Multiplicative non-gaussian model error estimation in data assimilation

JAMES | Journal of Advances in Modeling Earth Systems®

RESEARCH ARTICLE
10.1029/2021MS002564

Multiplicative Non-Gaussian Model Error Estimation in Data Assimilation

S. Pathiraja[1,2] **and P. J. van Leeuwen**[3,4]

[1]Institut für Mathematik, Universität Potsdam, Potsdam-Golm, Germany, [2]Now at School of Mathematics and Statistics, University of New South Wales, Sydney, NSW, Australia, [3]Department of Atmospheric Science, Colorado State University, Fort Collins, CO, USA, [4]Department of Meteorology, University of Reading, Reading, UK

Key Points:

- We address the challenging problem of model uncertainty quantification for data assimilation. The focus is on complex non-Gaussian errors
- Many existing methods for quantifying error due to sub-grid scale processes require full observations or knowledge of sub-grid scale process
- We propose a data-driven method that is capable of recovering complex error structures using only partial observations of resolved variables

Correspondence to:
S. Pathiraja,
s.pathiraja@unsw.edu.au

Citation:
Pathiraja, S., & van Leeuwen, P. J. (2022). Multiplicative non-Gaussian model error estimation in data assimilation. *Journal of Advances in Modeling Earth Systems*, *14*, e2021MS002564. https://doi.org/10.1029/2021MS002564

Received 10 APR 2021
Accepted 30 DEC 2021

Abstract Model uncertainty quantification is an essential component of effective data assimilation. Model errors associated with sub-grid scale processes are often represented through stochastic parameterizations of the unresolved process. Many existing Stochastic Parameterization schemes are only applicable when knowledge of the true sub-grid scale process or full observations of the coarse scale process are available, which is typically not the case in real applications. We present a methodology for estimating the statistics of sub-grid scale processes for the more realistic case that only partial observations of the coarse scale process are available. Model error realizations are estimated over a training period by minimizing their conditional sum of squared deviations given some informative covariates (e.g., state of the system), constrained by available observations and assuming that the observation errors are smaller than the model errors. From these realizations a conditional probability distribution of additive model errors given these covariates is obtained, allowing for complex non-Gaussian error structures. Random draws from this density are then used in actual ensemble data assimilation experiments. We demonstrate the efficacy of the approach through numerical experiments with the multi-scale Lorenz 96 system using both small and large time scale separations between slow (coarse scale) and fast (fine scale) variables. The resulting error estimates and forecasts obtained with this new method are superior to those from two existing methods.

Plain Language Summary Data Assimilation is an important statistical technique to optimally combine model simulations and observations based on their uncertainties. It is crucial for a wide range of areas from improving weather forecasts to target tracking. Characterizing uncertainty in both models and observations accurately is a fundamental part of effectively implementing data assimilation methods. Characterizing model uncertainty is notoriously difficult, here we propose a data driven method to quantify uncertainty in systems which are partially observed. We show the efficacy of the method over existing approaches through experiments on a two-scale atmospheric toy model.

1. Introduction

Model uncertainty quantification is one of the central challenges in successfully utilizing any data assimilation method; the optimal combination of forecasts and measurements is critically dependent on the uncertainties assigned to each. Model errors can arise from a range of sources, including but not limited to: model discretization errors in space and time, unresolved sub-grid processes, and uncertainties in model forcing or input data. The lack of complete high resolution and high quality verification data makes model error estimation difficult in most real world applications.

Early methods for characterizing model error in a variety of applications involved estimating the size and spatial structure of the missing physics (Van Leeuwen, 1999, 2001). In the realm of ensemble data assimilation, inflation and localization techniques which involve modifying the sample covariance (e.g., Anderson & Anderson, 1999; Hamill et al., 2001; Houtekamer & Mitchell, 2001) have been used extensively. These were initially developed as a heuristic remedy for filter divergence in ensemble Kalman filtering, but the increase in forecast variance associated with inflation also has the added benefit of at least partially accounting for forecast model errors. In additive inflation (Anderson & Anderson, 1999) the diagonal of the forecast covariance matrix is increased by some additive term $\lambda > 0$ whilst in multiplicative inflation (Anderson, 2001), all elements of the covariance matrix are multiplied by a $\lambda > 1$. The inflation parameter can either be manually tuned, or more objective adaptive inflation factor estimation can be used (Anderson, 2007; Liang et al., 2012; Miyoshi, 2011). A more explicit treatment

of model error involves estimating a forecast bias term, which can be considered as stochastic or deterministic. This can be estimated from the difference in mean analyses and forecasts (e.g., Dee, 1995; Saha, 1992), using the difference between 2 forecast models of differing resolution (Hamill & Whitaker, 2005) or online within the data assimilation system by incorporating a constant additive term to be updated alongside the system states (e.g., Dee & Da Silva, 1998). However, these methods are limited in that the focus is only on the first two moments of the model error distribution.

More recently, there has been renewed interest in off-line estimation of model error statistics using analysis increments (i.e., difference between forecast and analysis) from a data assimilation run (Mitchell & Carrassi, 2015; Rodwell & Palmer, 2007). These approaches rely heavily on a Gaussianity assumption. Another line of research is in time-varying model error estimation (Brasseur et al., 2005; for a marine biogeochemistry model) or in time varying parameter estimation (Pathiraja et al., 2018a, 2018b) which is particularly useful when one knows apriori that a model parameter is non-constant (e.g., land cover in a hydrologic model). However these approaches cannot capture more general model structural errors.

In the variational Data Assimilation literature, model error is often considered by formulating the forecast model as a weak constraint in the optimization problem (often referred to as Weak constraint 4D-Var; Tremolet, 2006; Zupanski, 1997). One approach to achieve this is through Long-window 4D-Var, where an additive model error term is incorporated as a control variable in the 4D-Var formulation, with initial conditions for each window held fixed (Fisher et al., 2005, 2011; Tremolet, 2006). These approaches require apriori specification of the model error covariance matrix, while our task is to estimate the characteristics of the model error (Zhu et al., 2017). Estimated the model error covariance online in a particle filter in a 1000-dimensional Lorenz 1996 model. The advantage of a particle filter is, similar to long-window 4D-Var, that the error covariance of the state plays no role in the estimation. The method needs a first guess of the model error covariance matrix, which is then updated over time, with the restriction that only an additive second order moment is estimated.

More complex methods involve estimating a parameterization on-line using linear regression on a pre-defined large set of potential functional forms (e.g., Lang et al., 2016). Furthermore, off-line methods from machine learning, such as Relevant Vector Machines (Bishop, 2006) and Bayesian Symbolic Regression (Jin et al., 2020) have been explored to find structural model errors, and hence missing physics. These methods define a set of basic functions and build model equations from these that fit the data best. All these methods have in common that they do need to define a set of functional relations, after which the fitting is performed, which limits the freedom of structure of model errors. Bonavita and Laloyaux (2020) provide an in-depth investigation of how best to integrate machine learning for model error estimation into existing data assimilation methods. See also the recent work of (Brajard et al., 2021) where neural networks are used to emulate what is viewed as model error from a data assimilation run.

A related line of research involves stochastic parameterization and model reduction methods to account for model errors associated with unresolved sub-grid scale processes in data assimilation (e.g., Berry & Harlim, 2014; Lu et al., 2017; Mitchell & Carrassi, 2015; Mitchell & Gottwald, 2012). This is particularly relevant in weather and climate modeling where the system dynamics evolve on a wide range of spatial and temporal scales. Such model reduction methods involve modeling or parameterizing sub-grid scale processes in a more computationally tractable fashion than solving the true sub-grid differential equations. Often this is achieved through either deterministic or stochastic parameterizations which aim to capture the mean effects of small scale processes on the resolved variables. Several studies have demonstrated the superiority of stochastic over purely deterministic parameterizations in this regard (Buizza et al., 1999; Palmer, 2001).

Methods for stochastic parameterizations of multi-scale systems vary widely; from homogenization methods that are suited to systems with large time scale separations (Pavliotis & Stuart, 2008; Wouters et al., 2016) to fitting stochastic models to sub-grid tendencies (e.g., Arnold et al., 2013; Wilks, 2005). Methods of the form of the latter include that of Crommelin and Vanden-Eijnden (2008), who proposed utilizing a Conditional Markov Chain to represent the evolution of sub-grid tendencies given the state of the resolved variable. The transition matrices are estimated using realizations of the true sub-grid tendencies. Kwasniok (2012) explored a similar approach whereby a clustering algorithm was used to develop a cluster-weighted Markov chain to represent the sub-grid tendencies. Arnold et al. (2013) extended the work of Wilks (2005) by examining the potential of autoregressive error models to effectively parameterize sub-grid tendencies in the multi-scale Lorenz 96 system.

This was further extended in Gagne et al. (2020) where a Generative Adversarial Network was trained on data of sub-grid tendencies and coarse scale variables from the 2 scale Lorenz 96 model. Lu et al. (2017) have proposed using a non-Markovian non-linear autoregressive moving average model to characterize model error. All of the aforementioned approaches require knowledge of the sub-grid scale equations, representative data of the sub-grid tendencies and/or full observations of the resolved variables. This reduces their applicability for more realistic data assimilation applications where knowledge of the sub-grid scale processes is unavailable and the resolved variables are only partially observed.

We propose a methodology for model uncertainty estimation that is specifically designed for partially observed systems and does not require knowledge of the sub-grid scale processes. The method is suited to systems where a locality and homogeneity assumption can be invoked, as this is used to regularize the ill-posed problem of estimating model errors from partial observations. In such systems, errors due to sub-grid scale processes are dependent only on neighboring states instead of the full resolved state vector, and the error statistics are the same at each location in space, or over larger parts of state space with similar physics.

The approach first approximates the conditional probability distribution of additive model errors given some informative covariates (e.g., the state of the system). This density is calculated from estimated model error realizations. These are obtained during a training phase by minimizing their variance conditioned on the informative covariates, constrained by available observations. Samples from the estimated distribution can then be combined with forward model simulations to generate a forecast distribution in any ensemble data assimilation framework. The distribution estimate is nonparametric, allowing for the characterization of highly non-Gaussian errors. We demonstrate its efficacy through numerical experiments with the multi-scale Lorenz 96 system. The forecast model in the assimilation experiment is the single layer Lorenz 96, so that model errors arise from the unresolved high frequency fast variables. The proposed approach is compared to two benchmark methods in terms of the ability to recover the true model error structure and the impact on assimilation and forecast quality.

The remainder of this paper is structured as follows. In Section 2, we discuss data assimilation methods and the Ensemble Transform Kalman Filter (ETKF), which is adopted as the assimilation algorithm in this study. Methods for estimating model uncertainty in partially observed systems are discussed in Section 3. The details of the proposed method are provided, along with a long window 4D-Var formulation and an ensemble analysis increment based method, both of which are adopted as benchmarks. In Section 4 we describe the numerical experiments with the multi-scale Lorenz 96' system. We conclude with a summary of the main outcomes and possibilities for future work in Section 5.

2. Data Assimilation Methods

The general problem setting considered in this study is described as follows. Suppose the system of interest can be represented by the following discrete time continuous state space equation:

$$\boldsymbol{x}_j = M\left(\boldsymbol{x}_{j-1}\right) + \boldsymbol{\eta}_j \tag{1}$$

where $\boldsymbol{x}_{j-1} \in \mathbb{R}^{N_x}$ is the true state vector at time $j-1$; $M : \mathbb{R}^{N_x} \rightarrow \mathbb{R}^{N_x}$ is a Markov Order 1 forecast model; and $\boldsymbol{\eta}_j \in \mathbb{R}^{N_x}$ is an additive model error at time j capturing deficiencies in the forecast model M.

Noisy partial observations of the state $\boldsymbol{x}_j$ are available, given by the following:

$$\boldsymbol{y}_j = \mathbf{H}\boldsymbol{x}_j + \boldsymbol{\epsilon}_j \tag{2}$$

where $\mathbf{H} : \mathbb{R}^{N_x} \rightarrow \mathbb{R}^{N_y}$ is a $N_y \times N_x$ matrix consisting of 1's and 0's only (i.e., state components are either directly observed or not at all), $\boldsymbol{y}_j \in \mathbb{R}^{N_y}$ is the vector of observations at time j, and $\boldsymbol{\epsilon}_j \in \mathbb{R}^{N_y}$ is the observation noise at time j, assumed to be temporally uncorrelated Gaussian with zero mean and known covariance matrix $\mathbf{R} \in \mathbb{R}^{N_y \times N_y}$. In this study, we focus on the case $N_y < N_x$, that is, the state vector is partially observed. These observations are available at a coarser temporal resolution than the model forecast time step. Throughout the manuscript, the notation $\boldsymbol{v}\,[k]$ is used to refer to the kth element of some vector $\boldsymbol{v}$; $\mathbf{A}\,[k, l]$ refers to the element at the kth row and lth column of some matrix $\mathbf{A}$.

The aim of data assimilation is to optimally combine observations and prior information (usually from a numerical model, e.g., Equation 1) based on their respective uncertainties. We focus on ensemble based data assimilation methods due to their suitability for dealing with non-Gaussian model errors, which is a particular focus of the proposed method. The standard discrete time Kalman filter provides the optimal posterior (in the minimum variance sense) for the special case of linear forecast model and observation operator, and for zero mean temporally uncorrelated Gaussian process and observation noise. Ensemble Kalman Filter (EnKF) methods (amongst others) have been developed for high-dimensional systems where the full covariance matrix is too large to store in a computer, with an additional benefit for the more general case of non-linear and non-Gaussian problems encountered in many applications. The (ETKF; Bishop et al., 2001; Wang et al., 2004) has been widely adopted particularly in meteorological data assimilation due to its computational efficiency and accuracy in high dimensional systems with small ensemble sizes when localization is applied. We will therefore use this as the data assimilation method in our numerical experiments in Section 4.

