Foundations of Data Science
doi:10.3934/fods.2021011

ITERATIVE ENSEMBLE KALMAN METHODS: A UNIFIED PERSPECTIVE WITH SOME NEW VARIANTS

Neil K. Chada

Applied Mathematics and Computational Science
King Abdullah University of Science and Technology
Thuwal, 23955, Saudi Arabia

Yuming Chen and Daniel Sanz-Alonso*

Department of Statistics
University of Chicago
Chicago, IL 60637, USA

(Communicated by Elaine Spiller)

Abstract. This paper provides a unified perspective of iterative ensemble Kalman methods, a family of derivative-free algorithms for parameter reconstruction and other related tasks. We identify, compare and develop three subfamilies of ensemble methods that differ in the objective they seek to minimize and the derivative-based optimization scheme they approximate through the ensemble. Our work emphasizes two principles for the derivation and analysis of iterative ensemble Kalman methods: statistical linearization and continuum limits. Following these guiding principles, we introduce new iterative ensemble Kalman methods that show promising numerical performance in Bayesian inverse problems, data assimilation and machine learning tasks.

1. **Introduction.** This paper provides an accessible introduction to the derivation and foundations of iterative ensemble Kalman methods, a family of derivative-free algorithms for parameter reconstruction and other related tasks. The overarching theme behind these methods is to iteratively update via Kalman-type formulae an ensemble of candidate reconstructions, aiming to bring the ensemble closer to the unknown parameter with each iteration. The ensemble Kalman updates approximate derivative-based nonlinear least-squares optimization schemes without requiring gradient evaluations. Our presentation emphasizes that iterative ensemble Kalman methods can be naturally classified in terms of the nonlinear least-squares objective they seek to minimize and the derivative-based optimization scheme they approximate through the ensemble. This perspective allows us to identify three subfamilies of iterative ensemble Kalman methods, creating unity into the growing literature on this subject. Our work also emphasizes two principles for the derivation and analysis of iterative ensemble Kalman methods: statistical linearization

2020 *Mathematics Subject Classification.* Primary: 37C10, 60G35; Secondary: 65N20.

Key words and phrases. Ensemble Kalman methods, inverse problems, statistical linearization, continuum limits.

NKC is supported by KAUST baseline funding. DSA is thankful for the support of NSF and NGA through the grant DMS-2027056. The work of DSA was also partially supported by the NSF Grant DMS-1912818/1912802.

* Corresponding author: Daniel Sanz-Alonso.

and continuum limits. Following these principles we introduce new iterative ensemble Kalman methods that show promising numerical performance in Bayesian inverse problems, data assimilation and machine learning tasks.

We consider the application of iterative ensemble Kalman methods to the problem of reconstructing an unknown $u \in \mathbf{R}^d$ from corrupt data $y \in \mathbf{R}^k$ related by

$$y = h(u) + \eta, \tag{1.1}$$

where η represents measurement or model error and h is a given map. A wide range of inverse problems, data assimilation and machine learning tasks can be cast into the framework (1.1). In these applications the unknown u may represent, for instance, an input parameter of a differential equation, the current state of a time-evolving signal and a regressor, respectively. Ensemble Kalman methods were first introduced as filtering schemes for sequential data assimilation [20, 21, 45, 51, 54] to reduce the computational cost of the Kalman filter [34]. Their use for state and parameter estimation and inverse problems was further developed in [4, 43, 47, 59]. The idea of *iterating* these methods was considered in [15, 18]. Iterative ensemble Kalman methods are now popular in inverse problems and data assimilation; they have also shown some potential in machine learning applications [26, 25, 38].

Starting from an initial ensemble $\{u_0^{(n)}\}_{n=1}^N$, iterative ensemble Kalman methods use various ensemble-based empirical means and covariances to update

$$\{u_i^{(n)}\}_{n=1}^N \to \{u_{i+1}^{(n)}\}_{n=1}^N, \tag{1.2}$$

until a stopping criteria is satisfied; the unknown parameter u is reconstructed by the mean of the final ensemble. The idea is analogous to classical Kalman methods and optimization schemes which, starting with a *single* initialization u_0 use evaluations of derivatives of h to iteratively update $u_i \to u_{i+1}$ until a stopping criteria is met. The initial ensemble is viewed as an input to the algorithm, obtained in a problem-dependent fashion. In Bayesian inverse problems and machine learning it may be obtained by sampling a prior, while in data assimilation the initial ensemble may be a given collection of particles that approximates the prediction distribution. In either case, it is useful to view the initial ensemble as a sample from a probability distribution. It is important to note that there is no time variable involved in the reconstruction task (1.1); however, we will often think of the iteration index $i \in \mathbb{N}$ in (1.2) as an artificial time index, since this allows us to interpret the evolution of iterates as arising from discretization of differential equations, and thereby to gain theoretical understanding.

There are two main computational benefits in updating an *ensemble* of candidate reconstructions rather than a single estimate. First, the ensemble update can be performed without evaluating derivatives of h, effectively approximating them using *statistical linearization*. This is important in applications where computing derivatives of h is expensive, or where the map h needs to be treated as a black-box. Second, the use of empirical rather than model covariances can significantly reduce the computational cost whenever the ensemble size N is smaller than the dimension d of the unknown parameter u. A further advantage of the ensemble approach is that, for problems that are not strongly nonlinear, the spread of the ensemble may contain meaningful information on the uncertainty in the reconstruction.

1.1. **Overview: Three subfamilies.**

This paper identifies, compares and further develops three subfamilies of iterative ensemble Kalman methods to implement the ensemble update (1.2). Each subfamily employs a different Kalman-type formulae,

Objective	Optimization	Derivative Method	Ensemble Method	New Variant
J_{TP}	GN	IExKF (2.1)	IEKF (3.1)	IEKF-SL (4.1)
J_{DM}	LM	LM-DM (2.2)	EKI (3.2)	EKI-SL (4.2)
J_{TP}	LM	LM-TP (2.3)	TEKI (3.3)	TEKI-SL (4.3)

TABLE 1. Roadmap to the algorithms considered in this paper. We use the abbreviations GN and LM for Gauss-Newton and Levenberg-Marquardt. The numbers in parenthesis represent the subsection in which each algorithm is introduced.

determined by a choice of objective to minimize and a derivative-based optimization scheme to approximate with the ensemble. All three approaches impose some form of regularization, either explicitly through the choice of the objective, or implicitly through the choice of the optimization scheme. Incorporating regularization is essential in parameter reconstruction problems encountered in applications, which are typically under-determined or ill-posed [44, 54].

The first subfamily considers a *Tikhonov-Phillips* objective associated with the parameter reconstruction problem (1.2), given by

$$\mathsf{J}_{\text{TP}}(u) := \frac{1}{2}|y - h(u)|_R^2 + \frac{1}{2}|u - m|_P^2, \tag{1.3}$$

where R and P are symmetric positive definite matrices that model, respectively, the data measurement precision and the level of regularization —incorporated explicitly through the choice of objective— and m represents a background estimate of u. Here and throughout this paper we use the notation $|v|_A^2 := |A^{-1/2}v|^2 = v^T A^{-1} v$ for symmetric positive definite A and vector v. The ensemble is used to approximate a Gauss-Newton method applied to the Tikhonov-Phillips objective J_{TP}. Algorithms in this subfamily were first introduced in geophysical data assimilation [1, 15, 18, 24, 41, 52] and were inspired by iterative, derivative-based, extended Kalman filters [5, 6, 33]. Extensions to more challenging problems with strongly nonlinear dynamics are considered in [53, 63]. In this paper we will use a new Iterative Ensemble Kalman Filter (IEKF) method as a prototypical example of an algorithm that belongs to this subfamily.

The second subfamily considers the *data-misfit* objective

$$\mathsf{J}_{\text{DM}}(u) := \frac{1}{2}|y - h(u)|_R^2. \tag{1.4}$$

When the parameter reconstruction problem is ill-posed, minimizing J_{DM} leads to unstable reconstructions. For this reason, iterative ensemble Kalman methods in this subfamily are complemented with a Levenberg-Marquardt optimization scheme that implicitly incorporates regularization. The ensemble is used to approximate a regularizing Levenberg-Marquardt optimization algorithm to minimize J_{DM}. Algorithms in this subfamily were introduced in the applied mathematics literature [29, 30] building on ideas from classical inverse problems [27]. Recent theoretical work has focused on developing continuous-time and mean-field limits, as well as various convergence results [9, 8, 14, 28, 38, 55]. Methodological extensions based on Bayesian hierarchical techniques were introduced in [10, 11] and the incorporation of constraints has been investigated in [2, 12]. In this paper we will use the Ensemble Kalman Inversion (EKI) method [30] as a prototypical example of an algorithm that belongs to this subfamily.

The third subfamily, which has emerged more recently, combines explicit regularization through the Tikhonov-Phillips objective and an implicitly regularizing optimization scheme [14, 13]. Precisely, a Levenberg-Marquardt scheme is approximated through the ensemble in order to minimize the Tikhonov-Phillips objective J_{TP}. In this paper we will use the Tikhonov ensemble Kalman inversion method (TEKI) [13] as a prototypical example.

To conclude this overview we note that while in this paper we will only consider least-squares objectives, iterative ensemble Kalman methods that use other regularizers have been recently proposed [38, 40, 57]. As well as this, we will restrict our attention to ensemble methods and their similarities with derivative-based methods. Iterative variants of other data assimilation methods such as 3DVAR and 4DVAR may be of interest [42, 46, 54], but are outside the scope of this paper.

1.2. **Statistical linearization, continuum limits and new variants.** Each subfamily of iterative ensemble Kalman methods stems from a derivative-based optimization scheme. However, there is substantial freedom as to how to use the ensemble to approximate a derivative-based method. We will focus on randomized-maximum likelihood implementations [24, 37, 54], rather than square-root or ensemble adjustment approaches [4, 62, 23]. Two principles will guide our derivation and analysis of ensemble methods: the use of statistical linearization [63] and their connection with gradient descent methods through the study of continuum limits [55].

The idea behind statistical linearization is to approximate the gradient of h using pairs $\left\{\left(u_i^{(n)}, h(u_i^{(n)})\right)\right\}_{n=1}^N$ in such a way that if h is linear and the ensemble size N is sufficiently large, the approximation is exact. As we shall see, this idea tacitly underlies the derivation of all the ensemble methods considered in this paper, and will be explicitly employed in our derivation of new variants. Statistical linearization has also been used within Unscented Kalman filters, see e.g. [63].

Differential equations have long been important in developing and understanding optimization schemes [48], and investigating the connections between differential equations and optimization is still an active area of research [58, 60, 64]. In the context of iterative ensemble methods, continuum limit analyses arise from considering small length-steps and have been developed primarily in the context of EKI-type algorithms [9, 56]. While the derivative-based algorithms that motivate the ensemble methods result in an ODE continuum limit, the ensemble versions lead to a system of SDEs. Continuum limit analyses are useful in at least three ways. First, they unveil the gradient structure of the optimization schemes. Second, viewing optimization schemes as arising from discretizations of SDEs lends itself to design of algorithms that are easy to tune: the length-step is chosen to be small and the algorithms are run until statistical equilibrium is reached. Third, a simple linear-case analysis of the SDEs may be used to develop new algorithms that satisfy certain desirable properties. Our new iterative ensemble Kalman methods will be designed following these observations.

While our work advocates the study of continuum limits as a useful tool to design ensemble methods, continuum limits cannot fully capture the full richness and flexibility of discrete-based implementable algorithms, since different algorithms may result in the same SDE continuum limit. This insight suggests that it is not only the study of differential equations, but also their *discretizations*, that may contribute to the design of iterative ensemble Kalman algorithms.

1.3. **Main contributions and outline.** In addition to providing a unified perspective of the existing literature, this paper contains several original contributions. We highlight some of them in the following outline and refer to Table 1 for a summary of the algorithms considered in this paper.

- In Section 2 we review three iterative derivative-based methods for nonlinear least-squares optimization. The ensemble-based algorithms studied in subsequent sections can be interpreted as ensemble-based approximations of the derivative-based methods described in this section. We also derive informally ODE continuum limits for each method, which unveils their gradient flow structure.
- In Section 3 we describe the idea of statistical linearization. We review three subfamilies of iterative ensemble methods, each of which has an update formula analogous to one of the derivate-based methods in Section 2. We analyze the methods when $h(u) = Hu$ is linear by formally deriving SDE continuum limits that unveil their gradient structure. A novelty in this section is the introduction of the IEKF method, which is similar to, but different from, the iterative ensemble method introduced in [63].
- The material in Section 4 is novel to the best of our knowledge. We introduce new variants of the iterative ensemble Kalman methods discussed in Section 3 and formally derive their SDE continuum limit. We analyze the resulting SDEs when $h(u) = Hu$ is linear. The proposed methods are designed to ensure that (i) no parameter tuning or careful stopping criteria are needed; and (ii) the ensemble covariance contains meaningful information of the uncertainty in the reconstruction in the linear case, avoiding the ensemble collapse of some existing methods.
- In Section 5 we include an in-depth empirical comparison of the performance of the iterative ensemble Kalman methods discussed in Sections 3 and 4. We consider four problem settings motivated by applications in Bayesian inverse problems, data assimilation and machine learning. Our results illustrate the different behavior of some methods in small noise regimes and the benefits of avoiding ensemble collapse.
- Section 6 concludes and suggests some open directions for further research.

2. **Derivative-based optimization for nonlinear least-squares.** In this section we review the Gauss-Newton and Levenberg-Marquardt methods, two derivative-based approaches for optimization of nonlinear least-squares objectives of the form

$$\mathsf{J}(u) = \frac{1}{2}|r(u)|^2. \tag{2.1}$$

We derive closed formulae for the Gauss-Newton method applied to the Tikhonov-Phillips objective J_{TP}, as well as for the Levenberg-Marquardt method applied to the data-misfit objective J_{DM} and the Tikhonov-Phillips objective J_{TP}. These formulae are the basis for the ensemble, derivative-free methods considered in the next section.

As we shall see, the search directions of Gauss-Newton and Levenberg-Marquardt methods are found by minimizing a linearization of the least-squares objective. It is thus instructive to consider first linear least-squares optimization before delving into the nonlinear setting. The following well-known result, that we will use extensively, characterizes the minimizer μ of the Tikhonov-Phillips objective J_{TP} in the case of linear $h(u) = Hu$.

Lemma 2.1. *It holds that*

$$\frac{1}{2}|y - Hu|_R^2 + \frac{1}{2}|u - m|_P^2 = \frac{1}{2}|u - \mu|_C^2 + \beta, \tag{2.2}$$

where β does not depend on u, and

$$C^{-1} = H^T R^{-1} H + P^{-1}, \tag{2.3}$$

$$C^{-1}\mu = H^T R^{-1} y + P^{-1} m. \tag{2.4}$$

Equivalently,

$$\mu = m + K(y - Hm), \tag{2.5}$$

$$C = (I - KH)P, \tag{2.6}$$

where K is the Kalman gain matrix given by

$$K = PH^T(HPH^T + R)^{-1} = CH^T R^{-1}. \tag{2.7}$$

Proof. The formulae (2.3) and (2.4) follows by matching linear and quadratic coefficients in u between

$$\frac{1}{2}|u - \mu|_C^2 \qquad \text{and} \qquad \frac{1}{2}|u - m|_P^2 + \frac{1}{2}|y - h(u)|_R^2. \tag{2.8}$$

The formulae (2.5) and (2.6) as well as the equivalent expressions for the Kalman gain K in Equation (2.7) can be obtained using the matrix inversion lemma [54]. □

Bayesian Interpretation. Lemma 2.1 has a natural statistical interpretation. Consider a statistical model defined by likelihood $y|u \sim \mathcal{N}(Hu, R)$ and prior $u \sim \mathcal{N}(m, P)$. Then Equation (2.2) shows that the posterior distribution is Gaussian, $u|y \sim \mathcal{N}(\mu, C)$, and Equations (2.3)-(2.4) characterize the posterior mean and precision (inverse covariance). We interpret Equation (2.5) as providing a closed formula for the posterior *mode*, known as the *maximum a posteriori* (MAP) estimator.

More generally, the generative model

$$\begin{aligned} u &\sim \mathcal{N}(m, P), \\ y|u &\sim \mathcal{N}(h(u), R), \end{aligned} \tag{2.9}$$

gives rise to a posterior distribution on $u|y$ with density proportional to $\exp\big(-\mathsf{J}_{\mathrm{TP}}(u)\big)$. Thus, minimization of $\mathsf{J}_{\mathrm{TP}}(u)$ corresponds to maximizing the posterior density under the model (2.9).

