Beating the Curse of Dimensionality of Sequential Monte Carlo for Bayesian Inverse Problems in Nonlinear Fluids

by

G. A. Robinson

B.S., University of California, Davis, 2013

M.S., University of Colorado, Boulder, 2017

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Applied Mathematics

2019

This thesis entitled:
Beating the Curse of Dimensionality of Sequential Monte
Carlo for Bayesian Inverse Problems in Nonlinear Fluids
written by G. A. Robinson
has been approved for the Department of Applied Mathematics

_____

Prof. Ian Grooms

_____

Prof. Will Kleiber

Date _____

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.

Robinson, G. A. (Ph.D., Applied Mathematics)

Beating the Curse of Dimensionality of Sequential Monte Carlo for Bayesian Inverse Problems in Nonlinear Fluids

Thesis directed by Prof. Ian Grooms

It is often desirable to estimate the distribution over quantities arising from inverse problems, but extant methods for estimating the state and uncertainty of dynamical systems are either computationally intractable for problem sizes in the domain of fluid dynamics, or they are provably inconsistent for nonlinear problems. Strategies to improve upon the bias those methods experience, while balancing for computational scalability, would therefore be helpful in obtaining good uncertainty estimates for inverse problems in nonlinear fluids. This research introduces an approximation to the sequential state estimation problem for spatially-extended dynamics that improves the dimensional scaling of the provably-consistent sequential importance resampling algorithm (SIR): we assume that observation errors are correlated with a strange spatial structure, having a spectrum that grows in the progression toward small scales. This decreases the ensemble size required to attain an accurate distributional estimate with SIR. Next we develop a fast implementation of our error model that is compatible with scattered observations, as required for numerical weather forecast, using a multiresolution approximation to the inverse of an elliptic differential operator with properties we prescribe of a covariance operator to make SIR more tractable. Finally, we combine our error model with a hybrid of SIR and the ensemble square root filter (ESRF) and apply it to a toy model in a class widely used to test meteorological data assimilation methods. The hybrid still suffers from the ESRF's inconsistency, but the SIR step helps by presenting ESRF with a prior that is closer to Gaussian. Our error model allows SIR to do more of that mitigation. Relative to a state-of-the-art ESRF, this improved the continuous ranked probability score by 15% and the root mean squared error by 10% in both the posterior and forecast. This improvement is substantial, comparable to the last 15 years of improvement in operational 6-day weather forecasts.

Dedication

To public education.

Acknowledgements

I am indebted to all those in APPM who have graced me with the time, trust, and wisdom they have shared, especially Anne Dougherty, Bob Easton, Chris Ketelsen, Greg Beylkin, Gunnar Martinsson, Jem Corcoran, Jim Curry, Juan Restrepo, Keith Julien, Mark Hoeffer, Stephen Becker, Sujeet Bhat, and Tomoko Matsuo. This entire work was performed in collaboration with Ian Grooms, and Will Kleiber collaborated on chapter 2.

My way to this study was paved by many others who worked to see me thrive, including Alex Perry, Alice Cheung, Behrouz Touri, Christian Burke, D. Eric Smith, Jim Crutchfield, Liz Bradley, Msgr. John Barry, Karen MacClune, Katrina Minck, Ken MacClune, Markus Luty, Matt Presley, Rob Maier, and Robert Spier. Greg Kuperberg showed me the beauty of mathematics as a creative process and inspired me to study it. Judy Simon, surely among the world's most skilled and caring teachers, indelibly marked my early personal development for the better.

Thank you to Nikki Sanderson for being a constant fixture of support throughout graduate school. The camaraderie of my classmates, the CRC, Picklebric, and Rad-ish Collective were essential, as well as the enduring support from fellow alumni of the UC Davis Tri-Cooperatives. Thank you to my family for always urging me to follow inspiration, especially Harold and LaVon Hall for giving me examples to follow in the pursuit of studies in science and math.

Most of all, I thank my advisor, Ian Grooms. Beyond his excellence as a scholar and teacher, he demonstrates profound virtue that makes working with him a joy. I will forever be glad for what he taught me of mathematical sciences and scholarship, and I will treasure even more his lessons by flawless example on how to be a mentor.

Contents

Chapter

Tables

Table

Figures

Figure

Chapter 1

Motivation and background

Inverse problems in nonlinear fluids include the use of observations to estimate the system's state, to infer its dynamical parameters, and to choose between physical models to describe its evolution. A simple point estimate can sometimes be a satisfactory solution to an inverse problem, but other times it is crucial to have a principled estimate of uncertainty. For example, an uncertainty estimate for the state of a weather pattern enables the choice between multiple evacuation plans, or between different times to sow or harvest a crop, in a manner that weighs risks and rewards according to an estimate of their probability.

It can also be scientifically helpful to determine a probability distribution over a collection of physical models, given observations, rather than merely rejecting models in the manner of frequentist statistics. There are important open questions in the study of nonlinear fluids, such as atmospheric and oceanographic dynamics, that beg for a probabilistic treatment of that form. One such motivating example is the search for the physical mechanisms behind a phenomenon, such as a weather pattern, for which several plausible models exist that cannot decisively be ruled out. In principle, Bayesian statistics provides a way to put a probability distribution on those models, which would at least help prioritize experimental time and expense.

Unfortunately, Bayesian treatment of nonlinear fluids is challenging. This is because nonlinear fluids must be observed in high resolution to resolve their rich personality, making them inherently high-dimensional systems, and nonlinearity introduces non-Gaussian probability distributions. All extant state estimators that scale well with dimension are proven to be inconsistent — they converge

to the wrong answers — in the face of non-Gaussian distributions. Strategies to improve upon the bias those methods display would therefore be helpful in obtaining good uncertainty estimates for inverse problems of nonlinear fluids.

Rooted in this desire to apply Bayesian uncertainty quantification toward understanding the physics of nonlinear fluids, this research seeks to wrangle nonlinear fluids inverse problems into the reach of consistent Bayesian data assimilation methods. The system in which this research found original inspiration is the Madden-Julian Oscillation (MJO), a multiscale nonlinear atmospheric oscillation associated with an eastward travelling wave of heavy precipitation in the Indian tropics. The MJO's phenomenology has been extensively studied in the last half-century, but the physical mechanisms for its initiation and propagation remain a mystery. A long litany of conceptual models produce simulations that bear some realistic semblance of the MJO, and selecting between those models has proven challenging. It has been impossible to reject some of the models with the traditional, frequentist view.

One standard approach for gaining traction on such a conundrum would be to run a Markov Chain Monte Carlo (MCMC) sampler on the joint probability space that includes all models under consideration. "All models" here really means exactly that: each discrete hypothesis for how the MJO works, plus all the hypothesis's (probably continuous) parameters, plus all the state descriptions of each discrete physical model. The multinomial marginal distribution over the discrete hypotheses (each being a complicated compound model) can then serve as a posterior on the hypothesis, finally yielding understandable description of uncertainty about the underlying physics. This approach scales poorly in the number of random variables, unfortunately, and that presents an obstacle for the purpose of inverse problems in fluids.[1]

The MJO is a good example of that obstacle. It covers a large part of the globe, but it depends on small-scale convective systems on the order of kilometers in size. One episode of the event lasts for about a month, but its small-scale dynamics with timescales on the order of hours

[1] Not to mention that it is complicated to estimate the marginal. The most common approach, the harmonic mean estimator, is notoriously prone to undetectable instability. Good approximation often requires bespoke estimators crafted for a specific statistical problem (Neal, 2008).

either drive or are driven by the slow large-scale dynamics (depending on the etiological model). Thus, it is reasonable to incorporate a half-billion observations in an effort to solve an inverse problem on it. That is far beyond the reach of naive MCMC.

Fluids are inherently high-dimensional, and virtually all nonlinear fluid phenomena have strong multiscale coupling. Consequently, their dynamics usually cannot be treated as separable into isolated low-dimensional dynamics. That means they must be handled as high-dimensional objects, which scuttles the possibility of straightforward methods, mentioned above, that are typically employed for Bayesian comparison between models of low dimensionality.

We are inspired, therefore, to seek a data assimilation approach to estimate the state of the system. Data assimilation is the process of using observations and a dynamical model to sequentially improve an initial estimate of the system state. However, options for performing data assimilation are limited when one wants to do principled model selection, since it is attractive for that purpose to seek a method with rigorous guarantee of convergence to the correct posterior distribution of the system state. The "particle filter" is the unique option with this property that is flexible enough to assimilate data with non-Gaussian priors, nonlinear dynamics, and nonlinear observations.

Chapter 2 explains the nature of challenges in using the particle filter in high dimension, and then introduces a way to improve the particle filter's large-scale uncertainty quantification by smoothing away small scales from the observations. Smoothing away small scales from the observations, before assimilating as if the smoothed observations have uncorrelated errors, is equivalent to inflating the presumed observation error variance of small-scale components. We will also describe why selectively inflating observation error variance is plausibly acceptable for application to geophysical fluids: essentially, it is because small scales of geophysical fluids mix rapidly and have little predictive value. Applying this method of smoothing data to a toy model consisting of a discretized linear stochastic partial differential equation, for which the posterior is determined analytically for comparison, we find that the smoother improves the continuous ranked probability score (CRPS) by as much as 25%. This is a large improvement. We also found that the root mean squared error of the particle filter's mean did not substantially change.

