Journal of Computational Physics 463 (2022) 111262

Contents lists available at ScienceDirect

Journal of Computational Physics

www.elsevier.com/locate/jcp

Iterated Kalman methodology for inverse problems

Daniel Zhengyu Huang *, Tapio Schneider, Andrew M. Stuart

California Institute of Technology, Pasadena, CA, United States of America

ARTICLE INFO

Article history:
Received 1 February 2021
Received in revised form 26 April 2022
Accepted 27 April 2022
Available online 4 May 2022

Keywords:
Inverse problem
Derivative-free optimization
Extended Kalman methods
Ensemble Kalman methods
Unscented Kalman methods
Interacting particle systems

ABSTRACT

This paper is focused on the optimization approach to the solution of inverse problems. We introduce a stochastic dynamical system in which the parameter-to-data map is embedded, with the goal of employing techniques from nonlinear Kalman filtering to estimate the parameter given the data. The extended Kalman filter (which we refer to as ExKI in the context of inverse problems) can be effective for some inverse problems approached this way, but is impractical when the forward map is not readily differentiable and is given as a black box, and also for high dimensional parameter spaces because of the need to propagate large covariance matrices. Application of ensemble Kalman filters, for example use of the ensemble Kalman inversion (EKI) algorithm, has emerged as a useful tool which overcomes both of these issues: it is derivative free and works with a low-rank covariance approximation formed from the ensemble. In this paper, we work with the ExKI, EKI, and a variant on EKI which we term unscented Kalman inversion (UKI).

The paper contains two main contributions. Firstly, we identify a novel stochastic dynamical system in which the parameter-to-data map is embedded. We present theory in the linear case to show exponential convergence of the mean of the filtering distribution to the solution of a regularized least squares problem. This is in contrast to previous work in which the EKI has been employed where the dynamical system used leads to algebraic convergence to an unregularized problem. Secondly, we show that the application of the UKI to this novel stochastic dynamical system yields improved inversion results, in comparison with the application of EKI to the same novel stochastic dynamical system.

The numerical experiments include proof-of-concept linear examples and various applied nonlinear inverse problems: learning of permeability parameters in subsurface flow; learning the damage field from structure deformation; learning the Navier-Stokes initial condition from solution data at positive times; learning subgrid-scale parameters in a general circulation model (GCM) from time-averaged statistics.

1. Introduction

1.1. Overview

This paper is devoted to optimization approaches to calibrating models with observational data. The basic problem is formulated as recovering unknown model parameters $\theta \in \mathbb{R}^{N_\theta}$ from noisy observation $y \in \mathbb{R}^{N_y}$ given by

[*] Corresponding author.
E-mail addresses: dzhuang@caltech.edu (D.Z. Huang), tapio@caltech.edu (T. Schneider), astuart@caltech.edu (A.M. Stuart).

https://doi.org/10.1016/j.jcp.2022.111262
0021-9991/© 2022 Elsevier Inc. All rights reserved.

$$y = \mathcal{G}(\theta) + \eta; \tag{1}$$

here $\mathcal{G}$ denotes the parameter-to-data map which, for the applications we have in mind, generally requires solving partial differential equations, and $\eta \sim \mathcal{N}(0, \Sigma_\eta)$ denotes the Gaussian observation error. Consider now the stochastic dynamical system

$$\text{evolution:} \qquad \theta_{n+1} = \alpha\theta_n + (1-\alpha)r_0 + \omega_{n+1}, \qquad \omega_{n+1} \sim \mathcal{N}(0, \Sigma_\omega), \tag{2a}$$

$$\text{observation:} \qquad y_{n+1} = \mathcal{G}(\theta_{n+1}) + \nu_{n+1}, \qquad \nu_{n+1} \sim \mathcal{N}(0, \Sigma_\nu). \tag{2b}$$

We assume that the artificial evolution error covariance $\Sigma_\omega \succ 0$, the artificial observation error covariance $\Sigma_\nu \succ 0$, and the regularization parameter $\alpha \in (0, 1]$, whilst r_0 is an arbitrary vector.[1] We study methods to determine θ from y given by (1) by employing filtering methods to find θ_n given $Y_n := \{y_\ell\}_{\ell=1}^n$, in the setting where $y_\ell \equiv y$ for all $\ell \in \mathbb{N}$.

Note that dynamical system (2a) for θ_n has, for $\alpha \in (0, 1)$, statistical equilibrium given by the Gaussian $\mathcal{N}(r_0, (1-\alpha^2)^{-1}\Sigma_\omega)$. The output of this statistical model is then repeatedly exposed to the observations, expressed via (2b) with y_{n+1} set to the data y, and hence it is intuitive that filtering methods will deliver an estimate of θ solving (1) as $n \to \infty$. Such a method, in the special case $\alpha = 1, \Sigma_\omega = 0, \Sigma_\nu = \Sigma_\eta$, is the basis of the ensemble Kalman inversion (EKI) algorithm as proposed in [1]. The two main takeaway messages of this paper are firstly to highlight the benefits of choosing $\alpha \in (0, 1)$ and $\Sigma_\omega \succ 0$, and secondly to demonstrate that application of the unscented Kalman filter improves on the ensemble Kalman filter, leading to unscented Kalman inversion (UKI).

The primary issue with the choice $\alpha = 1$ is that it leads to over-fitting for problems in which $N_\theta > N_y$, as shown in [1]. One approach to deal with this is to use an adaptive modification of the basic EKI algorithm, based on an analogy with the Levenberg-Marquardt algorithm, as developed in [2]; however, this leads to a need for stopping criterion and the area is still being developed [3]. Another approach is to build Tikhonov regularization directly into the inverse problem, before applying a filtering algorithm to (2) with $\alpha = 1, \Sigma_\omega = 0$, an approach introduced in [4]. However, this leads to an algorithm which requires the inversion of covariance matrices on spaces of dimension $N_\theta + N_y$ which is undesirable for many problems concerning inference about fields, where $N_\theta \gg 1$. This issue is removed if the continuum limit of the algorithm is used [4]. However, practical experience with using time-steppers for continuum limits of ensemble Kalman filtering algorithms is in its infancy and current implementations of the methods in [4–8] are not competitive with algorithms which start directly from a discrete time formulation.

Central to both the optimization and probabilistic approaches to inversion is the regularized objective function $\Phi_R(\theta)$ defined by

$$\Phi_R(\theta) := \Phi(\theta) + \frac{1}{2}\|\Sigma_0^{-\frac{1}{2}}(\theta - r_0)\|^2, \tag{3a}$$

$$\Phi(\theta) := \frac{1}{2}\|\Sigma_\eta^{-\frac{1}{2}}(y - \mathcal{G}(\theta))\|^2, \tag{3b}$$

where $\Sigma_\eta \succ 0$ normalizes the model-data misfit Φ by means of the known error statistics of the noise, prior mean r_0 encodes prior information about θ, and prior covariance $\Sigma_0 \succ 0$ normalizes the prior information. We will connect the parameters of (2) for θ to a form of regularization of the inverse problem. In this context it is worth noticing that, for linear problems, the implied Tikhonov regularization has implied mean r_0, whilst the implied covariance Σ_0 of the regularization term is defined implicitly via limit of an iterative procedure. Parameter $\alpha \in (0, 1)$ controls the size of the regularization effect; and when $\alpha = 1$ the regularization effect disappears, along with dependence of (2a) on r_0. Thus $\alpha = 1$ is useful primarily for over-determined problems.

1.2. Our contributions

We make the following contributions to the study of the solution of inverse problems by means of filtering methods:

- we introduce a filtering-based approach to solving the inverse problem (1), based on the novel stochastic dynamical system formulation (2);
- by studying linear problems we demonstrate that the methodology induces a form of Tikhonov regularization and we prove an exponential convergence of the algorithm to the minimizer of the Tikhonov-regularized problem, in the linear case;
- we introduce a Gaussian approximation for the filtering distribution defined by (2) and, from it, derive extended Kalman, ensemble Kalman and unscented Kalman (ExKI, EKI and UKI respectively) algorithms for the inverse problem (1), applicable in the general nonlinear case;

[1] We write $A \succ 0$ when A is strictly positive-definite, and will also write $A \prec B$ when $B - A$ is strictly positive-definite and $A \preceq B$ when $B - A$ is positive semi-definite.

- the algorithms are tested on a wide range of problems, including linear test problems, inversion for spatial fields in a variety of continuum mechanics applications, and the learning of parameters in chaotic dynamical systems, using time-averaged data;
- we show that UKI outperforms EKI, with both employed in the context of the stochastic dynamical model (2), for a wide range of inverse problems with unknown parameter space of moderate dimension.

Taken together, the theoretical framework we develop and the numerical results we present show that the UKI, applied to the stochastic dynamical system (2), is a competitive methodology for solving inverse problems and parameter estimation problems defined by an expensive black-box forward model; indeed the UKI is shown to outperform the EKI in settings where the number of parameters N_θ is of moderate size and the black-box is not readily differentiable so that ExKI methods are not applicable. Other ensemble filters, such as the ensemble adjustment and ensemble transform Kalman filters could also be used in place of unscented Kalman filters, and similar performance is to be expected. This issue is explored in detail in [9] where ideas introduced in this paper are developed further in order to approximate the Bayesian posterior distribution for inverse problem (1). We note that, as with the use of most nonlinear variants of the Kalman filter, rigorous justification beyond the linear setting is not currently available, but that our numerical results demonstrate effectiveness in a wide range of nonlinear inverse problems. The use of interacting particle systems to solve inverse problems with multimodal distributions, far from Gaussian, is considered in [10] and the derivation of mean-field limits of ensemble Kalman methods for inversion, viewed as interacting particle systems is established in [11,12].

We conclude this introductory section with a deeper literature review relating to the contributions we make in this paper, in Subsection 1.3. Then, in Section 2 we introduce a conceptual algorithm based on a Gaussian approximation of the filtering distribution associated with (2); we then derive the ExKI, UKI, and EKI algorithms as approximations to this conceptual Gaussian algorithm. In Section 3 we study the methodology for linear problems, obtaining insight into the regularization conferred by (2a); we study the relationship of the methodology to other gradient-based optimization techniques; we derive continuous-time limits in the nonlinear setting. Section 4 describes variants on the basic conceptual algorithm that may be useful in some settings, and in Section 5 we present numerical results demonstrating the performance of the inversion methodology introduced in this paper. The code relating to numerical experiments presented in Section 5 is accessible online:

https://github.com/Zhengyu-Huang/InverseProblems.jl

1.3. Literature review

The focus of this paper is mainly on derivative-free inversion by means of iterative techniques aimed at solving the optimization problem defined by minimization of Φ_R, or variants of this problem [13]. However, even in the optimization setting, the methods introduced in this paper are closely related to iterative methods applied in Bayesian (probabilistic) inversion. In the Bayesian approach to the inverse problem (1) [14,15] the posterior distribution is given by

$$\mu(d\theta) = \frac{1}{Z}\exp\bigl(-\Phi(\theta)\bigr)\mu_0(d\theta), \tag{4}$$

where $\mu_0 = \mathcal{N}(r_0, \Sigma_0)$ is the prior and μ is the posterior. A commonly adopted iterative approach to solving the problem of sampling from μ is the finite time approach known as sequential Monte Carlo (SMC) – see [16,17], and [18] for applications to inverse problems. The basic idea, upon which there are many variants, is to consider the sequence of measures μ_n defined by

$$\mu_{n+1}(d\theta) = \frac{1}{Z_n}\exp\bigl(-h\Phi(\theta)\bigr)\mu_n(d\theta). \tag{5}$$

Note, then, that if $Nh = 1$ it follows that $\mu_N = \mu$. Each step $\mu_n \mapsto \mu_{n+1}$ may be approximated by a particle-based filtering algorithm, leading to a variety of algorithms used in practice, involving a fixed finite number of steps N. Furthermore, continuous-time limits of this methodology may also be derived by taking $N \to \infty$ and $h \to 0$ with $Nh = 1$, giving insight into the algorithms; see [19,8].

On the other hand, if $h = 1$ is fixed and the measures μ_n are studied in the limit $n \to \infty$, they will tend to concentrate on minimizers of Φ, restricted to the support of μ_0, as the following identity shows:

$$\mu_n(d\theta) = \frac{1}{\bigl(\Pi_{\ell=0}^{n-1} Z_\ell\bigr)}\exp\bigl(-n\Phi(\theta)\bigr)\mu_0(d\theta). \tag{6}$$

This corresponds to an infinite time approach.

The finite time approach was developed for probabilistic problems; the infinite time approach is focused on optimization. This paper will build on the latter, optimization, approach to the problem. However, we note that, other than restriction of μ_n to the support of μ_0, regularization is lost in this approach since it focuses on minimizing $\Phi(\cdot)$ and not $\Phi_R(\cdot)$. To introduce regularization we consider the iteration

$$\mu_{n+1}(d\theta) = \frac{1}{Z_n} \exp\bigl(-\Phi(\theta)\bigr) P_n \mu_n(d\theta). \tag{7}$$

To address the issue of regularization, we will choose P_n to be the Markov kernel associated with a first-order autoregressive (AR1) process as defined by (2a); it is thus independent of $n : P_n \equiv P$. The resulting dynamic on measures μ_n defined by (7) corresponds to the filtering distribution for $\theta_n|Y_n$ defined by the stochastic dynamical system (2). We note that within SMC P_n is also introduced in a similar fashion in (5), but in that context it is chosen to be a μ_n-invariant Markov kernel so that $P_n\mu_n = \mu_n$, typically from MCMC; in this setting P_n is indeed n-dependent. Note that μ_n is not invariant with respect to P with the AR1 choice we make: thus the introduction of P_n in our setting differs from its use in SMC; this is because we are solving an optimization problem via iteration over n, and not the sampling problem which morphs the prior at time $n = 0$ into the posterior at time $n = N$. The specific choice of P_n made in our work, namely the Markov kernel P defined by an AR1 process, is made in order to regularize the iterative optimization approach to inversion encapsulated in (6). Once we apply particle methods, the presence of P plays the role of avoiding ensemble collapse [5,6,4]. We also note that, in contrast to SMC, the initial measure μ_0 in (7) does not need to be the prior distribution – it may be chosen arbitrarily, although a natural choice is the stationary measure for the AR1 process.

In the case where $\mathcal{G}$ is linear, (7) delivers a sequence of measures, which are defined through a Kalman filter. Our analysis of the underlying filtering problem in Subsection 3.1, which considers the linear Gaussian setting, thus constitutes an analysis of the Kalman filter for a specific state-space model with a specific choice of data. In order to deal with a range of cases, including exponential convergence, algebraic convergence and divergence of the mean/covariances of the filter, we introduce an explicit unified analysis of the Kalman filter in our setting. We note, however, that this is a well-trodden field and that variants on some of our results can be obtained from the existing literature [20,21].

The method we introduce and study in this paper arises from the application of ideas from Kalman filtering to the problem of approximating the distribution of $\theta_n|Y_n$. The Kalman filter itself applies to the case of linear $\mathcal{G}$ [22,23]. When $\mathcal{G}$ is nonlinear the methods can be generalized by use of the extended Kalman filter (ExKF) [24] which is based on linearization and application of Kalman methodology. However this method suffers from two drawbacks which hamper its application in many large-scale applications: (a) it requires a derivative of the forward map $\mathcal{G}(\cdot)$; and (b) the approach scales poorly to high dimensional parameter spaces where $N_\theta \gg 1$, because of the need to sequentially update covariances in $\mathbb{R}^{N_\theta \times N_\theta}$. Thus, despite an early realization that Kalman-based methods could be useful for large-scale filtering problems arising in the geosciences [25], the methods did not become practical in this context until the work of Evensen [26]. This revolutionary paper introduced the ensemble Kalman filter (EnKF) the essence of which is to avoid the linearization of the dynamics and sequential updating of the covariance, and instead use a low-rank approximation of the covariance found by maintaining an ensemble of estimates for $\theta_n|Y_n$ at every step n. These ensemble Kalman methods have been widely adopted in the geosciences, not only because they are effective for high dimensional parameter spaces, but also because they are derivative-free, requiring only $\mathcal{G}$ as a black box. Their use in the solution of inverse problems via iterative methods was pioneered in subsurface inversion [27,28] where the perspective of fixing $h \ll 1$ and iterating until $n = N = 1/h$ was used, so that μ_N is viewed as an approximation of the posterior, provided μ_0 is chosen as the prior. These papers thus view the ensemble methodology as a way of sampling from the posterior and have elements in common with SMC; this idea is also implicit in the paper [19], which is focused on data assimilation, and addresses the solution of a Bayesian inverse problem each time new data is received.

In [1] the Kalman methodology for inversion was revisited from the optimization perspective, based on fixing $h = 1$ and iterating in n, leading to an algorithm we will refer to as ensemble Kalman inversion (EKI). The paper [2] introduced a novel approach to regularizing the iterative method, by drawing an analogy with the Levenberg-Marquardt algorithm (LMA) [29]; see also [3]. Subsequent variants on the iterative optimization approach demonstrate how to introduce Tikhonov regularization into the EKI algorithm [4] and the paper [6] shows that adding noise to the iteration can lead to approximate Bayesian inversion, a method we will refer to as ensemble Kalman sampling (EKS) and which is further analyzed in [7,30]. The EKS provides a different approach to the problem of Bayesian inversion from the ones pioneered in [27,28] since it does not require starting with draws from the prior μ_0, but instead relies on ergodicity and iteration to large n; the methods in [27,28] must be started with draws from the prior μ_0 and iterated for precisely $n = 1/h$ steps, and are hence more rigid in their requirements. Since the ensemble methods do not, in general, accurately approximate the true posterior distribution [31,32] outside Gaussian scenarios, the derivative-free optimization perspective is arguably a more natural avenue within which to analyze ensemble inversion. However recent work demonstrates how a derivative-free multiscale stochastic sampling method can usefully take the output of EKS as a preconditioner for a method which provably approximates the true posterior distribution [33]; in that context, the EKS is central to making the method efficient. Furthermore, in recent interesting work, it has been shown how to reweight ensemble Kalman methods to recover statistical consistency in the non-Gaussian setting [8]; however computation of the weights requires gradients of $\mathcal{G}$ and hence is not practical for many of the problems where ensemble methods are most useful.

Within the control theory literature, and parallel to the development of the ensemble Kalman filter, the unscented Kalman filter (UKF) was introduced [34,35]. Like the ensemble Kalman methods, this method also sidesteps the need to sequentially update the derivative of the forward model as part of the covariance update; but, in the primary difference from ensemble Kalman methods, particles (sigma points) are chosen deterministically, and a quadrature rule is applied within a Gaussian approximation of the filter. This paper is to establish a framework for the development of unscented Kalman methods for inverse problems, based on (2): we formalize and demonstrate the power of unscented Kalman inversion (UKI) techniques. We also formalize extended Kalman inversion (ExKI) as a general purpose methodology for parameter learning and derive ExKI, UKI, and UKI as different approximations of a conceptual Gaussian methodology for the (in general non-Gaussian) filtering problem defined by (2).

Inverse and parameter estimation problems are ubiquitous in engineering and scientific applications. Applications that motivate this work include global climate model calibration [36–38], material constitutive relation calibration [39–41], seismic inversion in geophysics [42,43], and medical tomography [44,45]. These problems are generally highly nonlinear, may feature multiple scales, and may include chaotic and turbulent phenomena. Moreover, the observational data is often noisy and the inverse problem may be ill-posed. We note, also, that a number of inverse problems of interest may involve a moderate number of unknown parameters N_θ, yet may involve the solution of a very expensive forward model $\mathcal{G}$ depending on those parameters; furthermore, $\mathcal{G}$ may not be differentiable with respect to the parameters, or may be complex to differentiate as it is given as a black box.

In the nonlinear setting of state estimation, there are three primary types of Kalman filters [46–48]: the extended Kalman filter (ExKF), the unscented Kalman filter (UKF), and the ensemble Kalman filter (EnKF). The use of Kalman based methodology as a non-intrusive iterative method for parameter estimation originates in the papers [49,50] which were based on the ExKF, hence requiring derivative $d\mathcal{G}$, and its adjoint, to propagate covariances; the use of derivative-free ensemble methods was then developed systematically in the papers [27,28], in the SMC context, followed by the iterate for optimization EKI approach [1]. Derivative-free ensemble inversion and parameter estimation are particularly suitable for complex multiphysics problems requiring coupling of different solvers, such as fluid-structure interaction [51–54] and general circulation models [55] and methods containing discontinuities such as the immersed/embedded boundary method [56–59] and adaptive mesh refinement [60,61]. Furthermore, derivative-free ensemble inversion and parameter estimation has been demonstrated to be effective in the context of forward models defined by chaotic dynamical systems [62] where adjoint-based methods fail to deliver meaningful sensitivities [63,64]. These wide-ranging potential applications form motivation for developing other derivative-free Kalman based inversion and parameter estimation techniques, and in particular, the unscented Kalman methods developed here.

There is already some work in which unscented Kalman methods are used for parameter inversion. Extended, ensemble and unscented Kalman inversions have been applied to train neural networks [49,50,35,65] and EKI has been applied in the oil industry [66,27,28]. Dual and joint Kalman filters [67,35] have been designed to simultaneously estimate the unknown states and the parameters [67,68,35,69,70] from noisy sequential observations. However, whilst the EKI has been systematically developed and analyzed as a general purpose methodology for the solution of inverse and parameter estimation problems, the same is not the case for UKI.

Continuous-time limits and gradient flow structure of the EKI have been introduced and studied in [19,71,5,72,73,11,12]. This work led to the development of variants on the EKI, such as the Tikhonov-EKI (TEKI) [4] and the EKS [6]. We will develop study of continuous-time limits for the UKI, and variants including an unscented Kalman sampler (UKS), in this paper. There are interesting links to the Levenberg–Marquardt Algorithm (LMA) [74,29], as introduced in [2] and developed further in [75,76,3]. We will further refine the idea, which provides insights into understanding and improving the nonlinear Kalman inversion methodology as introduced here.

Finally, we mention that there are other derivative-free optimization techniques which are based on interacting particle systems, but are not Kalman based. Rather these methods are based on consensus-forming mean-field models, and their particle approximations, leading to consensus-based optimization [77] and consensus-based sampling [78]. The paper [33] also provides an alternative derivative-free approach to optimization and sampling for inverse problems, using ideas from multiscale dynamical systems.

2. Nonlinear Kalman inversion algorithms

Recall that the basic approach to inverse problems that we adopt in this paper is to pair the parameter-to-data relationship encoded in (1) with a stochastic dynamical system for the parameter, resulting in (2). We then employ techniques from filtering to approximate the distribution μ_n of $\theta_n|Y_n$. A useful way to think of updating μ_n is through the prediction and analysis steps [79,80]: $\mu_n \mapsto \hat{\mu}_{n+1}$, and then $\hat{\mu}_{n+1} \mapsto \mu_{n+1}$, where $\hat{\mu}_{n+1}$ is the distribution of $\theta_{n+1}|Y_n$. In Subsection 2.1 we first introduce a Gaussian approximation of the analysis step, leading to an algorithm which maps the space of Gaussian measures into itself at each step of the iteration; it is not implementable in general, but it is a useful conceptual algorithm. Subsection 2.2 shows how this algorithm can be made practical, for low to moderate dimension N_θ and assuming that $d\mathcal{G}$ is available, by means of the ExKF, a form of linearization of the conceptual algorithm; we refer to this as ExKI. In Subsection 2.3 we show how the UKI algorithm may be derived by applying a quadrature rule to evaluate certain integrals appearing in the conceptual Gaussian approximation. Subsection 2.4 connects the conceptual algorithm with the EKI, an approach in which ensemble approximation of the integrals is used.

2.1. Gaussian approximation

This conceptual algorithm maps Gaussians into Gaussians, and henceforth it is referred to as the Gaussian Approximation Algorithm (GAA). Assume that $\mu_n \approx \mathcal{N}(m_n, C_n)$. The GAA is a mapping from (m_n, C_n) into (m_{n+1}, C_{n+1}) which reduces to the Kalman filter in the linear setting. The algorithm proceeds by determining the joint distribution of $\theta_{n+1}, y_{n+1}|Y_n$, assuming that $\theta_n|Y_n$ is Gaussian $\mathcal{N}(m_n, C_n)$. We then project[2] this joint distribution onto a Gaussian by computing its mean and covariance. And finally, we compute the conditional distribution of this joint Gaussian on observed y_{n+1} to obtain a Gaussian approximation $\mathcal{N}(m_{n+1}, C_{n+1})$ to μ_{n+1}, the distribution of $\theta_{n+1}|Y_{n+1}$.