2.1. **Gauss-Newton optimization of Tikhonov-Phillips objective.** In this subsection we introduce two ways of writing the Gauss-Newton update applied to the Tikhonov-Phillips objective J_{TP}. We recall that the Gauss-Newton method applied to a general least-squares objective $\mathsf{J}(u) = \frac{1}{2}|r(u)|^2$ is a line-search method which, starting from an initialization u_0, sets

$$u_{i+1} = u_i + \alpha_i v_i, \qquad i = 0, 1, \dots$$

where v_i is a search direction defined by

$$v_i = \arg\min_v \mathsf{J}_i^\ell(v), \qquad \mathsf{J}_i^\ell(v) := \frac{1}{2}|r'(u_i)v + r(u_i)|^2, \tag{2.10}$$

where $\alpha_i > 0$ is a length-step parameter whose choice will be discussed later. In order to apply the Gauss-Newton method to the Tikhonov-Phillips objective, we write J_{TP} in the standard nonlinear least-squares form (2.1). Note that

$$\begin{aligned}\mathsf{J}_{\mathrm{TP}}(u) &= \frac{1}{2}|u-m|_P^2 + \frac{1}{2}|y-h(u)|_R^2 \\ &= \frac{1}{2}|z-g(u)|_Q^2,\end{aligned}$$

where

$$z := \begin{bmatrix} y \\ m \end{bmatrix}, \qquad g(u) := \begin{bmatrix} h(u) \\ u \end{bmatrix}, \qquad Q := \begin{bmatrix} R & 0 \\ 0 & P \end{bmatrix}. \tag{2.11}$$

Therefore we have

$$\mathsf{J}_{\mathrm{TP}}(u) = \frac{1}{2}|r_{\mathrm{TP}}(u)|^2, \qquad r_{\mathrm{TP}}(u) := Q^{-1/2}\big(z - g(u)\big). \tag{2.12}$$

The following result is a direct consequence of Lemma 2.1.

Lemma 2.2 ([6]). *The Gauss-Newton method applied to the Tikhonov-Phillips objective J_{TP} admits the characterizations:*

$$u_{i+1} = u_i + \alpha_i C_i\Big\{H_i^T R^{-1}\big(y - h(u_i)\big) + P^{-1}(m - u_i)\Big\}, \tag{2.13}$$

and

$$u_{i+1} = u_i + \alpha_i\Big\{K_i\big(y - h(u_i)\big) + (I - K_i H_i)(m - u_i)\Big\}, \tag{2.14}$$

where $H_i = h'(u_i)$ and

$$\begin{aligned}K_i &= PH_i^T(H_iPH_i^T + R)^{-1}, \\ C_i &= (I - K_iH_i)P.\end{aligned}$$

Proof. The search direction v_i of Gauss-Newton for the objective J_{TP} is given by

$$v_i = \arg\min_v \mathsf{J}_{\mathrm{TP},i}^{\ell}(v) \tag{2.15}$$

$$= \arg\min_v \frac{1}{2}\big|r_{\mathrm{TP}}'(u_i)v + r_{\mathrm{TP}}(u_i)\big|^2 \tag{2.16}$$

$$= \arg\min_v \frac{1}{2}\big|z - g(u_i) - g'(u_i)v\big|_Q \tag{2.17}$$

$$= \arg\min_v \left\{\frac{1}{2}\big|y - h(u_i) - h'(u_i)v\big|_R^2 + \frac{1}{2}\big|v - (m - u_i)\big|_P^2\right\}. \tag{2.18}$$

Applying Lemma 2.1, using formulae (2.4) and (2.6), we deduce that

$$v_i = C_i\Big\{H_i^T R^{-1}\big(y - h(u_i)\big) + P^{-1}(m - u_i)\Big\},$$

which establishes the characterization (2.13). The equivalence between (2.13) and (2.14) follows from the identity (2.7), which implies that $C_iH_i^TR^{-1} = K_i$ and $C_iP^{-1} = I - K_iH_i$. □

We refer to the Gauss-Newton method with constant length-step $\alpha_i = \alpha$ applied to J_{TP} as the Iterative Extended Kalman Filter (IExKF) algorithm. IExKF was developed in the control theory literature [33] without reference to the Gauss-Newton optimization method; the agreement between both methods was established in [6]. In order to compare IExKF with an ensemble-based method in Section 3, we summarize it here.

Algorithm 1 Iterative Extended Kalman Filter (IExKF)

1: **Input**: Initialization $u_0 = m$, length-step α.
2: For $i = 0, 1, \dots$ do:

Set $K_i = PH_i^T(H_iPH_i^T + R)^{-1}, \qquad H_i = h'(u_i)$.
Set $C_i = (I - K_iH_i)P$.
Set

$$u_{i+1} = u_i + \alpha\Big\{K_i\big(y - h(u_i)\big) + (I - K_iH_i)(m - u_i)\Big\}, \tag{2.19}$$

or, equivalently,

$$u_{i+1} = u_i + \alpha C_i\Big\{H_i^TR^{-1}(y - h(u_i)) + P^{-1}(m - u_i)\Big\}, \tag{2.20}$$

3: **Output**: $u_1, u_2, \dots$

The next proposition shows that in the linear case, if α is set to 1, IExKF finds the minimizer of the objective (1.3) in one iteration, and further iterations still agree with the minimizer.

Proposition 1. *Suppose that $h(u) = Hu$ is linear and $\alpha = 1$. Then the output of Algorithm 1 satisfies*

$$u_i = \mu, \qquad i = 1, 2, \dots$$

where μ is the minimizer of the Tikhonov-Phillips objective (1.3).

Proof. In the linear case we have

$$H_i = H, \qquad K_i = K = PH^T(HPH^T + R)^{-1}, \qquad i = 0, 1, \dots$$

Therefore, update (2.19) simplifies as

$$u_{i+1} = m + K(y - Hm), \quad i = 0, 1, \dots$$

This implies that, for all $i \geq 1$, it holds that $u_i = \mu$ with μ defined in Equation (2.5). □

Choice of Length-Step. When implementing Gauss-Newton methods, it is standard practice to perform a line search in the direction of v_i to adaptively choose the length-step α_i. For instance, a common strategy is to guarantee that the Wolfe conditions are satisfied [16, 49]. In this paper we will instead simply set $\alpha_i = \alpha$ for some fixed value of α, and we will follow a similar approach for all the derivative-based and ensemble-based algorithms we consider. There are two main motivations for doing so. First, it is appealing from a practical viewpoint to avoid performing a line search for ensemble-based algorithms. Second, when α is small each derivative-based algorithms we consider can be interpreted as a discretization of an ODE system, while the ensemble-based methods arise as discretizations of SDE systems. These ODEs and SDEs allow us to compare and gain transparent understanding of the gradient structure of the algorithms. They will also allow us to propose some new variants of existing ensemble Kalman methods. We next describe the continuum limit structure of IExKF.

Continuum Limit. It is not hard to check that the term in brackets in the update (2.20)

$$H_i^T R^{-1}(y - h(u_i)) + P^{-1}(m - u_i)$$

is the negative gradient of $\mathsf{J}_{\mathrm{TP}}(u)$, which reveals the following gradient flow structure in the limit of small length-step α :

$$\begin{aligned} \dot{u} &= C(t)\Big\{h'\big(u(t)\big)^T R^{-1}\big(y - h\big(u(t)\big)\big) + P^{-1}\big(m - u(t)\big)\Big\} \\ &= -C(t)\mathsf{J}'_{\mathrm{TP}}\big(u(t)\big), \end{aligned} \tag{2.21}$$

with preconditioner

$$C(t) := \left(h'\big(u(t)\big)^T R^{-1} h'\big(u(t)\big) + P^{-1}\right)^{-1}.$$

We remark that in the linear case, $C(t) \equiv C$, where C is the posterior covariance given by (2.6), which agrees with the inverse of the Hessian of the Tikhonov Phillips objective.

2.2. **Levenberg-Marquardt optimization of data-misfit objective.** In this subsection we introduce the Levenberg-Marquardt algorithm and describe its application to the data misfit objective J_{DM}. We recall that the Levenberg-Marquardt method applied to a general least-squares objective $\mathsf{J}(u) = \frac{1}{2}|r(u)|^2$ is a trust region method which, starting from an initialization u_0, sets

$$u_{i+1} = u_i + v_i, \qquad i = 0, 1, \dots$$

where

$$v_i = \arg\min_v \mathsf{J}_i^\ell(v), \quad \text{s.t. } |v|_P^2 \le \delta_i, \qquad \mathsf{J}_i^\ell(v) := \frac{1}{2}|r'(u_i)v + r(u_i)|^2.$$

Similar to Gauss-Newton methods, the increment v_i is defined as the minimizer of a linearized objective, but now the minimization is constrained to a ball $\{|v|_P^2 \le \delta_i\}$ in which we *trust* that the objective can be replaced by its linearization. The increment can also be written as

$$v_i = \arg\min_v \mathsf{J}_i^{\mathrm{UC}}(v),$$

where

$$\mathsf{J}_i^{\mathrm{UC}}(v) = \mathsf{J}_i^\ell(v) + \frac{1}{2\alpha_i}|v|_P^2. \tag{2.22}$$

The parameter $\alpha_i > 0$ plays an analogous role to the length-step in Gauss-Newton methods. Note that the Levenberg-Marquardt increment is the unconstrained minimizer of a *regularized* objective. It is for this reason that we say that Levenberg-Marquardt provides an implicit regularization.

We next consider application of the Levenberg-Marquardt method to the data-misfit objective J_{DM}, which we write in standard nonlinear least-squares form:

$$\mathsf{J}_{\mathrm{DM}}(u) = \frac{1}{2}|r_{\mathrm{DM}}(u)|^2, \qquad r_{\mathrm{DM}}(u) := R^{-1/2}\big(y - h(u)\big). \tag{2.23}$$

Lemma 2.3. *The Levenberg-Marquardt method applied to the data misfit objective* J_{DM} *admits the following characterization:*

$$u_{i+1} = u_i + K_i\Big\{y - h(u_i)\Big\}, \tag{2.24}$$

where

$$K_i = \alpha_i P H_i^T(\alpha_i H_i P H_i^T + R)^{-1}, \qquad H_i = h'(u_i).$$

Proof. Note that the increment v_i is defined as the unconstrained minimizer of

$$\begin{aligned} \mathsf{J}^{\mathrm{UC}}_{\mathrm{DM},i}(v) &= \frac{1}{2}|r'_{\mathrm{DM}}(u_i)v + r_{\mathrm{DM}}(u_i)|^2 + \frac{1}{2\alpha_i}|v|^2_P \\ &= \frac{1}{2}|y - h(u_i) - h'(u_i)v|^2_R + \frac{1}{2\alpha_i}|v|^2_P. \end{aligned} \tag{2.25}$$

The result follows from Lemma 2.1. □

Similar to the previous section, we will focus on implementations with constant length-step $\alpha_i = \alpha$, which leads to the following algorithm.

Algorithm 2 Iterative Levenberg-Marquardt with Data Misfit (ILM-DM)

1: **Input**: Initialization $u_0 = m$, length-step α.
2: For $i = 0, 1, \dots$ do:

Set $K_i = \alpha P H_i^T (\alpha H_i P H_i^T + R)^{-1}, \qquad H_i = h'(u_i).$
Set

$$u_{i+1} = u_i + K_i\Big\{y - h(u_i)\Big\}. \tag{2.26}$$

3: **Output**: $u_1, u_2, \dots$

When $\alpha = 1$, the following linear-case result shows that ILM-DM, i.e. Algorithm 2, reaches the minimizer of J_{TP} in one iteration. However, in contrast to IExKF, further iterations of ILM-DM will typically worsen the optimization of J_{TP}, and start moving towards minimizers of J_{DM}.

Proposition 2. *Suppose that $h(u) = Hu$ is linear and $\alpha = 1$. Then the output of Algorithm 2 satisfies*

$$u_1 = \arg\min_u \mathsf{J}_{\mathrm{TP}}(u),$$

where J_{TP} is the Tikhonov-Phillips objective (1.3).

Proof. The proof is identical to that of Proposition 1, noting that in the linear case $u_{i+1} = u_i + K(y - Hu_i)$. □

Example 2.1 (Convergence of ILM-DM with invertible observation map)**.** Suppose that $H \in \mathbf{R}^{d\times d}$ is invertible and $\alpha = 1$. Then, writing

$$u_{i+1} = (I - KH)u_i + Ky$$

and noting that $\rho(I - KH) < 1$ [3], it follows that $u_i \to u^*$, where u^* is the unique solution to $y = Hu$. That is, the iterates of ILM-DM converge to the unique minimizer of the data misfit objective J_{DM}.

Choice of Length-Step. When implementing Levenberg-Marquardt algorithms, the length-step parameter α_i is often chosen adaptively, based on the objective. However, similar to Section 2.1, we fix $\alpha_i = \alpha$ to be a small value which leads to an ODE continuum limit.

Continuum Limit. We notice that, in the limit of small length-step α, update (2.26) can be written as

$$u_{i+1} = u_i + \alpha P H_i^T R^{-1}\Big\{y - h(u_i)\Big\}.$$

The term $H_i^T R^{-1}\Big\{y - h(u_i)\Big\}$ is the negative gradient of $\mathsf{J}_{\mathrm{DM}}(u)$, which reveals the following gradient flow structure

$$\begin{aligned}\dot{u} &= Ph'\big(u(t)\big)^T R^{-1}\Big\{y - h\big(u(t)\big)\Big\} \\ &= -P\mathsf{J}'_{\mathrm{DM}}(u),\end{aligned} \tag{2.27}$$

where the preconditioner P is interpreted as the prior covariance in the Bayesian framework.

2.3. **Levenberg-Marquardt optimization of Tikhonov-Phillips objective.** In this subsection we describe the application of the Levenberg-Marquardt algorithm to the Tikhonov-Phillips objective J_{TP}.

Lemma 2.4. *The Levenberg-Marquardt method applied to the Tikhonov-Phillips objective J_{TP} admits the following characterization:*

$$u_{i+1} = u_i + K_i\Big\{z - g(u_i)\Big\},$$

where

$$K_i = \alpha_i P G_i^T(\alpha_i G_i P G_i^T + Q)^{-1}, \qquad G_i = g'(u_i),$$

recalling that g is defined in (2.11).

Proof. Note that the increment v_i is defined as the unconstrained minimizer of

$$\mathsf{J}^{\mathrm{UC}}_{\mathrm{TP},i}(v) = \mathsf{J}^{\ell}_{\mathrm{TP},i}(v) + \frac{1}{2\alpha_i}|v|_P^2 \tag{2.28}$$

$$= \frac{1}{2}|z - g(u_i) - g'(u_i)v|_Q^2 + \frac{1}{2\alpha_i}|v|_P^2, \tag{2.29}$$

which has the same form as Equation (2.25) replacing y with z, h with g, and R with Q. □

Setting $\alpha_i = \alpha$ leads to the following algorithm.

Algorithm 3 Iterative Levenberg-Marquardt with Tikhonov-Phillips (ILM-TP)

1: **Input**: Initialization $u_0 = m$, length-step α.
2: For $i = 0, 1, \ldots$ do:

Set $K_i = \alpha P G_i^T(\alpha G_i P G_i^T + Q)^{-1}, \qquad G_i = g'(u_i).$
Set

$$u_{i+1} = u_i + K_i\Big\{z - g(u_i)\Big\}. \tag{2.30}$$

3: **Output**: $u_1, u_2, \ldots$

Proposition 3. *Suppose that $h(u) = Hu$ is linear and $\alpha = 1$. The output of Algorithm 3 satisfies*

$$u_1 = \arg\min_u \Big\{\mathsf{J}_{\mathrm{TP}}(u) + \frac{1}{2}|u - m|^2\Big\}. \tag{2.31}$$

Objective	Optimization	Derivative Method	Ensemble Method
J_{TP}	GN	IExKF	IEKF
J_{DM}	LM	ILM-DM	EKI
J_{TP}	LM	ILM-TP	TEKI

TABLE 2. Summary of the main algorithms in Sections 2 and 3.

Proof. The result is a corollary of Proposition 2. To see this, note that $\mathsf{J}_{\mathrm{TP}}(u) = \frac{1}{2}|z - g(u)|_Q^2$ can be viewed as a data-misfit objective with data z, forward model $g(u)$ and observation matrix Q. Then, ILM-TP can be interpreted as applying ILM-DM to J_{TP}, and the objective $\mathsf{J}_{\mathrm{TP}}(u) + \frac{1}{2}|u - m|^2$ as its Tikhonov-Phillips regularization. □

Example 2.2 (Convergence of ILM-TP with linear invertible observation map.)**.** It is again instructive to consider the case where $H \in \mathbf{R}^{d\times d}$ is invertible. Then, following the same reasoning as in Example 2.1 we deduce that the iterates of ILM-TP converge to the unique minimizer of J_{TP}.

Continuum Limit. In the limit of small length-step α, we can derive the gradient flow structure of update (2.30). This is similar to Section 2.2, with J_{DM} replaced by J_{TP}. Precisely, the gradient flow of ILM-TP is given by

$$\begin{aligned}\dot{u} &= Pg'\big(u(t)\big)^T Q^{-1}\Big\{z - g\big(u(t)\big)\Big\}\\ &= -P\mathsf{J}'_{\mathrm{TP}}(u).\end{aligned}$$

3. **Ensemble-based optimization for nonlinear least-squares.** In this section we review three subfamilies of iterative methods that update an ensemble $\{u_i^{(n)}\}_{n=1}^N$ employing Kalman-based formulae, where $i = 0, 1, \ldots$ denotes the iteration index and N is a fixed ensemble size. Each ensemble member $u_i^{(n)}$ is updated by optimizing a (random) objective $\mathsf{J}_i^{(n)}$ defined using the current ensemble $\{u_i^{(n)}\}_{n=1}^N$ and/or the initial ensemble $\{u_0^{(n)}\}_{n=1}^N$. The optimization is performed without evaluating derivatives by invoking a *statistical linearization* of a Gauss-Newton or Levenberg-Marquardt algorithm. In analogy with the previous section, the three subfamilies of ensemble methods we consider differ in the choice of the objective and in the choice of the optimization algorithm. Table 2 summarizes the derivative and ensemble methods considered in the previous and the current section.

