Well-posedness and accuracy of the ensemble Kalman filter in discrete and continuous time

IOP Publishing | London Mathematical Society Nonlinearity

Nonlinearity **27** (2014) 2579–2603 doi:10.1088/0951-7715/27/10/2579

Well-posedness and accuracy of the ensemble Kalman filter in discrete and continuous time

D T B Kelly[1], **K J H Law**[2] **and A M Stuart**[1]

[1] Mathematics Institute, University of Warwick, Coventry CV4 7AL, UK
[2] Computer Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology, Thuwal 23955-6900, Kingdom of Saudi Arabia

E-mail: dtbkelly@gmail.com

Received 11 October 2013, revised 26 March 2014
Accepted for publication 13 August 2014
Published 22 September 2014

Recommended by B Eckhardt

Abstract

The ensemble Kalman filter (EnKF) is a method for combining a dynamical model with data in a sequential fashion. Despite its widespread use, there has been little analysis of its theoretical properties. Many of the algorithmic innovations associated with the filter, which are required to make a useable algorithm in practice, are derived in an *ad hoc* fashion. The aim of this paper is to initiate the development of a systematic analysis of the EnKF, in particular to do so for small ensemble size. The perspective is to view the method as a state estimator, and not as an algorithm which approximates the true filtering distribution. The perturbed observation version of the algorithm is studied, without and with variance inflation. Without variance inflation well-posedness of the filter is established; with variance inflation accuracy of the filter, with respect to the true signal underlying the data, is established. The algorithm is considered in discrete time, and also for a continuous time limit arising when observations are frequent and subject to large noise. The underlying dynamical model, and assumptions about it, is sufficiently general to include the Lorenz '63 and '96 models, together with the incompressible Navier–Stokes equation on a two-dimensional torus. The analysis is limited to the case of complete observation of the signal with additive white noise. Numerical results are presented for the Navier–Stokes equation on a two-dimensional torus for both complete and partial observations of the signal with additive white noise.

Keywords: data assimilation, ensemble Kalman filter, stochastic analysis
Mathematics Subject Classification: 62M20, 93E11, 60G35, 60H15

1. Introduction

In recent years the ensemble Kalman filter (EnKF) [8] has become a widely used methodology for combining dynamical models with data. The algorithm is used in oceanography, oil reservoir simulation and weather prediction [4, 9, 15, 25], for example. The basic idea of the method is to propagate an ensemble of particles to describe the distribution of the signal given data, employing empirical second order statistics to update the distribution in a Kalman-like fashion when new data is acquired. Despite the widespread use of the method, its behaviour is not well understood. In contrast with the ordinary Kalman filter, which applies to linear Gaussian problems, it is difficult to find a mathematical justification for EnKF. The most notable progress in this direction can be found in [18, 22], where it is proved that, for linear dynamics, the EnKF approximates the usual Kalman filter in the large ensemble limit. This analysis is however far from being useful for practitioners who typically run the method with small ensemble size on nonlinear problems. Furthermore there is an accumulation of numerical evidence showing that the EnKF, and related schemes such as the extended Kalman filter, can 'diverge' with the meaning of 'diverge' ranging from simply losing the true signal through to blow-up [11, 13, 19]. The aim of our work is to try and build mathematical foundations for the analysis of the EnKF, in particular with regards to well-posedness (lack of blow-up) and accuracy (tracking the signal over arbitrarily long time-intervals). To make progress on such questions it is necessary to impose structure on the underlying dynamics and we choose to work with dissipative quadratic systems with energy-conserving nonlinearity, a class of problems which has wide applicability [21] and which has proved to be useful in the development of filters [20].

In section 2 we set out the filtering problem that will be considered in the article. In section 3 we derive the perturbed observation form of the EnKF and demonstrate how it links to the randomized maximum likelihood (RML) method which is widely used in oil reservoir simulation [25]. We also introduce the idea of variance inflation, widely used in many practical implementations of the EnKF [1]. Section 4 contains theoretical analyses of the perturbed observation EnKF, without and with variance inflation. Without variance inflation we are able only to prove bounds which grow exponentially in the discrete time increment underlying the algorithm (theorem 4.2); with variance inflation we are able to prove filter accuracy and show that, in mean square with respect to the noise entering the algorithm, the filter is uniformly close to the true signal for all large times, provided enough inflation is employed (theorem 4.4). These results, and in particular the one concerning variance inflation, are similar to the results developed in [3] for the 3DVAR filter applied to the Navier–Stokes equation and for the 3DVAR filter applied to the Lorenz '63 model in [17], as well as the similar analysis developed in [24] for the 3DVAR filter applied to globally Lipschitz nonlinear dynamical systems. In section 5 we describe a continuous time limit in which data arrives very frequently, but is subject to large noise. If these two effects are balanced appropriately a stochastic (partial) differential equation limit is found and it is instructive to study this limiting continuous time process. This idea was introduced in [2] for the 3DVAR filter applied to the Navier–Stokes equation, and also analysed for the 3DVAR filter applied to the Lorenz '63 model in [LSS14], and is here employed for the EnKF filter. The primary motivation for the continuous time limit is to obtain insight into the mechanisms underlying the EnKF; some of these mechanisms are more transparent in continuous time. In section 6 we analyse the well-posedness of the continuous time EnKF (theorem 6.2). Section 7 contains numerical experiments which illustrate and extend the theory, and section 8 contains some brief concluding remarks.

Throughout the sequel we use the following notation. Let $\mathcal{H}$ be a separable Hilbert space with norm $|\cdot|$ and inner product $\langle\cdot,\cdot\rangle$. We will use the notation $\mathcal{L}(\mathcal{H},\mathcal{K})$ to denote the space

of linear operators with domain $\mathcal{H}$ and range $\mathcal{K}$. For a linear operator C on $\mathcal{H}$, we will write $C \geqslant 0$ (resp. $C > 0$) when C is self-adjoint and positive semi-definite (respectively, positive definite). Given $C > 0$, we will denote

$$|\cdot|_C \stackrel{\text{def}}{=} \left|C^{-1/2}(\cdot)\right|.$$

Unless otherwise stated, we will use $\boldsymbol{E}$ to refer to the expectation operator, taken over all random elements.

2. Set-up

2.1. Filtering distribution

We assume that the observed dynamics are governed by an evolution equation

$$\frac{\mathrm{d}u}{\mathrm{d}t} = F(u) \tag{1}$$

which generates a one-parameter semigroup $\Psi_t : \mathcal{H} \to \mathcal{H}$. We also assume that $\mathcal{K} \subset \mathcal{H}$ is another Hilbert space, which acts as the *observation space*. We assume that noisy observations are made in $\mathcal{K}$ every h time units and write $\Psi = \Psi_h$. We define $u_j = u(jh)$ for $j \in \mathbb{N}$ and, assuming that u_0 is uncertain and modelled as Gaussian distributed, we obtain

$$u_{j+1} = \Psi(u_j), \quad \text{with } u_0 \sim N(m_0, C_0)$$

for some initial mean m_0 and covariance C_0. We are given the observations

$$y_{j+1} = Hu_{j+1} + \Gamma^{1/2}\xi_{j+1}, \quad \text{with } \xi_j \sim N(0, I) \text{ i.i.d.},$$

where $H \in \mathcal{L}(\mathcal{H}, \mathcal{K})$ is the *observation operator* and $\Gamma \in \mathcal{L}(\mathcal{K}, \mathcal{K})$ with $\Gamma \geqslant 0$ is the covariance operator of the observational noise; the i.i.d. noise sequence $\{\xi_j\}$ is assumed independent of u_0. The aim of filtering is to approximate the distribution of u_j given $Y_j = \{y_\ell\}_{\ell=1}^j$ using a sequential update algorithm. That is, given the distribution $u_j|Y_j$ as well as the observation y_{j+1}, find the distribution of $u_{j+1}|Y_{j+1}$. We refer to the sequence $\boldsymbol{P}(u_j|Y_j)$ as the *filtering distribution*.

2.2. Assumptions

To write down the EnKF as we do in section 3, and indeed to derive the continuum limit of the EnKF, as we do in section 5, we need make no further assumptions about the underlying dynamics and observation operator other than those made above. However, in order to analyse the properties of the EnKF, as we do in sections 4 and 6, we will need to make structural assumptions and we detail these here. It is worth noting that the assumptions we make on the underlying dynamics are met by several natural models used to test data assimilation algorithms. In particular, the 2D Navier–Stokes equations on a torus, as well as both Lorenz '63 and '96 models, satisfy assumption 2.3 [20, 21, 26].

Assumption 2.3 (Dynamics model). *Suppose there is some Banach space $\mathcal{V}$, equipped with norm $\|\cdot\|$, that can be continuously embedded into $\mathcal{H}$. We assume that (1) has the form*

$$\frac{\mathrm{d}u}{\mathrm{d}t} + \mathcal{A}u + \mathcal{B}(u, u) = f, \tag{2}$$

where $\mathcal{A} : \mathcal{H} \to \mathcal{H}$ is an unbounded linear operator satisfying

$$\langle \mathcal{A}u, u\rangle \geqslant \lambda\|u\|^2, \tag{3}$$

for some $\lambda > 0$, $\mathcal{B}$ is a symmetric bilinear operator $\mathcal{B} : \mathcal{V} \times \mathcal{V} \to \mathcal{H}$ and time dependent forcing $f : \mathbb{R}_+ \to \mathcal{H}$. We furthermore assume that $\mathcal{B}$ satisfies the identity

$$\langle \mathcal{B}(u, u), u \rangle = 0, \tag{4}$$

for all $u \in \mathcal{H}$ and also

$$\langle \mathcal{B}(u, v), v \rangle \leqslant c\|u\|\|v\||v| \,, \tag{5}$$

for all $u, v \in \mathcal{H}$, where $c > 0$ depends only on the bilinear form. We assume that the equation (2) has a unique weak solution for all $u(0) \in \mathcal{H}$, and generates a one-parameter semigroup $\Psi_t : \mathcal{V} \to \mathcal{V}$ which may be extended to act on $\mathcal{H}$. Finally we assume that there exists a global attractor $\Lambda \subset \mathcal{V}$ for the dynamics, and constant $R > 0$ such that for any initial condition $u_0 \in \Lambda$, we have that $\sup_{t \geqslant 0} \|u(t)\| \leqslant R$. ■

Remark 2.4. In the finite dimensional case the final assumption on the existence of a global attractor does not need to be made as it is a consequence of the preceding assumptions made. To see this note that

$$\frac{1}{2}\frac{\mathrm{d}|u|^2}{\mathrm{d}t} + \lambda\|u\|^2 \leqslant \langle f, u \rangle. \tag{6}$$

The continuous embedding of $\mathcal{V}$, together with the Cauchy–Schwarz inequality, implies the existence of a strictly positive constant ϵ such that

$$\frac{1}{2}\frac{\mathrm{d}|u|^2}{\mathrm{d}t} + \epsilon|u|^2 \leqslant \frac{1}{2\delta}|f|^2 + \frac{\delta}{2}|u|^2 \tag{7}$$

for all $\delta > 0$. Choosing $\delta = \epsilon$ gives the existence of an absorbing set and hence a global attractor by theorem 1.1 in chapter I of [26]. In infinite dimensions the existence of a global attractor in $\mathcal{V}$ follows from the techniques in [26] for the Navier–Stokes equation by application of more subtle inequalities relating to the bilinear operator $\mathcal{B}$—see section 2.2 in chapter 3 of [26]. Other equations arising in dissipative fluid mechanics can be treated similarly. ■

Whilst the preceding assumptions on the underlying dynamics apply to a range of interesting models arising in applications, the following assumptions on the observation model are rather restrictive; however we have been unable to extend the analysis without making them. We will demonstrate, by means of numerical experiments, that our results extend beyond the observation scenario employed in the theory

Assumption 2.5 (Observation model). *The system is completely observed so that $\mathcal{K} = \mathcal{H}$ and $H = I$. Furthermore the i.i.d. noise sequence $\{\xi_j\}$ is white so that $\xi_1 \sim N(0, \Gamma)$ with $\Gamma = \gamma^2 I$.* ■

The following consequence of assumption 2.3 will be useful to us.

