An exploration of the equivalent weights particle filter

M. Ades*, P. J. van Leeuwen

Department of Meteorology, University of Reading, Reading, UK

*Correspondence to: M. Ades, Department of Meteorology, University of Reading, Earley Gate, PO Box 243, Reading, RG6 6BB, UK. E-mail: m.ades@pgr.reading.ac.uk

Particle filters are fully non-linear data assimilation techniques that aim to represent the probability distribution of the model state given the observations (the posterior) by a number of particles. In high-dimensional geophysical applications the number of particles required by the sequential importance resampling (SIR) particle filter in order to capture the high probability region of the posterior, is too large to make them usable. However particle filters can be formulated using proposal densities, which gives greater freedom in how particles are sampled and allows for a much smaller number of particles. Here a particle filter is presented which uses the proposal density to ensure that all particles end up in the high probability region of the posterior probability density function. This gives rise to the possibility of non-linear data assimilation in large dimensional systems. The particle filter formulation is compared to the optimal proposal density particle filter and the implicit particle filter, both of which also utilise a proposal density. We show that when observations are available every time step, both schemes will be degenerate when the number of independent observations is large, unlike the new scheme. The sensitivity of the new scheme to its parameter values is explored theoretically and demonstrated using the Lorenz 1963 model.

Key Words: data assimilation; particle filtering; proposal densities

Received . . .

Citation:

1. Introduction

The most comprehensive solution to the data assimilation problem is to establish the full probability density function (pdf) of the state of a system. Theoretically it is understood how to represent the full pdf, called the posterior, if it is a standard density such as a Gaussian. However, this is rarely the case with the nonlinear model equations and observation operators of most geoscience applications. Instead, most present day data assimilation schemes follow one of two options. One is to make Gaussian and linear assumptions that lead to a Gaussian posterior, as is done in the ensemble Kalman filter (EnKF) and its variants (Evensen 1994; Burgers *et al.* 1998; Bishop *et al.* 2001; Anderson 2001; Whitaker and Hamill 2002). The other, which is used by schemes such as 4D-Var (Talagrand and Courtier 1987), is to search only for the maximum of the full posterior pdf assuming a Gaussian prior. Unfortunately the methods used in searching for the maximum can not guarantee that the global, rather than a local, maximum is reached. Furthermore, uncertainty estimates using the Hessian are not necessarily appropriate as the Hessian measures the local curvature. Hence its inverse is not necessarily a good measure of the width of the posterior when the pdf is non-Gaussian. Ideally a data assimilation scheme is required that gives at least some understanding of the full posterior pdf.

Particle filters are data assimilation schemes that provide a partial solution to this problem. They are fully nonlinear and represent the full posterior pdf through an ensemble of model runs, called particles, weighted on their proximity to observations. Unfortunately they suffer from what is termed the 'curse of dimensionality'. Due to computer limitations only relatively few particles are available to represent the full posterior pdf. As the dimension of the state increases it is unlikely that these few particles, when moving at random through state space, will end up close to the potentially large number of observations (Snyder *et al.* 2008). Consequentially, only one or two particles end up providing information about the probability region of interest. One option would be to increase the number of particles until sufficiently many will always end up close to all observations. This is an unrealistic scenario unless computational power and efficiency drastically increase. A more immediate solution is to develop a scheme that ensures that only a few particles, when compared to the number required by the SIR filter, are able to provide a representation of the high probability regions of the full posterior pdf.

The equivalent weights particle filter has already been introduced by Van Leeuwen (2010, 2011). It uses a proposal density to guide all particles towards the observations and initial results are very exciting. The scheme was able to track the true solution of the chaotic three-dimensional Lorenz-63 model with only partial observations of the state vector using just three particles. In comparison, the SIR filter with 20 particles will still fail to capture the transition to the opposite wing of the Lorenz butterfly attractor for the chosen parameter settings. The scheme also scales extremely well. When applied to the both the 40-dimensional and 1000-dimensional Lorenz 95 model, 20 particles were sufficient to capture the behaviour of the true solution where as hundreds to thousands of particles are needed with traditional particle filters (Van Leeuwen 2011). This shows there is the potential for particle filters to represent the full posterior pdf with only relatively few particles in large dimensional geoscience applications.

In this paper we explore the equivalent weights particle filter in greater depth. We show how it relates to alternative schemes which also utilise proposal densities and how the proposal density can be designed to ensure that particles have certain properties. In particular we look at how, by combining different choices of proposal density, the particle filter can result in the majority of particles contributing significant information about the probability area of the posterior pdf local to the observations. This

results in a much smaller ensemble being required to give an effective representation of the posterior pdf and leads to a scheme which will scale well to high-dimensional systems. Using the Lorenz (1963) model we demonstrate how the representation of the posterior is sensitive to choices made as part of the specified proposal densities. Provided the right choices are made, the ability of the scheme to capture the essence of the posterior using just 20 particles is demonstrated.

2. Equivalent weights particle filter

2.1. *Particle filters*

We start by reviewing the basic particle filter and how it can be adapted to allow for sampling from a proposal density. In the most general form of the particle filter, the posterior pdf can be considered as the distribution of possible model trajectories over a period of time given vectors of observations of the state. Using Bayes theorem it can then be written:

$$p(x^{0:n}|y^{1:n}) = \frac{p(y^{1:n}|x^{0:n})p(x^{0:n})}{p(y^{1:n})} \tag{1}$$

where $x^{0:n}$ is a sequence of k-dimensional states of the system $(x^0, x^1, ..., x^n)$ that gives the model trajectory over the n time steps and $y^{1:n}$ is the set of observation vectors. The probability distribution of the observations given a particular trajectory, $p(y^{1:n}|x^{0:n})$, is the likelihood and is generally considered to be a known Gaussian distribution, although it is not restricted to this. The model prior, $p(x^{0:n})$, is the distribution of the model trajectories before the observations are taken into account.

If it is assumed that the observations at different times are independent and that $p(y^j|x^{0:n}) = p(y^j|x^j)$, then the posterior pdf becomes:

$$\begin{aligned} &p(x^{0:n}|y^{1:n}) \\ &= \frac{p(y^{1:n}|x^{0:n})p(x^{0:n})}{p(y^{1:n})} \\ &= \frac{p(y^n|x^n)}{p(y^n)}\frac{p(y^{n-1}|x^{n-1})}{p(y^{n-1})}...\frac{p(y^1|x^1)}{p(y^1)}p(x^{0:n}). \end{aligned} \tag{2}$$

Particle filters represent the model prior via Monte Carlo methods as an ensemble of system state trajectories or particles. Hence $p(x^{0:n})$ is given by the sum of delta functions positioned at the state trajectories chosen as the particles:

$$p(x^{0:n}) = \frac{1}{N}\sum_{i=1}^{N}\delta(x^{0:n} - x_i^{0:n}). \tag{3}$$

Using the particle representation in Bayes theorem gives:

$$\begin{aligned} p(x^{0:n}|y^{1:n}) &= \sum_{i=1}^{N}\frac{p(y^n|x_i^n)}{p(y^n)}...\frac{p(y^1|x_i^1)}{p(y^1)}\frac{1}{N}\delta(x^{0:n} - x_i^{0:n}) \\ &= \sum_{i=1}^{N} w_i^n...w_i^1\frac{1}{N}\delta(x^{0:n} - x_i^{0:n}) \end{aligned} \tag{4}$$

where the weights w_i^j are given by

$$w_i^j = \frac{p(y^j|x_i^j)}{p(y^j)} = \frac{Np(y^j|x_i^j)}{\sum_{i=1}^{N}p(y^j|x_i^j)}. \tag{5}$$

Since $p(y^j) = \int p(y^j|x^j)p(x^j)dx^j = \frac{1}{N}\sum_{i=1}^{N}p(y^j|x_i^j)$ ensures normalised weights we have that $w_i^j \leq 1, \forall j = 1, ..., n$. So for large numbers of observations the product of the weights over all time steps can become very low. In general it is found that only a few observations are required before one ensemble member takes all the weight leaving the rest to have negligible weight (Doucet *et al.* 2001; Gordon *et al.* 1993). This means that only this one particle has statistical significance in estimating the posterior pdf, which then effectively becomes a delta function centered on that state trajectory. This is known as filter degeneracy and means the advantage of the particle filter in allowing for a

full representation of the posterior pdf is lost. One way to avoid filter degeneracy is to use resampling. Every time an observation becomes available the weights are calculated. Those particles with very low weights are then abandoned whilst particles with high weight are kept and multiplied, so that once again all particles have equal weight. Particle filters that apply resampling are known as sequential importance resampling (SIR) filters and several ways of implementing resampling are available, see for example (Gordon *et al.* 1993; Lui and Chen 1998; Kitagawa 1996).

Since resampling results in the trajectories of some particles being curtailed, we can no longer consider particles across the entire time window. Instead state trajectories are now considered over the period between observations, so the posterior pdf becomes the probability distribution of these shortened trajectories given the latest observation vector:

$$\begin{aligned} p(x^{n-r:n}|y^n) &= \frac{p(y^n|x^n)p(x^{n-r:n})}{\int p(y^n|x^n)p(x^{n-r:n})dx^{n-r:n}} \\ &= \sum_{i=1}^{N} w_i^n \frac{1}{N}\delta(x^{n-r:n} - x_i^{n-r:n}) \end{aligned} \quad (6)$$

where w_i^n is as above and there are r steps from the previous vector of observations. Unfortunately resampling does not entirely solve the problem of filter degeneracy. It is still a serious issue in particle filtering whilst computer capabilities limit the number of particles that can be used (Snyder *et al.* 2008). The equivalent weights particle filter attempts to solve this problem.

2.2. *Equivalent weights particle filter: basic ideas*

The SIR filter gives a method for calculating the full posterior pdf via the particle states and their associated weights. However, in general it is of more interest to determine the expectation value of a function of the state vector:

$$\begin{aligned} &\overline{g(x^{n-r:n})} \\ &= \int g(x^{n-r:n})p(x^{n-r:n}|y^n)dx^{n-r:n} \\ &= \frac{1}{A}\int g(x^{n-r:n})p(y^n|x^{n-r:n})p(x^{n-r:n})dx^{n-r:n} \end{aligned} \quad (7)$$

where A is a normalisation factor. This integral can be written as

$$\begin{aligned} \overline{g(x^{n-r:n})} =& \frac{1}{A}\int g(x^{n-r:n})p(y^n|x^{n-r:n})\frac{p(x^{n-r:n})}{q(x^{n-r:n})} \\ &\times q(x^{n-r:n})dx^{n-r:n} \end{aligned} \quad (8)$$

where $q(x^{n-r:n})$ is another probability distribution called the proposal density. This will not effect the expected value provided that the support of $q(x^{n-r:n})$ is larger than or at least equal to the support of $p(x^{n-r:n})$, thus avoiding division by zero.

Rather than sampling particles from $p(x^{n-r:n})$, we now sample instead from $q(x^{n-r:n})$ which can be chosen arbitrarily. Replacing $q(x^{n-r:n})$ in Eq. (8) above with our Monte Carlo representation we find that:

$$\begin{aligned} \overline{g(x^{n-r:n})} &= \frac{1}{A}\int g(x^{n-r:n})p(y^n|x^{n-r:n})\frac{p(x^{n-r:n})}{q(x^{n-r:n})} \\ &\quad\times \frac{1}{N}\sum_{i=1}^{N}\delta(x^{n-r:n} - x_i^{n-r:n})dx^{n-r:n} \\ &= \frac{1}{A}\sum_{i=1}^{N} g(x_i^{n-r:n})p(y^n|x_i^{n-r:n})\frac{p(x_i^{n-r:n})}{q(x_i^{n-r:n})} \\ &= \sum_{i=1}^{N} w_i^n g(x_i^{n-r:n}) \end{aligned} \quad (9)$$

where the weights now include an extra term:

$$w_i^n = \frac{1}{A}p(y^n|x_i^{n-r:n})\frac{p(x_i^{n-r:n})}{q(x_i^{n-r:n})}. \quad (10)$$

Since the posterior pdf consists of both the particle trajectories and their associated weights, the extra term in the weight can be viewed as compensating for any change in

the trajectories due to not sampling directly from the model prior.

Other than requiring the support of $q(x_i^{n-r:n})$ to be at least equal to that of $p(x_i^{n-r:n})$, no other restrictions are placed on the proposal density. However since the aim is to improve the likelihood, it would seem logical to include the information from the observations as part of the density. This can be done in several ways and, although this paper does consider alternatives, its main focus is based on the ideas that follow.

2.3. A sequential approach to the proposal density

Before discussing how we include observation information in the proposal density it is necessary to note the sequential nature of the model prior. Exploiting the Markov chain property of the evolution of the model trajectory (e.g. Van Leeuwen (2009)) it is possible to write

$$p(x^{n-r:n}) = p(x^n|x^{n-1})p(x^{n-1}|x^{n-2})... \\ ...p(x^{n-r+1}|x^{n-r})p(x^{n-r}). \quad (11)$$

Each probability density $p(x^j|x^{j-1})$ is called a transition density and relates to the probability of moving from one state to the next in time. If $f(x_i^{j-1})$ are the discretised model equations applied to the model state at the previous time and $d\beta_i^j$ is the stochastic error representing unknown terms in the model equations, then the model state at the new time step is given by:

$$x_i^j = f(x_i^{j-1}) + d\beta_i^j. \quad (12)$$

Although the distribution of the $d\beta_i^j$ can be chosen appropriately for the model, if it is assumed to be Gaussian with mean zero and covariance Q then the expected value of the new model state given the old model state x_i^{j-1} is $f(x_i^{j-1})$ with covariance Q. Hence the transition density is also distributed as a Gaussian, $p(x^j|x_i^{j-1}) \sim N(f(x^{j-1}), Q)$. Obviously if the distribution of the $d\beta_i^j$ is chosen differently, the related transition density will also change.

