Correlation between System and Observation Errors in Data Assimilation

TYRUS BERRY AND TIMOTHY SAUER

George Mason University, Fairfax, Virginia

(Manuscript received 5 November 2017, in final form 15 May 2018)

ABSTRACT

Accurate knowledge of two types of noise, system and observational, is an important aspect of Bayesian filtering methodology. Traditionally, this knowledge is reflected in individual covariance matrices for the two noise contributions, while correlations between the system and observational noises are ignored. We contend that in practical problems, it is unlikely that system and observational errors are uncorrelated, in particular for geophysically motivated examples where errors are dominated by model and observation truncations. Moreover, it is shown that accounting for the cross correlations in the filtering algorithm, for example in a correlated ensemble Kalman filter, can result in significant improvements in filter accuracy for data from typical dynamical systems. In particular, we discuss the extreme case where the two types of errors are maximally correlated relative to the individual covariances.

1. Introduction

Consider a discrete time nonlinear dynamical system with state variable $\mathbf{x}_i \in \mathbb{R}^N$ and observations $\mathbf{y}_i \in \mathbb{R}^M$ given by

$$\mathbf{x}_{i+1} = f(\mathbf{x}_i, \boldsymbol{\omega}_i), \tag{1}$$

$$\mathbf{y}_{i+1} = h(\mathbf{x}_{i+1}, \boldsymbol{\nu}_{i+1}), \tag{2}$$

where $\boldsymbol{\omega}_i$ is called the system or dynamical noise (or the stochastic forcing) and $\boldsymbol{\nu}_{i+1}$ is called the observation noise. In practice these noise terms are needed to account for model mismatch, truncation errors caused by differing resolutions, and stochastic terms such as instrument errors. There has been considerable recent interest in the implications of these different sources of error (e.g., Satterfield et al. 2017; Hodyss and Nichols 2015; Van Leeuwen 2015; Janjić et al. 2018).

On the other hand, most filtering algorithms are designed based on a separation of dynamical and observation noise. The nonlinear filtering literature typically considers noise $\boldsymbol{\omega} \sim \mathcal{N}(\mathbf{0}, \mathbf{Q})$ and $\boldsymbol{\nu} \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$, allowing correlation within type, but tends to dismiss correlation between $\boldsymbol{\omega}$ and $\boldsymbol{\nu}$. In this article, we argue the importance of modeling correlations between system and observation noise. Specifically, we will consider Kalman filters and ensemble Kalman filters (EnKF) that are derived based on the assumption that

$$\begin{bmatrix} \boldsymbol{\omega}_i \\ \boldsymbol{\nu}_{i+1} \end{bmatrix} \sim \mathcal{N}(\mathbf{0}, \mathbf{C}), \quad \text{where} \quad \mathbf{C} = \begin{bmatrix} \mathbf{Q} & \mathbf{S} \\ \mathbf{S}^{\mathrm{T}} & \mathbf{R} \end{bmatrix}, \tag{3}$$

and $\mathbf{C}$ is assumed to be symmetric and positive semidefinite. The $N \times M$ matrix $\mathbf{S}$ contains the cross covariances between the variables $\boldsymbol{\omega}_i$ and $\boldsymbol{\nu}_{i+1}$.

In order for $\mathbf{C}$ to be positive semidefinite, both $\mathbf{Q}$ and $\mathbf{R}$ must be positive semidefinite. The classical viewpoint corresponds to simply extracting $\mathbf{Q}$ and $\mathbf{R}$ and ignoring $\mathbf{S}$. However, it is reasonable to expect that in many physical systems, the truncations causing the noise in the state of the system would also affect the sensor or observation system. The goal of this paper is to establish that 1) correlations between system and observation errors are likely to be common in applied data assimilation problems caused by truncation of infinite-dimensional solutions and 2) incorporating these correlations can dramatically improve filter results.

If we consider the true state to be a function of space and time evolving in an infinite-dimensional function space, then the truncated true state and the observation are essentially two different finite-dimensional projections of this infinite-dimensional space (Dee 1995; Janjić and Cohn 2006; Oke and Sakov 2008). In section 2 we will show that the errors between the exact projections and the finite-dimensional approximations are correlated for generic observations. Error correlations

Corresponding author: Timothy Sauer, tsauer@gmu.edu

DOI: 10.1175/MWR-D-17-0331.1

arising from model truncation have been previously observed (Mitchell and Daley 1997; Hamill and Whitaker 2005; Liu and Rabier 2002). In particular, models are often composed of discrete dynamics occurring at points of a two- or three-dimensional grid. Remote observations by satellite or radiosonde can be viewed as integrations over a region including several grid points. Here, we provide a general framework to explain correlations between these quantities. We also consider other sources of error such as model mismatch and instrument error, and we show that significant correlations persist except in the case where instrument error dominates (since this error will be modeled as white noise that is uncorrelated with the state).

In section 3, a correlated version of the EnKF is developed that takes the correlations in (3) into account, and recovers the Kalman equations for linear systems. In section 4, the correlated unscented Kalman filter (CUKF; an unscented version of the EnKF) is applied to examples from section 2. Using an appropriate $\mathbf{S}$ substantially improves output accuracy, while a filter that ignores the cross-correlation matrix $\mathbf{S}$ can lead to dramatically suboptimal results.

In section 5 we investigate the effect of correlations in greater detail. First, we show that in the linear case when $M = N$, for any $\mathbf{Q}$ and $\mathbf{R}$ there exists a "maximal" $\mathbf{S}$ for which perfect recovery of the state variables is possible. Second, we demonstrate examples of perfect recovery in nonlinear systems with such maximal $\mathbf{S}$. In these examples, if one ignores a maximal $\mathbf{S}$ (setting $\mathbf{S} = 0$ in the filter while the true $\mathbf{S}$ is maximal), variables that would have been perfectly recovered by using the true $\mathbf{S}$ will instead be estimated with variance on the order of the entries in $\mathbf{R}$ (e.g., see Fig. 6).

2. Correlation between system and observation errors

To understand how correlations arise in applied data assimilation, we must first leave behind the idealized scenario described in (1) and (2). Following Dee (1995) and Janjić and Cohn (2006), we describe the true evolution and observation processes by replacing the discrete solution $\mathbf{x}_i \in \mathbb{R}^N$ with an infinite-dimensional solution $\tilde{\mathbf{x}}(z, t)$, which has some regularity (continuity or differentiability) in the spatial variable z and temporal variable t. We should note that the following analysis is very similar to Satterfield et al. (2017), except that we consider an infinite-dimensional solution $\tilde{\mathbf{x}}$ instead of a high-resolution solution [which is denoted $\mathbf{x}_H^t$ in Satterfield et al. (2017)].

The discrete time solution and the observations can be viewed as projections of this continuous solution. The desired finite-dimensional discrete time solution

$$\mathbf{x}_i = \Pi_i(\tilde{\mathbf{x}})$$

is a projection of the continuous solution. Meanwhile, the finite-dimensional discrete time observations

$$\mathbf{y}_i^o = \mathscr{H}_i(\tilde{\mathbf{x}}) + \boldsymbol{\varepsilon}_i^I$$

are given by another projection of the solution $\mathscr{H}_i$, plus an instrument noise term $\boldsymbol{\varepsilon}_i^I$. Define the system error as the local truncation error (LTE) of the discrete solver f, namely,

$$\mathbf{w}_i \equiv \mathbf{x}_{i+1} - f(\mathbf{x}_i) = \Pi_{i+1}(\tilde{\mathbf{x}}) - f[\Pi_i(\tilde{\mathbf{x}})],$$

and define the observation error by

$$\boldsymbol{\varepsilon}_{i+1}^o \equiv \mathbf{y}_{i+1}^o - \hat{h}(\mathbf{x}_{i+1}) = \mathscr{H}_{i+1}(\tilde{\mathbf{x}}) - \hat{h}[\Pi_{i+1}(\tilde{\mathbf{x}})] + \boldsymbol{\varepsilon}_{i+1}^I,$$

where $\mathscr{H}_i$ is the true observation projection and $\hat{h}$ is the approximate discrete observation function. Letting h be a consistent discretization of the true observation projection [as in (5)], we further decompose the observation error in terms of the representation error

$$\boldsymbol{\varepsilon}_{i+1}^R \equiv \mathscr{H}_{i+1}(\tilde{\mathbf{x}}) - h[\Pi_{i+1}(\tilde{\mathbf{x}})],$$

and the observation model error

$$\boldsymbol{\varepsilon}_{i+1}^H \equiv h[\Pi_{i+1}(\tilde{\mathbf{x}})] - \hat{h}[\Pi_{i+1}(\tilde{\mathbf{x}})],$$

so that the total observation error becomes

$$\boldsymbol{\varepsilon}_{i+1}^o = \boldsymbol{\varepsilon}_{i+1}^R + \boldsymbol{\varepsilon}_{i+1}^H + \boldsymbol{\varepsilon}_{i+1}^I.$$

The above definitions are very similar to those found in Eqs. (1)–(7) in Satterfield et al. (2017) except that we have replaced their high-resolution observation function $\mathbf{H}_L$ and the truncation smoother $\mathbf{S}_{\text{sc}}$ with finite-dimensional projections of infinite-dimensional spaces, namely, $\mathscr{H}$ and Π, respectively. Projections from infinite-dimensional spaces were also considered by Janjić et al. (2018) who also considered additional terms in the decomposition of the observation error. Since our main focus is the correlations between system and observation error we will restrict our attention to the three sources of observation error listed above. We should note that while the observational errors can be formally decomposed as above, the individual components may be correlated (especially the model error and representation error terms) so in general we do not expect a corresponding decomposition of the observation error covariance matrix.

For simplicity we assume that $\mathbf{w}_i$, $\mathbf{v}_{i+1}$ are both mean zero and define the error variances to be

$$\mathbf{Q} = \lim_{T\to\infty} \frac{1}{T} \sum_{i=1}^{T} \mathbf{w}_i \mathbf{w}_i^{\mathrm{T}} \quad \text{and} \quad \mathbf{R} = \lim_{T\to\infty} \frac{1}{T} \sum_{i=1}^{T} \boldsymbol{\varepsilon}_i^o (\boldsymbol{\varepsilon}_i^o)^{\mathrm{T}}.$$

We briefly note that, while it may be reasonable to assume that the representation and observation model errors are uncorrelated from the instrument error $\boldsymbol{\varepsilon}^I$, their cross covariance is

$$\lim_{T\to\infty} \frac{1}{T} \sum_{i=1}^{T} \boldsymbol{\varepsilon}_i^R (\boldsymbol{\varepsilon}_i^H)^{\mathrm{T}} = \lim_{T\to\infty} \frac{1}{T} \sum_{i=1}^{T} \{\mathscr{H}_i(\tilde{\mathbf{x}}) - h[\Pi_i(\tilde{\mathbf{x}})]\} \times \{h[\Pi_i(\tilde{\mathbf{x}})] - \hat{h}[\Pi_i(\tilde{\mathbf{x}})]\}^{\mathrm{T}},$$

which seems unlikely to be zero. However, our main concern here is the cross covariance between system and observation error, which is defined as

$$\begin{aligned} \mathbf{S} &= \lim_{T\to\infty} \frac{1}{T} \sum_{i=1}^{T} \mathbf{w}_i (\boldsymbol{\varepsilon}_{i+1}^o)^{\mathrm{T}} \\ &= \lim_{T\to\infty} \frac{1}{T} \sum_{i=1}^{T} \Big[\{\Pi_{i+1}(\tilde{\mathbf{x}}) - f[\Pi_i(\tilde{\mathbf{x}})]\} \{\mathscr{H}_{i+1}(\tilde{\mathbf{x}}) - h[\Pi_{i+1}(\tilde{\mathbf{x}})]\}^{\mathrm{T}} \\ &\quad + \{\Pi_{i+1}(\tilde{\mathbf{x}}) - f[\Pi_i(\tilde{\mathbf{x}})]\} \{h[\Pi_{i+1}(\tilde{\mathbf{x}})] - \hat{h}[\Pi_{i+1}(\tilde{\mathbf{x}})]\}^{\mathrm{T}} \Big] \end{aligned} \tag{4}$$

(assuming uncorrelated instrument errors). In this general setting it is already puzzling why one would assume that $\mathbf{S} = 0$. By averaging over the full time series, we are defining global covariance matrices that are fixed in time. More generally, one could also consider time-varying covariance matrices that are either localized in time or in state space. However, this would only change the indices of the averages above and in most situations one should expect nonzero $\mathbf{S}$ matrices. In the next section we will show explicit examples of substantial correlation between system and observation errors in many practical situations. By imposing additional assumptions on the observation model (viz., that it is linear and local in both space and time) we will be able to show that the correlations are close to maximal. When these assumptions on the observation model are satisfied, we expect $\mathbf{S}$ to be very important to the filtering problem except when the observation errors are dominated by instrument error.

a. Evaluation and averaging projections

We will assume that the finite-dimensional dynamics f and observation function h are *consistent*, meaning that the errors $\mathbf{w}_i$ and $\boldsymbol{\varepsilon}_{i+1}^R$ go to zero in the limit of small discretization parameters Δz in space and Δt in time. As an example, consider the case when the evolution of the full solution $\tilde{\mathbf{x}}$ is governed by a PDE

$$\frac{\partial}{\partial t} \tilde{\mathbf{x}} = \mathscr{F}(\tilde{\mathbf{x}}),$$

and consider the projection of the state onto a grid $\mathbf{z} = \{z_j\}$ at time t_i, namely,

$$\Pi_i(\tilde{\mathbf{x}}) = \tilde{\mathbf{x}}(\mathbf{z}, t_i).$$

If we assume that the solution $\tilde{\mathbf{x}}$ has $n+1$ continuous derivatives in space and $m+1$ in time, we can use a solver that is order n in space and order m in time to obtain the system error

$$\mathbf{w}_i = \tilde{\mathbf{x}}(\mathbf{z}, t_{i+1}) - f[\tilde{\mathbf{x}}(\mathbf{z}, t_i)] = a_i \Delta t^m + b_i \Delta z^n + \text{h.o.t.},$$

where for simplicity we assume a uniform spatial grid Δz in each dimension. The coefficients a_i and b_i depend on the derivatives of $\tilde{\mathbf{x}}$ within Δz and Δt of $(\mathbf{z}, t_i)$.