The ETKF is an extension of the original EnKF proposed by Evensen (1994). It belongs to the class of ensemble square root filters which operate on the square root of the forecast and analysis error covariance rather than the full covariance matrices (Tippett et al., 2003; Vetra-Carvalho et al., 2018). Such methods use a deterministic transformation to map the forecast ensemble to the analysis ensemble, whose statistics are consistent with the Kalman filter update. As noted by Tippett et al. (2003), the linear transformation is not uniquely defined, and is the main distinguishing factor between different ensemble square root methods. Here we present the method of Wang et al. (2004), which is an updated version of the original ETKF proposed by Bishop et al. (2001) that ensures the filter is unbiased. A single cycle of the ETKF is summarized below.

A forecast ensemble at time j (denoted $\mathbf{X}_j^f$) is generated by propagating the analysis ensemble from the previous time through Equation 3:

$$\boldsymbol{x}_j^{f,i} = M\left(\boldsymbol{x}_{j-1}^{a,i}\right) + \boldsymbol{\eta}_j^i \quad \forall \quad i \in \{1, \ldots, n\} \tag{3}$$

$$\mathbf{X}_j^f = \left[\boldsymbol{x}_j^{f,1}, \quad \ldots \quad , \boldsymbol{x}_j^{f,n} \right] \in \mathbb{R}^{N_x \times n} \tag{4}$$

where the superscripts f and a denote the forecast and analysis, respectively. The crucial strength of EnKFs is that one can avoid the explicit calculation of the state covariance matrices. In the ETKF, the update is achieved by writing the analysis ensemble deviation matrix $\mathbf{X}_j^{a'}$ in terms of the forecast ensemble deviation matrix $\mathbf{X}_j^{f'}$ as

$$\mathbf{X}_j^{a'} = \mathbf{X}_j^{f'}\mathbf{T} \tag{5}$$

where $\mathbf{X}_j^{f'} := \mathbf{X}_j^f - \overline{\boldsymbol{x}}_j^f \boldsymbol{k}^T \in \mathbb{R}^{N_x \times n}$, $\overline{\boldsymbol{x}}_j^f$ is the ensemble mean, $\boldsymbol{k}$ is a vector of ones and $\mathbf{T} \in \mathbb{R}^{n \times n}$ is a transformation matrix given by

$$\mathbf{T} = \mathbf{U}\left(\mathbf{I} + \boldsymbol{\Sigma}\boldsymbol{\Sigma}^T\right)^{-1/2}\mathbf{U}^T \tag{6}$$

where $\mathbf{U}$ and $\boldsymbol{\Sigma}$ arise from the SVD of the scaled forecast ensemble observation deviation matrix $\mathbf{W}$, that is,

$$\mathbf{W} := \frac{1}{\sqrt{n-1}}\left(\left[\mathbf{X}_j^{f'}\right]^T \mathbf{H}^T \mathbf{R}^{-1/2}\right) = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T. \tag{7}$$

This approach transforms the computations to ensemble space which significantly reduces the required number of operations whenever $n \ll N_y$ (as is typically the case in real world geophysical applications). Finally, the SVD of $\mathbf{W}$ is utilized to efficiently calculate the analysis ensemble mean $\overline{\boldsymbol{x}}_j^a$:

$$\overline{\boldsymbol{x}}_j^a = \overline{\boldsymbol{x}}_j^f + \frac{1}{\sqrt{n-1}}\mathbf{X}_j^{f'}\mathbf{U}\left(\boldsymbol{\Sigma}^T\boldsymbol{\Sigma} + \mathbf{I}\right)^{-1}\boldsymbol{\Sigma}\mathbf{V}^T\mathbf{R}^{-1/2}\left(\boldsymbol{y}_j - \mathbf{H}\overline{\boldsymbol{x}}_j^f\right) \tag{8}$$

We will use this data assimilation method in our data assimilation experiments described in Section 4, after we have derived an expression for the model error distribution, as detailed in the next section.

3. Estimating Model Uncertainty

In the following, we propose a method for estimating model uncertainty in partially observed systems where knowledge of the unresolved processes is unavailable. The approach is specifically for use with Monte-Carlo based sequential filtering techniques such as Ensemble Kalman methods and Particle methods. Existing methods for accounting for model errors that are amenable to the partially observed setting are also discussed in Section 3.3. We first clarify some important notation that will be used from here onwards.

3.1. Notation

The notation $\boldsymbol{v}_{t_1:t_2}$ is used to indicate the sequence of vectors $\{\boldsymbol{v}_j\}_{j=t_1,t_1+1,\ldots,t_2}$. Unless otherwise stated, the subscript notation is reserved for the time index and the superscript bracket notation $x^{(i)}$ is used to indicate the ith iteration in an iterative optimization method. The bracket notation $[i]$ is used to indicate the ith element of a vector or set. Similarly $[i, j]$ is used to indicate the element in the ith row and jth column of a matrix and $[i, \cdot]$ is used to indicate the ith row. The hat notation ^ is used to indicate an estimate of a variable. We also let S^u and S^o denote the set of indices of the unobserved and observed grid points, respectively. The shorthand notation $\boldsymbol{v}\,[S^o]$ is used to denote a vector whose ith entry is given by $\boldsymbol{v}\,[S^o\,[i]]$.

3.2. Proposed Method

The proposed method utilizes a training period to obtain estimates of model errors using an optimization procedure and knowledge of some informative covariates (e.g., the state of the system). Estimates are generated by minimizing the conditional sum of squared deviations of the model errors given the covariates, constrained by available observations. These estimates are then used to build a conditional model error probability density using kernel density estimation, which allows for the characterization of potentially non-Gaussian features. In actual data assimilation experiments such as in Section 4 below, model errors are drawn from this conditional distribution. Since the density estimation is computed off-line the cost of incorporating uncertainty in this fashion is kept to a minimum.

The following assumptions are required for the method:

1. The system states are directly but partially observed, that is, $\mathbf{H}$ takes the form as described in Section 2
2. The additive error at time j and grid point k, $\boldsymbol{\eta}_j\,[k]$, is dependent on some informative covariates captured in the matrix $\boldsymbol{Z}_j$ of size $N_x \times N_c$ where N_c is the number of covariates. For instance, if it is assumed that the error depends only on the state at the previous time and same location, then $\boldsymbol{Z}_j = \boldsymbol{x}_{j-1}$. Other possibilities include $\boldsymbol{Z}_j[k,\cdot] = \left[\boldsymbol{x}_{j-1}[k], \boldsymbol{x}_{j-1}[k-1], \boldsymbol{x}_j[k+1]\right]$ if the error is expected to depend on the states in a neighborhood of the grid point, and $\boldsymbol{Z}_j[k,\cdot] = \left[\boldsymbol{x}_{j-1}[k], \boldsymbol{x}_{j-2}[k]\right]$ if a longer temporal dependence on the states is expected
3. The magnitude of the measurement errors is small in comparison to the magnitude of the model errors, that is, $\|\boldsymbol{\varepsilon}_j\| \ll \|\boldsymbol{\eta}_j\|$
4. Additive error statistics are the same in time and space, that is, $p\,(\boldsymbol{\eta}_j\,[k]|\boldsymbol{Z}_j\,[k,\cdot]) \equiv p\,(\boldsymbol{\eta}_m\,[l]|\boldsymbol{Z}_m\,[l,\cdot])\ \forall\, l, k \in \{1, 2, \ldots, N_x\}, j, m \in \{1, 2, \ldots, T\}$

The key advantages of the proposed approach are it (a) allows for the estimation of complex error structures with minimal apriori knowledge and partial observations; (b) requires no assumptions or specification of a parametric error distribution (e.g., Gaussian errors) and considers the full range of moments (not just bias and covariance); (c) computes all error statistics from data, without the need for numerical tuning; and (d) has sufficient flexibility to incorporate a range of covariates that influence error processes, which will generally be problem dependent.

The aforementioned assumptions could be relaxed for larger scale realistic applications. For instance, in many applications the model grid and observation locations do not align perfectly. Standard interpolation procedures are not likely to be problematic for Assumption (a), so long as the quantities are directly observed. The flexibility of Assumption (b) is a strength of the method. A brief exploration of the impact of Assumption (c) is given in Section 4.3.2, showing that this assumption can be relaxed to a large extent. As mentioned in the Introduction, Assumption (d) could be relaxed by dividing the study area into smaller groups based on certain physical characteristics so that assumption four is valid within each group. These group sizes has to be chosen such that the sample size is sufficiently large.

The methodology consists of two main steps and is discussed in detail for the remainder of this section.

3.2.1. Step 1. Offline Additive Error Estimation

Given a training period of length T time steps, the aim is to estimate the sequence of errors $\boldsymbol{\eta}_{1:T}$ under the assumptions stated above. To this end we solve a constrained optimization problem where the objective function is of conditional sum of squares type:

$$\hat{\boldsymbol{\eta}}_{1:T} = \underset{\boldsymbol{\eta}_{1:T}}{\operatorname{argmin}} \sum_{k=1}^{N_x} \sum_{j=1}^{T} \left(\boldsymbol{\eta}_j[k] - \hat{m}\left(\boldsymbol{Z}_j[k,\cdot]\right)\right)^2 \tag{9}$$

subject to the constraints

$$\begin{aligned} \boldsymbol{y}_j &= \mathbf{H}\boldsymbol{x}_j \quad \forall\ j = 1, 2, \cdots T \\ \boldsymbol{x}_j &= M\left(\boldsymbol{x}_{j-1}\right) + \boldsymbol{\eta}_j. \end{aligned} \tag{10}$$

$\hat{m}\left(\boldsymbol{Z}_j[k,\cdot]\right)$ is the Nadaraya–Watson Kernel estimator of $\mathbb{E}\left(\boldsymbol{\eta}_j[k] | \boldsymbol{Z}_j[k,\cdot]\right)$, given by

$$\hat{m}\left(\boldsymbol{Z}_j[k,\cdot]\right) := \frac{\sum_{l=1}^{N_x}\sum_{i=1}^{T} K_b\left(|\boldsymbol{Z}_i[l,\cdot] - \boldsymbol{Z}_j[k,\cdot]|\right)\boldsymbol{\eta}_i[l]}{\sum_{l=1}^{N_x}\sum_{i=1}^{T} K_b\left(|\boldsymbol{Z}_i[l,\cdot] - \boldsymbol{Z}_j[k,\cdot]|\right)} \tag{11}$$

where K_b is a kernel function with bandwidth b, both of which must be selected. Common choices for the kernel function could be a Gaussian, Uniform or Epanechnikov kernel. Such regularizers are also used in semi-supervised learning where they guide the learning method to find models that respect some underlying structure of the samples. The Levenberg-Marquardt (Levenberg, 1944; Marquardt, 1963) algorithm is used as the minimizer.

Optimizing $\boldsymbol{\eta}_{1:T}$ can be prohibitively expensive especially for large T and N_x. We therefore use a sequential optimization technique over a sliding time window of length τ, as is also employed in Long window weak-constraint 4D-Var (Tremolet, 2006) and particle smoothing methods (Sarkka, 2013). For a given time t, initial condition estimate $\hat{\boldsymbol{x}}_{t-1}$ and time window length τ, the optimization problem Equations 9–11 is restricted to $j \in \{t, t+1, \ldots, t+\tau\}$ instead of $j \in \{1, 2, \ldots, T\}$. This process is then repeated by sliding the window of length τ forward one time step, so that $\boldsymbol{\eta}_{t+1:t+\tau+1}$ is optimized, where the existing estimates from the previous optimization step are used as an initial guess. The sliding window procedure allows one to avoid specifying the background error covariance matrix or including the initial condition $\hat{\boldsymbol{x}}_{t-1}$ in the optimization (Tremolet, 2006) as is needed in standard 4D-Var.

The optimization window τ must not be so large that the optimization procedure is computationally infeasible, but large enough to ensure enough points to approximate the conditional variance. Furthermore, it should be large enough so that that inclusion of a new observation at the end of the time window does not influence the initial condition.

This step of estimating model errors given the observations in the training period is summarized in Algorithm 1 and 2. A pictorial representation of the optimization over a single time window is provided in Figure 1. It shows the estimated errors at various stages of the iterative minimization process for the numerical experiment considered in Section 4.

