

University of Warwick institutional repository: <http://go.warwick.ac.uk/wrap>

A Thesis Submitted for the Degree of PhD at the University of Warwick

<http://go.warwick.ac.uk/wrap/56806>

This thesis is made available online and is protected by original copyright.

Please scroll down to view the document itself.

Please refer to the repository record for this item for information to help you to cite it. Our policy information is available from the repository home page.

AUTHOR: **Damon McDougall** DEGREE: **Ph.D.**

TITLE: **Assimilating Eulerian and Lagrangian Data to Quantify Flow Uncertainty in Testbed Oceanography Models**

DATE OF DEPOSIT:

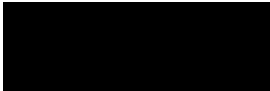
I agree that this thesis shall be available in accordance with the regulations governing the University of Warwick theses.

I agree that the summary of this thesis may be submitted for publication.

I **agree** that the thesis may be photocopied (single copies for study purposes only).

Theses with no restriction on photocopying will also be made available to the British Library for microfilming. The British Library may supply copies to individuals or libraries, subject to a statement from them that the copy is supplied for non-publishing purposes. All copies supplied by the British Library will carry the following statement:

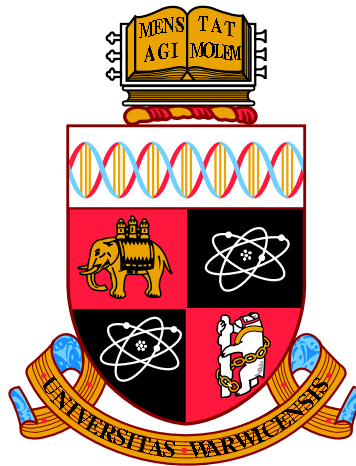
“Attention is drawn to the fact that the copyright of this thesis rests with its author. This copy of the thesis has been supplied on the condition that anyone who consults it is understood to recognise that its copyright rests with its author and that no quotation from the thesis and no information derived from it may be published without the author’s written consent.”

AUTHOR’S SIGNATURE: 

USER’S DECLARATION

1. I undertake not to quote or make use of any information from this thesis without making acknowledgement to the author.
2. I further undertake to allow no-one else to use this thesis while it is in my care.

DATE	SIGNATURE	ADDRESS
.....		
.....		
.....		
.....		
.....		



**Assimilating Eulerian and Lagrangian Data to
Quantify Flow Uncertainty in Testbed
Oceanography Models**

by

Damon McDougall

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Mathematics Institute

August 2012

THE UNIVERSITY OF
WARWICK

Contents

Acknowledgements	iii
Declarations	vi
Abstract	vii
Notation	viii
Chapter 1 Background and preliminaries	1
1.1 History of data assimilation	1
1.1.1 Bayesian data assimilation	2
1.1.2 Variational data assimilation	5
1.2 Flavours of data assimilation	6
1.2.1 Filtering and smoothing	6
1.2.2 Eulerian and Lagrangian data assimilation	8
1.3 Markov chain Monte Carlo methods	9
1.3.1 Adaptive burn-in	11
1.3.2 Metastability	13
1.3.3 A note on random numbers	15
1.4 Rigorous mathematical setting	18
1.4.1 Regularity of random fields	20
1.5 Thesis summary	23
Chapter 2 Data assimilation for the advection equation	27
2.1 Overview	27
2.2 Sampling the initial condition	29
2.2.1 Varying step-size and observational error	30
2.2.2 Varying the seed and sample size	34
2.2.3 Varying the number of observations	37

2.3	Sampling the wave velocity	38
2.3.1	Simulated annealing	44
2.4	Wavespeed mismatch	46
2.4.1	Sampling the initial condition with model error	50
2.5	Sampling the joint	53
2.5.1	Seeding nearby the truth	53
2.5.2	Slices of the objective function	54
2.5.3	Seeding the wave velocity	57
2.6	Modifying the likelihood	58
2.6.1	Sampling the initial condition	60
2.6.2	Sampling the wave velocity	64
2.7	Conclusions	68
Chapter 3	Data assimilation for controlled testbed ocean drifters	73
3.1	Overview	73
3.2	Time-independent flow	75
3.2.1	Naïve control strategy	76
3.2.2	A posteriori control strategy	95
3.3	Periodic time-dependent disturbances	99
3.3.1	Naïve control for time-dependent flow model	100
3.3.2	Time-dependent a posteriori control	104
3.4	Conclusions	111
Chapter 4	Data assimilation for optimally controlled testbed ocean drifters	114
4.1	Overview	114
4.2	Derivation of control theory	115
4.2.1	Optimal feedback control	118
4.3	Specific use-case	120
4.4	Application to data assimilation	124
4.5	Conclusions	126
Chapter 5	Discussion	129

Acknowledgements

First and foremost, I would like to thank both of my Ph.D supervisors, Professor Chris Jones and Professor Andrew Stuart. They have not only provided invaluable advice over the past four years, but have been an increasing source of encouragement, support and understanding. Our feedback-oriented working relationship has pledged staggeringly useful input, without which this doctoral work could not have been completed. I feel tremendously privileged to have worked with two exceptionally well-established scientists.

I have been fortunate to have had many inspirational scientific discussions with a plethora of different people. These people all deserve thanks and they are, Tom Bellskey, Graham Cox, Sean Crowell, Masoumeh Dashti, Kody Law, Igor Mezić, Lewis Mitchell, Richard Moore, Blane Rhoads, Naratip Santitissadeekorn, Elaine Spiller and David White.

Practically, the computational resources managed by the University of Warwick's Centre for Scientific Computing have been paramount in the presentation of the numerical material throughout this thesis. Thousands of hours of CPU time have been utilised to complete this work and it does not go without profound gratitude. Furthermore, I am very thankful for the funding provided by NERC and EPSRC to undertake this research.

On a personal level, I must thank Sarah Chandler, Anna Clugston, Simon Cotter, Martha Dellar, Andrew Duncan, Chris Cantwell, David Holmes, Dave Howden, Dave Moxey, Tom Ranner, Matt Reynolds and Yuan Xiang Zhang for making it a fun few

years. Moreover, Andrew, Dave, Tom and Lewis deserve thanks, and probably a few beers, for the daunting task of proofreading my work and taking immense pleasure in telling me when I am wrong. Together with my girlfriend, Andrea Overbay, they have provided an invaluable support network within which I sought solace where I had problems, and delight where I succeeded.

Lastly, a debt of thanks is owed to my family. My parents, Tom and Eve, have provided the requisite chromosomes needed to complete this work and, through challenging me, my brother, Steven, has helped improve my skill of communicating research level mathematics.

Mum and Dad, I have always appreciated your constant encouragement. The emotional support you have given me will forever be appreciated. I feel privileged to be your son and to share this great achievement with you. Though you have often said, “I will never be able to understand what you do,” I am sure you will be proud with what I have *done*.

To Mum and Dad

Declarations

Parts of the mathematical groundwork in Stuart [2010], which form the basis of infinite dimensional Bayesian inverse problems, have been adapted for section 1.4.

The numerical studies and discussion done in sections 2.4 and 2.5.3, and related conclusions in section 2.7, have been published jointly with Lee and Stuart in Lee *et al.* [2011]. My primary contributions to the paper were on the numerical side, and it is this aspect of the paper which I concentrate on in this thesis.

The efforts in chapter 3 are not yet published, but are a work in preparation with Jones in McDougall & Jones [2012].

The basis of chapter 4 concerns optimally controlled ocean gliders, the algorithm for which was taken from Rhoads *et al.* [2010]. The underpinning theory for this [Bryson Jr. & Ho, 1975] heavily inspired the mathematical presentation in section 4.2.

This work has not been submitted for a degree at any other university and, with the exception of the four cases made above, I declare this research to be my own work unless otherwise stated.

Abstract

Data assimilation is the act of merging observed data into a mathematical model. This act enables scientists from a wide range of disciplines to make predictions. For example, predictions of ocean circulations are needed to provide hurricane disaster maps. Alternatively, using ocean current predictions to adequately manage oil spills has significant practical applications. Predictions are uncertain and this uncertainty is encoded into a posterior probability distribution. This thesis aims to explore two overarching aspects of data assimilation, both of which address the influence of the mathematical model on the posterior distribution.

The first aspect we study is model error. Error is always present in mathematical models. Therefore, characterising posterior flow information as function of model error is paramount in understanding the practical implications of predictions. In a model describing advective transport, we make observations of the underlying flow at fixed locations. We characterise the mean of the posterior distribution as a function of the error in the advection velocity parameter. When the error is zero, the model is perfect and we reconstruct the true underlying flow. Partial recovery of the true underlying flow occurs when the error is rational, the denominator of which dictates the number of Fourier modes present in the reconstruction. An irrational error leads to retrieval only of the spatial mean of the flow.

The second aspect we study is the control of ocean drifters. Commonplace in oceanography is the collection of ocean drifter positions. Ocean drifters are devices that sit on the surface of the ocean and move with the flow, transmitting their position via GPS to stations on land. Using drifter data, it is possible to obtain a posterior on the underlying flow. This problem, however, is highly underdetermined. Through controlling an ocean drifter, we attempt to improve our knowledge of the underlying flow. We do this by instructing the drifter to explore parts of the flow currently uncharted, thereby obtaining fresh observations. The efficacy of a control is determined by its effect on the variance of the posterior distribution. A smaller variance is interpreted as a better understanding of the flow. We show a systematic reduction in variance can be achieved by utilising controls that allow the drifter to navigate new or ‘interesting’ flow structures, a good example of which are eddies.

Notation

S^1 : $\{x \in \mathbb{R}^2 \mid \|x\| = 1\}$

\mathbb{T}^2 : $\mathbb{R}^2/\mathbb{Z}^2$

\mathbb{K}^2 : $\mathbb{Z}^2 \setminus \{(0,0)\}$

$L^2(\mathbb{T}^2)$: $\{f: \mathbb{T}^2 \rightarrow \mathbb{R} \mid \int_{\mathbb{T}^2} |f|^2 < \infty\}$ (also denoted $L^2_{\text{per}}(\mathbb{T}^2)$)

\mathcal{H} : $\{f \in L^2(\mathbb{T}^2) \mid \int_{\mathbb{T}^2} f = 0\}$

H^s or H^s_{per} : $\{f \in \mathcal{H} \mid \sum_k \lambda_k^s |\langle f, \phi_k \rangle|^2 < \infty\}$, where $\{\lambda_k, \phi_k\}$ are eigenvalues/eigenvectors of the Laplacian that form a basis for \mathcal{H}

η : Observational error

σ^2 : Variance of observational error

μ : Prior standard deviation

I_n : $n \times n$ identity matrix

α : When used as an exponent, it refers to a regularity parameter. When used as a function, $\alpha(\cdot, \cdot)$, it refers to an acceptance probability

μ_0 : The prior measure

μ^y : The posterior measure with observed data y

Δ : The Laplacian differential operator

Chapter 1

Background and preliminaries

1.1 History of data assimilation

Consider a physical system describing some physical quantity of interest. Given noisy observations of the system's state over time, the aim is to estimate the state of that system at some future time. This is a hard problem. For large weather systems, this problem has been looked at for decades and is still an active area of research. Estimating a future atmospheric or oceanic state is an endeavour that does not benefit solely scientists. The general public seek information in this regard and depend on the scientific community to produce predictions that are accurate, informative and actionable. Predictions regarding natural disasters are useful for national emergency services to mitigate potential fatalities. Predictions of weather in the short term aid in making safe and informed travel decisions. Predictions on a longer timespan, such as seasonal states for example, help companies execute profitable business manoeuvres. Predictions affect people's lives.

Data assimilation is the act of merging observations of some quantity into a mathematical representation of a physical system [Kalnay, 2002]. The result is an objective estimate of the state, which can be propagated through the model to obtain a prediction. There are many ways of utilising information from both observed data and model output, and this is reflected by the diverse history of data assimilation.

The data assimilation story starts in the 1950s. The early work of Charney [1951] was one of the first in state estimation and prediction. Prediction is an entirely model-based paradigm and can be executed without making physical observations. To obtain predictions that hold scientific value, it is worth formalising how to con-

struct a ‘good’ initial condition to the model. Constructing such an initial condition will involve a mapping of the state in *physical* space to a state in *model* space. One of the first to consider such mappings was Bergthorsson & Doos [1955] who explored interpolation of observational data onto a grid. Least-squares fitting [Gilchrist & Cressman, 1954; Cressman, 1959] also fits well within this objective. Methods on data analysis were used extensively before they inevitably merged with models to form what is termed today as *data assimilation*. This is the act of taking observed data and a model to produce, in some sense, the ‘best possible’ prediction.

Rudolf Emil Kálmán is probably the considered the forefather of state estimation. Most of the techniques all stem from one set of equations developed by Kalman in 1960, the Kalman Filter [Kalman, 1960]. One of the earliest physical applications of the Kalman filter was in trajectory estimation, leading to its implementation in the Apollo navigation computer; a key development in the Apollo program in the United States. It is worth noting that Swerling [Swerling, 1958] and Bucy [Kalman & Bucy, 1961] also contributed to a lot of the early theory.

1.1.1 Bayesian data assimilation

The Kalman filter can be derived explicitly from the following setup. We are given a linear process model on some state $x_k \in \mathbb{R}^n$,

$$x_{k+1} = Mx_k, \quad M \in \mathbb{R}^{n \times n}, \quad k \in \mathbb{N},$$

and noisy observations of the state,

$$y_k = Hx_k + \eta_k, \quad \eta_k \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, R),$$

at times $k = 1, \dots, L$. Here H is linear, called the *observation operator*, and η_k is the *observation error*. The matrix R is the observation error covariance matrix. We then define two terms, $m_{k|k}$ and $C_{k|k}$, the state estimate at time k given all the data up to and including time k , and the error covariance of that state, respectively. The filter then proceeds in two steps, a prediction (or forecasting) step and an update (or analysis) step. The prediction step is as follows,

$$m_{k+1|k} = Mm_{k|k}, \tag{1.1a}$$

$$C_{k+1|k} = MC_{k|k}M^\top. \tag{1.1b}$$

The variables $m_{k|k}$ and $C_{k|k}$ are often referred to as the analysis mean and analysis covariance. Furthermore, the variables $m_{k+1|k}$ and $C_{k+1|k}$ are referred to as the forecast mean and forecast covariance. The prediction formulae 1.1 indicate that the estimate of the state at time $k + 1$ is simply the current state estimate propagated under the model M . Moreover, to obtain the covariance about the predicted state, one conjugates the current analysis covariance matrix by a forward model integration. These formulae are fairly intuitive. The interesting step is the update, where the analysis mean and covariance at time $k + 1$ using a new observation have the following form,

$$m_{k+1|k+1} = m_{k+1|k} + C_{k+1|k} H^\top (H C_{k+1|k} H^\top + R)^{-1} (y_{k+1} - H m_{k+1|k}), \quad (1.2a)$$

$$C_{k+1|k+1} = (I - \underbrace{C_{k+1|k} H^\top (H C_{k+1|k} H^\top + R)^{-1} H}_{\text{Kalman gain matrix}}) C_{k+1|k}. \quad (1.2b)$$

As we shall see later not only do (1.1) and (1.2) give a good state estimate for a linear system with Gaussian observation error, they are *exactly* the mean and covariance of Gaussian distributions about those states. They are called the prior distribution and posterior distribution respectively. The work of Kalman would form the basis of most data assimilation techniques for the next fifty years.

Data assimilation can be utilised in any model, though its most common application is to numerical weather prediction (NWP). Several theoretical and numerical developments have taken place since the 1960s that have made data assimilation a more numerically tractable and stable process.

Big weather models, of $\mathcal{O}(10^7)$ degrees of freedom, used in institutions such as the Met Office or ECMWF are certainly far from linear. It is therefore necessary to be concerned about how to implement (1.1)–(1.2). The *Extended Kalman filter* (EKF) [Sorenson, 1960; Jazwinski, 1970] deals with this nonlinearity, albeit in a somewhat naïve fashion, by linearising the M and H operators about the current mean and covariance. In practice, computing and storing these linearisations, especially in a weather system with millions of variables, is a costly process. Instead of computing the mean and covariance exactly, as in (1.1)–(1.2), representing the distributions discretely by an *ensemble* of members helps to alleviate the high dimensional burden. Analogous formulae may be derived now for the *ensemble mean* and *ensemble covariance* when the ensemble is designed to represent a Gaussian distribution. This is referred to as the *Ensemble Kalman filter* (EnKF), originally developed by Evensen [1994, 2006]; Houtekamer & Mitchell [1998]. The EnKF has been the subject of a large research audience since its initial inception in 1994, despite the fact it is

theoretically only applicable to linear and Gaussian models. Lots of variants have been devised to mitigate certain undesirable properties of the EnKF. For example, the *localised ensemble Kalman filter* [Anderson, 2003; Ott *et al.*, 2004; Baek *et al.*, 2006] addresses problems relating to the rank-deficiency of the ensemble covariance matrix. The square root formulation of the Kalman filter addresses computational speed and stability issues that arise when computing the covariance matrix [Carlson, 1973]. In 2001, the *ensemble transform Kalman filter* (ETKF) was devised with the aim of speeding up the computation of the forecast covariance matrix; a desirable quality. It is even more useful when using this to assess the efficacy of locations of future observation stations, as described in Bishop *et al.* [2001]. The *local ensemble transform Kalman filter* [Ott *et al.*, 2004] allows massively parallel data assimilation schemes to take place by partitioning the physical domain into smaller chunks, paving the way for even faster computation.

In all of these filters, some approximation is made. We either approximate the model by linearising it about the current analysis mean, or we approximate the analysis and forecast distribution by two moments; a mean and covariance. One last method we have not mentioned is the *particle filter*. Methods like the particle filter and other particle method variants are not new methods designed for the purpose of solving the data assimilation problem, they are existing statistical techniques for sampling unknown probability distributions. Sampling probability distributions is a task physicists, chemists and biologists have been executing decades before statisticians started proving theoretical results regarding their behaviour. Section 1.3 explores this topic in more detail.

Particle filters [Doucet *et al.*, 2001] are similar to the ensemble Kalman filter in the sense that distributions are approximated by a finite sum of delta functions: particles. The forecast distribution is obtained by propagating each one of these particles through the full, possibly nonlinear, model. The more particles used to approximate the distribution, the more useful calculated statistics become. As a compromise for the increased statistical value, one must solve the full model for each particle. The difference between the ensemble Kalman filter and the particle filter becomes apparent at the assimilation step. Instead of computing the relevant Kalman filter equation (1.2) as in the ensemble Kalman filter, each particle is *weighted* proportionally by its distance to the position of the observation. There are many types of particle filters and there is a large community of scientists researching them to improve their performance. When we talk of *the* particle filter, usually we refer to the Bootstrap particle filter [Doucet *et al.*, 2001]. In high dimensional state

spaces, it is a well-known fact that particle filters perform extremely poorly [Bickel *et al.*, 2008]. For an in-depth overview of the particle filter background material, see Doucet *et al.* [2001]. Recent advances in this field mean that problems in high dimensions can, to a certain degree, be mitigated [Leeuwen, 2010].

Note that all of the above methods are either ensemble based methods or methods derived directly from the original Kalman filter, both of which require the calculation of a, potentially, very large covariance matrix.

1.1.2 Variational data assimilation

Everything discussed hitherto has approached the problem from a Bayesian perspective, involving the computation of aspects of some distribution of interest. There are other approaches to addressing the problem of assimilating observations into a model that is utilised more heavily within the circle of numerical weather prediction. *Variational* methods; involving the (nonlinear) minimisation of some cost functional, is one such approach. It is worth noting here that Bayesian and variational approaches are very closely related. The cost functional is exactly the negative logarithm of the posterior probability density function.

A first variational approach to data assimilation, due to Hoke & Anthes [1976], employed a nudging term to push predictions towards physical observations. Although crude, methods like this are not entirely uncommon and exist today [Leeuwen, 2010; Vanden-Eijnden & Weare, 2012] for use in highly nonlinear problems. Optimal interpolation was the next important scientific tool for numerical weather prediction and Lorenc [Lorenc, 1986] has been a figurehead in implementing the cornerstone data assimilation techniques such as optimal interpolation, 3D-Var and 4D-Var into the Met Office’s operational numerical weather prediction schemes.

The two main methods used today are 3D-Var (introduced into numerical weather prediction by Lorenc *et al.* [2000] and also used at NCAR in Boulder, Colorado [Barker, 2004]) and 4D-Var. The 3D-Var method minimises the cost functional over state vectors x with a single observation vector y ,

$$J(x) = \frac{1}{2} \|H(x) - y\|_R^2 + \frac{1}{2} \|x\|_B^2, \quad (1.3)$$

where R is called the observation error covariance matrix and B is called the background, or prior, error covariance matrix. The 4D-Var method [Bengtsson, 1975; Lewis & Derber, 1985; Lorenc, 1986; Le Dimet & Talagrand, 1986; Talagrand &

Courtier, 1987] incorporates a sequential flavour and instead minimises the cost functional with observations that come in at certain times,

$$J(x) = \frac{1}{2} \sum_{k=1}^N \|H_k(x_k) - y_k\|_R^2 + \frac{1}{2} \|x\|_B^2. \quad (1.4)$$

Between observations, one executes a forecasting step.

There are several issues with both of these when the operators H and M are non-linear or when the observational noise is non-Gaussian. In these cases, one may use Gauss-Newton minimisation with the *incremental 4D-Var* proposed by Courtier *et al.* [1994] to achieve a reduction in computational cost by an order of magnitude over standard 4D-Var. Operationally, numerical weather prediction models are invariably highly nonlinear, necessitating the use of approximate methods [Lawless *et al.*, 2005a,b; Lawless & Nichols, 2006].

The last, and likely most important problem in data assimilation arises when the model M does not generate the data y that is being assimilated [Nichols & Griffith, 1996, 2000]. It is important to understand the role of getting the model wrong, as so frequently happens in the physical realm of numerical weather prediction. It is the job of mathematical modellers to construct such a model that is close to reality *and* computationally cheap to run. Examples of explorations of toy models can be found in Apte *et al.* [2008a], Cotter *et al.* [2009] Smith *et al.* [2009], Smith *et al.* [2011] and Lee *et al.* [2011].

1.2 Flavours of data assimilation

Many different types of data assimilation exist and it is important to illustrate their differences. There are too many to list them all here, though we highlight the most influential, explaining the differences in their methodology and goals. We also give real-world examples of their use when appropriate.

1.2.1 Filtering and smoothing

The two terms ‘filtering’ and ‘smoothing’ are sometimes used interchangeably. However, they are different approaches to data assimilation and we illustrate the distinction between them below.

Say we are given a model and we are interested in the state of a system at time k . Say we are also given a vector of observations $(y_1^\top, \dots, y_K^\top)^\top$. The distribution of interest is, $\mathbb{P}(x_k | \{y_j\}_{j=0}^l)$. When $l = k$, this is called the analysis distribution, when $l < k$ it is called the forecast distribution, and when $l = K$ it is called the smoothing distribution. Filtering is the process of obtaining the analysis distribution iteratively.

1. Start with $\mathbb{P}(x_k | \{y_j\}_{j=0}^k)$
2. Integrate model to obtain $\mathbb{P}(x_{k+1} | \{y_j\}_{j=0}^k)$
3. Assimilate to obtain $\mathbb{P}(x_{k+1} | \{y_j\}_{j=0}^{k+1})$
4. Set $k \leftarrow k + 1$. Go to 1.

Smoothing is the process of estimating the state using all possible data. Note, once an estimate to the state at time k , x_k , has been obtained, all future state realisations are determined by solving the model with initial condition x_k . When the model is deterministic, it is often the case that smoothers will estimate the initial condition x_0 , since the state path is thus entirely determined. Figure 1.1 illustrates, in a cartoon, the qualitative differences between filtering and smoothing. We show here the mean of the forecast and analysis distributions. Notice that the filtering path is not ‘smooth’. The jumps in this path are when the analysis step takes place. The green dots indicate the *analysis mean*.

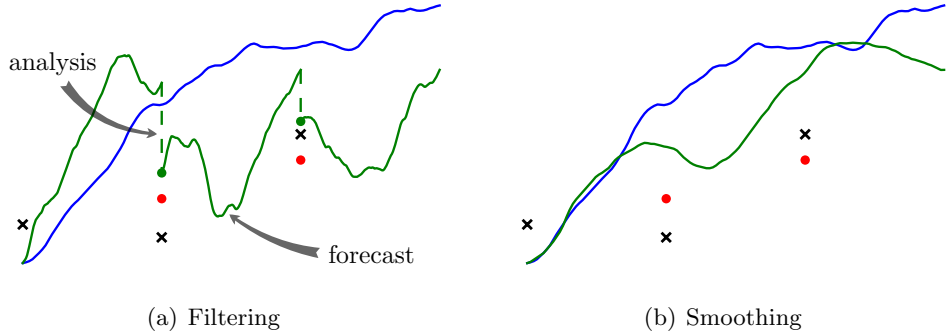


Figure 1.1: Two data assimilation strategies: filtering and smoothing. The blue path is the prior state. The black crosses indicate the true state of the system.

The red dots are observations. In the case of filtering, the green path is the forecast and the green dots are *analyses*. In the case of smoothing, the green path is the analysis given all the observations.

Lastly, as a sanity check for consistency between filtering and smoothing, the filtered

analysis distribution at the final time $t = K$ and the smoothing distribution on the final state are exactly the same, $\mathbb{P}\left(x_K|\{y_j\}_{j=0}^K\right)$. An excellent overview of filtering and statistical techniques to solve these inverse problems can be found in Wikle & Berliner [2007]. This work sets up the Bayesian framework for data assimilation methods and gives derivations of the Kalman and related sequential data assimilation approaches, including particle filter methods. For the beginner, it is an informative introduction to the underpinning theory.

Filtering, by construction, is an ‘on-line’ process. Operationally, it necessitates a model integration to generate a forecast. When a new observation is made, computational linear algebra and minimisation techniques are executed to give the best estimate of the state given all of the observations up to and including the new one. Due to its sequential nature, filtering is the more commonly used data assimilation strategy, not only for its interest to scientists within numerical weather prediction centres, but also to the dynamical systems community. Dynamical systems are also prevalent in the next two comparisons on types of data; Eulerian observations and Lagrangian observations.

1.2.2 Eulerian and Lagrangian data assimilation

Eulerian observations are point observations of a field. These are utilised extensively in chapter 2 of this thesis and an example of their usage is given below. We noisily observe a velocity field, v , of a fluid at *fixed* points in space and time,

$$y_{jk} = v(x_j, t_k) + \eta_{jk}, \quad \eta_{jk} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2).$$

The observation stations in space are fixed and do not change in time. Lagrangian observations, on the other hand, are observation stations that move with the fluid,

$$y_k = z(t_k) + \eta_k, \quad \eta_k \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2 I_2),$$

where z is the position of the observation station and satisfies,

$$\dot{z} = v(z, t). \tag{1.5}$$

The most common use-case of these observations are in representing positions of passive tracers in a fluid.

Eulerian data assimilation and Lagrangian data assimilation are done no differently.

Their differing nomenclature seeks only to emphasise the distinction on the type of observations being assimilated. The same data assimilation techniques are applied with either type of observation by appending the model state with the equations for the motion of the particles (1.5) [Kuznetsov *et al.*, 2003]. The difference between the two setups is that even if the model governing the evolution of v is linear, the full coupled model of the flow and the tracers is often nonlinear. Assimilation of data into nonlinear models is an advanced topic that will be covered in chapter 3 of this thesis with application to a two dimensional kinematic travelling wave model.

The Lagrangian observation set-up fits in extremely well within the oceanographic objective of assimilating observations of buoys, drifters, tracers or gliders. Estimating ocean flows and buoy trajectories has a long history, very similar to that of NWP and the general Kalman filter literature. First, a comparison of forecast errors in a barotropic open ocean model can be found in [Robinson & Haidvogel, 1981], with emphasis on how forecasts are sensitive to boundary information. Applications of the full Kalman filter with Lagrangian observations can be seen as early as 1982 [Barbieri & Schopf, 1982; Miller, 1986; Parrish & Cohn, 1985; Carter, 1989]. For a variational least-squares approach to eddy estimation, the reader is directed to [Robinson & Leslie, 1985]. A standard mathematical framework for assimilating Lagrangian observations appeared in 2003 [Kuznetsov *et al.*, 2003]. Finally, Robel *et al.* [2011] exposes a novel approach to ocean current observations involving the treatment of sea turtles as Lagrangian observers.

1.3 Markov chain Monte Carlo methods

Markov chain Monte Carlo (MCMC) methods are a class of computational techniques for drawing samples from a unknown target distribution. Employing MCMC methods for the case where the target distribution is the posterior distribution, these methods can be used to do data assimilation. We give a general outline for MCMC methods in the finite dimensional case, we then generalise the approach to an infinite dimensional setting. Infinite dimensional Monte Carlo schemes are employed heavily throughout this thesis.

Often, we want to know the shape of a distribution and moments are not enough to characterise it entirely. In that case, samples from the distribution are useful since one can compute as many moments as is desired. Conclusions can then be drawn if the target distribution has an irregular structure. In most practical cases the target

distribution is hard to sample directly, and an approximate method must be utilised. The idea is to make draws from a distribution that is easier to probe, and apply a transformation to map the resulting samples to the target distribution. More specifically, we wish to construct a sequence $\{x_k\}_{k \in \mathbb{N}}$, called a Markov chain, such that each x_k looks like it was drawn from the target distribution, $p(x)$. We dream up a distribution that is easy to sample, called the *proposal* distribution, $q(x_k, \cdot)$, which may depend on the current state of the chain. The proposal distribution specifies the probability of transitioning to the next term in the sequence. Given the current state in the Markov chain is x_k , we draw $z \sim q(x_k, \cdot)$ and propose z as the next state in the sequence. We then accept z as the next state with probability,

$$\alpha(x_k, z) = \min \left\{ 1, \frac{p(z)q(z, x_k)}{p(x_k)q(x_k, z)} \right\}. \quad (1.6)$$

The next state in the Markov chain is then,

$$x_{k+1} = \begin{cases} z & \text{with probability } \alpha(x_k, z) \\ x_k & \text{with probability } 1 - \alpha(x_k, z). \end{cases}$$

The Markov chain with acceptance probability (1.6) is called the Metropolis-Hastings chain. This acceptance probability was chosen so that the resulting Markov chain satisfies a *detailed balance* property with respect to the unknown stationary measure $p(x)$. Therefore, if it is the case that the Markov chain is also *ergodic*, then the samples converge to the unique invariant distribution $p(x)$.

Metropolis-Hastings samplers are characterised by their proposal distributions. Some of the most common choices are:

- Gibbs sampler (in two dimensions): Let $x = (x_1, x_2)$. Choose a component (1 or 2) of x and propose either $z = (y, x_2)$ or $z = (x_1, y)$, respectively, where y is sampled from either $p(y|x_2)$ or $p(y|x_1)$, respectively.
- Independence sampler: Labelled as such not because it produces independent samples from p , but because the proposal distribution $q(x, z) = q(z)$ is independent of the current state in the chain.
- Random walk sampler: Here the proposal distribution is chosen to be symmetric, $q(x, z) = q(z, x)$.

The Gibbs sampler is popular because it turns out that the acceptance probability is exactly 1. However, it performs extremely poorly when the target distribution is highly correlated in its components. The independence sampler does not exhibit

good computational performance, but due to the simplicity of the proposal density its theoretical properties are well understood. Random walk Metropolis-Hastings samplers have nice convergence properties and are, as a consequence, used extensively in practice. We shall use a random walk Metropolis-Hastings sampler for our purposes. If the current state in the Markov chain is x_k , a typical random walk proposal will look like,

$$w = (1 - \beta^2)^{\frac{1}{2}} x_k + \beta \xi, \quad \xi \sim \mathcal{N}(0, C), \quad (1.7)$$

where C is a specified covariance operator (see 1.4.1). The tuneable parameter β is called the *random walk step size*, or the *proposal step size*. If $\beta = 0$, the Markov chain does not propose a jump in the state space and instead proposes $w = x_k$. Any reasonable acceptance probability calculation should almost surely give an acceptance probability of 1 in this case. When $\beta = 1$, the Markov chain takes no information of its current state and the acceptance probability should be some non-negative quantity less than or equal to 1.

Using this approach, one can draw samples from the posterior distribution, obtaining its shape exactly. This is of use when the posterior distribution is not a Gaussian and cannot be uniquely determined by its first and second moments. The application of MCMC methods to data assimilation is widespread. For examples of their use, see Cotter *et al.* [2012, 2009, 2010, 2011]; Stuart [2010]; Lee *et al.* [2011]; Apte *et al.* [2008b, 2007, 2008a]; Herbei *et al.* [2008]; Herbei & McKeague [2009]; McKeague *et al.* [2005]; Michalak [2003]; Kaipio *et al.* [2000]; Kaipio & Somersalo [2007]; Mosegaard & Tarantola [1995].

Metropolis-Hastings samplers have been researched extensively to assess and improve their performance greatly [Roberts, 1997; Roberts & Rosenthal, 1998, 2001; Beskos *et al.*, 2009] since their development by Metropolis *et al.* [1953] and Hastings [1970]. In 2005, the scientific community saw the introduction of adaptive approaches that drop the Markovian assumption but maintain the ergodicity property with improved convergence times [Atchadé & Rosenthal, 2005; Atchadé, 2006].

1.3.1 Adaptive burn-in

It is well understood that, initially, Markov chains sample poorly. To ameliorate this in practice, some large integer M is chosen and the first M samples from the chain are discarded. Though the choice of M is important, it depends heavily

on the seed of the chain and can be very hard to determine objectively. As a consequence it is often chosen by trial and error. For example, if the seed lies in the tail of the posterior distribution, it will take a longer time for the chain to reach stationarity than when seeded at a mode. Figure 1.2 illustrates this effect. Once

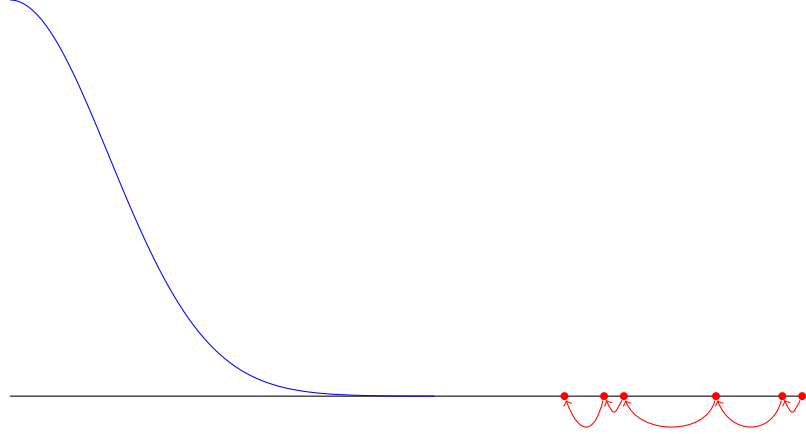


Figure 1.2: Initial transient behaviour of Markov chains

the chain has reached stationarity, there are theoretical results to justify tuning the proposal step size (1.7) to obtain an acceptance probability of roughly 25% [Roberts, 1997]. Though it is possible to tune the step size adaptively [Atchadé & Rosenthal, 2005; Atchadé, 2006], if it is not done during the burn-in the chain no longer satisfies the Markovian assumption and ergodicity is not guaranteed. To prevent this from happening we tune β adaptively only during the burn-in. There are many ways to tune β , and tuning too quickly during the burn-in has the result of a step size that is chosen to optimally sample parts of the state space where the chain exhibits transient behaviour and this is undesirable. The method we employ tunes more slowly. Given a period, T iterations, and a continuous range of acceptance probabilities, (α_1, α_2) , we adjust β by adding or subtracting a small increment, β_{inc} , to an initial β_0 according to the procedure outlined in figure 1.3. As discussed above, the smaller β is, the larger the resulting acceptance probability will be. This converse of this is also true. Our procedure aims to find a suitable β that will consistently give an average acceptance probability in the range (α_1, α_2) . This admissible range is chosen to be 20%-30%, consistent with the theory presented in Roberts [1997]. Values of β_{inc} and β_0 are choices that need to be made and throughout this work we choose them by trial and error. To see this method in action, figure 1.4(a) shows acceptance probabilities (blue dots) and their empirical mean (red line) during the burn-in period for some non-Gaussian distribution.

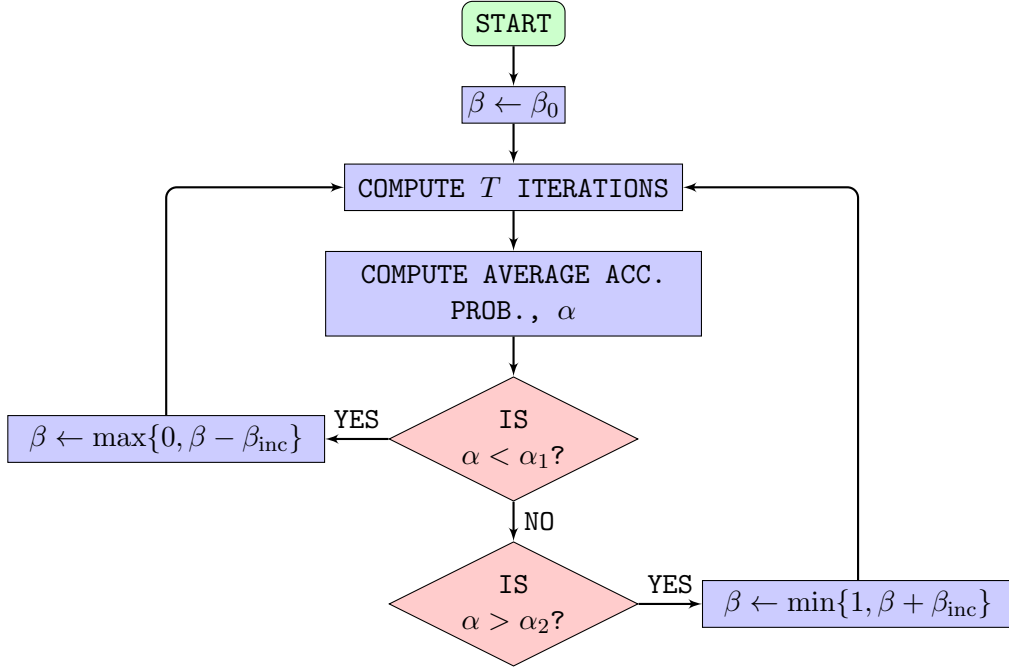
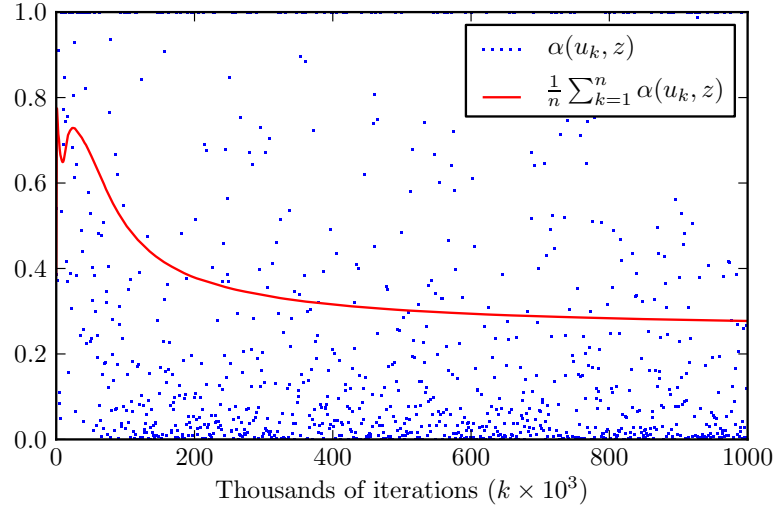


Figure 1.3: Flow chart detailing the adaptive scheme for β

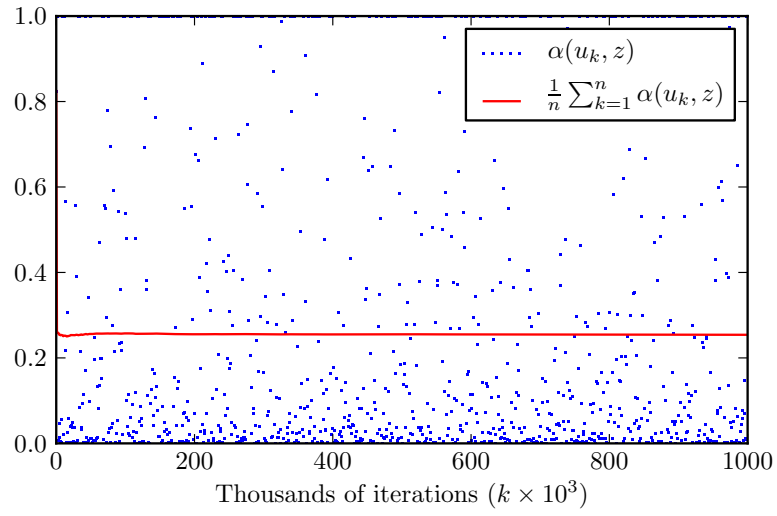
Observe that the initial step-size, β_0 , is too small and so the sampler takes sub-optimal jumps in the state space. The adaptive scheme adjusts the step-size to reach and maintain an acceptance probability of between 20% and 30%. After the burn-in, the adaptive scheme stops and the step-size is continued into the sampling part. Figure 1.4(b) shows that the acceptance probability maintains a steady value of about 25% after the burn-in.

1.3.2 Metastability

Looking at the form of the proposal for a random walk (1.7), it is clear that when $\beta \neq 1$, moves are based on the current state of the Markov chain. Within the context of nonlinear least square minimisation techniques, local moves often lead to problems when the cost functional has many local minima. The same problem occurs here within the context of Monte Carlo sampling. If the modes of the target distribution are far away from each other in the state space, one may observe *metastability*. Metastability is when the Markov chain gets stuck in one mode, sampling it well, and is completely oblivious to other modes in the state space that could possibly contribute a lot of probability mass (figure 1.5).



(a) Behaviour of the adaptive step-size scheme for $\beta_{\text{inc}} = 10^{-4}$



(b) After burn-in the acceptance probability settles

Figure 1.4: Acceptance probabilities for adaptive step-size scheme

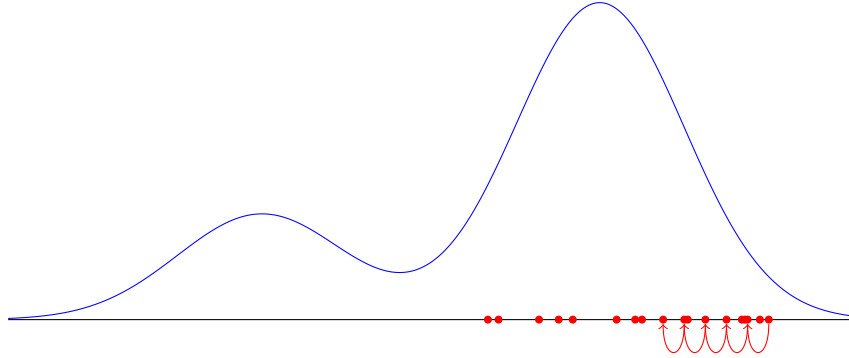


Figure 1.5: Illustration of metastability in MCMC samplers

1.3.3 A note on random numbers

Monte Carlo methods require the use of randomly generated numbers. Any correlation within the generated variates can lead to an impeded convergence speed and severely bias computed moments.

