A Bayesian Approach to Multivariate Adaptive Localization in Ensemble-Based Data Assimilation with Time-Dependent Extensions

Andrey A Popov[1] and Adrian Sandu[1]

[1]Department of Computer Science, Virginia Tech, 2202 Kraft Drive, Blacksburg, VA 24060

Correspondence: Andrey A Popov (apopov@vt.edu)

Abstract. Ever since its inception, the Ensemble Kalman Filter has elicited many heuristic methods that sought to correct it. One such method is localization—the thought that 'nearby' variables should be highly correlated with 'far away' variables not. Recognizing that correlation is a time-dependent property, adaptive localization is a natural extension to these heuristics. We propose a Bayesian approach to adaptive Schur-product localization for the DEnKF, and extend it to support multiple radii of influence. We test both the empirical validity of (multivariate) adaptive localization, and of our approach. We test a simple toy problem (Lorenz'96), extending it to a multivariate model, and a more realistic geophysical problem (1.5 Layer Quasi-Geostrophic). We show that the multivariate approach has great promise on the toy problem, and that the univariate approach leads to improved filter performance for the realistic geophysical problem.

1 Introduction

Data assimilation (Asch et al., 2016; Law et al., 2015; Evensen, 2009; Reich and Cotter, 2015) fuses information from the model forecast states (1) and observations (3) in order to obtain an improved estimation of the truth at any given point in time. Data assimilation approaches include the Ensemble Kalman Filters (EnKF) (Evensen, 1994, 2009; Constantinescu et al., 2007b) that rely on Gaussian assumptions, particle filters for non-Gaussian distributions (Van Leeuwen et al., 2015; Attia et al., 2017; Attia and Sandu, 2015), and variational approaches, rooted in control theory (Le Dimet and Talagrand, 1986; Sandu et al., 2005; Carmichael et al., 2008)

The EnKF is an important family of data assimilation techniques that propagate both the mean and covariance of the state uncertainty (2) through the model (1) using a Monte-Carlo approach. While large dynamical systems of interest have a large number of modes along which errors can grow, the number of ensemble members used to characterize uncertainty remains relatively small due to computational costs. As a result, inaccurate correlation estimates obtained through Monte Carlo sampling can profoundly affect the filter results. Techniques such as covariance localization and inflation have been developed to alleviate these problems (Petrie and Dance, 2010).

Localization techniques take advantage of the fundamental property of geophysical systems that correlations between variables decrease with spatial distance (Kalnay, 2003; Asch et al., 2016). This prior knowledge is used to scale ensemble-estimated covariances between distant variable such as to reduce inaccurate, spurious correlations. The prototypical approach to local-

ization is a Schur-product-based tapering of the covariance matrix (Bickel et al., 2008); theoretical results ensure that the covariance matrices estimated using small ensembles sampled from a multivariate normal probability distribution, upon tapering, approach quickly the true covariance matrix. A formal theory of localization (Flowerdew, 2015) has been attempted, though a true multivariate theory is still out of our grasp. Practical implementations of localization rely on restricting the information flow, either in state space or in observation space, to take place within a given "radius of influence" (Hunt et al., 2007). Some variants of EnKF like the Maximum Likelihood Ensemble Filter (MLEF) (Zupanski, 2005) reduce the need for localization, while others use localization in order to efficiently parallelize the analysis cycles in space (Nino-Ruiz et al., 2015).

The performance of the EnKF algorithms critically depends on the correct choice of localization radii (a.k.a, the decorrelation distances), since values that are too large fail to correct for spurious correlations, while values that are too small throw away important correlation information. However, the physical values of the spatial decorrelation scales are not known apriori, and they change with the temporal and spatial location. At the very least the decorrelation scales depend on the current atmospheric flow. In atmospheric chemistry systems, because of the drastic difference in reactivity, each chemical species has its own individual localization radius (Constantinescu et al., 2007a). Multi-scale schemes (Buehner and Shlyaeva, 2015) for localization are an immediate necessity. Clearly, the widely used approach of estimating decorrelation distances from historical ensembles of simulations, and then using a constant averaged value throughout the space and time domain, lead to a suboptimal performance of Ensemble Kalman filters.

Adaptive localization schemes seek to estimate decorrelation distances from the data, such as to optimize the filter performance according to some criteria. One approach to adaptive localization utilizes an ensemble of ensembles to detect and mitigate spurious correlations (Anderson, 2007). Relying purely on model dynamics and foregoing reliance on spatial properties of the model, the method is very effective for small scale systems, but its applicability to large-scale geophysical problems is unclear. There is, however, evidence that optimal localization depends more on ensemble size and observation properties than on model dynamics (Kirchgessner et al., 2014), and that adaptive approaches whose correlation functions follow these dynamics do not show a significant improvement over conventional static localization (Bishop and Hodyss, 2011). Methods such as sampling error correction (Anderson, 2012) take advantage of these properties to build correction factors and apply them as an ordinary localization scheme. A different approach uses the inherent properties of correlation matrices to construct Smoothed ENsemble COrrelations Raised to a Power (SENCORP) (Bishop and Hodyss, 2007) matrices, that in the limiting case remove all spurious correlations. This method relies purely on the statistical properties of correlation matrices, and ignores the model dynamics and the spatial and temporal dependencies it defines. A more theoretically sound approach (Ménétrier et al., 2015) would try to create a localization matrix such that the end result would better approximate the asymptotic infinite-ensemble covariance matrix. A recent approach considers machine learning algorithms to capture hidden properties in the propagation model that affect the localization parameters (Moosavi et al., 2018).

This work develops a Bayesian framework to dynamically learn the parameters of the Schur-product based localization from the ensemble of model states and the observations during the data assimilation in geophysical systems. Specifically, the localization radii are considered random variables described by parametrized distributions, and are retrieved as part of the assimilation step together with the analysis states. One of the primary goals of this paper is to develop ways in which such

an approach could be extended to both multivariate and time-dependent 4D-esque cases. We prove the approach's empirical validity through a type of idealized variance minimization that has access to the true solution (which we call an Oracle). We then show that the approach provides a more stable result with a much larger initial radius guess. The exploration of the idea is done through the use of several test problems such as that of the Lorenz'96 problem, a multivariate variant thereof that we introduce specifically for this paper, and a more realistic Quasi-Geostrophic model to showcase the applicability of the method to scenarios more in line with operational ones.

The paper is organized as follows. Sect. 2 reviews background material for Ensemble Kalman filtering and Schur-product localization. Sect. 4 describes the proposed theoretical framework for localization and the resulting optimization problems for maximum likelihood solutions. Numerical results with three test problems reported in Sect. 5 provide empirical validation of the proposed approach.

2 Background

We consider a computational model that approximates the evolution of a physical dynamical system such as the atmosphere:

$$\boldsymbol{x}_{i+1} = \mathcal{M}_{i,i+1}(\boldsymbol{x}_i) + \boldsymbol{\xi}_{i,i+1}, \tag{1}$$

The (finite-dimensional) state of the model $\mathbf{x}_i \in \mathbb{R}^n$ at time t_i approximates the (finite-dimensional projection of the) physical true state $\mathbf{x}_i^{\mathrm{t}} \in \mathbb{R}^n$. The computational model (1) is inexact, and we assume the model error to be additive and normally distributed, $\boldsymbol{\xi}_{i,i+1} \sim \mathcal{N}(\boldsymbol{0}, \mathbf{Q}_i)$.

The initial state of the model is also not precisely known, and to model this uncertainty we consider that it is a random variable drawn from a specific probability distribution:

$$\boldsymbol{x}_0 \sim \mathcal{N}(\boldsymbol{x}_0^{\mathrm{t}}, \mathbf{P}_0^{\mathrm{b}}). \tag{2}$$

Consider also observations of the truth,

$$\boldsymbol{y}_{i+1} = \mathcal{H}(\boldsymbol{x}_{i+1}^{\mathrm{t}}) + \boldsymbol{\eta}_{i+1}, \tag{3}$$

that are corrupted by normal observation errors $\boldsymbol{\eta}_{i+1} \sim \mathcal{N}(\boldsymbol{0}, \mathbf{R}_{i+1})$. We consider here the case with a linear observation operator, $\mathcal{H} := \mathbf{H}$.

Consider and ensemble of N states $\mathbf{x} \in \mathbb{R}^{n \times N}$ sampling the probability density that describes the uncertainty in the state at a given time moment (the time subscripts are omitted for simplicity of notation). The ensemble mean, ensemble anomalies, and ensemble covariance are

$$\bar{\boldsymbol{x}} = \frac{1}{N}\mathbf{x}\boldsymbol{1}_N, \quad \mathbf{X} = \mathbf{x} - \bar{\boldsymbol{x}}\boldsymbol{1}_N^{\mathsf{T}}, \quad \mathbf{P} = \frac{1}{N-1}\mathbf{X}\mathbf{X}^{\mathsf{T}}, \tag{4}$$

respectively. Typically N is smaller than the number of positive Lyapunov exponents in our dynamical system, and much smaller than the number of states (Bergemann and Reich, 2010). Consequently, the ensemble statistics (4) are marred by considerable sampling errors. The elimination of sampling errors, manifested as spurious correlations in the covariance matrix (Evensen, 2009), leads to the need for localization.

2.1 Kalman analysis

The mean and the covariance are propagated first through the forecast step. Specifically, each ensemble member is advanced to the current time using the model (1) to obtain the ensemble forecast $\mathbf{x}^{\mathrm{f}}$ (with mean $\bar{\boldsymbol{x}}^{\mathrm{f}}$ and covariance $\mathbf{P}^{\mathrm{f}}$) at the current time. A covariance inflation step $\mathbf{X}^{\mathrm{f}} \leftarrow \alpha \mathbf{X}^{\mathrm{f}}$, $\alpha > 1$, can be applied to prevent filter divergence (Anderson, 2001).

