WARWICK
THE UNIVERSITY OF WARWICK

Original citation:
Schillings, Claudia and Stuart, A. M.. (2017) Analysis of the ensemble Kalman filter for inverse problems. SIAM Journal on Numerical Analysis, 55 (3). pp. 1264-1290.

Permanent WRAP URL:
http://wrap.warwick.ac.uk/96051

Copyright and reuse:
The Warwick Research Archive Portal (WRAP) makes this work by researchers of the University of Warwick available open access under the following conditions. Copyright © and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable the material made available in WRAP has been checked for eligibility before being made available.

Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

Publisher's statement:
© 2017 SIAM Journal on Numerical Analysis
http://dx.doi.org/10.1137/16M105959X
A note on versions:
The version presented here may differ from the published version or, version of record, if you wish to cite this item you are advised to consult the publisher's version. Please see the 'permanent WRAP URL' above for details on accessing the published version and note that access may require a subscription.

For more information, please contact the WRAP Team at: wrap@warwick.ac.uk

warwick.ac.uk/lib-publications

arXiv:1602.02020v2 [math.NA] 20 Sep 2016

ANALYSIS OF THE ENSEMBLE KALMAN FILTER FOR INVERSE PROBLEMS

CLAUDIA SCHILLINGS* AND ANDREW M. STUART†

Abstract. The ensemble Kalman filter (EnKF) is a widely used methodology for state estimation in partial, noisily observed dynamical systems, and for parameter estimation in inverse problems. Despite its widespread use in the geophysical sciences, and its gradual adoption in many other areas of application, analysis of the method is in its infancy. Furthermore, much of the existing analysis deals with the large ensemble limit, far from the regime in which the method is typically used. The goal of this paper is to analyze the method when applied to inverse problems with fixed ensemble size. A continuous-time limit is derived and the long-time behavior of the resulting dynamical system is studied. Most of the rigorous analysis is confined to the linear forward problem, where we demonstrate that the continuous time limit of the EnKF corresponds to a set of gradient flows for the data misfit in each ensemble member, coupled through a common pre-conditioner which is the empirical covariance matrix of the ensemble. Numerical results demonstrate that the conclusions of the analysis extend beyond the linear inverse problem setting. Numerical experiments are also given which demonstrate the benefits of various extensions of the basic methodology.

Key words. Bayesian Inverse Problems, Ensemble Kalman Filter, Optimization

AMS subject classifications. 65N21, 62F15, 65N75

1. Introduction.

The Ensemble Kalman filter (EnKF) has had enormous impact on the applied sciences since its introduction in the 1990s by Evensen and coworkers; see [11] for an overview. It is used for both data assimilation problems, where the objective is to estimate a partially observed time-evolving system sequentially in time [17], and inverse problems, where the objective is to estimate a (typically distributed) parameter appearing in a differential equation [25]. Much of the analysis of the method has focussed on the large ensemble limit [24, 23, 13, 20, 10, 22]. However the primary reason for the adoption of the method by practitioners is its robustness and perceived effectiveness when used with small ensemble sizes, as discussed in [2, 3] for example. It is therefore important to study the properties of the EnKF for fixed ensemble size, in order to better understand current practice, and to suggest future directions for development of the method. Such fixed ensemble size analyses are starting to appear in the literature for both data assimilation problems [19, 28] and inverse problems [15, 14, 16]. In this paper we analyze the EnKF for inverse problems, adding greater depth to our understanding of the basic method, as formulated in [15], as well as variants on the basic method which employ techniques such as variance inflation and localization (see [21] and the references therein), together with new ideas (introduced here) which borrow from the use of sequential Monte Carlo (SMC) method for inverse problems introduced in [18].

Let $\mathcal{G} : \mathcal{X} \to \mathcal{Y}$ be a continuous mapping between separable Hilbert spaces $\mathcal{X}$ and $\mathcal{Y}$. We are interested in the inverse problem of recovering unknown u from observation y where

$$y = \mathcal{G}(u) + \eta.$$

Here η is an observational noise. We are typically interested in the case where the in-

*Mathematics Institute, University of Warwick, Coventry CV4 7AL, UK (c.schillings@warwick.ac.uk).

†Mathematics Institute, University of Warwick, Coventry CV4 7AL, UK (a.m.stuart@warwick.ac.uk).

version is ill-posed on $\mathcal{Y}$, i.e. one of the following three conditions is violated: existence of solutions, uniqueness, stability. In the linear setting, we can think for example of a compact operator violating the stability condition. Indeed throughout we assume in all our rigorous results, without comment, that $\mathcal{Y} = \mathbb{R}^K$ for $K \in \mathbb{N}$ except in a few particular places where we explicitly state that $\mathcal{Y}$ is infinite dimensional. A key role in such inverse problems is played by the least squares functional

$$\Phi(u; y) = \frac{1}{2}\|\Gamma^{-\frac{1}{2}}(y - \mathcal{G}(u))\|_{\mathcal{Y}}^2.$$

Here $\Gamma > 0$ normalizes the model-data misfit and often knowledge of the covariance structure of typical noise η is used to define Γ.

When the inverse problem is ill-posed, infimization of Φ in $\mathcal{X}$ is not a well-behaved problem and some form of regularization is required. Classical methods include Tikhonov regularization, infimization over a compact ball in $\mathcal{X}$ and truncated iterative methods [9]. An alternative approach is Bayesian regularization. In Bayesian regularization (u, y) is viewed as a jointly varying random variable in $\mathcal{X} \times \mathcal{Y}$ and, assuming that $\eta \sim N(0, \Gamma)$ is independent of $u \sim \mu_0$, the solution to the inverse problem is the $\mathcal{X}-$valued random variable $u|y$ distributed according to measure

$$\mu(du) = \frac{1}{Z}\exp\big(-\Phi(u; y)\big)\mu_0(du), \tag{1}$$

where Z is chosen so that μ is a probability measure:

$$Z := \int_{\mathcal{X}} \exp\big(-\Phi(u; y)\big)\mu_0(du).$$

See [7] for details concerning the Bayesian methodology.

The EnKF is derived within the Bayesian framework and, through its ensemble properties, is viewed as approximating the posterior distribution on the random variable $u|y$. However, except in the large sample limit for linear problems [23, 13, 20] there is little to substantiate this viewpoint; indeed the paper [10] demonstrates this quite clearly by showing that for nonlinear problems the large ensemble limit does not approximate the posterior distribution. In [22], a related result is proved for the EnKF in the context of data assimilaion; in the large ensemble size limit the EnKF is proved to converge to the mean-field EnKF, which provides the optimal linear estimator of the conditional mean, but does not reproduce the filtering distribution, except in the linear Gaussian case. A different perspective on the EnKF is that it constitutes a derivative-free optimization technique, with the ensemble used as a proxy for derivative information. This optimization viewpoint was adopted in [15, 14] and is the one we take in this paper: through analysis and numerical experiments we study the properties of the EnKF as a regularization technique for minimization of the least-squares misfit functional Φ at fixed ensemble size. We do, however, use the Bayesian perspective to derive the algorithm, and to suggest variants of it.

In section 2 we describe the EnKF in its basic form, deriving the algorithm by means of ideas from SMC applied to the Bayesian inverse problem, together with invocation of a Gaussian approximation. Section 3 describes continuous time limits of the method, leading to differential equations, and in section 4 we study properties of the differential equations derived in the linear case and, in particular, their long-time behavior. Using this analysis we obtain clear understanding of the sense in which the EnKF is a regularized optimization method for Φ. Indeed we show in section 3 that

the continuous time limit of the EnKF corresponds to a set of preconditioned gradient flows for the data misfit in each ensemble member. The common preconditioner is the empirical covariance matrix of the ensemble which thereby couples the ensemble members together and renders the algorithm nonlinear, even for linear inverse problems. Section 5 is devoted to numerical studies which illustrate the foregoing theory for linear problems, and which also demonstrate that similar ideas apply to nonlinear problems. In section 6 we discuss variants of the basic EnKF method, in particular the addition of variance inflation, localization or the use of random search methods based on SMC, within the ensemble method; all of these methods break the invariant subspace property of the basic EnKF proved in [15] and we explore numerically the benefits of doing so.

2. The EnKF for Inverse Problems. Here we describe how to derive the iterative EnKF as an approximation of the SMC method for inverse problems. Recall the posterior distribution μ given by (1) and define the probability measures μ_n by, for $h = N^{-1}$,

$$\mu_n(du) \propto \exp\bigl(-nh\Phi(u;y)\bigr)\mu_0(du).$$

The measures μ_n are intermediate measures defined via likelihoods scaled by the step size $h = N^{-1}$. It follows that $\mu_N = \mu$ the desired measure on $u|y$. Then

$$\mu_{n+1}(du) = \frac{1}{Z_n}\exp\bigl(-h\Phi(u;y)\bigr)\mu_n(du), \tag{2}$$

for

$$Z_n = \int \exp(-h\Phi(u;y))\mu_n(du).$$

Denoting by L_n the nonlinear operator corresponding to application of Bayes' theorem to map from μ_n to μ_{n+1} we have

$$\mu_{n+1} = L_n\mu_n\ . \tag{3}$$

We have introduced an articial discrete time dynamical system which maps the prior μ_0 into the posterior $\mu_N = \mu$. A heuristic worthy of note is that although we look at the data y at each of N steps, the effective variance is amplified by $N = 1/h$ at each step, compensating for the redundant, repeated use of the data. The idea of SMC is to approximate μ_n by a weighted sum of Dirac masses: given a set of particles and weights $\{u_n^{(j)}, w_n^{(j)}\}_{j=1}^J$ the approximation takes the form

$$\mu_n \simeq \sum_{j=1}^{J} w_n^{(j)}\delta_{u_n{}^{(j)}},$$

with $\delta_{u_n{}^{(j)}}$ denoting the delta-Dirac mass located at $u_n{}^{(j)}$. The method is defined by the mapping of the particles and weights at time n to those at time $n+1$. The method is introduced for Bayesian inverse problems in [18] where it is used to study the problem of inferring the initial condition of the Navier-Stokes equations from data. In [4] the method is applied to the inverse problem of determining the coefficient of a divergence form elliptic PDE from linear functionals of the solution; furthermore the method is also proved to converge in the large particle limit $J \to \infty$.

In practice the SMC method can perform poorly. This happens when the weights $\{w_n^{(j)}\}_{j=1}^J$ degenerate in that one of the weights takes a value close to one and all others

are negligible. The EnKF aims to counteract this by always seeking an approximation in the form

$$\mu_n \simeq \frac{1}{J}\sum_{j=1}^{J} \delta_{u_n}^{(j)}$$

and thus

$$\mu_{n+1} \simeq \frac{1}{J}\sum_{j=1}^{J} \delta_{u_{n+1}}^{(j)} .$$

The method is defined by the mapping of the particles at time n into those at time $n+1$. Let $u_n = \{u_n^{(j)}\}_{j=1}^J$. Then using equation (25) in [15] with $\Gamma \mapsto h^{-1}\Gamma$ shows that this mapping of particles has the form

$$u_{n+1}^{(j)} = u_n^{(j)} + C^{up}(u_n)(C^{pp}(u_n) + h^{-1}\Gamma)^{-1}\big(y_{n+1}^{(j)} - \mathcal{G}(u_n^{(j)})\big), \quad j = 1, \cdots, J, \tag{4}$$

where

$$y_{n+1}^{(j)} = y + \xi_{n+1}^{(j)}$$

and, for $u = \{u^{(j)}\}_{j=1}^J$, we define the operators C^{pp} and C^{up} by

$$\begin{aligned} C^{pp}(u) &= \frac{1}{J}\sum_{j=1}^{J}\big(\mathcal{G}(u^{(j)}) - \overline{\mathcal{G}}\big) \otimes \big(\mathcal{G}(u^{(j)}) - \overline{\mathcal{G}}\big), \\ C^{up}(u) &= \frac{1}{J}\sum_{j=1}^{J}\big(u^{(j)} - \overline{u}\big) \otimes \big(\mathcal{G}(u^{(j)}) - \overline{\mathcal{G}}\big), \\ \overline{u} &= \frac{1}{J}\sum_{j=1}^{J} u^{(j)}, \quad \overline{\mathcal{G}} = \frac{1}{J}\sum_{j=1}^{J} \mathcal{G}(u^{(j)}). \end{aligned} \tag{5}$$

We will consider both the cases where $\xi_{n+1}^{(j)} \equiv 0$ and where, with respect to both j and n, the $\xi_{n+1}^{(j)}$ are i.i.d. random variables distributed according to $N(0, h^{-1}\Gamma)$. We can unify by considering the i.i.d. family of random variables $\xi_{n+1}^{(j)} \sim N(0, h^{-1}\Sigma)$ and focussing exclusively on the cases where $\Sigma = \Gamma$ and where $\Sigma = 0$. The theoretical results will be solely derived for the setting $\Sigma = 0$, i.e. no artificial noise will be added to the observational data.