3.2.2. Step 2. Conditional PDF Estimation and Sample Generation

The resulting sample of additive error and states estimates $\hat{\boldsymbol{\eta}}_{1:T}$ and $\hat{\boldsymbol{x}}_{0:T-1}$ from Algorithm 1 is now used to derive the conditional probability density for example, $p\left(\boldsymbol{\eta}_j[k] | \boldsymbol{x}_{j-1}[k]\right)$ for a given grid point k. Kernel conditional density estimation methods (Hall et al., 2004; Hyndman et al., 1996) are well suited to such a task, although they are generally data-intensive and suffer from the curse of dimensionality. However, they are sufficient for the class of problems considered herein where the locality assumption greatly reduces the dimension of the response variable and covariates. We adopt the method of Hayfield and Racine (2008) as implemented in the np package in *R*. For a set of N data points $\{\boldsymbol{x}_i, y_i\}_{i=1:N}$ for covariate $\boldsymbol{x} \in \mathbb{R}^d$ and response variable $y \in \mathbb{R}$, a Kernel estimate of the conditional density is constructed as

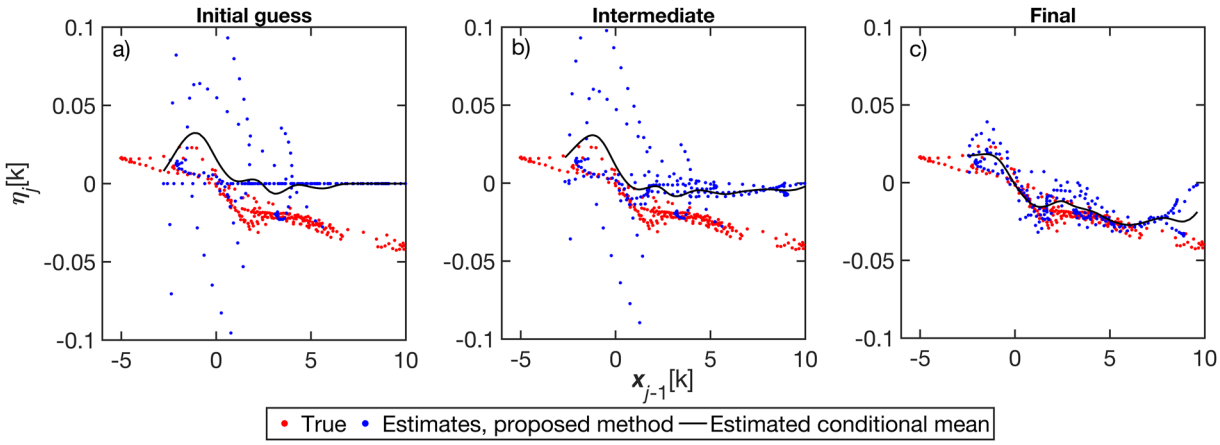

Figure 1. Representative example of the minimization process for a single window. An iterative minimization algorithm is used starting with an initial guess of zero for unobserved variables. Results are shown at various stages (a - initial guess, (b) intermediate and (c) final). The aim is to minimize the deviation of the errors from the nonparametric estimate of the conditional mean, subject to the constraint that the estimated states match the observations. Notice how the spread of the errors gradually becomes smaller from (a) to (c).

$$\hat{p}(y|\boldsymbol{x}) = \frac{\sum_{i=1}^{N} K_{b_y}(y - y_i) K_{b_x}(|\boldsymbol{x} - \boldsymbol{x}_i|)}{\sum_{i=1}^{N} K_{b_x}(|\boldsymbol{x} - \boldsymbol{x}_i|)}$$

where K_b is a user specified Kernel function with bandwidth b and b_x and b_y refer to the bandwidths selected for the covariates and response variable, respectively.

As mentioned earlier, it is also possible to include additional covariates that strongly influence the errors at the current time (for example, $\boldsymbol{\eta}_{j-1}[k]$ to capture serial dependence, as demonstrated in the numerical experiments in Section 4. The set of covariates is likely to be problem dependent; prior knowledge of the system is required to select them appropriately.

3.3. Benchmark Methods

The stochastic parameterization methods for multi-scale systems discussed in Section 1 (e.g., Arnold et al., 2013; Crommelin & Vanden-Eijnden, 2008; Kwasniok, 2012; Lu et al., 2017; Wilks, 2005) require knowledge of the sub-grid scale processes and/or fully observed resolved variables. These approaches are therefore inapplicable for the problem setting considered here. In the remainder of this section, two existing data assimilation based methods that are amenable to our problem setting are discussed. They are also adopted as benchmarks for comparison with the proposed approach.

3.3.1. B1 - Analysis Increment Based Method

Several researchers have investigated the potential of using analysis increments from a data assimilation run to characterize model errors, see for example, (Leith, 1978; Li et al., 2009; Mitchell & Carrassi, 2015). We adopt the recently proposed ETKF-TV of Mitchell and Carrassi (2015) as a representative method of such approaches (hereafter referred to as Method B1). Their method consists of estimating a Gaussian model error distribution by calculating the mean and covariance of the analysis increments over a so-called reanalysis period, via:

$$\delta \boldsymbol{x}_j^a = \frac{1}{n} \sum_{i=1}^{n} \left(\boldsymbol{x}_j^{ai} - \boldsymbol{x}_j^{fi} \right) \tag{14}$$

$$\overline{\boldsymbol{b}} = \frac{1}{T} \sum_{j=1}^{T} \delta \boldsymbol{x}_j^a \tag{15}$$

$$\overline{\mathbf{P}} = \frac{1}{T-1} \sum_{j=1}^{T} \left[\delta \boldsymbol{x}_j^a - \overline{\boldsymbol{b}} \right] \left[\delta \boldsymbol{x}_j^a - \overline{\boldsymbol{b}} \right]^T \tag{16}$$

Algorithm 1. Model Error Estimation Over Training Period With Sliding Window

1: Set:
- window size, τ
- initial state, $\hat{\boldsymbol{x}}_0$
- total time series length, T
- initial guess for errors on unobserved variables, $\boldsymbol{\gamma}_{1:T}^{(0)}$

2: for $t = 1$: T do
3: if $t \leq T - \tau - 1$ then
4: $\boldsymbol{\gamma}_{t:t+\tau}^{(t)} \leftarrow$kernelopt $\left(\hat{\boldsymbol{x}}_{t-1}, \boldsymbol{\gamma}_{t:t+\tau}^{(t-1)}\right)$
5: $\boldsymbol{\gamma}_{t:t+1+\tau}^{(t)} \leftarrow \left\{\boldsymbol{\gamma}_{t:t+\tau}^{(t)}, \boldsymbol{\gamma}_{t+\tau+1}^{(0)}\right\}$
6: $\hat{\boldsymbol{\eta}}_t\left[S^u\right] \leftarrow \boldsymbol{\gamma}_t^{(t)}$
7: else
8: $\hat{\boldsymbol{\eta}}_t\left[S^u\right] \leftarrow \boldsymbol{\gamma}_t^{(T-\tau-1)}$
9: end if
10: $\hat{\boldsymbol{\eta}}_t\left[S^o\right] \leftarrow \boldsymbol{y}_t - \mathbf{H}M\left(\hat{\boldsymbol{x}}_{t-1}\right)$
11: $\hat{\boldsymbol{x}}_t \leftarrow M\left(\hat{\boldsymbol{x}}_{t-1}\right) + \hat{\boldsymbol{\eta}}_t$
12: end for
13: return $\hat{\boldsymbol{\eta}}_{1:T}; \hat{\boldsymbol{x}}_{0:T-1}$

where $\boldsymbol{x}_j^{ai}$ and $\boldsymbol{x}_j^{fi}$ refer to the ith ensemble member at time j obtained from the reanalysis assimilation run for the analysis and forecast respectively. Note Equations 14-16 are derived assuming the analysis interval length is the same in the reanalysis and experimental run (as is the case in this study). This model error distribution is then used to draw model error samples for the the actual data assimilation experiments. Since this estimate of the model error is corrupted by various sources of error (including from the data assimilation algorithm used to generate the analysis increments), the authors include a tuning parameter α leading to a model forecast of the form:

$$\boldsymbol{x}_j^{fi} = M\left(\boldsymbol{x}_{j-1}^{ai}\right) + \alpha\boldsymbol{\eta}_j^i \tag{17}$$

$$\boldsymbol{\eta}_j^i \sim N\left(\overline{\boldsymbol{b}}, \overline{\mathbf{P}}\right) \tag{18}$$

3.3.2. B2 -Error Estimation Using Long Window Weak Constraint 4D-Var

As discussed in Section 3.2, the proposed method for estimation of additive errors relies on ideas from Long window weak-constraint 4D-Var (Tremolet, 2006) to avoid specification of the background covariance matrix. However, it differs in the specification of the cost function, as the 4D-Var method provides the least squares solution for the model error control variable. The second benchmark (Method B2) is taken to be the same as the proposed approach, but with Step 1 (see Section 3.2.1) replaced by Long window weak constraint 4D-Var estimates for the model error. The probability density estimation (Step 2) remains unchanged. It is worth noting that this is not exactly a "standard" method in its entirety, but is investigated to examine the benefit of the conditional sum of squared deviation minimization aspect of the proposed method. The long window weak constraint 4D-Var method is discussed below.

In variational data assimilation model errors are accounted for using weak constraint 4D-Var. In the formulation where the initial state and model errors are considered as control variables, this amounts to minimizing the following cost function over a time window of length τ (Tremolet, 2006):

Algorithm 2. kernelopt

1: Set:
- state estimate at time $t-1$, $\hat{\boldsymbol{x}}_{t-1}$
- best guess for unobserved errors for time t to $t + \tau$, $\boldsymbol{\gamma}_{t:t+\tau}^{(t-1)}$
- stopping criterion J_{stop}
- Kernel function and bandwidth K_b and b respectively
- maximum no. of iterations *maxiter* (to prevent excessive computations when convergence to reaching J_{stop} is too slow)

2: Initialize:
- $m \leftarrow 1$
- $\boldsymbol{\gamma}_{t:t+\tau}^{(t)} \leftarrow \boldsymbol{\gamma}_{t:t+\tau}^{(t-1)}$

3: while $m < maxiter$ do
4: for $j = t: t+\tau$ do
5: $\hat{\boldsymbol{\eta}}_j\,[S^o] \leftarrow \boldsymbol{y}_j - \mathbf{H}M\,(\hat{\boldsymbol{x}}_{j-1})$
6: $\hat{\boldsymbol{\eta}}_j\,[S^u] \leftarrow \boldsymbol{\gamma}_j^{(t)}$
7: $\hat{\boldsymbol{x}}_j \leftarrow M\,(\hat{\boldsymbol{x}}_{j-1}) + \hat{\boldsymbol{\eta}}_j$
8: end for
9: Calculate: $\hat{\boldsymbol{Z}}_j[k,\cdot]$ using $\hat{\boldsymbol{x}}_{t-1:t+\tau}$ for all j, k
10:

$$\hat{m}(\hat{\boldsymbol{Z}}_j[k,\cdot]) \leftarrow \frac{\sum_{l=1}^{N_x}\sum_{i=t}^{t+\tau} K_b\left(\left|\hat{\boldsymbol{Z}}_i[l,\cdot] - \hat{\boldsymbol{Z}}_j[k,\cdot]\right|\right)\hat{\eta}_i[l]}{\sum_{l=1}^{N_x}\sum_{i=t}^{t+\tau} K_b\left(\left|\hat{\boldsymbol{Z}}_i[l,\cdot] - \hat{\boldsymbol{Z}}_j[k,\cdot]\right|\right)} \tag{12}$$

11:

$$J\left(\gamma_{t:t+\tau}^{(t)}\right) \leftarrow \sum_{k=1}^{N_x}\sum_{j=t}^{t+\tau}\left(\hat{\eta}_j\,[k] - \hat{m}\left(\hat{\boldsymbol{Z}}_j\,[k,\cdot]\right)\right)^2 \tag{13}$$

12: if $J > J_{stop}$ then
13: Calculate: new guess $\boldsymbol{\gamma}_{t:t+\tau}^{(t)}$ based on $\boldsymbol{J}\left(\boldsymbol{\gamma}_{t:t+\tau}^{(t)}\right)$ as per chosen optimization scheme
14: $m \leftarrow m + 1$
15: else
16: $m \leftarrow maxiter$
17: end if
18: end while
19: return $\boldsymbol{\gamma}_{t:t+\tau}^{(t)}$

$$J(\boldsymbol{x}_0,\boldsymbol{\eta}) = \underbrace{\frac{1}{2}(\boldsymbol{x}_0-\boldsymbol{x}_b)^T\mathbf{B}^{-1}(\boldsymbol{x}_0-\boldsymbol{x}_b)}_{J_B} + \underbrace{\frac{1}{2}\sum_{j=1}^{\tau}\boldsymbol{\eta}_j^T\mathbf{Q}_j^{-1}\boldsymbol{\eta}_j}_{J_Q} + \underbrace{\frac{1}{2}\sum_{j=0}^{\tau}\left(\mathbf{H}\boldsymbol{x}_j-\boldsymbol{y}_j\right)^T\mathbf{R}^{-1}\left(\mathbf{H}\boldsymbol{x}_j-\boldsymbol{y}_j\right)}_{J_O} \tag{19}$$

where $\boldsymbol{x}_b$ is the background estimate of the initial state $\boldsymbol{x}_0$; $\mathbf{B}$ is the background error covariance matrix associated with $\boldsymbol{x}_b$; $\mathbf{Q}$ is the model error covariance matrix; and $\boldsymbol{x}_j = M(\boldsymbol{x}_{j-1}) + \boldsymbol{\eta}_j$ for $j \in \{1, \ldots\tau\}$. The assimilation cycle is repeated by then considering the next assimilation window $\{\tau+1, 2\tau\}$. However, in the long window approach, minimization is performed by shifting the interval by one observation interval rather than the full assimilation window of length τ. This allows one to neglect the background term J_B from the cost function after a suitable

warm up period. The estimate $\boldsymbol{x}_b$ would have already converged due to the many iterations of the minimization algorithm from the overlapping windows, meaning its uncertainty is negligible in comparison to the other terms. The window length should also be chosen to be sufficiently long, such that the inclusion of a new observation at the end of the time window does not affect the initial state (this is relevant for the proposed approach also).

