Graphical Abstract

Model-free Kalman filtering in embedded space

Sean P. McGowan, William S. P. Robertson, John Maclean, Sanjeeva Balasuriya

Highlights

Model-free Kalman filtering in embedded space

Sean P. McGowan, William S. P. Robertson, John Maclean, Sanjeeva Balasuriya

- A general method for developing model-free data assimilation algorithms is presented
- Adaptations of the Extended, Ensemble, and Unscented Kalman filters are derived
- Embeddings leverage reconstructed dynamics for prediction and implicit localisation
- The coupling of local embeddings allows global dynamics to be consistent and stable
- Model-free algorithms offer higher accuracy in settings of sparse observation

Model-free Kalman filtering in embedded space

Sean P. McGowan[a,*], William S. P. Robertson[b], John Maclean[a], Sanjeeva Balasuriya[a]

[a]*School of Computer and Mathematical Sciences, University of Adelaide, Adelaide, 5005, SA, Australia*
[b]*School of Electrical and Mechanical Engineering, University of Adelaide, Adelaide, 5005, SA, Australia*

Abstract

Estimating the state of a system when only incomplete and imprecise information is available is an important problem across many disciplines. This is critical for initialising numerical models for forecasting, filtering measurement noise, estimating unobserved states, and interpolating between infrequent observations. In settings of chaotic dynamics, noisy observations and imperfect model structure present significant challenges for accurate state estimation. Standard methods of data assimilation are guaranteed to find optimal estimates but often assume that the underlying model is perfect. Adaptations that replace the model dynamics with purely data-driven representations using delay embedding have shown promise in settings where accurate models are unavailable. This paper describes a general method of strengthening these existing methods by performing data assimilation directly in delay embedding space that enhances the stability, accuracy, and data efficiency of the filtering algorithms. Data-driven extensions of the Extended, Ensemble, and Unscented Kalman filters are derived and compared in detail against the traditional model-based algorithms in a variety of settings. The presented methods offer improved results in settings of partial observation and high nonlinearity through reconstruction of the global state by coupling local data-driven representations of dynamics.

Keywords: Kalman filtering, data assimilation, spatiotemporal chaos, Takens embedding, data-driven dynamical systems

*Corresponding author at: School of Computer and Mathematical Sciences, University of Adelaide, Adelaide, 5005, SA, Australia. Email address: sean.mcgowan@adelaide.edu.au

1. Introduction

The ability to accurately predict chaotic and high-dimensional dynamical systems is dependent on both the accuracy of observations and the accuracy of the underlying model used for prediction. In perfect model scenarios, perturbations of initial conditions will result in divergence from true trajectories, and in perfect observation scenarios, inaccuracies in model structure will do the same. Methods from the study of data assimilation seek to find a balance between these two sources of error by combining observational data with models to minimise error variance or some appropriate cost function (Asch et al., 2016). The standard approach can be simply explained as making a forecast with a provided model and updating this forecast with information from observational data, resulting in an improved estimate of the state of the system. However, many of these techniques rely on the accuracy of the model, with significant model imperfection leading to a substantial decrease in performance (Toth and Peña, 2007). Under certain assumptions, a model of a system may be substituted by dynamics reconstructed from data following Takens Embedding Theorem (Takens, 1981). Following recent research directions of incorporating Takens delay embedding into data assimilation algorithms, this paper presents work in combining Kalman filtering algorithms with data-driven representations of dynamics to strengthen the estimation of states in settings where accurate models are unavailable.

The field of data assimilation aims to provide accurate forecasts and parameter estimates of a dynamical system by combining all available information of the system, including noisy observations of state variables and model predictions. The celebrated Kalman filter (Kalman and Bucy, 1961) allows newly available measurements of a dynamical system to correct previous estimates. The Kalman filter can be derived from a Bayesian perspective, where the updates of the filter are facilitated by estimation of the forecast state distribution given the observations and model forecast (Asch et al., 2016). Many practical data assimilation methods for dynamical systems are adaptations of the Kalman filter, which provide optimal solutions in settings of linear and Gaussian assumptions. These conditions are often at best approximations of true behaviour, yet the utility of these approaches is still shown by their reduced computational cost, convergence, and improved accuracy in practical settings.

Conventional data assimilation approaches demonstrate accuracy and stability issues when models become increasingly nonlinear and high-dimensional to match the dynamic detail of the systems of interest, for example, increasing the spatial resolution of oceanic models to resolve eddies. Performance is further degraded due to limitations of computational resources, incomplete understanding of dynamics across space and time scales, and parameter misspecification. This model error can originate from many sources simultaneously, and therefore, methods to mitigate the effects of model imperfection are important. Many existing methods seek to either characterise model error by their statistics, fit stochastic models to describe the model error beyond Gaussian assumptions (Harlim, 2017) (stochastic parameterisation (Berner et al., 2017) for example), or use machine learning to learn model error corrections (Farchi et al., 2021) or assimilation schemes (Revach et al., 2022; Bocquet et al., 2024). These techniques often rely on a provided dynamic model of the system of interest and therefore are still limited by the imperfection of the modelled dynamics.

Methods that substitute the underlying model with a data-driven evolution of the system hope to circumvent many of the issues associated with model imperfection by discovering the dynamics directly from observational data. Combinations of data assimilation with neural networks (Gottwald and Reich, 2021; Tomizawa and Sawada, 2021; Penny et al., 2022), Dynamic Mode Decomposition (Falconer et al., 2023), and sparse identification of dynamics (Rosafalco et al., 2024) have all demonstrated the utility of machine learning methods in data assimilation problems. A promising approach to learning the dynamics of an observational system for use in data assimilation is analogue-based methods, where the state evolution is facilitated through finding 'analogues' of the current state in a historical dataset and using their future states for the forecast. Analogue-based techniques offer an interpretable method for data-driven prediction by being grounded in the idea that similar states evolve in similar ways, rather than relying on black-box modelling. This idea has been implemented to extend the Ensemble Kalman filter and particle filter into model-free data assimilation algorithms (Hamilton et al., 2016; Lguensat et al., 2017; Gottwald and Reich, 2021; Wang and Jiang, 2025) which show considerable improvement compared to model-based methods when available models are imperfect. In settings of partial observation for high-dimensional chaotic systems, these previous investigations demonstrated that the incorporation of delay embedded states in the training data can aid in reconstructing the dynamics through the Takens Embedding Theorem (Takens, 1981). For spatiotemporal systems, prediction using only locally embedded states was shown to be more effective than dealing with the global state of the system (Lguensat et al., 2017).

In this paper, rather than using delay embedding prediction techniques to substitute the forecast step within the data assimilation scheme as in previous studies, we additionally consider the update of the attractor state via data assimilation directly in the locally embedded space of observations. This framework allows global spatiotemporal dynamics to be represented by coupled spatially-local systems, implicitly introduces localisation into the data assimilation schemes, and facilitates estimation of local low-dimensional tangent linear operators which all serve to make the algorithms computationally cheaper in high-dimensional settings. Previously, only the Ensemble Kalman filter and particle filter had been studied as potential data-driven filters; however, by considering the dynamics in embedded space, we show that many dynamic filters can be adapted into model-free counterparts. The effectiveness of different flavours of nonlinear filters in a specific setting must be determined through simulation and cannot be determined a priori and therefore, in this work, our goal is to derive and compare model-free extensions of the Extended, Ensemble, and Unscented Kalman filters. They are validated on a simulated example in a range of observational settings and on an empirical dataset to assess their performance in a range of practical situations. They are also compared against models with varying levels of model imperfection to determine contexts in which it is preferable to substitute the dynamic model with this model-free approach.

The paper is organised as follows. The background theory of Kalman filtering and its nonlinear extensions are described in Section 2, followed by a discussion of how we adapted these algorithms into data-driven, model-free settings. The algorithms for the Extended, Ensemble, and Unscented Kalman filters are outlined in both model-based and model-free scenarios. The considerations for effective model-free filtering are explained in this section, with a focus on how spatially-local embeddings can be coupled to transmit information globally. Section 3 presents the results of the developed filtering algorithms on the multiscale Lorenz 96 system and an example using meteorological data. The results of the multiscale Lorenz 96 system provide a comprehensive comparative analysis across a wide range of observational settings. Future directions and limitations of the presented approaches to data-driven filtering are finally discussed in Section 4.

2. Methods

This section introduces the standard Kalman filter as well as three of its nonlinear extensions. The method of adapting these algorithms to their model-free counterparts is then explained with a discussion of their additional considerations for implementation. The following is a description of the estimation problem adapted from Asch et al. (2016).

Consider the evolution of a model state $\boldsymbol{x}_k \in \mathbb{R}^N$ at discrete time k through the model $\mathcal{M}_k : \mathbb{R}^N \to \mathbb{R}^N$ given by

$$\boldsymbol{x}_k = \mathcal{M}_k(\boldsymbol{x}_{k-1}) + \boldsymbol{w}_k,$$

as well as observation of the true system $\boldsymbol{y}_k$ at time k with observation operator $\mathcal{H}_k : \mathbb{R}^N \to \mathbb{R}^N$,

$$\boldsymbol{y}_k = \mathcal{H}_k(\boldsymbol{x}_k) + \boldsymbol{v}_k.$$

Here, $\boldsymbol{w}_k$ is the process noise at time k with $\boldsymbol{w}_k \sim \mathcal{N}(0, \boldsymbol{Q}_k)$, and $\boldsymbol{v}_k$ is the measurement noise with $\boldsymbol{v}_k \sim \mathcal{N}(0, \boldsymbol{R}_k)$. If the noise covariance matrices $\boldsymbol{Q}_k$ and $\boldsymbol{R}_k$ do not change in time, they are denoted $\boldsymbol{Q}$ and $\boldsymbol{R}$.

The *estimation* problem is: given an estimate of the state, $\boldsymbol{x}_k$, how can this be updated optimally (in the sense that it minimises error variances) based on the current observation $\boldsymbol{y}_k$? The filter cycle iteratively determines prior estimates of the state (the forecast state $\boldsymbol{x}^f$) by applying the model operator to the posterior estimate at the previous time step (the analysis state $\boldsymbol{x}^a$). The forecast step is given by

$$\boldsymbol{x}_k^f = \mathcal{M}_k(\boldsymbol{x}_{k-1}^a),$$

after which the forecast is corrected using the Kalman gain, $\boldsymbol{K}_k$, and the innovation vector, $\boldsymbol{y}_k - \mathcal{H}_k(\boldsymbol{x}_k^f)$ in the update step

$$\boldsymbol{x}_k^a = \boldsymbol{x}_k^f + \boldsymbol{K}_k\big(\boldsymbol{y}_k - \mathcal{H}_k(\boldsymbol{x}_k^f)\big). \tag{1}$$

The uncertainty in both analysis and forecast states are described by the analysis and forecast error covariance matrices that are given by $\boldsymbol{P}_k^a = \mathbb{E}[(\boldsymbol{x}_k^a - \boldsymbol{x}_k)(\boldsymbol{x}_k^a - \boldsymbol{x}_k)^\mathsf{T}]$ and $\boldsymbol{P}_k^f = \mathbb{E}[(\boldsymbol{x}_k^f - \boldsymbol{x}_k)(\boldsymbol{x}_k^f - \boldsymbol{x}_k)^\mathsf{T}]$. The standard Kalman filter provides optimal analysis estimation of the state when the model and observation operators are linear, represented as matrices $\boldsymbol{M}_k$ and $\boldsymbol{H}_k$. The optimal Kalman gain in this situation is

$$\boldsymbol{K}_k^* = \boldsymbol{P}_k^f \boldsymbol{H}_k^\mathsf{T}\big(\boldsymbol{H}_k \boldsymbol{P}_k^f \boldsymbol{H}_k^\mathsf{T} + \boldsymbol{R}_k\big)^{-1}, \tag{2}$$

with error covariance matrices of

$$\boldsymbol{P}_k^f = \boldsymbol{M}_k \boldsymbol{P}_{k-1}^a \boldsymbol{M}_k^\top + \boldsymbol{Q}_k$$

and

$$\boldsymbol{P}_k^a = (\boldsymbol{I} - \boldsymbol{K}_k \boldsymbol{H}_k) \boldsymbol{P}_k^f. \tag{3}$$

This is the standard Kalman filter, which may be implemented in settings of model linearity and Gaussian noise assumptions.

In practical settings, adaptations of the Kalman filter are needed when there are nonlinearities in the model dynamics to appropriately represent the forecast state distribution. Two methods for dealing with nonlinearities can be broadly categorised as: linearising through the tangent linear model, and sampling of the state space to approximate probability density functions. In this paper, three nonlinear extensions to the Kalman filter, widely used in navigation, control, geophysics, and weather, were investigated: the Extended, Ensemble, and Unscented Kalman filters. Figure 1 summarizes how each of these filters determines the forecast state $\boldsymbol{x}^f$ and the corresponding forecast error covariance matrix $\boldsymbol{P}^f$ given the analysis state and error covariance matrix. The following sections describe these filters in detail, followed by an explanation of how these methods may be adapted into their data-driven counterparts.

2.1. Extended Kalman filter

In settings of nonlinearity, the Kalman filter may be adapted through the linearisation of the dynamics. The Extended Kalman filter (ExKF) is one method that provides a first-order approximation of the nonlinear dynamics via the tangent linear operator, or Jacobian. The state distribution is evolved via the tangent dynamics to estimate the forecast state distribution.