Now consider the associated observation operator $\mathscr{H}$. Rather than sampling at an instantaneous time, most observation modes have an associated time constant, and an average value of an interval $[t_i - \delta, t_i + \delta]$ with some weight function Ψ is returned. Similarly, the observation may involve multiple spatial grid points, as in the case of satellite observations involving radiative transfer that explicitly integrate over the entire vertical grid. Even for very local observations, the true observing system may be located between grid points, thereby involving interpolation between grid points. Thus, we assume the true observation has the following form:

$$\mathscr{H}_{i+1}(\tilde{\mathbf{x}}) = \int_{|s-t_i|<\delta} \int_{\|\mathbf{w}-\mathbf{z}\|<\varepsilon} \tilde{\mathbf{x}}(\mathbf{w}, s) \Psi(\mathbf{w}, s)\, d\mathbf{w}\, ds,$$

meaning that a consistent observation function h should be a quadrature rule for approximating this integral. Assuming that the discrete observation function has order $q \le n$ convergence in space and order $r \le m$ in time, the representation error is

$$\boldsymbol{\varepsilon}_{i+1}^R = \mathscr{H}_{i+1}(\tilde{\mathbf{x}}) - h[\mathbf{x}(\mathbf{z}, t_{i+1})] = c_{i+1} \Delta t^r + d_{i+1} \Delta z^q + \text{h.o.t.}, \tag{5}$$

where the coefficients depend on the derivatives of $\tilde{\mathbf{x}}$ within Δz and Δt of $(\mathbf{z}, t_{i+1})$.

The situation described above is common in applications, namely, where the discrete solution is given by (or equivalent to) evaluation on a grid and the true observation operator is a local weighted average of the full solution. In this case, we find the cross covariance of the system and observation errors to be (excluding higher-order terms)

$$\mathbf{S} = \lim_{T\to\infty} \sum_{i=1}^{T} (a_i \Delta t^m + b_i \Delta z^n)(c_{i+1}\Delta t^r + d_{i+1}\Delta z^q)$$

and in the limit of small Δz and Δt, the derivatives in a, b, c, d will all be evaluated at points less that $(\Delta z, \Delta t)$ apart. Therefore, up to higher-order terms, a, b, c, d can be rewritten in terms of derivatives evaluated at the same point $(\mathbf{z}, t_i)$. Notice in particular that when $m = r$, the coefficients a and c are the same up to a scalar, and similarly when $n = q$ the coefficients b and d are linear combinations of the same order derivatives. While it is possible for these terms to combine so as to exactly cancel when averaged over time, the correlation will often be nonzero.

As a special case, consider the situation when both the system and observation errors are dominated by the same single variable (either time or one of the spatial variables). In this case, the leading-order terms would differ only by a constant, so that up to higher-order terms the system error and observation error would be multiples of one another. This not only implies that $\mathbf{S} \neq 0$ but also, as we will show in section 5, that the system and observation errors are *maximally correlated* up to higher-order corrections, so that S is as large as possible relative to the individual variances. For example, in the case of satellite observations of radiative transfer, the true observation integrates over the entire vertical component of the atmosphere, whereas the integral may be very localized in time and the horizontal variables. This would suggest a relatively large error in terms of vertical Δz (depending on the number of vertical grid points and the order of the quadrature rule used to estimate the radiative transfer) even if the model was perfectly specified. If the vertical model error also dominated the model error then we would expect a high correlation of the model and observation errors.

Similar to the above analysis, we can also consider the observation model error to be a difference of quadrature rules given by the observation function h and an approximate function $\hat{h}$. With these assumptions we find

$$\boldsymbol{\varepsilon}^H_{i+1} = h[\mathbf{x}(\mathbf{z}, t_{i+1})] - \hat{h}[\mathbf{x}(\mathbf{z}, t_{i+1})] = \sum_j (\alpha_j - \hat{\alpha}_j)\mathbf{x}(z_j, t_{i+1}),$$

where α and $\hat{\alpha}$ are the quadrature weights for h and $\hat{h}$, respectively. Since both observation functions are assumed to be local, we again obtain an error in terms of Δz and Δt according to the order of agreement between h and $\hat{h}$, which will lead to maximal correlations with the truncation errors. In practice, either the model error or the representation error could be the dominant term, but in either case we find nontrivial correlation with the system errors. In what follows, we focus on examples where representation error dominates, since this term will be the easiest to estimate in practice. The importance of treating correlated errors will hold in either case. We now turn to some concrete examples.

Even for general nonlinear observations functions h and approximate observation functions $\hat{h}$, we expect the correlation given in (4) to be nonzero, since setting (4) equal to zero imposes a nontrivial constraint on the system. The analysis in this section shows that for linear observations that have a local structure in space and time, truncation error can be expected to be nearly maximally correlated with both observation model error and representation error. When the observations are nonlinear or not local in space and time we still expect error correlations to be present, although if not close to maximal, they may not be crucial to the filtering problem.

b. *Time-averaged observations of an ODE*

First consider the case when the true solution $\tilde{\mathbf{x}}$ is discrete in space, meaning that $\tilde{\mathbf{x}} = \tilde{\mathbf{x}}(t) \in \mathbb{R}^N$ is a vector evolving continuously in time. Assume that the true evolution of $\tilde{\mathbf{x}}$ is governed by an ODE

$$\tilde{\mathbf{x}}' = \mathscr{F}(\tilde{\mathbf{x}}),$$

and the projection $\mathbf{x}_i = \Pi_i(\tilde{\mathbf{x}}) = \tilde{\mathbf{x}}(t_i)$ is evaluation at a discrete time grid $\{t_i\}$. If the discrete evolution operator f is Euler's method, the system error is

$$\begin{aligned}\mathbf{w}_i &= \mathbf{x}_{i+1} - f(\mathbf{x}_i)\\ &= \mathbf{x}_{i+1} - [\mathbf{x}_i + \Delta t\mathscr{F}(\mathbf{x}_i)]\\ &= \frac{1}{2}(\Delta t)^2\tilde{\mathbf{x}}''(t_i) + \mathscr{O}[(\Delta t)^3]\\ &= \frac{1}{2}(\Delta t)^2\tilde{\mathbf{x}}''(t_{i+1}) + \mathscr{O}[(\Delta t)^3],\end{aligned} \tag{6}$$

where we have used $\tilde{\mathbf{x}}''(t_{i+1}) = \tilde{\mathbf{x}}''(t_i) + \mathscr{O}(\Delta t)$.

We assume the true observation is an unweighted average over a short interval $[t_{i+1} - \delta, t_{i+1} + \delta]$, namely,

$$\mathscr{H}_{i+1}(\tilde{\mathbf{x}}) = \frac{1}{2\delta}\int_{t_{i+1}-\delta}^{t_{i+1}+\delta} \tilde{\mathbf{x}}(s)\, ds.$$

Consider the discrete observation function h to be a consistent quadrature rule using the grid points falling within the interval $[t_{i+1} - \delta, t_{i+1} + \delta]$. In particular

when $\delta \leq \Delta t$ we find $h(\mathbf{x}_{i+1}) = \mathbf{x}_{i+1}$ and the representation error is

$$\boldsymbol{\varepsilon}^R_{i+1} = \mathscr{H}_{i+1}(\tilde{\mathbf{x}}) - h(\mathbf{x}_{i+1}) = \frac{1}{2\delta}\int_{t_{i+1}-\delta}^{t_{i+1}+\delta} \tilde{\mathbf{x}}(s)\, ds - \mathbf{x}_{i+1}.$$

Expanding $\tilde{\mathbf{x}}(s)$ in a Taylor series centered at t_{i+1} and cancelling odd terms we find

$$\begin{aligned}\boldsymbol{\varepsilon}^R_{i+1} &= \frac{1}{2\delta}\int_{t_{i+1}-\delta}^{t_{i+1}+\delta} \mathbf{x}_{i+1} + \frac{(s-t_{i+1})^2}{2}\tilde{\mathbf{x}}''(t_{i+1}) \\ &\quad + \mathscr{O}[(s-t_{i+1})^4]\, ds - \mathbf{x}_{i+1} \\ &= \frac{\delta^2}{6}\tilde{\mathbf{x}}''(t_{i+1}) + \mathscr{O}(\delta^4). \end{aligned} \tag{7}$$

Comparing (6) and (7) shows that up to leading order, $\boldsymbol{\varepsilon}^R_{i+1}$ and $\mathbf{w}_i$ are directly proportional, meaning that if the leading-order terms are nonzero, they are correlated.

If a backward Euler method were used instead of a forward Euler method, the leading term of the system error would be $-(\Delta t^2/2)\tilde{\mathbf{x}}''(t_{i+1})$, which is negatively correlated with $\mathbf{v}_i = (\delta^2/6)\tilde{\mathbf{x}}''(t_{i+1})$. Moreover, if the observations arose from an asymmetric average (e.g., over $[t_{i+1}-\delta, t_{i+1}]$) then the observation error would be in terms of the first derivative of $\tilde{\mathbf{x}}$ instead of the second derivative; however, these different derivatives are still likely to be correlated since

$$\tilde{\mathbf{x}}'' = \frac{d}{dt}\tilde{\mathbf{x}}' = \frac{d}{dt}\mathscr{F}(\tilde{\mathbf{x}}) = \mathbf{D}\mathscr{F}(\tilde{\mathbf{x}})\tilde{\mathbf{x}}'.$$

If higher-order methods were used, then the errors would involve higher-order derivatives, which are still highly likely to be correlated.

c. *Estimation of the full covariance matrix*

The correlations described above will be illustrated in two simple examples. To show the effects most clearly, we assume a perfect model. We begin with a simple ODE solver.

Consider using forward Euler on the Lorenz-63 system (Lorenz 1963)

$$\begin{aligned}\dot{x}_1 &= \sigma(x_2 - x_1) \\ \dot{x}_2 &= x_1(\rho - x_3) - x_2 \\ \dot{x}_3 &= x_1 x_2 - \beta x_3, \end{aligned} \tag{8}$$

where $\sigma = 10$, $\rho = 28$, and $\beta = 8/3$. To demonstrate the correlation derived above, we first used a higher-order integrator [Runge–Kutta fourth order (RK4) with a 0.005 time step] to produce a finely sampled ground truth signal. To define the truncated model, we used a forward Euler method with $\Delta t = 0.1$. We used a 21-point composite trapezoid rule to approximate the integrated observation with $\delta = 0.05$. In Figs. 1a and 1b we show the correlation between the system error, in this case the local truncation error (LTE) of the Euler solver, and the observation error, in this case only the representation error, as a function of time.

In Fig. 1c we show the estimated covariance matrix $\mathbf{C}$, which reveals the strong correlations ($\mathbf{S} \neq 0$) between the system and observation errors. The covariance matrix $\mathbf{C}$ is estimated by concatenating the system and observation errors at each step into a six-dimensional vector, and then computing the empirical covariance matrix of these vectors (averaged over $T = 12\,000$ discrete time steps). The estimated $\mathbf{C}$ matrix will be used in section 4 in a nonlinear filter, and this method can be used to estimate the $\mathbf{C}$ matrix for general problems as long as one can afford a long offline run using a very fine discretization. Figures 1d–f show the same phenomenon for a more accurate solver, RK4 with a time step of $\Delta t = 0.05$. Although the system errors are much smaller, the correlation with observation errors are still evident.

The difference between the positive correlations in Figs. 1b and 1c and negative correlations in Figs. 1e and 1f will have a noticeable effect in filter accuracy, as shown in section 4a. In section 5, we will establish a theory explaining this disparity in the linear case.

As a second example, consider spatiotemporal dynamics $\tilde{\mathbf{x}}(z, t)$ given by the Kuramoto–Sivashinsky PDE (Kuramoto and Tsuzuki 1976; Sivashinsky 1977)

$$\tilde{\mathbf{x}}_t = -\tilde{\mathbf{x}}_{zzzz} - \tilde{\mathbf{x}}_{zz} - \tilde{\mathbf{x}}\tilde{\mathbf{x}}_z, \tag{9}$$

defined on a periodic domain with length $L = 100$. For simplicity, we will use an explicit method applying RK4 in time and second-order finite-difference formulas for the spatial derivatives. While an implicit method would be stable for much larger values of Δt, we will later see that the filter will be able to stably recover the signal from noisy observations even for large Δt (the filter uses the observations to stabilize what would otherwise be an unstable numerical scheme). To obtain a high-resolution "ground truth" signal we use a grid with 512 equally spaced spatial grid steps and a time step of 10^{-4}.