In summary, the B2 method is defined as being the same as the proposed approach, but with the cost function in Equation 9 replaced by minimization of the following cost function for any given time t:

$$J\left(\boldsymbol{\eta}_{t:t+\tau}, \hat{\boldsymbol{x}}_{t-1}\right) = \underbrace{\frac{1}{2}\sum_{j=t}^{t+\tau} \boldsymbol{\eta}_j^T \mathbf{Q}_j^{-1} \boldsymbol{\eta}_j}_{J_Q} + \underbrace{\frac{1}{2}\sum_{j=t}^{t+\tau} \left(\mathbf{H}\boldsymbol{x}_j - \boldsymbol{y}_j\right)^T \mathbf{R}^{-1} \left(\mathbf{H}\boldsymbol{x}_j - \boldsymbol{y}_j\right)}_{J_O} \tag{20}$$

where $\boldsymbol{x}_j = M(\boldsymbol{x}_{j-1}) + \boldsymbol{\eta}_j$ for $j \in \{t, \ldots t+\tau\}$ and the initial condition is given by $\hat{\boldsymbol{x}}_{t-1}$. The estimated model error distribution is derived in the same way as for the proposed method, as detailed below.

The Levenberg-Marquardt algorithm is again used as the minimizer, for the sake of comparison with the proposed approach. Notice that unlike the proposed approach, the errors in the entire state vector (not just unobserved states) must be optimized.

4. Numerical Experiments

4.1. Multi-Scale Lorenz 96

Here we investigate the efficacy of the proposed method and benchmarks discussed in Section 3 through synthetic experiments using the multi-scale Lorenz 96 model. This system has been used extensively as a toy model of the atmosphere to test new algorithms and to study model errors due to unresolved sub-grid processes. It consists of a coupled system of equations representing the evolution of an atmospheric quantity discretized over a latitude circle at different scales:

$$\frac{d\boldsymbol{X}[k]}{dt} = -\boldsymbol{X}[k-1](\boldsymbol{X}[k-2] - \boldsymbol{X}[k+1]) - \boldsymbol{X}[k] + F + \boldsymbol{U}[k];\ k \in \{1, \ldots, N_x\} \tag{21}$$

$$\xi\frac{d\boldsymbol{V}[l,k]}{dt} = -\boldsymbol{V}[l+1,k](\boldsymbol{V}[l+2,k] - \boldsymbol{V}[l-1,k]) - \boldsymbol{V}[l,k] + h_z\boldsymbol{X}[k]; \quad l \in \{1, \ldots, N_z\} \tag{22}$$

The $\{\boldsymbol{X}[k]\}_{k=1}^{N_x}$ variables represent quantities evolving continuously in time on a coarse spatial scale with low-frequency large amplitude fluctuations, where the subscript k refers to the kth grid point on the latitude circle. Each $\boldsymbol{X}[k]$ variable is coupled to N_z small-scale variables $\boldsymbol{V}[l,k]$ that are characterized by a high frequency and relatively small amplitude evolution. The variables are driven by a quadratic term that models advection, a linear damping, constant forcing (F) and coupling terms that link the two scales. The system is subject to periodic boundary conditions, so that $\boldsymbol{X}[k] = \boldsymbol{X}[k+N_x]$, $\boldsymbol{V}[l,k] = \boldsymbol{V}[l,k+N_x]$ and $\boldsymbol{V}[l+N_z,k] = \boldsymbol{V}[l,k+1]$. The effect of the unresolved fast variables on the slow variables is denoted by the so-called sub-grid tendency $\boldsymbol{U}[k]$:

$$\boldsymbol{U}[k] = \frac{h_x}{N_z}\sum_{l=1}^{N_z} \boldsymbol{V}[l,k] \tag{23}$$

we use the formulation of the Lorenz 96 Equations 21–22 as provided in (Fatkullin & Vanden-Eijnden, 2004) which makes the time-scale separation between the slow and fast variables (measured by ξ) explicit. Note that this formulation is equivalent to the system originally proposed by Lorenz with the following parameter conversions: $\xi = \frac{1}{c}$ where c = time scale ratio; $h_x = \frac{-hcN_z}{b^2}$ where b = spatial scale ratio and h is the coupling constant; and $h_z = h$.

The behavior of the system can vary considerably depending on the values assigned to the parameters in Equations 21-22. We consider two dynamical regimes to study the robustness of the proposed approach to different model error structures, summarized in Table 1. We first consider a case with large time scale separation ($\xi \approx 0.008$) studied by Fatkullin and Vanden-Eijnden (2004). The sub-grid tendency has a complex non-linear dependence on the resolved variable, making it of interest to this study. However, such large time scale separations

Table 1
Multi-Scale Lorenz 96 Parameters for the Two Different Case Studies.

	Parameter	Case Study 1	Case Study 2
Lorenz 96 parameters	ξ	$\frac{1}{128} \approx 0.008$	0.7
	h_x	−0.8	−2
	h_z	1	1
	N_z	128	20
	N_x	9	9
	F	10	14
Observation density	Observation frequency (MTU)	0.02	0.04
	Location of Observed $\boldsymbol{X}[k]$	$S^o = \{3, 4, 8, 9\}$	$S^o = \{1, 2, 5, 6\}$

Note. That 1 MTU $= \frac{1}{\Delta t}$ time steps and the model equations are discretized with $\Delta t = 8 \times 10^{-4}$

are not representative of the real atmosphere. We therefore consider a second case that has a much smaller time scale separation ($\xi \approx 0.7$) and stronger coupling ($h_x = -2$) than cases considered in previous studies (where ξ is typically 0.4 or 0.5 and $h_x = 1$; e.g., Arnold et al., 2013; Crommelin & Vanden-Eijnden, 2008; Lu et al., 2017). Smaller values of ξ (i.e., larger time scale separation) are generally considered more difficult to parameterize, and the larger magnitude of the coupling term amplifies the effect of model errors. In both case studies, the dynamics are chaotic and give rise to complex non-Gaussian conditional error densities, as shown by the variation of $\boldsymbol{U}[k]$ with $\boldsymbol{X}[k]$ in Figure 2.

4.2. Experimental Setup

The available forecast model is the single scale Lorenz 96 model Equation 24, where the forcing term is known perfectly but knowledge of the sub-grid processes $\boldsymbol{V}[l, k]$ is unavailable:

$$\frac{d\boldsymbol{X}[k]}{dt} = -\boldsymbol{X}[k-1](\boldsymbol{X}[k-2] - \boldsymbol{X}[k+1]) - \boldsymbol{X}[k] + F; \quad k \in \{1, \ldots, N_x\} \tag{24}$$

our aim is to first characterize the uncertainty in model simulations due to missing physics that is, the subgrid term in Equation 23, where the resolved variables are partially observed. Then we study the effects of uncertainty characterization on forecasts and assimilation.

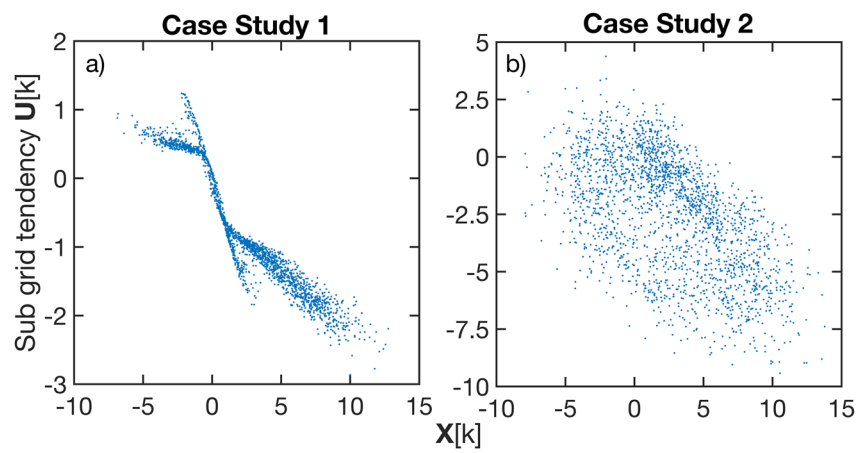

Figure 2. Sub-grid tendencies for the two different regimes of the multi-scale Lorenz 96 system considered in this study: (a) Case 1 - large time scale separation; (b) Case 2 - small time scale separation. For both cases, points are sampled at an interval of 0.3 MTU.

4.2.1. Training Period

A truth run for the training period was first generated by numerically integrating the full multi-scale system Equations 21-22 using a fourth-order Runge-Kutta scheme with time step $\Delta t = 0.0008$. Similar to (Arnold et al., 2013), we use MTU to denote model time units, where 1MTU $= \frac{1}{\Delta t}$.

Partial observations of the resolved slow variables were then developed by perturbing the true values with zero mean, temporally and spatially uncorrelated Gaussian noise:

$$\begin{aligned} \boldsymbol{y}_j &= \mathbf{H}\boldsymbol{x}_j + \boldsymbol{\varepsilon}_j \\ \boldsymbol{\varepsilon}_j &\sim N(0, \mathbf{R}) \end{aligned} \tag{25}$$

where $\mathbf{H}$ is a non-square matrix with $\mathbf{H}\,[i, S^o\,[i]] = 1$, for all $i = \{1, 2, \ldots, N_y\}$ and 0 otherwise (see details of S^o in Table 1), $\boldsymbol{x}_j$ is the true state at time j where $\boldsymbol{x}_j\,[k]$ is equivalent to the time discretized value of $\boldsymbol{X}_k$ in Equation 21, and $\mathbf{R} = 10^{-7}\mathbf{I}_{N_y}$ where $\mathbf{I}_{N_y}$ is the identity matrix of size N_y. $\mathbf{R}$ was chosen such that measurement errors are negligible in comparison to model errors.

The resolved variables are partially observed in space in all experiments (approx. 50% observed, see Table 1), and observations are available at 0.02 and 0.04 MTU for Case Studies 1 and 2 respectively. Based on the work of Lorenz (2006), this corresponds to an observation interval of 2.5 and 5 hr respectively (1 MTU is approximately equivalent to 5 days). These interval lengths where chosen to reflect a realistic observation network whilst also maintaining complex non-Gaussian error structures.

The proposed approach was then applied to the training data. The forecast model Equation 24 was integrated using a time step of $\Delta t = 8 \times 10^{-4}$. We estimate the following probability densities:

$$\text{Case Study 1}: p(\eta_j\,[k]\,|x_{j-1}\,[k])$$

$$\text{Case Study 2}: p(\eta_j\,[k]\,|x_{j-1}\,[k]\,, x_{j-1}\,[k-1]\,, \eta_{j-1}\,[k])$$

Note the inclusion of the past value of the model error in Case Study 2, which is related to the presence of time correlations in errors in the Lorenz 96 system (as identified by e.g., Arnold et al. (2013)). In a real system, such choices would be informed by expert knowledge of the error processes.

Window lengths equal to $\tau = 25$ and 50 observation intervals were selected for Case Study 1 and 2, respectively. This was sufficiently long to capture a range of dynamical states and also longer than the system decorrelation time, so that the sliding window approach can be utilized to ignore the background term in the cost function (as discussed in Section 3.3.2). To ensure temporal independence the data for the nonparametric conditional density estimation was generated by sampling the estimated error and states at an interval of 0.3 MTU, where autocorrelation is approximately zero. A Gaussian Kernel function was adopted throughout using the data-driven bandwidth estimation procedure as detailed in (Hayfield & Racine, 2008). The np package in *R* was used for the bandwidth estimation and the in-built Levenberg-Marquardt algorithm in Matlab was used for optimization. To avoid issues related to bandwidth specification and data sparsity in high dimensions, outlier points in the covariate space were removed from the data used for density estimation in Case Study 2.

This training data was also used in the benchmark methods. For method B2 we used the same window length and density estimation algorithm as for the proposed approach. The process error covariance matrix $\mathbf{Q}$ was estimated by calculating the sample covariance of the true errors over the training period. For the B1 method, the inflation parameter used in the ETKF which provides the analysis increments was tuned based on the analysis Root Mean Squared Error (RMSE), whilst the correction factor α (see Equation 17) was selected by evaluating the spread versus RMSE relationships, as it has a greater impact on ensemble spread than accuracy. The optimal value was found to be 0.8; the fact that it is less than one is because the model error spread is overestimated due to the inability of the method to resolve the complicated non-Gaussian error structure (see Figure 3). Localization was not required due to the large ensemble size ($n = 1000$) relative to the state dimension.