If the dynamics of $\mathcal{M}_k$ can be approximated by the tangent linear operator at $\boldsymbol{x}_k$,

$$\boldsymbol{M}_{i,j}|_{\boldsymbol{x}_k} = \left.\frac{\partial \mathcal{M}_i}{\partial x_j}\right|_{\boldsymbol{x}_k}, \tag{4}$$

then the forecast error is

$$\boldsymbol{e}_k^f = \boldsymbol{x}_k^f - \boldsymbol{x}_k \approx \boldsymbol{M}_k \boldsymbol{e}_{k-1}^a - \boldsymbol{w}_k$$

and the forecast error covariance matrix becomes

$$\boldsymbol{P}_k^f = \boldsymbol{M}_k \boldsymbol{P}_{k-1}^a \boldsymbol{M}_k^\top + \boldsymbol{Q}_k. \tag{5}$$

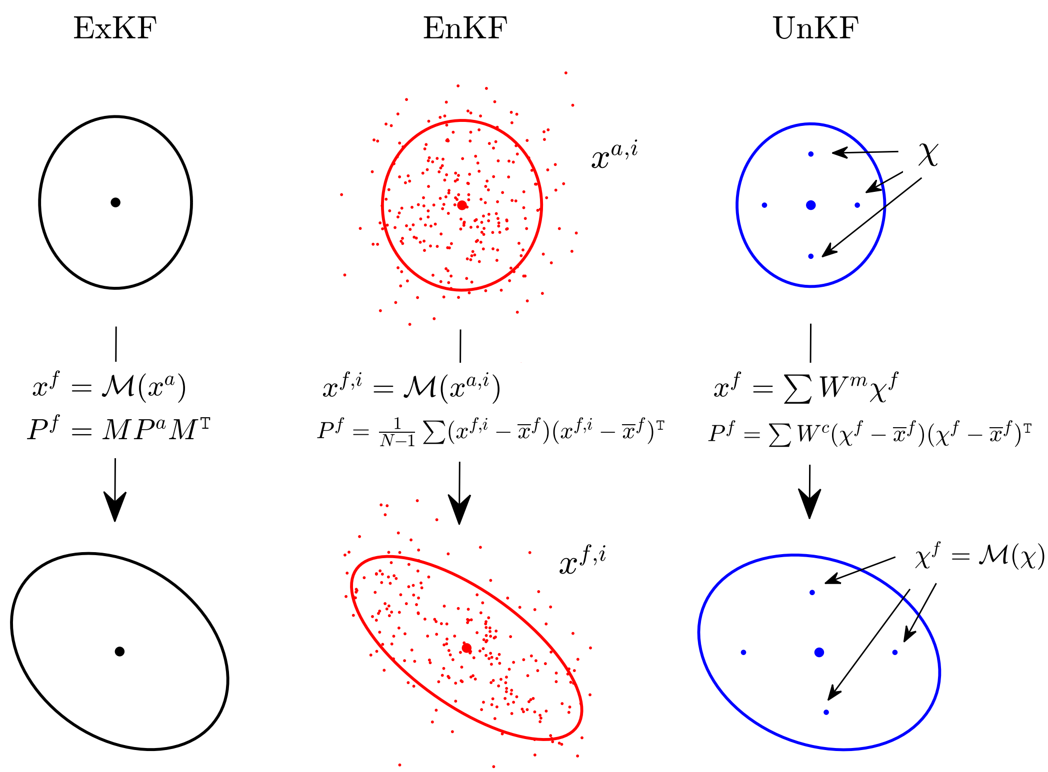

Figure 1: Examples of mean and covariance propagation for nonlinear extensions of the Kalman filter.

For continuous dynamics, the discrete tangent linear operator may be approximated using the matrix exponential. The expression for the optimal Kalman gain is the same as in the standard Kalman filter in Equation (2). The algorithm describing the process of the Extended Kalman filter is presented in Algorithm 4 in Appendix A. In this filter, the propagation of the state distribution through the dynamics is assumed to be approximated well by the first-order linearisation of the dynamics via the tangent linear operator. In practice, only slight deviations from linear or Gaussian assumptions are handled well by the ExKF (Asch et al., 2016), and so we turn our attention to sample points methods that avoid linearisation of the model.

2.2. *Ensemble Kalman filter*

Rather than approximating the forecast state distribution via tangent dynamics, the Ensemble Kalman filter (EnKF) instead approximates it by calculating the spread of an ensemble of perturbed states as a Monte Carlo approximation of the distribution (Evensen, 2009). This method avoids the calculation of the tangent linear operator and hence is more robust for highly nonlinear systems with high levels of noise.

Given an ensemble of forecast states $\boldsymbol{x}_k^{f,i} = \mathcal{M}_k(\boldsymbol{x}_{k-1}^{a,i}) + \boldsymbol{w}_k$ with $i = 1, \ldots, N_{\text{ens}}$ and $\boldsymbol{w}_k \sim \mathcal{N}(0, \boldsymbol{Q}_k)$, the forecast error covariance matrix may be estimated from the ensemble statistics as

$$\boldsymbol{P}_k^f = \frac{1}{N_{\text{ens}} - 1} \sum_{i=1}^{N_{\text{ens}}} \left(\boldsymbol{x}_k^{f,i} - \overline{\boldsymbol{x}}_k^f\right)\left(\boldsymbol{x}_k^{f,i} - \overline{\boldsymbol{x}}_k^f\right)^{\mathsf{T}},$$

where the ensemble average is given by

$$\overline{\boldsymbol{x}}_k^f = \frac{1}{N_{\text{ens}}} \sum_{i=1}^{N_{\text{ens}}} \boldsymbol{x}_k^{f,i}.$$

The optimal Kalman gain may then be computed using the same expression as in the standard Kalman filter, and with this gain the analysis on each ensemble member can be performed by

$$\boldsymbol{x}_k^{a,i} = \boldsymbol{x}_k^{f,i} + \boldsymbol{K}_k\left(\boldsymbol{y}_k^i + \boldsymbol{v}_k - \boldsymbol{H}_k \boldsymbol{x}_k^{f,i}\right)$$

where $\boldsymbol{v}_k \sim \mathcal{N}(0, \boldsymbol{R}_k)$.

The approximation of the state distribution given by the ensemble approaches the true distribution as the size of the ensemble tends to infinity. The errors resulting from the finite size of the ensemble can be handled by inflation and localisation of the covariance matrix Asch et al. (2016) which have been critical for the success of the EnKF in high-dimensional settings. The algorithm detailing the framework for filtering incoming observations using the Ensemble Kalman filter is shown in Algorithm 5 in Appendix A.

2.3. Unscented Kalman filter

Another method to handle nonlinearities is to take a deterministic sampling approach, where a minimal set of sampling points are used to describe the mean and covariance of the state distribution; this method is used by the Unscented Kalman filter (UnKF) (Julier and Uhlmann, 1997; Wan and Van Der Merwe, 2000). The selection of these points (known as sigma points) allows their evolution through the dynamics to provide an approximation of the mean and covariance of the forecast distribution.

The values of the sigma points χ^i and their corresponding weights $\boldsymbol{W}^i$, for $i = 1, \ldots, 2N+1$, is found through the Unscented Transform by

$$\chi_k^{a,i} = \begin{cases} \boldsymbol{x}_k^a & \text{for } i = 1 \\ \boldsymbol{x}_k^a + \left(\sqrt{(N+\lambda)\boldsymbol{P}_k^a}\right)_{i-1} & \text{for } i = 2, \ldots, N+1 \\ \boldsymbol{x}_k^a - \left(\sqrt{(N+\lambda)\boldsymbol{P}_k^a}\right)_{i-N-1} & \text{for } i = N+2, \ldots, 2N+1\,, \end{cases}$$

$$\boldsymbol{W}^{m,i} = \begin{cases} \dfrac{\lambda}{N+\lambda} & \text{for } i = 1 \\ \dfrac{1}{2(N+\lambda)} & \text{for } i = 2, \ldots, 2N+1\,, \end{cases}$$

$$\boldsymbol{W}^{c,i} = \begin{cases} \dfrac{\lambda}{N+\lambda} + (1 - \alpha^2 - \beta^2) & \text{for } i = 1 \\ \dfrac{1}{2(N+\lambda)} & \text{for } i = 2, \ldots, 2N+1\,, \end{cases}$$

where $\lambda = \alpha^2(N+\kappa) - N$. The parameter α determines the spread of points, κ is a secondary scaling parameter, and β can be used to include additional information of the underlying state distribution. The matrix square root is calculated via the Cholesky decomposition. Given the forecasts of sigma points $\chi_k^{f,i} = \mathcal{M}_k\left(\chi_k^{a,i}\right)$, the mean and covariance are approximated by weighted samplings

$$\begin{aligned} \overline{\boldsymbol{x}}_k^f &= \sum_{i=1}^{2N+1} \boldsymbol{W}^{m,i} \chi_k^{f,i} \\ \boldsymbol{P}_k^f &= \sum_{i=1}^{2N+1} \boldsymbol{W}^{c,i} \left(\chi_k^{f,i} - \overline{\boldsymbol{x}}_k^f\right)\left(\chi_k^{f,i} - \overline{\boldsymbol{x}}_k^f\right)^\mathsf{T} + \boldsymbol{Q}_k. \end{aligned}$$

This formulation substantially decreases the number of sample points required to accurately approximate the mean and error covariance matrices of the forecast state up to third order for any nonlinearity (Wan and Van Der Merwe, 2000).

The calculation of the Kalman gain for the Unscented Kalman filter first requires the calculation of the measurement covariance and cross covariance, $\boldsymbol{P}_k^{yy}$ and $\boldsymbol{P}_k^{xy}$

through

$$
\begin{aligned}
\boldsymbol{P}_k^{yy} &= \sum_{i=1}^{2N+1} \boldsymbol{W}^{c,i} \left(\boldsymbol{\chi}_k^{f,i} - \boldsymbol{H}_k \overline{\boldsymbol{x}}_k^f\right)\left(\boldsymbol{\chi}_k^{f,i} - \boldsymbol{H}_k \overline{\boldsymbol{x}}_k^f\right)^{\mathsf{T}} + \boldsymbol{R}_k \\
\boldsymbol{P}_k^{xy} &= \sum_{i=1}^{2N+1} \boldsymbol{W}^{c,i} \left(\boldsymbol{\chi}_k^{f,i} - \overline{\boldsymbol{x}}_k^f\right)\left(\boldsymbol{\chi}_k^{f,i} - \boldsymbol{H}_k \overline{\boldsymbol{x}}_k^f\right)^{\mathsf{T}}.
\end{aligned}
$$

From this, the Kalman gain can be calculated as

$$
\boldsymbol{K}_k = \boldsymbol{P}_k^{xy} (\boldsymbol{P}_k^{yy})^{-1}.
$$

The full algorithm for the Unscented Kalman filter is described in Algorithm 6 in Appendix A.

2.4. Kalman filtering in embedded space

In many applications of dynamic filtering, models become increasingly complex, high-dimensional, and nonlinear such as those governing the circulation of the atmosphere or ocean. The inevitable model error of these representations from imperfect modelling of subgrid processes, parameter misspecification, or structural imperfection results not only in inaccurate forecasting of analysis states but non-Gaussian process noise. These inaccuracies may substantially degrade the performance of the filter by offering states that are not representative of the underlying system. Further, in some cases, an observed system may be difficult or impossible to accurately model due to coupling with a large number of unobserved variables. In these settings, if sufficient data from the system of interest can be measured and if the dynamics can effectively be represented by embedding reconstructions of the data (Takens, 1981), it may be preferable to use the reconstructed dynamics to drive the model evolution and state update.

For spatiotemporal dynamics, as well as taking time delays at a particular grid point, spatial neighbours and their delays are also typically included as embedding coordinates, which is physically motivated by the fact that many real spatiotemporal systems evolve predominantly through local interaction. This approach admits dimension-reduced representations of dynamics on local scales to substitute high-dimensional models, effectively localising the dynamics without the need for explicit covariance localisation techniques as with model-based methods. For appropriate choice of time delay τ, number of time delays D, and number of spatial neighbours ν, the trajectory at grid point n can be embedded into $\mathbb{R}^M$, where $M = (2\nu + 1)(D + 1)$, with coordinates

$$\begin{aligned} z_{n,k} = \Phi_{\tau,D,\nu}(x_{n,k}) = [&x_{n,k}, x_{n,k-\tau}, \ldots, x_{n,k-D\tau}, x_{n-\nu,k}, \ldots, x_{n-\nu,k-D\tau}, \\ &\vdots \\ &x_{n-1,k}, \ldots, x_{n-1,k-D\tau}, x_{n+1,k}, \ldots, x_{n+1,k-D\tau}, \\ &\vdots \\ &x_{n+\nu,k}, \ldots, x_{n+\nu,k-D\tau}], \end{aligned}$$

where $\Phi_{\tau,D,\nu}(x_{n,k})$ denotes the local embedding at grid point n. The time delay and embedding dimension are typically optimised respectively through the mutual information (Fraser and Swinney, 1986) and false nearest neighbour methods (Kennel et al., 1992). It is often assumed throughout this section that the system exhibits periodic boundary conditions with $x_{n+N} = x_n$, which simplifies descriptions, but other forms of boundary conditions can be accounted for (e.g., assuming constant values for neighbours out of the domain as described in Isensee et al. (2020)). A visual example of this spatiotemporal embedding is shown in Figure 2.

