COMMUN. MATH. SCI.
Vol. 19, No. 7, pp. 1855–1880

ON THE CONTINUOUS TIME LIMIT OF ENSEMBLE SQUARE ROOT FILTERS*

THERESA LANGE† AND WILHELM STANNAT‡

Abstract. We provide a continuous time limit analysis for the class of ensemble square root filter algorithms with deterministic model perturbations. In the particular linear case, we specify general conditions on the model perturbations implying convergence of the empirical mean and covariance matrix towards their respective counterparts of the Kalman-Bucy filter. As a second main result we identify additional assumptions for the convergence of the whole ensemble towards solutions of the ensemble Kalman-Bucy filtering equations introduced in [J. de Wiljes, S. Reich, and W. Stannat, SIAM J. Appl. Dyn. Syst., 17(2):1152–1181, 2018]. The latter result can be generalized to nonlinear Lipschitz-continuous model operators. A striking implication of our results is the fact that the limiting equations for the ensemble members are universal for a large class of ensemble square root filters. This yields a mathematically rigorous justification for the analysis of these algorithms with the help of the ensemble Kalman-Bucy filter.

Keywords. Continuous time limit; Ensemble square root filter; Deterministic model perturbations.

AMS subject classifications. 60H35; 93E11; 60F99.

1. Introduction

Consider the optimal filtering problem in continuous time which consists of estimating the current state of a diffusion process

$$\mathrm{d}X_t = f(X_t)\,\mathrm{d}t + Q^{\frac{1}{2}}\,\mathrm{d}W_t, \qquad X_t \in \mathbb{R}^d, \tag{1.1}$$

using observations

$$\mathrm{d}Y_t = g(X_t)\,\mathrm{d}t + C^{\frac{1}{2}}\,\mathrm{d}V_t, Y_0 = 0, \qquad Y_t \in \mathbb{R}^p. \tag{1.2}$$

The processes W and V are independent standard Brownian motions, Q and C positive definite matrices, and f and g assumed to be Lipschitz-continuous. The solution of this problem is given by the posterior mean $\int x\pi_t(\mathrm{d}x)$, where

$$\pi_t(\mathrm{d}x) = \mathbb{P}[X_t \in \mathrm{d}x | \mathcal{Y}_t] \tag{1.3}$$

is the conditional distribution of X_t given $\mathcal{Y}_t := \{Y_s : s \le t\}$. In the last few decades a hoard of algorithms has been proposed to specify or approximate π_t.

In practice, however, observations are accumulated discretely in time rather than continuously. In a typical scenario, one observes a sequence of observations $Y_{t_k}, k = 1,2,3,...$, where $t_{k+1} = t_k + h$ for some $h > 0$, for which one solves the corresponding discrete-time filtering problem.

In the particular case of linear operators $f(x) = Ax$ and $g(x) = Gx$, this problem can be solved with the famous Kalman filter [7] computing the mean $\bar{x}$ and covariance matrix

*Received: March 25, 2020; Accepted (in revised form): April 02, 2021. Communicated by Grigorios A. Pavliotis.

†Fakultät für Mathematik, Universität Bielefeld, D-33501 Bielefeld, Germany (tlange@math.uni-bielefeld.de). https://ekvv.uni-bielefeld.de/pers_publ/publ/PersonDetail.jsp?personId=265230835

‡Institut für Mathematik, Technische Universität Berlin, Straße des 17. Juni 136, D-10623 Berlin and Bernstein Center for Computational Neuroscience, Philippstr. 13, D-10115 Berlin, Germany (stannat@math.tu-berlin.de).

P of the, in this case, Gaussian π according to the following recursion: given a current estimate $\bar{x}^a_{k-1}$, $\bar{x}^a_{k-1}$ is propagated forward according to the system equation to yield the forecast $\bar{x}^f_k$. If at time t_k, an observation Y_{t_k} is available, this will be used to update the current forecast to yield an improved estimate $\bar{x}^a_k$. When applied to the continuous-time setting, one uses (in the simplest case) the Euler-Maruyama time-discretization of (1.1) in the forecast step and updates each forecast using the observations in the form of $\Delta Y_k := Y_{t_k} - Y_{t_{k-1}}$. The precise evolution equations then read as follows:

Forecast:

$$\bar{x}^f_k = \bar{x}^a_{k-1} + hA\bar{x}^a_{k-1} \tag{1.4}$$

$$P^f_k = (\mathrm{Id} + hA)\, P^a_{k-1} (\mathrm{Id} + hA)^T + Q \tag{1.5}$$

Update:

$$\bar{x}^a_k = \bar{x}^f_k + K_k \left(\Delta Y_k - hG\bar{x}^f_k \right) \tag{1.6}$$

$$P^a_k = (\mathrm{Id} - hK_kG)\, P^f_k \tag{1.7}$$

$$K_k = P^f_k G^T \left(C + hGP^f_k G^T \right)^{-1} \tag{1.8}$$

It is well known that the Kalman filter admits a continuous-time analogue, the Kalman-Bucy filter [8], given by

$$\bar{x}_t = A\bar{x}_t \mathrm{d}t + P_t G^T C^{-1} (\mathrm{d}Y_t - G\bar{x}_t \mathrm{d}t) \tag{1.9}$$

$$\mathrm{d}P_t = \left(AP_t + P_t A^T + Q - P_t G^T C^{-1} G P_t \right) \mathrm{d}t. \tag{1.10}$$

In the nonlinear case, however, calculating the exact π_t is in general not possible necessitating approximative schemes. Ensemble Kalman filters (EnKF) form a class of second-order accurate Monte-Carlo algorithms approximating the conditional mean and covariance matrix with the help of the empirical mean and covariance matrix of an ensemble, and propagating the ensemble according to the nonlinear counterpart of the Kalman filtering equations. In the case of the popular stochastic EnKF (cf. [4,6]), for instance, each ensemble member is propagated according to

$$X^{(i),f}_{t_k} = X^{(i),a}_{t_{k-1}} + hf\left(X^{(i),a}_{t_{k-1}} \right) + Q^{\frac{1}{2}} \tilde{W}^{(i)}_k, \tag{1.11}$$

$$X^{(i),a}_{t_k} = X^{(i),f}_{t_k} + K_k \left(\Delta Y^{(i)}_k - hg\left(X^{(i),f}_{t_k} \right) \right), \qquad \Delta Y^{(i)}_k = \Delta Y_k + C^{\frac{1}{2}} \tilde{V}^{(i)}_k, \tag{1.12}$$

where $\tilde{W}^{(i)}_k, \tilde{V}^{(i)}_k \sim \mathcal{N}(0, h\mathrm{Id})$ are independent samples.

In our previous work [9], we were able to show the existence of a continuous time limit $h \to 0$ of (1.11) and (1.12) in the case of f and g being Lipschitz-continuous and bounded. Furthermore, we proved an even stonger convergence result in the case of a modified algorithm inspired by [15], replacing $\tilde{W}^{(i)}$ and $\tilde{V}^{(i)}$ by suitable deterministic perturbations. The filter in [15] is a so-called deterministic EnKF and the aim of this paper now is to generalize the latter result to the class of these filtering algorithms, in particular to the class of ensemble square root filters (ESRF).

This class of algorithms has been introduced in order to replace the additional noise $\tilde{V}^{(i)}$ to the observations, used in the stochastic EnKF to avoid that the empirical covariance matrix underestimates the true error covariance (cf. [4]). ESRF are widely

used in the geosciences since they were shown to numerically perform better than their stochastic counterpart (see e.g. [12,16,17]). The most popular ESRF algorithms are the ensemble adjustment Kalman filter (EAKF, see [1]), the ensemble transform Kalman filter (ETKF, see [3]), and the unperturbed EnKF (Whitaker, Hamill (2002), see [19]), as summarized in the survey paper [17].

The idea of the ESRF algorithms is the following: let $E_k^f := \left[X_{t_k}^{(i),f} - \bar{x}_k^f\right]_{i=1,\dots,M}$ denote the matrix of forecast deviations such that $P_k^f = \frac{1}{M-1} E_k^f \left(E_k^f\right)^T$ is the forecast covariance matrix. Then, in the case of linear observations, ESRF specify deterministic transformations of E_k^f such that the resulting covariance matrix P_k^a satisfies the Kalman Equation (1.7) exactly, i.e.

$$P_k^a \overset{!}{=} (\mathrm{Id} - hK_kG)\, P_k^f. \tag{1.13}$$

This paper is structured as follows: in the particular linear case and under appropriate assumptions on the deterministic model perturbations, we show in Section 3 that the ensemble equations lead to closed recursion formulas for the empirical mean and covariance matrix which, up to terms of order h^2, coincide with their respective Kalman filtering equations. Our main results concerning the convergence of the mean and covariance matrix towards their corresponding counterparts in the Kalman-Bucy filter are summarized in Theorem 3.1 and Theorem 3.2.

In Section 5, we then prove for the above three algorithms EAKF, ETKF, and Whitaker, Hamill (2002) the existence of the continuous time limit of the full ensemble $X_{t_k}^{(i),f}$ (respectively $X_{t_k}^{(i),a}$) towards the solution of the ensemble Kalman-Bucy filtering equations (cf. [2,5,14])

$$\mathrm{d}X_t^{(i)} = AX_t^{(i)}\mathrm{d}t + Q^{\frac{1}{2}}\hat{W}_t^{(i)}\mathrm{d}t + P_tG^TC^{-1}\left(\mathrm{d}Y_t - \frac{1}{2}G\left(X_t^{(i)} + \bar{x}_t\right)\mathrm{d}t\right) \tag{1.14}$$

as summarized in Theorem 5.1. Our analysis can be generalized to the case of nonlinear, Lipschitz-continuous model operators f. It is a striking fact that this continuous-time equation shows up as a universal limit of a broad class of deterministic filtering algorithms. As will be discussed in Section 6, this forms a powerful result in view of analyzing properties of the discrete-time counterparts.

1.1. Notation. In the following, we will abbreviate 'deterministic EnKF with deterministic model perturbations' by 'fully deterministic EnKF'. For any vector $x \in \mathbb{R}^n$ and matrix $A \in \mathbb{R}^{n\times m}$ let x^T resp. A^T denote the respective transpose. Further let $\|A\|_F$ denote the Frobenius norm and $\|A\|$ the operator norm of a matrix A. Also for a quadratic matrix A, let $\mathrm{tr}(A)$ denote the trace of A.

In the subsequent analysis we will use the notation $x_t \lesssim y_t$ for $x_t \leq Cy_t$ for some constant $C > 0$ independent of t (e.g. arising from the Cauchy-Schwarz inequality).

For the ensemble $\left\{X^{(i)}, i, ..., M\right\}$ of size M let

$$\bar{x} := \frac{1}{M}\sum_{i=1}^{M} X^{(i)} \tag{1.15}$$

denote the ensemble mean and

$$P := \frac{1}{M-1}\sum_{i=1}^{M}\left(X^{(i)} - \bar{x}\right)\left(X^{(i)} - \bar{x}\right)^T \tag{1.16}$$

the ensemble covariance matrix. Further define

$$\mathcal{V} := \frac{1}{M-1}\sum_{i=1}^{M}\left\|X^{(i)}-\bar{x}\right\|^2 = \mathrm{tr}(P). \tag{1.17}$$

As mentioned above, the discrete-time algorithms are carried out for the partition $0 = t_0 < \ldots < t_L = T$ with $t_{k+1} = t_k + h$, $h > 0$. In that case, if $t \in [t_k, t_{k+1})$ we use the notation

$$\eta(t) := t_k, \qquad \nu(t) := k, \qquad \eta_+(t) := t_{k+1} \qquad \nu_+(t) := k+1. \tag{1.18}$$

2. Deterministic model perturbations

As described in the introduction, the aim of ESRF algorithms is to transform the forecast ensemble in such a way that the covariance matrix of the resulting ensemble exactly satisfies (1.13) of the Kalman filter. The proposed transformations are deterministic in the sense that they do not introduce additional noise as in the case of the stochastic EnKF (see (1.12)). Observe that this way, ESRF algorithms less likely suffer from sampling errors in the analysis step. Inspired by the modified filter in [5], we reduce the total amount of noise in the algorithms even further by introducing deterministic model perturbations to replace the noise also in the forecast step. This ansatz can be interpreted as a form of inflation of the ensemble, a technique widely used in ensemble-based algorithms to increase the ensemble spread and prevent underestimation of the covariance matrix. Note that this way the only randomness present in the resulting algorithm comes from the initial ensemble and the observations.