The mean and covariance are then propagated through the analysis step, which fuses information from the forecast mean and covariance and from observations (3), to provide an analysis ensemble $\mathbf{x}^{\mathrm{a}}$ (with mean $\bar{\boldsymbol{x}}^{\mathrm{a}}$ and covariance $\mathbf{P}^{\mathrm{a}}$) at the same time. Here we consider the deterministic EnKF (DEnKF) (Sakov and Oke, 2008), which obtains the analysis as follows:

$$\bar{\boldsymbol{x}}^{\mathrm{a}} = \bar{\boldsymbol{x}}^{\mathrm{f}} + \mathbf{K}\boldsymbol{d}, \tag{5a}$$

$$\mathbf{K} = \mathbf{P}^{\mathrm{f}}\mathbf{H}^{\mathsf{T}}\left(\mathbf{H}\mathbf{P}^{\mathrm{f}}\mathbf{H}^{\mathsf{T}} + \mathbf{R}\right)^{-1}, \tag{5b}$$

$$\boldsymbol{d} = \boldsymbol{y} - \mathbf{H}\bar{\boldsymbol{x}}^{\mathrm{f}}, \tag{5c}$$

$$\mathbf{X}^{\mathrm{a}} = \mathbf{X}^{\mathrm{f}} - \frac{1}{2}\mathbf{K}\mathbf{H}\mathbf{X}^{\mathrm{f}}, \tag{5d}$$

$$\mathbf{x}^{\mathrm{a}} = \bar{\boldsymbol{x}}^{\mathrm{a}}\boldsymbol{1}_N^{\mathsf{T}} + \mathbf{X}^{\mathrm{a}}, \tag{5e}$$

where $\mathbf{K}$ is the Kalman gain matrix and $\boldsymbol{d}$ the vector of innovations. DEnKF is chosen for simplicity of implementation and ease of amending to support Schur product-based localization. However, the adaptive localization techniques discussed herein are general—they do not depend on this choice and are equally applicable to any EnKF algorithm.

2.2 Schur product localization

Covariance localization involves the Schur (element-wise) product between a symmetric positive semi-definite matrix $\boldsymbol{\rho}$ and the ensemble forecast covariance:

$$\mathbf{P}^{\mathrm{f}} \leftarrow \boldsymbol{\rho} \circ \mathbf{P}^{\mathrm{f}}, \quad \mathbf{P}^{\mathrm{f}}_{i,j} \leftarrow \rho_{i,j}\,\mathbf{P}^{\mathrm{f}}_{i,j}. \tag{6}$$

By the Schur product theorem (Schur, 1911), if $\boldsymbol{\rho}$ and $\mathbf{P}^{\mathrm{f}}$ are positive semi-definite, then their Schur product is positive semi-definite. The matrix $\boldsymbol{\rho}$ is chosen such that it reduces the sampling errors and brings the ensemble covariance closer to the true covariance matrix. We note that one can apply the localization in ways that mitigate the problem of storing full covariance matrices (Houtekamer and Mitchell, 2001). Efficient implementation aspects are not discussed further as they do not impact the adaptive localization approaches developed herein.

We seek to generate the entries of localization matrix $\boldsymbol{\rho}$ from a localization function $\ell : \mathbb{R}_{\geq 0} \to [0,1]$, used to represent the weights applied to each individual covariance:

$$\boldsymbol{\rho} = [\ell(d(i,j)/r)]_{1 \leq i,j \leq n}. \tag{7}$$

The function ℓ could be thought of as a regulator which brings the ensemble correlations in line with the physical correlations, which are often compactly supported functions (Gneiting, 2002). The metric d quantifies the physical distance between model

variables, such that $d(i, j)$ represents the spatial distance between the state-space variables x_i and x_j. The "radius of influence" parameter r represents the correlation spatial scale; the smaller r is the faster variables decorrelate with increasing distance.

If the spatial discretization is time-invariant, and ℓ and r are fixed, that the matrix $\boldsymbol{\rho}$ is also time invariant. The goal of the adaptive localization approach is to learn the best value of r dynamically from the ensemble.

A common localization function used in production software is due to Gaspari and Cohn (Gaspari and Cohn, 1999; Gneiting, 2002; Petrie, 2008). Here we use the Gaussian function

$$\ell(u) = \exp\left(-u^2/2\right) \tag{8}$$

to illustrate the adaptive localization strategy. However, our approach is general and can be used with any suitable localization function.

3   Extension to multivariate localization functions

It is intuitively clear that different physical effects propagate spatially at different rates, leading to different correlation distances. Consequently, different state-space variables should be analyzed using different radii of influence. This raises the additional question of how to localize the covariance of two variables when each of them is characterized by a different radius of influence. One approach (Roh et al., 2015) is to use a multivariate compactly supported function (Askey, 1973; Porcu et al., 2013) to calculate the modified error statistics. We however wish to take advantage of already developed univariate compactly supported functions.

We define the mapping operator $\mathfrak{r} : \mathbb{N}_n \to \mathbb{N}_g$ that assigns each state-space component to a group. All state variables assigned to the same group are analyzed using the same localization radius. Assuming a fixed mapping, the localization parameters are the radii values for each group $\boldsymbol{v} \in \mathbb{R}^g$. These values can be tuned independently during the adaptive algorithm. The corresponding prolongation operator $\mathfrak{p} : \mathbb{R}^g \to \mathbb{R}^n$ assigns one of the g radii values to each of the n state-space components:

$$\mathbf{r} = \mathfrak{p}(\boldsymbol{v}). \tag{9}$$

Setting $g = 1$ and $\mathfrak{p}(v) = v \boldsymbol{1}_n$, recovers the univariate localization approach.

Petrie (Petrie, 2008) showed that true square-root filters such as the LETKF (Hunt et al., 2007) are not amenable to Schur product-based localization, therefore they need to rely on sequentially assimilating every observation space variable. Here we wish to combine both the ease-of-use of Schur product-based localization and the utility of multivariate localization techniques.

To this end, we introduce a commutative, idempotent, binary operation, $m : \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$, that computes a "mean localization function value" such as to harmonize the effects of different values of the localization radii. More explicitly, given $0 \leq a \leq b$, m should have the properties that $m(a, a) = a$, $m(b, b) = b$, and $m(a, b) = m(b, a)$. We also impose the additional

common sense property that $a \leq m(a,b) \leq b$. We consider here several simple mean functions m, as follows:

$$m_{\min}(\lambda_i, \lambda_j) = \min\{\lambda_i, \lambda_j\}, \tag{10a}$$

$$m_{\max}(\lambda_i, \lambda_j) = \max\{\lambda_i, \lambda_j\}, \tag{10b}$$

$$m_{\mathrm{mean}}(\lambda_i, \lambda_j) = (\lambda_i + \lambda_j)/2\,, \tag{10c}$$

$$m_{\mathrm{sqrt}}(\lambda_i, \lambda_j) = \sqrt{\lambda_i \lambda_j}, \tag{10d}$$

$$m_{\mathrm{rms}}(\lambda_i, \lambda_j) = \sqrt{\lambda_i^2 + \lambda_j^2}\big/\sqrt{2}\,, \tag{10e}$$

$$m_{\mathrm{harm}}(\lambda_i, \lambda_j) = \frac{2\lambda_i \lambda_j}{\lambda_i + \lambda_j}. \tag{10f}$$

Assume that the variables x_i and x_j have the localization radii r_i and r_j, respectively. We extend the definition of the localization matrix $\boldsymbol{\rho}$ to account for multiple localization radii associated with different variables as follows:

$$\boldsymbol{\rho} = \Big[m\big(\ell(d(i,j)/r_i), \ell(d(j,i)/r_j)\big)\Big]_{1 \leq i,j \leq n} \tag{11}$$

The localization sub-matrix of state-space variables x_i and x_j:

$$\begin{bmatrix} \ell(0) & \rho_{i,j} \\ \rho_{j,i} & \ell(0) \end{bmatrix} \tag{12}$$

is a symmetric matrix for any mean function (10). When ℓ is a univariate compactly supported function, our approach implicitly defines its m-based multivariate counterpart.

One of the common criticisms of a distance assumption on the correlation of geophysical systems is that two variables in close proximity to each other might have very weak correlations. For example, in a model that takes into account the temperature and concentration of stationary cars at any given location, the two distinct types of information might not at all be correlated with each other. The physical distance between the two, however, is zero, and thus any single correlation function will take the value one and do not remove any spurious correlations. One can mitigate this problem by considering univariate localization functions for each pair of components $\ell_{i,j} = \ell_{j,i}$, and extending the localization matrix as follows:

$$\boldsymbol{\rho} = \big[m\big(\ell_{i,j}(d(i,j)/r_i), \ell_{j,i}(d(j,i)/r_j)\big)\big]_{1 \leq i,j \leq n}. \tag{13}$$

We will not analyze multiple localization functions in the paper.

4 Bayesian Approach

We denote the analysis step of the filter by:

$$\mathbf{x}^{\mathrm{a}} = \mathcal{A}(\mathbf{x}^{\mathrm{f}}, \boldsymbol{y}, \boldsymbol{v}), \tag{14}$$

where $\boldsymbol{v}$ are the variable localization and inflation parameters. In this paper, we look solely at varying localization, and will keep inflation constant.