Lemma 2.6. *Let assumption 2.3 hold. Then there is $\beta \in \mathbb{R}$ such that, for any $v_0 \in \Lambda$, $h > 0$ and $w_0 \in \mathcal{H}$,*

$$|\Psi_h(v_0) - \Psi_h(w_0)| \leqslant e^{\beta h}|v_0 - w_0|.$$

Proof. Let v, w denote the solutions of (2) with initial conditions v_0, w_0 respectively; define $e = v - w$. From the bilinearity of $\mathcal{B}$, it follows that

$$\frac{\mathrm{d}e}{\mathrm{d}t} + \mathcal{A}e + 2\mathcal{B}(v, e) - \mathcal{B}(e, e) = 0, \tag{8}$$

with $e(0) = v_0 - w_0$. Taking the inner-product with e, using (3), (4) and (5), and choosing $\delta = \lambda/(2cR)$, gives

$$\begin{aligned}\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}|e|^2 + \lambda\|e\|^2 &\leqslant 2c\|v\|\|e\||e| \\ &\leqslant 2cR\|e\||e| \\ &\leqslant cR(\delta\|e\|^2 + \delta^{-1}|e|^2) \\ &= \frac{\lambda}{2}\|e\|^2 + \frac{2}{\lambda}(cR)^2|e|^2.\end{aligned}$$

Thus

$$\frac{\mathrm{d}}{\mathrm{d}t}|e|^2 \leqslant \frac{4}{\lambda}(cR)^2|e|^2$$

and the desired result follows from an application of the Gronwall inequality. □

Remark 2.7. It should be emphasized that in lemma 2.6, we only require that one of the starting points v_0, w_0 lies in the attractor. This is a special advantage of assuming a quadratic structure for the nonlinearity. ■

3. The EnKF

3.1. The algorithm

The idea of the EnKF is to represent the filtering distribution through an ensemble of particles, to propagate this ensemble under the model to approximate the mapping $\boldsymbol{P}(u_j|Y_j)$ to $\boldsymbol{P}(u_{j+1}|Y_j)$ (refered to as *prediction* in the applied literature), and to update the ensemble distribution to include the data point Y_{j+1} by using a Gaussian approximation based on the second order statistics of the ensemble (refered to as *analysis* in the applied literature).

The prediction step is achieved by simply flowing forward the ensemble under the model dynamics, that is

$$\widehat{v}_{j+1}^{(k)} = \Psi(v_j^{(k)}), \qquad \text{for } k = 1 \dots K.$$

The analysis step is achieved by performing a randomized version of the Kalman update formula, and using the empirical covariance of the prediction ensemble to compute the Kalman gain. There are many variants on the basic EnKF idea and we will study the perturbed observation form of the method.

The algorithm proceeds as follows.

(i) Set $j = 0$ and draw an independent set of samples $\{v_0^{(k)}\}_{k=1}^K$ from $N(m_0, C_0)$.

(ii) (Prediction) Let $\widehat{v}_{j+1}^{(k)} = \Psi(v_j^{(k)})$ and define $\widehat{C}_{j+1}$ as the empirical covariance of $\{\widehat{v}_{j+1}^{(k)}\}_{k=1}^K$. That is,

$$\widehat{C}_{j+1} = \frac{1}{K}\sum_{k=1}^{K}(\widehat{v}_{j+1}^{(k)} - \bar{v}_{j+1}) \otimes (\widehat{v}_{j+1}^{(k)} - \bar{v}_{j+1}),$$

where $\bar{v}_{j+1} = \frac{1}{K}\sum_{k=1}^K \widehat{v}_{j+1}$ denotes the ensemble mean.

(iii) (Observation) Make an observation $y_{j+1} = Hu_{j+1} + \Gamma^{1/2}\xi_{j+1}$. Then, for each $k = 1 \dots K$, generate an artificial observation

$$y_{j+1}^{(k)} = y_{j+1} + \Gamma^{1/2}\xi_{j+1}^{(k)},$$

where $\xi_{j+1}^{(k)}$ are $N(0, I)$ distributed and pairwise independent.

(iv) (Analysis) Let $v_{j+1}^{(k)}$ be the minimizer of the functional

$$J(v) = \tfrac{1}{2}|y_{j+1}^{(k)} - Hv|_\Gamma^2 + \tfrac{1}{2}|\widehat{v}_{j+1}^{(k)} - v|_{\widehat{C}_{j+1}}^2.$$

(v) Set $j \mapsto j+1$ and return to step (ii).

The name 'perturbed observation EnKF' follows from the construction of the artificial observations $y_{j+1}^{(k)}$ which are found by perturbing the given observation with additional noise. The sequence of minimizers v_{j+1} can be written down explicitly by simply solving the quadratic minimization problem. This straightforward exercise yields the following result.

Proposition 3.2. *The sequence* $\{v_j^{(k)}\}_{j\geqslant 0}$ *is defined by the equation*

$$(I + \widehat{C}_{j+1}H^T\Gamma^{-1}H)v_{j+1}^{(k)} = \widehat{v}_{j+1}^{(k)} + \widehat{C}_{j+1}H^T\Gamma^{-1}y_{j+1}^{(k)},$$

for each $k = 1, \ldots, K$.

Hence, collecting the ingredients from the preceding, the defining equations of the EnKF are given by

$$(I + \widehat{C}_{j+1}H^T\Gamma^{-1}H)v_{j+1}^{(k)} = \Psi(v_j^{(k)}) + \widehat{C}_{j+1}H^T\Gamma^{-1}y_{j+1}^{(k)} \tag{9a}$$

$$y_{j+1}^{(k)} = y_{j+1} + \Gamma^{1/2}\xi_{j+1}^{(k)} \tag{9b}$$

$$\bar{v}_{j+1} = \frac{1}{K}\sum_{k=1}^{K}\Psi(v_j^{(k)}) \tag{9c}$$

$$\widehat{C}_{j+1} = \frac{1}{K}\sum_{k=1}^{K}\Big(\Psi(v_j^{(k)}) - \bar{v}_{j+1}\Big)\otimes\Big(\Psi(v_j^{(k)}) - \bar{v}_{j+1}\Big). \tag{9d}$$

There are other representations of the EnKF that are more algorithmically convenient, but the formulae (9) are better suited to our analysis.

3.3. Connection to RML

As pointed out in [10], the analysis step of EnKF can be understood in terms of the RML method widely used in oil reservoir history matching applications [25]. We will now briefly describe this method. Suppose that we have a random variable u and that $u \sim N(\widehat{m}, \widehat{C})$. Moreover, let G be some linear operator and suppose we observe

$$y = Gu + \xi \quad \text{where } \xi \sim N(0, \Gamma).$$

One can use Bayes' theorem to write down the conditional density $\boldsymbol{P}(u|y)$. In practice however, it is often sufficient (or sometimes even better) to simply have a collection of *samples* $\{u^{(k)}\}_{k=1}^K$ from the conditional distribution, rather than the density itself. RML is a method of taking samples from the prior $N(\widehat{m}, \widehat{C})$ and turning them into samples from the posterior. This is achieved as follows, given $\widehat{u}^{(k)} \sim N(\widehat{m}, \widehat{C})$ (samples from the prior), define $u^{(k)}$ for each $k = 1 \ldots K$ by $u^{(k)} = \operatorname{argmin}_u J^{(k)}(u)$ where

$$J^{(k)}(u) = \tfrac{1}{2}|y - Gu + \Gamma^{1/2}\xi^{(k)}|_\Gamma^2 + \tfrac{1}{2}|u - \widehat{u}^{(k)}|_{\widehat{C}}^2,$$

where $\xi^{(k)} \sim N(0, I)$ and independent of ξ. The $u^{(k)}$ are then draws from the posterior distribution of $u|y$ which is a Gaussian with mean m and covariance C. Since one can explicitly write down (m, C), it may be checked that the $u^{(k)}$ defined as above are independent random variables of the form $u^{(k)} = m + C^{1/2}\zeta^{(k)}$, where $\zeta^{(k)} \sim N(0, I)$ i.i.d. and are hence draws from the desired posterior, as we know show.

Proposition 3.4. *Assume that $\hat{C}$ is invertible. Then, in the above notation, we have that $u^{(k)} = m + C^{1/2}\zeta^{(k)}$, where $\zeta^{(k)} \sim N(0, I)$ i.i.d and (m, C) are defined by*

$$C^{-1} = \widehat{C}^{-1} + G^T\Gamma^{-1}G \tag{10}$$

$$C^{-1}m = G^T\Gamma^{-1}y + \widehat{C}^{-1}\widehat{m}. \tag{11}$$

In particular, $u^{(k)}$ is a sample from the posterior of $u|y$.

Proof. Firstly, note that C is invertible since $\hat{C}$ is invertible. Secondly, it is well known that the pair (m, C) defined by (10), (11) do indeed define the mean and covariance of the posterior. This can be easily verified by matching coefficients in the expression for the negative log-density

$$\tfrac{1}{2}|y - Gu|_\Gamma^2 + \tfrac{1}{2}|u - \widehat{m}|_{\widehat{C}}^2.$$

Hence, it suffices to verify that $u^{(k)} = m + C^{1/2}\zeta^{(k)}$. Since $\widehat{u}^{(k)} \sim N(\widehat{m}, \widehat{C})$, we can write $\widehat{u}^{(k)} = \widehat{m} + \widehat{C}\eta^{(k)}$, for $\eta^{(k)} \sim N(0, I)$ i.i.d. Moreover, by matching coefficients in $J^{(k)}$, we see that

$$\begin{aligned}(G^T\Gamma^{-1}G + \widehat{C}^{-1})u^{(k)} &= G^T\Gamma^{-1}(y + \Gamma^{1/2}\xi^{(k)}) + \widehat{C}^{-1}\widehat{u}^{(k)} \\ &= \left(G^T\Gamma^{-1}y + \widehat{C}^{-1}\widehat{m}\right) + \left(G^T\Gamma^{-1/2}\xi^{(k)} + \widehat{C}^{-1/2}\eta^{(k)}\right).\end{aligned}$$

Using (10), this can be rewritten as

$$C^{-1}u^{(k)} = \left(G^T\Gamma^{-1}y + \widehat{C}^{-1}\widehat{m}\right) + \left(G^T\Gamma^{-1/2}\xi^{(k)} + \widehat{C}^{-1/2}\eta^{(k)}\right).$$

Now, by (10) and (11) we have that

$$m = C\left(G^T\Gamma^{-1}y + \widehat{C}^{-1}\widehat{m}\right)$$

and moreover, we see that

$$\begin{aligned}&\boldsymbol{E}\left(C(G^T\Gamma^{-1/2}\xi^{(k)} + \widehat{C}^{-1/2}\eta^{(k)}) \otimes C(G^T\Gamma^{-1/2}\xi^{(k)} + \widehat{C}^{-1/2}\eta^{(k)})\right) \\ &\qquad = C\big(G^T\Gamma^{-1}G + \widehat{C}^{-1}\big)C = C.\end{aligned}$$

This completes the proof. □

The analysis step of perturbed observation EnKF fits into the above inverse problem framework, since we are essentially trying to find the conditional distribution of $u_{j+1}|Y_j$ given the observation y_{j+1}. Suppose we are given $\{v_j^{(k)}\}$ and think of this as a sample from an approximation to the distribution of $u_j|Y_j$. Then the ensemble $\{\widehat{v}_{j+1}^{(k)}\} = \{\Psi(v_j^{(k)})\}$ can be thought of as a sample from an approximation to the distribution of $u_{j+1}|Y_j$. Now, define $v_{j+1}^{(k)}$ using the RML method, minimizing the functional

$$J^{(k)}(v) = \tfrac{1}{2}|y_{j+1} - Hv + \xi_{j+1}^{(k)}|_\Gamma^2 + \tfrac{1}{2}|v - \widehat{v}^{(k)}|_{\widehat{C}_{j+1}}^2,$$

where $\xi_{j+1}^{(k)}$ are i.i.d. $N(0, \Gamma)$ and where the covariance $\widehat{C}_{j+1}$ is defined as the empirical covariance

$$\widehat{C}_{j+1} = \frac{1}{K}\sum_{k=1}^{K}(\widehat{v}_{j+1}^{(k)} - \bar{v}_{j+1}) \otimes (\widehat{v}_{j+1}^{(k)} - \bar{v}_{j+1}).$$

This is precisely the EnKF update step described in the algorithm above. There are several reasons that this update step only produces *approximate* samples from the filtering distribution.