In continuous time then, it seems to be natural to expect that the covariance of the continuous time limit, if it exists, satisfies (1.10) of the Kalman-Bucy filter. Using (1.13) and (1.11) with deterministic model perturbations of the form $hQ^{\frac{1}{2}}\hat{W}_k^{(i),h}$ the evolution equations of the forecast and update covariance matrix read as follows:

$$\begin{aligned} P_k^f = {} & (\mathrm{Id}+hA)\,P_{k-1}^a\,(\mathrm{Id}+hA)^T \\ & + \frac{h}{M-1}\sum_{i=1}^{M}(\mathrm{Id}+hA)^T\left(X_{t_{k-1}}^{(i),a}-\bar{x}_{k-1}^a\right)\left(\hat{W}_k^{(i),h}-\hat{w}_k^h\right)^T Q^{\frac{1}{2}} \\ & \qquad + Q^{\frac{1}{2}}\left(\hat{W}_k^{(i),h}-\hat{w}_k^h\right)\left(X_{t_{k-1}}^{(i),a}-\bar{x}_{k-1}^a\right)^T(\mathrm{Id}+hA)^T \\ & + \frac{h^2}{M-1}\sum_{i=1}^{M}Q^{\frac{1}{2}}\left(\hat{W}_k^{(i),h}-\hat{w}_k^h\right)\left(\hat{W}_k^{(i),h}-\hat{w}_k^h\right)^T Q^{\frac{1}{2}}, \end{aligned} \tag{2.1}$$

$$P_k^a = (\mathrm{Id}-hK_kG)\,P_k^f. \tag{2.2}$$

Therefore finding a choice of $\hat{W}_k^{(i),h}, i = 1,\ldots,M$, such that

$$P_k^f = P_{k-1}^f + h\left(AP_{k-1}^f + P_{k-1}^f A^T + Q - K_{k-1}GP_{k-1}^f\right) + O(h^2) \tag{2.3}$$

formally yields (1.10). Throughout the paper we assume the following properties of the model perturbations $\hat{W}_k^{(i),h}$:

Assumption 1. *It holds uniformly in* $k = 1,\ldots,L = \frac{T}{h}$

- $\frac{1}{M-1}\sum_{i=1}^{M}\left(X_{t_{k-1}}^{(i),a}-\bar{x}_{k-1}^a\right)\left(\hat{W}_k^{(i),h}-\hat{w}_k^h\right)^T Q^{\frac{1}{2}} = \frac{1}{2}Q$

- $\left\| \frac{1}{M-1} \sum_{i=1}^{M} Q^{\frac{1}{2}} \left(\hat{W}_k^{(i),h} - \hat{w}_k^h \right) \left(\hat{W}_k^{(i),h} - \hat{w}_k^h \right)^T Q^{\frac{1}{2}} \right\|_F \leq \kappa$
- *w.l.o.g.* $\hat{W}_k^{(i),h}$ *are centered, i.e.* $\hat{w}_k^h = 0$.

One can easily check that the resulting P_k^f satisfies the recursion (2.3) and as will be shown in the next section, converges to a continuous-time matrix-valued process satisfying (1.10).

EXAMPLE 2.1. *In [9], we discussed a fully deterministic EnKF using perturbations of the form*

$$\hat{W}_k^{(i),h} := \frac{1}{2} Q^{\frac{1}{2}} \left(P_{k-1}^a \right)^{-1} \left(X_{t_{k-1}}^{(i),a} - \bar{x}_{k-1}^a \right). \tag{2.4}$$

These satisfy Assumption 1. Indeed: it holds

$$\begin{aligned} &\frac{1}{M-1} \sum_{i=1}^{M} \left(X_{t_{k-1}}^{(i),a} - \bar{x}_{k-1}^a \right) \left(\hat{W}_k^{(i),h} - \hat{w}_k^h \right)^T Q^{\frac{1}{2}} \\ =&\frac{1}{2} P_{k-1}^a \left(P_{k-1}^a \right)^{-1} Q^{\frac{1}{2}} Q^{\frac{1}{2}} = \frac{1}{2} Q. \end{aligned} \tag{2.5}$$

Further it holds $\hat{w}_k^h = 0$ *and*

$$\frac{1}{M-1} \sum_{i=1}^{M} Q^{\frac{1}{2}} \left(\hat{W}_k^{(i),h} - \hat{w}_k^h \right) \left(\hat{W}_k^{(i),h} - \hat{w}_k^h \right)^T Q^{\frac{1}{2}} = \frac{1}{4} Q \left(P_{k-1}^a \right)^{-1} Q. \tag{2.6}$$

Furthermore, $\left\| \left(P_k^a \right)^{-1} \right\|_F$ *is bounded uniformly in* k *(see Appendix B).*

REMARK 2.1. One can replace Assumption 1 on the model perturbations $\hat{W}^{(i),h}$ by the following quadratic matrix equation

$$\begin{aligned} &\frac{1}{M-1} \sum_{i=1}^{M} (\mathrm{Id} + hA) \left(X_{t_{k-1}}^{(i),a} - \bar{x}_{k-1}^a \right) \left(\hat{W}_k^{(i),h} - \hat{w}_k^h \right)^T Q^{\frac{1}{2}} \\ &\qquad + Q^{\frac{1}{2}} \left(\hat{W}_k^{(i),h} - \hat{w}_k^h \right) \left(X_{t_{k-1}}^{(i),a} - \bar{x}_{k-1}^a \right)^T (\mathrm{Id} + hA)^T \\ &+ \frac{h}{M-1} \sum_{i=1}^{M} Q^{\frac{1}{2}} \left(\hat{W}_k^{(i),h} - \hat{w}_k^h \right) \left(\hat{W}_k^{(i),h} - \hat{w}_k^h \right)^T Q^{\frac{1}{2}} \\ =&Q + h\tilde{R}_k \end{aligned} \tag{2.7}$$

with rest term $\tilde{R}_k$ uniformly bounded in k. With the notation

$$\tilde{E}_{k-1}^a := \frac{1}{\sqrt{M-1}} (\mathrm{Id} + hA) \left[X_{t_{k-1}}^{(i),a} - \bar{x}_{k-1}^a \right]_{i=1,\dots,M}, \tag{2.8}$$

$$\mathcal{W}_k := \frac{1}{\sqrt{M-1}} Q^{\frac{1}{2}} \left[\hat{W}_k^{(i),h} - \bar{w}_k^h \right]_{i=1,\dots,M} \tag{2.9}$$

this yields the problem of solving

$$\tilde{E}_{k-1}^a \mathcal{W}_k^T + \mathcal{W}_k \left(\tilde{E}_{k-1}^a \right)^T + h \mathcal{W}_k \mathcal{W}_k^T = Q + h\tilde{R}_k. \tag{2.10}$$

In the particular case of $\tilde{R}_k \equiv 0$, if

$$\mathcal{W}_k = -\frac{1}{h}\tilde{E}^a_{k-1} \pm J_{k-1} \tag{2.11}$$

where J_{k-1} solves

$$J_{k-1}J_{k-1}^T = \frac{1}{h^2}(\mathrm{Id}+hA)P^a_{k-1}(\mathrm{Id}+hA)^T + \frac{1}{h}Q, \tag{2.12}$$

then solving (2.10) reduces to solving (2.12).

Looking back on Example 2.1, observe that (2.4) assumes $M \geq d+1$ necessary for P^a to be invertible. In (2.11), P^a need not have full rank since Q is already assumed to do so. However, for h small, the first term in (2.11) dominates and in case P^a is rank-deficient, it may cause numerical instabilities. Thus the case of $M \leq d$ needs to be treated with great care. Assuming $M \leq d$ in Example 2.1, the inverse in (2.4) should be replaced by the generalized inverse $\left(P^a_{k-1}\right)^\dagger$ yielding

$$\hat{W}^{(i),h}_k := \frac{1}{2}Q^{\frac{1}{2}}\left(P^a_{k-1}\right)^\dagger\left(X^{(i),a}_{t_{k-1}} - \bar{x}^a_{k-1}\right) \tag{2.13}$$

which gives in (2.1)

$$P^f_k = P^a_{k-1} + h\left(AP^a_{k-1} + P^a_{k-1}A^T + P^a_{k-1}\left(P^a_{k-1}\right)^\dagger Q + Q\left(P^a_{k-1}\right)^\dagger P^a_{k-1}\right) + O(h^2). \tag{2.14}$$

Motivated by this example, we presume that the quadratic matrix equation at the beginning of this remark should in general be replaced by imposing

$$(2.7) = \Pi_{k-1}Q + h\tilde{R}_k \tag{2.15}$$

where Π_{k-1} is the projection onto the span of the ensemble deviations such that

$$\begin{aligned} &\Pi_{k-1}\left(X^{(i),a}_{t_{k-1}} - \bar{x}^a_{k-1}\right) = X^{(i),a}_{t_{k-1}} - \bar{x}^a_{k-1}, \\ &\Pi_{k-1}v = 0 \qquad \forall v \in \mathrm{span}\left\{X^{(1),a}_{t_{k-1}} - \bar{x}^a_{k-1}, \ldots, X^{(M),a}_{t_{k-1}} - \bar{x}^a_{k-1}\right\}^\perp. \end{aligned} \tag{2.16}$$

Using stochastic perturbations instead of deterministic is another alternative in the case $M \leq d$ due to the regularizing effect of the noise. This shall, however, be discussed in a separate paper.

3. Continuous time limit I: mean and covariance matrix

Imposing Assumption 1 on the deterministic model perturbations $\hat{W}^{(i),h}$, we are now able to rigorously show that the resulting covariance process converges to a process P satisfying the Riccati Equation (1.10):

THEOREM 3.1. *Let P denote the continuous-time matrix-valued process satisfying (1.10) and let Assumption 1 hold. If*

$$\|P_0 - P^a_0\| \in O(h) \tag{3.1}$$

then

$$\sup_{t\in[0,T]}\left\|P_t - P^f_{\nu(t)}\right\| \in O(h) \qquad \textit{and} \qquad \sup_{t\in[0,T]}\left\|P_t - P^a_{\nu(t)}\right\| \in O(h). \tag{3.2}$$

First of all observe that $\|P_t\|$ is uniformly bounded in time on $[0,T]$. For the proof of Theorem 3.1, we further need the following lemma which is an immediate consequence of our above assumptions:

LEMMA 3.1. *Given Assumption 1, there exists a constant $0<p_T^{*,f}<\infty$ depending on $\|P_0^a\|$ and κ such that for all $k=1,\dots,L$ it holds*

$$\left\|P_k^f\right\| \le p_T^{*,f}. \tag{3.3}$$

This implies that also $\|K_k\|$ is uniformly bounded in k.

For the proof see Appendix A.1.

