Author's Accepted Manuscript
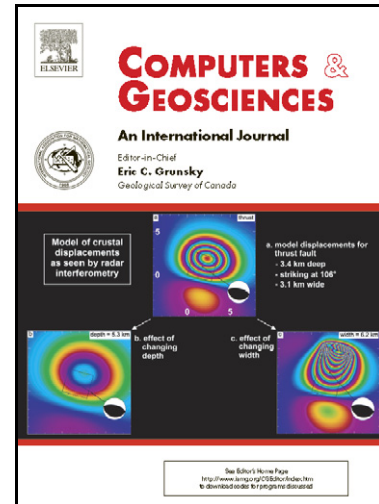
Efficient fully nonlinear data assimilation for geophysical fluid dynamics

Peter Jan van Leeuwen, Melanie Ades

Cite this article as: Peter Jan van Leeuwen and Melanie Ades, Efficient fully nonlinear data assimilation for geophysical fluid dynamics, *Computers & Geosciences,* http://dx.doi.org/10.1016/j.cageo.2012.04.015

Efficient fully nonlinear data assimilation for geophysical fluid dynamics

Peter Jan van Leeuwen and Melanie Ades

Data Assimilation Research Centre, Department of Meteorology, University of Reading, PO Box 243, Reading, RG6 6BB, United Kingdom

Abstract

A potential problem with Ensemble Kalman Filter is the implicit Gaussian assumption at analysis times. Here we explore the performance of a recently proposed fully nonlinear particle filter on a high-dimensional but simplified ocean model, in which the Gaussian assumption is not made. The model simulates the evolution of the vorticity field in time, described by the barotropic vorticity equation, in a highly nonlinear flow regime. While common knowledge is that particle filters are inefficient and need large numbers of model runs to avoid degeneracy, the newly developed particle filter needs only of the order of 10-100 particles on large scale problems. The crucial new ingredient is that the proposal density cannot only be used to ensure all particles end up in high-probability regions of state space as defined by the observations, but also to ensure that most of the particles have similar weights. Using identical twin experiments we found that the ensemble mean follows the truth reliably, and the difference from the truth is captured by the ensemble spread. A rank histogram is used to show that the truth run is indistinguishable from any of the particles, showing statistical consistency of the method.

Key words: Data Assimilation, Inverse Modeling, Particle Filter, Ensemble Kalman Filter

1. Introduction

26   Numerical models for simulation and prediction of the evolution of systems in
27 the geosciences are becoming ever more complex. While relatively simple linear
28 balances tend to dominate the systems at large scales, with increasing resolution
29 more and more nonlinear processes are involved. Furthermore, with the coupling
30 of many physical, chemical and biological systems extremely complex behaviour
31 with highly nonlinear feebacks has to be simulated.

32   To the extent that these flows are initial value problems our incomplete
33 knowledge of the exact initial conditions leads to incomplete knowledge of the
34 evolution of the system. This forces us to think in terms of uncertainty, which
35 can be described in probabilistic terms. The evolution equations for the related
36 probability densities have been known for decades (see e.g. Jazwinski, 1970).
37 If the system is Markovian, our present knowledge of the system in the form of
38 a probability density function evolves according to the Kolmogorov or Fokker-
39 Plank equation. This theory can be applied for small dimensional systems, but
40 the systems we study in the geosciences are not so.

41   When observations of the system are available, their information on the sys-
42 tem can be incorporated using Bayes Theorem, in which the prior probability
43 density function (pdf from now on), representing our prior knowledge, is mul-
44 tiplied by the likelihood, i.e. the probability density of the observations given
45 a specific model state. This then leads to the so-called posterior pdf, that de-
46 scribes our updated knowledge of the system. This process of updating the prior
47 pdf with observations is called *data assimilation*, and its goal is to determine
48 properties of this posterior pdf. It should be realised that this posterior pdf is
49 unlikely to be ever at our disposal in full because the size of the state space is
50 huge, typically 100 million for numerical weather prediction. We can only infer
51 statistical moments like mean, covariance, percentiles, and modes.

52   It is stressed here that the data-assimilation problem as specified above is a
53 multiplication problem and not an inverse problem: Bayes Theorem (see equa-
54 tion (1)) shows that one has to *multiply* the prior pdf with the likelihood to

55   obtain the posterior pdf. There is no inversion needed to obtain the poste-

56   rior. Also parameter estimation falls in this framework: the prior pdf of the

57   parameters is updated through multiplication with the likelihood to obtain the

58   posterior pdf of the parameters. Obviously, one needs the relation between the

59   parameters and the observations in the likelihood, and that typically involves

60   integrating a full numerical model, but that doesn't make the problem an in-

61   verse problem. The emphasis of this paper is on estimation of the pdf of the

62   model variables represented by a state vector, and not on that of parameters.

63   When the posterior pdf is unimodal or the majority of the posterior proba-

64   bility mass is concentrated around a mode of the posterior pdf it makes sense

65   to concentrate on the mode of the posterior pdf. The problem of finding the

66   mode is usually formulated as an inverse problem, i.e. a problem in which a

67   matrix has to be inverted, although there is no necessity to do so. Examples are

68   variational algorithms that try to find the mode by exploring the gradient of

69   the log of the posterior pdf. In the geosciences these methods are known as e.g.

70   gradient methods, 3DVar, 4DVar (Talagrand and Courtier, 1987), representer

71   method (Bennett, 1992), PSAS (Courtier, 1997), depending on details of the

72   solution method. The Ensemble Kalman filter (Evensen, 1994, Burgers et al.,

73   1998) is slightly different in that it tries to find the posterior mean (the least-

74   squares estimate, which is the mean by definition), but because of the linearity

75   assumptions in the Kalman filterit is assumed implicitly that the mean is close

76   to the mode. This has led to confusion that data assimilation is all about find-

77   ing this mode in the geophysical and the so-called inverse-problem communities,

78   and in some cases hampered progress to more nonlinear multimodal problems.

79   In this paper we propose solutions to highly nonlinear high-dimensional data-

80   assimilation problems. Our stating point is the particle filter (e.g. Gordon et al.,

81   1993), in which an ensemble of model runs is performed, representing our prior

82   knowledge of the system. Each ensemble member, or particle, is weighted with

83   its distance to observations when these become available. The distance norm

84   is determined by the value of the pdf of the observations given this particle,

85   so the likelihood of the observations given this particle. The weights are the

86   relative probabilistic weights of the particles, so e.g. the mean of the ensemble

87   now becomes a weighted mean in the posterior pdf.

88   It is well known that in systems with moderate dimensions, say of order

89   10 and higher, particle filters tend to be degenerate, meaning that the weights

90   vary too much. Typically after one or a few updates with observations the

91   relative weight of one particle is close to one, while that of all others is very

92   close to zero. This means that e.g. a weighted mean is in fact based on only one

93   particle, so all statistical information in the ensemble is lost. To prevent this

94   from happening several methods have been proposed, starting from resampling

95   (Gordon et al., 1993) to more complicated or approximating solutions (see e.g.

96   Doucet et al., 2001, and Van Leeuwen, 2009, for a review of applications in

97   the geosciences). None of the proposed methods is applicable to systems with

98   dimensions larger than say of order 100, without having to need millions of

99   particles, so millions of model integrations. As mentioned, our goal is perhaps

100   100 million dimensional systems, and this number keeps on increasing with the

101   size and speed of supercomputers.

102   In this paper we discuss a new particle filter methodology that is applicable

103   to systems of much higher dimension, and which up to the dimensions we tested

104   it on has perfect scaling, i.e. the number of particles is independent of the

105   dimension of the state vector. The secret is a proper use of the proposal den-

106   sity, that allows much more freedom than perhaps anticipated in earlier work.