The projection of the joint distribution of $\{\theta_{n+1}, y_{n+1}\}|Y_n$ onto a Gaussian distribution has the form

$$\mathcal{N}\Big(\begin{bmatrix}\widehat{m}_{n+1}\\ \widehat{y}_{n+1}\end{bmatrix}, \begin{bmatrix}\widehat{C}_{n+1} & \widehat{C}^{\theta y}_{n+1}\\ \widehat{C}^{\theta y}_{n+1}{}^{T} & \widehat{C}^{yy}_{n+1}\end{bmatrix}\Big); \tag{8}$$

we now define all the components of the mean and covariance. Note that, under (2a), $\hat{\mu}_{n+1}$ is also Gaussian if μ_n is Gaussian. The use of (2a) shows that

$$\begin{aligned}\widehat{m}_{n+1} &= \mathbb{E}[\theta_{n+1}|Y_n] = \alpha m_n + (1-\alpha) r_0,\\ \widehat{C}_{n+1} &= \mathrm{Cov}[\theta_{n+1}|Y_n] = \alpha^2 C_n + \Sigma_\omega.\end{aligned} \tag{9}$$

Then, with $\mathbb{E}$ denoting expectation with respect to $\theta_{n+1}|Y_n \sim \mathcal{N}(\widehat{m}_{n+1}, \widehat{C}_{n+1})$, we have

$$\begin{aligned}\widehat{y}_{n+1} =& \mathbb{E}[\mathcal{G}(\theta_{n+1})|Y_n],\\ \widehat{C}^{\theta y}_{n+1} =& \mathrm{Cov}[\theta_{n+1}, \mathcal{G}(\theta_{n+1})|Y_n],\\ \widehat{C}^{yy}_{n+1} =& \mathrm{Cov}[\mathcal{G}(\theta_{n+1})|Y_n] + \Sigma_\nu.\end{aligned} \tag{10}$$

Computing the conditional distribution of the joint Gaussian in (8) to find $\theta_{n+1}|\{Y_n, y_{n+1}\} = \theta_{n+1}|Y_{n+1}$ gives the following expressions for the mean m_{n+1} and covariance C_{n+1} of the approximation to μ_{n+1}:

$$\begin{aligned}m_{n+1} &= \widehat{m}_{n+1} + \widehat{C}^{\theta y}_{n+1}(\widehat{C}^{yy}_{n+1})^{-1}(y_{n+1} - \widehat{y}_{n+1}),\\ C_{n+1} &= \widehat{C}_{n+1} - \widehat{C}^{\theta y}_{n+1}(\widehat{C}^{yy}_{n+1})^{-1}\widehat{C}^{\theta y}_{n+1}{}^{T}.\end{aligned} \tag{11}$$

Equations (9), (10) and (11) define the GAA. As a method for solving the inverse problem (1), the GAA is implemented by assuming all observations $\{y_n\}$ are identical to y and iterating in n. With this assumption, we may write the algorithm as

$$(m_{n+1}, C_{n+1}) = F(m_n, C_n; \mathcal{G}, r_0, \Sigma_\omega), \tag{12}$$

noting that the mapping is dependent on $\mathcal{G}$ and on the mean and covariance of the assumed auto-regressive dynamics for $\{\theta_n\}$.[3]

In the setting where $\mathcal{G}$ is linear, the Gaussian ansatz used in the derivation of the conceptual algorithm is exact, the integrals appearing in (10) have closed form, and the algorithm reduces to the Kalman filter applied to (2), with a particular assumption on the data stream $\{y_n\}$. In Subsection 3.1 we will show, again in the setting where $\mathcal{G}$ is linear, that the mean of this iteration converges to a minimizer of Φ_R given by (3), in which the prior covariance of the regularization Σ_0 is defined by solution of a linear equation depending on the choices of α, Σ_ω, and Σ_ν, as well as on $\mathcal{G}$.

In the nonlinear setting, to make an implementable algorithm from the GAA encapsulated in equations (9) to (11), it is necessary to approximate the integrals appearing in (10). When extended, unscented and ensemble Kalman filters are applied, respectively, to make such approximation, we obtain the ExKI, UKI, and EKI algorithms. The extended, unscented, and ensemble approaches to this are detailed in the following three subsections. Underlying all of them is the following property of the GAA encapsulated in Proposition 1.

We recall the idea of affine invariance, introduced for MCMC methods in [82], motivated by the attribution of the empirical success of the Nelder-Mead algorithm [83] for optimization to a similar property; further development of the method in the context of sampling algorithms may be found in [84,7]. In words an iteration is affine invariant if an invertible linear transformation of the variable being iterated makes no difference to the algorithm and hence to the convergence properties of the algorithm; this has the desirable consequence that performance of the method is independent of the aspect ratio in highly anisotropic objective functions.

Consider the invertible mapping from $x \in \mathbb{R}^{N_\theta}$ to ${}^*x \in \mathbb{R}^{N_\theta}$ defined by ${}^*x = Ax + b$. Then define ${}^*\mathcal{G}(\theta) = \mathcal{G}\big(A^{-1}(\theta - b)\big)$, ${}^*r_0 = Ar_0 + b$ and ${}^*\Sigma_\omega = A\Sigma_\omega A^T$.

Proposition 1. *Define, for all $n \in \mathbb{Z}^{0+}$,*

$${}^*m_n = Am_n + b \qquad {}^*C_n = AC_nA^T.$$

[2] We refer to this as “projection” because it corresponds to finding the closest Gaussian p to the joint distribution of $\theta_{n+1}, y_{n+1}|Y_n$ with respect to variation in the second argument of the (nonsymmetric) Kullback-Leibler divergence [81][Theorem 4.5].

[3] F also depends on α and Σ_ν but we suppress this dependence for economy of notation; the highlighted dependence is what is relevant in Proposition 1.

Then

$$({}^*m_{n+1}, {}^*C_{n+1}) = F({}^*m_n, {}^*C_n; {}^*\mathcal{G}, {}^*r_0, {}^*\Sigma_\omega). \tag{13}$$

Proof. The proof is in Appendix A. □

The key observation of the previous theorem is that the same map F applies in the new coordinates. This establishes the property of affine invariance, noting that only $\mathcal{G}, r_0, \Sigma_\omega$ need to be transformed as the affine map applies only on the signal space for $\{\theta_n\}$ and not the observation space for $\{y_n\}$.

2.2. Extended Kalman inversion

Consider the GAA defined by equations (9) to (11). The ExKI algorithm follows from invoking the approximations

$$\mathcal{G}(\theta_{n+1}) \approx \mathcal{G}(\widehat{m}_{n+1}) + d\mathcal{G}(\widehat{m}_{n+1})(\theta_{n+1} - \widehat{m}_{n+1}) \tag{14}$$

in the analysis updates for the mean and covariance respectively. In particular both the mean and the covariances in (10) can be evaluated in closed form with the approximation (14). The approximations are valid if the fluctuations around the mean state are small, say of $\mathcal{O}(\epsilon) \ll 1$, and all the covariances are $\mathcal{O}(\epsilon^2)$. This results in the following algorithm:

- Prediction step:

$$\begin{aligned} \widehat{m}_{n+1} =& \alpha m_n + (1-\alpha) r_0, \\ \widehat{C}_{n+1} =& \alpha^2 C_n + \Sigma_\omega. \end{aligned} \tag{15}$$

- Analysis step:

$$\begin{aligned} &\widehat{y}_{n+1} = \mathcal{G}(\widehat{m}_{n+1}), \\ &\widehat{C}^{\theta y}_{n+1} = \widehat{C}_{n+1} d\mathcal{G}(\widehat{m}_{n+1})^T, \\ &\widehat{C}^{yy}_{n+1} = d\mathcal{G}(\widehat{m}_{n+1})\widehat{C}_{n+1} d\mathcal{G}(\widehat{m}_{n+1})^T + \Sigma_\nu, \\ &m_{n+1} = \widehat{m}_{n+1} + \widehat{C}^{\theta y}_{n+1}(\widehat{C}^{yy}_{n+1})^{-1}\big(y - \widehat{y}_{n+1}\big), \\ &C_{n+1} = \widehat{C}_{n+1} - \widehat{C}^{\theta y}_{n+1}(\widehat{C}^{yy}_{n+1})^{-1}\widehat{C}^{\theta y}_{n+1}{}^T. \end{aligned} \tag{16}$$

This is a map of the form (12), but with a different definition of F, now depending on $d\mathcal{G}$ as well as $\mathcal{G}$.

2.3. Unscented Kalman inversion

Like the ExKI, the UKI also approximates the GAA; but it approximates the integrals appearing in Equations (10) by means of deterministic quadrature rules which are exact when evaluating means and covariances of variables defined as linear transformations of the random variable in question. Both the ExKI and the UKI recover the Kalman filter when $\mathcal{G}$ is linear. We need the definition of the unscented transform [34,35]:

Definition 1 *(Modified unscented transform).* Consider Gaussian random variable $\theta \sim \mathcal{N}(m, C) \in \mathbb{R}^{N_\theta}$. Define the $2N_\theta + 1$ symmetric sigma points $\{\theta_j\}_{j=0}^{2N_\theta+1}$ by

$$\begin{aligned} \theta^0 &= m, \\ \theta^j &= m + c_j[\sqrt{C}]_j \quad (1 \le j \le N_\theta), \\ \theta^{j+N_\theta} &= m - c_j[\sqrt{C}]_j \quad (1 \le j \le N_\theta), \end{aligned} \tag{17}$$

where $[\sqrt{C}]_j$ is the jth column of the Cholesky factor of C. Let $\mathcal{G}_i, i = 1, 2$ denote any pair of real vector-valued functions on $\mathbb{R}^{N_\theta}$. Then the quadrature rule approximating the mean and covariance of the transformed variables $\mathcal{G}_1(\theta)$ and $\mathcal{G}_2(\theta)$ is given by

$$\mathbb{E}[\mathcal{G}_i(\theta)] \approx \mathcal{G}_i(\theta^0) \qquad \mathrm{Cov}[\mathcal{G}_1(\theta), \mathcal{G}_2(\theta)] \approx \sum_{j=1}^{2N_\theta} W_j^c (\mathcal{G}_1(\theta^j) - \mathbb{E}\mathcal{G}_1(\theta))(\mathcal{G}_2(\theta^j) - \mathbb{E}\mathcal{G}_2(\theta))^T. \tag{18}$$

Here these constant weights are, for any $a \in \mathbb{R}$,

$$c_j = a\sqrt{N_\theta}\ (j = 1, \cdots, N_\theta) \quad W_j^c = \frac{1}{2a^2 N_\theta}\ (j = 1, \cdots, 2N_\theta).$$

Lemma 1. *Let $\mathcal{G}_i, i = 1, 2$ denote any pair of real vector-valued functions on $\mathbb{R}^{N_\theta}$. If $\theta \sim \mathcal{N}(m, C)$ then*

$$\begin{aligned}
&\mathbb{E}[\mathcal{G}_i(\theta)] = \mathcal{G}_i(m) + \mathcal{O}(\|C\|), \\
&\mathrm{Cov}[\mathcal{G}_1(\theta), \mathcal{G}_2(\theta)] = \sum_{j=1}^{2N_\theta} W_j^c(\mathcal{G}_1(\theta^j) - \mathbb{E}\mathcal{G}_1(\theta))(\mathcal{G}_2(\theta^j) - \mathbb{E}\mathcal{G}_2(\theta))^T + \mathcal{O}(\|C\|^2);
\end{aligned}$$

thus the modified unscented transform is first and second order accurate in approximating means and covariances of $\mathcal{G}_1(\theta)$ and $\mathcal{G}_2(\theta)$ with respect to small $\|C\|$. Furthermore, if $\mathcal{G}_1$ and $\mathcal{G}_2$ are linear then the modified unscented transform is exact for these quantities.

Proof. The proof is in Appendix A. □

Remark 1. The first and second order high order error terms, appearing in the expressions for the mean and covariance respectively, depend on derivatives of $\mathcal{G}_i$ at m and hence, through these derivatives and through C, on the parameter dimension N_θ. The original unscented transform leads to second order accuracy in the mean as well as covariance [85]. The modification we employ here replaces the original second order approximation of the $\mathbb{E}[\mathcal{G}_i(\theta)]$ with its first order counterpart. We do this to avoid negative weights; it also has ramifications for the optimization process which we discuss in Remark 9. In this paper, the hyper-parameter is chosen to be $a = \min\{\sqrt{\dfrac{4}{N_\theta}}, 1\}$. We note that the papers [85,35,48], suggest using a small positive value of a. We find in the numerical examples considered in this paper that our proposed choice of a outperforms the choice $a = \min\{\sqrt{\frac{4}{N_\theta}}, 0.01\})$, which builds in the idea of using a small positive value of a.

Consider the algorithm defined by equations (9) to (11). By utilizing the aforementioned quadrature rule, we obtain the following UKI algorithm:

- Prediction step:

$$\begin{aligned}
\widehat{m}_{n+1} =& \alpha m_n + (1-\alpha) r_0, \\
\widehat{C}_{n+1} =& \alpha^2 C_n + \Sigma_\omega.
\end{aligned} \tag{19}$$

- Generate sigma points:

$$\begin{aligned}
&\widehat{\theta}_{n+1}^0 = \widehat{m}_{n+1}, \\
&\widehat{\theta}_{n+1}^j = \widehat{m}_{n+1} + c_j[\sqrt{\widehat{C}_{n+1}}]_j \quad (1 \le j \le N_\theta), \\
&\widehat{\theta}_{n+1}^{j+N_\theta} = \widehat{m}_{n+1} - c_j[\sqrt{\widehat{C}_{n+1}}]_j \quad (1 \le j \le N_\theta).
\end{aligned} \tag{20}$$

- Analysis step:

$$\begin{aligned}
&\widehat{y}_{n+1}^j = \mathcal{G}(\widehat{\theta}_{n+1}^j) \qquad \widehat{y}_{n+1} = \widehat{y}_{n+1}^0, \\
&\widehat{C}_{n+1}^{\theta y} = \sum_{j=1}^{2N_\theta} W_j^c(\widehat{\theta}_{n+1}^j - \widehat{m}_{n+1})(\widehat{y}_{n+1}^j - \widehat{y}_{n+1})^T, \\
&\widehat{C}_{n+1}^{yy} = \sum_{j=1}^{2N_\theta} W_j^c(\widehat{y}_{n+1}^j - \widehat{y}_{n+1})(\widehat{y}_{n+1}^j - \widehat{y}_{n+1})^T + \Sigma_\nu, \\
&m_{n+1} = \widehat{m}_{n+1} + \widehat{C}_{n+1}^{\theta y}(\widehat{C}_{n+1}^{yy})^{-1}(y - \widehat{y}_{n+1}), \\
&C_{n+1} = \widehat{C}_{n+1} - \widehat{C}_{n+1}^{\theta y}(\widehat{C}_{n+1}^{yy})^{-1}\widehat{C}_{n+1}^{\theta y}{}^T.
\end{aligned} \tag{21}$$

This is again a map of the form (12), but with a different definition of F; unlike the ExKF there is no dependence on $d\mathcal{G}$, only on $\mathcal{G}$.

2.4. Ensemble Kalman inversion

This method differs fundamentally from the ExKI and UKI in that it does not map the mean and covariance. Rather it works with a set of particles whose dynamics at each step is predicted using (2a) and then used to compute empirical approximations of covariances. These in turn are used in the analysis step. The entire algorithm maps the collection $\{\theta_n^j\}_{j=1}^J$ into $\{\theta_{n+1}^j\}_{j=1}^J$. However, in the large J limit the mean and covariance updates match those of the GAA.

Consider the algorithm defined by equations (9) to (11). The EKI approach to making this implementable is to work with an ensemble of parameter estimates and approximate the covariances $\widehat{C}_{n+1}^{\theta y}$ and $\widehat{C}_{n+1}^{yy}$ empirically:

- Prediction step:

$$\begin{aligned}
&\widehat{\theta}_{n+1}^j = \alpha\theta_n^j + (1-\alpha)r_0 + \omega_{n+1}^j,\\
&\widehat{m}_{n+1} = \frac{1}{J}\sum_{j=1}^J \widehat{\theta}_{n+1}^j.
\end{aligned} \tag{22}$$

- Analysis step:

$$\begin{aligned}
&\widehat{y}_{n+1}^j = \mathcal{G}(\widehat{\theta}_{n+1}^j) \qquad \widehat{y}_{n+1} = \frac{1}{J}\sum_{j=1}^J \widehat{y}_{n+1}^j,\\
&\widehat{C}_{n+1}^{\theta y} = \frac{1}{J-1}\sum_{j=1}^J (\widehat{\theta}_{n+1}^j - \widehat{m}_{n+1})(\widehat{y}_{n+1}^j - \widehat{y}_{n+1})^T,\\
&\widehat{C}_{n+1}^{yy} = \frac{1}{J-1}\sum_{j=1}^J (\widehat{y}_{n+1}^j - \widehat{y}_{n+1})(\widehat{y}_{n+1}^j - \widehat{y}_{n+1})^T + \Sigma_\nu,\\
&\theta_{n+1}^j = \widehat{\theta}_{n+1}^j + \widehat{C}_{n+1}^{\theta y}\left(\widehat{C}_{n+1}^{yy}\right)^{-1}(y - \widehat{y}_{n+1}^j - \nu_{n+1}^j),\\
&m_{n+1} = \frac{1}{J}\sum_{j=1}^J \theta_{n+1}^j.
\end{aligned} \tag{23}$$

Here the superscript $j = 1, \cdots, J$ is the ensemble particle index, $\omega_{n+1}^j \sim \mathcal{N}(0, \Sigma_\omega)$ and $\nu_{n+1}^j \sim \mathcal{N}(0, \Sigma_\nu)$ are independent and identically distributed random variables with respect to both j and n.

Remark 2. In [1], where the iterative EKI was introduced, a slightly different stochastic dynamical formulation is used, extending the parameter space to include the image of the parameters under $\mathcal{G}$ and then making a linear observation operator on the extended space. The resulting method reduces to our setting with $\alpha = 1$, $\Sigma_\omega = 0$, and $\Sigma_\nu = \Sigma_\eta$ in the preceding algorithm. In the next section we will demonstrate theoretically that choosing $\alpha \in (0, 1)$ and $\Sigma_\omega \succ 0$ is beneficial and hence that the version of EKI proposed in this paper is superior to that in [1].

3. Theoretical insights

Recall that we view the GAA as an underlying conceptual algorithm which gives insight into the ExKI, UKI, and EKI algorithms. The ExKI is itself an approximation of the GAA, found by linearizing $\mathcal{G}$ around the predictive mean and the UKI and EKI algorithms are approximations of the resulting ExKI. Thus study of the GAA and ExKI gives insights into the UKI and EKI algorithms. This section is devoted to such studies. In Subsection 3.1 we consider behaviour of the GAA in the linear setting. In Subsection 3.2, we show that the ExKI may be viewed as a generalization of the LMA for optimization. Subsection 3.3 exhibits an averaging property induced by the unscented approximation, indicating how this may help in solving problems with rough energy landscapes. And in Subsection 3.4 we study a continuous-time limit of the GAA, which may itself be approximated to obtain continuous-time limits of the ExKI, UKI, and EKI algorithms; this provides insight into the discrete algorithms as implemented in practice.

3.1. The linear setting

In the linear setting the stochastic dynamical system for state $\{\theta_n\}$ and observations $\{y_n\}$ is given by

$$\text{evolution:} \qquad \theta_{n+1} = \alpha\theta_n + (1-\alpha)r_0 + \omega_{n+1}, \qquad \omega_{n+1} \sim \mathcal{N}(0, \Sigma_\omega), \tag{24a}$$

$$\text{observation:} \qquad y_{n+1} = G\theta_{n+1} + \nu_{n+1}, \qquad \nu_{n+1} \sim \mathcal{N}(0, \Sigma_\nu). \tag{24b}$$

Thanks to the linearity, equations (10) reduce to

$$\widehat{y}_{n+1} = Gm_n, \quad \widehat{C}^{\theta y}_{n+1} = \widehat{C}_{n+1}G^T, \quad \text{and} \quad \widehat{C}^{yy}_{n+1} = G\widehat{C}_{n+1}G^T + \Sigma_\nu. \tag{25}$$

The update equations (11) become

$$\begin{aligned} \widehat{m}_{n+1} &= \alpha m_n + (1-\alpha)r_0, \\ \widehat{C}_{n+1} &= \alpha^2 C_n + \Sigma_\omega, \end{aligned} \tag{26}$$

and

$$m_{n+1} = \widehat{m}_{n+1} + \widehat{C}_{n+1}G^T(G\widehat{C}_{n+1}G^T + \Sigma_\nu)^{-1}\Big(y - G\widehat{m}_{n+1}\Big), \tag{27a}$$

$$C_{n+1} = \widehat{C}_{n+1} - \widehat{C}_{n+1}G^T(G\widehat{C}_{n+1}G^T + \Sigma_\nu)^{-1}G\widehat{C}_{n+1}. \tag{27b}$$

We have the following theorem about the convergence of the GAA in the setting of the linear forward model:

Theorem 1. *Assume that $\Sigma_\omega \succ 0$ and $\Sigma_\nu \succ 0$. Consider the iteration (26), (27) mapping (m_n, C_n) into (m_{n+1}, C_{n+1}). Assume further that $\alpha \in (0,1)$ or that $\alpha = 1$ and Range$(G^T) = \mathbb{R}^{N_\theta}$. Then the steady state equation of equation (27b)*

$$C_\infty^{-1} = G^T\Sigma_\nu^{-1}G + (\alpha^2 C_\infty + \Sigma_\omega)^{-1} \tag{28}$$

has a unique solution $C_\infty \succ 0$. The pair (m_n, C_n) converges exponentially fast to limit (m_∞, C_∞). Furthermore the limiting mean m_∞ is the minimizer of the Tikhonov regularized least squares functional Φ_R given by

$$\Phi_R(\theta) := \frac{1}{2}\|\Sigma_\nu^{-\frac{1}{2}}(y - G\theta)\|^2 + \frac{1-\alpha}{2}\|\widehat{C}_\infty^{-\frac{1}{2}}(\theta - r_0)\|^2, \tag{29}$$

where

$$\widehat{C}_\infty = \alpha^2 C_\infty + \Sigma_\omega. \tag{30}$$

Proof. The proof is in Appendix A. □

Remark 3. When $\alpha \in (0,1)$, the exponential convergence rates of the mean and covariance are independent of the condition number of $G^T\Sigma_\nu^{-1}G$. Furthermore, $\widehat{C}_\infty$ is bounded above and below:

$$\Sigma_\omega \preceq \widehat{C}_\infty \preceq \frac{\Sigma_\omega}{1-\alpha^2},$$

since $0 \preceq C_\infty \preceq \alpha^2 C_\infty + \Sigma_\omega$.

Remark 4. Despite the clear parallels between equation (29) and Tikhonov regularization [13], there is an important difference: the matrix $\widehat{C}_\infty$ defining the implied prior covariance in the regularization term depends on the forward model. This may be seen by noting that it is defined by (30) in terms of the steady state covariance C_∞ satisfying (28). To get some insight into the implications of this, we consider the over-determined linear system in which $G^T\Sigma_\eta^{-1}G$ is invertible and we may define

$$C_* = (G^T\Sigma_\eta^{-1}G)^{-1}. \tag{31}$$

If we choose the artificial evolution and observation error covariances

$$\Sigma_\nu = 2\Sigma_\eta, \tag{32a}$$

$$\Sigma_\omega = \big(2 - \alpha^2\big)C_*, \tag{32b}$$

then straightforward calculation with (28), (30) shows that

$$C_\infty = C_*, \quad \widehat{C}_\infty = 2C_*.$$

From (29) it follows that

$$\Phi_R(\theta) = \frac{1}{4}\left\|\Sigma_\eta^{-\frac{1}{2}}(y - G\theta)\right\|^2 + \frac{(1-\alpha)}{4}\left\|\Sigma_\eta^{-\frac{1}{2}}(Gr_0 - G\theta)\right\|^2. \tag{33}$$

This calculation clearly demonstrates the dependence of the second (regularization) term on the forward model and that choosing $\alpha \in (0,1]$ allows different weights on the regularization term. In contrast to Tikhonov regularization, the regularization term (33) scales similarly with respect to G as does the data misfit, providing a regularization between the prior mean r_0 and an overfitted parameter $\theta^* : y = G\theta^*$. Therefore, despite the differences from standard Tikhonov regularization, the implied regularization resulting from the proposed stochastic dynamical system is both interpretable and controllable; in particular, the single parameter α measures the balance between prior and the overfitted solution.

Remark 5. Theorem 1 holds for any Kalman inversions that fulfill equations (9) and (25) exactly, which include ExKI, UKI, and these square root Kalman inversions [86,9], but not the EKI.

We contrast Theorem 1 with the behaviour of the filtering distribution for the stochastic dynamical system used in the derivation of the standard form of the EKI [1], which corresponds to the choices $\alpha = 1$, $\Sigma_\omega = 0$, and $\Sigma_\nu = \Sigma_\eta$. To study this case we will assume that $C_0 \succ 0$ and define

$$C_n' = C_0^{-\frac{1}{2}} C_n C_0^{-\frac{1}{2}}, \quad m_n' = C_0^{-\frac{1}{2}} m_n, \quad G' = G C_0^{\frac{1}{2}}, \quad S = (G')^T \Sigma_\nu^{-1} G'. \tag{34a}$$

We note that the nullspace of S is equal to the nullspace of G', and that the nullspace of G' is found from the nullspace of G by application of $C_0^{-\frac{1}{2}}$. Let Q denote orthogonal projection onto the nullspace of S, and P the orthogonal complement of Q. We then have the following characterization of the filtering distribution for the stochastic dynamical system underlying the form of the EKI introduced in [1].

Theorem 2. *Assume that $\alpha = 1$, $\Sigma_\omega = 0$ and consider the iteration (26), (27) mapping (m_n, C_n) into (m_{n+1}, C_{n+1}). Assume further that $\Sigma_\nu \succ 0$ and that $C_0 \succ 0$. Then $C_n \succ 0$ for all $n \in \mathbb{N}$ and*

$$(C_n')^{-1} = I + nS, \tag{35a}$$

$$(I + nS) m_n' = m_0' + n (G')^T \Sigma_\nu^{-1} y. \tag{35b}$$

Thus, as $n \to \infty$, with S^+ denoting the Moore-Penrose pseudo-inverse of S,

$$n^{-1} P (C_n')^{-1} = S + \mathcal{O}(n^{-1}), \qquad Q (C_n')^{-1} = Q, \tag{36a}$$

$$P m_n' = S^+ (G')^T \Sigma_\nu y + \mathcal{O}(n^{-1}), \qquad Q m_n' = Q m_0'. \tag{36b}$$

Proof. The proof is in Appendix A. □

Remark 6. Consider Theorem 2, in which $\alpha = 1$ and $\Sigma_\omega = 0$, and note that $S \succ 0$ in $P\mathbb{R}^{N_\theta}$. The theorem shows that the covariance of the filtering distribution of the stochastic dynamical system underlying the original implementation of EKI exhibits collapse to zero at algebraic rate in the observed subspace $P\mathbb{R}^{N_\theta}$, and is unchanged in the unobserved subspace $Q\mathbb{R}^{N_\theta}$. The mean converges algebraically slowly at rate $\mathcal{O}(n^{-1})$ in $P\mathbb{R}^{N_\theta}$ and is unchanged in $Q\mathbb{R}^{N_\theta}$.

Remark 7. Theorem 1-2 suggests the importance of choosing $\alpha \in (0, 1)$ and $\Sigma_\omega \succ 0$ in the stochastic dynamical systems that we propose here, as this ensures exponential convergence of the filtering distribution to a regularized least squares problem. However, if the forward operator has empty null-space, the situation arising when the inversion problem is well-determined or over-determined, then $\alpha = 1$ may be chosen but it is again important to ensure $\Sigma_\omega \succ 0$ to avoid the algebraic convergence exhibited in Theorem 2. In the case $\alpha \in (0, 1)$ Theorem 1 demonstrates the regularization which underlies the proposed iterative method. In the case $\alpha = 1$, the regularization term vanishes.