Chapter 3 introduces a novel means of smoothing observations scattered in $\mathbb{R}^n$ that makes use of a multiresolution approximation to the bound-state Helmholtz kernel, which is then applied to a Guassian radial basis function interpolant of the observed data. This may seem like a circuitous approach for implementing a smoother, but it comes with the attractive advantage of linear asymptotic complexity in both time and memory. It also admits an easily tunable spectrum, a feature that is highly desirable in the application to improving the performance of particle filtering for uncertainty quantification of large-scales in geophysical dynamics. The effect of the smoother is demonstrated on an analytically tractable configuration of equally-spaced observations on a circle embedded in $\mathbb{R}^2$, and compared to an analogous (but not identical) Fourier spectrum of the bound-state Helmholtz kernel. Finally, the smoother is confirmed to behave appropriately on real atmospheric temperature data scattered on a portion of the globe. Effective sample sizes are computed for single importance sampling updates to these temperature data, using a real operational weather forecast as a prior, and it is shown that the smoother indeed increases the effective sample size substantially.

Practical geophysical applications must contend with a wide range of spatial scales that may be intractable even after smoothing away unpredictable or irrelevantly small scales. Hence, chapter 4 explores the application of our smoothing approach to particle filtering in the context of a hybridized filter that "bridges" particle filtering with the ensemble square root filter (ESRF) to reap some of each method's respective benefits. This involves treating the likelihood as two factors: $L = L^{1-\alpha} \cdot L^{\alpha}$. We assimilate with the particle filter using $L^{1-\alpha}$ as a likelihood. The resampled output of the particle filter step is then used as a prior for the ESRF step, which uses the complementary factor L^{α} as a likelihood with which to assimilate the observations. This approach is designed for situations having both highly non-Guassian forecast distributions (usually due to strong nonlinearity or long forecast time) and highly Gaussian posterior distributions (as in the case of dense linear observations with small Gaussian error). In a test problem based closely on the Lorenz-'96 dynamics with observations separated by approximately 6 days, we observe a 15% improvement in analysis CRPS and a 10% improvement in analysis root mean squared error

compared to a state-of-the-art ensemble square root filter stabilized with localization, inflation, and a mean-preserving random rotation. Similar improvements were also observed for the forecast, which is a particularly interesting result. A 10% improvement in RMSE is comparable to the improvements in operational numerical weather forecasts in the last 15 years.

Chapter 2

Improving particle filter performance by smoothing observations

2.1 Introduction

Particle filters are a class of ensemble-based methods for solving sequential Bayesian estimation problems. They are uniquely celebrated due to their provable convergence to the correct posterior distribution in the limit of an infinite number of particles, with minimal constraints on prior and likelihood (Crisan and Doucet, 2002). Processes that are nonlinear and non-Gaussian can be filtered in this flexible framework, with rigorous assurances of asymptotically correct uncertainty quantification. These advantages stand in contrast to ensemble Kalman filters that lack convergence guarantees for nonlinear or non-Gaussian problems, and to variational methods that provide a point estimate but do not quantify uncertainty in the common case where the Hessian of the objective is unavailable.

The simplest form of a particle filter is descriptively called sequential importance sampling (SIS). We briefly describe the algorithm here to fix notation and terminology, and recommend Doucet et al. (2001) for a gentler introduction.

SIS begins by approximating the prior probability distribution with density $p(\mathbf{x}_{j-1})$ at discrete time $j-1$ as a weighted ensemble of N_e members $\left\{\left(\mathbf{x}_{j-1}^{(i)}, w_{j-1}^{(i)}\right)\right\}$, where the weights $w_{j-1}^{(i)}$ are related to the prior probabilities of the corresponding states $\mathbf{x}_{j-1}^{(i)}$. The superscript (i) indexes the collection of particles, and the sum of the weights is one. This kind of approximation, an **importance sample**, is an ensemble drawn from one distribution that is easy to sample and then reweighted to represent another distribution of interest.

The distribution of interest is the Bayesian posterior at discrete time j, which is proportional to the product of the prior at time $j-1$, $p\left(\mathbf{x}_{j-1}\right)$, the transition kernel $p(\mathbf{x}_j|\mathbf{x}_{j-1})$, and the likelihood $p(\mathbf{y}_j|\mathbf{x}_j)$. SIS evolves the samples from time $j-1$ to time j according to a **proposal kernel** that takes the generic form $p\left(\mathbf{x}_j^{(i)}|\mathbf{x}_{0:j-1}^{(i)}, \mathbf{y}_j\right)$. The weights are updated to reflect the difference between the proposal kernel and the Bayesian posterior at time j:

$$w_j^i \propto w_{j-1}^i \frac{p\left(\mathbf{y}_j|\mathbf{x}_j^{(i)}\right) p\left(\mathbf{x}_j^{(i)}|\mathbf{x}_{j-1}^{(i)}\right)}{p\left(\mathbf{x}_j^{(i)}|\mathbf{x}_{0:j-1}^{(i)}, \mathbf{y}_j\right)}. \tag{2.1}$$

The proposal kernel is often set to equal the transition kernel, which simplifies the ratio in (2.1) so that the weights are proportional to the likelihood: $w_j^i \propto w_{j-1}^i \cdot p\left(\mathbf{y}_j|\mathbf{x}_j^{(i)}\right)$. The proportionality constant is chosen so that the weights sum to one. (Some authors, e.g. van Leeuwen (2010), integrate out dependence on x_{j-1}; we instead follow the convention of Doucet et al. (2001).)

Despite its attractive qualities, particle filtering is unpopular in meteorological applications due to an especially vexing curse of dimensionality. The problem is that the importance sampling weights associated with system replicates (**particles**) have a tendency to develop **degeneracy** as the system dimension grows. That is to say, a single particle near the observation will have essentially all the sampling weight while the rest of the particles, bearing effectively zero weight, are ignored in the computation of ensemble statistics.

One can quantify the degree of degeneracy with an **effective sample size** (ESS), which is a heuristic measurement of the importance sample quality defined as

$$\mathrm{ESS}_j = \frac{1}{\sum_{i=1}^{N_e} \left(w_j^{(i)}\right)^2}. \tag{2.2}$$

The ESS ranges from one if a single weight is nonzero (which is the worst case), to N_e if all weights are equal. If the effective sample size becomes much smaller than the ensemble size, the filter is said to have **collapsed**. A simple approach to combat collapse is to resample the particles from time to time, eliminating particles with low weight and replicating particles with high weights. There are several common approaches to resampling (e.g. Doucet and Johansen, 2009), and by construction of this resampling step, all weights become uniform: $w_j^{(i)} \to 1/N_e$ [see also the more recent resampling

alternatives in Reich (2013) and Acevedo et al. (2017)]. The term 'particle filter' commonly implies an SIS filter with a resampling step, also known as Sequential Importance Resampling (SIR).

SIR particle filters are guaranteed to converge to the correct Bayesian posterior in the limit of an infinite number of particles, but the rate of convergence can be prohibitively slow for high-dimensional problems. The number of particles required to avoid collapse is typically exponential in a quantity related to the number of observations, as described by Bengtsson et al. (2008) and Snyder et al. (2008). For example, consider a system with Gaussian prior on $\mathbf{x}_j$ and with likelihood, conditional on $\mathbf{x}_j$,

$$\mathbf{y}_j|\mathbf{x}_j \sim \mathcal{N}(\mathbf{H}\mathbf{x}_j, \mathbf{R}) \tag{2.3}$$

where $\mathcal{N}(\mu, \mathbf{\Sigma})$ denotes a multivariate normal distribution with mean μ and covariance $\mathbf{\Sigma}$, $\mathbf{H}$ is a linear observation operator, and $\mathbf{R}$ is the covariance of the additive observation error. For this example Snyder et al. (2008) show that the number of particles N_e required to avoid collapse is on the order of $\exp\{\tau^2/2\}$ where

$$\tau^2 = \sum_{k=1}^{N_y} \lambda_k^2 \left(\frac{3}{2}\lambda_k^2 + 1 \right), \tag{2.4}$$

in which N_y is the dimension of the observations and λ_k^2 are eigenvalues of

$$\mathrm{cov}\left(\mathbf{R}^{-1/2}\mathbf{H}\mathbf{x}_j \right). \tag{2.5}$$

Chorin and Morzfeld (2013) also discuss the notion of 'effective dimension' and how it relates to particle filter performance. Agapiou et al. (2017) give precise, non-asymptotic results on the relationship between the accuracy of the particle filter, the number of particles, and the 'effective dimension' of the filtering problem in both finite and infinite dimensional dynamical systems. For simplicity of exposition we rely on the formulas quoted here from Snyder et al. (2008) and Snyder et al. (2015).

A number of methods developed to minimize degeneracy in high-dimensional problems utilize a proposal kernel that is different from the transition prior, using observations to guide proposals.