In the Bayesian framework, we consider the localization parameters to be random variables with an associated prior probability distribution. Specifically, we assume that each of the radii $v^{(j)}$ is distributed according to a univariate gamma distribution $v^{(j)} \sim \Gamma(\alpha^{(j)}, \beta^{(j)})$. The gamma probability density:

$$5 \quad f_\Gamma(v; \alpha, \beta) = \frac{\beta^\alpha v^{\alpha-1} e^{-\beta v}}{\Gamma(\alpha)}, \quad v, \alpha, \beta > 0 \tag{15}$$

has mean $\bar{v} = \alpha/\beta$ and variance $\mathrm{Var}(v) = \alpha/\beta^2$. We have chosen the gamma distribution as it is the maximum entropy probability distribution for a fixed mean (Singh et al., 1986), i.e., is the best distribution that one can choose without additional information. It is supported on the interval $(0, \infty)$ and has exactly two free parameters (allowing to control both the mean and variance).

10   The assimilation algorithm computes a posterior (analysis) probability density over the state space considering the probabilities of observations and parameters. We start with Bayes' identity:

$$\pi(\boldsymbol{x}, \boldsymbol{v} \mid \boldsymbol{y}) \propto \pi(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{v})\, \pi(\boldsymbol{x} \mid \boldsymbol{v})\, \pi(\boldsymbol{v})$$
$$= \pi(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{v})\, \pi(\boldsymbol{x} \mid \boldsymbol{v}, \boldsymbol{y})\, \pi(\boldsymbol{v})\, \pi(\boldsymbol{y}). \tag{16}$$

Note that $\pi(\boldsymbol{y})$ is a constant scaling factor. Here $\pi(\boldsymbol{v})$ represents the (prior) uncertainty in the parameters, $\pi(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{v})$ represents the uncertainty in the state for a specific value of the parameters, and $\pi(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{v})$ represents the likelihood of observations with

15   respect to both state and parameters. Under Gaussian assumptions on the state and observation errors equation (16) can be explicitly written out as:

$$\pi(\boldsymbol{x}, \boldsymbol{v} \mid \boldsymbol{y}) \propto \exp\left(-\frac{1}{2} \left\| \mathcal{A}(\mathbf{x}^{\mathrm{f}}, \boldsymbol{y}, \boldsymbol{v}) - \mathbf{x}^{\mathrm{f}} \right\|_{\mathbf{P}^{\mathrm{f}}_{(\boldsymbol{v})}{}^{-1}}^2 \right)$$
$$\exp\left(-\frac{1}{2} \left\| \boldsymbol{y} - \mathbf{H}\mathcal{A}(\mathbf{x}^{\mathrm{f}}, \boldsymbol{y}, \boldsymbol{v}) \right\|_{\mathbf{R}^{-1}}^2 \right)$$
$$f_{\mathcal{P}}(\boldsymbol{v}), \tag{17}$$

with $f_{\mathcal{P}}(\boldsymbol{v})$ given by equation (15). Note that, once the analysis scheme (14) has been chosen, the analysis state becomes a function of $\boldsymbol{v}$, and the conditional probability in (17) represents the posterior probability of $\boldsymbol{v}$.

The negative log likelihood of the posterior probability (17) is:

$$
\begin{aligned}
\mathcal{J}(\boldsymbol{v}) = -\log \Bigg[&\exp\left(-\frac{1}{2}\left\|\mathcal{A}(\mathbf{x}^{\mathrm{f}},\boldsymbol{y},\boldsymbol{v})-\mathbf{x}^{\mathrm{f}}\right\|^{2}_{\mathbf{P}^{\mathrm{f}}_{(\boldsymbol{v})}{}^{-1}}\right) \\
&\exp\left(-\frac{1}{2}\left\|\boldsymbol{y}-\mathbf{H}\mathcal{A}(\mathbf{x}^{\mathrm{f}},\boldsymbol{y},\boldsymbol{v})\right\|^{2}_{\mathbf{R}^{-1}}\right) \\
&\prod_{j=1}^{g} v^{(j),\alpha-1}\exp\left(-\beta^{(j)}v^{(j)}\right)\Bigg] \\
= \;&\tfrac{1}{2}\left\|\mathcal{A}(\mathbf{x}^{\mathrm{f}},\boldsymbol{y},\boldsymbol{v})-\mathbf{x}^{\mathrm{f}}\right\|^{2}_{\mathbf{P}^{\mathrm{f}}_{(\boldsymbol{v})}{}^{-1}} \\
&+ \tfrac{1}{2}\left\|\boldsymbol{y}-\mathbf{H}\mathcal{A}(\mathbf{x}^{\mathrm{f}},\boldsymbol{y},\boldsymbol{v})\right\|^{2}_{\mathbf{R}^{-1}} \\
&+ \sum_{j=1}^{g}\left(\beta^{(j)}v^{(j)}-\left(\alpha^{(j)}-1\right)\log\left(v^{(j)}\right)\right).
\end{aligned}
\tag{18}
$$

5 Through liberal use of the chain rule and properties of symmetric semi-positive definite matrices, the gradient of the cost function with respect to individual localization parameters reads:

$$
\begin{aligned}
\frac{\partial \mathcal{J}}{\partial v^{(j)}} = \;&\tfrac{1}{2}\left\|\mathbf{P}^{\mathrm{f}}_{(\boldsymbol{v})}{}^{-1}\left(\mathcal{A}(\mathbf{x}^{\mathrm{f}},\boldsymbol{y},\boldsymbol{v})-\mathbf{x}^{\mathrm{f}}\right)\right\|^{2}_{\frac{\partial \mathbf{P}^{\mathrm{f}}_{(\boldsymbol{v})}}{\partial v^{(j)}}} \\
&+ \frac{\partial \mathcal{A}(\mathbf{x}^{\mathrm{f}},\boldsymbol{y},\boldsymbol{v})}{\partial v^{(j)}}^{\mathsf{T}}\mathbf{P}^{\mathrm{f}}_{(\boldsymbol{v})}{}^{-1}\left(\mathcal{A}(\mathbf{x}^{\mathrm{f}},\boldsymbol{y},\boldsymbol{v})-\mathbf{x}^{\mathrm{f}}\right) \\
&- \left(\mathbf{H}\frac{\partial \mathcal{A}(\mathbf{x}^{\mathrm{f}},\boldsymbol{y},\boldsymbol{v})}{\partial v^{(j)}}\right)^{\mathsf{T}}\mathbf{R}^{-1}\left(\boldsymbol{y}-\mathbf{H}\mathcal{A}(\mathbf{x}^{\mathrm{f}},\boldsymbol{y},\boldsymbol{v})\right) \\
&+ \beta^{(j)}-\left(\alpha^{(j)}-1\right)\frac{1}{v^{(j)}},
\end{aligned}
\tag{19}
$$

where we took advantage of the properties of symmetric matrices. Note that without the assumption that parameters are independent the gradient would involve higher order tensors.

4.1 Solving the optimization problem

Under the assumptions that the analysis function (14) is based on DEnKF (5), and (the not-quite-correct assumption) that all ensemble members are i.i.d., we obtain:

$$
5 \quad \left\| \mathcal{A}(\mathbf{x}^{\mathrm{f}}, \boldsymbol{y}, \boldsymbol{v}) - \mathbf{x}_i^{\mathrm{f}} \right\|_{\mathbf{P}_{(\boldsymbol{v})}^{\mathrm{f}-1}}^2 = \sum_{e=1}^{N} \left\| \mathbf{K}_{(\boldsymbol{v})} \mathbf{z}^{(:,e)} \right\|_{\mathbf{P}_{(\boldsymbol{v})}^{\mathrm{f}-1}}^2
$$

$$
= \sum_{e=1}^{N} \left\| \mathbf{S}_{(\boldsymbol{v})}^{-1} \mathbf{z}^{(:,e)} \right\|_{\mathbf{H}\mathbf{P}_{(\boldsymbol{v})}^{\mathrm{f}} \mathbf{H}^{\mathsf{T}}}^2
$$

$$
\mathbf{z} = \boldsymbol{d} \boldsymbol{1}_N^{\mathsf{T}} - \tfrac{1}{2} \mathbf{H} \mathbf{X}^{\mathrm{f}}, \tag{20}
$$

$$
\left\| \boldsymbol{y} - \mathbf{H} \mathcal{A}(\mathbf{x}^{\mathrm{f}}, \boldsymbol{y}, \boldsymbol{v}) \right\|_{\mathbf{R}^{-1}}^2 = \sum_{e=1}^{N} \left\| \mathbf{g}_{(\boldsymbol{v})}^{(:,e)} \right\|_{\mathbf{R}^{-1}}^2
$$

$$
\mathbf{g}_{(\boldsymbol{v})} = \left(\mathbf{I} - \mathbf{H} \mathbf{K}_{(\boldsymbol{v})} \right) \boldsymbol{d} \boldsymbol{1}_N^{\mathsf{T}}
$$

$$
- \mathbf{H} \mathbf{X}^{\mathrm{f}}
$$

$$
+ \tfrac{1}{2} \mathbf{H} \mathbf{K}_{(\boldsymbol{v})} \mathbf{H} \mathbf{X}^{\mathrm{f}},
$$

$$
\mathbf{K}_{(\boldsymbol{v})} = \mathbf{P}_{(\boldsymbol{v})}^{\mathrm{f}} \mathbf{H}^{\mathsf{T}} \mathbf{S}_{(\boldsymbol{v})}^{-1},
$$

$$
\mathbf{S}_{(\boldsymbol{v})} = \mathbf{H} \mathbf{P}_{(\boldsymbol{v})}^{\mathrm{f}} \mathbf{H}^{\mathsf{T}} + \mathbf{R}. \tag{21}
$$