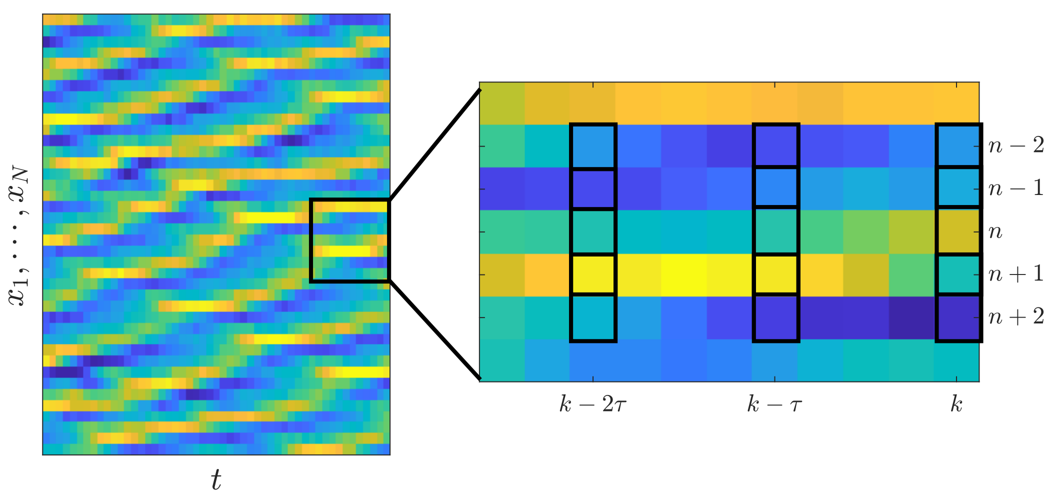

Figure 2: Example of spatiotemporal embedding at $x_{n,k}$. The black squares highlight the states that make up the embedded state, here using $\nu = 2$ spatial neighbours and $D = 2$ time delays of size $\tau = 4$ at each location to represent the local dynamics in $\mathbb{R}^{15}$.

The main idea behind model-free filtering is to substitute the model forecast step with a forecast made through these reconstructed attractors in embedded space. Rather than performing data assimilation on the model state $\boldsymbol{x}_k \in \mathbb{R}^N$ by substituting the model forecast step in the model-based filters with a model-free prediction using the reconstructed attractors, we consider data assimilation of the local state in embedded space, $\boldsymbol{z}_{n,k} \in \mathbb{R}^M$, at each grid point $n = 1, \ldots, N$,

$$\boldsymbol{z}^f_{n,k} = \mathcal{G}_n(\boldsymbol{z}^a_{n,k-1}), \quad x^f_{n,k} = (\Phi_{\tau,D,\nu})^{-1} \circ \mathcal{G}_n \circ \Phi_{\tau,D,\nu}(x^a_{n,k-1}).$$

This enables the embedded dynamics to be explicitly incorporated into the data assimilation scheme, while also implicitly introducing localisation. The embedding can then be inverted to retrieve the forecast state x^f_n (or its neighbours or delayed states). The dynamics of the reconstructed attractor at grid point n, $\mathcal{G}_n$, requires training data in the form of observations of the relevant delays and spatial neighbours that sufficiently traverse the attractor. In special cases of dynamical systems where the local dynamics at each grid point are equivalent, the attractor geometry and dynamics are equivalent and therefore $\mathcal{G}_1 = \mathcal{G}_2 = \cdots = \mathcal{G}_N = \mathcal{G}$. This allows observations at each grid point to be used for reconstruction of the attractor at any grid point for homogeneous systems, effectively reducing the required training dataset by a factor of N. The nearest neighbour method is used to forecast the embedded states, which finds the L closest states to the current embedded state with respect to the Euclidean distance, calculates the linear regression matrix which predicts the future state of each neighbour, and then uses this matrix to predict the analogue state. For the ExKF we additionally require an estimation of the tangent linear operator (in Equations (4) and (5)), which is often difficult to calculate for high-dimensional models. In this study, the tangent linear operator of the local embedding at each grid point, $\mathbf{G}_n$, is estimated numerically using a central finite difference scheme. The methods presented in this paper are agnostic to the choice of prediction scheme and more sophisticated techniques such as random feature maps (Gottwald and Reich, 2021), Dynamic Mode Decomposition (DMD) (Wang and Jiang, 2025), or Sparse Identification of Nonlinear Dynamics (SINDy) (Brunton et al., 2016) could substitute the forecast step and offer improved results. Figure 3 depicts a representation of the standard state update for model-based methods compared with the embedding update for model-free methods.

During the analysis step, the current (potentially partial and noisy) observations $\boldsymbol{y}_k$ are used to update each relevant coordinate at each grid point n with a Kalman gain $\boldsymbol{K}_{n,k} \in \mathbb{R}^{M \times M}$,

$$\boldsymbol{z}^a_{n,k} = \boldsymbol{z}^f_{n,k} + \boldsymbol{K}_{n,k}\big(\Phi_{\tau,D,\nu}(y_{n,k}) - \boldsymbol{H}_{n,k}\boldsymbol{z}^f_{n,k}\big), \qquad \text{for } n = 1, \ldots, N. \tag{6}$$

Crucially, the observations $\boldsymbol{y}_k$ and observation matrix $\boldsymbol{H}_{n,k}$ used to update the embedded states are dependent on all available information at the current time,

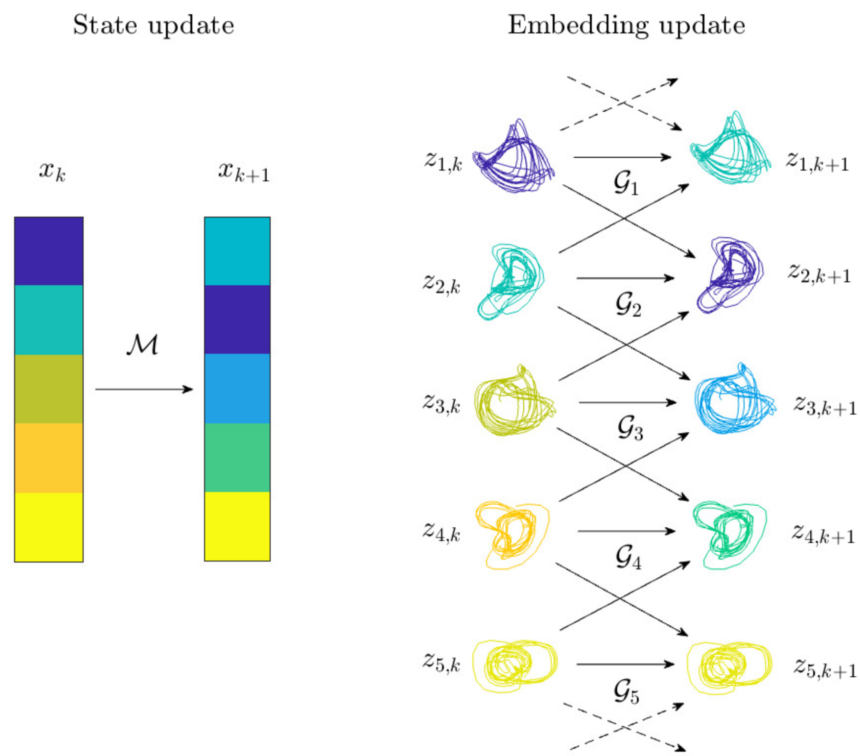

Figure 3: Comparison of state vector evolution and evolution of embedded states. The standard model-based state update evolves the global state vector through the modeled dynamics whereas the embedding update evolves each grid point based on the local reconstruction of the dynamics at that grid point and neighbouring points. For simplicity, the number of spatial neighbours shown in this example is $\nu = 1$ and the dotted lines represent couplings resulting from periodic boundary conditions. The colour of each grid point or local attractor represents the corresponding grid point state estimate.

including observations of neighbouring and past delayed states. The process noise and measurement noise covariance matrices, $\boldsymbol{Q}_{n,k}$ and $\boldsymbol{R}_{n,k}$, similarly describe the effect of these noise sources in the embedded space and are denoted $\boldsymbol{Q}$ and $\boldsymbol{R}$ if they do not vary in space and time. In practice, these covariance matrices must be estimated using adaptive algorithms (Akhlaghi et al., 2017). This Kalman gain is calculated for each grid point using the estimate of the forecast error covariance matrix in embedded space, $\boldsymbol{P}^f_{n,k}$, through selection of one of the three nonlinear Kalman methods previously introduced. By performing data assimilation in the embedded space, where we have access to a representation of the dynamics, we can accommodate many forms of filtering algorithms, not just sample point filters.

2.5. Local attractor coupling via nudging

Initial experimentation of the model-free filters showed that state estimation may often become unstable when the observations are sparse in space. One source of instability in the model-free scheme is the large degree of redundancy in the matrix of embedded states, where neighbouring attractors share states but may produce varying predictions for those states. One resolution to consider all embedded states is to nudge the values of the repeated entries to their mean

$$(x_{n,k})^l \rightarrow (x_{n,k})^l + \gamma\big(\overline{x_{n,k}} - (x_{n,k})^l\big) \tag{7}$$

where $(\cdot)^l$ refers to the l-th repetition of the variable in the embedding matrix $\boldsymbol{Z}_k = [\boldsymbol{z}_{1,k}, \cdots, \boldsymbol{z}_{N,k}]$, and the parameter $0 \leq \gamma \leq 1$ controls the intensity of the nudging. Values of γ close to 1 result in stronger nudging, with $\gamma = 0$ resulting in no nudging between embedded trajectories. States in each embedded vector that correspond to spatial neighbours will be updated by *their* spatial neighbours, imparting a longer range coupling. This coupling extends beyond the neighbours defined by parameter ν to include the attractors from grid points $n-2\nu, \ldots, n+2\nu$. The strength of this coupling diminishes for neighbours further away, with the immediate attractors at $n-1$ and $n+1$ having the strongest coupling by sharing $2\nu(D+1)$ state estimates and the furthest attractors at $n-2\nu$ and $n+2\nu$ sharing only $D+1$ state estimates. The embedding prediction at grid point n is therefore of the form

$$\boldsymbol{z}_{n,k+1} = \hat{\mathcal{G}}_n(\boldsymbol{z}_{n,k}) + \hat{\mathcal{G}}_{n-2\nu}(\boldsymbol{z}_{n-2\nu,k}) + \cdots + \hat{\mathcal{G}}_{n+2\nu}(\boldsymbol{z}_{n+2\nu,k}),$$

where the $\hat{\mathcal{G}}$ are defined by the parameter γ and the reconstructed attractors of the current grid point and neighbours $\mathcal{G}_n, \mathcal{G}_{n-2\nu}, \ldots, \mathcal{G}_{n+2\nu}$. This allows all available information to flow locally throughout the global state with the speed of information travel given by the simulation time step Δt and the number of spatial neighbours ν. In practice, this results in greater stability of the reconstructed dynamics and convergence to a global attractor while only considering spatially-local embeddings. Comparison between this coupling approach and others as well as analysis of the coupling parameter γ is shown in Section 3.

The approach of nudging the embedded states to ensure consistency across locally trained dynamics when the global dynamics are unknown is conceptually similar to the idea of diffusive coupling in dynamical systems. In certain settings, a consequence of diffusive coupling of subsystems of a dynamical system is synchronisation (Hale, 1997), where all subsystems have equal state as time goes to infinity (when subsystems are identical) or generalised synchronisation (Eroglu et al., 2017) (when subsystems are not identical). Generalised synchronisation describes the situation where the diffusive coupling of one subsystem onto another results in convergence of trajectories in the 'forced' subsystem. This case is the scenario of interest to us, where the coupling of all local subsystems to their neighbours produces a consistent global state. The presented nudging technique introduces a partial diffusive coupling on the locally trained dynamics that, with sufficiently strong coupling, allows self-organising of the global state through information propagation over space and time. The strength of diffusive coupling in our presented algorithms is controlled by the parameter γ.

The model-free adaptations of the three introduced nonlinear Kalman filters can now be described. At each time step $k = 2, \ldots, K$, the filtering algorithms are iterated over the local attractors at each grid point $n = 1, \ldots, N$. Prior estimates $\boldsymbol{z}^f_{n,k}$ for the embedded state are found by evolving the embedded dynamics at the previous analysis state $\boldsymbol{z}^a_{n,k-1}$. Depending on the choice of nonlinear Kalman filter, the forecast state distribution in embedded space is estimated by the tangent linear operator, ensemble approximation, or sample point approximation to estimate the forecast error covariance matrix $\boldsymbol{P}^f_{n,k}$. The analysis state is then found by calculating the Kalman gain and updating the forecast using the information available from observations of the current grid point, neighbouring grid points, and their delays. After iterating over all N grid points, the equivalent states of all local attractors are nudged towards their mean with strength γ. This is repeated for the next filter cycle at time $k+1$. Algorithms 1, 2, and 3 describe this process for the three data-driven Kalman filters applied in embedded space.