First of all, the distribution $u_{j+1}|Y_j$ is certainly not Gaussian in general, unless the dynamics are linear, hence the RML method becomes an approximation of samples. And secondly, since this distribution is not in general Gaussian, the choice of $\widehat{C}_{j+1}$ is another approximation.

Although the approximations outlined are clearly quite naive, the decision to use the empirical distribution instead of say the push-forward of the covariance $u_j|Y_j$ gives a huge advantage to the EnKF in terms of computational efficiency. Moreover, by avoiding linearization, the prediction ensemble exhibits more of the nonlinear dynamical effects present in the underlying model that are present in, say, the extended Kalman filter [14]. However the method as implemented is prone to failures of various kinds and a commonly used way of over-coming one of these, namely collapse of the particles onto a single trajectory, is to use variance inflation. We explain this next.

3.5. Variance inflation

The minimization step of the EnKF computes an update which is a compromise between the model predictions and the data. This compromise is weighted by the empirical covariance on the model and the fixed noise covariance on the data. The model typically allows for unstable (chaotic) divergence of trajectories, whilst the data tends to stabilize. Variance inflation is a technique of adding stability to the algorithm by increasing the size of the model covariance in order to weight the data more heavily. The form of variance inflation that we will study is found by shifting the forecast covariance $\widehat{C}$ by some positive definite matrix. That is, one sets

$$\widehat{C}_{j+1} \mapsto \widehat{C}_{j+1} + A,$$

in (9*a*). Here $A : \mathcal{H} \to \mathcal{H}$ is a linear operator with $A > 0$. Equation (9*a*) becomes

$$(I + (A + \widehat{C}_{j+1})H^T\tilde{\Gamma}^{-1}H)v_{j+1}^{(k)} = \widehat{v}_{j+1}^{(k)} + (A + \widehat{C}_{j+1})H^T\tilde{\Gamma}^{-1}y_{j+1}^{(k)}.$$

This has the effect of weighting the data more than the model. Furthermore, by adding a positive definite operator, one eliminates the null-space of $\widehat{C}_{j+1}$ (which will always be present if the number of ensemble members is smaller than the dimension of $\mathcal{H}$) effectively preventing the ensemble from becoming degenerate. Intuitively speaking, variance inflation ensures the spread of the ensemble is non-zero and prevents collapse onto a false trajectory. A natural choice is $A = \alpha^2 I$ where $\alpha \in \mathbb{R}$ and I is the identity operator. In the sequel it will become clear that variance inflation has the effect of strengthening a contractive term in the algorithm, leading to filter accuracy if α is chosen large enough.

4. Discrete-time estimates

In this section, we will derive long-time estimates for the discrete-time EnKF, under the assumption 2.3 and 2.5 on the dynamics and observation models respectively. We study the algorithm without and then with variance inflation. The technique is to consider evolution of the error between the filter and the true signal underlying the data. To this end we define

$$e_j^{(k)} = v_j^{(k)} - u_j. \tag{12}$$

Throughout this section we use $\boldsymbol{E}$ to denote expectation with respect to the independent i.i.d. noise sequences $\{\xi_j\}$ and $\{\xi_j^{(k)}\}$ and independent initial conditions u_0 and $v_0^{(k)}$.

4.1. Well-posedness without variance inflation

Theorem 4.2. *Let assumptions 2.3 and 2.5 hold and consider the algorithm (9). Then*

$$\boldsymbol{E}\left|e_j^{(k)}\right|^2 \leqslant e^{2\beta h j}\boldsymbol{E}\left|e_0^{(k)}\right|^2 + 2K\gamma^2\left(\frac{e^{2\beta h j}-1}{e^{2\beta h}-1}\right)$$

for any $j \geqslant 1$.

Proof. Firstly note that, under assumption 2.5, the update rule (9*a*) becomes

$$\left(I + \frac{1}{\gamma^2}\widehat{C}_{j+1}\right) v_{j+1}^{(k)} = \Psi(v_j^{(k)}) + \frac{1}{\gamma^2}\widehat{C}_{j+1} y_{j+1}^{(k)}.$$

Secondly note that the underlying signal satisfies

$$\left(I + \frac{1}{\gamma^2}\widehat{C}_{j+1}\right) u_{j+1} = \Psi(u_j) + \frac{1}{\gamma^2}\widehat{C}_{j+1}\Psi(u_j).$$

Thus, subtracting from (9*a*), we obtain

$$\left(I + \frac{1}{\gamma^2}\widehat{C}_{j+1}\right) \mathrm{e}_{j+1}^{(k)} = \Psi(v_j^{(k)}) - \Psi(u_j) + \frac{1}{\gamma^2}\widehat{C}_{j+1}(y_{j+1}^{(k)} - \Psi(u_j)).$$

Now, if we define r_1 and r_2 by

$$\left(I + \frac{1}{\gamma^2}\widehat{C}_{j+1}\right) r_1 = \Psi(v_j^{(k)}) - \Psi(u_j) \tag{13}$$

$$\left(I + \frac{1}{\gamma^2}\widehat{C}_{j+1}\right) r_2 = \frac{1}{\gamma^2}\widehat{C}_{j+1}(y_{j+1}^{(k)} - \Psi(u_j)), \tag{14}$$

then $\mathrm{e}_{j+1}^{(k)} = r_1 + r_2$. Moreover, since $\widehat{C}_{j+1}$ is symmetric and positive semi-definite, we have that

$$\left|\left(I + \frac{1}{\gamma^2}\widehat{C}_{j+1}\right)^{-1}\right| \leqslant 1 \quad \text{and} \quad \left|\left(I + \frac{1}{\gamma^2}\widehat{C}_{j+1}\right)^{-1} \frac{1}{\gamma^2}\widehat{C}_{j+1}\right| \leqslant 1.$$

Note also that $\widehat{C}_{j+1}$ has rank K and let P_{j+1} denote projection into the finite dimensional subspace orthogonal to the kernel of $\widehat{C}_{j+1}$. Then

$$\begin{aligned}\frac{1}{\gamma^2}\widehat{C}_{j+1}(y_{j+1}^{(k)} - \Psi(u_j)) &= \frac{1}{\gamma^2}\widehat{C}_{j+1} P_{j+1}(y_{j+1}^{(k)} - \Psi(u_j)) \\ &= \frac{1}{\gamma^2}\widehat{C}_{j+1} P_{j+1}(\xi_{j+1} + \xi_{j+1}^{(k)}).\end{aligned}$$

It follows from this and from lemma 2.6 that

$$|r_1| \leqslant \left|\Psi(v_j^{(k)}) - \Psi(u_j)\right| \leqslant \mathrm{e}^{\beta h}\left|\mathrm{e}_j^{(k)}\right|,$$

and

$$|r_2| \leqslant \left|y_{j+1}^{(k)} - \Psi(u_j)\right| = \left|P_{j+1}(\xi_{j+1} + \xi_{j+1}^{(k)})\right|.$$

Now, if we let $\mathcal{F}_j$ be the σ-algebra generated by $\{\mathrm{e}_1^{(k)}, \ldots, \mathrm{e}_j^{(k)}\}_{k=1}^K$ then, since r_2 has zero mean and is conditionally independent of r_1, we have

$$\boldsymbol{E}\left(\left|\mathrm{e}_{j+1}^{(k)}\right|^2 | \mathcal{F}_j\right) = |r_1|^2 + \boldsymbol{E}\left|P_{j+1}(\xi_{j+1} + \xi_{j+1}^{(k)})\right|^2 \leqslant \mathrm{e}^{2\beta h}\left|\mathrm{e}_j^{(k)}\right|^2 + 2K\gamma^2.$$

Here we have used the fact that P_{j+1} projects onto a space of dimension at most K. It follows that

$$\boldsymbol{E}\left|\mathrm{e}_{j+1}^{(k)}\right|^2 = \boldsymbol{E}\left(\boldsymbol{E}\left(\left|\mathrm{e}_{j+1}^{(k)}\right|^2 | \mathcal{F}_j\right)\right) \leqslant e^{2\beta h}\boldsymbol{E}\left|\mathrm{e}_j^{(k)}\right|^2 + 2K\gamma^2,$$

and the result follows from the discrete Gronwall inequality. □

The preceding result shows that the EnKF is well-posed and does not blow-up faster than exponentially. We now show that, with the addition of variance inflation, a stronger result can be proved, implying accuracy of the EnKF.

4.3. Accuracy with variance inflation

We will focus on the variance inflation technique with $A = \alpha^2 I$. In this setting, again assuming $H = I$ and $\Gamma = \gamma^2 I$, the EnKF ensemble is governed by the following update equations.

$$\left(I + \frac{\alpha^2}{\gamma^2} I + \frac{1}{\gamma^2}\widehat{C}_{j+1}\right) v_{j+1}^{(k)} = \Psi(v_j^{(k)}) + \left(\frac{\alpha^2}{\gamma^2} I + \frac{1}{\gamma^2}\widehat{C}_{j+1}\right) y_{j+1}^{(k)}. \tag{15}$$

We will now show that with variance inflation, one obtains much stronger long-time estimates than without it. In particular, provided the inflation parameter α is large enough, the ensemble stays within a bounded region of the truth, in a root-mean-square sense.

Theorem 4.4. *Let* $\{v^{(k)}\}_{k=1}^K$ *satisfy (15) and let* $e_j^{(k)} = v_j^{(k)} - u_j$. *Let* $\theta = \frac{\gamma^2}{\gamma^2+\alpha^2} e^{2\beta h}$, *then*

$$\boldsymbol{E}|e_j^{(k)}|^2 \leqslant \theta^j \boldsymbol{E}|e_0^{(k)}|^2 + 2K\gamma^2 \frac{1-\theta^j}{1-\theta},$$

for all $j \in \mathbb{N}$. *In particular, if* $\theta < 1$ *then*

$$\lim_{j\to\infty} \boldsymbol{E}|\mathrm{e}_j^{(k)}|^2 \leqslant \frac{2K\gamma^2}{1-\theta}.$$

Proof. The proof is almost identical to the proof of theorem 4.2. The only difference is that here we use the estimates

$$\left|\left(I + \frac{\alpha^2}{\gamma^2} I + \frac{1}{\gamma^2}\widehat{C}_{j+1}\right)^{-1}\right| \leqslant \frac{\gamma^2}{\alpha^2+\gamma^2} \quad \text{and}$$

$$\left|\left(I + \frac{\alpha^2}{\gamma^2} I + \frac{1}{\gamma^2}\widehat{C}_{j+1}\right)^{-1}\left(\frac{\alpha^2}{\gamma^2} I + \frac{1}{\gamma^2}\widehat{C}_{j+1}\right)\right| \leqslant 1.$$

Proceeding exactly as above, we obtain

$$\boldsymbol{E}|\mathrm{e}_{j+1}^{(k)}|^2 \leqslant \frac{\gamma^2}{\alpha^2+\gamma^2}\mathrm{e}^{\beta h}\boldsymbol{E}|\mathrm{e}_j^{(k)}|^2 + 2K^2\gamma^2 = \theta\boldsymbol{E}|\mathrm{e}_j^{(k)}|^2 + 2K^2\gamma^2,$$

and the result follows from the discrete Gronwall inequality. □

Remark 4.5. For a fixed model, choosing $\alpha^2 > \gamma^2(e^{2\beta h} - 1)$ will result in filter boundedness. Furthermore, if the observational noise standard deviation γ is small then choosing α large enough results in filter accuracy. That is, we can make long-time RMS error $\lim_{j\to\infty} \boldsymbol{E}|e_j^{(k)}|^2$ arbitrarily small. ■

Remark 4.6. The accuracy result implies that, with high probability, the ensemble is concentrated near the truth, for sufficiently large times. However, the perturbed observations typically prevent complete collapse and synchronization of the entire ensemble; instead the ensemble fluctuates around the truth with error scale set by the standard deviation of the observational noise. ■

5. Derivation of the continuous time limit

In this section we formally derive the continuous time scaling limits of the EnKF. The idea is to rearrange the update equation such that it resembles the *discretization* of a stochastic ODE/PDE; we will simply refer to this as an SDE, be it in finite or infinite dimensions. We shall see that non-trivial limits only arise in situations where the noise is rescaled.