Combining (18), (20), and (21) leads to the cost function form:

$$
10 \quad \mathcal{J}(\boldsymbol{v}) = \sum_{e=1}^{N} \left[\tfrac{1}{2} \left\| \mathbf{S}_{(\boldsymbol{v})}^{-1} \mathbf{z}^{(:,e)} \right\|_{\mathbf{H}\mathbf{P}_{(\boldsymbol{v})}^{\mathrm{f}} \mathbf{H}^{\mathsf{T}}}^2 + \tfrac{1}{2} \left\| \mathbf{g}_{(\boldsymbol{v})}^{(:,e)} \right\|_{\mathbf{R}^{-1}}^2 \right]
$$

$$
+ \sum_{j=1}^{g} \left(\beta^{(j)} v^{(j)} - \left(\alpha^{(j)} - 1 \right) \log \left(v^{(j)} \right) \right), \tag{22}
$$

which only requires the computation of the background covariance matrix $\mathbf{H}\mathbf{P}_{(\boldsymbol{v})}^{\mathrm{f}} \mathbf{H}^{\mathsf{T}}$ in observation space, thereby reducing the amount of computation and storage. The equivalent manipulations of the gradient lead to:

$$
\frac{\partial \mathcal{J}}{\partial v^{(j)}} = \sum_{e=1}^{N} \left[\tfrac{1}{2} \left\| \mathbf{S}_{(\boldsymbol{v})}^{-1} \mathbf{z}^{(e)} \right\|_{\mathbf{H} \frac{\partial \mathbf{P}_{(\boldsymbol{v})}^{\mathrm{f}}}{\partial v^{(j)}} \mathbf{H}^{\mathsf{T}}}^2 \right.
$$

$$
- \mathbf{z}^{\mathsf{T}} \mathbf{H} \frac{\partial \mathbf{P}_{(\boldsymbol{v})}^{\mathrm{f}}}{\partial v^{(j)}} \mathbf{H}^{\mathsf{T}} \mathbf{S}_{(\boldsymbol{v})}^{-2} \mathbf{H} \mathbf{P}_{(\boldsymbol{v})}^{\mathrm{f}} \mathbf{H}^{\mathsf{T}} \mathbf{S}_{(\boldsymbol{v})}^{-1} \mathbf{z}
$$

$$
\left. + \mathbf{H} \frac{\partial \mathbf{K}_{(\boldsymbol{v})}}{\partial v^{(j)}} \left(\tfrac{1}{2} \mathbf{H} \mathbf{X}^{\mathrm{f}} - \boldsymbol{d} \boldsymbol{1}_N^{\mathsf{T}} \right) \right]
$$

$$
15 \qquad\qquad + \beta^{(j)} - \left(\alpha^{(j)} - 1 \right) \frac{1}{v^{(j)}}, \tag{23}
$$

$$
\frac{\partial \mathbf{K}_{(\boldsymbol{v})}}{\partial v^{(j)}} = \frac{\partial \mathbf{P}_{(\boldsymbol{v})}^{\mathrm{f}}}{\partial v^{(j)}} \mathbf{H}^{\mathsf{T}} \mathbf{S}_{(\boldsymbol{v})}^{-1}
$$

$$
- \mathbf{P}_{(\boldsymbol{v})}^{\mathrm{f}} \mathbf{H}^{\mathsf{T}} \mathbf{H} \frac{\partial \mathbf{P}_{(\boldsymbol{v})}^{\mathrm{f}}}{\partial v^{(j)}} \mathbf{H}^{\mathsf{T}} \mathbf{S}_{(\boldsymbol{v})}^{-2}, \tag{24}
$$

where $\frac{\partial \mathbf{P}^{\mathrm{f}}_{(\boldsymbol{v})}}{\partial v^{(j)}} = \frac{\partial \boldsymbol{\rho}_{(\boldsymbol{v})}}{\partial v^{(j)}} \circ \mathbf{P}^{\mathrm{f}}$. Calculation of the gradient (23) requires computations only in observation space.

One will note that the form of our cost function (22) is similar to that of other hybrid approaches (Bannister, 2017) to data assimilation such as 3DEnVar (Hamill and Snyder, 2000). As a lot of similar computation is done, this technique could potentially be used as a preprocessing step in a larger hybrid data assimilation scheme.

The choice of using the DEnKF is a bit arbitrary. From the above, however, it is evident that a method that decouples the anomaly updates from the mean updates would most likely be more advantageous. A perturbed observation EnKF does not have this property, thus would incur significantly more computational effort to optimize the cost function. Extending this idea to true a square-root filter, like the ETKF, would require significant algebraic manipulation, and heuristics which are outside the scope of this paper.

4.2 Adaptive localization via a time-distributed cost function

We now seek to extend the 3D-Var-like cost function (18) to a time-dependent 4D-Var-like version. As the ensemble is essentially a reduced order model, we do not expect the accuracy of the forecast to hold over the long term as we might in traditional 4D-Var. We thereby propose a limited extension of the cost-function to include a small number K of additional future times. Assuming that we are currently at the i-th time moment, the extended cost function reads:

$$
\mathcal{J}_i(\boldsymbol{v}) = \sum_{e=1}^{N} \left[\tfrac{1}{2} \left\| \mathbf{S}^{-1}_{(\boldsymbol{v}),i} \mathbf{z}_i^{(:,e)} \right\|^2_{\mathbf{H}\mathbf{P}^{\mathrm{f}}_{(\boldsymbol{v}),i}\mathbf{H}^{\mathsf{T}}} \right.
$$
$$
+ \tfrac{1}{2} \left\| \mathbf{g}_{(\boldsymbol{v}),i}^{(:,e)} \right\|^2_{\mathbf{R}_i^{-1}}
$$
$$
\left. + \sum_{k=1}^{K} \tfrac{1}{2} \left\| \boldsymbol{y}_{i+k} - \mathbf{H}\mathbf{x}_{(\boldsymbol{v}),i+k}^{f,(:,e)} \right\|^2_{\mathbf{R}_{i+k}^{-1}} \right]
$$
$$
+ \sum_{j=1}^{g} \left(\beta_i^{(j)} v^{(j)} - \left(\alpha_i^{(j)} - 1 \right) \log \left(v^{(j)} \right) \right). \tag{25}
$$

One will notice that in the 4D part of the cost function the future forecasts are now dependent on $\boldsymbol{v}$ as they are obtained by running the model from $\mathbf{x}^{\mathrm{a}}_{i,(\boldsymbol{v})}$. Similar to some 4D ensemble approaches, the gradient computations can be approximated by the tangent linear model with the adjoint not being required. In fact all that is required is matrix vector products which can be approximated with finite differences (Tranquilli et al., 2017).

Various 4D-type approximation strategies are also applicable to this cost function extension, though are outside of the scope of this paper.

4.3 Algorithm

In practice, instead of dealing with the Gamma distribution parameters of α and β, we use the parameters $\bar{v}$ and $\mathrm{Var}(v)$, such that $\alpha = \frac{\bar{v}^2}{\mathrm{Var}(v)}$ and $\beta = \frac{\bar{v}}{\mathrm{Var}(v)}$. For the sake of simplicity we assume that all $\boldsymbol{v}$ are identically distributed, but this is not required for the algorithm to function. The initial guess for our minimization procedure is the vector of means. After minimizing the

cost function, the radii for different components will be different. These radii, along with the corresponding localization and m functions, are used to build the localization matrix $\boldsymbol{\rho}$. An outline is presented in Algorithm 1.

Algorithm 1 DEnKF Adaptive Localization Algorithm

Require: $[\bar{\upsilon}^{(j)}]_j, [\mathrm{Var}(\upsilon^{(j)})]_j$

1: $\alpha^{(j)} \leftarrow \frac{\bar{\upsilon}^{(j),2}}{\mathrm{Var}(\upsilon^{(j)})}$

2: $\beta^{(j)} \leftarrow \frac{\bar{\upsilon}^{(j)}}{\mathrm{Var}(\upsilon^{(j)})}$

3: $\boldsymbol{\upsilon}_0^* \leftarrow [\bar{\upsilon}^{(j)}]_j$

4: $\boldsymbol{\upsilon}^* \leftarrow \mathrm{argmin}_{\boldsymbol{\upsilon}} \mathcal{J}(\boldsymbol{\upsilon})$ {Eq (22)}

5: $\boldsymbol{r}^* \leftarrow \mathfrak{p}(\boldsymbol{\upsilon}^*)$ {Eq (9)}

6: $\boldsymbol{\rho} \leftarrow [m(l_i, l_j)]_{1 \leq i,j \leq n}$ {Eq (13)}

7: **with**

8: $\quad l_i \leftarrow \ell(d(i,j)/r^{(i),*})$ {Possibly Eq (8)}

9: $\quad l_j \leftarrow \ell(d(i,j)/r^{(j),*})$

10: **return** $\boldsymbol{\rho}$

---

5 Numerical Experiments and Results

5   In order to validate our methodology we carry out twin experiments under the assumption of identical perfect dynamical systems for both the truth and the model. The analysis accuracy is measured by the spatio-temporally averaged root mean square error:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n \cdot n_t} \sum_{i=1}^{n_t} \sum_{j=1}^{n} \left([\boldsymbol{x}_i^{\mathrm{t}}]_j - [\bar{\boldsymbol{x}}_i^{\mathrm{a}}]_j\right)^2}, \tag{26}$$

where n_t is the number of data assimilation cycles (the number of analysis steps).

10   For each of the models we repeat the experiments with different values of the inflation constant α and report the best (minimal RMSE) results.

All initial value problems used as were independently implemented

5.1 Oracles

We will make use of oracles to empirically evaluate the performance of the multivariate approach to Schur product localization.