Of all possible proposals that depend only on the previous system state and the present observations, there exists an optimal proposal that minimizes both the variance of the weights and the number of particles required to avoid degeneracy (Doucet et al., 2000, Snyder et al., 2015). It is typically impractical to sample from that optimal proposal. The various methods proposed to minimize weight degeneracy in practice include the implicit particle filter (Chorin et al., 2010, Chorin and Tu, 2009, 2012, Morzfeld et al., 2012), and the equivalent weights particle filter (Ades and Van Leeuwen, 2015, 2013, van Leeuwen, 2010). Snyder et al. (2015) have shown that improved proposals can reduce the number of particles required to avoid collapse, but the number is still prohibitive for meteorological applications. Another approach to improving the performance of particle filters uses 'localization.' Localization reduces the effective number of observations (and therefore the required number of particles) by breaking the assimilation into a sequence of smaller subsets. Localization can also improve the performance of particle filters (Morzfeld et al., 2017, Penny and Miyoshi, 2016, Poterjoy, 2016, Rebeschini and Van Handel, 2015), but breaks convergence guarantees. Other methods improve the filter results by making the observation error model state dependent (Okamoto et al., 2014, Zhu et al., 2016).

This paper describes a different but compatible approach for improving the dimensional scaling of particle filters by smoothing observations before proceeding as though the observations are uncorrelated; equivalently, we increase the small-scale variance in the error model. The goal of doing so is to achieve more desirable dimensional scaling. Whereas changing the proposal kernel allows particle filtering to sample a given posterior more efficiently, manipulating the observation model changes the posterior itself. This may seem to vitiate convergence guarantees at least as badly as localization does. After all, it is possible that localized particle filters and EnKFs converge to some distribution in the large ensemble limit. However, convergence results are still an open problem for EnKFs and localized particle filters. In any case, the limiting distribution of a localized filter is not the true Bayesian filter, and the nature of the bias in the limiting distribution is unknown. By contrast, we can guarantee convergence to a surrogate distribution with bias that can be described and controlled.

The key insight motivating our approach is evident in (2.5): increasing the observation error variance for any eigenvector of $\mathbf{R}$ correspondingly decreases the number of particles required. The challenge is to make the problem less expensive to sample with a particle filter, while still accurately incorporating observations on the most physically relevant large scales. This paper describes an analytically transparent and computationally efficient method that reduces the number of particles required to avoid collapse by increasing the observation error variance at small scales.

2.2 Theory

In this section we develop intuition by considering the observation error model (2.3) in the special case where $\mathbf{R}$ and $\mathrm{cov}(\mathbf{x}_j)$ are Fourier diagonalizable and $\mathbf{H} = \mathbf{I}$. Writing eigenvalues of $\mathbf{R}$ as γ_k^2 with k an integer wavenumber from 1 to N_y, and the eigenvalues of $\mathrm{cov}(\mathbf{x}_j)$ as σ_k^2, the matrix in (2.5) has eigenvalues

$$\lambda_k^2 = \sigma_k^2/\gamma_k^2. \tag{2.6}$$

The effects of aliasing complicate the Fourier scale analysis of filtering when observations are not available at every grid point, especially when the observation grid is irregular (Majda and Harlim, 2012, Chapter 7).

Recall from the introduction that Snyder et al.'s estimate (2.4) of the ensemble size required depends on the system covariance, the observing system, and the observation error covariance. Let us ground the theoretical discussion with general comments about the nature of these quantities in operational numerical weather prediction. Typically the model physics are reasonably well-known and held fixed, so we take $\mathrm{cov}(\mathbf{x}_j)$ to be given.[1] The observing system, like the dynamical model, is typically given and fixed. The observation error covariance, in contrast both to the dynamical model and the observing system, is often a crude heuristic approximation that is easier to modify.

[1] One can in principle design physical models to make an assimilation problem more tractable to a particle filter, homologous to the approach we describe that alters the observation model. We do not consider that in this article because the theory scantly differs and the praxis is much more problem dependent. The related representation errors, arising from a mismatch between the length scales resolvable by the numerical model and the length scales present in the observations, are difficult to quantify but are presumably spatially correlated.

Observation error is frequently taken to have no spatial correlation, for example $\mathbf{R} \propto \mathbf{I}$ in the case of distant identical thermometers, in which case $\{\gamma_k\}$ are constant. Otherwise the observation error may have strong spatial correlations, as may be expected of satellite observations biased by a spatially smooth distribution of unobserved atmospheric particulates, in which case $\gamma_k \to 0$ rapidly for large k.

2.2.1 Impact of observation error model on number of particles required

The following hypothetical examples demonstrate how the observation error model can affect the number of particles required for particle filtering. We first use Snyder's asymptotic arguments to estimate the particle filter ensemble size required to reconstruct a Bayesian posterior with a correlated observation error model, whose realizations are continuous with probability one, and contrast this with the ensemble size required under the approximation that observation errors are spatially uncorrelated. Making this approximation decreases the particle filter ensemble size required to reconstruct the Bayesian posterior. This progression is designed set the stage for our method; we show that using a peculiar choice of $\mathbf{R}$, possessing a growing spectrum, naturally extends the approximation of correlated errors with uncorrelated errors. Our method decreases the number of particles required to approximate the posterior regardless of whether the true errors are correlated or uncorrelated.

Fields whose correlations gradually decrease with distance have decaying spectra, i.e. $\gamma_k^2 \to 0$ at small scales. This has a detrimental effect on the effective dimensionality of the problem. Suppose, for example, that observation error variances $\gamma_k^2 = k^{-4}$ and system covariance $\sigma_k^2 = k^{-2}$. Then eigenvalues of (2.5) are $\lambda_k^2 = k^2$ and

$$\tau^2 \approx \int_1^{N_y} k^2 \left(\frac{3}{2}k^2 + 1\right) \mathrm{d}k \sim \frac{3}{10} N_y^5 \tag{2.7}$$

where the sum in (2.4) has been approximated by an integral. In this example the effective dimensionality of the problem increases extremely rapidly as the number of observations grows. A similar argument can be used to show that if σ_k^2 decays sufficiently faster than γ_k at small scales

(large k), then the effective dimensionality of the system remains bounded in the continuum limit.

When the spatial correlation of the observation error is unknown, it is not uncommon to use a spatially-uncorrelated (i.e. diagonal) observation error model. This approximation is also popular because it is computationally convenient in ensemble Kalman filters, where it enables serial assimilation (Bishop et al., 2001, Houtekamer and Mitchell, 2001, Whitaker and Hamill, 2002). For observations with correlated errors, such as swaths of remotely sensed data, approximating the errors as spatially uncorrelated changes the posterior relative to a more accurate observation error model with correlations; the approximation seems to work well enough in practice. The spatially uncorrelated approximation, compared to error models with continuous realizations, also makes particle filtering easier. When the error is spatially uncorrelated, γ_k^2 does not decay to zero at small scales. Repeating the asymptotic argument in the preceding paragraph with constant $\gamma_k^2 = 1$ implies $\lambda_k^2 = k^{-2}$, so

$$\tau^2 \approx \int_1^{N_y} k^{-2} \left(\frac{3}{2} k^{-2} + 1 \right) \mathrm{d}k \sim \frac{3}{2} \tag{2.8}$$

in the continuum limit. This illustrates that the number of particles required to avoid collapse can be significantly reduced by changing the spatial correlations in the observation error model, and in practice the filter results are still acceptably accurate.

Our proposal is take this approximation a step further: we let observation error covariance grow without bound in the progression to small scales. This model of the observation error, possessing a spectrum bounded away from zero, is called a **generalized random field** (GRF) and has peculiar properties described in section 2.2.2. Despite those peculiarities of GRFs which complicate analysis of the continuum limit, the finite dimensional vector of observational errors can be treated as a multivariate Gaussian random vector.

In the following subsections we discuss the impact of this observation error model on the posterior, and various numerical methods for constructing and implementing the associated particle filter. We find the theory to be more intuitive in terms of this covariance framework than working with smoothing operators, but the final subsection will make the equivalence precise.

2.2.2 Properties of generalized random fields

Generalized random fields (GRFs) are discussed at length in Yaglom (1987), and a few extra details can be found in Gelfand and Vilenkin (1964). A GRF whose Fourier spectrum is not integrable at small scales has infinite variance. The prototypical example is a spatially-uncorrelated field, whose spectrum is flat.

A GRF is not defined pointwise. Rather than being defined pointwise, or 'indexed by spatial location,' it is indexed by rapidly decaying test functions (often taken to be elements of a Schwartz space). This is perhaps best explained by reference to an ordinary random field. If $Z(\boldsymbol{x})$ is a random field that is defined pointwise and $\phi(\boldsymbol{x})$ is a test function then we can define a new, 'function indexed' random field $Z(\phi)$ using the expression

$$Z(\phi) = \int Z(\boldsymbol{x})\phi(\boldsymbol{x})\mathrm{d}x.$$

If the field Z is not defined pointwise, it may still be indexed by test functions.

The concept of a covariance function for an ordinary random field can be generalized to a GRF. The resulting object is a 'covariance kernel' which can be a generalized function, i.e. an element of the dual of a Schwartz space. The prototypical covariance kernel is the so-called Dirac delta function which is not, in fact, a function.