The derivation of the EnKF as presented here relies on a Gaussian approximation, which can be interpreted as a linearization of the nonlinear operator L_n in the following way: in the large ensemble size limit, the EnKF estimate corresponds to the best linear estimator of the conditional mean. See [12] for a general discussion of Bayes linear methods and [22] for details in the context of data assimilation. Thus, besides the approximation of the measures μ_n by a J particle Dirac measure, there is an additional uncontrolled error resulting from the Gaussian approximation, and analyzed in [10]. In [27], the EnKF in combination with an annealing process is used to account for nonlinearities in the forward problem by weight-correcting the EnKF. For data assimilation problems, similar techniques can be applied to improve the performance of the EnKF in the nonlinear regime, see e.g. [5] and the references therein for more details.

In summary, except in the Gaussian case of linear problems, there is no convergence to μ_n as $J \to \infty$. Our focus, then, is on understanding the properties of the algorithm for fixed J, as an optimization method; we do not study the approximation of the measure μ. In this context we also recall here the invariant subspace property of the EnKF method, as established in [15]:

LEMMA 1. *If $\mathcal{S}$ is the linear span of $\{u_0^{(j)}\}_{j=1}^J$ then $u_n^{(j)} \in \mathcal{S}$ for all $(n,j) \in \mathbb{Z}^+ \times \{1,\dots,J\}$.*

3. Continuous Time Limit. Here we study a continuous time limit of the EnKF methodology as applied to inverse problems; this limit arises by taking the parameter h, appearing in the incremental formulation (2) of the Bayesian inverse problem (1), to zero. We proceed purely formally, with no proofs of the limiting process, as our main aim in this paper is to study the behavior of the continuous time limits, not to justify taking that limit. However we note that the invariant subspace property of Lemma 1 means that the desired limit theorems are essentially finite dimensional and standard methods from numerical analysis may be used to establish the limits. In the next section all the theoretical results are derived under the assumption that $\mathcal{G}$ is linear and $\Sigma = 0$. However in this section we derive the continuous time limit in a general setting, before specifying to the linear noise-free case.

3.1. The Nonlinear Problem. Recall the definition of the operator C^{up} given by (5). We recall that $u_n = \{u_n^{(j)}\}_{j=1}^J$, and assume that $u_n \approx u(nh)$ in (4) in the limit $h \to 0$. The update step of the EnKF (4) can be written in the form of a time-stepping scheme:

$$\begin{aligned} u_{n+1}^{(j)} &= u_n^{(j)} + hC^{up}(u_n)(hC^{pp}(u_n)+\Gamma)^{-1}\big(y - \mathcal{G}(u_n^{(j)})\big) \\ &\quad + hC^{up}(u_n)(hC^{pp}(u_n)+\Gamma)^{-1}\xi_{n+1}^{(j)} \\ &= u_n^{(j)} + hC^{up}(u_n)(hC^{pp}(u_n)+\Gamma)^{-1}\big(y - \mathcal{G}(u_n^{(j)})\big) \\ &\quad + h^{\frac{1}{2}}C^{up}(u_n)(hC^{pp}(u_n)+\Gamma)^{-1}\sqrt{\Sigma}\zeta_{n+1}^{(j)}\,, \end{aligned}$$

where $\zeta_{n+1}^{(j)} \sim \mathcal{N}(0,I)$ i.i.d.. If we take the limit $h \to 0$ then this is clearly a tamed Euler-Maruyama type discretization of the set of coupled Itô SDEs

$$\frac{\mathrm{d}u^{(j)}}{\mathrm{d}t} = C^{up}(u)\Gamma^{-1}(y - \mathcal{G}(u^{(j)})) + C^{up}(u)\Gamma^{-1}\sqrt{\Sigma}\frac{dW^{(j)}}{dt}. \tag{6}$$

Using the definition of the operator C^{up} we see that

$$\frac{\mathrm{d}u^{(j)}}{\mathrm{d}t} = \frac{1}{J}\sum_{k=1}^{J}\big\langle \mathcal{G}(u^{(k)}) - \overline{\mathcal{G}}, y - \mathcal{G}(u^{(j)}) + \sqrt{\Sigma}\frac{dW^{(j)}}{dt}\big\rangle_\Gamma\big(u^{(k)} - \overline{u}\big), \tag{7}$$

where

$$\overline{u} = \frac{1}{J}\sum_{j=1}^{J} u^{(j)}, \quad \overline{\mathcal{G}} = \frac{1}{J}\sum_{j=1}^{J}\mathcal{G}(u^{(j)})$$

and $\langle\cdot,\cdot\rangle_\Gamma = \langle\Gamma^{-\frac{1}{2}}\cdot,\Gamma^{-\frac{1}{2}}\cdot\rangle$ with $\langle\cdot,\cdot\rangle$ the inner-product on $\mathcal{Y}$. The $W^{(j)}$ are independent cylindrical Brownian motions on $\mathcal{X}$. The construction demonstrates that, provided a solution to (7) exists, it will satisfy a generalization of the subspace property of Lemma 1 to continuous time because the vector field is in the linear span of the ensemble itself.

3.2. The Linear Noise-Free Problem. In this subsection we study the linear inverse problem, for which $\mathcal{G}(\cdot) = A\cdot$ for some $A \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$. We also restrict attention to the case where $\Sigma = 0$. Then the continuous time limit equation (7) becomes

$$\frac{\mathrm{d}u^{(j)}}{\mathrm{d}t} = \frac{1}{J}\sum_{k=1}^{J}\langle A(u^{(k)} - \overline{u}), y - Au^{(j)}\rangle_{\Gamma}\big(u^{(k)} - \overline{u}\big), \quad j = 1, \cdots, J. \tag{8}$$

For $u = \{u^{(j)}\}_{j=1}^{J}$ we define the empirical covariance operator

$$C(u) = \frac{1}{J}\sum_{k=1}^{J}\big(u^{(k)} - \overline{u}\big) \otimes \big(u^{(k)} - \overline{u}\big).$$

Then equation (8) may be written in the form

$$\frac{\mathrm{d}u^{(j)}}{\mathrm{d}t} = -C(u)D_u\Phi(u^{(j)}; y) \tag{9}$$

where, in this linear case,

$$\Phi(u; y) = \frac{1}{2}\|\Gamma^{-\frac{1}{2}}(y - Au)\|^2. \tag{10}$$

Thus each particle performs a preconditioned gradient descent for $\Phi(\cdot; y)$, and all the individual gradient descents are coupled through the preconditioning of the flow by the empirical covariance $C(u)$. This gradient flow is thus nonlinear, even though the forward map is linear. Using the fact that C is positive semi-definite it follows that

$$\frac{\mathrm{d}}{\mathrm{d}t}\Phi\big(u(t); y\big) = \frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{2}\|\Gamma^{-\frac{1}{2}}(y - Au)\|^2 \leq 0.$$

This gives an a priori bound on $\|Au(t)\|_{\Gamma}$, but does not give global existence of a solution when $\Gamma^{-\frac{1}{2}}A$ is compact. However, global existence may be proved as we show in in the next section, using the invariant subspace property.

4. Asymptotic Behavior in the Linear Setting. In this section we study the differential equations (8). Although derivation of the continuous time limit suggests stopping the integration at time $T = 1$ it is nonetheless of interest to study the dynamical system in the long-time asymptotic $T \to \infty$ as this sheds light on the mechanisms at play within the ensemble methodology, and points to possible improvements of the algorithm.

In the first subsection 4.1 we study the case where the data is noise-free. Theorem 2 shows existence of a solution satisfying the subspace property; Theorem 3 shows collapse of all ensemble members towards their mean value at an algebraic rate; Theorem 4 decomposes the error, in the image space under A, into an error which decays to zero algebraically (in a subspace determined by the initial ensemble) and an error which is constant in time (in a complementary space); Corollary 5 transfers these results to the state space of the unknown under additional assumptions. We assume throughout this section that the forward operator is a bounded, linear operator. The convergence analysis presented in Theorem 4 additionally requires the operator to be injective. For compact operators, the convergence result in the observational space does not imply convergence in the state space. However, assuming the forward operator is boundedly invertible, the generalization is straightforward. Note that this assumption is typically not fulfilled in the inverse setting, but opens up the perspective to use the EnKF as a linear solver. Subsection 4.2 briefly discusses the noisy data case where the results are analogous to those in the preceding subsection.

4.1. The Noise-Free Case. In proving the following theorems we will consider the situation where the data y is the image of a truth $u^\dagger \in \mathcal{X}$ under A. It is then useful to define

$$e^{(j)} = u^{(j)} - \bar{u}, \quad r^{(j)} = u^{(j)} - u^\dagger$$
$$E_{lj} = \langle Ae^{(l)}, Ae^{(j)} \rangle_\Gamma, \quad R_{lj} = \langle Ar^{(l)}, Ar^{(j)} \rangle_\Gamma, \quad F_{lj} = \langle Ar^{(l)}, Ae^{(j)} \rangle_\Gamma.$$

We view the last three items as entries of matrices E, R and F. The resulting matrices $E, R, F \in \mathbb{R}^{J\times J}$ satisfy the following: (i) E, R are symmetric; (ii) we may factorize $E(0) = X\Lambda(0)X^T$ where X is an orthogonal matrix whose columns are the eigenvectors of $E(0)$, and $\Lambda(0)$ is a diagonal matrix of corresponding eigenvalues; (iii) if $\mathsf{l} = (1, \ldots, 1)^T$, then $E\mathsf{l} = F\mathsf{l} = 0$. Note that $e^{(j)}$ measures deviation of the j^{th} ensemble member from the mean of the entire ensemble, and $r^{(j)}$ measures deviation of the j^{th} ensemble member from the truth $u^\dagger$ underlying the data. The matrices E, R and F contain information about these deviation quantities, when mapped forward under the operator A. The following theorem establishes existence and uniqueness of solutions to (8).

THEOREM 2. *Assume that y is the image of a truth $u^\dagger \in \mathcal{X}$ under A. Let $u^{(j)}(0) \in \mathcal{X}$ for $j = 1, \ldots, J$ and define $\mathcal{X}_0$ to be the linear span of the $\{u^{(j)}(0)\}_{j=1}^J$. Then equation (8) has a unique solution $u^{(j)}(\cdot) \in C([0,\infty); \mathcal{X}_0)$ for $j = 1, \ldots, J$.*

Proof. It follows from (8) that

$$\frac{\mathrm{d}u^{(j)}}{\mathrm{d}t} = -\frac{1}{J}\sum_{k=1}^{J} F_{jk}e^{(k)} \tag{11}$$

The right-hand side of equation (9) is locally Lipschitz in u as a mapping from $\mathcal{X}_0$ to $\mathcal{X}_0$. Thus local existence of a solution in $C([0,T); \mathcal{X}_0)$ holds for some $T > 0$, since $\mathcal{X}_0$ is finite dimensional. Thus it suffices to show that the solution does not blow-up in finite time. To this end we prove in Lemma 7, which is presented in the Appendix 7, that the matrices $E(t)$ and $F(t)$ are bounded by a constant depending on initial conditions, but not on time t. Using the global bound on F it follows from (11) that u is globally bounded by a constant depending on initial conditions, and growing exponentially with t. Global existence for u follows and the theorem is complete. ∎

The following theorem shows that all ensemble members collapse towards their mean value at an algebraic rate; and it demonstrates that the collapse slows down linearly as ensemble size increases.

THEOREM 3. *Assume that y is the image of a truth $u^\dagger \in \mathcal{X}$ under A. Let $u^{(j)}(0) \in \mathcal{X}$ for $j = 1, \ldots, J$. Then the matrix valued quantity $E(t)$ converges to 0 for $t \to \infty$ and, indeed $\|E(t)\| = \mathcal{O}(Jt^{-1})$.*

Proof. Lemma 7, which is presented in the Appendix 7, shows that the quantity $E(t)$ satisfies the differential equation

$$\frac{\mathrm{d}}{\mathrm{d}t}E = -\frac{2}{J}E^2$$

with initial condition $E(0) = X\Lambda(0)X^\top$, $\Lambda(0) = \{\lambda_0^{(1)}, \ldots, \lambda_0^{(J)}\}$ and orthogonal X. Using the eigensystem X gives the solution $E(t) = X\Lambda(t)X^\top$, where the entries of the diagonal matrix $\Lambda(t)$ are given by $(\frac{2}{J}t + \frac{1}{\lambda_0^{(j)}})^{-1}$ if $\lambda_0^{(j)} \neq 0$ and 0 otherwise. Then, the claim immediately follows from the explicit form of the solution $E(t)$. ∎

We are now interested in the long-time behavior of the residuals with respect to the truth. Theorem 4 characterizes the relation between the approximation quality of the initial ensemble and the convergence behavior of the residuals.