Given an ensemble $\{u_i^{(n)}\}_{n=1}^N$ we use the following notation for ensemble empirical means

$$m_i = \frac{1}{N}\sum_{n=1}^N u_i^{(n)}, \qquad h_i = \frac{1}{N}\sum_{n=1}^N h(u_i^{(n)}),$$

and empirical covariances

$$P_i^{uu} = \frac{1}{N}\sum_{n=1}^{N}(u_i^{(n)} - m_i)(u_i^{(n)} - m_i)^T,$$
$$P_i^{uy} = \frac{1}{N}\sum_{n=1}^{N}\big(u_i^{(n)} - m_i\big)\big(h(u_i^{(n)}) - h_i\big)^T,$$
$$P_i^{yy} = \frac{1}{N}\sum_{n=1}^{N}\big(h(u_i^{(n)}) - h_i\big)\big(h(u_i^{(n)}) - h_i\big)^T.$$

Two overarching themes that underlie the derivation and analysis of the ensemble methods studied in this section and the following one are the use of a statistical linearization to avoid evaluation of derivatives, and the study of continuum limits. We next introduce these two ideas.

Statistical Linearization. If $h(u) = Hu$ is linear, we have

$$P_i^{uy} = P_i^{uu}H^T,$$

which motivates the approximation in the general nonlinear case

$$h'(u_i^{(n)}) \approx (P_i^{uy})^T(P_i^{uu})^{-1} =: H_i, \qquad n = 1, \dots, N, \tag{3.1}$$

where here and in what follows $(P_i^{uu})^{-1}$ denotes the pseudoinverse of P_i^{uu}. Notice that (3.1) can be regarded as a linear least-squares fit of pairs $\big\{\big(u_i^{(n)}, h(u_i^{(n)})\big)\big\}_{n=1}^N$ normalized around their corresponding empirical means m_i and h_i. We remark that in order for the approximation in (3.1) to be accurate, the ensemble size N should not be much smaller than the input dimension d.

Continuum Limit. We will gain theoretical understanding by studying continuum limits. Specifically, each algorithm includes a length-step parameter $\alpha > 0$, and the evolution of the ensemble for small α can be interpreted as a discretization of an SDE system. We denote by $\{u^{(n)}(t)\}_{n=1}^N$ the sample paths of the underlying SDE. For each $1 \le n \le N$, we have $u_0^{(n)} = u^{(n)}(0)$, and we view $u_i^{(n)}$ as an approximation of $u^{(n)}(t)$ for $t = \alpha i$. Similarly as above, we define $P^{uu}(t), P^{uy}(t), P^{yy}(t)$ as the corresponding empirical covariances at time $t \ge 0$.

Remark 3.1. For the subsequent algorithms we will employ random perturbations $y_i^{(n)}$ of the original data y. Randomly perturbing the data is common practice for ensemble methods to ensure the correct statistics in the large ensemble limit under linearity assumptions [39].

3.1. Ensemble Gauss-Newton optimization of Tikhonov-Phillips objective.

3.1.1. *Iterative Ensemble Kalman Filter.* Given an ensemble $\{u_i^{(n)}\}_{n=1}^N$, consider the following Gauss-Newton update for each n:

$$u_{i+1}^{(n)} = u_i^{(n)} + \alpha v_i^{(n)}, \tag{3.2}$$

where $\alpha > 0$ is the length-step, and $v_i^{(n)}$ is the minimizer of the following (linearized) Tikhonov-Phillips objective (cf. equation (2.18))

$$\mathsf{J}_{\mathrm{TP},i}^{(n)}(v) = \frac{1}{2}\big|y_i^{(n)} - h(u_i^{(n)}) - H_i v\big|_R^2 + \frac{1}{2}\big|u_0^{(n)} - u_i^{(n)} - v\big|_{P_0^{uu}}^2, \qquad y_i^{(n)} \sim \mathcal{N}(y, \alpha^{-1}R). \tag{3.3}$$

Notice that we adopt the statistical linearization (3.1) in the above formulation. Applying Lemma 2.1, the minimizer $v_i^{(n)}$ can be calculated as

$$v_i^{(n)} = C_i\Big\{H_i^T R^{-1}\big(y_i^{(n)} - h(u_i^{(n)})\big) + (P_0^{uu})^{-1}\big(u_0^{(n)} - u_i^{(n)}\big)\Big\}, \tag{3.4}$$

or, in an equivalent form,

$$v_i^{(n)} = K_i\big(y_i^{(n)} - h(u_i^{(n)})\big) + (I - K_i H_i)(u_0^{(n)} - u_i^{(n)}), \tag{3.5}$$

where

$$C_i = \big(H_i^T R^{-1} H_i + (P_0^{uu})^{-1}\big)^{-1},$$
$$K_i = P_0^{uu} H_i^T (H_i P_0^{uu} H_i^T + R)^{-1}.$$

Combining (3.2) and (3.5) leads to the Iterative Ensemble Kalman Filter (IEKF) algorithm.

Algorithm 4 Iterative Ensemble Kalman Filter (IEKF)

1: **Input**: Initial ensemble $\{u_0^{(n)}\}_{n=1}^N$ sampled independently from the prior, length-step α.
2: For $i = 0, 1, \dots$ do:

Set $K_i = P_0^{uu} H_i^T (H_i P_0^{uu} H_i^T + R)^{-1}$, $\quad H_i = (P_i^{uy})^T (P_i^{uu})^{-1}$.
Draw $y_i^{(n)} \sim \mathcal{N}(y, \alpha^{-1} R)$ and set

$$u_{i+1}^{(n)} = u_i^{(n)} + \alpha\Big\{K_i\big(y_i^{(n)} - h(u_i^{(n)})\big) + (I - K_i H_i)\big(u_0^{(n)} - u_i^{(n)}\big)\Big\}, \qquad 1 \le n \le N. \tag{3.6}$$

3: **Output**: Ensemble means $m_1, m_2, \dots$

We highlight that IEKF is a natural ensemble-based version of the derivative-based IExKF Algorithm 1 with update (2.19). Algorithm 4 is a slight modification of the iterative ensemble Kalman algorithm proposed in [63]. The difference is that [63] sets $H_i = (P_i^{uy})^T (P_0^{uu})^{-1}$ rather than $H_i = (P_i^{uy})^T (P_i^{uu})^{-1}$. Our modification guarantees that Algorithm 4 is well-balanced in the sense that if $\alpha = 1$, $u_0^{(n)} \sim \mathcal{N}(m, P)$ and $h(u) = Hu$ is linear, then the output of Algorithm 4 satisfies that, as $N \to \infty$,

$$m_i \to \mu, \qquad i = 1, 2, \dots$$

where μ is the minimizer of $\mathsf{J}_{\mathrm{TP}}(u)$ given in Equation (1.3). This is analogous to Proposition 1 for IExKF. A detailed explanation is included in Section 3.1.2 below.

Other statistical linearizations and approximations of the Gauss-Newton scheme are possible. We next give a high-level description of the method proposed in [52], one of the earliest applications of iterative ensemble Kalman methods for inversion in the petroleum engineering literature. Consider the alternative characterization of the Gauss-Newton update (3.4). However, instead of using a different preconditioner C_i for each step, [52] uses a fixed preconditioner $C_* = P_0^{uu} - P_0^{uy}(R + P_0^{yy})^{-1}(P_0^{uy})^T$. Note that C_* can be viewed as an approximation of C_0 :

$$C_* \approx P_0^{uu} - P_0^{uu} H_0^T (R + H_0 P_0^{uu} H_0^T)^{-1} H_0 P_0^{uu}$$
$$= \big(H_0^T R^{-1} H_0 + (P_0^{uu})^{-1}\big)^{-1} = C_0.$$

This leads to the following algorithm.

Algorithm 5 Iterative Ensemble Kalman Filter (IEKF-RZL)

1: **Input**: Initial ensemble $\{u_0^{(n)}\}_{n=1}^N$ sampled independently from the prior, length-step α.

Set $C_* = P_0^{uu} - P_0^{uy}\big(R + P_0^{yy}\big)^{-1}(P_0^{uy})^T$.

2: For $i = 0, 1, \dots$ do:

Set $H_i = (P_i^{uy})^T (P_i^{uu})^{-1}$

Draw $y_i^{(n)} \sim \mathcal{N}(y, \alpha^{-1}R)$ and set

$$u_{i+1}^{(n)} = u_i^{(n)} + \alpha\, C_* \Big\{ H_i^T R^{-1}\big(y_i^{(n)} - h(u_i^{(n)})\big) + (P_0^{uu})^{-1}\big(u_0^{(n)} - u_i^{(n)}\big) \Big\}, \qquad 1 \le n \le N.$$

3: **Output**: Ensemble means $m_1, m_2, \dots$

We note that IEKF-RZL, where RZL refers to the authors of the work [52], is a natural ensemble-based version of the derivative-based IExKF Algorithm 1 with update (2.20). We have empirically observed in a wide range of numerical experiments that Algorithm 4 is more stable than Algorithm 5, and we now give a heuristic argument for the advantage of Algorithm 4 in small noise regimes.

- The update formula in Algorithm 4 is motivated by the large N approximation
$$P_0^{uu} H_i^T (H_i P_0^{uu} H_i^T + R)^{-1} \approx P H_i^T (H_i P H_i^T + R)^{-1},$$
which only requires that P_0^{uu} is a good approximation of P.
- The update formula in Algorithm 5 may be derived by invoking a large N approximation of several terms. In particular, the error arising from the approximation
$$C_* H_i^T R^{-1} \approx C_0 H_i^T R^{-1},$$
gets amplified when R is small.

Empirical evidence of the instability of Algorithm 5 will be given in Section 5.1.

3.1.2. *Analysis of IEKF.* In the literature [52, 63], the length-step α is sometimes set to be 1. Here we state a simple observation about Algorithm 4 when $\alpha = 1$ and $h(u) = Hu$ is linear. We further assume that $H_i \equiv H$ for all i. Then $K_i \equiv K_0 := P_0^{uu} H^T (H P_0^{uu} H^T + R)^{-1}$, and the update (3.6) can be simplified as

$$\begin{aligned} u_{i+1}^{(n)} &= u_i^{(n)} + K_0(y_i^{(n)} - H u_i^{(n)}) + (I - K_0 H)(u_0^{(n)} - u_i^{(n)}) \\ &= u_0^{(n)} + K_0(y_i^{(n)} - H u_0^{(n)}), \end{aligned}$$

where $y_i^{(n)} \sim \mathcal{N}(y, R)$. If $\{u_0^{(n)}\}_{n=1}^N$ are sampled independently from the prior $\mathcal{N}(m, P)$ and we let $N \to \infty$, we have $K_0 \to K = PH^T(HPH^T + R)^{-1}$ and, by the law of large numbers, for any $i \ge 1$,

$$\begin{aligned} \frac{1}{N}\sum_{n=1}^N u_i^{(n)} &= (I - K_0 H) \cdot \frac{1}{N}\sum_{n=1}^N u_0^{(n)} + K_0 \cdot \frac{1}{N}\sum_{n=1}^N y_i^{(n)} \\ &\to (I - KH)m + Ky, \end{aligned}$$

which is the posterior mean (and mode) in the linear setting. In other words, in the large ensemble limit, the ensemble mean recovers the posterior mean *after one iteration.* However, while the choice $\alpha = 1$ may be effective in low dimensional (nonlinear) inverse problems, we do not recommend it when the dimensionality is high, as H_i might not be a good approximation of H. Thus, we introduce Algorithms 4 and 5 with a choice of length-step α, resembling the derivative-based Gauss-Newton method discussed in Section 2.1. Further analysis of a new variant of Algorithm 4 will be conducted in a continuum limit setting in Section 4.

Remark 3.2. Some remarks:

1. Although this will not be the focus of our paper, the IEKF Algorithm 4 (together with the EKI Algorithm 6 and TEKI Algorithm 7 to be discussed later) enjoys the 'initial subspace property' studied in previous works [30, 55, 13] by which, for any i and any initialization of $\{u_0^{(n)}\}_{n=1}^N$,
$$\operatorname{span}\big(\{u_i^{(n)}\}_{n=1}^N\big) \subset \operatorname{span}\big(\{u_0^{(n)}\}_{n=1}^N\big).$$
This can be shown easily for the IEKF Algorithm 4 by expanding the P_0^{uu} term in K_i in the update formula (3.6).
2. Since we assume a Gaussian prior $\mathcal{N}(m, P)$ on u, a natural idea is to replace $u_0^{(n)}$ by m and P_0^{uu} by P in the update formula (3.6). We pursue this idea in Section 4.1, where we introduce a new variant of Algorithm 4 and analyze it in the continuum limit setting. While the initial subspace property breaks down, this new variant is numerically promising, as shown in Section 5.

3.2. Ensemble Levenberg-Marquardt optimization of data-misfit objective.

3.2.1. *Ensemble Kalman Inversion.* Given an ensemble $\{u_i^{(n)}\}_{n=1}^N$, consider the following Levenberg-Marquardt update for each n:
$$u_{i+1}^{(n)} = u_i^{(n)} + v_i^{(n)}, \tag{3.7}$$
where $v_i^{(n)}$ is the minimizer of the following regularized (linearized) data-misfit objective (cf. equation (2.25))
$$\mathsf{J}_{\mathrm{DM},i}^{(n),\mathrm{UC}}(v) = \frac{1}{2}\big|y_i^{(n)} - h(u_i^{(n)}) - H_i v\big|_R^2 + \frac{1}{2\alpha}|v|_{P_i^{uu}}^2, \qquad y_i^{(n)} \sim \mathcal{N}(y, \alpha^{-1}R), \tag{3.8}$$
and $\alpha > 0$ will be regarded as a length-step. Notice that we adopt the statistical linearization (3.1) in the above formulation. Applying Lemma 2.1, we can calculate the minimizer $v_i^{(n)}$ explicitly:
$$v_i^{(n)} = (H_i^T R^{-1} H_i + \alpha^{-1}(P_i^{uu})^{-1})^{-1} H_i^T R^{-1}\big(y_i^{(n)} - h(u_i^{(n)})\big), \tag{3.9}$$
or, in an equivalent form,
$$v_i^{(n)} = P_i^{uu} H_i^T (H_i P_i^{uu} H_i^T + \alpha^{-1} R)^{-1}\big(y_i^{(n)} - h(u_i^{(n)})\big). \tag{3.10}$$
We combine (3.7) and (3.10), substitute $P_i^{uu} H_i^T = P_i^{uy}$, and make another level of approximation $H_i P_i^{uy} \approx P_i^{yy}$. This leads to the Ensemble Kalman Inversion (EKI) method [30].

We note that EKI is a natural ensemble-based version of the derivative-based ILM-DM Algorithm 2. However, an important difference is that the Kalman gain in ILM-DM only uses the iterates to update H_i and P is kept fixed. In contrast, the ensemble is used in EKI to update P_i^{uy} and P_i^{yy}.