Definition 1.3.1. *Given some interval $[a, b]$, let $A = \mathbb{N} \cap [a, b]$. A pseudo-random number generator is a function $f : A \rightarrow A$. A seed for the random number generator is some $x_0 \in A$. Random numbers are produced by successively applying f to obtain a sequence $x_n = f(x_{n-1})$, $n = 1, 2, \dots$*

It is a property of all pseudo-random number generators that there exists $N \in \mathbb{N}$ such that $x_N = x_0$. In other words, pseudo-random number generators are *periodic*, and the smallest such N is called the period. It is no surprise then, that random number generators will, by construction, never generate ‘truly’ random numbers. One can only hope their output *appears* to be random. For this to hold, f should at least have a large period. For further sanity checks on randomness, a set of statistical tests have been devised to analyse various aspects of the output of pseudo-random number generators [Marsaglia, 1996].

One of the most common random number generator algorithms used is the Mersenne-Twister generator [Matsumoto & Nishimura, 1998]. It has a period of $2^{19937} - 1$. At this juncture it is worth noting that, since the period of this generator is so large, none of the numerical illustrations presented throughout this thesis required the full period of the Mersenne-Twister algorithm, even in an infinite dimensional setting. We made the choice to use a Tausworthe random number generator, whose computation is given here for the sake of completeness. Random numbers are given

by,

$$x_n = s_n^1 \oplus s_n^2 \oplus s_n^3,$$

where,

$$\begin{aligned} s_{n+1}^1 &= (((s_n^1 \& 4294967294) \ll 12) \oplus (((s_n^1 \ll 13) \oplus s_n^1) \gg 19)), \\ s_{n+1}^2 &= (((s_n^2 \& 4294967288) \ll 4) \oplus (((s_n^2 \ll 2) \oplus s_n^2) \gg 25)), \\ s_{n+1}^3 &= (((s_n^3 \& 4294967280) \ll 17) \oplus (((s_n^3 \ll 3) \oplus s_n^3) \gg 11)). \end{aligned}$$

The operators used above are defined as,

$\&$: bit-wise AND

\oplus : bit-wise XOR

\ll : bit-shift left (multiplication by 2)

\gg : bit-shift right (division by 2, rounded down).

The Tausworthe generator presented here has a period of 2^{88} . This is noticeably smaller than that of the Mersenne-Twister algorithm, but it is a small price to pay given the greatly reduced computational cost involved in producing random variates with this method. The review in Jones [2010] is a notable work on the best practices of generating random numbers.

Random number generators like these produce uniformly distributed nonnegative integers between some constructed bounds. It is often the case that one wants random samples from the standard normal distribution. This can be achieved using a transformation that maps uniformly distributed variates to Gaussian distributed variates. The Box-Muller transform [Box & Muller, 1958] is such a transformation, and one of the most widely used ones.

Theorem 1.3.2 (Box-Muller transform). *Let U_1 and U_2 be two independent random variables drawn from the uniform distribution on $[0, 1]$ then*

$$Z_1 = \sqrt{-2 \log(U_1)} \cos(2\pi U_2), \tag{1.8}$$

$$Z_2 = \sqrt{-2 \log(U_1)} \sin(2\pi U_2), \tag{1.9}$$

are two independent random variables with standard normal distribution.

Proof. The proof of this theorem can be found in in Box & Muller [1958]. We give a version here for the sake of completeness.

Let $f_{Z_1, Z_2}(z_1, z_2)$ be the joint probability density function of the pair (Z_1, Z_2) . Writing $u_1 = h_1^{-1}(z_1, z_2)$ and $u_2 = h_2^{-1}(z_1, z_2)$, we let $f_{U_1, U_2}(h_1^{-1}(z_1, z_2), h_2^{-1}(z_1, z_2))$ be the joint probability density function of the pair (U_1, U_2) . We use a standard change of variables relation,

$$f_{Z_1, Z_2}(z_1, z_2) = f_{U_1, U_2}(h_1^{-1}(z_1, z_2), h_2^{-1}(z_1, z_2)) |\det(J)|, \quad (1.10)$$

where,

$$J = \begin{pmatrix} \frac{\partial u_1}{\partial z_1} & \frac{\partial u_1}{\partial z_2} \\ \frac{\partial u_2}{\partial z_1} & \frac{\partial u_2}{\partial z_2} \end{pmatrix}.$$

We invert (1.8)–(1.9) to obtain u_1 ,

$$\begin{aligned} z_1^2 + z_2^2 &= -2 \log(u_1) \cos^2(2\pi u_2) - 2 \log(u_1) \sin^2(2\pi u_2) \\ &= -2 \log(u_1), \\ \therefore u_1 &= \exp\left(-\frac{1}{2}(z_1^2 + z_2^2)\right). \end{aligned}$$

Similarly, to obtain U_2 ,

$$\begin{aligned} \frac{z_2}{z_1} &= \tan(2\pi u_2) \\ \therefore u_2 &= \frac{1}{2\pi} \arctan \frac{z_2}{z_1}. \end{aligned}$$

The Jacobian has determinant,

$$\begin{aligned} |\det(J)| &= \left| \det \begin{pmatrix} -z_1 \exp\left(-\frac{1}{2}(z_1^2 + z_2^2)\right) & -z_2 \exp\left(-\frac{1}{2}(z_1^2 + z_2^2)\right) \\ -\frac{z_2}{2\pi(z_1^2 + z_2^2)} & \frac{z_1}{2\pi(z_1^2 + z_2^2)} \end{pmatrix} \right| \\ &= \frac{z_1^2}{2\pi(z_1^2 + z_2^2)} \exp\left(-\frac{1}{2}(z_1^2 + z_2^2)\right) + \frac{z_2^2}{2\pi(z_1^2 + z_2^2)} \exp\left(-\frac{1}{2}(z_1^2 + z_2^2)\right) \\ &= \frac{1}{2\pi} \exp\left(-\frac{1}{2}(z_1^2 + z_2^2)\right). \end{aligned}$$

Finally, substituting into (1.10) yields,

$$\begin{aligned} f_{Z_1, Z_2}(z_1, z_2) &= \frac{1}{2\pi} \exp\left(-\frac{1}{2}(z_1^2 + z_2^2)\right) \mathbf{1}_{\mathbb{R}}(z_1) \mathbf{1}_{\mathbb{R}}(z_2) \\ &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z_1^2\right) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z_2^2\right) \\ &= f_{Z_1}(z_1) f_{Z_2}(z_2). \end{aligned}$$

This is exactly two one-dimensional Gaussian probability density functions in both z_1 and z_2 . \square

The Box-Muller transform is a useful technique in generating standard normal deviates from uniform deviates, but requires the calculation of the elementary functions \log , \sin and \cos . These are expensive functions to calculate numerically. An alternative method for computing standard normal random variables is the Ziggurat method Marsaglia & Tsang [2000], which is a much cheaper computational approach.

There are a plethora of random number generation methods freely available to download for use by the wider community. The work presented in this thesis heavily uses Monte Carlo methods to compute moments and, as a consequence, extremely high quality random numbers are needed. Both the Tausworthe and the Mersenne-Twister generators come with the GNU Scientific Library [Galassi *et al.*, 2011] and produce high quality random numbers, so the choice to use this library was an easy one to make.

1.4 Rigorous mathematical setting

Here we introduce the Bayesian mathematical setting in which we solve data assimilation problems. This initial set-up will be finite dimensional to give the reader a gentle introduction to the main concepts. Most of what follows is adapted from Stuart [2010]. The reader should seek this work for a more general framework than the one given below.

Let X and Y be Banach spaces equipped with norms $\|\cdot\|_X$ and $\|\cdot\|_Y$ respectively. The space X is the space where the state of the system lives, and Y is the space where the observations live. We are given the map between them,

$$y = \mathcal{G}(x) + \eta, \tag{1.11}$$

Here \mathcal{G} is the *model*, x is the *state*, y is the *observation* and η is the *observational error*. The aim is to find

$$x^* = \operatorname{argmin}_{x \in X} \frac{1}{2} \|\mathcal{G}(x) - y\|_Y^2.$$

This minimisation can be problematic. In particular, it may lead to minimising sequences whose limit does not live in X . Instead, a common technique to overcome this issue is to regularise the minimisation by a penalty term. A very popular choice is the Tikhonov regularisation,

$$x^* = \operatorname{argmin}_{x \in E} \frac{1}{2} \|\mathcal{G}(x) - y\|_Y^2 + \frac{1}{2\mu^2} \|x - m\|_E^2. \quad (1.12)$$

Here $(E, \|\cdot\|_E)$ is some Banach space contained in X , and μ is a *regularisation parameter*. Note that several choices must be made. Namely, the choice of the norms $\|\cdot\|_Y$ and $\|\cdot\|_E$ needs to be made clear, they may depend on the map \mathcal{G} and also the practical setting of the problem. So far, what we have presented in this subsection looks variational without mention of any probability measures. The Bayesian approach can intuitively be obtained by applying an exponential transformation to the functional (1.12). More explicitly, we can view it as a probability density function,

$$\mathbb{P}(x|y) \propto \exp \left(-\frac{1}{2} \|\mathcal{G}(x) - y\|_Y^2 - \frac{1}{2\mu^2} \|x - m\|_E^2 \right). \quad (1.13)$$

It is easy to see that minimising (1.12) is equivalent to maximising (1.13).

We now develop the Bayesian approach from first principles. If η in (1.11) has probability density p then

$$\mathbb{P}(y|x) = p(y - \mathcal{G}(x)).$$

This is called the likelihood distribution. Let $\mathbb{P}(x)$ be a prior probability distribution with associated prior measure μ_0 on the state x . This distribution represents a belief about what x looks like. By Bayes' formula, the posterior distribution $\mathbb{P}(x|y)$ with associated posterior measure, μ^y , is given by,

$$\begin{aligned} \mathbb{P}(x|y) &= \frac{\mathbb{P}(y|x)\mathbb{P}(x)}{\int \mathbb{P}(y|x)\mathbb{P}(x) \, dx} \\ &\propto \mathbb{P}(y|x)\mathbb{P}(x). \end{aligned}$$

Since all the measures here are on finite dimensional spaces we can write down the Radon-Nikodym derivative of the posterior measure (denoted μ^y) with respect

to the prior measure (denoted μ_0),

$$\frac{d\mu^y}{d\mu_0}(x) \propto \mathbb{P}(y|x). \quad (1.14)$$

In finite dimensions, one usually writes down integrals with respect to Lebesgue measure, and multiplication by some probability density q in the integrand is a change of measure from Lebesgue measure to the measure q . If X and Y are infinite dimensional, it is not possible to write down measures with respect to Lebesgue measure. In Bayes' rule, the most natural choice of the reference measure is the prior measure μ_0 . Bayes' rule then states that the Radon-Nikodym derivative of the posterior measure μ^y with respect to the prior measure μ_0 is proportional to the likelihood measure. This is exactly (1.14) and it is this form of Bayes' rule that generalises to infinite dimensional spaces. For a formal commentary on infinite dimensional Gaussian measures, see Bogachev [1998].

1.4.1 Regularity of random fields

When dealing with the case where X and also potentially Y are infinite dimensional Banach spaces, the question of how to draw from distributions on these spaces becomes a pertinent one. One should choose the prior measure μ_0 on X such that $\mu_0(X) = 1$, so any draws we compute from μ_0 should be sufficiently regular that they live in X almost surely. Since all the priors throughout this thesis will be Gaussian, we will explore regularity properties of draws from Gaussian distributions on function spaces in terms of the eigenvalues of some covariance operator. Furthermore, we will deal with covariance operators that are fractional powers of the Laplacian. The domain of the Laplacian will be the two-dimensional torus, $\mathbb{T}^2 \subset \mathbb{R}^2$, with periodic boundary conditions. We define $\mathcal{H} \subset L^2_{\text{per}}(\mathbb{T}^2)$ as,

$$\mathcal{H} := \left\{ u \in L^2(\mathbb{T}^2) \mid \int_{\mathbb{T}^2} u \, dx = 0 \right\},$$

the set of mean zero square integrable functions periodic on \mathbb{T}^2 . Let $\{\lambda_k, \phi_k\}$ form a countable orthonormal basis for the separable Hilbert space \mathcal{H} comprising of eigenfunctions and eigenvalues of the Laplacian, $-\Delta$. Let $\mathbb{K} = \mathbb{Z}^2 \setminus \{0, 0\}$, then for $u \in \mathcal{H}$ we can write,

$$u = \sum_{k \in \mathbb{K}} \langle u, \phi_k \rangle \phi_k.$$

From this we can define fractional powers of the Laplacian as,

$$(-\Delta)^\alpha u = \sum_{k \in \mathbb{K}} \lambda_k^\alpha \langle u, \phi_k \rangle \phi_k.$$

Now, for $s \in \mathbb{R}$, we may define the separable Hilbert spaces H_{per}^s by,

$$H_{\text{per}}^s := \left\{ u \in \mathcal{H} \mid \sum_{k \in \mathbb{K}} \lambda_k^s |\langle u, \phi_k \rangle|^2 < \infty \right\},$$

equipped with the norm,

$$\|u\|_s^2 = \sum_{k \in \mathbb{K}} \lambda_k^s |\langle u, \phi_k \rangle|^2.$$

Note that when $s = 0$, by Parseval's theorem u is square integrable and we get back the space $L_{\text{per}}^2(\mathbb{T}^2)$.

For the specific case of the Laplacian operator above, we have $\phi_k(x) = \exp(2\pi i k \cdot x)$ and $\lambda_k = 4\pi^2 |k|^2$. Now we wish to construct a random function that lives in H_{per}^s almost surely. For this we use the Karhunen-Loève expansion,

$$\xi(x) = \sum_{k \in \mathbb{K}} \frac{\gamma_k}{(4\pi^2 |k|^2)^{\alpha/2}} \exp(2\pi i k \cdot x), \quad \gamma_k \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1). \quad (1.15)$$

To show almost-sure regularity, we have the following theorem.

Theorem 1.4.1. *If $\alpha > 1 + s$ then $\xi \in H_{\text{per}}^s$ almost surely.*

Proof. It is sufficient to show $\mathbb{E}(\|\xi\|_s^2) < \infty$,

$$\begin{aligned} \mathbb{E}(\|\xi\|_s^2) &= \mathbb{E} \left(\sum_{k \in \mathbb{K}} (4\pi^2 |k|^2)^s \frac{|\gamma_k|^2}{|4\pi^2 |k|^2|^\alpha} \right) \\ &= \mathbb{E} \left(\sum_{k \in \mathbb{K}} (4\pi^2 |k|^2)^{s-\alpha} |\gamma_k|^2 \right) \\ &= \sum_{k \in \mathbb{K}} (4\pi^2 |k|^2)^{s-\alpha} \mathbb{E} |\gamma_k|^2 \\ &= \sum_{k \in \mathbb{K}} (4\pi^2 |k|^2)^{s-\alpha}. \end{aligned}$$

In two dimensions, this sum is finite since $s - \alpha < -1$. □

The numerical recipe for generating random draws with s weak derivatives is illustrated in algorithm 1.

Data: number of grid points in x and y directions: n_j, n_k

Data: regularity parameter: α

Result: random function with $\alpha - 1$ weak derivatives

```
1 for  $j \leftarrow 1$  to  $n_j$  do
2   for  $k \leftarrow 1$  to  $n_k$  do
3      $\gamma \leftarrow \text{RandomNormal}(0, 1)$ ;
4      $\hat{u}[j, k] \leftarrow \gamma / (4\pi^2(j^2 + k^2))^{\alpha/2}$ ;
5   end
6 end
7  $u \leftarrow \text{InverseFFT}(\hat{u})$ ;
8 return  $u$ 
```

Algorithm 1: Drawing random functions

Though all the theory above has been stated with only inverse powers of the Laplacian in mind, this is not the *only* choice of covariance operator available to us. Choosing a suitable covariance operator requires thought about what properties are needed in the prior distribution. The Laplacian operator is convenient here because its L^2 basis functions are periodic, preserving the boundary conditions imposed in the models we explore in this thesis. Furthermore, it is diagonal in Fourier space, making draws from the associated prior distribution cheap to construct. Other, invertible and trace-class, operators may be used. For example, to not restrict oneself to mean zero functions, the operator $(I + \Delta)$ can be implemented. Its basis functions are still periodic, preserving the modelling domain. As a general heuristic, the basis of eigenfunctions of the covariance operator should reflect modelling assumptions and assumptions in the structure of prior draws. A basis of Haar wavelets leads to prior draws with discontinuities, useful for preserving edges in images or shocks in ocean waves. Regularity of prior draws is controlled by how quickly the eigenvalues of the covariance operator decay. This can be adjusted by raising the covariance operator to some power.

The theory above is also all done in an infinite dimensional setting. Numerically and operationally, a finite dimensional approximation is made. In the case of the Karhunen-Loève expansion this approximation is done by truncation. A choice must be made in where to truncate, and this choice coincides with a modelling assumption – that there are no frequencies of order larger than the truncation wavenumber. If it is feasible that solutions to the inverse problem do in fact admit higher-order frequencies, it is necessary to rethink this assumption. Throughout this thesis the data and initial conditions are known and the truncation is chosen to be much larger than necessary to mitigate the effects of poor modelling assumptions. As a concrete

example, in the inverse problem for a two dimensional advection equation, we choose the initial condition to consist of a linear combination three sinusoidal functions. We truncate the Karhunen-Loève expansion at 2^5 terms, an order of magnitude larger than is required. Practically, the initial condition to one's problem is unknown. In this scenario, care and diligence are necessary traits in choosing appropriate prior assumptions.

1.5 Thesis summary

This thesis is divided into four chapters. The first chapter has two aims, the first of which is to give a brief overview of the history and types of data assimilation for the reader's benefit. This puts into perspective the aims of data assimilation. The second aim is to provide the necessary general framework in which the mathematical and numerical content resides.

The second chapter concerns the Bayesian inverse problem for a simple linear two dimensional advection partial differential equation with periodic boundary conditions. We divide this into several parts, each with its own purpose. First, we seek to find the initial condition of the linear advection equation from noisy Eulerian observations of the discretised field at a series of times. This is a linear problem and the associated posterior distribution is Gaussian, characterised uniquely by its first two moments. This case is explored as a sanity check that the numerical scheme set in place to probe the posterior distribution is functioning correctly. We explore the effects on the mixing properties of the Markov chain as a function of random walk step size and observational error.

Secondly, we seek to find the wave velocity parameter in the PDE. This is a non-Gaussian problem. We expose the problems associated with nonlinear data assimilation when utilising a Markov chain Monte Carlo sampling method to explore the posterior distribution, observing a multitude of metastable states. We attempt to solve the problems associated with metastability by implementing a simulated annealing method.

The keystone of the second chapter is the accountability of model error within data assimilation. This is an active research topic. We numerically characterise the shape of first moment of the posterior distribution as a function of the model/data mismatch; the wave velocity error. To complement these numerical results, we have quoted four theorems that were proven by Lee in Lee *et al.* [2011]. These give an

explicit and analytic characterisation with associated rates of convergence, the proofs to which are not provided here. The characterisation of the posterior mean in the limit of infinite observed data is as follows. If the wave velocity error is irrational the posterior mean is the spatial average of the true initial condition. A rational wave velocity error of $1/q$ results in a posterior mean constructed from every q^{th} Fourier mode. Finally, and trivially, if the wave velocity error is zero then the posterior mean is exactly the true initial condition. This work structurally identifies everything about the first moment of the posterior distribution in the advent of model error. We extend this work to the joint distribution on both the initial condition and the wave velocity, utilising a Metropolis-within-Gibbs method to probe the associated posterior. We solve the problem of Markov chain metastability by application of a least-squares technique on the data to obtain estimate of the wave velocity and use this to seed the MCMC scheme. As a result of this seeding procedure, we successfully overcome metastability and, in the large data limit, observe convergence of the posterior measure to a Dirac centred at the truth.

Lastly, and related to the issue of model error, we provide numerical results when a non-smooth likelihood norm is utilised over the initial condition. This problem is also non-Gaussian but with a linear forward operator. The non-Gaussianity arises from assuming the log-likelihood grows only linearly in the tails. This is equivalent to a doubly-exponential likelihood distribution of the data/model mismatch. The purpose of this section is then twofold: expose MCMC as a flexible tool that can deal easily with non-Gaussian infinite dimensional inverse problems; and show that by utilising a doubly-exponential likelihood, a larger proposal step is admissible. This leads to more efficient state space exploration.

The third chapter concerns Lagrangian data assimilation for controlled drifters, the underlying flow for which is a perturbed kinematic travelling wave model. This chapter is divided into two main parts. The focus of the first is the time-independent unperturbed part of the flow model. We measure performance of the addition of a ‘control’ to fluid flow drifters by looking at the posterior variance on the velocity field. We show two main results. When the fluid flow drifter is trapped in a recirculation regime, the magnitude of the control is the main player in pushing the drifter out of the eddy. We show that, for a relatively simple control, when the magnitude is large enough a significant reduction in the posterior variance is achieved. The second result illustrates the effect of using posterior information from a previous assimilation cycle. Here the control magnitude corresponds geometrically to the distance between the drifter and a hyperbolic fixed point of an eddy transport bar-

rier. As the control magnitude increases, the drifter gets closer to the fixed point and a decrease in variance is observed. The second part of the third chapter involves the perturbed time-periodic flow model. Applying the same series of controls as in the first part of the third chapter, we show two main results. On a high level, the first result illustrates robustness of the posterior variance with respect to the perturbation parameter. More specifically, its structure as a function of control magnitude is carried over from the time-independent flow model. Moreover, we observe an additional, and separate, decrease in posterior variance corresponding to the purely time-dependent part of the flow. The second result aims to fairly represent the effects of controlling drifters. If the passive drifter does a reasonable job of exploring ‘interesting’ flow structures, eddies and hyperbolic fixed points, for example, then it is sometimes better not to utilise any control strategy.

The fourth chapter partially extends the work set out in the third chapter, concerning the application of cheap-to-compute controls to a testbed kinematic travelling wave model. The effect of each control on the associated posterior distribution on the underlying flow is analysed for a geometric correspondence between flow structure and posterior variance. Pushing the drifter out of an eddy yielded a net gain in information on the flow. Instead, there could be more to gain by choosing a specific point in the domain where the drifter should end up. Moreover, minimisation of the effort needed to reach such a terminal point is seen as a more challenging but realistically practical goal. For example, to see a reduction in posterior variance, one possibility would be to control an ocean drifter to a local maximum of the posterior variance from a previous assimilation cycle. This allows for observations to be collected in a part of the flow we are uncertain about. An approach of this type cannot be executed by use of simple cheap-to-compute controls as in the third chapter. As soon as the drifter reaches the relevant part of the domain, the flow would instantly push it away. This chapter, comprised of three sections, aims to pose minimum-cost control strategies within the Bayesian framework for data assimilation as a basis for more complicated uncertainty quantification.

The first section introduces the theoretical nature of optimal control problems on a high level. Heavily inspired by Bryson Jr. & Ho [1975], we derive the Hamilton-Jacobi-Bellman (HJB) equation for an optimal feedback control with a general cost function. Hamilton-Jacobi-Bellman equations, though useful, are often difficult to solve directly. They involve a global pre-determined grid of points on which the optimal cost-to-go function is computed.

The second section sees the application of the theory to a specific problem framed in

an oceanographic context. Here we use a specific cost function, that of minimising the time to reach a terminal point in the domain. This is a practically inspired cost function in light of the results presented in the third chapter. Choosing the terminal point to be in a new flow regime and getting there in minimum time allows for the collection of observations to happen sooner. The practical implications of such an objective are very clear. We go further by applying an algorithm due to Rhoads [Rhoads *et al.*, 2010] to obtain necessary conditions for an extremum of the HJB equations; the Euler-Lagrange equations. From the point of view of implementation, the Euler-Lagrange equations relax the requirement that the cost-to-go surface be computed over the whole domain. A local method such as this gels well with the framework of data assimilation applied to problems in the ocean and the heavily localised observations thereof. This should be a stepping stone for executing more complicated control strategies than those explored in the third chapter.

The third section presents the necessary workflow to execute the minimum time control algorithm within a Bayesian framework. Implications of such a complicated control construction are illustrated here. More specifically, Markov chain Monte Carlo methods are a state-of-the-art method to solve problems in data assimilation, but typically require a large number of samples to adequately compute posterior moments. We show that this state-of-the-art method does not exhibit an avenue for which clever control methods can be computed cheaply. For each sample, ocean drifter positions are integrated over the, potentially multivalued, cost-to-go surface. We explain two approaches to making this cheaper: reducing the number of draws from the posterior distribution; and computing less trajectories of the Euler-Lagrange equations. This exposes a trade-off between sampling error and control error.

The numerical studies and discussion done in sections 2.4 and 2.5.3, and related conclusions in section 2.7, have been published jointly with Lee and Stuart in Lee *et al.* [2011]. The efforts in chapter 3 are not yet published, but are a work in preparation with Jones in McDougall & Jones [2012].

Chapter 2

Data assimilation for the advection equation

2.1 Overview

Throughout this chapter the model we study is the equation for linear advection in two dimensions. We observe the solution to this equation at fixed points in space for a series of times. Given these observations, the aim of this chapter is to explore three related inverse problems: solve the inverse problem on the initial condition, the wave velocity and the joint, respectively. We pose each of these inverse problems in a Bayesian framework, as described in section 1.4. The first case is addressed in section 2.2, it is a Gaussian problem and the associated posterior distribution is characterised uniquely by its first two moments. The inverse problem on the initial condition is solved to ensure that the numerical scheme set in place to probe the posterior distribution is functioning correctly. This is achieved by exploring the effects on the mixing properties of the Markov chain as a function of random walk step size and observational error. Recall that since this problem is Gaussian, sampling from the posterior distribution can be done exactly, without Markov chain Monte Carlo. We utilise an MCMC method, even in this Gaussian case, for two reasons. Firstly, samples from the posterior distribution are obtained by drawing from a simple Gaussian prior distribution whose covariance operator is diagonal in Fourier space. The posterior covariance does not exhibit any diagonal structure. Secondly, the posterior distribution is very high dimensional. As such, methods like the Kalman filter that are exact are also computationally intractable due to the necessary storage of a large covariance operator.

In section 2.3 we seek to identify the wave velocity parameter; this is a non-Gaussian Bayesian inverse problem due to the nonlinearity of the forward operator which maps the wave speed to the observations. We expose the reader to problems associated with nonlinear data assimilation when utilising a Markov chain Monte Carlo sampling method to explore the posterior distribution. We see the Markov chain exhibits metastability. We utilise a standard method, simulated annealing, to move the sampler to a different mode of the posterior distribution. This increases state space coverage but is computationally expensive.

Accountability of model error within data assimilation is illustrated in section 2.4 with four main theorems. The theorems, and associated proofs, are due to Lee [Lee *et al.*, 2011]. They explicitly characterise the shape of first moment of the posterior distribution explicitly as a function of the model error mismatch; the error in the wave velocity. The numerical commentary in this section is due to McDougall [Lee *et al.*, 2011] and justifies Lee’s theory. In the limit of zero observational error, if the wave velocity error is irrational the posterior mean is the spatial average of the true initial condition. A rational wave velocity error of $1/q$ results in a posterior mean constructed from every q^{th} Fourier mode of the true initial condition. Finally, if the wave velocity error is zero then the posterior mean is exactly the true initial condition that generated the data. This work characterises, in its entirety, the first moment of the posterior distribution in the advent of model error.

Usually, when one refers to ‘error’ in the data assimilation community, one of three possible things are being discussed: a) error in the model; b) error in the model *parameters*; or c) error in the *observations*. Model error is canonically represented by a stochastic term added on as an extra term in the PDE. Parameter error refers to any errors made in the *parameters* in the PDE and observation error refers to the errors made upon observing a certain quantity. This quantity may or may not be an output of the model. Since we do not explore the addition of white noise onto the PDEs presented in this thesis, the terms *model error* and *parameter error* are considered interchangeable.

Section 2.5 extends the application of assimilating with model error to the joint distribution on both the initial condition and the wave velocity. A Metropolis-within-Gibbs method is implemented to sample the associated posterior and we discuss several methods on how to seed the Markov chain to ameliorate metastability. Subsection 2.5.3 is a notable and novel method in achieving this by application of a least-squares technique on the noisy data to obtain a crude estimate of the wave speed. Although rudimentary, this approach leads to good convergence of the

posterior distribution to a Dirac measure centred on the truth in the large data limit.

Related to the issue of model error, section 2.6 provides numerical results when a non-smooth likelihood norm is utilised over the initial condition. This problem is non-Gaussian but admits a linear forward operator. The non-Gaussianity arises from assuming the log-likelihood grows only linearly in the tails. This is equivalent to a doubly-exponential likelihood distribution of the data/model mismatch. The purpose of this section is then twofold: expose MCMC as a flexible tool that can deal easily with non-Gaussian infinite dimensional inverse problems; and show that by use of a doubly-exponential likelihood, a larger proposal step is admissible, leading to more efficient state space exploration.

2.2 Sampling the initial condition

Suppose we are given a model to describe the time behaviour of some physical quantity, for example the propagation of a wave in a fluid. If we are given the initial condition then we can integrate the model to obtain all future states of the quantity of interest at any time we may specify. This is commonly referred to as ‘the forward problem’. For a linear advection model, the forward problem says that given the linear advection equation, including initial condition u and wave velocity c ,

$$\text{(PDE)} \quad \frac{\partial v}{\partial t} = c \cdot \nabla v, \quad t > 0, \text{ and} \quad (2.1a)$$

$$\text{(IC)} \quad v(x, 0) = u(x), \quad (2.1b)$$

find the advected field $v(\cdot, t)$ for $t > 0$.

The prior distribution

Suppose we do not know u exactly but are instead given two other pieces of information. The first piece is a *prior* on u . A prior distribution is a probability distribution that represents some initial belief about what u looks like. Practically this distribution is: a) given to us; b) constructed ourselves by utilising expert knowledge; or c) a guess. For our purposes, we will choose the prior to be the infinite dimensional Gaussian measure $\mathbb{P}(u) = \mathcal{N}(0, (-\Delta)^{-\alpha})$ and all of our numerical experiments in this chapter shall set $\alpha = 2$. To draw from the prior distribution, recall the discussion in 1.4.1.

The likelihood distribution

The second piece of information we are given are noisy observations $(y_{j,k})$ of the solution (v) to (2.1a) at points (x_j, t_k) for $j = 1, \dots, N$ and $k = 1, \dots, K$, so that

$$y_{j,k} = v(x_j, t_k) + \eta_{j,k}, \quad \eta_{j,k} \sim \mathcal{N}(0, \sigma^2), \quad (2.2a)$$

$$\rightsquigarrow y = \mathcal{G}(u) + \eta, \quad \eta \sim \mathcal{N}(0, \sigma^2 I_{JK}). \quad (2.2b)$$

For now, we are thinking of the wave velocity c as known. With the model and the data in hand, we seek the initial condition. This set-up is now complete and fits into the framework outlined in 1.4.

The posterior distribution

The solution to this inverse problem is a probability distribution $\mathbb{P}(u|y)$. Schematically, the posterior is proportional to $\mathbb{P}(y|u)\mathbb{P}(u)$, both of which are known distributions. The discussion in 1.4.1 outlines how to draw samples from the prior. To draw samples from the posterior, we implement the random walk Metropolis-Hastings algorithm illustrated in 1.3. Specifically, we draw samples, ξ , from the prior distribution using the Karhunen-Loève expansion (1.15) and construct a Markov chain $\{u_n\}_{n \in \mathbb{N}}$ whose invariant measure is the posterior.

In what follows, plots are provided giving evidence of the correctness of the code and robustness of the algorithm.

2.2.1 Varying step-size and observational error

Here we explore samples from $\mathbb{P}(u|y)$ where the true initial condition $u(x_1, x_2)$ is

$$u(x_1, x_2) = \sin(2\pi x_1) \cos(2\pi x_2). \quad (2.3)$$

The true wave velocity we will use is $c = (0.5, 1.0)$. By default, we will observe the solution at integer times. Note that, on the unit square with periodic boundary conditions, the solution to the advection equation is time-periodic with period $T = 2$. Hence, every second observation in time will be a repeated version of the field, with a different realisation of the noise added.

As one changes the value of β , the posterior density should remain unchanged and,

as such, one can test the algorithm code by varying β with all other parameters fixed. We will look at two computed quantities, the first of which is, $\|u_k\|_{L^2}^2$ as a function of sample number k (figures 2.1–2.3). We also look at the negative log-likelihood, $\Phi(u_k) := \frac{1}{2\sigma^2} \|\mathcal{G}(u_k) - y\|^2$ (figures 2.4–2.6).

We keep the number of observation points fixed at $N = 1024$, the number of observation times $K = 50$ and the number of iterations at 10^6 . Figures 2.1(a), 2.2(a) and 2.3(a) below show the qualitative difference in rate of convergence of the Metropolis-Hastings sampler for $\beta = 0.01, 0.02$ and 0.05 respectively. As we can see, the algorithm performs best in figure 2.1(a); the samples seem to explore the state space with few large periods of rejections, depicted by the blue line. The norm of the truth (2.3) is $\|u\|_{L^2}^2 = 0.25$ and is depicted by the green line. Comparing figure 2.1(a) with figures 2.2(a) and 2.3(a), we notice that the rate of convergence becomes considerably slower, with an increasing number of large periods of time where the algorithm rejects proposed samples. The chain gets stuck in certain areas of the state space. This behaviour can be explained by noticing that, as $\beta \rightarrow 1$, the resulting proposal (1.7) converges to a draw from the prior, retaining no information about the current state of the chain. Therefore, for larger β , draws from the proposal distribution are less likely to explain the observed data and are more likely to be rejected. Exploration of the state space can be improved, leading to a more efficient algorithm, by increasing the observational noise σ . There is a price to pay for this increase in performance. There are more possibilities from the prior that could explain noisier observations, this is entirely intuitive. As a consequence, the sampler may wander further away from the true initial condition. This can be seen by comparing figures 2.1(a) and 2.1(b). Fixing $\sigma = 0.1$, the behaviour of the sampler as we increase β (comparing 2.1(b) with figures 2.2(b) and 2.3(b)) is much less dramatic than when $\sigma = 0.01$. This is due to the sampler being able to explore the state space more easily when the observations are noisier.

Figures 2.4, 2.5 and 2.6 show $\Phi(u_k) := \frac{1}{2\sigma^2} \|\mathcal{G}(u_k) - y\|^2$ for the same choice of parameters as above. We see very similar behaviour here as well. As β increases (figures 2.4(a)–2.6(a)), more rejections occur as a consequence of taking larger jumps in the state space. We also see a trade-off; when the chain *does* accept a sample, we observe faster convergence to the posterior mode due to this larger jump. Additionally, as σ is increased (comparing figures 2.4(a) and 2.4(b)), the chain converges to a mode more quickly. The price to pay by increasing σ is exactly as in the first case, the chain will settle down in a mode where Φ is further away from zero; farther from the truth. In summary, the qualitative behaviour between these two trace

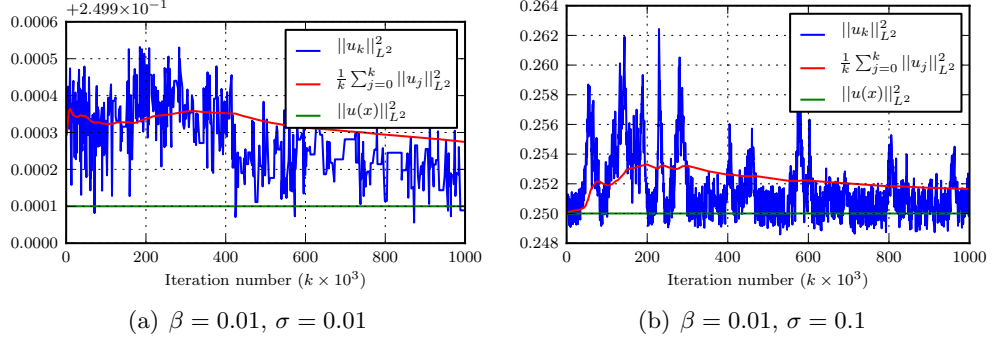


Figure 2.1: Trace plots showing effect of varying observational noise for $\beta = 0.01$. When the observational error is larger (right), the posterior is less tightly peaked and the sampler explores more of the state space.

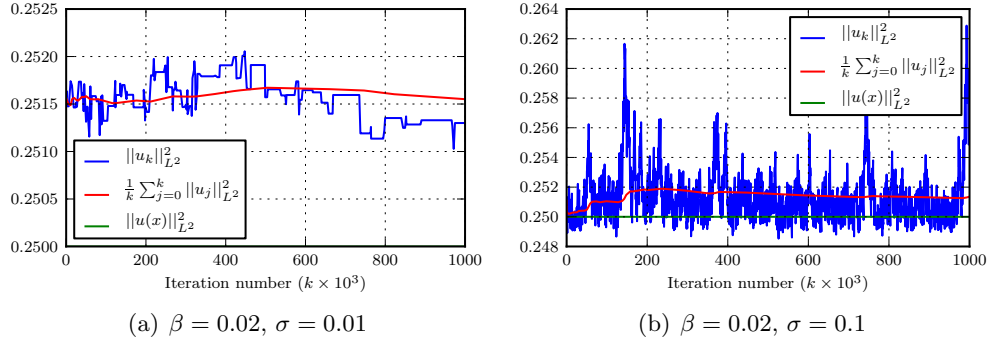


Figure 2.2: Trace plots showing effect of varying observational noise for $\beta = 0.02$. Comparing with figure 2.1, notice that in this case, where β is larger, the sampler ‘sticks’ more and samples the state space poorly (most noticeable on the left).

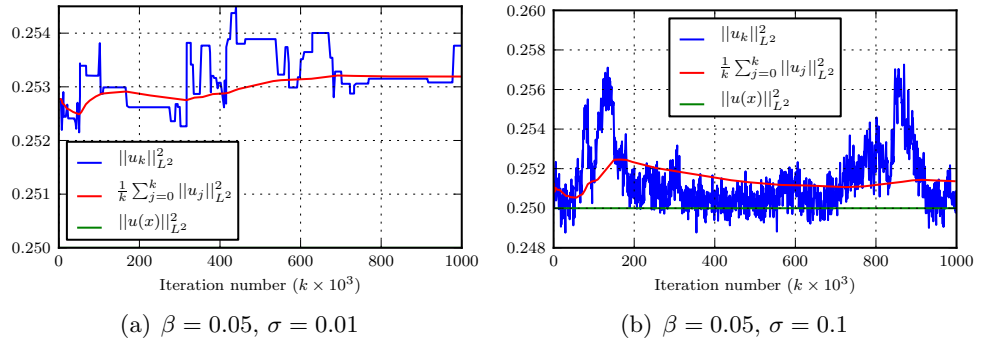


Figure 2.3: Trace plots showing effect of varying observational noise for $\beta = 0.05$. For this even larger value of β (comparing with figures 2.1 and 2.2), the sampler performs poorer still, rejecting lots of samples.

plots is very similar with convergence results easily analysed. In a practical set-

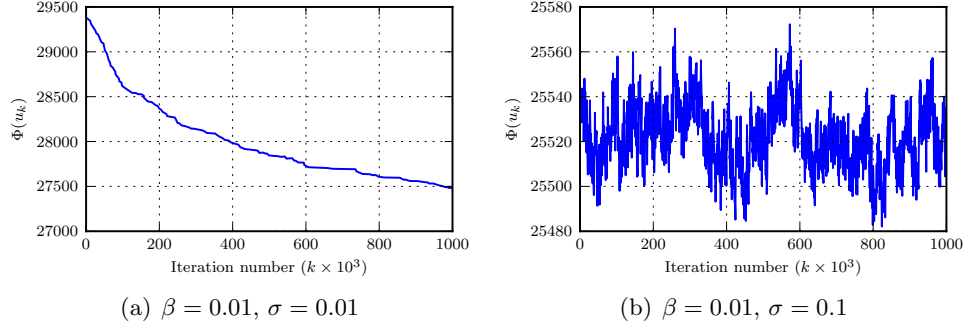


Figure 2.4: Trace plots showing effect of varying observational noise for $\beta = 0.01$. This figure is analogous to figure 2.1, but showing Φ instead of the acceptance probability.

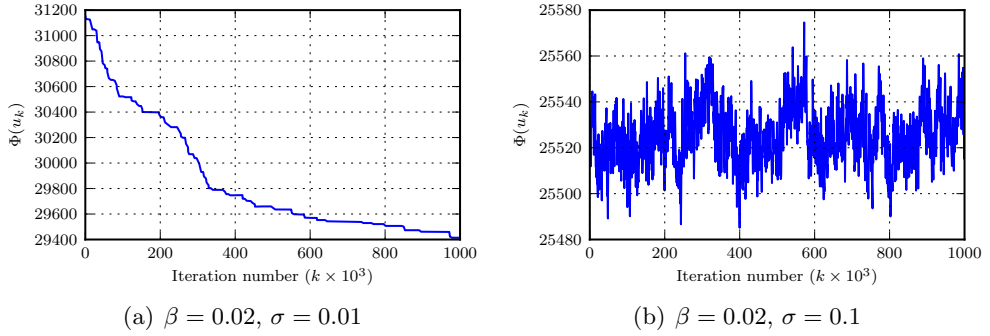


Figure 2.5: Trace plots showing effect of varying observational noise for $\beta = 0.02$. This figure is analogous to figure 2.2, but showing Φ instead of the acceptance probability.

ting, estimating σ is a hard problem and depends on the accuracy of the observing equipment. Theoretical and numerical advancements have been made in estimating σ in the context of sampling applied to imaging. The reader is directed to Cotter *et al.* [2012] for discussion on observational noise recovery through its treatment as a hyper-parameter. We have seen how varying σ affects the convergence properties of the Markov chain. Careful consideration is needed when making such adjustments. The Markov chain parameter β is something that can be tailored more frugally. As an example, in Pillai *et al.* [2012], analytical results are obtained concerning the limiting behaviour of the algorithm as $\beta \rightarrow 0$. In many practical applications, posterior distributions are potentially non-Gaussian and it may be the case that β should be varied adaptively according to the current average acceptance rate of the sampler.

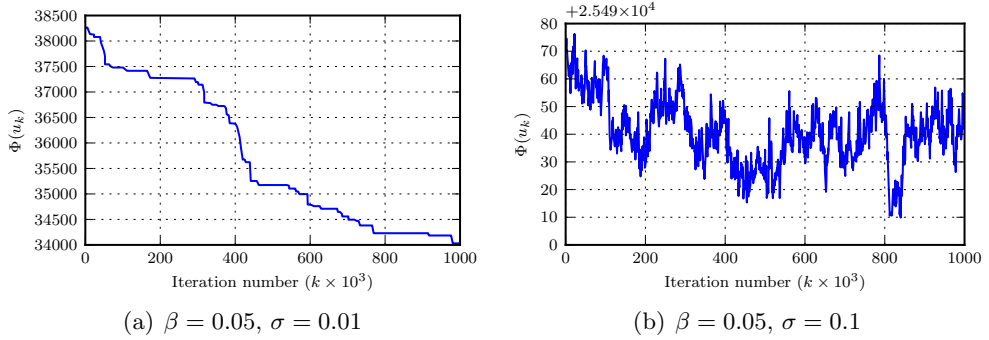
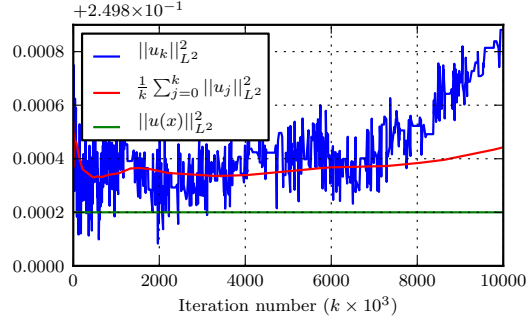


Figure 2.6: Trace plots showing effect of varying observational noise for $\beta = 0.05$. This figure is analogous to figure 2.3, but showing Φ instead of the acceptance probability.