First, observe from (9) that

$$\begin{aligned} v_{j+1}^{(k)} - v_j^{(k)} &= \widehat{v}_{j+1}^{(k)} - v_j^{(k)} - \widehat{C}_{j+1} H^T \Gamma^{-1} H v_{j+1}^{(k)} + \widehat{C}_{j+1} H^T \Gamma^{-1} y_{j+1}^{(k)} \\ &= \Psi_h(v_j^{(k)}) - v_j^{(k)} - \widehat{C}_{j+1} H^T \Gamma^{-1} H v_{j+1}^{(k)} + \widehat{C}_{j+1} H^T \Gamma^{-1} y_{j+1}^{(k)} \end{aligned}$$

Now, if we attempt to take $h \to 0$, then the third and fourth terms on the right hand side above will lead to divergences when added up, since they are $O(1)$. This can be avoided by choosing an appropriate rescaling for the noise sources. To this end, let $\Gamma = h^{-s}\Gamma$ for some $s > 0$, then we have

$$v_{j+1}^{(k)} - v_j^{(k)} = \Psi_h(v_j^{(k)}) - v_j^{(k)} - h^s \widehat{C}_{j+1} H^T \Gamma_0^{-1} H v_{j+1}^{(k)} + h^s \widehat{C}_{j+1} H^T \Gamma_0^{-1} y_{j+1}^{(k)}.$$

Now, if we define the primitive z of y by

$$z_{j+1}^{(k)} - z_j^{(k)} = h y_{j+1}^{(k)}$$

then we have coupled difference equations

$$v_{j+1}^{(k)} - v_j^{(k)} = \Psi_h(v_j^{(k)}) - v_j^{(k)} - h^s \widehat{C}_{j+1} H^T \Gamma_0^{-1} H v_{j+1}^{(k)} + h^{s-1} \widehat{C}_{j+1} H^T \Gamma_0^{-1} (z_{j+1}^{(k)} - z_j^{(k)}) \quad (16a)$$

$$z_{j+1}^{(k)} - z_j^{(k)} = h H u_{j+1} + h^{1-s/2} \Gamma_0^{1/2} (\xi_{j+1}^{(k)} + \xi_{j+1}). \quad (16b)$$

The final step is to find an SDE for which the above represents a reasonable numerical scheme. Of course, this depends crucially on the choice of scaling parameter s. In fact, it is not hard to show that the one non-trivial limiting SDEs corresponds to the choice $s = 1$. To see why this is the only valid scaling, notice that (16*a*) implies that $s \geqslant 1$, since otherwise the $O(h^s)$ terms would diverge when added up. Likewise, from the second equation we must have $1 - s/2 \geqslant 1/2$, for otherwise the stochastic terms would diverge when summed up, in accordance with the central limit theorem. Hence we must choose $s = 1$.

If we invoke the approximation

$$\Psi_h(v) - v \approx h F(v)$$

then, in the case $s = 1$, the system (16*a*) is a mixed implicit–explicit Euler–Maruyama type scheme for the SDE

$$\mathrm{d}v^{(k)} = F(v^{(k)})\,\mathrm{d}t - C(v) H^T \Gamma_0^{-1} H v^{(k)}\,\mathrm{d}t + C(v) H^T \Gamma_0^{-1} \mathrm{d}z^{(k)} \quad (17a)$$

$$\mathrm{d}z^{(k)} = H u\,\mathrm{d}t + \Gamma_0^{1/2} (\mathrm{d}W^{(k)} + \mathrm{d}B). \quad (17b)$$

Here $W^{(1)}, \ldots W^{(K)}$, B are pairwise independent cylindrical Wiener processes, arising as limiting processes of the discrete increments $\xi^{(1)}, \ldots \xi^{(K)}$, ξ respectively. We use v to denote the collection $\{v^{(k)}\}_{k=1}^K$ and the operator $C(v)$ is the empirical covariance of the particles defined as follows:

$$\bar{v} = \frac{1}{K} \sum_{k=1}^{K} v^{(k)} \quad (18a)$$

$$C(v) = \frac{1}{K} \sum_{k=1}^{K} \left(v^{(k)} - \bar{v} \right) \otimes \left(v^{(k)} - \bar{v} \right). \quad (18b)$$

Thus we have the system of SDEs (17) for $k = 1, \ldots, K$, coupled together through (18).

Remark 5.1. If we substitute the expression for $dz^{(k)}$ from (17*b*) into (17*a*) then we obtain

$$\mathrm{d}v^{(k)} = F(v^{(k)})\,\mathrm{d}t - C(v) H^T \Gamma_0^{-1} H (v^{(k)} - u)\,\mathrm{d}t + C(v) H^T \Gamma_0^{-1/2} (\mathrm{d}W^{(k)} + \mathrm{d}B). \quad (19)$$

Of course in practice the truth u is not known to us, but the equation (19) has a very clear structure which highlights the mechanisms at play in the EnKF. The equation is given by

the original dynamics with the addition of two terms, one which pulls the solution of each ensemble member back towards the true signal u, and a second which drives each ensemble member with a sum of two white noises, one independently chosen for each ensemble member (coming from the perturbed observations) and the second a common noise (coming from the noise in the data).

The stabilizing term, which draws the ensemble member back towards the truth, and the noise, both act only orthogonal to the null-space of the empirical covariance of the set of particles. The perturbed observations noise contribution will act to prevent the particles from synchronizing which, in their absence, could happen. If the particles were to synchronize then the covariance disappears and we simply obtain the original dynamics

$$\mathrm{d}v^{(k)} = F(v^{(k)})\,\mathrm{d}t \tag{20}$$

for each ensemble member.

In this context it is worth noting that another approach to the derivation of a continuous time limit is to never introduce the process z and only think of the equation for $v^{(k)}$. In particular, we have

$$\begin{aligned} v_{j+1}^{(k)} - v_j^{(k)} &= \Psi_h(v_j^{(k)}) - v_j^{(k)} - h^s \widehat{C}_{j+1} H^T \Gamma_0^{-1} H v_{j+1}^{(k)} \\ &\quad + h^s \widehat{C}_{j+1} H^T \Gamma_0^{-1} \left(H u_{j+1} + h^{-s/2} \Gamma_0^{1/2} (\xi_{j+1}^{(k)} + \xi_{j+1}) \right). \end{aligned}$$

In this case, we still must have $s \geqslant 1$ in order to get a limit, but there is no requirement for $s \leqslant 1$. However, it is easy to see that in the case $s > 1$, one obtains the trivial scaling limit (20) so that each ensemble member evolves according to the model dynamics and the data is not seen. Such scalings are of no interest since they do not elucidate the structure of the model/data trade-off which is the heart of the EnKF.. ■

5.2. Limits with variance inflation

With variance inflation, the update equation (16*a*) becomes

$$\begin{aligned} v_{j+1}^{(k)} - v_j^{(k)} &= \Psi_h(v_j^{(k)}) - v_j^{(k)} - h^s (A + \widehat{C}_{j+1}) H^T \Gamma_0^{-1} H v_{j+1}^{(k)} \\ &\quad + h^{s-1} (A + \widehat{C}_{j+1}) H^T \Gamma_0^{-1} (z_{j+1}^{(k)} - z_j^{(k)}) \\ z_{j+1}^{(k)} - z_j^{(k)} &= h H u_{j+1} + h^{1-s/2} \Gamma_0^{1/2} (\xi_{j+1}^{(k)} + \xi_{j+1}). \end{aligned} \tag{21}$$

By the same reasoning, it is clear that the only non-trivial continuous time limit is given by

$$\begin{aligned} \mathrm{d}v^{(k)} &= F(v^{(k)})\,\mathrm{d}t - (A + C(v)) H^T \Gamma_0^{-1} H v^{(k)}\,\mathrm{d}t + (A + C(v)) H^T \Gamma_0^{-1} \mathrm{d}z^{(k)} \\ \mathrm{d}z^{(k)} &= H u\,\mathrm{d}t + \Gamma_0^{1/2} (\mathrm{d}W^{(k)} + \mathrm{d}B). \end{aligned} \tag{22}$$

6. Continuous-time estimates

In this section we obtain long-time estimates for the continuous time EnKF, under assumptions 2.3 and 2.5. These are the same assumptions used in the discrete case and our results are analogous to theorems 4.2. Under assumptions 2.3 and 2.5, the continuous time EnKF equations (19) for the ensemble $v = \{v^{(k)}\}_{k=1}^K$ become

$$\begin{aligned} &\mathrm{d}v^{(k)}(t) + \left(\mathcal{A} v^{(k)}(t) + \mathcal{B}(v^{(k)}(t), v^{(k)}(t)) \right) \mathrm{d}t \\ &\quad = f - \frac{1}{\gamma^2} C(v)(v^{(k)}(t) - u(t))\,\mathrm{d}t + \frac{1}{\gamma} C(v)(\mathrm{d}W^{(k)}(t) + \mathrm{d}B(t)). \end{aligned} \tag{23}$$

We set $e^{(k)} = v^{(k)} - u$, and write e for $e = \{e^{(k)}\}_{k=1}^K$. Note that $C(v) = C(e)$ since shifting the origin does not change the empirical covariance. Thus we have

$$\mathrm{d}\mathrm{e}^{(k)}(t) + \left(\mathcal{A}\mathrm{e}^{(k)} + \mathcal{B}(\mathrm{e}^{(k)}, \mathrm{e}^{(k)}) + 2\mathcal{B}(\mathrm{e}^{(k)}, u)\right)\mathrm{d}t = -\frac{1}{\gamma^2}C(e)\mathrm{e}^{(k)}\mathrm{d}t + \frac{1}{\gamma}C(e)(\mathrm{d}W^{(k)}(t) + \mathrm{d}B(t)).$$

Remark 6.1. In the next theorem, we will analyse the growth properties of solutions to the SPDE (23), but to make the statement precise we must specify what we mean by a solution. We use the standard notion of a strong solution as found in [7]. In essence, we assume that the solution is strong enough so that all the terms in the integral expression

$$\begin{aligned} v^{(k)}(t) + \int_0^t \left(\mathcal{A}v^{(k)}(s) + \mathcal{B}(v^{(k)}(s), v^{(k)}(s))\right)\mathrm{d}s \\ = v^{(k)}(0) + \int_0^t \left(f - \frac{1}{\gamma^2}C(v(s))(v^{(k)}(s) - u(s))\right)\mathrm{d}s \\ + \frac{1}{\gamma}\int_0^t C(v(s))(\mathrm{d}W^{(k)}(s) + \mathrm{d}B(s)) \end{aligned} \tag{24}$$

are well defined and moreover, fall into the domain of Itô's formula. To be precise, we say that $v = \{v^{(k)}\}_{k=1}^K$ is a strong solution to (23) over the interval $[0, T]$ if v satisfies (24) ($\boldsymbol{P}$-a.s.) and moreover we have that

$$\int_0^T \left|\mathcal{A}v^{(k)}(t)\right| + \left|\mathcal{B}(v^{(k)}(t), v^{(k)}(t))\right| + \left|C(v)(v^{(k)}(t) - u(t))\right|\mathrm{d}t < \infty \quad \boldsymbol{P}\text{-a.s.}$$

and

$$\int_0^T \|C(v(t))\|_{HS}^2\,\mathrm{d}t < \infty \quad \boldsymbol{P}\text{-a.s}$$

where $\|\cdot\|_{HS}$ is the Hilbert–Schmidt norm. As can be seen in [7, theorem 4.17], these conditions are sufficient to utilize Itô's formula. We note that, in the case of (23), it is not unreasonable to assume the existence of strong solutions. Indeed, in finite dimensions any type of solution will be a strong solution and moreover, the very existence of a Lyapunov function, as obtained in the theorem below, is enough to guarantee global solutions [23]. In infinite dimensions this is not the case in general. However the fact that the noise is effectively finite dimensional, due to the presence of the finite rank covariance operator C, does mean that existence of strong solutions may well be established on a case-by-case basis in some infinite dimensional settings. However it is difficult to do this at the level of generality we study in this paper and hence we will not make any concrete statements concerning the existence of strong solutions, rather we will simply assume that one exists and is unique. ■

We may now state and prove the well-posedness estimate for this equation, analogous to theorem 4.2. We assume that (23) has a unique solution for all $t \geqslant 0$. In finite dimensions this is in fact a consequence of the mean square estimate provided by the theorem; since we have been unable to prove this in the rather general infinite dimensional setting, however, we make it an assumption. We let $(\Omega, \mathcal{F}, \boldsymbol{P})$ denote the probability space underlying the independent initial conditions and the driving Brownian motions $(\{W^{(k)}\}, B)$, and $\boldsymbol{E}$ denotes expectation with respect to this space.