Algorithm 6 Ensemble Kalman Inversion (EKI)

1: **Input**: Initial ensemble $\{u_0^{(n)}\}_{n=1}^N$, length-step α.
2: For $i = 0, 1, \dots$ do:

Set $K_i = P_i^{uy}(P_i^{yy} + \alpha^{-1}R)^{-1}$.
Draw $y_i^{(n)} \sim \mathcal{N}(y, \alpha^{-1}R)$ and set

$$u_{i+1}^{(n)} = u_i^{(n)} + K_i\Big\{y_i^{(n)} - h(u_i^{(n)})\Big\}, \qquad 1 \le n \le N. \tag{3.11}$$

3: **Output**: Ensemble means $m_1, m_2, \dots$

3.2.2. *Analysis of EKI.* In view of the definition of K_i, we can rewrite the update (3.11) as a time-stepping scheme (as similarly done in [55, 56]):

$$\begin{aligned} u_{i+1}^{(n)} &= u_i^{(n)} + \alpha P_i^{uy}(\alpha P_i^{yy} + R)^{-1}\big(y + \alpha^{-1/2}R^{1/2}\xi_i^{(n)} - h(u_i^{(n)})\big) \\ &= u_i^{(n)} + \alpha P_i^{uy}(\alpha P_i^{yy} + R)^{-1}\big(y - h(u_i^{(n)})\big) + \alpha^{1/2}P_i^{uy}(\alpha P_i^{yy} + R)^{-1}R^{1/2}\xi_i^{(n)}, \end{aligned} \tag{3.12}$$

where $\xi_i^{(n)} \sim \mathcal{N}(0, I)$ are independent. Taking the limit $\alpha \to 0$, we interpret (3.12) as a discretization of the SDE system

$$\mathrm{d}u^{(n)} = P^{uy}(t)R^{-1}\big(y - h(u^{(n)})\big)\,\mathrm{d}t + P^{uy}(t)R^{-1/2}\,\mathrm{d}W^{(n)}. \tag{3.13}$$

If $h(u) = Hu$ is linear, the SDE system (3.13) turns into

$$\mathrm{d}u^{(n)} = P^{uu}(t)H^TR^{-1}\big(y - Hu^{(n)}\big)\,\mathrm{d}t + P^{uu}(t)H^TR^{-1/2}\,\mathrm{d}W^{(n)}. \tag{3.14}$$

Proposition 4. *For the SDE system (3.13), assume $h(u) = Hu$ is linear, and suppose that the initial ensemble $\{u_0^{(n)}\}_{n=1}^N$ is drawn independently from a continuous distribution with finite second moments. Then, in the large ensemble size limit $N \to \infty$ (mean-field), the distribution of $u^{(n)}(t)$ has mean $\mathfrak{m}(t)$ and covariance $\mathfrak{C}(t)$, which satisfy*

$$\frac{\mathrm{d}\mathfrak{m}(t)}{\mathrm{d}t} = \mathfrak{C}(t)H^TR^{-1}\big(y - H\mathfrak{m}(t)\big), \tag{3.15}$$

$$\frac{\mathrm{d}\mathfrak{C}(t)}{\mathrm{d}t} = -\mathfrak{C}(t)H^TR^{-1}H\mathfrak{C}(t). \tag{3.16}$$

Furthermore, the solution can be computed analytically:

$$\mathfrak{m}(t) = \Big(\mathfrak{C}(0)^{-1} + tH^TR^{-1}H\Big)^{-1}\big(\mathfrak{C}(0)^{-1}\mathfrak{m}(0) + tH^TR^{-1}y\big), \tag{3.17}$$

$$\mathfrak{C}(t) = \Big(\mathfrak{C}(0)^{-1} + tH^TR^{-1}H\Big)^{-1}. \tag{3.18}$$

In particular, if $H \in \mathbf{R}^{k\times d}$ has full column rank (i.e., $d \le k$, $\mathrm{rank}(H) = d$), then, as $t \to \infty$,

$$\mathfrak{m}(t) \to (H^TR^{-1}H)^{-1}(H^TR^{-1}y), \tag{3.19}$$

$$\mathfrak{C}(t) \to 0. \tag{3.20}$$

If $H \in \mathbf{R}^{k\times d}$ has full row rank (i.e., $d \geq k$, $\mathrm{rank}(H) = k$), then, as $t \to \infty$,

$$H\mathfrak{m}(t) \to y, \tag{3.21}$$

$$H\mathfrak{C}(t)H^T \to 0. \tag{3.22}$$

Proof. The proof technique is similar to [22]. We will use that

$$\mathfrak{m}(t) = \lim_{N\to\infty} \mathbb{E}\big[u^{(n)}(t)\big],$$
$$\mathfrak{C}(t) = \lim_{N\to\infty} \mathbb{E}\big[e^{(n)}(t) \otimes e^{(n)}(t)\big],$$

where $e^{(n)}(t) := u^{(n)}(t) - \mathfrak{m}(t)$. First, note that (3.15) follows directly from (3.14) using that in the mean field limit $P^{uu}(t)$ can be replaced by $\mathfrak{C}(t)$. To obtain the evolution of $\mathfrak{C}(t)$, note that

$$\mathrm{d}\mathfrak{C}(t) = \lim_{N\to\infty} \mathbb{E}\Big[\mathrm{d}e^{(n)} \otimes e^{(n)} + e^{(n)} \otimes \mathrm{d}e^{(n)} + \mathrm{d}e^{(n)} \otimes \mathrm{d}e^{(n)}\Big],$$

where the last term accounts for the Itô correction. This simplifies as

$$\begin{aligned}\frac{\mathrm{d}\mathfrak{C}(t)}{\mathrm{d}t} &= \lim_{N\to\infty} \mathbb{E}\Big[-\mathfrak{C}(t)H^TR^{-1}H(e^{(n)} \otimes e^{(n)}) - (e^{(n)} \otimes e^{(n)})H^TR^{-1}H\mathfrak{C}(t) \\ &\quad + \mathfrak{C}(t)H^TR^{-1}H\mathfrak{C}(t)\Big] \\ &= -\mathfrak{C}(t)H^TR^{-1}H\mathfrak{C}(t),\end{aligned}$$

which gives Equation (3.16). To derive exact formulas for $\mathfrak{m}(t)$ and $\mathfrak{C}(t)$, we notice that

$$\frac{\mathrm{d}\mathfrak{C}(t)^{-1}}{\mathrm{d}t} = -\mathfrak{C}(t)^{-1}\frac{\mathrm{d}\mathfrak{C}(t)}{\mathrm{d}t}\mathfrak{C}(t)^{-1} = H^TR^{-1}H,$$

and

$$\frac{\mathrm{d}\big(\mathfrak{C}(t)^{-1}\mathfrak{m}(t)\big)}{\mathrm{d}t} = \frac{\mathrm{d}\mathfrak{C}(t)^{-1}}{\mathrm{d}t}\mathfrak{m}(t) + \mathfrak{C}(t)^{-1}\frac{\mathrm{d}\mathfrak{m}(t)}{\mathrm{d}t} = H^TR^{-1}y.$$

Then (3.17) and (3.18) follow easily.

If H has full column rank, then $H^TR^{-1}H$ is invertible and therefore, as $t \to \infty$,

$$\mathfrak{C}(t) = t^{-1}\big(t^{-1}\mathfrak{C}(0)^{-1} + H^TR^{-1}H\big)^{-1} \to 0$$

by continuity of the matrix inverse function. The limit of $\mathfrak{m}(t)$, (3.19), follows immediately from (3.17).

If H has full row rank, we make the following substitutions

$$\widetilde{\mathfrak{m}}(t) = R^{-1/2}H\mathfrak{m}(t), \qquad \widetilde{\mathfrak{C}}(t) = R^{-1/2}H\mathfrak{C}(t)H^TR^{-1/2}.$$

Then (3.15) and (3.16) can be transformed into

$$\frac{\mathrm{d}\widetilde{\mathfrak{m}}(t)}{\mathrm{d}t} = \widetilde{\mathfrak{C}}(t)\big(R^{-1/2}y - \widetilde{\mathfrak{m}}(t)\big), \tag{3.23}$$

$$\frac{\mathrm{d}\widetilde{\mathfrak{C}}(t)}{\mathrm{d}t} = -\widetilde{\mathfrak{C}}(t)^2. \tag{3.24}$$

Using the fact that $\widetilde{\mathfrak{C}}(0) = R^{-1/2}H\mathfrak{C}(0)H^TR^{-1/2}$ is invertible, we can solve these using the same technique as in the previous case:

$$\widetilde{\mathfrak{C}}(t) = \big(\widetilde{\mathfrak{C}}(0)^{-1} + t\big)^{-1}, \qquad \widetilde{\mathfrak{m}}(t) = \big(\widetilde{\mathfrak{C}}(0)^{-1} + t\big)^{-1}\big(\widetilde{\mathfrak{C}}(0)^{-1}\widetilde{\mathfrak{m}}(0) + tR^{-1/2}y\big).$$

As $t \to \infty$, we have $\widetilde{\mathfrak{m}}(t) \to R^{-1/2}y$ and $\widetilde{\mathfrak{C}}(t) \to 0$, which lead to (3.21) and (3.22). □

Remark 3.3. Some remarks:

1. In fact, (3.22) always holds, without rank constraints on H. The proof follows the similar idea as in [55]. This can be viewed from Equation (3.24), where we can perform eigenvalue decomposition $\widetilde{\mathfrak{C}}(0) = X\Lambda(0)X^T$, and show that $\widetilde{\mathfrak{C}}(t) = X\Lambda(t)X^T$, where $\Lambda(t)$ are diagonal matrices and $\Lambda(t) \to 0$ as $t \to \infty$. The statement that $H\mathfrak{C}(t)H^T \to 0$ (or $\mathfrak{C}(t) \to 0$ if H has full column rank) is referred to as 'ensemble collapse'. This can be interpreted as 'the images of all the particles under H collapse to a single point as time evolves'. Our numerical results in Section 5 show that the ensemble collapse phenomenon of Algorithm 6 is also observed in a variety of nonlinear examples and outside the mean-field limit, with moderate ensemble size N. These empirical results justify the practical significance of the linear continuum analysis in Proposition 4.
2. Under the same setting of Proposition 4, if we further require that the initial ensemble $\{u_0^{(n)}\}_{n=1}^N$ is drawn independently from the prior distribution $\mathcal{N}(m, P)$, then in the mean-field limit $N \to \infty$ we have $\mathfrak{m}(0) = m$ and $\mathfrak{C}(0) = P$, leading to
$$\begin{aligned} \mathfrak{m}(1) &= (P^{-1} + H^T R^{-1} H)^{-1}(P^{-1}m + H^T R^{-1} y), \\ \mathfrak{C}(1) &= (P^{-1} + H^T R^{-1} H)^{-1} \end{aligned}$$
which are the true posterior mean and covariance, respectively. However, we have observed in a variety of numerical examples (not reported here) that in nonlinear problems it is often necessary to run EKI up to times larger than 1 to obtain adequate approximation of the posterior mean and covariance. Providing a suitable stopping criteria for EKI is a topic of current research [56, 32] beyond the scope of our work.

3.3. Ensemble Levenberg-Marquardt optimization of Tikhonov-Phillips objective.

3.3.1. *Tikhonov Ensemble Kalman Inversion.* Recall that we define
$$z := \begin{bmatrix} y \\ m \end{bmatrix}, \qquad g(u) := \begin{bmatrix} h(u) \\ u \end{bmatrix}, \qquad Q := \begin{bmatrix} R & 0 \\ 0 & P \end{bmatrix}.$$
Then, given an ensemble $\{u_i^{(n)}\}_{n=1}^N$, we can define
$$g_i = \frac{1}{N}\sum_{n=1}^N g(u_i^{(n)}),$$
and empirical covariances
$$\begin{aligned} P_i^{zz} &= \frac{1}{N}\sum_{n=1}^N \big(g(u_i^{(n)}) - g_i\big)\big(g(u_i^{(n)}) - g_i\big)^T, \\ P_i^{uz} &= \frac{1}{N}\sum_{n=1}^N \big(u_i^{(n)} - m_i\big)\big(g(u_i^{(n)}) - g_i\big)^T. \end{aligned}$$
Furthermore, we define the statistical linearization G_i:
$$g'(u_i^{(n)}) \approx (P_i^{uz})^T (P_i^{uu})^{-1} =: G_i. \tag{3.25}$$
It is not hard to check that
$$G_i = \begin{bmatrix} H_i \\ I \end{bmatrix},$$

with H_i defined in (3.1).

Given an ensemble $\{u_i^{(n)}\}_{n=1}^N$, consider the following Levenberg-Marquardt update for each n:

$$u_{i+1}^{(n)} = u_i^{(n)} + v_i^{(n)},$$

where $v_i^{(n)}$ is the minimizer of the following regularized (linearized) Tikhonov-Phillips objective (cf. equation (2.29))

$$\mathsf{J}_{\mathrm{TP},i}^{(n),\mathrm{UC}}(v) = \frac{1}{2}\big|z_i^{(n)} - g(u_i^{(n)}) - G_i v\big|_Q^2 + \frac{1}{2\alpha}|v|_{P_i^{uu}}^2, \qquad z_i^{(n)} \sim \mathcal{N}(z, \alpha^{-1}Q), \tag{3.26}$$

and $\alpha > 0$ will be regarded as a length-step. We can calculate the minimizer $v_i^{(n)}$ explicitly, applying Lemma 2.1:

$$v_i^{(n)} = (G_i^T Q^{-1} G_i + \alpha^{-1}(P_i^{uu})^{-1})^{-1} G_i^T Q^{-1}\big(z_i^{(n)} - g(u_i^{(n)})\big), \tag{3.27}$$

or, in an equivalent form,

$$v_i^{(n)} = P_i^{uu} G_i^T (G_i P_i^{uu} G_i^T + \alpha^{-1} Q)^{-1}\big(z_i^{(n)} - g(u_i^{(n)})\big). \tag{3.28}$$

Similar to EKI, in Equation (3.28) we substitute $P_i^{uu} G_i^T = P_i^{uz}$, and make the approximation $G_i P_i^{uz} \approx P_i^{zz}$. This leads to Tikhonov Ensemble Kalman Inversion (TEKI), described in Algorithm 7.

Algorithm 7 Tikhonov Ensemble Kalman Inversion (TEKI)

1: **Input**: Initial ensemble $\{u_0^{(n)}\}_{n=1}^N$, length-step α.
2: For $i = 0, 1, \ldots$ do:

Set $K_i = P_i^{uz}(P_i^{zz} + \alpha^{-1}Q)^{-1}$.
Draw $z_i^{(n)} \sim \mathcal{N}(z, \alpha^{-1}Q)$ and set

$$u_{i+1}^{(n)} = u_i^{(n)} + K_i\Big\{z_i^{(n)} - g(u_i^{(n)})\Big\}, \qquad 1 \le n \le N. \tag{3.29}$$

3: **Output**: Ensemble means $m_1, m_2, \ldots$

We note that TEKI is a natural ensemble-based version of the derivative-based ILM-TP Algorithm 3. However, the Kalman gain in ILM-TP keeps P fixed, while in TEKI P_i^{uz} and P_i^{zz} are updated using the ensemble.

3.3.2. *Analysis of TEKI.* When α is small, we can rewrite the update (3.29) as a time-stepping scheme (as similarly done in [13]):

$$\begin{aligned} u_{i+1}^{(n)} &= u_i^{(n)} + \alpha P_i^{uz}(\alpha P_i^{zz} + Q)^{-1}\big(z + \alpha^{-1/2} Q^{1/2}\xi_i^{(n)} - g(u_i^{(n)})\big) \\ &= u_i^{(n)} + \alpha P_i^{uz}(\alpha P_i^{zz} + Q)^{-1}\big(z - g(u_i^{(n)})\big) + \alpha^{1/2} P_i^{uz}(\alpha P_i^{zz} + Q)^{-1} Q^{1/2}\xi_i^{(n)}, \end{aligned} \tag{3.30}$$

where $\xi_i^{(n)} \sim \mathcal{N}(0, I_{d+k})$ are independent. Taking the limit $\alpha \to 0$, we can interpret (3.30) as a discretization of an interacting particle SDE system

$$\mathrm{d}u^{(n)} = P^{uz}(t) Q^{-1}\big(z - g(u^{(n)})\big)\,\mathrm{d}t + P^{uz}(t) Q^{-1/2}\,\mathrm{d}W^{(n)}. \tag{3.31}$$

If $h(u) = Hu$ is linear, $g(u) = Gu$ is also linear and the SDE system (3.31) can be rewritten as

$$\begin{aligned} \mathrm{d}u^{(n)} &= P^{uz}(t)Q^{-1}\big(z - Gu^{(n)}\big)\,\mathrm{d}t + P^{uz}(t)Q^{-1/2}\,\mathrm{d}W^{(n)} \\ &= P^{uu}(t)G^TQ^{-1}\big(z - Gu^{(n)}\big)\,\mathrm{d}t + P^{uu}(t)G^TQ^{-1/2}\,\mathrm{d}W^{(n)}. \end{aligned} \tag{3.32}$$

Proposition 5. *For the SDE system (3.31), assume $h(u) = Hu$ is linear, and that the initial ensemble $\{u_0^{(n)}\}_{n=1}^N$ is made of independent samples from a distribution with finite second moments. Then, in the large ensemble limit (mean-field), the distribution of $u^{(n)}(t)$ has mean $\mathfrak{m}(t)$ and covariance $\mathfrak{C}(t)$, which satisfy:*

$$\frac{\mathrm{d}\mathfrak{m}(t)}{\mathrm{d}t} = \mathfrak{C}(t)G^TQ^{-1}(z - G\mathfrak{m}(t)), \tag{3.33}$$

$$\frac{\mathrm{d}\mathfrak{C}(t)}{\mathrm{d}t} = -\mathfrak{C}(t)G^TQ^{-1}G\mathfrak{C}(t). \tag{3.34}$$

Furthermore, the solution can be computed analytically:

$$\mathfrak{m}(t) = \Big(\mathfrak{C}(0)^{-1} + tG^TQ^{-1}G\Big)^{-1}\big(\mathfrak{C}(0)^{-1}\mathfrak{m}(0) + tG^TQ^{-1}z\big), \tag{3.35}$$

$$\mathfrak{C}(t) = \Big(\mathfrak{C}(0)^{-1} + tG^TQ^{-1}G\Big)^{-1}. \tag{3.36}$$

In particular, as $t \to \infty$,

$$\begin{aligned} \mathfrak{m}(t) &\to (G^TQ^{-1}G)^{-1}(G^TQ^{-1}z) \\ &= (H^TR^{-1}H + P)^{-1}(H^TR^{-1}y + P^{-1}m), \end{aligned} \tag{3.37}$$

$$\mathfrak{C}(t) \to 0. \tag{3.38}$$

Notice that $\mathfrak{m}(t)$ converges to the true posterior mean.

Proof. Equations (3.33)-(3.36) can be derived similarly as in the proof of Proposition 4, replacing H by G, R by Q, and y by z. Now since $G^TQ^{-1}G = H^TR^{-1}H + P$ is always invertible we have that, as $t \to \infty$,

$$\mathfrak{C}(t) = t^{-1}\big(t^{-1}\mathfrak{C}(0)^{-1} + G^TQ^{-1}G\big)^{-1} \to 0$$

by continuity of the matrix inverse function. The limit of $\mathfrak{m}(t)$ in (3.37) follows directly from (3.35). □

4. Ensemble Kalman methods: New variants.

In the previous section we discussed three popular subfamilies of iterative ensemble Kalman methods, analogous to the derivative-based algorithms in Section 2. The aim of this section is to introduce two new iterative ensemble Kalman methods which are inspired by the SDE continuum limit structure of the algorithms in Section 3. The two new methods that we introduce have in common that they rely on statistical linearization, and that the long-time limit of the ensemble covariance recovers the posterior covariance in a linear setting. This holds true even if the initial ensemble is not drawn from the prior distribution on the unknown.