Remark 8. The behaviour of the finite particle size EKI, in the case $\alpha = 1$, $\Sigma_\omega = 0$, is fully analyzed in [5]. Theorem 2 is a mean-field counterpart of that theory.

The following proposition is relevant to understanding some of the numerical experiments presented later in the paper and, taken together with Theorems 1 and 2, it also completes our analysis of the filtering distribution for the novel stochastic dynamical system introduced in this paper.

Proposition 2. *Assume that $\alpha = 1$ and $\Sigma_\omega \succ 0$ and consider the setting where the forward operator G has non-trivial null space (thus violating the assumption Range$(G^T) = \mathbb{R}^{N_\theta}$ in Theorem 1). Assume further that $\Sigma_\nu \succ 0$ and that $C_0 \succ 0$. Then $C_n \succ 0$ for all $n \in \mathbb{N}$ and m_n converges to a minimizer of $\frac{1}{2}\|\Sigma_\nu^{-\frac{1}{2}}(y - G\theta)\|^2$ exponentially fast. However C_n^{-1} converges to a singular matrix and hence $\|C_n\|$ diverges to $+\infty$; the rate of divergence is bounded by*

$$C_n \preceq C_0 + n\Sigma_\omega. \tag{37}$$

Proof. The proof is in Appendix A. □

3.2. ExKI: Levenberg–Marquardt connection

In the nonlinear setting, our numerical results will demonstrate the implicit regularization and linear (sometimes superlinear) convergence of ExKI and UKI. This desirable feature can be understood by the analogy with the Levenberg–Marquardt Algorithm (LMA). We focus this discussion on the particular case $\alpha = 1$ as we find that, for over-determined problems, this choice often produces the best results.

Consider the non-regularized nonlinear least-squares objective function Φ, defined in (3b). The key step in the Levenberg–Marquardt Algorithm (LMA) is to solve the minimization problem for (3b) by a preconditioned gradient descent procedure which maps θ_n to $\theta_n + \delta\theta_n$ and where $\delta\theta_n$ solves

$$(d\mathcal{G}(\theta_n)^T \Sigma_\nu^{-1} d\mathcal{G}(\theta_n) + \lambda_n \mathbb{I})\delta\theta_n = d\mathcal{G}(\theta_n)^T \Sigma_\nu^{-1}(y - \mathcal{G}(\theta_n)). \tag{38}$$

Here $\mathbb{I}$ is the identity matrix on $\mathbb{R}^{N_\theta}$ and λ_n is the (non-negative) damping factor, often chosen adaptively. Because of the damping matrix $\lambda_n \mathbb{I}$, the LMA is found to be more robust than the Gauss–Newton Algorithm and exhibits linear (or even superlinear) convergence in practice. The use of LMA for inverse problems is discussed in [29].

The ExKI procedure solves the optimization problem for (3b) by a different preconditioned gradient descent procedure, defined by the update

$$\Big(d\mathcal{G}(\theta_n)^T \Sigma_\nu^{-1} d\mathcal{G}(\theta_n) + (C_n + \Sigma_\omega)^{-1}\Big)\delta\theta_n = d\mathcal{G}(\theta_n)^T \Sigma_\nu^{-1}\left(y - \mathcal{G}(\theta_n)\right). \tag{39}$$

This may be viewed as a generalization of the LMA in which the adaptive damping term is now a matrix $C_n + \Sigma_\omega$ and the adaptation is automated through the covariance updates; furthermore this matrix is lower bounded (in the sense of quadratic forms) by Σ_ω, regardless of the adaptation through the covariance, ensuring some damping of the Gauss-Newton approximate Hessian. We may expect that the UKI and EKI, which approximate the linearization $d\mathcal{G}$ in the ExKI, to benefit from this generalized LMA. Connections between the LMA and EKI were first systematically explored in [2] and more recently in [76].

3.3. UKI: unscented approximation and averaging

Here we explain that the unscented transform may be viewed as smoothing the energy landscape of UKI, in comparison with ExKI; this helps to explain the improved behaviour of UKI over ExKI on rough landscapes, such as those we will show in section 5 when performing parameter estimation for chaotic differential equations. To understand this smoothing effect we first introduce a useful averaging property [87, Theorem 1].[4]

Lemma 2. *Let θ denote Gaussian random vector $\theta \sim \mathcal{N}(m, C) \in \mathbb{R}^{N_\theta}$. For any nonlinear function $\mathcal{G} : \mathbb{R}^{N_\theta} \to \mathbb{R}^{N_y}$, we define the associated averaged function $\mathcal{F}\mathcal{G} : \mathbb{R}^{N_\theta} \times \mathbb{R}^{N_\theta \times N_\theta}_{\succeq 0} \to \mathbb{R}^{N_y}$ and averaged gradient function $\mathcal{F}d\mathcal{G} : \mathbb{R}^{N_\theta} \times \mathbb{R}^{N_\theta \times N_\theta}_{\succeq 0} \to \mathbb{R}^{N_y \times N_\theta}$ as follows:*

$$\mathcal{F}\mathcal{G}(m, C) := \mathbb{E}[\mathcal{G}(\theta)] \qquad \mathcal{F}d\mathcal{G}(m, C) := \mathrm{Cov}[\mathcal{G}(\theta), \theta] \cdot C^{-1}. \tag{40}$$

Then we have $\dfrac{\partial \mathcal{F}\mathcal{G}(m, C)}{\partial m} = \mathcal{F}d\mathcal{G}(m, C).$

Proof. The proof is in Appendix A. □

Note that in the linear case $\mathcal{F}\mathcal{G}(m, C) = \mathcal{G}(m)$ and $\mathcal{F}d\mathcal{G}(m, C) = \mathcal{G}$; the averaged derivative is exact. This averaging procedure is useful to understand the conceptual GAA precisely because (40) may be used to express $\mathrm{Cov}[\mathcal{G}(\theta), \theta]$, which appears in the conceptual GAA, in terms of the averaged derivative $\mathcal{F}d\mathcal{G}(m, C)$. In order to use this idea in the context of the UKI it is useful to understand related averaging operations when the modified unscented transform (Definition 1) is employed to approximate Gaussian expectations. To this end we define, using (19)–(21),

$$\begin{aligned} \mathcal{F}_u\mathcal{G}_n &:= \widehat{y}_n, \\ \mathcal{F}_u d\mathcal{G}_n &:= \widehat{C}_n^{\theta y\, T} \widehat{C}_n^{-1}, \end{aligned} \tag{41}$$

noting that $\mathcal{F}_u\mathcal{G}_n$ and $\mathcal{F}_u d\mathcal{G}_n$ then correspond to approximation of (40) at step n of the algorithm, using the modified unscented transform from Definition 1.

Proposition 3. *The UKI algorithm (19)–(21) may be written in the following form:*

[4] In what follows, the suffix $_{\succeq 0}$ denotes positive semi-definite matrix and $\frac{\partial}{\partial m}$ denotes gradient with respect to m.

- *Prediction step:*

$$\begin{aligned}\widehat{m}_{n+1} =&\alpha m_n + (1-\alpha)r_0,\\ \widehat{C}_{n+1} =&\alpha^2 C_n + \Sigma_\omega.\end{aligned} \tag{42}$$

- *Analysis step:*

$$\begin{aligned}&\widehat{y}_{n+1} = \mathcal{F}_u\mathcal{G}_{n+1},\\ &\widehat{C}^{\theta y}_{n+1} = \widehat{C}_{n+1}\mathcal{F}_u d\mathcal{G}_{n+1}^T,\\ &\widehat{C}^{yy}_{n+1} = \mathcal{F}_u d\mathcal{G}_{n+1}\widehat{C}_{n+1}\mathcal{F}_u d\mathcal{G}_{n+1}^T + \Sigma_\nu + \widetilde{\Sigma}_{\nu,n+1},\\ &m_{n+1} = \widehat{m}_{n+1} + \widehat{C}^{\theta y}_{n+1}(\widehat{C}^{yy}_{n+1})^{-1}\big(y - \widehat{y}_{n+1}\big),\\ &C_{n+1} = \widehat{C}_{n+1} - \widehat{C}^{\theta y}_{n+1}(\widehat{C}^{yy}_{n+1})^{-1}\widehat{C}^{\theta y}_{n+1}{}^T.\end{aligned} \tag{43}$$

Here $\widetilde{\Sigma}_{\nu,n+1} \succeq 0$. Furthermore, $\|\widetilde{\Sigma}_{\nu,n+1}\| = \mathcal{O}(\|\widehat{C}_{n+1}\|^2)$ and $\widetilde{\Sigma}_{\nu,n+1} = 0$ when $\mathcal{G}$ is linear.

Proof. The proof is in Appendix A. □

Remark 9. Comparison of the original UKI algorithm (19)–(21) with its rewritten form (42)-(43) demonstrates that, in the regime where the covariance is small, or the forward model is linear, the UKI algorithm behaves like the ExKI algorithm (15)-(16) but with the nonlinear function $\mathcal{G}$ and its associated gradient $d\mathcal{G}$ having been averaged according to unscented approximations of the averaging operations defined in Lemma 2. From the preceding subsection, it follows that the UKI is also related to a modified LMA applied to an averaged objective function. Note that, by using the unscented approximation of the averaging procedure defined in Lemma 2, we essentially remove the averaging of $\mathcal{G}$ and retain it only on $d\mathcal{G}$. Averaging of the gradient $d\mathcal{G}$ alone will be demonstrated to have an important positive effect on parameter estimation for chaotic dynamical systems in Subsections 5.8, 5.9, and 5.10.

3.4. Continuous time limit

To derive a continuous-time limit we set $\alpha = 1 - \alpha_0 h$, $\Sigma_\omega \mapsto h\Sigma_\omega$, and $\Sigma_\nu \mapsto h^{-1}\Sigma_\nu$. The algorithm defined by Equations (9) to (11) then has the form of a first order accurate (in h) approximation of the dynamical system

$$\dot{m} = -\alpha_0(m - r_0) + C^{\theta y}\Sigma_\nu^{-1}\big(y - \mathbb{E}\mathcal{G}(\theta)\big), \tag{44a}$$

$$\dot{C} = -2\alpha_0 C + \Sigma_\omega - C^{\theta y}\Sigma_\nu{}^{-1}C^{\theta y}{}^T, \tag{44b}$$

where $\theta \sim \mathcal{N}(m, C)$, expectation $\mathbb{E}$ is with respect to this distribution and

$$C^{\theta y} = \mathbb{E}\Big(\big(\theta - m\big) \otimes \big(\mathcal{G}(\theta) - \mathbb{E}\mathcal{G}(\theta)\big)\Big).$$

This continuous-time dynamical system may be used as the basis for practical algorithms by discretizing in time, for example, using forward Euler with an adaptive time-step as in [65], and applying the same ideas used in the ExKF, UKI or EKI to approximate the expectations.

The steady state m_∞, C_∞ of the differential equations (44) are implicitly defined in a somewhat complicated fashion. However, any such steady state always has non-singular covariance as we now state and prove.

Lemma 3. *For any steady state (m_∞, C_∞) of equation (44), the steady covariance C_∞ is non-singular.*

Proof. The proof is in Appendix A. □

4. Variants on the basic algorithm

4.1. Enforcing constraints

Kalman inversion requires solving forward problems at every iteration. Failure of the forward problem to deliver physically meaningful solutions can lead to failure of the inverse problem. Adding constraints to the parameters (for example, dissipation is non-negative) significantly improves the robustness of Kalman inversion. Within the EKI there is a natural way to impose constraints, using the fact that each iteration of the algorithm may be interpreted as solving a set of coupled quadratic optimization problems, with coupling arising from empirical covariances. These optimization problems are readily appended with convex constraints, such as box (inequality) constraints [88]; see also [2,4]. The UKI does not have this optimization interpretation and so we adopt a different approach to enforcing box constraints.

In this paper there are occasions where we impose element-wise box constraints of the form

$$0 \le \theta \quad \text{or} \quad \theta_{min} \le \theta \le \theta_{max}.$$

These are enforced by change of variables writing $\theta = \varphi(\tilde{\theta})$ where, for example, respectively,

$$\varphi(\tilde{\theta}) = |\tilde{\theta}| \quad \text{or } \varphi(\tilde{\theta}) = \theta_{min} + \frac{\theta_{max} - \theta_{min}}{1 + |\tilde{\theta}|}.$$

The inverse problem is then reformulated as

$$y = \mathcal{G}(\varphi(\tilde{\theta})) + \eta,$$

and the UKI methods and variants are employed with $\mathcal{G} \mapsto \mathcal{G} \circ \varphi$.

4.2. Unscented Kalman sampler

Consider the following stochastic dynamical system, in which W is a standard unit Brownian motion in $\mathbb{R}^{N_\theta}$:

$$\dot{\theta} = C^{\theta y}\Sigma_\eta^{-1}\big(y - \mathcal{G}(\theta)\big) - C\Sigma_0^{-1}(\theta - r_0) + \sqrt{2}C^{\frac{1}{2}}\dot{W}, \tag{45a}$$

$$C^{\theta y} = \mathbb{E}\Big(\big(\theta - m\big) \otimes \big(\mathcal{G}(\theta) - \mathbb{E}\mathcal{G}(\theta)\big)\Big) \tag{45b}$$

and all expectations are computed under the law of θ, with respect to which the mean and covariance are denoted as m and C respectively. This Itò-McKean diffusion process can be approximated by an interacting particle system, and the law of θ approximated using the resulting empirical Gaussian approximation, leading to the EKS [6]; we now generalize this to an unscented version. First consider the following evolution equations for the mean and covariance of the Gaussian approximation to the law of θ:

$$\dot{m} = C^{\theta y}\Sigma_\eta^{-1}\big(y - \mathbb{E}\mathcal{G}(\theta)\big) - C\Sigma_0^{-1}(m - r_0), \tag{46a}$$

$$\dot{C} = -2C^{\theta y}\Sigma_\eta^{-1}C^{\theta y^T} - 2C\Sigma_0^{-1}C + 2C. \tag{46b}$$

Note that the expectations are computed under the law of (45) and so this is not, in general, a closed system for (m, C).

To obtain a closed system for (m, C), we consider a Gaussian evolving according to the equations (46), with matrix $C^{\theta y}$ again given by (45b), but now expectation $\mathbb{E}$ is computed with respect to the distribution $\mathcal{N}(m, C)$ so that a closed system for (m, C) is obtained. The UKS is defined by approximating the expectations in this system by use of an unscented transform.

In the case where $\mathcal{G}$ is linear and the solution is initialized at a Gaussian then the system (46) with expectations computed under $\mathcal{N}(m, C)$ is consistent with the solution of the Itò-McKean diffusion (45) governing θ – that latter has Gaussian distribution evolving according to (46). Furthermore, the analysis in [6] shows that then the system converges to the posterior distribution (4) at a rate $\exp(-t)$ independent of the problem being solved; this independence of the rate on the problem conditioning may be viewed as a consequence of affine invariance. We also mention that the analysis in [89] shows that, when initialized at a non-Gaussian, the Gaussian dynamics is an attractor. It is thus natural to consider using numerical simulations of (46) to generate approximate samples from the posterior distribution. Illustrative examples are presented in Appendix B.

5. Numerical results

In this section, we present numerical results for Kalman-based inversion using the proposed stochastic dynamical system equation (2).

5.1. Choice of hyperparameters

We make choices of Σ_ω and Σ_ν guided by the discussion in Remark 4. However, for general nonlinear problems C_* is not explicitly defined. Thus we modify the prescription given in (32) and instead choose

$$\Sigma_\nu = 2\Sigma_\eta \tag{47a}$$

$$\Sigma_\omega = \big(2 - \alpha^2\big)\gamma\mathbb{I} \tag{47b}$$

for some $\gamma > 0$. For over-determined problems, when the observational noise is absent or negligible, we take $\alpha = 1$. For under-determined problems, to avoid overfitting in the presence of noise, we generally choose $\alpha \in (0, 1)$; but we also present some under-determined problems with choice $\alpha = 1$ to demonstrate undesirable effects from doing so. In general, cross-validation should be invoked to determine an optimal choice of α. However in this paper, we have simply used the values $0.0, 0.5, 0.9, 1.0$ for illustrative purposes. To be concrete we initialize with $m_0 = r_0$ and $C_0 = \gamma\mathbb{I}$. Specific choices of r_0 and γ will differ between examples and will be spelled out in each case.

5.2. *Classes of problems studied*

For all applications, we focus mainly on the UKI; some comparisons between the UKI and EKI (specifically, as applied to the novel stochastic dynamical system (2) proposed here) are also presented; and computational difficulties inherent in the rough misfit landscape experienced by the ExKI for chaotic dynamical systems are illustrated, and are demonstrably overcome by deploying the UKI. The applications cover a wide range of problems. They include three categories:

1. Noiseless linear problems, where over-determined, under-determined, and well-determined systems are considered.
   - Linear 2-parameter model problem: this problem serves as a proof-of-concept example, which demonstrates the convergence of the mean and the covariance matrix discussed in Subsection 3.1. In this case, the UKI is exact, as a consequence of Lemma 1; numerics are performed using only the UKI.
   - Hilbert matrix problem: this problem illustrates the performance of the EKI and UKI when solving ill-conditioned inverse problems. The EKI suffers from divergence as it is iterated. However the UKI behaves well, again reflecting the exactness for linear problems, highlighted in Lemma 1, and the theory of Subsection 3.1 characterizing the behaviour of the filtering distribution in the linear setting.
2. Noisy field recovery problems, in which we add 0%, 1%, and 5% Gaussian random noise to the observation, as follows:

$$y_{obs} = y_{ref} + \epsilon \odot \xi, \quad \xi \sim \mathcal{N}(0, \mathbb{I}), \tag{48}$$

   where $y_{ref} = \mathcal{G}(\theta_{ref})$, $\epsilon = 0\% y_{ref}, 1\% y_{ref}$, and $5\% y_{ref}$, and $\odot$ denotes element-wise multiplication. It is important to distinguish between the added Gaussian random noise appearing in the data and the observation error model $\eta \sim \mathcal{N}(0, \Sigma_\eta)$ used in the development of the inversion algorithm; in essence we assume imperfect knowledge of the noise model.[5] Comparison of UKI and EKI is presented. EKI is shown to suffer from finite ensemble size effects, and in some cases diverges; in contrast, UKI behaves well. Thus we observe that what we have learned from the linear setting carries across to the setting of nonlinear inverse problems. This category of inversion for fields also serves to demonstrate the value of the Tikhonov regularization parameter $\alpha \in (0, 1)$ in the prevention of overfitting. We consider three examples, now listed.
   - Darcy flow problem: to find permeability parameters in subsurface flow from measurements of pressure (or piezometric head).
   - Damage detection problem: determining the damage field in an elastic body from displacement observations on the surface of the structure.
   - Navier-Stokes problem: we study a two dimensional incompressible fluid, using the vorticity-streamfunction formulation, and recover the initial vorticity from noisy observations of the vorticity field at later times.
3. Chaotic problems, in which the parameters are learned from time-averaged statistics. For these problems, which are over-determined, we demonstrate that choosing $\alpha = 1$ is satisfactory, relying on the implicit regularization inherent in the approximate LMA interpretation of ExKI and UKI, as discussed in Subsection 3.2. The three examples considered are now listed.
   - Lorenz63 model problem: we present a discussion of why adjoint based methods including ExKI, fail; we then demonstrate that the UKI succeeds. We attribute the success of the UKI to the averaging effect induced by the unscented transform and discussed in Subsection 3.3.
   - Multiscale Lorenz96 problem: we study a scale-separated setting, in which the closure for the fast dynamics is learned from time-averaged statistics.
   - Idealized general circulation model problem: this is a 3D Navier-Stokes problem with a hydrostatic assumption, and simple parameterized subgrid-scale models; we learn the parameters of the subgrid-scale model from time-averaged data. This problem demonstrates the potential of applying the UKI for large scale chaotic inverse problems.

5.3. *Linear 2-parameter model problem*

Consider the 2-parameter linear inverse problem to find $\theta \in \mathbb{R}^2$ from $y \in \mathbb{R}^{N_y}$ where $y = G\theta$ with $G \in \mathbb{R}^{N_y \times 2}$ and no noise is present in the data. We explore the following three scenarios corresponding to $N_y = 3, 2$ and 1:

- non-singular (well-determined) system (NS) $N_y = 2$

$$y = \begin{bmatrix} 3 \\ 7 \end{bmatrix} \quad G = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \quad \theta_{ref} = \begin{bmatrix} 1 \\ 1 \end{bmatrix};$$

[5] See section 7.1 of [90] for an example with a similar set-up; see also discussion around equation (55) in [91] where the additive Gaussian noise used in the data is carefully constructed to scale relative to the truth underlying it.

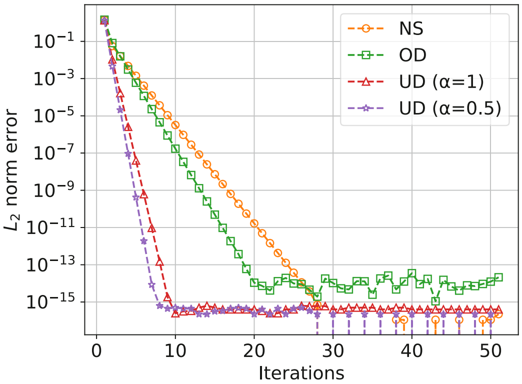


Fig. 1. L_2 error $\|m_n - \theta_{ref}\|_2$ of the linear 2-parameter model problem. NS: non-singular system, OD: over-determined system, UD: under-determined system.

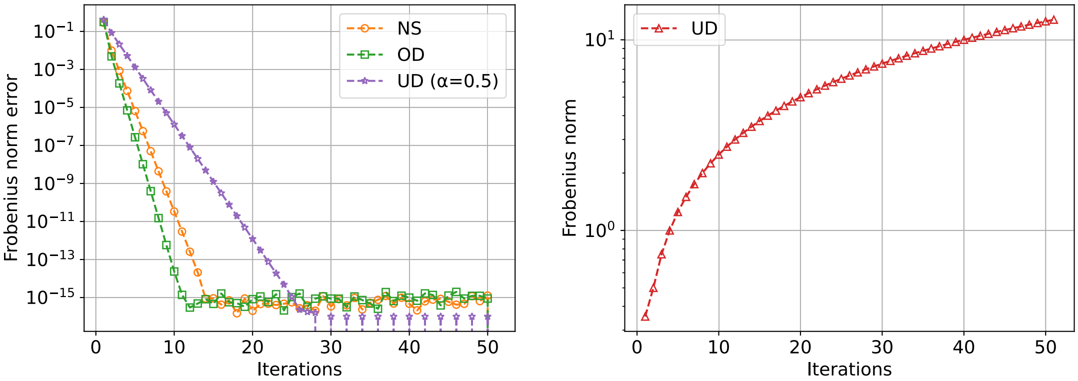


Fig. 2. Frobenius norm $\|C_n - C_\infty\|_F$ (left) for non-singular (NS) and over-determined (OD) systems, and $\|C_n\|_F$ (right) for the under-determined (UD) system of the linear 2-parameter model problem.

- over-determined system (OD) $N_y = 3$

$$y = \begin{bmatrix} 3 \\ 7 \\ 10 \end{bmatrix} \quad G = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix} \quad \theta_{ref} = \begin{bmatrix} 1/3 \\ 17/12 \end{bmatrix};$$

- under-determined system (UD) $N_y = 1$

$$y = \begin{bmatrix} 3 \end{bmatrix} \quad G = \begin{bmatrix} 1 & 2 \end{bmatrix} \quad \theta_{ref} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} + c \begin{bmatrix} 2 \\ -1 \end{bmatrix}, \quad c \in \mathbb{R}.$$

Since there is no noise in the data we have $\Sigma_\eta = 0$ and Φ is undefined. To proceed we apply our methodology as if $\Sigma_\eta = 0.1^2\mathbb{I}$, corresponding to a misspecified model. Then we may set

$$\theta_{ref} = \arg\min_\theta \Phi(\theta) = \arg\min_\theta \frac{1}{2}\|(y - G\theta)\|^2.$$

Note that for the OD and NS cases θ_{ref} is a single point, whereas in the UD case, θ_{ref} comprises a one-parameter ($c \in \mathbb{R}$) family of possible solutions.

We choose $r_0 = 0$, $\gamma = 0.5^2$ and also initialize the UKI at $\theta_0 \sim \mathcal{N}(0, \gamma\mathbb{I})$. In both the NS and OD cases $Range(G^T) = \mathbb{R}^{N_\theta}$ and so we set $\alpha = 1$, guided by Theorem 1. In the UD case $Range(G^T) \neq \mathbb{R}^{N_\theta}$ and we consider both $\alpha = 1$ and $\alpha = 0.5$, illustrating Proposition 2 and Theorem 1 respectively. The convergence of the parameter vectors $\{m_n\}$ is depicted in Fig. 1. In all scenarios, the mean vectors converge to a limiting value exponentially fast. In the cases of NS and OD this is as predicted by Theorem 1 and, since $\alpha = 1$, Φ_R and Φ coincide so that $m_\infty = \theta_{ref}$. For UD with $\alpha = 1$ and $\alpha = 0.5$, the mean vectors converge to $[0.6 \quad 1.2]^T$ and $[0.597 \quad 1.195]^T$ respectively, following Proposition 2 and Theorem 1. However, the limiting mean for $\alpha = 1$ depends on the initial conditions for the algorithm, whereas for $\alpha = 0.5$ it is uniquely determined. The convergence of the covariance matrices $\{C_n\}$ to C_∞ is depicted in Fig. 2, with NS, OD, and UD ($\alpha = 0.5$) on the left

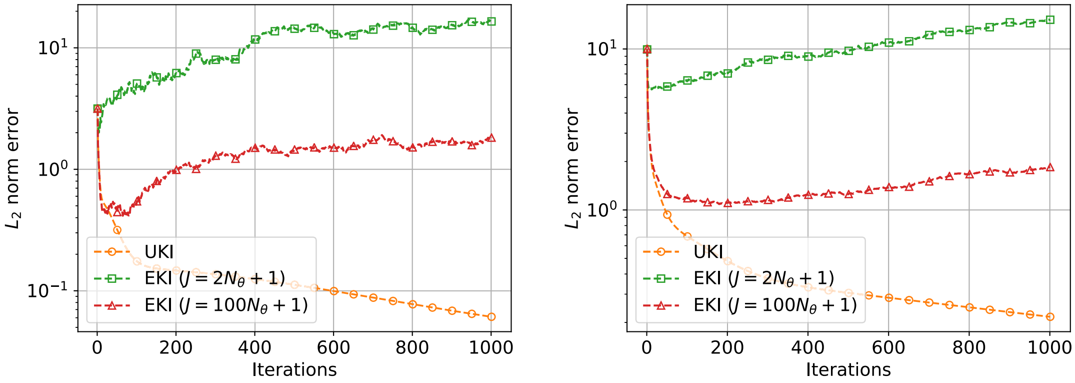


Fig. 3. L_2 error $\|m_n - \theta_{ref}\|_2$ of the Hilbert inverse problem with $N_\theta = 10$ (left) and $N_\theta = 100$ (right).

and UD ($\alpha = 1.0$) on the right. In the cases NS, OD, and UD ($\alpha = 0.5$), the estimated covariance matrices converge to the desired values (the steady state of equation (28), as predicted by Theorem 1). In the case UD, the covariance matrices $\{C_n\}$ diverge to $+\infty$ (see Proposition 2); nonetheless, this divergence of the covariance matrix does not affect the exponential convergence of the mean vector. In general, we advocate the use of $\alpha \in (0, 1)$ for under-determined problems and have set $\alpha = 1$ for problem UD here only to illustrate some of the issues that arise from doing so.