107   Typically, the proposal density has been used to steer the particles to high-

108   probability areas as defined by the observations in state space, but when the

109   number of independent observations is large, the relative weights of the particles

110   will vary enormously, leading to degeneracy. Here we exploit the fact that the

111   proposal density can in addition be used to obtain similar relative weights for the

112   particles, thus avoiding degeneracy. The method is introduced in Van Leeuwen

113   (2010), and Van Leeuwen (2011) discussed applications to systems of up to 1000

114   dimensions using only about 20 particles. In this paper, the method is outlined

115   and its performance on a geophysical system with about 65,000 dimensions is

116   demonstrated.

4

117   The paper is organised as follows. The next section discusses Particle Fil-
118   tering in general, followed by a section on the new method. It is highlighted
119   why other particle filter formulations fail, and how the new method can be suc-
120   cessful. Then the numerical model simulating the barotropic vorticity equation
121   is described, followed by initial results when applying the new particle filter to
122   that system. A concluding chapter closes the paper.

123   **2. Particle filtering**

The probability density function (pdf) of the state vector is represented, and
approximated, by a discrete set of delta functions centred around a set of model
states, called the particles. Using this representation of the prior pdf of the
model in Bayes theorem

$$p(x|y) = \frac{p(y|x)p(x)}{\int p(y|x)p(x)\,dx} \tag{1}$$

where x is the state vector, and y is the observation vector, one finds:

$$p(x|y) = \sum_{i=1}^{N} w_i \delta(x - x_i) \tag{2}$$

in which the weights w_i are related to how close each particle is to the observa-
tions:

$$w_i = \frac{p(y|x_i)}{\sum_{j=1}^{N} p(y|x_j)} \tag{3}$$

The density $p(y|x_i)$ is the likelihood, i.e. the probability density of the ob-
servations given the model state x_i. It is related to the fact that we cannot
make perfect observations, any observation comes with a measurement error,
and hence this density is the pdf of the errors in the observations due to the
measurement process. In the data-assimilation problem it is given, and often
assumed to be Gaussian:

$$p(y|x_i) = A \exp\left[-\frac{1}{2}(y - H(x_i))^T R^{-1}(y - H(x_i))\right] \tag{4}$$

124   in which $H(x_i)$ is the measurement operator, which projects the model state
125   onto the observation space, R is the observation error covariance matrix, and A
126   is a normalisation constant.

127   Unfortunately, weights vary wildly, and even when resampling is applied,

128   only a few particles will have relatively high weight, so will have any statistical

129   significance. This as called *filter degeneracy* and is a very serious problem in

130   standard particle filtering (Snyder et al, 2008). Several methods have been

131   proposed to solve this problem (see review for the geosciences by Van Leeuwen,

132   2009), but none of these is directly applicable to large-dimensional geophysical

133   problems.

134   To understand why this is the case, consider the following. Since the particles

135   x_i are not the evolution of the true system, the distance in observation space

136   between the observation y_j and the particle equivalent $H_j(x_i)$ will on average be

137   similar to or larger than a typical observation error (from the Cauchy-Swartz

138   inequality), so it can be expected that $(y_j - H_j(x_i))^T R_{jj}^{-1}(y_j - H_j(x_i))$ will

139   typically be similar to or larger than 1 for each observation y_j. Assuming M

140   independent observations, $(y - H(x_i))^T R^{-1}(y - H(x_i))$ is expected to be of

141   order M or larger. So, to start with, the likelihood for each particle will be

142   fairly small.

143   However, the particle filters works with relative weights, so we need to ad-

144   dress the variation of the likelihood with the particles. Let us assume that the

145   particles are drawn from a Gaussian with covariance B centred around the true

146   state. Clearly, the larger B, the larger the variation in the weights of the par-

147   ticles. Let us now assume, to illustrate the argument, that HBH^T is of similar

148   magnitude as R. In that case, the argument of the exponent in the likelihood

149   is a χ-squared variable with M degrees of freedom. Such a variable has mean

150   M, and standard deviation $\sqrt{2M}$. This means that the relative weights of the

151   particles differ by a factor $\exp\sqrt{2M}$. Assuming a moderate 50 independent ob-

152   servations, the weights will vary by a factor $\exp(50) \approx 5.0 \ 10^{21}$, so the particle

153   filter will be degenerate when the number of independent observations grows,

154   and serious improvement is needed.

6

3. The new method

The new method that will be explored consists of two ingredients. The first ingredient is that the particles are steered towards the future observations by choosing a specific form of model forcing that tends to pull the model towards the observations. This is an old idea in particle filtering, and has been explored in the Lorenz 1963 and 1996 models in Van Leeuwen (2010, 2011). Assume the model equation to be written as

$$x^n = f(x^{n-1}) + \beta^n \tag{5}$$

in which $f(..)$ denotes the deterministic part of the model and β^n is the stochastic part, and n is the time index. Instead of using this, the model equation is modified to:

$$x^n = f(x^{n-1}) + \hat{\beta}^n + K(y^m - H(x^{n-1})) \tag{6}$$

in which $\hat{\beta}^n$ is random forcing which might have different characteristics from the original random forcing, and y^m denotes future observations at time $m > n$. The main difference with the original model equation is the relaxation term that tends to pull the particle to the future observations y^m with a strength given by matrix K. This relaxation matrix will depend on the application, and an example is given below. This looks like cheating in the sense that the model forcing is not chosen from the probability density of the model error, but as something that we like better. Also, the different particles will have different strength of the 'pulling' term dependent on how far they are from the future observations, so we seem to loose control over the statistical meaning of each particle. However, this different forcing can be compensated for exactly by changing the relative weights of the particles.

In particle filter jargon, we have implemented a proposal transition density instead of using the original transition density. The original transition density is denoted as $p(x^n|x^{n-1})$ specifying how probable state x^n is given state x^{n-1} at the previous time step. For the original model equation (5) this density is

7

given by the pdf of β^n. If the β^n are Gaussian distributed as $N(0, Q)$, we find:

$$p(x^n|x^{n-1}) = N(f(x^{n-1}), Q) \tag{7}$$

The proposal transition density can be written as, assuming a Gaussian distribution for the $\hat{\beta}^n$ with covariance $\hat{Q}$:

$$q(x^n|x^{n-1}, y^m) = N(f(x^{n-1}) + K(y^m - H(x^{n-1})), \hat{Q}) \tag{8}$$

168   Also $\hat{Q}$ can be problem dependent. In the example discussed below we choose it
169   equal to Q. Note that the proposal transition density does depend on the future
170   observations. Furthermore, the relaxation term is part of the deterministic
171   proposal model, since the observations are given.

The question now is how the weights are affected when we arrive at the observations at time m. To this end, let us write the prior pdf at time m as:

$$\begin{aligned} p(x^m) &= \int p(x^m, x^{m-1}, ..., x^0) \, dx^{m-1}...dx^0 \\ &= \int p(x^m|x^{m-1})...p(x^1|x^0)p(x^0) \, dx^{0:m-1} \end{aligned} \tag{9}$$

in which we exploited the Markovian property of the model, and introduced the shorthand notation $dx^{n-1}...dx^0 = dx^{0:n-1}$. Furthermore, the previous set of observations was present at time 0 in this notation. The integrand can be multiplied and divided by the proposal transition densities to find:

$$p(x^m) = \int \frac{p(x^m|x^{m-1})}{q(x^m|x^{m-1}, y^m)}...\frac{p(x^1|x^0)}{q(x^1|x^0, y^m)} q(x^m|x^{m-1}, y^m)...q(x^1|x^0, y^m)p(x^0) \, dx^{0:n-1} \tag{10}$$

In the original model we draw random samples from $p(x^0)$ and from each of the $p(x^i|x^{i-1})$ as indicated above. Using the proposed model we draw samples from $p(x^0)$ and from the proposal transition densities $q(x^i|x^{i-1}, y^m)$. Doing the latter, realising that this creates delta functions for times 0 to $n-1$, we can perform the integrations and find for the prior at time m:

$$p(x^m) = \sum_{i=1}^{N} \hat{w}_i \delta(x - x_i) \tag{11}$$

8

in which the weights are given as:

$$\hat{w}_i = \frac{p(x_i^m|x_i^{m-1})}{q(x_i^m|x_i^{m-1},y^m)}\cdots\frac{p(x_i^1|x_i^0)}{q(x_i^1|x_i^0,y^m)} \qquad (12)$$

172　So where we had equally weighted particles in the standard particle filter for
173　the prior, we now have weighted particles. These weights are related to the fact
174　that we changed the model equations. They specify how probable the move
175　from x^n to x^{n-1} is in the original model, normalised by that probability in the
176　modified model.