The proposal density differs from the model prior since we choose it to be based on observational information. Whilst theoretically it is possible to have the proposal density dependent on all observations, to improve the likelihood of each particle only requires the proposal density to be based on the next available observation. In addition we choose the proposal density to have the same Markov property as the model prior so that $q(x^{n-r:n})$ becomes:

$$q(x^{n-r:n}, y^n) = q(x^n|x^{n-1}, y^n)q(x^{n-1}|x^{n-2}, y^n)... \\ ...q(x^{n-r+1}|x^{n-r}, y^n)q(x^{n-r}). \quad (13)$$

We choose this proposal transition density in a similar way to the model transition density. Like the model prior it is based on both the discretised model equations and includes a stochastic error. The difference lies in an additional term based on the future observation that works to relax the particle towards that observation. Hence the new model state at time j is now given by:

$$x_i^j = f(x_i^{j-1}) + B(\tau)(y^n - h(x_i^{j-1})) + d\hat{\beta}_i^j, \quad (14)$$

where $B(\tau)$ is a matrix specifying the strength of the relaxation dependent on the distance in time τ to the next observation and $h(x_i^{j-1})$ is a measurement operator which projects the model state on the observation y^n, but at time $j-1$. If the random error $d\hat{\beta}^j$ is again distributed as a Gaussian with mean zero and now with covariance $\hat{Q}$, then the proposal transition density is also Gaussian but now with mean $f(x_i^{j-1}) + B(\tau)(y^n - h(x_i^{j-1}))$. The proposal covariance $\hat{Q}$ can be chosen as one wishes but failing better knowledge is currently kept to be the same as Q.

This is a simple way to ensure that each particle ends up close to the future observation by giving it a small nudge at every time step. If a particle is close to the observation then its likelihood $p(y^n|x_i^n)$ is high and the hope is that it will now have significant weight in estimating the posterior density. However, as already stated, sampling from a proposal density rather than the model prior also effects the composition of the weights. Accounting for the sequential nature of both the model prior and the chosen proposal density the weights are now given by:

$$w_i^n = \frac{1}{A} p(y^n|x_i^n) \prod_{j=n-r+1}^{n} \frac{p(x_i^j|x_i^{j-1})}{q(x_i^j|x_i^{j-1}, y^n)}. \qquad (15)$$

By sampling x_i^j from $q(x^j|x_i^{j-1}, y^n)$ rather than $p(x^j|x_i^{j-1})$ at each time step, the probability $p(x_i^j|x_i^{j-1})$ is usually reduced. Hence $p(x_i^j|x_i^{j-1})$ is generally smaller than $q(x_i^j|x_i^{j-1}, y^n)$. Multiplied over several time steps this can lead to a very small value for $\prod_{j=n-r+1}^{n} \frac{p(x_i^j|x_i^{j-1})}{q(x_i^j|x_i^{j-1}, y^n)}$ and so, irrespective of the likelihood, the weight can become very small. Filter degeneracy is still occurring despite sampling from the chosen proposal density and an additional step is needed.

2.4. *Final proposal density*

The choice of proposal density considered so far has been designed to control the position of the particles and an associated change in weight has therefore been calculated. An equally valid alternative would be to decide upon the desired weight for a particle and then design the proposal density to ensure this weight. With this in mind it is informative to look at previous and current ideas for ensuring specific weights for particles before the 'equivalent weights' scheme proposed by Van Leeuwen (2010) is considered in more detail.

2.4.1. *Optimal proposal density*

This was discussed by Doucet *et al.* (2000) under the title of importance sampling and it naturally falls within the premise of proposal densities. Bocquet *et al.* (2010) also study the optimal proposal density and show how it outperforms the SIR filter on a Lorenz 95 example for a small ensemble size. The aim is to avoid filter degeneracy by using the proposal density to sample particles such that the variance of the weights are minimised. For simplicity let it be assumed that observations are available at every time step. If particles are then picked from a proposal density, the subsequent weight given to each particle is specified by:

$$w_i = \frac{1}{A} p(y^n|x_i^n) \frac{p(x_i^n|x_i^{n-1})}{q(x_i^n|x_i^{n-1}, y^n)} \qquad (16)$$

similar to the sequential approach outlined above. The suggested optimal proposal density is $q(x^n|x_i^{n-1}, y^n) = p(x^n|x_i^{n-1}, y^n)$. Using Bayes theorem and the fact that in general the systems of interest can be considered to be Markov,

$$p(x_i^n|x_i^{n-1}, y^n) = \frac{p(y^n|x^n)p(x^n|x^{n-1})}{p(y^n|x^{n-1})}. \qquad (17)$$

Hence the weights simplify to $w_i \propto p(y^n|x_i^{n-1})$. The optimal proposal density is considered to be optimal since there is zero variance in the weights for particles sampled from a proposal density $p(x^n|x_i^{n-1}, y^n)$ for a fixed value of x_i^{n-1}. However the value of x_i^{n-1} is different for each particle and so the zero variance in weights across the proposal density for a fixed x_i^{n-1} is not utilised. This can be seen if we return to the original specification of both the likelihood and model transition density. If $h(x)$ is linear (represented here by H) and hence both $p(x^n|x^{n-1})$ and $p(y^n|x^n)$ are Gaussian distributed, the weights using this

optimal proposal density are proportional to

$$\begin{aligned} w_i &\propto p(y^n|x_i^{n-1}) \\ &\propto \exp\left[-\frac{1}{2}(y^n - Hf(x_i^{n-1}))(HQH^T + R)^{-1}\right. \\ &\quad \left.\times(y^n - Hf(x_i^{n-1}))\right]. \end{aligned} \tag{18}$$

We wish to estimate the variance of this expression as the ensemble index i varies. The arguments of Snyder *et al.* (2008) can be applied here, but we chose to look at the simple case where both R and HQH^T are diagonal with respective variances given by V_y and V_x. The expression $(y^n - Hf(x_i^{n-1}))$ can be expanded to $((y^n - Hx_t^n) + Hd\beta_t^n + H(f(x_t^{n-1}) - f(x_i^{n-1}))$, where x_t^n denotes the true model state, and hence

$$\begin{aligned} -\log(w_i) &\propto \frac{1}{2(V_x + V_y)} \\ &\sum_{j=1}^{M}\left((y^n - Hx_t^n) + Hd\beta_t^n + H(f(x_t^{n-1}) - f(x_i^{n-1})\right)^2 \end{aligned} \tag{19}$$

where M is the number of independent observations. The variance of $-\log(w_i)$ is given by

$$\mathrm{var}[-\log(w_i)] \propto \frac{M}{2}\left(\frac{P_x^n}{V_x + V_y}\right)^2\left(1 + 2\left(\frac{V_y + V_x}{P_x^n}\right)\right) \tag{20}$$

where $P_x^n = HAP^{n-1}A^TH^T$ with A the linearised model equations and P^{n-1} the ensemble variance at time $n-1$ (see Appendix A for details). Assuming P^{n-1}, V_x and V_y are fixed, this shows that the variance of the weights is directly linked to M, the number of independent observations. Regardless of this optimal choice of proposal density, in reality filter degeneracy caused by significant variation in the weights will still be present in the large-dimensional geophysical systems of interest.

2.4.2. Implicit particle filter

Although Chorin and Tu (2009) adopt a different formulation in the implicit particle filter, for observations at every time step it is the same as the optimal proposal density (Chorin *et al.* 2010; Morzfeld *et al.* 2012) and so the variance of the weights will again increase with the number of independent observations. To see why this is true we start with the weights at observation time n given by (16) for the simplest case of observations available every time step. To try and ensure all samples have high weights, Chorin *et al.* (2010) wish to use the proposal density to sample particles from $p(y^n|x^n)p(x^n|x_i^{n-1}) = p(x^n|x_i^{n-1}, y^n)p(y^n|x_i^{n-1})$ (see Eq. (17)). We have that

$$\begin{aligned} &p(y^n|x^n)p(x^n|x_i^{n-1}) \\ &= \frac{1}{A}\exp\left[-\frac{1}{2}(y^n - Hx^n)^TR^{-1}(y^n - Hx^n)\right. \\ &\quad \left.-\frac{1}{2}(x^n - f(x_i^{n-1}))^TQ^{-1}(x^n - f(x_i^{n-1}))\right] \\ &= \frac{1}{A}\exp\left(-\frac{1}{2}(x^n - \hat{x}_i^n)^TP^{-1}(x^n - \hat{x}_i^n) - \phi_i\right) \\ &= \frac{1}{A}\exp\left(-\frac{1}{2}(x^n - \hat{x}_i^n)^TP^{-1}(x^n - \hat{x}_i^n)\right)\exp\left(-\phi_i\right) \\ &= p(x^n|x_i^{n-1}, y^n)p(y^n|x_i^{n-1}) \end{aligned} \tag{21}$$

where

$$P^{-1} = H^TR^{-1}H + Q^{-1} \tag{22}$$

$$\hat{x}_i^n = P(Q^{-1}f(x_i^{n-1}) + H^TR^{-1}y^n) \tag{23}$$

$$\begin{aligned} \phi_i = \frac{1}{2}&(y^n - Hf(x_i^{n-1}))^T(HQH^T + R)^{-1} \\ &\times(y^n - Hf(x_i^{n-1})) \end{aligned} \tag{24}$$

and $\hat{x}_i^n$ would be more commonly recognised in the Kalman filter like form

$$\begin{aligned} \hat{x}_i^n =& f(x_i^{n-1}) \\ &+ QH^T(HQH^T + R)^{-1}(y^n - Hf(x_i^{n-1})). \end{aligned} \tag{25}$$

Particles can be sampled from $p(x^n|x_i^{n-1}, y^n) \propto \exp\left(-\frac{1}{2}(x^n - \hat{x}_i^n)^T P^{-1}(x^n - \hat{x}_i^n)\right)$ by sampling a k-dimensional Gaussian reference variable ξ_i^n from $N(0, I)$ and transforming via $x_i^n = P^{\frac{1}{2}}\xi_i^n + \hat{x}_i^n$. This gives the proposal density $q(x^n|x_i^{n-1}, y^n) = p(x^n|x_i^{n-1}, y^n)$. The weights now simplify to $w_i^n \propto p(y^n|x_i^{n-1}) = \exp(-\phi_i)$. Comparing the definition of ϕ_i to Eq. (18), the equivalence of the two schemes when observations are available every time step is immediately apparent. Hence, like the optimal proposal density, the implicit particle filter cannot guarantee the avoidance of filter degeneracy in large-dimensional systems with large numbers of observations. When observations are not available every timestep, calculating the weights of the implicit particle filter is equivalent to finding a solution to the weak constraint 4D-Var problem with fixed initial conditions. It is currently unclear how the formulation over multiple time steps will change these estimates.

2.4.3. Equivalent weights

The equivalent weights proposal density, as specified in Van Leeuwen (2010), uses the same principles as the above schemes but tries to avoid the relation between the variance of the weights and the number of independent observations. The aim is to ensure equally significant particles are picked from the posterior density. It is assumed that the particles are already in the probability region of the posterior local to the observations through the relaxation term included in all the previous proposal densities. Hence the weight accrued in all time steps prior to that immediately preceding an observation is an inherent part of the scheme and we return to the specification of the weights at observation time given in (15).

Separating the weight accrued by each particle until the last step before an observation, the final weight of a particle is given by

$$w_i^n = \frac{1}{A}\left(\prod_{j=1}^{n-1}\frac{p(x_i^j|x_i^{j-1})}{q(x_i^j|x_i^{j-1}, y^n)}\right)\frac{p(x_i^n|x_i^{n-1})p(y^n|x_i^n)}{q(x_i^n|x_i^{n-1}, y^n)}$$
$$\propto w_i^{rest}\frac{p(x_i^n|x_i^{n-1})p(y^n|x_i^n)}{q(x_i^n|x_i^{n-1}, y^n)}. \quad (26)$$

The proposal density $q(x_i^n|x_i^{n-1}, y^n)$ in the last time step can be used to set $w_i^{rest}p(x_i^n|x_i^{n-1})p(y^n|x_i^n)$ equal to a constant. This is similar to the implicit particle filter but rather than randomly sampling particles we now pick a specific particle that will give us the desired weight. This is equivalent to solving the equality given by

$$-\log w_i^{rest} + \frac{1}{2}(x_i^n - f(x_i^{n-1}))^T Q^{-1}(x_i^n - f(x_i^{n-1}))$$
$$+ \frac{1}{2}(y^n - h(x_i^n))^T R^{-1}(y^n - h(x_i^n)) = C. \quad (27)$$

The value of C is chosen based on the maximum weight each particle can achieve (Van Leeuwen 2010). Unless it is chosen based on the lowest maximum weight, it will only be possible to solve this equality for a certain percentage of particles. However, that choice will lead to all weights becoming equal to that of the lowest. Instead a compromise can be reached between the percentage of particles kept and the value of the weight, with those discarded as unsolvable still returning via resampling. The effect of changing the value of C, and hence the percentage of particles kept, is discussed in Section 3.6.1.