THEOREM 4. *Assume that y is the image of a truth $u^\dagger \in \mathcal{X}$ under A and the forward operator A is one-to-one. Let $\mathcal{Y}^\parallel$ denote the linear span of the $\{Ae^{(j)}(0)\}_{j=1}^J$ and let $\mathcal{Y}^\perp$ denote the orthogonal complement of $\mathcal{Y}^\parallel$ in $\mathcal{Y}$ with respect to the inner product $\langle\cdot,\cdot\rangle_\Gamma$ and assume that the initial ensemble members are chosen so that $\mathcal{Y}^\parallel$ has the maximal dimension $\min\{J-1, \dim(\mathcal{Y})\}$. Then $Ar^{(j)}(t)$ may be decomposed uniquely as $Ar_\parallel^{(j)}(t) + Ar_\perp^{(j)}(t)$ with $Ar_\parallel^{(j)} \in \mathcal{Y}^\parallel$ and $Ar_\perp^{(j)} \in \mathcal{Y}^\perp$. Furthermore, for all $j \in \{1, \cdots, J\}$, $Ar_\parallel^{(j)}(t) \to 0$ as $t \to \infty$ and, for all $j \in \{1, \cdots, J\}$ and $t \geq 0$, $Ar_\perp^{(j)}(t) = Ar_\perp^{(j)}(0) = Ar_\perp^{(1)}(0)$.*

Proof. Lemma 8, which is presented in the Appendix 7, shows that the matrix L, the linear transformation which determines how to write $\{Ae^{(j)}(t)\}_{j=1}^J$ in terms of the coordinates $\{Ae^{(j)}(0)\}_{j=1}^J$, is invertible for all $t \geq 0$ and hence that the linear span of the $\{Ae^{(j)}(t)\}_{j=1}^J$ is equal to $\mathcal{Y}^\parallel$ for all $t \geq 0$. Lemma 8 also shows that $Ar^{(j)}(t)$ may be decomposed uniquely as $Ar_\parallel^{(j)}(t) + Ar_\perp^{(j)}(t)$ with $Ar_\parallel^{(j)} \in \mathcal{Y}^\parallel$ and $Ar_\perp^{(j)} \in \mathcal{Y}^\perp$ and that $Ar_\perp^{(j)}(t) = Ar_\perp^{(j)}(0)$ for all $t \geq 0$. It thus remains to show that $Ar_\parallel^{(j)}(t)$ converges to zero as $t \to \infty$.

From Lemma 8 we know that we may write

$$Ar^{(j)}(t) = \sum_{k=1}^{J} \alpha_k Ae^{(k)}(t) + Ar_\perp^{(1)}, \tag{12}$$

for some (j-dependent) coefficients $\alpha = (\alpha_1, \ldots, \alpha_J)^T \in \mathbb{R}^J$. Furthermore Lemma 8 shows that if we initially choose α to be orthogonal to the eigenvectors $x^{(k)},\ k = 1, \ldots, J - \tilde{J}$ of E with corresponding eigenvalues $\lambda^{(k)}(t) = 0,\ k = 1, \ldots, J - \tilde{J}$, then this property will be preserved for all time. This is since we have $QL^{-1}x^{(k)} = Qx^{(k)} = 0,\ k = 1, \ldots, J - \tilde{J}$ and the matrix Υ with j^{th} row given by α satisfies $Q = \Upsilon L$ so that $\Upsilon x^{(k)} = 0,\ k = 1, \ldots, J - \tilde{J}$. Finally we choose coordinates in which $Ar_\perp^{(j)}(0)$ is orthogonal to $\mathcal{Y}^\parallel$.

Define the seminorms

$$|\alpha|_1^2 := |E^{1/2}\alpha|^2$$
$$|\alpha|_2^2 := |E\alpha|^2 \ .$$

Note that that these norms are time-dependent because E is. On the subspace of $\mathbb{R}^J$ orthogonal to $\text{span}\{x^{(1)}, \ldots, x^{(J-\tilde{J})}\}$,

$$|\alpha|_2^2 \geq \lambda_{\min}(t)|\alpha|_1^2 \ ,$$

where $\lambda_{\min}(t) = (\frac{2}{J}t + \frac{1}{\lambda_0^{\min}})^{-1}$ is the minimal positive eigenvalue of E, see (25). Furthermore, for the Euclidean norm $|\cdot|$, we have

$$|\alpha|_1^2 \geq \lambda_{\min}(t)|\alpha|^2$$

on the subspace of $\mathbb{R}^J$ orthogonal to $\text{span}\{x^{(1)}, \ldots, x^{(J-\tilde{J})}\}$.

Now note that the following differential equation holds for the quantity $\|Ar^{(j)}\|_\Gamma^2$:

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|Ar^{(j)}\|_\Gamma^2 = -\frac{1}{J}\sum_{r=1}^{J}\langle Ar^{(j)}, Ae^{(r)}\rangle_\Gamma\langle Ar^{(j)}, Ae^{(r)}\rangle_\Gamma.$$

We also have

$$\begin{aligned}\sum_{r=1}^{J}\langle Ar^{(j)}, Ae^{(r)}\rangle_\Gamma^2 &= \sum_{r=1}^{J}\langle\sum_{k=1}^{J}\alpha_k Ae^{(k)}, Ae^{(r)}\rangle_\Gamma\langle\sum_{l=1}^{J}\alpha_l Ae^{(l)}, Ae^{(r)}\rangle_\Gamma \\ &= \sum_{k=1}^{J}\sum_{l=1}^{J}\alpha_k\left(\sum_{r=1}^{J}E_{kr}E_{lr}\right)\alpha_l \\ &= |\alpha|_2^2\,.\end{aligned}$$

Using (12), the norm of the residuals can be expressed in terms of the coefficient vector of the residuals as follows:

$$\begin{aligned}\|Ar^{(j)}\|_\Gamma^2 &= \langle\sum_{k=1}^{J}\alpha_k Ae^{(k)} + Ar_\parallel^{(j)}(0), \sum_{l=1}^{J}\alpha_l Ae^{(l)} + Ar_\parallel^{(j)}(0)\rangle_\Gamma \\ &= \sum_{k=1}^{J}\sum_{l=1}^{J}\alpha_k E_{kl}\alpha_l + \|Ar_\parallel^{(j)}(0)\|_\Gamma^2 \\ &= |\alpha|_1^2 + \|Ar_\parallel^{(j)}(0)\|_\Gamma^2\,.\end{aligned}$$

Thus, the coefficient vector satisfies the following differential equation

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}|\alpha|_1^2 = -\frac{1}{J}|\alpha|_2^2 \le -\frac{\lambda_{\min}(t)}{J}|\alpha|_1^2\,,$$

which gives

$$\frac{1}{|\alpha|_1^2}\frac{\mathrm{d}}{\mathrm{d}t}|\alpha|_1^2 \le -\frac{2}{J}\left(\frac{2}{J}t + \frac{1}{\lambda_0^{\min}}\right)^{-1}$$

Hence, using that $\ln|\alpha(t)|_1^2 - \ln|\alpha(0)|_1^2 \le \ln\frac{1}{\lambda_0^{\min}} - \ln\left(\frac{2}{J}t + \frac{1}{\lambda_0^{\min}}\right)$, the coefficient vector can be bounded by

$$|\alpha(t)|_1^2 \le \frac{1}{\lambda_0^{\min}}|\alpha(0)|_1^2\lambda_{\min}(t)\,.$$

In the Euclidean norm, we have

$$\lambda_{\min}(t)|\alpha(t)|^2 \le \frac{1}{\lambda_0^{\min}}|\alpha(0)|_1^2\lambda_{\min}(t)$$

and thus

$$|\alpha(t)|^2 \le \frac{1}{\lambda_0^{\min}}|\alpha(0)|_1^2 \le \frac{\lambda_0^{(J)}}{\lambda_0^{\min}}\|\alpha(0)\|^2\,.$$

Since α is bounded in time, and since the $e^{(k)} \to 0$ by Lemma 8, the desired result follows from (12). ∎

Corollary 5 generalizes the convergence results of the preceding theorem to the infinite dimensional setting, i.e. to the case $\dim \mathcal{Y} = \infty$ under the additional assumption that the forward operator A is boundedly invertible. This implies that A cannot be a compact operator. However, this result allows to use the EnKF as a linear solver, since the convergence results can be transferred to the state space under the stricter assumptions on A.

COROLLARY 5. *Assume that y is the image of a truth $u^\dagger \in \mathcal{X}$ under A, the forward operator A is boundedly invertible and the initial ensemble is chosen such that the subspace $\mathcal{X}_0 = \text{span}\{e^{(1)}, \dots, e^{(J)}\}$ has maximal dimension $J-1$. Then, there exists a unique decomposition of the residual $r^{(j)}(t) = r_\parallel^{(j)}(t) + r_\perp^{(j)}(t)$ with $r_\parallel^{(j)} \in \mathcal{X}_0$ and $r_\perp^{(j)} \in \mathcal{X}_1$, where $\mathcal{X} = \mathcal{X}_0 \bigoplus \mathcal{X}_1$. The two subspaces $\mathcal{X}_0$ and $\mathcal{X}_1$ are orthogonal with respect to the inner product $\langle \cdot, \cdot \rangle_{\mathcal{X}^\perp} := \langle \Gamma^{-\frac{1}{2}} A \cdot, \Gamma^{-\frac{1}{2}} A \cdot \rangle$. Then, $r_\parallel^{(j)}(t) \to 0$ as $t \to \infty$ and $r_\perp^{(j)}(t) = r_\perp^{(j)}(0) = r_\perp^{(1)}(0)$.*

Proof. The assumption that the forward operator is boundedly invertible ensures that the range of the operator is closed. Furthermore, the invertibility of the operator A allows to transfer results from the observational space directly to the state space. Thus, the same arguments as in the proof of Theorem 4 prove the claim. ∎

4.2. Noisy Observational Data. Very similar analyses to those in the previous subsection may be carried out in the case where the observational data $y^\dagger$ is polluted by additive noise $\eta^\dagger \in \mathbb{R}^K$:

$$y^\dagger = Au^\dagger + \eta^\dagger \, .$$

Global existence of solutions, and ensemble collapse, follow similarly to the proofs of Theorems 2 and 3. Theorem 4 is more complex to generalize since it is not the mapped residual $Ar^{(j)}$ which is decomposed into a space where it decays to zero and a space where it remains constant, but rather the quantity $\vartheta^{(j)} = Ar^{(j)} - \eta^\dagger$. Driving this quantity to zero of course leads to over-fitting. Furthermore, the generalization to the infinite dimensional setting as presented in Corollary 5 is no longer valid, since the noise may take the data out of the range of the forward operator.

5. Numerical Results. In this section we present numerical experiments both with and without data, illustrating the theory of the previous section for the linear inverse problem. We also study a nonlinear groundwater flow inverse problem, demonstrating that the theory in the linear problem provides useful insight for the nonlinear problem.

5.1. Linear Forward Model. We consider the one dimensional elliptic equation

$$-\frac{\mathrm{d}^2 p}{\mathrm{d}x^2} + p = u \quad \text{in } D := (0, \pi) \, , \; p = 0 \quad \text{in } \partial D \, , \tag{13}$$

where the uncertainty-to-observation operator is given by $\mathcal{G} = \mathcal{O} \circ G = \mathcal{O} \circ A^{-1}$ with $A = -\frac{\mathrm{d}^2}{\mathrm{d}x^2} + id$ and $D(A) = H^2(I) \cap H_0^1$. The observation operator $\mathcal{O}$ consists of $K = 2^4 - 1$ system responses at K equispaced observation points at $x_k = \frac{k}{2^4}, k = 1, \dots, 2^4 - 1$, $o_k(\cdot) := \delta(\cdot - x_k)$, i.e. $(O(\cdot))_k = o_k(\cdot)$. The forward problem (13) is solved numerically by a FEM using continuous, piecewise linear ansatz functions on a uniform mesh with meshwidth $h = 2^{-8}$ (the spatial discretization leads to a discretization of u, i.e. $u \in \mathbb{R}^{2^8 - 1}$).

The goal of the computation is to recover the unknown data u from noisy observations

$$y^\dagger = p + \eta = \mathcal{O}A^{-1}u^\dagger + \eta\,. \tag{14}$$

The measurement noise is chosen to be normally distributed, $\eta \sim \mathcal{N}(0, \gamma I)$, $\gamma \in \mathbb{R}$, $\gamma > 0$, $I \in \mathbb{R}^{K\times K}$. The initial ensemble of particles is chosen to be based on the eigenvalues and eigenfunctions $\{\lambda_j, z_j\}_{j\in\mathbb{N}}$ of the covariance operator C_0. Here $C_0 = \beta(A - id)^{-1}$, $\beta = 10$. Although we do not pursue the Bayesian interpretation of the EnKF, the reader interested in the Bayesian perspective may think of the prior as being $\mu_0 = N(0, C_0)$, which is a Brownian bridge. In all our experiments we set $u^{(j)}(0) = \sqrt{\lambda_j}\zeta_j z_j$ with $\zeta_j \sim \mathcal{N}(0,1)$ for $j = 1, \ldots, J$. Thus the j^{th} element of the initial ensemble may be viewed as the j^{th} term in a Karhunen-Loève (KL) expansion of a draw from μ_0 which, in this case, is a Fourier sine series expansion.

5.1.1. Noise-Free Observational Data. To numerically verify the theoretical results presented in section 4.1, we first restrict the discussion to the noise-free case, i.e. we assume that $\eta = 0$ in (14) and set $\Gamma = id$. The study summarized in Figures 1 - 4 shows the influence of the number of particles on the dynamical behavior of the quantities e and r, the matrix-valued quantities and the resulting EnKF estimate.