Algorithm 1 Model-free Extended Kalman filter

Choose τ, D, ν, γ, $\boldsymbol{Q}_{n,k}$, $\boldsymbol{R}_{n,k}$
Train $\mathcal{G}_n$ and $\boldsymbol{G}_n$ for $n = 1$ to N on available data
Initialise $\boldsymbol{Z}_1^a$, $\boldsymbol{P}_{1:N,1}^a$
for $k = 2$ to K **do**
  for $n = 1$ to N **do**
    $\boldsymbol{z}_{n,k}^f = \mathcal{G}_{n,k}(\boldsymbol{z}_{n,k-1}^a)$
    $\boldsymbol{P}_{n,k}^f = \boldsymbol{G}_{n,k}\boldsymbol{P}_{n,k-1}^a\boldsymbol{G}_{n,k}^\top + \boldsymbol{Q}_{n,k}$
    $\boldsymbol{K}_{n,k} = \boldsymbol{P}_{n,k}^f\boldsymbol{H}_{n,k}^\top(\boldsymbol{H}_{n,k}\boldsymbol{P}_{n,k}^f\boldsymbol{H}_{n,k}^\top + \boldsymbol{R}_{n,k})^{-1}$
    $\boldsymbol{z}_{n,k}^a = \boldsymbol{z}_{n,k}^f + \boldsymbol{K}_{n,k}(\Phi_{\tau,D,\nu}(y_{n,k}) - \boldsymbol{H}_{n,k}\boldsymbol{z}_{n,k}^f)$ (Equation 6)
    $\boldsymbol{P}_{n,k}^a = (\boldsymbol{I} - \boldsymbol{K}_{n,k}\boldsymbol{H}_{n,k})\boldsymbol{P}_{n,k}^f$
  end for
  Nudge equivalent states of $\boldsymbol{Z}_k^a$ towards their mean with γ (Equation 7)
end for

Algorithm 2 Model-free Ensemble Kalman filter

Choose τ, D, ν, γ, N_{ens}, μ, $\boldsymbol{Q}_{n,k}$, $\boldsymbol{R}_{n,k}$
Train $\mathcal{G}_n$ for $n = 1$ to N on available data
Initialise $[\boldsymbol{Z}_1^a]^i$ for $i = 1$ to N_{ens}
for $k = 2$ to K **do**
  for $n = 1$ to N **do**
    for $i = 1$ to N_{ens} **do**
      $\boldsymbol{z}_{n,k}^{f,i} = \mathcal{G}_{n,k}(\boldsymbol{z}_{n,k-1}^{a,i}) + \boldsymbol{w}_k$ where $\boldsymbol{w}_k \sim \mathcal{N}(0, \boldsymbol{Q}_{n,k})$
    end for
    $\overline{\boldsymbol{z}}_{n,k}^{f} = \frac{1}{N_{\text{ens}}} \sum_{i=1}^{N_{\text{ens}}} \boldsymbol{z}_{n,k}^{f,i}$
    $\boldsymbol{P}_{n,k}^{f} = \frac{1}{N_{\text{ens}}-1}([\boldsymbol{z}_{n,k}^{f,1:N_{\text{ens}}}] - \overline{\boldsymbol{z}}_{n,k}^{f})([\boldsymbol{z}_{n,k}^{f,1:N_{\text{ens}}}] - \overline{\boldsymbol{z}}_{n,k}^{f})^{\top}$
    $\boldsymbol{K}_{n,k} = \boldsymbol{P}_{n,k}^{f} \boldsymbol{H}_{n,k}^{\top} (\boldsymbol{H}_{n,k} \boldsymbol{P}_{n,k}^{f} \boldsymbol{H}_{n,k}^{\top} + \boldsymbol{R}_{n,k})^{-1}$
    for $i = 1$ to N_{ens} **do**
      $\boldsymbol{z}_{n,k}^{a,i} = \boldsymbol{z}_{n,k}^{f,i} + \boldsymbol{K}_{n,k}(\Phi_{\tau,D,\nu}(y_{n,k}) + \boldsymbol{v}_k - \boldsymbol{H}_{n,k}\boldsymbol{z}_{n,k}^{f,i})$ (Equation 6) where $\boldsymbol{v}_k \sim \mathcal{N}(0, \boldsymbol{R}_{n,k})$
    end for
    $\overline{\boldsymbol{z}}_{n,k}^{a} = \frac{1}{N_{\text{ens}}} \sum_{i=1}^{N_{\text{ens}}} \boldsymbol{z}_{n,k}^{a,i}, \quad \boldsymbol{z}_{n,k}^{a,i} = \overline{\boldsymbol{z}}_{n,k}^{a} + \mu(\overline{\boldsymbol{z}}_{n,k}^{a} - \boldsymbol{z}_{n,k}^{a,i})$
  end for
  for $i = 1$ to N_{ens} **do**
    Nudge equivalent states of $[\boldsymbol{Z}_k^a]^i$ towards their mean with γ (Equation 7)
  end for
end for

Algorithm 3 Model-free Unscented Kalman filter

Choose τ, D, ν, γ, α, β, κ, $\boldsymbol{Q}_{n,k}$, $\boldsymbol{R}_{n,k}$
Train $\mathcal{G}_n$ for $n = 1$ to N on available data
Calculate M, λ, $\boldsymbol{W}^{m,i}$, $\boldsymbol{W}^{c,i}$
Initialise $\boldsymbol{Z}_1^a$, $\boldsymbol{P}_{1:N,1}^a$
for $k = 2$ to K **do**
  for $n = 1$ to N **do**
    for $i = 1$ to $2M + 1$ **do**
      $\boldsymbol{\chi}_{n,k-1}^{a,i} = \left[\boldsymbol{z}_{n,k-1}^a, \boldsymbol{z}_{n,k-1}^a \pm \sqrt{(M+\lambda)\boldsymbol{P}_{n,k-1}^a}\right]$
      $\boldsymbol{\chi}_{n,k}^{f,i} = \mathcal{G}_{n,k}(\boldsymbol{\chi}_{n,k-1}^{a,i})$
    end for
    $\overline{\boldsymbol{z}}_{n,k}^f = \sum_{i=1}^{2M+1} \boldsymbol{W}^{m,i} \boldsymbol{\chi}_{n,k}^{f,i}$
    $\boldsymbol{P}_{n,k}^f = \sum_{i=1}^{2M+1} \boldsymbol{W}^{c,i} (\boldsymbol{\chi}_{n,k}^{f,i} - \overline{\boldsymbol{z}}_{n,k}^f)(\boldsymbol{\chi}_{n,k}^{f,i} - \overline{\boldsymbol{z}}_{n,k}^f)^\top + \boldsymbol{Q}_{n,k}$
    $\boldsymbol{P}_{n,k}^{yy} = \sum_{i=1}^{2M+1} \boldsymbol{W}^{c,i} (\boldsymbol{\chi}_{n,k}^{f,i} - \boldsymbol{H}_{n,k}\overline{\boldsymbol{z}}_{n,k}^f)(\boldsymbol{\chi}_{n,k}^{f,i} - \boldsymbol{H}_{n,k}\overline{\boldsymbol{z}}_{n,k}^f)^\top + \boldsymbol{R}_{n,k}$
    $\boldsymbol{P}_{n,k}^{xy} = \sum_{i=1}^{2M+1} \boldsymbol{W}^{c,i} (\boldsymbol{\chi}_{n,k}^{f,i} - \overline{\boldsymbol{z}}_{n,k}^f)(\boldsymbol{\chi}_{n,k}^{f,i} - \boldsymbol{H}_{n,k}\overline{\boldsymbol{z}}_{n,k}^f)^\top$
    $\boldsymbol{K}_{n,k} = \boldsymbol{P}_{n,k}^{xy}(\boldsymbol{P}_{n,k}^{yy})^{-1}$
    $\boldsymbol{z}_{n,k}^a = \boldsymbol{z}_{n,k}^f + \boldsymbol{K}_{n,k}(\Phi_{\tau,D,\nu}(y_{n,k}) - \boldsymbol{H}_{n,k}\boldsymbol{z}_{n,k}^{f,i})$ (Equation 6)
    $\boldsymbol{P}_{n,k}^a = \boldsymbol{P}_{n,k}^f - \boldsymbol{K}_{n,k}\boldsymbol{P}_{n,k}^{yy}\boldsymbol{K}_{n,k}^\top$
  end for
  Nudge equivalent states of $\boldsymbol{Z}_k^a$ towards their mean with γ (Equation 7)
end for

2.6. *Case studies*

The filters were tested on two systems—one simulated and one empirical. The multiscale Lorenz 96 system is a model of spatiotemporal chaos and was used to comprehensively study the performance of the filters on a range of observational settings. The empirical case study used a dataset of daily averaged meteorological variables to assess the effectiveness of the model-free filters in estimating real data.

2.6.1. *Lorenz 96*

The monoscale model of Lorenz (1996) (referred to as L96) is a set of differential equations frequently used as a model for the chaotic variability of weather and climate. The variables may represent some nondimensional atmospheric variables evaluated along equally spaced points on a latitude circle. The monoscale L96 system is defined by

$$\frac{\mathrm{d}x_n}{\mathrm{d}t} = (x_{n+1} - (1+\delta)x_{n-2})x_{n-1} - x_n + F, \quad \text{for } n = 1, \ldots, N,$$

where F is the constant forcing term which affects the irregularity of dynamics (Lorenz, 1996) and δ is an additional parameter which artificially introduces model error with $\delta = 0$ representing a perfect model scenario. This model error parameter influences the nonlinear transport properties of the system, resulting in model imperfection that is more realistic and challenging than the variation of the forcing term. Two forcing scenarios will be considered, $F = 10$ for fully developed chaotic behaviour for the majority of analysis, and $F = 8$ for the standard L96 experiment. The system has periodic boundary conditions and therefore $x_{n+N} = x_n$. In this work, the monoscale model is used within the model-based filters.

The multiscale extension of the L96 model (Lorenz, 1996) introduces a set of coupled variables, which may represent subgrid quantities that evolve with higher variability but smaller amplitude, in turn forcing the slower varying variables. This system is described by the differential equations,

$$\begin{aligned}
\frac{\mathrm{d}x_n}{\mathrm{d}t} &= (x_{n+1} - x_{n-2})x_{n-1} - x_n + F + \frac{h_x}{J}\sum_{m=1}^{M} y_{m,n} \\
\frac{\mathrm{d}y_{m,n}}{\mathrm{d}t} &= \frac{1}{\epsilon}\big((y_{m-1,n} - y_{m+2,n})y_{m+1,n} - y_{m,n} + h_y x_n\big),
\end{aligned}$$

for $n = 1, \ldots, N$ and $m = 1, \ldots, M$ where $\boldsymbol{x}$ are now referred to as the 'slow' variables, and $\boldsymbol{y}$ are referred to as the 'fast' variables. The system is selected to have $M = 10$ fast variables for each slow variable, and the number of slow variables is chosen to be $N = 40$ for visualising filtering examples and $N = 10$ when calculating performance statistics to increase computational speed. As well as the periodic boundary condition of $x_{n+N} = x_n$, the fast variables also require $y_{m,n\pm N} = y_{m,n}$ and $y_{m\pm M,n} = y_{m,n\pm 1}$. The coupling between slow and fast variables is chosen to be strong and therefore $h_x = -1$, $h_y = 1$, and $\epsilon = 0.1$. The multiscale L96 is used to generate observational data for the filter, where only the slow variables are observed. This is a typical situation when estimating the states of large chaotic systems such as the weather, where observational data of the variable of interest are available, but observations of other system variables (particularly subgrid processes or coupled systems) are not. Comparing the time scales of error growth between the atmosphere of the Earth and the Lorenz 96 system with a forcing of $F = 10$ results in the scaling that 1 time unit corresponds to 5 Earth days (Lorenz, 1996).

2.6.2. Meteorological data

The chosen empirical dataset is daily-averaged data of temperature (°C), relative humidity (%), and pressure (hPa) from the weather station at the Adelaide International Airport accessed through Weather Underground (Weather Undergound, 2024). The training set consists of 20 years of daily data from 2004 to 2023. For this example, the performance of the model-free filters is evaluated through the estimation of the daily-averaged temperature from partial and infrequent observations of humidity and pressure. Without access to the ground truth, the performance is quantified by the RMSE between the estimated states of the filters and the observational data. The governing equations of the atmosphere that depict the evolution of these three variables are very high-dimensional, depending on couplings to other weather variables and the broader spatiotemporal extent of the weather. However, it is assumed that the only data available are the estimates of these averaged quantities at a single point location, which is considerably limited in appropriately estimating the state of the weather.

The suitability of using model-free filters to estimate temperature T from relative humidity RH and pressure p is compared against two regression models. The model R1 assumes a linear relation between the daily-averaged variables and the model R2 is determined through theoretical relations between temperature, humidity and pressure. The derivation of these models is presented in Appendix B. These two regression models can be used to estimate the daily-averaged temperature from observations of the relative humidity and pressure but cannot, in this form, be used to predict their evolution between infrequent observations. Therefore, they will only be used as a comparison in the setting of daily observation.

3. Results

The performance of both model-based and model-free nonlinear dynamic filters was assessed and compared on a simulated system and an empirical system. The effects of coupling between spatially-local embeddings were investigated before implementation into the model-free filters. The simulated example of the multiscale Lorenz 96 was used to study the effectiveness of the filters in various scenarios, where the model-free filters were compared to model-based filters with varying degrees of model error. The performance of the filters was comprehensively analysed for ranges of observational setting and model error, and was averaged over many realisations to assess expected error. The empirical dataset was filtered with model-free filters to demonstrate the performance on partial, noisy (measurement and dynamic) observations in situations where an accurate model is not available.

3.1. Local attractor coupling

The nudging of equivalent states between neighbouring attractors results in a coupling between the embedded states, allowing local dynamics to reconstruct the global dynamics as proposed in Section 2. The resulting dynamics of this coupling were investigated in two ways: through the spread of equivalent state estimates in the embedding matrix, and through the properties of the global dynamics. For this analysis, the 40-dimensional L96 system was used as a spatiotemporal example. Four approaches for the coupling between neighbouring attractors were studied—predicting the global state $\boldsymbol{x}$ by the first element of the spatially-local embeddings at each grid point and reinitialising for future predictions (as used in previous studies), no coupling between states ($\gamma = 0$), nudging of shared states of neighbouring attractors to their mean ($0 < \gamma < 1$), and setting shared states to equal their mean ($\gamma = 1$). The effects of these approaches were determined through free-run simulations of the dynamics. These simulations were facilitated through the same training set of length 50 time units (250 days), corrupted by noise of variance $\sigma^2 = 0.1$, and using 30 nearest neighbours to predict embedded states.