15   An oracle is an idealized procedure that produces close to optimal results by making use of all the available information, some of which is unknown to the data assimilation system. In our case the oracle minimizes cost functions involving the true model state. Specifically, in an ideal filtering scenario one seeks to minimize the error of the analysis with respect to the truth, i.e., the cost function $\mathcal{J}(\boldsymbol{x}^{\mathrm{a}}) = \mathrm{RMSE}(\boldsymbol{x}^{\mathrm{a}} - \boldsymbol{x}^{\mathrm{t}})$. Our oracle will pick the best parameters, in this case the radii, that minimize the ideal

cost function $\mathcal{J}(\boldsymbol{v}) = \mathrm{RMSE}(\bar{\boldsymbol{x}}^{\mathrm{a}}_{(\boldsymbol{v})} - \boldsymbol{x}^{\mathrm{t}})$. This can be viewed as a ideal variance minimization of the state space in parameter space.

5.2 Lorenz'96

5 The 40 variable Lorenz model (Lorenz, 1996; Lorenz and Emanuel, 1998) is the first standard test employed. This problem is widely used in the testing of data assimilation algorithms.

5.2.1 Model Setup

The Lorenz'96 model equations:

$$\frac{dx_i}{dt} = (x_{i+1} - x_{i-2})x_{i-1} - x_i + F, \quad i = 1, \dots n, \tag{27}$$

10 are obtained through a coarse discretization of a forced Burger's equation with Newtonian damping (Reich and Cotter, 2015). We impose $x_0 = x_n$, $x_{-1} = x_{n-1}$, and $x_{n+1} = x_1$, where $n = 40$. We take the canonical value for the external forcing factor, $F = 8$. Using known techniques for dynamical systems (Parker and Chua, 2012) one can calculate that this particular system has 13 positive Lyapunov exponents (Strogatz, 2014) and one zero exponent, with a fractal dimension of approximately 27.1.

The initial conditions used for experiments are obtained by starting with

15
$$[\boldsymbol{x}_0]_i = \begin{cases} 8 & i \neq 20 \\ 8.008 & i = 20 \end{cases}, \tag{28}$$

and integrating (27) forward for one time unit in order to reach the attractor.

The physical distance between x_i and x_j is the shortest cyclic distance between any two state variables:

$$d(i,j) = \min\{|i - j|, |n + i - j|, |n + j - i|\}, \tag{29}$$

where the distance between two neighbors is one.

20 For the numerical experiments, we consider a perfect model and noisy observations. We take a six hours assimilation window (corresponding to $\Delta t_{\mathrm{obs}} = 0.05$ model time units), and calculate the RMSE on the assimilation cycle interval $[100, 1100]$ in the case of oracle testing, or $[5000, 55000]$ in the case of adaptive localization testing. Lorenz '96 is an ergodic system (Fatkullin and Vanden-Eijnden, 2004) and therefore its behavior in time for most initial conditions is the same as the behavior of all its possible phase spaces on any orbit around a strange attractor at any point in time, meaning that a long enough time averaged 25 run should be the same as a collection of shorter space averaged runs. We use 10 ensemble members (just under the number of positive Lyapunov exponents), and observe the thirty model variables $(2, 4, \dots, 18, 20, 21, \dots, 39, 40)$, with an observation error variance of one for each observed state.

5.2.2 Oracle Results

Figure 1 shows a visualization of multivariate oracle runs for the Lorenz'96 test problem using best constant radius, univariate oracle, and multivariate oracle utilizing the m-functions from (10). As one can see, the univariate oracle performs no better

than a constant radius, while several of the multivariate approaches provide much better results. This indicates that the problem is better suited for multivariate localization. The worst m-function results are given by $m_{\min}$. The other functions perform similarly, and in further experiments we will only consider m_{mean} as a representative and easy to implement option. We note that for Lorenz-96 with DEnKF and our experimental setup, no 3D univariate adaptive localization scheme beats the best constant localization radius.

We also test both the validity of arbitrarily grouping the radii and the validity of using a time-distributed cost function ((25)). Figure 2 presents results for arbitrary radii groupings and for a limited run 4D approach. There is significant benefit in using more radii groupings, but marginal benefit from the 4D approach.

5.2.3 Adaptive Localization Results

Adaptive localization results for Lorenz'96 are shown in figure 3. The optimal constant localization radius was found and a search of possible input mean and variance values was performed around it for the adaptive case. As accurately predicted by the oracle results, a univariate adaptive approach for this model cannot do better than the best univariate radius (figure 1), as no meaning reduction in error was detected.

5.3 Multivariate Lorenz'96

The canonical Lorenz'96 model is ill suited for multivariate adaptive localization as each variable in the problem behaves identically too all the others. This means that for any univariate localization scheme a constant radius is close to optimal.

5.3.1 Model Setup

We modify the problem in such a way that the average behavior remains very similar to that of the original model, but that instantaneous behavior requires different localization radii. In order to accomplish this we use the time-dependent forcing function that is different for each variable:

$$[\boldsymbol{F}(t)]_i = 8 + 4\cos\left(\omega\left(t + \frac{(i-1) \mod q}{q}\right)\right), \tag{30}$$

where $\omega = 2\pi$ (in the context of Lorenz'96 the equivalent period is 5 days), i is the variable index, and q is an integer factor of n. Here we set set $q = 4$. The different forcing values will create differences between the dynamics of different variables.

For each individual variable the forcing value cycles between 4 and 12, with an average value of 8, just like in the canonical Lorenz'96 formulation. If taken to be a constant, the forcing factor of 4 will make the equation lightly chaotic with only one positive Lyapunov exponent, whilst a constant value of 12 will make the dynamics to have about 15 positive Lyapunov exponents. Our modified system still has the same average behavior with 13 positive Lyapunov exponents. The mean doubling times of the two problems are also extremely similar at around 0.36. This is the ideal desired behavior. Figure 4 shows a comparison of numerically computed covariance matrices for this modified problem. This showcases that an adaptive approach to the instantaneous covariance is required.

Data assimilation will be carried out over the interval $[500, 5500]$. The rest of the problem setup is identical to that of the canonical Lorenz'96.

5.3.2 Adaptive Localization Results

5　Figure 5 shows results with the Multivariate Lorenz'96 model, with $g = q = 4$ radii groups, and the 4D parameter set to $K = 1$. As before, the results for the constant radius were their optimal for each given inflation value, while the adaptive results were obtained through a search of possible means and variances around that value. The maximal improvement is error is only about 8%, however this is leagues more than that of the univariate Lorenz '96. In the canonical Lorenz '96 there is no meaningful choice of grouping other than arbitrary, but in this case, the groups were chosen such that all related variables have the same
10　forcing from equation (30). The results show a significant improvement over the univariate case, especially for low inflation values. We note that the filter spin-uptakes significantly longer for the adaptive localization case than for the constant univariate case. Consequently, the assimilation cycles are chosen in the time interval $[500, 5500]$ units. An idea to mitigate this might be to run the filter with a constant radius for a few assimilation cycles before switching to the adaptive localization strategy, such as to allow the filter to quickly catch up with the shadow attractor trajectory.

15　Figure 6 demonstrates a sampling of the radii obtained through the adaptive 4D multivariate method. For the post-spinup case, the means (5.8449, 5.8669, 5.8441, 5.8602) of the radii for each grouping are fairly similar, however the variances (0.0300, 0.0731, 0.0295, 0.0821) differ wildly. The groupings will less observations, group 1 and 3, have much more conservative variances, while groups 3 and 4, which are observed fully, have much greater variances.

This gives us insight into a potential way of choosing multivariate localization groups. Based on some measure of the
20　observability of any given state-space variable, similarly 'observable' state-space variables should have similar radii.

Tightly coupled models like the multivariate Lorenz'96 have rapidly diverging solutions, and constraining them requires more information about the underlying dynamics. Incorporating future observations and adding degrees of freedom to the cost function increase the performance of our analysis. In the limiting case of one radius per variable and general information from the future one approaches a variant of 4DenVar, which is in principle superior to any pure filtering method.

25　## 5.4 Quasi-geostrophic model (QGSO)

The 1.5 layer quasi-geostrophic model of Sakov and Oke (Sakov and Oke, 2008), obtained by non-dimensionalizing potential vorticity, is given by the equations:

$$q_t = -\psi_x - \varepsilon J(\psi, q) - A \triangle^3 \psi + 2\pi \sin(2\pi y),$$
$$\triangle \psi - F\psi = q, \quad J(\psi, q) \equiv \psi_x q_y - \psi_y q_x. \tag{31}$$

30　The variable ψ can be thought of as the stream function, and the spatial domain is the square $(x, y) \in [0, 1]^2$. The constants are $F = 1600$, $\varepsilon = 10^{-5}$, and $A = 2 \times 10^{-11}$. We use homogeneous Dirichlet boundary conditions.

A second order central finite difference spatial discretization of the Laplacian operator $\triangle$ is performed over the interior of a 129×129 grid, leading to a model dimension $n = 127^2 = 16,129$. The time discretization is the canonical fourth-order explicit

14

Runge-Kutta method with a timestep of 1 time units. The Helmholtz equation on the right-hand side of (31) is solved by an offline pivoted sparse Cholesky decomposition. J is calculated via the canonical Arakawa approximation (Arakawa, 1966; Ferguson, 2008). The $\triangle^3$ operator is implemented by repeated application of our discrete Laplacian.