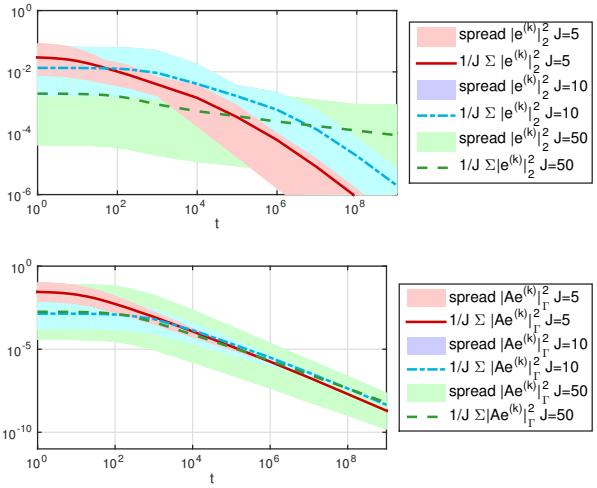

FIG. 1. *Quantities* $|e|_2^2$, $|Ae|_\Gamma^2$ *w.r. to time* t, $J = 5$ *(red),* $J = 10$ *(blue) and* $J = 50$ *(green),* $\beta = 10$, $K = 2^4 - 1$, *initial ensemble chosen based on KL expansion of* $C_0 = \beta(A - id)^{-1}$.

As shown in Theorem 3, the rate of convergence of the ensemble collapse is algebraic (cf. Figure 1) with a constant growing with larger ensemble size. Comparing the dynamical behavior of the residuals, we observe that, for the ensemble of size J=5, the estimate can be improved in the beginning, but reaches a plateau after a short time. Increasing the number of particles to $J = 10$ improves the accuracy of the estimate. For the ensemble size $J = 50$, Figure 2 shows the convergence of the projected residuals, i.e. the observations can be perfectly recovered. The same behavior can be observed by comparing the EnKF estimate with the truth and the observational data (cf. Figure 4).

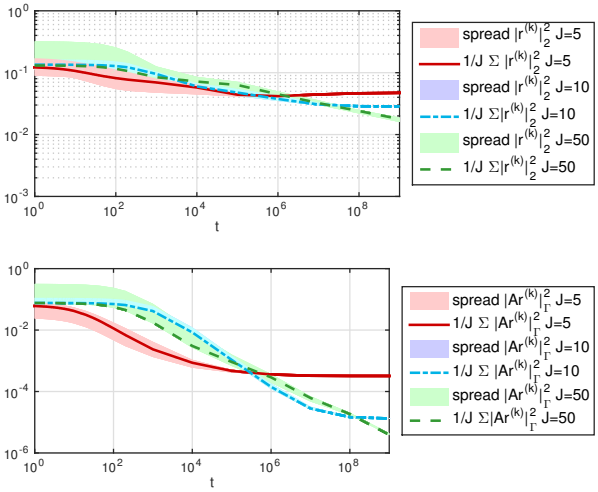

FIG. 2. *Quantities* $|r|_2^2$, $|Ar|_\Gamma^2$ *w.r. to time* t, $J = 5$ *(red),* $J = 10$ *(blue) and* $J = 50$ *(green),* $\beta = 10$, $K = 2^4 - 1$, *initial ensemble chosen based on KL expansion of* $C_0 = \beta(A - id)^{-1}$.

The results derived in this paper hold true for each particle, however, for the sake of presentation, the empirical mean of the quantities of interest is shown and the spread indicates the minimum and maximum deviations of the ensemble members from the empirical mean.

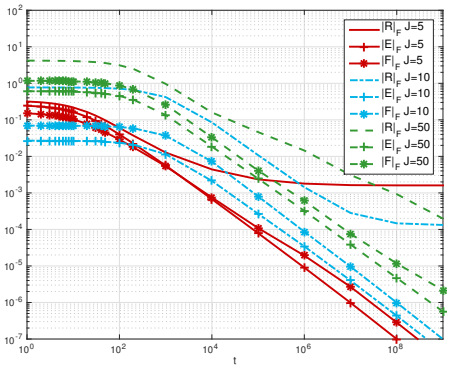

FIG. 3. *Quantities* $\|E\|_F$, $\|F\|_F$, $\|R\|_F$ *w.r. to time* t, $J = 5$ *(red),* $J = 10$ *(blue) and* $J = 50$ *(green),* $\beta = 10$, $K = 2^4 - 1$, *initial ensemble chosen based on KL expansion of* $C_0 = \beta(A - id)^{-1}$.

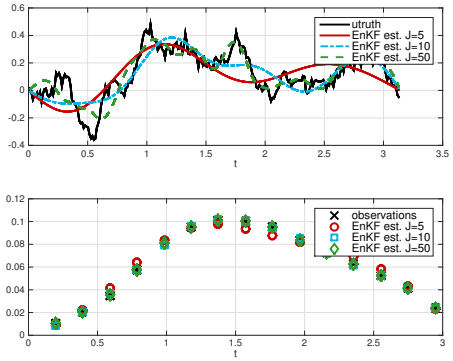

FIG. 4. *Comparison of the EnKF estimate with the truth and the observations,* $J = 5$ *(red),* $J = 10$ *(blue) and* $J = 50$ *(green),* $\beta = 10$, $K = 2^4 - 1$, *initial ensemble chosen based on KL expansion of* $C_0 = \beta(A - id)^{-1}$.

Due to the construction of the ensembles in the example, the subspace spanned by the ensemble of size 5 is a strict subset of the subspace spanned by the larger ensembles. Thus, due to Theorem 4, which characterizes the convergence of the residuals with respect to the approximation quality of the subspace spanned by the initial ensemble, the EnKF estimate can be substantially improved by controlling this subspace. As illustrated in Figure 2, the mapped residual of the ensemble of size 5 decreases

monotonically, but levels off after a short time. Similar convergence properties can be observed for $J = 10$. The same behavior is expected for the larger ensemble, when integrating over a larger time horizon. This can be also observed for the matrix-valued quantities depicted in Figure 3.

We will investigate this point further by comparing the performance of two ensembles, both of size 5: one based on the KL expansion and one chosen such that the contribution of $Ar_\perp^{(j)}(t)$ in Theorem 4 is minimized. Since we use artificial data, we can minimize the contribution of $Ar_\perp^{(j)}(t)$ by ensuring that $Ar^{(1)} = \sum_{k=1}^J \alpha_k Ae^{(k)}$ for some coefficients $\alpha_k \in \mathbb{R}$. Given $u^{(2)}, \dots, u^{(J)}$ and coefficients $\alpha_1, \dots, \alpha_J$, we define $u^{(1)} = (1-\alpha_1+\sum_{k=1}^J \alpha_k/J)^{-1}(u^\dagger - \alpha_1/J \sum_{j=2}^J u^{(j)} + \sum_{k=2}^J \alpha_k u^{(k)} - \alpha_k/J \sum_{j=2}^J u^{(j)})$, which gives the desired property of the ensemble. Note that this approach is not feasible in practice and has to be replaced by an adaptive strategy minimizing the contribution of $Ar_\perp^{(j)}(t)$. However this experiment serves to illusrate the important role of the initial ensemble in determining the error and is included for this reason. The convergence rate of the mapped residuals and of the ensemble collapse is algebraic in both cases, with rate 1 (in the squared Euclidean norm). Figure 5 shows the convergence of the projected residuals for the adaptively chosen ensemble. The decomposition of the residuals (cf. Figure 6) numerically verifies the presented theory, which motivates the adaptive construction of the ensemble. Methods to realize this strategy, in the linear and nonlinear case, will be addressed in a subsequent paper.

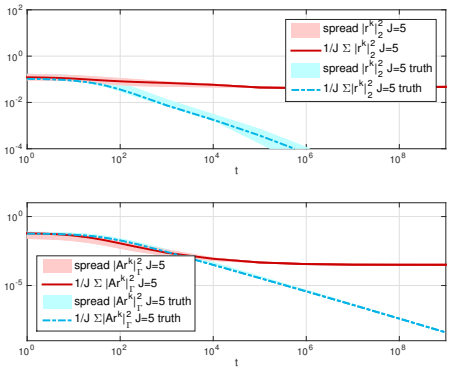

FIG. 5. *Quantities* $|r|_2^2$, $|Ar|_\Gamma^2$ *w.r. to time* t, $J = 5$ *based on KL expansion of* $C_0 = \beta(A - id)^{-1}$ *(red) and* $J = 5$ *minimizing the contribution of* $Ar_\perp^{(j)}(t)$ *(blue),* $\beta = 10$, $K = 2^4 - 1$.

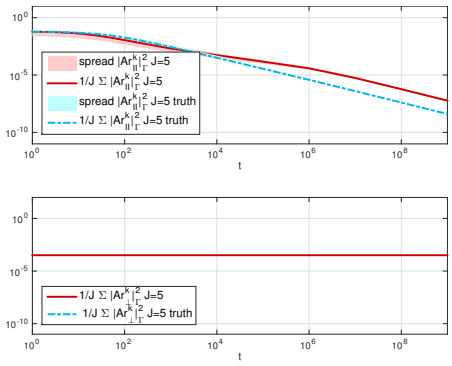

FIG. 6. *Quantities* $|Ar_\parallel^{(j)}|_\Gamma^2$, $|Ar_\perp^{(j)}|_\Gamma^2$ *w.r. to time* t, $J = 5$ *based on KL expansion of* $C_0 = \beta(A - id)^{-1}$ *(red) and* $J = 5$ *minimizing the contribution of* $Ar_\perp^{(j)}(t)$ *(blue),* $\beta = 10$, $K = 2^4 - 1$.

5.1.2. Noisy Observational Data. We will now allow for noise in the observational data, i.e. we assume that the data is given by $y^\dagger = \mathcal{O}A^{-1}u^\dagger + \eta^\dagger$, where $\eta^\dagger$ is a fixed realization of the random vector $\eta \sim \mathcal{N}(0, 0.01^2\,\mathrm{id})$. Note that the standard deviation is chosen to be roughly 10% of the (maximum of the) observed data. Besides the quantities e and r, the misfit $\vartheta^{(j)} = Au^{(j)} - y^\dagger = Ar^{(j)} - \eta^\dagger$ of each ensemble member is of interest, since, in practice, the residual is not accessible and the misfit is used to check for convergence and to design an appropriate stopping criterion.

Besides the two ensembles with 5 particles introduced in the previous section, we define an additional one minimizing the contribution of $\vartheta_\perp^{(j)}$ to the misfit, analogously to what was done in the adaptive initialization in the previous subsection, and mo-

tivated by the analogue of Theorem 4 in the noisy data case. Note that the design of an adaptive ensemble based on the decomposition of the projected residual is, in general, not feasible without explicit knowledge of the noise $\eta^\dagger$.

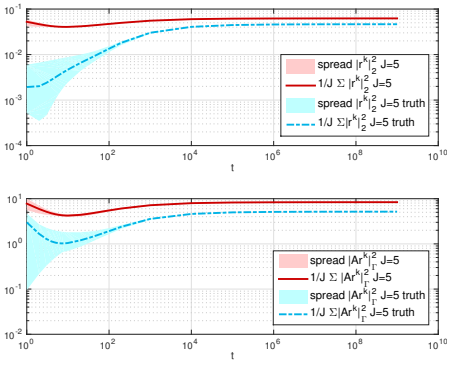

FIG. 7. *Quantities $|r|_2^2$, $|Ar|_\Gamma^2$ w.r. to t, $J = 5$ based on KL expansion of $C_0 = \beta(A - id)^{-1}$ (red), $J = 5$ adaptively chosen (blue), $\beta = 10$, $K = 2^4 - 1$, $\eta \sim \mathcal{N}(0, 0.01^2\,\mathrm{id})$.*

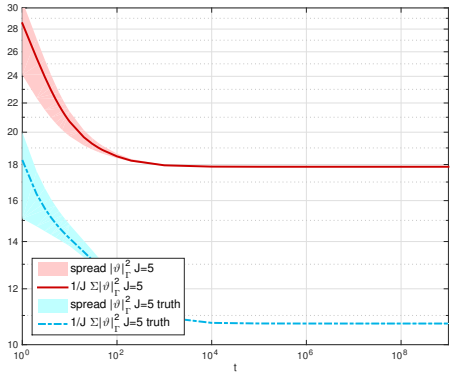

FIG. 8. *Misfit $|\vartheta|_2^2$ w.r. to time t, $J = 5$ based on KL expansion of $C_0 = \beta(A - id)^{-1}$ (red), $J = 5$ adaptively chosen (blue), $\beta = 10$, $K = 2^4 - 1$, $\eta \sim \mathcal{N}(0, 0.01^2\,\mathrm{id})$.*

Figure 7 illustrates the well-known overfitting effect, which arises without using appropriate stopping criteria. The method tries to fit the noise in the measurements, which results in an increase in the residuals. This effect is not seen in the misfit functional, cf. Figure 8 and Figure 9.