4.2.2. Assimilation Period

Model errors generated using the proposed method and two benchmarks were then assessed in assimilation experiments using the ETKF. The forecast model in the assimilation experiment was also the single scale Lorenz

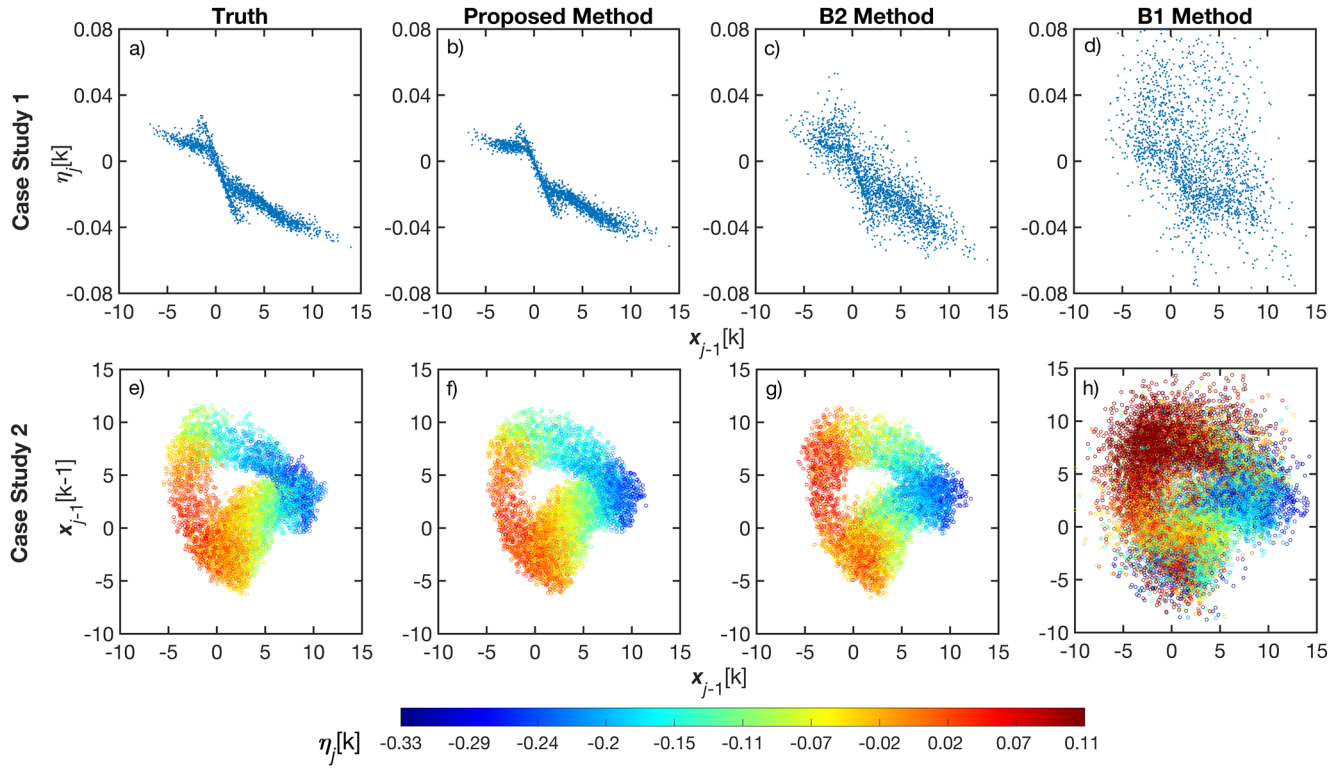

Figure 3. Sub-grid tendencies for the two different regimes of the multi-scale Lorenz 96 system considered in this study: (a) to (d) Case Study 1, that is,.large time scale separation; (b) to (h) Case Study 2, that is, small time scale separation. For both case studies, points are sampled at an interval of 0.3 MTU.

96 Equation 24; spatio-temporal observation frequency was the same and observations were generated also using Equation 25. Assimilation was undertaken for 30 independent runs of length 100 observation intervals with independent initial conditions. Truth runs were first generated using the same approach as for the training period. The initial conditions were generated by selecting 30 values on the attractor at intervals of 12,500 time steps, which is sufficient to ensure that the autocorrelation in the resolved variables is close to zero (also adopted by Arnold et al. (2013)). Perfect initial conditions were adopted in all experiments as the focus is on the effects of model error. Similarly, a large ensemble size ($n = 1000$) was adopted to minimize the effects of sampling error and to avoid the use of localization methods. Furthermore, we can avoid the use of inflation for the assimilation experiments with the model errors generated from the proposed method.

4.3. Results and Discussion

4.3.1. Model Error Estimation

In both case studies, the proposed approach recovers the true error estimates from partial observations more accurately than the benchmark methods. This is demonstrated qualitatively in Figure 3, which shows the sample set of additive errors $\boldsymbol{\eta}_j\,[k]$ against the spatial covariates (resolved variable at the previous observation time, $\boldsymbol{x}_{j-1}\,[k]$ for Case Study one; and $\boldsymbol{x}_{j-1}\,[k-1]$ and $\boldsymbol{x}_{j-1}\,[k]$ for Case Study 2). In Case Study 1, the B2 method manages to at least partially recover the non-linear relationship between $\boldsymbol{\eta}_j\,[k]$ and $\boldsymbol{x}_{j-1}\,[k]$, but is less precise than estimates from the proposed method (compare Figures 3b and 3c). In Case Study 2, it more closely reflects the true error structure, although an overestimation and underestimation of error values is apparent in key regions of the covariate space (compare Figures 3e–3g). Method B1 produces poor quality error estimates in both case studies; errors are grossly overestimated and the dependence structure between the errors and covariates is poorly represented.

The model error estimation techniques considered here can also be considered as stochastic parameterizations of the sub-grid dynamics. The ability of the methods to replicate key characteristics of the full 2-scale Lorenz 96 model when used in this manner is also assessed. For each case the single-layer Lorenz 96 system is run for

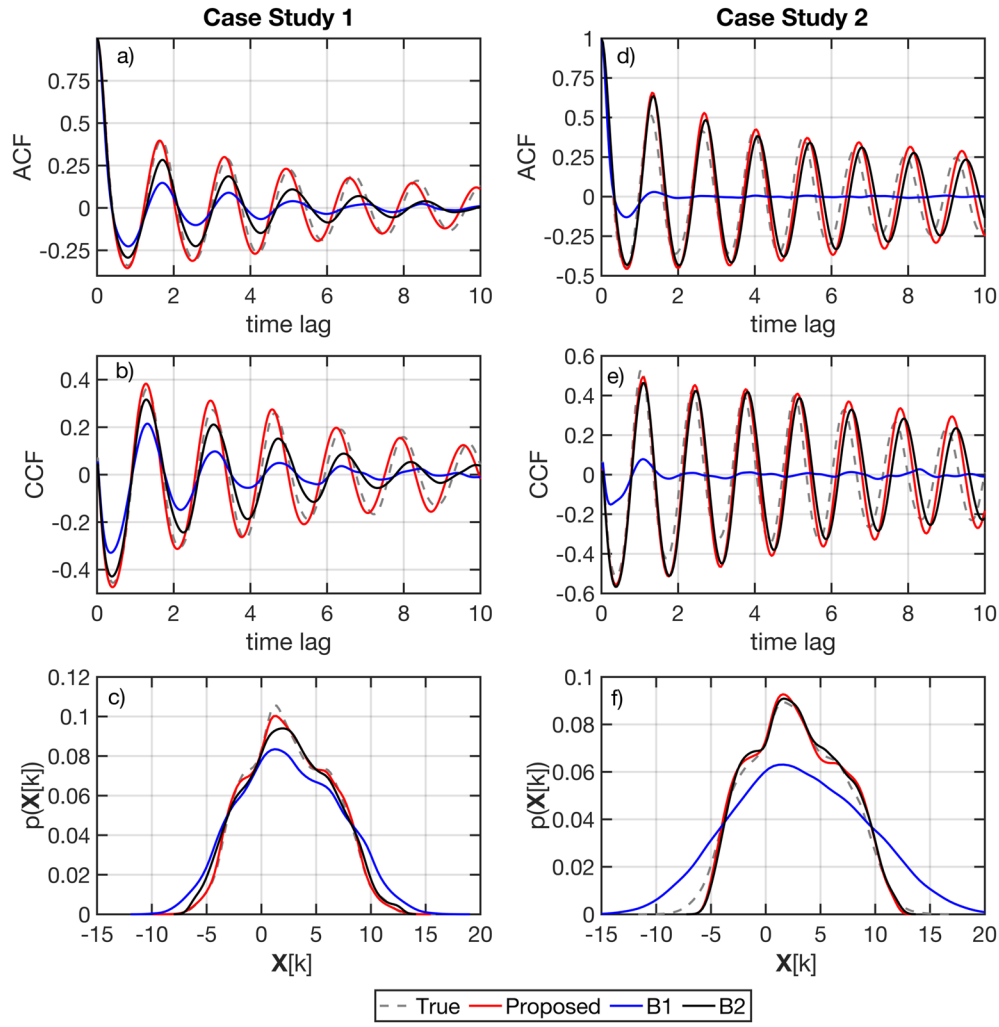

Figure 4. Autocorrelation function, cross-correlation function and marginal density of a resolved variable for both case studies using different parameterization approaches.

10^5 time steps with a $\Delta t = 8 \times 10^{-4}$, adding draws from the model error pdfs at the observation intervals used to construct these pdfs, that is, 0.02 and 0.04 MTU for Case Study 1 and 2, respectively. We calculate the autocorrelation function of X_k, the cross-correlation function between $\boldsymbol{X}\,[k]$ and $\boldsymbol{X}\,[k+1]$, and the marginal probability density of $\boldsymbol{X}\,[k]$ (see Figure 4). The correlation functions approximate the dynamical transitions of the slow variables whilst the marginal probability density approximates the invariant measure. Again, Method B1 performs poorly in all aspects, particularly in Case Study 2 where temporal correlations are not reproduced, meaning that the dynamical transitions are poorly represented. The results are similar to those from using inflation and localization only, in case studies with a similar time scale separation (e.g., Lu et al., 2017). Improvements of the proposed method over Method B2 are more distinct in Case Study 1 than in Case Study 2, consistent with the greater similarity in error estimates in this case study (see Figure 3). The Proposed Method reproduces all three features relatively accurately in both case studies, and even compares favorably with other methods that rely on data of the sub-grid processes (cf. Figures 5–7 in Crommelin and Vanden-Eijnden (2008) and Figure 1 in Lu et al. (2017)).

The superior performance of the proposed method is attributed to two aspects (a) the formulation of the cost function which aims to minimize the conditional sum of squared deviations of the estimated errors; and (b) optimization of errors over a time window (as is performed in traditional 4D-Var and smoothing methods). First, minimizing the conditional sum of squared deviations of the errors allows one to estimate more complex state dependent error structures, as opposed to the 4D-Var type approach in Method B2 where dependence information

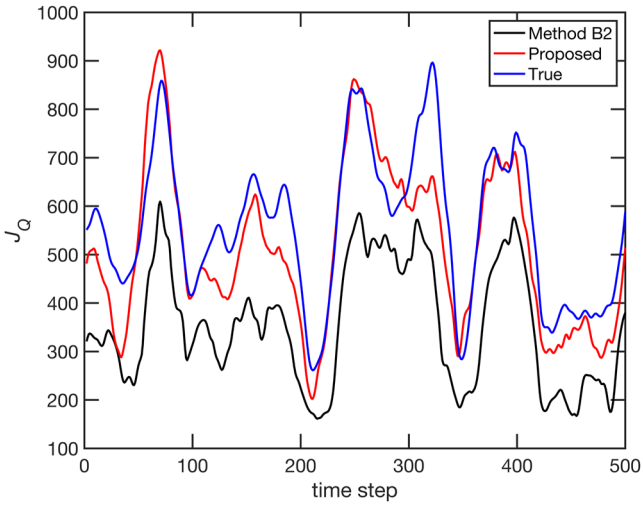

Figure 5. Snapshot of J_Q values (see Equation 20) for method B2, proposed and the true data for Case Study 2.

is not taken into account. The error estimates from Method B2 give J_Q terms (see Equation 20) that are most often lower than J_Q of the true data (see Figure 5), meaning that estimates are obtained by minimizing an inappropriate cost function for this setting. Furthermore, the proposed approach has the added benefit of avoiding the specification of a model noise covariance matrix, which is needed in Method B2.

Second, optimization over a time window allows one to more effectively constrain the range of possible errors in the partially observed setting, particularly when errors are time correlated. This partly explains the poor performance of Method B1, which is based on increments from a filter). Furthermore, the error estimates from Method B1 are heavily influenced by the quality of the assimilation algorithm (ETKF with inflation). Poorly specified prior uncertainty in the unobserved variables from the inflation procedure can lead to large incremental updates in observed variables at future times. This ultimately corrupts the error estimates, as the increments are now dominated by initial condition errors in the unobserved variables.

4.3.2. Model Error Estimation With Non-Negligble Observation Error

In the aforementioned experiment, negligible observation errors $\left(\mathbf{R} = 10^{-7}\mathbf{I}_{N_y}\right)$ were considered, consistent with assumption 3. This assumption is clearly a limitation for real world applications, and future work will examine how this assumption can be relaxed. As a first step in this direction, we examined the robustness of the procedure by repeating the error estimation procedure described in Section 3.2 for Case Study 1, but with larger observation error $\mathbf{R} = 2 \times 10^{-5}\mathbf{I}_{N_y}$. With this choice, the model error standard deviation is approximately 3 times the observation error standard deviation.

Figure 6 shows that although error estimates from the Proposed method are not as precise as in the previous experiment, they still capture the underlying error structure more effectively than method B2. Both are superior to method B1 even in the presence of negligible observation error (cf. Figure 3d). In cases where the observation error variance is of similar or greater magnitude than the model error variance, the performance of the Proposed method will degrade because of the hard constraint in Equation 10. Future research will involve developing methods to deal with this scenario whilst maintaining the flexibility of being able to detect non-Gaussian error structures.