To simulate a PDE integrator in practice, we truncate the model, applying the same RK4 solver with a reduced number of grid steps and a larger Δt. Let $P \leq 512$ be the number of spatial grid steps on $[0, L]$ and let Δz be the spatial step size, so that $P\Delta z = L$. We define an observation function that integrates in space as

$$\mathscr{H}_i(\tilde{\mathbf{x}})_j = \frac{1}{2\delta}\int_{z_j-\delta}^{z_j+\delta} \tilde{\mathbf{x}}(w, t_i)\, dw, \tag{10}$$

where δ defines the spatial region over which the observations are averaged. So for example, when $\delta = L/128$, we

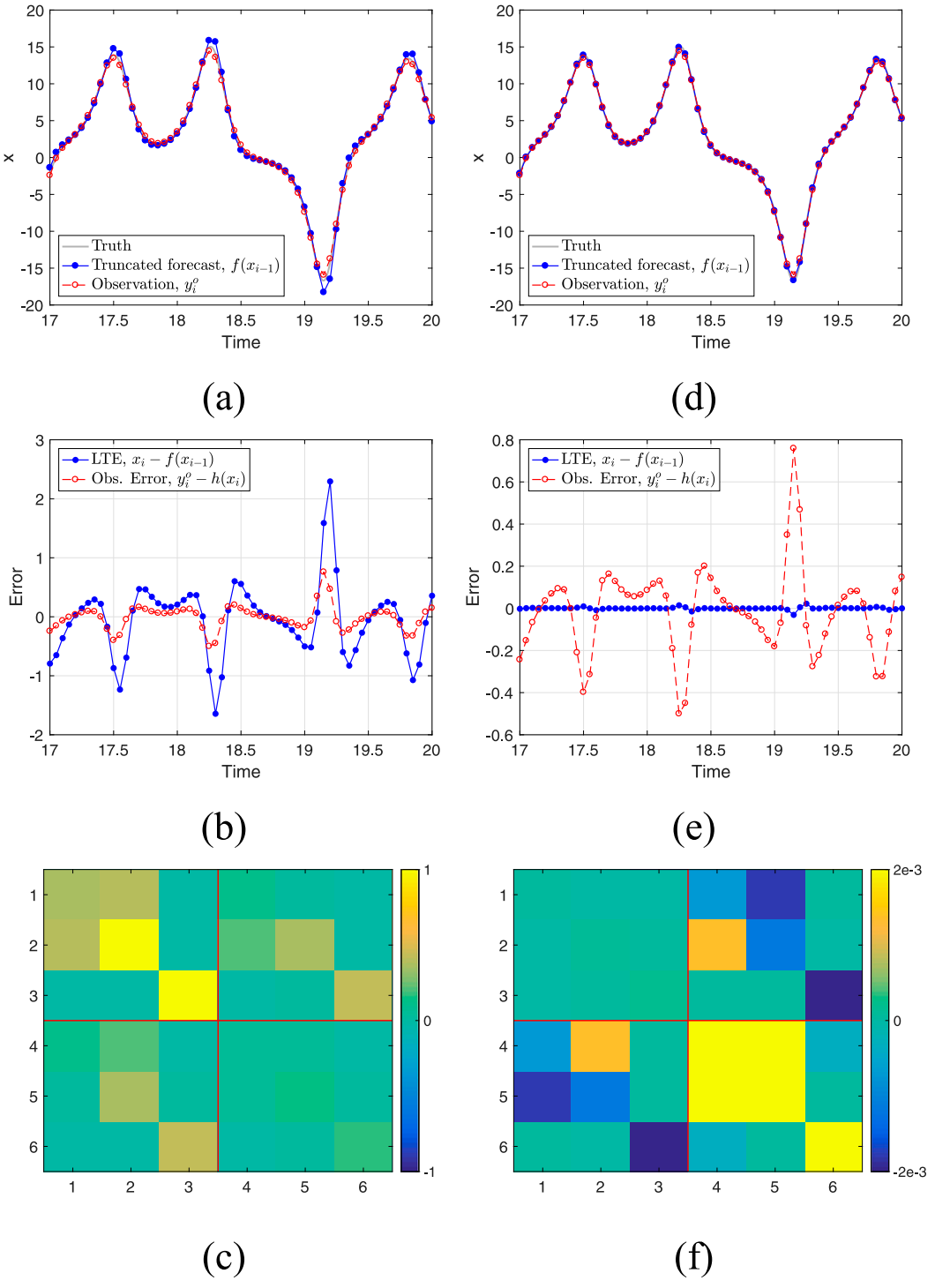

FIG. 1. Demonstrating correlated noise in the truncated L63 system. (a) Comparing the true x coordinate of L63 (gray) to a one-step forecast using the forward Euler method (blue, solid curve) with $\Delta t = 0.05$ and the integrated observation (red, dashed curve). (b) Comparing the system error (defined as LTE) to the observation error (only representation error in this example), note the correlation. (c) Empirical covariance matrices, $\mathbf{C}$ with red lines dividing the $\mathbf{Q}$, $\mathbf{S}$, $\mathbf{S}^{\mathrm{T}}$, and $\mathbf{R}$ blocks. (d)–(f) As in (a)–(c), but using the RK4 integrator with the same coarse time step $\Delta t = 0.05$. Color ranges in (c) and (f) are selected to emphasize the $\mathbf{S}$ matrix and may saturate for $\mathbf{Q}$ and $\mathbf{R}$. Notice positive correlations in (b),(c) and negative correlations in (e),(f).

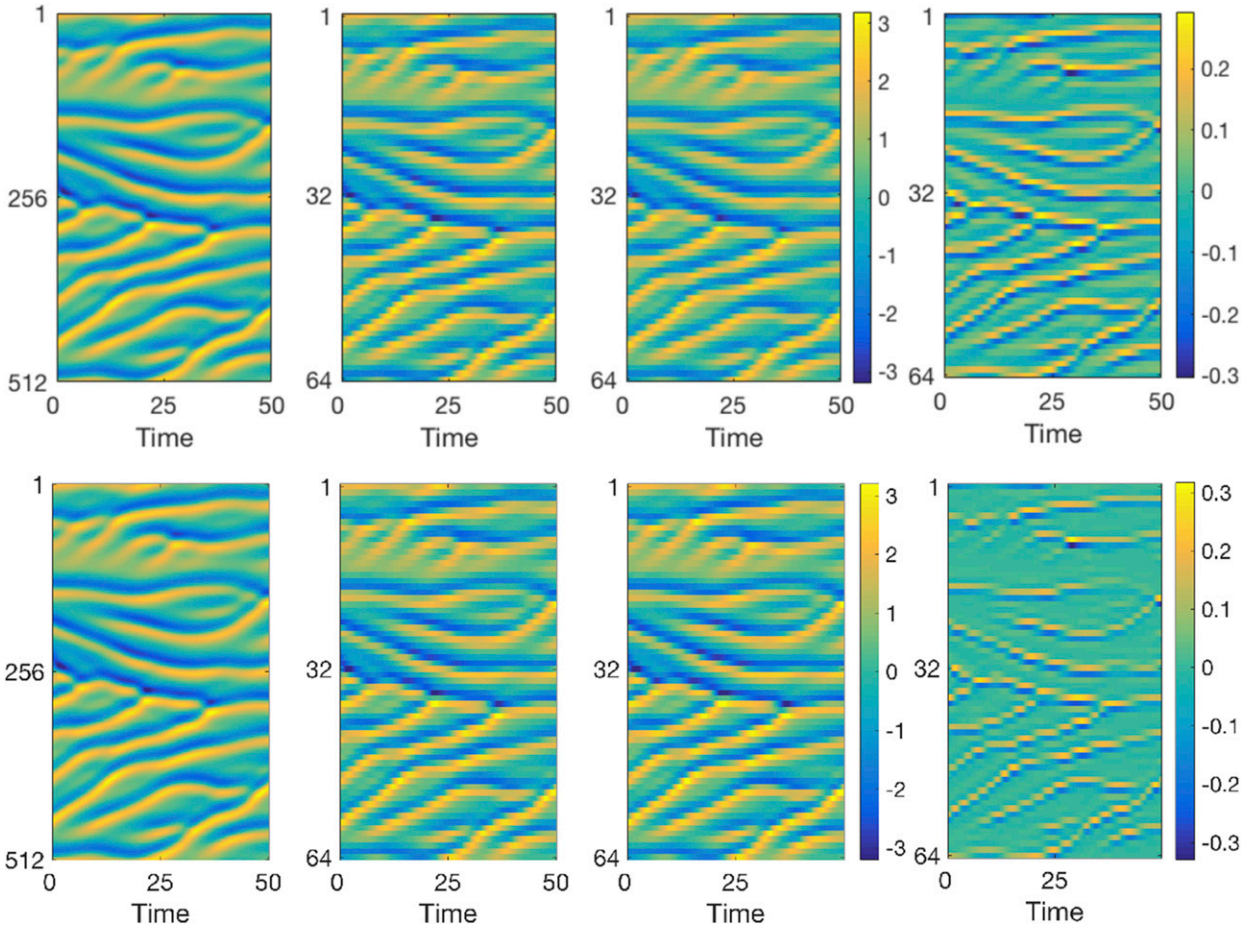

FIG. 2. (top) (left) Ground truth 512 gridpoint solution (middle left) the same solution decimated to 64 grid points (middle right) the observation, which integrates the leftmost solution over 9 grid points before truncating, and (right) the observation error, which is the difference between the middle two solutions. (bottom) (left),(middle left) As in (top). (middle right) The 1-step integrator output from the truncated model, using 64 grid points and $\Delta t = 0.1$. (right) System error, difference between the middle two solutions.

will first estimate the true observation $\mathscr{H}_i$ using a composite trapezoid rule with $(2\delta/\Delta z) + 1 = 9$ grid points from the full 512 gridpoint solution. If we consider a truncated model with $P = 128$ grid points, then the observation function will have to be estimated using only $(2\delta/\Delta z) + 1 = 3$ grid points. If we consider a truncated model with $P = 64$ grid points, we will estimate the observation functions using only a single grid point, in other words our coarse model for the observation function will be direct observation at each grid point. This is because the integration range δ has become smaller than our truncated Δz. In each case, the estimated observation function h is consistent with the true observation $\mathscr{H}_i$, so we are not considering observation model error yet but only representation error.

In Fig. 2 we compare the full resolution and truncated solutions for 64 grid points and $\Delta t = 0.1$. The observation representation errors (top right) are tightly correlated to the system errors, consisting of truncation errors from the one-step integrator (bottom right). Both errors are also correlated with the underlying truth solution.

It is helpful to view the empirical full covariance matrix $\mathbf{C}$ of the system plus observation errors that can be estimated from the data in Fig. 2. In Fig. 3 (left) we show the matrix $\mathbf{C}$ where the submatrices $\mathbf{Q}$, $\mathbf{S}$, $\mathbf{S}^{\mathrm{T}}$, and $\mathbf{R}$ have been spatially averaged [using the symmetry of (9) on the periodic domain], which reveals the strong correlation between the system and observation errors. In Fig. 3 (right), we plot the sorted eigenvalues of the empirically estimated $\mathbf{C}$ matrix (black, solid curve), which result purely from the correlation of the truncation error in the model and the representation error in the observation.

Next, we consider the case of a large observation model error and show that the observed correlations are still present in this case. While leaving the truncated

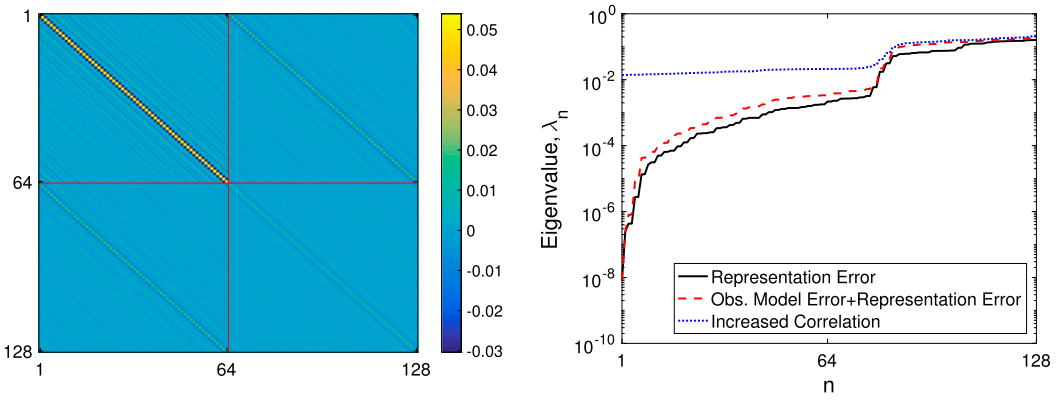

FIG. 3. For the Kuramoto–Sivashinsky model truncated onto 64 grid points with $\Delta t = 0.1$ we show (left) the spatially averaged **C** matrix, note that the cross covariance between dynamical truncation errors and observation representation errors has a larger magnitude than the variance of the observation errors and (right) the eigenvalues of **C** (black, solid curve) decay quickly. The presence of eigenvalues that are very close to zero indicates that the matrix **C** is close to maximally correlated as we will show in section 5. We also show the eigenvalues for a correlation matrix computed in the presence of both observational model error and representation error (red, dashed curve). Finally, we show the eigenvalues after the diagonal of the **S** matrix is increased by 50% (blue, dotted curve).

observation function unchanged, we changed the true observation function to compute a weighted spatial average of the four nearest grid points to each observed point. Explicitly, prior to applying the composite trapezoid rule, we first average each location with its four nearest neighbors with weights:

$$\tilde{\mathbf{x}}(z_j) = \frac{\tilde{\mathbf{x}}(z_j) + \tilde{\mathbf{x}}(z_{j+1})/2 + \tilde{\mathbf{x}}(z_{j-1})/2 + \tilde{\mathbf{x}}(z_{j+2})/4 + \tilde{\mathbf{x}}(z_{j-2})/4}{2.5},$$

which maintains the local structure but also implies that the quadrature rule used for the truncated state is no longer consistent with this new observation. In Fig. 3 (right), we plot the eigenvalues of the empirically estimated **C** matrix (red, dashed curve) for this new observation, which contains both observation model error and representation error. While the correlation is slightly further than maximal, the same strong decay of eigenvalues and almost singular behavior is present as in the case of representation error alone.

In Fig. 3 we should emphasize the presence of many small eigenvalues, indicating that the **C** matrix is close to singular. To show that these small eigenvalues come from the special structure of the correlated errors, we artificially increased the diagonal of the **S** submatrices by 50%, and the resulting eigenvalues are shown as the blue dashed curve in Fig. 3. In other words, by increasing the correlation beyond the true correlations we destroy the special ''almost rank deficient'' nature of this type of correlated error. In section 5, we focus on this phenomenon, which indicates that the system and observation errors are very close to being *maximally correlated*. We will show that the strong correlation between the system and observation errors has significant consequences for the ability to estimate the true state from the observations.

3. Filtering in the presence of correlations

In this section we review versions of the Kalman filter for linear and nonlinear dynamics, which include full correlations of system and observation errors. We begin with the linear formulas, and then discuss the unscented version of the ensemble Kalman filter for nonlinear models.

a. *The Kalman filter for correlated system and observation noise*

We begin by reviewing the Kalman update equations for a linear system with correlated noise (e.g., see Simon 2006). Assume the following model and observation equations:

$$\mathbf{x}_i = f(\mathbf{x}_{i-1}) + \mathbf{\Gamma}\boldsymbol{\omega}_{i-1}, \tag{11}$$

$$\mathbf{y}_i = h(\mathbf{x}_i) + \mathbf{J}\boldsymbol{\nu}_i, \tag{12}$$

where $\mathbf{F} = f$ and $\mathbf{H} = h$ represent the systems dynamics and linear observable, respectively; and $\mathbf{\Gamma}$ and **J**

are fixed matrices. Assume (3) represents the noise covariances.