One solution to this equality, where it is solvable, leads to new model states for each particle at observation time given by (Van Leeuwen 2010):

$$x_i^* = f(x_i^{n-1}) + \alpha_i K(y^n - Hf(x_i^{n-1})) \quad (28)$$

where H is now our linear observation operator. This is the same as Eq. (28) only now with the addition of α_i, which is used to ensure equal weights for the specified percentage of particles rather than the minimum weight given by Eq.

(28). α_i solves the quadratic equation (30) and is given by $\alpha_i = 1 \pm \sqrt{1 - b_i/a_i}$ in which

$$a_i = 0.5 d_i^T R^{-1} H K d_i \tag{29}$$

$$b_i = 0.5 d_i^T R^{-1} d_i - C - \log w_i^{rest} \tag{30}$$

$$K = Q H^T (H Q H^T + R)^{-1} \tag{31}$$

$$d_i = y^n - H f(x_i^{n-1}). \tag{32}$$

Previously the negative root has been chosen for α_i (Van Leeuwen 2010), however the choice of root can have a significant impact on the behaviour of the equivalent weights step and this is explored in more detail in Section 3.3.

Unfortunately, by picking a specific particle there is no stochastic dependency. This leads to a proposal transition density equivalent to a delta function centered on the deterministic value given by Eq. (31). The one restriction placed on the proposal density is that its support needs to be at least equal to that of the model prior. Since this is not true of a delta function an additional stochastic term is added to the equation:

$$x_i^n = f(x_i^{n-1}) + \alpha_i K (y^n - H f(x_i^{n-1})) + d\tilde{\beta}_i^n. \tag{33}$$

The final value for the weights is now given by:

$$\begin{aligned} w_i^n &= \frac{w_i^{rest} p(x_i^n | x_i^{n-1}) p(y^n | x_i^n)}{q(x_i^n | x_i^{n-1}, y^n)} \\ &\propto \frac{\exp(-C_i)}{q(x_i^n | x_i^{n-1}, y^n)}. \end{aligned} \tag{34}$$

where C_i is equal to C plus a perturbation due to the addition of the random error $d\tilde{\beta}_i^n$. The value of the weights using the optimal proposal density depends primarily on the differences between $p(y^n | x_i^{n-1})$ for each particle. In the equivalent weights particle filter, any difference in weights is now due to the distribution of the proposal density in this last step, and hence the distribution of $d\tilde{\beta}^n$.

If $d\tilde{\beta}^n$ is chosen to be Gaussian with mean zero and covariance $\tilde{Q}$ then once again the proposal transition density is Gaussian with mean now given by $f(x_i^{n-1}) + \alpha_i K(y^n - H f(x_i^{n-1})) = x_i^*$. In order to ensure that the expected change from this state is insignificant, the error covariance can be chosen to be small. For example it could be set proportional to the model covariance, $\tilde{Q} = \gamma^2 Q$, where γ is a small dimensionless number, as was originally suggested in Van Leeuwen (2010). However it transpires that regardless of the value of γ, choosing the proposal transition density as a Gaussian can lead to filter degeneracy.

If $\gamma d\beta_i^n = d\tilde{\beta}_i^n \sim N(0, \gamma^2 Q)$, where $d\beta_i^n \sim N(0, Q)$ and γ is sufficiently small, then the dominant terms when calculating the weights are:

$$-\log w_i^n \approx C - \frac{1}{2} d\beta_i^{nT} Q^{-1} d\beta_i^n \tag{35}$$

(see Appendix B for details). Since there is no dependence on γ in the above expression, the size of $d\beta_i^n$ controls the value of the weight. An outlier from $N(0, Q)$ will reduce $-\log w_i^n$ leading to a significantly larger weight, regardless of the value of γ. Filter degeneracy may once again occur.

Ideally, all weights should be approximately equal. This requires both $q(x^n | x^{n-1}, y^n)$ to be close to a constant, regardless of the model state sampled, and the model state not to change significantly. One distribution that would fulfil these criteria is a Uniform distribution over a small interval. Unfortunately, similar to the delta function of the deterministic solution, a Uniform distribution does not have support equal to or greater than the Gaussian distribution $p(x^n | x_i^{n-1})$.

A solution is to use a mixture density so that

$$\begin{aligned} &q(x_i^n | x_i^{n-1}, y^n) \\ &= (1 - \epsilon) Q^{\frac{1}{2}} U[x_i^* - \gamma_U I, x_i^* + \gamma_U I] + \epsilon N(x_i^*, \gamma_N^2 Q) \end{aligned} \tag{36}$$

where I is a k dimensional vector of ones with k the dimension of the system. Since a multivariate uniform distribution does not exist, this equates to choosing a uniform random variable for each dimension of the state from the range given by plus and minus γ_U. These uniform random variables are then multiplied by $Q^{\frac{1}{2}}$ to introduce correlations between these variables. Using a mixture density means the error can be drawn from either a Uniform distribution or a Gaussian distribution with the proportion controlled by the value of ϵ. Choosing ϵ to be small ensures that in general x_i^n are picked from a uniform distribution. In particular by relating ϵ to the size of the ensemble, for example $\epsilon = 0.001/N$, means that drawing x_i^n from the Gaussian distribution is very unlikely even as the ensemble size increases and hence equivalent weights can be assured for all particles. However unlike using a purely Uniform distribution, the possibility of picking from the Gaussian ensures continuous support across the entire space of x_i^n. Appendix D gives further guidance on the relation between γ_N, γ_U and ϵ.

2.4.4. Discussion

The ideal particle filter would concisely represent the posterior distribution $p(x^n|y^n)$ with only a few particles. In geophysical applications, with a large number of observations, the likelihood is very localised in state space. The prior distribution has a much larger spread due to the non-linearities present in most geophysical equations. In order to gain information about the posterior density in the state space local to the observations we wish to only sample particles from this region. We also require particles to have equal significance in order to avoid filter degeneracy. The optimal proposal density, the implicit particle filter and the equivalent weights particle filter all sample particles from proposal densities to try and ensure particles with these properties. The optimal proposal density aims to sample particles with minimally varying weights and the implicit particle filter from the region of the posterior local to observations, however the schemes are the same for a linear H and Gaussian distributed observation and model transition densities when observations are available at every time step. Furthermore the variance of the weights increases with the number of independent observations so both will ultimately suffer from filter degeneracy. The equivalent weights particle filter combines these two aims. It uses the proposal density in the majority of time steps to relax each particle towards the probability region of the posterior local to the observations and then the equivalent weights step ensures equally significant particles. Since the proposal density in the equivalent weights step is chosen as a mixture density, there is not the cancellation in the factors of the weight as occurs in the optimal proposal density. Hence the variance is unrelated to the number of independent observations and there is not the same potential for filter degeneracy to occur.

It is also worth noting here the differences in computational expense between the schemes. As already mentioned, calculating the weights of the implicit particle filter applied over multiple time steps between observations is equivalent to weak constraint 4D-Var with fixed initial conditions. This requires inner and outer loop iterations for each particle which gets increasingly costly as the dimension of the system and the number of particles increases. In contrast calculating the weights in the equivalent weights particle filter requires inverting two matrices, Q and $(HQH^T + R)$, and a subsequent matrix vector multiplication at each time step. Although the size of the matrices will increase with the dimension of the system, they must only be inverted once and are the same for each particle. Hence the main computational increase from the SIR filter is the matrix vector multiplication.

2.5. *Convergence of the scheme*

The final section in this theoretical consideration of the equivalent weights particle filter discusses the convergence of the scheme as the number of particles grows. With the SIR filter, as the number of particles approaches infinity, the posterior distribution becomes the true posterior distribution of the system (Doucet *et al.* 2001). It is desirable to retain this property as the SIR filter is adapted to make it usable for high-dimensional systems.

Theoretically, a particle filter which uses a proposal density will still converge to the true posterior as the number of particles approaches infinity. However the application of the equivalent weights scheme to only a specified percentage of particles in the final proposal density, means the convergence of the scheme is no longer clear. One solution would be to adapt the proposal densities so that as the number of particles grows the scheme reverts back to the SIR filter. For example this could involve including a factor of $1000/(1000+N)$, where N is the total number of particles, as part of the relaxation term in the majority of proposal densities and as part of α_i in the equivalent weights step. This would also require both the ϵ and γ_N of the mixture density to tend to one as N increases. This has not been implemented in the following example since here the aim is to demonstrate the ability of the equivalent weights particle filter to represent the posterior with only a few particles.

3. Application to the Lorenz-63 model

The efficiency of the equivalent weights particle filter to estimate the ensemble mean has already been demonstrated by applying it to the Lorenz (1963) model (Van Leeuwen 2010). Here a more detailed examination is made of how well the scheme manages to succeed in representing the posterior pdf with a relatively small number of particles. In particular the effects of changing some of the particular choices within the scheme are explored in detail.

3.1. *Model specification*

We use the standard parameters $\sigma = 10$, $\rho = 28$ and $\beta = 8/3$ and assume Gaussian distributions with a standard deviation of $\sqrt{2}$ for both the observation and initial model error. The starting point is (1.508870, -1.531271, 25.46091) and this is perturbed using the initial model standard deviation to form the ensemble of particles. The model error distribution is taken as a Gaussian with a standard deviation of $\sqrt{2}\Delta t$, where $\Delta t = 0.01$, multiplied by a correlation matrix $\widetilde{Q}$ which has 1 on the diagonal, 0.5 on the first sub- and super-diagonals and 0.25 on the second sub- and super-diagonals. A truth scenario is generated by solving the stochastic model with the above parameters using the Euler-Maruyama time discretisation scheme and values from this truth run are taken every 40 times steps. Random noise is added to these values using the standard deviation given above to generate the observations.

3.2. *Relaxation proposal density*

The equivalent weights particle filter uses two distinct proposal densities. The first relaxes the particles towards the observations, and hence to the localised region of the posterior, over the majority of time steps. As outlined in Section 2.3, we update the model state for these time steps according to

$$x_i^j = f(x_i^{j-1}) + B(\tau)(y^n - Hx_i^{j-1}) + d\hat{\beta}_i^j \qquad (37)$$

where τ increases linearly from zero at the previous observation time to one at n and with the random Gaussian error $d\hat{\beta}_i^j$ having the model standard deviation and correlation given above. The first choice to be explored is the function $B(\tau)$ and here we consider the effect changing this function has on the ensemble of particles.

The function $B(\tau)$ has two primary purposes. Its first purpose is to control the strength with which each particle is pulled towards the observations and its second is to spread

this information to unobserved variables. Spreading the information is achieved by including the model correlation matrix $\widetilde{Q}$ along with H^T and this should form part of the function regardless of how strong the relaxation is chosen to be. The simplest way of relaxing towards the observation would be to use a linear function that increases with τ, however applying this from the start of the analysis window immediately starts to restrict the movement of each particle as determined by the model equations. A more sensible choice would be to start the linear increase from half way between the two observations, thus allowing the ensemble more freedom to initially spread out due to the random forcing. Figure 1 shows the effect of two different choices for $B(\tau)$ on both the trajectories of the particles and their ability to represent the true posterior pdf. $B_1(\tau)$ uses the linear increase from half way between the two observations multiplied by 25 times the model correlation matrix. $B_2(\tau)$ is given by

$$B_2(\tau) = \begin{cases} 0 & \text{if } \tau < 0.6 \\ 40p(\tau)\widetilde{Q}H^T & \text{if } \tau \geq 0.6 \text{ and } \frac{f(x_i^j)-x_i^j}{h(x_i^j)-y^n} > 0 \\ 4p(\tau)\widetilde{Q}H^T & \text{if } \tau \geq 0.6 \text{ and } \frac{f(x_i^j)-x_i^j}{h(x_i^j)-y^n} < 0. \end{cases}$$

where $p(\tau) = -8.9(\tau - 0.9)^2 + 0.8$.

The main difference between the two versions of $B(\tau)$ lies in the trajectories of the particles. $B_1(\tau)$ always applies a relaxation, regardless of the movement already induced by model equations. Hence it can relax too strongly towards the observation. In contrast $B_2(\tau)$ is a conditional function that applies a much weaker relaxation towards the observation if the particle is already moving in that direction via the model equations. Hence the trajectories using $B_2(\tau)$ bear a much stronger resemblance to the truth (Figure 1).

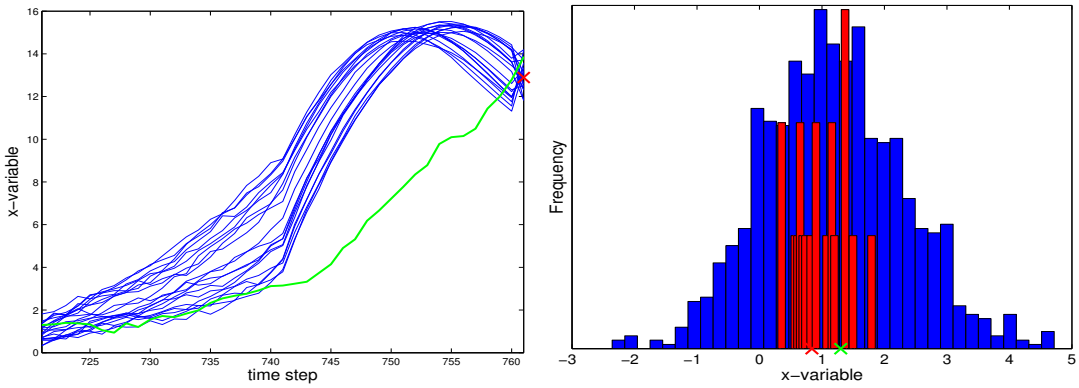
Another consequence of using a linear function is that since the relaxation is strongest immediately prior to an observation, the ensemble gets less diverse as the particles approach the analysis time. If instead the relaxation term decreases in the final few time steps then dispersiveness is re-introduced, although the particles still remain in the vicinity of the observation. This is realised in function $B_2(\tau)$ via $p(\tau)$, a negative quadratic over τ with a peak relaxation centred at 0.9 of the distance between observations. The factors of 40 and 4 included in $B_2(\tau)$ have been chosen rather arbitrarily, but the results seem insensitive to variations of the order of 10-20%.