Figure 4 presents the results of the free-run simulations using the four coupling methods. The first method, which was used in previous studies, shows instability later in the testing interval. This is due to slight errors introduced by noise which lead to large deviations from true values in grid points around x_{20}. As this method is biased towards one particular prediction of each state, this causes large growth of prediction errors that 'fan out' in time as neighboring attractors use these inaccurate states in their predictions. Initially, this method seems to be more accurate than

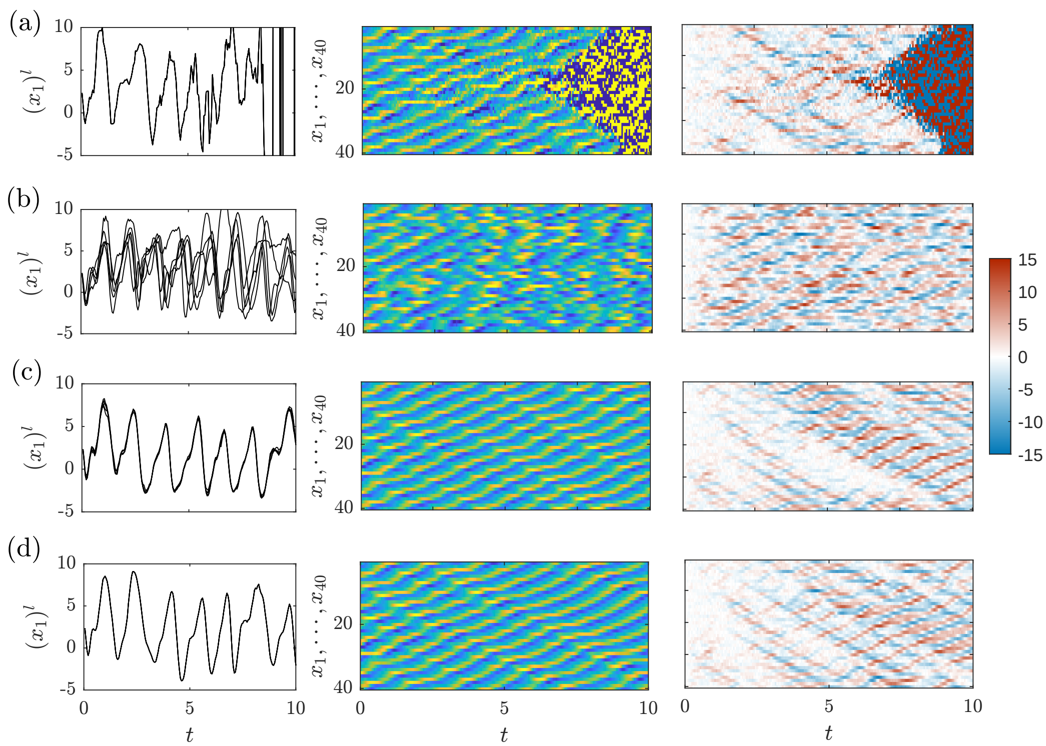

Figure 4: The effect of different coupling schemes for neighbouring local attractors on the resulting global dynamics for the L96 system. Plots on the left depict the variation of repeated entries of x_1 between neighbouring attractors, plots in the middle show the resulting global dynamics $x_1, \cdots, x_{40}$, and the plots on the right show the spatiotemporal error between the true dynamics, each for 10 time units (50 days) of the same testing interval. Coupling methods shown are (a) local dynamics to update global state, (b) uncoupled local dynamics, (c) intermediate coupling of local attractors with $\gamma = 0.5$, (d) strong coupling with $\gamma = 1$.

other approaches, presenting low error for the first time unit in the testing interval, but it was found to frequently become unstable for limited data and high noise in free-run simulations. The next approach was to let each local attractor update their state iteratively, with no coupling or reinitialisation introduced. This leads to unrealistic global dynamics as there is no coherence between neighbouring states.

Our alternative approach is to mimic the global dynamics by coupling neighbouring spatially-local embeddings rather than using these local dynamics to update each global state in a detached way. The examples shown in Figure 4 are for (c) an intermediate coupling ($\gamma = 0.5$) and (d) a strong coupling ($\gamma = 1$). For intermediate values of the coupling parameter γ, neighbouring attractors exhibit 'flocking' behaviour where equivalent states are pushed towards their mean, allowing variability but consistency among embedded states. Strong coupling between local attractors with $\gamma = 1$ results in the shared states of neighbouring attractors equaling their mean. The result is that each shared state between neighbouring attractors is equal. This attractor ensemble approach leads to greater stability of the free-run simulation compared to other methods, particularly in situations of low training data or high measurement noise, which is essential for data assimilation implementation.

Figure 5 shows the spread of prediction error for the four different methods averaged over 50 randomised testing intervals. Coupling improves the prediction, as the free-run simulation without coupling gives a large RMSE at a short time into the testing interval. The global method and the two local coupling methods give comparatively good predictions early into prediction, with the global method becoming unstable later in the interval. The local coupling approach, therefore, offers improvement in the accurate and stable prediction of the global dynamics.

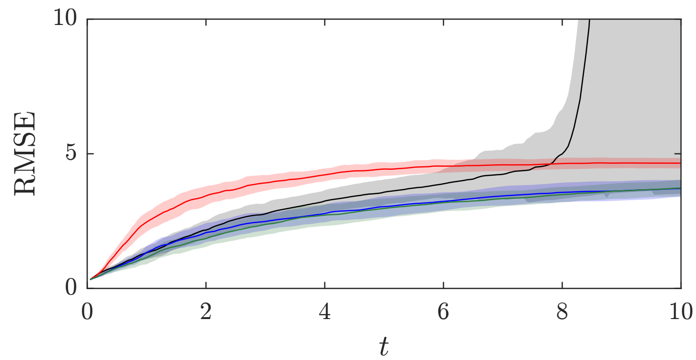

Figure 5: The average prediction error for different coupling schemes. The RMSE is shown for local dynamics to update global state (black), uncoupled local dynamics (red), intermediate coupling with $\gamma = 0.5$ (blue), and strong coupling with $\gamma = 1$ (green, overlapping with blue). The solid line shows the median RMSE with the shaded region representing the spread by the MAD.

The change in the spread of shared states in the embedded matrix for different coupling strengths is found by averaging the Median Absolute Deviation (MAD) of all shared states in the embedded matrix for different values of γ. The results are presented in Figure 6, which shows an inverse relation between the coupling parameter and state spread. For chaotic dynamical systems, it is expected that the best results will be attained through values of γ close, but not equal, to 1 as they allow shared states to 'flock' and stay in equivalent regions of the attractor while still allowing variability in prediction. Relating the nudging approach to diffusive coupling for synchronisation of chaotic systems, the smallest coupling strength needed to result in synchronisation of a set of coupled subsystems is known as the critical transition coupling (Eroglu et al., 2017). In this setting, a value of $\gamma \approx 0.5$ seems to be the critical transition coupling, yet more investigation is required to conclude more generally. In practice, this parameter must be tuned through robust methods such as cross-validation.

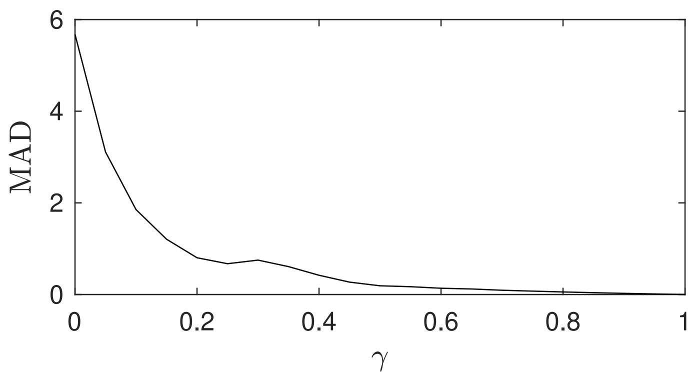

Figure 6: MAD between equivalent states in the embedding matrix of the L96 system after 10 time units for $0 \leq \gamma \leq 1$.

3.2. Lorenz 96

The embeddings for the L96 system at each grid point were constructed using $\nu = 2$ spatial neighbours and $D = 2$ delays resulting in an embedding dimension of 15. The length of time delay was calculated from minimising the average mutual information and was found to be $\tau = 0.2$ time units (1 day). For prediction on the embedded attractor, 100 nearest neighbours were used for interpolation. The nudging of repeated entries in each filter loop in the embedding matrix was facilitated with a coupling parameter of $\gamma = 0.9$. For the simulation of the Lorenz 96 system, time steps of $\Delta t = 0.05$ were used, which correspond to 6 hour time steps in the atmosphere. The amount of training data required was substantially reduced compared to other methods (500 time units or ~ 7 years in Hamilton et al. (2016) or 1000 time units or ~ 13 years in Lguensat et al. (2017)) needing only 50 time units (250 days). This reduction in training data is due to searching for analogues at different points in the domain and allowing predictions of states in neighbouring embeddings to stabilise the forecast. This approach relies on the statistically identical nature of the L96 variables and is not applicable for spatiotemporal systems in general.

The observational data given to the filters were of the slow variables $\boldsymbol{x}$ from simulations of the multiscale Lorenz 96 system with Gaussian measurement noise of variance σ^2 added. The training data used for the model-free filters were also corrupted by measurement noise of the same variance. To avoid spinup effects of the filters, their analysis states were initialised with the ground truth, and their analysis covariance matrices were initialised to the observation error covariance matrix $\boldsymbol{R}$. The robustness of data assimilation methods to poor initial ensembles is critical for their success and future work is required to establish the practicality of data-driven techniques in these settings. Initial investigation demonstrated good performance of the model-free filters when given inaccurate initial conditions. The ensemble size for both model-based and model-free EnKF was selected as 100, and the model-based filter incorporated covariance inflation with $\mu = 1.02$. The model error covariance matrix $\boldsymbol{Q}$ was set to $0.5\boldsymbol{I}$ for both model-based and model-free filters as found through initial experimentation. The observation error covariance matrix $\boldsymbol{R}$ was set to $\sigma^2\boldsymbol{I}$. For the UnKF, the parameter values were selected to be $\alpha = 10^{-3}$ for $N = 10$ or $\alpha = 10^{-2}$ for $N = 40$ for the model-based filter, $\alpha = 1$ for the model-free filter, and $\beta = 2$ and $\kappa = 0$ for both filters. For a fair comparison between the model-based and model-free filters, localisation is introduced to all model-based filters using the Gaspari-Cohn function (Gaspari and Cohn, 1999) with a radius of 5.

The model-based ExKF, EnKF, and UnKF using the monoscale L96 model were compared to their model-free counterparts in settings of noisy observations, infrequent observations, and sparse-in-space observations. The performance of the filters on noisy data was tested on observational data of increasing variance from $\sigma^2 = 0.1, 0.2, 0.5, 1, 2, 5$. The noise level was set to $\sigma^2 = 0.1$ for the remaining analysis. The performance on infrequent observations, or sparse-in-time observations, was investigated by increasing the time between observations between $T = 0.1, 0.2, 0.5, 1$ time units (12 hours, 1 day, 2.5 days, and 5 days). The effect of sparse spatial observations was studied by varying the number of grid points observed from every 2nd, 5th, and 10th grid point (represented by $S = 2, 5, 10$). The effect of model error on the model-based filters was assessed through the introduction of model imperfection through the parameter δ in the monoscale L96 model. This parameter was varied between $\delta = 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1$ to simulate increasingly severe model error.

Figures 7 and 8 show examples of extreme cases of spatially sparse or infrequent observations for the L96 system with $N = 40$. These examples highlight how the model-free filters outperform the model-based filters in highly sparse observational settings, which is studied in more detail in the following results. Figures 9 to 11

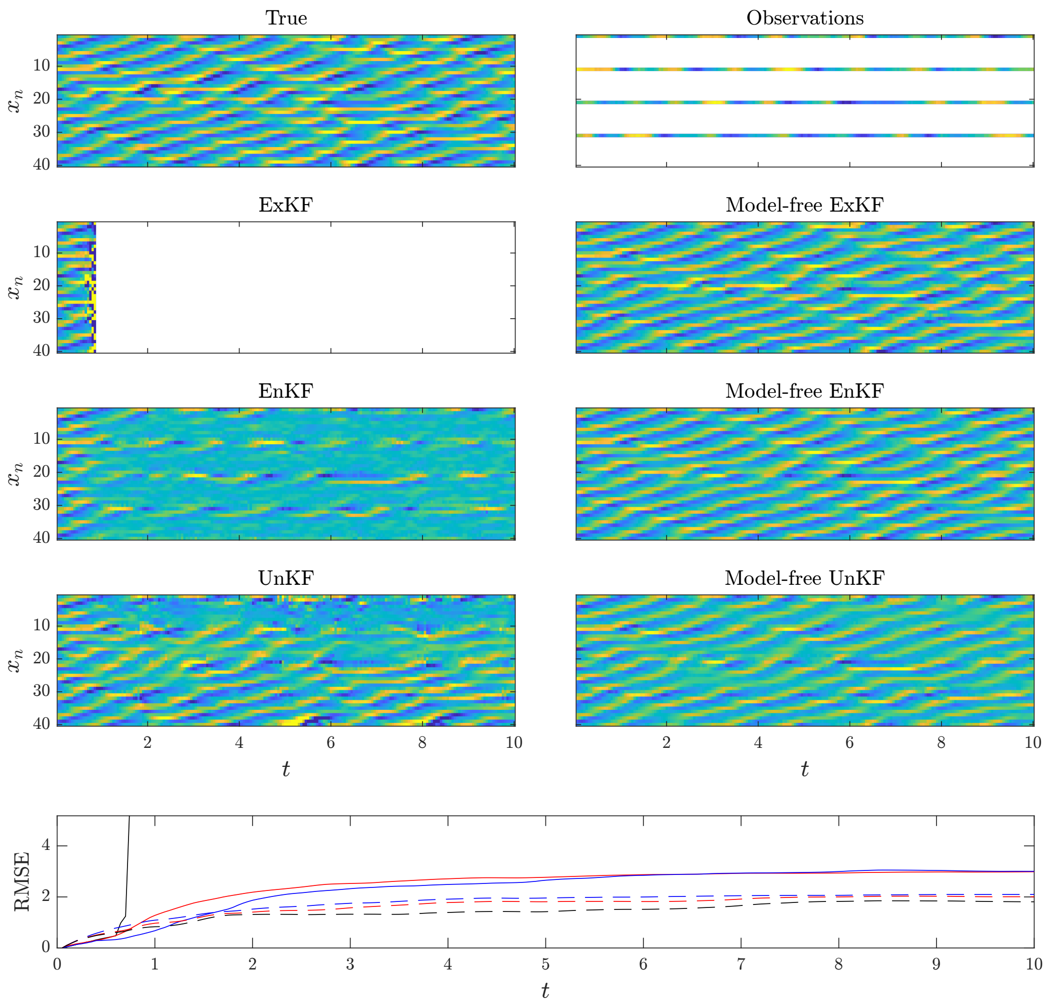

Figure 7: Filtering example for L96 with $N = 40$ with every 10th grid point observed. Error over time is depicted by the cumulative RMSE of ExKF (black), EnKF (red), and UnKF (blue), for model-based (solid line) and model-free (dashed line) filters.

present the performance of the different filtering methods when observational data were missing in various ways and subject to varying levels of measurement noise. The analysis of the expected performance of each filter was facilitated in a reduced domain of $N = 10$, with the median RMSE at the final time (5 time units or 25 days) representing the averaged error over 50 realisations.