Given the posterior estimate of the state $\mathbf{x}^a_{i-1}$ and covariance $\mathbf{P}^a_{i-1}$ at step $i-1$, the Kalman update for a linear system (Simon 2006; Bélanger 1974) is

$$\begin{aligned}
\mathbf{x}^b_i &= \mathbf{F}\mathbf{x}^a_{i-1}\\
\mathbf{y}^b_i &= \mathbf{H}\mathbf{x}^b_i\\
\mathbf{P}^b_i &= \mathbf{F}\mathbf{P}^a_{i-1}\mathbf{F}^{\mathrm{T}} + \boldsymbol{\Gamma}\mathbf{Q}\boldsymbol{\Gamma}^{\mathrm{T}}, \qquad (13)\\
\mathbf{K}_i &= (\mathbf{P}^b_i\mathbf{H}^{\mathrm{T}} + \boldsymbol{\Gamma}\mathbf{S}\mathbf{J}^{\mathrm{T}})(\mathbf{H}\mathbf{P}^b_i\mathbf{H}^{\mathrm{T}} + \mathbf{H}\boldsymbol{\Gamma}\mathbf{S}\mathbf{J}^{\mathrm{T}}\\
&\quad + \mathbf{J}\mathbf{S}^{\mathrm{T}}\boldsymbol{\Gamma}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}} + \mathbf{J}\mathbf{R}\mathbf{J}^{\mathrm{T}})^{-1}, \qquad (14)\\
\mathbf{x}^a_i &= \mathbf{x}^b_i + \mathbf{K}_i(\mathbf{y}_i - \mathbf{y}^b_i)\\
\mathbf{P}^a_i &= \mathbf{P}^b_i - \mathbf{K}_i(\mathbf{H}\mathbf{P}^b_i + \mathbf{J}\mathbf{S}^{\mathrm{T}}\boldsymbol{\Gamma}^{\mathrm{T}}), \qquad (15)
\end{aligned}$$

where $\mathbf{x}^b_i$ represents the forecast of the state given only the observations up to time $i-1$ and $\mathbf{P}^b_i$ represents the covariance of the forecast. Similarly, $\mathbf{y}^b_i$ represents the forecast of the ith observation given only the observations up to time $i-1$, and the difference between the observed variables and the forecast mapped into observation space,

$$\boldsymbol{\varepsilon}_i \equiv \mathbf{y}_i - \mathbf{y}^b_i\,,$$

is called the *innovation*. These innovations are often used to estimate the system and observation error as in Bélanger (1974), Mehra (1970, 1972), and Berry and Sauer (2013). The Kalman gain matrix $\mathbf{K}_i$ optimally combines the forecast $\mathbf{x}^b_i$ with the innovation to form the posterior estimate $\mathbf{x}^a_i$, which is the maximal likelihood and minimum variance estimator of the true state $\mathbf{x}_i$. The filter also produces the covariance matrix $\mathbf{P}^a_i$ of the estimator for use in the next filter step.

b. *The correlated unscented Kalman filter (CUKF)*

We now generalize the correlated system and observation noise filtering approach to nonlinear systems and we will show that for linear systems we recover exactly the equations above.

To apply the unscented Kalman filter we need to generate an unscented ensemble with the correct correlations. Since the noise realization is independent of the current state, we consider the concatenated state and noise vector as follows:

$$\begin{aligned}
\begin{bmatrix}\mathbf{x}_{i-1}\\ \boldsymbol{\omega}_{i-1}\\ \boldsymbol{\nu}_i\end{bmatrix} &\sim \mathcal{N}(\tilde{\mathbf{x}}^a_{i-1}, \tilde{\mathbf{P}}^a_{i-1})\\
&\equiv \mathcal{N}\left(\begin{bmatrix}\mathbf{x}^a_{i-1}\\ \mathbf{0}\\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\mathbf{P}^a_{i-1} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{Q} & \mathbf{S}\\ \mathbf{0} & \mathbf{S}^{\mathrm{T}} & \mathbf{R}\end{bmatrix}\right).
\end{aligned}$$

Notice that the concatenated state is $2N+M$ dimensional, and the joint covariance matrix is $(2N+M)\times(2N+M)$.

We then form the unscented ensemble, which is represented in a $(2N+M)\times(4N+2M+1)$ matrix

$$\begin{bmatrix}\mathbf{X}^a_{i-1}\\ \mathbf{W}_{i-1}\\ \mathbf{V}_i\end{bmatrix} \equiv \left(\tilde{\mathbf{x}}^a_{i-1}, \tilde{\mathbf{x}}^a_{i-1} + \sqrt{\alpha\tilde{\mathbf{P}}^a_{i-1}}, \tilde{\mathbf{x}}^a_{i-1} - \sqrt{\alpha\tilde{\mathbf{P}}^a_{i-1}}\right),$$

where $\mathbf{X}^a_{i-1}$ contains the first N rows of the ensemble, $\mathbf{W}_{i-1}$ contains the next N rows, and $\mathbf{V}_{i-1}$ contains the final M rows. We also define the associated ensemble weights as

$$\mathbf{w}_i = \begin{cases} 1 - \dfrac{N}{\alpha} & i = 1\\[2ex] \dfrac{1}{2\alpha} & i \neq 1\end{cases}$$

for $i = 2, \ldots, 2N+1$. The scalar α defines the scaling of the ensemble, which is often chosen to be $\alpha = N$ or $\alpha = N \pm 1$ although Julier and Uhlmann (2004) suggests $\alpha = 3$ [in the limit as $\alpha \to 0$ the unscented Kalman filter (UKF) approaches the extended Kalman filter (EKF)]. Notice that if the matrix $\mathbf{C}$ is constant, the square root of $\mathbf{C}$ can be computed offline and then $\sqrt{\tilde{\mathbf{P}}^a_{i-1}}$ can be formed at each step as the block diagonal matrix with blocks $\sqrt{\mathbf{P}^a_{i-1}}$ and $\sqrt{\mathbf{C}}$.

Now that we have generated an unscented ensemble with the correct correlations we can pass this ensemble through the nonlinear transformations defining

$$\begin{aligned}
\mathbf{X}^b_i &= f(\mathbf{X}^a_{i-1}, \mathbf{W}_{i-1})\,,\\
\mathbf{Y}^b_i &= h(\mathbf{X}^b_i, \mathbf{V}_i)\,,
\end{aligned}$$

where the nonlinear functions f and h are applied to each column of the ensemble matrices to form the forecast ensemble matrices $\mathbf{X}^b_i$ and $\mathbf{Y}^b_i$ that are $N\times(4N+2M+1)$ and $M\times(4N+2M+1)$, respectively. Now we can compute the following forecast statistics:

$$\begin{aligned}
\mathbf{x}^b_i &= \mathbf{X}^b_i\mathbf{w}\\
\mathbf{y}^b_i &= \mathbf{Y}^b_i\mathbf{w}\\
\mathbf{P}^x_i &= \sum_{j=2}^{2N+1} w_j[(\mathbf{X}^b_i)_{\cdot j} - \mathbf{x}^b_i][(\mathbf{X}^b_i)_{\cdot j} - \mathbf{x}^b_i]^{\mathrm{T}}\\
\mathbf{P}^y_i &= \sum_{j=2}^{2N+1} w_j[(\mathbf{Y}^b_i)_{\cdot j} - \mathbf{y}^b_i][(\mathbf{Y}^b_i)_{\cdot j} - \mathbf{y}^b_i]^{\mathrm{T}}\\
\mathbf{P}^{xy}_i &= \sum_{j=2}^{2N+1} w_j[(\mathbf{X}^b_i)_{\cdot j} - \mathbf{x}^b_i][(\mathbf{Y}^b_i)_{\cdot j} - \mathbf{y}^b_i]^{\mathrm{T}},
\end{aligned}$$

where $(\mathbf{X}^b_i)_{\cdot j}$ is the jth column of $\mathbf{X}^b_i$ (the jth ensemble member).

We can now define the unscented version of the Kalman update for correlated noise as follows:

$$\mathbf{K}_i = \mathbf{P}_i^{xy}(\mathbf{P}_i^{y})^{-1}$$
$$\mathbf{x}_i^a = \mathbf{x}_i^b + \mathbf{K}_i(\mathbf{y}_i - \mathbf{y}_i^b)$$
$$\mathbf{P}_i^a = \mathbf{P}_i^b - \mathbf{K}_i(\mathbf{P}_i^{xy})^{\mathrm{T}}. \quad (16)$$

Finally, as an implementation detail, the last equation should be computed as

$$\mathbf{P}_i^a = \mathbf{P}_i^b - \mathbf{K}_i\mathbf{P}_i^{y}\mathbf{K}_i^{\mathrm{T}}$$

in order to maintain numerical symmetry.

In appendix A we show the equivalence of the CUKF and Kalman filter (KF) for linear problems with correlated noise, which shows that the CUKF is a natural generalization to nonlinear problems with correlated errors. We note that the generalization of the CUKF approach to an ensemble square root Kalman filter (EnSQKF) is a straightforward extension of the same Kalman update formulas. Integrating correlated noise into other Kalman filters such as the EnKF and ensemble transform Kalman filter (ETKF) can also be achieved using the Kalman update for correlated noise. For large problems the covariance matrix would need to be localized in order to be a practical method; for example, the localized ensemble transform Kalman filter (LETKF) can be adapted to use the unscented ensembles used here Berry and Sauer (2013). A significant remaining task is generalizing the ensemble adjustment Kalman filter EAKF for additive system noise and correlated system and observation noise. Serial filters such as the EAKF cannot currently be applied even for additive system noise, which is not correlated to the observation noise, and instead these filters typically use inflation to try to account for system error. Generalizing the serial filtering approach to allow these more general types of inflation is a significant and important task and is beyond the scope of this article.

4. Filtering systems with truncation errors

In this section we apply the CUKF to truncated observations of the Lorenz-63 and Kuramoto–Sivashinsky systems as described in section 2. The dynamics and observations considered in this section have no added noise, so that the system errors arise only from truncation of the numerical solvers, and the observation errors arises only from local integration (representation error only). The CUKF will use the empirically estimated **C** matrices described in section 2, and we compare these to the filter results with the covariance matrix **C** modified by setting the **S** block equal to the **0** matrix, which we denote as UKF.

a. Example: Lorenz equations

First we consider the Lorenz-63 system in (8) with the observation described in section 2b. Using the same data generated in that example, we applied the CUKF and UKF. The estimates produced by these filters are shown in Fig. 4a (for the same time interval shown in Fig. 1). In Fig. 4b we show the errors between each filter's estimates and the truth, compared to the observation representation errors over the same time interval. The CUKF, which uses the full **C** matrix, obtains significantly superior estimates of the true state. Averaged over 6000 filter steps (after removing the initial filter transient) the root-mean-squared error (RMSE) of standard UKF estimates is 0.29 whereas that of the CUKF estimates is 0.16. Compared to the RMSE of the raw observations, which is 0.35, the UKF reduced the error by 17%, while the CUKF reduced the error by 54%.

We then repeated this experiment using the RK4 integrator instead of forward Euler with the same truncated time step of $\Delta t = 0.05$ and the results are shown in Figs. 4c and 4d. The RMSE of the UKF estimates with this integrator is 0.18 and the RMSE of the CUKF estimates is 0.16. Recall that the local truncation errors of RK4 were negatively correlated with the observation representation errors, resulting in relatively small differences between the UKF and CUKF for this example. The difference between positively and negatively correlated errors will be studied below in section 5. Notice that the CUKF with forward Euler obtains better estimates than the UKF using the far superior RK4 integrator. This shows that by using the correlations we can obtain better results with a much faster integrator. It also emphasizes the importance of the sign of the correlations, so that if possible one should select an integrator that yields errors that are positively correlated with observation representation errors (in the global average), possibly even if this requires using a lower-order method.

b. Example: Kuramoto–Sivashinsky

Next we consider filtering the observations of the Kuramoto–Sivashinsky model in (9) introduced in section 2c. Using a ground truth integrated with 512 spatial grid points and $\Delta t = 10^{-4}$ we consider truncated models with 64, 128, and 256 grid points and $\Delta t = 0.05$ (results were similar for $\Delta t = 0.1$ and $\Delta t = 0.025$). In Figs. 5a and 5b we compare the RMSE of the UKF estimates, CUKF estimates, and the observations for two different spatial integration widths $\delta = L/512$ and $\delta = L/128$. When $\delta = L/512$ the integral in (10) defining the true observation is estimated using the composite trapezoid rule on 3 grid points of the full 512 grid point solution and the

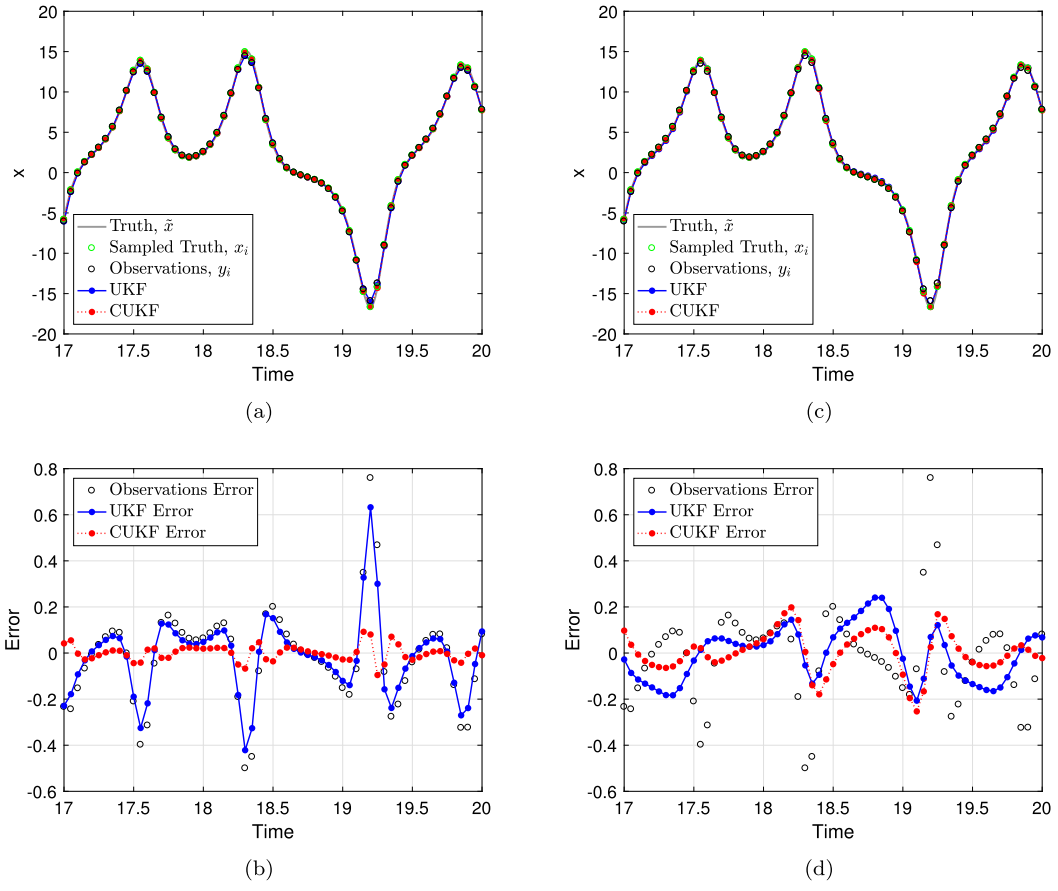

FIG. 4. (a) Comparison of the true solution, $\tilde{\mathbf{x}}$ (gray) and its discrete time samples $\mathbf{x}_i$ (black circles) and integrated observations $\mathbf{y}_i$ (green circles) with the UKF estimates (blue, solid) and CUKF estimates (red, dotted) over the same time interval shown in Fig. 1. (b) Errors computed by subtracting the true discretized signal from the observation (green circles), the UKF estimates (blue, solid), and the CUKF estimate (red, dotted). (c),(d) As in (a),(b), but using the RK4 integrator with the same $\Delta t = 0.05$.