Theorem 6.2. *Assume that (23) has a unique strong solution, in the sense of remark 6.1 and that assumptions 2.3 and 2.5 hold. We then have that*

$$\frac{1}{K}\sum_{k=1}^K \boldsymbol{E}\left|\mathrm{e}^{(k)}(t)\right|^2 \leqslant \left(\frac{1}{K}\sum_{k=1}^K \boldsymbol{E}\left|\mathrm{e}^{(k)}(0)\right|^2\right)\exp\left(\frac{4(cR)^2 t}{\lambda}\right),$$

where $c > 0$ is the constant appearing in (5). Moreover, we have that

$$\frac{1}{K}\sum_{k=1}^{K}\int_0^t \boldsymbol{E}\left\|\mathrm{e}^{(k)}(s)\right\|^2 \mathrm{d}s \leqslant \left(\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{E}\left|\mathrm{e}^{(k)}(0)\right|^2\right)\frac{1}{\lambda}\exp\left(\frac{4(cR)^2 t}{\lambda}\right).$$

Proof. Using Itô's formula, one can show that

$$\begin{aligned}\boldsymbol{E}\left|\mathrm{e}^{(k)}(t)\right|^2 = \boldsymbol{E}\left|\mathrm{e}^{(k)}(0)\right|^2 &+ \int_0^t -2\boldsymbol{E}\left\langle \mathrm{e}^{(k)}, \mathcal{A}\mathrm{e}^{(k)} + \mathcal{B}(\mathrm{e}^{(k)}, \mathrm{e}^{(k)}) + 2\mathcal{B}(\mathrm{e}^{(k)}, u)\right\rangle \mathrm{d}s \\ &+ \int_0^t \frac{-2}{\gamma^2}\boldsymbol{E}\left\langle \mathrm{e}^{(k)}, C(e)\mathrm{e}^{(k)}\right\rangle + \frac{2}{\gamma^2}\boldsymbol{E}\mathrm{tr}\left(C(e)^2\right)\mathrm{d}s. \end{aligned} \tag{25}$$

Now, if we let $\{\xi_i\}_{i\in I}$ be some orthonormal basis of $\mathcal{H}$, then we can simplify the above using the identity

$$\mathrm{tr}\left(C(e)^2\right) = \sum_{i\in I}\langle C(e)\xi_i, C(e)\xi_i\rangle.$$

By expanding the right $C(e)$, we obtain

$$\begin{aligned}\mathrm{tr}\left(C(e)^2\right) &= \sum_{i\in I}\frac{1}{K}\sum_{m=1}^{K}\left\langle \mathrm{e}^{(m)} - \bar{e}, \xi_i\right\rangle\left\langle \mathrm{e}^{(m)}, C(e)\xi_i\right\rangle \\ &= \frac{1}{K}\sum_{m=1}^{K}\sum_{i\in I}\left\langle \mathrm{e}^{(m)} - \bar{e}, \xi_i\right\rangle\left\langle C(e)\mathrm{e}^{(m)}, \xi_i\right\rangle \\ &= \frac{1}{K}\sum_{m=1}^{K}\left\langle \mathrm{e}^{(m)} - \bar{e}, C(e)\mathrm{e}^{(m)}\right\rangle \\ &= \frac{1}{K}\sum_{m=1}^{K}\left\langle \mathrm{e}^{(m)}, C(e)\mathrm{e}^{(m)}\right\rangle - \langle\bar{e}, C(e)\bar{e}\rangle.\end{aligned}$$

Substituting this into (25) and summing over $k = 1\ldots K$, we obtain

$$\begin{aligned}\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{E}\left|\mathrm{e}^{(k)}(t)\right|^2 = \frac{1}{K}\sum_{k=1}^{K}\boldsymbol{E}\left|\mathrm{e}^{(k)}(0)\right|^2 &\\ + \frac{1}{K}\sum_{k=1}^{K}\int_0^t -2\boldsymbol{E}\left\langle \mathrm{e}^{(k)}, \mathcal{A}\mathrm{e}^{(k)} + \mathcal{B}(\mathrm{e}^{(k)}, \mathrm{e}^{(k)}) + 2\mathcal{B}(\mathrm{e}^{(k)}, u)\right\rangle \mathrm{d}s &\\ - \frac{2}{\gamma^2}\int_0^t \boldsymbol{E}\langle\bar{e}, C(e)\bar{e}\rangle\,\mathrm{d}s.&\end{aligned}$$

And since $C(e)$ is positive semi-definite, we have

$$\begin{aligned}\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{E}\left|\mathrm{e}^{(k)}(t)\right|^2 \leqslant \frac{1}{K}\sum_{k=1}^{K}\boldsymbol{E}\left|\mathrm{e}^{(k)}(0)\right|^2 &\\ + \frac{1}{K}\sum_{k=1}^{K}\int_0^t -2\boldsymbol{E}\left\langle \mathrm{e}^{(k)}, \mathcal{A}\mathrm{e}^{(k)} + \mathcal{B}(\mathrm{e}^{(k)}, \mathrm{e}^{(k)}) + 2\mathcal{B}(\mathrm{e}^{(k)}, u)\right\rangle \mathrm{d}s.&\end{aligned}$$

Finally, using the assumptions on $\mathcal{A}$, $\mathcal{B}$, we have that

$$\begin{aligned}-\left\langle \mathrm{e}^{(k)}, \mathcal{A}\mathrm{e}^{(k)} + \mathcal{B}(\mathrm{e}^{(k)}, \mathrm{e}^{(k)}) + 2\mathcal{B}(\mathrm{e}^{(k)}, u)\right\rangle &\leqslant -\lambda\left\|\mathrm{e}^{(k)}\right\|^2 + 2\left|\left\langle \mathcal{B}(\mathrm{e}^{(k)}, u), \mathrm{e}^{(k)}\right\rangle\right| \\ &\leqslant -\lambda\left\|\mathrm{e}^{(k)}\right\| + 2c\left|\mathrm{e}^{(k)}\right|\left\|\mathrm{e}^{(k)}\right\|\|u\| \\ &\leqslant -\lambda\left\|\mathrm{e}^{(k)}\right\| + cR\left(\delta^{-1}\left|\mathrm{e}^{(k)}\right|^2 + \delta\left\|\mathrm{e}^{(k)}\right\|^2\right),\end{aligned}$$

recalling that $\sup_{t\geqslant 0}\|u(t)\|\leqslant R$. Putting this altogether, we have that

$$\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{E}\left|\mathrm{e}^{(k)}(t)\right|^2+\frac{1}{K}\sum_{k=1}^{K}\int_0^t 2\left(\lambda-cR\delta\right)\boldsymbol{E}\left\|\mathrm{e}^{(k)}(s)\right\|^2\mathrm{d}s$$
$$\leqslant\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{E}\left|\mathrm{e}^{(k)}(0)\right|^2+\frac{1}{K}\sum_{k=1}^{K}\int_0^t 2cR\delta^{-1}\boldsymbol{E}\left|\mathrm{e}^{(k)}(s)\right|^2\mathrm{d}s.$$

If we pick $\delta=\lambda/(2cR)$ then we obtain the estimate

$$\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{E}\left|\mathrm{e}^{(k)}(t)\right|^2+\frac{1}{K}\sum_{k=1}^{K}\int_0^t \lambda\boldsymbol{E}\left\|\mathrm{e}^{(k)}(s)\right\|^2\mathrm{d}s$$
$$\leqslant\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{E}\left|\mathrm{e}^{(k)}(0)\right|^2+\frac{1}{K}\sum_{k=1}^{K}\int_0^t \frac{4(cR)^2}{\lambda}\boldsymbol{E}\left|\mathrm{e}^{(k)}(s)\right|^2\mathrm{d}s,$$

and the result follows from Gronwall's inequality. As a consequence of this, we see that

$$\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\int_0^t \lambda\boldsymbol{E}\left\|\mathrm{e}^{(k)}(s)\right\|^2\mathrm{d}s\\
&\leqslant\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{E}\left|\mathrm{e}^{(k)}(0)\right|^2+\frac{1}{K}\sum_{k=1}^{K}\int_0^t \frac{4(cR)^2}{\lambda}\boldsymbol{E}\left|\mathrm{e}^{(k)}(s)\right|^2\mathrm{d}s\\
&\leqslant\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{E}\left|\mathrm{e}^{(k)}(0)\right|^2+\frac{4(cR)^2}{\lambda}\left(\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{E}\left|\mathrm{e}^{(k)}(0)\right|^2\right)\int_0^t\exp\left(\frac{4(cR)^2 s}{\lambda}\right)\mathrm{d}s\\
&=\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{E}\left|\mathrm{e}^{(k)}(0)\right|^2\exp\left(\frac{4(cR)^2 t}{\lambda}\right),
\end{aligned}$$

which proves the second result and hence the theorem. □

Remark 6.3. In this case of non-trivial H and Γ, the above argument does not work, but nevertheless it is still informative to see *why* it does not work. Indeed, if we apply the exact same argument to the case of arbitrary H, Γ, we still obtain the identity

$$\begin{aligned}
\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{E}\left|\mathrm{e}^{(k)}(t)\right|^2=\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{E}\left|\mathrm{e}^{(k)}(0)\right|^2&\\
+\frac{1}{K}\sum_{k=1}^{K}\int_0^t -2\boldsymbol{E}\left\langle\mathrm{e}^{(k)},\mathcal{A}\mathrm{e}^{(k)}+\mathcal{B}(\mathrm{e}^{(k)},\mathrm{e}^{(k)})+2\mathcal{B}(\mathrm{e}^{(k)},u)\right\rangle\mathrm{d}s&\\
-2\int_0^t\boldsymbol{E}\left\langle\bar{e},C(e)H^T\Gamma^{-1}H\bar{e}\right\rangle\mathrm{d}s.&
\end{aligned}$$

The reason we cannot proceed further is that even though $C(e)$ and $H\Gamma^{-1}H$ are themselves positive semi-definite and self adjoint, the same is not necessarily true for the product. ■

7. Numerical results

In this section we confirm the validity of the theorems derived in the previous sections for variants of the EnKF when applied to the dynamical system (2). Furthermore, we extend

our numerical explorations beyond the strict range of validity of the theory and, in particular, consider the case of partial observations. We conduct all of our numerical experiments in the case of the incompressible Navier–Stokes equation on a two- dimensional (2D) torus.

We observe not only well-posedness, but indeed *boundedness* of the ensemble for both complete and partial observations, over long time-scales compared with the natural variability of the dynamical system itself. However, the filter is always inaccurate when used without inflation. We thus turn to study the effect of inflation and note that our results indicate the filter can then always be made accurate, even in the case of partial observations, provided that sufficiently many low Fourier modes are observed. In the case that only the high Fourier modes are observed the filter cannot be made accurate with inflation.

7.1. Setup

Let $\mathbb{T}^2$ denote the 2D torus of side $L : [0, L) \times [0, L)$ with periodic boundary conditions. We consider the equations

$$\begin{aligned}&\partial_t u(x,t) - \nu\Delta u(x,t) + u(x,t)\cdot\nabla u(x,t) + \nabla p(x,t) = f(x)\\&\nabla\cdot u(x,t) = 0\\&u(x,0) = u_0(x)\end{aligned}$$

for all $x \in \mathbb{T}^2$ and $t \in (0,\infty)$. Here $u\colon \mathbb{T}^2 \times (0,\infty) \to \mathbb{R}^2$ is a time-dependent vector field representing the velocity, $p\colon \mathbb{T}^2 \times (0,\infty) \to \mathbb{R}$ is a time-dependent scalar field representing the pressure and $f\colon \mathbb{T}^2 \to \mathbb{R}^2$ is a vector field representing the forcing which we take as time-independent for simplicity. The parameter ν represents the viscosity. We assume throughout that u_0 and f have average zero over $\mathbb{T}^2$; it then follows that $u(\cdot, t)$ has average zero over $\mathbb{T}^2$ for all $t > 0$.