Proof. **(Proof of Theorem 3.1.)** It holds

$$\begin{aligned} P_{t_k} - P_k^f = P_0 - P_0^f \\ &+ \int_0^{t_k} A\left(P_s - P_{\nu(s)}^f\right) + \left(P_s - P_{\nu(s)}^f\right) A^T \\ &\quad - \left(P_s G^T C^{-1} G P_s - P_{\nu(s)}^f G^T \left(hGP_{\nu(s)}^f G^T + C\right)^{-1} GP_{\nu(s)}^f\right) \mathrm{d}s \\ &+ h^2 \sum_{j=0}^{k-1} R_j \end{aligned} \tag{3.4}$$

where

$$\begin{aligned} R_j := hAK_j GP_j^f A^T \\ &- \frac{1}{M-1}\sum_{i=1}^{M} Q^{\frac{1}{2}} \left(\hat{W}_{j+1}^{(i),h} - \hat{w}_{j+1}^h\right)\left(\hat{W}_{j+1}^{(i),h} - \hat{w}_{j+1}^h\right)^T Q^{\frac{T}{2}} \\ &- \left(AP_j^f A^T - AK_j GP_j^f - K_j GP_j^f A^T + AQ + QA^T\right). \end{aligned} \tag{3.5}$$

By Lemma 3.1 and Assumption 1 we obtain a constant $0<r_T^*<\infty$ such that $\|R_j\|<r_T^*$ for all $j=0,\dots,L-1$. Further it holds

$$\begin{aligned} &\left\|P_s G^T C^{-1} G P_s - P_{\nu(s)}^f G^T \left(hGP_{\nu(s)}^f G^T + C\right)^{-1} GP_{\nu(s)}^f\right\| \\ \le &\left\|P_s G^T C^{-1} G P_s - P_{\nu(s)}^f G^T C^{-1} GP_{\nu(s)}^f\right\| \\ &\quad + \left\|P_{\nu(s)}^f G^T \left(C^{-1} - \left(hGP_{\nu(s)}^f G^T + C\right)^{-1}\right) GP_{\nu(s)}^f\right\| \\ \le &\left\|P_s - P_{\nu(s)}^f\right\| \|G\|^2 \|C^{-1}\| \left(\|P_s\| + \left\|P_{\nu(s)}^f\right\|\right) \\ &\quad + \left\|P_{\nu(s)}^f\right\|^2 \|G\|^2 \left\|C^{-1} - \left(hGP_{\nu(s)}^f G^T + C\right)^{-1}\right\|. \end{aligned} \tag{3.6}$$

Using the Woodbury matrix identity, we can estimate

$$\left\|C^{-1} - \left(hGP_{\nu(s)}^f G^T + C\right)^{-1}\right\| \le h\left\|C^{-1}\right\|^2 \|G\|^2 \left\|P_{\nu(s)}^f\right\|. \tag{3.7}$$

Thus, again using Lemma 3.1, we obtain for a constant $\tilde{C} = \tilde{C}(T)$

$$\left\| P_{t_k} - P_k^f \right\| \leq \| P_0 - P_0^a \| + \tilde{C} \int_0^{t_k} \left\| P_s - P_{\nu(s)}^f \right\| \mathrm{d}s + hTr_T^*. \tag{3.8}$$

Since P_t satisfies the Riccati Equation (1.10), one can further show that

$$\sup_{t\in[0,T]} \left\| P_t - P_{\eta(t)} \right\| \in O(h) \tag{3.9}$$

which by a Gronwall argument yields

$$\sup_{t\in[0,T]} \left\| P_t - P_{\nu(t)}^f \right\| \in O(h). \tag{3.10}$$

Due to

$$P_k^a = (\mathrm{Id} - hK_kG)\, P_k^f, \tag{3.11}$$

this further yields by Lemma 3.1

$$\sup_{t\in[0,T]} \left\| P_t - P_{\nu(t)}^a \right\| \in O(h). \tag{3.12}$$

∎

This convergence result further implies convergence of the ensemble mean: recall that by Assumption 1 it holds $\hat{w}_k^h = 0$ which gives the following recursion:

$$\bar{x}_k^f = (\mathrm{Id} + hA)\, \bar{x}_{k-1}^a, \tag{3.13}$$

$$\bar{x}_k^a = \bar{x}_k^f + K_k \left(\Delta Y_k - hG\bar{x}_k^f \right). \tag{3.14}$$

For the proceeding analysis in this section and throughout the whole paper it is important to stress the following: the Euler-Maruyama time-discretization of the observation process Y

$$\Delta Y_k := Y_{t_k} - Y_{t_{k-1}} \approx hGX_{t_{k-1}} + C^{\frac{1}{2}} \left(V_{t_k} - V_{t_{k-1}} \right) \tag{3.15}$$

yields a discrete-time observation process with observation operator hG. Thus hG will be the modeling assumption on the observations. The actual observations used in the update, however, take the form

$$\Delta Y_k = Y_{t_k} - Y_{t_{k-1}} = \int_{t_{k-1}}^{t_k} GX_s^{\mathrm{ref}} \mathrm{d}s + C^{\frac{1}{2}} \left(V_{t_k} - V_{t_{k-1}} \right) \tag{3.16}$$

where X^{ref} is a reference trajectory of the continuous-time process X. We will assume that

$$\sup_{t\in[0,T]} \mathbb{E}\left[\left\| X_t^{\mathrm{ref}} \right\|^2 \right] < \infty. \tag{3.17}$$

Therefore, we will use the above approximate model of the observations to set up the filter but use (3.16) for the actual observations in the following analysis.

First of all note that it holds:

LEMMA 3.2. *The ensemble mean satisfies*

$$\sup_{t\in[0,T]} \mathbb{E}\left[\left\|\bar{x}^a_{\nu(t)}\right\|^2\right] < \infty. \tag{3.18}$$

For the proof see Appendix A.2.

This enables us to show:

THEOREM 3.2. *Let $(\bar{x}_t)_{t\in[0,T]}$ denote the continuous-time vector-valued process satisfying*

$$\mathrm{d}\bar{x}_t = A\bar{x}_t\mathrm{d}t + P_t G^T C^{-1}\left(\mathrm{d}Y_t - G\bar{x}_t\mathrm{d}t\right). \tag{3.19}$$

Then

$$\mathbb{E}\left[\sup_{t\in[0,T]}\left\|\bar{x}_t - \bar{x}^a_{\nu(t)}\right\|^2\right] \in O(h). \tag{3.20}$$

Proof. First of all, it holds

$$\mathbb{E}\left[\sup_{t\in[0,T]}\left\|\bar{x}_t - \bar{x}^a_{\nu(t)}\right\|^2\right] \lesssim \mathbb{E}\left[\sup_{t\in[0,T]}\left\|\bar{x}_t - \bar{x}_{\eta(t)}\right\|^2\right] + \mathbb{E}\left[\sup_{t\in[0,T]}\left\|\bar{x}_{\eta(t)} - \bar{x}^a_{\nu(t)}\right\|^2\right]. \tag{3.21}$$

The updated ensemble mean satisfies the following recursion

$$\bar{x}^a_k = \bar{x}^a_{k-1} + hA\bar{x}^a_{k-1} + K_k\left(\Delta Y_k - hG\bar{x}^a_{k-1} - h^2GA\bar{x}^a_{k-1}\right). \tag{3.22}$$

Thus we obtain

$$\begin{aligned}
\bar{x}_{\eta(t)} - \bar{x}^a_{\nu(t)} = \bar{x}_0 - \bar{x}^a_0 &\\
&+ \int_0^{\eta(t)} A\left(\bar{x}_s - \bar{x}^a_{\nu(s)}\right) - \left(P_s G^T C^{-1}\bar{x}_s - K_{\nu_+(s)} G\bar{x}^a_{\nu(s)}\right)\mathrm{d}s\\
&+ \int_0^{\eta(t)} \left(P_s G^T C^{-1} - K_{\nu_+(s)}\right)\mathrm{d}Y_s\\
&- h\int_0^{\eta(t)} GA\bar{x}^a_{\nu(s)}\mathrm{d}s.
\end{aligned} \tag{3.23}$$

Using (3.16), we can estimate via the Cauchy-Schwarz inequality

$$\begin{aligned}
&\left\|\bar{x}_{\eta(t)} - \bar{x}^a_{\nu(t)}\right\|^2\\
&\lesssim \|\bar{x}_0 - \bar{x}^a_0\|^2\\
&\quad + \eta(t)\int_0^{\eta(t)} \left(\|A\|^2 + \|P_s\|^2\|G\|^4\left\|C^{-1}\right\|^2\right)\left\|\bar{x}_s - \bar{x}^a_{\nu(s)}\right\|^2\\
&\qquad\qquad + \left\|K_{\nu_+(s)} - P_s G^T C^{-1}\right\|^2\|G\|^2\left(\left\|X^{\mathrm{ref}}_s\right\|^2 + \left\|\bar{x}^a_{\nu(s)}\right\|^2\right)\mathrm{d}s\\
&\quad + \left\|\int_0^{\eta(t)} \left(K_{\nu_+(s)} - P_s G^T C^{-1}\right) C^{\frac{1}{2}}\mathrm{d}V_s\right\|^2
\end{aligned}$$

$$+h^2\eta(t)\int_0^{\eta(t)}\|G\|^2\|A\|^2\left\|\bar{x}^a_{\nu(s)}\right\|^2\mathrm{d}s \tag{3.24}$$

thus it holds

$$\begin{aligned}
&\mathbb{E}\left[\sup_{t\in[0,T]}\left\|\bar{x}_{\eta(t)}-\bar{x}^a_{\nu(t)}\right\|^2\right]\\
\lesssim&\mathbb{E}\left[\left\|\bar{x}_0-\bar{x}^a_0\right\|^2\right]\\
&+T\int_0^T\mathbb{E}\left[\left(\|A\|^2+\|P_s\|^2\|G\|^4\left\|C^{-1}\right\|^2\right)\sup_{r\in[0,s]}\left\|\bar{x}_r-\bar{x}^a_{\nu(r)}\right\|^2\right]\\
&\qquad+\mathbb{E}\left[\left\|K_{\nu_+(s)}-P_sG^TC^{-1}\right\|^2\|G\|^2\left(\left\|X^{\mathrm{ref}}_s\right\|^2+\left\|\bar{x}^a_{\nu(s)}\right\|^2\right)\right]\mathrm{d}s\\
&+\mathbb{E}\left[\sup_{t\in[0,T]}\left\|\int_0^{\eta(t)}\left(K_{\nu_+(s)}-P_sG^TC^{-1}\right)C^{\frac{1}{2}}\mathrm{d}V_s\right\|^2\right]\\
&+h^2T\int_0^{\eta(t)}\|G\|^2\|A\|^2\mathbb{E}\left[\left\|\bar{x}^a_{\nu(s)}\right\|^2\right]\mathrm{d}s.
\end{aligned}\tag{3.25}$$

Observe now that we can estimate

$$\begin{aligned}
&\left\|K_{\nu_+(t)}-P_tG^TC^{-1}\right\|\\
\le&\left\|P^f_{\nu_+(t)}G^T\right\|\left\|\left(C+hGP^f_{\nu_+(t)}G^T\right)^{-1}-C^{-1}\right\|+\left\|P^f_{\nu_+(t)}-P_t\right\|\left\|G^TC^{-1}\right\|\\
\le&h\left\|P^f_{\nu_+(t)}\right\|^2\|G\|^3\|C^{-1}\|^2+\left\|P^f_{\nu_+(t)}-P_t\right\|\|G\|\|C^{-1}\|.
\end{aligned}\tag{3.26}$$

Thus by Lemma 3.1 we obtain

$$\sup_{t\in[0,T]}\left\|K_{\nu_+(t)}-P_tG^TC^{-1}\right\|\in O(h). \tag{3.27}$$

By assumption (3.17), this yields

$$\mathbb{E}\left[\left\|K_{\nu_+(s)}-P_sG^TC^{-1}\right\|^2\|G\|^2\left\|X^{\mathrm{ref}}_s\right\|^2\right]\in O(h^2) \tag{3.28}$$

and we further deduce by the L^p-maximal-inequality

$$\begin{aligned}
&\mathbb{E}\left[\sup_{t\in[0,T]}\left\|\int_0^{\eta(t)}\left(K_{\nu_+(s)}-P_sG^TC^{-1}\right)C^{\frac{1}{2}}\mathrm{d}V_s\right\|^2\right]\\
=&\int_0^T\mathbb{E}\left[\left\|\left(K_{\nu_+(s)}-P_sG^TC^{-1}\right)C^{\frac{1}{2}}\right\|^2\right]\mathrm{d}s\in O(h^2).
\end{aligned}\tag{3.29}$$

Finally using (3.17) and boundedness of $\|P_t\|$ on $[0,T]$, one can similarly show that

$$\mathbb{E}\left[\sup_{t\in[0,T]}\left\|\bar{x}_t-\bar{x}_{\eta(t)}\right\|^2\right]\in O(h). \tag{3.30}$$

Thus Lemma 3.2 and a Gronwall argument conclude the proof. ∎

4. Algorithms

In this section, we introduce the three ESRF algorithms EAKF, ETKF, and the unperturbed filter from [19] which we will focus on in this paper.