4.3.3. Forecast Skill

The superior error estimates from the proposed approach leads to improved forecasts compared to the benchmark methods. Representative results of one-step-ahead forecasts for both case studies are provided in Figure 7 and Figure 8. They show the relative histograms of the ensemble anomalies (forecast - truth) for both an observed (left

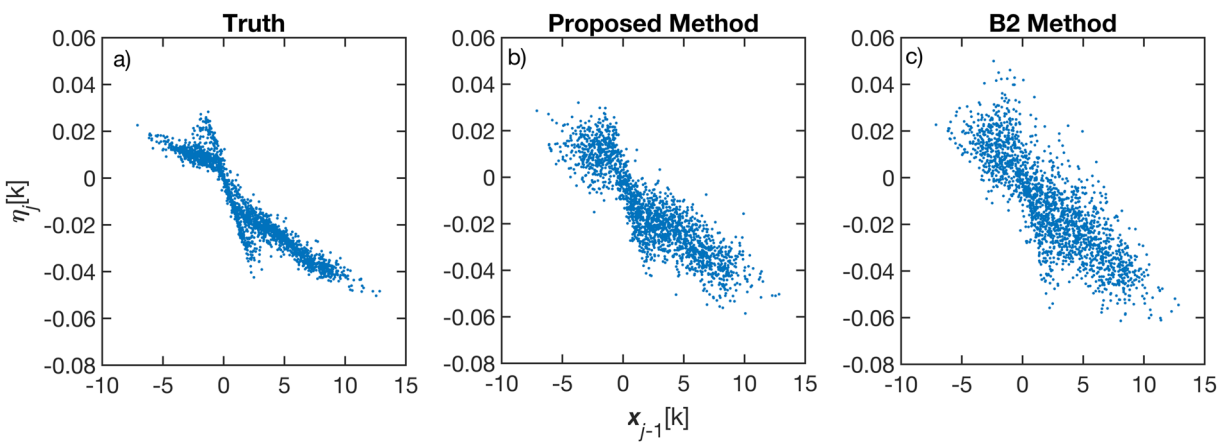

Figure 6. Sub-grid tendencies for Case Study 1 with increased observation error.

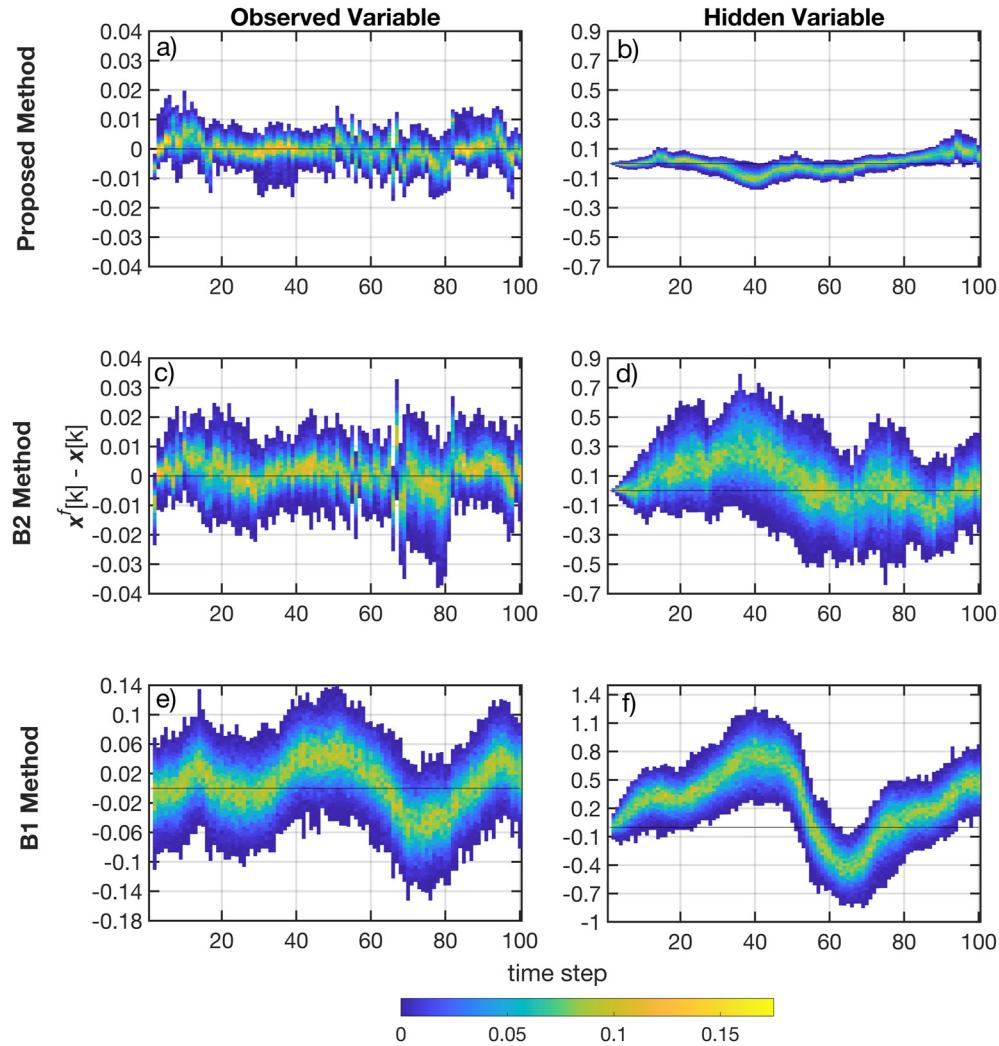

Figure 7. Relative histogram of anomalies $\left(x_i^f[k] - x_t[k]\right)$ for an observed variable (left column) and unobserved variable (right column) for the different methods for Case Study 1. Forecasts are one-step-ahead (in this case, 0.02 MTU).

column) and unobserved (right column) variable for a single assimilation run of 100 cycles. In both case studies, relatively large systematic errors can be seen when using Method B1 compared to the other approaches, which is unsurprising given the results in Figure 3. One step ahead forecasts of the observed variables are relatively similar between the proposed method and the B2 method, although the forecast variance is considerably lower, particularly in Case Study 1. This is a direct consequence of the more precise additive model error estimates obtained from the proposed approach.

Resolving bimodality of the transition errors allows one to generate more accurate analyses (hence initial conditions for subsequent time steps) and forecasts, even in an EnKF setting. This is demonstrated in Figure 9 where initial conditions and forecasts in Method B2 have greater variance than in the proposed method, as it is unable to precisely resolve the two modes of the error density. Differences between forecasts from the proposed and B2 method are much more pronounced for the hidden variables, where both bias and variance are much lower when using the proposed method in both case studies. The conditional sum of squared deviations minimization procedure allows for a more accurate representation of the spatial dependence structure. This means that information from observed variables is more effectively transferred to unobserved variables during the assimilation, thereby contributing to the improved forecasts seen for the Proposed Method compared to Method B2, through better initial conditions. Forecast skill is assessed quantitatively in the remainder of this section.

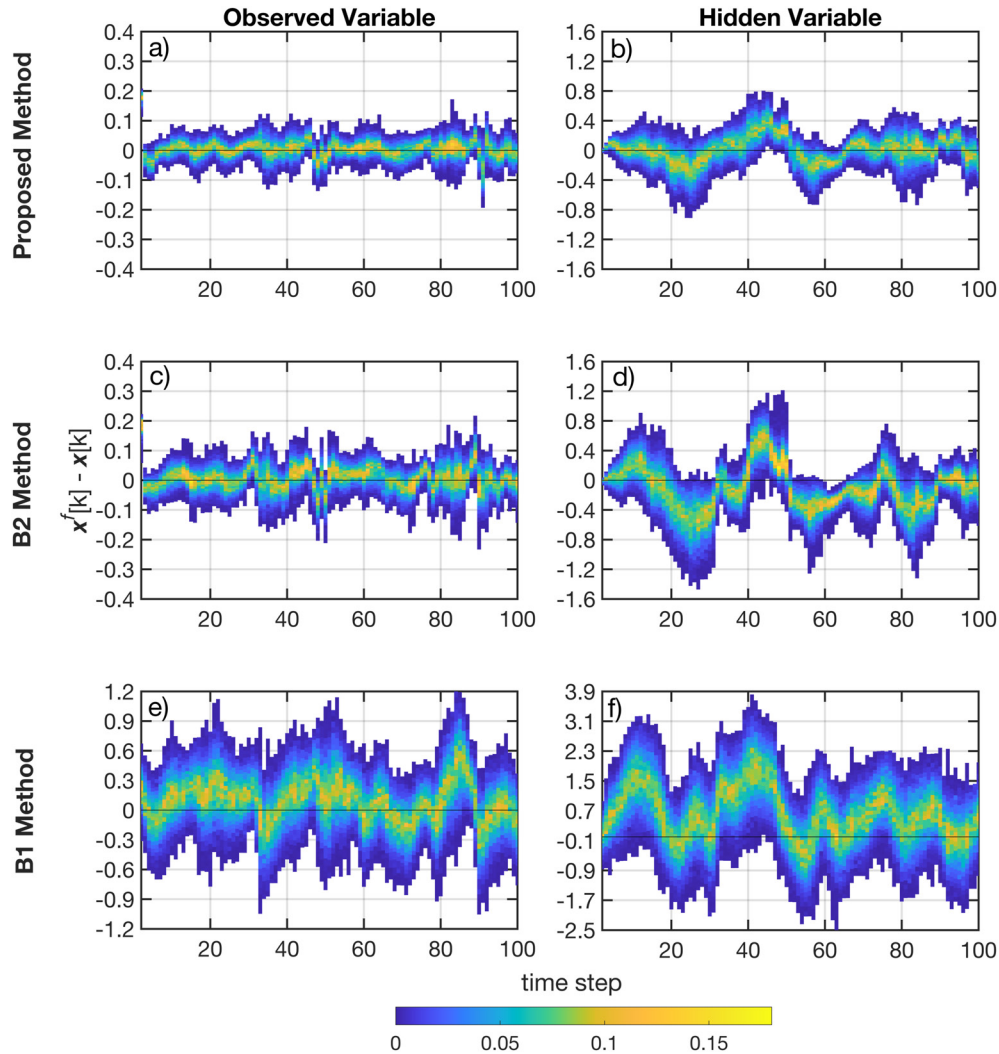

Figure 8. Relative histogram of anomalies $\left(x_i^f[k] - x_t[k]\right)$ for an observed variable (left column) and unobserved variable (right column) for the different methods for Case Study 2. Forecasts are one-step-ahead (in this case, 0.04 MTU).

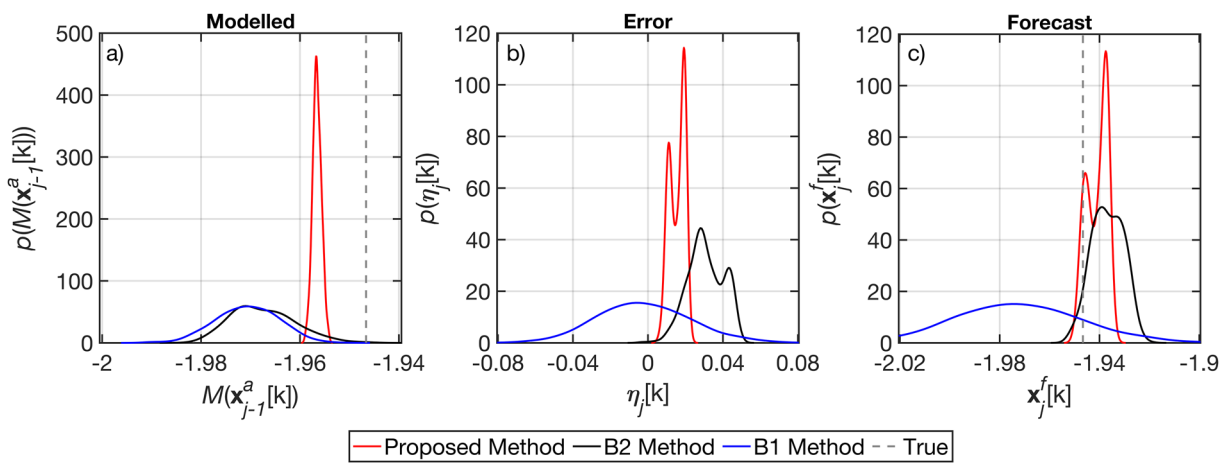

Figure 9. Example showing benefit of accounting for bimodal transition in an observed variable in Case Study 1 (shown for $t = 82$ in Figure 7).

A range of forecast metrics were considered to quantify forecast properties including reliability, resolution, accuracy and consistency. Reliability and resolution were quantified using the Continuous Ranked Probability Score (CRPS) and the (negative) Logarithmic Score (LS) given in Equations 26–28:

$$CRPS_j = \int_{-\infty}^{\infty} \left(F_j^f(y) - F_j^o(y)\right)^2 dy \tag{26}$$

$$F_j^o(y) = \begin{cases} 0 & y < y_j \\ 1 & y \geq y_j \end{cases} \tag{27}$$

$$LS_j = -\ln\left(p_j^f\left(y = y_j\right)\right) \tag{28}$$

where $F_j^f(y)$ is the empirical cumulative distribution function of the forecast of variable y at time j; $F_j^o(y)$ is the cumulative distribution function of the observations of y at time j; and $p_j^f\left(y = y_j\right)$ indicates the value of the forecast probability density function, evaluated at the observation value. For cases where only a single observation of y is available at each time, the Heaviside step function is used to characterize the cumulative distribution function of the observation (see Equation 27).

The CRPS is routinely adopted in forecasting studies, although it can be a poor statistic for complex forecast probability densities (see for example Smith et al. (2015) who showed that the CRPS can give misleadingly good scores to outcomes that fall in between two modes of a bimodal forecast density). Hence, the LS is also calculated, although it has the drawback of heavily penalizing forecasts in which the outcome falls outside the forecast range. Accuracy is measured by the RMSE, which is evaluated on the ensemble mean.