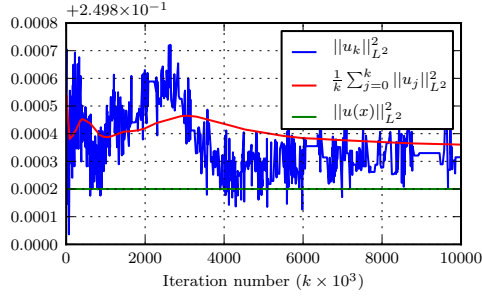
Under linearity and Gaussianity assumptions, the choice of β on the acceptance rate and state space exploration properties of Markov chains is well understood [Roberts, 1997; Beskos *et al.*, 2009; Atchadé & Rosenthal, 2005; Atchadé, 2006]. It is widely accepted that the optimal acceptance rate should be around 25% in high dimensional state spaces. Though not technically applicable, it is commonplace to apply these results in practice.

2.2.2 Varying the seed and sample size

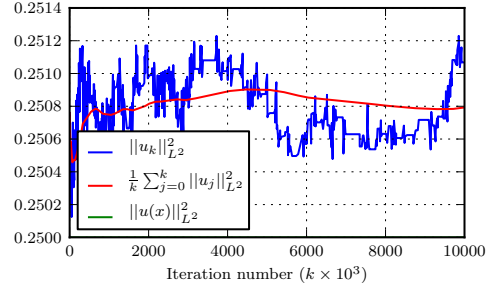
In this section we extend the sample size to 10^7 iterations and, keeping the same number of observation points and observational noise as before, we provide the three plots (figure 2.7). These plots show Markov chain realisations starting from different random draws from the prior. For Gaussian distributions the sampler should settle down in the same posterior mode. This illustrates that the implemented numerical algorithm is functioning correctly. As we can see, this robustness is achieved. Note that it is often the case one may see statistical artefacts in the sampler, especially when β is large. For example, in figure 2.7(a) it appears as though the chain wanders away from the truth. This is clearly undesirable behaviour. As a sanity check, we decrease β to 0.0005 (figure 2.8), and observe that for three different seeds (figures 2.8(a)–2.8(c)), these Markov chains explore the same posterior mode.



(a) $\beta = 0.01, \sigma = 0.01$

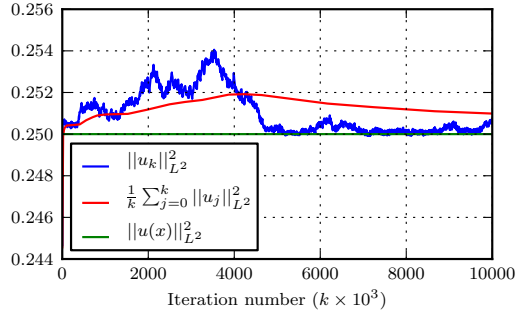


(b) $\beta = 0.01, \sigma = 0.01$

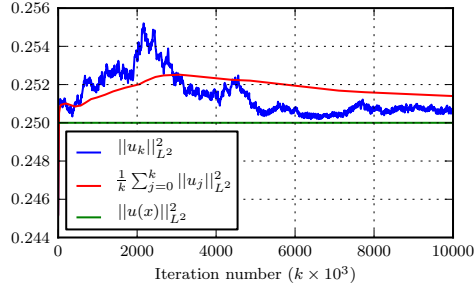


(c) $\beta = 0.01, \sigma = 0.01$

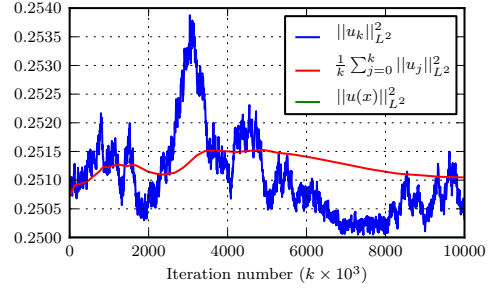
Figure 2.7: Trace plots showing effect of varying seed and lengthening run. Each of (a), (b) and (c) show the chain starting from a different seed. We see that the chain exhibits robustness, i.e., it doesn't explore a different mode. There are periods where the chain rejects a lot of samples and conclude it is necessary to decrease β .



(a) $\beta = 0.0005, \sigma = 0.01$



(b) $\beta = 0.0005, \sigma = 0.01$



(c) $\beta = 0.0005, \sigma = 0.01$

Figure 2.8: Trace plots showing effect of varying seed and lengthening run. Three differently seeded chains, with a smaller value of β than in figure 2.7. There are no noticeable periods where lots of samples are rejected.

2.2.3 Varying the number of observations

In this section we analyse the effect of varying the number of observations in two regimes. The first regime is to vary the number of observation locations, N , whilst keeping the number of observation times, K , fixed. The second is to vary K whilst keeping N fixed. In both cases, we track the behaviour of one Fourier coefficient as a function of sampler iteration number. The aim is to study which values it takes most often and how it explores the state space. This behaviour is shown in a histogram. In the first regime, the normalised histogram in figure 2.9(a) shows the real part of Fourier coefficient $(0, 1)$, $\Re(\hat{u}_{(0,1)})$, with N varying and $K = 50$ fixed. Figure 2.9(b) is a histogram of the same Fourier mode but for K varying and N fixed at 1024. The first point to note here is that each histogram neatly approximates a Gaussian; this is expected behaviour with linear model and Gaussian observation error. The second point to make is that the convergence of the histogram as one increases K is much slower than that induced by increasing N . This is a property of the model. The solution to the linear advection equation on a torus is periodic and it is the case that one observes the same values of the solution even though the solution is propagated in time (an effect called *aliasing*). The third and last point to note is that the real part of Fourier coefficient $(0, 1)$ of the true initial condition (2.3) is exactly zero. Both figures 2.9(a) and 2.9(b) show convergence to a Dirac measure centred at this value.

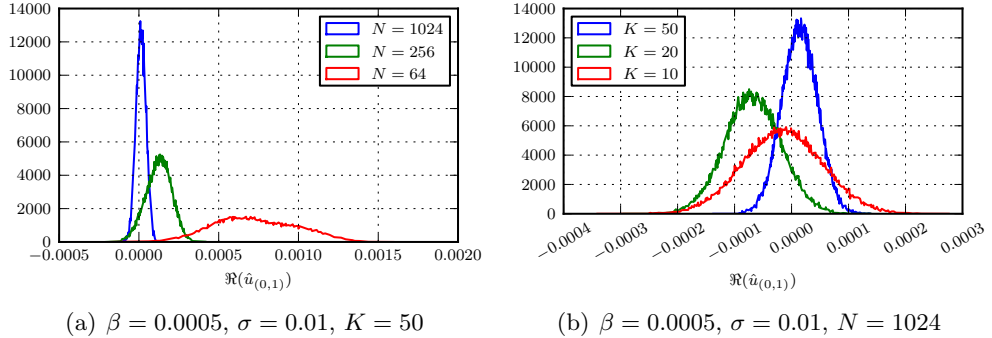


Figure 2.9: Histograms showing effect of varying the number of observations points, N , and times, K . Notice the convergence to a Dirac measure on the true value as the number of observations, in either space or time, increases.

2.3 Sampling the wave velocity

This section follows the ideas presented in Section 2.2, though here we approach the data assimilation problem from the perspective of recovering the wave speed, c . Throughout all of the numerical experiments shown below, u is fixed at the true initial condition (2.3). The set-up is exactly the same as in 2.2 with the exception that (2.2b) acknowledges that \mathcal{G} is instead a function of the wave velocity,

$$y = \mathcal{G}(c) + \eta, \quad \eta \sim \mathcal{N}(0, \sigma^2 I). \quad (2.4)$$

The number of spatial observations will be fixed at $N = 256$ and the number of temporal observations at $K = 50$. The aim is to sample the distribution $\mathbb{P}(c|y)$.

We will use the same form of proposal as in (1.7),

$$z_c = (1 - \beta_c^2)^{\frac{1}{2}} c_k + \beta_c \zeta, \quad \text{some } \beta_c \in (0, 1) \quad (2.5)$$

where $\zeta \sim \mathcal{N}(0, 0.1^2 \times I_2)$. Note that since $c \in \mathbb{R}^2$, (2.5) is a standard finite dimensional random walk proposal.

In section 2.2 the forward map $\mathcal{G}(u)$ is a *linear* function of u and hence the posterior distribution $\mathbb{P}(u|y)$ is Gaussian. Furthermore, its mean and covariance are given analytically by the Kalman filter formulae (1.1) and (1.2). A key difference to note here is that (2.4) is a *nonlinear* function of c . This plays an important role in the performance of the Markov chain since the posterior distribution on wave velocities given the observed data is no longer a Gaussian distribution. Non-Gaussian, and in particular multimodal, distributions can wreak havoc with Monte Carlo samplers (section 1.3).

In what follows, the sampler is run for 10^7 iterations after a burn-in period of 10^6 iterations. The number of spatial observations is set at $N = 256$ and the number of temporal observations is set to $K = 50$. The observational error is set to $\sigma^2 = 10^{-4}$. We now show some graphs of the samples of the wave velocity produced using the regime discussed above.

Figure 2.10 shows samples from $\mathbb{P}(c|y)$ where the Markov chain has been seeded with a random draw from $\mathcal{N}((0, 0), (0.1)^2)$. Subfigures 2.10(a), 2.10(c) and 2.10(e) show samples of the first component of the wave velocity, c_1 , for $\beta_c = 0.05, 0.02$ and 0.01 respectively. Subfigures 2.10(b), 2.10(d) and 2.10(f) show exactly the same thing but for the second component c_2 . Observe that on decreasing β_c to 0.01 the

sampler stays in roughly the same mode. However, the mode that is settled to in either component does not correspond to the true value of the wave velocity. This can be further understood by manually choosing a seed for the chain. Figure 2.11 achieves exactly this and depicts plots of c_1 and c_2 where we seed the Markov chain at $(0.7, 1.4)$. Notice that the posterior mode that is sampled does not coincide with the true value of the wave velocity, nor does it coincide with the mode in figure 2.10. We conclude that the posterior distribution is multimodal.

On decreasing β_c further, we investigate how close the seed, $(0.7, 1.4)$, is to a posterior mode. Figure 2.12(a) shows samples of c_1 from a Markov chain seeded at $(0.7, 1.4)$ but with $\beta_c = \sigma^2 = 10^{-4}$. Figure 2.12(b) shows the same but for c_2 . We observe that the posterior mode is around the seed of the chain. We conclude that the resulting log-likelihood on the data given some wave velocity exhibits many pits and valleys, with the width of each related to the value of β_c .

Observations are taken at integer times in all of the cases presented above. We have a wave velocity of $(0.5, 1.0)$ on the unit square with doubly periodic boundary conditions. Every observed solution is therefore one of two possible functions. This effect is called *aliasing*. It is the case that not much of the solution is observed in time and so aliasing effects present themselves in the log-likelihood, leading to many local minima. Figure 2.13 shows trace plots of the wave velocity for observation times in $\{0, 1, \dots, 49\}$ chosen uniformly at random. Even in this case, we see that the sampler stays around the seed of the chain and posterior samples are nowhere near the truth.

Finally, we explore seeding the Markov chain at the true wave velocity. Figure 2.14 depicts exactly this and shows samples of c_1 and c_2 from a Markov chain seeded at $(0.5, 1.0)$ with $\beta_c = 2 \times 10^{-4}$, 10^{-4} and 10^{-6} . Notice that in each of the three cases, the Monte Carlo sampler stays around the true value and good quality samples are achieved for these values of the random-walk step size.

In practice, knowing the true value of the quantity of interest that generated the data would nullify the effort of implementing these numerics to explore the posterior distribution. This is a cautionary tale when heavily multimodal distributions are involved. The multimodal nature of the posterior distribution on the wave velocity can be explicitly observed in figure 2.35. Though this figure is for a modified likelihood, the general structure of the energy landscape is similar. Multimodality is also seen in the trace plots of the log-likelihood. Since there is no model error here, there is a guaranteed mode around the true value of the wave velocity. Since the

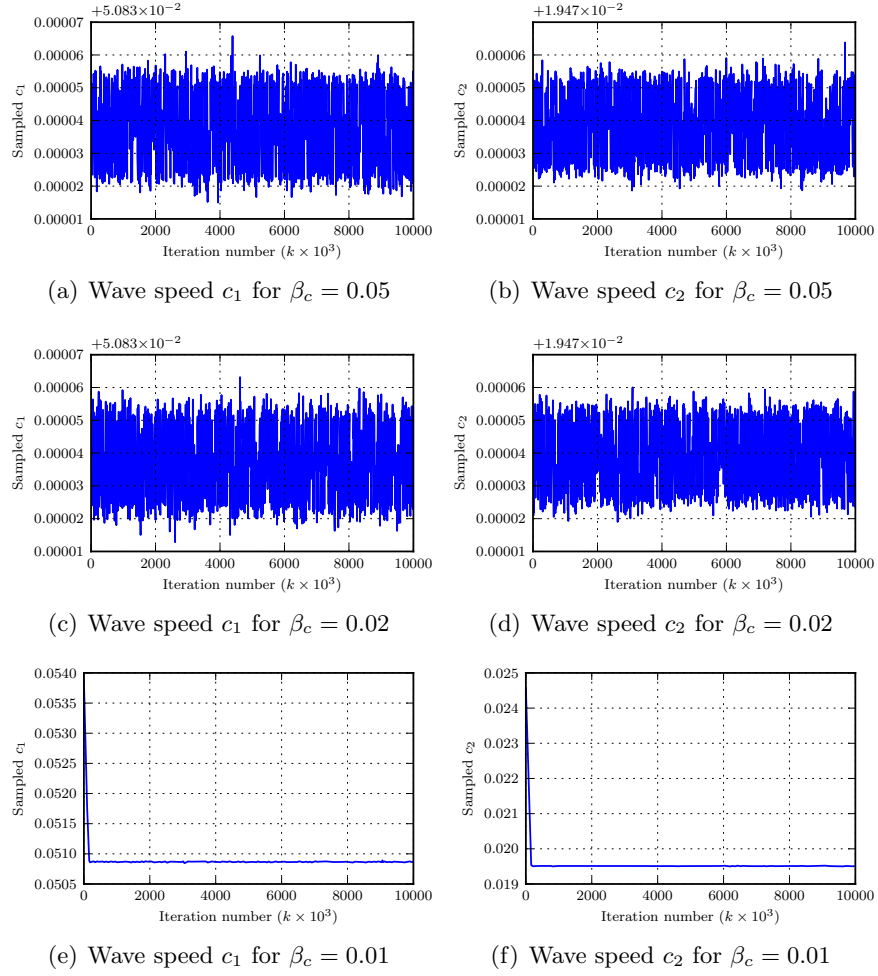


Figure 2.10: Plots of c with random MCMC seed drawn from $\mathcal{N}((0,0), (0.1)^2)$ for varying β_c . The chains all behave roughly identically, except for the case $\beta_c = 0.01$ where the random walk step size is now big enough to allow the chain to jump to a different mode.

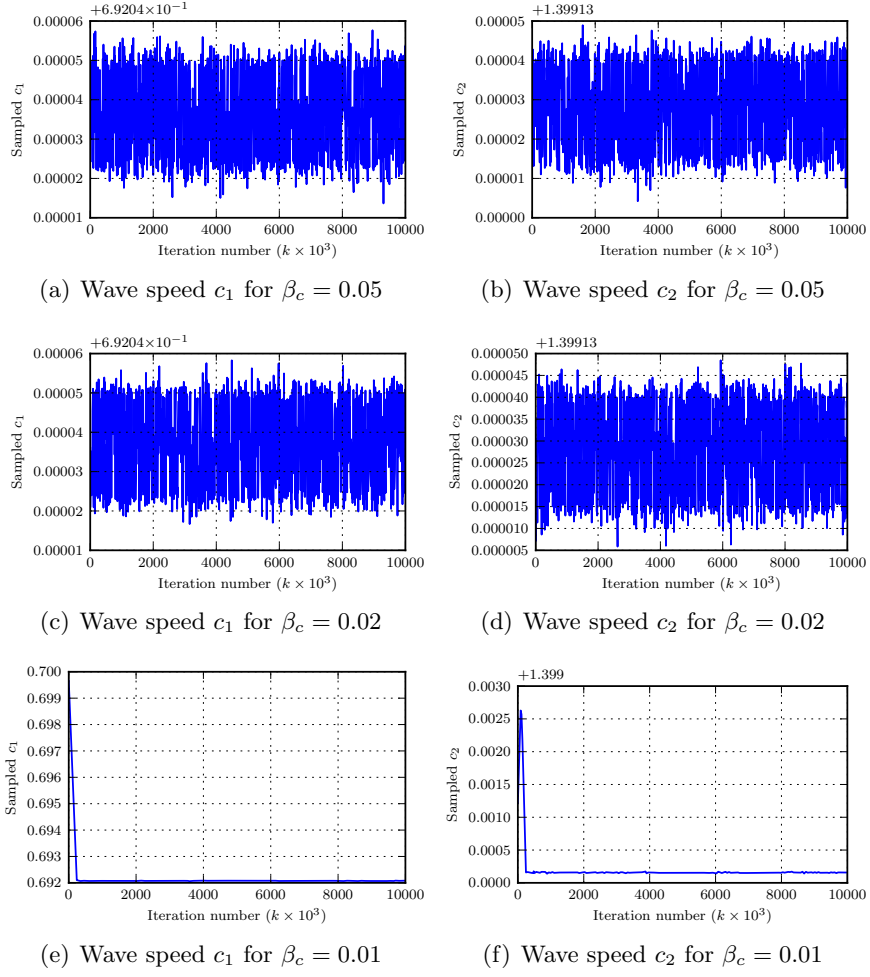


Figure 2.11: Plots of c with MCMC seed at $(0.7, 1.4)$ for varying β_c . A different seed than in figure 2.10, but the sampler behaviour is the same: identical until β_c is large enough to allow the chain to explore a new mode.

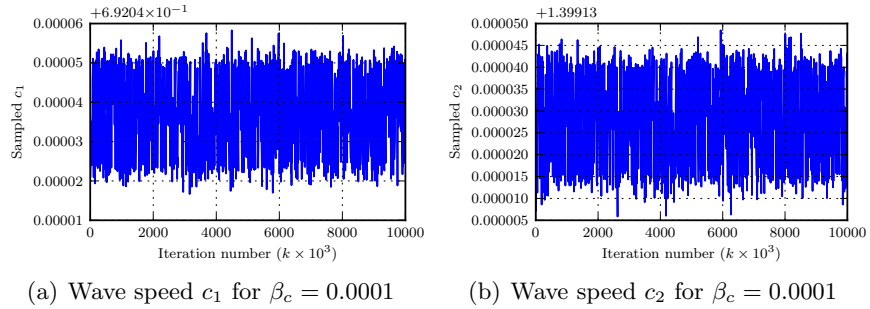


Figure 2.12: Plots of c with MCMC seed at $(0.7, 1.4)$ for varying β_c . The value of β_c is much smaller here than in figures 2.10 and 2.11. Observe that in this case, the chain essentially sits around the seed.

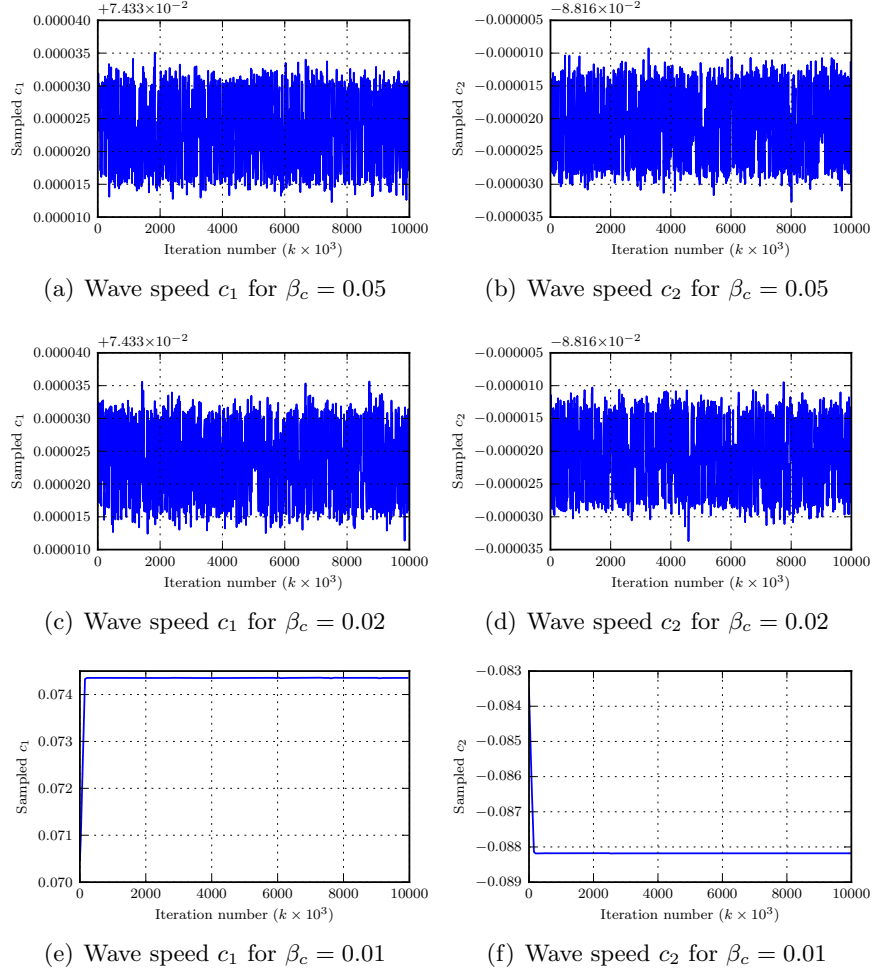


Figure 2.13: Plots of c with random MCMC seed drawn from $\mathcal{N}((0, 0), (0.1)^2)$ for randomly chosen observation times and varying β_c

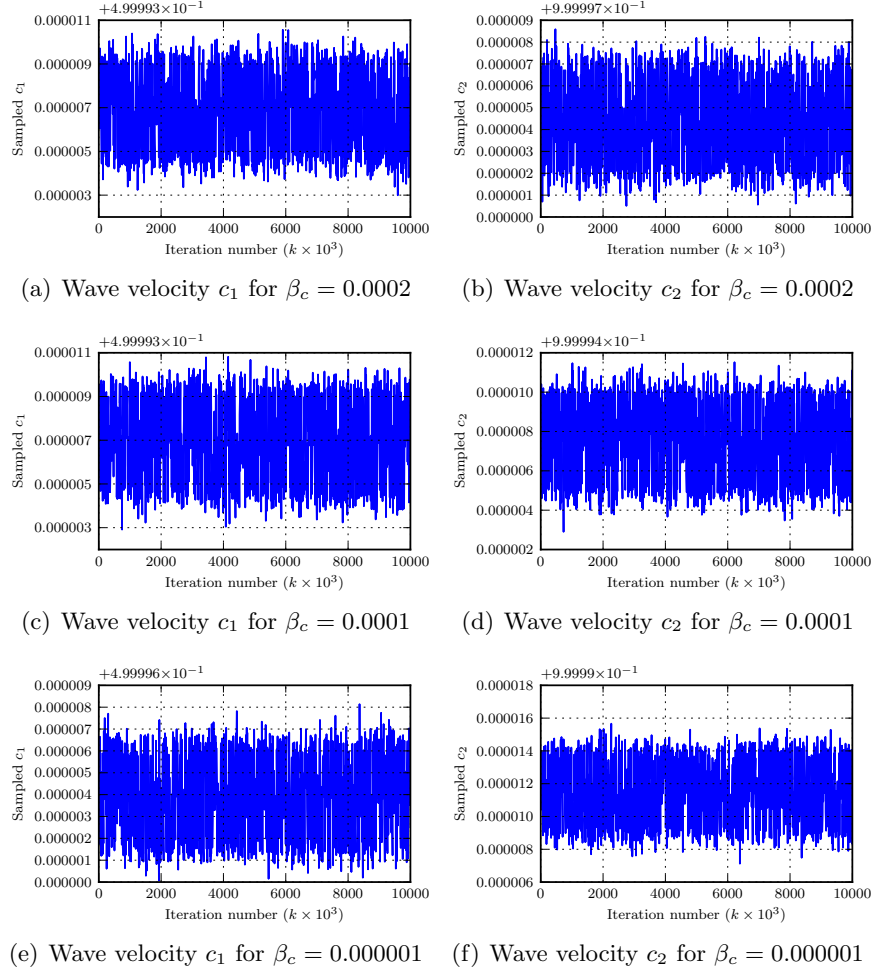


Figure 2.14: Plots of c with MCMC seed at $(0.5, 1.0)$ for varying β_c . Here the chain is seeded at the true value. It is here we observe that this posterior mode is robust to the different values of β_c .

sampler sees at least one other distinct mode (recall figure 2.13), we conclude the posterior is multimodal.

We now present a method that is designed to ameliorate metastability in Markov chain Monte Carlo samplers which does not require a priori knowledge of the truth.

2.3.1 Simulated annealing

Simulated annealing is a method of sampling multimodal probability distribution functions [Kirkpatrick *et al.*, 1983; Černý, 1985]. To set up, first note that

$$\exp\left(-\frac{1}{2\sigma^2}(x-\bar{x})^2\right)^{\frac{1}{T}} = \exp\left(-\frac{1}{2T\sigma^2}(x-\bar{x})^2\right),$$

and with some abuse of notation we can write,

$$\mathcal{N}(\bar{x}, \sigma^2)^{\frac{1}{T}} = \mathcal{N}(\bar{x}, T\sigma^2).$$

We will generalise this immediately to infinite dimensions, yielding,

$$\mathcal{N}(0, (-\Delta)^{-\alpha})^{\frac{1}{T}} = \mathcal{N}(0, T(-\Delta)^{-\alpha}).$$

We do not use the infinite dimensional case here, but write it here for observational purposes. We apply the scaling above to tailor the acceptance probability,

$$\begin{aligned} \alpha(u, c, z, c')^{\frac{1}{T}} &= \min\left\{1, \exp\left(\Phi(u, c) - \Phi(z, c')\right)\right\}^{\frac{1}{T}} \\ &= \min\left\{1, \exp\left(\frac{1}{T}\left(\Phi(u, c) - \Phi(z, c')\right)\right)\right\}, \end{aligned}$$

for $T > 0$.

The algorithm goes as follows. Choose M and $1 < T_1 < T_2 < \dots < T_M$ where $T_M \gg 1$, then,

1. Make d draws from $\mathbb{P}(c|y)$.

2. For $i = 1, \dots, M$:

Make d_i draws from $\mathbb{P}(c|y)^{\frac{1}{T_i}}$.

3. For $i = M - 1, \dots, 1$:

Make d_i draws from $\mathbb{P}(c|y)^{\frac{1}{T_i}}$.

4. Make d draws from $\mathbb{P}(c|y)$.

To sample $\mathbb{P}(c|y)^{\frac{1}{T}}$ the usual random walk Metropolis-Hastings algorithm is used, though recall that in this regime, draws from the prior, ζ , are taken from $\mathcal{N}(0, 0.1^2 \times I_2)^{\frac{1}{T}} = \mathcal{N}(0, 0.1^2 T \times I_2)$ and the acceptance probability $\alpha_T(u, c, z, c') := \alpha(u, c, z, c')^{\frac{1}{T}}$ is used.

Procedurally, this may be repeated as many times as the problem necessitates. We provide plots of samples of the wave velocity as a result of a simulated annealing scheme. We take 5×10^4 samples before starting the annealing scheme. We then temper the distribution immediately at $T_1 = 1000$ and use the formula $T_i = 1001 - i$ for $i = 2, \dots, 999$. For each i , we draw 1×10^3 samples. Linear cooling schemes like this are not necessarily the best, but we use it here as a test case for exploring other modes of the distribution. Once the cooling scheme ends, we draw a further 5×10^4 samples from the posterior distribution on the wave velocity. Figures 2.15(a) and 2.15(b) show samples of c_1 and c_2 respectively. The first 5×10^4 samples are before the tempering scheme and the latter 5×10^4 are after it. All other numerical parameters are exactly as in 2.3. The first thing to notice is that we successfully manage to explore a different mode of the posterior distribution after the tempering scheme has finished. Furthermore, although the first component stays reasonably close to zero, we manage to capture the true mode in the second component well. For interest, we also provide plots of the wave velocity c during the tempering stage of this algorithm (values $T_i \neq 1$) in figure 2.16.

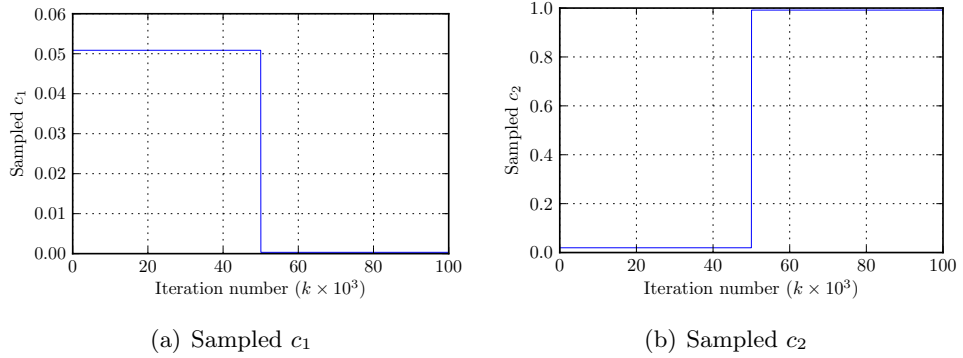


Figure 2.15: Plots of wave velocity samples before and after cooling. Notice that after the cooling scheme has ended, the chain has successfully navigated into a different posterior mode.

Notice that, immediately after we heat up the posterior distribution the second component of the wave velocity jumps close to the value 1. Meanwhile, the first

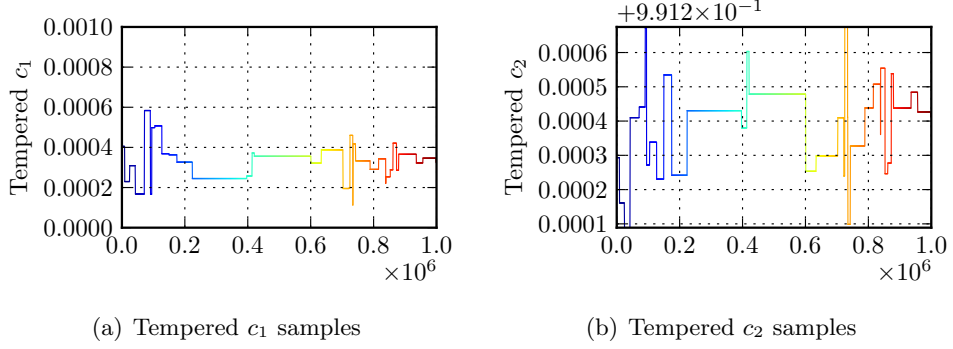


Figure 2.16: Plots of tempered wave velocity samples. The colours dictate the current temperature, starting at $T_1 = 1000$ and decreasing by 1 every one thousand samples.

component doesn't make a corresponding jump to 0.5. Also note that, during the tempering scheme here, the sample quality is poor. This suggests the energy landscape of the log-posterior is extremely rough and obtaining high quality samples is difficult. One solution is to increase the initial temperature T_1 , the choice of which is not something known a priori.

A lot of computational and research effort needs to be invested for an efficient heating/cooling scheme. Furthermore for any temperature greater than 1, the resulting samples come from the wrong distribution and must be discarded. In a practical setting where an extremely expensive forward model is assimilated, simulated annealing is a very wasteful practice. Even in the case of our simple linear forward model, the posterior distribution is rife with local modes. We conclude that cons of simulated annealing heavily outweigh its benefits.

2.4 Wavespeed mismatch

We take observations of a field that is propagated with wave velocity c' and initial condition fixed at the truth,

$$y_{j,k} = v(x_j, t_k) + \eta_{j,k}, \quad \eta_{j,k} \sim \mathcal{N}(0, \sigma^2), \quad (2.6a)$$

$$\rightsquigarrow y' = \mathcal{G}(u, c') + \eta, \quad \eta \sim \mathcal{N}(0, \sigma^2 I_{NK}), \quad (2.6b)$$

where $j \in \{1, 2, \dots, N\}$ and $k \in \{1, 2, \dots, K\}$. It is convenient to decompose y' as $y' = \{y_1, \dots, y_K\}$, where each y_k is an observation of the whole (discretised) solution

field u propagated through linear advection to time $t = t_k$.

The acceptance probability (1.6) in the random walk Metropolis step for the distribution on initial conditions requires a forward integration of some model; the model in which we wish to assimilate the data (2.6b). The model we choose to assimilate the data into is a difficult choice to make, especially in a more practical setting where, for example, we require a state estimate of the whole atmosphere over the UK. For our experiments we choose the two dimensional linear advection PDE, the wave velocity for which will be some c , and may differ from the ‘true’ wave velocity, c' , as in (2.1a). We present some theorems from Lee *et al.* [2011] where we characterise the shape of the posterior mean, $\mathbb{E}(u|y')$, as a function of $\delta c := c' - c$.

Theorem 2.4.1. *For the statistical model (2.1a), suppose that the data, $y' = \{y_1, \dots, y_K\}$, is created from (2.6b) with $c = c'$. Then, as $K \rightarrow \infty$, $\mathbb{E}(u|y') = m'_K \rightarrow u$ in the sense that*

$$\|m'_K - u\|_{L^2(\Omega'; H^s(\mathbb{T}^2))} = \mathcal{O}\left(K^{-\frac{1}{2}}\right), \quad (2.7a)$$

$$\|m'_K - u\|_{H^s(\mathbb{T}^2)} = o\left(K^{-\theta}\right) \quad \Omega' - a.s., \quad (2.7b)$$

for the probability space Ω' generating the true observation noise $\{\eta_k\}_{k \in \mathbb{N}}$, and for any non-negative $\theta < \frac{1}{2}$. Furthermore, the posterior covariance operator, \mathcal{C}_K , satisfies $\mathcal{C}_K \rightarrow 0$ in the sense that its operator norm from $L^2(\mathbb{T}^2)$ to $H^s(\mathbb{T}^2)$ satisfies

$$\|\mathcal{C}_K\|_{\mathbf{L}(L^2(\mathbb{T}^2); H^s(\mathbb{T}^2))} = \mathcal{O}(K^{-1}). \quad (2.8)$$

Theorem 2.4.1 says that, in the absence of model error, the posterior mean converges to the true initial condition (figure 2.17(b)) in the large data limit; expected behaviour.

Definition 2.4.2. *Given $p, q \in \mathbb{N}$, define*

$$\mathbb{K} := \left\{ k \in \mathbb{Z}^2 \mid \left(\frac{k_1}{p}, \frac{k_2}{q} \right) \in \mathbb{Z}^2 \right\}.$$

Now define $\mathcal{F}_{(p,q)} : L^2(\mathbb{T}^2) \rightarrow L^2(\mathbb{T}^2)$ as

$$\mathcal{F}_{(p,q)} f = \sum_{k \in \mathbb{K}} \langle f, \phi_k \rangle \phi_k.$$

This is the linear operator that projects onto every $(p, q)^{th}$ Fourier mode of f .

Definition 2.4.3. Let $f \in L^2(\mathbb{T}^2)$. Define the spatial mean of f ,

$$\langle f \rangle := \langle f, \phi_0 \rangle = \int_{\mathbb{T}^2} f(x) \, dx.$$

Theorem 2.4.4. For the statistical model (2.1a), suppose that the data, $y' = \{y_1, \dots, y_K\}$, is created from (2.6b) with $\delta c \neq 0 \pmod{(1,1)}$ (equivalently $\delta c \notin \mathbb{Z} \times \mathbb{Z}$). As $K \rightarrow \infty$,

1. if $\Delta t \delta c = (p'/p, q'/q) \in \mathbb{Q} \times \mathbb{Q}$ and $\gcd(p', p) = \gcd(q', q) = 1$, then $m_K \rightarrow \mathcal{F}_{(p,q)} u$ in the sense that

$$\|m_K - \mathcal{F}_{(p,q)} u\|_{L^2(\Omega'; H^s(\mathbb{T}^2))} = \mathcal{O}\left(K^{-\frac{1}{2}}\right), \quad (2.9a)$$

$$\|m_K - \mathcal{F}_{(p,q)} u\|_{H^s(\mathbb{T}^2)} = o\left(K^{-\theta}\right) \quad \Omega' - a.s., \quad (2.9b)$$

for any non-negative $\theta < 1/2$;

2. if $\Delta t \delta c \in \mathbb{R} \setminus \mathbb{Q} \times \mathbb{R} \setminus \mathbb{Q}$, then $m_K \rightarrow \langle u \rangle$ in the sense that

$$\|m_K - \langle u \rangle\|_{L^2(\Omega'; H^s(\mathbb{T}^2))} = o(1), \quad (2.10a)$$

$$\|m_K - \langle u \rangle\|_{H^s(\mathbb{T}^2)} = o(1) \quad \Omega' - a.s. \quad (2.10b)$$

Theorem 2.4.4 explicitly relates a rational wave velocity mismatch to the Fourier modes of the posterior mean. When the difference has denominator q , say, then the posterior mean is made up of every q^{th} Fourier mode of the true initial condition (figure 2.17(c)). Consequently, if the difference is irrational (equivalently, $q = 0$) then the posterior mean is just the first Fourier mode; the spatial average of the true initial condition (figure 2.17(d)).

Theorem 2.4.5. For the statistical model (2.1a), suppose that the data, $y' = \{y_1, \dots, y_K\}$, is created from (2.6b) with time-dependent wave velocity $\delta c(t) = c'(t) - c(t)$ satisfying $\int_0^t \delta c(s) \, ds = \alpha + \mathcal{O}(t^{-\beta})$. Then, as $K \rightarrow \infty$, $m_K \rightarrow u_\alpha := u(x - \alpha)$ in the sense that

$$\|m_K - u_\alpha\|_{L^2(\Omega'; H^s(\mathbb{T}^2))} = \mathcal{O}\left(K^{-\phi}\right), \quad (2.11a)$$

$$\|m_K - u_\alpha\|_{H^s(\mathbb{T}^2)} = o\left(K^{-\theta}\right) \quad \Omega' - a.s., \quad (2.11b)$$

for $\phi = 1/2 \wedge \beta$ and for any non-negative $\theta < \phi$.

When the wave velocities are time varying and decay sufficiently fast, theorem 2.4.5 states that the posterior mean is a spatial translation of the true initial condition (figure 2.17(e)). The size of this translation is given by the ‘area’ between the two wave velocity paths.

Theorem 2.4.6. *For the statistical model (2.1a), suppose that the data, $y' = \{y_1, \dots, y_K\}$, is created from (2.6b) with time-dependent wave velocities satisfying $\int_0^t c'(s; \omega') ds = \int_0^t c(t) ds - \varepsilon W(t)$ and $\varepsilon W(t)$ is the Wiener process with amplitude $\varepsilon > 0$. Then, as $K \rightarrow \infty$, $m_K \rightarrow \langle u \rangle$ in the sense that*

$$\|m_K - \langle u \rangle\|_{L^2(\Omega'; H^s(\mathbb{T}^2))} = \mathcal{O}\left(K^{-\frac{1}{2}}\right), \quad (2.12a)$$

$$\|m_K - \langle u \rangle\|_{H^s(\mathbb{T}^2)} = o\left(K^{-\theta}\right) \quad \Omega' - a.s., \quad (2.12b)$$

for any non-negative $\theta < 1/2$.

Lastly, the final result can be summarised as follows. When the time dependent wave velocities are time integrated and their difference is a Weiner process then the posterior mean, like theorem 2.4.4, is the spatial average of the true initial condition (figure 2.17(f)). The proofs of theorems 2.4.1–2.4.6 are due to Lee [Lee *et al.*, 2011] and are not presented here.

The purpose of this section is twofold: first to illustrate the preceding theorems with numerical experiments; and secondly, to show that relaxing the statistical model can avoid some of the lack of consistency problems that the theorems highlight. All of the numerical results we describe are based on using (2.1a) with some, possibly time-dependent, wave velocity c . The data is generated by (2.6b) with, possibly random, wave velocity $c'(t)$. In section 2.4.1 we illustrate theorems 2.4.1, 2.4.4, 2.4.5 and 2.4.6. In section 2.5.3 we will also describe a numerical method in which the state of the system *and* the wave velocity are learnt by combining the data and statistical model. Since this problem is inherently non-Gaussian we adopt from the outset a Bayesian approach which coincides with the Gaussian filtering approach when the wave velocity is fixed, but is sufficiently general to also allow for the wave velocity to be part of the unknown state of the system. In both cases we apply function space MCMC methods [Stuart, 2010] to sample the distribution of interest. Note, however, that the purpose of this section is *not* to determine the most efficient numerical methods, but rather to study the properties of the statistical distributions of interest.

For fixed wave velocity c , the statistical model (2.1a) with observations (2.6b) de-

defines a probability distribution $\mathbb{P}(u, y'|c)$. This is a Gaussian distribution and the conditional distribution $\mathbb{P}(u|y', c)$ is given by the measure $\mu_K = \mathcal{N}(m_K, \mathcal{C}_K)$, studied in Lee *et al.* [2011]. In our first set of numerical results (section 2.4.1), the wave velocity is considered known. We sample $\mathbb{P}(u|y', c)$ using the function space random walk from (1.7) [Cotter *et al.*, 2011]. In section 2.5, the wave velocity is considered an unknown constant. If we place a prior measure $\rho(c)$ on the wave velocity then we may define $\mathbb{P}(c, u, y') = \mathbb{P}(u, y'|c)\rho(c)$. We are then interested in the conditional distribution $\mathbb{P}(c, u|y')$ which is non-Gaussian. We adopt a Metropolis-within-Gibbs approach [Gilks *et al.*, 1995; Geweke & Tanizaki, 1999; Geweke, 2001] in which we sample alternately from $\mathbb{P}(u|c, y')$, which we do as in section 2.5, and $\mathbb{P}(c|u, y')$, which we sample using the analogous finite dimensional random walk Metropolis-Hastings algorithm. An analysis of the convergence properties of the Metropolis-within-Gibbs sampler can be found in Roberts & Rosenthal [2006].

Throughout the numerical simulations, we have truncated the Karhunen-Loève expansion (1.15) at 2^5 terms in the x -direction and 2^5 terms in the y -direction—representing the solution of the wave equation on a grid of $2^5 \times 2^5$ points. Observations are also taken on this grid. The observational noise is uncorrelated with variance $\sigma^2 = 10^{-4}$ at each grid point. The continuum limit of such a covariance operator satisfies weaker assumptions than those in Lee *et al.* [2011], but is used to illustrate the fact that the theoretical results can be generalised to such observations. Note also that the numerical results are performed with model error so that the aforementioned distributions are sampled with $\delta c \neq 0$ in (2.1a) and (2.6b).