4.1. Ensemble Adjustment/Transform Kalman Filter. The transformations of E_k^f in case of EAKF and ETKF are given by the following:

- EAKF: $E_k^a = A_k E_k^f$
- ETKF: $E_k^a = E_k^f T_k$

for matrices A_k and T_k specified below. Then using the update step (1.6) of the ensemble mean, one computes the updated ensemble members via

$$X_{t_k}^{(i),a} = X_{t_k}^{(i),a} - \bar{x}_k^a + \bar{x}_k^a = E_k^a e_i + \bar{x}_k^a \tag{4.1}$$

where $e_i \in \mathbb{R}^M$ with $(e_i)_j = \delta_{ij}, j = 1,...,M$. The structure of the transformation matrices and equivalence of both EAKF and ETKF has been summarized in [17] in terms of the singular value decomposition factors of the forecast covariance matrix, as well as in the appendix of [13] using basic linear algebra.

In our recent paper [10] using the following integral representation

$$\sqrt{P^{-1}} = \frac{1}{\sqrt{\pi}} \int_0^\infty \frac{1}{\sqrt{t}} e^{-tP} \mathrm{d}t \tag{4.2}$$

for any symmetric positive semi-definite matrix P, we were able to identify an equivalent analytic representation of these transformations of the following form: as derived in [10], A_k and T_k are given by

$$A_k = \sqrt{P_k^f} \left(\mathrm{Id} + h \sqrt{P_k^f} G^T C^{-1} G \sqrt{P_k^f} \right)^{-\frac{1}{2}} \sqrt{P_k^f}^{-1}, \tag{4.3}$$

$$T_k = \left(\mathrm{Id} + \frac{h}{M-1} \left(E_k^f \right)^T G^T C^{-1} G E_k^f \right)^{-\frac{1}{2}} \tag{4.4}$$

where $\sqrt{P_k^f}$ denotes the symmetric positive semidefinite square root of P_k^f and $\sqrt{P_k^f}^{-1}$ its pseudo inverse. These transformations are adjoint in the sense that it holds

$$A_k E_k^f = E_k^f T_k \tag{4.5}$$

which is the consequence of the following important integral representation

$$A_k E_k^f = \frac{1}{\sqrt{\pi}} \int_0^\infty \frac{e^{-t}}{\sqrt{t}} e^{-thP_k^f G^T C^{-1} G} \mathrm{d}t E_k^f = E_k^f T_k \tag{4.6}$$

(see [10] for an in-depth discussion). We introduce the following first-order expansion of the integral term

$$\frac{1}{\sqrt{\pi}} \int_0^\infty \frac{e^{-t}}{\sqrt{t}} e^{-thP_k^f G^T C^{-1} G} \mathrm{d}t = \mathrm{Id} - \frac{h}{2} P_k^f G^T C^{-1} G + R_k^h. \tag{4.7}$$

Then using (4.1), both EAKF and ETKF can be summarized as the following

Algorithm 1. (EAKF/ETKF)

$$X_{t_k}^{(i),f} = X_{t_{k-1}}^{(i),a} + hAX_{t_{k-1}}^{(i),a} + hQ^{\frac{1}{2}} \hat{W}_k^{(i),h}, \tag{4.8}$$

$$X_{t_k}^{(i),a} = X_{t_k}^{(i),f} - \frac{h}{2} P_k^f G^T C^{-1} G X_{t_k}^{(i),f} - h\left(K_k - \frac{1}{2} P_k^f G^T C^{-1}\right) G\bar{x}_k^f + K_k \Delta Y_k + R_k^h E_k^f e_i. \tag{4.9}$$

4.2. Whitaker, Hamill (2002). In [19], the authors approach the problem of omitting stochastic perturbations in a different way: similar as in the Kalman filter, the update step of the ensemble mean and the ensemble deviations should be of the form

$$\bar{x}_k^a = \bar{x}_k^f + K_k\left(\Delta Y_k - hG\bar{x}_k^f\right), \tag{4.10}$$

$$E_k^a = E_k^f + \tilde{K}_k\left((\Delta Y_k)' - hGE_k^f\right) \tag{4.11}$$

with K_k as in (1.8), where $\tilde{K}_k$ denotes the gain for the update of the ensemble deviations and in case of the stochastic EnKF, $(\Delta Y_k)'$ denotes the perturbations added to the actual observation ΔY_k. In the aim of avoiding such perturbations, setting $(\Delta Y_k)' = 0$ yields

$$E_k^a = \left(\mathrm{Id} - h\tilde{K}_k G\right) E_k^f. \tag{4.12}$$

This gives the correct covariance matrix only if $\tilde{K}_k$ solves

$$\left(\mathrm{Id} - h\tilde{K}_k G\right) P_k^f \left(\mathrm{Id} - h\tilde{K}_k G\right)^T \overset{!}{=} P_k^a = (\mathrm{Id} - hK_k G) P_k^f \tag{4.13}$$

which has the solution

$$\tilde{K}_k = P_k^f G^T \left(C + hGP_k^f G^T\right)^{-\frac{1}{2}} \left(\left(C + hGP_k^f G^T\right)^{\frac{1}{2}} + C^{\frac{1}{2}}\right)^{-1}. \tag{4.14}$$

For ease of notation denote in the following

$$\tilde{C}_k := \left(C + hGP_k^f G^T\right)^{-\frac{1}{2}} \left(\left(C + hGP_k^f G^T\right)^{\frac{1}{2}} + C^{\frac{1}{2}}\right)^{-1}. \tag{4.15}$$

Thus using (4.1), we can deduce the following algorithm:

Algorithm 2.(Whitaker, Hamill (2002))

$$X_{t_k}^{(i),f} = X_{t_{k-1}}^{(i),a} + hAX_{t_{k-1}}^{(i),a} + hQ^{\frac{1}{2}} \hat{W}_k^{(i),h}, \tag{4.16}$$

$$X_{t_k}^{(i),a} = X_{t_k}^{(i),f} + K_k\left(\Delta Y_k - hG\bar{x}_k^f\right) - h\tilde{K}_k G\left(X_{t_k}^{(i),f} - \bar{x}_k^f\right). \tag{4.17}$$

4.3. Summary. Comparing Algorithm 1 and 2, we obtain a unified structure of the update step of the following form

$$X_{t_k}^{(i),a} = X_{t_k}^{(i),f} - h\hat{K}_k G X_{t_k}^{(i),f} - h\left(K_k - \hat{K}_k\right) G\bar{x}_k^f + K_k \Delta Y_k + \mathcal{R}_k^h e_i \tag{4.18}$$

where for EAKF/ETKF we denote

$$\hat{K}_k := \frac{1}{2} P_k^f G^T C^{-1}, \quad \mathcal{R}_k^h := R_k^h E_k^f \tag{4.19}$$

and for the unperturbed filter

$$\hat{K}_k := \tilde{K}_k, \quad \mathcal{R}_k^h = 0. \tag{4.20}$$

LEMMA 4.1. *The following estimates hold:*

$$\left\|\hat{K}_k\right\| \leq \frac{1}{2}\left\|P_k^f\right\| \|G\| \left\|C^{-1}\right\| \tag{4.21}$$

and

$$\left\|\mathcal{R}_k^h\right\| \leq \frac{3h^2}{8}\left\|P_k^f\right\|^2 \left\|G^T C^{-1} G\right\|^2. \tag{4.22}$$

Proof. On (4.21): for EAKF/ETKF the estimate is straightforward. For the unperturbed filter use

$$C^{\frac{1}{2}} \leq \left(C + hGP_k^f G^T\right)^{\frac{1}{2}} \tag{4.23}$$

to deduce

$$\left\|\tilde{C}_k\right\| \leq \left\|\left(C + hGP_k^f G^T\right)^{-\frac{1}{2}}\right\| \left\|\left(\left(C + hGP_k^f G^T\right)^{\frac{1}{2}} + C^{\frac{1}{2}}\right)^{-1}\right\| \leq \frac{1}{2}\left\|C^{-\frac{1}{2}}\right\|^2 \tag{4.24}$$

which gives the claim.

On (4.22): the claim trivially holds true in the case of the unperturbed filter. For EAKF/ETKF observe that with $\Theta := G^T C^{-1} G$

$$\begin{aligned}
R_k^h &= \frac{1}{\sqrt{\pi}}\int_0^\infty \frac{e^{-t}}{\sqrt{t}}\left(e^{-thP_k^f\Theta} - \mathrm{Id} + thP_k^f\Theta\right)\mathrm{d}t \\
&= \frac{1}{\sqrt{\pi}}\int_0^\infty \frac{e^{-t}}{\sqrt{t}}\left(-\int_0^t e^{-shP_k^f\Theta} hP_k^f\Theta \mathrm{d}s + thP_k^f\Theta\right)\mathrm{d}t \\
&= \frac{1}{\sqrt{\pi}}\int_0^\infty \frac{e^{-1}}{\sqrt{t}}\left(\int_0^t\int_0^s e^{-rhP_k^f\Theta}\mathrm{d}r\mathrm{d}s\right)\left(hP_k^f\Theta\right)^2\mathrm{d}t \\
&= \frac{1}{\sqrt{\pi}}\int_0^\infty \frac{e^{-t}}{\sqrt{t}}\left(\int_0^t\int_0^s \sqrt{P_k^f}e^{-rh\sqrt{P_k^f}\Theta\sqrt{P_k^f}}\sqrt{P_k^f}\mathrm{d}r\mathrm{d}s\right)h^2\Theta P_k^f\Theta\mathrm{d}t.
\end{aligned} \tag{4.25}$$

The claim then follows by using $\left\|e^{-th\sqrt{P_k^f}\Theta\sqrt{P_k^f}}\right\| \leq 1$. ∎

4.4. Non-uniqueness of the transformations. As has been pointed out in [17], the above transformations are not unique. Indeed, for an orthogonal matrix $\mathcal{U}_k$ note that, for instance, in case of the ETKF the modified transformation $\hat{E}_k := E_k^f T_k \mathcal{U}_k$ also yields the correct covariance matrix since

$$\hat{P}_k := \frac{1}{M-1}\hat{E}_k\hat{E}_k^T = \frac{1}{M-1}E_k^f T_k \mathcal{U}_k\mathcal{U}_k^T T_k^T\left(E_k^f\right)^T = \frac{1}{M-1}E_k^f T_k T_k^T\left(E_k^f\right)^T = P_k^a. \tag{4.26}$$

Thus post-multiplying the transformed ensemble with an orthogonal matrix does not change the resulting covariance matrix. This issue was further exploited in [18] and [11]. In the latter, the authors elaborate more conditions on the matrix $\mathcal{U}_k$: for $\mathbf{1} = (1,\ldots,1)^T$ note that it holds $E_k^f\mathbf{1} = 0$. A transformation τ is called mean-preserving if after applying the transformation it still holds true $\tau\left(E_k^f\right)\mathbf{1} = 0$. This is a desirable property since

it is needed in the update step (4.1). The EAKF, the filter by Whitaker and Hamill and, due to Lemma 4.5, also the ETKF are mean-preserving. Thus an orthogonal post-multiplier $\mathcal{U}_k$ is appropriate in this sense, if it does not violate the mean-preserving property, i.e. satisfies $\tau\left(E_k^f\right)\mathcal{U}_k\mathbf{1}=0$. This is clearly the case if $\mathbf{1}$ is an eigenvector of $\mathcal{U}_k$. In the following section, after conducting the continuous time limit analysis for the unmodified algorithms, we further elaborate on a possible extension to orthogonal transformations in Section 5.4.