One final observation is in order. The majority of this paper explores the ability of the equivalent weights particle filter to represent the posterior pdf, since this is ultimately what particle filters are trying to achieve. Provided the relaxation proposal density ensures the particles are in the general vicinity of the observations, the exact choice of $B(\tau)$ has little notable effect on the representation of the posterior using only a small number of particles (Figure 1). The final equivalent weight proposal density, which is considered in the remainder of this paper, has a much stronger influence on the representation of the posterior pdf. However, since we wish to retain as much model information as possible moving into the equivalent weights step, we have chosen to use $B_2(\tau)$ throughout the rest of this paper.

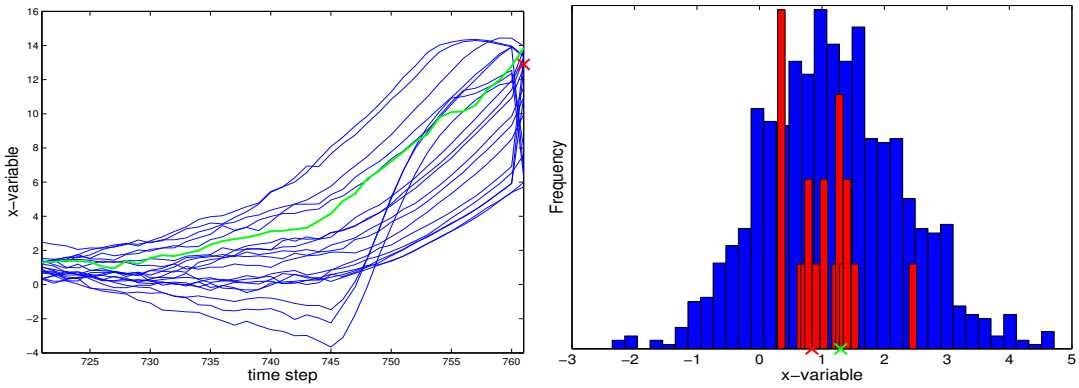
3.3. Choice of root for α_i

We now move on to considering the choices made as part of the equivalent weights proposal density. The first question to answer is whether the positive or negative version of $\alpha_i = 1 \pm \sqrt{1 - b_i/a_i}$ should be used, where α_i dictates the movement of each particle away from its minimum such that $-\log w_i$ attains the specified value for C (Section 2.4.3).

Before the question is answered, it is informative to look at the distribution of weight related to the use of the equivalent weights step. In the SIR filter the value of the weight is controlled by the likelihood, or the distance between a particle and the observations. However



(a) $B_1(\tau)$



(b) $B_2(\tau)$

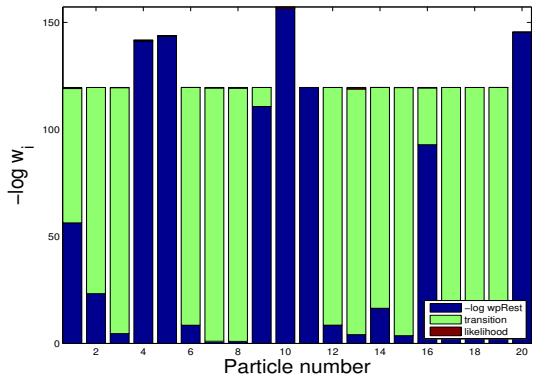
Figure 1. The left column shows the trajectories of particles (blue) compared to the truth (green) between time step 681 and 721 for two different versions of $B(\tau)$ (see text). The right column shows the posterior pdfs for the truth (blue) compared to that generated by the equivalent weights particle filter (red) using 20 particles with the observation (red cross) and truth run (green cross) marked. The truth pdf has been generated using 1000 particles and the standard particle filter. $B(\tau)$ is the only choice varying between the two different versions. Although the trajectories differ substantially when different versions of $B(\tau)$ are used, their representations of the posterior pdf are similar.

in the equivalent weights particle filter the weights include additional components and hence making the weights lower than optimal does not necessarily imply an increase in the distance to the observations. Ignoring for now the proposal density in the last time step, the weights at observation time in the equivalent weights particle filter are given by $w_i = w_i^{rest} p(x_i^n|x_i^{n-1}) p(y^n|x_i^n)$. The weights now depend on the weight accrued from the previous relaxation proposal densities w_i^{rest}, the final model transition density $p(x_i^n|x_i^{n-1})$ as well as the likelihood $p(y^n|x_i^n)$. Decreasing the value of the weights equates to increasing the value of minus the logarithm of the weights:

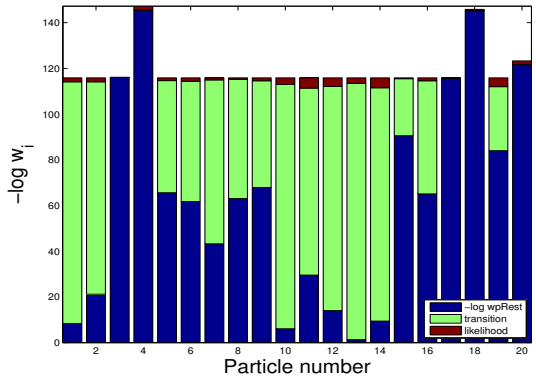
$$-\log w_i^{rest} + \frac{1}{2}(x_i^* - f(x_i^{n-1}))^T Q^{-1}(x_i^* - f(x_i^{n-1})) + \frac{1}{2}(y^n - Hx_i^*)^T R^{-1}(y^n - Hx_i^*). \quad (38)$$

where the three distinct terms relate to the three constituent parts of the weight and x_i^* is as defined in Eq. (31). It is clear from Eq. (41) that another alternative for decreasing the weight of a particle is to move x_i^* away from its deterministic position $f(x_i^{n-1})$.

For the Lorenz 63 system studied it was found that the move away from the deterministic position was more influential in reducing the weight than the distance to the observations, regardless of the version of α_i used. This is verified in Figure 2 which shows bars representing the value of $-\log w_i$ for each of 20 particles using both the positive and negative versions of α_i, with C chosen such that 80% of particles are kept. Each bar is divided into three sections; $-\log w_i^{rest}$ (blue), the value from the model transition (green) and the value from the likelihood (brown). The inclusion of 80% of particles is evident since 16 of



(a) *Positive version of* α_i



(b) *Negative version of* α_i

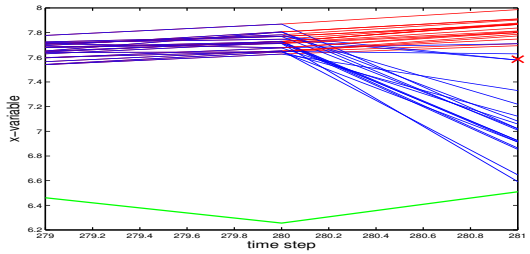
Figure 2. *The value of* $-\log w_i$ *of each particle using the two different versions of* α_i *at time step 641 in a 20 particle run of the Lorenz 63 model with one observation and C chosen so that 80% of particles can achieve that value. In each figure the bars represent the total value of* $-\log w_i$ *for each particle and the colours show how the value is divided between* $-\log w_i^{rest}$ *(blue), the transition (green) and the likelihood (brown). That 20% of the particles are unable to achieve* $-\log w_i = C$ *is evident in the 4 particles with bars higher than the majority. These will be resampled, returning as duplicates of random choices from amongst the 16 particles.*

the 20 particles have bars of equal height and hence the same value of $-\log w_i$. The remaining four particles have bars which are higher than the majority and which consist mainly of value coming from $-\log w_i^{rest}$. These will never be able to achieve the value of C and so will be resampled, returning as duplicates of random choices from the 16 particles. For the rest of the particles it is the transition value (green) that dominates in increasing the value of $-\log w_i$ to C. This is a significant point, since if the transition density has more impact than the likelihood then it is more important that the movement by particles is away from the deterministic position $f(x_i^{n-1})$ than the observations. This means particles can move towards the observations, and by assumption the truth, and still achieve the weight required, provided the movement is such that the particle is sufficiently far from its deterministic model state. Hence ideally α_i needs to be chosen such that the movement by the particles is towards the observations.

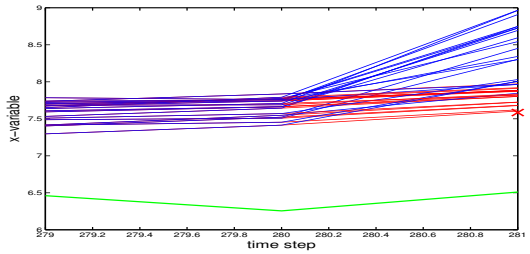
If the deterministic movement by each particle under the equivalent weights proposal density is given by Eq. (31) then a positive α_i is required to ensure movement towards the observation. The negative form of $\alpha_i = 1 - \sqrt{1 - b_i/a_i}$ requires $\sqrt{1 - b_i/a_i} < 1$ in order for α_i to be positive. For the scalar case of only one variable being observed, this can be related to a bound on the difference between C and $-\log w_i^{rest}$. Combined with the bound required for real roots this gives:

$$\frac{d_i^2}{2(V_x + V_y)} \leq C - \left(-\log w_i^{rest}\right) < \frac{d_i^2}{2V_y} \qquad (39)$$

where $d_i = y^n - Hf(x_i^{n-1})$ and V_x and V_y are the scalar values of HQH^T and R (see Appendix C for a derivation of this result). For V_y much larger than V_x, as is generally assumed, this gives a very small bound on the difference between C and $-\log w_i^{rest}$ for α_i to be positive. In general it was found that the difference between C and $-\log w_i^{rest}$ was outside these bounds with the negative version of α_i. In comparison the positive form of $\alpha_i = 1 + \sqrt{1 - b_i/a_i}$ will always be positive and hence ensure a move towards the observation. Although these bounds have only been determined for the scalar case, the argument for the positive form of α_i can be validated graphically for a greater number of observations (Figure 3). Similar behaviour was observed at every time step studied. Hence it can be concluded that in general the positive version of α_i leads to a move by particles towards the observation. Therefore the positive version of α_i is used from now on.



(a) *Positive version of* α_i



(b) *Negative version of* α_i

Figure 3. *Movement of the x-variable of each particle in a 20 particle run of the Lorenz 63 model at time step 281 for the positive (3(a)) and negative (3(b)) version of α_i. The red lines represent the movement by the particles for the purely deterministic model equations. The blue lines are the movement after equivalent weights has been applied and the green line the truth run from which the observation (red cross) was generated. The value of C was chosen so that 80% of particles could achieve that weight and all variables were observed.*

3.4. *Model balance issues*

The above section highlights the importance of the change in model state in achieving weights that are equivalent. This movement by each particle has positive and negative connotations for the performance and validity of the equivalent weights particle filter. Since the transition weight dominates, allowing movement by the particles closer to the observation, the movement often has positive impact on the distribution of the ensemble.

The potential downside to the movement induced by equivalent weights relates to possible balances in a system. Unfortunately it is not possible to explore this issue with the Lorenz 63 model and so further work needs to be undertaken to understand the exact impact introducing additional terms to the model equations has on known variable relationships. For the majority of time steps between observations, where the movement by a particle is increasingly relaxed towards the observations, it is likely that the strength of the relaxation can be controlled so that in general the model equations dominate and the balances are maintained. However the larger movement caused by the addition of equivalent weights can be more problematic, although it may not have quite the potential to be an issue that Figure 3 would imply. It was found in the Lorenz 63 system that as the number of variables observed was increased the movement seen in any one variable became much less distinctive. The Lorenz 63 system has only three variables and so at most only three observations. It may be that in a higher dimensional system, with observation vectors of much greater dimension, much smaller movements of the particles have a larger effect on the weights and so balance issues are less of a problem. Another alternative would be construct both $B(\tau)$ and Q in the equivalent weights step such that they project the model states onto a balanced manifold, similar to the B matrix in 4DVar or initialisation in the EnKF (Houtekamer and Mitchell 2005; Buehner *et al.* 2010). We will explore this issue in more detail using higher dimensional models in future papers.

3.5. *Mixture density*

The next parameters to be discussed relate to the full equivalent weights proposal density rather than its initial deterministic move. The full proposal density is given in Eq. (39) and has the associated parameters γ_U, γ_N and ϵ. The parameter ϵ controls the proportion of particles which are sampled from the Uniform distribution as opposed to the Gaussian distribution, γ_U the width of the uniform distribution and γ_N the variance of the Gaussian distribution. The idea is that by keeping ϵ small, the probability of sampling from the Gaussian is kept to a minimum and filter degeneracy is avoided. Figure 4 shows the effect on the distribution of maximum weights of increasing the value of ϵ so that particles start to be sampled from both distributions. The remaining two parameters, γ_U and γ_N, are kept constant with $\gamma_U = 10^{-5}$ and γ_N related

to γ_U via $\gamma_N = 2^{k/2}\epsilon(\pi^{k/2}(1-\epsilon))^{-1}\gamma_U^k$ (see Appendix D for details).