2.4.1 Sampling the initial condition with model error

Throughout, we use the wave velocity,

$$c = (0.5, 1.0), \quad (2.13)$$

in our statistical model. The true initial condition used to generate the data is,

$$u(x_1, x_2) = \sum_{k_1, k_2=1}^3 \sin(2\pi k_1 x_1) + \cos(2\pi k_2 x_2). \quad (2.14)$$

This function is displayed in figure 2.17(a). As a prior on u we choose the Gaussian $\mathcal{N}(0, (-\Delta)^{-2})$ where the domain of $-\Delta$ is $L^2(\mathbb{T}^2)$ with constants removed, so that it has zero spatial mean. We implement the MCMC method to sample from $\mathbb{P}(u|c, y')$

for a number of different data y' , corresponding to different choices of $c' = c'(t)$. We calculate the empirical mean of $\mathbb{P}(u|c, y')$, which approximates $\mathbb{E}(u|c, y')$. The results are shown in figures 2.17(b)–2.17(f). In all cases the Markov chain is burnt in for 10^6 iterations, and this transient part of the simulation is not used to compute moments of the conditioned measure $\mathbb{P}(u|c, y')$. After the burn in we proceed to iterate a further 10^7 times and use this information to compute the corresponding moments. The number of spatial observations is $N = 1024$. Observational error is set to $\sigma^2 = 10^{-4}$. The number of temporal observations is $K = 50$ unless otherwise stated.

In the perfect model scenario ($c = c'$), the empirical mean shown in figure 2.17(b) should fully recover the true initial condition u from theorem 2.4.1. Comparison with figure 2.17(a) shows that this is indeed the case. We now demonstrate the effect of model error in the form of a constant shift in the wave velocity: figure 2.17(c) and figure 2.17(d) show the empirical means when $c' - c = (1/2, 1/2) \in \mathbb{Q}^2$ and $c' - c = (1/e, 1/\pi) \in \mathbb{R}^2 \setminus \mathbb{Q}^2$, respectively. From theorem 2.4.4, the computed empirical distribution should be close to $\mathcal{F}_{(2,2)}u$, the $(2, 2)^{\text{th}}$ of (2.14), and $\langle u \rangle = 0$, respectively; this is indeed the case.

If we choose $c'(t)$ satisfying $\int_0^\infty (c'(s) - c) \, ds = (1/2, 1/2)$, then theorem 2.4.5 says that figure 2.17(e) should be close to a shift of u by $(1/2, 1/2)$, and this is exactly what we observe. In this case, we know from Lee *et al.* [2011] that although the smoother is in error, the filter should correctly recover the true $u_K := u(x, t_K)$ for large K . To illustrate this we compute $\|\mathbb{E}(u_K|c, y') - u'_K\|_{L^2(\mathbb{T}^2)}$ as a function of K and depict it in figure 2.18(a). This shows convergence to 0 as predicted. To obtain a rate of convergence, we compute the gradient of a log-log plot of figure 2.18(b). We observe the rate of convergence is close to $\mathcal{O}(K^{-2})$. Note that this is higher than the theoretical bound given in Lee *et al.* [2011]; this suggests that our convergence theorems do not exhibit sharp rates.

Finally, we examine the random $c'(t, \omega')$ cases. Figure 2.17(f) shows the empirical mean when $c'(t; \omega')$ is chosen such that $\int_0^t (c'(s; \omega') - c) \, ds = W(t)$ where $W(t)$ is a standard Brownian motion. Theorem 2.4.6 says that the computed empirical distribution should have mean close to $\langle u \rangle$, and this is again the case.

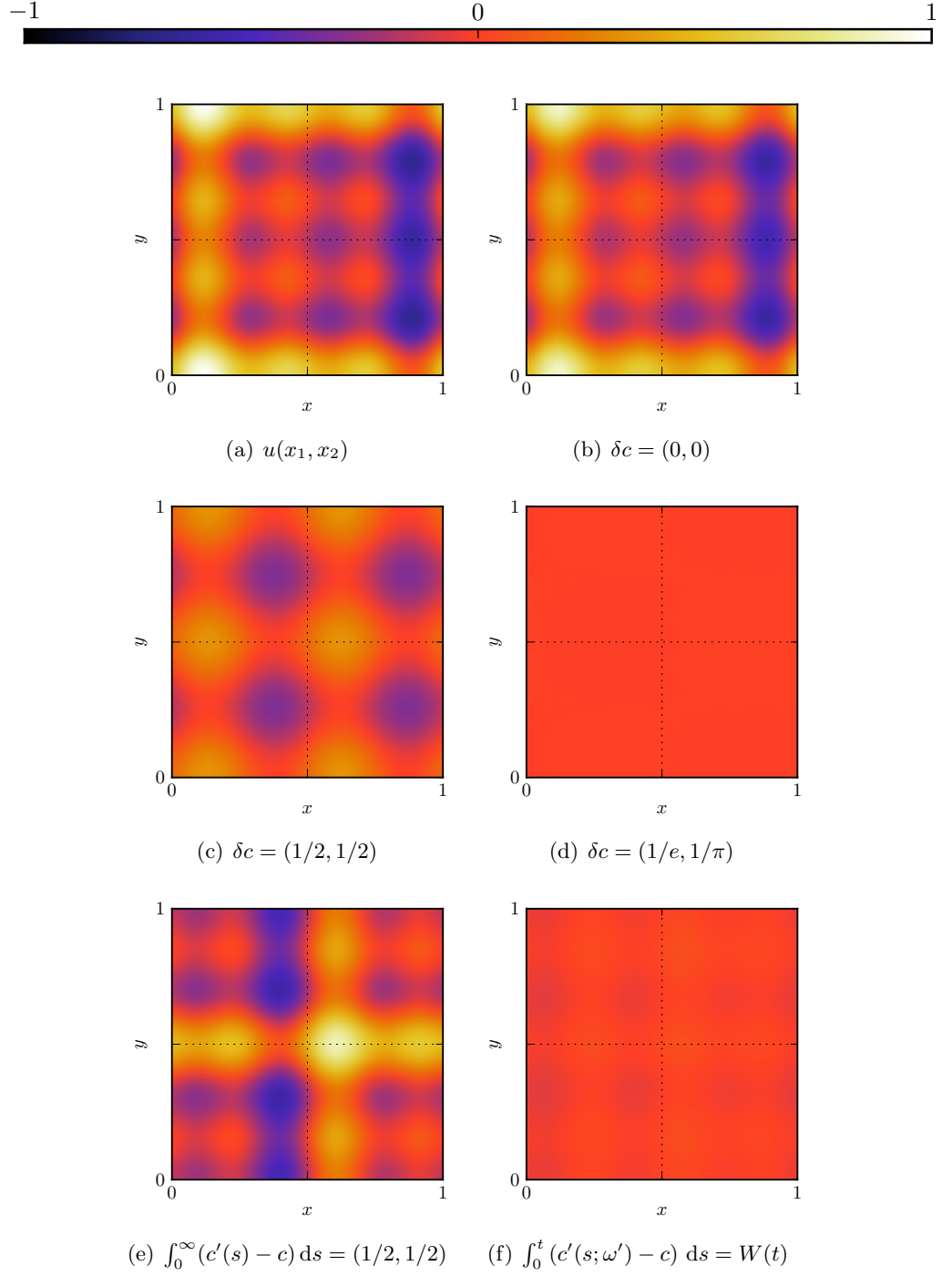


Figure 2.17: Figure 2.17(a) is the true initial condition. Figures 2.17(b) – 2.17(f) show the desired empirical mean of the smoothing $\mathbb{P}(u|y')$ for $\delta c = (0, 0)$, $\delta c = (1/2, 1/2)$, $\delta c \in \mathbb{R}^2 \setminus \mathbb{Q}^2$, $\int_0^\infty \delta c dt = (1/2, 1/2)$ and $\delta c = \dot{W}$ respectively.

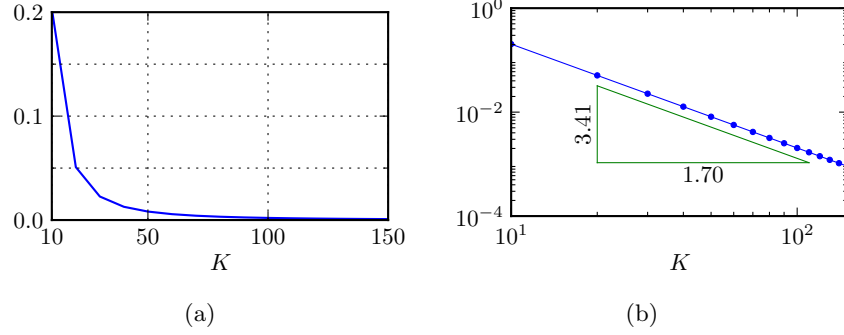


Figure 2.18: Plot 2.18(a) shows $\|\mathbb{E}(u_K|c, y') - u_K\|_{L^2(\mathbb{T}^2)}^2$ as a function of K , when $\int_0^\infty \delta c(s) ds = (1/2, 1/2)$. Its log-log plot, along with a least squares fit, is depicted in Plot 2.18(b), demonstrating quadratic convergence.

2.5 Sampling the joint

In this section we make several attempts to sample $\mathbb{P}(u, c|y)$. Namely, we seed nearby the truth and gradually move away; we seed at a draw from the prior; we view simulated annealing from the optimiser's perspective; and we seed using the observed data.

Sampling of the joint distribution is achieved by implementing a Metropolis-within-Gibbs method. The Metropolis steps are done on either the u or c components, and the Gibbs steps are done to transition *between* the u and c components. Formally, given c , we propose from $u|y, c$ according to

$$\frac{d\mu^{y,c}}{d\mu_0}(u) \propto \exp(-\Phi(u, c)),$$

and given u we propose from $c|y, u$ according to

$$\mathbb{P}(c|y, u) \propto \exp(-\Phi(u, c)),$$

where $\mu^{y,c}$ is the posterior measure on u given y and μ_0 is the prior measure on u .

2.5.1 Seeding nearby the truth

Here we show the behaviour of the posterior sampler if one seeds the Markov chain not at the truth, but ‘nearby’. Here we mean that for $\epsilon = 3, 2, 1$ we seed the chain

with

$$u_0 = \sum_{k=1}^3 \sin(2\pi(kx + 10^{-\epsilon})) + \cos(2\pi(ky + 10^{-\epsilon})) \quad (2.15)$$

$$c_0 = (0.5, 1.0) + (10^{-\epsilon}, 10^{-\epsilon}). \quad (2.16)$$

We burn-in the chain for 10^6 iterations and then proceed to draw 10^7 samples from the posterior distribution. The number of spatial and temporal observations is set to $N = 1024$ and $K = 50$, respectively. The observational error is $\sigma^2 = 10^{-4}$. Figure 2.19 illustrates the posterior mean for the values of ϵ noted above. Note that for $\epsilon = 2, 1$ the Markov chain falls into a local minimum that is not the true value and, as a result, the posterior mean on the initial condition is incorrect. This crystallises the multimodal nature of the posterior joint distribution.

2.5.2 Slices of the objective function

Here we look at the functional $\Phi(u, c)$. We expect that with c fixed at the true wave speed, $\Phi(u, c)$ will be a quadratic in u as the posterior distribution $\mathbb{P}(u|y, c)$ is a Gaussian distribution. Figure 2.20(a) illustrates this, with the red and blue curves showing how Φ varies in one of the Fourier coefficients of u . The Fourier coefficients we choose here are $\Re(\hat{u}_{(0,1)})$ (red line) and $\Im(\hat{u}_{(0,1)})$ (blue line). All other coefficients are fixed at their true values. Notice that Φ takes minima at 0.0 and 0.5 respectively. Figure 2.20(b) shows the behaviour of Φ when u is fixed at the truth and c_1 (first component of the wave velocity) is varied. Notice that the minimiser is $c_1 = 0.5$, which is the true value. Furthermore, observe that this graph is not a quadratic as the map,

$$c \mapsto u(x + ct), \quad (2.17)$$

is not linear in c for our choice of initial condition. Note also that Figure 2.20(b) exhibits many local minima, this explains the metastability of the Markov chain that samples $\mathbb{P}(c|y, u)$. Tempering the wave speed component of the posterior distribution results in an increase in variance of the modes. This corresponds to a scaling of the objective function and Figure 2.21 shows plots of Φ for a selection of temperatures.

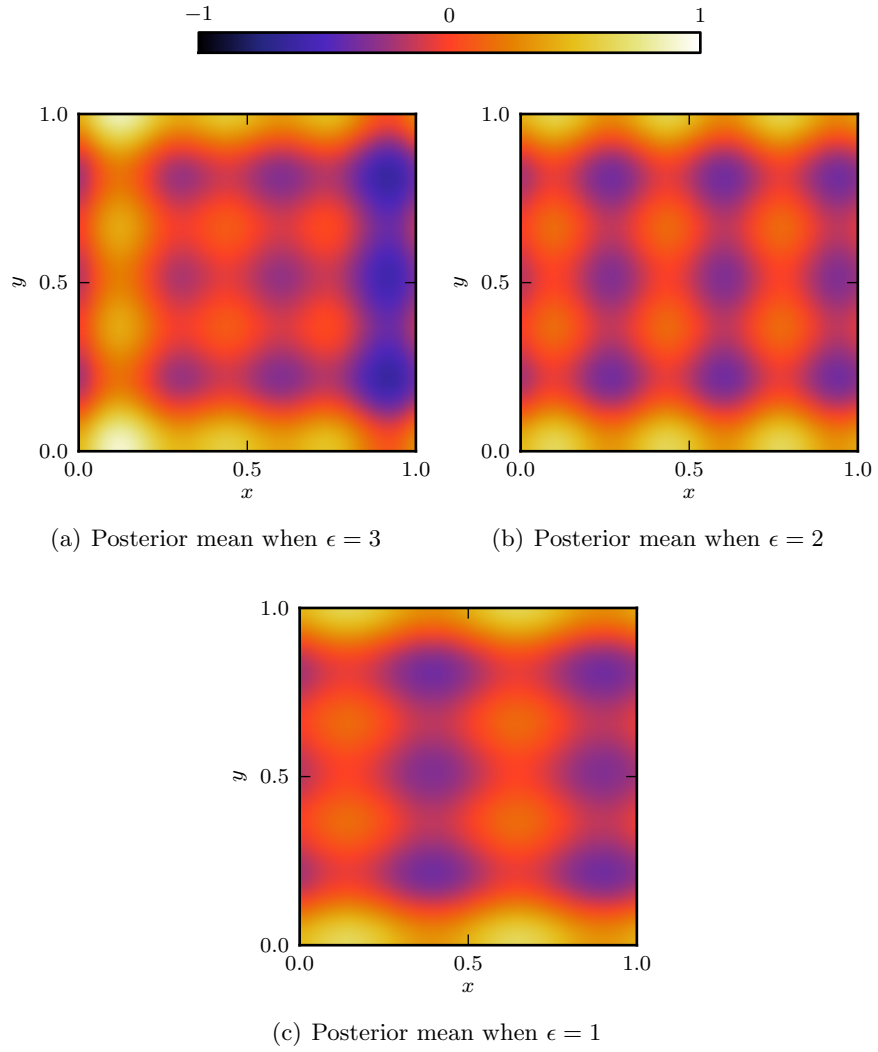


Figure 2.19: Plots of the posterior mean for varying ϵ . Notice the posterior mean is incorrect when we are too far away from the truth ($\epsilon = 1, 2$). When $\epsilon = 3$, we recover the true initial condition. This shows the posterior distribution is multimodal.

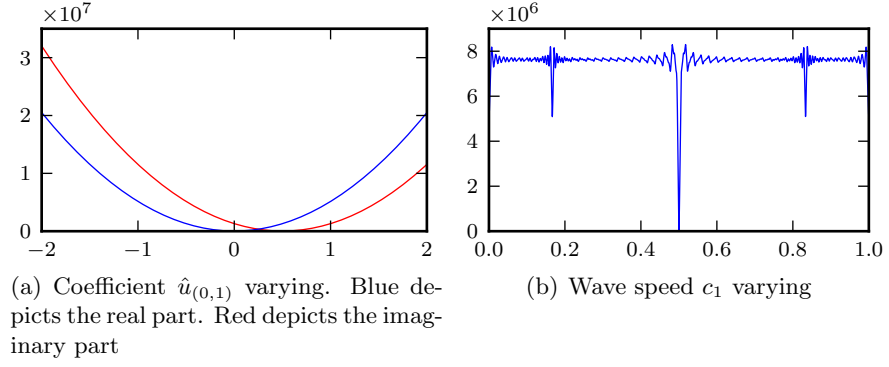


Figure 2.20: Plots of slices of the objective function Φ . Left: The forward map is linear. This is reflected by the quadratic nature of the log-likelihood. Right: The forward map is nonlinear. This is reflected by the very nonquadratic behaviour of the log-likelihood.

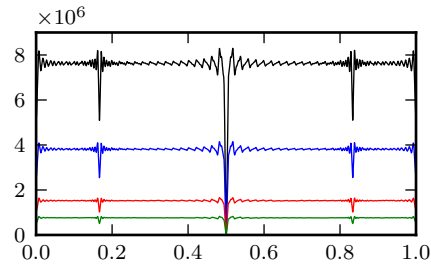


Figure 2.21: The black, blue, red and green lines correspond to Φ at temperatures $T = 1, 2, 5$ and 10 respectively

2.5.3 Seeding the wave velocity

In this section we will look at attempting to obtain a high quality estimate for the wave velocity by analysing the observational data rather than burning in the chain. We generate data from (2.1a), with $c' = c$ given by (2.13) and initial condition (2.14). We assume that neither the wave velocity nor the initial condition are known to us, and we attempt to recover them from given data.

The desired conditional distribution is multimodal with respect to c – recall that it is non-Gaussian – and care is required to seed the chain close to the desired value in order to avoid metastability. Although the algorithm does not have access to the true signal u , we do have noisy observations of it: y .

The solution to the advection model can be written in terms of the Fourier coefficients of the initial condition. Let v_j^k denote Fourier coefficient (j_1, j_2) of the solution to the two-dimensional advection equation at time t_k , then

$$\hat{v}_j^k = \hat{v}_j^{k-1} \exp(2\pi i c(t_k - t_{k-1})j \cdot x). \quad (2.18)$$

Assuming the observed data \hat{y}_j^k approximately advects with velocity \bar{c} , we write

$$\begin{aligned} \hat{y}_j^k &\approx \hat{y}_j^{k-1} \exp(2\pi i \bar{c}(t_k - t_{k-1})j \cdot x) \\ &= \hat{y}_j^{k-1} \lambda^{\bar{c}} \end{aligned}$$

where $\lambda = \exp(2\pi i(t_k - t_{k-1})j \cdot x)$. Taking logs yields,

$$\begin{aligned} \log \hat{y}_j^k &\approx \log \hat{y}_j^{k-1} \exp(2\pi i \bar{c}(t_k - t_{k-1})j \cdot x) \\ \therefore \log \hat{y}_j^k &\approx \log \hat{y}_j^{k-1} + 2\pi i \bar{c}(t_k - t_{k-1})j \cdot x \\ \therefore 2\pi i \bar{c}(t_k - t_{k-1})j \cdot x &\approx \log \left(\frac{\hat{y}_j^k}{\hat{y}_j^{k-1}} \right), \quad k = 1, \dots, K \end{aligned}$$

A formal way of stating this problem is to find \bar{c} such that $\|A\bar{c} - b\|_2^2$ is minimised, where

$$A = \begin{bmatrix} 2\pi i(t_1 - t_0)j \cdot x \\ 2\pi i(t_2 - t_1)j \cdot x \\ \vdots \\ 2\pi i(t_K - t_{K-1})j \cdot x \end{bmatrix}, \quad b = \begin{bmatrix} \log \left(\hat{y}_j^1 / \hat{y}_j^0 \right) \\ \log \left(\hat{y}_j^2 / \hat{y}_j^1 \right) \\ \vdots \\ \log \left(\hat{y}_j^K / \hat{y}_j^{K-1} \right) \end{bmatrix}. \quad (2.19)$$

This is a linear least-squares problem, for which there are numerous computational approaches to solve. We use a standard singular value decomposition method to obtain a solution. Since the observational data is noisy, this estimate is more accurate for small values of j and we choose $j = (1, 0)$ to estimate \bar{c}_1 and $j = (0, 1)$ to estimate \bar{c}_2 . Once the minimiser \bar{c} is obtained, this value is used as the seed – the initial value in (2.5) – in the Metropolis-Hastings sampler for the wave velocity. We continue sampling the joint distribution $\mathbb{P}(u, c|y)$ using the usual Metropolis-within-Gibbs method.

Figure 2.22 shows the marginal distribution for c computed with four different values of the data size ($K = 10, 50, 100, 1000$), in all cases with the Markov chain seeded as above. The results show that the marginal wave velocity distribution $\mathbb{P}(c|y')$ converges to a Dirac on the true value as the amount of data is increased. Although not shown here, the initial condition is also converging to a Dirac on the true value (2.13) in this limit. As for the numerical parameters used here, we use $N = 1024$ and $\sigma^2 = 10^{-4}$.

We round-off this subsection by mentioning related published literature. First we mention that, in a setting similar to ours, a scheme to approximate the true wave velocity is proposed which uses parameter estimation within 3D Var for the linear advection equation with constant velocity Smith *et al.* [2009], and non-constant velocity Smith *et al.* [2011]. These methodologies deal with the problem entirely in finite dimensions but are not limited to the linear dynamics. Secondly we note that, although a constant wave velocity parameter in the linear advection equation is a useful physical idealisation in some cases, it is a rigid assumption, making the data assimilation problem with respect to this parameter quite hard; this is manifest in the large number of samples required to estimate this constant parameter. A notable, and desirable, direction in which to extend this work numerically is to consider a time-dependent wave velocity as presented in theorems 2.4.5–2.4.6. For efficient filtering techniques to estimate time-dependent parameters, the reader is directed to Cohn [1997]; Dee [1996]; Baek *et al.* [2006]; Gershgorin *et al.* [2010].

2.6 Modifying the likelihood

Recall the likelihood functional given by,

$$\Phi(\cdot) := \frac{1}{2\sigma^2} \|\mathcal{G}(\cdot) - y\|^2. \quad (2.20)$$

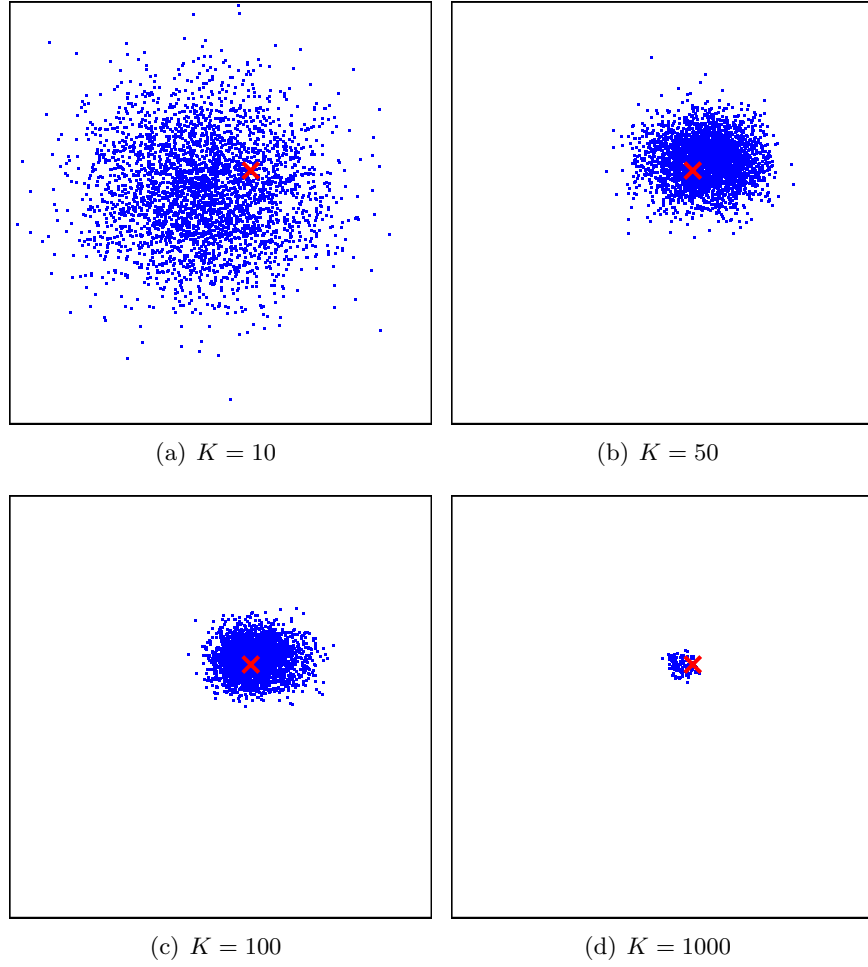


Figure 2.22: The marginal distribution of $\mathbb{P}(c, u|y')$ with respect to c are depicted on the square 1.4×10^{-4} by 10^{-4} . The red cross marks the true wave velocity $c = (0.5, 1.0)$. As the number of observations increases, the posterior measure converges to a Dirac on the true value.

Throughout this section, we harness a modified functional of the form,

$$\tilde{\Phi}(\cdot) := \frac{1}{2\sigma} \|\mathcal{G}(\cdot) - y\|. \quad (2.21)$$

The modified likelihood functional (2.21) is similar to (2.20), though we discard the square in the norm. This has the effect of imposing a double-exponential distribution around the truth, rather than a smooth mode in the probability distribution. We explore the effects on the sampler of the introduced cusp.

2.6.1 Sampling the initial condition

Illustrations of the L^2 norm of the sampled initial condition for the modified likelihood number are provided for comparison with figures 2.1–2.3. First, observe that good quality samples in figure 2.23 are achieved with a larger random walk step size than compared with figure 2.1. Throughout this subsection, we burn in the Markov chain for 10^4 iterations and take 10^6 draws from the posterior distribution.

One can observe the quality of a sampler heuristically by looking at a trace plot and noting how much of the state space is explored. Sample quality is said to be good if the sampler explores a large portion of the stationary measure. On the other hand, if the sampler looks like it is only sampling very close to one specific value, this could mean that either: a) the measure is very tight and highly peaked; or b) the implemented sampler is behaving poorly. In the case of a), nothing is wrong and the sampler correctly portrays the underlying stationary measure. In the case of b), more thought is required to determine the cause of the behaviour. For an extensive comparison of Markov Chain convergence diagnostics, see Cowles & Carlin [1996].

Varying step-size and observational error

Here the numerical parameters we use are $N = 1024$ and $K = 50$. Figures 2.23(a) and 2.23(b) show trace plots for $\sigma = 0.01$ and $\sigma = 0.1$ respectively. Notice we see the same behaviour when comparing 2.1(a) and 2.1(b) for the unmodified likelihood. For larger σ , higher quality samples are drawn due to better exploration of the state space. However, as a penalty for increasing σ , we obtain samples more towards the prior measure (further away from the truth). As one increases β from 0.1 to 0.2 (figure 2.24), we see a large degradation in sample quality, illustrated by the large numbers of samples that are rejected by (1.6). This is a result of taking larger steps in the state space by using less information of the current state in the sampler. There

is a notable improvement in sample quality on increasing σ from 0.01 to 0.1, which can be seen in figures 2.24(a) and 2.24(b). Note the effects from the prior make the norm of the samples smaller than the norm of the truth for larger σ . Finally, we present the case $\beta = 0.3$ in figure 2.25. The cases $\sigma = 0.01, 0.1$ are illustrated in figures 2.25(a) and 2.25(b). The case $\beta = 0.3$ is not a useful one in practice. The step size is too big, leading to a large number of rejections. After a burn-in of 10^4 the chain has not reached stationarity.

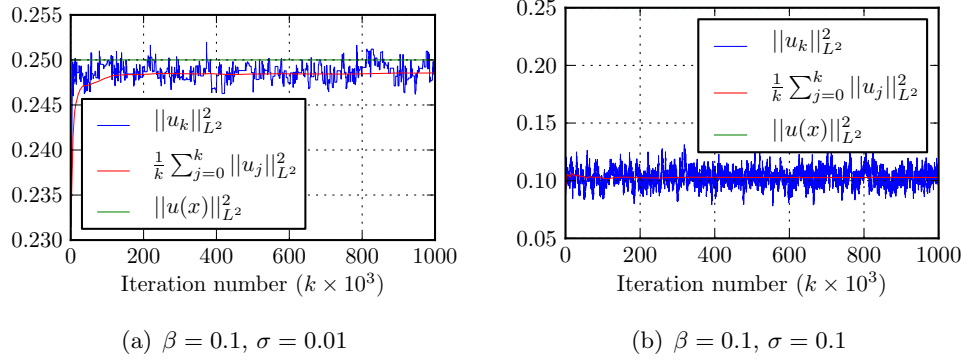


Figure 2.23: Trace plots showing effect of varying observational noise for $\beta = 0.1$. Notice that when the observations are worse (σ is larger), the sampler leans more towards the prior and further away from the truth.

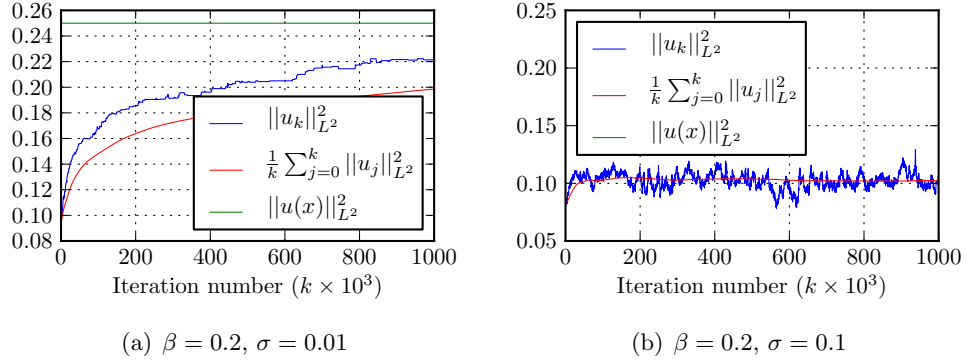


Figure 2.24: Trace plots showing effect of varying observational noise for $\beta = 0.2$. Here β is larger than in figure 2.23 and, as a result, the chain takes noticeably longer to converge to a posterior mode (more evident on the left).

Using the modified likelihood (2.21) can give better quality samples for large random walk step size than compared with the unmodified likelihood (2.20).

We also provide illustrations of the modified likelihood (2.21) as a function of sampler

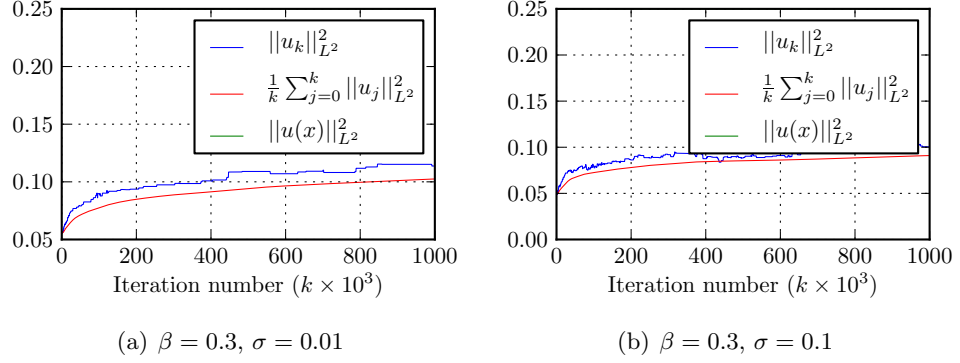


Figure 2.25: Trace plots showing effect of varying observational noise for $\beta = 0.3$. For this, even larger, value of β , convergence is slower still.

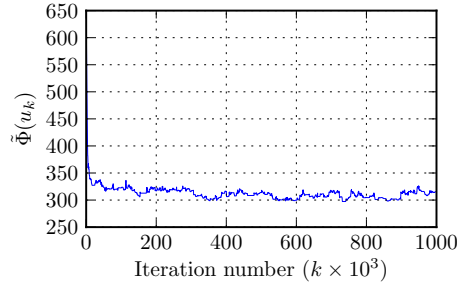
iteration number. These illustrations are useful for observing how far the mode is from the truth. Since (2.21) is a norm, it is always nonnegative. In the case where there is no observational error, the minimum value attainable in the space $H_{\text{per}}^{\alpha-1}(\mathbb{T}^2)$ is zero. It is attained when the true initial condition u satisfies

$$u = \operatorname{argmin}_{u^* \in H^{\alpha-1}} \tilde{\Phi}(u^*).$$

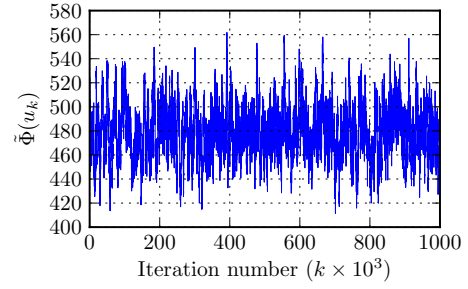
Illustrations are shown in figures 2.26, 2.27 and 2.28 for $\beta = 0.1, 0.2$ and 0.3 , respectively. For increasing β , observe the exact same behaviour as in figures 2.23–2.25, extended periods of rejections and poorer state-space exploration properties. Figures 2.26(a), 2.27(a) and 2.28(a) are for the case $\sigma = 0.01$ and figures 2.26(b), 2.27(b) and 2.28(b) illustrate the case $\sigma = 0.1$. Notice that increasing the observational noise has the same effect on the likelihood trace plots; we observe better sample quality.

Varying number of spatiotemporal observations

Using exactly the same set-up as in section 2.2.3, and with $\sigma^2 = 10^{-4}$, we track one Fourier coefficient from the posterior distribution and plots a histogram of its journey in state space. The illustrations (figure 2.29) the histograms converging to a Dirac measure centred at the truth.

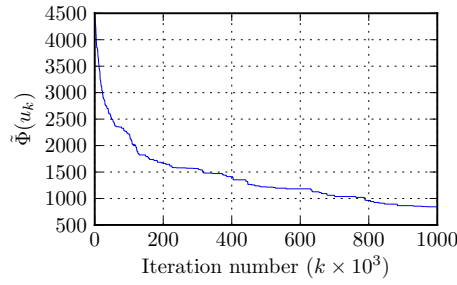


(a) $\beta = 0.1, \sigma = 0.01$

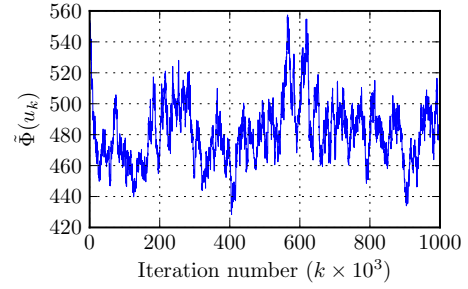


(b) $\beta = 0.1, \sigma = 0.1$

Figure 2.26: Trace plots showing effect of varying observational noise for $\beta = 0.1$. This is the analogue of figure 2.23, but for Φ instead of the acceptance probability.

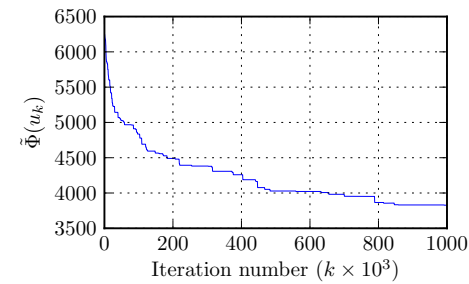


(a) $\beta = 0.2, \sigma = 0.01$

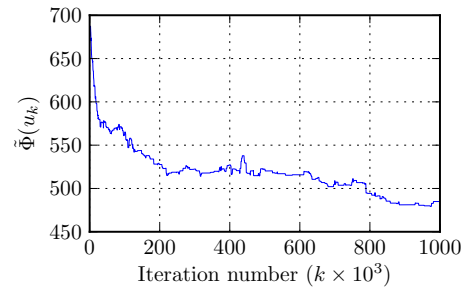


(b) $\beta = 0.2, \sigma = 0.1$

Figure 2.27: Trace plots showing effect of varying observational noise for $\beta = 0.2$. This is the analogue of figure 2.24, but for Φ instead of the acceptance probability.



(a) $\beta = 0.3, \sigma = 0.01$



(b) $\beta = 0.3, \sigma = 0.1$

Figure 2.28: Trace plots showing effect of varying observational noise for $\beta = 0.3$. This is the analogue of figure 2.25, but for Φ instead of the acceptance probability.

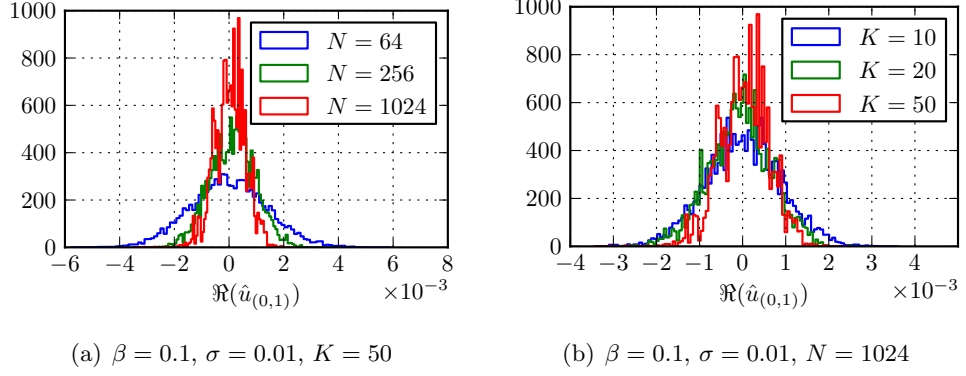


Figure 2.29: Histograms showing effect of varying the number of observations points, N , and times, K . Notice the convergence of the posterior measure to a Dirac measure centred on the true value as the number of observations is increased.

2.6.2 Sampling the wave velocity

Here we show some graphs of samples of the wave velocity produced using the same technique as in section 2.3. Throughout this subsection, we burn in the chain for 10^4 iterations and take 10^7 draws from the posterior distribution. The number of spatial observations is $N = 256$ and the number of temporal observations is $K = 50$. The observational error is set to $\sigma^2 = 10^{-4}$.

Figure 2.30 shows samples from $\mathbb{P}(c|y)$ where the Markov chain has been seeded with a random draw from $\mathcal{N}((0,0), (0.1)^2)$. Subfigures 2.30(a), 2.30(c) and 2.30(e) show samples of the first component of the wave velocity, c_1 , for $\beta_c = 0.5, 0.2$ and 0.1 respectively. Subfigures 2.30(b), 2.30(d) and 2.30(f) show the second component c_2 . Observe that for larger β_c , big initial jumps in the state space occur before the sampler settles down to the true wave velocity mod $(1,1)$. Recall that the best case scenario for recovering the true wave velocity is recovering it mod $(1,1)$, due to aliasing. As β_c is decreased notice that the sampler settles down to a local mode missed when taking bigger step sizes in the state space. Furthermore, this posterior mode (in both the cases $\beta_c = 0.02, 0.01$) does not correspond with the true wave velocity.

Figure 2.31 depicts the same quantity as Figure 2.30 except here we seed the Markov chain at $(0.7, 1.4)$. Notice that for the case $\beta_c = 0.05$, the step size is big enough so that exploration through the state space is minimised, and the posterior mode corresponding to the true wave velocity is reached. Once the sampler reaches this posterior mode, the step size is then too big to achieve high quality samples from

the invariant measure. As β_c decreases, higher quality samples are achieved with the added cost of not taking big enough jumps to be able to reach the posterior mode that corresponds to the true wave velocity. Note that in both the cases $\beta_c = 0.02, 0.01$, the posterior mode is around the seed of the Markov chain, $(0.7, 1.4)$. Figure 2.32(a) shows samples of c_1 from a Markov chain seeded again at $(0.7, 1.4)$ but with $\beta_c = \sigma^2 = 10^{-4}$ much smaller. Notice the similar behaviour as in the previous cases: the sampler settles down to a posterior mode near the seed and extremely high sample quality is obtained. Figure 2.32(b) shows the same trace plot but for c_2 .

Many posterior modes can be a result of aliasing. We attempt to remedy this by randomly choosing observation times instead of observing at exactly integer times. Figure 2.33 shows exactly this for observation times in $\{0, 1, \dots, 49\}$ chosen uniformly at random. Notice that the sampler still does not settle to the posterior mode corresponding to the true wave velocity. From this we deduce that the bumpy nature of the energy landscape is primarily due to the nonlinearity of the model, rather than the aliasing effects of advection on a domain with periodic boundary conditions.

Figure 2.34(a) shows samples of c_1 from a Markov chain seeded at the truth, $(0.5, 1.0)$, but with $\beta_c = \sigma^2 = 10^{-4}$. Figure 2.34(b) shows the same but for c_2 . Note how the sampler has settled to a posterior mode that corresponds exactly to the true wave velocity. Parameter values corresponding to good quality samples here are obtained through knowledge of the truth; small random walk step size and small observational noise. Knowledge of the truth is an undesirable thing to require to draw samples from the ‘true’ posterior mode. Instead, an a posteriori least-squares minimisation can be done (as in section 2.5.3) to obtain a ‘good’ Markov chain seed, rather than requiring the truth exactly.

One can observe metastability in the Markov chain given the evidence provided in figure 2.30, to see this more visually, figure 2.35 shows 1000 samples from the end of the chain corresponding to figures 2.30(e) and 2.30(f) on top of a plot of the functional $\Phi(u, c) := |\mathcal{G}(u, c) - y|_B$. Note that it is apparent that samples of the wave velocity are being drawn from a region where Φ attains a local minimum.