RMSE of the observation representation errors is 0.01, which is 0.3% of the signal variance. In Fig. 5a we see that for 64 grid points the UKF does not reduce the error much relative to the observation representation error, whereas the CUKF obtains a much better estimate. In fact, the CUKF error with 64 grid points is comparable to the UKF error with 256 grid points. When $\delta = L/128$ the integral in (10) is estimated using the composite trapezoid rule on 9 grid points of the full 512 grid point solution and the RMSE of the observation representation errors is 0.11, which is 3.5% of the signal variance. As shown in Fig. 5b, the CUKF still outperforms the UKF; however, the difference at 64 grid points is less significant since $\delta < \Delta z$ meaning that each integral is completely contained between grid points.

The results of both UKF are robust for large Δt until the numerical solver becomes extremely unstable, which occurred for $\Delta t = 0.1$ with 256 grid points, since more grid points generally require smaller Δt to stabilize the solver. However, we should note that the numerical solver is unstable even for 64 grid points with $\Delta t = 0.1$ and the filter is stabilizing the solver using the observations. We also examined the robustness of these results in the presence of additive Gaussian observation noise with covariance $\mathbf{R}$ in Fig. 5c. Notice that for small $\mathbf{R}$ the random noise is small and the errors from truncation dominate, meaning the correlations in $\mathbf{S}$ are significant. As the uncorrelated noise $\mathbf{R}$ is increased it eventually dominates the correlated part of the observation errors, so that UKF and CUKF have similar performance.

Finally, in Fig. 5d we show the effect of inflation in the UKF by adding a constant multiple of the identity to either $\mathbf{Q}$ or $\mathbf{R}$, and in each case the best performance is found when using no inflation. We also tried inflating the filter background covariance matrix by multiplying by a constant greater than one, and this also had very little effect as shown in Fig. 5d. These results indicate that

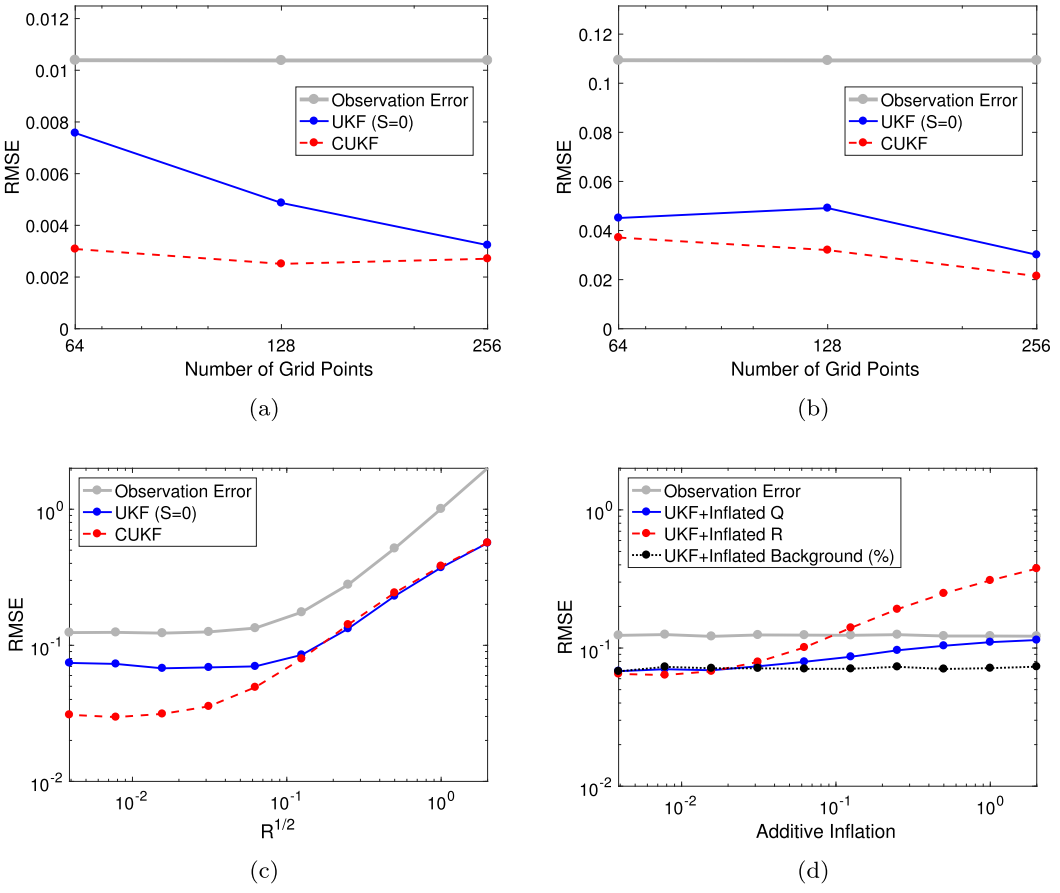

FIG. 5. (a),(b) Comparison of filter results using the UKF without correlations to filtering with correlations (CUKF) on the Kuramoto–Sivashinsky model truncated in space to 64, 128, and 256 grid points for observations integrated over (a) $\delta = L/512$ and (b) $\delta = L/128$. (c) For 64 grid points and $\Delta t = 0.1$, we show the robustness of the results after adding various levels of Gaussian instrument noise with variance $\mathbf{R}$ to the observations. (d) For the case $\mathbf{R} = 10^{-2}$ we test the UKF with inflation by adding the identity matrix times a constant to $\mathbf{Q}$ (blue, solid) or $\mathbf{R}$ (red, dashed). We also show the effect of inflating the filter background covariance (black, dotted) where the x axis indicates inflation percentage. In each case, inflation degraded the filter performance.

inflation cannot account for the correlated error. Since the $\mathbf{Q}$ and $\mathbf{R}$ used in the UKF were determined empirically to be optimal in this example, the only way to improve the performance is to account for correlations using the CUKF.

5. Maximally correlated random variables and perfect recoverability

In the previous section, the importance of using the full correlation matrix $\mathbf{C}$ was demonstrated, for system and observation errors that arise naturally from truncation and averaging that is common in geophysical modeling and filtering. In this section, we investigate the effects of cross correlation in a more systematic way. In particular, we identify the extreme case of *maximally correlated* random variables.

a. Maximum correlation

We begin by defining maximally correlated random variables.

DEFINITION 5.1 (MAXIMALLY CORRELATED RANDOM VARIABLES)

Let $\mathbf{X} \in \mathbb{R}^N$ and $\mathbf{Y} \in \mathbb{R}^M$ be random variables with covariances $\mathbf{Q} = \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^{\mathrm{T}}]$ and $\mathbf{R} = \mathbb{E}[(\mathbf{Y} - \mathbb{E}[\mathbf{Y}])(\mathbf{Y} - \mathbb{E}[\mathbf{Y}])^{\mathrm{T}}]$, respectively, and let $\mathbf{S} = \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{Y} - \mathbb{E}[\mathbf{Y}])^{\mathrm{T}}]$ be the cross covariance. We say that $\mathbf{X}$, $\mathbf{Y}$ are *maximally correlated* if the Schur complement of $\mathbf{R}$ in $\mathbf{C}$, namely $\mathbf{Q} - \mathbf{S}\mathbf{R}^{-1}\mathbf{S}^{\mathrm{T}}$, has minimal trace among all $N \times M$ matrices $\mathbf{S}$. In other words $\mathrm{trace}(\mathbf{Q} - \mathbf{S}\mathbf{R}^{-1}\mathbf{S}^{\mathrm{T}}) = \min_C\{\mathrm{trace}(\mathbf{Q} - \mathbf{C}\mathbf{R}^{-1}\mathbf{C}^{\mathrm{T}})\}$.

While it is not immediately obvious from the definition, Lemma B.1 in appendix B shows that the roles of

$\mathbf{X}$ and $\mathbf{Y}$ are symmetric, so that $\mathbf{S}$ also minimizes $\text{trace}(\mathbf{R} - \mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{S})$. The idea of maximally correlated random variables is that by choosing an appropriate $\mathbf{S}$ the $(N+M)\times(N+M)$ matrix $\mathbf{C}$ becomes rank deficient with rank N. Notice that we can make a linear change of the $\mathbf{X}$ variables,

$$\begin{bmatrix} \tilde{\mathbf{X}} \\ \tilde{\mathbf{Y}} \end{bmatrix} = \begin{bmatrix} \mathbf{I}_{N\times N} & -\mathbf{S}\mathbf{R}^{-1} \\ \mathbf{0} & \mathbf{I}_{M\times M} \end{bmatrix} \begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix},$$

so that $\tilde{\mathbf{Y}} = \mathbf{Y}$ is unchanged but the covariance matrix of $\tilde{\mathbf{X}}$, $\tilde{\mathbf{Y}}$ is

$$\begin{aligned} \tilde{\mathbf{C}} &\equiv \mathbb{E}\left[\begin{bmatrix} \tilde{\mathbf{X}} \\ \tilde{\mathbf{Y}} \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{X}} \\ \tilde{\mathbf{Y}} \end{bmatrix}^{\mathrm{T}} \right] \\ &= \begin{bmatrix} \mathbf{Q} - \mathbf{S}\mathbf{R}^{-1}\mathbf{S}^{\mathrm{T}} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{bmatrix}, \end{aligned} \tag{17}$$

and the new state variables $\tilde{\mathbf{X}}$ have covariance matrix $\mathbf{Q} - \mathbf{S}\mathbf{R}^{-1}\mathbf{S}^{\mathrm{T}}$ with minimal trace. In other words, the variables $\tilde{\mathbf{X}}$ have minimal variance among all possible choices of $\mathbf{S}$. According to the rank additivity formula of Guttman (1946), the rank of $\mathbf{C}$ is equal to the sum of the rank of $\mathbf{R}$ and the rank of its Schur complement $\mathbf{Q} - \mathbf{S}\mathbf{R}^{-1}\mathbf{S}^{\mathrm{T}}$, meaning that $\text{rank}(\mathbf{C}) = \text{rank}(\tilde{\mathbf{C}})$. Thus, by reducing the rank of the Schur complement we are actually choosing $\mathbf{S}$, which minimizes the rank of $\mathbf{C}$. Intuitively speaking, this choice of $\mathbf{S}$ minimizes the dimensionality of the joint noise process.

A simple example of maximal correlation is to consider the 2×2 case where $\mathbf{Q} = q$, $\mathbf{R} = r$, and $\mathbf{S} = s$ are scalars. By setting $s = \pm\sqrt{q}\sqrt{r}$, we find the Schur complement to be $q - sr^{-1}s = 0$. Moreover, with this choice of s the eigenvalues of $\mathbf{C}$ are $\{0, q + r\}$, so that $\text{rank}(\mathbf{W}) = 1$ is minimal over all possible choices of s. In general, when $N = M$ we can set $\mathbf{S} = \sqrt{\mathbf{Q}}\sqrt{\mathbf{R}}^{\mathrm{T}}$, where $\sqrt{\mathbf{Q}}\sqrt{\mathbf{Q}}^{\mathrm{T}} = \mathbf{Q}$ and $\sqrt{\mathbf{R}}\sqrt{\mathbf{R}}^{\mathrm{T}} = \mathbf{R}$ are matrix square roots (recall that matrix square roots are unique up to a choice of orthogonal matrix) and we find the Schur complement to be $\mathbf{Q} - \mathbf{S}\mathbf{R}^{-1}\mathbf{S}^{\mathrm{T}} = \mathbf{0}_{N\times N}$. When $N \neq M$ the formula for $\mathbf{S}$ is similar and is given in Lemma B.1, which is stated and proved in appendix B.

It follows from Lemma B.1 that given random variables $\mathbf{X} \in \mathbb{R}^N$ and $\mathbf{Y} \in \mathbb{R}^M$ with covariance matrices $\mathbf{Q}$, $\mathbf{R}$, respectively, there always exists an $N \times M$ matrix $\mathbf{S}$ such that the total covariance matrix $\mathbf{C}$ in (3) makes $\mathbf{X}$ and $\mathbf{Y}$ maximally correlated, and $\text{rank}(\mathbf{W}) = N$. This finding is striking, in the sense that if $\mathbf{X}$ and $\mathbf{Y}$ represent the system and observation errors of a dynamical system, respectively, and if they are maximally correlated, then the underlying noise/error process is actually only N dimensional, despite appearing $N + M$ dimensional. Since the observation errors are linear combinations of the system errors, up to an orthogonal transformation we can think of the process as *effectively having no observation noise.* We will make this rigorous below by showing that in the case of maximal correlations, the observation errors can be completely eliminated by filtering and the true state can be perfectly recovered.