Finally, to find the full posterior weights we use Bayes theorem to include
the likelihood, leading to

$$w_i \propto p(y^m|x^m)\frac{p(x_i^m|x_i^{m-1})}{q(x_i^m|x_i^{m-1},y^m)}\cdots\frac{p(x_i^1|x_i^0)}{q(x_i^1|x_i^0,y^m)} \qquad (13)$$

177　Making sure that all particles end up relatively close to the observations still
178　does not avoid wildly varying weights in large-dimensional systems. Clearly,
179　ending up close to the observations reduces the variance in the likelihood weights,
180　but the variance in the weights related to the proposal density are nonzero, and
181　can be substantial.

182　The second new ingredient is that we ensure that all posterior weights are of
183　equivalent size. This is achieved in two stages: first, use the scheme mentioned
184　above for all time steps up to time $n-1$ and perform a deterministic time step
185　with each particle that ensures that most of the particles have equal weight; and
186　secondly, add a very small random perturbation to ensure that Bayes theorem
187　is satisfied. There are many ways to accomplish both stages.

Let us assume that the observation errors and the errors in the model equa-
tions are Gaussian distributed. The weights can be written as:

$$w_i \propto p(y^m|x^m)\frac{p(x^m|x_i^{m-1})}{q(x^m|x_i^{m-1},y^m)}w_i^{rest} \qquad (14)$$

leaving the last time step open. w_i^{rest} contains the weights from all time steps
up to time $n-1$, which are now given (we have done all these steps). Ignoring
the proposal transition density part for the moment, making the weight of each

9

particle equal to $\exp(-C)$, say, leads to the following quadratic equation for particle x_i at time m:

$$\tfrac{1}{2}(x^m - f(x_i^m))^T Q^{-1}(x^m - f(x_i^m))+$$
$$\tfrac{1}{2}(y - Hx_i^m)^T R^{-1}(y - Hx_i^m) - \log(w_i^{rest}) = C \qquad (15)$$

188 Now any quadratic form has a minimum, and depending on the value for C this
189 equations has two, one, or zero real roots for a one dimensional system. Zero
190 roots means that the particle is unable to reach this specified weight $\exp(-C)$;
191 the w_i^{rest} factor for such a particle is too low. Clearly, we don't want the weight
192 of each particle to be the same as the worst particle. We have chosen here a
193 weight C such that 80% can reach it, and the other 20% will be ignored for now.
194 They will re-enter the ensemble via the resampling step later on.

Once C is chosen, an infinite number of solutions exist if the dimension of the system is larger than 1. A simple choice is to enforce

$$x_i^m = f(x_i^{m-1}) + \alpha_i M(y^m - H(f(x_i^{m-1}))) \qquad (16)$$

in which $M = QH^T(HQH^T + R)^{-1}$, Q is the error covariance of the model errors, and R is the error covariance of the observations. α_i is a scalar that is determined from equation (15), and we obtain for each α_i, (see Van Leeuwen, 2010, 2011)]:

$$\alpha_i = 1 - \sqrt{1 - b_i/a_i} \qquad (17)$$

195 in which $a_i = 0.5x_i^T R^{-1} H K z$ and $b_i = 0.5x_i^T R^{-1} x_i - C - \log w_i^{rest}$. Here
196 $z = y^m - H(f(x_i^{m-1}))$, C is the chosen weight level, and w_i^{rest} denotes the
197 relative weights of each particle i up to this time step, related to the proposal
198 density explained above.

199 Note that the last time step so far is a purely deterministic step: we have
200 chosen C, and directly calculated x_i^m. Of course, this last step towards the
201 observations cannot be fully deterministic, as can be seen from Eq. (13). A
202 deterministic proposal would mean that the proposal transition density q can
203 be zero while the target transition density p is non zero, leading to division by

10

204   zero: a deterministic move the transition density is a delta function. In the
205   example presented below the proposal transition density was chosen to be a
206   Gaussian. Since the weights have q in the denominator a draw from the tail of
207   a Gaussian could lead to a very high weight for a particle that is perturbed by a
208   relatively large amount, resulting in the opposite of the intended outcome. We
209   didn't encounter this problem in this experiment.

To avoid this potential problem q could be chosen in the last step before the
observations as a mixture density

$$q(x^m|x') = (1 - \gamma)U(-a, a) + \gamma N(0, a^2) \tag{18}$$

210   in which x' is the particle after the deterministic step outlined above. A draw
211   from this density would be performed as follows. First, we determine from which
212   density U, or N, we will draw the stochastic perturbation, e.g. by drawing u
213   from a uniform density $U[0, 1]$ and if $u < \gamma$ we draw from the normal density
214     $N(0, a^2)$, and we draw from the uniform density $U(-a, a)$ otherwise. By choos-
215   ing γ very small we most likely draw from the uniform density $U(-a, a)$. For
216   small a we can completely control the size of the stochastic perturbation to the
217   state vector. If by chance we have to choose from $N(0, a^2)$ we most likely draw
218   from near the peak of this Gaussian. It is very unlikely to draw from the Gaus-
219   sian and at the same time draw from the tail of that Gaussian. It is mentioned
220   that γ can be made dependent on the number of particles to control the number
221   of times we actually draw from the Gaussian, and keep that number small.

222   **4. The barotropic vorticity equation and statistical set up**

The barotropic vorticity equation describes how the vorticity field ζ changes
with time through advection of the vorticity field by the velocity field:

$$\frac{\partial q}{\partial t} + u\frac{\partial q}{\partial x} + v\frac{\partial q}{\partial y} = \beta \tag{19}$$

in which u the eastward and v the northward velocity, and in which we included
a random forcing β. The vorticity field is related to the velocity field as

$$\zeta = \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y} = 0 \tag{20}$$

11

Because the divergence of the horizontal velocity field is zero:

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0 \tag{21}$$

a streamfunction can be defined as

$$u = -\frac{\partial \psi}{\partial y} \quad v = \frac{\partial \psi}{\partial x} \tag{22}$$

Combining this with the evolution equation for the vorticity field leads to the following set of equations that have to be solved at every time step:

$$\frac{\partial q}{\partial t} - \frac{\partial \psi}{\partial y}\frac{\partial q}{\partial x} + \frac{\partial \psi}{\partial x}\frac{\partial q}{\partial y} = \beta$$
$$q = \frac{\partial^2 \psi}{\partial x^2} + \frac{\partial^2 \psi}{\partial y^2} \tag{23}$$

223 This set of equations is solved on a double periodic domain of 256 by 256 grid
224 points and grid spacing $\Delta x = \Delta y = 1/256$, leading to a state dimension of
225 close to 65,000. At each time step the vorticity field is updated using a semi-
226 Lagrangian scheme with time step $\Delta t = 0.04$, followed by an update of the
227 streamfunction via an inversion of the second equation using FFT's.

228   The stochastic term is chosen from a multivariate Gaussian with mean zero,
229 variance 0.01, and a Gaussian spatial correlation with decorrelation lengthscale
230 4 gridpoints. It is integrated using a simple Euler scheme. Because we are not
231 interested in the specific stochastic evolution, but in the overall properties of the
232 stochastic equation, the accuracy is $O(\Delta t)$. The initial condition was a random
233 vorticiy field with nondimensional amplitude 1 and spatial decorrelation length
234 of 10 grid points. This initial condition results in highly nonlinear turbulent
235 flow structures. Without the random forcing, the evolution of the system would
236 follow that of 2-D turbulence, cascading energy to the largest scales. However,
237 the random forcing keeps on injecting energy at smaller scales, so the flow
238 remains fully turbulent throughout the whole data assimilation experiment. The
239 whole experiment lasted 600 time steps.