Figure 9 presents the performance comparison between the model-based and model-free filters on the chaotic L96 system for increasing levels of measurement

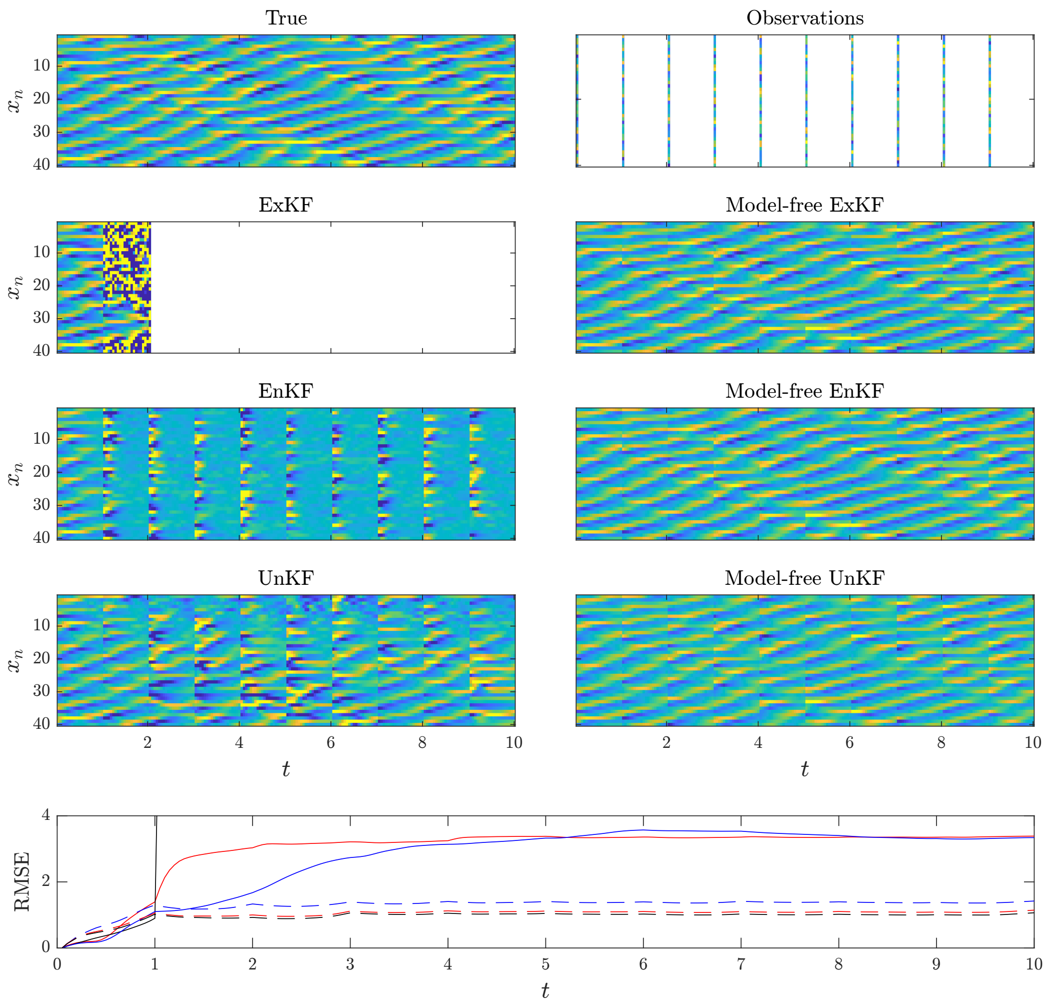

Figure 8: Filtering example for L96 with $N = 40$ with observations every 20th time step. Error over time is depicted by the cumulative RMSE of ExKF (black), EnKF (red), and UnKF (blue), for model-based (solid line) and model-free (dashed line) filters.

noise. The ExKF and UnKF performed similarly across the model-based and model-free versions and across noise levels, with the expected trends of higher noise and more significant model error both decreasing predictive skill. The model-based EnKF presented the most robust state estimates, offering high predictive accuracy across various noise levels for low model error. In this setting, where the model-free filters were trained on observations as noisy as those in the testing interval, they achieved surprisingly similar accuracy to the model-based filters. The model-free filters became preferable to the model-based filters when substantial model error was introduced.

The results for the tests on infrequent observations are shown in Figure 10. While all filters estimated the system states similarly for moderately infrequent observations, for settings where observations are unavailable for longer than 0.5 time units (2.5 days) the model-based filters often exhibited high error or instability. For the model-based filters, infrequent observations paired with the high nonlinearity of the L96 dynamics caused issues for the tangent linear approximation and sample point approximation of the state distribution. For example, the model-based EnKF started underestimating the system state due to the time between updates being large enough for ensemble members to travel to different regions of the attractor, known as ensemble blowup, resulting in high estimation error (as seen in Figure 8). The model-free EnKF, however, was not limited by this issue due to the significant coupling between embedded attractors enabling a 'flocking' of ensemble members.

The performance of the filters for observations with partial observation in space is shown in Figure 11. The nonlinear Kalman filters can effectively estimate and predict unobserved variables of a chaotic system when sufficient measurements of some variables are possible. This experiment on the L96 sought to investigate when the number of variables of a large chaotic spatiotemporal system is insufficient to estimate the state of the system effectively. The model-based ExKF became unstable when only one variable was measured, and performed poorly when the variation of every 5th grid point was available. The model-free ExKF performed much better, offering accurate estimates even when only one variable was measured. The model-based EnKF and UnKF offered stable estimates for all levels of sparsity when the model had low error, with high sparsity settings resulting in inaccurate estimation. The model-free filters substantially outperformed the model-based filters in this example, providing accurate state estimates even in extreme settings of data sparsity.

As an additional test of sparse observations that provides a more realistic example, a measurement setting that mimics the behavior of satellite-derived observations was tested. In this test case, it is assumed that observations are available at a subset of spatial grid points intermittently in time along spatial bands, resembling the incomplete spatial coverage provided by satellite observations. Figure 12 shows the results of this example, which correspond similarly to the spatially sparse tests. The model-free filters outperformed the model-based filters in providing accurate state estimation between spatially sparse observations.

The final experiment tested the long-term robustness of the model-free filters

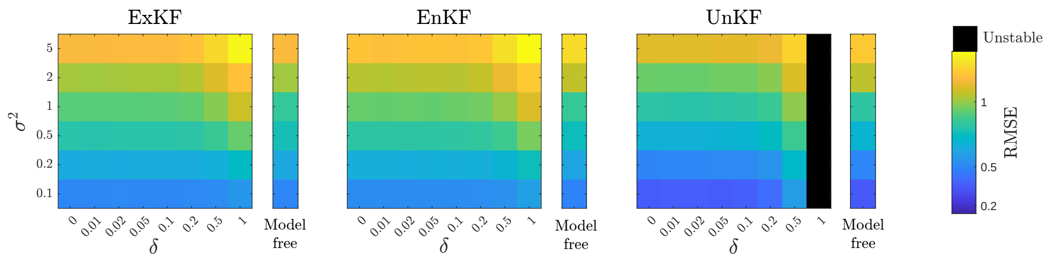

Figure 9: Performance of perfect and imperfect model-based filters and model-free filters on the L96 system for varying levels of measurement noise σ^2 and model error δ.

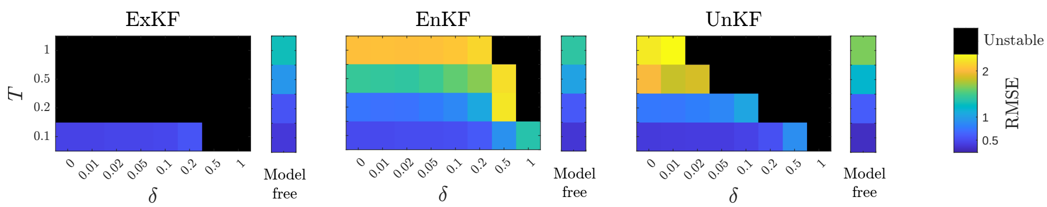

Figure 10: Performance of perfect and imperfect model-based filters on the L96 system and model-free filters for varying time between observations T and model error δ.

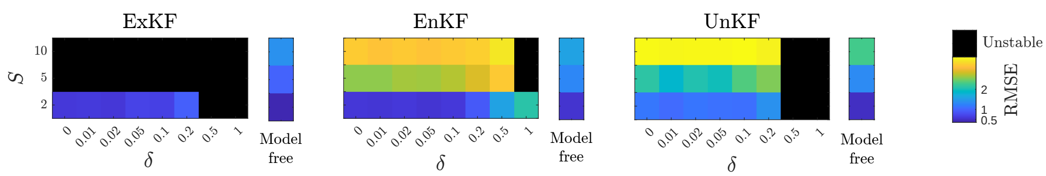

Figure 11: Performance of perfect and imperfect model-based filters and model-free filters on the L96 system for observations of every S-th grid point and model error δ.

by comparing the EnKF filters on the 'standard' L96 experimental configuration as outlined in Maclean and Van Vleck (2025). This experiment used the single scale L96 of dimension $N = 40$, forcing $F = 8$, simulation time step of 0.01, with full observation every 0.05 time units corrupted with measurement noise of variance $\sigma^2 = 1$. The model-based filter had access to the perfect model and therefore the process covariance matrix was set to $\boldsymbol{Q} = 0$. The model-free filter was trained with 50 time units of data, and since the data was corrupted by measurement noise, we set $\boldsymbol{Q} = 0.5\boldsymbol{I}$. A period of 1000 assimilation steps was computed and discarded, after which the RMSE and mean ensemble spread were measured for a further 10,000 steps. Figure 13 presents the moving average (over 20 observations) of these results. This experiment demonstrates that the model-free variant of

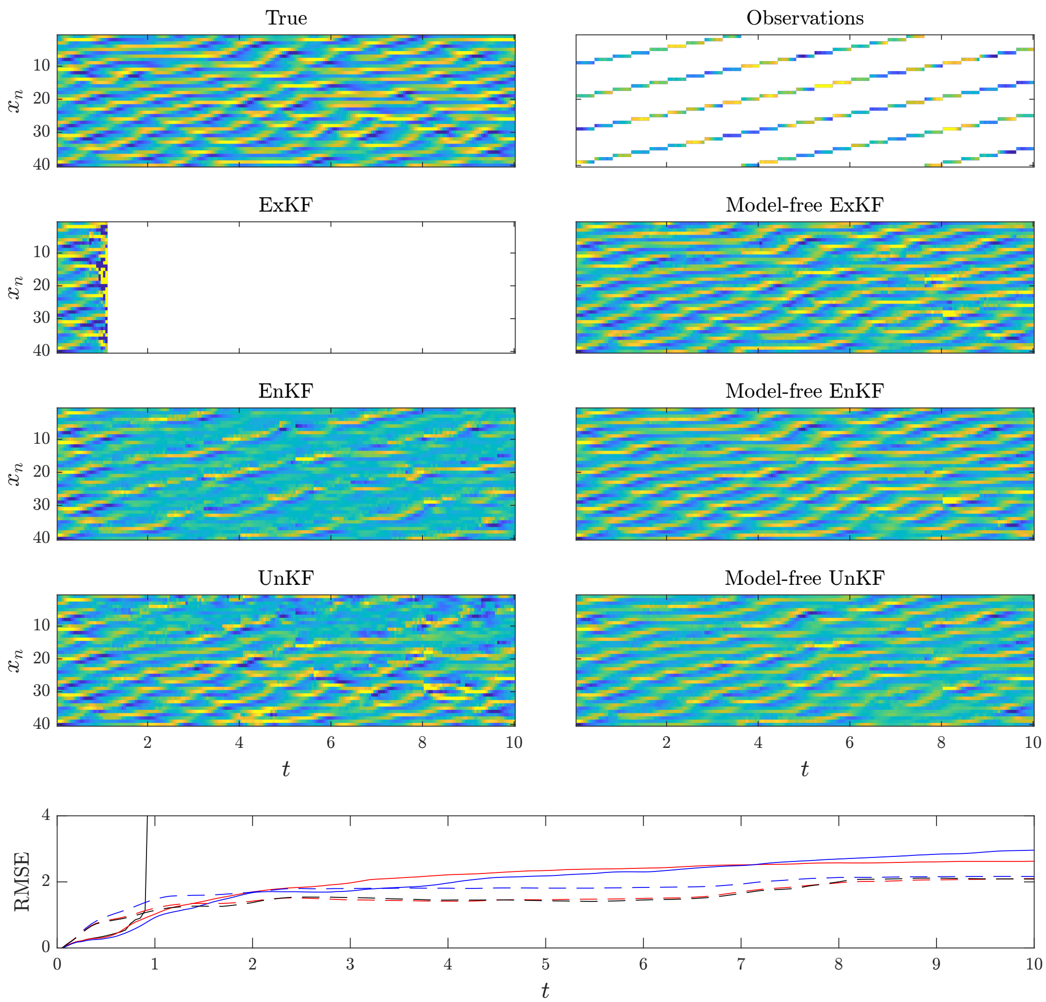

Figure 12: Filtering example for L96 with $N = 40$ with satellite-like observations. Error over time is depicted by the cumulative RMSE of ExKF (black), EnKF (red), and UnKF (blue), for model-based (solid line) and model-free (dashed line) filters.

the EnKF is robust in long-term assimilation and consistently reduces the mean observation error of the incoming observations by approximately 60%.