5   The time between consecutive observations is 5 time units, and the model is run for 3300 such cycles. The first 300 cycles, corresponding to the filter spin-up, are discarded, therefore the assimilation interval is $[300, 3300]$ time units. Observations are performed with a standard 300 component observation operator, as shown in Figure 7. An observation error variance of 4 is taken for each component. The physical distance between two component is defined as:

$$d(i,j) = \sqrt{(i_x - j_x)^2 + (i_y - j_y)^2},\tag{32}$$

10   with (i_x, i_y) and (j_x, j_y) are the spatial grid coordinates of the state space variables i and j, respectively.

    Our rough estimate of the number of positive Lyapunov exponents of this model is $1,451$, with a fractal dimension estimate of 6573.4, thus we will take a conservative 25 ensemble members whose initial states are derived from a random sampling of a long run of the model.

    This model has been tested extensively with both the DEnKF and with various localization techniques (Sakov and Bertino,
15   2011; Bergemann and Reich, 2010; Moosavi et al., 2018).

5.4.1   Adaptive Localization Results

The adaptive localization results for the Quasi-geostrophic problem are shown in Figure 8. As before a constant best univariate localization radius was calculated for each inflation value, and was used as a seed for varying the results in the adaptive case. An improvement of RMSE of up to about 33% can be seen for certain inflation values.

20   One can readily see that for certain values of the inflation factor the adaptive localization procedure results in significant reductions in analysis error, while for other values no significant benefits are observed.

    The empirical utility of the adaptive localization technique is further analyzed in Figure 9 which compares the error results of a suboptimal constant radius with that of an adaptive run with the mean parameter set to the same values as the constant ones. The adaptive results are—except in a few cases of filter divergence—always as good or better than their constant localization
25   counterparts, with an improvement in error as large as 33% with the same mean radius as the constant radius, and as much as 50% with different radii. Even in the case where the adaptive filter diverged, the constant localization filter diverged as well. This indicates that our localization scheme is significantly better than a corresponding sub-optimal constant scheme with the same parameters, as is typically the case in real-world production codes. This opens up the possibility of adapting existing systems that use a conservative suboptimal constant localization to an adaptive localization scheme.

30   Figure 10 shows a sample of the radii obtained by the adaptive algorithm. The initial period of algorithm spin-up is clearly visible. One notes that the adaptive scheme can discern—on average—whether the observations or the model are to be trusted more.

6 Conclusions

This paper proposes a novel Bayesian approach to adaptive Schur product-based localization. A multivariate approach is developed, where multiple radii corresponding to different types of variables are taken into account. The Bayesian approach is solved by constructing 3D and 4D cost functions, and minimizing them to obtain the maximum aposteriori estimates of the localization radii. We show that in the case of the DEnKF these cost functions and their gradients are computationally inexpensive to evaluate, and can be relatively easily implemented within existing frameworks. We provide a new approach for assessing the performance of adaptive localization approaches through the use of restricted cost function oracles.

The adaptive localization approach is tested using the Lorenz'96 and the Quasi-Geostrophic models. Somewhat surprisingly, the adaptivity produces better results for the larger Quasi-Geostrophic problem. This may be due to the ensemble analysis anomaly independence assumption made in Sect. 4.1, an assumption that holds better for a large system with sparse observations than for a small tightly coupled system with dense observations. The performance of the adaptive approach on the small, coupled Lorenz'96 system is increased by using multivariate and 4D extensions of the cost function.

We believe that the algorithm presented herein has a strong potential to improve existing geophysical data assimilation systems that use ensemble based filters such as the DEnKF. In order to avoid filter divergence in the long term, these systems often use a conservative localization radius and a liberal inflation factor. The QG model results indicate that, in such cases, our adaptive method outperforms the approach based on a constant localization. The new adaptive methodology can replace the existing approach with a relatively modest implementation effort.

Future work to extend the methodology includes finding good approximations of the probability distribution of the localization parameters, perhaps through a Machine Learning approach, and reducing the need of assumption that the ensemble members are independent identically distributed random variables.

Applying this methodology to a larger operational model, like the Weather Research and Forecasting Model (WRF) (Skamarock et al., 2008) would require minimal effort and is computationally feasible in the short term, and is a possible future direction.
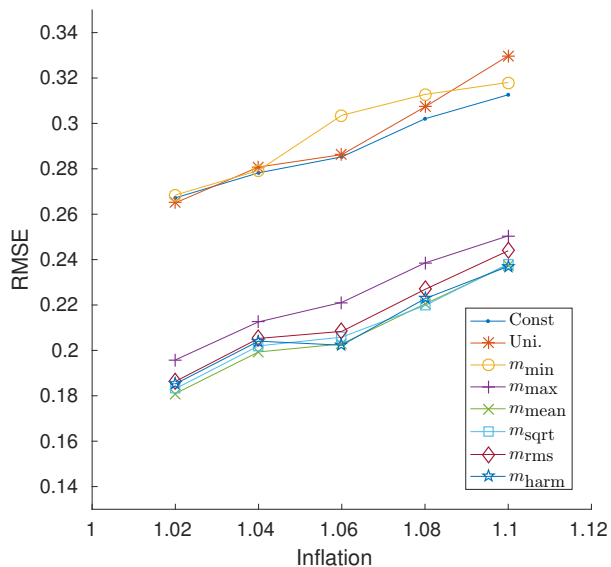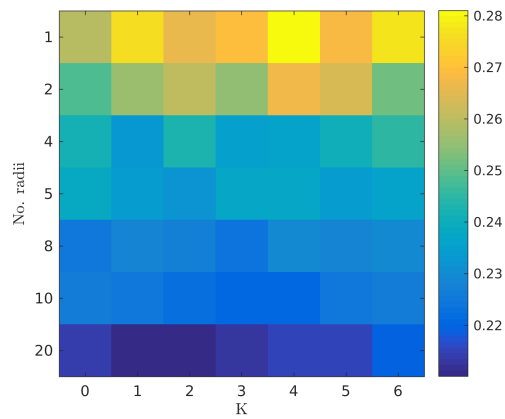
References

Anderson, J. L.: An ensemble adjustment Kalman filter for data assimilation, Monthly weather review, 129, 2884–2903, 2001.

Anderson, J. L.: Exploring the need for localization in ensemble data assimilation using a hierarchical ensemble filter, Physica D: Nonlinear Phenomena, 230, 99–111, 2007.

5

Anderson, J. L.: Localization and sampling error correction in ensemble Kalman filter data assimilation, Monthly Weather Review, 140, 2359–2371, 2012.

Arakawa, A.: Computational design for long-term numerical integration of the equations of fluid motion: Two-dimensional incompressible flow. Part I, Journal of Computational Physics, 1, 119–143, 1966.

10   Asch, M., Bocquet, M., and Nodet, M.: Data assimilation: methods, algorithms, and applications, SIAM, 2016.

Askey, R.: Radial characteristics functions, Tech. rep., Wisconsin University Madison Mathematics Research Center, 1973.

Attia, A. and Sandu, A.: A hybrid Monte Carlo sampling filter for non-gaussian data assimilation, AIMS Geosciences, 1, 41–78, 2015.

Attia, A., Ştefănescu, R., and Sandu, A.: The reduced-order hybrid Monte Carlo sampling smoother, International Journal for Numerical Methods in Fluids, 83, 28–51, 2017.

15   Bannister, R.: A review of operational methods of variational and ensemble-variational data assimilation, Quarterly Journal of the Royal Meteorological Society, 143, 607–633, 2017.

Bergemann, K. and Reich, S.: A localization technique for ensemble Kalman filters, Quarterly Journal of the Royal Meteorological Society, 136, 701–707, 2010.

Bickel, P. J., Levina, E., et al.: Regularized estimation of large covariance matrices, The Annals of Statistics, 36, 199–227, 2008.

20   Bishop, C. H. and Hodyss, D.: Flow-adaptive moderation of spurious ensemble correlations and its use in ensemble-based data assimilation, Quarterly Journal of the Royal Meteorological Society, 133, 2029–2044, 2007.

Bishop, C. H. and Hodyss, D.: Adaptive ensemble covariance localization in ensemble 4D-VAR state estimation, Monthly Weather Review, 139, 1241–1255, 2011.

Buehner, M. and Shlyaeva, A.: Scale-dependent background-error covariance localisation, Tellus A: Dynamic Meteorology and Oceanography, 67, 28 027, 2015.

25

Carmichael, G., Chai, T., Sandu, A., Constantinescu, E., and Daescu, D.: Predicting air quality: improvements through advanced methods to integrate models and measurements, Journal of Computational Physics, 227, 3540–3571, https://doi.org/10.1016/j.jcp.2007.02.024, http://dx.doi.org/10.1016/j.jcp.2007.02.024, 2008.

Constantinescu, E., Sandu, A., Chai, T., and Carmichael, G.: Ensemble-based chemical data assimilation. II: Covariance localization, Quarterly Journal of the Royal Meteorological Society, 133, 1245–1256, https://doi.org/10.1002/qj.77, http://dx.doi.org/10.1002/qj.77, 2007a.

30

Constantinescu, E., Sandu, A., Chai, T., and Carmichael, G.: Ensemble-based chemical data assimilation. I: General approach, Quarterly Journal of the Royal Meteorological Society, 133, 1229–1243, https://doi.org/10.1002/qj.76, http://dx.doi.org/10.1002/qj.76, 2007b.

Evensen, G.: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics, Journal of Geophysical Research: Oceans, 99, 10 143–10 162, 1994.

35   Evensen, G.: Data assimilation: the ensemble Kalman filter, Springer Science & Business Media, 2009.

Fatkullin, I. and Vanden-Eijnden, E.: A computational strategy for multiscale systems with applications to Lorenz 96 model, Journal of Computational Physics, 200, 605–638, 2004.