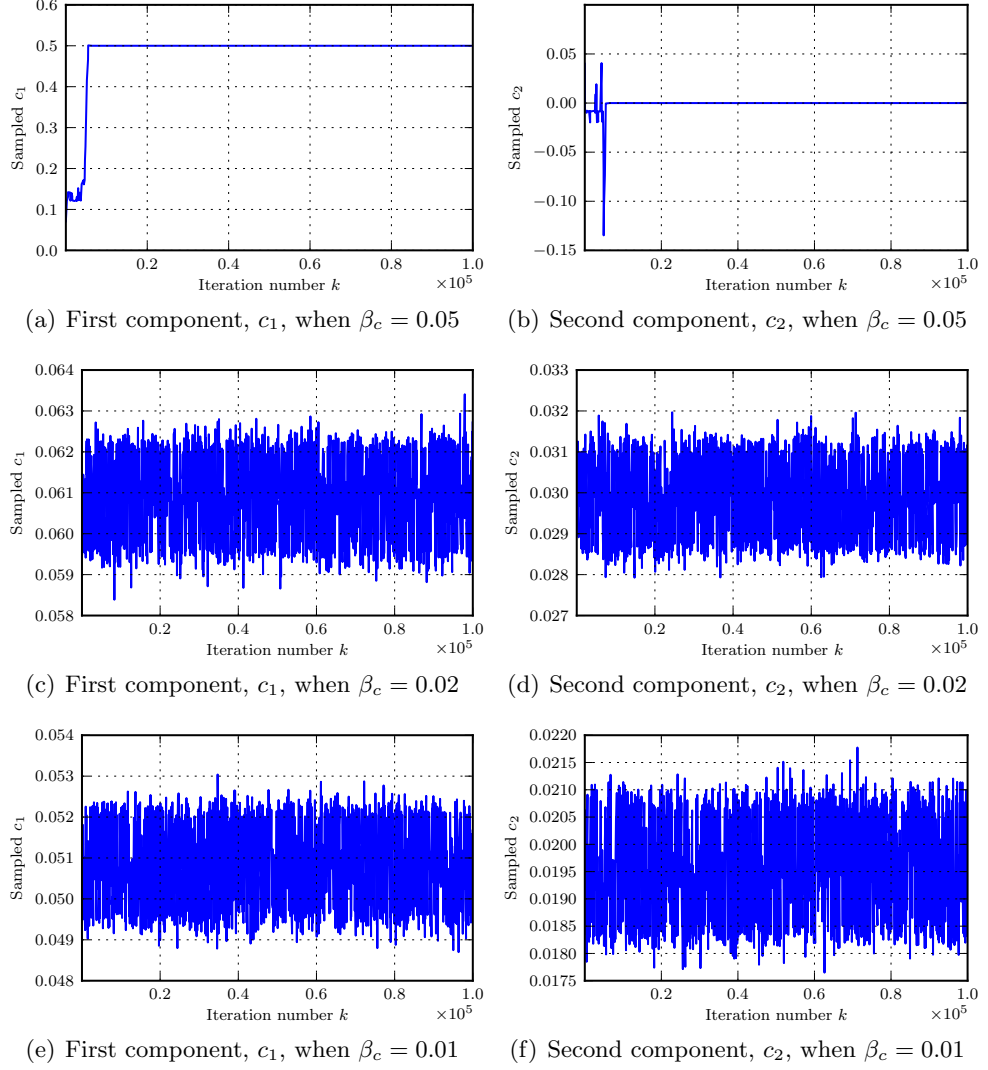
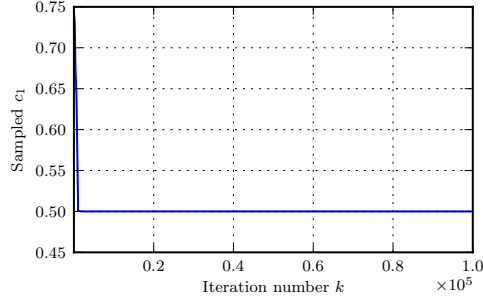
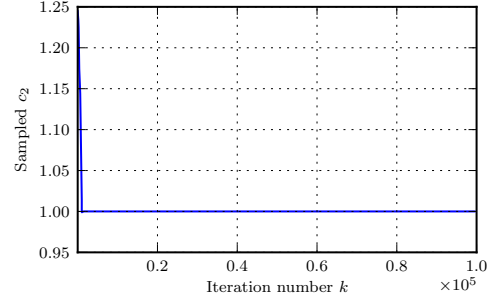


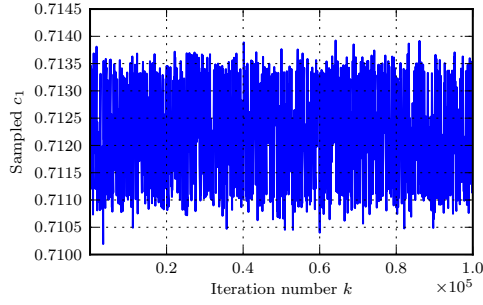
Figure 2.30: Plots of c with random MCMC seed drawn from $\mathcal{N}((0,0), (0.1)^2)$ for varying β_c . Notice that each chain roughly stays in one mode until β_c becomes large enough to allow the chain to jump to a different mode.



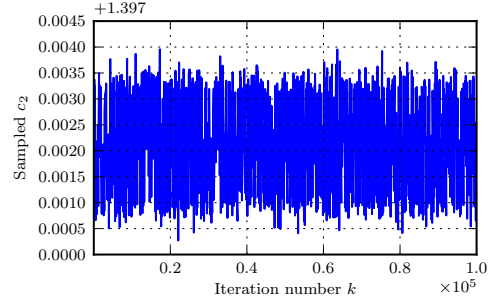
(a) First component, c_1 , when $\beta_c = 0.05$



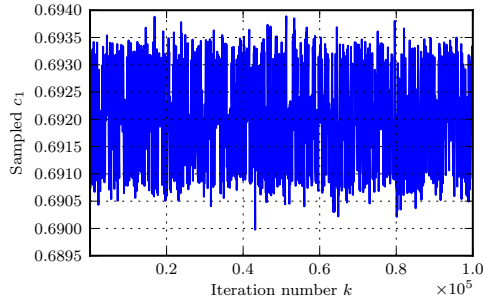
(b) Second component, c_2 , when $\beta_c = 0.05$



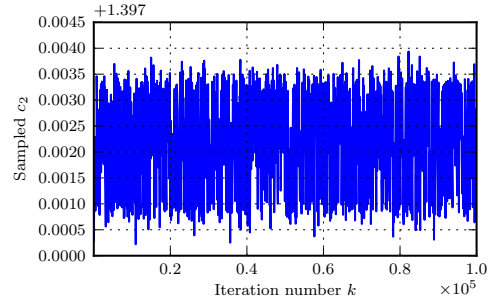
(c) First component, c_1 , when $\beta_c = 0.02$



(d) Second component, c_2 , when $\beta_c = 0.02$



(e) First component, c_1 , when $\beta_c = 0.01$



(f) Second component, c_2 , when $\beta_c = 0.01$

Figure 2.31: Plots of c with MCMC seed at $(0.7, 1.4)$ for varying β_c . . Even in the case where we hand-pick the seed, each chain roughly stays in one mode until β_c becomes large enough to allow the chain to jump to a different mode.

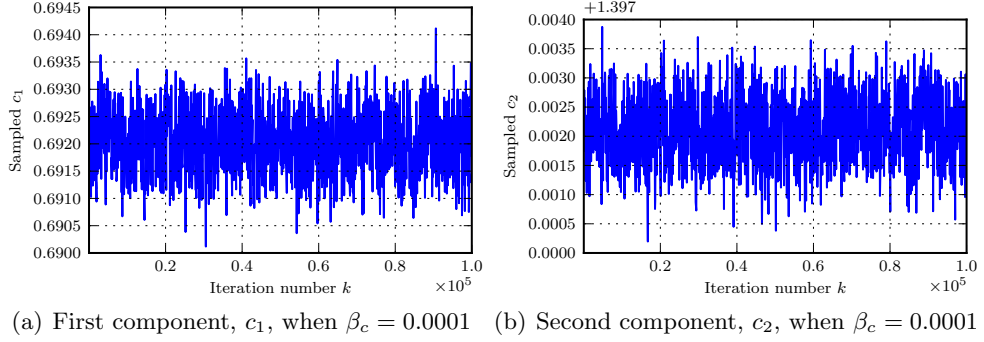


Figure 2.32: Plots of c with MCMC seed at $(0.7, 1.4)$ for varying β_c . Hand-picking the seed and drastically decreasing β_c compared to figure 2.31, we see that the chain samples around the seed, crystallising the multimodal nature of the posterior distribution.

2.7 Conclusions

Throughout this chapter we have studied the Bayesian inverse problem posed for the linear advection equation in two dimensions under a multitude of different scenarios. The Gaussian problem on the initial condition is studied in section 2.2. Varying the observational noise and MCMC proposal step size, we ensure that the numerical scheme set in place to probe the posterior distribution is functioning correctly. In general, MCMC samplers for Gaussian problems are largely unneeded due to their analytic formulation, however the case we have presented here a) is extremely high dimensional and storage of the associated covariance operator is intractable; and b) utilises a prior measure whose covariance is diagonal in Fourier space, making proposals cheap to compute. There is no ‘curse of dimensionality’ for MCMC samplers. We are restricted only by the convergence rate of $\mathcal{O}(\sqrt{n})$ set out by the Central Limit Theorem which underpins any statistical integration technique. Though slow, since the prior measure has a diagonal covariance operator, we consequently make a large number of draws from the posterior to mitigate sampling error.

Section 2.3 sees the application of a commonly used statistical technique in computational physics; simulated annealing. We apply this method to the posterior distribution on the wave velocity parameter, keeping the initial condition fixed at the truth. The sampler is successfully steered to a new mode in the posterior distribution, but is not guaranteed to be near the truth. The method is expensive, requiring several heating/cooling cycles to thoroughly explore the admissible states for the wave velocity. We execute only one cooling routine and have limited success

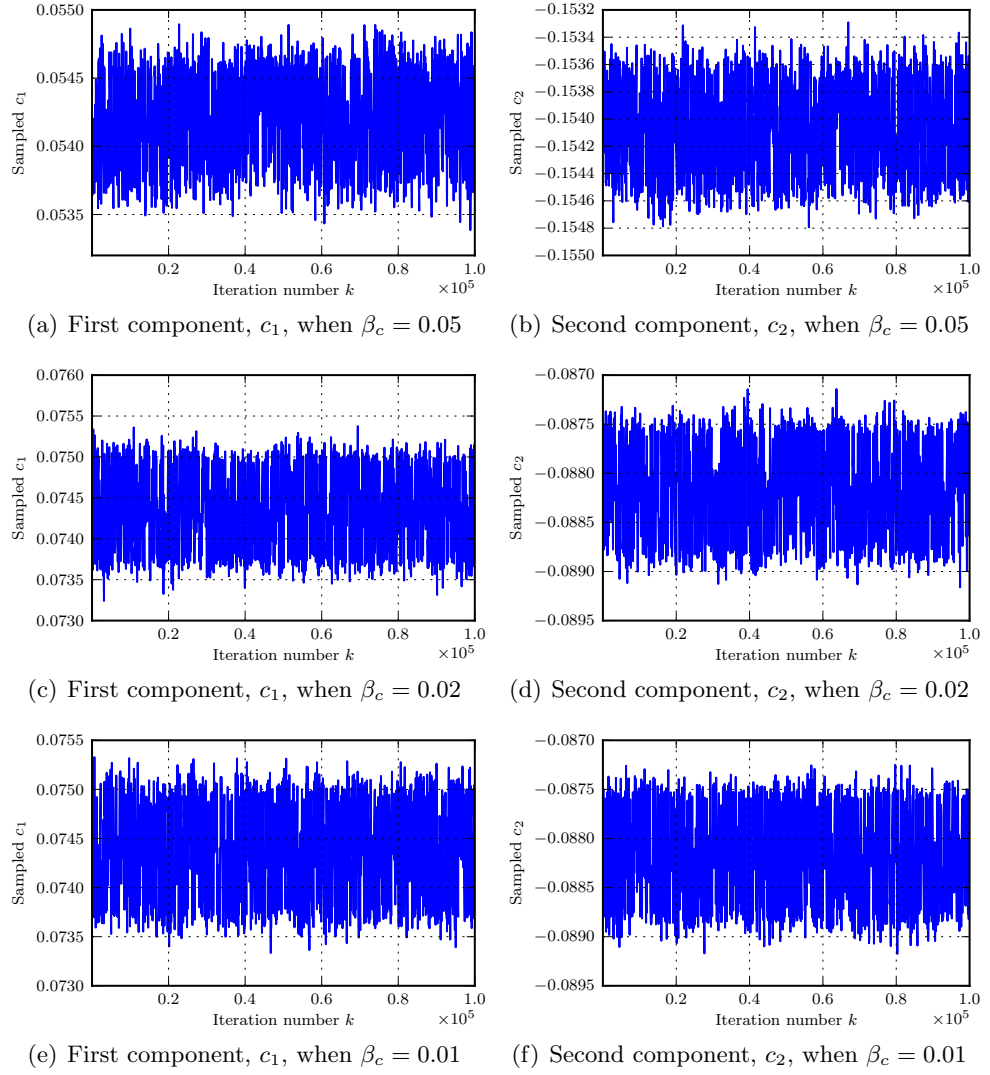
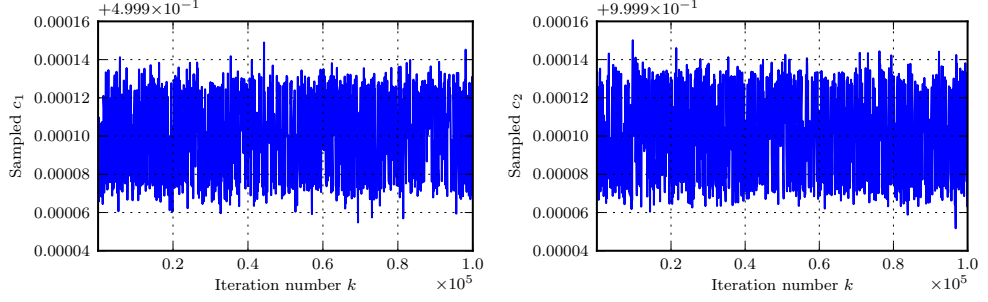


Figure 2.33: Plots of c with random MCMC seed drawn from $\mathcal{N}((0,0), (0.1)^2)$ for randomly chosen observation times and varying β_c



(a) First component, c_1 , when $\beta_c = 0.0001$ (b) Second component, c_2 , when $\beta_c = 0.0001$

Figure 2.34: Plots of c with MCMC seed at $(0.5, 1.0)$ for varying β_c .

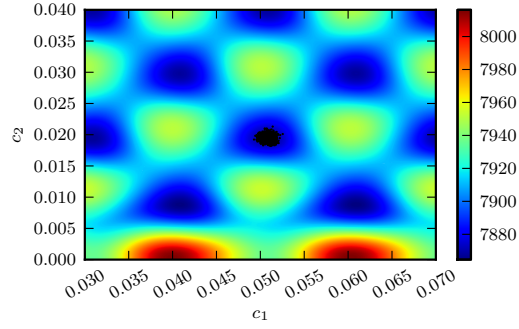


Figure 2.35: Plot of samples of c (black dots) on top of contour plot of $\Phi(u, c)$. Each blue region corresponds to a local minimum of the log-likelihood. This illustrates the multimodal landscape of the posterior distribution.

in obtaining good quality samples.

We have studied an infinite dimensional state estimation problem in the presence of model error in section 2.4. For the statistical model of an advection equation on a torus, with noisily observed functions in discrete time and space, the large data limit of the inverse problem recovers the truth in the perfect model scenario (theorem 2.4.1 and figure 2.17(b)). If the actual wave velocity differs from the true wave velocity in a time-integrable fashion then the solution to the inverse problem is in error by a constant phase shift (theorem 2.4.5 and figure 2.17(e)), determined by the integral of the difference in wave velocities. When the difference in wave velocities is constant the inverse problem solution recovers a subset of the Fourier modes of the truth in the large data limit (theorem 2.4.4, figure 2.17(c) and figure 2.17(d)). When the difference in wave velocities is a fluctuating random field, however small, the truth is not recovered in the large data limit. We have considered the dynamics as a hard constraint, and do not allow for the addition of mean zero Gaussian noise to the time evolution of the state. Adding such noise to the model is sometimes known as a weak constraint approach in the data assimilation community and the relative merits of hard and weak constraint approaches are widely debated; see Bennett [2002]; Apte *et al.* [2008b] for discussion and references. New techniques of analysis would be required to study the weakly constrained problem, because the inverse covariance does not evolve linearly as it does for the hard constraint problem we have explored here. We leave this for future study. There are a number of other ways in which the analysis and methods here could be generalised in order to obtain a deeper understanding of filtering methods for high dimensional systems. These include: (i) the study of dissipative model dynamics; (ii) the study of nonlinear wave propagation problems; (iii) the study of Lagrangian rather than Eulerian data (see chapter 3). Many other generalisations are also possible. For nonlinear systems, the key computational challenge is to find filters which can be justified, either numerically or analytically, and which are computationally feasible to implement. There is already significant research activity in this direction, and studying the effect of model/data mismatch will form an important part of the evaluation of these methods.

We extend the case of exploring model error to the joint distribution on both the PDE initial condition and the wave velocity parameter (section 2.5). We explore different methods of seeding the Markov chain, seeding at the truth and seeding nearby the truth. We look at the effect of these seeds on the posterior mean of the joint distribution and deduce the root cause of sampler metastability to be a poorly

selected random MCMC seed for the wave velocity. To solve this issue, a novel least-squares method approach is employed using the advection model on the noisy data as input. This results in good convergence of the posterior measure on the wave speed to a Dirac in the large data limit. Further questions involving the use of this approach arise. Firstly, mileage in utilising this approach on very noisy observed data may be very model-dependent. Secondly, a minimisation is easiest to do on the lower wave numbers since observational noise spoils higher frequency structures. One could extend this to a minimisation over all wave number components. This would be interesting further work. Lastly, recall that the dynamics are a hard constraint in this section. Further insight is needed if one is to apply the method of fitting to a model parameter in the advent of a soft model constraint.

Section 2.6 provides numerical results when a non-smooth likelihood norm is utilised over the initial condition. This problem is non-Gaussian with a linear forward operator. The non-Gaussianity arises from assuming the log-likelihood grows only linearly in the tails; a doubly-exponential likelihood distribution of the data/model mismatch. We exposed MCMC as a flexible tool, able to deal easily with non-Gaussian infinite dimensional inverse problems. With this modified likelihood formulation, larger steps in the state space can be taken leading to more efficient state space exploration. Efficient state space exploration is the bedrock of statistical methods, since the $\mathcal{O}(\sqrt{n})$ rate of convergence cannot be improved.

Chapter 3

Data assimilation for controlled testbed ocean drifters

3.1 Overview

In physical oceanographic settings, it is very common to use ocean drifters that move passively with the underlying ocean velocity field. Information regarding the drifter's latitude/longitude position in the ocean is communicated via satellite to oceanographic agencies on land. The observations are noisy and are used in data assimilation schemes in an attempt to estimate the underlying flow. More recently, ocean *gliders* have been designed to scour the Earth's oceans, including oceanic structures below the turbulent boundary layer. Gliders have the capability to control their roll, pitch and yaw underwater by shifting their internal battery to act as a counterweight. Operationally, their objective is to descend into a body of water and take measurements of quantities of interest during an ascent, yielding a vertical profile. Usually drifters are equipped with an array of sensors to measure quantities such as temperature, concentration of suspended solids, salinity, depth, the amount of fluorescent particles and current position. A good overview of some operational ocean apparatus can be found in Rudnick *et al.* [2004].

This chapter has two main components. In the first component (section 3.2) we explore the efficacy of utilising naïve control methods on ocean drifters in a time-independent testbed ocean model. Specifically, we construct rudimentary but effective 'control' procedures designed to push ocean drifters into uncharted flow regimes. The three cases of control we employ here are a purely latitudinal control; a control

of equal magnitude in both the x and y directions; and a control constructed using a posteriori information from a previous assimilation cycle. Harnessing information from the posterior distribution to dictate future modelling is of direct interest to the oceanographic populace. We achieve a sizeable reduction of the posterior variance in the mean flow direction for these three cases of control. We also see that on comparing the posterior variance for the latitudinal and bi-directional controls, similar structures arise when viewed as a function of control magnitude. Furthermore, in the third case, we use values of the control magnitude such that the drifter *always* leaves the recirculation regime. For most of these values the true drifter paths are very similar and so we see no visible benefit in the posterior variance. However, when the control magnitude is large enough such that the true path navigates towards a hyperbolic fixed point of the drifter equation in a new flow regime, we see a substantial reduction in variance. Hyperbolic fixed points of the drifter equations join transport barriers in the flow and act as a boundary to observations. Observing near these points outweighs the negative effects produced by polluting the observations with a large control size relative to the size of the flow. This gives a novel geometric correspondence between the control utilised here and the structure of the posterior variance as a function of control magnitude and, consequently, the structure of the true drifter path.

The second component of this chapter is section 3.3. This section adds a time-periodic disturbance to the testbed time-independent ocean model explored in the first component. We continue to analyse controlled ocean drifters in this time-periodic case using exactly the same three cases of control mentioned above. In the third case, the a posteriori control we construct in section 3.3.2 will be time-dependent. For the first two cases, controls that are purely latitudinal and controls that are bi-directional, we make a novel link to the unperturbed time-independent flow case. When the control magnitude is such that the drifter leaves the eddy in the *unperturbed* flow, the posterior variance on the initial condition for the time-periodic model decreases. In other words, we show *robustness* of posterior variance as a function of the perturbation parameter. This robustness also carries over to the transient within-eddy increase in variance also seen in the time-independent case. When employing a time-dependent a posteriori control, we see no overall net gain in posterior variance over the uncontrolled case. For our particular flow and drifter initial condition, it is the case that the uncontrolled drifter path explores a hyperbolic fixed point of an eddy in the time-dependent flow more effectively than the controlled path.

3.2 Time-independent flow

We take a very similar approach to the one outlined in section 2.2. We have a model to describe the time evolution of fluid flow. If we are given the initial condition for this model then we can integrate it to obtain all flow states at any future time. In its most general form, we will concern ourselves with an important kinematic travelling wave, v , [Samelson & Wiggins, 2006], the stream function of which will be denoted ψ ,

$$\psi(x, y, t) = -\pi y + \sin(2\pi x) \sin(2\pi y) + \varepsilon \psi_1(x, y, t), \quad (3.1)$$

where $x, y \in \mathbb{T}^2$. We view ψ_1 as a perturbation to the underlying large scale structure denoted by the first two terms. The parameter $\varepsilon > 0$ is a perturbation parameter. For now, we will take $\varepsilon = 0$ and model a time-independent flow, $v: \mathbb{R}^2 \rightarrow \mathbb{R}^2$ with periodic boundary conditions on \mathbb{T}^2 . Lagrangian data assimilation necessitates appending the fluid model with the evolution equation of tracers in the flow,

$$\text{Fluid model: } \frac{\partial v}{\partial t} = 0, \quad t > 0, \text{ and} \quad (3.2a)$$

$$\text{Drifter model: } \frac{dz}{dt} = v(z), \quad t > 0, \quad (3.2b)$$

with initial conditions,

$$\begin{aligned} v(x, 0) &= u(x), \quad x \in \mathbb{R}^2 \\ z(0) &= z_0. \end{aligned}$$

We will take the stream function of $u(x)$ to be (3.1) with $\varepsilon = 0$. Now we make observations, y_k , of positions of passive tracers in the fluid, z , at times t_k for $k = 1, \dots, K$.

$$y_k = z(t_k) + \eta_k, \quad \eta_k \sim \mathcal{N}(0, \sigma^2 I_2), \quad (3.3a)$$

$$\rightsquigarrow y = \mathcal{G}(u) + \eta, \quad \eta \sim \mathcal{N}(0, \sigma^2 I_{2K}). \quad (3.3b)$$

Assuming the initial conditions of the drifters are known, the goal is to find the initial condition to the *flow* equation (3.2a). This set-up fits into the framework outlined in section 1.4. The solution to this inverse problem is a probability distribution $\mathbb{P}(u|y)$. The prior distribution will be exactly as in section 2.2, and draws from it will be constructed exactly the same way (1.15).

3.2.1 Naïve control strategy

Figure 3.1 explicitly illustrates the stream function of u , the red crosses in which denote zeros of u . These are consequently fixed points of (3.2b). These points essentially constrain the flow structure. Notice the three main flow regimes; the *recirculation* regimes in the bottom-left and top-right, and the *meandering jet* regime that lies between them. The mean flow along streamlines is from left to right. So particles in the bottom-left recirculation regime flow clockwise, in the top-right recirculation regime they flow anticlockwise and in the meandering jet regime they flow from left to right. The associated vector field to ψ is given by,

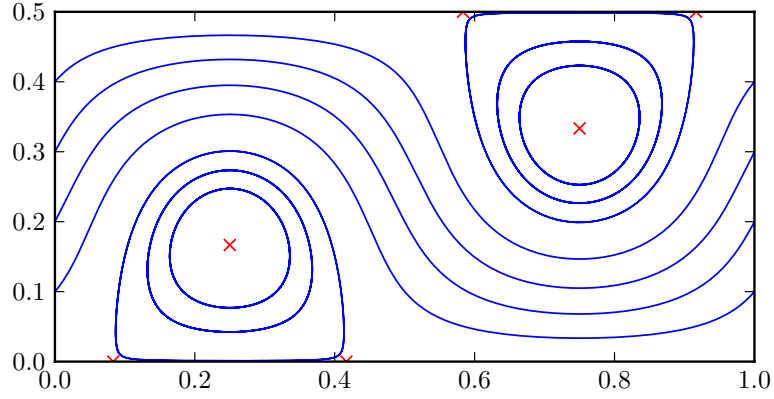


Figure 3.1: Stream function of initial condition (3.1). The blue lines indicate the stream lines and the red cross are zeros of the flow

$$v_1 = -\frac{\partial\psi}{\partial y} = \pi - 2\pi \sin(2\pi x) \cos(2\pi y) \quad (3.4)$$

$$v_2 = \frac{\partial\psi}{\partial x} = 2\pi \cos(2\pi x) \sin(2\pi y). \quad (3.5)$$

Solutions to (3.2b) will be trajectories that are encapsulated by the flow regime boundaries discussed above. Furthermore, these trajectories will be entirely contained within level sets of ψ . It is very often the case that large-scale eddy structures in the ocean trap ocean drifters, preventing them from exploring the state space, leaving a large area of unobserved locations. It is in this section that we explore the effect of forcing a drifter out of a trapped regime to assess the net information gain of observing over a larger portion of the domain. Our initial attempt at achieving

this is to make a slight modification to (3.2b),

$$\text{Fluid model: } \frac{\partial v}{\partial t} = 0, \quad t > 0, \text{ and} \quad (3.6a)$$

$$\text{Uncontrolled drifter: } \frac{dz}{dt} = v(z), \quad 0 < t < t_{K/2}, \quad (3.6b)$$

$$\text{Controlled drifter: } \frac{dz}{dt} = v(z) + f(z), \quad t_{K/2} \leq t \leq t_K. \quad (3.6c)$$

We solve this using a standard method in numerical ODEs: the explicit 4th order Runge-Kutta method. We set $K = 2 \times 10^4$ and observe at every time-step. The choice of the function f is of paramount importance, necessitating investigation as to which choice is the ‘best’. As a simple first attempt, we will choose,

$$f_1(x) = \zeta \quad (3.7)$$

$$f_2(x) = 0, \quad (3.8)$$

so that solutions to (3.6c) behave like those to (3.2b) with a naïve latitudinal control applied. The strength of this control is parameterised by ζ . Practically and operationally, it is the case that ocean drifters may be given a ‘mission plan’ to cruise a certain part of the ocean. There is no real constraint on travel direction in the mission plan, drifters (at depth) may adjust their roll, pitch and yaw to go to any specified location. The only practical constraint placed on oceanographic machinery is speed relative to the underlying flow. In our case this corresponds to the parameter ζ and practically this is usually about 30 cm/s [Rudnick *et al.*, 2004].

If the initial condition of (3.6c) lies within the bottom-left eddy regime, sufficiently large ζ will force the drifter out of the eddy and into the meandering jet, leading to a more diverse path along which we can observe. Figure 3.2 shows an integral curve solving (3.6c) when $\zeta = 0$. As mentioned above, the trajectory is contained entirely within a streamline of (3.1). Figure 3.3 shows an integral curve solving (3.6c) when $\zeta = 2$. Notice that at time $t = t_{K/2}$, the forcing is instantaneously switched on and the drifter can escape from the recirculation regime, cross a fluid transport boundary and enter the meandering jet regime. Notice also that the drifter comes back in to the recirculation regime exactly one period later. The path traced out by the drifter is contained entirely within a streamline of $\psi + f$. We wish to compare the posterior distributions parameterised by ζ , so we observe positions of drifters

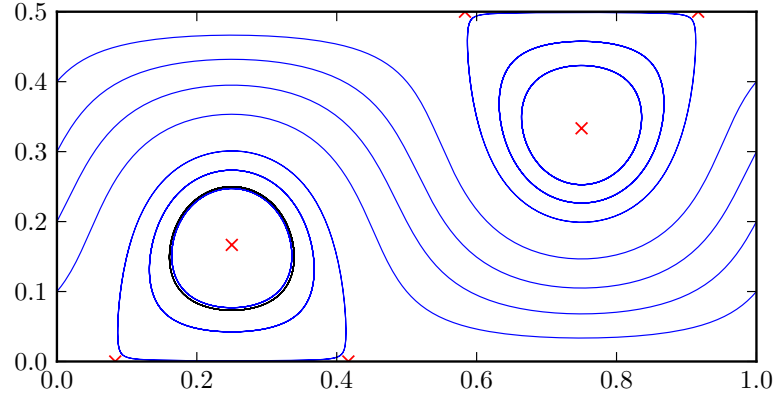


Figure 3.2: Solution to (3.6c) with $\zeta = 0$, $z_0 = (0.25, 0.25)$, depicted by the black curve. The stream lines (blue) and zeros (red crosses) are shown for reference

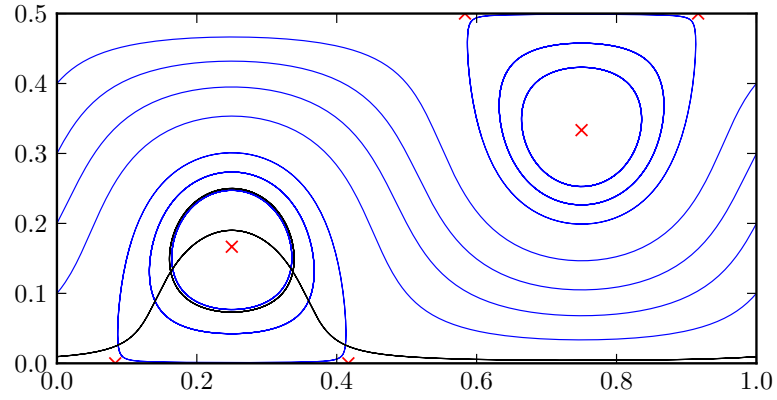


Figure 3.3: Solution to (3.6c) with $\zeta = 2$, $z_0 = (0.25, 0.25)$, depicted by the black curves. Passive model solutions coincide with stream lines. Controlled model solutions may escape the eddy. Stream lines (blue) and zeros (red crosses) are shown for reference

that trace out the paths governed by solutions to (3.6b)–(3.6c),

$$y_k = z(t_k) + \eta_k, \quad \eta_k \sim \mathcal{N}(0, \sigma^2 I_2), \text{ for } k = 1, \dots, K, \quad (3.9a)$$

$$\rightsquigarrow y^\zeta = \mathcal{G}(u) + \eta, \quad \eta \sim \mathcal{N}(0, I_{2K}). \quad (3.9b)$$

The superscript denotes the dependence of the data on the strength of the latitudinal control used. The reason for this is so that we can look at the resulting posterior distribution, $\mathbb{P}(u|y^\zeta)$, for various values of the parameter ζ .

Tempering the likelihood

Real drifters out in the ocean are communicating their position to land by GPS. We think of the x -coordinate in the domain of our model as being on the order of several kilometres long. GPS systems nowadays are accurate to within a few metres, depending on the number of satellites available for triangulation. As a result, a good value for the observational error of the positions of drifters is $\sigma^2 = 10^{-6}$. In other words, the observational data is very accurate. A penalty for making good observations is that the posterior distribution becomes very peaked. This has the same effect as extending the tails of the posterior probability distribution function and can significantly increase the time one needs to burn in the sampler. See figure 1.2 for an illustration of this effect. Figure 3.4 shows the trace plot of Φ , the negative log-likelihood. As we can see, even after burning in, the sampler is still making its way through the state space with no signs of settling down into a mode. Compare this with figure 3.5, which shows improved performance. The sampler still appears to not have settled down by the time the burn-in ends, but it is slowing down. Finally, figure 3.6 shows the desired behaviour, the sampler initially takes big jumps and, after some transient behaviour, samples a mode where samples lie a distance of about 10^4 , in the 2-norm, from the true initial condition. In each case mentioned above, the Markov chain was burnt in for 10^6 iterations. Afterwards, we draw 10^6 samples from the posterior distribution. The chain was also seeded from the true initial condition. At the true initial condition, the following approximation can be

made,

$$\begin{aligned}
\Phi(u) &= \frac{1}{2\gamma^2} \|\mathcal{G}(u) - y\|^2 \\
&= \frac{1}{2\gamma^2} \sum_{k=1}^K |\mathcal{G}_k(u) - y_k|^2 \\
&\approx \frac{1}{\gamma^2} \sum_{k=1}^K \mathcal{O}(\sigma^2) \\
&\approx \frac{1}{\gamma^2} \mathcal{O}(K\sigma^2).
\end{aligned}$$

Throughout the rest of this section, the burn-in period and sampling period will both be 10^6 iterations.

We have $K = 2 \times 10^4$ and observe at every time-step, so when $\gamma^2 = \sigma^2$, as in figure 3.4, the value of the Φ at the beginning of the burn-in should be $\Phi(u) \approx 2 \times 10^4$. This is exactly what we see. The cases $\gamma^2 = 10^2 \times \sigma^2$ and $\gamma^2 = 10^4 \times \sigma^2$ can be just as easily calculated and we obtain $\Phi(u) \approx 2 \times 10^2$ and $\Phi(u) \approx 2$, respectively. Compared to the scale, these are close to zero, as can be confirmed in figures 3.5 and 3.6 respectively.

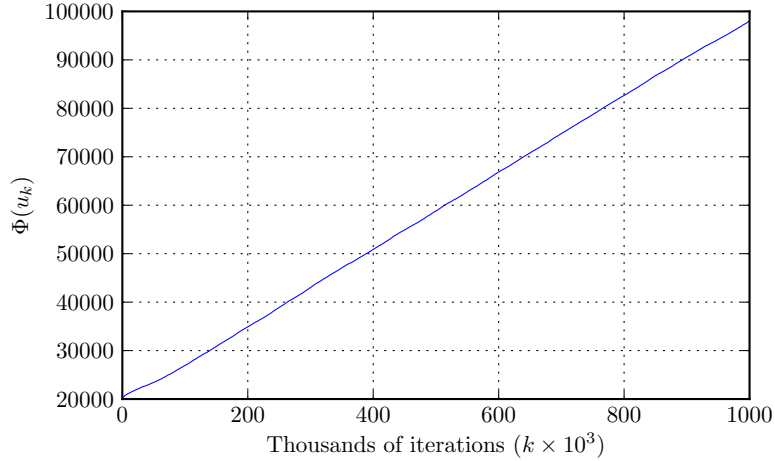


Figure 3.4: Negative log-likelihood during the burn-in for $\sigma = 10^{-6}$

From these results we learn that the nonlinearity and under-determined behaviour of this system are obstacles to the Monte Carlo sampler. To mitigate the effects of these obstacles, one option is to temper the likelihood to increase convergence time to a mode in the sampler. We choose $\gamma^2 = 10^4 \times \sigma^2$ for our purposes.

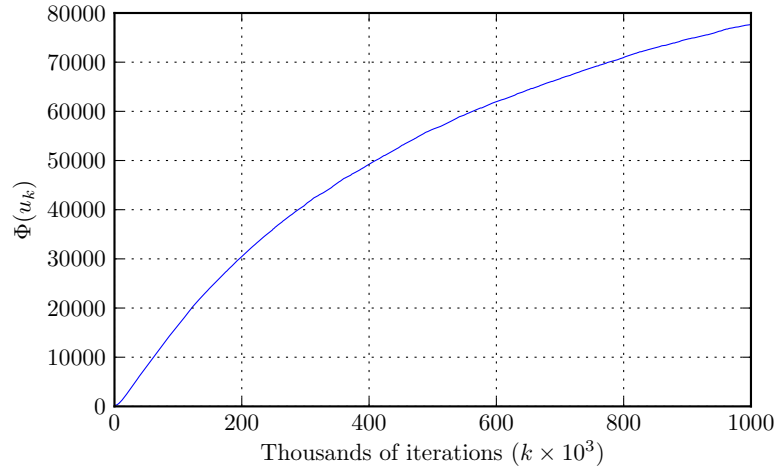


Figure 3.5: Negative log-likelihood during the burn-in for $\sigma = 10^{-4}$

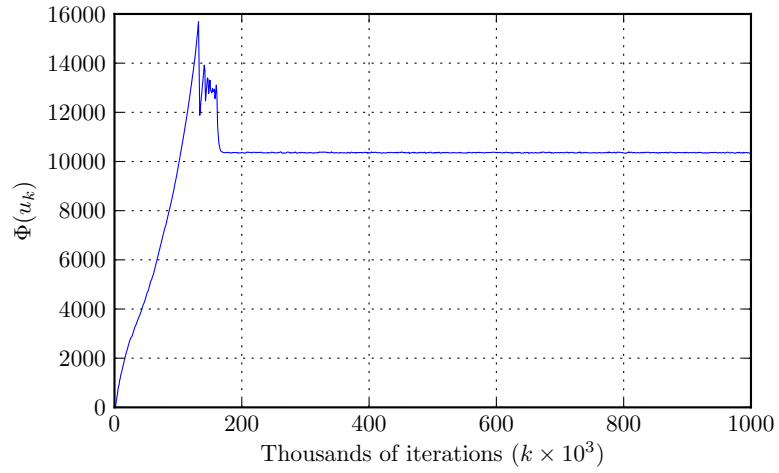


Figure 3.6: Negative log-likelihood during the burn-in for $\sigma = 10^{-2}$

Seeding from the prior

We keep the same numerical parameters as above, with $\sigma^2 = 10^{-6}$, for the rest of this section. Figure 3.7 shows the variance of the posterior distribution on the horizontal component of the velocity field with no control present, $\text{Var}(u_1|y^0)$. Compare this with figures 3.8(b) and 3.8(a), which show $\text{Var}(u_1|y^3)$ and $\text{Var}(u_1|y^2)$ respectively. Notice that, with the control switched on (figure 3.8), the variance around the area of the path of observations is smaller relative to the other parts of the domain. This is expected behaviour. What is unexpected is that, with a stronger control, the observation path is longer and so one would expect to see a reduction in variance in this region. The opposite is true. With a stronger control, one learns less about the horizontal component of the velocity field relative to ζ , in this case by an order of magnitude (see colour scale). This can be explained by looking at (3.6c). Notice here that we observe a path forced by f and then assimilate to get an estimate of v . So it is reasonable to think that as ζ increases to the point where the size of f relative to the size of v is big, one is mainly observing effects imposed by f . The latitudinal control is polluting the observations of v to the point where it does not matter what your estimate for v is, you will still get good observations under f .

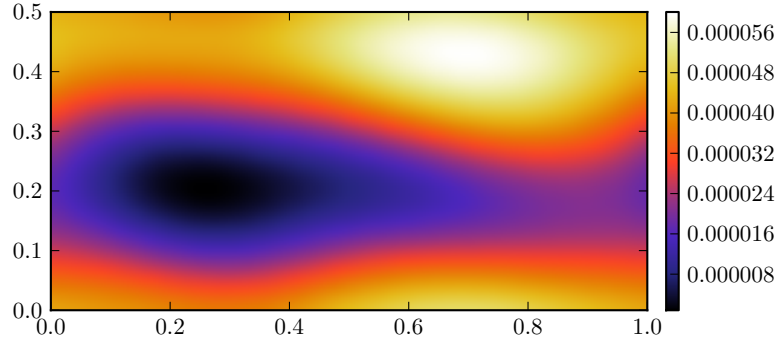


Figure 3.7: $\text{Var}(u_1|y^0)$. MCMC seeded with prior draw.

These calculations were done with a random seed for the Markov chain. Typically these are just random draws from the prior distribution. Unfortunately, since the model we assimilate into is not linear, it could be the case that the sampler sits in a different mode for each of these cases. Figure 3.9 shows where the mode is for the case $\zeta = 0$. Note the difference for the cases $\zeta = 2, 3$, given by figures 3.10 and 3.11, against the case $\zeta = 0$. It looks like the modes have shifted far away from the case $\zeta = 0$, but relative to each other are still quite close together.

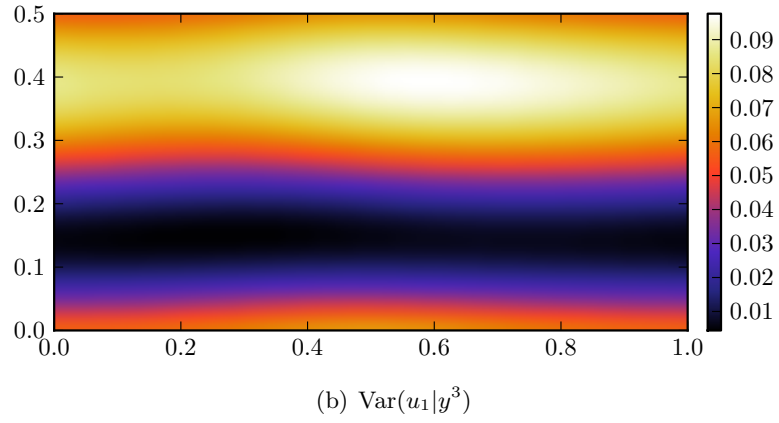
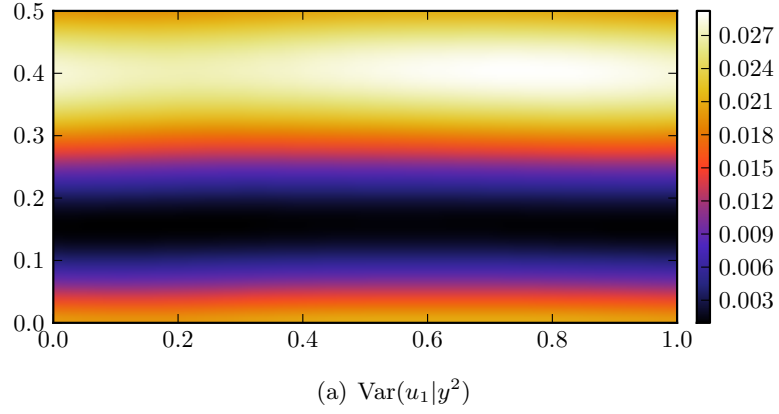


Figure 3.8: Posterior variances for $\zeta = 2, 3$. MCMC seeded with prior draws. Note the order of magnitude difference in the scales

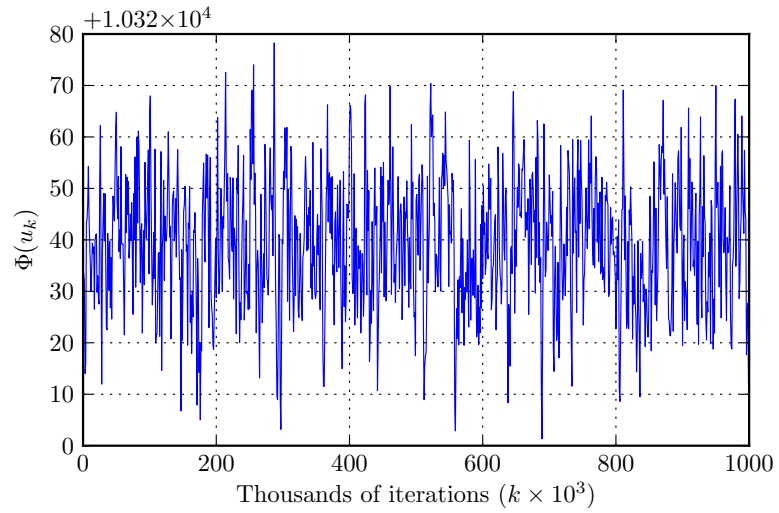


Figure 3.9: Trace plot of negative log-likelihood for $\zeta = 0$, $\sigma^2 = 10^{-6}$

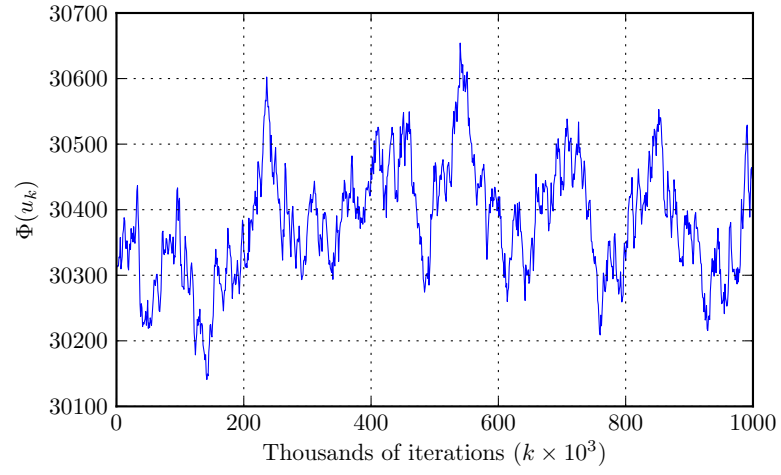


Figure 3.10: Trace plot of negative log-likelihood for $\zeta = 2$, $\sigma^2 = 10^{-6}$

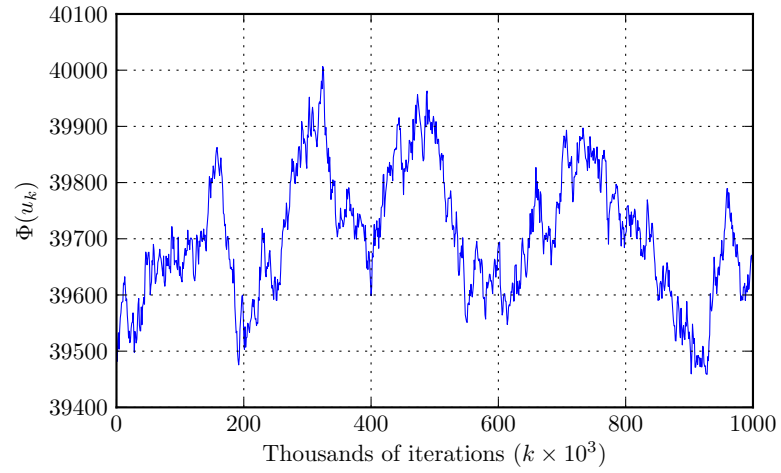


Figure 3.11: Trace plot of negative log-likelihood for $\zeta = 3$, $\sigma^2 = 10^{-6}$

Seeding from the truth

To be more objective in this approach to minimising the posterior variance, it is advisable to be more confident that the sampler is seeded from the same place in each case. The reason for this is that the numerical results should be *reproducible*. For the sake of convenience, we will pick the true initial condition u to be the seed for the Markov chain in each case. For the case where there is no control present, $\zeta = 0$, figure 3.6 shows a trace plot of the log-likelihood during the burn-in phase and figure 3.12 shows the negative log-likelihood after the burn-in phase. Notice the sampler stays in a mode and samples the state space well; the desired behaviour. Unfortunately, for the case $\zeta = 2$ this is a different story. As illustrated in figure 3.13, we see that the sampler is not steadily sampling one region of the state space, but is wandering in a near-linear fashion. In this case, the length of the burn-in is 10^7 iterations. We conclude that just seeding from the truth is not enough to achieve convergence. The next section describes another approach.