5.4. Hilbert matrix problem

In this example $N_y = N_\theta$. We define the Hilbert matrix $G \in R^{N_\theta \times N_\theta}$ by its entries

$$G_{i,j} = \frac{1}{i + j - 1}.$$

The condition number of G grows as $\mathcal{O}\Big((1+\sqrt{2})^{4N_\theta}/\sqrt{N_\theta}\Big)$ [92]. We consider the inverse problem of finding $\theta \in \mathbb{R}^{N_\theta}$ from $y \in \mathbb{R}^{N_y}$ where $y = G\theta_{ref}$ and we define $\theta_{ref} := \mathbb{1}$. The ill-conditioning of G makes the determination of θ from y difficult. Traditional linear solvers fail for such a problem.[6]

We consider two scenarios: $N_\theta = 10$ and $N_\theta = 100$. As in the previous linear case study we assume a model misspecification setting in which $\Sigma_\eta = 0.1^2\mathbb{I}$, even though the data itself contains no noise, and we take $\alpha = 1$. We set $r_0 = 0$ and $\gamma = 0.5^2$. Thus $\theta_0 \sim \mathcal{N}(0, 0.5^2\mathbb{I})$. Both UKI and EKI are applied. For the EKI, the ensemble sizes are set to $J = 2N_\theta + 1$ and $J = 100N_\theta + 1$. The convergence of the parameter vector m_n is depicted in Fig. 3. The UKI converges, but the convergence rate depends on the condition number of G, slowing as it grows. The EKI converges to a certain accuracy as fast as the UKI and then diverges. This divergence is related to the finite ensemble size, and is delayed by use of the larger J. Indeed in the mean-field limit $J = \infty$ the EKI will coincide with the UKI. This example clearly demonstrates the benefits of the UKI over the EKI.

5.5. Darcy flow problem

Consider the Darcy flow equation on the two-dimensional spatial domain $D = [0, 1]^2$. The forward model is to find the pressure field $p(x)$ in a porous medium defined by a positive permeability field $a(x, \theta)$:

$$\begin{aligned} -\nabla \cdot (a(x,\theta)\nabla p(x)) &= f(x), \qquad && x \in D, \\ p(x) &= 0, && x \in \partial D. \end{aligned}$$

For simplicity, we have imposed homogeneous Dirichlet boundary conditions on the pressure at the boundary ∂D. The fluid source field f is defined as

$$f(x_1, x_2) = \begin{cases} 1000 & 0 \leq x_2 \leq \frac{4}{6} \\ 2000 & \frac{4}{6} < x_2 \leq \frac{5}{6} \\ 3000 & \frac{5}{6} < x_2 \leq 1 \end{cases}.$$

[6] $G\backslash y$ in Julia leads to an L_2 error of 4250.142 for $N_\theta = 100$.

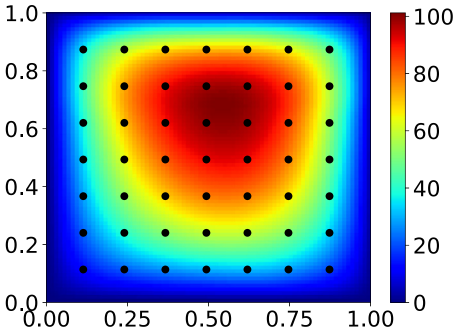


Fig. 4. The pressure field of the Darcy flow problem and the 49 equidistant pointwise measurements (black dots). (For interpretation of the colours in the figure(s), the reader is referred to the web version of this article.)

We study the inverse problems of finding a from noisy measurements of p. We place a prior on the permeability field $a(x,\theta)$ by assuming that $\log a(x,\theta)$ is a centred Gaussian with covariance

$$\mathsf{C} = (-\Delta + \tau^2)^{-d};$$

here $-\Delta$ denotes the Laplacian on D subject to homogeneous Neumann boundary conditions on the space of spatial-mean zero functions, $\tau > 0$ denotes the inverse length scale of the random field and $d > 0$ determines its regularity ($\tau = 3$ and $d = 2$ in the present study). See [4,93,6,94] for examples. The parameter θ represents the countable set of coefficients in the Karhunen-Loève (KL) expansion of the Gaussian random field:

$$\log a(x,\theta) = \sum_{l \in K} \theta_{(l)} \sqrt{\lambda_l}\, \psi_l(x), \tag{49}$$

where $K = \mathbb{Z}^{0+} \times \mathbb{Z}^{0+} \setminus \{0,0\}$, $\theta_{(l)} \sim \mathcal{N}(0,1)$ i.i.d. and the eigenpairs are of the form

$$\psi_l(x) = \begin{cases} \sqrt{2}\cos(\pi l_1 x_1) & l_2 = 0 \\ \sqrt{2}\cos(\pi l_2 x_2) & l_1 = 0 \\ 2\cos(\pi l_1 x_1)\cos(\pi l_2 x_2) & \text{otherwise} \end{cases}, \qquad \lambda_l = (\pi^2 |l|^2 + \tau^2)^{-d}.$$

The KL expansion equation (49) can be rewritten as a sum over $\mathbb{Z}^{0+}$ rather than a lattice:

$$\log a(x,\theta) = \sum_{k \in \mathbb{Z}^{0+}} \theta_{(k)} \sqrt{\lambda_k}\, \psi_k(x), \tag{50}$$

where the eigenvalues λ_k are in descending order. In practice, we truncate this sum to N_θ terms, based on the largest N_θ eigenvalues, and hence $\theta \in \mathbb{R}^{N_\theta}$. The forward problem is solved by a finite difference method on an 80×80 grid.

For the inverse problem, the observation y_{ref} consists of pointwise measurements of the pressure value $p(x)$ at 49 equidistant points in the domain (see Fig. 4). We generate a truth random field $\log a_{ref}(x)$ with $\theta \sim \mathcal{N}(0,\mathbb{I})$ in $\mathbb{R}^{256}$ (i.e. we use the first 256 KL modes) to construct the observation y_{ref}; different levels of noise are added to make data y_{obs} as explained in (48). Using this data, we consider two incomplete parameterization scenarios: solving for the first 32 KL modes ($N_\theta = 32$) and for the first 8 KL modes ($N_\theta = 8$). EKI and UKI are both applied. We take $r_0 = 0$ and $\gamma = 1$ so that $\theta_0 \sim \mathcal{N}(0,\mathbb{I})$. The observation error satisfies $\eta \sim \mathcal{N}(0,\mathbb{I})$. For the EKI, the ensemble size is set to be $J = 100$, which is larger than the number of σ-points used in UKI $(2N_\theta + 1)$.

For the $N_\theta = 32$ case, the convergence of the log-permeability fields $\log a(x, m_n)$ and the optimization errors (3) at each iteration for different noise levels are depicted in Fig. 5; the top row shows the relative L_2 errors in the estimate of $\log a$ and the bottom row shows the optimization errors (data-misfit), left to right corresponds to different noise levels in the data. Without explicit regularization ($\alpha = 1.0$), both UKI and EKI suffer from overfitting for noisy scenarios: the optimization errors keep decreasing, but the parameter errors show the "U-shape" characteristic of overfitting. Adding regularization ($\alpha = 0.5$) relieves the overfitting. The estimated log-permeability fields $\log a(x, m_n)$ at the 50th iteration and the truth random field are depicted in Fig. 6. Both UKI and EKI deliver similar results and these estimated log-permeability fields capture main features of the truth random field.

For the $N_\theta = 8$ case, the convergence of the log-permeability fields $\log a(x, m_n)$ and the optimization errors at each iteration for different noise levels are depicted in Fig. 7. Even without explicit regularization ($\alpha = 1.0$), none of these Kalman inversions suffer from overfitting. Both UKI and EKI lead to similar parameter errors and optimization errors. The

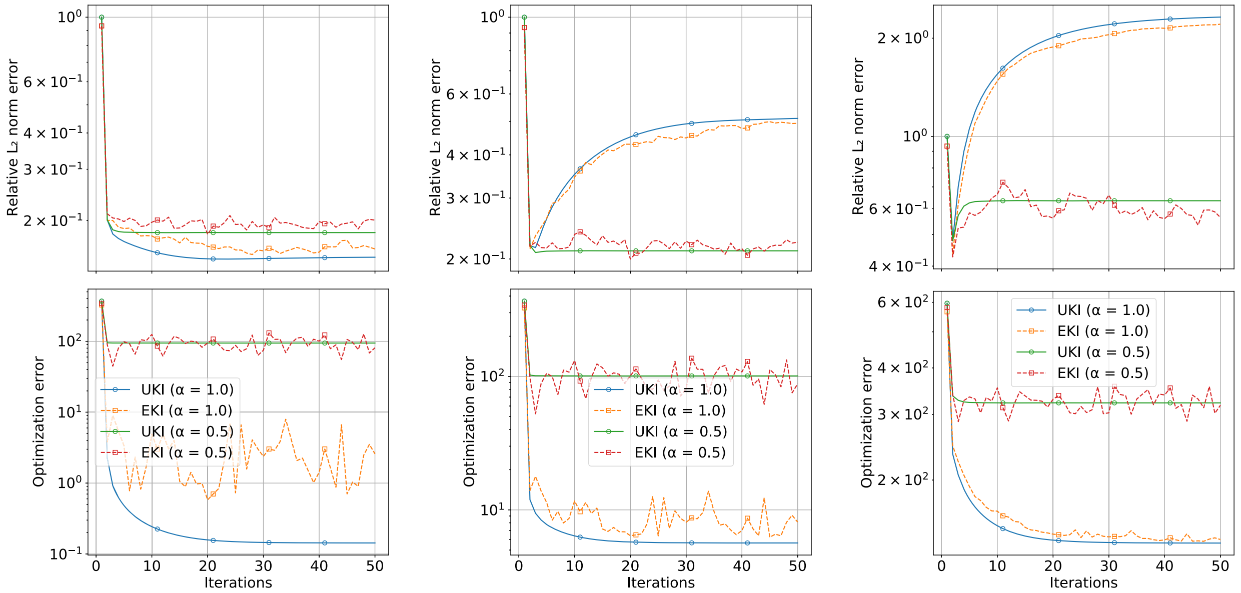


Fig. 5. Relative error $\frac{\|\log a(x, m_n) - \log a_{ref}(x)\|_2}{\|\log a_{ref}(x)\|_2}$ (top) and the optimization error $\frac{1}{2}\|\Sigma_\eta^{-\frac{1}{2}}(y_{obs} - \widehat{y}_n)\|^2$ (bottom) of the Darcy problem ($N_\theta = 32$) with different noise levels: noiseless (left), 1% error (middle), and 5% error (right).

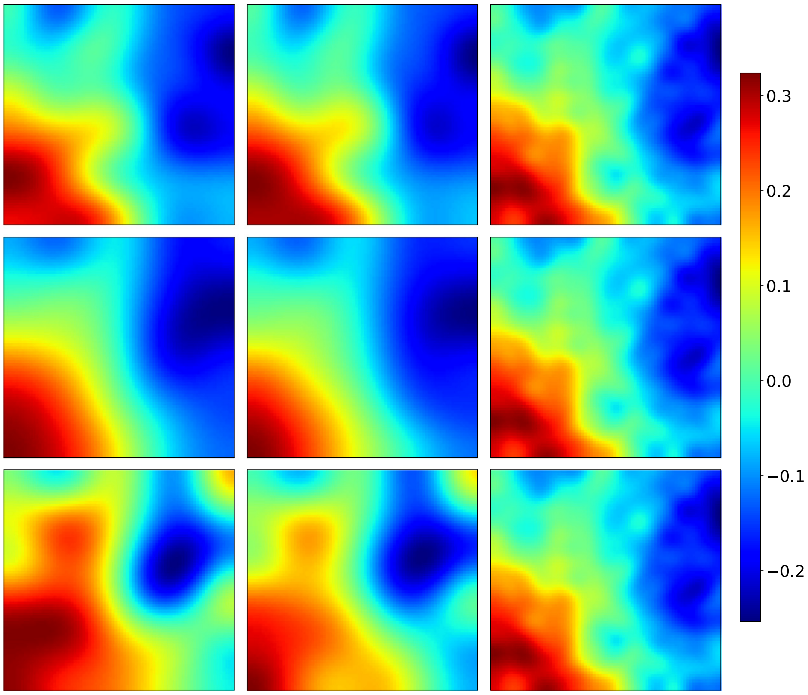


Fig. 6. Log-permeability fields $\log a(x, m_n)$ with $N_\theta = 32$ obtained by UKI, EKI, and the truth (left to right) for different noise levels: noiseless $\alpha = 1$ (top), 1% noise $\alpha = 0.5$ (middle), 5% noise $\alpha = 0.5$ (bottom).

estimated log-permeability fields $\log a(x, m_n)$ at the 50th iteration for different noise levels, obtained by the UKI and the truth random field, are depicted in Fig. 8. Comparing with the $N_\theta = 32$ case, all Kalman inversions with $N_\theta = 8$ perform better for the 5% noise scenario. This indicates the possibility of regularizing the inverse problem by reducing the parameter dimensionality.

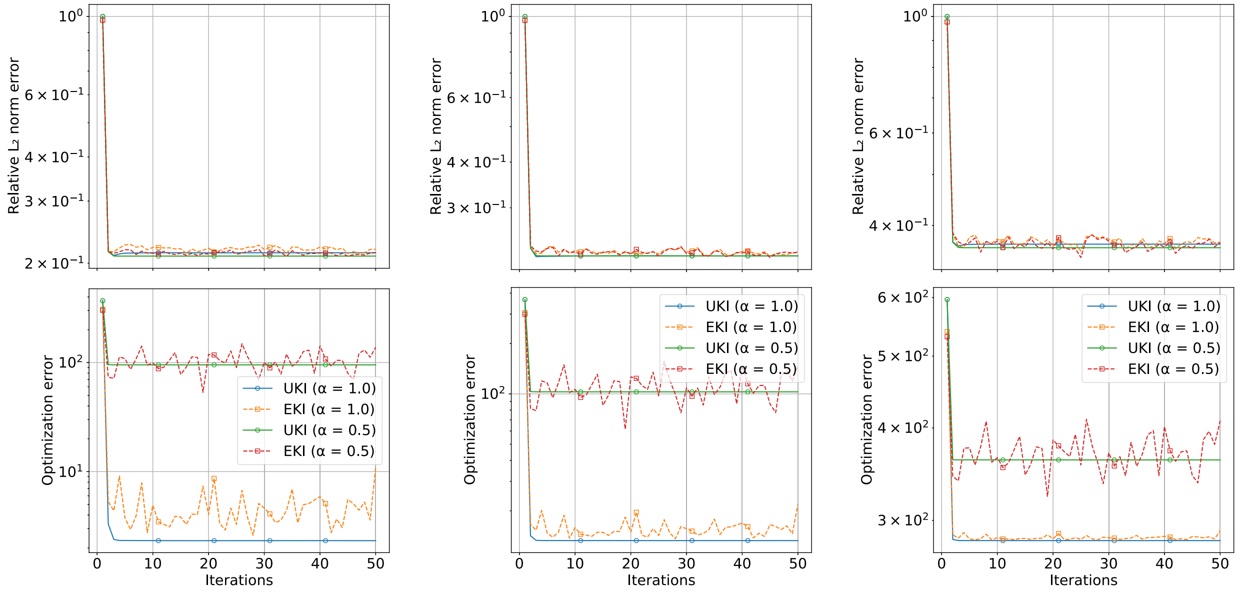


Fig. 7. Relative error $\frac{\|\log a(x,m_n) - \log a_{ref}(x)\|_2}{\|\log a_{ref}(x)\|_2}$ (top) and the optimization error $\frac{1}{2}\|\Sigma_\eta^{-\frac{1}{2}}(y_{obs} - \widehat{y}_n)\|^2$ (bottom) of the Darcy problem ($N_\theta = 8$) with different noise levels: noiseless (left), 1% error (middle), and 5% error (right).

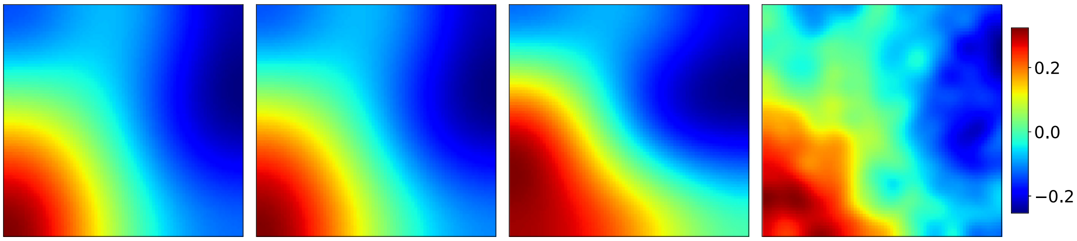


Fig. 8. Log-permeability fields $\log a(x, m_n)$ with $N_\theta = 8$ obtained by the UKI and the truth (right) for different noise levels: noiseless $\alpha = 1$ (left), 1% noise $\alpha = 1$ (middle-left), 5% noise $\alpha = 1$ (middle-right).

Finally we observe the smoothness, as a function of the iteration number, of the UKI in comparison to EKI. This may be seen in all the experiments undertaken in the Darcy flow example.

5.6. Damage detection problem

Consider a thin linear elastic arch-like plate, which is fixed on the bottom edges Γ_u. A traction boundary condition is applied on the top edge Γ_{t_1}, with distributed load $\bar{t} = (2, -20)$, and a traction free boundary condition is applied on the remaining edges Γ_{t_2}. See Fig. 9 The equations of linear elastostatics with plane stress assumptions are expressed in terms of the (Cauchy) stress tensor σ and take the form

$$\begin{aligned} \nabla \cdot \sigma + b &= 0 \text{ in } \Omega, \\ u &= 0 \text{ on } \Gamma_u, \\ \sigma \cdot n &= \bar{t} \text{ on } \Gamma_{t_1}, \\ \sigma \cdot n &= 0 \text{ on } \Gamma_{t_2}. \end{aligned} \tag{51}$$

Here u is the displacement vector, $b = 0$ is the body force vector, $\Omega \in \mathbb{R}^2$ is the bounded domain occupied by the plate. The strain tensor is

$$\varepsilon_{mn} = \frac{1}{2}\Big(\frac{\partial u_n}{\partial x_m} + \frac{\partial u_m}{\partial x_n}\Big). \tag{52}$$

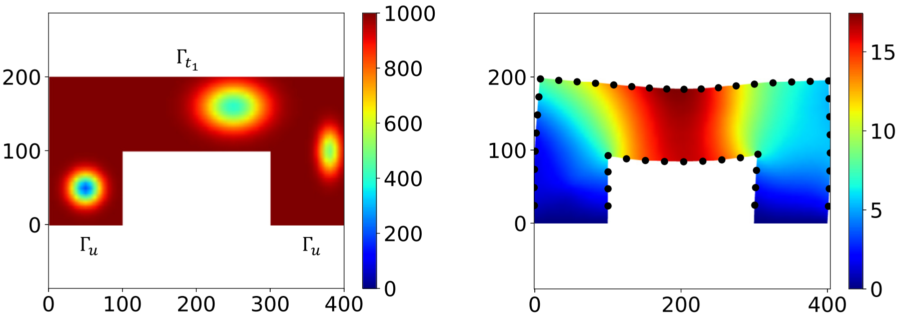


Fig. 9. The damaged Young's modulus (left) and the displacement magnitude field (right) with 46 measurement locations on the surface of the boundaries (black dots). The five unlabelled edges comprise Γ_{t_2}; see equation (51).

The linear constitutive relation between strain and stress is written as

$$\sigma_{ij} = \mathsf{C}_{ijmn}(E, \nu)\varepsilon_{mn}. \tag{53}$$

Here C_{ijmn} are the constitutive tensor components, which depend on the Young's modulus E and Poisson's ratio ν; throughout this study, we fix $\nu = 0.4$ and focus on learning the spatially-dependent damage information present in the field E. The damage is assumed to be isotropic elasticity-based damage with

$$E(x, \theta) = \big(1 - \omega(x, \theta)\big)E_0.$$

Throughout this study, we fix $E_0 = 1000$, and $\omega(x, \theta)$ is the scalar-valued damage variable, which varies between zero (no damage) to one (complete damage). The truth damage field (see Fig. 9-left) is

$$\omega_{ref}(x) = a_1 e^{-\frac{1}{2}(x-x_1)\Sigma_1^{-1}(x-x_1)} + a_2 e^{-\frac{1}{2}(x-x_2)\Sigma_2^{-1}(x-x_2)} + a_3 e^{-\frac{1}{2}(x-x_3)\Sigma_3^{-1}(x-x_3)},$$

$$a_1 = 0.8,\ a_2 = 0.6,\ a_3 = 0.5,$$

$$x_1 = \begin{bmatrix} 50 \\ 50 \end{bmatrix},\ x_2 = \begin{bmatrix} 250 \\ 160 \end{bmatrix},\ x_3 = \begin{bmatrix} 380 \\ 100 \end{bmatrix},\ \Sigma_1 = \begin{bmatrix} 200 & 0 \\ 0 & 200 \end{bmatrix},\ \Sigma_2 = \begin{bmatrix} 800 & 0 \\ 0 & 400 \end{bmatrix},\ \Sigma_3 = \begin{bmatrix} 100 & 0 \\ 0 & 400 \end{bmatrix},$$

and may be seen to exhibit three flaws. Noise is added to the observations on the boundary as in (48). The forward equation is solved by the finite element method with 384 quadratic quadrilateral elements (1649 nodes) using the NNFEM library [39,40].

For the inverse problem, the damage field is parameterized in terms of field $\theta(x)$ as follows

$$\omega(\theta(x)) = 0.9\frac{1 - e^{-\theta(x)}}{1 + 9e^{-\theta(x)}} \in (-0.1, 0.9).$$

Field $\theta(x)$ is itself discretized and represented by 24 quadratic quadrilateral elements ($N_\theta = 125$).[7] The observations are x_1 and x_2 displacements measured at 46 ($N_y = 92$) locations on the surface boundaries (see Fig. 9-right). We consider both $\alpha = 0.5$ and $\alpha = 1.0$, and we set $r_0 = 0$ and $\gamma = 1$. The UKI and EKI are both applied, initialized with $\theta_0 \sim \mathcal{N}(0, \mathbb{I})$. The observation error model used in the algorithm is $\eta \sim \mathcal{N}(0, 0.1^2\mathbb{I})$. For this problem the prior information $\omega(\theta = 0) = 0$ corresponds to an undamaged plate, and is expected to be reasonable for most of the domain. For the EKI, the ensemble size is set to $J = 500$, which is larger than the number of σ-points used in UKI ($2N_\theta + 1$).

The convergence of the damage field $\omega(\theta(x, m_n))$ and the optimization errors at each iteration are depicted in Fig. 10; the organization of the information is the same as in the Darcy flow example. In the noiseless scenario, the EKI exhibits divergence without regularization ($\alpha = 1.0$) due to the ill-posedness, however, the UKI converges.[8]

For noisy scenarios, the effect of overfitting is significant. At 1% noise level, setting $\alpha = 0.5$ eliminates overfitting; however at 5% noise level, setting $\alpha = 0.5$ does not eliminate overfitting. Therefore, the results obtained with $\alpha = 0.0$ are also reported for the 5% noise scenario. The estimated damaged Young's modulus fields $E(x, \theta)$ and the truth are depicted in Fig. 11. Both Kalman inversion methods perform comparably, and these three flaw areas are captured; however at 5% noise

[7] It is worth mentioning that increasing the parameter dimensionality by refining the parameter mesh exacerbates the ill-posedness and, therefore, deteriorates the performance of both Kalman inversions.

[8] We will see the same phenomenon in Subsection 5.7.

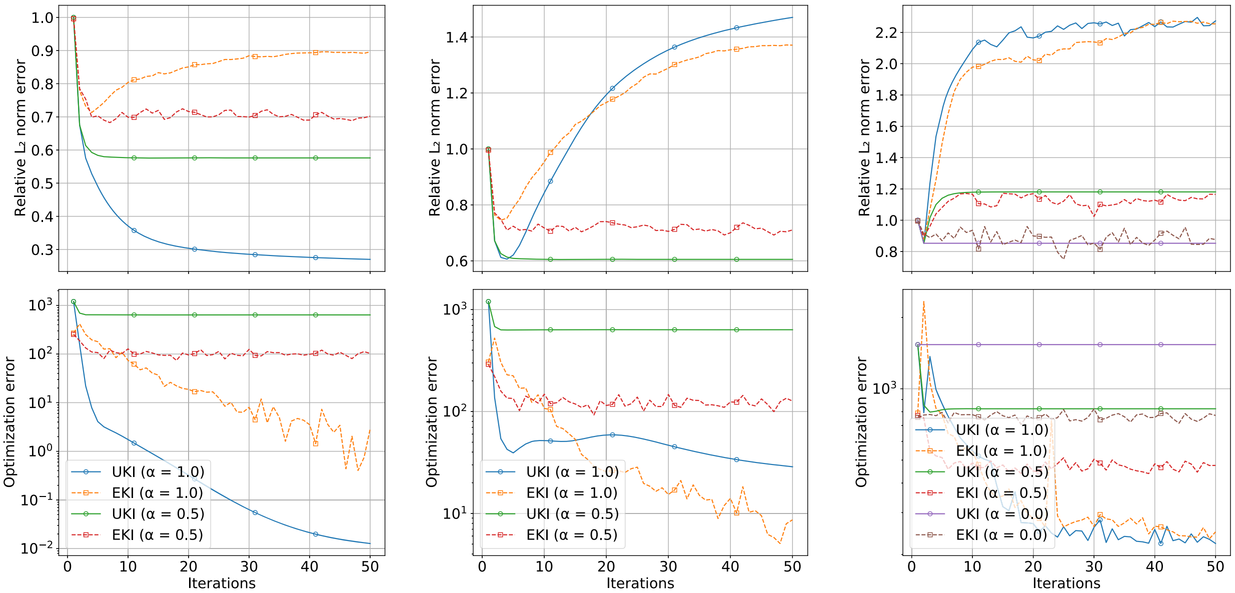


Fig. 10. Relative error $\dfrac{\|\omega(\theta(x, m_n)) - \omega_{ref}\|_2}{\|\omega_{ref}\|_2}$ (top) and the optimization error $\dfrac{1}{2}\|\Sigma_\eta^{-\frac{1}{2}}(y_{obs} - \widehat{y}_n)\|^2$ (bottom) of the damage detection problem with different noise levels: noiseless (left), 1% error (middle), and 5% error (right).