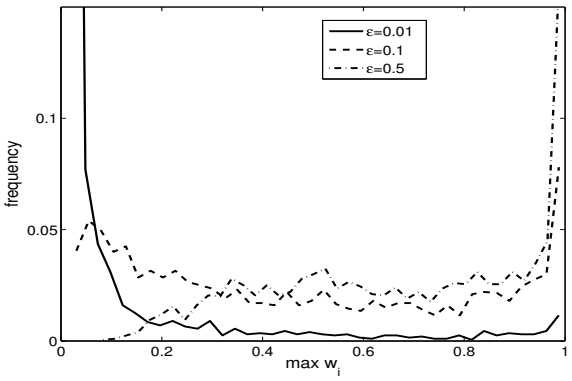


Figure 4. Empirical statistics of the maximum weights of the particles over 2000 observation times using 100 particles for varying values of ϵ. Increasing ϵ results in a greater number of particles being sampled from the Gaussian part of the mixture density, leading to more cases of filter degeneracy.

For $\epsilon = 0.01$ the density is rather balanced with only a few instances of filter degeneracy. As the value of ϵ is increased, the probability of sampling from the Gaussian part of the mixture density rises and the weights become more degenerate. This figure verifies the results from Appendix B, which shows a theoretical justification of why sampling from a Gaussian distribution in the final proposal density is likely to lead to filter degeneracy. To ensure that the probability of sampling from the Gaussian part of the mixture density is minimised, ϵ has been set equal to $0.001/N$ for the rest of the experiments, where N is the number of ensemble members.

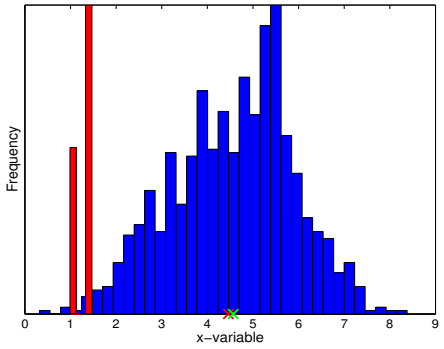
Changing the variable γ_U has no effect on the value of $q(x_i^n|x_i^{n-1}, y^n)$ (again see Appendix D). However, increasing γ_U results in particles being sampled from a Uniform distribution with a greater range. In turn this moves the particle further from the position required to ensure that $-\log\left[w_i^{rest}p(x_i^n|x_i^{n-1})p(y^n|x_i^n)\right] = C$. In effect γ_U controls the variance of $p(x_i^n|x_i^{n-1})p(y^n|x_i^n)$ and hence we wish to choose it to be small ($\gamma_U = 10^{-5}$) to ensure the weights are all equivalent.

3.6. Percentage of particles retained under equivalent weights

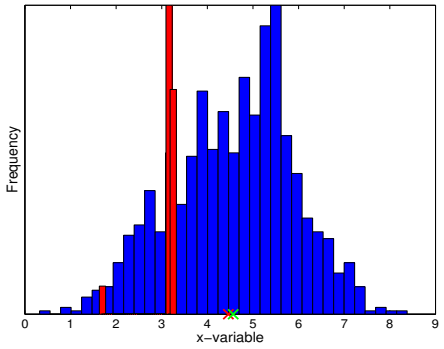
The conclusions reached so far on the choices of $B(\tau)$, α_i, γ_U, γ_N and ϵ are relatively generic. Although they all need to be verified and explored in higher dimensional models it is likely that a relaxation function that allows the model movement to dictate the strength of the pull, a positive α_i that is more likely to ensure the ensemble surrounds the observation or an ϵ chosen small enough that all particles have equivalent weights, will be beneficial choices regardless of the model. The next variable to be considered is the percentage of particles that should be retained under the equivalent weights step. Here the arguments are much less theoretical and so become more dependent on the Lorenz 63 model used to generate the results. Verifying the performance of the particle filter also becomes harder. Changing the percentage of particles retained effects the distribution of the posterior representation, and for a non-Gaussian distribution this is a difficult measure to judge. The advantage of using the Lorenz 63 model to explore this issue is that it is of small enough dimension to enable an approximation to the truth posterior to be generated using 1000 particles and the SIR filter. Here we use several different methods to draw conclusions on the equivalent weights particle filter. The first is a visual and statistical comparison of the individual posterior pdfs generated by the equivalent weights particle filter compared to the approximation of the true posterior generated by the SIR filter. However these only assess performance at individual time steps and so rank histograms (Hamill 2000) and root mean square errors have been used to verify the results over multiple observation times.

3.6.1. Representation of posterior pdfs

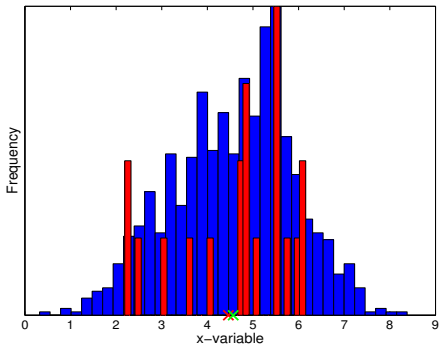
In general the equivalent weights particle filter shows a clear improvement over the SIR filter in capturing the essence of the posterior. Figure 5 shows the posterior pdfs of a



(a) SIR filter



(b) Relaxation only partile filter



(c) Equivalent weights particle filter

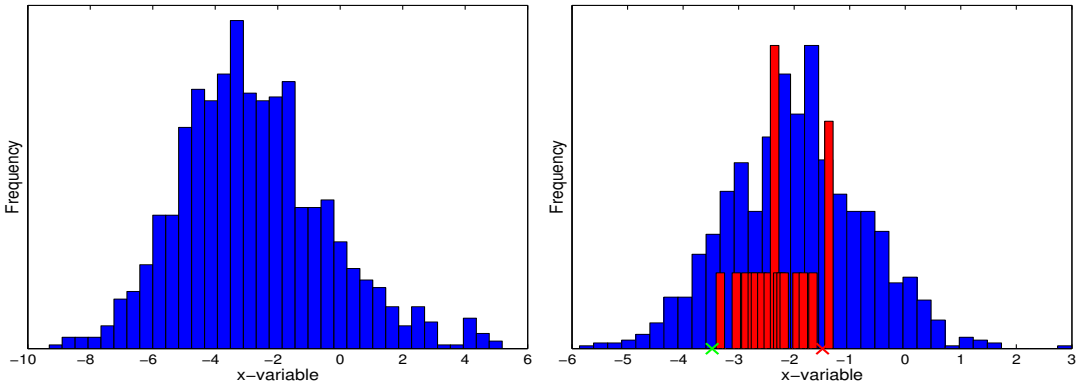
Figure 5. *The posterior probability density functions for the (a) SIR filter, (b) a particle filter which just relaxes towards the observations and (c) for the equivalent weight particle filter with 80% of particles kept, at time step 1321 and with observations of the x-variable only. The blue bars are the result of a 1000 particle run with the SIR filter and the red bars represent the result from each scheme with only 20 particles. The SIR filter clearly misses the mean given by the 1000 particle run. Just relaxing towards the obs brings the mean of the 20 particle posterior closer to the true pdf mean but in no way captures the spread of the pdf. The equivalent weights particle filter not only ends up close to the mean but also captures some of the spread of the true pdf.*

20 particle ensemble run using the SIR filter, a particle filter which just relaxes towards the observations and the effect of both relaxing and applying equivalent weights when 80% of particles are kept. The red bars show the histogram of the x-variable of the particles at time step 1801 and the blue bars the results from a SIR filter using 1000 particles. Applying the equivalent weights step provides a clear improvement over both the other schemes. The SIR filter with 20 particles fails to capture the mean of the 1000 particle pdf and suffers from filter degeneracy since only two particles are present after resampling. If equivalent weights is not included then the pdf gets closer to capturing the mean but filter degeneracy means that again none of the shape of the blue pdf is seen. In the equivalent weights scheme not only is the mean captured but also some of the spread of the pdf.

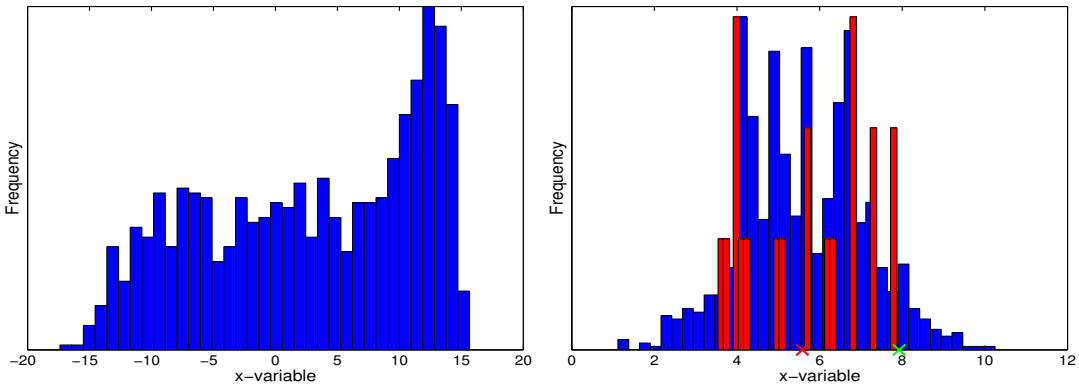
Particle filters are preferable over other data assimilation schemes since they allow for both multi-modal prior and posterior distributions. In the Lorenz 63 system it was found that when using 20 particles with the equivalent weights particle filter, the support of the 20 particles in general matched the support of the full 1000 particle run from the SIR filter, regardless of the shape of either the prior or posterior distributions. Although the full posterior seems to tend towards a Gaussian, Figure 6 shows various 'true' prior distributions at different time steps, concurrently with the full posterior and 20 particle representation. It is evident that not only does the equivalent weight particle filter ensure a good support in comparison to the full posterior distribution for a Gaussian prior, but also for skewed or multi-modal priors.

3.6.2. Increasing the percentage of particles retained

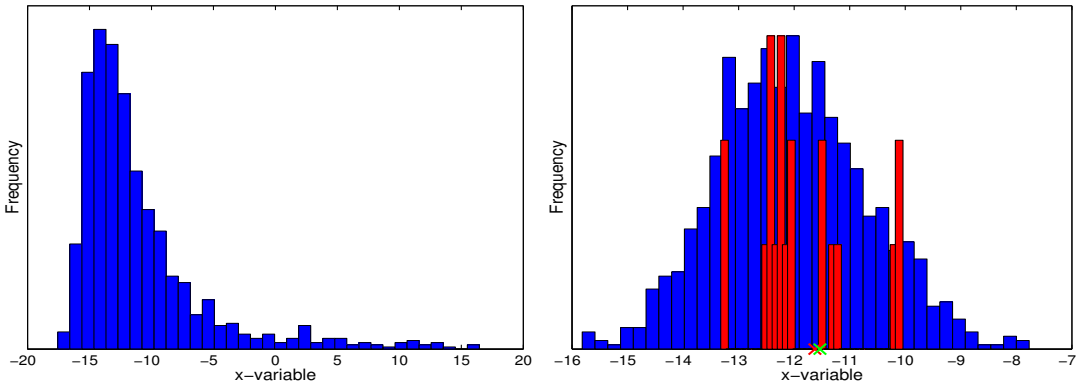
Although these results are very promising, the percentage of particles retained has an impact on the ability of the equivalent weights particle filter to effectively represent the posterior pdf. The higher the percentage of particles included under the equivalent weights scheme, the further each individual particle will have to move from its deterministic position to ensure its weight matches that specified. This effect is demonstrated in Figure 7, which shows the posterior pdfs using 20 particles for time step 1881 as the percentage of particles kept is increased. At



(a) *Time step 801 - prior* (b) *Time step 801 - posterior*



(c) *Time step 1041 - prior* (d) *Time step 1041 - posterior*

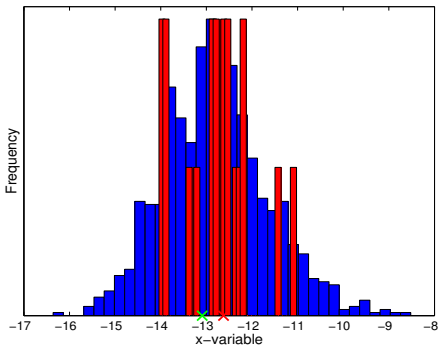


(e) *Time step 1481 - prior* (f) *Time step 1481 - posterior*

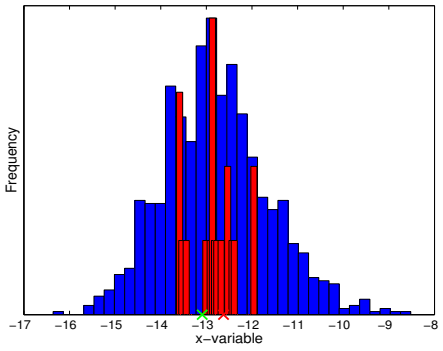
Figure 6. The prior and posterior pdfs at three different timesteps where observations of the x-variable only have been used. The blue bars represent the 'true' pdf and are calculated using 1000 particles and the SIR filter. The red bars on the posterior are from the equivalent weights particle filter using just 20 particles and with 80% of particle retained under equivalent weights. Regardless of the shape of the 'true' prior pdf, the equivalent weights particle filter is able to match the support of the full posterior pdf.