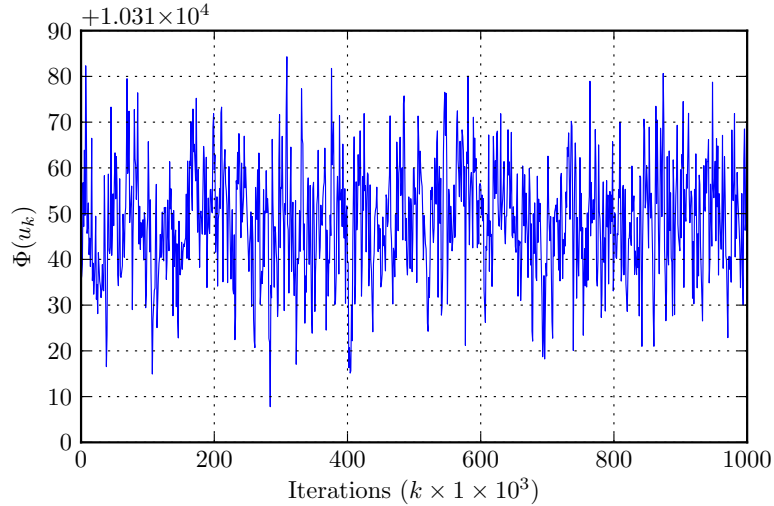


Figure 3.12: Trace plot of negative log-likelihood for $\zeta = 0$ and $\sigma^2 = 10^{-6}$ when MCMC sampler is seeded from the true initial condition.

Re-weighting the prior

For some values of ζ we still have problems getting the sampler to converge to a posterior mode, especially when it is seeded from the true initial condition. Another

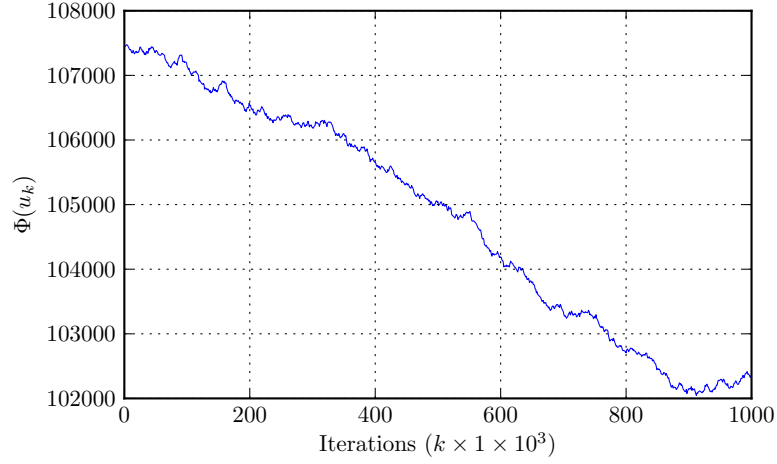


Figure 3.13: Trace plot of negative log-likelihood for $\zeta = 2$ and $\sigma^2 = 10^{-6}$ when MCMC sampler is seeded from the true initial condition.

approach to ameliorating this issue is to look again at Bayes' theorem,

$$\mathbb{P}(u|y^\zeta) = \mathbb{P}(y^\zeta|u)\mathbb{P}(u) \quad (3.10)$$

where the prior distribution we have been using is the standard normal distribution $\mathcal{N}(0, (-\Delta)^{-\alpha})$. Notice that the variance – the coefficient in front of the covariance operator – is 1. Such a small variance yields to an overly informative prior distribution and this is almost certainly something that we should be able to choose. We adjust it slightly so that draws from the prior have distribution, $\xi \sim \mathcal{N}(0, \mu^2(-\Delta)^{-\alpha})$. In what follows, $\alpha = 3$ so that draws from the prior lie in H^1 almost surely.

When the likelihood is tempered, the weight of the likelihood distribution compared to the prior distribution decreases and so the sampler will tend to converge more towards the mode prescribed by the prior. In this section we approach this from another perspective, decreasing the weight of the prior distribution compared to the likelihood. This is interpreted as trusting the data far more than trusting the prior. We have to make a choice of what value μ should take. Trace plots of the negative log-likelihood illustrate the behaviour of the Markov chain for various values of μ .

Figures 3.14–3.16 show plots of $\Phi(u)$ for $\sigma = 10^{-6}$ and $\mu = 10^4$ fixed and for $\zeta = 1, 2, 3$, respectively. Each of these plots is after a burn-in of 10^6 draws. Notice that in each case, the Markov chain appears to have converged to a mode. Also observe that, on comparing figure 3.16 with figure 3.13, which required a burn-in of 10^7 draws but was seeded from the truth, we see that the chain appears to have

converged to a mode after only 10^6 draws; an order of magnitude improvement in convergence time. In light of the sampler behaviour discussed above, from here on in all numerical results will utilise a Markov Chain whose seed is the true initial condition of the flow with untampered likelihood ($\sigma^2 = 10^{-6}$) and uninformative prior ($\mu = 10^4$). This is to ensure that any time a sampler is used, it will converge to a similar mode in each experiment.

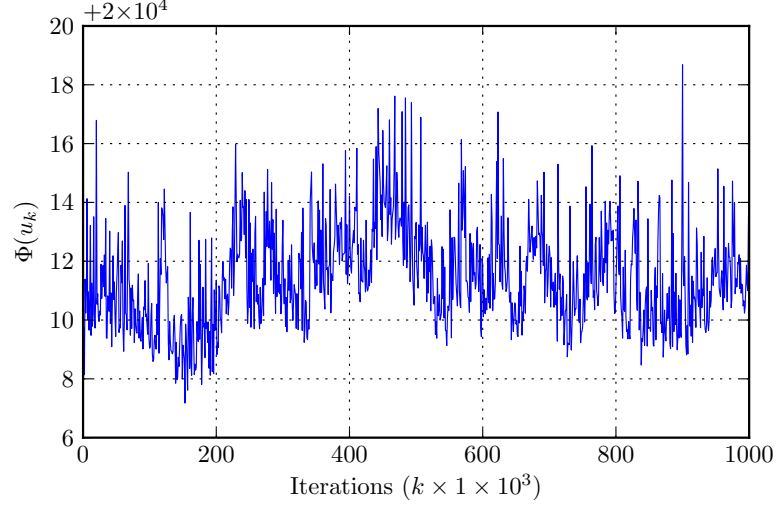


Figure 3.14: Negative log-likelihood when $\zeta = 0$, $\sigma^2 = 10^{-6}$ and $\mu^2 = 10^4$

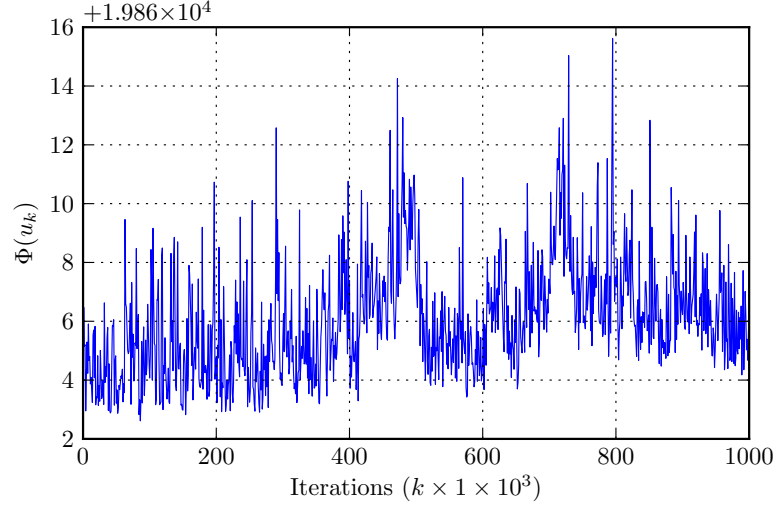


Figure 3.15: Negative log-likelihood when $\zeta = 1$, $\sigma^2 = 10^{-6}$ and $\mu^2 = 10^4$

Figure 3.20 shows the variance of the horizontal component of the flow as a function of control magnitude in the max norm, the L^1 norm and L^2 norm. Meanwhile,

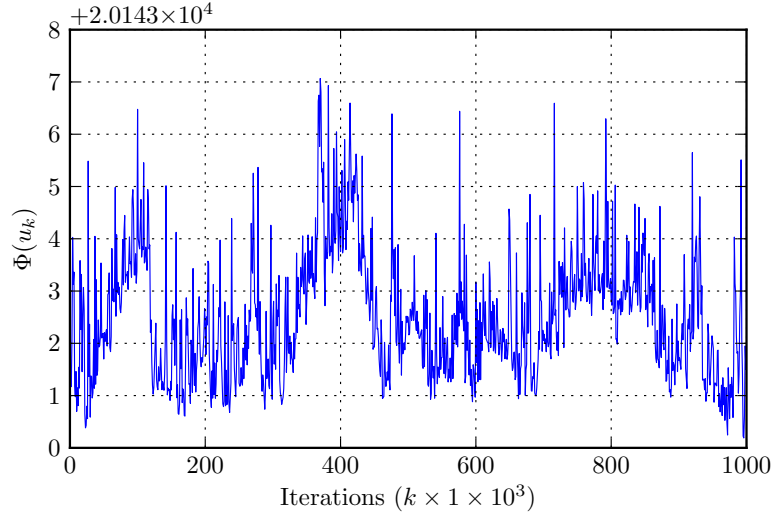


Figure 3.16: Negative log-likelihood when $\zeta = 2$, $\sigma^2 = 10^{-6}$ and $\mu^2 = 10^4$

figures 3.17–3.19 show the variance of the horizontal component of the velocity field, $\text{Var}(u|y^\zeta)$, for $\zeta = 0, 1, 2$. In figure 3.20, the norm of the variance is roughly the same for $\zeta = 0, 1$ and this is backed up by comparing 3.17 with 3.18, where we see that the structure of the variance looks very similar. The area where the variance is smallest is in roughly the same part of the domain. This is explained by the fact that when $\zeta = 1$, the drifter is still within the recirculation regime of the kinematic travelling wave and, as a consequence, the observed data is in the same area of the flow. Also observe that this patch of small variance appears to be slightly bigger in the case where $\zeta = 0$, with the colour scale being roughly the same in each case. Comparing both of these figures with the case $\zeta = 2$ in figure 3.19, we see that the colour scale has shrunk dramatically, and although the dark areas of the variance appear to have shrunk as compared with the previous two cases we see that the variance across the whole domain has decreased by an order of magnitude. Again, referring to figure 3.20, it is indeed the case that the variance drops off in magnitude as ζ increases. Recall that when $\zeta = 2$, the forcing is strong enough to push the drifter out of the recirculation regime and into the meandering jet, leading to observed data entering new parts of the domain.

The red hatched region in figure 3.20 corresponds to values of ζ that are too small to push the drifter out of the eddy. The green hatched region corresponds to values of ζ such that the drifter leaves the eddy. The transition value is not computed exactly and lies somewhere in $\zeta \in (1.5, 1.75)$, this corresponds to the blue hatched region. Experiments were done for $\zeta = 0, 0.25, 0.5, \dots, 2.75, 3$. The case $\zeta = 1.75$ was the

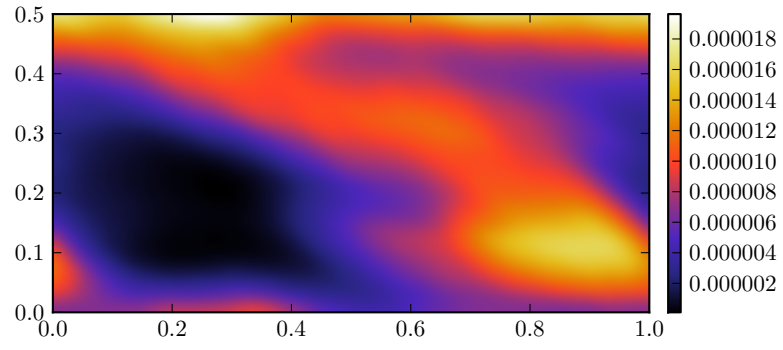


Figure 3.17: $\text{Var}(u_1|y^0)$ when $\sigma^2 = 10^{-6}$ and $\mu^2 = 10^4$

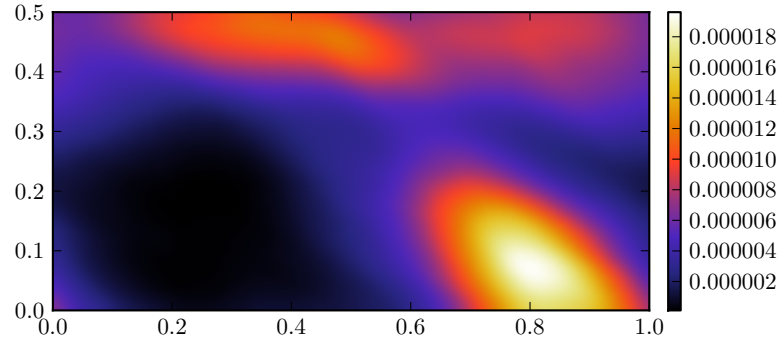


Figure 3.18: $\text{Var}(u_1|y^1)$ when $\sigma^2 = 10^{-6}$ and $\mu^2 = 10^4$

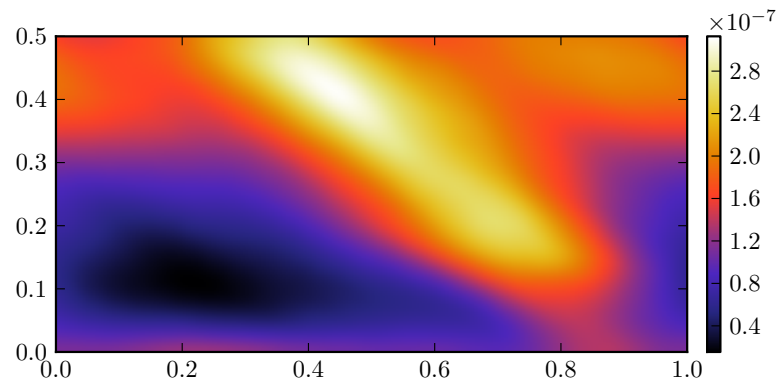


Figure 3.19: $\text{Var}(u_1|y^2)$ when $\sigma^2 = 10^{-6}$ and $\mu^2 = 10^4$

first experiment in which we observed the drifter leaving the recirculation regime. The blue line shows the maximum value of the variance over the domain $[0, 1] \times [0, 0.5]$. The red line and cyan line show the L^2 norm and L^1 norm, respectively. The minimum value of the variance is small enough to be difficult to see on the plot but remains consistently small, so it has been omitted for clarity reasons. There are some notable points to make here. Firstly, in the green region (where the drifter leaves the eddy) we see that the size of the variance decreases in all of our chosen norms. We have learned more about the flow around the truth by forcing the drifter to cross a transport boundary and enter a new flow regime. Secondly, in the red region (where the drifter does not leave the eddy) we see an initial increase in the size of the variance. There are many factors at play here. We will try to shed some light on them. Firstly, for small ζ , the controlled and uncontrolled paths along which we take observations are close. Their closeness and the size of σ^2 creates a delicate interplay between whether they are statistically indistinguishable or not. If they are indistinguishable up to two or three standard deviations, this could explain the increase and then decrease of the variance in the red region. Secondly, as ζ increases initially, we see that the controlled path gets pushed down near the elliptic stagnation point of the flow (see figure 3.21). If this region is an area where the flow is smaller in magnitude than the flow along the uncontrolled path, this is equivalent to an increase in the magnitude of the control relative to the underlying flow. This could lead to the observations becoming polluted by f . This is an effect we have seen before when seeding the Markov chain from the prior.

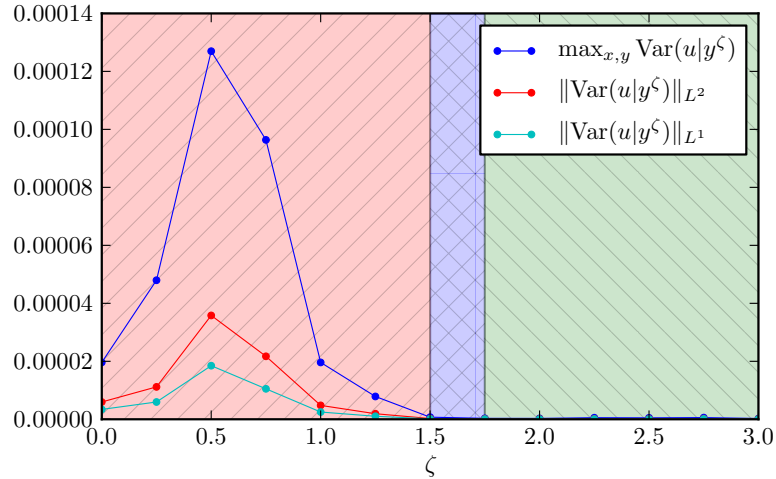


Figure 3.20: $\text{Var}(u|y^\zeta)$ for varying ζ

Exploring this further, we compute the mean magnitude of the flow along the con-

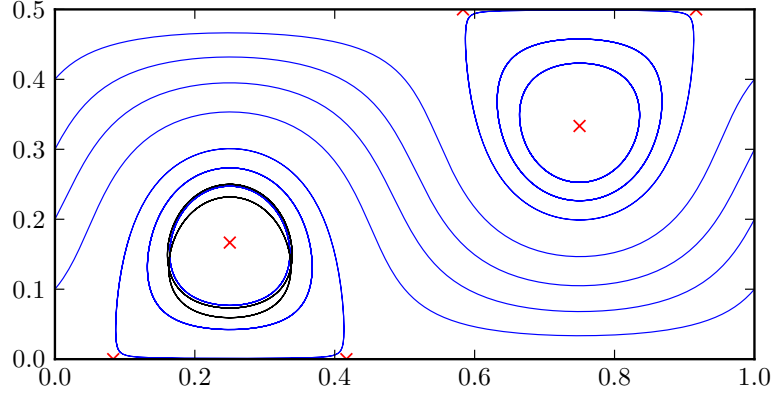


Figure 3.21: Passive path and controlled path for $\zeta = 0.5$. Notice the controlled path is pushed further down.

trolled path of the drifter. More formally, we solve (3.6b)–(3.6c) to obtain a set of points $\{z_k = z(t_k)\}_{k=1}^K$. Then we compute the mean flow magnitude as follows,

$$\langle v \rangle = \frac{2}{K} \sum_{k=K/2+1}^K \|v(z_k)\|. \quad (3.11)$$

This quantity is computed for each fixed ζ and the result is plotted in figure 3.22. The mean flow magnitude is given by the blue line in this figure and the red dotted line depicts the control magnitude. Notice the first three values of ζ for which the mean flow magnitude decreases in 3.22. This is equivalent to an increase in the magnitude of the control relative to the magnitude of the underlying flow and so the information gain from taking observations here decreases. This corresponds nicely with the first three values of ζ in figure 3.20 that show an increase in variance. Notice also that for the other values of ζ , the mean flow magnitude shows a mostly increasing trend, consistent with a decrease in the posterior variance.

Forcing in both coordinate directions

We now explore the effect on the posterior variance of the horizontal component of the velocity field when the forcing function is,

$$f(z) = \begin{pmatrix} \zeta \\ \zeta \end{pmatrix}. \quad (3.12)$$

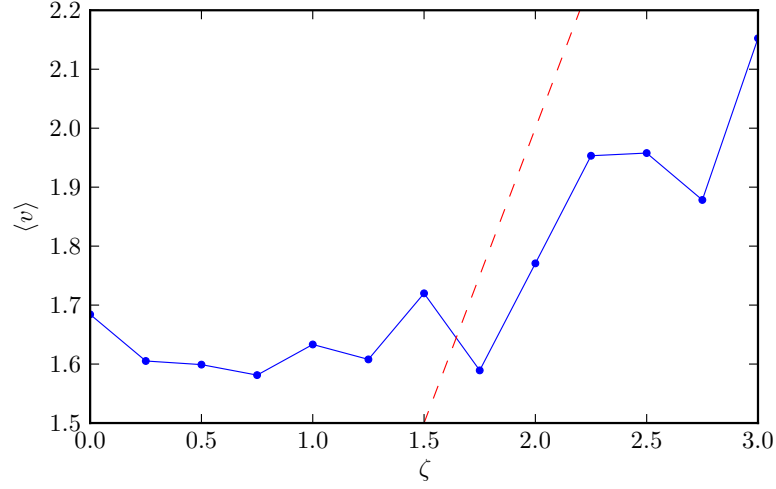


Figure 3.22: Mean flow magnitude (blue line) as a function of control magnitude (red dotted line)

The data, y , is still parameterised by ζ . We will investigate what happens as ζ is varied, just as we have seen in figure 3.20. Figures 3.23–3.25 show plots of the negative log-likelihood for the cases $\zeta = 0, 0.5, 1$, respectively. They all show that the Markov chain has settled down to a mode in the posterior distribution.

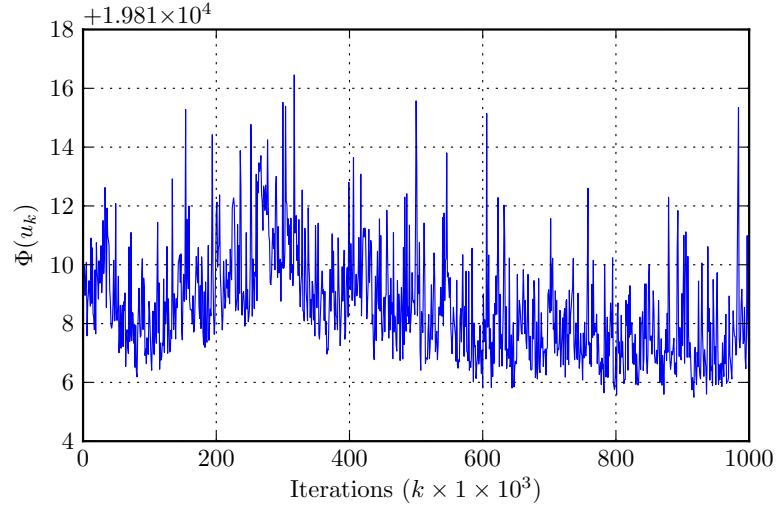


Figure 3.23: Negative log-likelihood when $\zeta = 0$, $\sigma^2 = 10^{-6}$ and $\mu^2 = 10^4$

Now we provide the analogue of figure 3.20 for the new forcing function (3.12). This is shown in 3.26. We see similar behaviour for the variance of the posterior distribution again. In the red hatched region, the values of ζ for which the drifter is

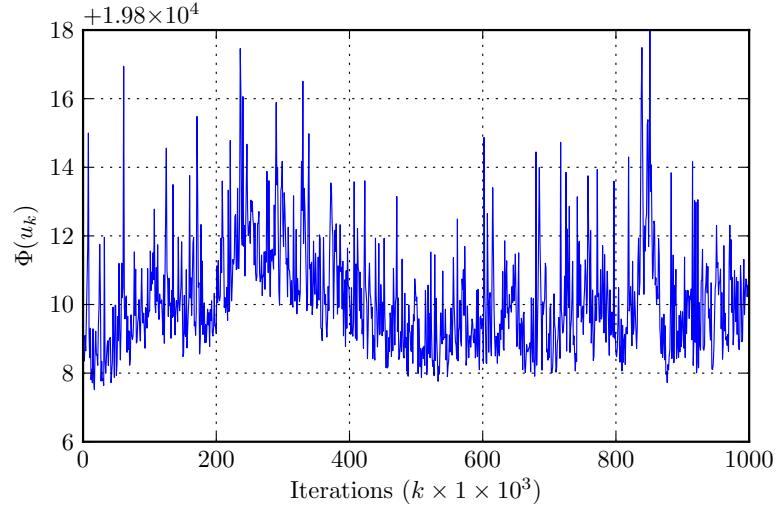


Figure 3.24: Negative log-likelihood when $\zeta = 0.5$, $\sigma^2 = 10^{-6}$ and $\mu^2 = 10^4$

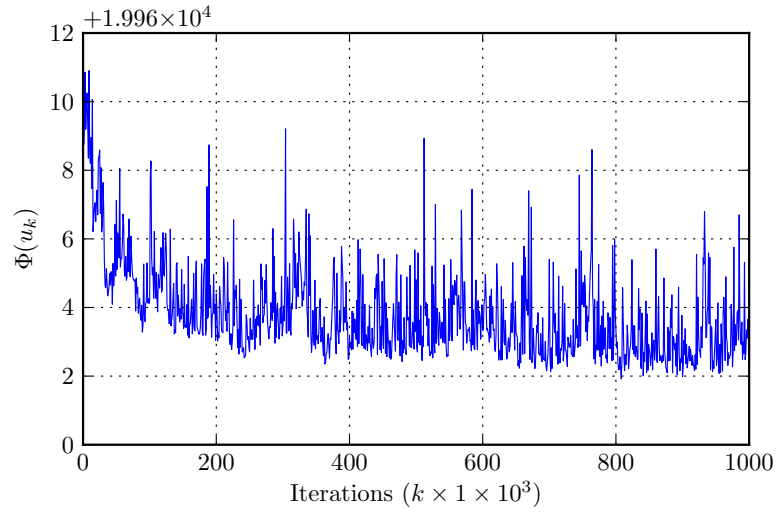


Figure 3.25: Negative log-likelihood when $\zeta = 1$, $\sigma^2 = 10^{-6}$ and $\mu^2 = 10^4$

not forced hard enough to leave the recirculation regime, we see an initial increase in the size of the posterior variance. Then we observe a decrease in posterior variance as ζ approaches a value large enough to push the drifter out of the eddy regime, depicted by the green hatched region. The blue hatched region denotes the range of values of ζ for which it is unknown whether the drifter leaves the eddy or not, due to lack of experiments.

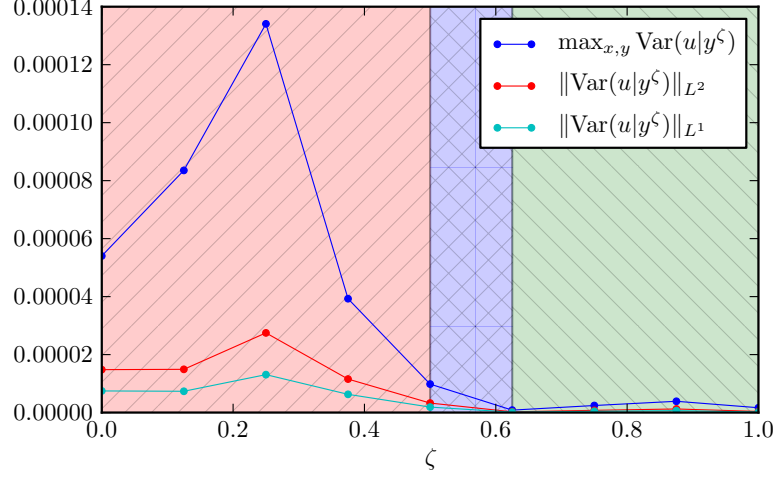


Figure 3.26: $\text{Var}(u|y^\zeta)$ for varying ζ

To explain the initial increase in the posterior variance within the red hatched region, we calculate the mean flow magnitude just as in (3.11). This is shown in figure 3.27. We see an initial period where the mean flow along the controlled path remains almost constant. As a consequence, the magnitude of the forcing increases relative to the magnitude of the flow. This pollutes the observations and leads to an increased posterior variance just as we have observed in the previous section. We also see the opposite effect; the big jump in flow magnitude at $\zeta = 0.5$ (and consequently when the drifter escapes the gyre) is attested as the cause of the decrease in posterior variance as we enter the green hatched region of figure 3.26.

The cases of forcing explored thus far are $f(z) = (\zeta, 0)^\top$ and $f(z) = (\zeta, \zeta)^\top$. The main results are summarised by referring to figures 3.20 and 3.26. In these two cases, we see strikingly similar structure of the posterior variance as a function of control magnitude. The initial increase in posterior variance within the eddy; decreasing posterior variance as the drifter's flow path approaches the transport boundary and small posterior variance (compared to the case $\zeta = 0$) once a new flow regime is being observed. Compare the values of ζ for which this behaviour occurs. Notice that the values of ζ in figure 3.20 are about three times larger than those in figure 3.26.

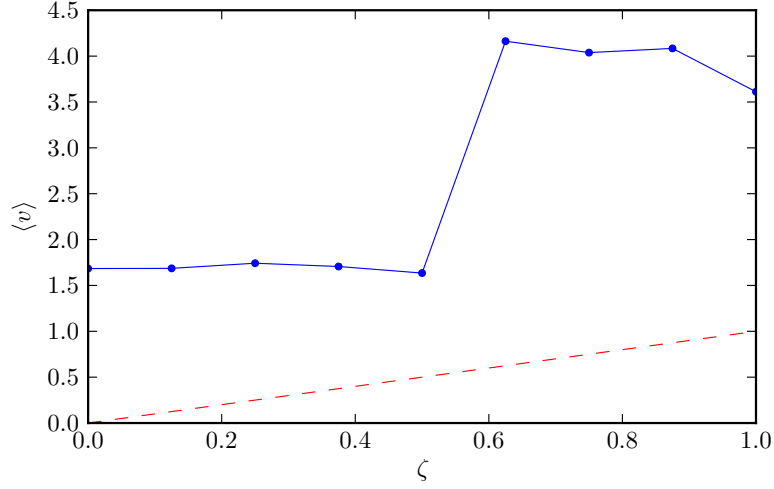


Figure 3.27: Mean flow magnitude (blue line) as a function of control magnitude (red dotted line)

One factor at play here is the relative magnitude of the controls in each case. For $\zeta = 1$, the control has magnitude 1 in the latitudinal case, and magnitude $\sqrt{2}$ in the bi-directional case. Even scaling the results in the bi-directional case by $\sqrt{2}$, notice that the value of ζ for which the drifter first leaves the eddy, is $\zeta = \frac{\sqrt{2}}{2}$ and this is still smaller than $\zeta = 1.5$ for the x -directional case. The final factor affecting the scaling is the dynamics of the system after the forcing has been applied. Controlling in only the latitudinal direction will require a larger magnitude force to push the drifter out of the eddy than when forcing in both the x and y directions simultaneously.

3.2.2 A posteriori control strategy

The illustrations presented in 3.2.1 only give a sense of how *simple* controls f affect the posterior distribution. Practically, it is of greater significance to explore how the posterior distribution is affected by a control that has been constructed using posterior information from a previous assimilation cycle. One assimilation cycle is done with the passive model equation,

$$\text{Fluid model: } \frac{\partial v}{\partial t} = 0, \quad t > 0, \quad \text{and} \quad (3.13a)$$

$$\text{Uncontrolled drifter: } \frac{dz}{dt} = v(z), \quad 0 < t < t_{K/2}, \quad (3.13b)$$

with observations,

$$y^1 = \mathcal{G}^1(v) + \eta, \quad \eta \sim \mathcal{N}(0, \sigma^2 I_K) \quad (3.14)$$

The posterior distribution, $\mathbb{P}(v, |y^1)$, is sampled and a control is constructed using the information about the distribution. Finally, the forced drifter model is utilised,

$$\text{Fluid model: } \frac{\partial v}{\partial t} = 0, \quad t > 0, \text{ and} \quad (3.15a)$$

$$\text{Controlled drifter: } \frac{dz}{dt} = v(z) + f(z), \quad t_{K/2} \leq t \leq t_K, \quad (3.15b)$$

with observations,

$$y^2 = \mathcal{G}^2(v) + \eta, \quad \eta \sim \mathcal{N}(0, \sigma^2 I_K) \quad (3.16)$$

and a final posterior distribution, $\mathbb{P}(v|y^1, y^2)$ is obtained for comparison. This methodology fits in neatly with the set-up presented in (3.6), where we separate the two different observation regimes.

Gradient of the posterior mean

In section 3.2.1 we concluded that crossing a transport boundary and entering a new flow regime has the desirable effect of reducing the posterior variance. Crossing into new flow regimes with a stationary flow as in (3.6) can be translated to travelling transversely against the streamlines of the underlying flow. For the recirculation regime located in the bottom-left area (3.1), particles in the fluid will move in a clockwise fashion. The gradient of the stream function will therefore point in towards the fixed point of (3.6b) at $z = (\frac{1}{4}, \frac{1}{6})$. The negative gradient of the stream function points towards the fixed point at $z = (\frac{3}{4}, \frac{1}{3})$. Therefore, to escape the recirculation regime we choose,

$$f(z) = -\zeta \nabla_z (\mathbb{E}(\psi|y^1)), \quad (3.17)$$

for the forced drifter model (3.15b), where ψ is the stream function of the flow v . The rationale behind this choice is that, if the posterior mean stream function is a good estimator of the flow, the drifter will be forced transversely with the stream lines and escape the recirculation regime and allow us to make observations in a new flow regime in (3.15b).

Figure 3.29 depicts the variance of the horizontal component as the strength of the control, ζ in (3.17), is varied. Note that we do not see the same behaviour as we do for the two naïve controls chosen in section 3.2.1. We see a large band of values

of ζ for which the posterior variance oscillates, leading to a lack of information gain in the knowledge of the flow. From about $\zeta = 0.5$ to $\zeta = 0.55$, we see a structurally significant reduction in posterior variance where we have a sustained gain in information about the underlying flow field. This is attributed to a drifter path that explores an ‘interesting’ part of the flow where a lot of information can be obtained from observations. To explore the geometric correspondence between the variance reduction for $\zeta = 0.5$ to $\zeta = 0.55$, we show figure 3.28. This figure presents the true path of the drifter for $\zeta = 0.3, \dots, 0.55$. The light pink path corresponds to a value of $\zeta = 0.3$ and the purple path corresponds to $\zeta = 0.55$. Notice that as ζ increases, the true path forms a kink and forms a trajectory close to the zero of the flow at $(x, y) = (\frac{7}{12}, \frac{1}{2})$. Just as we have seen in section 3.2.1, we observe a transient period in the posterior variance until we utilise a control for which the true path explores new aspects of the flow compared with other ‘nearby’ controls. Interestingly, also note that we observe this reduction in variance despite the true path navigating near a zero of the flow, where we also satisfy the fact the size of the control is large in comparison to the flow (see figure 3.30). In this case, a logical conclusion here would be that the information gain from observing near an interesting flow structure heavily outweighs the information loss in polluting the observations with such a control. The cost of polluting the observed data can be seen by computing the most structurally significant reduction in the posterior variance and comparing this with figure 3.26, for example. By ‘most structurally significant’ we loosely mean the most dramatic reduction that leads to the most benefit in knowledge of the underlying flow. In this example, this occurs between $\zeta = 0.52$ and $\zeta = 0.55$, where it is approximately 3×10^{-5} . In the case of the bi-directional control, where the relative size of the flow *increases* for the values of ζ that give a reduction in variance (see figures 3.26 and 3.27), it occurs between $\zeta = 0.25$ and $\zeta = 0.625$ where it is approximately 1.5×10^{-4} . This is about an order of magnitude bigger, crystallising the tradeoff between polluting the observed data versus exploring ‘interesting’ parts of the flow. If the posterior mean is a good estimator of the underlying flow, utilising a control of this nature is beneficial if the drifters navigates close to a hyperbolic fixed point of the passive drifter model equation.

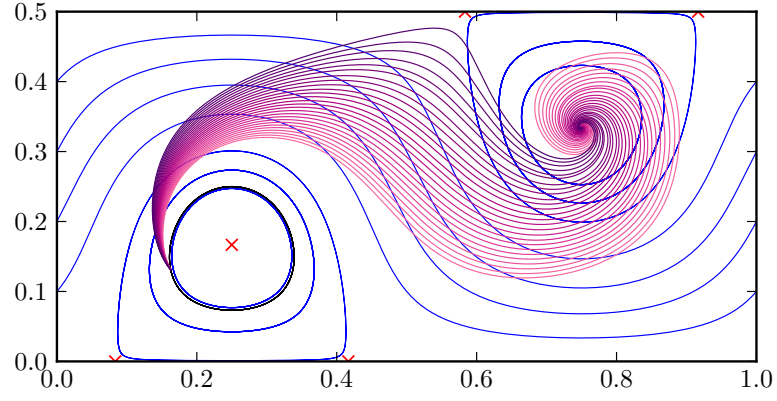


Figure 3.28: True drifter paths for a range of values of ζ . The light pink path corresponds to $\zeta = 0.3$ and the dark purple path corresponds to $\zeta = 0.55$

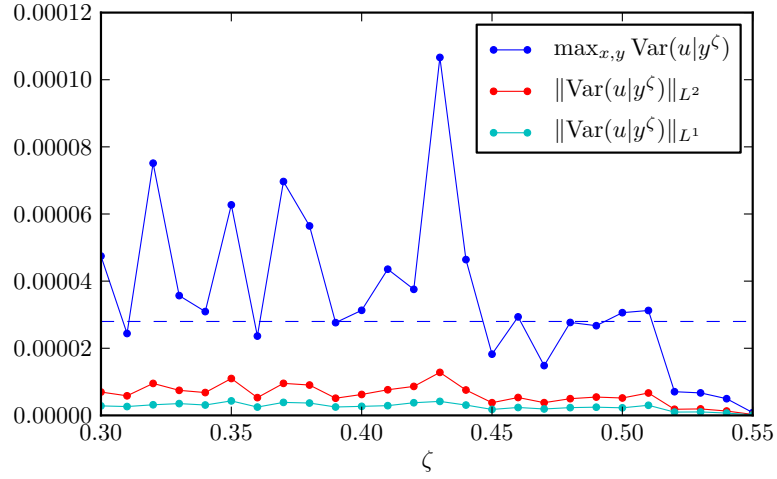


Figure 3.29: $\text{Var}(u|y^\zeta)$ for varying ζ

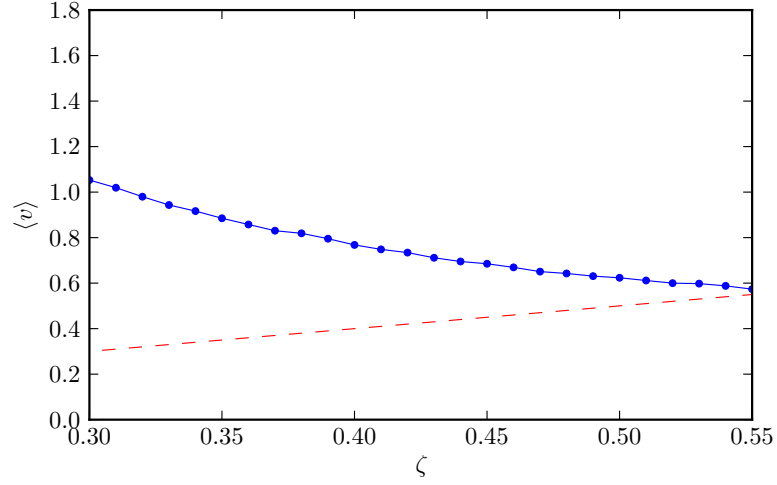


Figure 3.30: Mean flow magnitude (blue line) as a function of control magnitude (red dotted line)

3.3 Periodic time-dependent disturbances

We use the same formulation for the stream function as in (3.1) but we assign $\varepsilon = 0.3$ and ψ_1 to be periodic in time. We choose,

$$\psi_1(x, y, t) = \sin(2\pi x - \pi t) \sin(4\pi y). \quad (3.18)$$

This is the form of an oscillatory disturbance as presented in Samelson & Wiggins [2006]. The disturbance presented in (3.18) is periodic with period $T = 2$. Put this into the form of (3.6) to obtain,

$$\text{Fluid: } \frac{\partial v}{\partial t} = 0.3 \begin{pmatrix} 4\pi^2 \cos(2\pi x - \pi t) \cos(4\pi y) \\ 2\pi^2 \sin(2\pi x - \pi t) \sin(4\pi y) \end{pmatrix}, \quad t > 0, \text{ and} \quad (3.19a)$$

$$\text{Uncontrolled: } \frac{dz}{dt} = v(z), \quad 0 < t < t_{K/2}, \quad (3.19b)$$

$$\text{Controlled: } \frac{dz}{dt} = v(z) + f(z), \quad t_{K/2} \leq t \leq t_K. \quad (3.19c)$$

The model is fixed in v and so can be integrated directly,

$$\begin{aligned}
v(x, y, t) &= u(x) + \int_0^t \frac{\partial v}{\partial s} ds \\
&= u(x) + 0.3 \times 2\pi \int_0^t \left(\frac{2 \cos(2\pi x - \pi s) \cos(4\pi y)}{\sin(2\pi x - \pi s) \sin(4\pi y)} \right) ds \\
&= u(x) + 0.3 \times 2\pi \left(\frac{-2 \sin(2\pi x - \pi t) \cos(4\pi y) + 2 \sin(2\pi x) \cos(4\pi y)}{\cos(2\pi x - \pi t) \sin(4\pi y) - \cos(2\pi x) \sin(4\pi y)} \right)
\end{aligned}$$

Here u denotes the initial condition. To generate the data we use,

$$u(x, y) = \begin{pmatrix} \pi - 2\pi \sin(2\pi x) \cos(2\pi y) - 0.3 \times 4\pi \sin(2\pi x) \cos(4\pi y) \\ 2\pi \cos(2\pi x) \sin(2\pi y) + 0.3 \times 2\pi \cos(2\pi x) \sin(4\pi y) \end{pmatrix},$$

which gives a closed expression for the solution of the model at time t for (3.19),

$$v(x, y, t) = \begin{pmatrix} \pi - 2\pi \sin(2\pi x) \cos(2\pi y) - 0.3 \times 4\pi \sin(2\pi x - \pi t) \cos(4\pi y) \\ 2\pi \cos(2\pi x) \sin(2\pi y) + 0.3 \times 2\pi \cos(2\pi x - \pi t) \sin(4\pi y) \end{pmatrix}. \quad (3.20)$$

3.3.1 Naïve control for time-dependent flow model

Just as in the naïve control for the time-independent case in section 3.2.1, we make observations of positions of drifters z at various times, as in (3.9). We also use the same numerical parameters for all the experiments. The burn-in period and sampling period will both be set to 10^6 iterations. We make $K = 2 \times 10^4$ observations, one at every time-step, with error $\sigma^2 = 10^{-6}$. The prior variance is set to $\mu = 10^4$. The control f will likewise have the same form as the previous section,

$$f(z) = \begin{pmatrix} \zeta \\ 0 \end{pmatrix}.$$

The posterior distribution takes a similar form, as well. It is parameterised by the strength of the control, ζ , $\mathbb{P}(u|y^\zeta)$. Here the first half of the elements of y^ζ are observations of the passive motion of the ocean drifter. Observations of the controlled positions are in the second half of y^ζ . We wish to understand the effects on the posterior distribution as ζ is varied. We choose the same values for ζ as in the time independent case. To assure the reader that the implemented random walk Metropolis-Hastings method converges to a posterior mode for this control, we present figure 3.31 which shows the negative log-likelihood as a function of sample

number. Notice the sampler behaviour does not exhibit any noticeable trend and hovers around a mode corresponding to approximately $\Phi(u_k) = 2 \times 10^4$. The effect

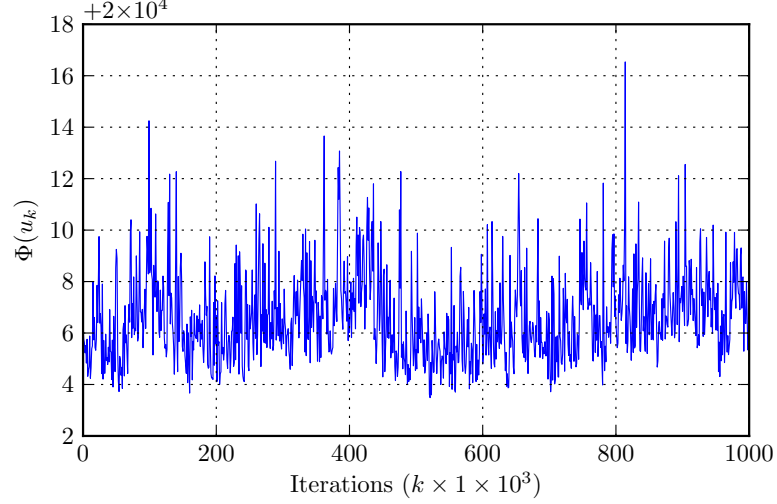


Figure 3.31: Trace plot of the negative log-likelihood for $\zeta = 1$ and $\sigma^2 = 10^{-6}$

on the posterior variance of the horizontal component of the initial condition u is shown in 3.32.