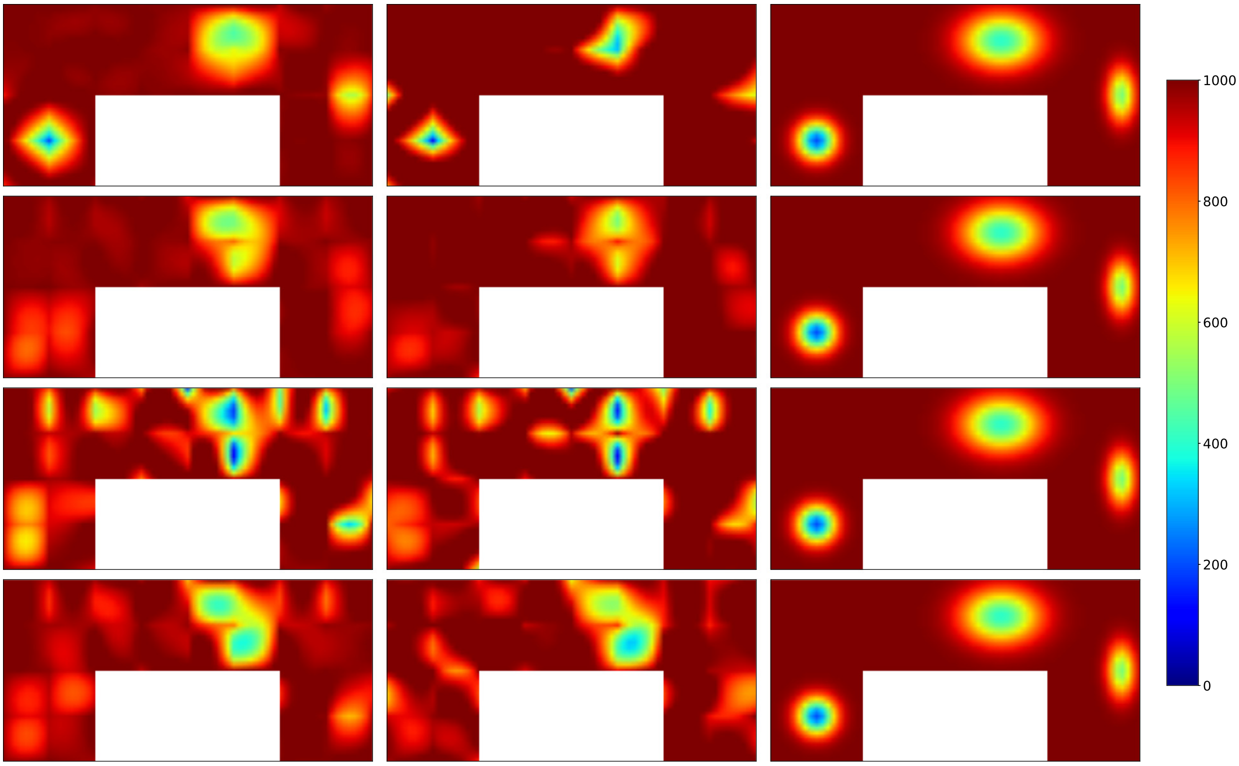


Fig. 11. Damaged Young's modulus fields $(1 - \omega(x, m_n))E_0$ obtained by UKI, EKI, and the truth (left to right) at different noise levels: noiseless $\alpha = 1$, 1% noise $\alpha = 0.5$, 5% noise $\alpha = 0.5$, and 5% noise $\alpha = 0$ (top to bottom).

level noticeable bias is visible in the flaws to the left and right of the domain. As in the Darcy flow case, the convergence histories of the UKI are smoother than for the EKI.

5.7. Navier-Stokes problem

We consider the 2D Navier-Stokes equation on a periodic domain $D = [0, 2\pi] \times [0, 2\pi]$:

$$\begin{aligned}
&\frac{\partial v}{\partial t} + (v \cdot \nabla)v + \nabla p - \nu \Delta v = 0,\\
&\nabla \cdot v = 0,
\end{aligned}$$

with initial condition chosen to imply the conservation law

$$\frac{1}{4\pi^2}\int v = v_b.$$

Here v and p denote the velocity vector and the pressure, $\nu = 0.01$ denotes the dynamic viscosity, and $v_b = (2\pi, 2\pi)$ denotes the non-zero mean background velocity. The forward problem is rewritten in the vorticity-streamfunction $(\omega - \psi)$ formulation:

$$\begin{aligned}
&\frac{\partial \omega}{\partial t} + (v \cdot \nabla)\omega - \nu \Delta \omega = 0,\\
&\omega = -\Delta\psi \qquad \frac{1}{4\pi^2}\int \psi = 0,\\
&v = \Big(\frac{\partial \psi}{\partial x_2}, -\frac{\partial \psi}{\partial x_1}\Big) + v_b,
\end{aligned}$$

and solved by the pseudo-spectral method [95] on a 128×128 grid. To eliminate aliasing error, the Orszag 2/3-Rule [96] is applied and, therefore there are 85^2 Fourier modes (padding with zeros). Time-integration is performed using the Crank–Nicolson method with $\Delta T = 2.5 \times 10^{-4}$.

We study the problem of recovering the initial vorticity field from measurements at positive times. We parameterize this field as $\omega_0(x, \theta)$, defined by parameters $\theta \in \mathbb{R}^{N_\theta}$, and modeled a priori as a Gaussian field with covariance operator $\mathsf{C} = \Delta^{-2}$, subject to periodic boundary conditions, on the space of spatial-mean zero functions. The KL expansion of the initial vorticity field is given by

$$\omega_0(x, \theta) = \sum_{l \in K} \theta_{(l)}^c \sqrt{\lambda_l}\psi_l^c + \theta_{(l)}^s \sqrt{\lambda_l}\psi_l^s, \tag{54}$$

where $K = \{(k_x, k_y) | k_x + k_y > 0 \text{ or } (k_x + k_y = 0 \text{ and } k_x > 0)\}$, and the eigenpairs are of the form

$$\psi_l^c(x) = \frac{\cos(l \cdot x)}{\sqrt{2}\pi} \qquad \psi_l^s(x) = \frac{\sin(l \cdot x)}{\sqrt{2}\pi} \qquad \lambda_l = \frac{1}{|l|^4},$$

and $\theta_{(l)}^c, \theta_{(l)}^s \sim \mathcal{N}(0, 2\pi^2)$ i.i.d. The KL expansion equation (54) can be rewritten as a sum over $\mathbb{Z}^{0+}$ rather than a lattice:

$$\omega_0(x, \theta) = \sum_{k \in \mathbb{Z}^{0+}} \theta_{(k)} \sqrt{\lambda_k}\psi_k(x), \tag{55}$$

where the eigenvalues λ_k are in descending order.

For the inverse problem, we recover the initial condition, specifically the initial vorticity field of the Navier-Stokes equation, given pointwise observations y_{ref} of the vorticity field at 16 equidistant points ($N_y = 32$) at $T = 0.25$ and $T = 0.5$ (see Fig. 12). The observations y_{obs} are defined as in (48). The initial vorticity field $\omega_{0,ref}$ is generated with all 85^2 Fourier modes, and the first $N_\theta = 100$ KL modes of equation (55) are recovered. We take $\alpha = 1.0$ and $\alpha = 0.9$, and fix $r_0 = 0$ and $\gamma = 10$. Both UKI and EKI are applied with $\theta_0 \sim \mathcal{N}(0, 10\mathbb{I})$ and the observation error assumed for inversion purposes is $\eta \sim \mathcal{N}(0, \mathbb{I})$. For the EKI, the ensemble size is set to be $J = 201$, which equals the number of σ-points in UKI $(2N_\theta + 1)$.

The convergence of the initial vorticity field $\omega_0(x, m_n)$ and the optimization errors for different noise levels at each iteration are depicted in Fig. 13; the organization of the figure is the same as in the Darcy case. In all scenarios, the UKI outperforms EKI. Moreover, without regularization ($\alpha = 1.0$), EKI exhibits slight divergence. This inverse problem is not sensitive to added Gaussian random noise, and the behaviour of any given Kalman inversion, with respect to different noise levels, are almost indistinguishable. The estimated initial vorticity fields $\omega_0(x, m_n)$ at the 50th iteration for different noise levels obtained by the Kalman inversions and the truth random field are depicted in Fig. 14. Both Kalman inversions capture main features of the truth random initial field, but not the detailed small features, due to the irreversibility of the diffusion process ($\nu = 0.01$).

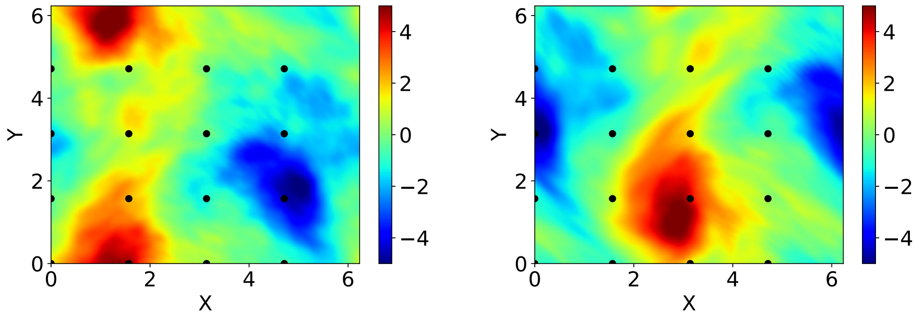


Fig. 12. The vorticity fields of the Navier-Stokes problem and the 16 equidistant pointwise measurements (black dots) at two observation times ($T = 0.25$ and $T = 0.5$).

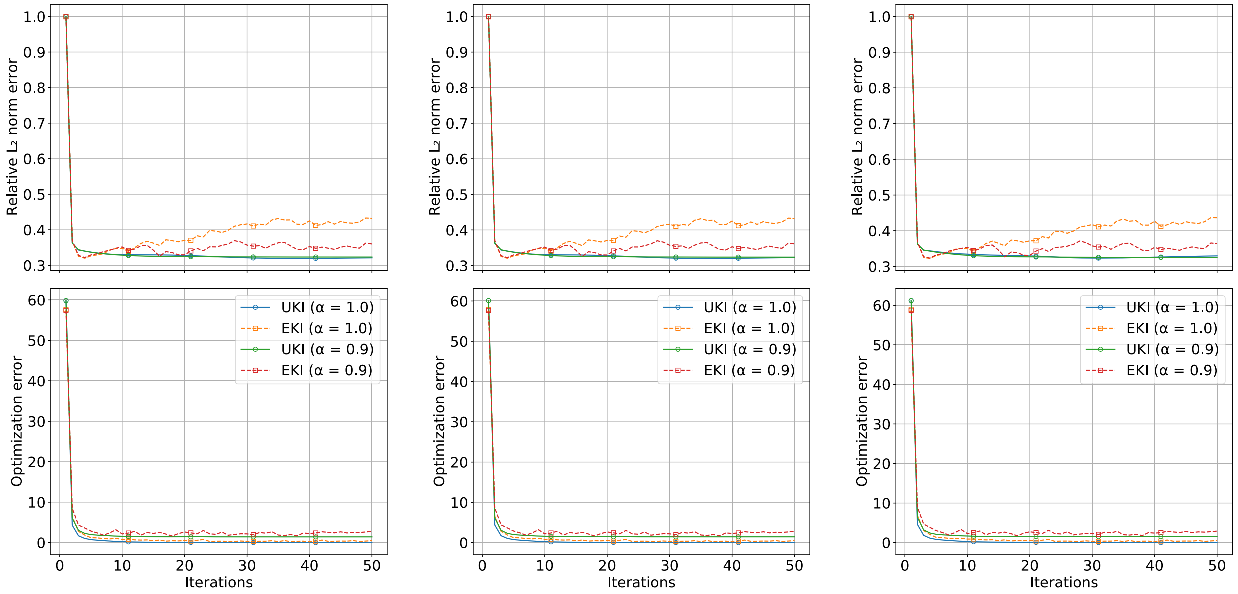


Fig. 13. Relative error $\frac{\|\omega_0(x,m_n)-\omega_{0,ref}\|_2}{\|\omega_{0,ref}\|_2}$ (top) and the optimization error $\frac{1}{2}\|\Sigma_\eta^{-\frac{1}{2}}(y_{obs}-\widehat{y}_n)\|^2$ (bottom) of the Navier-Stokes problem with different noise levels: noiseless (left), 1% error (middle), and 5% error (right).

5.8. Lorenz63 model problem

Consider the Lorenz63 system, a simplified mathematical model for atmospheric convection [97]:

$$\frac{dx_1}{dt} = \sigma(x_2 - x_1),$$
$$\frac{dx_2}{dt} = x_1(r - x_3) - x_2,$$
$$\frac{dx_3}{dt} = x_1x_2 - \beta x_3;$$

the system is parameterized by $\sigma, r, \beta \in \mathbb{R}_+$. We consider learning various subsets of these parameters from time-averaged data. To be concrete, the observation consists of the time-average of the various moments over time windows of size $T = 20$, with an initial spin-up period $T = 30$ to eliminate the influence of the initial condition; if $f : \mathbb{R}^3 \mapsto \mathbb{R}$ computes a moment, then we define

$$\overline{f(x)} = \frac{1}{20}\int_{30}^{50} f\big(x(t)\big)dt. \tag{56}$$

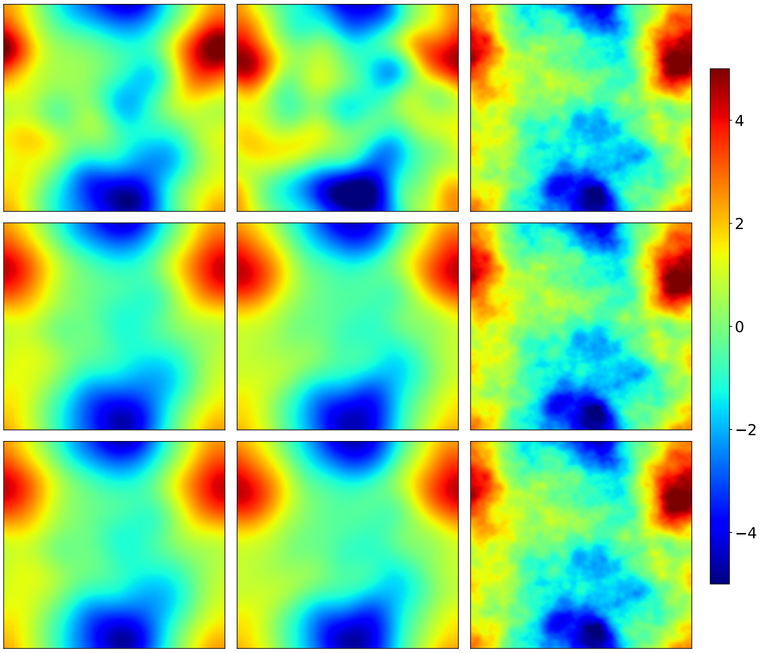


Fig. 14. Initial vorticity fields $\omega_0(x, m_n)$ recovered by UKI, EKI, and the truth (left to right) for different noise levels: noiseless $\alpha = 1$, 1% noise $\alpha = 0.9$, 5% noise $\alpha = 0.9$ (top to bottom).

We view this as an approximation of the ergodic average

$$\mathbb{E} f(x) = \lim_{\tau\to\infty} \frac{1}{\tau} \int_0^\tau f\big(x(t)\big)dt.$$

If the observation operator comprises finite time averages of the form (56) for a collection of moments $f(x)$ then we may reformulate the inverse problem as

$$y = \mathcal{G}(r) + \eta \tag{57}$$

with η a Gaussian which may be estimated from a long time trajectory (we use $T = 200$) by appealing to the central limit theorem [98]. In this interpretation $\mathcal{G}$ is the ergodic average. Note, however, that when we run any algorithm we will only use finite-time average approximations of $\mathcal{G}$.

The truth observation is computed with parameters $(\sigma,\ r,\ \beta) = (10,\ 28,\ 8/3)$ over a time window of size $T = 200$, also with an initial spin-up period $T = 30$. To estimate the statistics of η we split the observation time-series into 10 windows of size $T = 20$ and compute covariance of the observation error η following [62]. We set $r_0 = 5.0\mathbb{1}$ and $\gamma = 1$. The UKI is initialized with $\theta_0 \sim \mathcal{N}(5.0\mathbb{1}, \mathbb{I})$, and α is set to 1.

We start with the following one-parameter inverse problem with fixed $\sigma = 10$ and $\beta = 8/3$:

$$y = \mathcal{G}(r) + \eta \quad \text{with} \quad y = \overline{x_3}. \tag{58}$$

The UKI is applied, and the estimated r and the associated 3-σ confidence intervals at each iteration are depicted in Fig. 15. The confidence intervals give an indication of the evolving covariance C_n. The estimation of r at the 20th iteration is $r \sim \mathcal{N}(28.03, 0.22)$.

The landscape of $\mathcal{G}$ and sensitivity of $\mathcal{G}(\cdot)$ with respect to the input for observations, derived from chaotic problems such as equation (58), are widely studied [63,64]. We study them further, here, and the results are depicted in Fig. 16. The function $\mathcal{G}$ is characterized by a sudden change at $r \approx 22$ and the landscape is highly oscillatory for $r > 22$; furthermore, the sensitivity $d\mathcal{G}(r)$ computed with the discrete adjoint method blows up:

$$|d\mathcal{G}(r)| \propto \mathcal{O}(e^{\lambda T}),$$

with the value of the exponent λ consistent with the first global Lyapunov exponent [63,99]. This illustrates the challenges inherent in parameter estimation and sensitivity analyses for chaotic systems. In particular, the ExKI method suffers from the large derivatives of $\mathcal{G}$. Based on Lemma 2, it is natural to study the landscape of the averaged function $\mathcal{F}\mathcal{G}$ and its associated

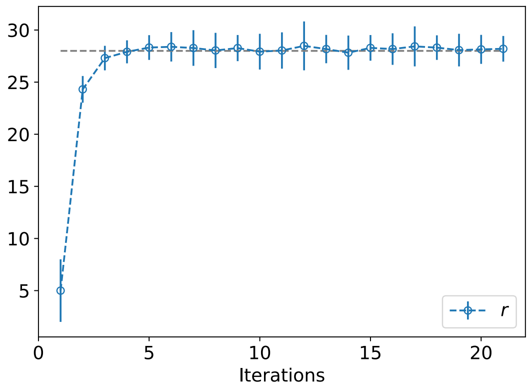


Fig. 15. Convergence of the 1-parameter Lorenz63 inverse problem with UKI ($\alpha = 1.0$); the true parameter value is represented by the dashed grey line.

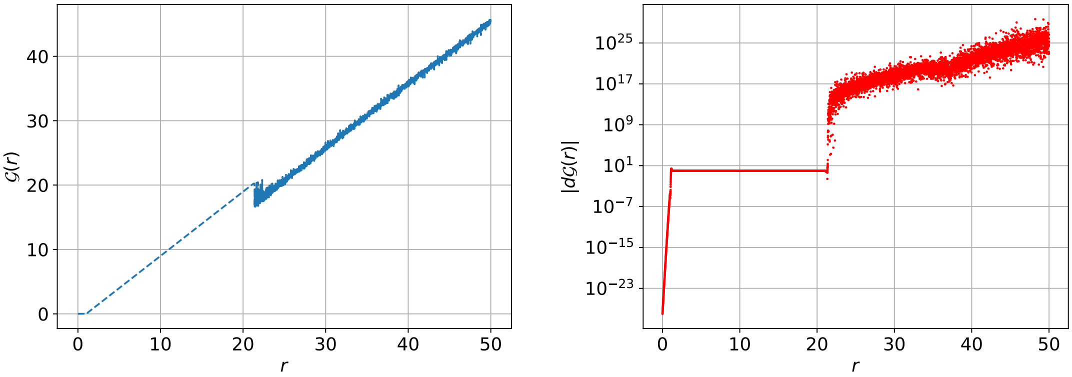


Fig. 16. Landscape (left) and sensitivity (right) of $\mathcal{G}$ in the 1-parameter Lorenz63 inverse problem equation (58).

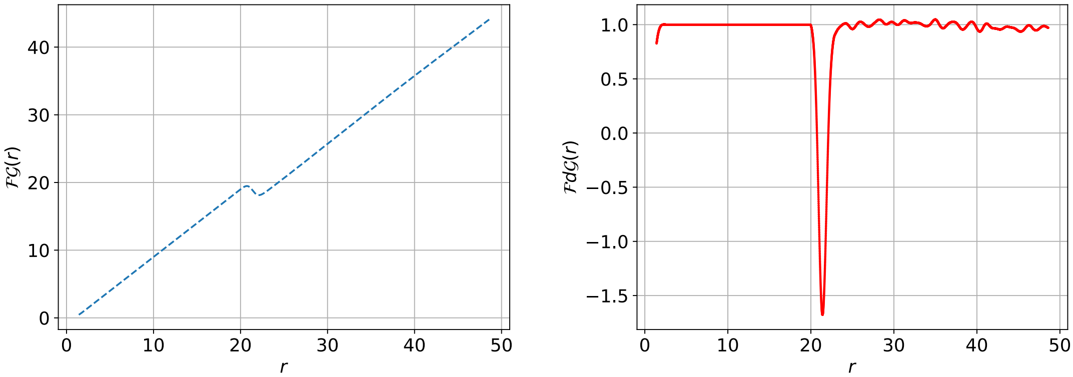


Fig. 17. Landscape (left) and sensitivity (right) of $\mathcal{F}\mathcal{G}$ in the 1-parameter Lorenz63 inverse problem equation (58) smoothed and viewed by UKI.

gradient $\mathcal{F}d\mathcal{G}$, with the standard deviation $\sigma_r = \sqrt{0.22}$ fixed; this gives an indication of the landscape as perceived by the UKI. In particular, we have:

$$\mathcal{F}\mathcal{G}(r) = \int \mathcal{G}(x)\frac{1}{\sqrt{2\pi}\sigma_r}e^{-\frac{(x-r)^2}{2\sigma_r^2}}dx, \qquad \mathcal{F}d\mathcal{G}(r) = \frac{\int (x-r)(\mathcal{G}(x)-\mathcal{G}(r))\frac{1}{\sqrt{2\pi}\sigma_r}e^{-\frac{(x-r)^2}{2\sigma_r^2}}dx}{\int (x-r)^2\frac{1}{\sqrt{2\pi}\sigma_r}e^{-\frac{(x-r)^2}{2\sigma_r^2}}dx}.$$

These functions are depicted in Fig. 17, which should be compared with Fig. 16. We see that $\mathcal{F}\mathcal{G}$ is smooth (except the transition point), and $\mathcal{F}d\mathcal{G}$ does not suffer from blow-up in the way $d\mathcal{G}$ does; furthermore, $\mathcal{F}d\mathcal{G}$ represents the averaged gradient $\overline{d\mathcal{G}(r)} \approx 0.96$ well, away from the blow-up regions. This explains why the adjoint/gradient-based methods, including ExKI, fail, but the UKI succeeds for this chaotic inverse problem.

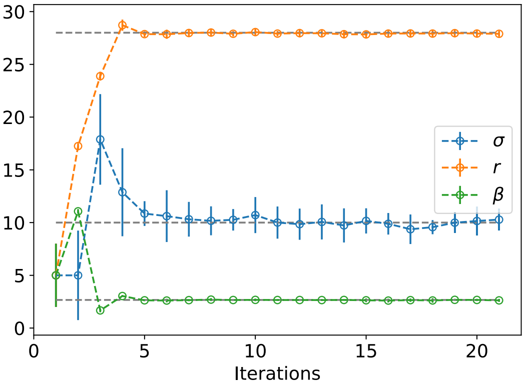


Fig. 18. Convergence of the 3-parameter Lorenz63 inverse problem with UKI ($\alpha = 1.0$); true parameter values are represented by dashed grey lines.

Next, we consider a three-parameter inverse problem, using the ideas in Subsection 4.1. Let $\theta = (\theta_{(1)}, \theta_{(2)}, \theta_{(3)})$ and let $(\sigma, r, \beta) = (|\theta_{(1)}|, |\theta_{(2)}|, |\theta_{(3)}|)$. The map $\mathcal{G}(\theta)$ is found by computing time-averages of all three components of x, as described above, for given input parameter θ. The use of the modulus helps ensure solution trajectories which do not blow-up. We have

$$y = \mathcal{G}(\theta) + \eta \quad \text{with} \quad y = (\overline{x_1}, \overline{x_2}, \overline{x_3}, \overline{x_1^2}, \overline{x_2^2}, \overline{x_3^2}). \tag{59}$$

All other aspects of the setup are the same as the aforementioned one-parameter inverse problem. The estimated parameters and associated 3-σ confidence intervals for each component at each iteration are depicted in Fig. 18. The estimation of the parameters at the 20th iteration is

$$(\sigma \quad r \quad \beta) = (10.28 \quad 27.90 \quad 2.63)$$

For both scenarios, the UKI converges efficiently, thanks to the linear (or superlinear) convergence rate of the LMA and the averaging property.