70-80%, the equivalent weights particle filter provides a relatively good match to the posterior pdf of the SIR filter. However as the percentage increases to 100% the effect of the larger movement is seen in the reduced numbers of particles close to the observation. This verifies that making all weights equal to that of the worst particle, and hence retaining 100% of particles, does not in general lead to the best representation of the posterior. Including less than 70% of particles has not currently been examined since the idea is to ensure equivalent weights for the greatest possible number of particles.

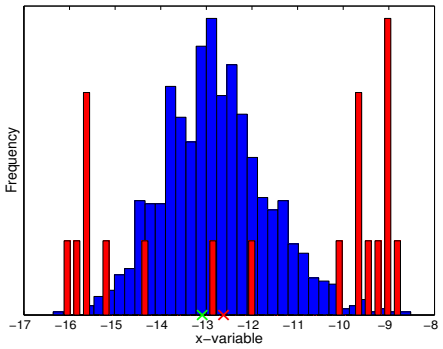
Table I shows the statistics of the true posterior pdf compared to the posterior pdfs produced by the equivalent weights particle filter at the same time step. As would be expected, the statistics verify the results seen in the Figure 7. In particular the larger movement by the particles, when



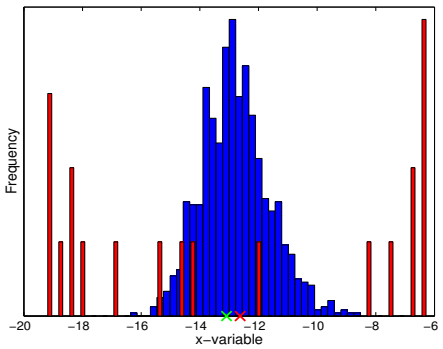
(a) 70% kept



(b) 80% kept



(c) 90% kept



(d) 100% kept

Figure 7. *Posterior pdfs at time step 1881 with one observation for an increasing percentage of particles kept in the equivalent weights scheme. The blue bars come from a 1000 particle run with the SIR filter and the red bars are from the equivalent weights scheme with 20 particles. As the percentage of particles kept reaches 100% the pdf becomes less similar to the 1000 particle run due to the increased movement of the particles.*

	mean	var	skewness	kurtosis
Truth	-12.78	1.32	0.33	3.25
70%	-12.90	0.43	0.11	2.68
80%	-12.95	0.58	-0.48	2.85
90%	-12.37	11.58	-0.18	1.52
100%	-12.24	20.03	0.04	1.43

Table I. The statistics of the posterior pdfs at time step 1881 with one observation and 20 particles for an increasing percentage of particles kept in the equivalent weights scheme. Although the mean stays fairly stationary, the variance increases and the kurtosis decreases as the percentage of particles retained moves towards 100%.

a higher percentage are retained, can be seen through the increase in variance and decrease in kurtosis whilst the mean stays fairly stationary.

The additional movement caused by retaining a greater percentage of particles can also be seen over multiple time steps in the rank histograms, which score where the truth ranks in the ensemble (Figure 8). The standard and relaxation only particle filter schemes clearly have a U-shape, indicating that the ensemble is under-dispersive. When 70% of particles are kept through equivalent weights a U-shape is still clearly evident. As the percentage of particles retained is increased the ensemble becomes more uniform. When 100% are kept the histogram has a humped shape indicating possible over spreading of the ensemble. As the percentage of particles that are kept increases, the ensemble becomes more likely to split around the observation (Figure 7) and hence the truth. This is reflected over multiple time steps by the over-dispersion seen in the rank histograms.

The results shown here over all percentages might initially be considered to be in contrast to those given in Figure 7, which shows a good spread of the ensemble around the truth, regardless of the percentage of particles retained. However Figure 7 shows just one time step specifically chosen to demonstrate the movement by particles, whereas Figure 8 is a measure over 80,000 time steps and shows a more generalised trend.

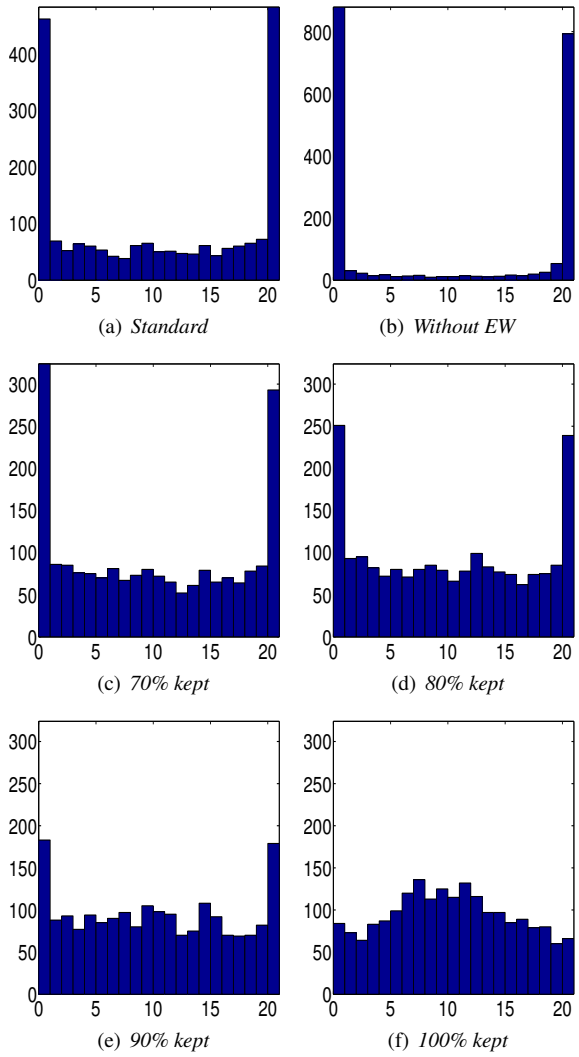


Figure 8. *Rank histograms for the x-variable with one observation run over 80,000 time steps for different Particle filter schemes and compared to the truth. Both the standard and relaxation only particle filters are under dispersive as shown by the distinctive U shape. Although this shape can still be seen for the equivalent weights scheme with 70% of particles kept, as the percentage of particles increases the spread improves until with 100% the truth is more likely to fall in the centre of the ensemble.*

3.6.3. Increasing the size of the ensemble

As the size of the ensemble is increased, the conclusions drawn on the effect of retaining a greater percentage of particles change very little. Since all particles relax towards the observation, increasing the number of ensemble members does not lead to the same increase in variance as would be expected when using the SIR filter. This can be seen pictorially in Figure 9, which shows the posterior pdfs when 80% of particles are retained under the equivalent weights step as the number of particles is increased, again

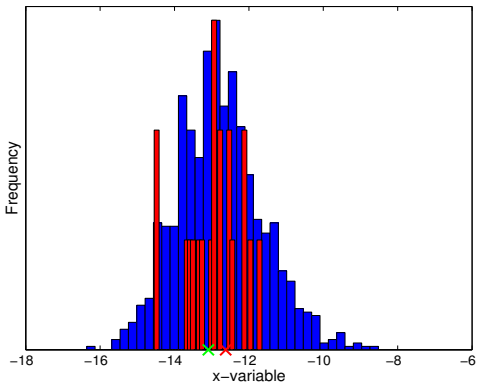
%	N	mean	var	skewness	kurtosis
	Truth	-12.78	1.32	0.33	3.25
70	20	-12.90	0.43	0.11	2.68
	100	-12.82	1.09	1.76	7.48
	1000	-12.90	0.64	0.88	5.23
80	20	-12.95	0.58	-0.48	2.85
	100	-12.48	1.66	2.07	7.73
	1000	-12.77	0.79	1.47	7.34
90	20	-12.37	11.58	-0.18	1.52
	100	-12.11	1.89	2.19	8.55
	1000	-11.55	5.92	-0.82	2.15
100	20	-12.24	20.03	0.04	1.43
	100	-11.06	29.24	-0.48	1.44
	1000	-11.89	40.41	-0.12	1.18

Table II. The statistics of the posterior pdfs at time step 1881 with one observation and 20 particles for an increasing percentage and number of particles in the equivalent weights scheme. No discernible pattern is created by increasing the number of particles (N). The biggest effect on the statistics of the equivalent weights posterior pdfs is still from changing the percentage of particles retained.

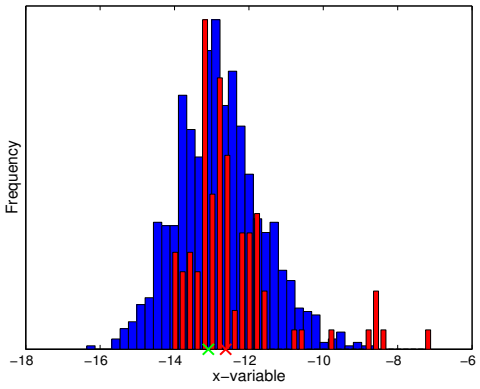
at time step 1881. Studying the statistics for the same time step as both the percentage and number of particles is increased (Table II) gives a similar pattern. Apart from the 100 particle run with 90% of particles retained, the significant changes in variance are seen as the percentage of particles is increased rather than with greater ensemble size across a particular percentage of particles. Less clear is the reduction in kurtosis with percentage rather than number of particles, although this is at a minimum when 100% of particles are retained. Interestingly the mean stays relatively consistent regardless of percentage or ensemble size. This is evident not just for time step 1881, but also over all times steps and all variables by considering the root mean square errors (Table III). Although the root mean square errors increase when 100% of particles are retained, there is no discernible pattern for 70-90% of particles or as the the ensemble size gets larger.

3.6.4. Changing the statistics, frequency and number of observations

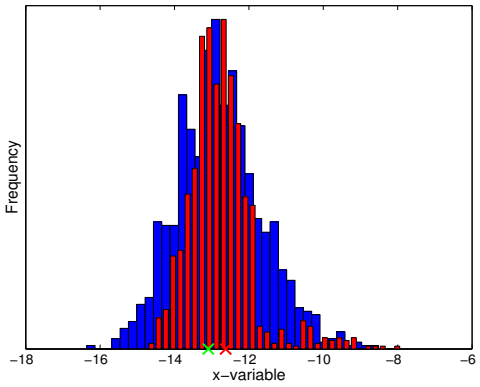
The majority of the discussion in the previous sections has focussed on the effect changing the percentage of particles retained has on the posterior pdf, since the trajectories of the particles remain unaffected. Changing the frequency



(a) 20 particles



(b) 100 particles



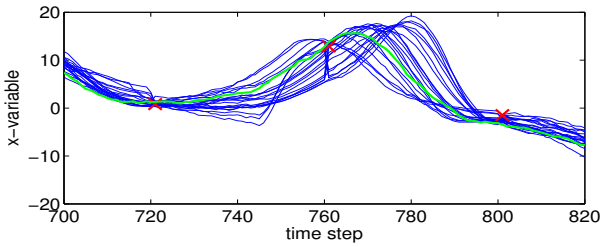
(c) 1000 particles

Figure 9. *Posterior pdfs at time step 1881 with one observation for an increasing number of ensemble members. The blue bars come from a 1000 particle run with the SIR filter and the red bars are from the equivalent weights scheme with 80% of particles kept.*

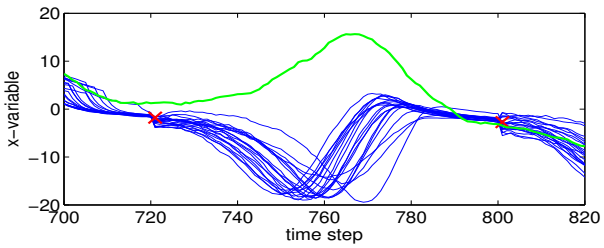
N	70%	80%	90%	100%
20	20.44	18.35	20.90	23.85
100	20.70	19.30	20.02	22.46
1000	19.15	19.44	18.70	25.41

Table III. Root mean square error over a 2000 model time step run comparing all three variables to the truth generated as part of the twin experiment. For comparison the rmse of the 1000 SIR filter truth distribution is 15.98. No pattern is apparent as the percentage retained or number of particles is increased, although retaining 100% of particles does lead to a worse rmse.

of observations, however, has the largest impact on the trajectories of the particles. Doubling the distance between observations from 40 to 80 time steps means that there is now only one or two observations per cycle of the Lorenz 63 system. If one observation has a slight but significant perturbation from the truth at the start of a cycle, then the particles can all move to the opposing wing of the butterfly. The lack of observations over the remaining duration of the cycle can lead to the particles remaining separated from the truth until the next observation (Figure 10).



(a) Observations every 40 timesteps



(b) Observations every 80 timesteps

Figure 10. The trajectories of 20 particles (blue) compared to the true trajectrory (green) with observations (red) every 40 (a) and every 80 (b) timesteps. If observations are available every 40 timesteps then the particle trajectories are able to follow the truth whereas with observations every 80 timesteps the particles miss the true transition to the opposite side of the Lorenz attractor.

Returning to the posterior pdfs, it is evident that changing the frequency of the observations has little effect on the statistics of the posterior pdfs (Figure 11). There are larger differences in variance between the true distribution and the equivalent weights ensemble posterior as the frequency of observations increases, specifically there are more cases of the ensemble variance being larger than than the true variance, but no obvious relationship emerges. The differences in mean, skewness and kurtosis are also clearly effected by a greater number of timesteps between observations, however again no relationship is evident.