Statistical consistency is characterized using RMS Error versus RMS Spread diagnostic plots, which has been adopted in similar studies (see e.g., Arnold et al., 2013). Ensemble forecasts are considered statistically consistent if the expected ensemble variance equals the expected squared ensemble mean error (assuming unbiasedness and a large enough ensemble size). We separate forecasts into 10 equally populated bins according to their forecast variance, and the mean square spread and mean square error are calculated for each bin prior to taking the square root.

We used the forecast skill score (*FSS*) to quantify the relative improvement of the proposed approach over the benchmark methods, defined as:

$$FSS = \frac{Score_{\mathrm{Pr}} - Score_{Be}}{Score_{Pe} - Score_{Be}} \tag{29}$$

where $Score_{\mathrm{Pr}}$ indicates the forecast score of the proposed method; $Score_{Be}$ indicates the forecast score of the reference method (i.e., Method B1 or B2); and $Score_{Pe}$ indicates the score associated to a perfect forecast (e.g., a perfect forecast has FSS = 1). A skill score of 0.5 means that the proposed approach provides a 50% improvement over the benchmark, whilst a negative score indicates a degradation in performance.

Overall, the proposed method was found to outperform the benchmark methods in all forecast metrics considered across a range of lead times, in both case studies. This is demonstrated in Figure 10 which shows the space and time averaged forecast score against lead time. Forecasts from the proposed approach have better reliability, resolution and accuracy scores than the benchmarks, and are significantly more skilful at longer lead times (e.g., 0.6 MTU, or approximately 3 days). The observed improvements are robust to different dynamical regimes, as indicated in Figure 11 which shows the mean and standard deviation of skill scores computed over the 30 independent simulations. Relative improvements are greatest when comparing to Method B1, where the proposed approach offers a 70% improvement on average based on the RMSE and CRPS, although a sizable improvement of 30% is still apparent when comparing to Method B2 in Case Study 2 (see Figure 11). Forecast ensembles from the proposed approach also have better consistency properties, as shown by the RMS Error versus RMS Spread diagnostic plots (Figure 12) where the points lie closer to the diagonal in the proposed approach.

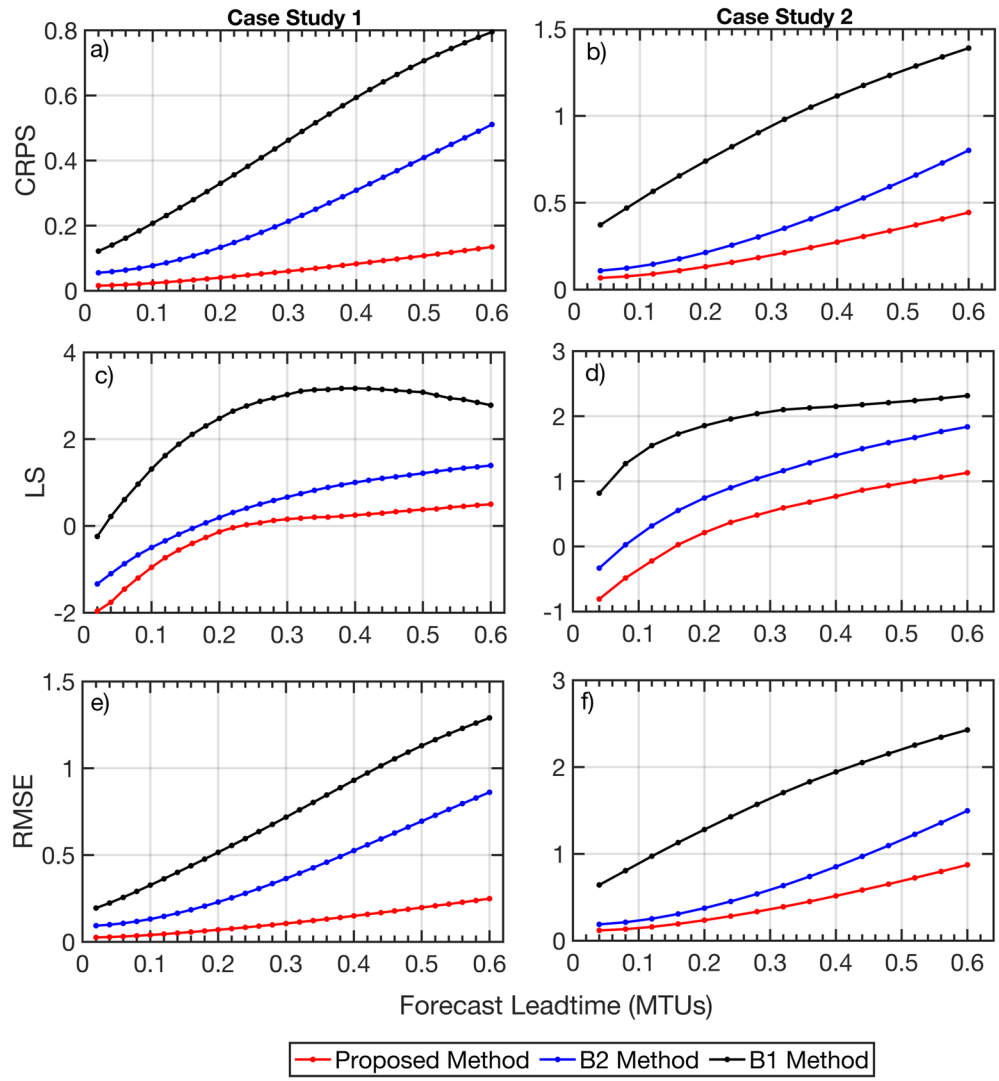

Figure 10. Forecast scores against lead time for both case studies. Scores are presented as averages across space (i.e., over all k variables) and across all simulations. Lower values indicate better performance.

5. Conclusions

Characterizing model error is critical to ensure ensemble Data Assimilation methods produce high quality forecasts and analyses. Accounting for model errors due to unresolved scales is particularly of interest in weather and climate modeling. Numerous stochastic parameterization methods have been proposed for this purpose, although such methods generally rely on data or knowledge of the sub-grid scale processes and/or require observations of all resolved variables. We develop a method that is suited to the more realistic condition where the resolved variables are only partially observed and knowledge of the sub-grid processes is unavailable. It allows for the estimation of complex error structures which depend on known covariates (e.g., state); requires no assumptions or specification or a parametric error distribution (e.g., Gaussian errors); considers the full range of statistical moments (not just bias and covariance); and avoids the need for numerical tuning typical of inflation and localization methods.

The efficacy of the method is demonstrated through numerical experiments on the multi-scale Lorenz 96 model. Comparisons are made to two existing methods that use data assimilation to estimate model errors offline, as these are amenable to the partially observed setting: (a) where the errors are assumed to be Gaussian with mean and covariance estimated from a sample of analysis increments; and (b) where model errors are estimated using

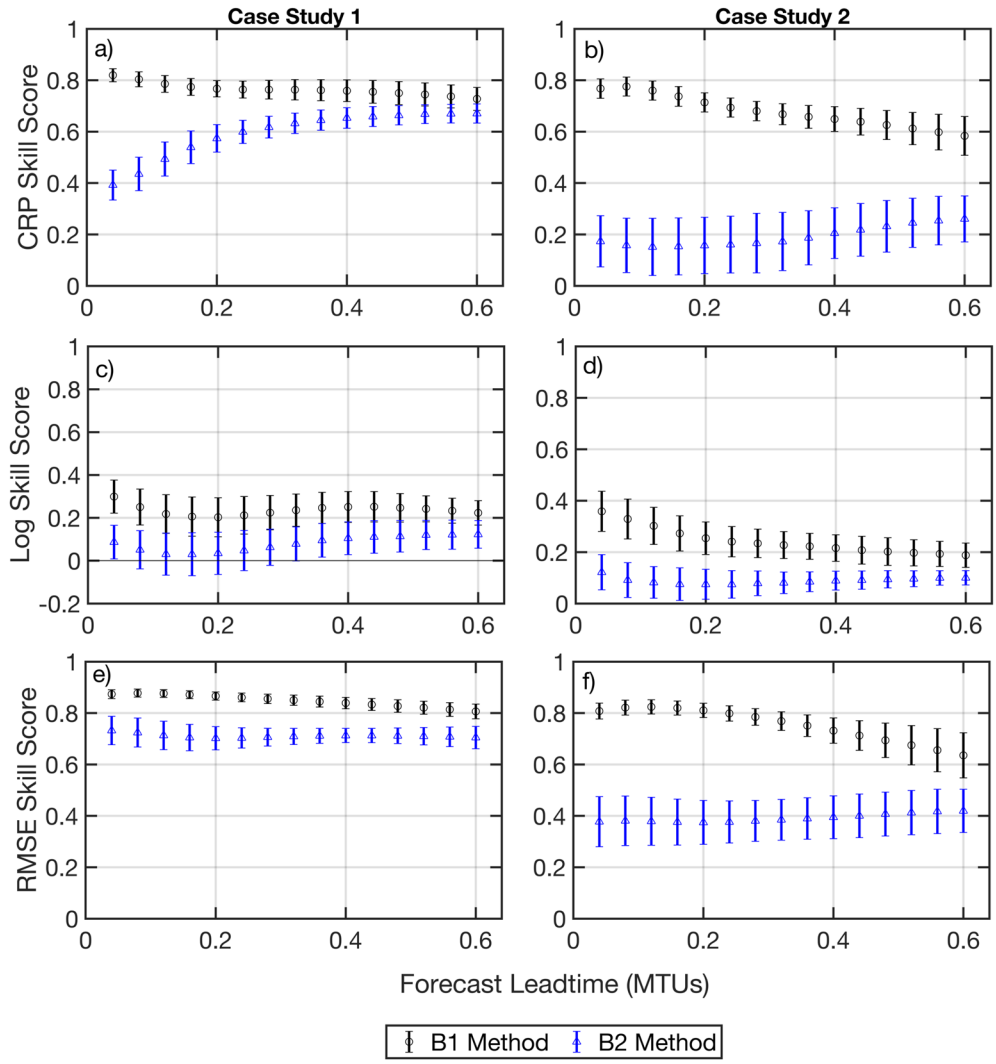

Figure 11. Forecast skill scores against lead time for both case studies. Skill scores are first averaged across space (i.e., over all k variables) and time within each independent simulation. The average of all such values over the 30 independent simulations is shown in the plot (square and triangle markers), as well as the standard deviation. More positive skill scores indicate greater relative improvement of the Proposed method compared to the benchmark method.

long window weak constraint 4D-Var. The proposed approach is shown to recover model errors more precisely than the benchmark methods, thereby making it a more effective parameterization of the sub-grid processes. It is also particularly useful for cases with highly non-Gaussian errors, as considered in this study. Assimilation experiments with the ETKF show that the proposed approach leads to improved forecasts in terms of accuracy, reliability, resolution and consistency. The conditional sum of squares minimization procedure in the proposed method also allows complex error structures to be estimated more precisely than with the least squares type 4D-Var approach. The advantages of accounting for complex state dependent error relationships are also clearly demonstrated by the considerably poorer performance of the constant mean and covariance Gaussian error method.

The proposed method is suited to multi-scale systems where a locality and homogeneity assumption can be made, that is, where errors are influenced by neighboring states instead of the full state vector and the error statistics are the same at each location in space or in parts of the state space with similar dynamics. These assumptions help regularize the ill-posed problem of estimating model errors from partial observations. Future work will investigate systems where such assumptions are inapplicable, although it is expected that other simplifying assumptions would be needed. Finally, the method was applied to a case with negligible observation error, with some

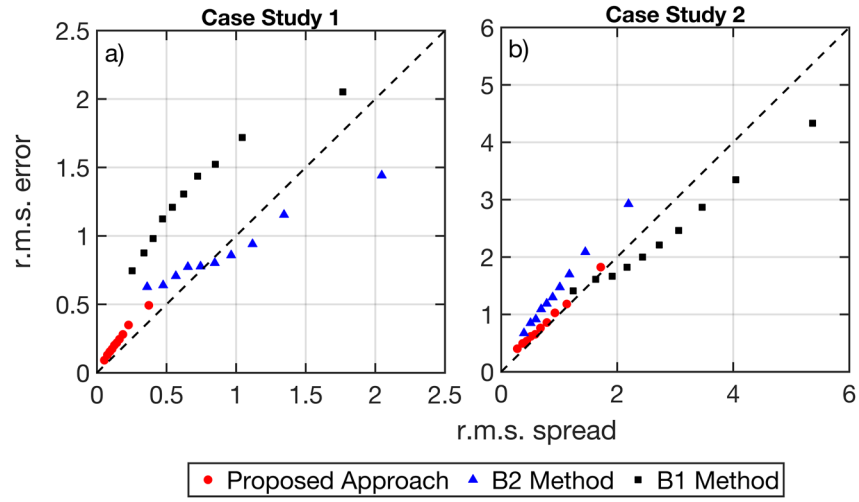

Figure 12. Forecast r.m.s error versus r.m.s spread for both case studies as a measure of statistical consistency. Results are provided for a forecast lead time of 0.6 MTU (≈3 days).

preliminary work including more prominent observation error. Subsequent work will consider the more complex case of estimating model errors from noisy observations.

Data Availability Statement

The implementation of the proposed method and benchmarks on the Lorenz 96 application detailed in Section 4 can be accessed at https://doi.org/10.5281/zenodo.5820227.