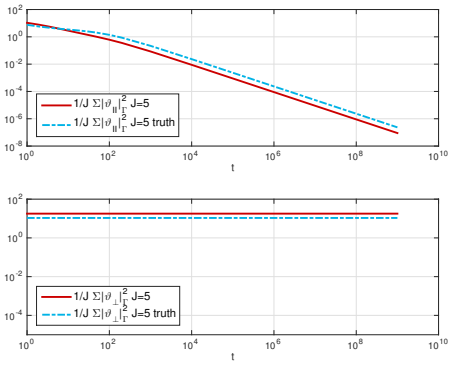

FIG. 9. *Quantities $|\vartheta_\parallel^{(j)}|_\Gamma^2$, $|\vartheta_\perp^{(j)}|_\Gamma^2$ w.r. to time t, $J = 5$ based on KL expansion of $C_0 = \beta(A - id)^{-1}$ (red), $J = 5$ minimizing the contribution of $Ar_\perp^{(j)}(t)$ (blue), $\beta = 10$, $K = 2^4 - 1$, $\eta \sim \mathcal{N}(0, 0.01^2\,\mathrm{id})$.*

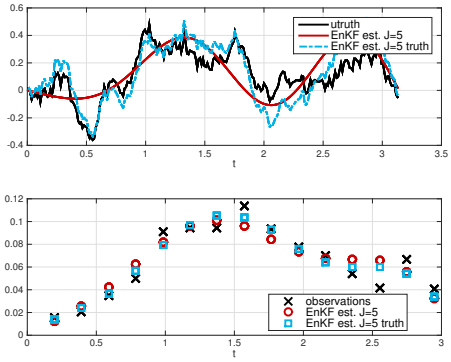

FIG. 10. *Comparison of the EnKF estimate with the truth and the observations, $J = 5$ based on KL expansion of $C_0 = \beta(A-id)^{-1}$ (red), $J = 5$ minimizing the contribution of $Ar_\perp^{(j)}(t)$ (blue), $\beta = 10$, $K = 2^4 - 1$, $\eta \sim \mathcal{N}(0, 0.01^2\,\mathrm{id})$.*

However, the comparison of the EnKF estimates to the truth reveals, in Figure 10, the strong overfitting effect and suggests the need for a stopping criterion. The Bayesian setting itself provides a so-called a priori stopping rule, i.e. the SMC viewpoint motivates a stopping of the iterations at time $T = 1$. Another common choice in the deterministic optimization setting is the discrepancy principle, which accounts for the realization of the noise by checking the following condition $\|\mathcal{G}(\bar{u}(t)) - y\|_\Gamma \leq \tau$, where $\tau > 0$ depends on the dimension of the observational space.

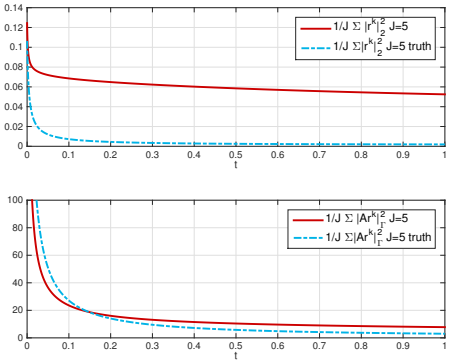

FIG. 11. *Quantities* $|r|_2^2$, $|Ar|_\Gamma^2$ *w.r. to time* t, $J = 5$ *based on KL expansion of* $C_0 = \beta(A - id)^{-1}$ *(red),* $J = 5$ *minimizing the contribution of* $Ar_\perp^{(j)}(t)$ *(blue),* $\beta = 10$, $K = 2^4 - 1$, $\eta \sim \mathcal{N}(0, 0.01^2\,\mathrm{id})$, *Bayesian stopping rule.*

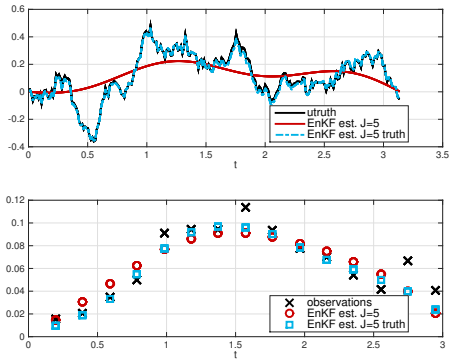

FIG. 12. *Comparison of the EnKF estimate with the truth and the observations,* $J = 5$ *based on KL expansion of* $C_0 = \beta(A-id)^{-1}$ *(red),* $J = 5$ *minimizing the contribution of* $Ar_\perp^{(j)}(t)$ *(blue),* $\beta = 10$, $K = 2^4 - 1$, $\eta \sim \mathcal{N}(0, 0.01^2\,\mathrm{id})$, *Bayesian stopping rule.*

Figures 11 - 12 show the results obtained by employing the Bayesian stopping rule, i.e. by integrating up to time $T = 1$. The adaptively chosen ensemble leads to much better results in the case of the Bayesian stopping rule. Since we do not expect to have explicit knowledge of the noise, the adaptive strategy as presented above is in general not feasible. However an adaptive choice of the ensemble according to the misfit may lead to a strong overfitting effect, as shown in Figures 13 - 14.

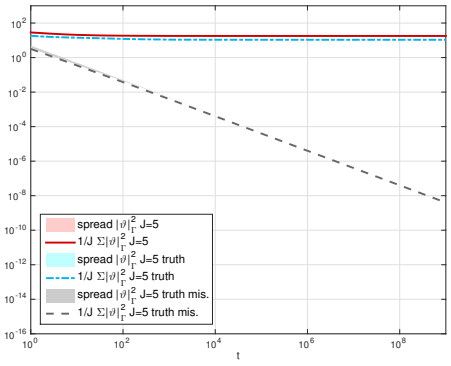

FIG. 13. *Misfit* $|\vartheta|_\Gamma^2$ *w.r. to time* t, $J = 5$ *based on KL expansion of* $C_0 = \beta(A - id)^{-1}$ *(red),* $J = 5$ *adaptively chosen (blue),* $J = 5$ *minimizing the contribution of* $\vartheta_\perp$ *w.r. to misfit (gray),* $\beta = 10$, $K = 2^4 - 1$, $\eta \sim \mathcal{N}(0, 0.01^2\,\mathrm{id})$.

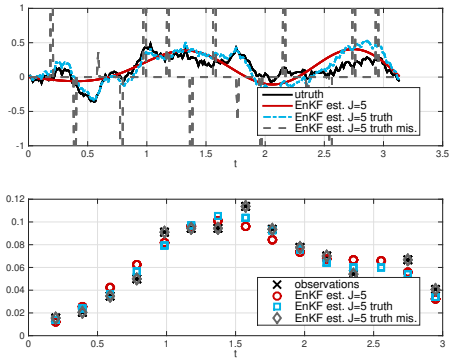

FIG. 14. *Comparison of the EnKF estimate with the truth and the observations,* $J = 5$ *based on KL expansion of* $C_0 = \beta(A - id)^{-1}$ *(red),* $J = 5$ *adaptively chosen (blue),* $J = 5$ *minimizing the contribution of* $\vartheta_\perp$ *w.r. to misfit (gray),* $\beta = 10$, $K = 2^4 - 1$, $\eta \sim \mathcal{N}(0, 0.01^2\,\mathrm{id})$.

The overfitting effect is still present in the small noise regime as shown below. The standard deviation of the noise is reduced by 100, i.e. $\eta \sim \mathcal{N}(0, 0.001^2\,\mathrm{id})$.

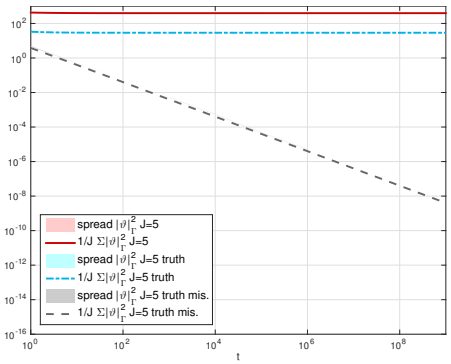

FIG. 15. *Misfit $|\vartheta|^2_\Gamma$ w.r. to time t, $J = 5$ based on KL expansion of $C_0 = \beta(A - id)^{-1}$ (red), $J = 5$ adaptively chosen (blue), $J = 5$ minimizing the contribution of $\vartheta_\perp$ w.r. to misfit (gray), $\beta = 10$, $K = 2^4 - 1$, $\eta \sim \mathcal{N}(0, 0.001^2\,\mathrm{id})$.*

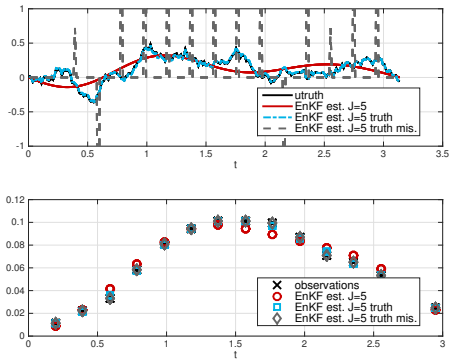

FIG. 16. *Comparison of the EnKF estimate with the truth and the observations, $J = 5$ based on KL expansion of $C_0 = \beta(A - id)^{-1}$ (red), $J = 5$ adaptively chosen (blue), $J = 5$ minimizing the contribution of $\vartheta_\perp$ w.r. to misfit (gray), $\beta = 10$, $K = 2^4 - 1$, $\eta \sim \mathcal{N}(0, 0.001^2\,\mathrm{id})$.*

The ill-posedness of the problem leads to the instabilities of the identification problem and requires the use of an appropriate stopping rule.

5.2. Nonlinear Forward Model. To investigate the numerical behavior of the EnKF for nonlinear inverse problems, we consider the following two-dimensional elliptic PDE:

$$-\operatorname{div}(e^u \nabla p) = f \quad \text{in } D := (-1, 1)^2\,, \; p = 0 \quad \text{in } \partial D.$$

We aim to find the log permeability u from 49 observations of the solution p on a uniform grid in D. We choose $f(x) = 100$ for the experiments. The mapping from u to these observations is now nonlinear. Again we work in the noise-free case, and take $\Gamma = I$, $\Sigma = 0$ and solve (6) to estimate the unknown parameters. The prior is assumed to be Gaussian with covariance operator $C_0 = (-\triangle)^{-2}$, employing homogeneous Dirichlet boundary conditions to define the inverse of $-\triangle$. We use a FEM approximation based on continuous, piecewise linear ansatz functions on a uniform mesh with meshwidth $h = 2^{-4}$. The initial ensemble of size 5 and 50 is chosen based on the KL expansion of C_0 in the same way as in the previous subsection.

The results given in Figure 17 and Figure 18 show a similar behavior as in the linear case. The approximation quality of the subspace spanned by the initial ensemble clearly influences, also in the nonlinear example, the accuracy of the estimate. Taking a look at the EnKF estimate in this nonlinear setting, we observe a satisfactory approximation of the truth and a perfect match of the observational data in the case of the larger ensemble, cf. Figure 19.

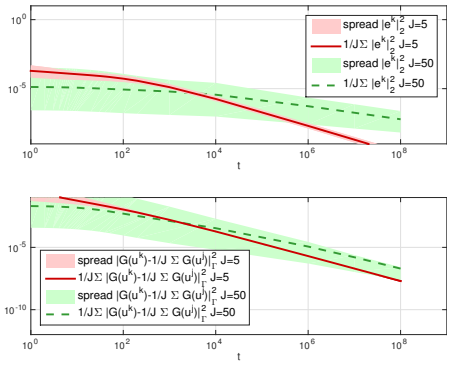

FIG. 17. *Quantities* $|e^{(k)}|_2^2$, $|\mathcal{G}(u^{(k)}) - \frac{1}{J}\sum_{j=1}^{J}\mathcal{G}(u^{(j)}|_\Gamma^2$ *w.r. to time* t, $J = 5$ *(red) and* $J = 50$ *(green), initial ensemble chosen based on KL expansion of* $C_0 = (-\triangle)^{-2}$.

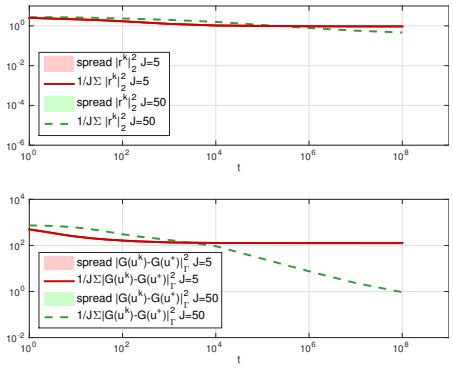

FIG. 18. *Quantities* $|r^{(k)}|_2^2$ $|\mathcal{G}(u^{(k)}) - \mathcal{G}(u^\dagger)|_\Gamma^2$ *w.r. to time* t, $J = 5$ *(red) and* $J = 50$ *(green), initial ensemble chosen based on KL expansion of* $C_0 = (-\triangle)^{-2}$.