5. Continuous time limit II: ensemble members

Throughout this section, we assume that the deterministic model perturbations $\hat{W}^{(i),h}$ satisfy Assumption 1. First observe that formally taking the continuous time limit in Algorithm 1 then yields the following coupled system of differential equations with suitably defined model perturbations $\hat{W}_t^{(i)}$:

$$\mathrm{d}X_t^{(i)} = AX_t^{(i)}\mathrm{d}t + Q^{\frac{1}{2}}\hat{W}_t^{(i)}\mathrm{d}t + P_tG^TC^{-1}\left(\mathrm{d}Y_t - \frac{1}{2}G\left(X_t^{(i)}+\bar{x}_t\right)\mathrm{d}t\right). \tag{5.1}$$

Example 5.1. *In [9], we were able to show a continuous time limit result for the case of perturbations of the form (2.4) yielding (5.1) with model perturbations*

$$\hat{W}_t^{(i)} := \frac{1}{2}Q^{\frac{1}{2}}P_t^{-1}\left(X_t^{(i)}-\bar{x}_t\right). \tag{5.2}$$

Assuming that the processes $\left(\hat{W}_t^{(i)}\right)_{t\geq 0}$ are continuous and fulfill

Assumption 2. *It holds:*

- $\frac{1}{M-1}\sum_{i=1}^M\left(X_t^{(i)}-\bar{x}_t\right)\left(\hat{W}_t^{(i)}-\hat{w}_t\right)^TQ^{\frac{1}{2}}=\frac{1}{2}Q$
- *w.l.o.g.* $\hat{W}_t^{(i)}$ *are centered.*

One can easily check that this yields the correct structure of the first and second moments, i.e. corresponding to the processes (5.1), the ensemble mean $\bar{x}$ and the covariance matrix P satisfy the Kalman-Bucy filter Equations (1.9) and (1.10), respectively. Further we obtain that (5.2) is an exemplary choice of such perturbations.

However, it is not clear whether Assumption 1 and Assumption 2 already yield convergence of the model perturbations. We therefore further impose

Assumption 3. *There exists a constant* $R_T>0$ *such that*

$$\sum_{i=1}^M\left\|\hat{W}_{\nu_+(t)}^{(i),h}-\hat{W}_t^{(i)}\right\|^2 \leq R_T\left(h^2+\sum_{i=1}^M\left\|X_{\eta(t)}^{(i),a}-X_t^{(i)}\right\|^2\right). \tag{5.3}$$

Example 5.2. *Recall Examples 2.1 and 5.1. For these choices of model perturbation it holds*

$$\begin{aligned}&\sum_{i=1}^M\left\|\hat{W}_{\nu_+(t)}^{(i),h}-\hat{W}_t^{(i)}\right\|^2\\ =&\sum_{i=1}^M\left\|\frac{1}{2}Q\left(\left(P_{\nu(t)}^a\right)^{-1}\left(X_{\eta(t)}^{(i),a}-\bar{x}_{\nu(t)}^a\right)-P_t^{-1}\left(X_t^{(i)}-\bar{x}_t\right)\right)\right\|^2\end{aligned}$$

$$\lesssim \|Q\|^2 \left((M-1) \left\| \left(P^a_{\nu(t)}\right)^{-1} - P_t^{-1} \right\|^2 \mathcal{V}_t + \left\| \left(P^a_{\nu(t)}\right)^{-1} \right\|^2 \sum_{i=1}^M \left\| X^{(i),a}_{\eta(t)} - X^{(i)}_t \right\|^2 \right)$$
$$\leq R_T \left(h^2 + \sum_{i=1}^M \left\| X^{(i),a}_{\eta(t)} - X^{(i)}_t \right\|^2 \right) \tag{5.4}$$

by Theorem 3.1, Appendix B and boundedness of $\left\|P_t^{-1}\right\|$ *and* $\mathcal{V}_t$ *on* $[0,T]$ *(as shown in [9]).*

In the following, we assume that there exists a pathwise unique strong solution $X^{(i)}$ to (5.1) which is almost surely continuous and satisfies

$$\sup_{t\in[0,T]} \mathbb{E}\left[\left\| X^{(i)}_t \right\|^2 \right] < \infty. \tag{5.5}$$

In case of Example 5.1, see the argumentation in [5] on existence of such solutions. By using (3.16) and assumption (3.17), we obtain that

$$\mathbb{E}\left[\sup_{t\in[0,T]} \sum_{i=1}^M \left\| X^{(i)}_t - X^{(i)}_{\eta(t)} \right\|^2 \right] \in O(h). \tag{5.6}$$

Then under the above assumptions, the main result of this paper now reads as follows:

THEOREM 5.1. *Consider Algorithm 1 or Algorithm 2, let* $\left(X^{(i)}_t\right)_{t\in[0,T]}$ *be the unique strong solution to (5.1) and let* $\|P^a_0\|$ *be bounded uniformly in* ω*. If*

$$\mathbb{E}\left[\sum_{i=1}^M \left\| X^{(i),a}_0 - X^{(i)}_0 \right\|^2 \right] \in O(h), \tag{5.7}$$

then it holds

$$\mathbb{E}\left[\sup_{t\in[0,T]} \sum_{i=1}^M \left\| X^{(i),a}_{\eta(t)} - X^{(i)}_t \right\|^2 \right] \in O(h). \tag{5.8}$$

The proof of Theorem 5.1 will now be given in the following sections.

5.1. Preliminaries. By Assumption 1 and using the same analysis as in [5], one can show that it holds (recall that $\hat{W}^{(i)}_k$ are centred)

$$\sum_{i=1}^M \left\| Q^{\frac{1}{2}} \hat{W}^{(i),h}_k \right\|^2 \leq \sqrt{M}(M-1) \left\| \frac{1}{M-1} \sum_{i=1}^M Q^{\frac{1}{2}} \left(\hat{W}^{(i),h}_k - \hat{w}^h_k \right) \left(\hat{W}^{(i),h}_k - \hat{w}^h_k \right)^T Q^{\frac{1}{2}} \right\|_F$$
$$\leq \sqrt{M}(M-1)\kappa. \tag{5.9}$$

With the above, we obtain the following result:

LEMMA 5.1. *For both Algorithm 1 and Algorithm 2 it holds*

$$\sup_{t\in[0,T]} \mathbb{E}\left[\sum_{i=1}^M \left\| X^{(i),a}_{\eta(t)} \right\|^2 \right] < \infty. \tag{5.10}$$

For the proof see Appendix A.3.

5.2. The continuous time limit. Using (4.18), observe that the difference process satisfies

$$\begin{aligned}
&X^{(i),a}_{\eta(t)} - X^{(i)}_{\eta(t)} \\
&= X^{(i),a}_0 - X^{(i)}_0 + \int_0^{\eta(t)} A\left(X^{(i),a}_{\eta(s)} - X^{(i)}_s\right) + Q^{\frac{1}{2}}\left(\hat{W}^{(i),h}_{\nu_+(s)} - \hat{W}^{(i)}_s\right) \\
&\qquad + \left(K_{\nu_+(s)} - P_s G^T C^{-1}\right) G X^{\mathrm{ref}}_s \\
&\qquad - \left(\hat{K}_{\nu_+(s)} G X^{(i),f}_{\eta_+(s)} + \left(K_{\nu_+(s)} - \hat{K}_{\nu_+(s)}\right) G \bar{x}_{\nu_+(s)} - \frac{1}{2} P_s G^T C^{-1} G\left(X^{(i)}_s + \bar{x}_s\right)\right) \mathrm{d}s \\
&\quad + \int_0^{\eta(t)} \left(K_{\nu_+(s)} - P_s G^T C^{-1}\right) C^{\frac{1}{2}} \mathrm{d}V_s + \mathcal{R}^h_{\nu(t)}.
\end{aligned} \tag{5.11}$$

With

$$\begin{aligned}
&\hat{K}_{\nu_+(s)} G X^{(i),f}_{\eta_+(s)} + \left(K_{\nu_+(s)} - \hat{K}_{\nu_+(s)}\right) G \bar{x}_{\nu_+(s)} - \frac{1}{2} P_s G^T C^{-1} G\left(X^{(i)}_s + \bar{x}_s\right) \\
&= \hat{K}_{\nu_+(s)} G\left(X^{(i),f}_{\eta_+(s)} - X^{(i)}_s\right) + \left(\hat{K}_{\nu_+(s)} - \frac{1}{2} P_s G^T C^{-1}\right) G X^{(i)}_s \\
&\quad + \left(K_{\nu_+(s)} - \hat{K}_{\nu_+(s)}\right) G\left(\bar{x}^f_{\nu_+(s)} - \bar{x}_s\right) + \left(K_{\nu_+(s)} - \hat{K}_{\nu_+(s)} - \frac{1}{2} P_s G^T C^{-1}\right) G \bar{x}_s
\end{aligned}$$

and the Cauchy-Schwarz-inequality we estimate

$$\begin{aligned}
&\sum_{i=1}^M \left\| X^{(i),a}_{\eta(t)} - X^{(i)}_{\eta(t)} \right\|^2 \\
&\lesssim \sum_{i=1}^M \left\| X^{(i),a}_0 - X^{(i)}_0 \right\|^2 + \eta(t) \int_0^{\eta(t)} \|A\|^2 \sum_{i=1}^M \left\| X^{(i),a}_{\eta(s)} - X^{(i)}_s \right\|^2 + \|Q\| \sum_{i=1}^M \left\| \hat{W}^{(i)}_{\nu_+(s)} - \hat{W}^{(i)}_s \right\|^2 \\
&\quad + M \left\| K_{\nu_+(s)} - P_s G^T C^{-1} \right\|^2 \|G\|^2 \left\| X^{\mathrm{ref}}_s \right\|^2 \\
&\quad + \left(\left\| \hat{K}_{\nu_+(s)} \right\|^2 + \left\| K_{\nu_+(s)} - \hat{K}_{\nu_+(s)} \right\|^2 \right) \|G\|^2 \sum_{i=1}^M \left\| X^{(i),f}_{\eta_+(s)} - X^{(i),s}_s \right\|^2 \\
&\quad + \left(\left\| \hat{K}_{\nu_+(s)} - \frac{1}{2} P_s G^T C^{-1} \right\|^2 \right. \\
&\quad \left. + \left\| K_{\nu_+(s)} - \hat{K}_{\nu_+(s)} - \frac{1}{2} P_s G^T C^{-1} \right\|^2 \right) \|G\|^2 \sum_{i=1}^M \left\| X^{(i)}_s \right\|^2 \mathrm{d}s \\
&\quad + M \left\| \int_0^{\eta(t)} \left(K_{\nu_+(s)} - P_s G^T C^{-1}\right) C^{\frac{1}{2}} \mathrm{d}V_s \right\|^2 + M \left\| \sum_{k=1}^{\nu(t)} \mathcal{R}^h_k \right\|^2.
\end{aligned} \tag{5.12}$$

Recall from the proof of Theorem 3.2 that it holds

$$\mathbb{E}\left[\left\| K_{\nu_+(s)} - P_s G^T C^{-1} \right\|^2 \|G\|^2 \left\| X^{\mathrm{ref}}_s \right\|^2 \right] \in O(h^2) \tag{5.13}$$

and

$$\mathbb{E}\left[\sup_{t\in[0,T]} \left\| \int_0^{\eta(t)} \left(K_{\nu_+(s)} - P_s G^T C^{-1}\right) C^{\frac{1}{2}} \mathrm{d}V_s \right\|^2 \right] \in O(h^2). \tag{5.14}$$

Furthermore by Lemma 3.1 and Lemma 4.1, $\|K_k\|$ and $\left\|\hat{K}_k\right\|$ are uniformly bounded in k, and again from Lemma 4.1 we obtain

$$\left\|\mathcal{R}_k^h\right\| \in O(h^2) \tag{5.15}$$

uniformly in k. It remains to investigate the differences of the Kalman gains. We claim that it holds

$$\sup_{t\in[0,T]} \left\|\hat{K}_{\nu_+(t)} - \frac{1}{2}P_t G^T C^{-1}\right\| \in O(h) \tag{5.16}$$

and

$$\sup_{t\in[0,T]} \left\|K_{\nu_+(t)} - \hat{K}_{\nu_+(t)} - \frac{1}{2}P_t G^T C^{-1}\right\| \in O(h). \tag{5.17}$$