Define

$$\mathsf{T} := \left\{\text{trigonometric polynomials } u : \mathbb{T}^2 \to \mathbb{R}^2 \;\middle|\; \nabla\cdot u = 0,\ \int_{\mathbb{T}^2} u(x)\,\mathrm{d}x = 0\right\}$$

and $\mathcal{H}$ as the closure of T with respect to the norm in $(L^2(\mathbb{T}^2))^2 = L^2(\mathbb{T}^2, \mathbb{R}^2)$. We let $P : (L^2(\mathbb{T}^2))^2 \to \mathcal{H}$ denote the Leray–Helmholtz orthogonal projector. Given $m = (m_1, m_2)^{\mathrm{T}}$, define $m^\perp := (m_2, -m_1)^{\mathrm{T}}$. Then an orthonormal basis for (a complexified) $\mathcal{H}$ is given by $\psi_m\colon \mathbb{T}^2 \to \mathbb{C}^2$, where

$$\psi_m(x) := \frac{m^\perp}{|m|}\exp\Big(\frac{2\pi i m\cdot x}{L}\Big) \tag{26}$$

for $m \in \mathbb{Z}^2 \setminus \{0\}$. Thus for $u \in \mathcal{H}$ we may write

$$u(x) = \sum_{m\in\mathbb{Z}^2\setminus\{0\}} u_m\psi_m(x)$$

where, since u is a real-valued function, we have the reality constraint $u_{-m} = -\overline{u_m}$. We define the projection operators $\mathcal{P}_\lambda : \mathcal{H} \to \mathcal{H}$ and $\mathcal{Q}_\lambda : \mathcal{H} \to \mathcal{H}$ for $\lambda \in \mathbb{N} \cup \{\infty\}$ by

$$\big(\mathcal{P}_\lambda u\big)(x) = \sum_{|2\pi m|^2 < \lambda L^2} u_m\psi_m(x), \qquad \mathcal{Q}_\lambda = I - \mathcal{P}_\lambda.$$

Below we will choose the observation operator H to be $\mathcal{P}_\lambda$ or $\mathcal{Q}_\lambda$.

We define $A = -\nu P\Delta$ the Stokes operator, and, for every $s \in \mathbb{R}$, define the Hilbert spaces $\mathcal{H}^s$ to be the domain of $A^{s/2}$. We note that A is diagonalized in $\mathcal{H}$ in the basis comprised of the $\{\psi_m\}_{m\in\mathbb{Z}^2\setminus\{0\}}$. We denote by $|\cdot|$ the norm on $\mathcal{H} := \mathcal{H}^0$.

Applying the projection P to the Navier–Stokes equation for $f = Pf$ we may write it as an ODE in $\mathcal{H}$ as in (2), with $\mathcal{B}(u, v)$ the *symmetric* bilinear form defined by

$$\mathcal{B}(u, v) = \tfrac{1}{2}P(u \cdot \nabla v) + \tfrac{1}{2}P(v \cdot \nabla u)$$

for all $u, v \in \mathcal{V}$. See [5] for details of this formulation of the Navier–Stokes equation as an ODE in $\mathcal{H}$.

We fix the domain size $L = 2$. The forcing in f is taken to be $f \propto \nabla^{\perp}\Psi$, where $\Psi = \cos(\pi k_f \cdot x)$ and $\nabla^{\perp} = J\nabla$ with J the canonical skew-symmetric matrix. The parameters $(\nu, k_f, |f|)$ in (2), where k_f is the wavevector of the forcing frequency, are fixed throughout all of the experiments shown in this paper at values which yield a chaotic regime; specifically we take $(\nu, k_f, |f|) \approx \{0.01, (5, 5), 10\}$.

Our first step in constructing a numerical experiment is to compute the true solution u solving equation (2). The true initial condition u_0 is randomly drawn from $N(0, \nu^2 A^{-2})$. The true solution is computed with $M = 32^2$ non-zero Fourier coefficients. Padding is included in the spectral domain to avoid aliasing, resulting in $4M$-dimensional FFTs. For all the experiments presented below, we will then begin with an initial ensemble which is far from the truth, in order to probe the accuracy and stability of the filter for the given parameters. Specifically we let $m_0 \sim N(u(0), \beta\frac{4\pi^2\nu}{L^2}A^{-1})$ and $v_0^{(k)} \sim N(m_0, (\beta/25)\frac{4\pi^2\nu}{L^2}A^{-1})$ with $\beta = 0.25$. Throughout, $\gamma = 0.01$. We use the notation $m(t)$ to denote the mean of the ensemble.

The method used to approximate the forward model is a modification of a fourth-order Runge–Kutta method, ETD4RK [6], in which the Stokes semi-group is computed exactly by working in the incompressible Fourier basis $\{\psi_m(x)\}_{m\in\mathbb{Z}^2\setminus\{0\}}$, and Duhamel's principle (variation of constants formula) is used to incorporate the nonlinear term. We use a time-step of $dt = 0.005$. Spatially, a Galerkin spectral method [12] is used, in the same basis, and the convolutions arising from products in the nonlinear term are computed via FFTs.

Before proceeding with the numerical experiments concerning the EnKF, it is instructive to run an experiment in which $H = 0$ so that the ensemble evolves according to the underlying attractor with no observations taken into account. This is shown in figure 1. Notice that the statistics of the ensemble remain well-behaved. This is in stark contrast to the case of evolving extended Kalman filter without observations, in which case the covariance will have an exponential growth rate corresponding asymptotically to the Lyapunov exponents of the attractor. Figure 1 sets a reference scale against which subsequent experiments, which include observations, should be compared. The results of all experiments are summarized in table 1, in terms of root mean square error of the mean with respect to the truth:

$$RMSE = \sqrt{\frac{1}{N}\sum_{j=1}^{N}|m_j - u(t_j)|^2}, \qquad m_j = \frac{1}{K}\sum_{k=1}^{K} v^{(k)}(t_j).$$

7.2. Discrete time

Here we explore the case of discrete-time observations by means of numerical experiments, illustrating the results of section 4. We consider the cases $H = \mathcal{P}_\lambda$ with $\lambda = \infty$ so that all Fourier modes represented on the grid are observed, as well as both $H = \mathcal{P}_\lambda$ and $H = \mathcal{Q}_\lambda$ with $\lambda < \infty$.

7.2.1. Full observations. Here we consider observations made at all numerically resolved, and hence observable, wavenumbers in the system; hence $M = 32^2$, not including padding in the

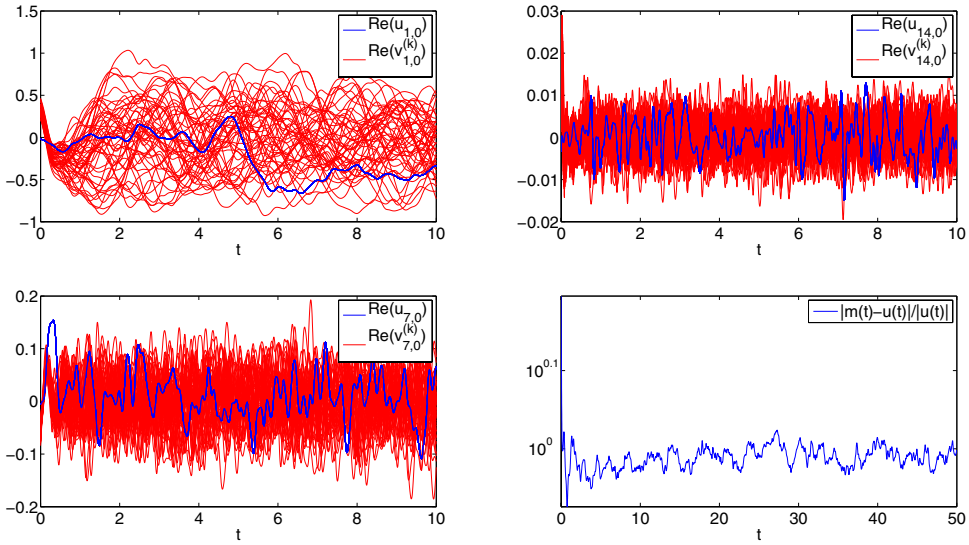

Figure 1. Trajectories of various modes (at observation times) of the estimators $v^{(k)}$ and the signal u are depicted above along with the relative error in the L^2 norm, $|m-u|/|u|$, for $H = \mathcal{P}_\lambda$, with $\lambda = 0$—i.e. nothing is observed.

Table 1. Root mean squared error for each of the cases investigated here. Acronyms indicate the following cases: free evolution $H = 0$ (Z), discrete observation (D), continuous observation (C), inflation (I), full observation $H = \mathcal{P}_\lambda$ with $\lambda = \infty$ (F), partial inner observation $H = \mathcal{P}_\lambda$ with $|k_\lambda| = 5$ (IN), and partial outer observation $H = \mathcal{Q}_\lambda$ with $|k_\lambda| = 5$ (OUT).

Method	RMSE
Z	2.1217
D–F	2.7357
D–F–I	0.2144
D–IN–I	0.1851
D–OUT–I	2.7693
C–F	2.7584
C–F–I	0.3899
C–IN–I	0.5166
C–OUT–I	2.7863

spectral domain which avoids aliasing. So, effectively we approximate the case $H = \mathcal{P}_\lambda$ where $\lambda = \infty$. Observations of the full-field are made every $J = 20$ time-steps. In figure 2 there is no variance inflation and, whilst typical ensemble members remain bounded on the time-scales shown, the error between the ensemble mean and the truth is $\mathcal{O}(1)$; indeed comparison with figure 1 shows that the error in the mean is in fact *worse* than that of an ensemble evolving without access to data. This may be attributed to the fact that the ensemble loses track of the signal, and its spread collapses to zero. The result is that the error between the mean of this ensemble and the signal is more like the average error between any two observation-free trajectories, rather than the error between a given observation-free trajectory (the signal) and the mean over observation-free trajectories as in the bottom right panel of figure 1. The averaging of the latter leads to marginally smaller error. This can be observed in the table 1.

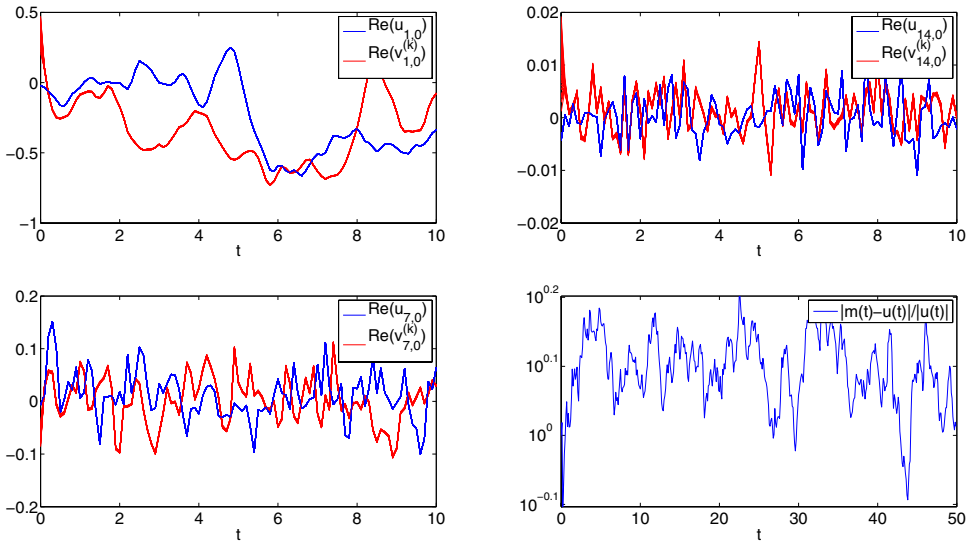

Figure 2. Discrete-time observations, without inflation. Trajectories of various modes (at observation times) of the estimators $v^{(k)}$ and the signal u are depicted above along with the relative error in the L^2 norm, $|m-u|/|u|$, for $H = \mathcal{P}_\lambda$, with $\lambda = \infty$.