3.3. Meteorological data

For the empirical example of filtering meteorological observations, a training dataset of length 10 years with the following year being a spinup period, followed by a year for the testing set was randomly chosen from the 20 year batch of data.

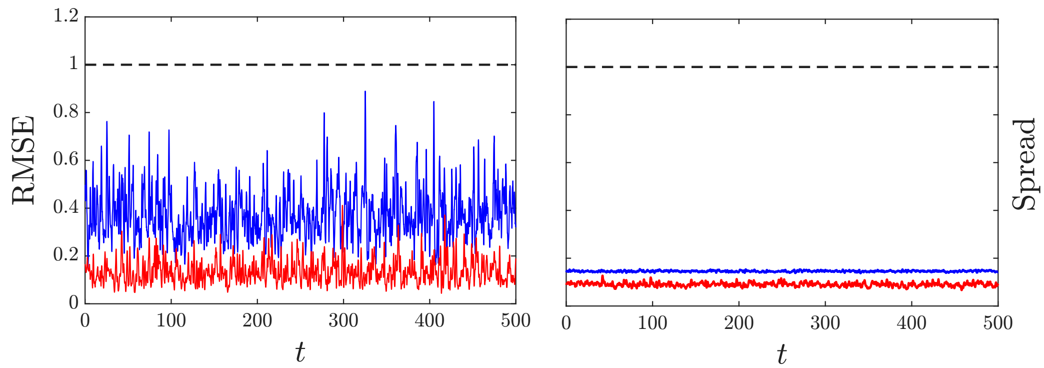

Figure 13: Long-time experiment of model-free (blue) and model-based (red) Ensemble Kalman filters. The moving average over 20 observations of the RMSE (left) and ensemble spread (right) of both filters are compared with the mean observation error (black solid line).

A time delay of length $\tau = 1$ day was used to represent the high variability of the data, with parameters of $D = 2$ and $\nu = 1$ chosen to couple the available variables sufficiently. The prediction of the variables used 100 of the nearest neighbours for interpolation. The ensemble filter used 100 ensemble members with no inflation, and the unscented filter used parameter values of $\alpha = 1, \beta = 2, \kappa = 0$. The coupling of spatially-local embeddings was chosen to be strong with a coupling parameter of $\gamma = 1$ selected to ensure the stability of the filters. In this case, the adaptive estimation of $\boldsymbol{Q}$ and $\boldsymbol{R}$ was facilitated by the covariance matching approach laid out by Akhlaghi et al. (2017), with distinct forgetting factors chosen to be $\alpha_Q = 0.8$ and $\alpha_R = 0.5$. It was observed that the daily averaged temperature exhibited a higher variance in the summer than in the winter, and therefore, the offline estimates were parameterised by day of the year for use in the partially observed setting.

The accuracy of the filter estimates was compared across varying data infrequency from observations every day to every 7 days. The results were averaged over 20 realisations on randomised datasets to provide the median and MAD of the expected error. The regression models R1 and R2 cannot provide estimates between observations in an infrequent data setting, as they do not have forecasting capability. The dynamics as reconstructed through the embedded attractors, however, are not limited by infrequent observations and may be investigated. Figure 14 shows the results of the expected error spread for the model-free filters across different levels of data infrequency and the regression models at daily observations.

The regression model informed by theoretical atmospheric relations (R2) performed slightly better than the linear model. This minor improvement emphasises the idea that models based on linear assumptions are not well-suited for tasks of estimating nonlinearly evolving variables. The model-free filters yielded slightly larger median RMSE values for daily observations as they slightly underestimated the peaks and troughs of the temperature variation, whereas the regression model often overestimated. This is a limitation in comparing the RMSE between the estimated states and observations in the absence of ground truth. The UnKF performed the worst of the model-free filters, showing a spike in RMSE when observations became sparse in time. The ExKF and EnKF performed well in estimating and predicting the temperature from infrequent observations of humidity and pressure, presenting RMSE values comparative to the daily observation setting up to measurements every three days.

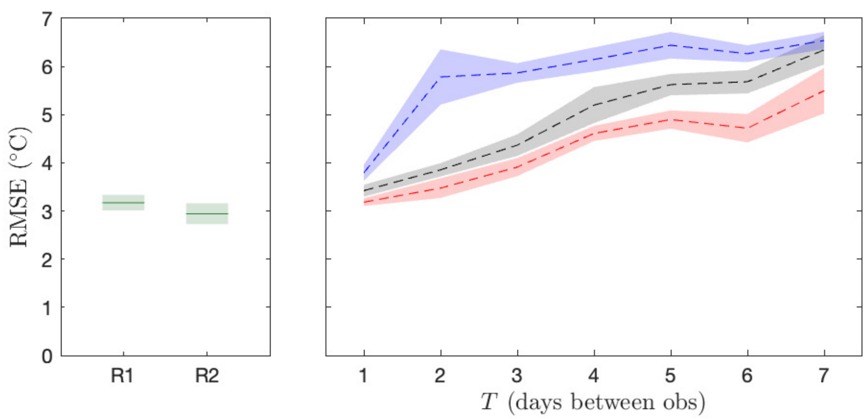

Figure 14: Performance of model-free filters on estimating temperature from observations of relative humidity and pressure at different levels of observation infrequency measured by median RMSE of temperature for ExKF (black), EnKF (red), and UnKF (blue). The performance of regression models R1 and R2 (green) are shown by the median RMSE of temperature for daily observations (equivalent to $T = 1$). MAD is shown by the shaded regions.

Figure 15 shows an example of the model-free filters estimating the daily averaged temperature over a year with observations of daily averaged humidity and pressure every three days. The ExKF and EnKF estimate the temperature surprisingly well, even between observations of humidity and pressure, indicating that the reconstructed dynamics possess accurate forecast capability. The peaks and troughs of the unavailable, measured temperature time series curve are well estimated by these filters, and in situations of poor estimation of these high-frequency dynamics, the overall trend is still well captured.

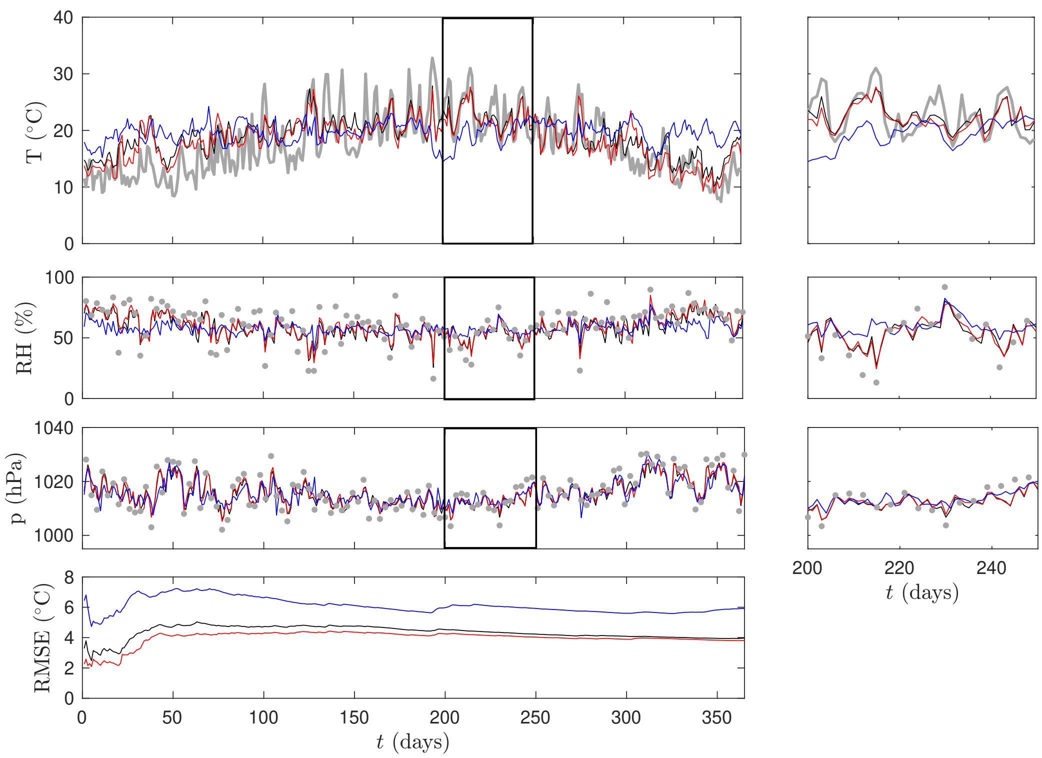

Figure 15: Performance of model-free filters on estimating temperature from observations of relative humidity and pressure every three days. The curves shown are the unavailable time series of temperature (grey solid line), the observations provided to the filters (grey dots), ExKF (black), EnKF (red), and UnKF (blue). Final RMSE value for model-free filters is 3.97 (ExKF), 3.81 (EnKF), 5.92 (UnKF).

4. Discussion

This paper presented the development and analysis of model-free dynamic filters for data assimilation in the absence of an accurate model and a general way of constructing data-driven filters using delay embeddings. Three nonlinear adaptations of the Kalman filter were introduced, and the considerations for extending them into model-free settings were discussed. The model-free filters were compared against their model-based counterparts on a simulated system and an empirical system. The effects of measurement noise, partial observation in space and time, and model error were comprehensively studied for each filter.

Rather than considering the Kalman updates in the global state space as in previous studies, we considered updating the state of embedded attractors at each global grid point in their embedded space. This allows the calculation of the embedded tangent linear operator for the ExKF but also introduces additional considerations for reconstructing global dynamics from local embeddings. The main difference between prior investigations into model-free Kalman filtering is that in this work, the global dynamics are reconstructed through the coupling of local attractors, which offers improved stability, accuracy, and data-efficiency in settings of high noise and low data. An interesting direction for future research is assessing how well the emerging global dynamics from the coupled local dynamics represent the true dynamics of the spatiotemporal system and how well they scale with dimension. Comparing the effects of covariance approximation using explicit localisation in the model-based filters with the implicit localisation in the model-free setting may also shed light on effective localisation strategies.

The effects of measurement noise on the performance of state estimation were shown to be similar for the model-based and model-free filters, a surprising result considering the training data for the model-free methods was equally noisy. This may be due to the catalog of observational data of the same noise level being better suited for updating predictions in these noisy regimes. The experiments concerning sparsity in space and time highlighted the utility of model-free filtering in situations of partial observation, showing considerable improvement over the model-based methods particularly when compared to filters using imperfect models. Our presented model-free filters mitigate ensemble blowup through the resulting reduced dimension and flocking nature of embedded trajectories facilitated through the coupling of states between neighbours. This allows the model-free adaptations of the filter to outperform the model-based filters in several tested scenarios of sparse observations. In settings where models do not exist, such as the meteorological example, the model-free filters show utility in describing the nonlinear evolution of dynamics in substantially-partial observed settings.

Despite the improvement in estimation shown for model-free filters in certain observational settings, there are several limitations of the proposed methods. The reconstruction of dynamics from observations requires training data for all variables intended to be estimated, and therefore filtering of unobserved variables or spatial locations is impossible without access to complete data. One approach would be to include model output data that provides complete spatial coverage within the training dataset for the model-free filters. This would present several solutions for applications where observations are limited: supplementing the observational data with imperfect model data, hybrid coupling of data-driven and model-based schemes, or using a suite of models to generate data for use by the filters. This would be particularly important for systems that exhibit extreme events or offer substantially sparse observations.

The presented analysis shows the promise of model-free methods in the data assimilation of chaotic systems where accurate models are not available. Applications of these techniques may aid in state estimation for high-dimensional, nonlinear, and partially observed systems where accurate models are unavailable. These methods, enabled through the Takens Embedding Theorem, offer interpretability over other data-driven techniques for prediction, allowing informed and improved state estimation in settings of spatiotemporal chaos.

CRediT authorship contribution

SPM: Conceptualization, Software, Writing – original draft. **WSPR:** Supervision, Writing – review and editing. **JM** Supervision, Writing – review and editing. **SB:** Supervision, Writing – review and editing.

Data availability

The developed data-driven filtering code used to produce the figures in this research is available at `github.com/dankesean/model-free-filtering`.

Acknowledgements

SPM acknowledges with thanks support from Australian Government Research Training Program Scholarships. SB acknowledges partial funding from the Australian Research Council (grant DP200101764).

Appendix A. Nonlinear Kalman filter algorithms

The standard model-based nonlinear Kalman filters are presented in Algorithms 4, 5, and 6. In Algorithm 5, the vectors $\boldsymbol{x}_k^{a,i}$ and $\boldsymbol{x}_k^{a,i}$ for $i = 1, \ldots, N_{\text{ens}}$ are placed in respective matrices $\boldsymbol{X}_k^a$ and $\boldsymbol{X}_k^f$.