240   The vorticity field was observed every 50 time steps on every gridpoint, giv-
241 ing about 65,000 observations every time step. The observations were obtained

12

242   from a truth run and independent random measurement noise with standard

243   deviation 0.05 was added to each observation. This should be compared to the

244   typical nondimensional vorticity values of about 1. We determined the decor-

245   relation timescale τ of the system by averaging the correlation time series at

246   several points in the field, and taking τ as the time scale at which the correlation

247   was $1/e$. We found a decorrelation time scale of this system of about 26 time

248   steps. Since we observe the system every 50 time steps this is an extremely hard

249   nonlinear data assimilation problem.

250   Only 24 particles were used to track the posterior pdf. Because we observe

251   the full state vector we chose K in the relaxation term as a scalar with maximum

252   value $K = 0.1$. (In general, when only part of the state vector is observed, or

253   when H is a nonlinear function, K will be a matrix.) The random forcing

254   covariance $\hat{Q}$ was the same as in the original model Q. This value for K is equal

255   to the standard deviation of the model errors, chosen such that the relaxation

256   term will be of that order of magnitude. Furthermore, K was chosen to vary

257   linearly from zero to its maximum value between observation times. This allows

258   the ensemble to spread out due to the random forcing initially, and pulling

259   harder and harder towards the new observation the closer the system comes to

260   the new observation time. No tuning has been applied in this example of the

261   new particle filter, the reasonable values used for the parameters in the scheme

262   applied to the Lorenz 1993 and 1996 systems (see Van Leeuwen, 2010,2011) have

263   been implemented directly. More detailed experiments in which the sensitivity

264   of the results to these specific choices for K will be described in another paper

265   in preparation.

266   **5. Results**

267   Here a few initial results using the new particle filter with equivalent weights

268   are shown. Figure 1 shows the vorticity field at time 50, and figure 2 the mean of

269   the particles at that time. The two field are almost identical to the eye, showing

270   that the new method is able to track the truth in this highly nonlinear regime.

13

271   Figure 3 shows the vorticity field at time 600, and its particle filter counterpart

272   is shown in figure 4. Again the close tracking is very encouraging.

273   Figure 5 shows the absolute value of the difference between the ensemble

274   mean and the truth run at time 50. This can be compared to the standard

275   deviation in the ensemble in figure 6. Figures 7 and 8 show the same, but now

276   after 600 time steps. Although the spread around the truth is underestimated

277   at several locations, it is over estimated elsewhere, and the averages over the

278   fields are almost equal. Given the statistical nature of these estimates, this is

279   satisfactory.

280   A check on the workings of the equivalent weights scheme is to visualise the

281   weights before resampling. Figure 9 shows that the weights are distributed as

282   they should: they display small variance around the equal weight value $1/20$

283   for the 80% of the 24 particles. Note that the particles with zero weight had

284   too small weight to be included in the equivalent weight scheme, and will be

285   resampled from the rest. Because the weights vary so little the weights can

286   be used back in time, generating a smoother solution for this high-dimensional

287   problem with only 24 particles. The results presented here refer to the filter

288   solution only.

289   One of the questions one could ask is if these results could have been obtained

290   with one of the standard scheme's used in meteorology or oceanography, like

291   4DVar or variants of the EnKF. When concentrating on the mean this might be

292   so, but clearly the structure of the full pdf cannot be reconstructed with these

293   methods. An example is depicted in figure 10, which shows the posterior pdf

294   for the vorticity value at a certain point after 600 time steps. The non-Gaussian

295   structure, hinting at bimodality, cannot be captured by any of these traditional

296   methods.

297   Variational methods like 4DVar typically provide no error estimate because

298   that is too expensive for large-dimensional problems like encountered in e.g.

299   numerical weather prediction. From a scientific point of view this is not satis-

300   factory. Furthermore, in a situation like depicted in figure 10 the usefulness of

301   just the modal value would be limited, and also an error estimate based on the

302 Hessian, so the local curvature, has limited significance. Finally, given that the
303 observation times are about two decorrelation time scales apart, 4DVar might
304 struggle to convergence, but that is not tested here.

305   Ensemble Kalman filter methods do assume Gaussian prior and posterior
306 densities. It might be possible by tuning inflation factors and localisation func-
307 tions to obtain a proper evolution of the ensemble mean, comparable with the
308 true evolution of the system. The ensemble generated with an EnKF might
309 show bimodal structures, but it is unclear if these are real since the update does
310 assume Gaussian pdf's, leading to a posterior pdf that does have the correct
311 posterior mean and covariance (under the Gaussian assumption). The actual
312 positions of the ensemble members differ in the different variants of the ensem-
313 ble Kalman filter and have no direct statistical meaning. Furthermore, it is
314 unclear if the ensemble covariance would have any real scientific meaning for
315 these highly non-Gaussian pdf's. The ensemble spread is perhaps more used
316 for tuning the system to ensure the correct evolution of the mean via inflation
317 and localisation, than representing actual covariances in these highly nonlinear
318 systems.

319   It is stressed here that no tuning has been applied in this example of the
320 new particle filter, the reasonable values used for the parameters in the scheme
321 applied to the Lorenz 1993 and 1996 systems (see Van Leeuwen, 2010, 2011)
322 have been implemented directly.

323   An important issue is the quality of the scheme to infer the full posterior pdf.
324 We have seen that the mean is close to the truth, but that could be due to e.g.
325 extreme relaxation, so that all particles are very close to the observations, and
326 so close to the truth in this high dimensional system. To investigate the quality
327 of the ensemble we calculated a rank histogram using the ensemble values at
328 every 50th time step and at every 4th grid point in each row and column of the
329 field, assuming they were close to independent. For each time instance we rank
330 the value of the truth in each gridpoint in the ensemble values at that gridpoint.
331 This is done through ranking the values for the ensemble members from low to
332 high, and determining where the truth lies in this ranking. The rank histogram

15

333   is constructed by adding a value of 1 to that bin in which the truth falls, e.g. bin

334   4 is increased by 1 if the truth ranks between ensemble member 3 and 4. This

335   is repeated for each gridpoint as mentioned above, and for each time instance,

336   generating one rank histogram. The result is depicted in figure 11.

337   The second way to generate a rank histogram is to rank the observations in

338   the measured ensemble members perturbed by the normal measurement error.

339   This is the method of choice when the truth is not available, as in any real

340   situation. Figure 11 shows the ranking of the truth, but both methods give

341   similar results.

342   In the ideal case any of the particles could act as the truth, resulting in a

343   uniform histogram. A low bias of the particles would yield a histogram in which

344   the truth is biased to the higher rankings, so a histogram with higher bins to

345   the right, and vice versa. An under dispersive ensemble will give rise to a truth

346   value that is either lower or higher than the typical ensemble, resulting in a

347   U-shaped histogram. Finally, an over dispersive ensemble leads to a histogram

348   with a hump in the middle. Although the present histogram in figure 11 is

349   not uniform, it is close to it given the statistical noise. One could do a proper

350   confidence interval test, but that is not attempted here. It is remarkable to

351   see how flat the histogram is, realising the high dimension of the system, the

352   long interval between observations, and the fact that we only use 24 particles.

353   The hump in the middle of the histogram might indicate an over dispersive

354   ensemble, but the peaks at the end of the interval tend to show the opposite,

355   an under dispersive ensemble. Or perhaps we see an over dispersive ensemble

356   with biases in both directions. This little discussion shows that weakness of

357   the histogram, it can be nearly flat for several reasons, not all of them positive.

358   But, not withstanding that, the results are encouraging.

359   The results so far are encouraging and a much more detailed analysis of the

360   present results, looking e.g. more closely at the posterior pdf's, the sensitivity

361   to the observation uncertainty, and the spatial and temporal frequency of the

362   observations. This will be reported on in a future paper.