Indeed: in the case of EAKF/ETKF, the claim follows from the decompositions

$$\hat{K}_{\nu_+(t)} - \frac{1}{2}P_t G^T C^{-1} = \frac{1}{2}\left(P_{\nu_+(t)} - P_t\right) G^T C^{-1} \tag{5.18}$$

and

$$\begin{aligned} &K_{\nu_+(t)} - \hat{K}_{\nu_+(t)} - \frac{1}{2}P_t G^T C^{-1} \\ &= P^f_{\nu_+(t)} G^T \left(\left(C + hGP^f_{\nu_+(t)}G^T\right)^{-1} - C^{-1}\right) + \frac{1}{2}\left(P^f_{\nu_+(t)} - P_t\right) G^T C^{-1} \end{aligned} \tag{5.19}$$

together with Lemma 3.1 and Theorem 3.1. For the unperturbed filter we decompose the differences in such a way that

$$\begin{aligned} &\hat{K}_{\nu_+(t)} - \frac{1}{2}P_t G^T C^{-1} \\ &= \frac{1}{2}\left(P^f_{\nu_+(t)} - P_t\right) G^T C^{-1} + P^f_{\nu_+(t)} G^T \left(\tilde{C}_{\nu_+(t)} - \frac{1}{2}C^{-1}\right) \\ &= \frac{1}{2}\left(P^f_{\nu_+(t)} - P_t\right) G^T C^{-1} \\ &\quad + P^f_{\nu_+(t)} G^T \tilde{C}_{\nu_+(t)} \left(C^{\frac{1}{2}}\left(C^{\frac{1}{2}} - \left(C + hGP^f_{\nu_+(t)}G^T\right)^{\frac{1}{2}}\right) - hGP^f_{\nu_+(t)}G^T\right)\frac{1}{2}C^{-1} \end{aligned} \tag{5.20}$$

and

$$\begin{aligned} K_{\nu_+(t)} - \hat{K}_{\nu_+(t)} &= P^f_{\nu_+(t)} G^T \left(C + hGP^f_{\nu_+(t)}G^T\right)^{-\frac{1}{2}} \\ &\quad \times \left(\left(C + hGP^f_{\nu_+(t)}G^T\right)^{-\frac{1}{2}} - \left(\left(C + hGP^f_{\nu_+(t)}G^T\right)^{\frac{1}{2}} + C^{\frac{1}{2}}\right)^{-1}\right) \\ &= K_{\nu_+(t)} C^{\frac{1}{2}} \left(\left(C + hGP^f_{\nu_+(t)}G^T\right)^{\frac{1}{2}} + C^{\frac{1}{2}}\right)^{-1} \end{aligned} \tag{5.21}$$

which gives

$$K_{\nu_+(t)} - \hat{K}_{\nu_+(t)} - \frac{1}{2}P_t G^T C^{-1}$$

$$
\begin{aligned}
&= \left(K_{\nu_+(t)} - P_t G^T C^{-1}\right) C^{\frac{1}{2}} \left(\left(C + hGP^f_{\nu_+(t)} G^T\right)^{\frac{1}{2}} + C^{\frac{1}{2}}\right)^{-1} \\
&\quad + P_t G^T C^{-1} \left(C^{\frac{1}{2}} \left(\left(C + hGP^f_{\nu_+(t)} G^T\right)^{\frac{1}{2}} + C^{\frac{1}{2}}\right)^{-1} - \frac{1}{2}\mathrm{Id}\right) \\
&= \left(K_{\nu_+(t)} - P_t G^T C^{-1}\right) C^{\frac{1}{2}} \left(\left(C + hGP^f_{\nu_+(t)} G^T\right)^{\frac{1}{2}} + C^{\frac{1}{2}}\right)^{-1} \\
&\quad + \frac{1}{2} P_t G^T C^{-1} \left(C^{\frac{1}{2}} - \left(C + hGP^f_{\nu_+(t)} G^T\right)^{\frac{1}{2}}\right) \left(\left(C + hGP^f_{\nu_+(t)} G^T\right)^{\frac{1}{2}} + C^{\frac{1}{2}}\right)^{-1}.
\end{aligned} \tag{5.22}
$$

Using

$$
\left(C + hGP^f_k G^T\right)^{\frac{1}{2}} + C^{\frac{1}{2}} \geq 2C^{\frac{1}{2}} \geq 2\sqrt{\lambda_{\min}(C)}\mathrm{Id} \tag{5.23}
$$

with $\lambda_{\min}(C)$ the smallest eigenvalue of C, which yields

$$
\left\| \left(C + hGP^f_k G^T\right)^{\frac{1}{2}} - C^{\frac{1}{2}} \right\| \leq \frac{1}{2\sqrt{\lambda_{\min}(C)}} \left\| C + hGP^f_k G^T - C \right\| = \frac{h}{2\sqrt{\lambda_{\min}(C)}} \left\| GP^f_k G^T \right\|, \tag{5.24}
$$

together with Lemma 3.1 and Theorem 3.1 then yields the claim.
Note that by (5.9) it holds

$$
\begin{aligned}
&\sum_{i=1}^{M} \left\| X^{(i),f}_{\eta_+(t)} - X^{(i)}_t \right\|^2 \\
&\lesssim \sum_{i=1}^{M} \left\| X^{(i),a}_{\eta(t)} - X^{(i)}_t \right\|^2 + h^2 \|A\|^2 \left(\sum_{i=1}^{M} \left\| X^{(i),a}_{\eta(t)} \right\|^2 \right) + h^2 \sqrt{M}(M-1)\kappa
\end{aligned} \tag{5.25}
$$

where in expectation, the first h^2-term is bounded uniformly in time by Lemma 5.1. Thus

$$
\begin{aligned}
&\mathbb{E}\left[\sup_{t\in[0,T]} \sum_{i=1}^{M} \left\| X^{(i),a}_{\eta(t)} - X^{(i)}_{\eta(t)} \right\|^2 \right] \\
&\lesssim \mathbb{E}\left[\sum_{i=1}^{M} \left\| X^{(i),a}_0 - X^{(i)}_0 \right\|^2 \right] + T \int_0^T \mathbb{E}\left[\sup_{r\in[0,s]} \sum_{i=1}^{M} \left\| X^{(i),a}_{\eta(r)} - X^{(i)}_r \right\|^2 \right] \mathrm{d}s + O(h^2).
\end{aligned} \tag{5.26}
$$

Together with (5.6) this then proves the claim of Theorem 5.1 by a Gronwall argument.

5.3. Extension to general Lipschitz-continuous model operators. In the nonlinear case where

$$
\mathrm{d}X_t = f(X_t)\mathrm{d}t + Q^{\frac{1}{2}} \mathrm{d}W_t
$$

yielding a forecast ensemble of the form

$$
X^{(i),f}_{t_k} = X^{(i),a}_{t_{k-1}} + hf\left(X^{(i),a}_{t_{k-1}}\right) + hQ^{\frac{1}{2}} \hat{W}^{(i),h}_k, \tag{5.27}
$$

we do not obtain a recursion of the covariance matrices in closed form and consequently no a priori continuous time limit result as in Theorem 3.1. Nevertheless, for Lipschitz-continuous f we can show a continuous time limit of the ensemble members which then in turn yields the convergence of the covariance matrices. Observe the following:

$$\begin{aligned}
&\left\|P^f_{\nu(t)} - P_t\right\| \\
&= \left\| \frac{1}{M-1}\sum_{i=1}^{M}\left(X^{(i),f}_{\eta(t)} - \bar{x}^f_{\nu(t)}\right)\left(X^{(i),f}_{\eta(t)} - \bar{x}^f_{\nu(t)}\right)^T - \left(X^{(i)}_t - \bar{x}_t\right)\left(X^{(i)}_t - \bar{x}_t\right)^T \right\| \\
&\leq 2\left(\left(\mathcal{V}^f_{\nu(t)}\right)^{\frac{1}{2}} + (\mathcal{V}_t)^{\frac{1}{2}}\right)\left(\frac{1}{M-1}\sum_{i=1}^{M}\left\|X^{(i),f}_{\eta(t)} - X^{(i)}_t\right\|^2\right)^{\frac{1}{2}} \qquad (5.28)
\end{aligned}$$

which yields

$$\left\|P^f_{\nu(t)} - P_t\right\|^2 \leq 8\left(\mathcal{V}^f_{\nu(t)} + \mathcal{V}_t\right)\left(\frac{1}{M-1}\sum_{i=1}^{M}\left\|X^{(i),f}_{\eta(t)} - X^{(i)}_t\right\|^2\right) \qquad (5.29)$$

with $\mathcal{V}^f_k$ and $\mathcal{V}_t$ as defined in (1.17). Similar to [5], one can show that $\mathcal{V}_t$ is bounded uniformly on $[0,T]$. Since it holds by (5.9) that

$$\begin{aligned}
\mathcal{V}^f_k &= \frac{1}{M-1}\sum_{i=1}^{M}\left\|X^{(i),a}_{t_{k-1}} - \bar{x}^a_{k-1} + h\left(f\left(X^{(i),a}_{t_{k-1}}\right) - \bar{f}^a_{k-1}\right) + hQ^{\frac{1}{2}}\hat{W}^{(i)}_k\right\|^2 \\
&\leq \left(1 + h + 2h(Lf)_+ + 8h^2\|f\|^2_{\mathrm{Lip}}\right)\mathcal{V}^a_{k-1} + h(1+2h)\sqrt{M}(M-1)\kappa \qquad (5.30)
\end{aligned}$$

and

$$\mathcal{V}^a_k = \mathrm{tr}(P^a_k) = \mathrm{tr}\left(P^f_k - hK_kGP^f_k\right) = \mathrm{tr}(P^f_k) - \left\|P^f_k G^T C^{-\frac{1}{2}}\right\|^2_F \leq \mathrm{tr}(P^f_k) = \mathcal{V}^f_k, \qquad (5.31)$$

we thus deduce that both $\mathcal{V}^a_k$ and $\mathcal{V}^f_k$ are bounded uniformly in k by a Gronwall argument. This further yields that $\left\|P^f_k\right\|$ and $\|P^a_k\|$ are bounded uniformly in k.

Therefore assume existence of a strong solution to

$$\mathrm{d}X^{(i)}_t = f\left(X^{(i)}_t\right)\mathrm{d}t + Q^{\frac{1}{2}}\hat{W}^{(i)}_t\mathrm{d}t + P_tG^TC^{-1}\left(\mathrm{d}Y_t - \frac{1}{2}G\left(X^{(i)}_t + \bar{x}_t\right)\mathrm{d}t\right) \qquad (5.32)$$

with

$$\sup_{t\in[0,T]}\mathbb{E}\left[\left\|X^{(i)}_t\right\|^2\right] < \infty, \qquad (5.33)$$

then following similar steps as seen in the proof of Theorem 5.1 we easily deduce:

THEOREM 5.2. *Consider Algorithm 1 or Algorithm 2, let* $\left(X^{(i)}_t\right)_{t\in[0,T]}$ *be the unique strong solution to (5.32) and let* $\|P^a_0\|$ *be bounded uniformly in* ω. *If*

$$\mathbb{E}\left[\sum_{i=1}^{M}\left\|X^{(i),a}_0 - X^{(i)}_0\right\|^2\right] \in O(h), \qquad (5.34)$$

then it holds

$$\mathbb{E}\left[\sup_{t\in[0,T]}\sum_{i=1}^{M}\left\|X^{(i),a}_{\eta(t)} - X^{(i)}_t\right\|^2\right] \in O(h). \qquad (5.35)$$

5.4. Extension to orthogonal transformations. As pointed out in Section 4.4, also post-multiplying with an orthogonal matrix $\mathcal{U}_k$ is a valid transformation where, additionally, $\mathbf{1}=(1,...,1)^T$ is an eigenvector of $\mathcal{U}_k$ such that the transformation remains mean-preserving. An example is to be found in [18] where the authors propose a revised version of the ETKF in terms of singular value decompositions: originally in [3], the transformation reads

$$T_k = U(I+h\Sigma)^{-\frac{1}{2}} \tag{5.36}$$

as a result of the singular value decomposition

$$\left(E_k^f\right)^T G^T C^{-1} G E_k^f = U\Sigma V^T. \tag{5.37}$$

This formulation was then modified in [18] via

$$T_k = U(I+h\Sigma)^{-\frac{1}{2}} U^T \tag{5.38}$$

which, apart from other properties as discussed in [18], gives again a mean-preserving ETKF.