Since the relaxation proposal density ensures the particles end up close to the observations, regardless of their frequency, the ability of the equivalent weights particle filter to replicate the true posterior pdf does not appear to be related to the number of timesteps between observations. Increasing the percentage of particles retained under the equivalent weights step still leads to the same increase in variance and distinctive splitting behaviour demonstrated in Figure 7, regardless of the frequency of observations.

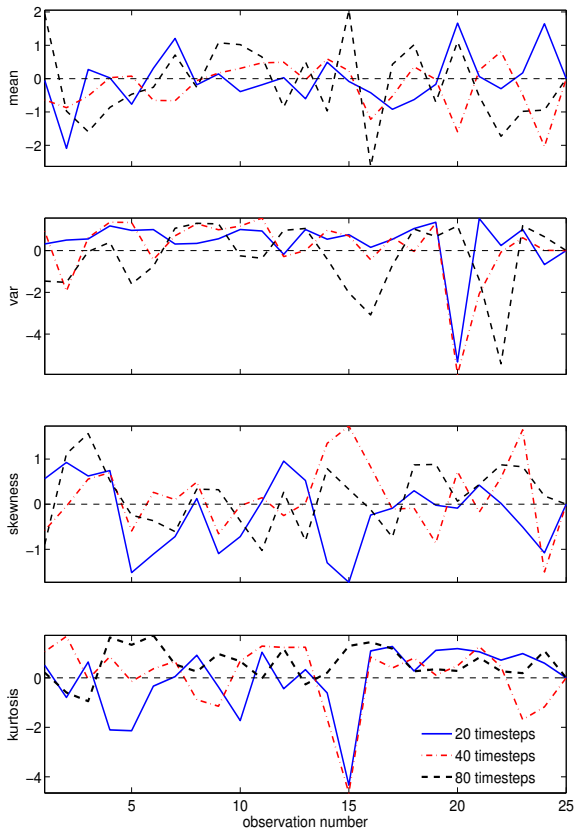


Figure 11. The mean, variance, skewness and kurtosis of the true distribution minus the ensemble distribution over 25 observation times for observations every 20, 40 and 80 timesteps. No clear relationship is seen between the number of timesteps between observations and the ability of the equivalent weights particle filter to replicate the true distribution.

Changing the size of the standard deviation of the observations also appears to bear little relationship with the ability of the equivalent weights to represent the posterior. Since we return to observations every 40 time steps and R plays no part in the relaxation proposal density, the trajectories of the particles are unaffected by any changes to observation variance. Similarly changing the variance in R has no effect on the equivalent weights proposal density, the value of α_i simply changes to compensate for the reduction or increase in observation variance and no change is seen in x_i^*. There is a slight effect to the value of C, which each particle is trying to achieve (Eq. (30)), but this was not noted to have a significant impact on the posterior distribution. There was some evidence to suggest that retaining a greater percentage of particles provides a closer match to the increased true variance of the posterior as a result of a higher observation variance. This would be due to the increased movement, and hence ensemble variance, as the percentage of particles kept is increased. However since the link is tenuous and the splitting behaviour associated with retaining 100% of particles was still observed at a number of observation time steps, this is not discussed further.

Increasing the number of variables that are being observed has a much more noticeable effect on the dispersiveness of the posterior pdf, both in terms of the variances for individual posterior pdfs and for the rank histograms over multiple time steps. With increasing numbers of observed variables, the variance at individual time steps is consistently closer to the variance of the true posterior pdf over the first 50 observation times (Figure 12). This is true regardless of the percentage of particles retained, although keeping 100% of particles still leads to a greater ensemble variance compared to the true variance at an increased number of time steps when all three variables are observed. Over 80,000 time steps this reduction in variance leads to a U-shape in the rank histograms regardless of the percentage of particles retained. This supports the arguments made in Section 3.4, that increasing the number of variables observed decreases the movement required by each individual particle which is encouraging in relation to balances. It does, however, imply that the ensemble becomes less dispersive with increasing numbers of observations. One possible method of increasing the spread of the ensemble would be to change

the proposal density at the majority of time steps. For example the relaxation term could be reduced or the random error $d\beta^j$ increased.

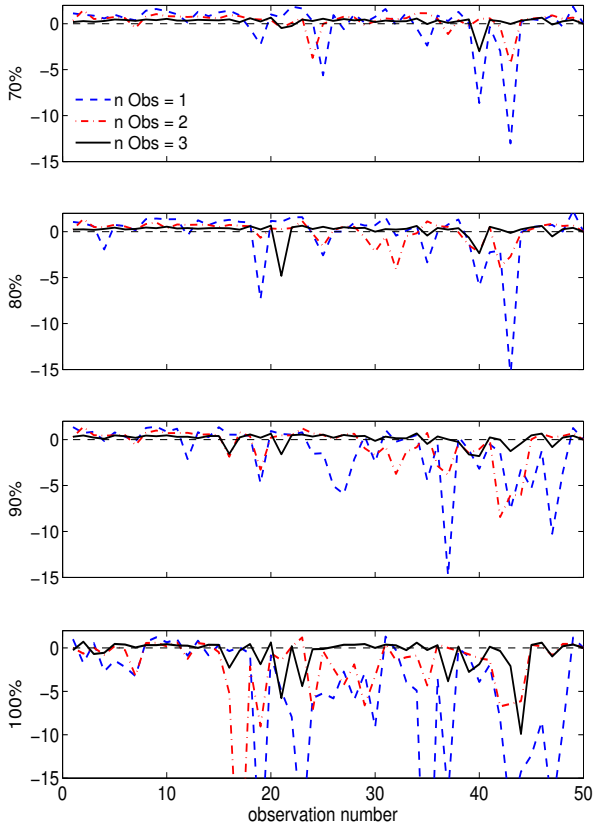


Figure 12. The 'true' variance from the SIR filter using 1000 particles minus the variance from the 20 particle equivalent weights particle filter as both the percentage of particles retained and the number of variables observed are increased. The black dashed zero line shows where the 'true' and 'ensemble' variance are the same, points above this imply the ensemble is less variable than the truth and below the ensemble is more variable. As the percentage of particles retained is increased, the equivalent weights particle filter ensemble tends to become more variable than the truth regardless of the number of variables observed however observing three variables consistently leads to less variance. In order to keep the scales the same and retain the detail in the lower percentage cases, some of the peaks have been omitted in the 100% case.

3.6.5. Discussion

In this section of the paper the performance of the equivalent weights particle filter is examined in detail as the various parts of the scheme are changed. The necessity of choosing the equivalent weights proposal density as a mixture density, rather than a Gaussian, is demonstrated as well as the benefit of choosing the positive version of α_i. The change in the trajectories of the particles from using a simple $B(\tau)$ to a more complex one is also shown. It is argued that, although the choices made for these parts of the equivalent weights particle filter need to be explored in more detail in higher dimensional models, the conclusions drawn are likely to be generic regardless of the model to which the scheme is applied.

The percentage of particles retained under the equivalent weights proposal density is expected to be much more dependent on the model and observation specifics. For the Lorenz 63 system of equations, the percentage of particles retained has the greatest effect on the variance of posterior pdfs regardless of the number of ensemble members or the frequency or variance of observations. The effect is only reduced when the number of variables observed is increased.

It is clear with the Lorenz 63 system that if the main objective is to represent the posterior pdf using just a few particles then it is preferable to ensure equivalent weights for only 70-80% of particles. However there is also an argument for including 100% of particles. If 100% of particles are kept then all particles have equivalent weights at observation time. This removes the need for resampling, since there will no longer be particles with small weights to be abandoned. As a result the posterior pdf, $p(x^{n-r:n}|y^n)$, no longer needs to be based only on the model trajectories between time steps and we can now return to the full posterior $p(x^{0:n}|y^{1:n})$ based on whole model trajectories and all observations. This provides an incentive for keeping all particles despite the negative impact on representing the posterior pdf. Since a greater number of observed variables leads to less movement, it may be that in a higher dimensional model 100% of particles can be retained without negatively impacting the posterior pdf.

4. Conclusion and Discussion

The equivalent weights particle filter was introduced in Van Leeuwen (2010) where its efficiency in capturing the truth with only minimal particles in the Lorenz-63 model was

demonstrated. This article explores the behaviour of the scheme in more depth. It looks at how the proposal density is used in the equivalent weights particle filter to ensure that equally significant particles are sampled from the region of the posterior local to the observations. It is shown how the equivalent weights particle filter is similar to both the optimal proposal density and implicit particle filters through its use of proposal densities. However the optimal proposal density, and equivalently the implicit particle filter under certain conditions, will still suffer from filter degeneracy when a large number of observations are present. The equivalent weights particle filter avoids this relationship and ensures equivalent weights by sampling from a specific mixture density.

The ability of the equivalent weights particle filter to represent the posterior distribution at individual time steps is shown using the Lorenz-63 model. Provided that certain choices are made as part of the mixture density, then using just 20 particles, the scheme is able to not only capture the mean of the posterior but also some of the spread of the distribution. Examining the behaviour of the scheme as the number of particles is increased and as the number, frequency and variance of the observations is changed, shows little or no relationship to the ability of the equivalent weights particle filter to represent the posterior pdf. Much more significant is the effect of varying the percentage of particles for which equivalent weights are assured. Increasing the percentage of particles has a noticeable impact on the representation of the posterior and the ensemble statistics. This allows tuning of the scheme when the truth is known but will lead to uncertainties when the true state of the system is unavailable.

One considerable benefit of the scheme is the ease with which it can be implemented in large-dimensional problems. We believe that the sensitivity seen in the Lorenz-63 system will not be so apparent as the dimension increases. The equivalent weights particle filter ensures equally significant particles by reducing the weight of some particles until the specified percentage have equivalent weights. The reduction in weight is achieved by changing the model states of the particles. In a higher-dimensional system the movement by each particle can be distributed over a larger number of dimensions and so less impact will be seen in individual variables. This will lead to a smaller impact on the representation of the posterior distribution. Work is currently being carried out to investigate this with the barotropic vorticity equation solved over a large (approx 65,500) dimensional grid.

References

Anderson JL. 2001. An ensemble adjustmant filter for data assimilation. *Monthly Weather Review* **129**: 2884–2903.

Bishop CH, Etherton BJ, Majumdar SJ. 2001. Adaptive sampling with the ensemble transform kalman filter. part i: Theoretical aspects. *Monthly Weather Review* **129**: 420–436.

Bocquet M, Pires CA, Wu L. 2010. Beyond gaussian statistical modeling in geophysical data assimilation. *Monthly Weather Review* **138**: 2997–3023.

Buehner M, Houtekamer P, Charette C, Mitchell HL, He B. 2010. Intercomparison of variational data assimilation and the ensemble kalman filter for global deterministic nwp. part i: description and single-observation experiments. *Monthly Weather Review* **138**: 1550–1566.

Burgers G, van Leeuwen PJ, Evensen G. 1998. Analysis scheme in the ensemble kalman filter. *Monthly Weather Review* **126**(6): 1719–1724.

Chorin AJ, Morzfeld M, Tu X. 2010. Interpolation and iteration for nonlinear filters. *Communications in Applied Mathematics and Computational Science* **5**: 221–240.

Chorin AJ, Tu X. 2009. Implicit sampling for particle filters. *Proceedings of the National Academy of Sciences* **106**(41): 17 249–17 254.

Doucet A, de Freitas N, Gordon N. 2001. *Sequential monte-carlo methods in practice*. Springer-Verlag.

Doucet A, Godsill S, Andrieu C. 2000. On sequential monte carlo sampling methods for bayesian filtering. *Statistics and Computing* **10**: 197–208.

Evensen G. 1994. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error

statistics. *Journal of Geophysical Research* **99**: 10 143–10 162.

Gordon NJ, Salmond DJ, Smith AF. 1993. Novel approach to nonlinear/non-gaussian bayesian state estimation. *IEE proceedings* **140**: 107–113.

Hamill TM. 2000. Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review* **129**: 550–560.

Houtekamer PL, Mitchell HL. 2005. Ensemble kalman filtering. *Quarterly Journal of the Royal Meteorological Society* **131**: 3269–3289.

Kitagawa G. 1996. Monte-carlo filter and smoother for non-gaussian non-linear state-space models. *Journal of Computational and Graphical Statistics* **10**: 253–259.

Lui JS, Chen R. 1998. Sequential monte-carlo methods for dynamical systems. *Journal of the American Statistical Association* **90**: 567–576.

Morzfeld M, Tu X, Atkins E, Chorin AJ. 2012. A random map implementation of implicit filters. *Journal of Computational Physics* **231**: 2049–2066.

Snyder C, Bengtsson T, Bickel P, Anderson J. 2008. Obstacles to high-dimensional particle filtering. *Monthly Weather Review* **136**: 4629–4640.

Talagrand O, Courtier P. 1987. Variational assimilation of meteorological observations with the adjoint voriticity equation. i. theory. *Quarterly Journal of the Royal Meteorological Society* **113**: 1311–1328.

Van Leeuwen PJ. 2009. Particle filtering in geophysical systems. *Monthly Weather Review* **137**: 4089–4114.

Van Leeuwen PJ. 2010. Nonlinear data assimilation in geosciences: an extremely efficient particle filter. *Quarterly Journal of the Royal Meteorological Society* **136**: 1991–1999.