6. Conclusions and discussion

The effectiveness of a new particle filter that exploits the proposal density and allows small ensemble sizes has been demonstrated on the highly nonlinear 65,000 dimensional barotropic vorticity equation that simulates ocean eddy processes. It was shown using identical twin experiments that the ensemble mean closely follows the truth, and that the ensemble spread is a good measure of the difference between the two. The nonlinear character of the problem is highlighted by studying the posterior pdf's, which often tend to show bimodal behaviour. Finally, a rank histogram for the whole experiment was shown to be close to uniform, indicating that the statistics of the ensemble is sound.

The advantage of this method is the enormous freedom in the two steps that make up the new method. The first adds terms to the model equations that force the model towards the future observations. The simple additive terms allow easy implementation in any simulation code for atmosphere or ocean, or more generally any computer code that simulates a Markov process. But also more sophisticated proposals can be used, like methods that optimise paths on each particle, e.g. a weak-constraint 4DVar solution on each particle. Note that the 4DVar would be special in the sense that the initial condition of the 4DVar is fixed, the particle position at time zero, but a model error term has to be included. Furthermore, since a 4DVar is a deterministic solution a random perturbation has to be added to each time step after the full 4DVar solution has been obtained.

The second crucial step allows the weights to be almost equal. Without this step the particle filter would still be degenerate with a large number of independent observations in the present settings. Also here much freedom exists in how this term is implemented. We replaced the search for the intersection of a hyperplane and the pdf in the 65,000 dimensional space by a simple line search, but many other possibilities can be explored. There is an interesting connection with new developments in rare event simulation using Monte-Carlo methods. Also there good proposal densities are essential, and advances have

17

393 been made that allow simulation with minimal Monte-Carlo statistical errors

394 (see e.g. Vanden Eijnden and Weare, 2012). These links will be pursued in

395 future work.

396 One of the main questions is why this particle filter works in this high-

397 dimensional system with only 24 particles. The reason is not entirely clear

398 yet, but is most likely related to the following. First, one has to realise that

399 there is no inherent problem related to the size of the space spanned by a small

400 number of the particles with a high number of independent observations. (This

401 would be the case for an ensemble Kalman filter.) The clearest examples are

402 variational methods like 4DVar that are able to absorb all observations in a

403 single model run. In the present implementation of the particle filter we do

404 not run a complete 4DVar on each particle but a very crude approximation to

405 that through the relaxation term. (One could run a 4DVar on each particle,

406 as mentioned above, which is what the implicit particle filter of Chorin and

407 Tu (2009) does, but that would be much more expensive, although probably

408 better.) This, however is not enough to avoid filter divergence of the particle

409 filter, i.e. the fact that the likelihood and proposal weights vary too much, with

410 one particle getting a weight close to one, and the others all weights close to zero,

411 when the number of independent observations is large. For that a scheme like

412 the equivalent-weights step is needed to allow for the majority of the particles

413 to have very similar weights, thus avoiding degeneracy. The actual dimension

414 of the manifold on which the dynamics happens will be (much) smaller than

415 the 65,000. The barotropic vorticity dynamics exhibits spatial and temporal

416 coherency in which the smaller-scale motions tend to be slaved to the larger

417 scales. (However, it should be realised that the small-scale random forcing does

418 destroy this coupling to some extent.) It should be realised that of interest is

419 the dimension of the dynamics given the observations, which will be different

420 from that of the dynamical manifold of a free run. Exploring this fact is a very

421 exciting research direction in which the data assimilation community and the

422 dynamical systems community will have to work closely together. It will be

423 clear, however that the dimension of this manifold will be much higher than 24.

18

424   Some variant of the EnKF could be used as proposal density for the particle

425   filter, allowing e.g. for localisation, which is not straightforward in particle

426   filtering (see e.g. Van Leeuwen, 2009, Papadakis et al, 2010, Prakash et al,

427   2011). This might allow us to ignore the relaxation scheme at each time step.

428   The localised EnKF scheme could then be followed by the equivalent weights

429   scheme. This is one direction of further research.

430   One of the main advantages of this particle filter scheme is that no reference

431   is made to the covariance of the model state. It is well known that 4DVar

432   stands or falls with the quality of the covariance of the initial state, the so-

433   called B matrix. An enormous research effort has been spent, and is still spent

434   on improving this B matrix. Also Ensemble Kalman Filters rely on the accuracy

435   of the ensemble covariance matrix. This is why so much effort has gone into,

436   and is still going into better inflation and localisation schemes. All these issues

437   play no role in particle filtering.

438   It is well realised in the geoscienes community that errors in the model

439   equations have to be included in the data-assimilation schemes. However, a

440   proper statistical description of these errors is hard to come by. Even if it is

441   assumed that the errors are Gaussian distributed, the mean, related to a model

442   bias, and its covariance need to be specified, which is not easy. But that doesn't

443   mean we should not go forward, especially when we realise that this will be the

444   proper way to model improvement. As soon as an estimate of the statistical

445   properties of the model errors is obtained, implementation in ensemble data

446   assimilation methods like EnKf and particle filters is relatively easy because

447   random realisations for these error estimates can be added directly to each

448   ensemble member (and similar for multiplicative errors). Much more research

449   is needed to come up with efficient implementations in variational methods.

450   So, particle filters like the one explored here force us to consider where we are

451   weakest: the errors in the model equations, and these particle filters are not

452   distracted by problems in covariance structures in the model states themselves.

453   Although the results presented here might be promising, much more research

454   is needed before questions on suitability for e.g. numerical weather forecasting

19

455   can be answered. For example, we observed the full state vector at observation
456   times, which is never the case for any real application from the geosciences.
457   We are working hard on partially observed systems now. On the other hand,
458   we observed the system at twice the decorrelation time scale, which makes the
459   problem extremely hard since the information from previous observations is lost
460   to a very large extent, showing the robustness of the method.

461   Another critical issue is to what extent the method pulls the system out
462   of quasi balanced states. In numerical weather prediction tremendous progress
463   was made when models were forced to stay close to balanced states, greatly
464   suppressing artificial gravity waves that ruined the forecasts. It should be re-
465   alised that as soon as we accept a statistical description of model errors model
466   balances will be perturbed. So the question is if and how the proposal den-
467   sity will perturb the model balances more than just the random forcing. By
468   keeping the stochastic part of the proposal density of similar magnitude as the
469   original transition density that part should not add extra perturbations. The
470   deterministic relaxation term can grow quite large, but if that becomes problem-
471   atic we can restrict its size to some maximum value without problem. Another
472   option is to project the relaxation terms to some sort of slow manifold, as is
473   done in high-resolution numerical weather prediction ensemble Kalman filter
474   applications. However, when the dynamics is strongly nonlinear it is unclear
475   what the actual balances are. Finally, the essential equivalent weights step can
476   be large too. Also here we could limit the size of the deterministic move, but
477   this might destroy the possibility for majority of the weights to be equivalent.
478   Also, projection on a slow manifold might help here too, with the same caveat
479   as above. More research into these aspects are needed, and will no doubt be
480   problem dependent.

481   **Acknowledgements**

20

484   two anonymous reviewers are thanked for critical comments that improved the

485   paper substantially.

486   **REFERENCES**

487   Bennett, A., 1992, Inverse Methods in Physical Oceanography, Cambridge Uni-

488   versity Press, Cambridge.

489

490   Burgers, G., Van Leeuwen, P.J., and Evensen, G., 1998, Analysis Scheme in the

491   Ensemble Kalman Filter, Monthly Weather Review, 126, 1719-1724.

492

493   Chorin, A.J., and X. Tu, 2009, Implicit sampling for particle filters, PNAS, 106,

494   17249-17254.

495

496   Courtier, P., 1997, Dual formulation of four-dimensional variational assimila-

497   tion, Quarterly Journal of the Royal Meteorological Society, 123(B), 2449-2461.

498

499   Doucet, A., De Freitas,N., and Gordon, 2001, N., Sequential Monte-Carlo meth-

500   ods in practice, Springer, Berlin.

501

502   Evensen, G., 1994, Sequential data assimilation with a nonlinear quasi-geostrophic

503   model using Monte-Carlo methods to forecast error statistics, Journal of Geo-

504   physical Research, 99, 10143-10162.