5.9. Multiscale Lorenz96 problem

Consider the multi-scale Lorenz96 system, a simplified mathematical model for the midlatitude atmosphere [100], with K slow variables $X^{(k)}$ which are each coupled with J fast variables $Y^{(j,k)}$, given by:

$$\begin{aligned} \frac{dX^{(k)}}{dt} &= -X^{(k-1)}(X^{(k-2)} - X^{(k+1)}) - X^{(k)} + F - \frac{hc}{b}\sum_{j=1}^{J} Y^{(j,k)}, \\ \frac{dY^{(j,k)}}{dt} &= -cbY^{(j+1,k)}(Y^{(j+2,k)} - Y^{(j-1,k)}) - cY^{(j,k)} + \frac{hc}{b}X^k. \end{aligned} \tag{60}$$

To close the system, it is appended with the cyclic boundary conditions $X^{(k+K)} = X^{(k)}$, $Y^{(j,k+K)} = Y^{(j,k)}$ and $Y^{(j+J,k)} = Y^{(j,k+1)}$. The time scale separation is parameterized by the coefficient c and the large-scales are subjected to external forcing F. We choose here as parameters $K = 8$, $J = 32$, $F = 20$, $c = b = 10$ and $h = 1$ as in [101–104]. As time-integrator, we use the 4th-order Runge Kutta method with $\Delta T = 5 \times 10^{-3}$.

Our goal is to learn the closure model $\psi(X)$ of the fast dynamics for a reduced model of the form

$$\frac{dX^{(k)}}{dt} = -X^{(k-1)}(X^{(k-2)} - X^{(k+1)}) - X^{(k)} + F + \psi(X^{(k)}).$$

The closure model $\psi : D \subset \mathbb{R} \mapsto \mathbb{R}$ is parameterized by the finite element method with cubic Hermite polynomials. The domain is set to be $D = [-20, 20]$ and decomposed into 5 elements and, therefore, $N_\theta = 12$.

For the inverse problem, the observations consist of the time-average of the first and second moments of $X^{(1)}, X^{(2)}, X^{(3)}$, and $X^{(4)}$ over a time window of size $T = 1000$ and, therefore $N_y = 14$. The same central limit theorem arguments are used to formulate the problem as in the Lorenz63 model. The truth observation y_{ref} is computed with the multiscale chaotic system equation (60) with a random initial condition $X^{(k)} \sim \mathcal{N}(0, 1)$ and $Y^{(j,k)} \sim \mathcal{N}(0, 0.01^2)$. And 1%, 2%, and 5% Gaussian random noises are added to the observation following equation (48).

We set $r_0 = 0$ and $\gamma = 1$; the UKI is thus initialized with $\theta_0 \sim \mathcal{N}(0, \mathbb{I})$. The observation error is set to be $\eta = \mathcal{N}(0, \mathrm{diag}\{0.05^2 y_{obs} \odot y_{obs}\})$, and we take $\alpha = 1$, since the system is over-determined. Moreover, these simulations start with another random initialization of $X^{(k)} \sim \mathcal{N}(0, 1)$. The learned closure models at the 20th iteration are reported in

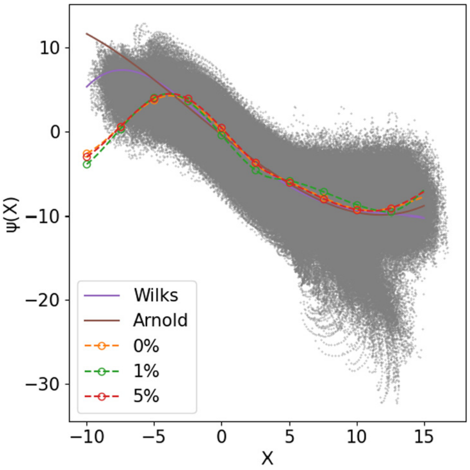


Fig. 19. Closure terms $\psi(X)$ for the multi-scale Lorenz96 system obtained from the truth (grey dots) and polynomial data-fitting by Wilks [102] and Arnold [103], compared with what is learned using the UKI approach ($\alpha = 1$) with different noise levels.

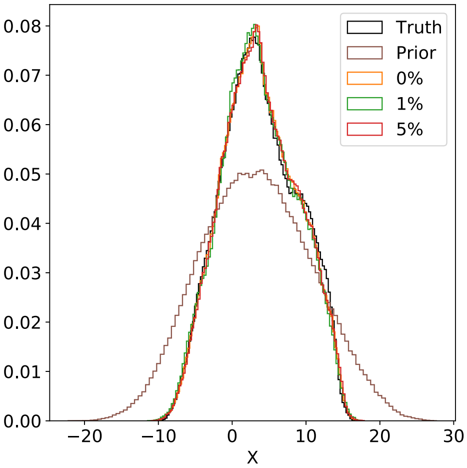


Fig. 20. Empirical probability density functions of the slow variables $X^{(k)}$ obtained from the full multi-scale Lorenz96 system (Truth), the initial closure model (Prior), and the closure models learned by the UKI ($\alpha = 1$) at different noise levels.

Fig. 19. The estimated empirical probability density functions of the slow variables are reported in Fig. 20. For all scenarios, although the learned closure models show non-trivial variability with respect to those published in [102,103] at the left most extreme of D, the predicted probability density functions match well with the reference, obtained from a full multiscale simulation. It is worth mentioning this problem is not sensitive with respect to the added Gaussian random noise.

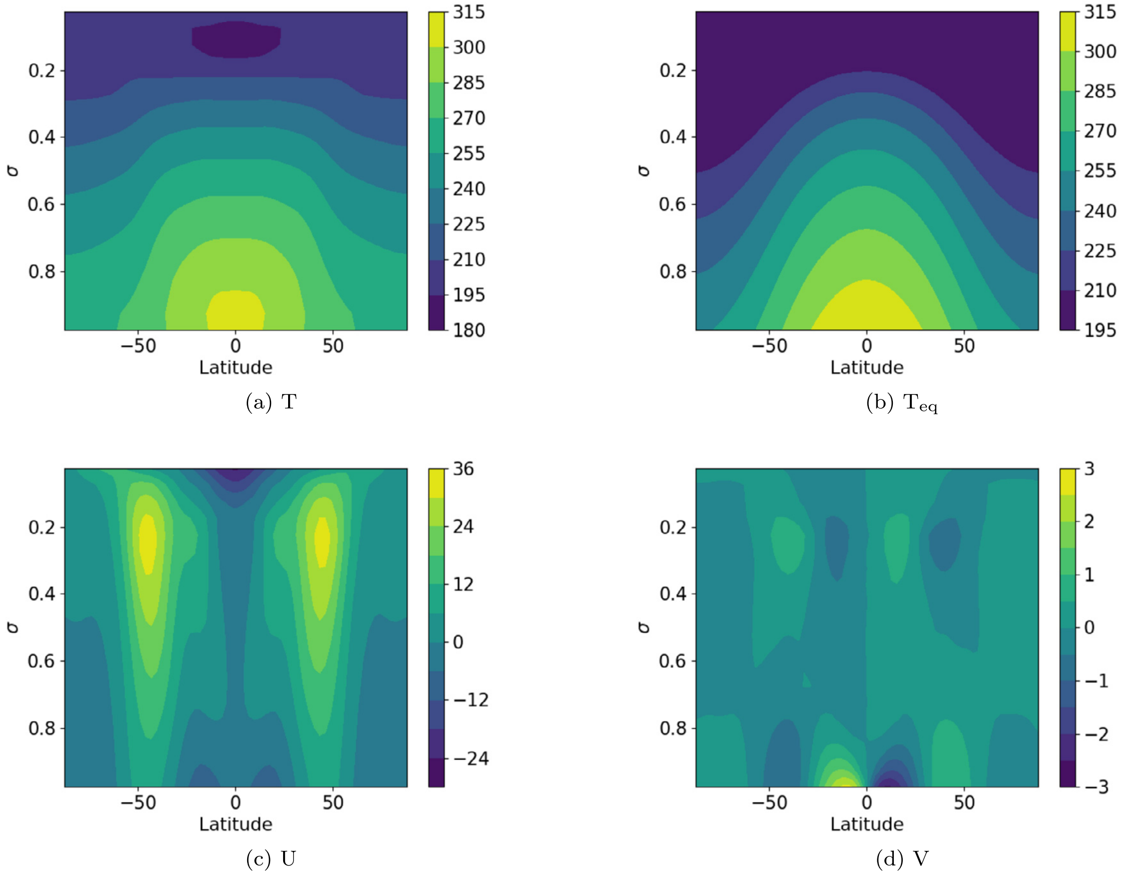
5.10. Idealized general circulation model

Finally, we consider an idealized general circulation model. The model is based on the 3D Navier-Stokes equations, making the hydrostatic and shallow-atmosphere approximations common in atmospheric modeling. Specifically, we test UKI on the well-known Held-Suarez test case [105], in which a detailed radiative transfer model is replaced by Newtonian relaxation of temperatures toward a prescribed "radiative equilibrium" $T_{\rm eq}(\phi, p)$ that varies with latitude ϕ and pressure p. Specifically, the thermodynamic equation for temperature T

$$\frac{\partial T}{\partial t} + \dots = Q$$

(dots denoting advective and pressure work terms) contains a diabatic heat source

$$Q = -k_T(\phi, p, p_s)\big(T - T_{\rm eq}(\phi, p)\big),$$



(a) T

(b) $\mathrm{T_{eq}}$

(c) U

(d) V

Fig. 21. Zonal mean profile of temperature (a), radiative equilibrium temperature (b), zonal wind velocity (c), and meridional wind velocity (d), all from a 1000-day average. The horizontal coordinate is latitude and the vertical coordinate is the nondimensional σ coordinate of the model.

with relaxation coefficient (inverse relaxation time)

$$k_T = k_a + (k_s - k_a)\max\Big(0, \frac{\sigma - \sigma_b}{1-\sigma_b}\Big)\cos^4\phi.$$

Here, $\sigma = p/p_s$, which is pressure p normalized by surface pressure p_s, is the vertical coordinate of the model, and

$$T_{\text{eq}} = \max\Big\{200K, \Big[315K - \Delta T_y \sin^2\phi - \Delta\theta_z \log\Big(\frac{p}{p_0}\Big)\cos^2\phi\Big]\Big(\frac{p}{p_0}\Big)^{\kappa}\Big\}$$

is the equilibrium temperature profile ($p_0 = 10^5$ Pa is a reference surface pressure and $\kappa = 2/7$ is the adiabatic exponent). Default parameters are

$$k_a = (40\text{ day})^{-1}, \qquad k_s = (4\text{ day})^{-1}, \qquad \Delta T_y = 60\text{ K}, \qquad \Delta\theta_z = 10\text{ K}.$$

For the numerical simulations, we use the spectral transform method in the horizontal, with T42 spectral resolution (triangular truncation at wavenumber 42, with 64 × 128 points on the latitude-longitude transform grid); we use 20 vertical levels equally spaced in σ. With the default parameters, the model produces an Earth-like zonal-mean circulation, albeit without moisture or precipitation. A single jet is generated with maximum strength of roughly 30 m s^{-1} near 45° latitude (Fig. 21).

Our inverse problem is constructed to learn parameters in the Newtonian relaxation term Q:

$$(k_a,\ k_s,\ \Delta T_y,\ \Delta\theta_z).$$

We do so in the presence of the following constraints:

$$0\text{ day}^{-1} < k_a < 1\text{ day}^{-1}, \qquad k_a < k_s < 1\text{ day}^{-1} + k_a, \qquad 0\text{ K} < \Delta T_y, \qquad 0\text{ K} < \Delta\theta_z.$$

Conceptually, the setting is identical to that for the Lorenz63 example. We use the same overline notation to denote averaging, which here in addition to the time average in the Lorenz models also includes a zonal average over longitude (because

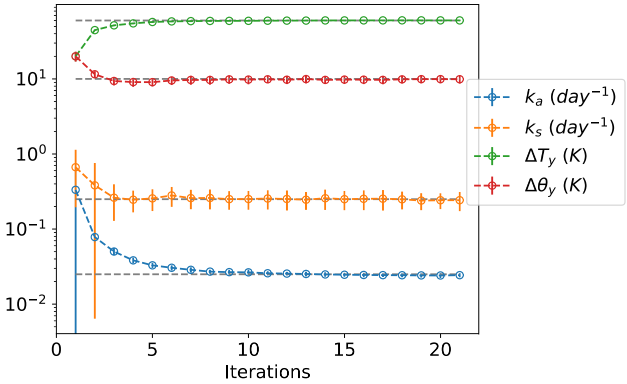


Fig. 22. Convergence of the idealized general circulation model inverse problem with UKI ($\alpha = 1.0$). The true parameter values are represented by dashed grey lines.

the model is statistically symmetric under rotations around the planet's spin axis), and we apply the same central limit theorem arguments to formulate the inverse problem. To incorporate the imposition of the constraints, the inverse problem is formulated as follows (see Subsection 4.1 for details):

$$y = \mathcal{G}(\theta) + \eta \quad \text{with} \quad \mathcal{G}(\theta) = \overline{T}(\phi, \sigma) \tag{61}$$

with the parameter transformation

$$\theta : (k_a, k_s, \Delta T_y, \Delta\theta_z) = \Big(\frac{1}{1 + |\theta_{(1)}|},\ \frac{1}{1 + |\theta_{(1)}|} + \frac{1}{1 + |\theta_{(2)}|},\ |\theta_{(3)}|, |\theta_{(4)}|\Big). \tag{62}$$

The observation mapping is defined by mapping from the unknown θ to the 200-day zonal mean of the temperature as a function of latitude (ϕ) and height (σ), after an initial spin-up of 200 days. The truth observation is the 1000-day zonal mean of the temperature (see Fig. 21-a), after an initial spin-up 200 days to eliminate the influence of the initial condition. Because the truth observations come from an average 5 times as long as the observation window used for parameter learning, the chaotic internal variability of the model introduces noise in the observations. As for the Lorenz63 setting, the central limit theorem may be invoked to model the observation error from internal variability.

To perform the inversion, we set $r_0 = [2 \text{ day},\ 2 \text{ day},\ 20 \text{ K},\ 20 \text{ K}]^T$ and $\gamma = 1$. Thus UKI is initialized with $\theta_0 \sim \mathcal{N}\Big(r_0,\ \mathbb{I}\Big)$. Within the algorithm, we assume that the observation error satisfies $\eta \sim \mathcal{N}(0 \text{ K}, 3^2\mathbb{I} \text{ K}^2)$. Because the problem is over-determined, we set $\alpha = 1$. The estimated parameters and associated 3-σ confidence intervals for each component at each iteration are depicted in Fig. 22. The estimation of model parameters at the 20th iteration is

$$\begin{pmatrix} k_a & k_s & \Delta T_y & \Delta\theta_z \end{pmatrix} = \begin{pmatrix} 0.0243 \text{ day}^{-1} & 0.243 \text{ day}^{-1} & 60.2 \text{ K} & 9.91 \text{ K} \end{pmatrix}.$$

UKI converges to the true parameters in fewer than 10 iterations with 9 σ-points, demonstrating the potential of applying UKI for large-scale inverse problems.

6. Conclusion

We introduced a novel stochastic dynamical system, into which an arbitrary inverse problem may be embedded as an observation operator; by applying filtering methods to this stochastic dynamical system we obtain methods to solve inverse problems. In the linear case, we have demonstrated that this approach leads to an unusual Tikhonov regularized least squares solution, with prior covariance depending on the forward model, and a tunable parameter in the stochastic dynamical system determining the level of regularization. We have also introduced unscented Kalman inversion (UKI) and shown that it outperforms the EKI, when applied to the same novel stochastic dynamical system. As well as outperforming EKI, UKI shares its advantages: it is derivative-free, black-box, embarrassingly parallel, and robust. Our numerical results demonstrate its theoretical properties and its applicability; in particular, it is demonstrated to outperform the EKI on large scale problems in which the number of unknown parameters is small. Because the methodology constitutes a novel approach to parameter estimation, there are many avenues for future research, including applications of the method, methodological improvements and extensions, and theoretical analysis.

CRediT authorship contribution statement

Daniel Zhengyu Huang: Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. **Tapio Schneider:** Conceptualization, Funding acquisition, Writing – original draft. **Andrew M. Stuart:** Conceptualization, Funding acquisition, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the generosity of Eric and Wendy Schmidt by recommendation of the Schmidt Futures program and by the National Science Foundation (NSF, award AGS-1835860). A.M.S. was also supported by the Office of Naval Research (award N00014-17-1-2079). The authors thank Sebastian Reich and anonymous reviewers for helpful comments on an earlier draft.

Appendix A. Proof of theorems

Proof of Proposition 1. An affine transformation is an invertible mapping from R^{N_θ} to R^{N_θ} of the form ${}^*x = Ax + b$. When we apply the following affine transformation

$$
{}^*m_n = Am_n + b \qquad {}^*C_n = AC_nA^T \quad \text{with} \quad {}^*r_0 = Ar_0 + b \qquad {}^*\Sigma_\omega = A\Sigma_\omega A^T,
$$

keep y_n and Σ_ν unchanged, and define ${}^*\mathcal{G}(\theta) = \mathcal{G}\big(A^{-1}(\theta - b)\big)$. We prove

$$
{}^*m_{n+1} = Am_{n+1} + b \qquad {}^*C_{n+1} = AC_{n+1}A^T. \tag{A.1}
$$

Equation (9) leads to

$$
{}^*\widehat{m}_{n+1} = \alpha {}^*m_n + (1-\alpha){}^*r_0 = A\widehat{m}_{n+1} + b \qquad {}^*\widehat{C}_{n+1} = \alpha^2 {}^*C_n + {}^*\Sigma_\omega = A\widehat{C}_{n+1}A^T. \tag{A.2}
$$

Therefore, the distribution of ${}^*\theta_{n+1}|Y_n \sim \mathcal{N}({}^*\widehat{m}_{n+1}{}^*\widehat{C}_{n+1})$ is the same as $A\theta_{n+1} + b|Y_n$ and equation (10) becomes

$$
{}^*\widehat{y}_{n+1} = \widehat{y}_{n+1} \qquad {}^*\widehat{C}^{\theta y}_{n+1} = A\widehat{C}^{\theta y}_{n+1} \qquad {}^*\widehat{C}^{yy}_{n+1} = \widehat{C}^{yy}_{n+1}. \tag{A.3}
$$

Finally, equation (11) leads to

$$
\begin{aligned}
&{}^*m_{n+1} = {}^*\widehat{m}_{n+1} + {}^*\widehat{C}^{\theta y}_{n+1}({}^*\widehat{C}^{yy}_{n+1})^{-1}(y_{n+1} - {}^*\widehat{y}_{n+1}) = Am_{n+1} + b,\\
&{}^*C_{n+1} = {}^*\widehat{C}_{n+1} - {}^*\widehat{C}^{\theta y}_{n+1}({}^*\widehat{C}^{yy}_{n+1})^{-1}{}^*\widehat{C}^{\theta y}_{n+1}{}^{T} = AC_{n+1}A^T. \quad \square
\end{aligned} \tag{A.4}
$$

Proof of Lemma 1. In this proof recall that $\theta \sim \mathcal{N}(m, C)$. When both $\mathcal{G}_1 = G_1$ and $\mathcal{G}_2 = G_2$ are linear transformations, we have

$$
\begin{aligned}
\mathbb{E}[\mathcal{G}_i(\theta)] &= G_i\mathbb{E}[\theta] = G_i m = \mathcal{G}_i(m),\\
\mathrm{Cov}[\mathcal{G}_1(\theta), \mathcal{G}_2(\theta)] &= G_1\mathrm{Cov}[\theta,\theta]G_2^T = G_1CG_2^T,\\
\sum_{j=1}^{2N_\theta} W_j^c(\mathcal{G}_1(\theta^j) - \mathbb{E}\mathcal{G}_1(\theta))(\mathcal{G}_2(\theta^j) - \mathbb{E}\mathcal{G}_2(\theta))^T &= \frac{1}{2a^2N_\theta}\sum_{j=1}^{2N_\theta}(G_1\cdot\theta^j - G_1m)(G_2\cdot\theta^j - G_2m)^T\\
&= \frac{1}{2a^2N_\theta}\sum_{j=1}^{2N_\theta} 2G_1c_j[\sqrt{C}]_jc_j[\sqrt{C}]_jG_2^T\\
&= G_1CG_2^T.
\end{aligned}
$$

In the following we use $\nabla^k\mathcal{G}_i$ to denote the k^{th} derivative of $\mathcal{G}_i$ evaluated at m. For the nonlinear case, Taylor's expansion of $\mathcal{G}_i(\cdot)$ at m is then

$$
\mathcal{G}_i(\theta) = \mathcal{G}_i(m) + \nabla\mathcal{G}_i\delta\theta + \frac{1}{2}\nabla^2\mathcal{G}_i\delta\theta\otimes\delta\theta + \frac{1}{6}\nabla^3\mathcal{G}_i\delta\theta\otimes\delta\theta\otimes\delta\theta + O(\|\delta\theta\|^4) \quad \text{with} \quad \delta\theta = \theta - m.
$$

The mean approximation is thus first-order accurate:

$$\mathbb{E}\mathcal{G}_i(\theta) = \mathcal{G}_i(m) + O(\|C\|).$$

The covariance approximation is second-order accurate:

$$\begin{aligned}
\mathrm{Cov}[\mathcal{G}_1(\theta), \mathcal{G}_2(\theta)] &= \mathbb{E}\,(\mathcal{G}_1(\theta) - \mathbb{E}\mathcal{G}_1(\theta))\,(\mathcal{G}_2(\theta) - \mathbb{E}\mathcal{G}_2(\theta))^T \\
&= \mathbb{E}\left(\nabla\mathcal{G}_1\delta\theta + \frac{1}{2}\nabla^2\mathcal{G}_1(\delta\theta\otimes\delta\theta - C)\right)\left(\nabla\mathcal{G}_2\delta\theta + \frac{1}{2}\nabla^2\mathcal{G}_2(\delta\theta\otimes\delta\theta - C)\right)^T + \mathcal{O}(\|C\|^2) \\
&= \nabla\mathcal{G}_1 C \nabla\mathcal{G}_2^T + \mathcal{O}(\|C\|^2),
\end{aligned}$$

whilst we also have

$$\begin{aligned}
&\sum_{j=1}^{2N_\theta} W_j^c(\mathcal{G}_1(\theta^j) - \mathbb{E}\mathcal{G}_1(\theta))(\mathcal{G}_2(\theta^j) - \mathbb{E}\mathcal{G}_2(\theta))^T \\
&\quad= \sum_{j=1}^{2N_\theta} W_j^c(\mathcal{G}_1(\theta^j) - \mathcal{G}_1(m))(\mathcal{G}_2(\theta^j) - \mathbb{E}\mathcal{G}_2(m))^T \\
&\quad= \sum_{j=1}^{N_\theta} W_j^c(\nabla\mathcal{G}_1 c_j[\sqrt{C}]_j + \frac{1}{2}\nabla^2\mathcal{G}_1 c_j[\sqrt{C}]_j \otimes c_j[\sqrt{C}]_j)(\nabla\mathcal{G}_2 c_j[\sqrt{C}]_j + \frac{1}{2}\nabla^2\mathcal{G}_2 c_j[\sqrt{C}]_j \otimes c_j[\sqrt{C}]_j)^T \\
&\quad+ \sum_{j=1}^{N_\theta} W_j^c(-\nabla\mathcal{G}_1 c_j[\sqrt{C}]_j + \frac{1}{2}\nabla^2\mathcal{G}_1 c_j[\sqrt{C}]_j \otimes c_j[\sqrt{C}]_j)(-\nabla\mathcal{G}_2 c_j[\sqrt{C}]_j + \frac{1}{2}\nabla^2\mathcal{G}_2 c_j[\sqrt{C}]_j \otimes c_j[\sqrt{C}]_j)^T \\
&\qquad+ \mathcal{O}(\|C\|^2) \\
&\quad= \frac{1}{2a^2 N_\theta}\sum_{j=1}^{N_\theta} 2\nabla\mathcal{G}_1 c_j[\sqrt{C}]_j c_j[\sqrt{C}]_j \nabla\mathcal{G}_2^T + \mathcal{O}(\|C\|^2) \\
&\quad= \nabla\mathcal{G}_1 C \nabla\mathcal{G}_2^T + \mathcal{O}(\|C\|^2). \quad \square
\end{aligned}$$

Proof of Theorem 1. In this proof we let $\mathcal{B}$ denote the Banach space of matrices in $\mathbb{R}^{N_\theta \times N_\theta}$ equipped with the operator norm induced by the Euclidean norm on $\mathbb{R}^{N_\theta}$. Furthermore, we let $\mathcal{L}$ denote the Banach space of bounded linear operators from $\mathcal{B}$ into itself, equipped with the standard induced operator norm. For simplicity we consider the case $r_0 = 0$; a change of origin may be used to extend to the case $r_0 \neq 0$. We first prove that the precision operators converge: $\lim_{n\to\infty} C_n^{-1} = C_\infty^{-1}$; we then study behaviour of the mean sequence $\{m_n\}_{n\in\mathbb{Z}^+}$. For both the precision and the mean we first study $\alpha \in (0,1)$ and then $\alpha = 1$. In what follows it is useful to note [80][Theorem 4.1] that the mean and covariance update equations (27) can be rewritten as

$$\begin{aligned}
C_{n+1}^{-1} &= G^T\Sigma_\nu^{-1}G + (\alpha^2 C_n + \Sigma_\omega)^{-1}, \\
C_{n+1}^{-1}m_{n+1} &= G^T\Sigma_\nu^{-1}y + (\alpha^2 C_n + \Sigma_\omega)^{-1}\alpha m_n;
\end{aligned} \tag{A.5}$$

furthermore the iteration for the covariance remains in the cone of positive semi-definite matrices [80][Theorem 4.1]. Since $\Sigma_\omega \succ 0$, the sequence $\{C_n^{-1}\}$ is bounded:

$$G^T\Sigma_\nu^{-1}G \preceq C_n^{-1} \preceq G^T\Sigma_\nu^{-1}G + \Sigma_\omega^{-1}, \quad \forall n \in \mathbb{Z}_+. \tag{A.6}$$

Introducing $(C_n')^{-1} := \Sigma_\omega^{\frac{1}{2}} C_n^{-1} \Sigma_\omega^{\frac{1}{2}}$, we may rewrite the covariance update equation (A.5) in the form

$$(C_{n+1}')^{-1} = \Sigma_\omega^{\frac{1}{2}} G^T\Sigma_\nu^{-1}G\Sigma_\omega^{\frac{1}{2}} + \left(\alpha^2 C_n' + \mathbb{I}\right)^{-1}. \tag{A.7}$$

We define the map

$$f(X;\alpha) = \Sigma_\omega^{\frac{1}{2}} G^T\Sigma_\nu^{-1}G\Sigma_\omega^{\frac{1}{2}} + \left(\alpha^2 X^{-1} + \mathbb{I}\right)^{-1} \tag{A.8}$$

noting that then $(C_{n+1}')^{-1} = f\big((C_n')^{-1};\alpha\big)$. This iteration is well-defined for C_n' in $\mathcal{B}$ satisfying (A.6) and hence for the iteration (A.5).