Subsection 4.1 contains a new variant of IEKF which in addition to recovering the posterior covariance, it also recovers the posterior mean in the long-time limit. Subsection 4.2 introduces a new variant of the EKI method. Finally, Subsection 4.3 highlights the gradient structure of the algorithms in this and the previous section, shows that our new variant of IEKF can also be interpreted as a modified TEKI algorithm, and sets our proposed new methods into the broader literature.

4.1. **Iterative Ensemble Kalman Filter with statistical linearization.** In some of the literature on iterative ensemble Kalman methods [52, 63], the length-step (α in Algorithms 4 and 5) is set to be 1. Although this choice of length-step allows to recover the true posterior mean in the linear case after one iteration (see Section 3.1.2), it leads to numerical instability in complex nonlinear models. Alternative ways to set α include performing a line-search that satisfies Wolfe's condition, or using other ad-hoc line-search criteria [24]. These methods allow α to be adaptively chosen throughout the iterations, but they introduce other hyperparameters that need to be selected manually.

Our idea here is to slightly modify Algorithm 4 so that in the linear case its continuum limit has the true posterior as its invariant distribution. In this way we can simply choose a small enough α and run the algorithm until the iterates reach a statistical equilibrium, avoiding the need to specify suitable hyperparameters and stopping criteria. Our empirical results show that this approach also performs well in the nonlinear case.

In the update Equation (3.6), we replace each of the $u_0^{(n)}$ by a perturbation of the prior mean m, and we replace P_0^{uu} by the prior covariance matrix P in the definition of K_i. Details can be found below in our modified algorithm, which we call IEKF-SL.

Algorithm 8 Iterative Ensemble Kalman Filter with Statistical Linearization (IEKF-SL)

1: **Input**: Initial ensemble $\{u_0^{(n)}\}_{n=1}^N$, step size α.
2: For $i = 0, 1, \dots$ do:

$$\text{Set } K_i = PH_i^T(H_iPH_i^T + R)^{-1}, \qquad H_i = (P_i^{uy})^T(P_i^{uu})^{-1}.$$
$$\text{Draw } y_i^{(n)} \sim \mathcal{N}(y, 2\alpha^{-1}R),\ m_i^{(n)} \sim \mathcal{N}(m, 2\alpha^{-1}P) \text{ and set}$$
$$u_{i+1}^{(n)} = u_i^{(n)} + \alpha\Big\{K_i\big(y_i^{(n)} - h(u_i^{(n)})\big) + (I - K_iH_i)\big(m_i^{(n)} - u_i^{(n)}\big)\Big\}, \qquad 1 \le n \le N. \tag{4.1}$$

3: **Output**: Ensemble means $m_1, m_2, \dots$

It is natural to regard the update (4.1) as a time-stepping scheme. We rewrite it in an alternative form, in analogy to (3.4):

$$\begin{aligned} u_{i+1}^{(n)} &= u_i^{(n)} + \alpha C_i\Big\{H_i^TR^{-1}\big(y_i^{(n)} - h(u_i^{(n)})\big) + P^{-1}\big(m_i^{(n)} - u_i^{(n)}\big)\Big\} \\ &= u_i^{(n)} + \alpha C_i\Big\{H_i^TR^{-1}\big(y - h(u_i^{(n)})\big) + P^{-1}\big(m - u_i^{(n)}\big) + (H_i^TR^{-1}\zeta + P^{-1}\eta)\Big\}, \end{aligned} \tag{4.2}$$

where

$$C_i = \big(H_i^TR^{-1}H_i + P^{-1}\big)^{-1}, \quad \zeta \sim \mathcal{N}(0, 2\alpha^{-1}R), \quad \eta \sim \mathcal{N}(0, 2\alpha^{-1}P).$$

We interpret Equation (4.2) as a discretization of the SDE system

$$\mathrm{d}u^{(n)} = C(t)\Big(H(t)^TR^{-1}\big(y - h(u^{(n)})\big) + P^{-1}\big(m - u^{(n)}\big)\Big)\,\mathrm{d}t + \sqrt{2C(t)}\,\mathrm{d}W^{(n)}, \tag{4.3}$$

where

$$H(t) = \big(P^{uy}(t)\big)^T \big(P^{uu}(t)\big)^{-1},$$
$$C(t) = \big(H(t)^T R^{-1} H(t) + P^{-1}\big)^{-1}.$$

The diffusion term can be derived using the fact that

$$C_i(H_i^T R^{-1}\zeta + P^{-1}\eta) \sim \mathcal{N}\big(0, 2\alpha^{-1} C_i (H_i^T R^{-1} H_i + P^{-1}) C_i\big) = \mathcal{N}(0, 2\alpha^{-1} C_i).$$

If $h(u) = Hu$ is linear and the empirical covariance $P^{uu}(t)$ has full rank for all t, then $H(t) \equiv H$, $C(t) \equiv C = (P^{-1} + H^T R^{-1} H)^{-1}$, and the SDE system can be decoupled and further simplified:

$$\begin{aligned} \mathrm{d}u^{(n)} &= C\Big(H^T R^{-1}\big(y - Hu^{(n)}\big) + P^{-1}\big(m - u^{(n)}\big)\Big)\,\mathrm{d}t + \sqrt{2C}\,\mathrm{d}W \\ &= \Big(-u^{(n)} + C(H^T R^{-1} y + P^{-1} m)\Big)\,\mathrm{d}t + \sqrt{2C}\,\mathrm{d}W. \end{aligned} \tag{4.4}$$

Proposition 6. *For the SDE system (4.3), assume $h(u) = Hu$ is linear and $H(t) \equiv H$ holds. Assume that the initial ensemble $\{u_0^{(n)}\}_{n=1}^N$ is made of independent samples from a continuous distribution with finite second moments. Then, for $1 \le n \le N$, the mean $\mathfrak{m}(t)$ and covariance $\mathfrak{C}(t)$ of $u^{(n)}(t)$ satisfy*

$$\frac{\mathrm{d}\mathfrak{m}(t)}{\mathrm{d}t} = -\mathfrak{m}(t) + C(H^T R^{-1} y + P^{-1} m), \tag{4.5}$$
$$\frac{\mathrm{d}\mathfrak{C}(t)}{\mathrm{d}t} = -2\mathfrak{C}(t) + 2C. \tag{4.6}$$

Furthermore, as $t \to \infty$,

$$\begin{aligned} \mathfrak{m}(t) &\to C(H^T R^{-1} y + P^{-1} m) \\ &= (P^{-1} + H^T R^{-1} H)^{-1}(H^T R^{-1} y + P^{-1} m), \end{aligned} \tag{4.7}$$
$$\mathfrak{C}(t) \to C = (P^{-1} + H^T R^{-1} H)^{-1}. \tag{4.8}$$

In other words, $\mathfrak{m}(t)$ and $\mathfrak{C}(t)$ converge to the true posterior mean and covariance, respectively.

Proof. It is clear that, for fixed $t > 0$, $\{u^{(n)}(t)\}_{n=1}^N$ are independent and identically distributed. The evolution of the mean follows directly from (4.4), and the evolution of the covariance follows from (4.4) by applying Itô's formula. It is then straightforward to derive (4.7) and (4.8) from (4.5) and (4.6), respectively. □

4.2. **Ensemble Kalman inversion with statistical linearization.** Recall that in the formulation of EKI, we define a regularized (linearized) data-misfit objective (3.8), where we have a regularizer on v with respect to the norm $|\cdot|_{P_i^{uu}}$. However, in view of Proposition 4, under a linear forward model $h(u) = Hu$, the particles $\{u_i^{(n)}\}_{i=1}^n$ will 'collapse', meaning that the empirical covariance of $\{Hu_i^{(n)}\}_{n=1}^N$ will vanish in the large i limit. One possible solution to this is 'covariance inflation', namely to inject certain amount of random noise after each ensemble update. However, this requires ad-hoc tuning of additional hyperparameters. An alternative approach to avoid the ensemble collapse is to modify the regularization term in the Levenberg-Marquardt formulation (3.8). The rough idea is to consider another regularizer on $H_i v$ in the data space, as we describe in what follows.

We define a new regularized data-misfit objective, slightly different from (3.8):

$$\mathsf{J}_{\mathrm{DM},i}^{(n),\mathrm{UC}}(v) = \frac{1}{2}\big|y_i^{(n)} - h(u_i^{(n)}) - H_i v\big|_R^2 + \frac{1}{2\alpha}|v|_{C_i}^2 \qquad y_i^{(n)} \sim \mathcal{N}(y, 2\alpha^{-1}R), \tag{4.9}$$

where

$$C_i = (P^{-1} + H_i^T R^{-1} H_i)^{-1}, \tag{4.10}$$

and H_i is defined in (3.1). The regularization term can be decomposed as

$$|v|^2_{C_i} = |v|^2_P + |H_i v|^2_R.$$

The first term can be regarded as a regularization on v with respect to the prior covariance P. The second term can be regarded as a regularization on $H_i v$ with respect to the noise covariance R. Applying Lemma 2.1, we can calculate the minimizer of (4.9):

$$\begin{aligned} v_i^{(n)} &= (H_i^T R^{-1} H_i + \alpha^{-1} C_i^{-1})^{-1} H_i^T R^{-1} \big(y_i^{(n)} - h(u_i^{(n)})\big) \\ &= \big(\alpha^{-1} P^{-1} + (1+\alpha^{-1}) H_i^T R^{-1} H_i\big)^{-1} H_i^T R^{-1} \big(y_i^{(n)} - h(u_i^{(n)})\big) \\ &= \alpha P H_i^T \big((1+\alpha) H_i P H_i^T + R\big)^{-1} \big(y_i^{(n)} - h(u_i^{(n)})\big), \end{aligned} \tag{4.11}$$

where the second equality follows from the definition of C_i, and the third equality follows from the matrix identity (2.7). This leads to Algorithm 9.

Algorithm 9 Ensemble Kalman Inversion with Statistical Linearization (EKI-SL)

1: **Input**: Initial ensemble $\{u_0^{(n)}\}_{n=1}^N$, step size α.
2: For $i = 0, 1, \dots$ do:

Set $K_i = \alpha P H_i^T \big((1+\alpha) H_i P H_i^T + R\big)^{-1}$, $\quad H_i = (P_i^{uy})^T (P_i^{uu})^{-1}$.
Draw $y_i^{(n)} \sim \mathcal{N}(y, 2\alpha^{-1} R)$ and set

$$u_{i+1}^{(n)} = u_i^{(n)} + K_i \Big\{ y_i^{(n)} - h(u_i^{(n)}) \Big\}, \qquad 1 \le n \le N. \tag{4.12}$$

3: **Output**: Ensemble means $m_1, m_2, \dots$

For small $\alpha > 0$, we interpret the update (4.12) as a discretization of the coupled SDE system

$$\begin{aligned} \mathrm{d}u^{(n)} &= P H(t)^T \big(H(t) P H(t)^T + R\big)^{-1} \Big(\big(y - h(u^{(n)})\big)\,\mathrm{d}t + \sqrt{2} R^{1/2}\,\mathrm{d}W \Big) \\ &= C(t) H(t)^T R^{-1} \big(y - h(u^{(n)})\big)\,\mathrm{d}t + \sqrt{2} C(t) H(t)^T R^{-1/2}\,\mathrm{d}W, \end{aligned} \tag{4.13}$$

where

$$\begin{aligned} H(t) &= \big(P^{uy}(t)\big)^T \big(P^{uu}(t)\big)^{-1}, \\ C(t) &= \big(P^{-1} + H(t)^T R^{-1} H(t)\big)^{-1}. \end{aligned}$$

For our next result we will work under the assumption that $H(t) \equiv H$ is constant, which in particular requires that the empirical covariance $P^{uu}(t)$ has full rank for all t. Importantly, under this assumption $C(t) \equiv C = (P^{-1} + H^T R^{-1} H)^{-1}$and the SDE system is decoupled:

$$\mathrm{d}u^{(n)} = C H^T R^{-1} \big(y - H u^{(n)}\big)\,\mathrm{d}t + \sqrt{2} C H^T R^{-1/2}\,\mathrm{d}W^{(n)}. \tag{4.14}$$

Proposition 7. *For the SDE system (4.13), assume $h(u) = Hu$ is linear and $H(t) \equiv H$ holds. Assume that the initial ensemble $\{u_0^{(n)}\}_{n=1}^N$ is made of independent*

samples from a continuous distribution with finite second moments. Then the mean $\mathfrak{m}(t)$ *and covariance* $\mathfrak{C}(t)$ *of* $u^{(n)}(t)$ *satisfy*

$$\frac{\mathrm{d}\mathfrak{m}(t)}{\mathrm{d}t} = CH^TR^{-1}(y - H\mathfrak{m}(t)), \tag{4.15}$$

$$\frac{\mathrm{d}\mathfrak{C}(t)}{\mathrm{d}t} = -CH^TR^{-1}H\mathfrak{C}(t) - \mathfrak{C}(t)H^TR^{-1}HC + 2CH^TR^{-1}HC. \tag{4.16}$$

In particular, if $H \in \mathbf{R}^{k\times d}$ *has full column rank (i.e.,* $d \le k$, $\mathrm{rank}(H) = d$*), then, as* $t \to \infty$,

$$\mathfrak{m}(t) \to (H^TR^{-1}H)^{-1}(H^TR^{-1}y), \tag{4.17}$$

$$\mathfrak{C}(t) \to C = (P^{-1} + H^TR^{-1}H)^{-1}. \tag{4.18}$$

If $H \in \mathbf{R}^{k\times d}$ *has full row rank (i.e.,* $d \ge k$, $\mathrm{rank}(H) = k$*), then, as* $t \to \infty$,

$$H\mathfrak{m}(t) \to y, \tag{4.19}$$

$$H\mathfrak{C}(t)H^T \to HCH^T. \tag{4.20}$$

Proof. Note that here, in contrast to the setting of Proposition 4, the distribution of $u^{(n)}(t)$ does not depend on the ensemble size N, and we have simply $\mathfrak{m}(t) = \mathbb{E}\big[u^{(n)}(t)\big]$ and $\mathfrak{C}(t) = \mathbb{E}\left[e^{(n)}(t) \otimes e^{(n)}(t)\right]$, with $e^{(n)}(t) := u^{(n)}(t) - \mathfrak{m}(t)$. The evolution of $\mathfrak{m}(t)$ follows directly from (4.14). To obtain the evolution of $\mathfrak{C}(t)$, we use a similar technique as in the proof of Proposition 4. Applying Itô's formula,

$$\begin{aligned}\frac{\mathrm{d}\mathfrak{C}(t)}{\mathrm{d}t} &= \mathbb{E}\Big[- CH^TR^{-1}H(e^{(n)} \otimes e^{(n)}) - (e^{(n)} \otimes e^{(n)})H^TR^{-1}HC + 2CH^TR^{-1}HC\Big] \\ &= -CH^TR^{-1}H\mathfrak{C}(t) - \mathfrak{C}(t)H^TR^{-1}HC + 2CH^TR^{-1}HC,\end{aligned}$$

which recovers (4.16).

If H has full column rank, then $H^TR^{-1}H$ is invertible, and (4.17) follows immediately from (4.15). By setting the right-hand side of (4.16) to 0, and using the fact that it has a unique solution, we derive (4.18).

If H has full row rank, then (4.19) follows immediately from (4.15). Next, the substitutions

$$\widetilde{\mathfrak{C}}(t) = R^{-1/2}H\mathfrak{C}(t)H^TR^{-1/2}, \qquad \widetilde{C} = R^{-1/2}HCH^TR^{-1/2}$$

allow to transform (4.16) into

$$\frac{\mathrm{d}\widetilde{\mathfrak{C}}(t)}{\mathrm{d}t} = -\widetilde{C}\widetilde{\mathfrak{C}}(t) - \widetilde{\mathfrak{C}}(t)\widetilde{C} + 2\widetilde{C}^2. \tag{4.21}$$

Since $\widetilde{C}$ is invertible, by setting the right-hand side of (4.21) to 0 we deduce that $\widetilde{\mathfrak{C}}(t) \to \widetilde{C}$ as $t \to \infty$, which recovers (4.20). □

4.3. **Gradient structure and discussion.**

LM Algorithms in the Continuum Limit. Levenberg-Marquardt algorithms have a natural gradient structure in the continuum limit. This was shown in Equation (2.27), where the preconditioner P corresponds to the regularizer $|\cdot|_P$ that is used in the Levenberg-Marquardt algorithm (2.22). Ensemble-based Levenberg-Marquardt algorithms also possess a gradient structure. To see this, we consider an update $u_{i+1}^{(n)} = u_i^{(n)} + v_i^{(n)}$, where $v_i^{(n)}$ is the unconstrained minimizer of the same objective as (4.9)

$$\mathsf{J}_{\mathrm{DM},i}^{(n),\mathrm{UC}}(v) = \frac{1}{2}\big|y_i^{(n)} - h(u_i^{(n)}) - H_iv\big|_R^2 + \frac{1}{2\alpha}|v|_{S_i}^2 \qquad y_i^{(n)} \sim \mathcal{N}(y, 2\alpha^{-1}R),$$

except that we allow any positive (semi)definite matrix S_i to act as a 'regularizer'. The resulting continuum limit is given by the SDE system

$$\mathrm{d}u^{(n)} = S(t)H(t)^T R^{-1}\big(y - h(u^{(n)})\big)\,\mathrm{d}t + \sqrt{2}S(t)H(t)^T R^{-1/2}\,\mathrm{d}W^{(n)}, \tag{4.22}$$

where $u^{(n)}(t), S(t), H(t)$ are continuous time analogs of $u_i^{(n)}, S_i, H_i$. Notice that $H(t)^T R^{-1}(y - h(u^{(n)}))$ is exactly $-\mathsf{J}'_{\mathrm{DM}}(u^{(n)})$, so that we may rewrite (4.22) as

$$\dot{u}^{(n)} = -S(t)\mathsf{J}'_{\mathrm{DM}}(u^{(n)}) + \sqrt{2}S(t)H(t)^T R^{-1/2}\dot{W}^{(n)}, \tag{4.23}$$

which is a perturbed gradient descent with preconditioner $S(t)$. We recall that $S(t) = P^{uu}(t)$ in EKI and $S(t) = C(t) = \big(P^{-1} + H(t)^T R^{-1} H(t)\big)^{-1}$ in EKI-SL. Other choices of $S(t)$ are possible and will be studied in future work.