Now consider the case when $\mathbf{Q}$ and $\mathbf{R}$ are the system and observation covariances, respectively. From the previous lemma we can see that the easiest way to obtain maximally correlated processes is when the observation errors are linear combinations of the system errors (since this implies that $\mathbf{C}$ has rank N). So, returning to the discussion in section 2a, we can now see that when the leading-order terms in the system and observation errors only differ by a constant multiple they will be maximally correlated up to higher-order terms. More generally, whenever the $N \times M$ matrix $\mathbf{C}$ has rank N, the system and observation errors are maximally correlated. In particular, the small eigenvalues in Fig. 3 indicate that the system and observation errors are close to maximally correlated.

b. *Perfect recoverability in maximally correlated linear systems*

Consider the linear system of the form (1) and (2), where

$$\begin{aligned} \mathbf{x}_{i+1} &= f(\mathbf{x}_i, \boldsymbol{\omega}_i) = \mathbf{F}\mathbf{x}_i + \boldsymbol{\Gamma}\boldsymbol{\omega}_i \\ \mathbf{y}_{i+1} &= h(\mathbf{x}_{i+1}, \boldsymbol{\nu}_{i+1}) = \mathbf{H}\mathbf{x}_{i+1} + \mathbf{J}\boldsymbol{\nu}_{i+1}, \end{aligned}$$

and assume the noise is generated by

$$\begin{bmatrix} \boldsymbol{\omega}_i \\ \boldsymbol{\nu}_{i+1} \end{bmatrix} \sim \mathcal{N}(\mathbf{0}, \mathbf{C}), \quad \text{where} \quad \mathbf{C} = \begin{bmatrix} \mathbf{Q} & \mathbf{S} \\ \mathbf{S}^{\mathrm{T}} & \mathbf{R} \end{bmatrix}$$

as in (3). In this section we will show that when the covariance matrices $\mathbf{Q}$ and $\mathbf{R}$ are maximally correlated, meaning that $\mathbf{S}$ is chosen as in Lemma B.1, the state variables become perfectly recoverable, meaning that the limiting variance of the Kalman filter estimates of those variables is zero. Of course, in real applications we do not get to choose $\mathbf{S}$. Our purpose here is to demonstrate the maximal effect that $\mathbf{S}$ can have on the ability to estimate random variables. As a consequence, if the true $\mathbf{S}$ were maximal and one instead used a suboptimal filter with $\mathbf{S} = 0$, the relative loss of accuracy would be ''infinite'' (since perfect reconstruction was possible with the true $\mathbf{S}$). Although the results in this section only apply to linear filtering problems, in section 5c we will show similar empirical results for nonlinear filtering problems.

We show the effect that the maximal correlation $\mathbf{S}$ has on the stationary posterior covariance $\mathbf{P}$ of a Kalman filter. Without loss of generality, in this section we will assume $\mathbf{\Gamma} = \mathbf{I}_{N\times N}$ and $\mathbf{J} = \mathbf{I}_{M\times M}$ since we may replace $\mathbf{C}$ by $\tilde{\mathbf{C}} = \mathbf{BCB}^{\mathrm{T}}$ where $\mathbf{B}$ is block diagonal with blocks $\mathbf{\Gamma}$ and $\mathbf{J}$. Substituting (13) and (14) into (15) and setting $\mathbf{P}_i^a = \mathbf{P}_{i-1}^a = \mathbf{P}$, we find the discrete time algebraic Riccati equation (DARE):

$$\mathbf{P} = \mathbf{FPF}^{\mathrm{T}} + Q - (\mathbf{FPF}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}} + \mathbf{QH}^{\mathrm{T}} + \mathbf{S})(\mathbf{HFPF}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}} + \mathbf{HQH}^{\mathrm{T}} + \mathbf{HS} + \mathbf{S}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}} + \mathbf{R})^{-1}(\mathbf{HFPF}^{\mathrm{T}} + \mathbf{HQ} + \mathbf{S}^{\mathrm{T}}). \tag{18}$$

If (18) has a solution that is stabilizing, meaning that all the eigenvalues of $(\mathbf{I} - \mathbf{KH})\mathbf{F}$ are inside the unit circle [where $\mathbf{K} = \mathbf{K}_\infty$ is defined by (14) using the solution $\mathbf{P}$], then this solution is unique and is the limiting covariance matrix of the Kalman filter as shown in Ran and Vreugdenhil (1988). We can now state the following result, the proof can be found in appendix C.

THEOREM 5.2

Assume that all the eigenvalues of the matrix

$$[\mathbf{I} - \mathbf{HS}(\mathbf{HS} + \mathbf{R})^{-1}]\mathbf{F}$$

lie inside the unit circle. Then the limiting covariance matrix $\mathbf{P}$ of a Kalman filtering problem with maximally correlated noise processes is zero when $M \geq N$. In other words, all state variables are perfectly recoverable. When $M < N$, if the general stability condition on $(\mathbf{I} - \mathbf{KH})\mathbf{F}$ is met, the limiting covariance matrix $\mathbf{P}$ is zero when projected onto the top M eigenvectors of $\mathbf{Q}$.

Notice that the Kalman filter has an asymmetry between $\mathbf{Q}$ and $\mathbf{R}$, which is not present in the definition of maximal correlation, because of their differing roles in the dynamics. The consequence of this asymmetry is seen in the stability condition. For simplicity, consider the case $N = M = 1$, when $\mathbf{HS}$ is positive the stability condition is met for all $|\mathbf{F}| < 1$, and even for some $|\mathbf{F}| > 1$ since $|(1 - \mathbf{HS}(\mathbf{HS} + \mathbf{R})^{-1})| < 1$. Conversely, when $\mathbf{HS}$ is negative, we find $|[1 - \mathbf{HS}(\mathbf{HS} + \mathbf{R})^{-1}]| > 1$ and stability of $\mathbf{F}$ is no longer sufficient.

To demonstrate this result, we applied the numerical DARE solver implemented by MATLAB to a linear system with $\mathbf{F} = \lambda\mathbf{I}_{N\times N}$, where λ will be varied to demonstrate the effect of stability and $\mathbf{H} = \mathbf{I}_{N\times N}$, $\mathbf{Q} = \mathbf{I}_{N\times N}$, and $\mathbf{R} = 4\mathbf{I}_{N\times N}$. To show how the filter estimates improve as $\mathbf{S}$ approaches the maximal choice, we let $\mathbf{S} = (2 - \delta_j)\mathbf{I}_{N\times N}$ for $\delta_j = 2^{-j}$ and $j = 0, 1, \ldots, 18$. In this case, the correlation in $\mathbf{S}$ is positive and we find that the stabilization criterion is

$$\mathrm{eig}([\mathbf{I} - \mathbf{HS}(\mathbf{HS} + \mathbf{R})^{-1}]\mathbf{F}) = [1 - 2/(2 + 4)]\lambda < 1$$

if and only if $\lambda < 1.5$. In Fig. 6a we plot the mean of the diagonal elements of the numerical solution $\mathbf{P}$ to (18) against the trace of the Schur complement $\mathrm{trace}(\mathbf{Q} - \mathbf{SR}^{-1}\mathbf{S}^{\mathrm{T}}) = 4N\delta_j - N\delta_j^2$ for all values of j. The different curves correspond to different values of λ chosen near 1.5. Notice that for $\lambda < 1.5$, as the noise approached maximally correlated, the filter estimates approach the true state up to the limits of numerical precision. When $\lambda > 1.5$, $\mathbf{P} = 0$ is still a solution of the DARE but is no longer stabilizing and so the filter converges to a covariance matrix that has variances greater than zero.

To show the effect of negative correlation, we next consider the case $\mathrm{S} = -(2 - \delta_j)\mathrm{I}_{N\times N}$ for $\delta_j = 2^{-j}$ and $j = 0, 1, \ldots, 18$. In this case, the correlation in $\mathbf{S}$ is negative and we find that the stabilization criterion is

$$\mathrm{eig}([\mathbf{I} - \mathbf{HS}(\mathbf{HS} + \mathbf{R})^{-1}]\mathbf{F}) = [1 + 2/(-2 + 4)]\lambda < 1$$

if and only if $\lambda < 0.5$. Notice that in this case the dynamics are required to be stable ($\lambda < 1$) in order to stabilize the $\mathbf{P} = 0$ solution, whereas with positive correlations the dynamics could be unstable ($\lambda > 1$). In Fig. 6b we plot the mean of the diagonal elements of the numerical solution $\mathbf{P}$ to (18) against the trace of the Schur complement $\mathrm{trace}(\mathbf{Q} - \mathbf{SR}^{-1}\mathbf{S}^{\mathrm{T}}) = 4N\delta_j - N\delta_j^2$ for all values of j. The different curves correspond to different values of λ chosen near 0.5. Notice that for $\lambda < 0.5$, as the noise approached maximally correlated, the filter estimates approach the true state up to the limits of numerical precision. When $\lambda > 0.5$, $\mathbf{P} = 0$ is still a solution of the DARE but is no longer stabilizing and so the filter converges to a covariance matrix with variances greater than zero.

Finally, we note that a standard form for the DARE used in numerical solvers, such as MATLAB, is

$$\mathbf{0} = \mathbf{A}^{\mathrm{T}}\mathbf{PA} - \mathbf{E}^{\mathrm{T}}\mathbf{PE} - (\mathbf{A}^{\mathrm{T}}\mathbf{PB} + \hat{\mathbf{S}})(\mathbf{B}^{\mathrm{T}}\mathbf{PB} + \hat{\mathbf{R}})^{-1} \times (\mathbf{B}^{\mathrm{T}}\mathbf{PA} + \hat{\mathbf{S}}^{\mathrm{T}}) + \hat{\mathbf{Q}},$$

and (18) can be put in this form by setting $\mathbf{E} = \mathbf{I}$, $\mathbf{A} = \mathbf{F}^{\mathrm{T}}$, $\mathbf{B} = \mathbf{F}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}$, $\hat{\mathbf{Q}} = \mathbf{Q}$, $\hat{\mathbf{S}} = \mathbf{QH}^{\mathrm{T}} + \mathbf{S}$, and $\hat{\mathbf{R}} = \mathbf{HQH}^{\mathrm{T}} + \mathbf{HS} + \mathbf{S}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}} + \mathbf{R}$.

c. *Examples of UKF and perfect recovery in nonlinear systems*

In this section we will apply the UKF to synthetic datasets generated with nonlinear dynamics where the system and observation errors are Gaussian distributed pseudorandom numbers. A surprising result is that

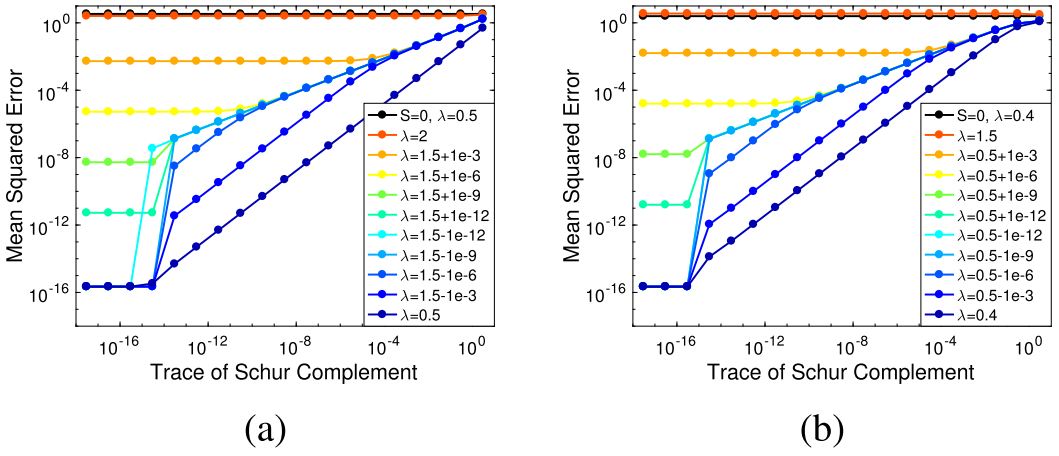

FIG. 6. Mean-squared error of filter estimates for linear models $\mathbf{F} = \lambda \mathbf{I}_{N\times N}$ with (a) positive correlations and (b) negative correlations. Black curve is based on the filter using $\mathbf{S} = 0$, all other curves use the true $\mathbf{S}$. Notice that we obtain perfect recovery to the limit of numerical precision when the stability criterion is satisfied: (a) $\lambda < 3/2$ for positive correlation and (b) $\lambda < 1/2$ for negative correlation.

despite the nonlinearity, we still obtain perfect recovery up to numerical precision for maximally correlated errors. Moreover, in analogy to the linear case, perfect recovery is not possible when the instabilities in the nonlinear dynamics become sufficiently strong.

We first consider the Lorenz-63 system introduced above (8). We consider the discrete time dynamics $\mathbf{x}_{i+1} = f(\mathbf{x}_i)$ to be given by applying the RK4 solver with $\Delta t = 0.01$ to the chaotic vector field in (8) and the direct observation function $h(\mathbf{x}_{i+1}) = \mathbf{x}_{i+1}$. We artificially add substantial system and observation noise of covariance $\mathbf{Q} = \mathbf{I}_{3\times 3}$ and $\mathbf{R} = 4\mathbf{I}_{3\times 3}$, respectively. According to the remarks preceding Lemma B.1, the system and observation noise are maximally correlated when $\mathbf{S} = 2\mathbf{I}_{3\times 3}$, which implies the Schur complement of $\mathbf{Q}$ and $\mathbf{R}$ is the zero matrix. To test the recovery of the deterministic variables x, y, z, we set $\mathbf{S} = (2 - \delta_j)\mathbf{I}_{3\times 3}$ for $\delta_j = 2^{-j}$ and $j = 0, 1, \ldots, 18$, implying that δ_j is the trace of the Schur complement. A time series of Lorenz-63 was produced with noise specified from $\mathbf{Q}$, $\mathbf{R}$, and $\mathbf{S}$ and the UKF algorithm of section 3a was applied. Results for RMSE of the recovered variables x, y, z are shown in Fig. 7a. In the limit as $\delta \to 0$ and $\mathbf{S}$ approaches maximal correlation, we find perfect recovery of the true state using the CUKF algorithm with the true covariance matrix $\mathbf{C}$, as foreshadowed by the linear case. We repeated this experiment for the alternative parameter value $\sigma = 350$ in (8), which yields a globally attracting periodic orbit, and obtained very similar results, also shown in Fig. 7a.