We first consider $\alpha \in (0,1)$. Then equation (A.7) leads to

$$C'_{n+1} \preceq \alpha^2 C'_n + \mathbb{I} \preceq \frac{1-\alpha^{2n+2}}{1-\alpha^2}\mathbb{I} + \alpha^{2n+2}C'_0 \preceq \frac{1}{1-\alpha^2}\mathbb{I} + \alpha^{2n+2}C'_0, \tag{A.9}$$

and hence there exists $\epsilon_0 \in (0, 1-\alpha)$ such that, for n is sufficiently large, we have

$$(C'_{n+1})^{-1} \succeq (1-\alpha^2-\epsilon_0)\mathbb{I}. \tag{A.10}$$

Let $\mathcal{M} \subset \mathcal{B}$ denote the set of matrices $B \in \mathcal{B}$ satisfying $B \succeq (1-\alpha^2-\epsilon_0)\mathbb{I}$. Then $\mathcal{M}$ is absorbing and forward invariant under f. Thus to show the existence of a globally exponentially attracting steady state it suffices to show that $f(\cdot;\alpha)$ is a contraction on $\mathcal{M}$.[9] The derivative of $f(\cdot;\alpha) : \mathcal{M} \mapsto \mathcal{M}$ is the element $Df(X;\alpha) \in \mathcal{L}$ defined by its action on $\Delta X \in \mathcal{B}$ as follows:

$$Df(X;\alpha)\Delta X = \alpha^2 (X+\alpha^2\mathbb{I})^{-1}\Delta X (X+\alpha^2\mathbb{I})^{-1}. \tag{A.11}$$

Thus

$$\begin{aligned}\|Df(X;\alpha)\Delta X\| =& \alpha^2 \left\|(X+\alpha^2\mathbb{I})^{-1}\Delta X(X+\alpha^2\mathbb{I})^{-1}\right\|.\\ \leq& \frac{\alpha^2}{(1-\epsilon_0)^2}\|\Delta X\|.\end{aligned}$$

Therefore, since $\alpha \in (0, 1-\epsilon_0)$,

$$\sup_{X\in\mathcal{M}} \|Df(X;\alpha)\|_{\mathcal{L}} < 1$$

and f is a contraction map on $\mathcal{M}$. This establishes the exponential convergence of $\{(C'_n)^{-1}\}$. Finally, the sequence $\{C_n^{-1}\}$ converges exponentially fast to C_∞^{-1}, the non-singular fixed point of equation (A.5); Equation (A.10) indicates that C_∞^{-1} is indeed non-singular.

When $\alpha = 1$ define mapping $f(X) = f(X;1)$ so that

$$(C'_{n+1})^{-1} = f\big((C'_n)^{-1}\big).$$

The derivative $Df(X) \in \mathcal{L}$ is defined by its action on $\Delta X \in \mathcal{B}$ as follows:

$$Df(X)\Delta X =(\mathbb{I}+X)^{-1}\Delta X(I+X)^{-1}. \tag{A.12}$$

Thus, using the lower bound from (A.6) and $\mathrm{Range}(G^T) = \mathbb{R}^{N_\theta}$,

$$\begin{aligned}\|Df(X)\Delta X\| \leq& \left\|(\mathbb{I}+X)^{-1}\right\|^2 \|\Delta X\|\\ \leq& \left\|\left(\mathbb{I}+\Sigma_\omega^{\frac{1}{2}}G^T\Sigma_\nu^{-1}G\Sigma_\omega^{\frac{1}{2}}\right)^{-1}\right\|^2 \|\Delta X\|\\ \leq&(1+\epsilon_1)^{-2}\|\Delta X\|,\end{aligned} \tag{A.13}$$

where $\epsilon_1 > 0$. Therefore, f is a contraction map on the whole of $\mathcal{B}$ and the sequence $\{C_n^{-1}\}$ converges. This completes the proof of exponential convergence of $\{C_n^{-1}\}$ to a limit; the sequence $\{C_n^{-1}\}$ converges to C_∞^{-1}, the fixed point of equation (A.5), viewed as a mapping on precision matrices. That $C_\infty \succ 0$ follows from (A.6). Because the convergence is global, the result also establishes the uniqueness of the steady state of equation (28).

We now prove that the mean $\{m_n\}$ converges exponentially fast to m_∞. Using (A.5) the update equation (27) of m_n can be rewritten as

$$m_{n+1} = \alpha(\mathbb{I} - C_{n+1}G^T\Sigma_\nu^{-1}G)m_n + C_{n+1}G^T\Sigma_\nu^{-1}y. \tag{A.14}$$

Thus convergence to m_∞ satisfying

$$m_\infty = \alpha(\mathbb{I} - C_\infty G^T\Sigma_\nu^{-1}G)m_\infty + C_\infty G^T\Sigma_\nu^{-1}y \tag{A.15}$$

[9] The use of contraction mapping arguments to study convergence of the Kalman filter is widespread, sometimes applied to the covariance and not the precision [20], and sometimes using Riemannian metric space structure on positive-definite matrices, rather than the vector space structure used here [21].

is determined by the spectral radius of $\alpha(\mathbb{I} - C_{n+1}G^T\Sigma_\nu^{-1}G)$. The matrix $\mathbb{I} - C_{n+1}G^T\Sigma_\nu^{-1}G$ has real spectrum; this may be established by showing the same for $\mathbb{I} - C_{n+1}(G^T\Sigma_\nu^{-1}G + \delta\mathbb{I})$, for $\delta > 0$, and letting $\delta \to 0$. If $\alpha \in (0,1)$, using equation (A.6), it follows that

$$\rho(\alpha\mathbb{I} - \alpha C_{n+1}G^T\Sigma_\nu^{-1}G) \leq \alpha < 1$$

and, by using a vector norm on $\mathbb{R}^{N_\theta}$ in which the induced operator norm on $\mathbb{I} - C_{n+1}G^T\Sigma_\nu^{-1}G$ is less than one, it follows that $\{m_n\}$ converges exponentially fast to m_∞. If $\alpha = 1$ then we use the fact that $B := G^T\Sigma_\nu^{-1}G$ is symmetric and that $B \succ 0$. From this it follows that $I - C_{n+1}B$ has the same spectrum as $I - B^{\frac{1}{2}}C_{n+1}B^{\frac{1}{2}}$. Using the upper bound on C_{n+1}^{-1} appearing in (A.6) we deduce that

$$\begin{aligned}\rho(\mathbb{I} - C_{n+1}B) &= \rho\Big(\mathbb{I} - B^{\frac{1}{2}}C_{n+1}B^{\frac{1}{2}}\Big)\\ &\leq 1 - \rho\Big(B^{\frac{1}{2}}\big(B + \Sigma_\omega^{-1}\big)^{-1}B^{\frac{1}{2}}\Big)\\ &= 1 - \epsilon_2,\end{aligned}$$

for some $\epsilon_2 \in (0,1)$. Since the spectral radius of $I - C_{n+1}B$ is less than one, there is again a norm on $\mathbb{R}^{N_\theta}$ in which the operator norm on $I - C_{n+1}B$ is less than one and exponential convergence follows. Equation (A.15) can be rewritten as

$$\begin{aligned}0 &= C_\infty\Big(G^T\Sigma_\nu^{-1}(y - Gm_\infty) + (1-\alpha)(G^T\Sigma_\nu^{-1}G - C_\infty^{-1})m_\infty\Big)\\ &= C_\infty\Big(G^T\Sigma_\nu^{-1}(y - Gm_\infty) - (1-\alpha)\widehat{C}_\infty^{-1}m_\infty\Big).\end{aligned}$$

Finally we note that m_∞ is the minimizer of equation (29). □

Proof of Theorem 2. In this setting where $\alpha = 1$ and $\Sigma_\omega = 0$ it follows from (A.5) that

$$\begin{aligned}C_{n+1}^{-1} &= G^T\Sigma_\nu^{-1}G + C_n^{-1},\\ C_{n+1}^{-1}m_{n+1} &= G^T\Sigma_\nu^{-1}y + C_n^{-1}m_n;\end{aligned}\tag{A.16}$$

so that

$$\begin{aligned}C_n^{-1} &= nG^T\Sigma_\nu^{-1}G + C_0^{-1},\\ C_n^{-1}m_n &= nG^T\Sigma_\nu^{-1}y + C_0^{-1}m_0.\end{aligned}\tag{A.17}$$

This demonstrates that if C_0 is positive definite so is C_n for all $n \in \mathbb{N}$. In the variables (34) we obtain

$$\begin{aligned}(C_n')^{-1} &= n(G')^T\Sigma_\nu^{-1}G' + I,\\ (C_n')^{-1}m_n' &= n(G')^T\Sigma_\nu^{-1}y + m_0'.\end{aligned}\tag{A.18}$$

This gives (35) and the proof is completed by applying the projections P and Q, noting that $PG' = G'$ and $QS = 0$, $QG' = 0$. □

Proof of Proposition 2. In this setting recall that we have $\alpha = 1$ and $\Sigma_\omega \succ 0$. The covariance update equation (27b) can be rewritten as

$$\begin{aligned}C_{n+1}^{-1} &= G^T\Sigma_\nu^{-1}G + (C_n + \Sigma_\omega)^{-1},\\ C_{n+1}^{-1}m_{n+1} &= G^T\Sigma_\nu^{-1}y + (C_n + \Sigma_\omega)^{-1}m_n.\end{aligned}\tag{A.19}$$

Since $\Sigma_\omega \succ 0$, the sequence $\{C_n^{-1}\}$ is bounded: $G^T\Sigma_\nu^{-1}G \preceq C_n^{-1} \preceq G^T\Sigma_\nu^{-1}G + \Sigma_\omega^{-1}$ and $C_n \succ 0$. Let us denote

$$C_n' = \Sigma_\omega^{-\frac{1}{2}}C_n\Sigma_\omega^{-\frac{1}{2}}, \quad m_n' = \Sigma_\omega^{-\frac{1}{2}}m_n, \quad G' = G\Sigma_\omega^{\frac{1}{2}}, \quad S = (G')^T\Sigma_\nu^{-1}G'.$$

First we prove the convergence of $\{C_n^{-1}\}$. Note that the update equation (A.19) becomes

$$(C_{n+1}')^{-1} = f\left((C_n')^{-1}\right)\tag{A.20}$$

where

$$f(X) = S + \left(X^{-1} + \mathbb{I}\right)^{-1}.$$

We note that the nullspace of S is equal to the nullspace of G'. Now consider the $\mathrm{Ker}(G') \otimes \mathrm{Range}(G'^T)$ decomposition of the vector space, and the corresponding orthogonal projections P and Q. Constraining on $\mathrm{Ker}(G')$, we have

$$(C'_{n+1})^{-1} = \big(C'_n + \mathbb{I}\big)^{-1} \prec (C'_n)^{-1}. \tag{A.21}$$

Since the sequence $\{(C'_n)^{-1}\}$ is strictly decreasing in the cone of positive-semidefinite matrices it must have limit 0. Therefore, we have $\lim\limits_{n\to\infty}(C'_n)^{-1} = 0$ on $\mathrm{Ker}(G')$. Constraining on $\mathrm{Range}({G'}^T)$, where $S \succ 0$, the update function (A.20) satisfies

$$\begin{aligned}\left\|\frac{df(X)}{dX}\Delta X\right\| &= \left\|(\mathbb{I}+X)^{-1}\Delta X(I+X)^{-1}\right\| \\ &\leq \left\|(\mathbb{I}+X)^{-1}\right\|^2 \|\Delta X\| \\ &\leq \left\|(\mathbb{I}+S)^{-1}\right\|^2 \|\Delta X\| \\ &\leq (1+\epsilon_1)^{-2}\|\Delta X\|,\end{aligned} \tag{A.22}$$

where $\epsilon_1 > 0$. Therefore, equation (A.20) is a contraction map on $\mathrm{Range}({G'}^T)$, which leads to the convergence of $(C'_n)^{-1}$ on that space. Combining the convergence of $(C'_n)^{-1}$ on both subspaces, we deduce that C_n^{-1} converges to a singular matrix. We conclude the analysis of the covariance by noting that Equation (A.19) leads to $C_{n+1} \preceq C_n + \Sigma_\omega$, which implies that $C_{n+1} \preceq C_0 + n\Sigma_\omega$ as required for (37).

Now we establish the convergence of $\{m'_n\}$. The update equation of m'_n can be rewritten as

$$m'_{n+1} = m'_n + C'_{n+1}{G'}^T\Sigma_\nu^{-1}y - C'_{n+1}Sm'_n. \tag{A.23}$$

Consider the $\mathrm{Range}({G'}^T)\otimes\mathrm{Ker}(G')$ decomposition $m'_n = Pm'_n + Qm'_n$, noting that in these coordinates the update equation can be rewritten as

$$Pm'_{n+1} = Pm'_n + PC'_{n+1}{G'}^T\Sigma_\nu^{-1}y - PC'_{n+1}SPm'_n, \tag{A.24a}$$
$$Qm'_{n+1} = Qm'_n + QC'_{n+1}{G'}^T\Sigma_\nu^{-1}y - QC'_{n+1}SPm'_n. \tag{A.24b}$$

Now consider the operator $P - PC'_{n+1}SP$ constrained to apply on $\mathrm{Range}({G'}^T)$. On this space $S\succ 0$ and, with $\mathbb{I} - S^{\frac{1}{2}}C'_{n+1}S^{\frac{1}{2}}$ also viewed as acting on $\mathrm{Range}({G'}^T)$,

$$\begin{aligned}\rho(P - PC'_{n+1}SP) &= \rho\Big(\mathbb{I} - S^{\frac{1}{2}}C'_{n+1}S^{\frac{1}{2}}\Big) \\ &\leq 1 - \rho\Big(S^{\frac{1}{2}}(S+\mathbb{I})^{-1}S^{\frac{1}{2}}\Big) \\ &= 1-\epsilon_0,\end{aligned} \tag{A.25}$$

where $\epsilon_0 \in (0,1)$. Hence, we deduce that $\{Pm'_n\}$ converges exponentially to $\theta_{ref} := S^+{G'}^T\Sigma_\nu^{-1}y$ where S^+ denotes the Moore-Penrose inverse of S. Since $SS^+ = \mathbb{I}$ on $\mathrm{Range}({G'}^T)$, the update equation (A.24b) for Qm'_n may be written as

$$Qm'_{n+1} = Qm'_n + QC'_{n+1}S(\theta_{ref} - Pm'_n). \tag{A.26}$$

It follows from (37) that $\|QC'_{n+1}S\|$ is bounded above by a function which grows linearly in n in any norm. Furthermore $\lim\limits_{n\to\infty} Pm'_n - \theta_{ref} = 0$ exponentially fast. Hence we deduce the exponential convergence of $\{Qm'_n\}$ to a limit, depending on Qm'_0. Therefore, $\{m_n\}$ converges exponentially fast to a stationary point of $\frac{1}{2}\|\Sigma_\nu^{-\frac{1}{2}}(y - G\theta)\|^2$. □

Proof of Lemma 2.

$$\begin{aligned}\frac{\partial\mathcal{F}\mathcal{G}(m,C)}{\partial m} &= \frac{\partial\mathbb{E}[\mathcal{G}(\theta)]}{\partial m} \\ &= \int \mathcal{G}(\theta)\frac{1}{\sqrt{(2\pi)^{N_\theta}|C|}}\exp\big(-\frac{1}{2}\|C^{-\frac{1}{2}}(\theta-m)\|^2\big)\big(C^{-1}(\theta-m)\big)^T d\theta \\ &= \int \mathcal{G}(\theta)(\theta-m)^T\frac{1}{\sqrt{(2\pi)^{N_\theta}|C|}}\exp\big(-\frac{1}{2}\|C^{-\frac{1}{2}}(\theta-m)\|^2\big)d\theta\cdot C^{-1} \\ &= \int \big(\mathcal{G}(\theta) - \mathbb{E}\mathcal{G}(\theta)\big)(\theta-m)^T\frac{1}{\sqrt{(2\pi)^{N_\theta}|C|}}\exp\big(-\frac{1}{2}\|C^{-\frac{1}{2}}(\theta-m)\|^2\big)d\theta\cdot C^{-1} \\ &= \mathcal{F}d\mathcal{G}(m,C). \quad \square\end{aligned} \tag{A.27}$$

Proof of Proposition 3. From equation (41) we have

$$\begin{aligned}\widehat{y}_{n+1} &= \mathcal{F}_u\mathcal{G}_{n+1}, \\ \widehat{C}^{\theta y}_{n+1} &= \widehat{C}_{n+1}\mathcal{F}_u d\mathcal{G}^T_{n+1}.\end{aligned} \tag{A.28}$$

In what follow we use the modified unscented transform Definition 1, and specifically its use to derive (20) and (21). First note that

$$\widehat{m}_{n+1} = \widehat{\theta}^0_{n+1}, \quad \widehat{y}_{n+1} = \mathcal{G}(\widehat{\theta}^0_{n+1}) = \widehat{y}^0_{n+1}, \quad \text{and} \quad w = W^c_1 = W^c_2 = \cdots = W^c_{2N_\theta}.$$

Now define the matrices

$$\begin{aligned}
\mathcal{Y}_1 &= [\widehat{y}^1_{n+1} - \widehat{y}_{n+1} \quad \widehat{y}^2_{n+1} - \widehat{y}_{n+1} \quad \cdots \quad \widehat{y}^{N_\theta}_{n+1} - \widehat{y}_{n+1}],\\
\mathcal{Y}_2 &= [\widehat{y}^{N_\theta+1}_{n+1} - \widehat{y}_{n+1} \quad \widehat{y}^{N_\theta+2}_{n+1} - \widehat{y}_{n+1} \quad \cdots \quad \widehat{y}^{2N_\theta}_{n+1} - \widehat{y}_{n+1}],\\
\Theta &= [\widehat{\theta}^1_{n+1} - \widehat{m}_{n+1} \quad \widehat{\theta}^2_{n+1} - \widehat{m}_{n+1} \quad \cdots \quad \widehat{\theta}^{N_\theta}_{n+1} - \widehat{m}_{n+1}].
\end{aligned}$$

Then we have

$$\widehat{C}^{\theta y}_{n+1} = \sum_{j=1}^{2N_\theta} W^c_j (\widehat{\theta}^j_{n+1} - \widehat{m}_{n+1})(\widehat{y}^j_{n+1} - \widehat{y}_{n+1})^T = w\Theta(\mathcal{Y}_1^T - \mathcal{Y}_2^T), \tag{A.29a}$$

$$\widehat{C}^{yy}_{n+1} = \sum_{j=1}^{2N_\theta} W^c_j (\widehat{y}^j_{n+1} - \widehat{y}_{n+1})(\widehat{y}^j_{n+1} - \widehat{y}_{n+1})^T + \Sigma_\nu = w(\mathcal{Y}_1\mathcal{Y}_1^T + \mathcal{Y}_2\mathcal{Y}_2^T) + \Sigma_\nu, \tag{A.29b}$$

$$\widehat{C}_{n+1} = \sum_{j=1}^{2N_\theta} W^c_j (\widehat{\theta}^j_{n+1} - \widehat{m}_{n+1})(\widehat{\theta}^j_{n+1} - \widehat{m}_{n+1})^T = 2w\Theta\Theta^T. \tag{A.29c}$$

Equation (A.29c) follows from the definition of the sigma points (20). Since $\widehat{C}_{n+1} \succeq \Sigma_\omega \succ 0$, the matrix $\Theta \in \mathbb{R}^{N_\theta \times N_\theta}$ is non-singular. Thus we have

$$\begin{aligned}
\mathcal{F}_u d\mathcal{G}_{n+1}\widehat{C}_{n+1}\mathcal{F}_u d\mathcal{G}_{n+1}^T &= \widehat{C}^{\theta y\ T}_{n+1}\widehat{C}^{-1}_{n+1}\widehat{C}_{n+1}\widehat{C}^{-1}_{n+1}\widehat{C}^{\theta y}_{n+1}\\
&= \widehat{C}^{\theta y\ T}_{n+1}\widehat{C}^{-1}_{n+1}\widehat{C}^{\theta y}_{n+1}\\
&= w(\mathcal{Y}_1 - \mathcal{Y}_2)\Theta^T\left(2w\Theta\Theta^T\right)^{-1}\Theta(\mathcal{Y}_1^T - \mathcal{Y}_2^T)w\\
&= \frac{w}{2}(\mathcal{Y}_1\mathcal{Y}_1^T + \mathcal{Y}_2\mathcal{Y}_2^T - \mathcal{Y}_1\mathcal{Y}_2^T - \mathcal{Y}_2\mathcal{Y}_1^T).
\end{aligned} \tag{A.30}$$

Using equation (A.30) in equation (A.29b) yields

$$\widehat{C}^{yy}_{n+1} = \mathcal{F}_u d\mathcal{G}_{n+1}\widehat{C}_{n+1}\mathcal{F}_u d\mathcal{G}_{n+1}^T + \Sigma_\nu + \widetilde{\Sigma}_{\nu,n+1}, \tag{A.31}$$

where

$$\widetilde{\Sigma}_{\nu,n+1} := \frac{w}{2}(\mathcal{Y}_1 + \mathcal{Y}_2)(\mathcal{Y}_1 + \mathcal{Y}_2)^T.$$

We note that $\widetilde{\Sigma}_{\nu,n+1}$ is positive semi-definite. Furthermore, the i-th column of $\mathcal{Y}_1 + \mathcal{Y}_2$ satisfies

$$\begin{aligned}
\widehat{y}^i_{n+1} + \widehat{y}^{i+N_\theta}_{n+1} - 2\widehat{y}_{n+1} &= \mathcal{G}(\widehat{m}_{n+1} + c_i[\sqrt{\widehat{C}_{n+1}}]_j) + \mathcal{G}(\widehat{m}_{n+1} - c_i[\sqrt{\widehat{C}_{n+1}}]_j) - 2\mathcal{G}(\widehat{m}_{n+1})\\
&\approx \frac{d^2\mathcal{G}(\widehat{m}_{n+1})}{d^2\theta} : [\sqrt{\widehat{C}_{n+1}}]_j \otimes [\sqrt{\widehat{C}_{n+1}}]_j.
\end{aligned} \tag{A.32}$$

Hence $\widetilde{\Sigma}_{\nu,n+1} = 0$ when $\mathcal{G}$ is linear; otherwise $\|\widetilde{\Sigma}_{\nu,n+1}\| = \mathcal{O}(\|\widehat{C}^2_{n+1}\|)$, a second order term with small covariance $\widehat{C}_{n+1}$. □

Proof of Lemma 3. If the steady state C of equation (44b) is singular, then $\exists v \in R^{N_\theta}$ s.t. $v^T C v = 0$. We have

$$\begin{aligned}
\left(v^T C^{\theta y} u\right)^2 &= \left(\mathbb{E}[v^T(\theta - m) \otimes (\mathcal{G}(\theta) - \mathcal{G}(m))u]\right)^2\\
&\leq \mathbb{E}[v^T(\theta - m) \otimes (\theta - m)v]\mathbb{E}[u^T(\mathcal{G}(\theta) - \mathcal{G}(m)) \otimes (\mathcal{G}(\theta) - \mathcal{G}(m))u]\\
&= 0,
\end{aligned}$$

for any $u \in R^{N_y}$. This implies that $v^T C^{\theta y} = 0$, and therefore,

$$-2\alpha_0 v^T C v - v^T C^{\theta y}\Sigma_\nu{}^{-1}C^{\theta y^T} v = 0,$$

which contradicts the assumption that $\Sigma_\omega \succ 0$. □

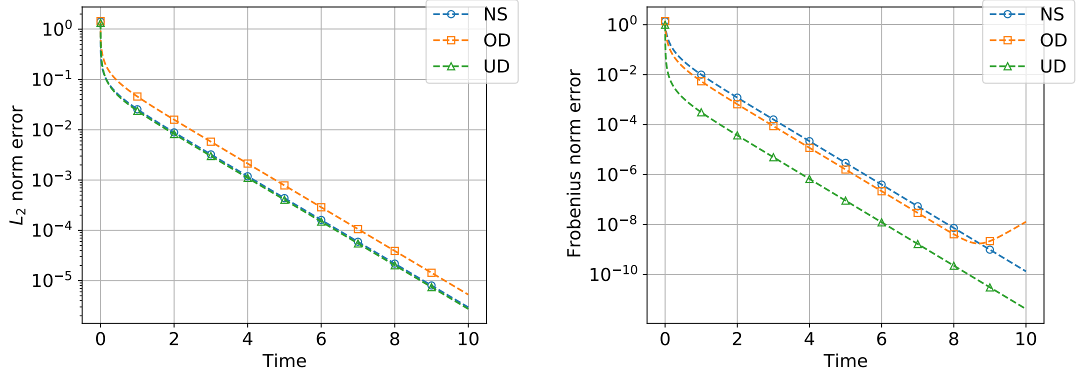


Fig. B.23. L_2 error $\|m_n - m_{ref}\|_2$ (left) and Frobenius norm $\|C_n - C_{ref}\|_F$ (right) obtained by UKS for non-singular (NS), over-determined (OD), and under-determined (UD) systems of the linear 2-parameter model problem.

Appendix B. Illustrative examples for UKS

The primary focus of the paper is on using the UKI for optimization purposes. However the basic ingredients of the method, and the dynamical system (46) in particular, can also be used to perform approximate posterior sampling from the measure μ given by (4). In the case where μ is Gaussian, the posterior is exactly captured by the steady state of these equations; when the posterior is not Gaussian, then only an approximation is obtained. To illustrate the UKS, we consider, in Subsection Appendix B.1, application to three linear inverse problems from Subsection 3.1, for which the posterior is Gaussian if the prior is Gaussian; and then give a simple example of application to a non-Gaussian posterior in Subsection Appendix B.2.