Gauss-Newton Algorithms in the Continuum Limit. Gauss-Newton algorithms also have a natural gradient structure in the continuum limit. As we have shown in Equation (2.21), the Gauss-Newton method applied to a Tikhonov-Phillips objective can be regarded as a gradient flow with a preconditioner that is the inverse Hessian matrix of the objective. Ensemble-based Gauss-Newton methods also possess a similar gradient structure. Recall that in Equation (4.3) we formulate the continuum limit of IEKF-SL:

$$\mathrm{d}u^{(n)} = C(t)\Big(H(t)^T R^{-1}\big(y - h(u^{(n)})\big) + P^{-1}\big(m - u^{(n)}\big)\Big)\,\mathrm{d}t + \sqrt{2C(t)}\,\mathrm{d}W^{(n)}. \tag{4.24}$$

Notice that the drift term is exactly $-C(t)\mathsf{J}'_{\mathrm{TP}}(u^{(n)})$, so we may rewrite (4.24) as

$$\dot{u}^{(n)} = -C(t)\mathsf{J}'_{\mathrm{TP}}(u^{(n)}) + \sqrt{2C(t)}\dot{W}^{(n)}, \tag{4.25}$$

which is a perturbed gradient descent with preconditioner $C(t)$, the inverse Hessian of J_{TP}.

A Unified View of Levenberg-Marquardt and Gauss-Newton Algorithms in the Continuum Limit. As a conclusion of above discussion, in the continuum limit Levenberg-Marquardt algorithms (e.g., EKI, EKI-SL) are (perturbed) gradient descent methods, with a preconditioner determined by the regularizer used in the Levenberg-Marquardt update step. Gauss-Newton algorithms (e.g., IEKF, IEKF-SL) are also (perturbed) gradient descent, with a preconditioner determined by the inverse Hessian of the objective.

Interestingly, there are cases when the two types of algorithms coincide, even with the same amount of perturbation in the gradient descent step. A way to see this is to set the Levenberg-Marquardt regularizer to be the inverse Hessian of the objective. In equation (4.23), if we consider the Tikhonov-Phillips objective (i.e., replace J_{DM} by J_{TP}, H by G, Q by R), and set $S(t) = C(t)$, then (4.23) and (4.25) will coincide, leading to the same SDE system. This is the reason why we do not introduce a 'TEKI-SL' algorithm, as it is identical to IEKF-SL.

Relationship to Ensemble Kalman Sampler (EKS). Another algorithm of interest is the Ensemble Kalman Sampler (EKS) [22, 17]. Although not discussed in this work, the EKS update is similar to (4.24). In fact, if we replace $C(t)$ by the ensemble covariance $P^{uu}(t)$, and use the fact that $P^{uu}(t)H(t)^T = P^{uy}(t)$ by definition, we immediately recover the evolution equation of EKS:

$$\mathrm{d}u^{(n)} = \Big(P^{uy}(t)^T R^{-1}\big(y - h(u^{(n)})\big) + P^{-1}\big(m - u^{(n)}\big)\Big)\,\mathrm{d}t + \sqrt{2P^{uu}(t)}\,\mathrm{d}W^{(n)}.$$

Then an Euler-Maruyama discretization is performed to compute the update formula in discrete form. However, as we find out in several numerical experiments, the length-step α needs to be chosen carefully. If the noise R has a small scale, EKS often blows up in the first few iterations, while other algorithms with the same length-step do not. Also, if we consider a linear forward model $h(u) = Hu$, EKS still requires a mean field assumption $N \to \infty$ in order for it to converge to the posterior distribution, while IEKF-SL approximately needs $N \approx d$ particles where d is the input dimension.

5. Numerical examples.

In this section we provide a numerical comparison of the iterative ensemble Kalman methods introduced in Sections 3 and 4. Our experiments highlight the variety of applications that have motivated the development of these methods, and illustrate their use in a wide range of settings. In order to provide a comparison between all variants, we will assess the performance in various ways. First, the relative error at artificial time $t > 0$ is defined as

$$\text{relative error} = \frac{|m(t) - u^\dagger|}{|u^\dagger|},$$

where $m(t)$ denotes the mean of the ensemble at time t. Plots of the evolution of the data misfit $\mathsf{J}_{\mathrm{DM}}\big(m(t)\big)$ and Tikhonov-Phillips $\mathsf{J}_{\mathrm{TP}}\big(m(t)\big)$ objectives will also be provided. We will also assess the performance through reconstruction plots, which show the ensemble mean and several ensemble quantiles at the last iteration. Additionally, for the first experiment we compare each variant through the evolution of the Frobenius norm $\|P^{uu}(t)\|_F$ of the ensemble covariance.

5.1. Elliptic boundary value problem.

In this subsection we consider a simple nonlinear Bayesian inverse problem originally presented in [19], where the unknown parameter is two dimensional and the forward map admits a closed formula. These features facilitate both the visualization and the interpretation of the solution.

5.1.1. *Problem setup.* Consider the elliptic boundary value problem

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}x}\left(\exp(u_1)\frac{\mathrm{d}}{\mathrm{d}x}p(x)\right) = 1, \quad x \in (0,1), \\ p(0) = 0, \qquad p(1) = u_2. \end{aligned} \tag{5.1}$$

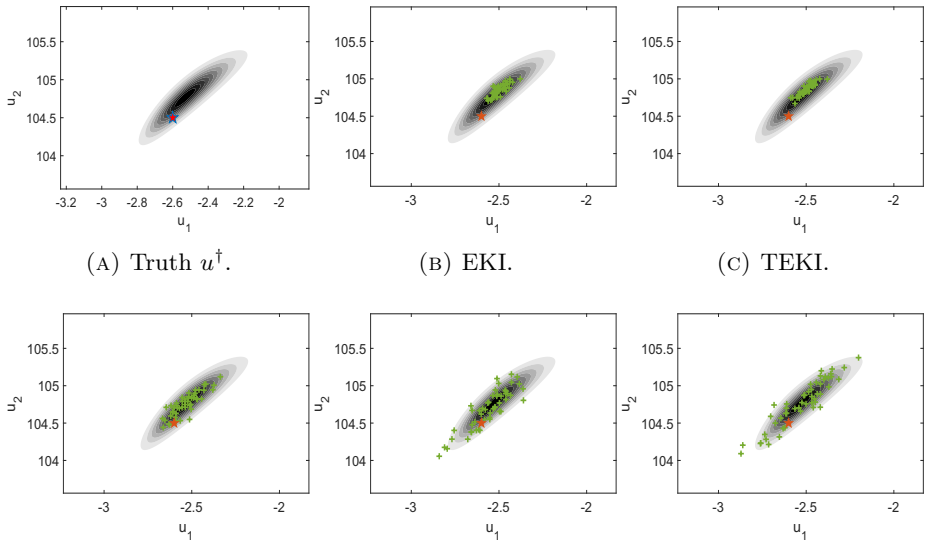
It can be checked that (5.1) has an explicit solution

$$p_u(x) = u_2 x - \frac{1}{2}\exp(-u_1)(x^2 - x). \tag{5.2}$$

We seek to recover $u = (u_1, u_2)^T$ from noisy observation of p_u at two points $x_1 = 0.25$ and $x_2 = 0.75$. Precisely, we define $h(u) := \big(p_u(x_1), p_u(x_2)\big)^T$ and consider the inverse problem of recovering u from data y of the form

$$y = h(u) + \eta, \tag{5.3}$$

where $\eta \sim \mathcal{N}(0, \gamma^2 I_2)$. We set a Gaussian prior on the unknown parameter $u \sim \mathcal{N}(0,1) \times \mathcal{N}(100, 16)$. In our numerical experiments we let the true parameter be $u^\dagger = (-2.6, 104.5)^T$ and use it to generate synthetic data $y = h(u^\dagger) + \eta$ with noise level $\gamma = 0.1$.



(A) Truth $u^\dagger$. (B) EKI. (C) TEKI.

(D) IEKF. (E) EKI-SL. (F) IEKF-SL.

FIGURE 1. Ensemble members (green) after 100 iterations, with truth $u^\dagger$ (red star) and contour plot of (unnormalized) posterior density.

5.1.2. *Implementation details and numerical results.* We set the ensemble size to be $N = 50$. The initial ensemble $\{u_0^{(n)}\}_{n=1}^N$ is drawn independently from the prior. The length-step α is fixed to be 0.1 for all methods. We run each algorithm up to time $T = 10$, which corresponds to 100 iterations. We emphasize that the results here are not sensitive to the choice of α, provided that it is taken to be sufficiently small. The range of suitable length-steps will, in general, depend on the strength of the nonlinearity of h and the dimension of the problem. For the choice of T, stopping criteria for convergence could be added to the algorithms. Here for simplicity we set T manually and let all algorithms run for a sufficient amount of time.

We plot the level curves of the posterior density of $u = (u_1, u_2)^T$ in Figure 1. The forward model is approximately linear, as can be seen from the contour plots in Figure 1 or directly from Equation (5.2). Hence it can be used to validate the claims we made in previous sections.

Figure 2 compares the time evolution of the empirical covariance of the different methods. The IEKF-RZL algorithm consistently blows up in the first few iterations due to the small noise which, as discussed in Section 3.1, is an important drawback. Due to the poor performance of IEKF-RZL in small noise regimes, we will not include this method in subsequent comparisons. The EKI and TEKI plots show the 'ensemble collapse' phenomenon discussed in Sections 3.2 and 3.3. Note that the size of the empirical covariances of EKI and TEKI decreases monotonically, and by the 100-th iteration their size is one order of magnitude smaller than that of the new variants IEKF-SL and EKI-SL. The ensemble collapse of EKI and TEKI can also be seen in Figure 1, where we plot all the ensemble members after 100 iterations. In contrast, Figure 2 shows that the size of the empirical covariances of the new variants IEKF-SL and EKI-SL stabilizes after around 40 iterations, and Figure 1 suggests that the spread of the ensemble matches that of the posterior.

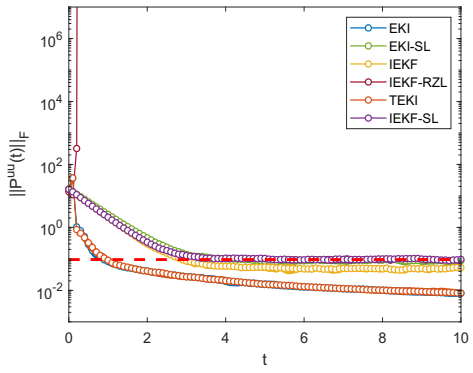


FIGURE 2. Evolution of the Frobenius norm of the ensemble covariance $P^{uu}(t)$. For reference, we also plot the Frobenius norm of the true posterior covariance (red dashed line). The norm of IEKF-RZL blows up after a few iterations. The norms of the EKI and TEKI are almost identical and monotonically decreasing. The norms of the new variants EKI-SL and IEKF-SL are similar and stabilize after around 40 iterations. The norm of IEKF lies between those of the old and new variants.

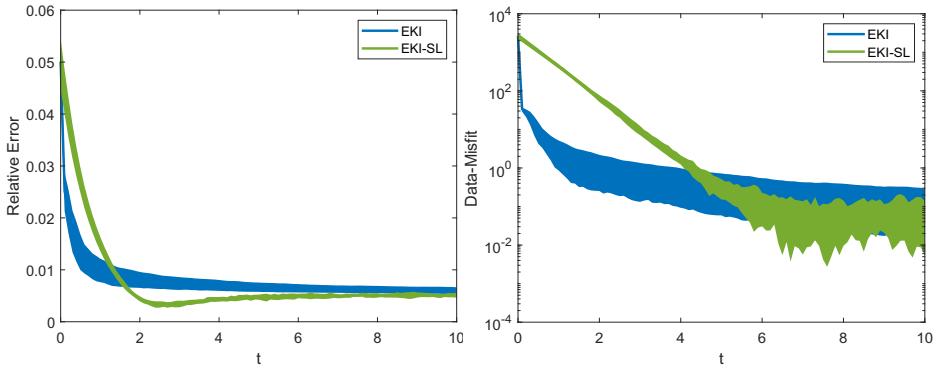


FIGURE 3. EKI & EKI-SL: Relative errors and data misfit w.r.t time t.

These results are in agreement with the theoretical results derived in Sections 4.2 and 4.1 in a linear setting. Although we have not discussed the IEKF method when α is small, its ensemble covariance does not collapse in the linear setting, as can be seen in Figure 2.

In Figures 3 and 4 we show the performance of the different methods along the full iteration sequence. Here and in subsequent numerical examples we use two performance assessments. First, the relative error, defined as $|m(t) - u^\dagger|/|u^\dagger|$ where $m(t)$ is the ensemble's empirical mean, which evaluates how well the ensemble mean approximates the truth. Second, we assess how each ensemble method performs in terms of its own optimization objective. Precisely, we report the data-misfit objective $\mathrm{J}_{\mathrm{DM}}\big(m(t)\big)$ for EKI and EKI-SL, and the Tikhonov-Phillips objective $\mathrm{J}_{\mathrm{TP}}\big(m(t)\big)$ for IEKF, TEKI and IEKF-SL. We run 10 trials for each algorithm, using different

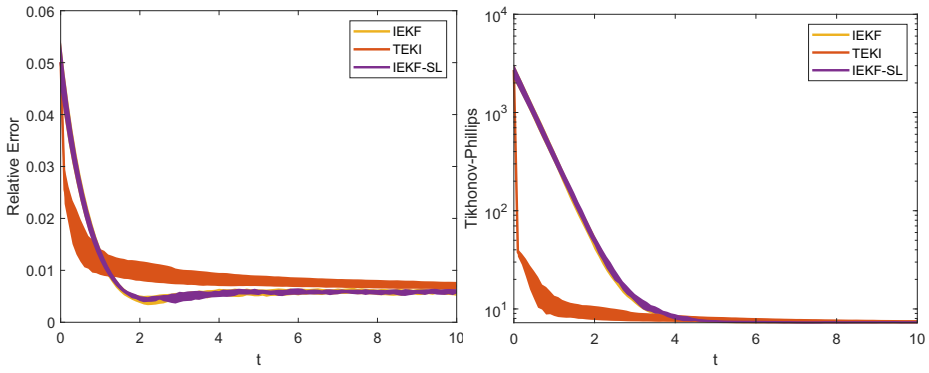


FIGURE 4. TEKI, IEKF & IEKF-SL: Relative errors and Tikhonov-Phillips objective w.r.t time t.

ensemble initializations (drawn from the same prior), and generate the error bars accordingly. Since this is a simple toy problem, all methods perform well. However, we note that the first iterations of EKI and TEKI reduce the objective faster than other algorithms.

5.2. **High-dimensional linear inverse problem.** In this subsection we consider a linear Bayesian inverse problem from [30]. This example illustrates the use of iterative ensemble Kalman methods in settings where both the size of the ensemble and the dimension of the data are significantly smaller than the dimension of the unknown parameter.