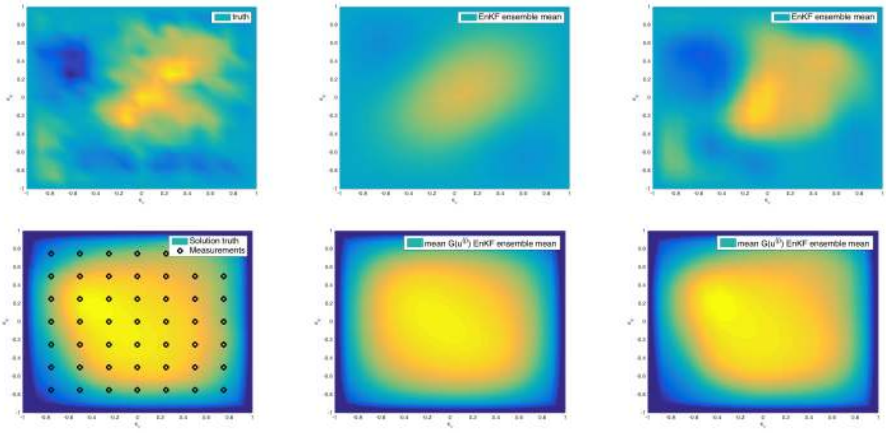

FIG. 19. *Comparison of the truth (left above) and the EnKF estimate w.r. to* x, *J=5 (middle above) ,* $J = 50$ *(right above) and comparison of the forward solution* $G(u^\dagger)$ *(left below) and the estimated solutions of the forward problem* $J = 5$ *(middle below),* $J = 50$ *(right below).*

6. Variants on EnKF. In this section we describe three variants on the EnKF, all formulated in continuous time in order to facilitate comparison with the preceding studies of the standard EnKF in continuous time. The first two methods, variance inflation and localization, are commonly used by practitioners in both the filtering and inverse problem scenarios [8, 1]. The third method, random search, is motivated by the SMC derivation of the EnKF for inverse problems, and is new in the context of the EnKF. For all three methods we provide numerical results which illustrate the behavior of the EnKF variant, in comparison with the standard method. In the following, we focus on the linear case with $\Sigma = 0$. The methods from the first two subsections have generalizations in the general nonlinear setting, and indeed are widely used in data assimilation and, to some extent, in geophysical inverse problems; but we present them in a form tied to the equation (9) which is derived in the linear case.

The method in the final subsection is implemented through an entirely derivative-free MCMC method, and is hence automatically defined, as is, for nonlinear as well as linear inverse problems.

6.1. Variance Inflation. The empirical covariances C^{up} and C^{pp} all have rank no greater than $J-1$ and hence are rank deficient whenever the number of particles J is less than the dimension of the space X. Variance inflation proceeds by correcting such rank deficiencies by the addition of self-adjoint, strictly positive operators. A natural variance inflation technique is to add a multiple of the prior covariance C_0 to the empirical covariance which gives rise to the equations

$$\frac{\mathrm{d}u^{(j)}}{\mathrm{d}t} = -\big(\alpha C_0 + C(u)\big) D_u\Phi(u^{(j)};y), \quad j=1,\dots,J\,, \tag{15}$$

where Φ is as defined in (10). Taking the inner-product in X with $D_u\Phi(u^{(j)};y)$ we deduce that

$$\frac{\mathrm{d}\Phi(u^{(j)};y)}{\mathrm{d}t} \le -\alpha \|C_0^{\frac{1}{2}} D_u\Phi(u^{(j)};y)\|^2\,,$$

This implies that all $\omega-$limit points of the dynamics are contained in the critical points of $\Phi(\cdot;y)$.

6.2. Localization. Localization techniques aim to remove spurious long distance correlations by modifying the covariance operators C^{up} and C^{pp}, or directly the Kalman gain $C^{up}_{n+1}(C^{pp}_{n+1}+h^{-1}\Gamma)^{-1}$. Typical convolution kernels reducing the influence of distant regions are of the form

$$\begin{aligned}&\rho : D\times D \to \mathbb{R}\\ &\rho(x,y) = \exp(-|x-y|^r)\,,\end{aligned}$$

where $D\subset\mathbb{R}^d$, $d\in\mathbb{N}$ denotes the physical domain and $|\cdot|$ is a suitable norm in D, $r\in\mathbb{N}$, cf. [21]. The continuous time limit in the linear setting then reads as

$$\frac{\mathrm{d}u^{(j)}}{\mathrm{d}t} = -C^{\mathrm{loc}}(u) D_u\Phi(u^{(j)};y), \quad j=1,\dots,J\,, \tag{16}$$

where $C^{\mathrm{loc}}(u)\phi(x) = \int_D \phi(y)k(x,y)\rho(x,y)\mathrm{d}y$ with k denoting the kernel of $C(u)$ and $\phi\in\mathcal{X}$.

6.3. Randomized Search. We notice that the mapping on probability measures given by (3) may be replaced by

$$\mu_{n+1} = L_n P_n \mu_n\,.$$

where P_n is any Markov kernel which preserves μ_n. For example we may take P_n to be the pCN method [6] for measure μ_n. One step of the pCN method for given particle $u_n^{(j)}$ in iteration n is realized by

- Propose $v_n^{(j)} = \sqrt{(1-\beta^2)}u_n^{(j)} + \beta\iota^{(j)}$, $\iota^{(j)}\sim\mathcal{N}(0,C_0)$.
- Set $\tilde{u}_n^{(j)} = v_n^{(j)}$ with probability $a(u_n^{(j)}, v_n^{(j)})$.
- Set $\tilde{u}_n^{(j)} = u_n^{(j)}$ otherwise

assuming the prior is Gaussian, i.e. $\mathcal{N}(0,C_0)$. The acceptance probability is given by

$$a(u_n^{(j)}, v_n^{(j)}) = \min\{1, \exp(nh\Phi(u_n^{(j)}) - nh\Phi(v_n^{(j)}))\}.$$

The particles $\tilde{u}_n^{(j)}$ are used to approximate the measure $\tilde{\mu}_n = P_n\mu_n$, which is then mapped to μ_{n+1} by the application of Bayes' theorem, i.e. $\mu_{n+1} = L_n\tilde{\mu}_n$.

Using the continuous-time diffusion limit arguments from [26, Theorem 4], which apply in the nonlinear case, and combining with the continuous time limits described for the EnKF earlier in this paper, we obtain

$$\begin{aligned}\frac{\mathrm{d}u^{(j)}}{\mathrm{d}t} = \frac{1}{J}\sum_{k=1}^{J}\left\langle \mathcal{G}(u^{(k)}) - \overline{\mathcal{G}}, y - \mathcal{G}(u^{(j)})\right\rangle_{\Gamma}\left(u^{(k)} - \overline{u}\right) \\ -u^{(j)} - tC_0 D_u\Phi(u^{(j)}; y) + \sqrt{2C_0}\frac{\mathrm{d}W^{(j)}}{\mathrm{d}t}\,.\end{aligned} \tag{17}$$

Although the limiting equation involves gradients of Φ, and hence adjoints for the forward model, the discrete time implementation above avoids the gradient computation by using the accept-reject step, and remains a derivative free optimizer.

6.4. Numerical Results. In the following, to illustrate behavior of the EnKF variants, we present numerical experiments for the linear forward problem in the noise-free case: (13) and (14) with $\eta = 0$. The performance of the EnKF variants is compared to the basic algorithms shown in Figure 2 and Figure 4.

6.4.1. Inflation. We investigate the numerical behavior of variance inflation of the form given in (15) with $\alpha = 0.01$. Figures 20 and 21 show that the variance inflated method becomes a preconditioned gradient flow, which, in the linear case, leads to fast convergence of the projected iterates. It is noteworthy that in this case there is very little difference in behavior between ensemble sizes of 5 and 50.

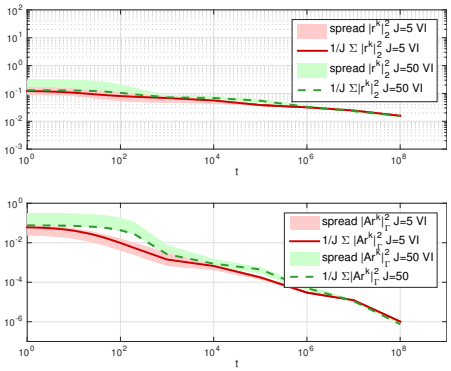

FIG. 20. *Quantities $|r|_2^2$, $|Ar|_\Gamma^2$ w.r. to time t, $J = 5$ with variance inflation (red) and $J = 50$ with variance inflation (green), $\beta = 10$, $K = 2^4 - 1$, initial ensemble chosen based on KL expansion of $C_0 = \beta(A - id)^{-1}$.*

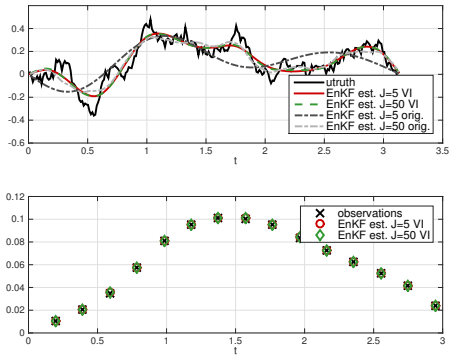

FIG. 21. *Comparison of the EnKF estimate with the truth and the observations, $J = 5$ with variance inflation (red) and $J = 50$ with variance inflation (green), $\beta = 10$, $K = 2^4 - 1$, initial ensemble chosen based on KL expansion of $C_0 = \beta(A - id)^{-1}$.*

6.4.2. Localization. We consider a localization of the form given by equations (16), (16) with $r = 2$ and Euclidean norm inside the cut-off kernel. Figures 22 - 23 clearly demonstrate the improvement by the localization technique, which can overcome the linear span property and thus, leads to better estimates of the truth.

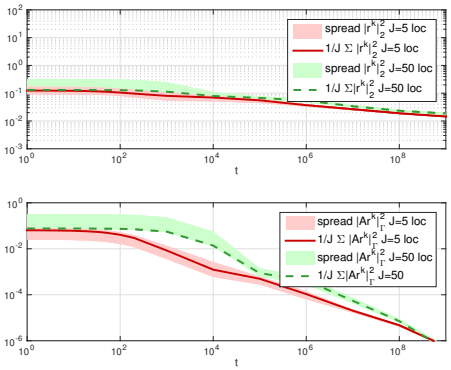

FIG. 22. *Quantities* $|r|_2^2$, $|Ar|_\Gamma^2$ *w.r. to time* t, $J = 5$ *with localization (red) and* $J = 50$ *with localization (green),* $\beta = 10$, $K = 2^4 - 1$, *initial ensemble chosen based on KL expansion of* $C_0 = \beta(A - id)^{-1}$.

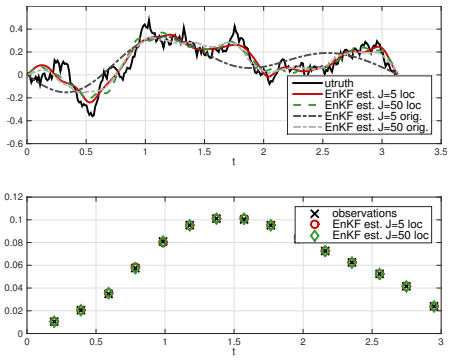

FIG. 23. *Comparison of the EnKF estimate with the truth and the observations,* $J = 5$ *with localization (red) and* $J = 50$ *with localization (green),* $\beta = 10$, $K = 2^4 - 1$, *initial ensemble chosen based on KL expansion of* $C_0 = \beta(A - id)^{-1}$.

6.4.3. Randomized Search. We investigate the behavior of randomized search for the linear problem with $\mathcal{G}(\cdot) = A\cdot$. For the numerical solution of the continuous limit (17), we employ a splitting scheme with a linearly implicit Euler step, namely

$$\begin{aligned} \tilde{u}_{n+1}^{(j)} &= \sqrt{1-2h}u_n^{(j)} + \sqrt{2hC_0}\zeta_n \\ Ku_{n+1}^{(j)} &= \tilde{u}_{n+1}^{(j)} + h(C(\tilde{u}_{n+1})A^*\Gamma^{-1}y^\dagger + nhC_0A^*\Gamma^{-1}y^\dagger)\,, \end{aligned}$$

where $\zeta_n \sim N(0, id)$ and $K := I + h(C(\tilde{u}_{n+1})A^*\Gamma^{-1}A + nhC_0A^*\Gamma^{-1}A)$. In all numerical experiments reported we take $h = 2^{-8}$. Figure 24 and Figure 25 show that the randomized search leads to an improved performance compared to the original EnKF method. Due to the fixed step size and the resulting high computational costs, the solution is computed up to time $T = 100$. In order to accelerate the numerical solution of the limit (17), implicit schemes can be considered. Note that the limit requires the computation of the gradients, which is in practice undesirable. However, the limit reveals from a theoretical point of view important structure, whereas the discrete version is more suitable for applications. The advantage of the randomized search is apparent.

6.4.4. Summary. The experiments show a similar performance for all discussed variants. The variance inflation technique and the localization variant both lead to gradient flows, which are, in the noise-free case, favorable due to the fast convergence. On the other hand, these strategies also accelerate the convergence of the ensemble to the mean (ensemble collapse) and may be considered less desirable for this reason. The randomized search preserves by construction the spread of the ensemble. A similar regularization effect is achieved by perturbing the observational data, see (6). The variants all break the subspace property of the original version, which results in an improvement in the estimate.