The observation error covariance model advocated in this article can be conceptualized in two ways. It can be thought of as an approximation to a GRF where the spectrum has been truncated at the smallest resolvable scale on the grid. Alternatively, one can assume that observations are not taken at infinitesimal points in space, but rather that the observing instrument senses over a small region of space via some test function ϕ. The value of the GRF for an observation is thus indexed by the allowed test functions ϕ rather than the spatial location of the observation. A more rigorous treatment of our discretization of generalized random fields appears in chapter 4.

2.2.3 Effect of a generalized random field likelihood on posterior

The performance advantage, described above, does not come for free. Changing the observation error model changes the posterior. To demonstrate how our choice of error model affects the posterior, consider again a fully Gaussian system for which the system covariance $\mathrm{cov}(\mathbf{x}_j)$ has the same eigenvectors as the presumed observation error covariance $\mathbf{R}$, and where the observation operator is the identity. Let σ_k^2 be eigenvalues of $\mathrm{cov}(\mathbf{x}_j)$ and γ_k^2 be eigenvalues of $\mathbf{R}$, indexed by k in the diagonalizing basis with index k increasing towards small scales. Let $\hat{\boldsymbol{x}}_k$ and $\hat{\boldsymbol{y}}_k$ denote the projection of the prior mean and observations onto the k^{th} eigenvector, respectively. Then the posterior mean of $p(\hat{\boldsymbol{x}}_k|\hat{\boldsymbol{y}}_k)$ is

$$\hat{\boldsymbol{x}}_k + \frac{\sigma_k^2}{\sigma_k^2 + \gamma_k^2}(\hat{\boldsymbol{y}}_k - \hat{\boldsymbol{x}}_k). \tag{2.9}$$

In order for the posterior mean to be accurate at large scales, it will be necessary to design an observation error model with realistic variance at large scales; we return to this point in section 22.2.4. Clearly, if $\gamma_k^2 \to \infty$ at small scales then the posterior mean will equal the prior mean at small scales. If the filter tends to ignore small-scale information, then the small-scale part of the prior mean will eventually tend towards the climatological small-scale mean, which is often zero since climatological means are often large-scale. This observation error model can therefore be expected to have a smoothing effect on the posterior mean.

This is the price to be paid for reducing the effective dimensionality of the system, but the price is not too high. Small scales are inherently less predictable than large scales, so loss of small-scale observational information may not significantly damage the accuracy of forecasts. Practical implementations will need to balance between ignoring enough observational information to avoid particle collapse and keeping enough to avoid filter divergence (i.e. the filter wandering away from the true state of the system).

In the same example as above, the eigenvalues of the posterior covariance are

$$\xi_k^2 = \frac{\sigma_k^2 \gamma_k^2}{\sigma_k^2 + \gamma_k^2}.$$

As noted above, in order for the posterior variance to be accurate at large scales, it will be necessary to design an observation error model with realistic variance at large scales. At small scales we argue that ξ_k^2 is small (using the notation $\ll 1$) regardless of the behavior of γ_k^2. This is because the state $\boldsymbol{x}$ is associated with a viscous fluid model whose solutions should be continuous. A GRF error model with $1 \ll \gamma_k^2$ will lead to a posterior variance close to the prior variance at small scales: $\xi_k^2 \approx \sigma_k^2 \ll 1$. A more realistic error model with $\gamma_k^2 \ll 1$ will lead to a much smaller posterior variance, but in either case $\xi_k^2 \ll 1$. This argument suggests that the GRF approach should not have a detrimental effect on the posterior variance when applied to atmospheric or oceanic dynamics, provided that the observation error variance at large scales is realistic.

2.2.4 Constructing GRF Covariances

In the context of an SIR particle filter using the standard proposal with a nonlinear observation error model of the form

$$\boldsymbol{y}_j = \boldsymbol{H}(\boldsymbol{x}_j) + \boldsymbol{\eta}_j$$

where $\boldsymbol{\eta}_j \sim \mathcal{N}(0, \mathbf{R})$ is the observation error, the incremental weights are computed using

$$w_j^{(i)} \propto w_{j-1}^{(i)} \exp\left\{-\frac{1}{2}\left(\boldsymbol{y}_j - \boldsymbol{H}(\boldsymbol{x}_j^i)\right)^T \mathbf{R}^{-1}\left(\boldsymbol{y}_j - \boldsymbol{H}(\boldsymbol{x}_j^i)\right)\right\}.$$

The goal of this section is to describe two methods for defining an observation error covariance $\mathbf{R}$ that has the increasing variance prescribed above, and that allow for rapid computation of the weights. First, we will suppose that the true observation error variance is known, and we will scale it out so that we are dealing only with the error correlation matrix. If $\mathbf{R}_0$ is a diagonal matrix with elements that are the observational error variances, then we will let

$$\mathbf{R} = \mathbf{R}_0^{1/2} \mathbf{C}\, \mathbf{R}_0^{1/2}$$

and we will model the matrix $\mathbf{C}$.

There is a well-known connection between stationary Gaussian random fields and elliptic stochastic partial differential equations (Lindgren et al., 2011, Rue and Held, 2005) that allows

fast approximation of likelihoods. Specifically, the inverse of the covariance matrix of a discretized random field can in some cases be identified with the discretization of a self-adjoint elliptic partial differential equation (PDE). The connection extends in a natural way to generalized Gaussian random fields, with the caveat that the covariance matrix rather than its inverse is identified with the discretization of an elliptic PDE. For example, the matrix $\mathbf{C}$ can be constructed as a discretization of the operator

$$\left(1 - \ell^2 \Delta\right)^{\kappa}, \tag{2.10}$$

in which Δ is the Laplacian operator, $\ell > 0$ is a tuning parameter with dimensions of length, and $\kappa > 0$ controls the rate of growth of eigenvalues. Both the continuous differential operator and its discretization have positive spectra with eigenvalues growing in wavenumber. The parameter $\ell > 0$ controls the range of scales with eigenvalues close to 1. For length scales longer than ℓ the eigenvalues are close to 1 and the observation error model is similar to the commonly-used diagonal, uncorrelated observation error model. The large-scale observation error is correct, meaning that the posterior will also be correct at large scales. For length scales smaller than ℓ the observation error variance grows at a rate determined by κ, rapidly rolling off the influence of small scales.

Taking the matrix $\mathbf{C}$ to be a discretization of an elliptic PDE permits efficient application of the inverse, as required in computing the weights, by means of sparse solvers. It is also possible to construct $\mathbf{C}^{-1}$ directly as the discretization of the integral operator that corresponds to the inverse of this PDE, also enabling fast algorithms that have no limitation to regular observation grids. These kinds of methods will be explored more fully elsewhere.

An alternative to the PDE based approach for modeling $\mathbf{C}$ is to simply smooth the observations. Let the smoothing operator be a matrix $\mathbf{S}$, and the smoothed observations be denoted $\boldsymbol{y}_s$. Then the observation model

$$\boldsymbol{y}_S = \mathbf{S}\mathbf{R}_0^{-1/2}\boldsymbol{y}_j = \mathbf{S}\mathbf{R}_0^{-1/2}\boldsymbol{H}(\boldsymbol{x}_j) + \boldsymbol{\eta}_s$$

where the smoothed observation errors are assumed to have independent, unit-variance errors,

implies incremental importance weights of the form

$$w_j^{(i)} \propto w_{j-1}^{(i)} \times \exp\left\{-\frac{1}{2}(\boldsymbol{y}_j - \boldsymbol{H}(\boldsymbol{x}_j^i))^T \mathbf{R}_0^{-1/2} \mathbf{S}^T \mathbf{S} \mathbf{R}_0^{-1/2} (\boldsymbol{y}_j - \boldsymbol{H}(\boldsymbol{x}_j^i))\right\}.$$

If a smoothing operator $\mathbf{S}$ is available, this is equivalent to setting $\mathbf{C}^{-1} = \mathbf{S}^T\mathbf{S}$. As long as the smoothing operator leaves large scales nearly unchanged while attenuating small scales, the impact on the effective sample size and on the posterior will be as described in the foregoing subsections. If it is possible to construct $\mathbf{S}$ to *project* onto a large-scale subspace, it would be equivalent to setting certain eigenvalues of the observation error covariance to infinity.

2.3 Experimental Configuration

To illustrate the effects of a GRF likelihood in a simple example, we apply an SIR particle filter to a 1-dimensional linear stochastic partial differential equation,

$$\frac{du}{dt} = \left(-b - c\frac{d}{dx} + \nu\frac{d^2}{dx^2}\right) u + F_t, \tag{2.11}$$

where $b, c, \nu \in \mathbb{R}^+$ are constant scalars and F is a time-dependent stochastic forcing that is white in time and correlated in space with a form described below. The domain is periodic, with length 2π. Such models have been used to test filtering algorithms by Majda and Harlim (2012). In Fourier space this model can be represented as the Itô equation

$$d\hat{u} = -(b + ikc + \nu k^2)\hat{u}\ dt + \zeta dW, \tag{2.12}$$

where $\hat{u}$ is the Fourier coefficient at wavenumber k, ζ is the noise amplitude, and dW is a standard circularly symmetric complex white noise. The coefficients are $b = 1$, $c = 2\pi$, and $\nu = 1/9$. To mimic turbulence in many physical models, we choose a stochastic forcing F_t that decays linearly for large wavenumbers. Specifically, let

$$\zeta^2 = 1/(1 + |k|) \tag{2.13}$$

such that the variance of the noise is one half of its maximum at wavenumber 1. This configuration (11-13) is chosen to possess a fairly limited range of active wavenumbers so that the particle filtering problem is tractable.

