, A., Milstein, G. & Schoenmakers, J. 2005 Simulation of the transport of particles in coastal waters using forward and reverse time diffusion. *Adv. Water Res.* **28** (9), 927 – 938.

Spivakovskaya, D., Heemink, A. W. & Deleersnijder, E. 2007 Lagrangian modelling of multi-dimensional advection-diffusion with space-varying diffusivities: theory and idealized test cases. *Ocean Dynamics* **57**, 189–203.

Stohl, A., Forster, C., Frank, A., Seibert, P. & Wotawa, G. 2005 Technical note: The Lagrangian particle dispersion model FLEXPART version 6.2. *Atmos. Chem. Phys.* **5**, 2461–2474.

Taylor, G. I. 1953 Dispersion of soluble matter in solvent flowing slowly through a tube. *Proc. R. Soc. Lond. A* **219**, 186–203.

Thomson, D. J. 1987 Criteria for the selection of stochastic models of particle trajectories in turbulent flows. *J. Fluid Mech.* **180**, 529—556.

Tocino, A. & Ardanuy, R. 2002 Runge-Kutta methods for numerical solution of stochastic differential equations. *J. Comp. Appl. Math.* **138**, 219–241.

Vanden-Eijnden, E. & Weare, J. 2012 Rare event simulation for small noise diffusions. *Comm. Pure Appl. Math.* **65**, 1770–1803.

Wand, M. P. & Jones, M. C. 1994 *Kernel Smoothing*. London, UK: Chapman and Hall, 224pp.

Weiss, J. B. & Knobloch, E. 1989 Mass transport by modulated travelling waves. *Phys. Rev. A* **40**, 2579–2589.

Weiss, J. B. & Provenzale, A. 2008 *Transport and Mixing in Geophysical Flows*. Berlin Heidelberg: Springer.

Zimmermann, S., Koumoutsakos, P. & Kinzelbach, W. 2001 Simulation of pollutant transport using a particle method. *J. Comp. Phys.* **173** (1), 322–347.