505

506   Gordon, N.J., D.J. Salmond, and A.F.M. Smith., 1993, Novel approach to

507   nonlinear/non-Gaussian Bayesian state estimation, IEEE Proceedings-F, 140,

508   107-113.

509

510   Jazwinki, A.H., 1970, Stochastic processes and filtering theory, Academic Press,

511   376p.

512   Papadakis, N., E.Memin, A. Cuzol, and N.Gengembre, 2010, Data assimilation

513   with the weighted ensemble Kalman filter, Tellus, 62, 673-697.

514 Prakask J., S.C.Patwardhan, and S.L.Shah, 2011, On the choice of importance
515 distributions for unconstrained and constrained state estimation using particle
516 filters, J. Process Control, 21, 3-16.

517

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

520

521 Talagrand, O., and P. Courtier, 1987, Variational assimilation of meteorological
522 observations with the adjoint vorticity equation. I: Theory, Quarterly Journal
523 of the Royal Meteorological Society, 113, 1311-1328.

524

525 Vanden Eijnden E. and J. Weare, 2012, Rare event simulation with vanishing
526 error for small noise diffusions, Communications Pure Applied Math., in press.
527 Van Leeuwen, P.J., 2009, Particle Filtering in Geosciences, Monthly Weather
528 Review, 137, 4089-4114.

529

530 Van Leeuwen, P.J., 2010, Nonlinear Data Assimilation in geosciences: an ex-
531 tremely efficient particle filter, Quarterly Journal of the Royal Meteorological
532 Society., 136, 1991-1996.

533

534 Van Leeuwen, P.J. 2011, Efficient non-linear Data Assimilation in Geophysical
535 Fluid Dynamics, Computers and Fluids, doi:10,1016/j.compfluid.2010.11.011.

536 **List of Figures**

22

Figure 1: Snap shot of the vorticity field of the truth at time 50. Note the highly chaotic state of the field.
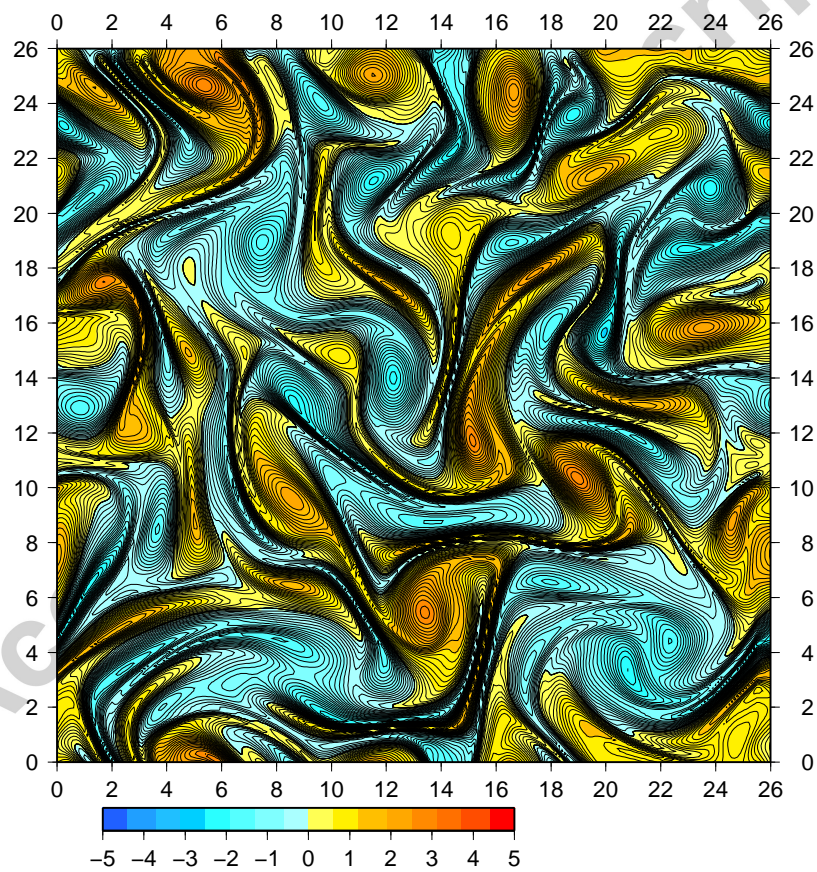
Figure 2: Snap shot of the vorticity field of the mean of the particle filter mean at time 50. Compare with figure 1 and note the close to perfect tracking.
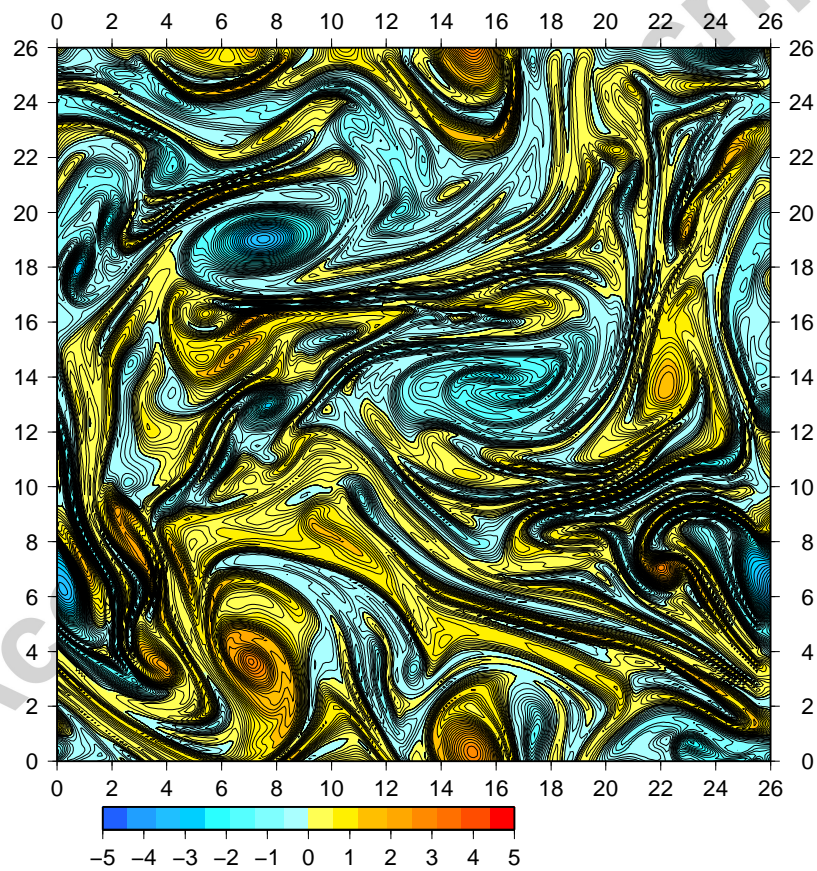
Figure 3: Snap shot of the vorticity field of the truth at time 600. Note again the highly chaotic state of the field.
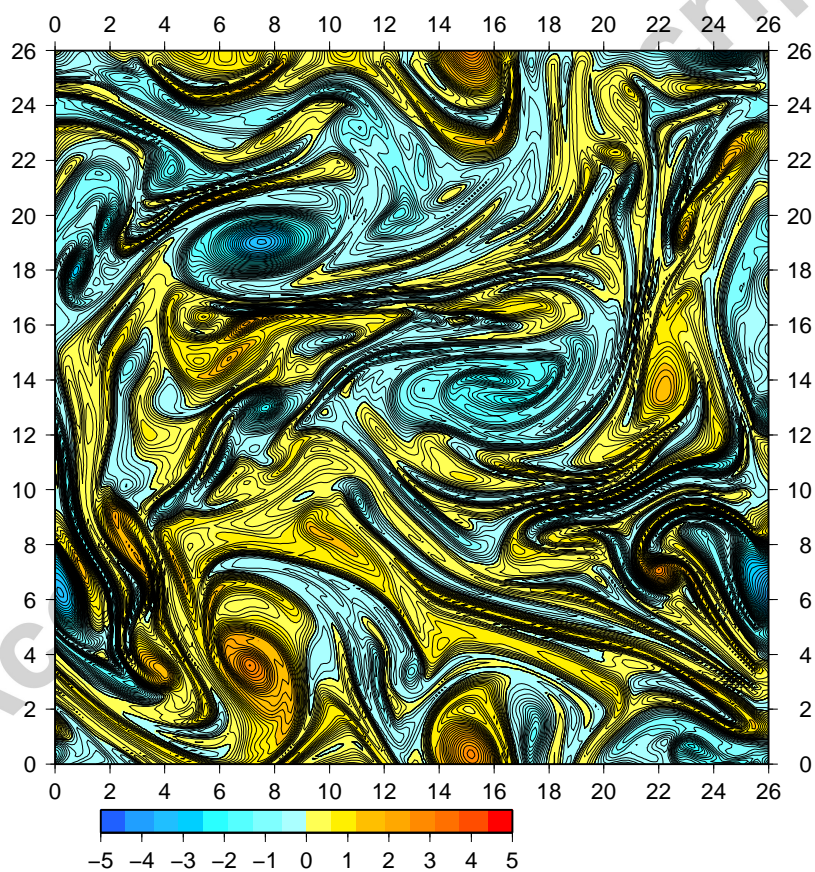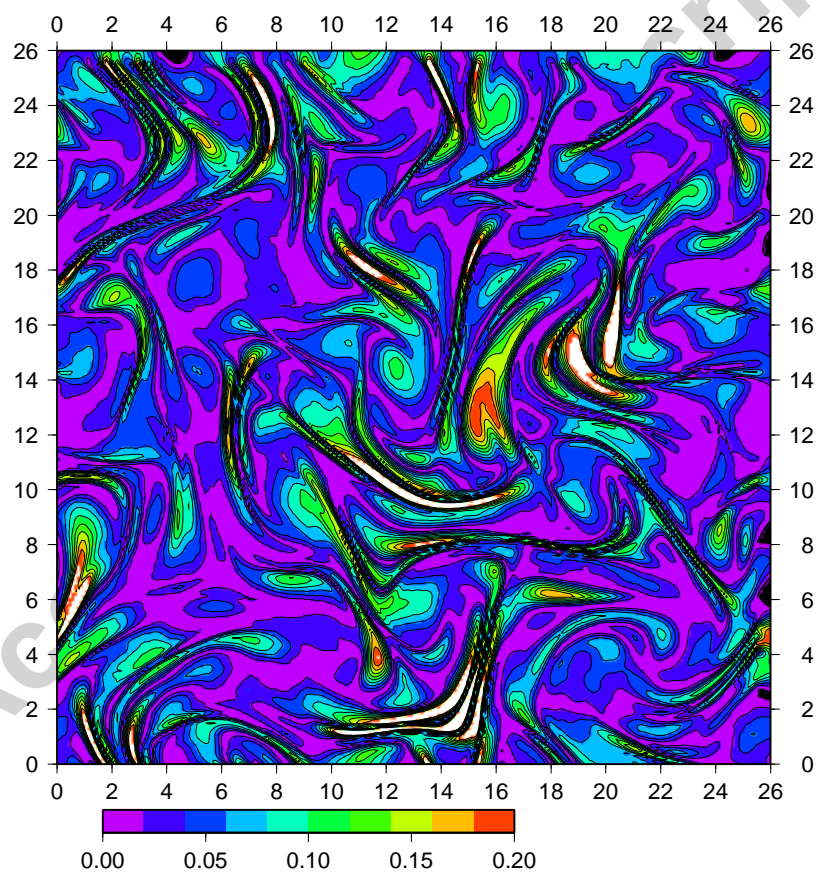
Figure 4: Snap shot of the vorticity field of the mean of the particle filter mean at time 600. Compare with figure 3 and note the close to perfect tracking.