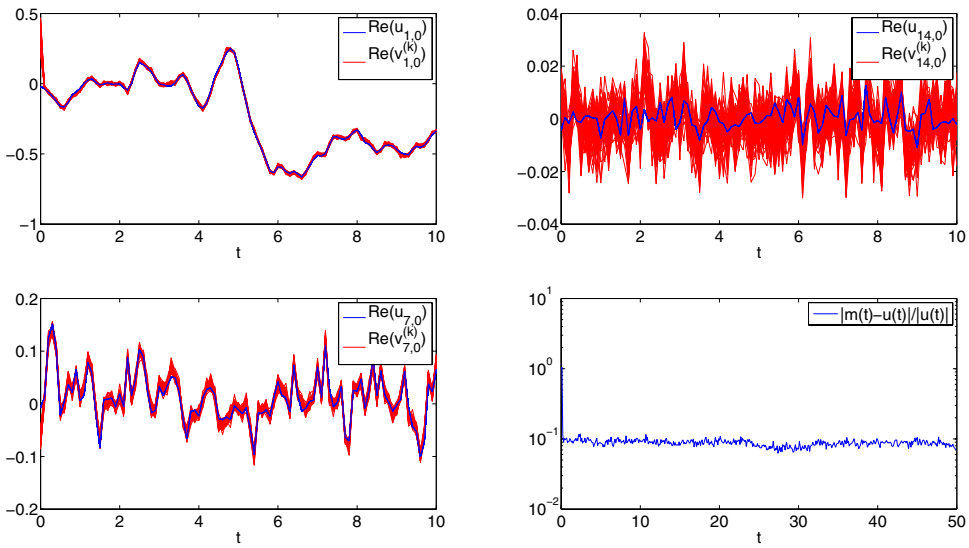

Figure 3. Discrete-time observations, with inflation. Trajectories of various modes (at observation times) of the estimators $v^{(k)}$ and the signal u are depicted above along with the relative error in the L^2 norm, $|m-u|/|u|$, for $H = \mathcal{P}_\lambda$, with $\lambda = \infty$.

Using variance inflation removes this problem and filter accuracy is obtained: see figure 3. The inflation parameter is chosen as $\alpha^2 = 0.0025$.

7.2.2. Partial observations. In this section, the observations are again made every $J = 20$ time-steps, but we will now consider observing only projections inside and outside a ring of

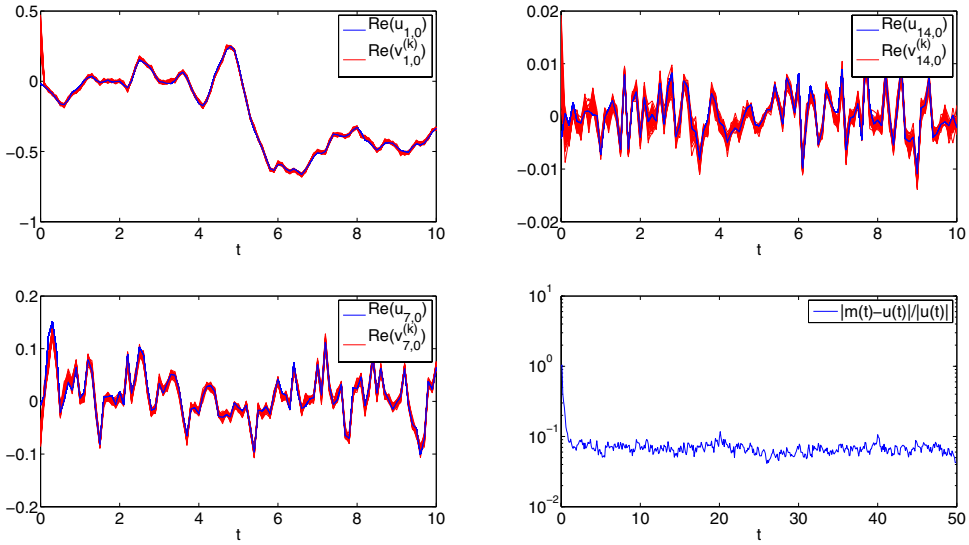

Figure 4. Discrete-time observations, with inflation. Trajectories of various modes (at observation times) of the estimators $v^{(k)}$ and the signal u are depicted above along with the relative error in the L^2 norm, $|m-u|/|u|$, for $H=\mathcal{P}_\lambda$, with $|k_\lambda|=5$ and $J=20$.

radius $|k_\lambda|=5$ in Fourier space. In other words, we consider two cases $H=\mathcal{P}_\lambda$ and $H=\mathcal{Q}_\lambda$ where $\lambda=\pi^2|k_\lambda|^2$ and $|k_\lambda|=5$. This is outside the validity of the theory which considers only full observations. Inflation is used in both cases. The inflation parameter is again chosen as $\alpha^2=0.0025$. Figure 4 shows that when observing all Fourier modes inside a ring of radius $|k_\lambda|=5$ the filter is accurate over long time-scales. In contrast, figure 5 shows that observing all Fourier modes outside a ring of radius $|k_\lambda|=5$ does not provide enough information to induce accurate filtering.

7.3. Continuous time

In this section we study the SPDE (23), and its relation to the underlying truth governed by (2), by means of numerical experiments. We thereby illustrate and extend the results of section 6. We invoke a split-step scheme to solve equation (23), in which for each ensemble member $v^{(k)}$ one step of numerical integration of the Navier–Stokes equation (2) is composed with one step of numerical integration of the stochastic process

$$\begin{aligned}
&\mathrm{d}v^{(k)}+\frac{1}{\gamma^2}C(v)(v^{(k)}(t)-u(t))\,\mathrm{d}t=\frac{1}{\gamma}C(v)(\mathrm{d}W^{(k)}(t)+\mathrm{d}B(t))\\
&v^{(k)}(0)=v_0^{(k)},\quad m=\frac{1}{K}\sum_{k=1}^{K}v_0^{(k)}\quad C(v)=\frac{1}{K}\sum_{k=1}^{K}(v_0^{(k)}-m)\otimes(v_0^{(k)}-m)
\end{aligned}\tag{27}$$

at each step. The Navier–Stokes equation 2 itself is solved by the method described in section 7.1. The stochastic process is also diagonalized in the Fourier basis (26) and then time-approximated by the Euler–Maruyama scheme [16]. We consider the cases $H=\mathcal{P}_\lambda$ with $\lambda=\infty$ so that all Fourier modes represented on the grid are observed, as well as both $H=\mathcal{P}_\lambda$ and $H=\mathcal{Q}_\lambda$ with $\lambda<\infty$.

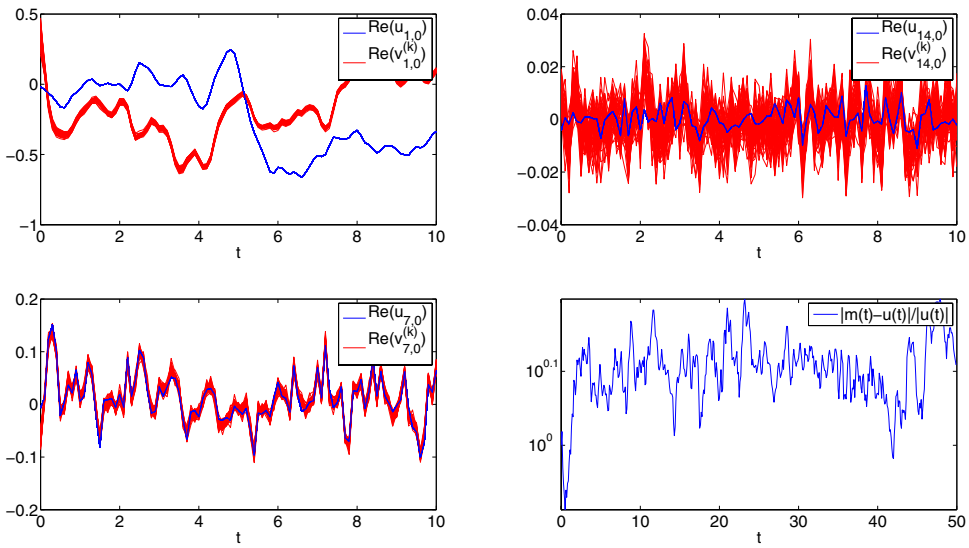

Figure 5. Discrete-time observations, with inflation. Trajectories of various modes (at observation times) of the estimators $v^{(k)}$ and the signal u are depicted above along with the relative error in the L^2 norm, $|m-u|/|u|$, for $H = \mathcal{Q}_\lambda$, with $|k_\lambda| = 5$ and $J = 20$.

7.3.1. Full observations. Here we consider observations made at all numerically resolved, and hence observable, wavenumbers in the system, as in discrete time. So, effectively we approximate the case $H = \mathcal{P}_\lambda$ where $\lambda = \infty$. Figure 6 shows that, without inflation, the ensemble remains bounded, but the mean is inaccurate, on the time-scales of interest. In contrast figure 7 demonstrates that inflation leads to accurate reconstruction of the truth via the ensemble mean. The inflation parameter is chosen as $\alpha^2 = 0.00025$.

7.3.2. Partial observations. Here we consider two cases again, as in section 7.2.2 $H = \mathcal{P}_\lambda$ and $H = \mathcal{Q}_\lambda$ where $\lambda = \pi^2|k_\lambda|^2$, with $|k_\lambda| = 5$. Inflation is used in both cases and the inflation parameter is again chosen as $\alpha^2 = 0.00025$. As for discrete time-observations we see that observing inside a ring in Fourier space leads to filter accuracy (figure 8) whilst observing outside yields only filter boundedness (figure 9.)

8. Conclusions

We have developed a method for the analysis of the EnKF. Instead of viewing it as an algorithm designed to accurately approximate the true filtering distribution, which it cannot do, in general, outside Gaussian scenarios and in the large ensemble limit, we study it as an algorithm for signal estimation in the finite (possibly small) ensemble limit. We show well-posedness of the filter and, when suitable variance inflation is used, mean-square asymptotic accuracy in the large-time limit. These positive results about the EnKF are encouraging and serve to underpin its perceived effectiveness in applications. On the other hand it is important to highlight that our analysis applies only to fully observed dynamics and interesting open questions remain concerning the partially observed case. In this regard it is important to note that the filter divergence observed in [11, 19] concerns partially observed models. Thus more analysis remains to be done in this area. The tools introduced herein may be useful in this regard.

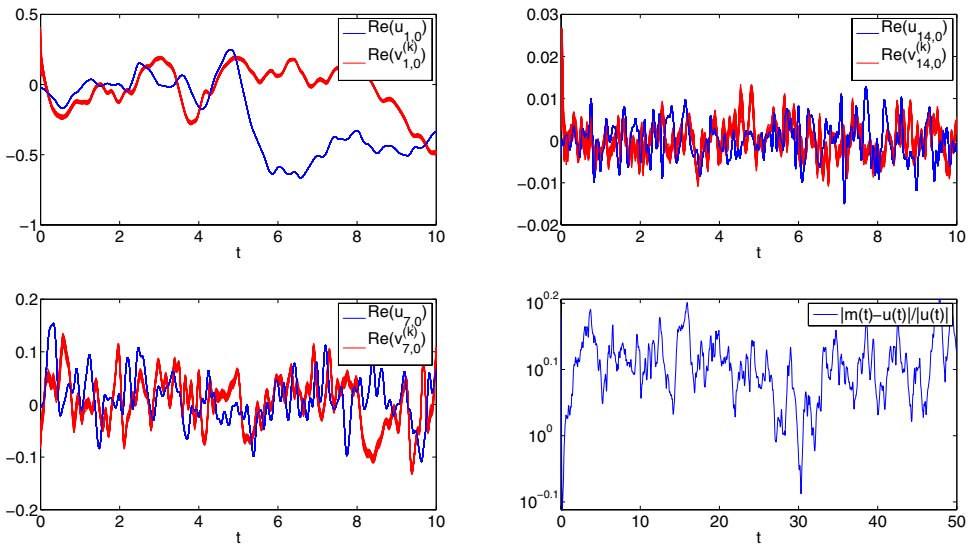

Figure 6. Continuous-time observations, without inflation. Trajectories of various modes of the estimators $v^{(k)}$ and the signal u are depicted above along with the relative error in the L^2 norm, $|v^{(1)} - u|/|u|$, for $H = \mathcal{P}_\lambda$, with $\lambda = \infty$.