The model admits an analytical solution to which we can compare experimental results. Since the dynamic is linear and Fourier coefficients are independent, it follows that each Fourier mode evolves as an Ornstein-Uhlenbeck process independent of all other modes. This means we can efficiently propagate the system by sampling directly from the Gaussian distribution available in closed form for each Fourier coefficient (Øksendal, 2003):

$$\hat{u}_{t+\Delta t} = \hat{u}_t e^{-\theta_k \ \Delta t} + \zeta \sqrt{\frac{1 - e^{-2\theta_{r,k}\Delta t}}{2\theta_{r,k}}} \chi_t, \tag{2.14}$$

where $\theta_k = d + ikc + \nu k^2$, $\theta_{r,k}$ is the real part of θ_k, and χ_t is a standard circularly symmetric complex normal random variable. The initial condition for the experiment is drawn from the stationary distribution, obtained as the limit $\Delta t \to \infty$ in (2.14), which for each wavenumber is a circularly symmetric complex normal random number of standard deviation $1/\sqrt{2(1+|k|)\theta_{r,k}}$.

A particular solution, hereafter called the 'true system state' solution is computed at 2048 equally spaced points in the 2π-periodic spatial domain, and at 101 equally-spaced points in the time interval $[0, 4]$ (the initial condition being at $t = 0$). From this solution, synthetic observations are generated at every 32^{nd} spatial location (except as otherwise noted) by adding samples from a stationary zero-mean multivariate normal distribution with variance 0.36 and correlations of the form $\exp\{-|\delta/0.06|\}$ where δ is the distance between observations. There are thus 64×100 total observations (there are no observations of the initial condition).

The standard deviation of the observational error is 0.6, while the pointwise climatological standard deviation of the system is about 0.8. This is a very high observational noise level; we set the observational noise this high because the theoretical estimates of the required ensemble size are extremely large for smaller observational noise. Observational noise levels in meteorological applications are not usually this high relative to the climatological variability of the system. Despite this high level of noise, the observing system is dense enough in space and time that the filter is able to recover an accurate estimate of the system.

The GRF observation error covariance, used only for assimilation, is constructed as the periodic tridiagonal matrix formed by the second-order centered finite difference approximation

to the operator $0.36(1 - \ell^2\partial_x^2)$. The diagonal elements (the observation error variance) are all $0.36(1 + 2(\ell/\delta)^2)$ where δ is the distance between observations; the elements corresponding to nearest-neighbor covariances are all $0.36(1 - (\ell/\delta)^2)$. When $\ell = 0$ the observation error covariance is diagonal. The local observation error variances increase when ℓ increases, and the nearest-neighbor covariances decrease and can even become negative. The eigenvectors of this matrix are discrete Fourier modes. When ℓ increases, the variance increases for all Fourier modes except the constant mode, which remains at this baseline variance 0.36. Experiments are run with 101 values of ℓ^2 equally spaced in the interval $[0, 1]$. The GRF observation error covariance is not used to generate the synthetic observations.

Assimilation experiments are run with an SIR particle filter to test how the GRF observation error model impacts its performance. An ensemble size of $N_e = 400$ is used, except as noted otherwise. The SIR particle filter is configured to resample using the standard multinomial resampling algorithm Doucet et al. (2001). The ESS is tracked before resampling. Resampling reduces the information content of the ensemble by eliminating some particles and replicating others; to avoid unnecessary loss of information, resampling is only performed whenever the effective sample size (ESS) falls below $N_e/2$.

Two quantities are used to evaluate the effect of the GRF error model on the particle filter's performance. The first is the root mean squared error between the particle filter's posterior mean and the true system state, where the mean is taken over the spatial domain. The second is the continuous ranked probability score (Gneiting and Raftery, 2007, Hersbach, 2000, CRPS). This measures the accuracy of the posterior distribution associated with the particle filter's weighted ensemble. The score is non-negative; a score of zero is perfect, and smaller scores are better. It is more common to compare the RMSE to the ensemble spread, a function of the ensemble covariance trace (Fortin et al., 2014), but the CRPS is a more precise way to describe the quality of a probabilistic estimate. The CRPS is computed at every point of the spatial and temporal grid of 2048×100 points. We compute the CRPS for a range of different $N_y \in (16, 32, 64, 128)$ in order to probe the effects of changing the number of observations. All assimilation runs with the same

N_y use the same observations.

We will gauge particle filter performance with the GRF likelihood by comparing it to the reference case of a particle filter computed using a spatially-uncorrelated likelihood. In some cases we will also want to compare the particle filter estimate to the true Bayesian posterior. Though one of the main reasons for using a particle filter is that it works in nonlinear, non-Gaussian problems, a benefit of experimenting with a linear Gaussian problem is that the exact solution to the optimal filtering problem can be computed for this comparison using the Kalman filter. In particular, the Kalman filter provides the exact posterior covariance $\mathbf{P}_k$,

$$\mathbf{K}_k = \mathbf{P}_{k|k-1}\mathbf{H}^T \left(\mathbf{R} + \mathbf{H}\mathbf{P}_{k|k-1}\mathbf{H}^T\right)^{-1}$$
$$\mathbf{P}_k = (\mathbf{I} - \mathbf{K}_k\mathbf{H})\mathbf{P}_{k|k-1},$$

which allows us to estimate the number of particles required to avoid filter degeneracy a priori (without running the particle filter) using (2.4) and (2.5). The prior covariance at time k is denoted $\mathbf{P}_{k|k-1}$ in the above formulas.

2.4 Results

We compute τ^2 from the Kalman filter results at $t = 4$, the end of the assimilation window. This gives an approximation to the steady-state filtering problem because the posterior covariance converges exponentially to a limiting covariance (Chui and Chen, 2009). This process is repeated for each of eleven ℓ^2 linearly distributed between 0 and 1 and the results are plotted in the first panel of Figure 1. Note that the $\ell^2 = 0$ case is a spatially-uncorrelated observation error model. We observe a dramatic reduction in the theoretical number of particles required to avoid filter collapse. The theory of Bengtsson et al. (2008) and Snyder et al. (2008) predicts that the spatially-uncorrelated noise model requires on the order of 10^{26} particles to avoid collapse in this simple 1-dimensional PDE with 2048 Fourier modes. As ℓ^2 increases from 0 to 1, the number of required particles drops rapidly to about 8,000. In fact, as shown below, the SIR particle filter performs well with $\ell^2 = 1$ for an ensemble size of 400.

Reducing τ^2 by increasing ℓ^2 is a result of increasing the observation variance, and the chosen form of the surrogate observation error model is designed to increase the variance primarily for small scales while leaving large scales intact. The impact on the posterior is visualized in the second panel of Figure 1. This panel shows the time-average RMSE of the particle filter mean of the first 50 Fourier modes, normalized by the climatological standard deviation of each Fourier coefficient, for $\ell^2 \in (0, 0.04, 0.4)$. Here we observe that increasing ℓ^2 primarily increases the posterior variance at small scales, as designed.

The distribution of ESS throughout the 100 assimilation cycles is plotted in Figure 2.2 for various values of ℓ^2. The box plots are constructed from the time series of ESS over all 100 assimilation cycles. In this proxy for the quality of uncertainty quantification achieved by the particle filter, we observe approximately a tenfold increase in median ESS with $\ell^2 = 0.3$ and a thirty-fold increase in median ESS with $\ell^2 = 1$ compared to $\ell^2 = 0$. The ESS averages only 10–20% of N_e when $\ell^2 = 1$, with occasional collapses. This is not inconsistent with the theory, which requires N_e of about 8000 to avoid collapse, yet still shows the significant improvements from using a GRF likelihood with relatively small ensembles. The results below suggest that the particle filter can give an accurate probabilistic estimate of the system state even when the ESS is a small percentage of the ensemble size.

Next we consider how the root mean square error (RMSE) of the particle filter posterior mean from the true system state depends on ℓ. Figure 2.3 shows box plots of the RMSE as a function of ℓ^2. The box plots are constructed from the RMSE time series for the final 90 assimilation time steps in each experiment. The RMSE appears fairly insensitive to ℓ^2. The median RMSE for all cases remains below the observation error standard deviation of 0.6. These results demonstrate that the particle filter remains a fairly accurate point estimator – both when the filter is collapsed while ℓ is small, and when the posterior is substantially over-dispersed due to large ℓ. The Kalman filter using the true observation model, which is the optimal filter in the best case scenario for this problem, achieves a median RMSE of 0.32.