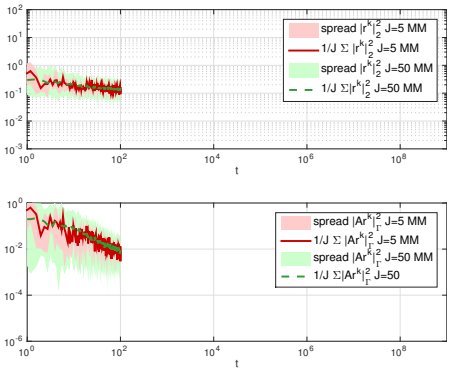

FIG. 24. *Quantities* $|r|_2^2$, $|Ar|_\Gamma^2$ *w.r. to time* t, $J = 5$ *with randomized search (red) and* $J = 50$ *with randomized search (green),* $\beta = 10$, $K = 2^4 - 1$, *initial ensemble chosen based on KL expansion of* $C_0 = \beta(A - id)^{-1}$.

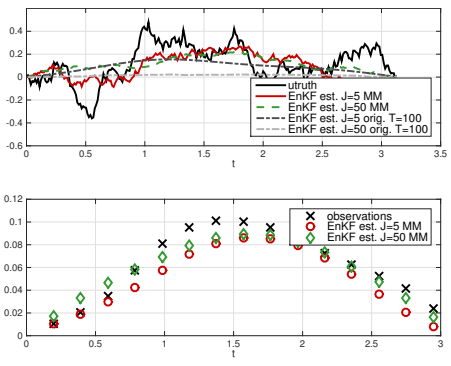

FIG. 25. *Comparison of the EnKF estimate with the truth and the observations,* $J = 5$ *with randomized search (red) and* $J = 50$ *with randomized search (green),* $\beta = 10$, $K = 2^4 - 1$, *initial ensemble chosen based on KL expansion of* $C_0 = \beta(A - id)^{-1}$.

We study the same test case as in section 5.1.2, with the same realization of the measurement noise (cf. Figure 11 and Figure 12), to allow for a comparison of the three methods introduced in this section. Combining the various techniques with the Bayesian stopping rule for noisy observations, we observe the following behavior given in Figure 26 and Figure 27.

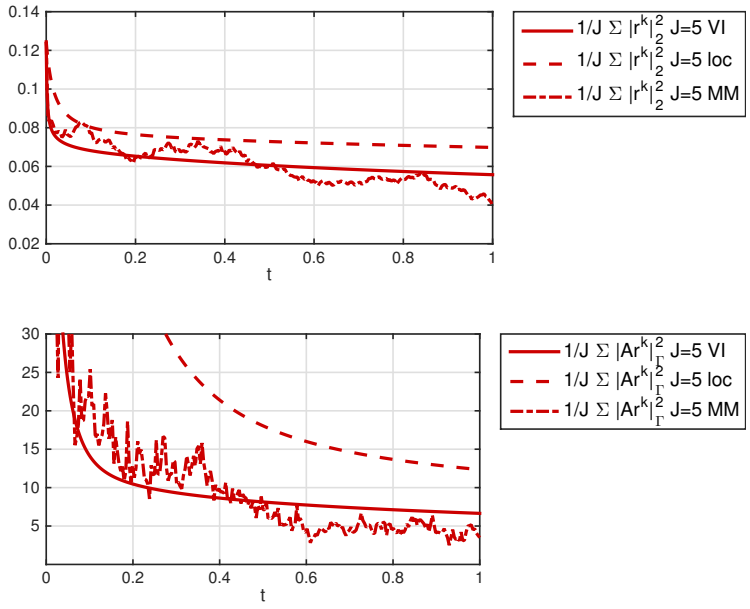

FIG. 26. *Quantities* $|r|_2^2$, $|Ar|_\Gamma^2$ *w.r. to time* t, $J = 5$ *(red) for the discussed variants,* $\beta = 10$, $\beta = 10$, $K = 2^4 - 1$, *initial ensemble chosen based on KL expansion of* $C_0 = \beta(A - id)^{-1}$.

In both figures, the abbreviation VI refers to variance inflation, loc denotes the localization technique and MM stands for the randomized search (Markov mixing). The randomized search clearly outperforms the two other strategies and leads to a

better estimate of the unknown data. In Figures 26 and 27, one path of the solution of (17) is shown, similar performance can be observed for further paths. This strategy has the potential to significantly improve the performance of the EnKF and will be investigated in more details in subsequent papers.

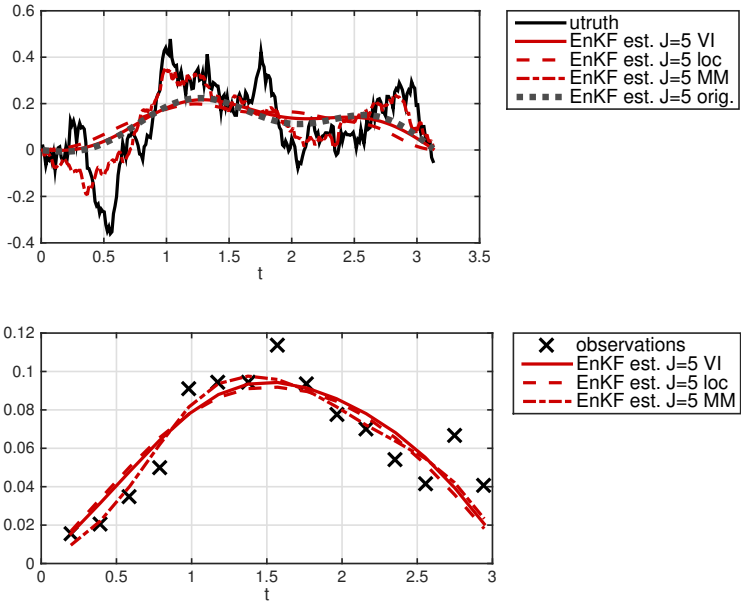

FIG. 27. *Comparison of the EnKF estimate with the truth and the observations, $J = 5$ (red) for the discussed variants, $\beta = 10$, $K = 2^4 - 1$, initial ensemble chosen based on KL expansion of $C_0 = \beta(A - id)^{-1}$.*

7. Conclusions. Our analysis and numerical studies for the ensemble Kalman filter applied to inverse problems demonstrate several interesting properties: (i) the continuous time limit exhibits structure that is hard to see in discrete time implementations used in practice; (ii) in particular, for the linear inverse problem, it reveals an underlying gradient flow structure; (iii) in the linear noise-free case the method can be completely analyzed and this leads to a complete understanding of error propagation; (iv) numerical results indicate that the conclusions observed for linear problems carry over to nonlinear problems; (v) that the widely used localization and inflation techniques can improve the method, but that the (introduced here for the first time) use of ideas from SMC hold considerable promise for further improvement; (vi) that importing stopping criteria and other regularization techniques is crucial to the effectiveness of the method, as highlighted by the work of Iglesias [14, 16]. Our future work in this area, both theoretical and computational, will reflect, and build on, these conclusions.

Acknowledgments Both authors are grateful to Dean Oliver for helpful advice, and to the EPSRC Programme Grant EQUIP for funding of this research. AMS is also grateful to DARPA and to ONR for funding parts of this research.

REFERENCES

[1] J. Anderson, *An adaptive covariance inflation error correction algorithm for ensemble filters*, Tellus A, 59 (2007), pp. 210–224.

[2] K. Bergemann and S. Reich, *A localization technique for ensemble Kalman filters*, Quarterly Journal of the Royal Meteorological Society, 136 (2010), pp. 701–707.

[3] K. Bergemann and S. Reich, *A mollified ensemble Kalman filter*, Quarterly Journal of the Royal Meteorological Society, 136 (2010), pp. 1636–1643.

[4] A. Beskos, A. Jasra, E. Muzaffer, and A. Stuart, *Sequential Monte Carlo methods for Bayesian elliptic inverse problems*, arXiv preprint arXiv:1412.4459, (2014).

[5] M. Bocquet and P. Sakov, *An iterative ensemble Kalman smoother*, Quarterly Journal of the Royal Meteorological Society, 140 (2014), pp. 1521–1535.

[6] S. Cotter, G. Roberts, A. Stuart, and D. White, *MCMC methods for functions: modifying old algorithms to make them faster*, Statistical Science, 28 (2013), pp. 424–446.

[7] M. Dashti and A. Stuart, *The Bayesian approach to inverse problems*, arXiv preprint arXiv:1302.6989, (2014).

[8] A. ELSheikh, C. Pain, F. Fang, J. Gomes, and I. Navon, *Parameter estimation of subsurface flow models using iterative regularized ensemble Kalman filter*, Stochastic Environmental Research and Risk Assessment, 27 (2013), pp. 877–897.

[9] H. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems*, vol. 375, Springer Science & Business Media, 1996.

[10] O. Ernst, B. Sprungk, and H. Starkloff, *Analysis of the ensemble and polynomial chaos Kalman filters in Bayesian inverse problems*, arXiv preprint arXiv:1504.03529, (2015).

[11] G. Evensen, *The ensemble Kalman filter: Theoretical formulation and practical implementation*, Ocean dynamics, 53 (2003), pp. 343–367.

[12] M. Goldstein and D. Wooff, *Bayes linear statistics, theory and methods*, vol. 716, John Wiley & Sons, 2007.

[13] S. Gratton, J. Mandel, et al., *On the convergence of a non-linear ensemble Kalman smoother*, arXiv preprint arXiv:1411.4608, (2014).

[14] M. Iglesias, *Iterative regularization for ensemble data assimilation in reservoir models*, Computational Geosciences, (2014), pp. 1–36.

[15] M. Iglesias, K. Law, and A. Stuart, *Ensemble Kalman methods for inverse problems*, Inverse Problems, 29 (2013), p. 045001.

[16] M. A. Iglesias, *A regularizing iterative ensemble Kalman method for pde-constrained inverse problems*, arXiv preprint arXiv:1505.03876, (2015).

[17] E. Kalnay, *Atmospheric Modeling, Data Assimilation, and Predictability*, Cambridge university press, 2003.

[18] N. Kantas, A. Beskos, and A. Jasra, *Sequential Monte Carlo methods for high-dimensional inverse problems: a case study for the Navier–Stokes equations*, SIAM/ASA Journal on Uncertainty Quantification, 2 (2014), pp. 464–489.

[19] D. Kelly, K. Law, and A. Stuart, *Well-posedness and accuracy of the ensemble Kalman filter in discrete and continuous time*, Nonlinearity, 27 (2014), p. 2579.

[20] E. Kwiatkowski and J. Mandel, *Convergence of the square root ensemble Kalman filter in the large ensemble limit*, SIAM/ASA Journal on Uncertainty Quantification, 3 (2015), pp. 1–17.

[21] K. Law, A. Stuart, and K. Zygalakis, *Data Assimilation: A Mathematical Introduction*, Springer, 2015.

[22] K. J. Law, H. Tembine, and R. Tempone, *Deterministic mean-field ensemble Kalman filtering*, SIAM Journal on Scientific Computing, 38 (2016), pp. A1251–A1279.

[23] F. Le Gland, V. Monbet, and V.-D. Tran, *Large sample asymptotics for the ensemble Kalman filter*, Research Report RR-7014, INRIA, 2009, https://hal.inria.fr/inria-00409060.

[24] J. Li and D. Xiu, *On numerical properties of the ensemble Kalman filter for data assimilation*, Computer Methods in Applied Mechanics and Engineering, 197 (2008), pp. 3574–3583.

[25] D. Oliver, A. Reynolds, and N. Liu, *Inverse theory for petroleum reservoir characterization and history matching*, Cambridge University Press, 2008.

[26] N. S. Pillai, A. M. Stuart, and A. H. Thiéry, *Noisy gradient flow from a random walk in Hilbert space*, Stochastic Partial Differential Equations: Analysis and Computations, 2 (2014), pp. 196–232.

[27] A. S. Stordal and A. H. Elsheikh, *Iterative ensemble smoothers in the annealed importance sampling framework*, Advances in Water Resources, 86 (2015), pp. 231–239, doi:10.1016/j.advwatres.2015.09.030.

[28] X. Tong, A. Majda, and D. Kelly, *Nonlinear stability and ergodicity of ensemble based*

Kalman filters, arXiv preprint arXiv:1507.08307v1, (2015).

Appendix.