The UKS equations (46) can be discretized by the following semi-implicit scheme

$$\begin{aligned} m_{n+1} - m_n &= h\Big(C^{\theta y}\Sigma_\eta^{-1}\big(y - \mathbb{E}\mathcal{G}(\theta)\big) - C\Sigma_0^{-1}(m_{n+1} - r_0)\Big),\\ C_{n+1} - C_n &= h\Big(-2C^{\theta y}\Sigma_\eta^{-1}C^{\theta y^T} - 2C_n\Sigma_0^{-1}C_n + 2C_{n+1}\Big), \end{aligned} \tag{B.1}$$

with a fixed time-step. The integrals defining $C^{\theta y}$ and $\mathbb{E}\mathcal{G}(\theta)$ are explicitly approximated by the modified unscented transform (see Definition 1) using the Gaussian $\mathcal{N}(m_n, C_n)$. Integration could also be performed using an adaptive time-step, as in [62]; however more work is needed to develop efficient methods stemming from the UKS as formulated here.

B.1. Linear 2-parameter model problem

The linear 2-parameter model problems discussed in Section 5.3 are used with prior

$$r_0 = 0 \quad \text{and} \quad \Sigma_0 = \mathbb{I}.$$

Therefore, the posterior distribution is $\mu \sim \mathcal{N}(m_{ref}, C_{ref})$, where

$$m_{ref} = \Big(\Sigma_0^{-1} + G^T\Sigma_\eta^{-1}G\Big)^{-1}\Big(G^T\Sigma_\eta^{-1}y + \Sigma_0^{-1}r_0\Big) \quad \text{and} \quad C_{ref} = \Big(\Sigma_0^{-1} + G^T\Sigma_\eta^{-1}G\Big)^{-1}. \tag{B.2}$$

The UKS is initialized with $\theta_0 \sim \mathcal{N}(r_0, \Sigma_0)$. The convergence of the UKS, in terms of the posterior mean and covariance errors for $t \in [0, 10]$, is reported in Fig. B.23. Both mean and covariance converge to the posterior mean and covariance. However, even with the semi-implicit scheme the maximum time step that allows for stable simulation is $h = 5 \times 10^{-5}$.

B.2. Nonlinear 2-parameter model problem

The following Bayesian logistic regression problem is considered,

$$y = \frac{1}{1 + \exp(\theta_{(1)} + \theta_{(2)}x)} + \eta.$$

Here $N_\theta = 2$ and $N_y = 1$, and hence this is an under-determined problem. The prior distribution $\mathcal{N}(r_0, \Sigma_0)$ satisfies

$$r_0 = [1 \quad 1]^T \quad \text{and} \quad \Sigma_0 = \mathbb{I}.$$

The observation data $y_{ref} = 0.08$ is generated at $x = \frac{1}{2}$, with observation error $\eta \sim \mathcal{N}(0, 0.1^2)$ and $\theta_{ref} = [2 \quad 2]^T$.

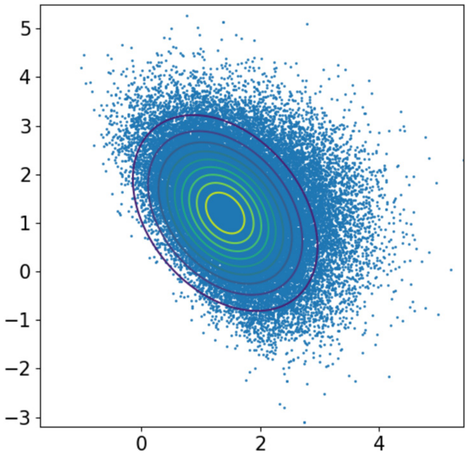


Fig. B.24. Contour plot: posterior distributions obtained by UKS at $t = 10$; blue dots: reference posterior distribution obtained by MCMC for the nonlinear 2-parameter model problem. x-axis is for $\theta_{(1)}$ and y-axis is for $\theta_{(2)}$.

The UKS is initialized with $\theta_0 \sim \mathcal{N}(r_0, \Sigma_0)$. The posterior distributions obtained by the UKS at $t = 10$ with a time step $h = 5 \times 10^{-5}$ and Markov chain Monte Carlo method (MCMC) with a step size 1.0 and 5×10^6 samples (with a 10^6 sample burn-in period) are presented in Fig. B.24. The estimated posterior distributions are in reasonably good agreement, but of course not as accurate as in the linear setting in the previous subsection, because of a Gaussian approximation being made to a non-Gaussian distribution. Specifically, the posterior mean and covariance estimated by the UKS are

$$[1.41 \quad 1.20]^T \quad \text{and} \quad \begin{bmatrix} 0.526 & -0.235 \\ -0.235 & 0.884 \end{bmatrix},$$

whilst the posterior mean and covariance estimated by the MCMC are

$$[1.62 \quad 1.31]^T \quad \text{and} \quad \begin{bmatrix} 0.619 & -0.254 \\ -0.254 & 1.00 \end{bmatrix}.$$

References

[1] Marco A. Iglesias, Kody J.H. Law, Andrew M. Stuart, Ensemble Kalman methods for inverse problems, Inverse Probl. 29 (4) (2013) 045001.

[2] Marco A. Iglesias, A regularizing iterative ensemble Kalman method for pde-constrained inverse problems, Inverse Probl. 32 (2) (2016) 025002.

[3] Marco Iglesias, Yuchen Yang, Adaptive regularisation for ensemble Kalman inversion, Inverse Probl. 37 (2) (2021) 025008.

[4] Neil K. Chada, Andrew M. Stuart, Xin T. Tong, Tikhonov regularization within ensemble Kalman inversion, SIAM J. Numer. Anal. 58 (2) (2020) 1263–1294.

[5] Claudia Schillings, Andrew M. Stuart, Analysis of the ensemble Kalman filter for inverse problems, SIAM J. Numer. Anal. 55 (3) (2017) 1264–1290.

[6] Alfredo Garbuno-Inigo, Franca Hoffmann, Wuchen Li, Andrew M. Stuart, Interacting Langevin diffusions: gradient structure and ensemble Kalman sampler, SIAM J. Appl. Dyn. Syst. 19 (1) (2020) 412–441.

[7] Alfredo Garbuno-Inigo, Nikolas Nüsken, Sebastian Reich, Affine invariant interacting Langevin dynamics for Bayesian inference, SIAM J. Appl. Dyn. Syst. 19 (3) (2020) 1633–1658.

[8] Zhiyan Ding, Qin Li, Jianfeng Lu, Ensemble Kalman inversion for nonlinear problems: weights, consistency, and variance bounds, arXiv preprint, arXiv:2003.02316, 2020.

[9] Daniel Zhengyu Huang, Jiaoyang Huang, Sebastian Reich, Andrew M. Stuart, Efficient derivative-free Bayesian inference for large-scale inverse problems, arXiv preprint, arXiv:2204.04386, 2022.

[10] Sebastian Reich, Simon Weissmann, Fokker–Planck particle systems for Bayesian inference: computational approaches, SIAM/ASA J. Uncertain. Quantificat. 9 (2) (2021) 446–482.

[11] Zhiyan Ding, Qin Li, Ensemble Kalman inversion: mean-field limit and convergence analysis, Stat. Comput. 31 (1) (2021) 1–21.

[12] Zhiyan Ding, Qin Li, Ensemble Kalman sampler: mean-field limit and convergence analysis, SIAM J. Math. Anal. 53 (2) (2021) 1546–1578.

[13] Heinz Werner Engl, Martin Hanke, Andreas Neubauer, Regularization of Inverse Problems, vol. 375, Springer Science & Business Media, 1996.

[14] Jari Kaipio, Erkki Somersalo, Statistical and Computational Inverse Problems, vol. 160, Springer Science & Business Media, 2006.

[15] Masoumeh Dashti, Andrew M. Stuart, The Bayesian approach to inverse problems, arXiv preprint, arXiv:1302.6989, 2013.

[16] Pierre Del Moral, Arnaud Doucet, Ajay Jasra, Sequential Monte Carlo samplers, J. R. Stat. Soc., Ser. B, Stat. Methodol. 68 (3) (2006) 411–436.

[17] Nicolas Chopin, Omiros Papaspiliopoulos, et al., An Introduction to Sequential Monte Carlo, vol. 4, Springer, 2020.

[18] Alexandros Beskos, Ajay Jasra, Ege A. Muzaffer, Andrew M. Stuart, Sequential Monte Carlo methods for Bayesian elliptic inverse problems, Stat. Comput. 25 (4) (2015) 727–737.

[19] Sebastian Reich, A dynamical systems framework for intermittent data assimilation, BIT Numer. Math. 51 (1) (2011) 235–249.

[20] Peter Lancaster, Leiba Rodman, Algebraic Riccati Equations, Clarendon Press, 1995.

[21] Philippe Bougerol, Kalman filtering with random coefficients and contractions, SIAM J. Control Optim. 31 (4) (1993) 942–959.

[22] Rudolph Emil Kalman, A new approach to linear filtering and prediction problems, J. Basic Eng. 82 (1) (Mar 1960) 35–45.

[23] Harold Wayne Sorenson, Kalman Filtering: Theory and Application, IEEE, 1985.

[24] Andrew H. Jazwinski, Stochastic Processes and Filtering Theory, Courier Corporation, 2007.
[25] Michael Ghil, S. Cohn, John Tavantzis, K. Bube, Eugene Isaacson, Applications of estimation theory to numerical weather prediction, in: Dynamic Meteorology: Data Assimilation Methods, Springer, 1981, pp. 139–224.
[26] Geir Evensen, Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics, J. Geophys. Res., Oceans 99 (C5) (1994) 10143–10162.
[27] Yan Chen, Dean S. Oliver, Ensemble randomized maximum likelihood method as an iterative ensemble smoother, Math. Geosci. 44 (1) (2012) 1–26.
[28] Alexandre A. Emerick, Albert C. Reynolds, Investigation of the sampling performance of ensemble-based methods with a simple reservoir model, Comput. Geosci. 17 (2) (2013) 325–350.
[29] Martin Hanke, A regularizing Levenberg-Marquardt scheme, with applications to inverse groundwater filtration problems, Inverse Probl. 13 (1) (1997) 79.
[30] Nikolas Nüsken, Sebastian Reich, Note on interacting Langevin diffusions: gradient structure and ensemble Kalman sampler by Garbuno-Inigo, Hoffmann, Li and Stuart, arXiv preprint, arXiv:1908.10890, 2019.
[31] Kody JH Law, Andrew M. Stuart, Evaluating data assimilation algorithms, Mon. Weather Rev. 140 (11) (2012) 3757–3782.
[32] Oliver G. Ernst, Björn Sprungk, Hans-Jörg Starkloff, Analysis of the ensemble and polynomial chaos Kalman filters in Bayesian inverse problems, SIAM/ASA J. Uncertain. Quantificat. 3 (1) (2015) 823–851.
[33] G.A. Pavliotis, A.M. Stuart, U. Vaes, Derivative-free Bayesian inversion using multiscale dynamics, arXiv preprint, arXiv:2102.00540, 2021.
[34] Simon J. Julier, Jeffrey K. Uhlmann, Hugh F. Durrant-Whyte, A new approach for filtering nonlinear systems, in: Proceedings of 1995 American Control Conference-ACC'95, vol. 3, IEEE, 1995, pp. 1628–1632.
[35] Eric A. Wan, Rudolph Van Der Merwe, The unscented Kalman filter for nonlinear estimation, in: Proceedings of the IEEE 2000 Adaptive Systems for Signal Processing, Communications, and Control Symposium (Cat. No. 00EX373), Ieee, 2000, pp. 153–158.
[36] Mrinal K. Sen, Paul L. Stoffa, Global Optimization Methods in Geophysical Inversion, Cambridge University Press, 2013.
[37] Tapio Schneider, Shiwei Lan, Andrew Stuart, Joao Teixeira, Earth system modeling 2.0: a blueprint for models that learn from observations and targeted high-resolution simulations, Geophys. Res. Lett. 44 (24) (2017) 12–396.
[38] Oliver R.A. Dunbar, Alfredo Garbuno-Inigo, Tapio Schneider, Andrew M. Stuart, Calibration and uncertainty quantification of convective parameters in an idealized gcm, arXiv preprint, arXiv:2012.13262, 2020.
[39] Daniel Z. Huang, Kailai Xu, Charbel Farhat, Eric Darve, Learning constitutive relations from indirect observations using deep neural networks, J. Comput. Phys. (2020) 109491.
[40] Kailai Xu, Daniel Z. Huang, Eric Darve, Learning constitutive relations using symmetric positive definite neural networks, J. Comput. Phys. 428 (2020) 110072.
[41] Philip Avery, Daniel Z. Huang, Wanli He, Johanna Ehlers, Armen Derkevorkian, Charbel Farhat, A computationally tractable framework for nonlinear dynamic multiscale modeling of membrane fabric, arXiv preprint, arXiv:2007.05877, 2020.
[42] Brian H. Russell, Introduction to Seismic Inversion Methods, SEG Books, 1988.
[43] Carey Bunks, Fatimetou M. Saleck, S. Zaleski, G. Chavent, Multiscale seismic waveform inversion, Geophysics 60 (5) (1995) 1457–1473.
[44] Johannes Töger, Matthew J. Zahr, Nicolas Aristokleous, Karin Markenroth Bloch, Marcus Carlsson, Per-Olof Persson, Blood flow imaging by optimal matching of computational fluid dynamics to 4d-flow data, Magn. Reson. Med. (2020).
[45] Flávio Celso Trigo, Raul Gonzalez-Lima, Marcelo Britto Passos Amato, Electrical impedance tomography using the extended Kalman filter, IEEE Trans. Biomed. Eng. 51 (1) (2004) 72–81.
[46] Dan Simon, Optimal State Estimation: Kalman, H Infinity, and Nonlinear Approaches, John Wiley & Sons, 2006.
[47] François Auger, Mickael Hilairet, Josep M. Guerrero, Eric Monmasson, Teresa Orlowska-Kowalska, Seiichiro Katsura, Industrial applications of the Kalman filter: a review, IEEE Trans. Ind. Electron. 60 (12) (2013) 5458–5471.
[48] Huazhen Fang, Ning Tian, Yebin Wang, MengChu Zhou, Mulugeta A. Haile, Nonlinear Bayesian estimation: from Kalman filtering to a broader horizon, IEEE/CAA J. Autom. Sin. 5 (2) (2018) 401–417.
[49] Sharad Singhal, Lance Wu, Training multilayer perceptrons with the extended Kalman algorithm, in: Advances in Neural Information Processing Systems, 1989, pp. 133–140.
[50] Gintaras V. Puskorius, Lee A. Feldkamp, Decoupled extended Kalman filter training of feedforward layered networks, in: IJCNN-91-Seattle International Joint Conference on Neural Networks, vol. 1, IEEE, 1991, pp. 771–777.
[51] Zhengyu Huang, Philip Avery, Charbel Farhat, Jason Rabinovitch, Armen Derkevorkian, Lee D. Peterson, Simulation of parachute inflation dynamics using an Eulerian computational framework for fluid-structure interfaces evolving in high-speed turbulent flows, in: 2018 AIAA Aerospace Sciences Meeting, 2018, p. 1540.
[52] Daniel Z. Huang, P-O. Persson, Matthew J. Zahr, High-order, linearly stable, partitioned solvers for general multiphysics problems based on implicit–explicit Runge–Kutta schemes, Comput. Methods Appl. Mech. Eng. 346 (2019) 674–706.
[53] Daniel Z. Huang, Philip Avery, Charbel Farhat, Jason Rabinovitch, Armen Derkevorkian, Lee D. Peterson, Modeling, simulation and validation of supersonic parachute inflation dynamics during Mars landing, in: AIAA Scitech 2020 Forum, 2020, p. 0313.
[54] Daniel Z. Huang, Will Pazner, Per-Olof Persson, Matthew J. Zahr, High-order partitioned spectral deferred correction solvers for multiphysics problems, J. Comput. Phys. (2020) 109441.
[55] Alistair Adcroft, Whit Anderson, V. Balaji, Chris Blanton, Mitchell Bushuk, Carolina O. Dufour, John P. Dunne, Stephen M. Griffies, Robert Hallberg, Matthew J. Harrison, et al., The gfdl global ocean and sea ice model om4. 0: model description and simulation features, J. Adv. Model. Earth Syst. 11 (10) (2019) 3167–3211.
[56] Charles S. Peskin, Numerical analysis of blood flow in the heart, J. Comput. Phys. 25 (3) (1977) 220–252.
[57] Marsha Berger, Michael Aftosmis, Progress towards a Cartesian cut-cell method for viscous compressible flow, in: 50th AIAA Aerospace Sciences Meeting Including the New Horizons Forum and Aerospace Exposition, 2012, p. 1301.
[58] Daniel Z. Huang, Dante De Santis, Charbel Farhat, A family of position- and orientation-independent embedded boundary methods for viscous flow and fluid–structure interaction problems, J. Comput. Phys. 365 (2018) 74–104.
[59] Daniel Z. Huang, Philip Avery, Charbel Farhat, An embedded boundary approach for resolving the contribution of cable subsystems to fully coupled fluid-structure interaction, Int. J. Numer. Methods Eng. (2020).
[60] Marsha J. Berger, Phillip Colella, et al., Local adaptive mesh refinement for shock hydrodynamics, J. Comput. Phys. 82 (1) (1989) 64–84.
[61] Raunak Borker, Daniel Huang, Sebastian Grimberg, Charbel Farhat, Philip Avery, Jason Rabinovitch, Mesh adaptation framework for embedded boundary methods for computational fluid dynamics and fluid-structure interaction, Int. J. Numer. Methods Fluids 90 (8) (2019) 389–424.
[62] Emmet Cleary, Alfredo Garbuno-Inigo, Shiwei Lan, Tapio Schneider, Andrew M. Stuart, Calibrate, emulate, sample, J. Comput. Phys. 424 (2020) 109716.
[63] Daniel J. Lea, Myles R. Allen, Thomas W.N. Haine, Sensitivity analysis of the climate of a chaotic system, Tellus, Ser. A Dyn. Meteorol. Oceanogr. 52 (5) (2000) 523–532.
[64] Qiqi Wang, Rui Hu, Patrick Blonigan, Least squares shadowing sensitivity analysis of chaotic limit cycle oscillations, J. Comput. Phys. 267 (2014) 210–224.
[65] Nikola B. Kovachki, Andrew M. Stuart, Ensemble Kalman inversion: a derivative-free technique for machine learning tasks, Inverse Probl. 35 (9) (2019) 095005.

[66] Dean S. Oliver, Albert C. Reynolds, Ning Liu, Inverse Theory for Petroleum Reservoir Characterization and History Matching, Cambridge University Press, 2008.

[67] Eric A. Wan, Alex T. Nelson, Neural dual extended Kalman filtering: applications in speech enhancement and monaural blind signal separation, in: Neural Networks for Signal Processing VII. Proceedings of the 1997 IEEE Signal Processing Society Workshop, IEEE, 1997, pp. 466–475.

[68] Alexander G. Parlos, Sunil K. Menon, A. Atiya, An algorithmic approach to adaptive state filtering using recurrent neural networks, IEEE Trans. Neural Netw. 12 (6) (2001) 1411–1432.

[69] J.H. Gove, D.Y. Hollinger, Application of a dual unscented Kalman filter for simultaneous state and parameter estimation in problems of surface-atmosphere exchange, J. Geophys. Res., Atmos. 111 (D8) (2006).

[70] David J. Albers, Matthew Levine, Bruce Gluckman, Henry Ginsberg, George Hripcsak, Lena Mamykina, Personalized glucose forecasting for type 2 diabetes using data assimilation, PLoS Comput. Biol. 13 (4) (2017) e1005232.

[71] Kay Bergemann, Sebastian Reich, An ensemble Kalman-Bucy filter for continuous data assimilation, Meteorol. Z. 21 (3) (2012) 213–219.

[72] Claudia Schillings, Andrew M. Stuart, Convergence analysis of ensemble Kalman inversion: the linear, noisy case, Appl. Anal. 97 (1) (2018) 107–123.

[73] Zhiyan Ding, Qin Li, Ensemble Kalman inversion: mean-field limit and convergence analysis, arXiv preprint, arXiv:1908.05575, 2019.

[74] Bradley M. Bell, Frederick W. Cathey, The iterated Kalman filter update as a Gauss-Newton method, IEEE Trans. Autom. Control 38 (2) (1993) 294–297.

[75] Neil K. Chada, Xin T. Tong, Convergence acceleration of ensemble Kalman inversion in nonlinear settings, arXiv preprint, arXiv:1911.02424, 2019.

[76] Neil K. Chada, Yuming Chen, Daniel Sanz-Alonso, Iterative ensemble Kalman methods: a unified perspective with some new variants, arXiv preprint, arXiv:2010.13299, 2020.

[77] José A. Carrillo, Young-Pil Choi, Claudia Totzeck, Oliver Tse, An analytical framework for consensus-based global optimization method, Math. Models Methods Appl. Sci. 28 (06) (2018) 1037–1066.

[78] J.A. Carrillo, F. Hoffmann, A.M. Stuart, U. Vaes, Consensus-based sampling, Stud. Appl. Math. 148 (3) (2022) 1069–1140.

[79] Sebastian Reich, Colin Cotter, Probabilistic Forecasting and Bayesian Data Assimilation, Cambridge University Press, 2015.

[80] Kody Law, Andrew Stuart, Kostas Zygalakis, Data Assimilation, Springer, Cham, Switzerland, 2015.

[81] Daniel Sanz-Alonso, Andrew M. Stuart, Armeen Taeb, Inverse problems and data assimilation, arXiv preprint, arXiv:1810.06191, 2018.

[82] Jonathan Goodman, Jonathan Weare, Ensemble samplers with affine invariance, Commun. Appl. Math. Comput. Sci. 5 (1) (2010) 65–80.

[83] John A. Nelder, Roger Mead, A simplex method for function minimization, Comput. J. 7 (4) (1965) 308–313.

[84] Benedict Leimkuhler, Charles Matthews, Jonathan Weare, Ensemble preconditioning for Markov chain Monte Carlo simulation, Stat. Comput. 28 (2) (2018) 277–290.

[85] Simon Julier, Jeffrey Uhlmann, Hugh F. Durrant-Whyte, A new method for the nonlinear transformation of means and covariances in filters and estimators, IEEE Trans. Autom. Control 45 (3) (2000) 477–482.

[86] Michael K. Tippett, Jeffrey L. Anderson, Craig H. Bishop, Thomas M. Hamill, Jeffrey S. Whitaker, Ensemble square root filters, Mon. Weather Rev. 131 (7) (2003) 1485–1490.

[87] Lehel Csató, Manfred Opper, Sparse on-line Gaussian processes, Neural Comput. 14 (3) (2002) 641–668.

[88] David J. Albers, Paul-Adrien Blancquart, Matthew E. Levine, Elnaz Esmaeilzadeh Seylabi, Andrew Stuart, Ensemble Kalman methods with constraints, Inverse Probl. 35 (9) (2019) 095007.

[89] J.A. Carrillo, U. Vaes, Wasserstein stability estimates for covariance-preconditioned Fokker–Planck equations, arXiv preprint, arXiv:1910.07555, 2019.

[90] Lassi Roininen, Janne M.J. Huttunen, Sari Lasanen, Whittle-Matérn priors for Bayesian statistical inversion with applications in electrical impedance tomography, Inverse Probl. Imaging 8 (2) (2014) 561.

[91] Marco A. Iglesias, Kody J.H. Law, Andrew M. Stuart, Evaluation of Gaussian approximations for data assimilation in reservoir models, Comput. Geosci. 17 (5) (2013) 851–885.

[92] Bernhard Beckermann, The condition number of real Vandermonde, Krylov and positive definite Hankel matrices, Numer. Math. 85 (4) (2000) 553–577.

[93] Matthew M. Dunlop, Marco A. Iglesias, Andrew M. Stuart, Hierarchical Bayesian level set inversion, Stat. Comput. 27 (6) (2017) 1555–1584.

[94] Nicholas H. Nelsen, Andrew M. Stuart, The random feature model for input-output maps between Banach spaces, arXiv preprint, arXiv:2005.10224, 2020.

[95] Jan S. Hesthaven, Sigal Gottlieb, David Gottlieb, Spectral Methods for Time-Dependent Problems, vol. 21, Cambridge University Press, 2007.

[96] Steven A. Orszag, G.S. Patterson Jr., Numerical simulation of three-dimensional homogeneous isotropic turbulence, Phys. Rev. Lett. 28 (2) (1972) 76.

[97] Edward N. Lorenz, Deterministic nonperiodic flow, in: The Theory of Chaotic Attractors, Springer, 2004, pp. 25–36.

[98] Wael Bahsoun, Ian Melbourne, Marks Ruziboev, Variance Continuity for Lorenz Flows, Annales Henri Poincare, vol. 21, Springer, 2020, pp. 1873–1892.

[99] Jan Frøyland, Knut H. Alfsen, Lyapunov-exponent spectra for the Lorenz model, Phys. Rev. A 29 (5) (1984) 2928.

[100] Edward N. Lorenz, Predictability: a problem partly solved, in: Proc. Seminar on Predictability, vol. 1, 1996.

[101] Ibrahim Fatkullin, Eric Vanden-Eijnden, A computational strategy for multiscale systems with applications to Lorenz 96 model, J. Comput. Phys. 200 (2) (2004) 605–638.

[102] Daniel S. Wilks, Effects of stochastic parametrizations in the Lorenz'96 system, Q. J. R. Meteorol. Soc., J. Atmos. Sci. Appl. Meteorol. Phys. Oceanogr. 131 (606) (2005) 389–407.

[103] H.M. Arnold, I.M. Moroz, T.N. Palmer, Stochastic parametrizations and model uncertainty in the Lorenz'96 system, Philos. Trans. R. Soc. A, Math. Phys. Eng. Sci. 371 (1991) (2013) 20110479.

[104] Georg A. Gottwald, Sebastian Reich, Supervised learning from noisy observations: combining machine-learning techniques with data assimilation, arXiv preprint, arXiv:2007.07383, 2020.

[105] Isaac M. Held, Max J. Suarez, A proposal for the intercomparison of the dynamical cores of atmospheric general circulation models, Bull. Am. Meteorol. Soc. 75 (10) (1994) 1825–1830.