5.2.1. *Problem setup.* Consider the one dimensional elliptic equation

$$\begin{aligned} -\frac{\mathrm{d}^2 p}{\mathrm{d}x^2} + p &= u, \quad x \in (0,\pi), \\ p(0) &= p(\pi) = 0. \end{aligned} \tag{5.4}$$

We seek to recover u from noisy observation of p at $k = 2^4 - 1$ equispaced points $x_j = \frac{j}{2^4}\pi$. We assume that the data is generated from the model

$$y_j = p(x_j) + \eta_j, \quad j = 1, \dots, k, \tag{5.5}$$

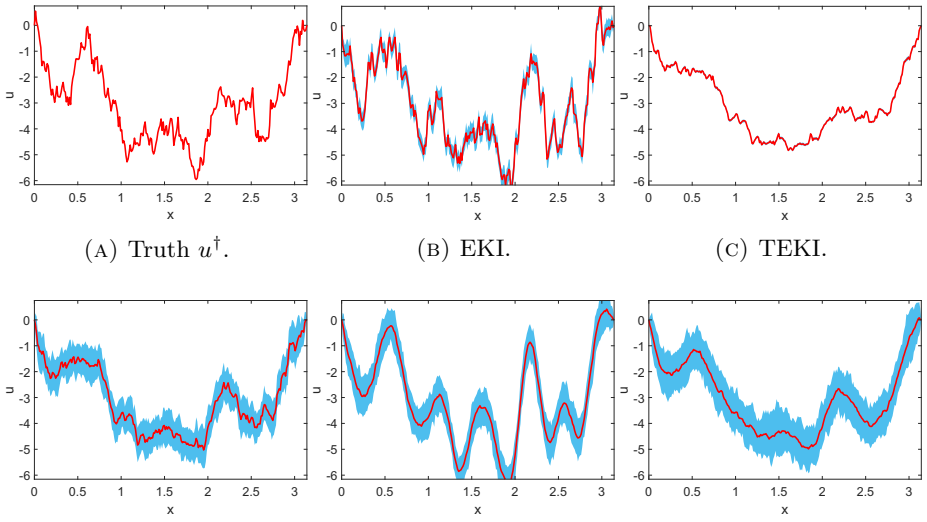
where $\eta_j \sim \mathcal{N}(0, \gamma^2)$ are independent. By defining $A = -\frac{\mathrm{d}^2}{\mathrm{d}x^2} + id$ and letting $\mathcal{O}$ be the observation operator defined by $\big(\mathcal{O}(p)\big)_j = p(x_j)$, we can rewrite (5.5) as

$$y = h(u) + \eta, \quad \eta \sim \mathcal{N}(0, \gamma^2 I_k),$$

where $h = \mathcal{O} \circ A^{-1}$. The forward problem (5.4) is solved on a uniform mesh with meshwidth $w = 2^{-8}$ by a finite element method with continuous, piecewise linear basis functions. We assume that the unknown parameter u has a Gaussian prior distribution, $u \sim \mathcal{N}(0, C_0)$ with covariance operator $C_0 = 10(A - id)^{-1}$ with homogeneous Dirichlet boundary conditions. This prior can be interpreted as the law of a Brownian bridge between 0 and π. For computational purposes we view u as a random vector in $\mathbf{R}^{2^8}$ and the linear map $h(u)$ is represented by a matrix $H \in \mathbf{R}^{(2^4-1)\times 2^8}$. The true parameter $u^\dagger$ is sampled from this prior (cf. Figure 5), and is used to generate synthetic observation data $y = Hu^\dagger + \eta$ with noise level $\gamma = 0.01$.

5.2.2. *Implementation details and numerical results.* We set the ensemble size to be $N = 50$. The initial ensemble $\{u_0^{(n)}\}_{n=1}^N$ is drawn independently from the prior. The length-step α is fixed to be 0.05 for all methods. We run each algorithm up to time $T = 30$, which corresponds to 600 iterations.

We note that the linear inverse problem considered here has input dimension $2^8 = 256$, which is much larger than the ensemble size N. Combining the results from Figure 5, 6 and 7, EKI and EKI-SL clearly overfit the data. In contrast, TEKI over-regularizes the data, and we easily notice the ensemble collapse. The IEKF and IEKF-SL lie in between, and approximate the truth slightly better.



(A) Truth $u^\dagger$. (B) EKI. (C) TEKI.

(D) IEKF. (E) EKI-SL. (F) IEKF-SL.

FIGURE 5. Ensemble mean (red) at the final iteration, with 10, 90-quantiles (blue).

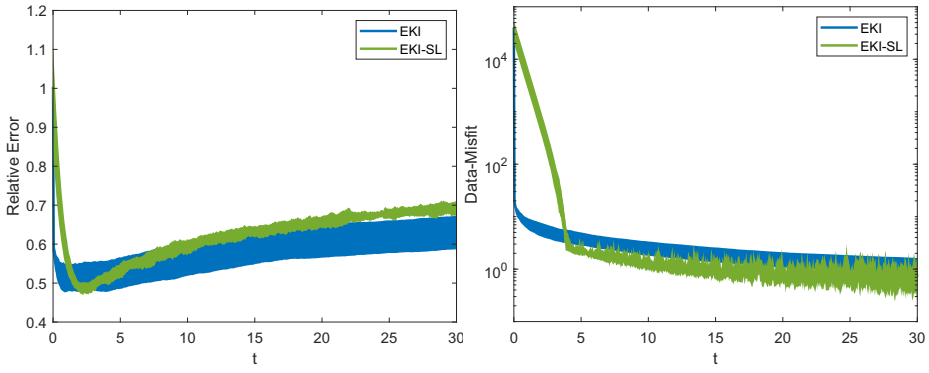


FIGURE 6. EKI & EKI-SL: Relative errors and data misfit w.r.t time t.

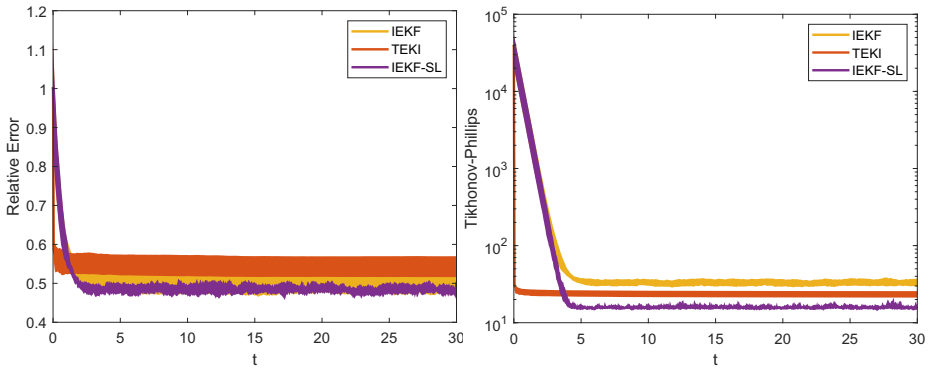


FIGURE 7. IEKF, TEKI & IEKF-SL: Relative errors and Tikhonov-Phillips objective w.r.t time t.

5.3. **Lorenz-96 model.** In this subsection we investigate the use of iterative ensemble Kalman methods to recover the initial condition of the Lorenz-96 system from partial and noisy observation of the solution at two positive times. The experimental framework is taken from [14] and is illustrative of the use of iterative ensemble Kalman methods in numerical weather forecasting.

5.3.1. *Problem setup.* Consider the dynamical system

$$\begin{aligned} \frac{dz_l}{dt} &= z_{l-1}(z_{l+1} - z_{l-2}) - z_l + F, \quad l = 1, \dots, d, \\ z_0 &= z_d, \quad z_{d+1} = z_1, \quad z_{-1} = z_{d-1}. \end{aligned} \tag{5.6}$$

Here z_l denotes the l^{th} coordinate of the current state of the system and F is a constant forcing with default value of $F = 8$. The dimension d is often chosen as 40. We want to recover the initial condition

$$u := (z_1(0), \dots, z_d(0))^T$$

of (5.6) from noisy partial measurements at discrete times $\{s_i\}_{i=1}^I$:

$$y_{i,j} = u_{l_j}(s_i) + \eta_{i,j}, \tag{5.7}$$

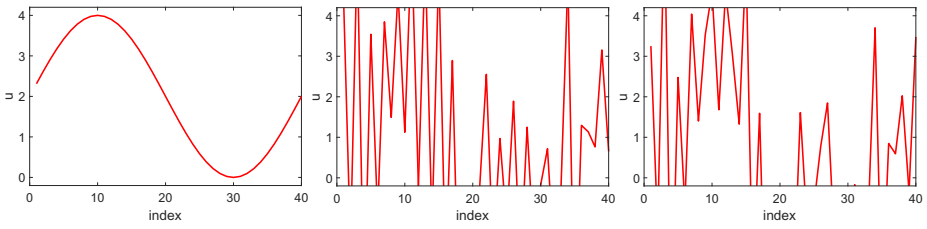
where $\{l_j\}_{j=1}^J \subset \{1, 2, \dots, d\}$ is the subset of observed coordinates, and $\eta_{i,j} \sim \mathcal{N}(0, \gamma^2)$ are assumed to be independent. In our numerical experiments we set $I = 2$, $J = 20$, $\{s_1, s_2\} = \{0.3, 0.6\}$, $\{l_j\}_{j=1}^{20} = \{1, 3, 5, \dots, 39\}$. We set the prior on u to be a Gaussian $\mathcal{N}(0, 2I_d)$. The true parameter $u^\dagger \in \mathbf{R}^{40}$ is shown in Figure 8, and is used to generate the observation data $\{y_{i,j}\}$ according to (5.7) with noise level $\gamma = 0.01$.

5.3.2. *Implementation details and numerical results.* We set the ensemble size to be $N = 50$. The initial ensemble $\{u_0^{(n)}\}_{n=1}^N$ is drawn independently from the prior. The length-step α is fixed to be 0.05 for all methods. We run each algorithm up to time $T = 30$, which corresponds to 600 iterations.

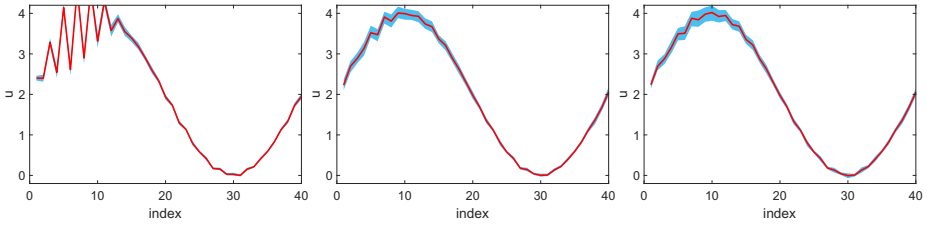
This is a moderately high dimensional nonlinear problem, where the forward model involves a black-box solver of differential equations. Figure 8 clearly indicates an ensemble collapse for the EKI and TEKI methods. Although they are still capable of finding a descending direction of the loss direction (see Figures 9 and 10), iterates get stuck in a local minimum. In contrast, we can see the advantage

of IEKF, EKI-SL and TEKI-SL in this setting: intuitively, a broader spread of the ensemble allows these methods to 'explore' different regions in the input space, and thereby to find a solution with potentially lower loss.

We notice that in practice 'covariance inflation' is often applied in EKI and TEKI algorithms, by manually injecting small random noise after each ensemble update. In general, this technique will 'force' a non-zero empirical covariance, prevent ensemble collapse and boost the performence of EKI and TEKI. However, the amount of noise injected is an additional hyperparameter that should be chosen manually, which should depend on the input dimension, observation noise, etc. Here we give a fair comparison of the different methods introduced, under the same time-stepping setting, with as few hyperparameters as possible.



(A) Truth $u^\dagger$. (B) EKI. (C) TEKI.



(D) IEKF. (E) EKI-SL. (F) IEKF-SL.

FIGURE 8. Ensemble mean (red) at the final iteration, with 10, 90-quantiles (blue).

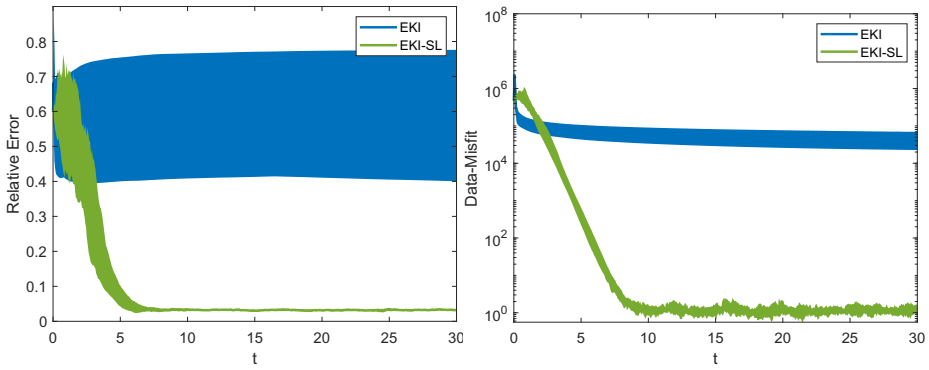


FIGURE 9. EKI & EKI-SL: Relative errors and data misfit w.r.t time t.

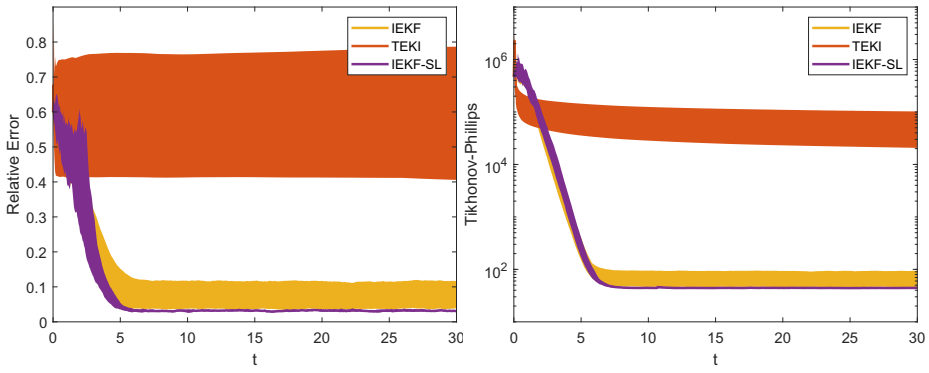


FIGURE 10. IEKF, TEKI & IEKF-SL: Relative errors and Tikhonov-Phillips objective w.r.t time t.

5.4. **High-dimensional nonlinear regression.** In this subsection we consider a nonlinear regression problem with a highly oscillatory forward map introduced in [26], where the authors investigate the use of iterative ensemble Kalman methods to train neural networks without back propagation.

5.4.1. *Problem setup.* We consider a nonlinear regression problem

$$y = h(u) + \eta, \quad \eta \sim \mathcal{N}(0, \gamma^2 I_k),$$

where $u \in \mathbf{R}^d, y \in \mathbf{R}^k$, and h is defined by

$$h(u) := Au + \sin(cBu), \tag{5.8}$$

where $A, B \in \mathbf{R}^{k\times d}$ are random matrices with independent $\mathcal{N}(0,1)$ entries. We set $d = 200$, $k = 150$ and $c = 20$. We want to recover u from y. We assume that the unknown u has a Gaussian prior $u \sim \mathcal{N}(0, 4I_d)$. The true parameter $u^\dagger$ is set to be $2 \cdot \mathbf{1}$, where $\mathbf{1}$ is the all-one vector. Observation data is generated as $y = h(u^\dagger) + \eta$.

By definition, h is highly oscillatory, and we may expect that the loss function, either Tikhonov-Phillips or data misfit objective, will have many local minima. Figure 11 visualizes the Tihonov-Phillips objective function $\mathsf{J}_{\mathrm{TP}}(u)$ with respect to two randomly choosen coordinates while other coordinates are fixed to value of 2. The data misfit objective exhibits a similar behavior.

5.4.2. *Implementation details and numerical results.* We set the ensemble size to be $N = 200$. The initial ensemble $\{u_0^{(n)}\}_{n=1}^N$ is drawn independently from the prior. The length-step α is fixed to be 0.05 for all methods. We set $\gamma = 0.01$ for the observation noise. We run each algorithm up to time $T = 30$, which corresponds to 600 iterations.

We notice that this is a difficult problem, due to its high dimensionality and nonlinearity. All methods except for IEKF-SL are not capable of reconstructing the truth. In particular, from Figures 12, 13 and 14, both EKI and TEKI do a poor job with relative error larger than 1, while IEKF and EKI-SL have slightly better performance. It is worth noticing from Figure 14 that IEKF-SL has a larger Tikhonov-Phillips objective function, with a much lower relative error. This may suggest that other types of regularization objectives can be used, other that the Tikhonov-Phillips objective, to solve this problem.

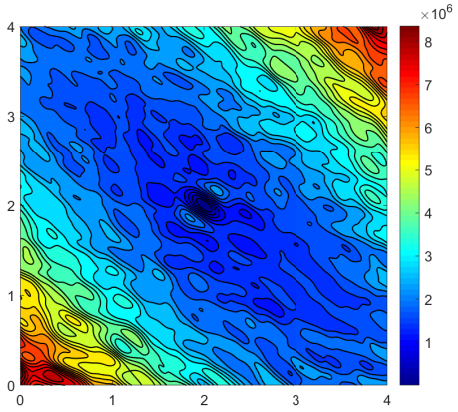
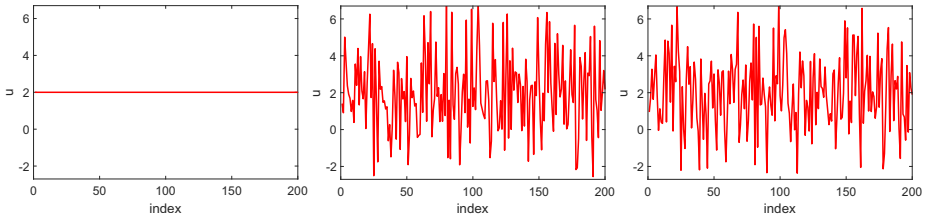
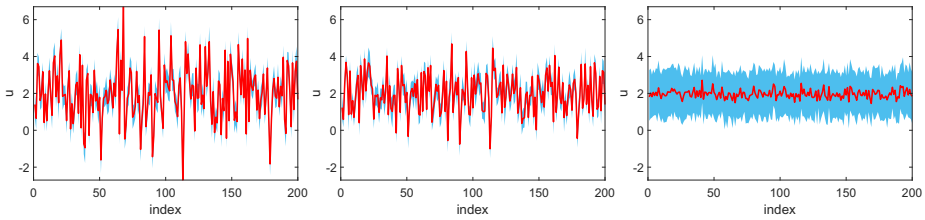


FIGURE 11. Tikhonov-Phillips objective function with respect to two randomly chosen coordinates.