Note that the red hatched region corresponds to values of ζ that are too small to push the glider out of the eddy *in the unperturbed case* $\varepsilon = 0$. The green hatched region corresponds to values of ζ for which the glider leaves the eddy, this is also in the unperturbed case. The transition value is not computed exactly and lies somewhere in $\zeta \in (1.5, 1.75)$, this corresponds to the blue hatched region. Experiments were done for $\zeta = 0, 0.25, 0.5, \dots, 2.75, 3$. In the case $\varepsilon = 0$, the value $\zeta = 1.75$ was the first experiment in which we observed the glider leaving the recirculation regime. The blue line shows the maximum value of the variance over the domain $[0, 1] \times [0, 0.5]$. The red line and cyan line show the L^2 norm and L^1 norm, respectively. There are some notable points to make. Firstly, in the red region (where the glider does not leave the eddy in the unperturbed case) we see a sizeable reduction of posterior variance in the max norm as the the green hatched region is approached. To establish a connection in uncertainty quantification between the time independent and time-periodic case is of great scientific interest. Many more dynamically consistent models are time-dependent. Our results here indicate that underlying time-stationary features can be observed through use of a control in a time-dependent flow. Use of a control in a real-world setting to tease out underlying large-scale and time-independent flow features would help improve our understand-

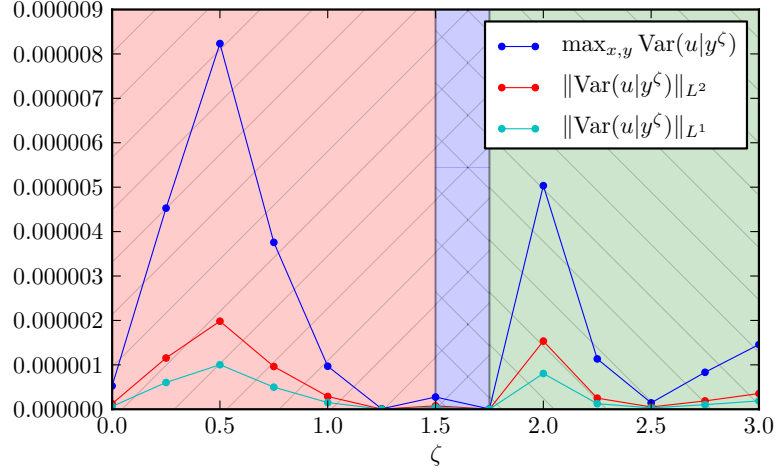


Figure 3.32: $\text{Var}(u|y^\zeta)$ for varying ζ in a time-periodic flow. The red hatched region corresponds to values of ζ for which the control magnitude is not large enough to push the drifter out of the eddy for the *unperturbed* system ($\varepsilon = 0$). The green hatched region corresponds to values of ζ for which the drifter is pushed out of the eddy. The blue region indicates the values of ζ for which there are no experiments conducted. This region contains the transition case.

ing of physical processes inherent in nature. This is exactly the connection we make here, but on a more rudimentary playing field. We observe the effects of an underlying time-independent eddy when utilising a control in a perturbed time-dependent flow. Note that as ζ increases and progresses further into the green hatched region, the posterior variance repeats the increasing/decreasing structure induced by the eddy that we observed in the red hatched region. These new effects introduced into the green region are purely from the time-dependent nature of the moving eddy. The reason for their presence is much the same as in the time-independent case; observations trapped within an eddy regime. For further assurance, compare figures 3.33 and 3.34. Notice that for the case $\zeta = 2.0$, the variance is higher and this is attributed to the extra loop the true trajectory takes within the eddy. This can be seen just north-east of the stationary point depicted by the red cross. The true initial condition of the unperturbed PDE is also plotted for reference.

We have learned more about the flow around the truth by forcing the glider into the meandering jet flow regime. The benefits of such a control occur at exactly the same place as in the time-independent case; as the drifter leaves the eddy in the unperturbed flow. However, extra care is required when the flow is time-dependent and the eddy moves. One cannot simply apply the same control techniques as is evidenced by the extra bump in variance in the green hatched region. Of particular

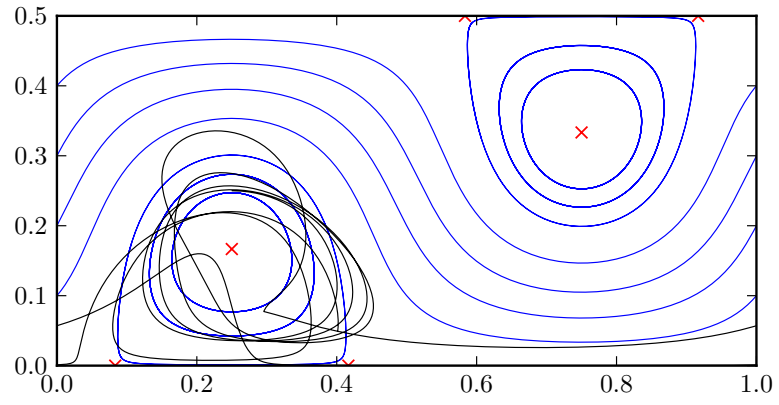


Figure 3.33: True drifter path with latitudinal forcing magnitude and $\zeta = 1.75$

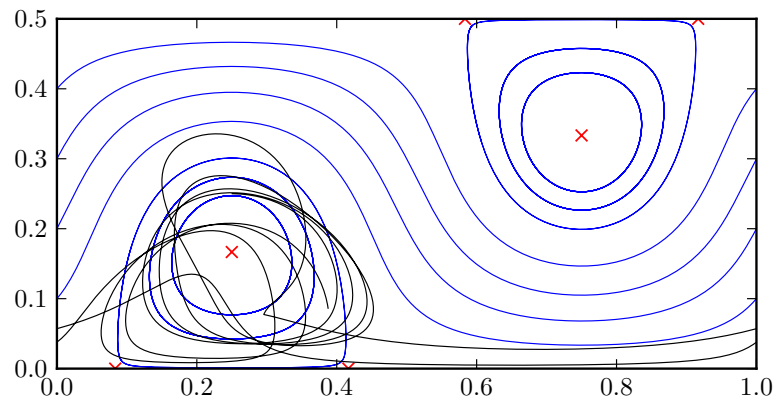


Figure 3.34: True drifter path with latitudinal forcing magnitude and $\zeta = 2.0$

use would be extra eddy-tracking information to construct an a posteriori control to keep the variance small.

Forcing in both coordinate directions

We now explore the effect on the posterior variance of the horizontal component of the velocity field when the forcing function is,

$$f(z) = \begin{pmatrix} \zeta \\ \zeta \end{pmatrix}.$$

Note that the data, y , is still parameterised by ζ . We will investigate what happens as ζ is varied, just as we have seen in figure 3.32. An analogue for figure 3.32 for the new forcing function is shown in figure 3.35. We see similar behaviour for the variance of the posterior distribution. Again, the red hatched region corresponds to values of ζ that are not big enough to push the drifter out of the recirculation regime in the *unperturbed* case. Just as in figure 3.32, we see the unperturbed eddy affecting the variance of the posterior distribution on the flow in the classic ‘bump’ fashion. We observe a reduction in posterior variance as ζ approaches a value large enough to push the glider out of the eddy regime (in the case $\varepsilon = 0$), depicted by the green hatched region. The blue hatched region denotes the range of values of ζ for which it is unknown whether the glider leaves the eddy or not, due to lack of experiments. In the green hatched region, the time-dependent flow effects take over and push the variance up. Again, a connection of uncertainty quantification is made between the time-independent case and the case where the flow is perturbed by a time-periodic disturbance, this connection lies entirely within the red hatched region.

3.3.2 Time-dependent a posteriori control

The illustrations presented in section 3.3.1 give a sense of how simple controls f affect the posterior distribution and links structure in information gained to large-scale structures in the time-independent case 3.2. Practically, it is of greater significance to explore how the posterior distribution is affected by a control that has been constructed using posterior information from a previous assimilation cycle. One

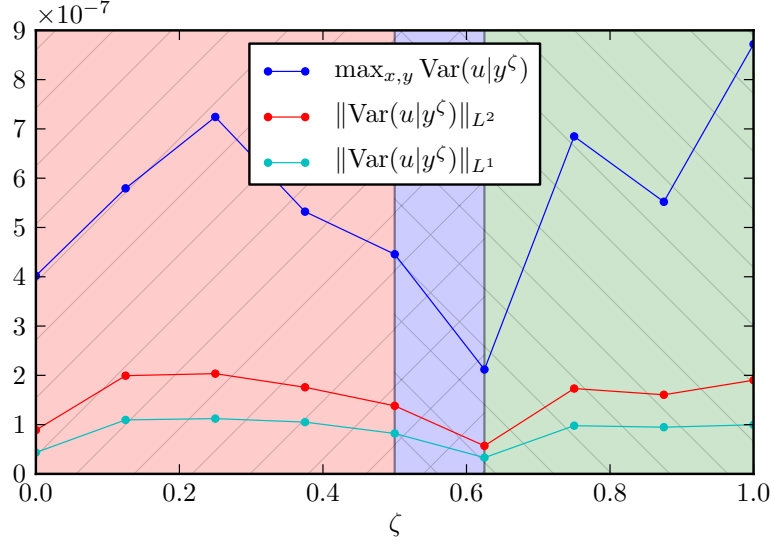


Figure 3.35: $\text{Var}(u|y^\zeta)$ for varying ζ in a time-periodic flow

assimilation cycle is done with the passive model equation,

$$\text{Fluid model: } \frac{\partial v}{\partial t} = 0.3 \begin{pmatrix} 4\pi^2 \cos(2\pi x - \pi t) \cos(4\pi y) \\ 2\pi^2 \sin(2\pi x - \pi t) \sin(4\pi y) \end{pmatrix}, \quad t > 0, \quad (3.21a)$$

$$\text{Uncontrolled drifter: } \frac{dz}{dt} = v(z), \quad 0 < t < t_{K/2}, \quad (3.21b)$$

with observations,

$$y^1 = \mathcal{G}^1(v) + \eta, \quad \eta \sim \mathcal{N}(0, \sigma^2 I_K) \quad (3.22)$$

The posterior distribution, $\mathbb{P}(v, |y^1)$, is sampled and a control is constructed using the information from these samples. Finally, the forced glider model is utilised,

$$\text{Fluid model: } \frac{\partial v}{\partial t} = 0.3 \begin{pmatrix} 4\pi^2 \cos(2\pi x - \pi t) \cos(4\pi y) \\ 2\pi^2 \sin(2\pi x - \pi t) \sin(4\pi y) \end{pmatrix}, \quad t > 0, \quad (3.23a)$$

$$\text{Controlled drifter: } \frac{dz}{dt} = v(z, t) + f(z, t), \quad t_{K/2} \leq t \leq t_K, \quad (3.23b)$$

with observations,

$$y^2 = \mathcal{G}^2(v) + \eta, \quad \eta \sim \mathcal{N}(0, \sigma^2 I_K) \quad (3.24)$$

and a final posterior distribution, $\mathbb{P}(v|y^1, y^2)$ is obtained for comparison. This methodology fits in neatly with the set-up presented in (3.19), where we separate the two different observation regimes.

Gradient of the posterior mean

In section 3.3.1 we conclude that crossing a transport boundary and entering a new flow regime in the unperturbed case has the desirable effect of reducing the posterior variance. Moreover, on comparison with section 3.2.1 we also conclude that this reduction is robust to time-periodic perturbations of the flow equations. Crossing into new flow regimes with a stationary flow as in (3.6) can be translated to travelling transversely against the streamlines of the underlying flow. For the recirculation regime located in the bottom-left area of the unperturbed flow (figure 3.1), particles in the fluid will move in an clockwise fashion. The gradient of the stream function will therefore point in towards the fixed point of (3.6b) at $z = (\frac{1}{4}, \frac{1}{6})$. The negative gradient of the stream function points towards the fixed point at $z = (\frac{3}{4}, \frac{1}{3})$. In the time-periodic case the negative gradient will not direct the drifter to the same fixed points as it did in section 3.2.2, but to fixed point of the full, perturbed flow (3.19). Therefore, to escape the recirculation regime, we choose,

$$f(z, t) = -\zeta \nabla_z (\mathbb{E}(\psi(t)|y^1)), \quad (3.25)$$

for the forced glider model (3.23b), where $\psi(t)$ is the stream function of $v(t)$. The rationale behind this choice is that, if the posterior mean stream function is a good estimator of the stream function for the true flow, the glider will be forced transversely with the stream lines and escape the recirculation regime and allow us to make observations in a new flow regime in (3.23b). We seek to compare this control in the time-periodic setting to the same case of control as in section 3.2.2, perhaps in search for a robustness result.

Operationally, construction of the control (3.25) involves computing an expectation of the flow at time t . The samples obtained from the posterior distribution after the passive model cycle has finished are of the initial condition of the model (3.21b). Generally, given a linear model,

$$\begin{aligned} \frac{\partial v}{\partial t} &= \mathcal{L}v, \\ v(0) &= v_0, \end{aligned}$$

we write its solution at time t as,

$$v(t) = e^{\mathcal{L}t} v_0. \quad (3.26)$$

Formula (3.26) can also be used to propagate posterior samples through a model,

$$\mathbb{E}(v(t)|y^1) = e^{\mathcal{L}t}\mathbb{E}(v_0|y^1).$$

This is exactly the formula we will use to compute posterior samples with increasing time t for use in constructing the control f in (3.23b). Figure 3.36 provides an analog of figure 3.29 associated with the unperturbed case. This depicts the variance of the horizontal component as the strength of the control, ζ , is varied. The first thing to

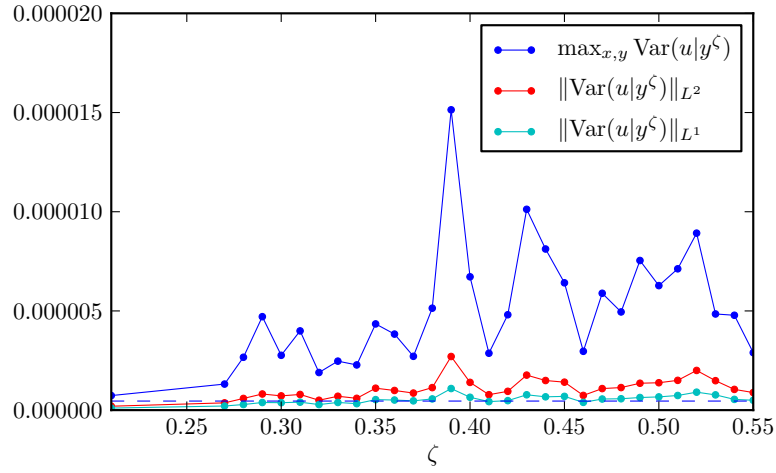


Figure 3.36: $\text{Var}(u|y^\zeta)$ for varying ζ

note is that we do not see the same behaviour as we do for the two naïve controls chosen in section 3.2.1. Nor do we see similar structures when compared with figure 3.29. For each value of ζ , it is the case that the true path navigates to the time-dependent eddy surrounding the zero of the flow at the point $(x, y) = (\frac{3}{4}, \frac{1}{3})$. The second thing to note is that for all of these values of control magnitude, the smaller values tend to do better than the larger ones. This can be explained by figures 3.37–3.39 which show the true drifter paths for the values $\zeta = 0.21, 0.27, 0.39$. In each of these plots, the corresponding posterior variance gets larger with ζ . The unperturbed initial condition is shown in each plot for reference. Notice that, just as in section 3.2.2, the variance is lower in the cases $\zeta = 0.21$ and $\zeta = 0.27$ because the true path is navigating towards one of the hyperbolic fixed points of the eddy. A novel connection is established between the behaviour of these two controls in both the time-independent case and the time-periodic case.

The last thing to note in this case is the variance on the flow of controlled system does worse than the uncontrolled system where $\zeta = 0$. This can also be explained by

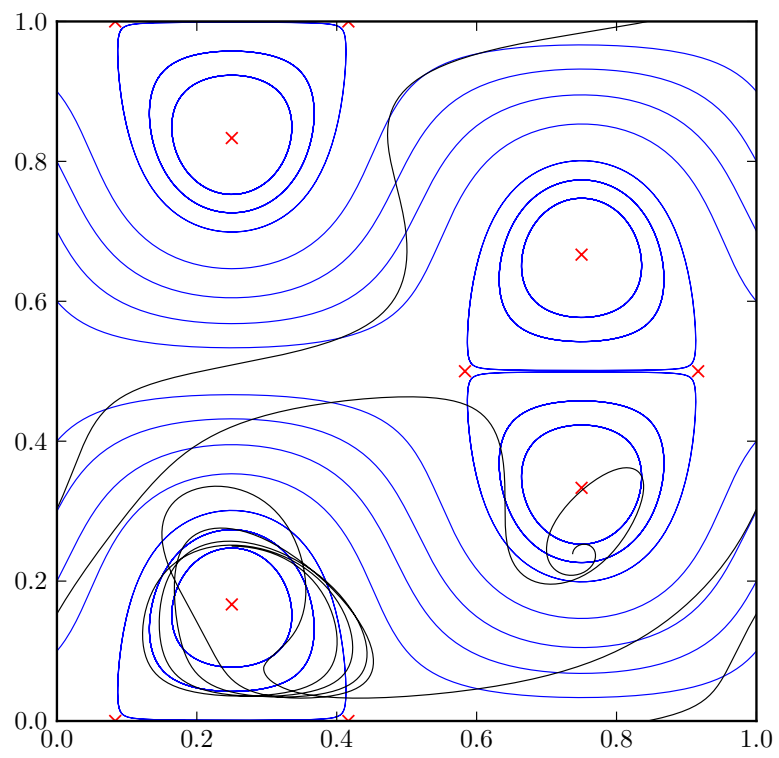


Figure 3.37: True drifter path for $\zeta = 0.21$

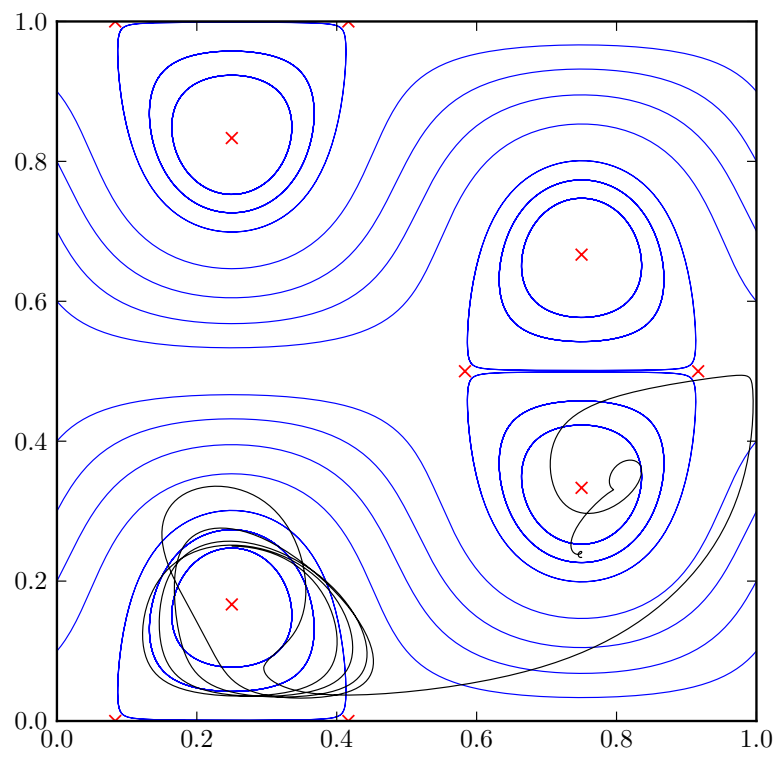


Figure 3.38: True drifter path for $\zeta = 0.27$

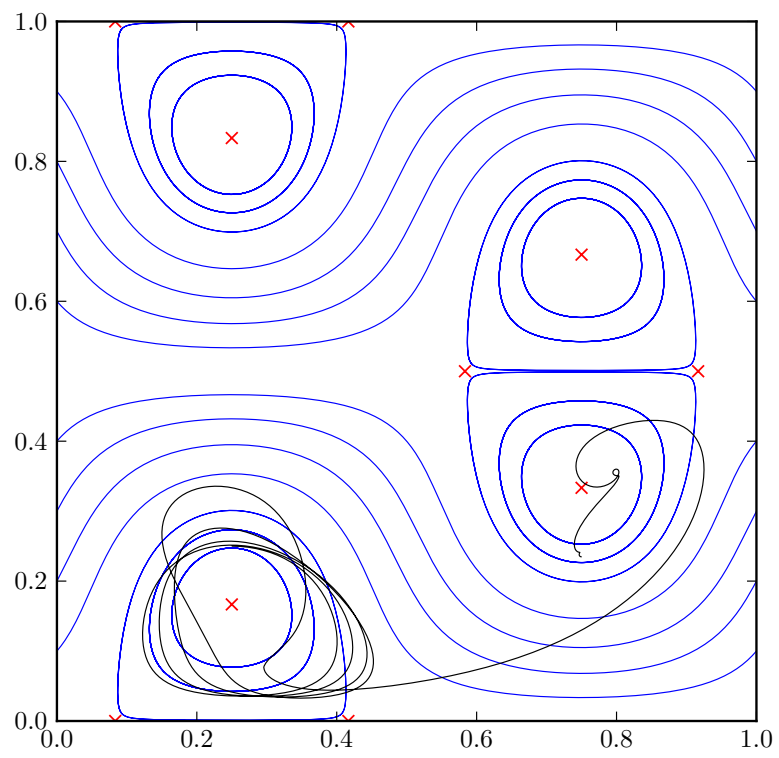


Figure 3.39: True drifter path for $\zeta = 0.39$

looking at the true drifter path (figure 3.40). Observe that in the uncontrolled case, the passive path does a more effective job of navigating towards fixed point of the eddy it starts in, thereby leading to a smaller variance than compared with the cases $\zeta > 0$. One should not conclude that exploring just one eddy is better than exploring both, it is that one should explore parts of the flow where the information gained in flow structure outweighs the loss through pollution of the observations. Parts of the flow where this occurs is around hyperbolic fixed points. These ‘pin down’ possible initial conditions to the model, leading to a decreased posterior variance.

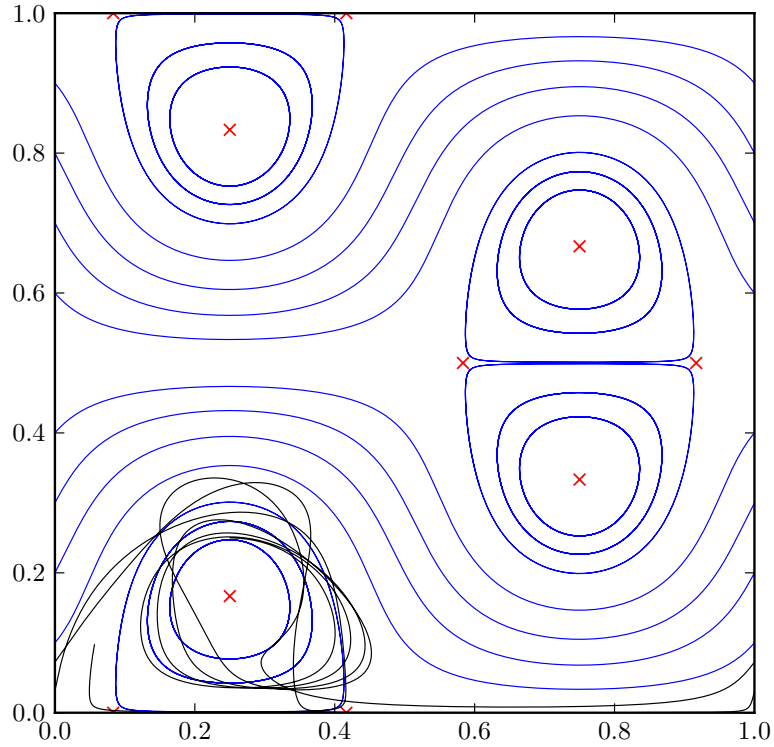


Figure 3.40: True drifter path for $\zeta = 0$

3.4 Conclusions

In this chapter we have measured the performance of two naïve control methods, and one a posteriori control method. We have done so by observing their influence on the posterior variance in the mean flow direction. Section 3.2 does this for a time-independent testbed ocean model; a kinematic travelling wave. Each control is designed to push ocean drifters into uncharted flow regimes. The three cases of control we employ here are a purely latitudinal control; a control of equal magnitude

in both the x and y directions; and the gradient of the posterior mean constructed using a posteriori information from a previous assimilation cycle. We show a sizeable reduction of the posterior variance in the mean flow direction for these three cases of control. We also see that on comparing the posterior variance for the latitudinal and bi-directional controls, similar structures arise when viewed as a function of control magnitude, which dictates when the drifter leaves the eddy and is the main influence on the posterior information. In the case of the a posteriori control, the drifter leaves the eddy for all the values of control magnitude we have chosen. Here we observe the variance reduction occurring when the true drifter path approaches a hyperbolic fixed point on the transport barrier of the eddy in the upper-right of the domain. This is evidence that oceanic transport barriers heavily influence posterior information and sets up a novel geometric correspondence between the flow structure and the posterior variance.

Section 3.3 adds a time-periodic disturbance to the testbed time-independent ocean model explored in the first component. We apply exactly the same control methods as in section 3.2. In the third case, the a posteriori control we construct in section 3.3.2 is time-dependent. For the first two cases, controls that are purely latitudinal and controls that are bi-directional, we show *robustness* of posterior variance as a function of the perturbation parameter. When the control magnitude is such that the drifter leaves the eddy in the *unperturbed* flow, we see reduction in the posterior variance on the initial condition for the time-periodic flow. When employing a time-dependent a posteriori control, we see no overall net gain in posterior variance over the uncontrolled case. For our particular flow and drifter initial condition, it is the case that the uncontrolled drifter path explores a hyperbolic fixed point of an eddy in the time-dependent flow more effectively than the controlled path. This reiterates the efficacy of control strategies and their influence on the path along which observations are made.

There are a number of ways in which this work could be generalised in order to obtain a deeper understanding of the effects controlled ocean drifters have on flow uncertainty. For example, the study of non-periodic model dynamics; (ii) the use of information from the posterior *variance*; (iii) more elaborate control strategies (see chapter 4). Many other generalisations are also possible. Non-periodic models are more dynamically consistent with regards to their approximation of larger ocean models. We have seen the application of posterior knowledge in the construction of a control, though only through use of the mean. The variance of the underlying flow could be used in a similar fashion, perhaps to control ocean drifters towards

an area of large variance. This could have a similar affect on the posterior distribution as the method of controlling a drifter into a new, unexplored flow regime. Moreover, controls could be constructed to better reflect reality. Ocean gliders have a limited amount of battery power. Utilising this knowledge in designing a mission plan to optimise a glider's lifespan certainly has its practical applications. Controls that minimise the pollution of the observed data is also desirable. Throughout this chapter, we have only used information from one previous assimilation cycle. Constructing and executing a posteriori control strategies is a paradigm well suited to that of a Kalman or particle filter; updating the control every time an analysis step is performed. This is left for future discussion.

Chapter 4

Data assimilation for optimally controlled testbed ocean drifters

4.1 Overview

In this chapter we partially extend the work set out in chapter 3. Chapter 3 concerned the application of cheap-to-compute controls to a testbed kinematic travelling wave model. The effect of each control on the associated posterior distribution on the underlying flow was analysed for a geometric correspondence between flow structure and posterior variance. Pushing the drifter out of an eddy yielded a net gain in information on the flow. Instead, there could be more to gain by choosing a specific point in the domain where the drifter should end up. Moreover, minimisation of the effort needed to reach such a terminal point is seen as a more challenging but realistically practical goal. For example, to see a reduction in posterior variance, one possibility would be to control an ocean drifter to a local maximum of the posterior variance from a previous assimilation cycle. This allows for observations to be collected in a part of the flow we are uncertain about. An approach of this type cannot be executed by use of simple cheap-to-compute controls as in chapter 3. As soon as the drifter reaches the relevant part of the domain, the flow would instantly push it away. This chapter aims to pose minimum-cost control strategies within the Bayesian framework for data assimilation as a basis for more complicated uncertainty quantification.

Section 4.2 introduces the theoretical nature of optimal control problems on a high level. Heavily inspired by Bryson Jr. & Ho [1975], we derive the Hamilton-Jacobi-

Bellman (HJB) equation for an optimal feedback control with a general cost function. Hamilton-Jacobi-Bellman equations, though useful, are often difficult to solve directly. They involve a global pre-determined grid of points on which the optimal cost-to-go function is computed.

Section 4.3 sees the application of the theory to a specific problem framed in an oceanographic context. Here we use a specific cost function, that of minimising the time to reach a terminal point in the domain. This is a practically inspired cost function in light of the results presented in chapter 3. Choosing the terminal point to be in a new flow regime and getting there in minimum time allows for the collection of observations to happen sooner. The practical implications of such an objective are very clear. We go further by applying an algorithm due to Rhoads [Rhoads *et al.*, 2010] to obtain necessary conditions for an extremum of the HJB equations; the Euler-Lagrange equations. From the point of view of implementation, the Euler-Lagrange equations relax the requirement that the cost-to-go surface be computed over the whole domain. A local method such as this gels well with the framework of data assimilation applied to problems in the ocean and the heavily localised observations thereof. This should be a stepping stone for executing more complicated control strategies than those explored in chapter 3.

Section 4.4 presents the necessary workflow to execute the minimum time control algorithm within a Bayesian framework. Implications of such a complicated control construction are illustrated here. More specifically, Markov chain Monte Carlo methods are a state-of-the-art method to solve problems in data assimilation, but typically require a large number of samples to adequately compute posterior moments. We show that this state-of-the-art method does not exhibit an avenue for which clever control methods can be computed cheaply. For each sample, ocean drifter positions are integrated over the, potentially multivalued, cost-to-go surface. We explain two approaches to making this cheaper: reducing the number of draws from the posterior distribution; and computing less trajectories of the Euler-Lagrange equations. This exposes a trade-off between sampling error and control error.

4.2 Derivation of control theory

In this section we provide a very general introduction to optimal control theory from within the context of dynamic programming. This introduction is largely adapted

from Bryson Jr. & Ho [1975], the theory in which we later apply to a testbed scenario involving ocean drifters.

Given some differential equation,

$$\dot{x} = g(x, h, t), \quad (4.1)$$

and some cost functional,

$$J = \psi(x(t_f), t_f) + \int_{t_0}^{t_f} L(x, h, t) dt, \quad (4.2)$$

with terminal boundary conditions $\psi(x(t_f), t_f)$, we wish to find the function $h(t)$ that minimises J . Here, we call $x \in \mathbb{R}^n$ the *state* variable and $h \in \mathbb{R}^m$ is called the *control* variable. Finding such a control that minimises (4.2) subject to (4.1) is called a *constrained optimisation* problem. The first term in (4.2) is some terminal cost at the final time t_f . Constrained optimisation problems like this can be rewritten to combine both the constrained and the cost functional into a new, *constrained*, cost functional,

$$\bar{J} = \psi(x(t_f), t_f) + \int_{t_0}^{t_f} \left(L(x, h, t) + \lambda^\top (g(x, h, t) - \dot{x}) \right) dt. \quad (4.3)$$

Notice that minimisers of (4.2) subject to (4.1), should they exist, are also minimisers of (4.3). The variables λ are often called *Lagrange multipliers*. We define the Hamiltonian by,

$$H(x, h, \lambda, t) = L(x, h, t) + \lambda^\top g(x, h, t), \quad (4.4)$$

and integrate the last term of (4.3) by parts to obtain,

$$\bar{J} = \psi(x(t_f), t_f) + \lambda^\top(t_0)x(t_0) - \lambda^\top(t_f)x(t_f) + \int_{t_0}^{t_f} H(x, h, \lambda, t) + \dot{\lambda}^\top x dt. \quad (4.5)$$

This is the full constrained cost functional written in terms of a Hamiltonian. The question now is how to compute an optimal control from this functional. We shall step through a classical technique in the calculus of variations. To do this, look at the first variation of (4.5) with respect to h , i.e, let h be the optimal control and

compute the following,

$$\begin{aligned}
\left(\frac{d\bar{J}(h + \varepsilon\eta)}{d\varepsilon}\right)_{\varepsilon=0} &= \left(\frac{d\psi(x(t_f), t_f)}{d\varepsilon}\right)_{\varepsilon=0} + \left(\frac{d\lambda^\top(t_0)x(t_0)}{d\varepsilon}\right)_{\varepsilon=0} - \left(\frac{d\lambda^\top(t_f)x(t_f)}{d\varepsilon}\right)_{\varepsilon=0} \\
&\quad + \int_{t_0}^{t_f} \left(\frac{dH(x, h, \lambda, t)}{d\varepsilon}\right)_{\varepsilon=0} + \left(\frac{d\dot{\lambda}^\top x}{d\varepsilon}\right)_{\varepsilon=0} dt \\
&= \frac{\partial\psi(x(t_f), t_f)}{\partial x(t_f)} \left(\frac{dx(t_f)}{d\varepsilon}\right)_{\varepsilon=0} \\
&\quad + \lambda^\top(t_0) \left(\frac{dx(t_0)}{d\varepsilon}\right)_{\varepsilon=0} - \lambda^\top(t_f) \left(\frac{dx(t_f)}{d\varepsilon}\right)_{\varepsilon=0} \\
&\quad + \int_{t_0}^{t_f} \left(\frac{\partial H}{\partial x} + \dot{\lambda}^\top\right) \left(\frac{dx}{d\varepsilon}\right)_{\varepsilon=0} + \frac{\partial H}{\partial h} \left(\frac{d(h + \varepsilon\eta)}{d\varepsilon}\right)_{\varepsilon=0} dt \\
&= \left(\frac{\partial\psi}{\partial x} - \lambda^\top(t_f)\right) \left(\frac{dx(t_f)}{d\varepsilon}\right)_{\varepsilon=0} \\
&\quad + \lambda^\top(t_0) \left(\frac{dx(t_0)}{d\varepsilon}\right)_{\varepsilon=0} \\
&\quad + \int_{t_0}^{t_f} \left(\frac{\partial H}{\partial x} + \dot{\lambda}^\top\right) \left(\frac{dx}{d\varepsilon}\right)_{\varepsilon=0} + \frac{\partial H}{\partial h} \eta dt.
\end{aligned}$$

To avoid calculating $\left(\frac{dx}{d\varepsilon}\right)_{\varepsilon=0}$, we make the following choice,

$$\dot{\lambda}^\top = -\frac{\partial H}{\partial x}, \quad (4.6)$$

$$\lambda^\top(t_f) = \frac{\partial\psi(x(t_f), t_f)}{\partial x(t_f)}, \quad (4.7)$$

which gives,

$$\left(\frac{d\bar{J}(h + \varepsilon\eta)}{d\varepsilon}\right)_{\varepsilon=0} = \lambda^\top(t_0) \left(\frac{dx(t_0)}{d\varepsilon}\right)_{\varepsilon=0} + \int_{t_0}^{t_f} \frac{\partial H}{\partial h} \eta dt.$$

Therefore, $\lambda^\top(t_0)$ is exactly the gradient of \bar{J} with respect to the initial condition. And, for an extremum of \bar{J} , this quantity is zero which implies,

$$\frac{\partial H}{\partial h} = 0. \quad (4.8)$$

Equations (4.6)–(4.8) form the *Euler-Lagrange* equations. So to compute an optimal

control h , we solve the following system of equations,

$$\begin{aligned}\dot{x} &= g(x, h, t), \\ \dot{\lambda}^\top &= -\frac{\partial H}{\partial x},\end{aligned}$$

where,

$$\begin{aligned}x(t_0) &\text{ is given,} \\ \lambda^\top(t_f) &= \frac{\partial \psi(x(t_f), t_f)}{\partial x(t_f)},\end{aligned}$$

and h satisfies $\frac{\partial H}{\partial h} = 0$. Since the boundary conditions above are mixed (some given at time t_0 and some given at time t_f), this is a two-point boundary value problem. We now continue, with the aim of applying some of this theory to optimal *feedback* control.

4.2.1 Optimal feedback control

In this subsection we extend some of the theory presented in section 4.1 to optimal feedback control. This is also adapted from Bryson Jr. & Ho [1975]. The functional of interest is the *optimal* cost-to-go. This is similar to (4.5) but posed from an arbitrary initial point (x, t) at an optimal control. It is defined as,

$$J^o(x, t) = \min_h \left\{ \psi(x(t_f), t_f) + \int_t^{t_f} L(x, h, t) dt \right\} \quad (4.9)$$

with boundary condition,

$$J^o(x, t) = \psi(x(t_f), t_f) \quad \text{on the hypersurface } \Psi(x, t) = 0.$$

Assume J^o exists and is twice continuously differentiable. Suppose the system starts at (x, t) and proceeds for a time Δt using a non-optimal control. The state will evolve to the point,

$$(x + g(x, h, t)\Delta t, t + \Delta t).$$

Suppose it then continues with optimal control for times larger than Δt . To first order, the optimal cost-to-go function is given by,

$$J^o(x + g(x, h, t)\Delta t, t + \Delta t) + L(x, h, t)\Delta t =: J^1(x, t). \quad (4.10)$$

Since the optimal control was used in the interval $(t, t + \Delta t)$, we have,

$$J^o(x, t) \leq J^1(x, t)$$

Choosing the optimal control yields a minimum of the right-hand side of the previous equation and we obtain,

$$J^o(x, t) = \min_h \{J^o(x + g(x, h, t)\Delta t, t + \Delta t) + L(x, h, t)\Delta t\} \quad (4.11)$$

Using the smoothness of J^o , Taylor expand (4.11) around (x, t) to get,

$$\begin{aligned} J^o(x, t) &= \min_h \left\{ J^o(x, t) + \frac{\partial J^o}{\partial x} g(x, h, t)\Delta t + \frac{\partial J^o}{\partial t} \Delta t + L(x, h, t)\Delta t + \mathcal{O}(\Delta t^2) \right\} \\ \Rightarrow 0 &= \min_h \left\{ \frac{\partial J^o}{\partial x} g(x, h, t)\Delta t + \frac{\partial J^o}{\partial t} \Delta t + L(x, h, t)\Delta t + \mathcal{O}(\Delta t^2) \right\} \\ \Rightarrow 0 &= \min_h \left\{ \frac{\partial J^o}{\partial x} g(x, h, t) + \frac{\partial J^o}{\partial t} + L(x, h, t) + \mathcal{O}(\Delta t) \right\} \end{aligned}$$

Let $\Delta t \rightarrow 0$ to obtain,

$$\frac{\partial J^o}{\partial t} = - \min_h \left\{ L(x, h, t) + \frac{\partial J^o}{\partial x} g(x, h, t) \right\} \quad (4.12)$$

This is the Hamilton-Jacobi-Bellman equation. On comparing the right-hand side of (4.12) with the right-hand side of (4.4) it is easy to see that,

$$\lambda^\top = \frac{\partial J^o}{\partial x}, \quad (4.13)$$

along an optimal trajectory. This is justified by differentiating (4.13) and noting its equivalence with (4.6) at an extremum. Furthermore, from (4.12) and (4.4) it follows that,

$$\frac{\partial J^o}{\partial t} = - \min_h \{H(x, h, \lambda, t)\}. \quad (4.14)$$

The work of Crandall & Lions [1983] deals with the notion of a *viscosity solution*. Viscosity solutions are required to prove well-posedness results concerning the HJB equations.