Ferguson, J.: A numerical solution for the barotropic vorticity equation forced by an equatorially trapped wave, Master's thesis, University of Victoria, 2008.

Flowerdew, J.: Towards a theory of optimal localisation, Tellus A: Dynamic Meteorology and Oceanography, 67, 25 257, 2015.

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

Gneiting, T.: Compactly supported correlation functions, Journal of Multivariate Analysis, 83, 493–508, 2002.

Hamill, T. M. and Snyder, C.: A hybrid ensemble Kalman filter–3D variational analysis scheme, Monthly Weather Review, 128, 2905–2919, 2000.

10   Houtekamer, P. L. and Mitchell, H. L.: A sequential ensemble Kalman filter for atmospheric data assimilation, Monthly Weather Review, 129, 123–137, 2001.

Hunt, B. R., Kostelich, E. J., and Szunyogh, I.: Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter, Physica D: Nonlinear Phenomena, 230, 112–126, 2007.

Kalnay, E.: Atmospheric modeling, data assimilation and predictability, Cambridge university press, 2003.

15   Kirchgessner, P., Nerger, L., and Bunse-Gerstner, A.: On the choice of an optimal localization radius in ensemble Kalman filter methods, Monthly Weather Review, 142, 2165–2175, 2014.

Law, K., Stuart, A., and Zygalakis, K.: Data assimilation: a mathematical introduction, vol. 62, Springer, 2015.

Le Dimet, F.-X. and Talagrand, O.: Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects, Tellus A, 38, 97–110, 1986.

20   Lorenz, E. N.: Predictability: A problem partly solved, in: Proc. Seminar on predictability, vol. 1, 1996.

Lorenz, E. N. and Emanuel, K. A.: Optimal sites for supplementary weather observations: Simulation with a small model, Journal of the Atmospheric Sciences, 55, 399–414, 1998.

Ménétrier, B., Montmerle, T., Michel, Y., and Berre, L.: Linear filtering of sample covariances for ensemble-based data assimilation. Part I: Optimality criteria and application to variance filtering and covariance localization, Monthly Weather Review, 143, 1622–1643, 2015.

25   Moosavi, A., Attia, A., and Sandu, A.: A Machine Learning Approach to Adaptive Covariance Localization, ArXiv e-prints, 2018.

Nino-Ruiz, E. D., Sandu, A., and Deng, X.: A parallel ensemble Kalman filter implementation based on modified Cholesky decomposition, in: Proceedings of the 6th Workshop on Latest Advances in Scalable Algorithms for Large-Scale Systems, p. 4, ACM, 2015.

Parker, T. S. and Chua, L.: Practical numerical algorithms for chaotic systems, Springer Science & Business Media, 2012.

Petrie, R.: Localization in the ensemble Kalman filter, MSc Atmosphere, Ocean and Climate University of Reading, 2008.

30   Petrie, R. E. and Dance, S. L.: Ensemble-based data assimilation and the localisation problem, Weather, 65, 65–69, 2010.

Porcu, E., Daley, D. J., Buhmann, M., and Bevilacqua, M.: Radial basis functions with compact support for multivariate geostatistics, Stochastic environmental research and risk assessment, 27, 909–922, 2013.

Reich, S. and Cotter, C.: Probabilistic forecasting and Bayesian data assimilation, Cambridge University Press, 2015.

Roh, S., Jun, M., Szunyogh, I., and Genton, M. G.: Multivariate localization methods for ensemble Kalman filtering, Nonlinear Processes in
35   Geophysics, 22, 723, 2015.

Sakov, P. and Bertino, L.: Relation between two common localisation methods for the EnKF, Computational Geosciences, 15, 225–237, 2011.

Sakov, P. and Oke, P. R.: A deterministic formulation of the ensemble Kalman filter: an alternative to ensemble square root filters, Tellus A, 60, 361–371, 2008.
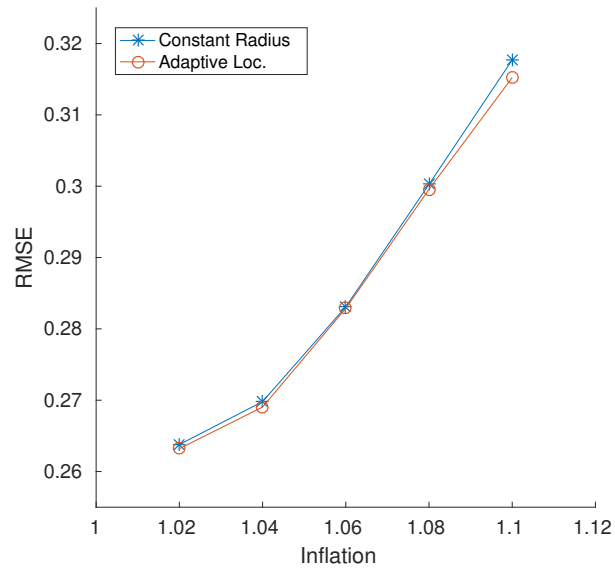
Sandu, A., Daescu, D., Carmichael, G., and Chai, T.: Adjoint sensitivity analysis of regional air quality models, Journal of Computational Physics, 204, 222–252, https://doi.org/10.1016/j.jcp.2004.10.011, http://dx.doi.org/10.1016/j.jcp.2004.10.011, 2005.

Schur, J.: Bemerkungen zur Theorie der beschränkten Bilinearformen mit unendlich vielen Veränderlichen., Journal für die reine und ange-
5   wandte Mathematik, 140, 1–28, http://eudml.org/doc/149352, 1911.

Singh, V. P., Rajagopal, A., and Singh, K.: Derivation of some frequency distributions using the principle of maximum entropy (POME), Advances in Water Resources, 9, 91–106, 1986.

Skamarock, W. C., Klemp, J. B., Dudhia, J., Gill, D. O., Barker, D. M., Duda, M. G., Huang, X., Wang, W., and Powers, J. G.: A description of the advanced research WRF version 3, Tech. rep., National Center For Atmospheric Research, 2008.

10   Strogatz, S. H.: Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering, Westview press, 2014.

Tranquilli, P., Glandon, S. R., Sarshar, A., and Sandu, A.: Analytical Jacobian-vector products for the matrix-free time integration of partial differential equations, Journal of Computational and Applied Mathematics, 310, 213–223, 2017.

Van Leeuwen, P. J., Cheng, Y., and Reich, S.: Nonlinear data assimilation, vol. 2, Springer, 2015.

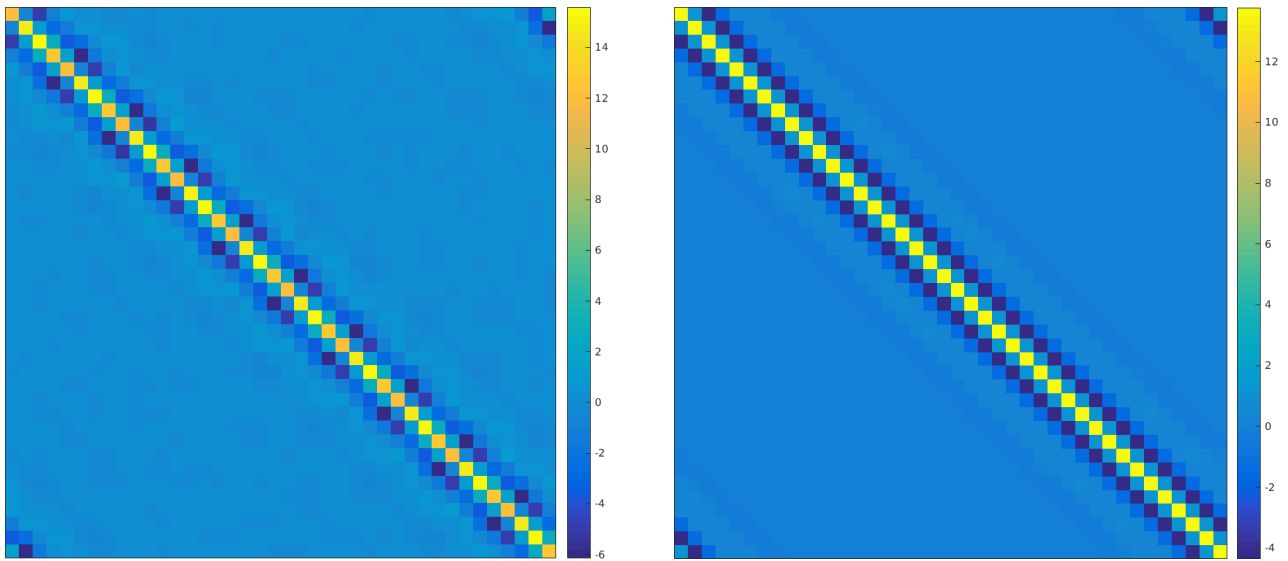Zupanski, M.: Maximum likelihood ensemble filter: Theoretical aspects, Monthly Weather Review, 133, 1710–1726, 2005.

Figure 1. GENERALISED MEAN ORACLES FOR LORENZ'96. Comparison of the various m-based multivariate localization techniques with that of standard univariate localization. Results obtained using the Lorenz'96 model over the assimilation cycles $[100, 1100]$. The particular localization scheme used is an adaptive oracle search that minimizes the error of the Schur-product localized analysis with respect to the truth. Each of the forty of variables is given an independent radius in the multivariate case. The schemes (10) that closely mirror an unbiased mean, namely m_{mean}, m_{sqrt}, m_{rms}, and m_{harm}, yield the best results, while the conservative scheme m_{min} performs no better than a univariate approach.