LEMMA 6. *The deviations from the mean $e^{(j)}$ and the deviations from the truth $r^{(j)}$ satisfy*

$$\frac{\mathrm{d}e^{(j)}}{\mathrm{d}t} = -\frac{1}{J}\sum_{k=1}^{J} E_{jk}e^{(k)} = -\frac{1}{J}\sum_{k=1}^{J} E_{jk}r^{(k)} \tag{18}$$

and

$$\frac{\mathrm{d}r^{(j)}}{\mathrm{d}t} = -\frac{1}{J}\sum_{k=1}^{J} F_{jk}e^{(k)} = -\frac{1}{J}\sum_{k=1}^{J} F_{jk}r^{(k)}. \tag{19}$$

Proof. Recall (8):

$$\frac{\mathrm{d}u^{(j)}}{\mathrm{d}t} = \frac{1}{J}\sum_{k=1}^{J}\langle A(u^{(k)}-\overline{u}), y - Au^{(j)}\rangle_{\Gamma}\big(u^{(k)}-\overline{u}\big), \quad j = 1,\cdots,J.$$

From this it follows that

$$\frac{\mathrm{d}\overline{u}}{\mathrm{d}t} = -\frac{1}{J}\sum_{k=1}^{J}\langle A(\overline{u}-u^{\dagger}), Ae^{(k)}\rangle_{\Gamma}e^{(k)}.$$

Hence (18) follows, with the second identity following from the fact that $E\mathsf{l} = 0$. Since $u^{\dagger}$ is time-independent we also have that (19) follows, with the second identity now following from the fact that $F\mathsf{l} = 0$. ∎

LEMMA 7. *Assume that y is the image of a truth $u^{\dagger} \in \mathcal{X}$ under A. The matrices E and F satisfy the equations*

$$\frac{\mathrm{d}}{\mathrm{d}t}E = -\frac{2}{J}E^2, \quad \frac{\mathrm{d}}{\mathrm{d}t}F = -\frac{2}{J}FE, \quad \frac{\mathrm{d}}{\mathrm{d}t}R = -\frac{2}{J}FF^T.$$

As a consequence both E and F satisfy a global-in-time a priori bound, depending only on initial conditions. Explicitly we have the following. For the orthogonal matrix X defined through the eigendecomposition of $E(0)$ it follows that

$$E(t) = X\Lambda(t)X^{\top} \tag{20}$$

with $\Lambda(t) = diag\{\lambda^{(1)}(t),\ldots,\lambda^{(J)}(t)\}$, $\Lambda(0) = diag\{\lambda_0^{(1)},\ldots,\lambda_0^{(J)}\}$ and

$$\lambda^{(j)}(t) = \Big(\frac{2}{J}t + \frac{1}{\lambda_0^{(j)}}\Big)^{-1}, \tag{21}$$

if $\lambda_0^{(j)} \neq 0$, otherwise $\lambda^{(j)}(t) = 0$. The matrix R satisfies $\mathrm{Tr}\big(R(t)\big) \leq \mathrm{Tr}\big(R(0)\big)$ for all $t \geq 0$, and $F_{ij} \to 0$ at least as fast as $\frac{1}{\sqrt{t}}$ as $t \to \infty$ for each i, j and, in particular, is bounded uniformly in time.

Proof. The first equation may be derived as follows:

$$\begin{aligned}\big(\frac{\mathrm{d}}{\mathrm{d}t}E\big)_{ij} &= \frac{\mathrm{d}}{\mathrm{d}t}\langle Ae^{(i)}, Ae^{(j)}\rangle_\Gamma \\ &= -\frac{1}{J}\sum_{k=1}^{J} E_{ik}\langle Ae^{(k)}, Ae^{(j)}\rangle_\Gamma - \frac{1}{J}\sum_{k=1}^{J} E_{jk}\langle Ae^{(i)}, Ae^{(k)}\rangle_\Gamma \\ &= -\frac{2}{J}\sum_{k=1}^{J} E_{ik}E_{kj}\end{aligned}$$

as required. The second equation follows similarly:

$$\begin{aligned}\big(\frac{\mathrm{d}}{\mathrm{d}t}F\big)_{ij} &= \langle A\frac{\mathrm{d}}{\mathrm{d}t}r^{(i)}, Ae^{(j)}\rangle_\Gamma + \langle Ar^{(i)}, A\frac{\mathrm{d}}{\mathrm{d}t}e^{(j)}\rangle_\Gamma \\ &= -\frac{1}{J}\sum_{k=1}^{J} F_{ik}E_{kj} - \frac{1}{J}\sum_{k=1}^{J} F_{ik}E_{kj}\ ,\end{aligned} \tag{22}$$

as required; here we have used the fact that $F_{kj} - E_{kj}$ is independent of k and hence, since $F\mathbb{1} = 0$,

$$\sum_{k=1}^{J} F_{ik}E_{kj} = \sum_{k=1}^{J} F_{ik}F_{kj}.$$

Due to the symmetry (and positive semidefiniteness) of E, $E(0)$ is diagonalizable by orthogonal matrices, that is $E(0) = X\Lambda(0)X^\top$, where $\Lambda(0) = \mathrm{diag}\{\lambda_0^{(1)}, \ldots, \lambda_0^{(J)}\}$. The solution of the ODE for $E(t)$ is therefore given by

$$E(t) = X\Lambda(t)X^\top \tag{23}$$

with $\Lambda(t)$ satisfying the following decoupled ODE

$$\frac{\mathrm{d}\lambda^{(j)}}{\mathrm{d}t} = -\frac{2}{J}(\lambda^{(j)})^2\ . \tag{24}$$

The solution of (24) is thus given by

$$\lambda^{(j)}(t) = \big(\frac{2}{J}t + \frac{1}{\lambda_0^{(j)}}\big)^{-1}\ , \tag{25}$$

if $\lambda_0^{(j)} \neq 0$, otherwise $\lambda^{(j)}(t) = 0$. The behavior of R is described by

$$\begin{aligned}\big(\frac{\mathrm{d}}{\mathrm{d}t}R\big)_{ij} &= \langle A\frac{\mathrm{d}}{\mathrm{d}t}r^{(i)}, Ar^{(j)}\rangle_\Gamma + \langle Ar^{(i)}, A\frac{\mathrm{d}}{\mathrm{d}t}r^{(j)}\rangle_\Gamma \\ &= -\frac{1}{J}\sum_{k=1}^{J} F_{ik}F_{jk} - \frac{1}{J}\sum_{k=1}^{J} F_{jk}F_{ik}\ ,\end{aligned}$$

and thus

$$\frac{\mathrm{d}}{\mathrm{d}t}R = -\frac{2}{J}FF^\top.$$

Taking the trace of this identity gives

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{Tr}(R) = -\frac{2}{J}\|F\|_{\mathrm{Fr}}^2,$$

where $\|\cdot\|_{\mathrm{Fr}}$ is the Frobenius norm. The bound on the trace of R follows.

By the Cauchy-Schwartz inequality, we have

$$F_{ij}^2 = \langle Ar^{(i)}, Ae^{(j)}\rangle_\Gamma^2 \leq |Ar^{(i)}|_\Gamma^2 \cdot |Ae^{(j)}|_\Gamma^2 \leq C|Ae^{(j)}|_\Gamma^2 \,,$$

and hence, $F_{ij} \to 0$ at least as fast as $\frac{1}{\sqrt{t}}$ as $t \to \infty$ as required. ∎

LEMMA 8. *Assume that y is the image of a truth $u^\dagger \in \mathcal{X}$ under A and the forward operator A is one-to-one. Then*

$$Ae^{(j)}(t) = \sum_{k=1}^{J} \ell_{jk}(t) Ae^{(k)}(0),$$
$$Ar^{(j)}(t) = \sum_{k=1}^{J} q_{jk}(t) Ae^{(k)}(0) + \rho^{(j)}(t)$$

where the matrices $L = \{\ell_{jk}\}$ and $Q = \{q_{jk}\}$ satisfy

$$\frac{\mathrm{d}L}{\mathrm{d}t} = -\frac{1}{J} EL, \qquad \frac{\mathrm{d}Q}{\mathrm{d}t} = -\frac{1}{J} FL,$$

and $\rho^{(j)}(t) = \rho^{(j)}(0) = \rho^{(1)}(0)$ is the projection of $Ar^{(j)}(0)$ into the subspace which is orthogonal in $\mathcal{Y}$ to the linear span of $\{Ae^{(k)}(0)\}_{k=1}^J$, with respect to the inner product $\langle\cdot,\cdot\rangle_\Gamma$. As a consequence

$$L(t) = X\Omega(t)X^\top \tag{28}$$

with $\Omega(t) = diag\{\omega^{(1)}(t), \ldots, \omega^{(J)}(t)\}$, $\Omega(0) = I$ and

$$\omega^{(j)}(t) = \Big(\frac{2}{J}\lambda_0^{(j)} t + 1\Big)^{-\frac{1}{2}} . \tag{29}$$

We also assume that the rank of the subspace spanned by the vectors $\{Ae^{(j)}(t)\}_{j=1}^J$ is equal to $\tilde{J}$ and that (after possibly reordering the eigenvalues) $\lambda^{(1)}(t) = \ldots = \lambda^{(J-\tilde{J})}(t) = 0$ and $\lambda^{(J-\tilde{J}+1)}(t), \ldots, \lambda^{(J)}(t) > 0$. It then follows that $\omega^{(1)}(t) = \ldots = \omega^{(J-\tilde{J})}(t) = 1$. Furthermore, $L(t)x^{(k)} = L(t)^{-1}x^{(k)} = x^{(k)}$ for all $t \geq 0$, $k = 1, \ldots, J - \tilde{J}$, where $x^{(k)}$ are the columns of X. Without loss of generality we may assume that $Q(t)x^{(k)} = 0$ for all $t \geq 0$, $k = 1, \ldots, J - \tilde{J}$.

Proof. Differentiating expression (26a) and substituting in (18) from Lemma 6 gives

$$\sum_{m=1}^{J} \frac{\mathrm{d}\ell_{jm}}{\mathrm{d}t} Ae^{(m)}(0) = -\frac{1}{J}\sum_{k=1}^{J}\sum_{m=1}^{J} E_{jk}\ell_{km} Ae^{(m)}(0).$$

Reordering the double summation on the right-hand side and re-arranging we obtain

$$\sum_{m=1}^{J} \Big(\frac{\mathrm{d}\ell_{jm}}{\mathrm{d}t} + \frac{1}{J}\sum_{k=1}^{J} E_{jk}\ell_{km}\Big) Ae^{(m)}(0) = 0.$$

This is satisfied identically if equation (27a) holds. By uniqueness choosing the $Ae^{(j)}(t)$ to be defined in this way gives the unique solution for their time evolution.

Now we differentiate expression (26b) and substitute into (19) from Lemma 6. A similar analysis to the preceding yields

$$\sum_{m=1}^{J}\Big(\frac{\mathrm{d}Q_{jm}}{\mathrm{d}t}+\frac{1}{J}\sum_{k=1}^{J}F_{jk}\ell_{km}\Big)Ae^{(m)}(0)+\frac{\mathrm{d}\rho^{(j)}}{\mathrm{d}t}=0.$$

Again this can be satisfied identically if equation (27b) holds and if $\rho^{(j)}(t)$ is the constant function as specified above. By uniqueness we have the desired solution. The independence of $\rho^{(j)}(t)$ with respect to j, i.e. $\rho^{(j)}(0)=\rho^{(1)}(0)$, follows from the fact that $\rho^{(j)}(0)$ is the function inside the norm $\|\cdot\|_\Gamma$ which is found by choosing the vector $q_j:=\{q_{jk}\}_{k=1}^J$ so as to minimize the functional

$$\|Ar^{(j)}(0)-\sum_{k=1}^{J}q_{jk}(0)Ae^{(k)}(0)\|_\Gamma.$$

From the definition of the $r^{(j)}$ and $e^{(j)}$ this is equivalent to determining the function inside the norm $\|\cdot\|_\Gamma$ found by choosing the vector $q_j:=\{q_{jk}\}_{k=1}^J$ so as to minimize the functional

$$\|Au^{(j)}(0)-\sum_{k=1}^{J}q_{jk}(0)A(u^{(k)}(0)-\bar{u}(0))-Au^\dagger\|_\Gamma.$$

This in turn is equivalent to determining the function inside the norm $\|\cdot\|_\Gamma$ found by choosing the vector $\tilde{q}:=\{\tilde{q}_k\}_{k=1}^J$ so as to minimize the functional

$$\|\sum_{k=1}^{J}\tilde{q}_kAu^{(k)}(0)-Au^\dagger\|_\Gamma$$

and is hence independent of j. Our assumptions on the span of $\{Ae^{(j)}(t)\}_{j=1}^J$ imply that $E(0)$ has exactly $J-\tilde{J}$ zero eigenvalues, corresponding to eigenvectors $\{x^{(k)}\}_{k=1}^{J-\tilde{J}}$ with the property that

$$\sum_{j=1}^{J}x_j^{(k)}Ae^{(j)}(0)=0.$$

(One of these vectors $x^{(k)}$ is of course I so that $\tilde{J}\geq 1$.) As a consequence we also have

$$E(0)x^{(k)}=F(0)x^{(k)}=0\quad k=1,\cdots,J-\tilde{J}. \tag{30}$$

The fact that $L(t)x^{(k)}=x^{(k)}$ for all $t\geq 0$, $k=1,\ldots,\tilde{J}$ is immediate from the fact that $L=X\Omega X^\top$, because $x^{(k)}$ is the eigenvector corresponding to eigenvalue $\omega^{(k)}(t)=1$ $k=1,\ldots,\tilde{J}$; an identical argument shows the same for $L(t)^{-1}$. The property that $Q(t)x^{(k)}=0$ for all $t\geq 0$, $k=1,\ldots,\tilde{J}$ follows by choosing $Q(0)$ so that $Q(0)x^{(k)}=0$, which is always possible because the $x^{(k)}$ are eigenvectors with corresponding eigenvalues $\lambda^{(k)}=0$, and then noting that $Q(t)x^{(k)}=0$ for all time because $F(t)L(t)x^{(k)}=F(t)x^{(k)}=0$ for all $t\geq 0$. The last item is zero because $Ex^{(k)}=0$ and because $\frac{\mathrm{d}}{\mathrm{d}t}F=-\frac{2}{J}FE$; we also use that $F(0)x^{(k)}=0$ from (30).