A natural generalization of the above is to use the transformation

$$\tilde{T}_k := T_k \mathcal{U}_k, \qquad \mathcal{U}_k = \mathcal{U}\left(\mathbb{X}_k^f\right), \tag{5.39}$$

where $\mathcal{U}$ is a function of the underlying ensemble $\mathbb{X}_k^f = \left(X_{t_k}^{(i),f}\right)_{i=1,...,M}$ taking values in the set of orthogonal matrices that are mean-preserving. As we have already argued in Section 4.4, (5.39) yields the same covariance matrix as the original transformed ensemble thus Lemma 3.1 carries over immediately.

Let E^a now denote the ensemble resulting from the algorithm using (5.39). Writing out the previously analyzed algorithms in this setting gives the following evolution equation

$$\begin{aligned} E_k^a &= E_{k-1}^a \mathcal{U}_k + h\left(\left(A - \frac{1}{2}P_k^f G^T C^{-1} G\right) E_{k-1}^a + \mathcal{W}_k\right)\mathcal{U}_k + O\left(h^2\right) \\ &= E_{k-1}^a + E_{k-1}^a\left(\mathcal{U}_k - \mathrm{Id}\right) + h\left(\left(A - \frac{1}{2}P_k^f G^T C^{-1} G\right) E_{k-1}^a + \mathcal{W}_k\right)\mathcal{U}_k + O\left(h^2\right). \end{aligned} \tag{5.40}$$

For the abstract ansatz that, when applied to E_{k-1}^a, the matrices $\mathcal{U}_k$ evolve according to

$$\mathcal{U}_k = \mathrm{Id} + h\mathcal{R}_k^h + O\left(h^2\right) \tag{5.41}$$

such that $\frac{1}{h}\left(\mathcal{U}_k - \mathrm{Id}\right)$ converges to some continuous-time process $\mathcal{R}$ in a sense to be specified, a similar analysis of Equation (5.40) on existence of a continuous time limit as in Section 5, should apply. This analysis, however, is beyond the scope of this paper.

6. Discussion

We want to highlight here the main aspects of Section 5 and especially Theorem 5.1: the statement (5.8) fully characterizes the continuous time limit of the analyzed filtering algorithms by specifying the sense of convergence as well as the rate of convergence. Interestingly, Theorem 5.1 gives the same limit result for EAKF, ETKF, and Whitaker,

Hamill (2002) all together. This suggests that (5.1) forms a universal limiting ensemble in the sense specified by (5.8) of the class of ESRF algorithms with deterministic model perturbations and in general fully deterministic EnKF. Indeed, consider for instance the deterministic EnKF in [15] in which the proposed transformation of the ensemble deviations E_k^f yields (1.7) with an additional h^2-term. Following the analysis in Section 5 together with using deterministic model perturbations satisfying Assumption 1, one can easily show convergence of the ensemble coming from this filter towards solutions of the ensemble Kalman-Bucy filter (5.1) in the sense of Theorem 5.1.

These results come in handy in the property analysis: in [5], the authors demonstrated in the fully-observed case (i.e. $G = \mathrm{Id}$) that (5.1) together with deterministic model perturbations of the form (5.2), is stable and accurate. Their results easily extend to the case of general deterministic model perturbations fulfilling Assumption 2. Therefore by Theorem 5.1, these properties now carry over to the discrete-time counterparts as they are independent of h by construction. This yields the powerful conclusion that by analyzing one continuous-time equation we immediately analyze a whole class of discrete-time algorithms.

7. Conclusion and Outlook

In this paper, we showed the existence of a continuous time limit of a broad class of ensemble square root filtering algorithms with deterministic model perturbations. In the linear setting, we derived general conditions on these perturbations which enabled us to show convergence of the empirical mean and covariance matrix towards their respective counterparts in the Kalman-Bucy filter in the sense that locally uniformly in time the distance to their continuous-time counterpart decays to zero at rate h. Under further assumptions, we showed for three exemplary algorithms the existence of an ensemble solving the ensemble Kalman-Bucy filtering Equations (5.1) such that the ensemble-mean-square error between the discrete-time and continuous-time ensemble converges to zero locally uniformly in time in expectation at rate h. As shown, this result further holds in the case of nonlinear, Lipschitz-continuous model operators.

An important general observation coming from this analysis is the universality of the limiting ensemble, i.e. we obtain the same limit for all ESRF and furthermore for all fully deterministic EnKF algorithms.

Along the above analysis, we identified and discussed suitable assumptions on the deterministic model perturbations to yield the above convergence results. However, these assumptions were motivated by the aim of approximating the Riccati Equation (1.10) satisfied by the covariance matrices of the ensemble Kalman-Bucy filter which implicitly requires an invertibility condition on the ensemble covariance matrices. We shortly discussed possible generalizations involving projected versions of the Riccati equation, or stochastic perturbations instead.

The latter will form the next step in conducting continuous time limit analyses for these algorithms. In the setting of [9] where the model operator f and observation operator g were assumed to be Lipschitz-continuous and bounded, we were able to show the existence of a continuous time limit using a forecast step of the form

$$X_{t_k}^{(i),f} = X_{t_{k-1}}^{(i),a} + hf\left(X_{t_{k-1}}^{(i),a}\right) + Q^{\frac{1}{2}}\left(W_{t_k}^{(i)} - W_{t_{k-1}}^{(i)}\right) \tag{7.1}$$

as prescribed by the Euler-Maruyama time-discretizations where $W^{(i)}$ are independent

standard Brownian motions. Convergence then holds in the sense that

$$\sup_{t\in[0,T]} \mathbb{E}\left[\sum_{i=1}^{M}\left\|X_{\eta(t)}^{(i),a}-X_t^{(i)}\right\|^2\right]\in O(h) \tag{7.2}$$

where

$$\begin{aligned}\mathrm{d}X_t^{(i)} &= f\left(X_t^{(i)}\right)\mathrm{d}t+Q^{\frac{1}{2}}\mathrm{d}W_t^{(i)}\\ &\quad+\frac{1}{M-1}E_t\mathcal{G}_t^TC^{-1}\left(\mathrm{d}Y_t-\frac{1}{2}\left(g\left(X_t^{(i)}\right)+\bar{g}_t\right)\mathrm{d}t\right)\end{aligned} \tag{7.3}$$

for the appropriate choice of initial conditions (for notation see [9]). The proof, though, highly relies on the boundedness assumption on f and g. Thus, extending the analysis to the above setting with f Lipschitz-continuous and g linear is still a work in progress.

In [9] and in this paper, we used the Euler-Maruyama time-discretization to formulate the filtering algorithms. An interesting extension would be to consider different discretization schemes, e.g. implicit Euler or higher-order Taylor expansions, and to analyze resulting limiting equations and further implications on structure or properties of these algorithms. Again, these are considerations for future research.

Acknowledgements. The research of Theresa Lange and Wilhelm Stannat has been partially funded by Deutsche Forschungsgemeinschaft (DFG) - SFB1294/1 - 318763901.

Appendix A. Proof of the Lemmas.

A.1. Proof of Lemma 3.1. First of all, note that it holds for any $k=1,...,L$ in the sense of symmetric positive semidefinite matrices

$$\begin{aligned}P_k^a &= (\mathrm{Id}-hK_kG)P_k^f = P_k^f-hP_k^fG^T\left(hGP_k^fG^T+C\right)^{-1}GP_k^f\\ &\leq P_k^f.\end{aligned} \tag{A.1}$$

This yields by Assumption 1

$$\begin{aligned}P_k^f &= (\mathrm{Id}+hA)P_{k-1}^a(\mathrm{Id}+hA)^T+h(\mathrm{Id}+hA)Q+hQ(\mathrm{Id}+hA)^T\\ &\quad+\frac{h^2}{M-1}\sum_{i=1}^{M}Q^{\frac{1}{2}}\left(\hat{W}_k^{(i),h}-\hat{w}_k^h\right)\left(\hat{W}_k^{(i),h}-\hat{w}_k^h\right)^TQ^{\frac{T}{2}}\\ &\leq (\mathrm{Id}+hA)P_{k-1}^f(\mathrm{Id}+hA)^T+h(\mathrm{Id}+hA)Q+hQ(\mathrm{Id}+hA)^T\\ &\quad+\frac{h^2}{M-1}\sum_{i=1}^{M}Q^{\frac{1}{2}}\left(\hat{W}_k^{(i),h}-\hat{w}_k^h\right)\left(\hat{W}_k^{(i),h}-\hat{w}_k^h\right)^TQ^{\frac{T}{2}}\end{aligned} \tag{A.2}$$

and

$$\left\|P_k^f\right\|\leq\left(1+2h\|A\|+h^2\|A\|\right)\left\|P_{k-1}^f\right\|+2h\left(1+h\|A\|\right)\|Q\|+h^2\left\|Q^{\frac{1}{2}}\right\|^2\kappa \tag{A.3}$$

which, with a Gronwall argument, yields the claim.

A.2. Proof of Lemma 3.2. Since the updated mean satisfies the recursion

$$\bar{x}^a_{k+1} = \bar{x}^a_k + hA\bar{x}^a_k + K_{k+1}\left(\Delta Y_{k+1} - hG\bar{x}^a_k - h^2GA\bar{x}^a_k\right), \tag{A.4}$$

we can estimate

$$\begin{aligned}\left\|\bar{x}^a_{k+1}\right\|^2 \le{}& \left\|\mathrm{Id} + hA - hK_{k+1}G - h^2K_{k+1}GA\right\|^2 \left\|\bar{x}^a_k\right\|^2 + \left\|K_{k+1}\right\|^2 \left\|\Delta Y_{k+1}\right\|^2 \\ &+ 2\left\langle\left(\mathrm{Id} + hA - hK_{k+1}G - h^2K_{k+1}GA\right)\bar{x}^a_k, K_{k+1}\Delta Y_{k+1}\right\rangle.\end{aligned} \tag{A.5}$$

Observe that by (3.16) it holds

$$\begin{aligned}\mathbb{E}\left[\left\langle\bar{x}^a_k, K_{k+1}\Delta Y_{k+1}\right\rangle\right] &= \int_{t_k}^{t_{k+1}} \mathbb{E}\left[\left\langle\bar{x}^a_k, K_{k+1}GX^{\mathrm{ref}}_s\right\rangle\right] \mathrm{d}s \\ &\le h\mathbb{E}\left[\left\|\bar{x}^a_k\right\|^2\right] + h\int_{t_k}^{t_{k+1}} \mathbb{E}\left[\left\|K_{k+1}\right\|^2\left\|G\right\|^2\left\|X^{\mathrm{ref}}_s\right\|^2\right] \mathrm{d}s\end{aligned} \tag{A.6}$$

where due to (3.17) the last summand is in $O(h^2)$. Thus by boundedness of $\|K_k\|$ uniformly in k, we obtain an estimate of the form

$$\mathbb{E}\left[\left\|\bar{x}^a_{k+1}\right\|^2\right] \le \left(1 + hC^{(1)}(h)\right)\mathbb{E}\left[\left\|\bar{x}^a_k\right\|^2\right] + hC^{(2)}(h) \tag{A.7}$$

which, by a Gronwall argument, yields the claim.