Acknowledgments
The research of S. Pathiraja has been partially funded by Deutsche Forschungsgemeinschaft - SFB1294/1–318763901 and by the UNSW Faculty of Engineering Postdoctoral Writing Fellowship. PJvL acknowledges support from the EU-funded ERC grant CUNDA under number 694509. We gratefully acknowledge Professor G. Gottwald and Professor S. Reich for insightful discussions during the development of this work.

References

Anderson, J. L. (2001). An ensemble adjustment kalman filter for data assimilation. *Monthly Weather Review*, *129*(12), 2884–2904. https://doi.org/10.1175/1520-0493(2001)129<2884:aeakff>2.0.co;2

Anderson, J. L. (2007). An adaptive covariance inflation error correction algorithm for ensemble filters. *Tellus*, *59*(2), 210–224. https://doi.org/10.1111/j.1600-0870.2006.00216.x

Anderson, J. L., & Anderson, S. L. (1999). A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. *Monthly Weather Review*, *127*(12), 2741–2758. https://doi.org/10.1175/1520-0493(1999)127<2741:amciot>2.0.co;2

Arnold, H., Moroz, I., & Palmer, T. (2013). Stochastic parametrizations and model uncertainty in the Lorenz' 96 system. *Philosophical Transactions of the Royal Society A*, *371*(1991).

Berry, T., & Harlim, J. (2014). Linear theory for filtering nonlinear multiscale systems with model error. *Proceedings of the Royal Society A*, *470*(2167), 20140168. https://doi.org/10.1098/rspa.2014.0168

Bishop, C. H., Etherton, B. J., & Majumdar, S. J. (2001). Adaptive sampling with the ensemble transform kalman filter. Part I: Theoretical aspects. *Monthly Weather Review*, *129*(3), 420–436. https://doi.org/10.1175/1520-0493(2001)129<0420:aswtet>2.0.co;2

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Bonavita, M., & Laloyaux, P. (2020). Machine learning for model error inference and correction. *Journal of Advances in Modeling Earth Systems*, *12*(12), 1–22. https://doi.org/10.1029/2020MS002232

Brajard, J., Carrassi, A., Bocquet, M., & Bertino, L. (2021). Combining data assimilation and machine learning to infer unresolved scale parametrization. *Philosophical Transactions of the Royal Society A: Mathematical, Physical & Engineering Sciences*, *379*(2194), 20200086. https://doi.org/10.1098/rsta.2020.0086

Brasseur, P., Bahurel, P., Bertino, L., Birol, F., Brankart, J.-M., Ferry, N., et al. (2005). Data assimilation for marine monitoring and prediction: The MERCATOR operational assimilation systems and the MERSEA developments. *Quarterly Journal of the Royal Meteorological Society*, *131*(613), 3561–3582. https://doi.org/10.1256/qj.05.142

Buizza, R., Miller, M., Palmer, T. N., & Milleer, M. (1999). Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, *125*(560), 2887–2908. https://doi.org/10.1002/qj.49712556006

Crommelin, D., & Vanden-Eijnden, E. (2008). Subgrid-scale parameterization with conditional Markov chains. *Journal of the Atmospheric Sciences*, *65*(8), 2661–2675. https://doi.org/10.1175/2008JAS2566.1

Dee, D., & Da Silva, A. (1998). Data Assimilation in the presence of forecast bias. *Quarterly Journal of the Royal Meteorological Society*, *124*(545), 269–295. https://doi.org/10.1016/B978-0-12-088759-0.00022-5

Dee, D. P. (1995). Online estimation of error covariance parameters for atmospheric data assimilation. *Monthly Weather Review*, *123*(4), 1128–1145. https://doi.org/10.1175/1520-0493(1995)123<1128:oleoec>2.0.co;2

Evensen, G. (1994). Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research*, *99*(C5), 10143. https://doi.org/10.1029/94JC00572

Fatkullin, I., & Vanden-Eijnden, E. (2004). A computational strategy for multiscale systems with applications to Lorenz 96 model. *Journal of Computational Physics*, *4121*(2), 605–638. https://doi.org/10.1016/j.jcp.2004.04.013

Fisher, M., Leutbecher, M., & Kelly, G. a. (2005). On the equivalence between Kalman smoothing and weak-constraint four-dimensional variational data assimilation. *Quarterly Journal of the Royal Meteorological Society*, *131*(613), 3235–3246. https://doi.org/10.1256/qj.04.142

Fisher, M., Trémolet, Y., Auvinen, H., Tan, D., & Poli, P. (2011). Weak-constraint and long window 4DVAR. *Technical memorandum*, *655*, 47.

Gagne, D. J., Christensen, H. M., Subramanian, A. C., & Monahan, A. H. (2020). Machine learning for stochastic parameterization: Generative adversarial networks in the Lorenz '96 model. *Journal of Advances in Modeling Earth Systems*, *12*(3). https://doi.org/10.1029/2019MS001896

Hall, P., Racine, J., & Li, Q. (2004). Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association*, *99*(468), 1015–1026. https://doi.org/10.1198/016214504000000548

Hamill, T., & Whitaker, J. (2005). Accounting for the error due to unresolved scales in ensemble data assimilation. *Monthly Weather Review*, *133*(11), 3132–3147. https://doi.org/10.1175/mwr3020.1

Hamill, T., Whitaker, J., & Snyder, C. (2001). Distance-dependent filtering of background error covariance estimates in an ensemble kalman filter. *Monthly Weather Review*, *129*(11), 2776–2790. https://doi.org/10.1175/1520-0493(2001)129<2776:ddfobe>2.0.co;2

Hayfield, T., & Racine, J. (2008). Nonparametric econometrics: The np package. *Journal of Statistical Software*, *27*(5). https://doi.org/10.18637/jss.v027.i05

Houtekamer, P., & Mitchell, H. (2001). A sequential ensemble kalman filter for atmospheric data assimilation. *American Meteorological Society*, *129*(1), 123–137. https://doi.org/10.1175/1520-0493(2001)129<0123:asekff>2.0.co;2

Hyndman, R. J., Bashtannyk, D. M., & Grunwald, G. K. (1996). Estimating and visualizing conditional densities. *Journal of Computational & Graphical Statistics*, *5*(4), 315–336. https://doi.org/10.2307/1390887

Jin, Y., Fu, W., Kang, J., Guo, J., & Gu, J. (2020). *Bayesian symbolic regression*. arXiv:1910.08892 [stat.ME].

Kwasniok, F. (2012). Data-based stochastic subgrid-scale parametrization: An approach using cluster-weighted modelling. *Philosophical Transactions of the Royal Society A: Mathematical, Physical & Engineering Sciences*, *370*(1962), 1061–1086. https://doi.org/10.1098/rsta.2011.0384

Lang, M., Van Leeuwen, P. J., & Browne, P. (2016). A systematic method of parameterisation estimation using data assimilation. *Tellus, Series A: Dynamic Meteorology and Oceanography*, *68*(1), 29012. https://doi.org/10.3402/tellusa.v68.29012

Leith, C. E. (1978). Predictability of climate. *Nature*, *276*(5686), 352–355. https://doi.org/10.1038/276352a0

Levenberg, K. (1944). A method for the solution of certain problems in least-squares. *Quarterly of Applied Mathematics*, *2*, 164–168. https://doi.org/10.1090/qam/10666

Li, H., Kalnay, E., Miyoshi, T., & Danforth, C. M. (2009). Accounting for model errors in ensemble data assimilation. *Monthly Weather Review*, *137*(10), 3407–3419. https://doi.org/10.1175/2009MWR2766.1

Liang, X., Zheng, X., Zhang, S., Wu, G., Dai, Y., & Li, Y. (2012). Maximum likelihood estimation of inflation factors on error covariance matrices for ensemble Kalman filter assimilation. *Quarterly Journal of the Royal Meteorological Society*, *138*(662), 263–273. https://doi.org/10.1002/qj.912

Lorenz, E. (2006). Predictability - A problem partly solved. In T. Palmer, & R. Hagedorn (Eds.), *Predictability of weather and climate*. Cambridge University Press.

Lu, F., Tu, X., & Chorin, A. J. (2017). Accounting for model error from unresolved scales in ensemble Kalman filters by stochastic parametrization. *Monthly Weather Review*, *145*(9), 3709–3723. https://doi.org/10.1175/MWR-D-16-0478.1

Marquardt, D. (1963). An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics*, *11*(2), 431–441. https://doi.org/10.1137/0111030

Mitchell, L., & Carrassi, A. (2015). Accounting for model error due to unresolved scales within ensemble Kalman filtering. *Quarterly Journal of the Royal Meteorological Society*, *141*(689), 1417–1428. https://doi.org/10.1002/qj.2451

Mitchell, L., & Gottwald, G. a. (2012). Data assimilation in Slow?Fast systems using homogenized climate models. *Journal of the Atmospheric Sciences*, *69*(4), 1359–1377. https://doi.org/10.1175/JAS-D-11-0145.1

Miyoshi, T. (2011). The Gaussian approach to adaptive covariance inflation and its implementation with the local ensemble transform kalman filter. *Monthly Weather Review*, *139*(5), 1519–1535. https://doi.org/10.1175/2010MWR3570.1

Palmer, T. (2001). A nonlinear dynamical perspective on model error: A proposal for non-local stochastic-dynamic parameterization in weather and climatic prediction models. *Quarterly Journal of the Royal Meteorological Society*, *127*(572).

Pathiraja, S., Anghileri, D., Burlando, P., Sharma, A., Marshall, L., & Moradkhani, H. (2018a). Insights on the impact of systematic model errors on data assimilation performance in changing catchments. *Advances in Water Resources*, *113*, 202–222. https://doi.org/10.1016/j.advwatres.2017.12.006

Pathiraja, S., Anghileri, D., Burlando, P., Sharma, A., Marshall, L., & Moradkhani, H. (2018b). Time-varying parameter models for catchments with land use change: The importance of model structure. *Hydrology and Earth System Sciences*, *22*(5), 2903–2919. https://doi.org/10.5194/hess-22-2903-2018

Pavliotis, G. A., & Stuart, A. M. (2008). *Multiscale methods: Averaging and homogenization*. Springer-Verlag. https://doi.org/10.1017/CBO9781107415324.004

Rodwell, M., & Palmer, T. (2007). Using numerical weather prediction to assess climate models. *Quarterly Journal of the Royal Meteorological Society*, *133*, 129–146. https://doi.org/10.1002/qj

Saha, S. (1992). Response of the NMC MRF model to systematic-error correction within integration. *Monthly Weather Review*, *120*(2), 345–360. https://doi.org/10.1175/1520-0493(1992)120<0345:rotnmm>2.0.co;2

Sarkka, S. (2013). *Bayesian filtering and smoothing*. Retrieved from http://ebooks.cambridge.org/ref/id/CBO9781139344203

Smith, L. A., Suckling, E. B., Thompson, E. L., Maynard, T., & Du, H. (2015). Towards improving the framework for probabilistic forecast evaluation. *Climatic Change*, *132*(1), 31–45. https://doi.org/10.1007/s10584-015-1430-2

Tippett, M. K., Anderson, J. L., Bishop, C. H., Hamill, T. M., & Whitaker, J. S. (2003). Ensemble square root filters. *Monthly Weather Review*, *131*(7), 1485–1490. https://doi.org/10.1175/1520-0493(2003)131<1485:ESRF>2.0.CO;2

Tremolet, Y. (2006). Accounting for an imperfect model in 4D-Var. *Quarterly Journal of the Royal Meteorological Society*, *132*(621), 2483–2504. https://doi.org/10.1256/qj.05.224

Van Leeuwen, P. J. (1999). The time-mean circulation in the agulhas region determined with the ensemble smoother. *Journal of Geophysical Research*, *104*, 1393–1404. https://doi.org/10.1029/1998jc900012

Van Leeuwen, P. J. (2001). An ensemble smoother with error estimates. *Monthly Weather Review*, *129*(4), 709–728. https://doi.org/10.1175/1520-0493(2001)129<0709:aeswee>2.0.co;2

Vetra-Carvalho, S., van Leeuwen, P. J., Nerger, L., Barth, A., Altaf, U., Brasseur, P., & Beckers, J.-M. (2018). State-of-the-art stochastic data assimilation methods for high-dimensional non-gaussian problems. *Tellus*. https://doi.org/10.1080/16000870.2018.144536

Wang, X., Bishop, C. H., & Julier, S. J. (2004). Which is better, an ensemble of positive-negative pairs or a centered spherical simplex ensemble. *Bulletin of the American Meteorological Society*, *132*(7), 2823–2829. https://doi.org/10.1175/1520-0493(2004)132<1590:wibaeo>2.0.co;2

Wilks, D. (2005). Effects of stochastic parameterizations in the Lorenz 96 system. *Quarterly Journal of the Royal Meteorological Society*, *131*(606), 389–407. https://doi.org/10.1256/qj.04.03

Wouters, J., Dolaptchiev, S., Lucarini, V., & Achatz, U. (2016). Parameterization of stochastic multiscale triads. *Nonlinear Processes in Geophysics*, *23*(6), 435–445. https://doi.org/10.5194/npg-23-435-2016

Zhu, M., van Leeuwen, P. J., & Zhang, W. (2017). Estimating model-error covariances using particle filters. *Quarterly Journal of the Royal Meteorological Society*, *144*(713), 1310–1320. https://doi.org/10.1002/qj.3132

Zupanski, D. (1997). A general weak constraint applicable to operational 4DVAR data assimilation systems. *Monthly Weather Review*, *125*(9), 2274–2292. https://doi.org/10.1175/1520-0493(1997)125<2274:agwcat>2.0.co;2