The use of a GRF likelihood clearly reduces the incidence of collapse in the particle filter,

with mild detriment to the RMSE. The RMSE measures a spatially-integrated squared error, which can mask errors at small scales. The arguments of section 22.2.3 suggest that the GRF posterior mean will be inaccurate primarily at small scales. We visualize the severity of this effect in Figure 5,which compares the true state (red) to the posterior mean (blue) and to ensemble members (gray) for four different values of ℓ^2: 0 (diagonal error model), 0.2, 0.4, and 0.6. The ensemble members are shaded according to their weight: weights near 1 yield black lines while weights near 0 yield faint gray lines. At $\ell^2 = 0$ there are few ensemble members visible, reflecting the fact that the ESS is small. Nevertheless, the posterior mean is reasonably close to the true state. As ℓ^2 increases, the number of visible ensemble members increases (reflecting increasing ESS), and the posterior mean becomes smoother. Although the posterior mean at $\ell^2 = 0.6$ is smoother than the true system state, the individual ensemble members are not overly smooth; they are instantiations of the dynamical model and are, as such, qualitatively similar to the true state.

The foregoing results have shown that the GRF observation error model improves the ESS without substantially damaging the RMSE, and that the posterior mean is smoother than the true state but the individual ensemble members (particles) are not too smooth. We finally test whether the uncertainty quantification afforded by the particle filter is improved by using a GRF observation error model. To this end we compute the CRPS at each point of the spatio-temporal grid of 2048 $\times$ 100 points. The median CRPS is computed using all 204,800 spatio-temporal grid points for 101 values of ℓ^2 equally spaced between 0 and 1. The result is shown in Fig. 2.4. Median CRPS with $N_y = 64$ improves from about 0.27 at $\ell^2 = 0$ to 0.22 at $\ell^2 = 0.3$, and then remains steady or slightly increases at larger ℓ^2.[2] Some sampling variability is still evident in the median CRPS, with occasional values as low as 0.21.

Varying the number of observations, also shown in Figure 2.4, displays additional interesting behavior about the distributional estimate the particle filter provides. In each N_y case we explored, there is a choice of ℓ^2 that improves the particle filter CRPS. The differences in optimal ℓ^2 empha-

[2] For comparison, the ensemble spread simultaneously improves by a factor of about 2, going from a time-averaged 36% of RMSE when $\ell^2 = 0$ to 71% RMSE when $\ell^2 = 1$.

sizes that the optimal parameter depends not only on the active scales in the underlying physics, but also on the resolution of the data.

There is less information to spare about physically important scales when observations are sparse (cf. $N_y = 16$), in which case there is only a narrow window of suitable choices for $\ell^2 \approx 0.12$ before the smoothing effect deteriorates the predictive quality of the particle filter by over-suppressing active scales in the observations.

On the other hand, dense observations provide more abundant small-scale information that makes the particle filtration more susceptible to collapse. This necessitates a larger choice of ℓ^2 to achieve optimal particle filter performance. In this case, the more abundant information in denser observations can compensate for the injury we do to the surrogate posterior by more aggressively smoothing away small scales. Indeed the most dramatic improvement in the particle filter's uncertainty quantification occurs for $N_y = 128$. Here the particle filter greatly struggles for small ℓ^2, where we observe a CRPS over 0.29; however when $\ell^2 \approx 0.7$ the CRPS dips under 0.22, competitive with that of all other observation models considered here. This suggests that smoothing is particularly helpful in improving the particle filter's overall probabilistic estimate when observations are dense.

The CRPS results show that the particle filter's uncertainty quantification is improved by the GRF likelihood: a 25% decrease (improvement) in CRPS is comparable to the improvement achieved by various statistical post-processing techniques for ensemble forecasts (Feldmann et al., 2015, Kleiber et al., 2011a,b, Scheuerer and Büermann, 2014). Somewhat surprisingly, the CRPS significantly improves moving from $\ell^2 = 0$ to $\ell^2 = 0.1$ despite the fact that the ESS remains quite small. Overall, these CRPS results suggest that even small improvements in ESS can substantially improve the quality of the probabilistic state estimate. They also confirm that improving the ESS due to increasing ℓ^2 must be considered in balance against the consequent departure from the true posterior; the CRPS does not improve at large ℓ^2, even though the ESS improves, because the surrogate posterior becomes less realistic.

Figure 6 demonstrates how SIR uncertainty quantification depends on ensemble size. The

figure shows a kernel density estimate of CRPS over all 2048 grid points and all 100 timesteps, for varying number of particles $N_p \in (100, 200, 400, 800, 1600)$. The CRPS mode remains unchanged, but the mean decreases as the distribution concentrates around the mode primarily at the expense of mass in the tail. The weak dependence of CRPS on ensemble size underscores the appeal of improving UQ by other means.

2.5 Conclusions

We have demonstrated theoretically (in the framework of Bengtsson et al. (2008) and Snyder et al. (2008)) and in a simple experiment that the number of particles required to avoid collapse in a particle filter can be significantly reduced through a judicious construction of the observation error model. This observation error model has large observation error variance at small scales, which reduces the effective dimensionality and focuses attention on the more dynamically-relevant large scales. This observation error model is equivalent to smoothing observations before proceeding as though the observations are uncorrelated. The cost of this approach is that it alters the posterior, leading to a smoother posterior mean. In practice, a balance will need to be found between avoiding collapse and retaining as much observational information as possible.

An observation error model whose variance increases at small scales is associated with a so-called generalized random field (GRF). This connection allows for rapidly applying the covariance matrix's inverse (which is required to compute the particle weights) using fast numerical methods for self-adjoint elliptic partial differential equations. The method can also be implemented by smoothing the observations before assimilating them, and then assimilating the smoothed observations with an assumption of independent errors. Both of these avenues are amenable to serial processing of observations, as required by certain parallel implementations (e.g. Anderson and Collins, 2007). All of these approaches are compatible with periodic or aperiodic domains.

The results of the one-dimensional stochastic partial differential equation show that this approach improves the 'effective sample size' (ESS), which measures how well the weights are balanced between the particles, by an order of magnitude. The root mean squared error of the

particle filter's posterior mean is not significantly impacted by the approach. One of the main motivations for using particle filters is that they provide meaningful uncertainty estimates even in problems with nonlinear dynamics and observations, and non-Gaussian distributions. Thus, the continuous ranked probability score (CRPS) is used to test the quality of the particle filter's associated probability distribution. The GRF observation error model improves the CRPS by as much as 25%, which is a large improvement, comparable to results obtained by statistical post-processing of the ensemble (e.g. Feldmann et al., 2015, Kleiber et al., 2011a,b, Scheuerer and Büermann, 2014). This improvement in CRPS is obtained even when the effective sample size (ESS) is less than 20 out of 400, which shows that good probabilistic state estimation can be achieved even with ESS much less than the ensemble size. The theoretical results suggest that an ensemble size on the order of 8000 is required to avoid collapse in this example problem. Good results are obtained with an ensemble size of 400, even though the ensemble does collapse from time to time.

The theory of Snyder et al. (2008) estimates the ensemble size required to avoid collapse, which is unrealistically large for typical meteorological applications using standard observation error models. Using a GRF observation error model increases the ESS for a fixed ensemble size, making it easier to achieve the goal of avoiding collapse. The approach advocated here may still prove insufficient to enable particle filtering of weather, ocean, and climate problems; the minimum required ensemble size will be reduced, but may still be impractically large. Happily, the method is entirely compatible with approaches based on altered proposals (Ades and Van Leeuwen, 2015, Chorin and Tu, 2009, van Leeuwen, 2010) and with localization methods (Penny and Miyoshi, 2016, Poterjoy, 2016, Rebeschini and Van Handel, 2015). The method is also compatible with ensemble Kalman filters and with variational methods, but it is not clear whether the approach would yield any benefit there.

Indeed, dynamics of extratropical synoptic scales are often assumed to be approximately linear and are easily estimated with an ensemble Kalman filter. But ensemble Kalman filters do not provide robust uncertainty quantification in the face of nonlinear observation operators or

nonlinear dynamics, e.g. at synoptic scales in the tropics. In contrast, the method proposed here has the potential to provide robust uncertainty quantification even with nonlinear dynamics and observations. However, it is still unknown in what contexts our peculiar error model damages the posterior more severely than approximating the system as linear and Gaussian for the sake of assimilating data with ensemble Kalman filters. We expect performance comparison to be context-dependent, and hope future work will help reveal how to balance advantages and disadvantages that are relevant in practice.