Figure 5: Snap shot of the absolute value of the mean-truth misfit at time 50. Note the highly irregularity of the field, reflecting the statistical nature of the estimate.
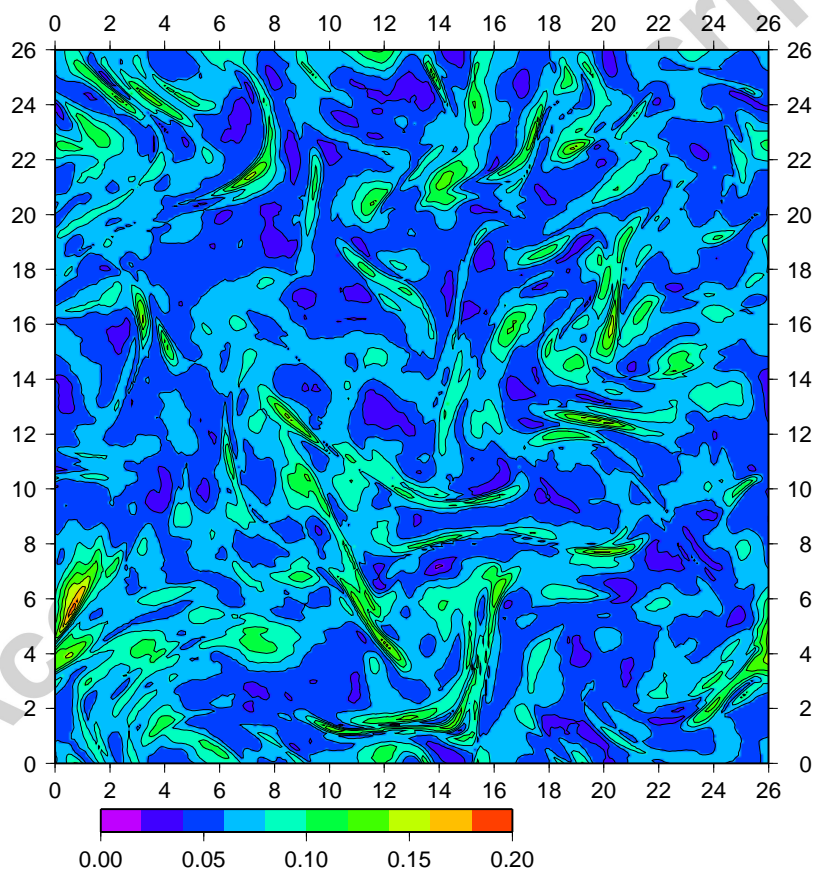
28

Figure 6: Snap shot of the absolute value of the standard deviation in the ensemble at time 50 for comparison with figure 5. The ensemble underestimates the spread at several locations, but averaged over the field it is slightly higher, 0.074 versus 0.056.
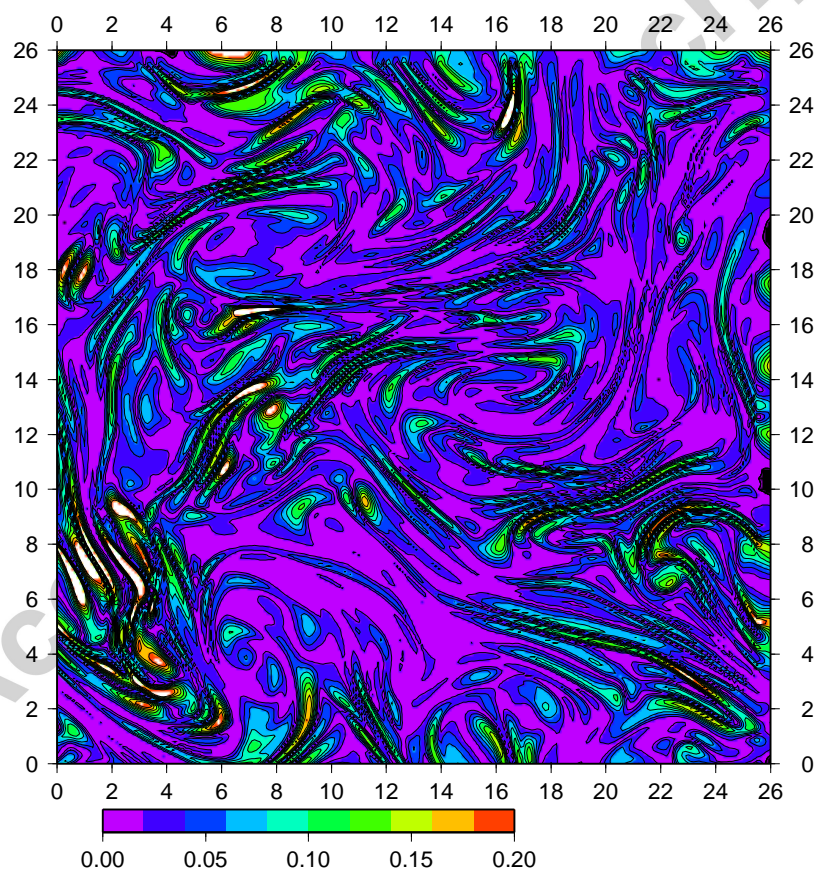
29