Van Leeuwen PJ. 2011. Efficient nonlinear data-assimilation in geophysical fluid dynamics. *Computers and Fluids* **46**: 52–58.

Whitaker JS, Hamill TH. 2002. Ensemble data assimilation without perturbed observations. *Monthly Weather Review* **130**: 1913–1923.

A. The variance of weights using the optimal proposal density

Using the optimal proposal density, the weights are given by

$$w_i = A \exp\left[-\frac{1}{2}(y^n - Hf(x_i^{n-1}))(HQH^T + R)^{-1} \times (y^n - Hf(x_i^{n-1}))\right]. \quad (40)$$

We can expand $y^n - Hf(x_i^{n-1})$ to $y^n - Hx_t^n + H(x_t^n - f(x_i^{n-1}))$, where x_t^n is the true state at time n. We can also write $x_t^n = f(x_t^{n-1}) + d\beta_t^n$ to give $y^n - Hx_t^n + Hd\beta_t^n + H(f(x_t^{n-1}) - f(x_i^{n-1}))$. To obtain an order of magnitude estimate for the variance in the weights, we also assume that both the observation R and model errors HQH^T are uncorrelated, with variances V_y and V_x respectively, to give

$$-\log(w_i) = \frac{1}{2(V_x + V_y)} \sum_{j=1}^{M} \left[y_j^n - H_j x_t^n + H_j d\beta_t^n + H_j(f(x_t^{n-1}) - f(x_i^{n-1}))\right]^2 \quad (41)$$

We are interested in the variance as the ensemble member index i changes. To make the notation simpler, we group the constants y_j^n, $H_j x_t^n$ and $H_j d\beta_t^n$ into the variable α_j. We also assume that $f(x_t^{n-1})$ is linear and so represent the model equations by the matrix A. Hence we are interested in the variance of

$$-\log(w_i) = \frac{1}{2(V_x + V_y)} \sum_{j=1}^{M} \left[\alpha_j + H_j A(x_t^{n-1} - x_i^{n-1})\right]^2. \quad (42)$$

If we assume $(x_t^{n-1} - x_i^{n-1})$ is Gaussian distributed then (45) is close to a non-central χ^2_M distribution. To make it non-central χ^2_M distributed requires normalisation by the variance of $\left[\alpha_j + H_j A(x_t^{n-1} - x_i^{n-1})\right] =$

$H_j A P^{n-1} A^T H_j^T = (P_x^n)_j$:

$$-\log(w_i) = \frac{P_x^n}{2(V_x+V_y)} \sum_{j=1}^{M} \frac{\left[\alpha_j + H_j A(x_t^{n-1} - x_i^{n-1})\right]^2}{(P_x^n)_j} \quad (43)$$

where P^{n-1} is the variance of the ensemble at time $n-1$. Using the properties of non-central χ^2_M distributions the variance of $-\log(w_i)$ is now given by $a^2 2(M+2\lambda)$ where $a = P_x^n/2(V_x+V_y)$ and $\lambda = (\sum_{j=1}^M \alpha_j^2)/P_x^n$. We have that $\sum_{j=1}^M \alpha_j^2 = \sum_{j=1}^M \left[(y_j^n - H_j x_t^n) + H_j d\beta_t^n)\right]^2$. The observations y^n are given and so the estimate of variance depends on these specific observations. However, since the observations are drawn from a Gaussian with covariance R, for a large enough number of observations we can approximate the realisation of $\sum_{j=1}^M (y_j^n - H_j x_t^n)^2$ with MV_y and expect that $\sum_{j=1}^M (y_j^n - H_j x_t^n)$ to be zero on average. Similarly $\sum_{j=1}^M (H_j d\beta_t^n)^2 = MV_x$. Hence $\lambda = (MV_y + MV_x)/P_x^n$ and the full variance of $-\log w_i$ is given by:

$$\frac{M}{2}\left(\frac{P_x^n}{V_x+V_y}\right)^2 \left(1 + 2\left(\frac{V_y+V_x}{P_x^n}\right)\right). \quad (44)$$

B. Derivation of dominant terms when sampling from a Gaussian under the equivalent weights step

Let $x_i^* = f(x_i^{n-1}) + \alpha_i K(y^n - Hf(x_i^{n-1}))$, then it satisfies:

$$-\log w_i^{rest} + \frac{1}{2}(x_i^* - f(x_i^{n-1}))^T Q^{-1}(x_i^* - f(x_i^{n-1})) + \frac{1}{2}(y^n - Hx_i^*)^T R^{-1}(y^n - Hx_i^*) = C. \quad (45)$$

Stochastic error can now be added to this value so that

$$x_i^n = f(x_i^{n-1}) + \alpha_i K(y^n - Hf(x_i^{n-1})) + d\tilde{\beta}_i^n = x_i^* + \gamma d\beta_i^n \quad (46)$$

where $\gamma d\beta_i^n = d\tilde{\beta}_i^n \sim N(0, \gamma^2 Q)$ with γ controlling the width of the variance and $d\beta_i^n \sim N(0,Q)$. Then minus the logarithm of the final weights at observation time is given by

$$\begin{aligned}
&-\log w_i \\
&= -\log\left(\frac{w_i^{rest} p(x_i^n|x_i^{n-1}) p(y^n|x_i^n)}{q(x_i^n|x_i^{n-1}, y^n)}\right) \\
&\propto -\log w_i^{rest} \\
&\quad + \frac{1}{2}((x_i^* + \gamma d\beta_i^n) - f(x_i^{n-1}))^T Q^{-1} \\
&\qquad \times ((x_i^* + \gamma d\beta_i^n) - f(x_i^{n-1})) \\
&\quad + \frac{1}{2}(y^n - H(x_i^* + \gamma d\beta_i^n))^T R^{-1}(y^n - H(x_i^* + \gamma d\beta_i^n)) \\
&\quad - \frac{1}{2}(x_i^n - x_i^*)^T(\gamma^2 Q)^{-1}(x_i^n - x_i^*) \\
&\propto -\log w_i^{rest} \\
&\quad + \frac{1}{2}((x_i^* - f(x_i^{n-1})) + \gamma d\beta_i^n)^T Q^{-1} \\
&\qquad \times ((x_i^* - f(x_i^{n-1})) + \gamma d\beta_i^n) \\
&\quad + \frac{1}{2}(y^n - Hx_i^* - \gamma H d\beta_i^n)^T R^{-1} \\
&\qquad \times (y^n - Hx_i^* - \gamma H d\beta_i^n) \\
&\quad - \frac{1}{2}(\gamma d\beta_i^n)^T(\gamma^2 Q)^{-1}(\gamma d\beta_i^n)
\end{aligned} \quad (47)$$

$$\begin{aligned}
&\propto C + \frac{1}{2}(x_i^* - f(x_i^{n-1}))^T Q^{-1}\gamma d\beta_i^n \\
&\quad + \frac{1}{2}\gamma d\beta_i^{nT} Q^{-1}(x_i^* - f(x_i^{n-1})) + \frac{1}{2}\gamma^2 d\beta_i^{nT} Q^{-1} d\beta_i^n \\
&\quad - \frac{1}{2}(y^n - Hx_i^*)^T R^{-1}\gamma H d\beta_i^n \\
&\quad - \frac{1}{2}(\gamma H d\beta_i^n)^T R^{-1}(y^n - Hx_i^*) \\
&\quad + \frac{1}{2}(H\gamma d\beta_i^n)^T R^{-1}(H\gamma d\beta_i^n) - \frac{1}{2}(d\beta_i^n)^T Q^{-1}(d\beta_i^n).
\end{aligned} \quad (48)$$

γ controls the width of the variance and so is chosen to be very small. Hence all terms including γ do not significantly contribute to $-\log w_i$ and the dominant terms are

$$-\log w_i \approx C - \frac{1}{2} d\beta_i^{nT} Q^{-1} d\beta_i^n \quad (49)$$

C. Scalar analysis of α_i

If there is only one observation and so y^n is a scalar rather than a vector, then R consists simply of the singular value V_y, HQH^T is V_x and $d_i = y^n - Hf(x_i^{n-1})$ is also a scalar. Hence

$$\begin{aligned} b_i &= 0.5 d_i^T R^{-1} d_i - C - \log w_i^{rest} \\ &= \frac{d_i^2}{2V_y} - C - \log w_i^{rest} \end{aligned} \tag{50}$$

and

$$\begin{aligned} a_i &= 0.5 d_i^T R^{-1} HK d_i \\ &= 0.5 d_i^T R^{-1} HQH^T (HQH^T + R)^{-1} d_i \\ &= \frac{d_i^2 V_x}{2V_y(V_y + V_x)}. \end{aligned} \tag{51}$$

In order for $\alpha_i = 1 \pm \sqrt{1 - b_i/a_i}$ to have real roots requires

$$\begin{aligned} 1 - \frac{b_i}{a_i} &\geq 0 \\ \implies a_i &\geq b_i \\ \implies \frac{d_i^2 V_x}{2V_y(V_y + V_x)} &\geq \frac{d_i^2}{2V_y} - C - \log w_i^{rest} \\ \implies C - \left(-\log w_i^{rest}\right) &\geq \frac{(V_x + V_y) di^2 - d_i^2 V_x}{2V_y(V_x + V_y)} \\ \implies C - \left(-\log w_i^{rest}\right) &\geq \frac{di^2}{2(V_x + V_y)} \end{aligned} \tag{52}$$

In addition if $\alpha_i = 1 - \sqrt{1 - b_i/a_i}$ is going to be positive then $\sqrt{1 - b_i/a_i}$ must be less than one which implies that $b_i > 0$, since we already know from above that $a_i \geq b_i$.

$$\begin{aligned} 0 &< b_i \\ \implies 0 &< \frac{d_i^2}{2V_y} - C - \log w_i^{rest} \\ \implies C - \left(-\log w_i^{rest}\right) &< \frac{d_i^2}{2V_y} \end{aligned} \tag{53}$$

So for the negative version of α_i to be both positive and real requires

$$\frac{di^2}{2(V_x + V_y)} \leq C - \left(-\log w_i^{rest}\right) < \frac{d_i^2}{2V_y} \tag{54}$$

whereas for the positive version to be positive and real simply requires

$$\frac{di^2}{2(V_x + V_y)} \leq C - \left(-\log w_i^{rest}\right). \tag{55}$$

D. Relative values of γ_U and γ_N

Sampling from a mixture density implies that a particle can be sampled either from a Gaussian distribution or a uniform distribution. In general, with particle filters, the normalisation factors of the distributions are not considered as they are the same for every particle and hence cancel in the normalisation of the particles. However if particles are sampled from two different densities, with two different normalisation factors, then this needs to be accounted for in the weights. The weights at observation time are given by

$$w_i = \frac{w_i^{rest} p(x_i^n | x_i^{n-1}) p(y^n | x_i^n)}{q(x_i^n | x_i^{n-1}, y^n)}. \tag{56}$$

If a particle is sampled from the Gaussian part of the proposal density $q(x_i^n | x_i^{n-1}, y^n)$ then its weight is given by

$$w_i = \frac{w_i^{rest} p(x_i^n | x_i^{n-1}) p(y^n | x_i^n)}{\frac{\epsilon}{(2\pi)^{k/2} |\gamma_N^2 Q|^{1/2}} \exp(-\frac{1}{2} \gamma_U d\beta_i^n (\gamma_U^2 Q)^{-1} \gamma_U d\beta_i^n)} \tag{57}$$

whereas if it is sampled from the Uniform part its weight is

$$w_i = \frac{|Q|^{1/2} (2\gamma_U)^k}{1 - \epsilon} w_i^{rest} p(x_i^n | x_i^{n-1}) p(y^n | x_i^n) \tag{58}$$

Ignoring both $w_i^{rest} p(x_i^n | x_i^{n-1}) p(y^n | x_i^n)$ and $\exp(-\frac{1}{2} \gamma_U d\beta_i^n (\gamma_U^2 Q)^{-1} \gamma_U d\beta_i^n)$, since variations in these values are due to the samples chosen and so are uncontrollable, we instead focus on the normalisation constants for the different densities. It can be assumed that

the majority of particles will be sampled from the Uniform distribution since we choose the value of ϵ to ensure this. Hence we can divide the weight of all particles by the normalisation constant $\frac{|Q|^{1/2}(2\gamma_U)^k}{1-\epsilon}$ so that it no longer needs to be calculated for the majority of particles. This leads to particles sampled from the Gaussian now having weight due to the normalisation constants given by

$$\frac{(2\pi)^{k/2}|\gamma_N^2 Q|^{1/2}}{\epsilon}\frac{(1-\epsilon)}{|Q|^{1/2}(2\gamma_U)^k}. \tag{59}$$

Ideally we want this value to be less than or at least equal to one to negate the additional weight a particle has from $\exp(-\frac{1}{2}\gamma_U d\beta_i^n(\gamma_U^2 Q)^{-1}\gamma_U d\beta_i^n)$, since this has the potential to cause filter degeneracy. For Eq. (62) to equal one requires

$$\gamma_N = \frac{2^{k/2}\epsilon}{\pi^{k/2}(1-\epsilon)}\gamma_U^k. \tag{60}$$

Choosing γ_N smaller than this value may solve the issues surrounding sampling from the Gaussian distribution. However we have not explored this any further since we assume that ϵ is chosen small enough that it is highly unlikely that a particle will be sampled from the Gaussian part of the mixture density.