Figure 2. 4D ORACLE FOR LORENZ'96. Plot of the RMSE for a radius oracle. The y-axis represents the number of independent radii values (groups of the model state components). The assimilation cycle interval is $[100, 1100]$, a fixed inflation value of $\alpha = 1.02$ and the function m_{mean} are used. As can be seen, there is significant benefit in using more radii groupings, but marginal benefit from the 4D approach.

Figure 3. LORENZ'96 ADAPTIVE LOCALIZATION RESULTS Comparison of the best univariate localization radius results with their corresponding adaptive localization counter parts. $\bar{\upsilon}$ was varied by additive factors of one of $-1, -0.5, +0, +0.5, +1$ with respect to the best univariate radius with $\text{Var}(\upsilon)$ chosen to be one of $1/8, 1/4, 1/2, 1, 2$. As was predicted, no adaptive scheme that takes into account only currently available information can do well for Lorenz'96.
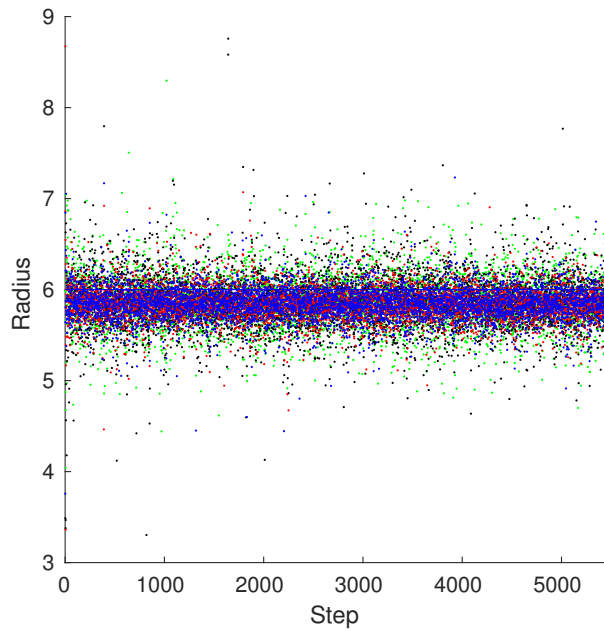
Figure 4. Comparison of ensemble covariance matrices for the multivariate Lorenz'96 equations for a single time-step (left) with that for time-averaged run (right) for a large number of ensemble members. There is considerable difference between the size of non-diagonal entries for different state variables over the course of a single step, but this difference disappears after averaging. This indicates that for this problem the best constant univariate radius is the same as for the canonical Lorenz'96 model, but that instantaneous adaptive radii are different.

Figure 5. MULTIVARIATE LORENZ'96 4D ADAPTIVE LOCALIZATION. A Gauss localization function is employed. The inflation is constant, and we test values α from 1.02 to 1.1, represented on the x-axis. The localization radius is varied in increments of 0.5 over the range $[0.5, 16]$ for the constant case. The minimal errors are shown in the graph, and the corresponding radii are used as mean inputs into our adaptive algorithm. For the adaptive case we choose four arbitrary groupings of radii ($g = 4$) with the mean function m_{mean}. The parameters $\bar{v}$ take one of three possible values, the optimal constant radius -1, $+0$, and $+1$. The parameters $\text{Var}(v)$ take one value from $\{1/4, 1, 4\}$. The 4D variable is set to $K = 1$ to look at an additional observation in the future.
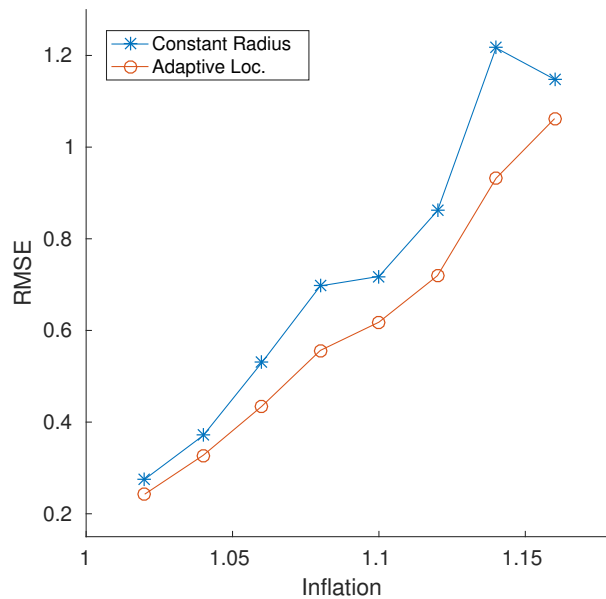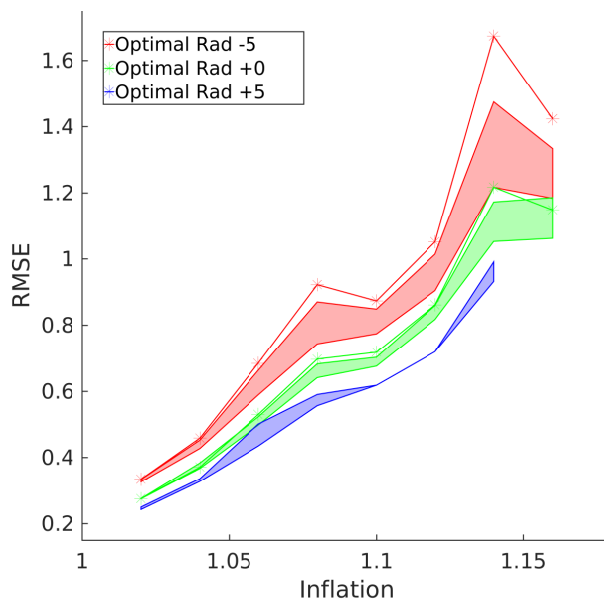
Figure 6. For the configuration in Figure 5 with $\alpha = 1.02$, $\bar{v} = 6$, and $\text{Var}(v) = 1$. Each unique marker type represents a unique radius grouping with a red · representing the first group, a green · representing the second, a blue · representing the third, and a black · representing the fourth. Groups two and four contain variables that were observed fully, and groups one and three contain variables that were observed partially. One will note the tendency of the fully observed radii to be less conservative in their deviation from the mean. The graph has been truncated to only show radii from 3 to 9 in order to better capture the general tendancies.
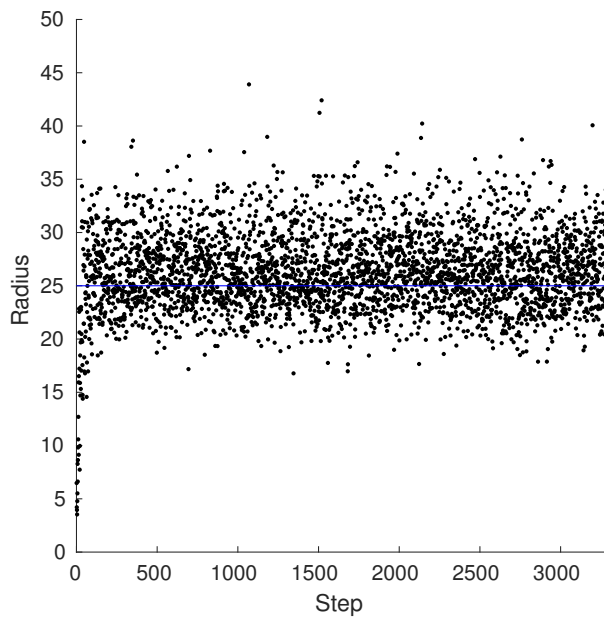
Figure 7. QUASI-GEOSTROPHIC MODEL. A typical model state of the 1.5 layer Quasi-Geostrophic model. The left panel shows the stream function values, with the red dots representing the locations of the variables that are observed. The right panel shows the corresponding vorticity.

Figure 8. Quasi-Geostrphic Model Adaptive Localization. A Gauss localization function is used. The inflation factor is kept constant, and we test α values from 1.02 to 1.16 (represented on the x-axis). The constant localization radius varies in increments of 5 over the range $[5, 45]$. Only the best results are plotted, and are used as the mean seeds for the adaptive algorithm. For the adaptive case we vary the parameters $\bar{v}$ to by taking one of three possible values, differing from the optimal constant radius by -5, 0, and $+5$. The adaptive $\mathrm{Var}(v)$ takes one of the values $1/2$, 1, 2, and 4. As can be seen, for higher values of inflation, the adaptive localization approach performs stellarly.

Figure 9. QUASI-GEOSTRPHIC MODEL ADAPTIVE LOCALIZATION RAW RMSE. A better representation of how well the adaptive localization scheme works is by showing its consistency. The green line represents the same optimal constant localization radius as in Figure 8. The red line represents the error for the radius -5 units below, and the blue line, if it had not experienced filter divergence would have represented +5. The correspondingly colored areas represent the ranges of error of the adaptive localization scheme obtained by fixing the mean, $\bar{v}$, to be that of the constant scheme and ranging over the variance. As can be seen, the adaptive scheme generally outperformed the constant scheme for almost all ranges of the variance and even managed to not suffer from filter divergence for some values of the variance in the +5 case, thereby providing evidence for the fact that an overestimation of the variance coupled with a useful estimate of the mean, might produce more useful results than the corresponding constant counterpart.

Figure 10. QGSO SAMPLE RADII. For the configuration in Figure 8 with $\alpha = 1.08$, $\bar{v} = 25$, and $\text{Var}(v) = 4$. Each dot represents a radius, with the line representing the mean. One will notice that during the spin-up time the algorithm is much more conservative with the radius of influence signaling that there should be an over-reliance on the observations instead of the model prior. After the spin-up time however, the algorithm tends to select radii greater than the mean, signifying a greater confidence in the observations.