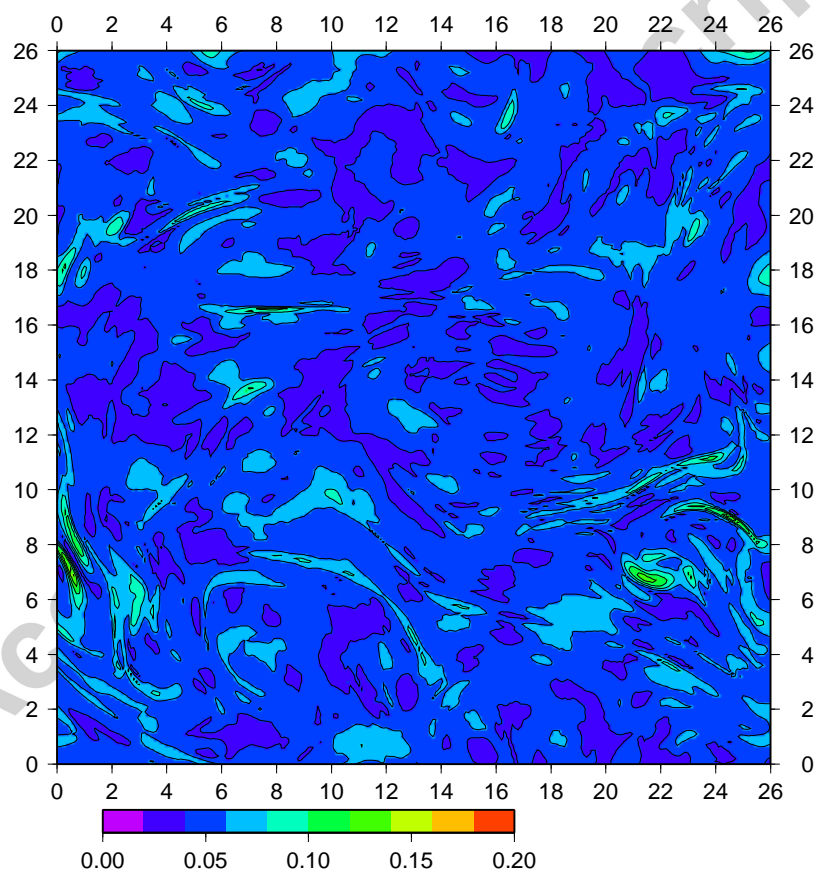Figure 7: Snap shot of the absolute value of the mean-truth misfit at time 600.

Figure 8: Snap shot of the absolute value of the standard deviation in the ensemble at time 600 for comparison with figure 7
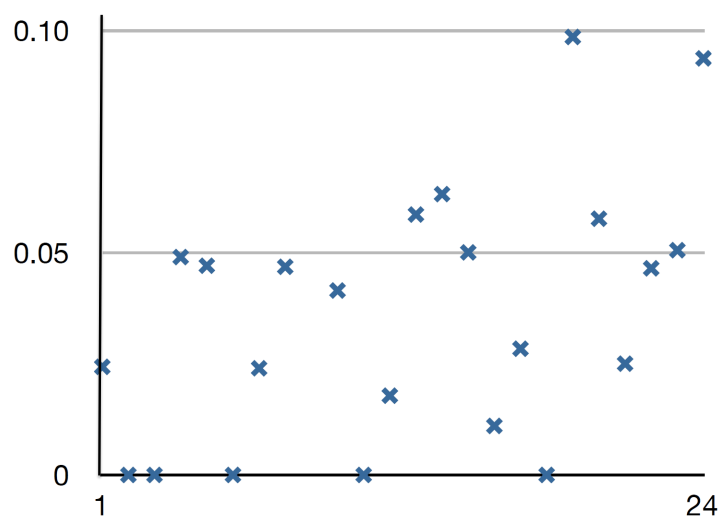
Figure 9: Weights distribution of the particles before resampling. All weights cluster around 0.05, which is close to 1/24 for uniform weights (using 24 particles). The 5 particles with weights zero will be resampled. Note that the other particles form the smoother estimate.
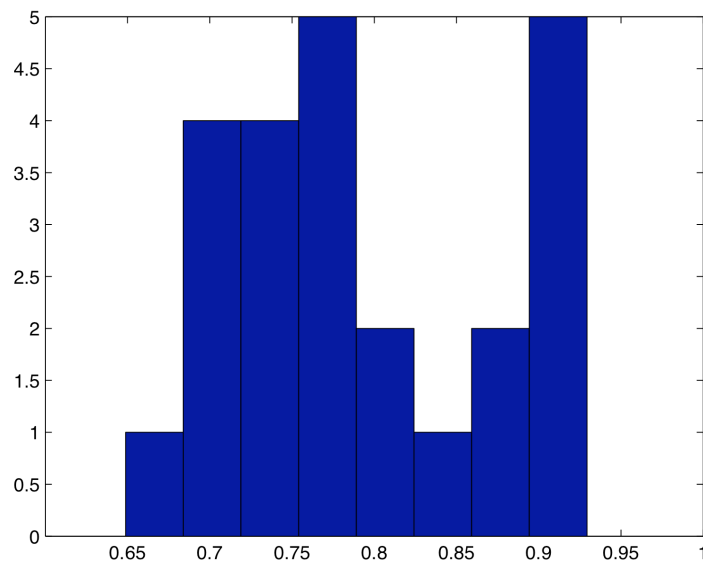
Figure 10: Estimate of the posterior pdf of the vorticity value at point (200,200) after 600 time steps. The bimodal structure shows that variational methods that look for the mode of this pdf have little meaning, and also methods based on the EnKF will not be able to represent this structure accurately.
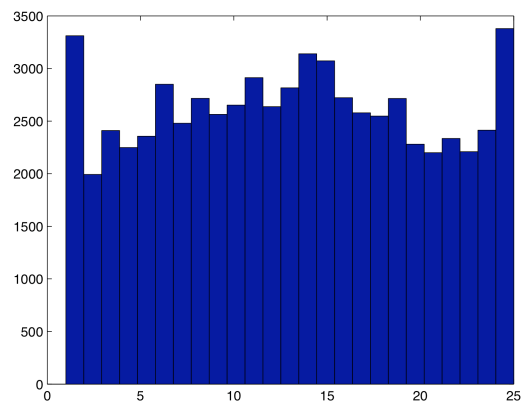
Figure 11: Rank histogram of how the truth ranks in the ensemble. The nearly uniform distribution shows that each particle could act as the truth, suggesting the ensemble is of good quality. (But see discussion in text.)

34

Department of Meteorology

University of Reading
Earley Gate, PO Box 243
Reading RG6 6BB

phone +44 (0)118 378 5578
fax   +44 (0)118 378 8905
email p.j.vanleeuwen@reading.ac.uk

5 January 2012

Highlights:

1. First application of fully nonlinear data assimilation method to complex geophysical fluid dynamical system.
2. First application of full particle filter to 65,000 dimensional system
3. Detailed discussion of advantages and disadvantages of the new method