(A) Truth $u^\dagger$. (B) EKI. (C) TEKI.



(D) IEKF. (E) EKI-SL. (F) IEKF-SL.

FIGURE 12. Ensemble mean (red) at the final iteration, with 10, 90-quantiles (blue).

6. **Conclusions and open directions.** In this paper we have provided a unified perspective of iterative ensemble Kalman methods and introduced some new variants. These new variants include a statistical linearization of both EKI and the IEnKF. Our numerical experiments suggest that the IEnKF-SL does not suffer from the overfitting of data. Furthermore and more interestingly, that for high-dimensional nonlinear problems, the IEnKF-SL may outperform other known methodologies. This is a promising result which has potential for other highly nonlinear inverse problems. However, for linear inverse problems, all variants discussed, new and known, perform well. As stated the advantage of such new variants is that

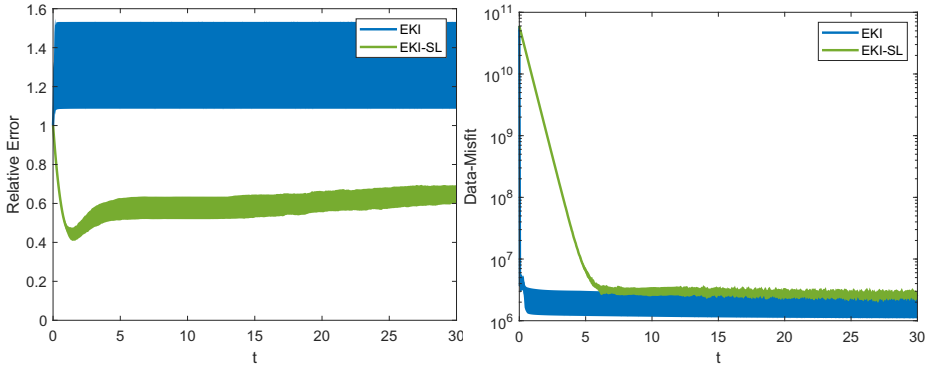


FIGURE 13. EKI & EKI-SL: Relative errors and data misfit w.r.t time t.

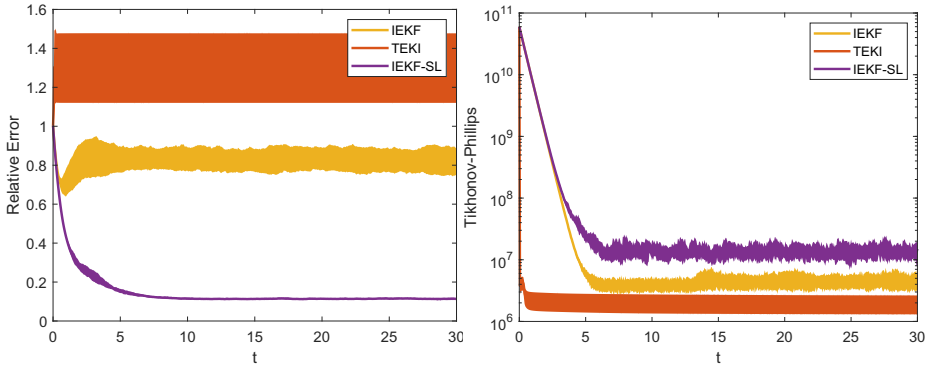


FIGURE 14. IEKF, TEKI & IEKF-SL: Relative errors and Tikhonov-Phillips objective w.r.t time t.

no parameter tuning is required. We hope that our work will stimulate further research in this active area, and we conclude with a list of some open directions.

- Continuum limits have been formally derived in our work. The rigorous derivation and analysis of SDE continuum limits, possibly in nonlinear settings, deserves further research.
- We have advocated the analysis of continuum limits for the understanding and design of iterative ensemble Kalman methods, but it would also be desirable to develop a framework for the analysis of discrete, implementable algorithms, and to further understand the potential benefits of various discretizations of a given continuum SDE system.
- From a theoretical viewpoint, it would be desirable to further analyze the convergence and stability of iterative ensemble methods with small or moderate ensemble size. While mean-field limits can be revealing, in practice the ensemble size is often not sufficiently large to justify the mean-field assumption. It would also be important to further analyze these questions in mildly nonlinear settings.
- An important methodological question, still largely unresolved, is the development of adaptive and easily implementable line search schemes and stopping

criteria for ensemble-based optimization schemes. An important work in this direction is [32].

- Another avenue for future methodological research is the development of iterative ensemble Kalman methods that are sparse-promoting, considering alternative regularizations beyond the least-squares objectives discussed in our paper [38, 40, 57].
- Ensemble methods are cheap in comparison to derivative-based optimization methods and Markov chain Monte Carlo sampling algorithms. It is thus natural to use ensemble methods to build ensemble preconditioners or surrogate models to be used within more expensive but accurate computational approaches.
- Finally, a broad area for further work is the application of iterative ensemble Kalman filters to new problems in science and engineering.

Acknowledgments. NKC is supported by KAUST baseline funding. DSA is thankful for the support of NSF and NGA through the grant DMS-2027056. The work of DSA was also partially supported by the NSF Grant DMS-1912818/1912802.

REFERENCES

[1] S. I. Aanonsen, G. Nævdal, D. S. Oliver, A. C. Reynolds and B. Vallès, The ensemble Kalman filter in reservoir engineering–A review, *SPE J.*, **14** (2009), 393–412.
[2] D. J. Albers, P.-A. Blancquart, M. E. Levine, E. Esmaeilzadeh Seylabi and A. Stuart, Ensemble Kalman methods with constraints, *Inverse Problems*, **35** (2019), 28pp.
[3] B. D. O. Anderson and J. B. Moore, *Optimal Filtering*, Information and System Sciences Series, (1979).
[4] J. L. Anderson, An ensemble adjustment Kalman filter for data assimilation, *Monthly Weather Review*, **129** (2001), 2884–2903.
[5] B. M. Bell, The iterated Kalman smoother as a Gauss–Newton method, *SIAM J. Optim.*, **4** (1994), 626–636.
[6] B. M. Bell and F. W. Cathey, The iterated Kalman filter update as a Gauss-Newton method, *IEEE Trans. Automat. Control*, **38** (1993), 294–297.
[7] D. P. Bertsakas, Incremental least squares method and the extended Kalman filter, *SIAM J. Optim*, **6** (1996), 807–822.
[8] D. Blömker, C. Schillings and P. Wacker, A strongly convergent numerical scheme from ensemble Kalman inversion, *SIAM J. Numer. Anal.*, **56** (2018), 2537–2562.
[9] D. Blömker, C. Schillings, P. Wacker and S. Weissmann, Well posedness and convergence analysis of the ensemble Kalman inversion, *Inverse Problems*, **35** (2019), 32pp.
[10] N. K. Chada, Analysis of hierarchical ensemble Kalman inversion, preprint, `arXiv:1801.00847`.
[11] N. K. Chada, M. A. Iglesias, L. Roininen and A. M. Stuart, Parameterizations for ensemble Kalman inversion, *Inverse Problems*, **34** (2018), 31pp.
[12] N. K. Chada, C. Schillings and S. Weissmann, On the incorporation of box-constraints for ensemble Kalman inversion, *Foundations of Data Science*, **1** (2019), 433–456.
[13] N. K. Chada, A. M. Stuart and X. T. Tong, Tikhonov regularization within ensemble Kalman inversion, *SIAM J. Numer. Anal.*, **58** (2020), 1263–1294.
[14] N. K. Chada and X. T. Tong, Convergence acceleration of ensemble Kalman inversion in nonlinear settings, preprint, `arXiv:1911.02424`.
[15] Y. Chen and D. S. Oliver, Ensemble randomized maximum likelihood method as an iterative ensemble smoother, *Mathematical Geosciences*, **44** (2012), 1–26.
[16] J. E. Dennis Jr. and R. B. Schnabel, Numerical methods for unconstrained optimization and nonlinear equations, Classics in Applied Mathematics, 16, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1996.
[17] Z. Ding and Q. Li, Ensemble Kalman sampler: Mean-field limit and convergence analysis, *SIAM J. Math. Anal.*, **53** (2021), 1546–1578.

[18] A. A. Emerick and A. C. Reynolds, Ensemble smoother with multiple data assimilation, *Computers & Geosciences*, **55** (2013), 3–15.
[19] O. G. Ernst, B. Sprungk and H.-J. Starkloff, Analysis of the ensemble and polynomial chaos Kalman filters in Bayesian inverse problems, *SIAM/ASA J. Uncertain. Quantif.*, **3** (2015), 823–851.
[20] G. Evensen, *Data Assimilation. The Ensemble Kalman Filter*, Springer-Verlag, Berlin, 2009.
[21] G. Evensen and P. J. Van Leeuwen, Assimilation of geosat altimeter data for the Agulhas current using the ensemble Kalman filter with a quasigeostrophic model, *Monthly Weather Review*, **124** (1996), 85–96.
[22] A. Garbuno-Inigo, F. Hoffmann, W. Li and A. M. Stuart, Interacting Langevin diffusions: Gradient structure and ensemble Kalman sampler, *SIAM J. Appl. Dyn. Syst.*, **19** (2020), 412–441.
[23] I. Grooms, A note on the formulation of the Ensemble Adjustment Kalman Filter, preprint, arXiv:2006.02941.
[24] Y. Gu and D. S. Oliver, An iterative ensemble Kalman filter for multiphase fluid flow data assimilation, *Spe Journal*, **12** (2007), 438–446.
[25] P. A. Guth, C. Schillings and S. Weissmann, Ensemble Kalman filter for neural network based one-shot inversion, preprint, arXiv:2005.02039.
[26] E. Haber, F. Lucka and L. Ruthotto, Never look back - A modified EnKF method and its application to the training of neural networks without back propagation, preprint, arXiv:1805.08034.
[27] M. Hanke, A regularizing Levenberg-Marquardt scheme, with applications to inverse groundwater filtration problems, *Inverse Problems,* **13** (1997), 79–95.
[28] M. Herty and G. Visconti, Kinetic methods for inverse problems, *Kinet. Relat. Models*, **12** (2019), 1109–1130.
[29] M. A. Iglesias, A regularising iterative ensemble Kalman method for PDE-constrained inverse problems, *Inverse Problems*, **32** (2016), 45pp.
[30] M. A. Iglesias, K. J. H. Law and A. M. Stuart, Ensemble Kalman methods for inverse problems, *Inverse Problems*, **29** (2013), 20pp.
[31] M. A. Iglesias, K. Lin, S. Lu and A. M. Stuart, Filter based methods for statistical linear inverse problems, *Commun. Math. Sci.*, **15** (2017), 1867–1895.
[32] M. A. Iglesias and Y. Yang, Adaptive regularisation for ensemble Kalman inversion, *Inverse Problems*, **37** (2021), 40pp.
[33] A. H. Jazwinski, *Stochastic Processes and Filtering Theory*, Courier Corporation, 2007.
[34] R. E. Kalman, A new approach to linear filtering and prediction problems, *Trans. ASME Ser. D. J. Basic Engrg.*, **82** (1960), 35–45.
[35] E. Kalnay, S. K. Park, Z.-X. Pu and J. Gao, Application of the quasi-inverse method to data assimilation, *Monthly Weather Review*, **128** (2000), 864–875.
[36] B. Katltenbacher, A. Neubauer and O. Scherzer, *Iterative Regularization Methods for Nonlinear Ill-Posed Problems*, Radon Series on Computational and Applied Mathematics, 6, Walter de Gruyter GmbH & Co. KG, Berlin, 2008.
[37] D. T. B. Kelly, K. J. H. Law and A. M. Stuart, Well-posedness and accuracy of the ensemble Kalman filter in discrete and continuous time, *Nonlinearity*, **27** (2014), 2579–5604.
[38] N. B. Kovachki and A. M. Stuart, Ensemble Kalman inversion: A derivative-free technique for machine learning tasks, *Inverse Problems*, **35** (2019), 35pp.
[39] W. G. Lawson and J. A. Hansen, Implications of stochastic and deterministic filters as ensemble-based data assimilation methods in varying regimes of error growth, *Monthly Weather Review*, **132** (2004), 1966–1981.
[40] Y. Lee, l_p regularization for ensemble Kalman inversion, preprint, arXiv:2009.03470.
[41] G. Li and A. C. Reynolds, An iterative ensemble Kalman filter for data assimilation, SPE Annual Technical Conference and Exhibition, Society of Petroleum Engineers, Anaheim, CA, 2007.
[42] A. C. Lorenc, Analysis methods for numerical weather prediction, *Quart. J. Roy. Meteor. Soc.*, **112** (1986), 1177–1194.
[43] R. J. Lorentzen, K.-K. Fjelde, J. Froyen, A. C. V. M. Lage, G. Nævdal and E. H. Vefring, Underbalanced drilling: Real time data interpretation and decision support, SPE/IADC Drilling Conference, Amsterdam, Netherlands, 2001.
[44] S. Lu and S. V. Pereverzev, Multi-parameter regularization and its numerical realization, *Numer. Math.*, **118** (2001), 1–31.

[45] A. J. Madja and J. Harlim, *Filtering Complex Turbulent Systems*, Cambridge University Press, Cambridge, 2012.
[46] J. Mandel, E. Bergou and S. Gratton, 4DVAR by ensemble Kalman smoother, preprint, arXiv:1304.5271.
[47] G. Nævdal, T. Mannseth and E. H Vefring, Instrumented wells and near-well reservoir monitoring through ensemble Kalman filter, *Proceedings of 8th European Conference on the Mathematics of Oil Recovery, Freiberg, Germany*, 2001.
[48] A. S. Nemirovsky and D. B. Yudin, *Problem Complexity and Method Efficiency in Optimization*, Wiley-Interscience Series in Discrete Mathematics, John Wiley & Sons, Inc., New York, 1983.
[49] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer Series in Operations Research and Financial Engineering, Springer, New York, 2006.
[50] D. Oliver, A. C. Reynolds and N. Liu, *Inverse Theory for Petroleum Reservoir Characterization and History Matching*, Cambridge University Press, 2008.
[51] S. Reich and C. J. Cotter, Ensemble filter techniques for intermittent data assimilation, in *Large Scale Inverse Problems*, Radon Ser. Comput. Appl. Math., 13, De Gruyter, Berlin, 2013, 91–134.
[52] A. C. Reynolds, M. Zafari and G. Li, Iterative forms of the ensemble Kalman filter, ECMOR X-10th European Conference on the Mathematics of Oil Recovery, 2006.
[53] P. Sakov, D. S. Oliver and L. Bertino, An Iterative EnKF for strongly nonlinear systems, *Monthly Weather Review*, **140** (2012), 1988–2004.
[54] D. Sanz-Alonso, A. Stuart and A. Taeb, Data assimilation and inverse problems, preprint, arXiv:1810.06191.
[55] C. Schillings and A. M. Stuart, Analysis of the ensemble Kalman filter for inverse problems, *SIAM J. Numer. Anal.*, **55** (2017), 1264–1290.
[56] C. Schillings and A. M. Stuart, Convergence analysis of ensemble Kalman inversion: The linear, noisy case, *Appl. Anal.*, **97** (2018), 107–123.
[57] T. Schneider, A. M. Stuart and J.-L. Wu, Ensemble Kalman inversion for sparse learning of dynamical systems from time-averaged data, preprint, arXiv:2007.06175.
[58] B. Shi, S. S. Du, M. I. Jordan and W. J. Su, Understanding the acceleration phenomenon via high-resolution differential equations, preprint, arXiv:1810.08907.
[59] J. A. Skjervheim, G. Evensen, J. Hove and J. G. Vabø, An ensemble smoother for assisted history matching, SPE Reservoir Simulation Symposium, The Woodlands, TX, 2001.
[60] W. Su, S. Boyd and E. J. Candès, A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights, *J. Mach. Learn. Res.*, **17** (2016), 43pp.
[61] A. Tarantola, *Inverse Problem Theory and Methods for Model Parameter Estimation*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2005.
[62] M. K. Tippett, J. L. Anderson, C. H. Bishop, T. M. Hamill and J. S. Whitaker, Ensemble square root filters, *Monthly Weather Review*, **131** (2003), 1485-1490.
[63] S. Ungarala, On the iterated forms of Kalman filters using statistical linearization, *J. Process Control*, **22** (2012), 935–943.
[64] A. Wibisono, A. C. Wilson and M. I. Jordan, A variational perspective on accelerated methods in optimization, *Proc. Natl. Acad. Sci. USA*, **113** (2016), E7351–E7358.

Received October 2020; revised January 2021.

E-mail address: neilchada123@gmail.com
E-mail address: ymchen@uchicago.edu
E-mail address: sanzalonso@uchicago.edu