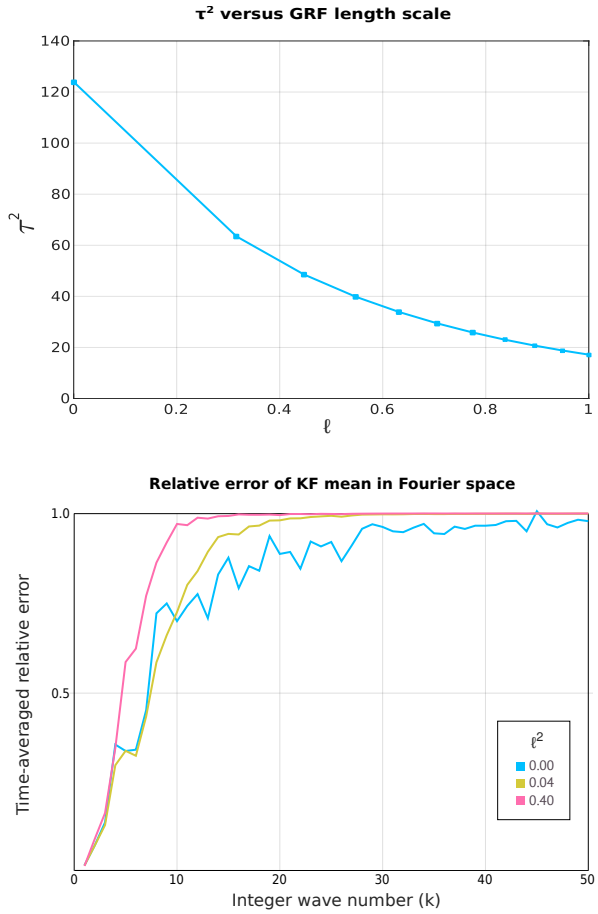


Figure 2.1: The left panel shows τ^2 (2.4) for different values of GRF length scale ℓ. Because the number of particles required to avoid degeneracy increases exponentially in $\tau^2/2$, the observed decrease in τ^2 as we roll off scales greater than ℓ indicates a reduced computational burden in using particle filtering for uncertainty quantification. Similarly, the decrease suggests that for fixed computation cost one may be able to mitigate the variance underestimation that tends to plague particle filters in high dimensions. Although the ordinate in this figure is ℓ to make direct contact with the length scale, all other figures are given in terms of ℓ^2 to relate more directly to the spectrum of the GRF likelihood. The panel on the right shows the RMS error in the Kalman Filters posterior mean, in Fourier space, normalized by the climatological standard deviation of each Fourier coefficient for different values of ℓ^2. Here we see how the error in the posterior mean, considered as a function of wavenumber, approaches the climatological standard deviation more rapidly when ℓ^2 is larger. It is exactly this posterior variance increase at small scales that underpins our approach: a posterior with larger total variance is easier for a particle filter to sample, while keeping the posterior accurate at large scales is key in forecast.

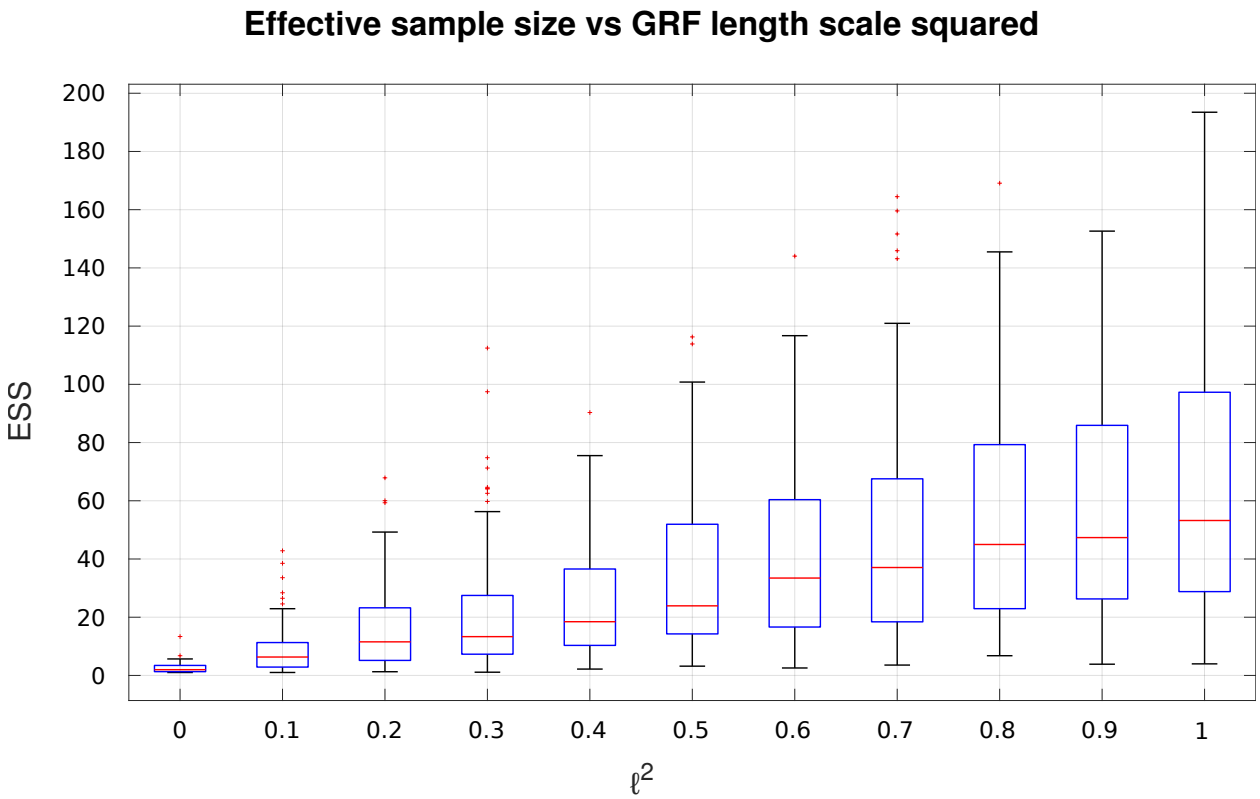


Figure 2.2: Effective sample size (2.2) distributions for different values of ℓ^2 from 0 to 1. Each box represents the middle 50% quantile, a central line representing the median, and the whiskers span the data not considered outliers by the 1.5×IQR rule.

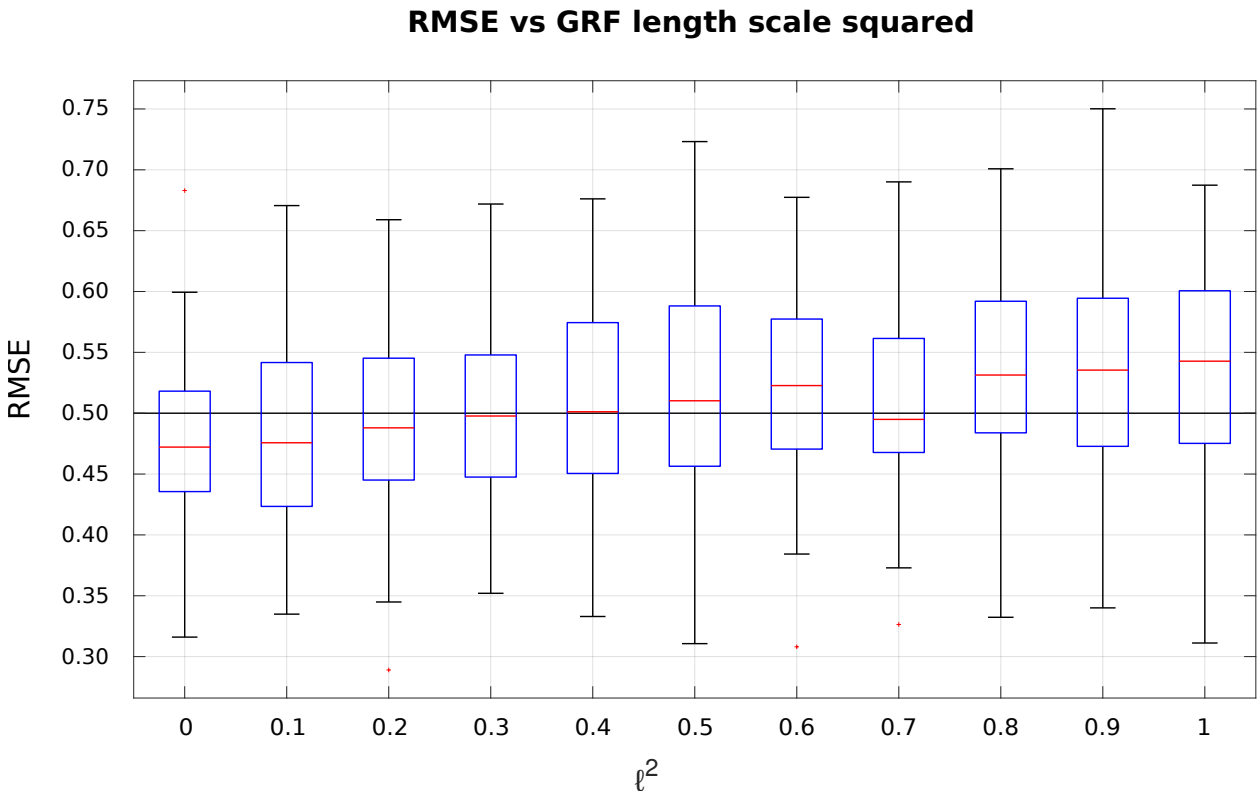


Figure 2.3: Root mean squared error (RMSE) between the truth and the posterior mean, using 11 different values of ℓ^2 from 0 to 1. The first category, with $\ell^2 = 0$, corresponds to the uncorrelated observation error model. The RMSE using GRF likelihoods, i.e. $\ell^2 > 0$, does not dramatically suffer in comparison to that of the white likelihood that is more common in operational practice. In exchange for this small cost in RMSE, using the GRF likelihood comes with notable gain in the accuracy of uncertainty quantification. Each box represents the middle 50% quantile, a central line representing the median, and the whiskers span the data not considered outliers by the 1.5×IQR rule. The horizontal line at 0.5 serves only to guide the eye.