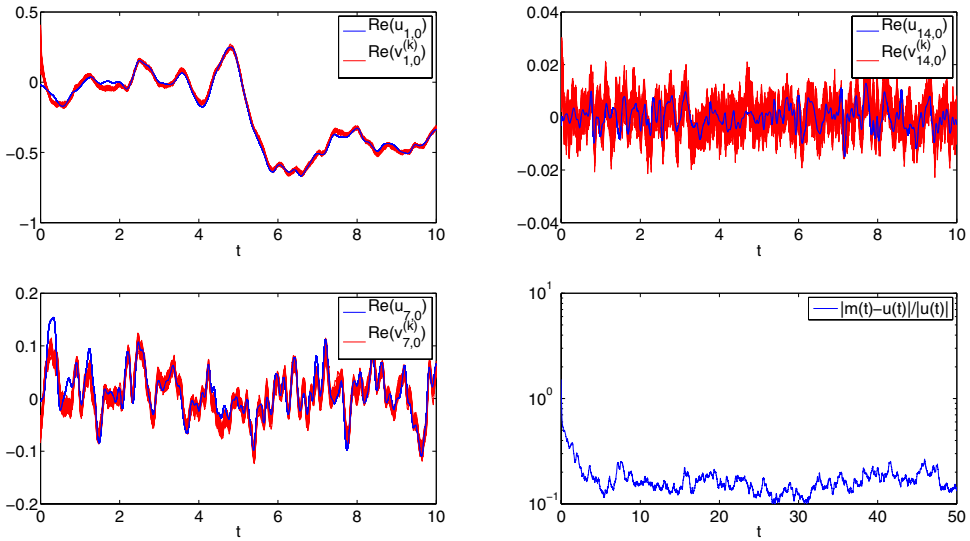

Figure 7. Continuous-time observations, with inflation. Trajectories of various modes of the estimators $v^{(k)}$ and the signal u are depicted above along with the relative error in the L^2 norm, $|v^{(1)} - u|/|u|$, for $H = \mathcal{P}_\lambda$, with $\lambda = \infty$.

A second important direction in which the analysis could usefully be extended is the class of models to which it applies. We have studied dissipative quadratic dynamical systems with energy conserving nonlinearities. These are of direct relevance in the atmospheric sciences [15] but more general models will be required for subsurface applications such as those arising in oil reservoir simulation [25]. The theoretical results have been confirmed

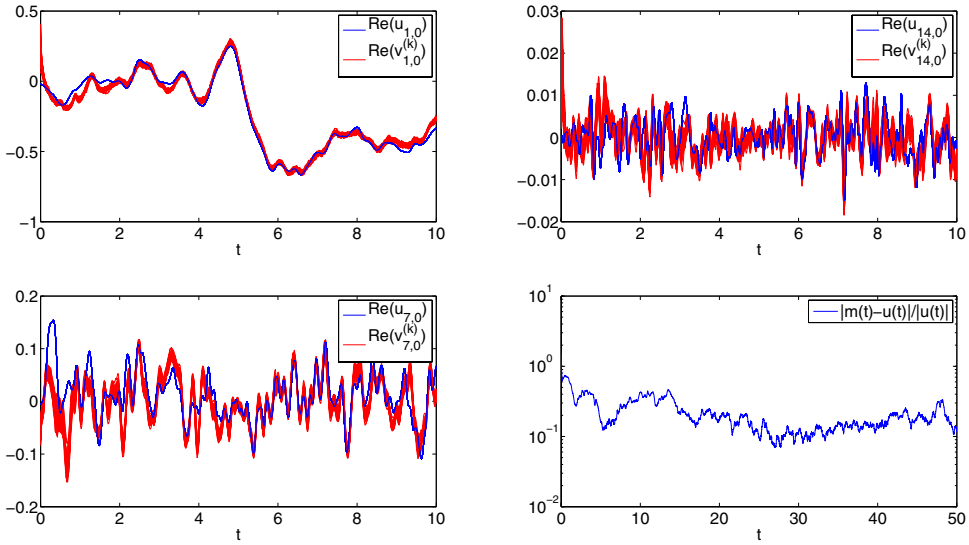

Figure 8. Continuous-time observations, with inflation. Trajectories of various modes of the estimators $v^{(k)}$ and the signal u are depicted above along with the relative error in the L^2 norm, $|v^{(1)} - u|/|u|$, for $H = \mathcal{P}_\lambda$, with $|k_\lambda| = 5$.

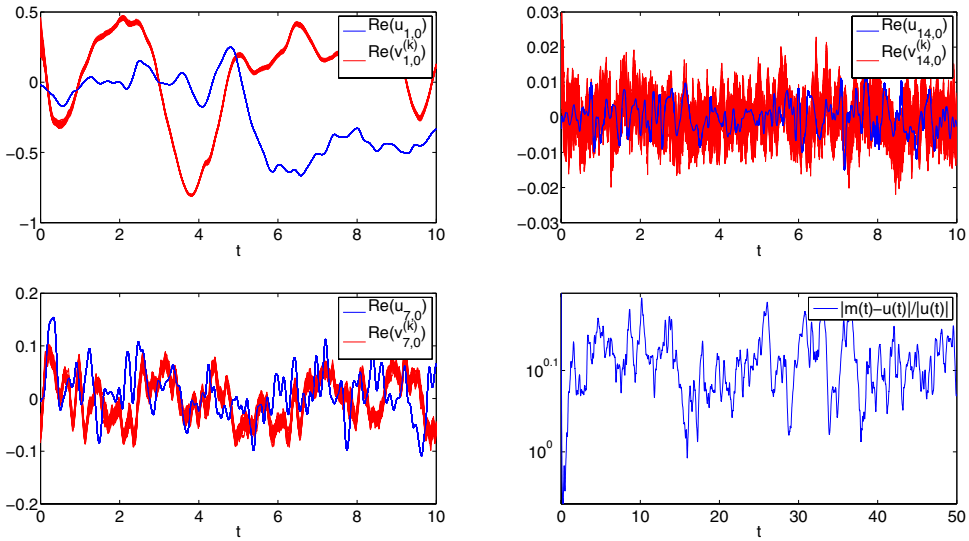

Figure 9. Continuous-time observations, with inflation. Trajectories of various modes of the estimators $v^{(k)}$ and the signal u are depicted above along with the relative error in the L^2 norm, $|v^{(1)} - u|/|u|$, for $H = \mathcal{Q}_\lambda$, with $|k_\lambda| = 5$.

with numerical simulations of the Navier–Stokes equation on a torus. These numerical results demonstrate two interesting potential extensions of our theory: (i) to strengthen well-posedness to obtain boundedness of trajectories, at least in mean square; (ii) to extend well-posedness and accuracy results to certain partial observation scenarios. Furthermore we highlight the fact that our results have assumed exact solution of the underlying differential equation model;

understanding how filtering interacts with numerical approximations, and potentially induces numerical instabilities, is a subject which requires further investigation; this issue is highlighted in [11].

Acknowledgments

The authors are grateful to A J Majda for helpful discussions concerning this work. DTBK is supported by ONR grant N00014-12-1-0257. The work of AMS is supported by ERC, EPSRC, ESA and ONR. KJHL was supported by the King Abdullah University of Science and Technology (KAUST) and is a member of the KAUST Strategic Research Initiative Center for Uncertainty Quantification.

References

[1] Anderson J L 2007 An adaptive covariance inflation error correction algorithm for ensemble filters *Tellus* A **59** 210–24
[2] Bloemker D, Law K J H, Stuart A M and Zygalalkis K 2013 Accuracy and stability of the continuous-time 3DVAR filter for the Navier–Stokes equation *Nonlinearity* **26** 2193
[3] Brett C E A, Lam K F, Law K J H, McCormick D S, Scott M R and Stuart A M 2012 Accuracy, stability of filters for dissipative PDEs *Phys. D: Nonlinear Phenom.* **245** 34–45
[4] Burgers G, Van P J and Evensen G 1998 On the analysis scheme in the ensemble Kalman filter *Mon. Weather Rev.* **126** 1719–24
[5] Constantin P and Foiaş C 1988 *Navier–Stokes Equations* (Chicago, IL: University of Chicago Press)
[6] Cox S M and Matthews P C 2002 Exponential time differencing for stiff systems *J. Comput. Phys.* **176** 430–55
[7] DaPrato G and Zabczyk J 1992 Stochastic equations in infinite dimensions *Encyclopedia of Mathematics and its Applications* vol 44 (Cambridge: Cambridge University Press)
[8] Evensen G 2006 *Data Assimilation: The Ensemble Kalman Filter* (Berlin: Springer)
[9] Evensen G and Van Leeuwen P J 2000 An ensemble Kalman smoother for nonlinear dynamics *Mon. Weather Rev.* **128** 1852–67
[10] Evensen G, Burgers G and van Leeuwen P J 1998 Analysis scheme in the ensemble kalman filter *Mon. Weather Rev.* **126** 1719–24
[11] Gottwald G A and Majda A J 2013 A mechanism for catastrophic filter divergence in data assimilation for sparse observation networks *Nonlinear Process. Geophys.* **20** 705–12
[12] Hesthaven J S, Gottlieb S and Gottlieb D 2007 *Spectral Methods for Time-dependent Problems* vol 21 (Cambridge: Cambridge University Press)
[13] Ide K, Kuznetsov L and Jones C K R T 2002 Lagrangian data assimilation for point-vortex system *J. Turbul.* **3** 53
[14] Jazwinski A H 1970 *Stochastic Processes and Filtering Theory* vol 63 (New York: Academic)
[15] Kalnay E 2003 *Atmospheric Modeling, Data Assimilation and Predictability* (Cambridge: Cambridge University Press)
[16] Kloeden P E and Platen E 1992 Numerical solution of stochastic differential equations *Applications of Mathematics* vol 23 (Berlin: Springer)
[17] Law K J H, Shukla A and Stuart A M 2014 Analysis of the 3DVAR filter for the partially observed Lorenz '63 model *Discrete Continuous Dyn. Syst.* A **34** 1061–78
[18] François Le Gland *et al* 2011 Large sample asymptotics for the ensemble Kalman filter *Oxford Handbook of Nonlinear Filtering* ed D Crisan and B Rozovskii (Oxford: Oxford University Press) pp 598–631
[19] Majda A J and Harlim J 2008 Catastrophic filter divergence in filtering nonlinear dissipative systems *Commun. Math. Sci.* **8** 27–43
[20] Majda A J and Harlim J 2012 *Filtering Complex Turbulent Systems* (Cambridge: Cambridge University Press)
[21] Majda A J and Wang X 2006 *Nonlinear Dynamics and Statistical Theories for Geophysical Flows* (Cambridge: Cambridge University Press)
[22] Mandel J, Cobb L and Beezley J D 2011 On the convergence of the ensemble Kalman filter. *Appl. Math.* **56** 533–41

[23] Mao X 1997 *Stochastic Differential Equations, their Applications (Horwood Publishing Series in Mathematics and Applications)* (Chichester: Horwood)
[24] Moodey A J F, Lawless A S, Potthast R W E and van Leeuwen P J 2013 Nonlinear error dynamics for cycled data assimilation methods *Inverse Problems* **29** 025002
[25] Oliver D S, Reynolds A C and Liu N 2008 *Inverse Theory for Petroleum Reservoir Characterization and History Matching* (Cambridge: Cambridge University Press)
[26] Temam R 1997 Infinite-dimensional dynamical systems in mechanics, physics *Applied Mathematical Sciences* vol 68, 2nd edn (New York: Springer)