4.3 Specific use-case

Just as in section 3.2.2, the controlled ocean drifter set up is as follows,

$$\text{Fluid model: } \frac{\partial v}{\partial t} = 0, \quad t > 0, \text{ and} \quad (4.15a)$$

$$\text{Uncontrolled drifter: } \frac{dz}{dt} = v(z), \quad 0 < t < t_{K/2}, \quad (4.15b)$$

$$\text{Controlled drifter: } \frac{dz}{dt} = v(z) + f(z), \quad t_{K/2} \leq t \leq t_K. \quad (4.15c)$$

Instead of choosing f as in chapter 3, we will choose it to be such that the ocean drifter is forced to go to some ‘terminal point’ in the domain in minimum time (see section 4.2). As briefly mentioned in section 4.1, the terminal point could be chosen by assimilating the observations made from the passive system and looking at, for example, local maxima of the resulting posterior variance. This a posteriori control would force an ocean drifter to an area of the domain where there is large uncertainty in the flow and allowing observations to be made there. This strategy is beyond the scope of chapter 3. Furthermore, the ocean glider will have some practical maximum speed relative to the underlying flow, s , so that $|f(z)| \leq s$. The natural choice of the admissible control set is therefore $h(t) \in \{h \in \mathbb{R}^2 \mid |h| \leq s\}$. The cost function (4.3) for the minimum time case is $L(x, h, t) = 1$ and the optimal cost-to-go function (4.9) is the time-to-go, T . As a reminder, the model, or constraint equations (4.1), for controlled ocean drifters are $\dot{x} = v(x, t) + h(t)$, so $g(x, h, t) = v(x, t) + h(t)$. Substituting into (4.12) gives,

$$\frac{\partial T}{\partial t} = - \min_h \{ [v(x, t) + h(t)] \cdot \nabla T(x, t) + 1 \}, \quad (4.16)$$

where $\lambda = \nabla T$ along the optimal path. This is exactly as in Rhoads *et al.* [2010]. Substituting (4.16) into (4.6) we obtain,

$$\begin{aligned} \frac{d\lambda^\top}{dt} &= - \frac{\partial}{\partial x} ([v(x, t) + h(t)] \cdot \lambda + 1) \\ &= - \lambda^\top \frac{\partial v}{\partial x} \\ \Rightarrow \frac{d\lambda}{dt} &= - \left[\frac{\partial v}{\partial x} \right]^\top \lambda. \end{aligned} \quad (4.17)$$

From Pontryagin’s Maximum Principle [Boltyanskii *et al.*, 1956; Pontryagin *et al.*, 1962] we also have that the optimal control, $h(t)$, minimises the Hamiltonian,

$$H(x, h, \lambda, t),$$

$$\operatorname{argmin}_h ([v(x, t) + h(t)] \cdot \lambda + 1) = \operatorname{argmin}_h h(t) \cdot \lambda$$

The value of $h(t)$ that minimises this expression is exactly the vector that is parallel to λ but points in the opposite direction with maximum size. That is,

$$h(t) = -s \frac{\lambda(t)}{\|\lambda(t)\|}. \quad (4.18)$$

Furthermore, as in Rhoads *et al.* [2010], the free final time end condition gives us,

$$\begin{aligned} 0 &= H(x(t_f), h(t_f), \lambda(t_f), t_f) \\ &= [v(x(t_f), t_f) + h(t_f)] \cdot \lambda(t_f) + 1 \\ &= v(x(t_f), t_f) \cdot \lambda(t_f) - s \frac{\lambda(t_f)}{\|\lambda(t_f)\|} \cdot \lambda(t_f) + 1 \\ &= v(x(t_f), t_f) \cdot \lambda(t_f) - s \|\lambda(t_f)\| + 1 \\ &= \|\lambda(t_f)\| v(x(t_f), t_f) \cdot n_f - s \|\lambda(t_f)\| + 1, \quad \text{where } n_f := \frac{\lambda(t_f)}{\|\lambda(t_f)\|}, \\ &= v(x(t_f), t_f) \cdot n_f - s + \frac{1}{\|\lambda(t_f)\|} \\ \therefore \|\lambda(t_f)\| &= \frac{1}{s - v(x(t_f), t_f) \cdot n_f} \\ \therefore \lambda(t_f) &= \frac{1}{s - v(x(t_f), t_f) \cdot n_f} n_f. \end{aligned} \quad (4.19)$$

Notice that the free final time end condition gives us a terminal boundary condition, $\lambda(t_f)$, parameterised by $n_f \in S^1$. Coupled with the drifter model equations, equation (4.18) and equation (4.17), we have the following Euler-Lagrange equations,

$$\begin{pmatrix} \dot{x} \\ \dot{\lambda} \end{pmatrix} = \begin{pmatrix} v(x, t) - s \frac{\lambda(t)}{\|\lambda(t)\|} \\ - \left[\frac{\partial v}{\partial x} \right]^\top \lambda \end{pmatrix} \quad (4.20)$$

For each $n_f \in S^1$, solving (4.20) with initial condition $x(t_0)$ yields a *locally* optimal trajectory $x(t)$ terminating at the point $x(t_f)$ such that $t_f - t_0$ is minimised locally. We now give a slightly simpler version of the more sophisticated algorithm in Rhoads *et al.* [2010] that solves this system to obtain an approximate global optimal control.

The result of algorithm 2 is a set of points $\{x(0), x(\Delta t), \dots, x(T_{\max})\}$ for each n_f . This is a total of $T_{\max} \times \Delta t \times n$ points in the plane. To compute $\nabla_x T(x, t)$ among

Data: number of extremals to compute, n
Data: maximum allotted time to wait, T_{\max}
Data: terminal point, x_f
Data: set of angles θ_f such that $n_f = (\cos(\theta_f), \sin(\theta_f))$
1 Initialise $\{\theta_f\} = \{2\pi k/n \mid k = 1, \dots, n\}$;
2 Let $x(T_{\max}) = x_f$;
3 Compute $\lambda(T_{\max})$ from (4.19);
4 **for** $j \leftarrow 1$ **to** n **do**
5 Integrate (4.20) backwards in time;
6 **end**

Algorithm 2: Computing locally extremal trajectories

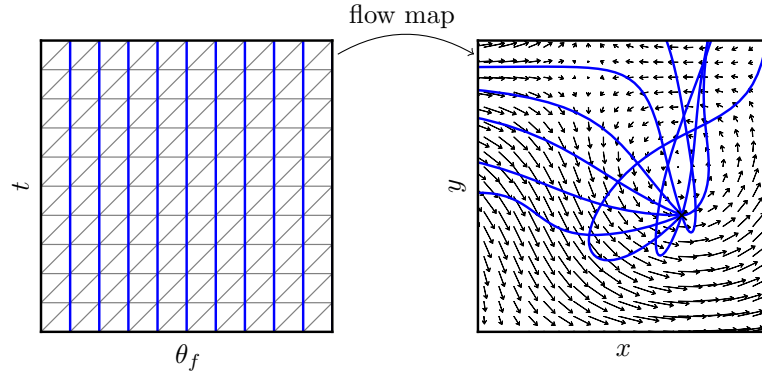


Figure 4.1: Euler Lagrange ODE flow map from extremals in the triangulation to locally optimal trajectories in physical space

these points and find a global extremal we could apply a triangulation algorithm on them all and use linear interpolation on each triangle. This is not a good approach since each trajectory is only locally optimal. Points close together could belong on different extremals and have very different time-to-go values. Instead, we triangulate the (θ_f, t) domain and use the flow map defined by (4.20) to map to the physical (x, y) domain. Figure 4.1 illustrates this mapping. Once the extremals have been computed, one can think of the ‘time’ direction as coming out of the page in the (x, y) physical space plot of figure 4.1. To obtain the globally optimal time-to-go *surface*, look at this from below. The surface may be multivalued, with extremals lying on top of each other at different time-to-go values, but looking from below they will be obscured by the globally optimal extremals. Figure 4.2 shows an example time-to-go surface. Black indicates smaller time-to-go. Notice the time-to-go at the terminal point $(\frac{3}{4}, \frac{1}{3})$ is zero. The green line depicts the minimum time trajectory from the recirculation regime of the vector field in the lower-left of the domain to the

elliptic fixed point of the recirculation regime in the upper-right. Use of a control like this yields paths that can cross transport boundaries that would have usually been impassable if only the passive drifter equation had been integrated. Minimum time controls of this sort are important within the context of data assimilation on the ocean surface since transporting a drifter to a terminal point as quickly as possible means that one can start taking observations of a new flow regime sooner. This may not be the best choice of control. For example, ocean gliders have a limited amount of battery power and so a minimum power usage control may be of more interest regarding ocean data assimilation when utilising gliders. Exactly the same method as the one presented above can be applied in this case, all that is needed is a modification to the cost function $L(x, h, t)$. The issue of convergence arises when

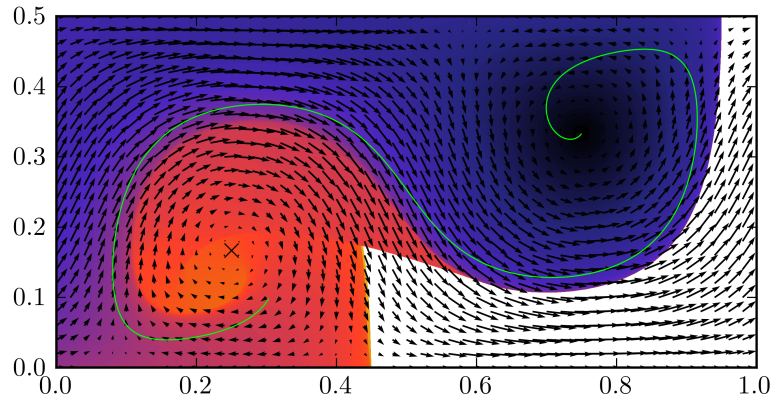


Figure 4.2: Stream function of initial condition (3.1)

one asks how many extremals should be computed. A standard convergence study is employed as a sanity check to ensure the associated algorithm, and numerical code, are functioning as expected. Take some maximal number of extremals, N , and compute an optimal path p down the gradient of the time-to-go surface to some fixed terminal point, x_f . Compute a set of K optimal paths $\{p_i\}_{i=1}^K$ obtained by using N_i extremals, where $N_i < N \ \forall i \in \{1, \dots, K\}$. Figure 4.3 illustrates $\|p_i - p\|_{L^2}^2$ for $i = 1, \dots, K$. The number of extremals was doubled for each i , up to a maximum of 2^{13} . A vertex of the triangulation in (θ_f, t) space was placed every 100 timesteps except for the last two cases, where a vertex was placed every 10 timesteps. Observe solid convergence to the most finely resolved path, at a rate of roughly second order.

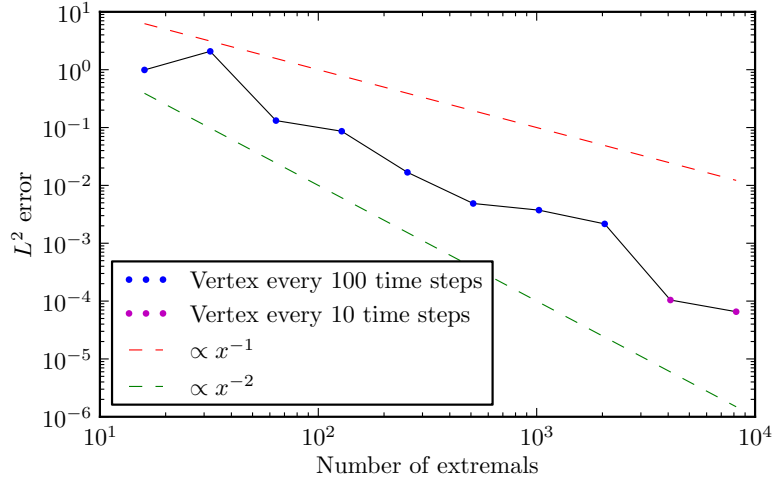


Figure 4.3: Convergence of the optimal path in a stationary double gyre

4.4 Application to data assimilation

There are two main approaches to minimising the time-to-go cost functional. The first is to solve the Hamilton-Jacobi-Bellman equation (4.16) – a PDE – directly. The second is to compute the Euler-Lagrange equations (4.20) associated to the cost functional, resulting in a set of first-order ODEs. Either of these approaches yields the same result, their difference is mainly in the relative ease of finding a solution numerically. Notice that, solving the Hamilton-Jacobi-Bellman equation (4.16) directly requires gridding the domain to sufficiently small resolution and applying a method suited to two-point boundary value PDEs. This is inferior compared to solving the Euler-Lagrange family of ODEs for the following reasons. Solving a family of ODEs is embarrassingly parallel since none of the local extremals depends on any other extremal. Furthermore, solving the HJB equation directly may not necessarily be the best approach to resolving discontinuities in the surface. The Euler-Lagrange method is superior in this case as viewing the time-to-go surface from below produces much more pronounced discontinuities. Lastly, and more pertinently, the Euler-Lagrange family of ODEs is more suited to data assimilation. In this context, we can solve the Euler-Lagrange ODEs whilst only using partial knowledge of the flow. Solving the full HJB equation requires knowledge of the flow over the whole domain and this information may not be available. If the flow is only known in some eddy (due to the existence of a transport boundary), then it is advantageous to use the Euler-Lagrange method on this local flow information with the aim of finding a minimum time control to push an ocean drifter out. The choice of a

minimum time control, as opposed to minimisation of an alternative cost function, is sensible in the situation where one needs to observe new information as quickly as possible. In this case, computing a minimum time control with terminal point in some new flow regime is a natural strategy to adopt. In chapter 3 we show that pushing an ocean drifter into a new flow regime yields an advantageous reduction in posterior variance on the underlying flow field. Furthermore, we show that, in the case of the a posteriori control case outlined in section 3.2.2, pushing too hard does not present any benefit. Thus, the setting in which drifters have a maximum speed corresponding to exactly to the control magnitude that yields maximum reduction in posterior variance is practical and novel. Extending this setting by applying a minimum time control whose maximum magnitude is the same as the maximum glider speed yields the best case scenario of balance between flow exploration and observations in a new flow regime. Motivation for application of minimum time control strategies should now be clear.

Table 4.1 categorises the time it takes to a) compute some number of locally extremal trajectories; and b) compute the globally optimal time-to-go path. As the number of extremals increases the time-to-go surface becomes more finely resolved and as a consequence, the more optimal the final globally extremal trajectory will be. Therefore, it is important to resolve the surface finely.

Table 4.1: Timings for extremal trajectories in a stationary kinematic wave

Extremals	Time for extremals (s)	Time for path (s)
2^5	0.32	6.61
2^6	0.65	16.02
2^7	1.31	33.81
2^8	2.60	42.59
2^9	5.17	151.30
2^{10}	10.34	250.57
2^{11}	20.66	717.84

Recall the two-part setup of constructing a posteriori control strategies outlined in section 3, equation (3.6). The idea is to initially solve the passive glider model to collect some data and assimilate it into a flow model. The next step is to sample the posterior distribution and utilise the information of the samples to construct a control. The purpose of using a minimum time control in this setting would be so that one can observe a new flow regime as soon as possible. Given the results presented in chapter 3, we conclude that observing an uncharted flow regime as quickly as possible whilst using a conservative control magnitude is the method that best

suits minimising the posterior variance on the flow. Integrating down the time-to-go surface requires computing the gradient of the surface on the triangle which contains the current drifter location. Since the surface can be multivalued in many cases, the unique coordinates of the drifter in physical space may correspond to several points in the (θ_f, t) -plane. Figure 4.4 shows some locally extremal trajectories (solutions of (4.20)) for the case of the same flow as in 4.2. Multivaluedness occurs around the area of the eddy in the lower-left part of the (x, y) -plane—notice the tangle of Euler-Lagrange trajectories sitting on top of each other. As a result it is not clear how to apply commonly used search algorithms on the (θ_f, t) triangulation to locate the drifter. One now relies on a linear search to find the triangle containing drifter on the best current extremal. As the number of local extremals increases, the number of triangles and amount of computation time increases. One forward integration of the model is required per posterior sample. Therefore, to obtain the same number of samples as we saw in section 3.2 ($\mathcal{O}(10^6)$), we require $\mathcal{O}(10^6)$ model integrations. Integrating the forward model on a candidate time-independent flow sample will require, in the best case of 32 extremals and a very poorly resolved time-to-go surface, roughly 7 seconds of computation time. As a result, drawing 10^6 samples will require just over 2.5 months of computational effort. To reduce the computational time, it is necessary to reduce the number of samples drawn from the posterior distribution which increases sampling error in computing the posterior moments. For a time-dependent flow, the situation is worse. The method of Rhoads *et al.* [2010] must be executed for every time step of the model. In this case, drawing just 10^3 samples from the posterior could take approximately two years. Note the computational cost is mostly in the integration of the optimal path, not the local extremals. The calculation of trajectories of the Euler-Lagrange equations can be done concurrently since every extremal is independent of every other extremal. We have shown that, given the current state of the art in data assimilation techniques, applying the method of Rhoads *et al.* [2010] yields a method that is impractically slow. Parallelisation is imperative, and is trivial in the case of solving for the extremals, this is not the case for the computation of the optimal path.

4.5 Conclusions

Advanced methods from control theory hold out the potential to harness a posteriori information from a previous assimilation cycle more systematically. As a concrete example, to reduce uncertainty in the underlying flow it may be advantageous to

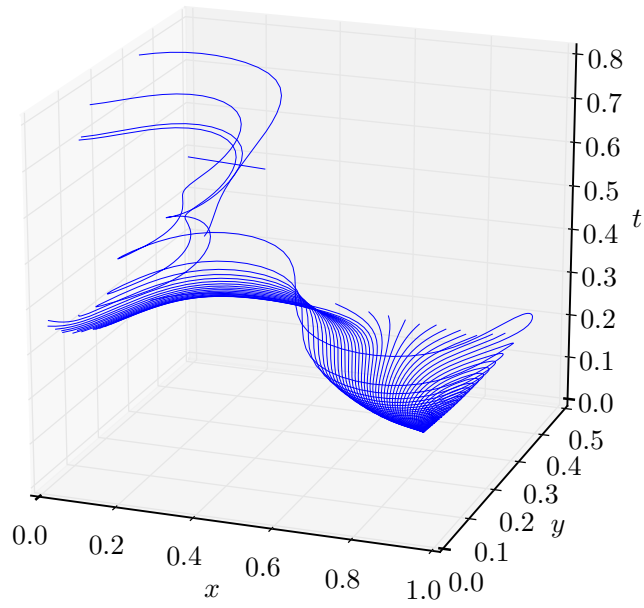


Figure 4.4: The x -direction corresponds to latitude; the y -direction corresponds to longitude. Each of the blue lines corresponds to a solution of the Euler-Lagrange equations (4.20) with a different initial condition for λ . The t -axis denotes the time-to-go to get to the terminal point in the centre of the eddy in the upper-right of the (x, y) -plane

push an ocean drifter towards a local maximum of the posterior variance. This allows observations to be collected in an area of the domain where the uncertainty is large. Utilising the strategy set out in chapter 3 by only influencing the magnitude of the control is not necessarily the most effective way to achieve this goal. Other bumps and valleys in the variance affect the drifter path, as does the underlying flow. More elegant control over the drifter’s position is one option towards better utilising a posteriori flow information. This chapter has set the basis for more complex control strategies than those presented in chapter 3.

We pose an optimal feedback control strategy that has more influence on the drifter’s position in a testbed oceanographic setting (section 4.3). By utilising the the method of Rhoads *et al.* [2010], a cheaper procedure for solving the Hamilton-Jacobi-Bellman equation, we show convergence of the optimal path as more Euler-Lagrange trajectories (figure 4.1) are computed (figure 4.3), giving confidence that our implementation of the algorithm is functioning correctly.

Section 4.4 presents the necessary workflow to execute a general control methodology within a Bayesian framework for data assimilation. This provides a concrete basis for which more elaborate control strategies can be engineered. Note that Markov chain Monte Carlo methods, a state-of-the-art method to solve problems in data assimilation, typically require a large number of samples to adequately compute posterior moments. Consequently, we show that this does not exhibit an avenue for which clever control methods can be computed *cheaply*. For each sample, ocean drifter positions are integrated over the, potentially multivalued, cost-to-go surface. We explain two approaches to making this cheaper; reducing the number of draws from the posterior distribution; or, computing less trajectories of the Euler-Lagrange equations. This exposes a trade-off between sampling error and control error.

Chapter 5

Discussion

On a high level, this thesis has addresses the influence of the mathematical model on the posterior distribution from two different settings. One involving assimilation of Eulerian data, the other Lagrangian data. In both of these problems, we employ Markov chain Monte Carlo methods to probe the associated posterior measure. This measure sits on a finite dimensional approximation of an infinite dimensional space. Each posterior sample requires a forward-run of the model. In our toy problems, this model is extremely cheap to compute and, as a consequence, it is possible to draw lots of posterior samples to saturate the state space. In practice, the model is nonlinear, very high dimensional and expensive to execute. Fully three dimensional turbulent Navier-Stokes equations are one such example, and are dynamically consistent with reality. Asking for a well resolved posterior measure in this context is ambitious enough to require more sophisticated approaches to sampling. Advances have been made to improve the computational cost in the context of Monte Carlo sampling, see Cui *et al.* [2011]; Giles [2008, 2006].

We study the Bayesian inverse problem for the linear advection equation in two dimensions. We make Eulerian observations of the underlying flow and, for the inverse problem on the initial condition, we characterise the mean of the posterior distribution as a function of model error. Here, the model error is in the advection velocity parameter. When the error is zero, the model is perfect and reconstruction of the true underlying flow is possible. Partial recovery of the true underlying flow is occurs when the error is rational, the denominator of which dictates the number of Fourier modes present in the reconstruction. An irrational error leads to retrieval only of the spatial mean of the flow. If the actual wave velocity is time-dependent and differs from the true wave velocity in a time-integrable fashion then the solution to

the inverse problem is in error by a constant phase shift, determined by the integral of the difference in wave velocities.

We also consider the inverse problem on both the initial condition and the wave velocity parameter, resulting in a non-Gaussian posterior joint distribution. We deduce the root cause of sampler metastability to be a poorly chosen MCMC seed for the wave velocity. To solve this issue, some kind of prior knowledge or expertise is required about the system. Alternatively, some heuristic pre-assimilation analysis with the data to obtain an estimate is advised. In our case, a novel least-squares method approach is employed using the advection model on the noisy data as input. This results in good convergence of the posterior measure on the wave speed to a Dirac in the large data limit.

Lastly, we provide numerical results when a non-smooth likelihood norm is utilised over the initial condition. This problem is non-Gaussian with a linear forward operator. The non-Gaussianity arises from assuming the log-likelihood grows only linearly in the tails; a doubly-exponential likelihood distribution of the data/model mismatch. We expose MCMC as a flexible tool, able to deal easily with non-Gaussian infinite dimensional inverse problems. With this modified likelihood formulation, larger steps in the state space can be taken leading to more efficient state space exploration.

Applying aspects of control theory to controlling an ocean drifter, we provide several methods to improve our knowledge of the underlying flow. We do this by placing an ocean drifter in an eddy and enforcing several control procedures, instructing the drifter to explore parts of the flow currently uncharted and obtaining fresh observations of new flow structures. The efficacy of each control we utilise is determined by its resulting effect on the variance of posterior distribution. A smaller variance is interpreted as a better understanding of the flow. We do this for a time-independent testbed ocean model; a kinematic travelling wave. We employ three cases of control: a purely latitudinal control; a control of equal magnitude in both coordinate directions; and the gradient of the posterior mean constructed using a posteriori information from a previous assimilation cycle. We show a sizeable reduction of the posterior variance for each case of control. Furthermore, on comparing the posterior variance for the latitudinal and bi-directional controls, similar structures arise when viewed as a function of control magnitude. In the case of the a posteriori control, we show a reduction in variance when the drifter approaches a hyperbolic fixed point on the transport barrier of an eddy. This is evidence that oceanic transport barriers heavily influence posterior information and sets up a novel geometric correspondence

between the flow structure and the posterior variance.

We add a time-periodic disturbance to the testbed time-independent ocean model above, applying exactly the same control methods. For the first two cases: purely latitudinal controls; and bi-directional controls, we show *robustness* of posterior variance as a function of the perturbation parameter. When the control magnitude is such that the drifter leaves the eddy in the *unperturbed* flow, we see reduction in the posterior variance on the initial condition for the time-periodic flow. For the time-dependent a posteriori control, we see no overall net gain in posterior variance over the uncontrolled case. For our particular flow and drifter initial condition, the uncontrolled drifter explores a hyperbolic fixed point of an eddy in the time-dependent flow more effectively than the controlled path. This exposes the efficacy of control strategies and their influence on the path along which observations are made.

We extend our own work regarding rudimentary controls to the case of an *optimal* control, allowing ourselves to choose a specific point in the domain where the drifter should end up. Here, optimality is taken in a minimum-time sense. Minimisation of the *effort* needed to reach a terminal point is seen as a more challenging but equally realistic goal. We pose minimum-time control strategies within a Bayesian framework for data assimilation as a basis for more involved uncertainty quantification.

We introduce the theoretical nature of optimal control problems on a high level. We derive the Hamilton-Jacobi-Bellman (HJB) equation for an optimal feedback control with a general cost function. Hamilton-Jacobi-Bellman equations, though useful, are often difficult to solve directly. They involve a global pre-determined grid of points on which the optimal cost-to-go function is computed. We then see the application of this theory to a specific problem framed in an oceanographic context. Here we use a specific cost function, that of minimising the time to reach a terminal point in the domain. This is a practically inspired cost function in light of the results presented in the third chapter. Choosing the terminal point to be in a new flow regime and getting there in minimum time allows for the collection of observations to happen sooner. The practical implications of such an objective are very clear. We go further by applying an algorithm due to Rhoads [Rhoads *et al.*, 2010] to obtain necessary conditions for an extremum of the HJB equations—the Euler-Lagrange equations. From the point of view of implementation, the Euler-Lagrange equations relax the requirement that the cost-to-go surface be computed over the whole domain. A local method such as this gels well with the framework of data assimilation applied to problems in the ocean and the heavily localised observations thereof. This should

be a stepping stone for executing more complicated control strategies than those explored in the third chapter.

Finally, we show that this state-of-the-art method does not exhibit an avenue for which clever control methods can be computed cheaply. For each sample, ocean drifter positions are integrated over the, potentially multivalued, cost-to-go surface. We explain two approaches to making this cheaper: reducing the number of draws from the posterior distribution; and computing less trajectories of the Euler-Lagrange equations. This exposes a trade-off between sampling error and control error.

The work in this thesis contains substantial advances in the field of data assimilation and, in particular, furthers the understanding of model error in inverse problems. Regarding the work done in characterising the posterior mean for an advection equation, it is the case that the work here will influence a class of operational data assimilation techniques called ‘model reduction’. Advection is a prominent component in large ocean and weather models. These models are too large to assimilate directly and a reduced model is utilised in their place. Error made in reduced models, as is shown in this work, can lead to an incorrect posterior distribution. This work also shows the benefits of coupling a flow model to a controlled but otherwise passive drifter. The rewards here, however, are not well understood when the control is, in some sense, optimal or when the underlying flow model is nonlinear and chaotic. Both of these scenarios are ubiquitous in natural operational settings and this is an avenue ripe for further exploration.

The numerical studies and discussion characterising the shape of the posterior mean in the advent of model error, along with the novel data-oriented Markov Chain seeding methodology is work published jointly with Lee and Stuart in Lee *et al.* [2011]. The efforts in chapter 3 are not yet published, but are a work in preparation with Jones in McDougall & Jones [2012].

Bibliography

- J L Anderson (2003). *A Local Least Squares Framework for Ensemble Filtering*. Monthly Weather Review **131** (4), 634–642.
- A Apte, M Hairer, A M Stuart and J Voss (2007). *Sampling the posterior: An approach to non-Gaussian data assimilation*. Physica D: Nonlinear Phenomena **230** (1-2), 50–64.
- A Apte, C K R T Jones and A M Stuart (2008)a. *A Bayesian approach to Lagrangian data assimilation*. Tellus A **60** (2), 336–347a.
- A Apte, C K R T Jones, A M Stuart and J Voss (2008)b. *Data assimilation: Mathematical and statistical perspectives*. International Journal for Numerical Methods in Fluids **56**, 1033–1046b.
- Y F Atchadé (2006). *An Adaptive Version for the Metropolis Adjusted Langevin Algorithm with a Truncated Drift*. Methodology and Computing in Applied Probability **8** (2), 235–254.
- Y F Atchadé and J S Rosenthal (2005). *On adaptive Markov chain Monte Carlo algorithms*. Bernoulli **11** (5), 815–828.
- S-J Baek, B R Hunt, E Kalnay, E Ott and I Szunyogh (2006). *Local ensemble Kalman filtering in the presence of model bias*. Tellus **58** (3), 293–306.
- R W Barbieri and P S Schopf (1982). *Oceanographic applications of the Kalman filter*.
- D M Barker (2004). *A Three-Dimensional Variational Data Assimilation System for MM5: Implementation and Initial Results*. American Meteorological Society **132**, 897–914.
- L Bengtsson (1975). *4-dimensional assimilation of meteorological observations*. World Meteorological Organization.

- A F Bennett (2002). *Inverse modeling of the ocean and atmosphere*. Cambridge University Press.
- P Bergthorsson and B R Doos (1955). *Numerical Weather Map Analysis*. Tellus **7** (3).
- A Beskos, G O Roberts and A M Stuart (2009). *Optimal scalings for local Metropolis-Hastings chains on nonproduct targets in high dimensions*. The Annals of Applied Probability **19** (3), 863–898.
- P Bickel, B Li and T Bengtsson (2008). *Sharp failure rates for the bootstrap particle filter in high dimensions*. IMS Collections: Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh **3**, 318–329.
- C H Bishop, B J Etherton and S J Majumdar (2001). *Adaptive Sampling with the Ensemble Transform Kalman Filter. Part I: Theoretical Aspects*. Monthly Weather Review **129** (3), 420–436.
- V I Bogachev (1998). *Gaussian Measures*. American Mathematical Society.
- V G Boltyanskii, R V Gamkrelidze and L S Pontryagin (1956). *Towards a theory of optimal processes*. Reports Acad. Sciences USSR **110** (1).
- G E P Box and Mervin E Muller (1958). *A note on the generation of random normal deviates*. The Annals of Mathematical Statistics **29** (2), 610—611.
- A E Bryson Jr. and Y-C Ho (1975). *Applied Optimal Control: Optimization, estimation, and control*. Hemisphere Publishing Corporation.
- N A Carlson (1973). *Fast triangular formulation of the square root filter*. AIAA **11**, 1259–1265.
- E F Carter (1989). *Assimilation of Lagrangian data into a numerical model*. Dynamics of Atmospheres and Oceans **13** (3-4), 335–348.
- J G Charney (1951). *Dynamic forecasting by numerical process*. Compendium of Meteorology, American Meteorological Society **6**, 470–482.
- S E Cohn (1997). *An Introduction to Estimation Theory*. Tech. rep.
- S L Cotter, M Dashti, J C Robinson and A M Stuart (2009). *Bayesian inverse problems for functions and applications to fluid mechanics*. Inverse Problems **25** (11), 115008.

- S L Cotter, M Dashti and A M Stuart (2010). *Approximation of Bayesian inverse problems for PDEs*. SIAM Journal of Numerical Analysis **48** (1), 322–345.
- S L Cotter, M Dashti and A M Stuart (2011). *Variational data assimilation using targetted random walks*. International Journal for Numerical Methods in Fluids **68**, 403–421.
- S L Cotter, G O Roberts, A M Stuart and D White (2012). *MCMC Methods for functions: Modifying old algorithms to make them faster*.
- P Courtier, J-N Thépaut and A Hollingsworth (1994). *A strategy for operational implementation of 4D-Var, using an incremental approach*. Quarterly Journal of the Royal Meteorological Society **120** (519), 1367–1387.
- M K Cowles and B P Carlin (1996). *Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review*. Journal of the American Statistical Association **91** (434), 883–904.
- M G Crandall and P L Lions (1983). *Viscosity Solutions of Hamilton-Jacobi Equations*. Transactions of the American Mathematical Society **277** (1), 1–42.
- G P Cressman (1959). *An operational objective analysis system*. Monthly Weather Review **87** (10), 594–5.
- T Cui, C Fox and M J O’Sullivan (2011). *Bayesian calibration of a large-scale geothermal reservoir model by a new adaptive delayed acceptance Metropolis Hastings algorithm*. Water Resources Research **47** (10).
- D Dee (1996). *Data assimilation in the presence of forecast bias*. Tech. Rep. December 1996.
- A Doucet, N de Freitas and N Gordon (2001). *Sequential Monte Carlo Methods In Practice*.
- G Evensen (1994). *Sequential data assimilation with a nonlinear quasi-geostrophic moel using Monte Carlo methods to forecast error statistics*. Journal of Geophysical Research **99**, 10143–10162.
- G Evensen (2006). *Data Assimilation: The Ensemble Kalman Filter*. Springer.
- M Galassi, J Davies, J Theiler, B Gough, G Jungman, P Alken, M Booth and F Rossi (2011). *GNU Scientific Library Reference Manual*, 3rd Edition.

- B Gershgorin, J Harlim and A J Majda (2010). *Test models for improving filtering with model errors through stochastic parameter estimation*. Journal of Computational Physics **229** (1), 1–31.
- J Geweke (2001). *Bayesian estimation of state-space models using the Metropolis-Hastings algorithm within Gibbs sampling*. Computational Statistics & Data Analysis **37** (2), 151–170.
- J Geweke and H Tanizaki (1999). *On Markov chain Monte Carlo methods for nonlinear and non-Gaussian state-space models*. Communications in Statistics-Simulation and Computation **28** (4), 867–894.
- B Gilchrist and G P Cressman (1954). *An experiment in objective analysis*. Tellus **6** (4).
- M B Giles (2006). *Improved multilevel Monte Carlo convergence using the Milstein scheme*. Monte Carlo and Quasi-Monte Carlo Methods, 343–358.
- M B Giles (2008). *Multilevel Monte Carlo Path Simulation*. Operations Research **56** (3), 607–617.
- W R Gilks, N G Best and K K C Tan (1995). *Adaptive Rejection Metropolis Sampling within Gibbs Sampling*. Applied Statistics **44** (4), 455—472.
- W K Hastings (1970). *Monte Carlo sampling methods using Markov chains and their applications*. Biometrika **57** (1), 97–109.
- R Herbei and I McKeague (2009). *Hybrid Samplers for Ill-Posed Inverse Problems*. Scandinavian Journal of Statistics **36** (4), 839—853.
- R Herbei, I W McKeague and K G Speer (2008). *Gyres and Jets: Inversion of Tracer Data for Ocean Circulation Structure*. Journal of Physical Oceanography **38** (6), 1180–1202.
- J E Hoke and R A Anthes (1976). *The initialization of numerical models by a dynamic-initialization technique (fluid flow models for wind forecasting)*. Monthly Weather Review **104**, 1551–1556.
- P L Houtekamer and H L Mitchell (1998). *Data Assimilation Using an Ensemble Kalman Filter Technique*. Monthly Weather Review **126**, 796–811.
- J D Hunter (2007). *Matplotlib: A 2D Graphics Environment*. Computing in Science and Engineering **9** (3), 90–95.

- A H Jazwinski (1970). *Stochastic processes and filtering theory*. Academic Pr.
- D Jones (2010). *Good Practice in (Pseudo) Random Number Generation for Bioinformatics Applications*.
- J Kaipio and E Somersalo (2007). *Statistical inverse problems: Discretization, model reduction and inverse crimes*. Journal of Computational and Applied Mathematics **198** (2), 493–504.
- J P Kaipio, V Kolehmainen, E Somersalo and M Vauhkonen (2000). *Statistical inversion and Monte Carlo sampling methods in electrical impedance tomography*. Inverse problems **16** (5), 1487.
- R E Kalman (1960). *A New Approach to Linear Filtering and Prediction Problems*. Journal of Basic Engineering **82** (Series D), 35–45.
- R E Kalman and R S Bucy (1961). *New results in linear filtering and prediction theory*. Journal of Basic Engineering **83**, 95–107.
- E Kalnay (2002). *Atmospheric modeling, data assimilation and predictability*. Cambridge University Press.
- S Kirkpatrick, C D Gelatt and M P Vecchi (1983). *Optimization by simulated annealing*. Science **220** (4598), 671–680.
- L Kuznetsov, K Ide and C K R T Jones (2003). *A Method for Assimilation of Lagrangian Data*. Monthly Weather Review **131**, 2247–2260.
- A S Lawless, S Gratton and N K Nichols (2005)a. *An investigation of incremental 4D-Var using non-tangent linear models*. Tech. rep.a.
- A S Lawless, S Gratton and N K Nichols (2005)b. *Approximate iterative methods for variational data assimilation*. Tech. rep.b.
- A S Lawless and N K Nichols (2006). *Inner-Loop Stopping Criteria for Incremental Four-Dimensional Variational*. Monthly Weather Review **134**, 3425–3435.
- F-X Le Dimet and O Talagrand (1986). *Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects*. Tellus A (38A), 97–110.
- W Lee, D McDougall and A M Stuart (2011). *Kalman filtering and smoothing for linear wave equations with model error*. Inverse Problems **27** (9), 095008.

- P J V Leeuwen (2010). *Nonlinear data assimilation in geosciences: an extremely efficient particle filter*. Quarterly Journal of the Royal Meteorological Society **136** (653), 1991–1999.
- J M Lewis and J C Derber (1985). *The use of adjoint equations to solve a variational adjustment problem with advective constraints*. Tellus A (37A), 309–322.
- A C Lorenc (1986). *Analysis methods for numerical weather prediction*. Quarterly Journal of the Royal Meteorological Society **112**, 1177–1194.
- A C Lorenc, S P Ballard, R S Bell, N B Ingleby, P L F Andrews, D M Barker, J R Bray, A M Clayton, T Dalby, D Li, T J Payne and F W Saunders (2000). *The Met. Office global three-dimensional variational data assimilation scheme*. Quarterly Journal of the Royal Meteorological Society **126** (570), 2991–3012.
- G Marsaglia (1996). *The Diehard Battery of Tests of Randomness*. Tech. rep.
- G Marsaglia and W W Tsang (2000). *The Ziggurat Method for Generating Random Variables*. Journal of Statistical Software, 1–7.
- M Matsumoto and T Nishimura (1998). *Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator*. ACM Transactions on Modeling and Computer Simulation **8** (1), 3–30.
- D McDougall and C K R T Jones (2012). *Decreasing flow uncertainty in Lagrangian data assimilation through drifter control*. In preparation.
- I W McKeague, G Nicholls, K Speer and R Herbei (2005). *Statistical inversion of South Atlantic circulation in an abyssal neutral density layer*. Journal of Marine Research **63** (4), 683–704.
- N Metropolis, A W Rosenbluth, M N Rosenbluth, A H Teller and E Teller (1953). *Equation of State Calculations by Fast Computing Machines*. The Journal of Chemical Physics **21** (6), 1087.
- A M Michalak (2003). *A method for enforcing parameter nonnegativity in Bayesian inverse problems with an application to contaminant source identification*. Water Resources Research **39** (2), 1–14.
- R N Miller (1986). *Toward the Application of the Kalman Filter to Regional Open Ocean Modeling*. Journal of Physical Oceanography **16**, 72–86.

- K Mosegaard and A Tarantola (1995). *Monte Carlo sampling of solutions to inverse problems*. Journal of Geophysical Research **100** (B7), 12431—12447.
- N K Nichols and A K Griffith (1996). *Accounting for model error in data assimilation using adjoint methods*. Computational Differentiation: Techniques, Applications and Tools, 195—204.
- N K Nichols and A K Griffith (2000). *Adjoint methods in data assimilation for estimating model error*. Flow, turbulence and combustion **65** (3), 469—488.
- E Ott, B R Hunt, I Szunyogh, A V Zimin, E J Kostelich, M Corazza, E Kalnay, D J Patil and J A Yorke (2004). *A local ensemble Kalman filter for atmospheric data assimilation*. Tellus A **56** (5), 415–428.
- D F Parrish and S E Cohn (1985). *A Kalman filter for a two-dimensional shallow-water model*. In: 7th Conference on Numerical Weather Prediction, Montreal, Canada. pp. 1–8.
- N S Pillai, A M Stuart and A H Thiéry (2012). *Optimal proposal design for random walk type Metropolis algorithms with Gaussian random field priors*, 1–42.
- L S Pontryagin, V G Boltyanskii, R V Gamkrelidze and E Mishchenko (1962). *The mathematical theory of optimal processes (International series of monographs in pure and applied mathematics)*. Interscience Publishers.
- B Rhoads, I Mezić and A Poje (2010). *Minimum Time Feedback Control of Autonomous Underwater Vehicles*. In: 49th IEEE Conference on Decision and Control. pp. 5828–5834.
- A A Robel, M Susan Lozier, S F Gary, G L Shillinger, H Bailey and S J Bograd (2011). *Projecting uncertainty onto marine megafauna trajectories*. Deep Sea Research Part I: Oceanographic Research Papers **58** (12), 915–921.
- G O Roberts (1997). *Weak convergence and optimal scaling of random walk Metropolis Algorithms*. Annals of Applied Probability **7** (1), 110–120.
- G O Roberts and J S Rosenthal (1998). *Optimal scaling of discrete approximations to Langevin diffusions*. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **60** (1), 255–268.
- G O Roberts and J S Rosenthal (2001). *Optimal scaling for various Metropolis-Hastings algorithms*. Statistical Science **16** (4), 351–367.

- G O Roberts and J S Rosenthal (2006). *Harris recurrence of Metropolis-within-Gibbs and trans-dimensional Markov chains*. The Annals of Applied Probability **16** (4), 2123–2139.
- A R Robinson and D B Haidvogel (1981). *Dynamical Forecast Experiments with a Barotropic Open Ocean Model*. Journal of Physical Oceanography **10**, 1928.
- A R Robinson and W G Leslie (1985). *Estimation and Prediction of Oceanic Eddy Fields*. Progress in Oceanography **14**, 485–510.
- D L Rudnick, R E Davis, C C Eriksen, D M Fratantoni and M J Perry (2004). *Underwater Gliders for Ocean Research*. Marine Technology Society Journal **38** (2), 73–84.
- R M Samelson and S Wiggins (2006). *Lagrangian transport in geophysical jets and waves: the dynamical systems approach*. Springer Verlag.
- P J Smith, S L Dance, M J Baines, N K Nichols and T R Scott (2009). *Variational data assimilation for parameter estimation: application to a simple morphodynamic model*. Ocean Dynamics **59** (5), 697–708.
- P J Smith, S L Dance and N K Nichols (2011). *A hybrid data assimilation scheme for model parameter estimation: Application to morphodynamic modelling*. Computers & Fluids **46** (1), 436–441.
- H W Sorenson (1960). *Kalman filtering: theory and application*. IEEE.
- A M Stuart (2010). *Inverse Problems: A Bayesian Perspective*. In: Acta Numerica. pp. 1–107.
- P Swerling (1958). *A proposed stagewise differential correction procedure for satellite tracking and prediction*. P-1292 Rand Corporation **8**.
- O Talagrand and P Courtier (1987). *Variational assimilation of meteorological observations with the adjoint vorticity equation. I: Theory*. Quarterly Journal of the Royal Meteorological Society **113**, 1311–1328.
- E Vanden-Eijnden and J Q Weare (2012). *Data assimilation in the low noise, accurate observation regime with application to the kuroshio current*. Arxiv preprint arXiv:1202.4952.
- V Černý (1985). *Thermodynamical approach to the Traveling Salesman problem: An efficient simulation algorithm*. Journal of Optimization Theory and Applications **45** (1), 41–51.

C Wikle and L Berliner (2007). *A Bayesian tutorial for data assimilation*. Physica D: Nonlinear Phenomena **230** (1-2), 1–16.