A.3. Proof of Lemma 5.1. The recursion

$$X^{(i),f}_{t_k} = X^{(i),a}_{t_{k-1}} + hAX^{(i),a}_{t_{k-1}} + hQ^{\frac{1}{2}}\hat{W}^{(i),h}_k, \tag{A.8}$$

$$X^{(i),a}_{t_k} = X^{(i),f}_{t_k} - h\hat{K}_kGX^{(i),f}_{t_k} - h\left(K_k - \hat{K}_k\right)G\bar{x}^f_k + K_k\Delta Y_k + \mathcal{R}^h_k e_i \tag{A.9}$$

together with Lemma 4.1 yields an estimate of the following form

$$\begin{aligned}\sum_{i=1}^M \left\|X^{(i),a}_{t_{k+1}}\right\|^2 \le{}& (1 + h\mathcal{C}_1(h))\sum_{i=1}^M \left\|X^{(i),a}_{t_k}\right\|^2 + h\mathcal{C}_2(h) \\ &+ 2M\left\langle\bar{x}^a_k, K_{k+1}\Delta Y_{k+1}\right\rangle + MC_3\left\|K_{k+1}\Delta Y_{k+1}\right\|^2.\end{aligned} \tag{A.10}$$

Using (3.17) we obtain

$$\begin{aligned}\mathbb{E}\left[\left\|K_{k+1}\Delta Y_{k+1}\right\|^2\right] &\le 2\mathbb{E}\left[\left\|K_{k+1}\right\|^2\left(\left\|\int_{t_k}^{t_{k+1}} GX^{\mathrm{ref}}_s \mathrm{d}s\right\|^2 + \left\|C^{\frac{1}{2}}\right\|^2\left\|V_{t_{k+1}} - V_{t_k}\right\|^2\right)\right] \\ &\le \bar{C}_1h^2 + \bar{C}_2h\end{aligned} \tag{A.11}$$

for some constants $\bar{C}_1, \bar{C}_2$ independent of h and uniform in k. Thus in total this yields with (A.6)

$$\mathbb{E}\left[\sum_{i=1}^M \left\|X^{(i),a}_{t_{k+1}}\right\|^2\right] \le \left(1 + h\tilde{\mathcal{C}}_1(h)\right)\mathbb{E}\left[\sum_{i=1}^M \left\|X^{(i),a}_{t_k}\right\|^2\right] + h\tilde{\mathcal{C}}_2(h) \tag{A.12}$$

which, by a Gronwall argument, yields the claim.

Appendix B. Bounds for the modified filter. The modified filter analyzed in [9] using deterministic model perturbations of the form (2.4) reads as follows:

Algorithm 3.

$$X_{t_k}^{(i),f} = X_{t_{k-1}}^{(i),a} + hAX_{t_{k-1}}^{(i),a} + \frac{h}{2}Q\left(P_{k-1}^a\right)^{-1}\left(X_{t_{k-1}}^{(i),a} - \bar{x}_{k-1}^a\right), \tag{B.1}$$

$$X_{t_k}^{(i),a} = X_{t_k}^{(i),f} + K_k\left(\Delta Y_k - \frac{h}{2}G\left(X_{t_k}^{(i),f} + \bar{x}_k^f\right)\right). \tag{B.2}$$

The same analysis as conducted in the main part of this paper also applies for this algorithm in the case of the above particular choice of perturbations due to the following results:

LEMMA B.1. *It holds for each* $k = 1, ..., L$*:*

- *in the sense of symmetric positive semidefinite matrices*

$$P_k^a \le P_k^f \tag{B.3}$$

- *for* h *small enough there exists a constant* $0 < p_T^{*,a} < \infty$ *such that*

$$\left\|(P_k^a)^{-1}\right\| \le p_T^{*,a} \tag{B.4}$$

- *for* h *small enough there exists a constant* $0 < p_T^{*,f} < \infty$ *such that*

$$\left\|P_k^f\right\| \le p_T^{*,f}. \tag{B.5}$$

Proof. On (B.3): using (1.8) for K_k we obtain

$$\begin{aligned}
P_k^a &= \left(\mathrm{Id} - \frac{h}{2}K_kG\right)P_k^f\left(\mathrm{Id} - \frac{h}{2}K_kG\right)^T \\
&= P_k^f - hP_k^fG^T\left(C + hGP_k^fG^T\right)^{-1}GP_k^f \\
&\quad + \frac{h^2}{4}P_k^fG^T\left(C + hGP_k^fG^T\right)^{-1}GP_k^fG^T\left(C + hGP_k^fG^T\right)^{-1}GP_k^f \\
&= P_k^f - \frac{3}{4}hP_k^fG^T\left(C + hGP_k^fG^T\right)^{-1}GP_k^f \\
&\quad - \frac{h}{4}P_k^fG^T\left(C + hGP_k^fG^T\right)^{-1}C^{-1}\left(C + hGP_k^fG^T\right)^{-1}GP_k^f \\
&\le P_k^f.
\end{aligned} \tag{B.6}$$

On (B.4): observe that by the Woodbury matrix identity it holds

$$\begin{aligned}
P_k^a &= P_k^f - \frac{h}{2}K_kGP_k^f - \frac{h}{2}P_k^fG^TK_k^T + \frac{h^2}{4}K_kGP_k^fG^TK_k^T \\
&\ge P_k^f - hK_kGP_k^f \\
&= P_k^f - hP_k^fG^T\left(C + hGP_k^fG^T\right)^{-1}GP_k^f \\
&= \left(\left(P_k^f\right)^{-1} + hG^TC^{-1}G\right)^{-1}
\end{aligned} \tag{B.7}$$

$$\Rightarrow (P_k^a)^{-1} \le \left(P_k^f\right)^{-1} + hG^TC^{-1}G. \tag{B.8}$$

Further one can estimate

$$\begin{aligned} P_k^f &\geq (\mathrm{Id}+hA)\,P_{k-1}^a\,(\mathrm{Id}+hA)^T + hQ + \frac{h^2}{2}AQ + \frac{h^2}{2}QA^T \\ &= (\mathrm{Id}+hA)\left(P_{k-1}^a + \frac{h}{2}Q\right)(\mathrm{Id}+hA)^T + \frac{h}{2}\left(Q - h^2 AQA^T\right). \end{aligned} \tag{B.9}$$

If

$$h^2 < \frac{\lambda_-(Q)}{\lambda_+(Q)\|A\|^2},$$

where λ_- and λ_+ denote smallest and largest eigenvalues, respectively, then $Q - h^2 AQA^T$ is positive semidefinite and by choice of Q it holds

$$P_k^f \geq (\mathrm{Id}+hA)\,P_{k-1}^a\,(\mathrm{Id}+hA)^T. \tag{B.10}$$

If therefore

$$h < \min\left(\frac{1}{\|A\|}, \sqrt{\frac{\lambda_-(Q)}{\lambda_+(Q)\|A\|^2}}\right) = \sqrt{\frac{\lambda_-(Q)}{\lambda_+(Q)}}\,\frac{1}{\|A\|} =: h^*,$$

then

$$\begin{aligned} (P_k^a)^{-1} &\leq \left(P_k^f\right)^{-1} + hG^T C^{-1} G \\ &\leq (\mathrm{Id}+hA)^{-T}\left(P_{k-1}^a\right)^{-1}(\mathrm{Id}+hA)^{-1} + hG^T C^{-1} G \\ &\leq \frac{1}{(1-h\|A\|)^2}\left(P_{k-1}^a\right)^{-1} + hG^T C^{-1} G \\ &\leq \frac{1}{(1-h\|A\|)^{2k}}\,(P_0^a)^{-1} + \left(\sum_{j=0}^{k-1}\frac{1}{(1-h\|A\|)^{2j}}\right) hG^T C^{-1} G. \end{aligned} \tag{B.11}$$

For any $h < h^*$ and any $0 \leq j \leq L$ observe that it holds

$$\frac{1}{(1-h\|A\|)^{2j}} \leq e^{2T\frac{\|A\|}{1-h^*\|A\|}} =: \alpha_T,$$

thus

$$(P_k^a)^{-1} \leq \alpha_T\,(P_0^a)^{-1} + T\alpha_T G^T C^{-1} G \tag{B.12}$$

which yields the bound

$$\left\|(P_k^a)^{-1}\right\| \leq \alpha_T\left\|(P_0^a)^{-1}\right\| + T\alpha_T\|G\|^2\lambda_+(C^{-1}) =: p_T^{a,*}.$$

On (B.5): it holds by using (B.4)

$$\left\|P_k^f\right\| \leq \left(1 + 2h\|A\| + h^2\|A\|^2\right)\left\|P_{k-1}^a\right\| + h\|Q\| + O(h^2) \tag{B.13}$$

thus by (B.3) and a Gronwall argument we obtain that $\|P_k^f\|$ is bounded uniformly in k. ∎

REFERENCES

[1] J.L. Anderson, *An ensemble adjustment Kalman filter for data assimilation*, Mon. Weather Rev., 129(12):2884–2903, 2001. 1

[2] K. Bergemann and S. Reich, *An ensemble Kalman-Bucy filter for continuous data assimilation*, Meteorol. Z., 21:213–219, 2012. 1

[3] C.H. Bishop, B.J. Etherton, and S.J. Majumdar, *Adaptive sampling with the ensemble transform Kalman filter. Part I: Theoretical aspects*, Mon. Weather Rev., 129(3):420–436, 2001. 1, 5.4

[4] G. Burgers, P.J. van Leeuwen, and G. Evensen, *Analysis scheme in the ensemble Kalman filter*, Mon. Weather Rev., 126:1719–1724, 1998. 1, 1

[5] J. de Wiljes, S. Reich, and W. Stannat, *Long-time stability and accuracy of the ensemble Kalman–Bucy filter for fully observed processes and small measurement noise*, SIAM J. Appl. Dyn. Syst., 17(2):1152–1181, 2018. 1, 2, 5, 5.1, 5.3, 6

[6] G. Evensen, *Sequential data assimilation with a nonlinear quasi-geostrohpic model using Monte Carlo methods to forecast error statistics*, J Geophys. Res., 99:10143–10162, 1994. 1

[7] R. Kalman, *A new approach to linear filtering and prediction problems*, J. Basic Eng., 82(1):35–45, 1960. 1

[8] R. Kalman and R. Bucy, *New results in liner filtering and prediction theory*, J. Basic Eng., 83:95–108, 1961. 1

[9] T. Lange and W. Stannat, *On the continuous time limit of the ensemble Kalman filter*, Math. Comp., 90:233–265, 2021. 1, 2.1, 5.1, 5.2, 7, 7, B

[10] T. Lange and W. Stannat, *Mean field limit of ensemble square root filters - discrete and continuous time*, Found. Data Sci., 2021. 4.1, 4.1, 4.1

[11] D.M. Livings, S.L. Dance, and N.K. Nichols, *Unbiased ensemble square root filters*, Phys. D, 237:1021–1028, 2008. 4.4

[12] L. Nerger, T. Janji, J. Schroeter, and W. Hiller, *A unification of ensemble square root filters*, Mon. Weather Rev., 140:2335–2345, 2012. 1

[13] E. Ott, B.R. Hunt, I. Szungogh, A.V. Zimin, E.J. Kostelich, M. Corazza, E. Kalnay, D.J. Patil, and J.A. Yorke, *A local ensemble Kalman filter for atmospheric data assimilation*, Tellus A, 56:415–428, 2004. 4.1

[14] S. Reich, *A dynamical systems framework for intermittent data assimilation*, BIT Numer. Anal., 51:235–249, 2011. 1

[15] P. Sakov and P. Oke, *A deterministic formulation of the ensemble Kalman filter: an alternative to ensemble square root filters*, Tellus A, 60(2):361–371, 2008. 1, 6

[16] A.Y. Sun, A. Morris, and S. Mohanty, *Comparison of deterministic ensemble Kalman filters for assimilating hydrological data*, Adv. Water Resour., 32(2):280–292, 2009. 1

[17] M.K. Tippett, J.L. Anderson, C.H. Bishop, T.M. Hamill, J.S. Whitaker, *Ensemble square root filters*, Mon. Weather Rev., 131:1485–1490, 2003. 1, 4.1, 4.4

[18] X. Wang, C.H. Bishop, and S.J. Julier, *Which is better, an ensemble of positive-negative pairs or centered spherical simplex ensemble?* Mon. Weather Rev., 132:1590–1605, 2004. 4.4, 5.4, 5.4, 5.4

[19] J.S. Whitaker and T.M. Hamill, *Ensemble data assimilation without perturbed observations*, Mon. Weather Rev., 130(7):1913–1924, 2002. 1, 4, 4.2