Since perfect recovery in the linear case depended on the degree of stability of the dynamics, we next investigate the effect of the Lyapunov exponents of a chaotic dynamical system on the ability to obtain perfect recovery. We consider the chaotic Lorenz-96 system, a 40-dimension ODE given by

$$\dot{\mathbf{x}}_i = \mathbf{x}_{i-1}\mathbf{x}_{i+1} - \mathbf{x}_{i-1}\mathbf{x}_{i-2} - \mathbf{x}_i + F, \tag{19}$$

where F determines the size and number of the positive Lyapunov exponents of the chaotic dynamics Lorenz (1996). Using $\mathbf{Q} = \mathbf{I}_{40\times 40}$, $\mathbf{R} = 4\mathbf{I}_{40\times 40}$, the maximal correlation occurs when $\mathbf{S} = 2\mathbf{I}_{40\times 40}$. We set $\mathbf{S} = (2 - \delta_j)\mathbf{I}_{40\times 40}$ for $\delta_j = 2^{-j}$ and $j = 0, 1, \ldots, 18$ and generated time series with noise from $\mathbf{Q}$, $\mathbf{R}$, and $\mathbf{S}$ as above. The CUKF algorithm was applied to recover the 40-dimensional state. In Fig. 7b we show that for $F = 6$ we obtain perfect recovery in the case of maximal correlation between system and observation noise. However, as F increases, the system becomes more strongly chaotic and the perfect recovery breaks down. Notice that as in the linear case, the failure of perfect recovery occurs very sharply between $F = 7$ and $F = 9$. This suggests that in analogy to the linear result, some form of stability condition is likely necessary for perfect recovery.

6. Discussion

Approximating a dynamical system on a grid is pervasive in geophysical data assimilation applications. For dynamical processes, time is usually handled in discrete fashion. We have shown that correlation between system and observation errors should be expected when the system errors derive from local truncation errors from differential equation solvers, both in discrete time and on a spatial grid, and when observational error is dominated by either observation model error or representation error.

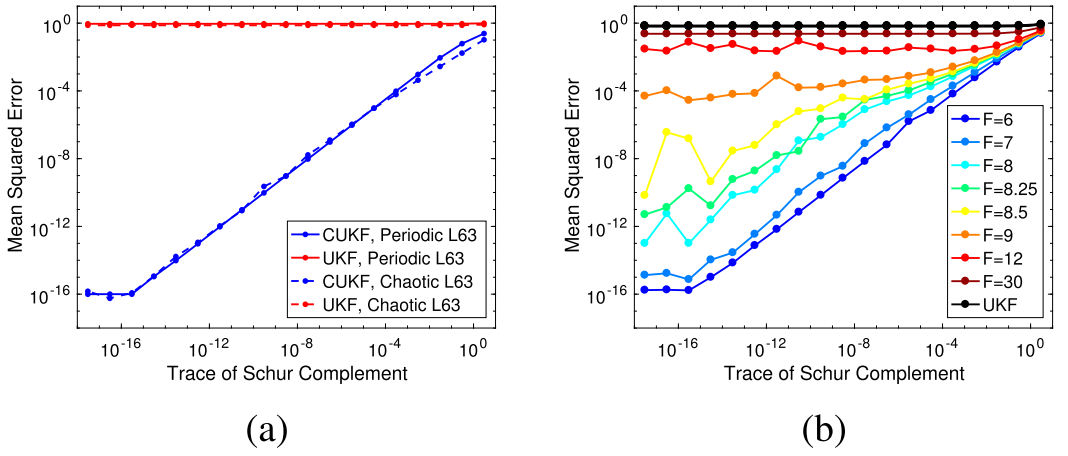

FIG. 7. Mean-squared error of filter estimates with positively correlated noise for (a) L63 in periodic and chaotic parameter regimes and (b) L96 dynamical systems for various values of the forcing parameter. Black curve is based on the filter using $\mathbf{S} = 0$ (UKF), and all other curves use the true $\mathbf{S}$ (CUKF).

In section 3, we introduced an approach to the ensemble Kalman filter that accounts for the correlations between system and observation errors. In particular, we showed that for spatiotemporal problems, extending the covariance matrix to allow cross correlations can reduce filtering error as much as a significant increase in grid resolution. Of course, obtaining more precise estimates of the truth with much coarser discretization allows faster runtimes and/or larger ensembles to be used.

Correlations are most significant when other independent sources of observation and system error are small compared to the truncation error. Of course, other sources of error, such as model error, may influence both the state and the observations leading to further significant correlations, but for simplicity we focus on correlations arising in the perfect model scenario. It is reasonable to expect that in many physical systems, the noise affecting the state of the system would also affect the sensor or observation system.

The generalization of the CUKF to an ensemble square root Kalman filter (EnSQKF) is a straightforward extension. However, it remains to extend the ensemble adjustment Kalman filter EAKF for additive system noise to correlations between system and observation noise. An EAKF formulation is critical for situations when the ensemble size is necessarily much smaller than either the state or observation dimensions (N and M, respectively). This situation is common when the covariance matrices, which are explicitly used in the UKF approach above do not fit in memory. A significant challenge in this formulation is that we cannot appropriately inflate the ensemble since we assume the full correlation matrix $\mathbf{C}$ is of maximum rank, and any inflation of the small ensemble would only match the inflation in the subspace spanned by the ensemble. A promising alternative is to follow the approach of Whitaker and Hamill (2002) and design an alternative gain matrix $\tilde{\mathbf{K}}_i$ such that the analysis ensemble $\tilde{\mathbf{X}}_i^a$ has the same covariance as applying the Kalman gain $\mathbf{K}$ to an appropriately inflated ensemble.

In this article, we have not dealt with the question of real-time estimation of the full covariance matrix $\mathbf{C}$. The importance of correctly specifying the $\mathbf{Q}$ and $\mathbf{R}$ matrices was first demonstrated for the Kalman filter in Mehra (1970, 1972) and for nonlinear filters in Berry and Sauer (2013). We consider sequential methods for estimation of the full covariance matrix in parallel with filtering to be a fruitful area of future research.

Acknowledgments. We thank three reviewers whose helpful suggestions led to a much improved paper. This research was partially supported by National Science Foundation Grant DMS1723175.

APPENDIX A

Equivalence of CUKF and KF for Linear Problems with Correlated Errors

To justify our definition of the CUKF in section 3b, we will show that for linear systems the update in (16) is equivalent to the Kalman filter equations given in section 3a. We can define the covariance of the forecast by expanding the innovation as

$$\begin{aligned}\boldsymbol{\varepsilon}_i &\equiv \mathbf{y}_i - \mathbf{y}_i^b = \mathbf{H}(\mathbf{x}_i - \mathbf{x}_i^b) + \mathbf{J}\boldsymbol{\nu}_i \\ &= \mathbf{HF}(\mathbf{x}_{i-1} - \mathbf{x}_{i-1}^a) + \mathbf{H}\boldsymbol{\Gamma}\boldsymbol{\omega}_{i-1} + \mathbf{J}\boldsymbol{\nu}_i, \end{aligned} \tag{A1}$$

and writing $\mathbf{H}\boldsymbol{\Gamma}\boldsymbol{\omega}_{i-1} + \mathbf{J}\boldsymbol{\nu}_i = [\mathbf{H}\boldsymbol{\Gamma}, \mathbf{J}]\begin{bmatrix}\boldsymbol{\omega}_{i-1}\\ \boldsymbol{\nu}_i\end{bmatrix}$ we find

$$\begin{aligned}\mathbf{P}_i^y &\equiv \mathbb{E}[\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i^{\mathrm{T}}]\\ &= \mathbf{H}\mathbf{F}\mathbf{P}_{i-1}^a\mathbf{F}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}} + [\mathbf{H}\boldsymbol{\Gamma}, \mathbf{J}]\mathbf{W}[\mathbf{H}\boldsymbol{\Gamma}, \mathbf{J}]^{\mathrm{T}}\\ &= \mathbf{H}\mathbf{F}\mathbf{P}_{i-1}^a\mathbf{F}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}} + \mathbf{H}\boldsymbol{\Gamma}\mathbf{Q}\boldsymbol{\Gamma}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}} + \mathbf{H}\boldsymbol{\Gamma}\mathbf{S}\mathbf{J}^{\mathrm{T}}\\ &\quad + \mathbf{J}\mathbf{S}^{\mathrm{T}}\boldsymbol{\Gamma}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}} + \mathbf{J}\mathbf{R}\mathbf{J}^{\mathrm{T}},\end{aligned}$$

where we recall that $\mathbb{E}\left[\begin{bmatrix}\boldsymbol{\omega}_{i-1}\\ \boldsymbol{\nu}_i\end{bmatrix}\begin{bmatrix}\boldsymbol{\omega}_{i-1}\\ \boldsymbol{\nu}_i\end{bmatrix}^{\mathrm{T}}\right] = \mathbf{W}$ and $\mathbb{E}[(\mathbf{x}_{i-1} - \mathbf{x}_{i-1}^a)(\mathbf{x}_{i-1} - \mathbf{x}_{i-1}^a)^{\mathrm{T}}] = \mathbf{P}_{i-1}^a$. Notice that

$$\mathbf{H}\mathbf{P}_i^b\mathbf{H}^{\mathrm{T}} = \mathbf{H}\mathbf{F}\mathbf{P}_{i-1}^a\mathbf{F}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}} + \mathbf{H}\boldsymbol{\Gamma}\mathbf{Q}\boldsymbol{\Gamma}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}},$$

so that

$$\begin{aligned}\mathbf{P}_i^y = \mathbf{H}\mathbf{P}_i^b\mathbf{H}^{\mathrm{T}} + \mathbf{H}\boldsymbol{\Gamma}\mathbf{S}\mathbf{J}^{\mathrm{T}} + \mathbf{H}\boldsymbol{\Gamma}\mathbf{Q}\boldsymbol{\Gamma}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}\\ + \mathbf{J}\mathbf{S}^{\mathrm{T}}\boldsymbol{\Gamma}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}} + \mathbf{J}\mathbf{R}\mathbf{J}^{\mathrm{T}},\end{aligned}$$

which implies that we can rewrite the Kalman gain equation as

$$\mathbf{K}_i = (\mathbf{P}_i^b\mathbf{H}^{\mathrm{T}} + \boldsymbol{\Gamma}\mathbf{S}\mathbf{J}^{\mathrm{T}})(\mathbf{P}_i^y)^{-1}.$$

Similarly, we can define the cross correlation between the state and observation as

$$\begin{aligned}\mathbf{P}_i^{xy} &\equiv \mathbb{E}[(\mathbf{x}_i - \mathbf{x}_i^b)(\mathbf{y}_i - \mathbf{y}_i^b)^{\mathrm{T}}]\\ &= \mathbb{E}[(\mathbf{F}(\mathbf{x}_{i-1} - \mathbf{x}_{i-1}^a) + \boldsymbol{\Gamma}\boldsymbol{\omega}_{i-1})(\mathbf{H}\mathbf{F}(\mathbf{x}_{i-1} - \mathbf{x}_{i-1}^a) + \mathbf{H}\boldsymbol{\Gamma}\boldsymbol{\omega}_{i-1} + \mathbf{J}\boldsymbol{\nu}_i)^{\mathrm{T}}]\\ &= \mathbf{F}\mathbf{P}_{i-1}^a\mathbf{F}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}} + \boldsymbol{\Gamma}\mathbf{Q}\boldsymbol{\Gamma}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}} + \boldsymbol{\Gamma}\mathbf{S}\mathbf{J}^{\mathrm{T}}\\ &= \mathbf{P}_i^b\mathbf{H}^{\mathrm{T}} + \boldsymbol{\Gamma}\mathbf{S}\mathbf{J}^{\mathrm{T}},\end{aligned}$$

and finally we can write the Kalman gain as

$$\mathbf{K}_i = \mathbf{P}_i^{xy}(\mathbf{P}_i^y)^{-1},$$

which agrees with the definition used in (16) for our version of the unscented Kalman filter.

APPENDIX B

Maximal Correlation when $N \neq M$

In section 5 we showed how to define the maximal correlation matrix $\mathbf{C}$ when $N = M$. In the following lemma we derive the formula when $N \neq M$.

Lemma B.1

Let $\mathbf{Q}$, $\mathbf{R}$ be symmetric positive-definite matrices with eigendecompositions $\mathbf{Q} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathrm{T}}$ and $\mathbf{R} = \mathbf{V}\mathbf{L}\mathbf{V}^{\mathrm{T}}$. Denote diagonal entries by $\Lambda_{1,1} \geq \Lambda_{2,2} \geq \cdots \geq \Lambda_{N,N}$ and $L_{1,1} \geq L_{2,2} \geq \cdots \geq L_{M,M}$.

1. If $N \geq M$, let $\tilde{\mathbf{U}}$ be the first M columns of $\mathbf{U}$ and $\tilde{\boldsymbol{\Lambda}}$ be the first $M \times M$ block of $\boldsymbol{\Lambda}$ and let $\tilde{\mathbf{V}} = \mathbf{V}$ and $\tilde{\mathbf{L}} = \mathbf{L}$.
2. If $N \leq M$, let $\tilde{\mathbf{U}} = \mathbf{U}$ and $\tilde{\boldsymbol{\Lambda}} = \boldsymbol{\Lambda}$ and let $\tilde{\mathbf{V}}$ be any N columns of $\mathbf{V}$ and $\tilde{\mathbf{L}}$ the corresponding $N \times N$ block of $\mathbf{L}$.