Algorithm 4 Extended Kalman filter

Choose $\boldsymbol{Q}_k$, $\boldsymbol{R}_k$, and localisation matrix ρ
Initialise $\boldsymbol{x}_1^a$, $\boldsymbol{P}_1^a$
for $k = 2$ to K **do**
  $\boldsymbol{x}_k^f = \mathcal{M}_k(\boldsymbol{x}_{k-1}^a)$
  $\boldsymbol{P}_k^f = \boldsymbol{M}_k \boldsymbol{P}_{k-1}^a \boldsymbol{M}_k^\mathsf{T} + \boldsymbol{Q}_k$
  $\boldsymbol{K}_k = (\rho \circ (\boldsymbol{P}_k^f \boldsymbol{H}_k^\mathsf{T}))(\boldsymbol{H}_k \boldsymbol{P}_k^f \boldsymbol{H}_k^\mathsf{T} + \boldsymbol{R}_k)^{-1}$
  $\boldsymbol{x}_k^a = \boldsymbol{x}_k^f + \boldsymbol{K}_k(\boldsymbol{y}_k - \boldsymbol{H}_k \boldsymbol{x}_k^f)$
  $\boldsymbol{P}_k^a = (\boldsymbol{I} - \boldsymbol{K}_k \boldsymbol{H}_k)\boldsymbol{P}_k^f$
end for

Algorithm 5 Ensemble Kalman filter

Choose N_{ens}, inflation parameter μ, $\boldsymbol{Q}_k$, $\boldsymbol{R}_k$, and localisation matrix ρ
Initialise $\boldsymbol{X}_1^a$
for $k = 2$ to K **do**
  for $i = 1$ to N_{ens} **do**
    $\boldsymbol{x}_k^{f,i} = \mathcal{M}_k(\boldsymbol{x}_{k-1}^{a,i}) + \boldsymbol{w}_k$ where $\boldsymbol{w}_k \sim \mathcal{N}(0, \boldsymbol{Q}_k)$
  end for
  $\overline{\boldsymbol{x}}_k^f = \frac{1}{N_{\text{ens}}} \sum_{i=1}^{N_{\text{ens}}} \boldsymbol{x}_k^{f,i}$
  $\boldsymbol{P}_k^f = \frac{1}{N_{\text{ens}}-1}(\boldsymbol{X}_k^f - \overline{\boldsymbol{x}}_k^f)(\boldsymbol{X}_k^f - \overline{\boldsymbol{x}}_k^f)^\mathsf{T}$
  $\boldsymbol{K}_k = (\rho \circ (\boldsymbol{P}_k^f \boldsymbol{H}_k^\mathsf{T}))(\boldsymbol{H}_k \boldsymbol{P}_k^f \boldsymbol{H}_k^\mathsf{T} + \boldsymbol{R}_k)^{-1}$
  for $i = 1$ to N_{ens} **do**
    $\boldsymbol{x}_k^{a,i} = \boldsymbol{x}_k^{f,i} + \boldsymbol{K}_k(\boldsymbol{y}_k + \boldsymbol{v}_k - \boldsymbol{H}_k \boldsymbol{x}_k^{f,i})$ where $\boldsymbol{v}_k \sim \mathcal{N}(0, \boldsymbol{R}_k)$
  end for
  $\overline{\boldsymbol{x}}_k^a = \frac{1}{N_{\text{ens}}} \sum_{i=1}^{N_{\text{ens}}} \boldsymbol{x}_k^{a,i}$
  $\boldsymbol{x}_k^{a,i} = \overline{\boldsymbol{x}}_k^a + \mu(\overline{\boldsymbol{x}}_k^a - \boldsymbol{x}_k^{a,i})$
end for

Algorithm 6 Unscented Kalman filter

Choose α, β, κ, $\boldsymbol{Q}_k$, $\boldsymbol{R}_k$, and localisation matrix ρ
Calculate λ, $\boldsymbol{W}^{m,i}$, $\boldsymbol{W}^{c,i}$
Initialise $\boldsymbol{x}_1^a$, $\boldsymbol{P}_1^a$
for $k = 2$ to K **do**
    for $i = 1$ to $2N+1$ **do**
        $\boldsymbol{\chi}_{k-1}^{a,i} = \left[\boldsymbol{x}_{k-1}^a, \boldsymbol{x}_{k-1}^a \pm \sqrt{(N+\lambda)\boldsymbol{P}_{k-1}^a}\,\right]$
        $\boldsymbol{\chi}_k^{f,i} = \mathcal{M}_k(\boldsymbol{\chi}_{k-1}^{a,i})$
    end for
    $\overline{\boldsymbol{x}}_k^f = \sum_{i=1}^{2N+1} \boldsymbol{W}^{m,i}\boldsymbol{\chi}_k^{f,i}$
    $\boldsymbol{P}_k^f = \sum_{i=1}^{2N+1} \boldsymbol{W}^{c,i}(\boldsymbol{\chi}_k^{f,i} - \overline{\boldsymbol{x}}_k^f)(\boldsymbol{\chi}_k^{f,i} - \overline{\boldsymbol{x}}_k^f)^{\mathsf{T}} + \boldsymbol{Q}_k$
    $\boldsymbol{P}_k^{yy} = \sum_{i=1}^{2N+1} \boldsymbol{W}^{c,i}(\boldsymbol{\chi}_k^{f,i} - \boldsymbol{H}_k\overline{\boldsymbol{x}}_k^f)(\boldsymbol{\chi}_k^{f,i} - \boldsymbol{H}_k\overline{\boldsymbol{x}}_k^f)^{\mathsf{T}} + \boldsymbol{R}_k$
    $\boldsymbol{P}_k^{xy} = \sum_{i=1}^{2N+1} \boldsymbol{W}^{c,i}(\boldsymbol{\chi}_k^{f,i} - \overline{\boldsymbol{x}}_k^f)(\boldsymbol{\chi}_k^{f,i} - \boldsymbol{H}_k\overline{\boldsymbol{x}}_k^f)^{\mathsf{T}}$
    $\boldsymbol{K}_k = (\rho \circ \boldsymbol{P}_k^{xy})(\boldsymbol{P}_k^{yy})^{-1}$
    $\boldsymbol{x}_k^a = \boldsymbol{x}_k^f + \boldsymbol{K}_k(\boldsymbol{y}_k - \boldsymbol{H}_k\boldsymbol{x}_k^{f,i})$
    $\boldsymbol{P}_k^a = \boldsymbol{P}_k^f - \boldsymbol{K}_k\boldsymbol{P}_k^{yy}\boldsymbol{K}_k^{\mathsf{T}}$
end for

Appendix B. Meteorological regression models

The standard multiple linear regression model

$$T = \beta_0 + \beta_1 RH + \beta_2 p,$$

is tested as a simple model that assumes a linear relation between the daily-averaged variables, this is referred to as R1. The second regression model is determined through theoretical relations between temperature, humidity and pressure.

For moist air, relative humidity is the amount of water in the air relative to the amount of water the air can hold and is given by

$$RH = \frac{r}{r_w}, \tag{B.1}$$

where r is the mixing ratio of water to air. The saturation mixing ratio r_w may be given by its relationship with pressure p and the equilibrium water vapor pressure e_w,

$$r_w = \frac{1}{\epsilon}\left(\frac{e_w}{p}\right), \tag{B.2}$$

where $\epsilon \approx 0.622$ (Trenberth, 1992). The equilibrium water vapor pressure e_w (in hPa) is related to temperature T (in Kelvin) by the Clausius-Clapeyron equation (Trenberth, 1992)

$$\log e_w \approx 9.4 - \frac{2.35 \times 10^3}{T}. \tag{B.3}$$

Combining Equations (B.1), (B.2), and (B.3), and assuming r is constant (no evaporation or condensation) we get a relation of the form

$$\frac{1}{T} = \beta_0 + \beta_1 \log RH + \beta_2 \log p,$$

that is referred to as R2. Multiple linear regression can then be used to fit these parameters, from the same training dataset used to train the model-free filters.

References

Akhlaghi, S., Zhou, N., Huang, Z., 2017. Adaptive adjustment of noise covariance in Kalman filter for dynamic state estimation, in: 2017 IEEE Power & Energy Society General Meeting, IEEE. pp. 1–5.

Asch, M., Bocquet, M., Nodet, M., 2016. Data Assimilation: Methods, Algorithms, and Applications. SIAM.

Berner, J., Achatz, U., Batté, L., Bengtsson, L., Cámara, A.D.L., Christensen, H.M., Colangeli, M., Coleman, D.R., Crommelin, D., Dolaptchiev, S.I., Franzke, C.L., Friederichs, P., Imkeller, P., Järvinen, H., Juricke, S., Kitsios, V., Lott, F., Lucarini, V., Mahajaajaajan, S., Palmer, T.N., Penland, C., Sakradzijaja, M., Storch, J.S.V., Weisheimer, A., Weniger, M., Williams, P.D., Yano, J.I., 2017. Stochastic parameterization toward a new view of weather and climate models. Bulletin of the American Meteorological Society 98, 565–587. doi:10.1175/BAMS-D-15-00268.1.

Bocquet, M., Farchi, A., Finn, T.S., Durand, C., Cheng, S., Chen, Y., Pasmans, I., Carrassi, A., 2024. Accurate deep learning-based filtering for chaotic dynamics by identifying instabilities without an ensemble. Chaos: An Interdisciplinary Journal of Nonlinear Science 34.

Brunton, S.L., Proctor, J.L., Kutz, J.N., 2016. Sparse identification of nonlinear dynamics with control (sindyc). IFAC-PapersOnLine 49, 710–715.

Eroglu, D., Lamb, J.S., Pereira, T., 2017. Synchronisation of chaos and its applications. Contemporary Physics 58, 207–243.

Evensen, G., 2009. The ensemble Kalman filter for combined state and parameter estimation. IEEE Control Systems Magazine 29, 83–104.

Falconer, S.A., Lloyd, D.J., Santitissadeekorn, N., 2023. Combining dynamic mode decomposition with ensemble Kalman filtering for tracking and forecasting. Physica D: Nonlinear Phenomena 449, 133741.

Farchi, A., Laloyaux, P., Bonavita, M., Bocquet, M., 2021. Using machine learning to correct model error in data assimilation and forecast applications. Quarterly Journal of the Royal Meteorological Society 147, 3067–3084.

Fraser, A.M., Swinney, H.L., 1986. Independent coordinates for strange attractors from mutual information. Physical review A 33, 1134.

Gaspari, G., Cohn, S.E., 1999. Construction of correlation functions in two and three dimensions. Quarterly Journal of the Royal Meteorological Society 125, 723–757.

Gottwald, G.A., Reich, S., 2021. Supervised learning from noisy observations: Combining machine-learning techniques with data assimilation. Physica D: Nonlinear Phenomena 423, 132911.

Hale, J.K., 1997. Diffusive coupling, dissipation, and synchronization. Journal of Dynamics and Differential Equations 9, 1–52.

Hamilton, F., Berry, T., Sauer, T., 2016. Ensemble Kalman filtering without a model. Physical Review X 6, 1–12. doi:10.1103/PhysRevX.6.011021.

Harlim, J., 2017. Model error in data assimilation, in: Nonlinear and Stochastic Climate Dynamics. Cambridge University Press, pp. 276–317.

Isensee, J., Datseris, G., Parlitz, U., 2020. Predicting spatio-temporal time series using dimension reduced local states. Journal of Nonlinear Science 30, 713–735. doi:10.1007/s00332-019-09588-7.

Julier, S.J., Uhlmann, J.K., 1997. New extension of the Kalman filter to nonlinear systems, in: Signal Processing, Sensor Fusion, and Target Recognition VI, Spie. pp. 182–193.

Kalman, R.E., Bucy, R.S., 1961. New results in linear filtering and prediction theory. Journal of Basic Engineering 83, 95–108.

Kennel, M.B., Brown, R., Abarbanel, H.D., 1992. Determining embedding dimension for phase-space reconstruction using a geometrical construction. Physical review A 45, 3403.

Lguensat, R., Tandeo, P., Ailliot, P., Pulido, M., Fablet, R., 2017. The analog data assimilation. Monthly Weather Review 145, 4093–4107. doi:10.1175/MWR-D-16-0441.1.

Lorenz, E.N., 1996. Predictability: A problem partly solved, in: Proc. Seminar on Predictability, Reading.

Maclean, J., Van Vleck, E.S., 2025. Decomposition of likelihoods and techniques for multi-scale data assimilation. Discrete and Continuous Dynamical Systems-B 30, 2159–2182.

Penny, S.G., Smith, T.A., Chen, T.C., Platt, J.A., Lin, H.Y., Goodliff, M., Abarbanel, H.D., 2022. Integrating recurrent neural networks with data assimilation for scalable data-driven state estimation. Journal of Advances in Modeling Earth Systems 14, e2021MS002843.

Revach, G., Shlezinger, N., Ni, X., Escoriza, A.L., Van Sloun, R.J., Eldar, Y.C., 2022. KalmanNet: Neural network aided Kalman filtering for partially known dynamics. IEEE Transactions on Signal Processing 70, 1532–1547.

Rosafalco, L., Conti, P., Manzoni, A., Mariani, S., Frangi, A., 2024. EKF-SINDy: Empowering the extended Kalman filter with sparse identification of nonlinear dynamics. arXiv preprint arXiv:2404.07536 .

Takens, F., 1981. Detecting strange attractors in turbulence, in: Dynamical Systems and Turbulence, Warwick 1980: proceedings of a symposium held at the University of Warwick 1979/80, Springer. pp. 366–381.

Tomizawa, F., Sawada, Y., 2021. Combining ensemble Kalman filter and reservoir computing to predict spatiotemporal chaotic systems from imperfect observations and models. Geoscientific Model Development 14, 5623–5635.

Toth, Z., Peña, M., 2007. Data assimilation and numerical forecasting with imperfect models: The mapping paradigm. Physica D: Nonlinear Phenomena 230, 146–158.

Trenberth, K.E., 1992. Climate system modeling. Cambridge University Press.

Wan, E.A., Van Der Merwe, R., 2000. The unscented Kalman filter for nonlinear estimation, in: Proceedings of the IEEE 2000 Adaptive Systems for Signal Processing, Communications, and Control Symposium, pp. 153–158.

Wang, Z., Jiang, L., 2025. Model free data assimilation with takens embedding. Journal of Computational and Applied Mathematics 460, 116399.

Weather Undergound, 2024. Adelaide, South Australia, Australia Weather History. `https://www.wunderground.com/history/monthly/au/adelaide/YPAD`.