Then $\mathrm{trace}(\mathbf{Q} - \mathbf{S}\mathbf{R}^{-1}\mathbf{S}^{\mathrm{T}})$ is minimized over all $N \times M$ matrices $\mathbf{S}$ by

$$\mathbf{S} = \tilde{\mathbf{U}}\tilde{\boldsymbol{\Lambda}}^{1/2}\mathbf{G}\tilde{\mathbf{L}}^{1/2}\tilde{\mathbf{V}}^{\mathrm{T}}$$

for any orthogonal matrix $\mathbf{G}$. Moreover, for any maximal $\mathbf{S}$ we have $\mathrm{rank}(\mathbf{Q} - \mathbf{S}\mathbf{R}^{-1}\mathbf{S}^{\mathrm{T}}) = \max\{N - M, 0\}$ and $\mathrm{trace}(\mathbf{Q} - \mathbf{S}\mathbf{R}^{-1}\mathbf{S}^{\mathrm{T}}) = \sum_{i=M+1}^{N}\Lambda_{i,i}$.

Proof

Notice that

$$\begin{aligned}\mathbf{S}\mathbf{R}^{-1}\mathbf{S}^{\mathrm{T}} &= \tilde{\mathbf{U}}\tilde{\boldsymbol{\Lambda}}^{1/2}\mathbf{G}\tilde{\mathbf{L}}^{1/2}\tilde{\mathbf{V}}^{\mathrm{T}}\mathbf{V}\mathbf{L}^{-1}\mathbf{V}^{\mathrm{T}}\tilde{\mathbf{V}}\tilde{\mathbf{L}}^{1/2}\mathbf{G}^{\mathrm{T}}\tilde{\boldsymbol{\Lambda}}^{1/2}\tilde{\mathbf{U}}^{\mathrm{T}}\\ &= \tilde{\mathbf{U}}\tilde{\boldsymbol{\Lambda}}\tilde{\mathbf{U}}^{\mathrm{T}},\end{aligned} \tag{B1}$$

so that $\mathbf{Q} - \mathbf{S}\mathbf{R}^{-1}\mathbf{S} = \mathbf{U}\hat{\boldsymbol{\Lambda}}\mathbf{U}^{\mathrm{T}}$, where $\hat{\boldsymbol{\Lambda}}$ replaces the first M diagonal entries of $\boldsymbol{\Lambda}$ with zeros if $N > M$ and $\hat{\boldsymbol{\Lambda}} = 0$ if $N \leq M$.

APPENDIX C

Proof of Theorem 5.2

Proof

It suffices to show that $\mathbf{P} = 0$ is a solution to the DARE. Let $\mathbf{R}^{1/2}$ be the unique symmetric square root of $\mathbf{R}$ and let $\mathbf{R}^{-1/2}$ be its inverse. Setting $\mathbf{A} = \mathbf{H}\mathbf{S}\mathbf{R}^{-1/2} + \mathbf{R}^{1/2}$, notice that

$$\begin{aligned}\mathbf{I} &= \mathbf{A}^{\mathrm{T}}(\mathbf{A}\mathbf{A}^{\mathrm{T}})^{-1}\mathbf{A}\\ &= \mathbf{A}^{\mathrm{T}}(\mathbf{H}\mathbf{S}\mathbf{R}^{-1}\mathbf{S}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}} + \mathbf{H}\mathbf{S} + \mathbf{S}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}} + \mathbf{R})^{-1}\mathbf{A},\end{aligned}$$

and multiplying on the left by $\mathbf{S}\mathbf{R}^{-1/2}$ and on the right by $\mathbf{R}^{-1/2}\mathbf{S}^{\mathrm{T}}$ we have

$$\mathbf{SR}^{-1}\mathbf{S} = (\mathbf{SR}^{-1}\mathbf{S}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}} + \mathbf{S})(\mathbf{HSR}^{-1}\mathbf{S}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}} + \mathbf{R})^{-1} + (\mathbf{HS} + \mathbf{S}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}} + \mathbf{R}).$$

Finally, since $\mathbf{S}$ is maximal with $M \geq N$ we have $\mathbf{Q} = \mathbf{SR}^{-1}\mathbf{S}^{\mathrm{T}}$ by Lemma B.1, which implies that

$$\mathbf{Q} = (\mathbf{QH}^{\mathrm{T}} + \mathbf{S})(\mathbf{HQH}^{\mathrm{T}} + \mathbf{HS} + \mathbf{S}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}} + \mathbf{R})^{-1}(\mathbf{HQ} + \mathbf{S}^{\mathrm{T}}).$$

Since the previous equation is exactly like our (18) with $\mathbf{P} = 0$, this shows that $\mathbf{P} = 0$ is a solution to the DARE. Moreover, since $\mathbf{P} = 0$ implies that the limiting Kalman gain is

$$\begin{aligned}\mathbf{K} &= (\mathbf{QH}^{\mathrm{T}} + \mathbf{S})(\mathbf{HQH}^{\mathrm{T}} + \mathbf{HS} + \mathbf{S}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}} + \mathbf{R})^{-1}\\ &= \mathbf{SR}^{-1/2}\mathbf{A}^{\mathrm{T}}(\mathbf{AA}^{\mathrm{T}})^{-1}\\ &= \mathbf{SR}^{-1/2}\mathbf{A}^{-1}\\ &= \mathbf{SR}^{-1/2}(\mathbf{HSR}^{-1/2} + \mathbf{R}^{1/2})^{-1}\\ &= \mathbf{S}(\mathbf{HS} + \mathbf{R})^{-1},\end{aligned}$$

where $\mathbf{A}$ is invertible since it is the square root of an invertible matrix. Thus, the stabilizing condition is that the matrix

$$(\mathbf{I} - \mathbf{HK})\mathbf{F} = [\mathbf{I} - \mathbf{HS}(\mathbf{HS} + \mathbf{R})^{-1}]\mathbf{F}$$

has eigenvalues inside the unit circle. Since $\mathbf{P} = 0$ solves the DARE and is stabilizing, it is the limiting covariance matrix of the Kalman filtering problem.

In the case when $M < N$, recall that $\mathbf{Q} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathrm{T}}$ is the eigendecomposition from Lemma B.1 and $\tilde{\mathbf{U}}$ contains the first M columns of $\mathbf{U}$ so that $\mathbf{Q} = \tilde{\mathbf{U}}\tilde{\boldsymbol{\Lambda}}\tilde{\mathbf{U}} + \hat{\mathbf{Q}}$. Since $\mathbf{S}$ is maximal we have that $\mathbf{Q} - \hat{\mathbf{Q}} = \mathbf{SR}^{-1}\mathbf{S}^{\mathrm{T}}$ and multiplying both sides by $\tilde{\mathbf{U}}^{\mathrm{T}}$ on the left and $\tilde{\mathbf{U}}$ on the right we find $\tilde{\boldsymbol{\Lambda}} = \tilde{\mathbf{U}}^{\mathrm{T}}\mathbf{SR}^{-1}\mathbf{S}^{\mathrm{T}}\tilde{\mathbf{U}} = \tilde{\mathbf{S}}\mathbf{R}^{-1}\tilde{\mathbf{S}}^{\mathrm{T}}$, where $\tilde{\mathbf{S}} = \tilde{\mathbf{U}}^{\mathrm{T}}\mathbf{S}$. Similarly, setting $\tilde{\mathbf{P}} = \tilde{\mathbf{U}}^{\mathrm{T}}\mathbf{P}\tilde{\mathbf{U}}$, $\tilde{\mathbf{F}} = \tilde{\mathbf{U}}^{\mathrm{T}}\mathbf{F}\tilde{\mathbf{U}}$, and $\tilde{\mathbf{H}} = \mathbf{H}\tilde{\mathbf{U}}$ we can rewrite the DARE (again multiplying both sides by $\tilde{\mathbf{U}}^{\mathrm{T}}$ on the left and $\tilde{\mathbf{U}}$ on the right). The result is precisely the DARE from (18) with $\mathbf{P}$, $\mathbf{F}$, $\mathbf{H}$, $\mathbf{S}$, $\mathbf{Q}$ replaced by $\tilde{\mathbf{P}}$, $\tilde{\mathbf{F}}$, $\tilde{\mathbf{H}}$, $\tilde{\mathbf{S}}$, $\tilde{\boldsymbol{\Lambda}}$, respectively. Since $\tilde{\boldsymbol{\Lambda}} = \tilde{\mathbf{S}}\mathbf{R}^{-1}\tilde{\mathbf{S}}^{\mathrm{T}}$, we have reduced to the case above, so $\tilde{\mathbf{P}} = 0$ is a solution of this DARE. In other words $\mathbf{P}$ satisfying $\tilde{\mathbf{U}}^{\mathrm{T}}\mathbf{P}\tilde{\mathbf{U}} = 0$ is a solution of the DARE, and projecting the DARE onto the eigenvectors of $\mathbf{U}$ orthogonal to $\tilde{\mathbf{U}}$ would yield another DARE, which would need to be satisfied with a nonzero solution. Moreover, the resulting Kalman gain and stability condition become nontrivial in this case, but if the stability condition for the DARE is met, then we find a limiting covariance matrix, which is zero when projected onto the top M eigenvectors of $\mathbf{Q}$.

REFERENCES

Bélanger, P. R., 1974: Estimation of noise covariance matrices for a linear time-varying stochastic process. *Automatica*, **10**, 267–275, https://doi.org/10.1016/0005-1098(74)90037-5.

Berry, T., and T. Sauer, 2013: Adaptive ensemble Kalman filtering of non-linear systems. *Tellus*, **65A**, 20331, https://doi.org/10.3402/tellusa.v65i0.20331.

Dee, D. P., 1995: On-line estimation of error covariance parameters for atmospheric data assimilation. *Mon. Wea. Rev.*, **123**, 1128–1145, https://doi.org/10.1175/1520-0493(1995)123<1128:OLEOEC>2.0.CO;2.

Guttman, L., 1946: Enlargement methods for computing the inverse matrix. *Ann. Math. Stat.*, **17**, 336–343, https://doi.org/10.1214/aoms/1177730946.

Hamill, T. M., and J. S. Whitaker, 2005: Accounting for the error due to unresolved scales in ensemble data assimilation: A comparison of different approaches. *Mon. Wea. Rev.*, **133**, 3132–3147, https://doi.org/10.1175/MWR3020.1.

Hodyss, D., and N. Nichols, 2015: The error of representation: Basic understanding. *Tellus*, **67A**, 24822, https://doi.org/10.3402/tellusa.v67.24822.

Janjić, T., and S. E. Cohn, 2006: Treatment of observation error due to unresolved scales in atmospheric data assimilation. *Mon. Wea. Rev.*, **134**, 2900–2915, https://doi.org/10.1175/MWR3229.1.

——, and Coauthors, 2018: On the representation error in data assimilation. *Quart. J. Roy. Meteor. Soc.*, https://doi.org/10.1002/qj.3130, in press.

Julier, S. J., and J. K. Uhlmann, 2004: Unscented filtering and nonlinear estimation. *Proc. IEEE*, **92**, 401–422, https://doi.org/10.1109/JPROC.2003.823141.

Kuramoto, Y., and T. Tsuzuki, 1976: Persistent propagation of concentration waves in dissipative media far from thermal equilibrium. *Prog. Theor. Phys.*, **55**, 356–369, https://doi.org/10.1143/PTP.55.356.

Liu, Z.-Q., and F. Rabier, 2002: The interaction between model resolution, observation resolution and observation density in data assimilation: A one-dimensional study. *Quart. J. Roy. Meteor. Soc.*, **128**, 1367–1386, https://doi.org/10.1256/003590002320373337.

Lorenz, E. N., 1963: Deterministic nonperiodic flow. *J. Atmos. Sci.*, **20**, 130–141, https://doi.org/10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2.

——, 1996: Predictability—A problem partly solved. *Proc. Seminar on Predictability*, Vol. 1, Reading, United Kingdom, ECMWF, 18 pp., https://www.ecmwf.int/sites/default/files/elibrary/1995/10829-predictability-problem-partly-solved.pdf.

Mehra, R., 1970: On the identification of variances and adaptive Kalman filtering. *IEEE Trans. Autom. Control*, **15**, 175–184, https://doi.org/10.1109/TAC.1970.1099422.

——, 1972: Approaches to adaptive filtering. *IEEE Trans. Autom. Control*, **17**, 693–698, https://doi.org/10.1109/TAC.1972.1100100.

Mitchell, H. L., and R. Daley, 1997: Discretization error and signal/error correlation in atmospheric data assimilation. *Tellus*, **49A**, 32–53, https://doi.org/10.3402/tellusa.v49i1.12210.

Oke, P. R., and P. Sakov, 2008: Representation error of oceanic observations for data assimilation. *J. Atmos. Oceanic Technol.*, **25**, 1004–1017, https://doi.org/10.1175/2007JTECHO558.1.

Ran, A., and R. Vreugdenhil, 1988: Existence and comparison theorems for algebraic Riccati equations for continuous- and discrete-time systems. *Linear Algebra Appl.*, **99**, 63–83, https://doi.org/10.1016/0024-3795(88)90125-5.

Satterfield, E., D. Hodyss, D. D. Kuhl, and C. H. Bishop, 2017: Investigating the use of ensemble variance to predict observation error of representation. *Mon. Wea. Rev.*, **145**, 653–667, https://doi.org/10.1175/MWR-D-16-0299.1.

Simon, D., 2006: *Optimal State Estimation: Kalman, H Infinity, and Nonlinear Approaches.* Wiley-Interscience, 552 pp.

Sivashinsky, G., 1977: Nonlinear analysis of hydrodynamic instability in laminar flames I. Derivation of basic equations. *Acta Astronaut.*, **4**, 1177–1206, https://doi.org/10.1016/0094-5765(77)90096-0.

Van Leeuwen, P. J., 2015: Representation errors and retrievals in linear and nonlinear data assimilation. *Quart. J. Roy. Meteor. Soc.*, **141**, 1612–1623, https://doi.org/10.1002/qj.2464.

Whitaker, J. S., and T. M. Hamill, 2002: Ensemble data assimilation without perturbed observations. *Mon. Wea. Rev.*, **130**, 1913–1924, https://doi.org/10.1175/1520-0493(2002)130<1913:EDAWPO>2.0.CO;2.