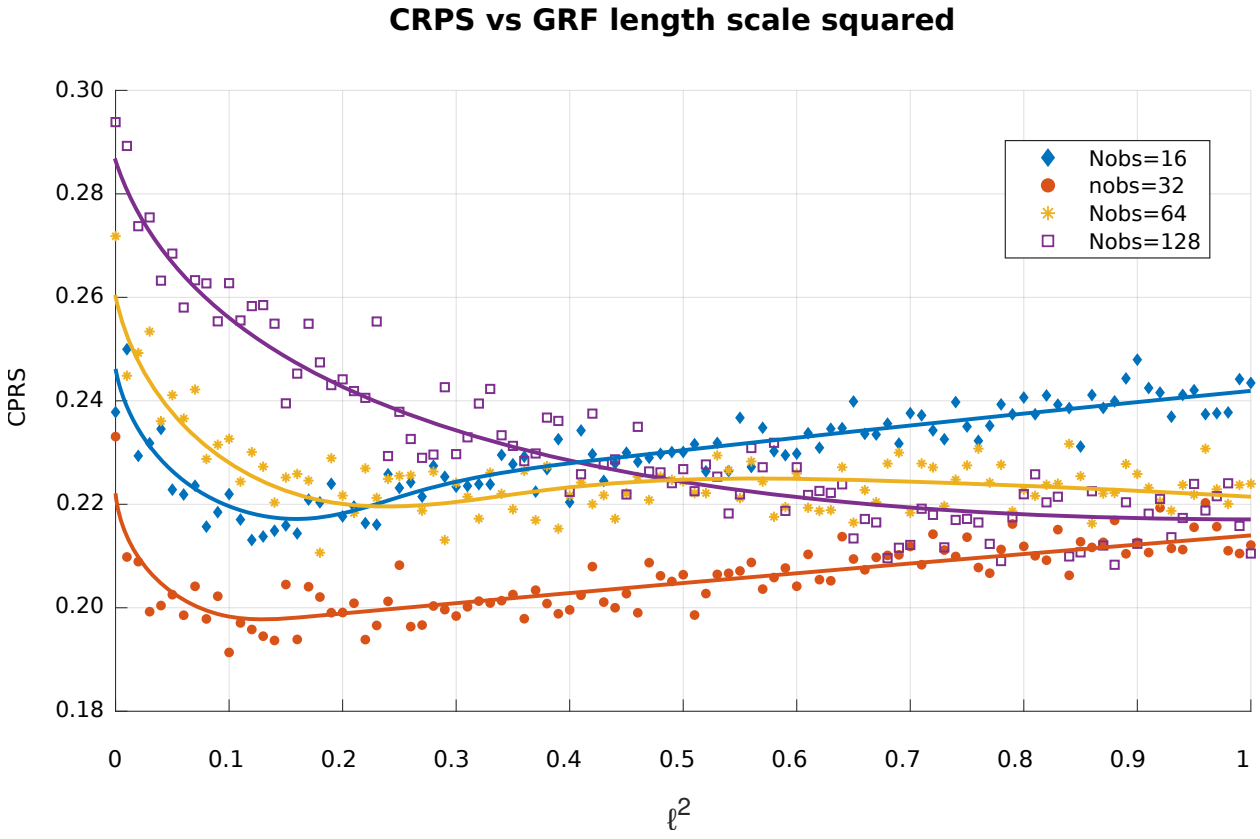


Figure 2.4: Continuous ranked probability score median over all time steps and grid locations, shown as a function of ℓ^2. Each point plotted represents a particle filter assimilation run, with the same true and observed data, for different values of squared GRF length scale ℓ^2. Each marker style represents different numbers of observations, demonstrating how the particle filter is sensitive to the number of observations. The traces are spline approximations of the data that serve to guide the eye. In each N_y case we explored, there is a choice of ℓ^2 that improves the particle filter CRPS. This plot emphasizes that the optimal choice of ℓ^2 depends not only on the active scales in the underlying physics, but also on the resolution of the data. There is less information to spare about physically important scales when observations are sparse (cf. $N_y = 16$), in which case there is only a narrow window of suitable choices for $\ell^2 \approx 0.12$ before the smoothing effect deteriorates the predictive quality of the particle filter. On the other hand, dense observations provide more abundant small-scale information that necessitates a larger choice of ℓ^2 to achieve optimal particle filter performance. Fortunately, the more abundant information in denser observations can compensate for the injury we do to the surrogate posterior by more aggressively smoothing away small scales.

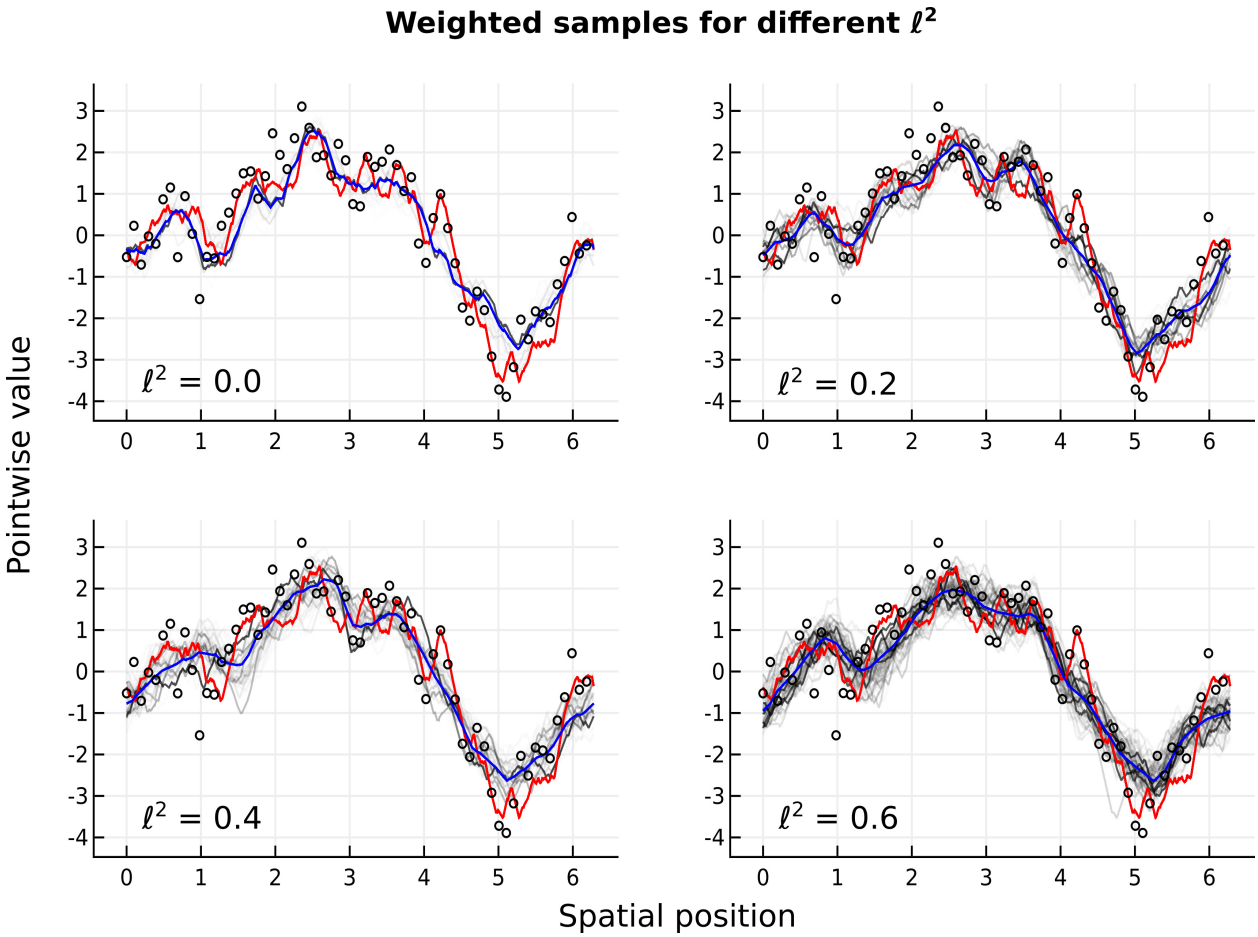


Figure 2.5: Pictured are the true state (red trace), PF mean (blue trace), observations (black circles), and samples from the posterior visually weighted with darkness proportional to sample weight (gray traces) for different values of $\ell^2 \in (0.0, 0.2, 0.4, 0.6)$ from left to right and top to bottom. This panel demonstrates again how a small change to the likelihood can substantially improve the problem of underestimating variance, and that this effect comes with diminishing marginal returns as the surrogate model yields progressively smoother estimates of the posterior mean. Observe also that the samples are all realistic instantiations of the physical process, rather than overly smooth estimates. The assimilation time shown here was chosen to exhibit monotonic improvement in ℓ^2, which is the time-averaged behavior; due to the probabilistic nature of particle filtering, there is an abundance of times when there is not such monotonic improvement.

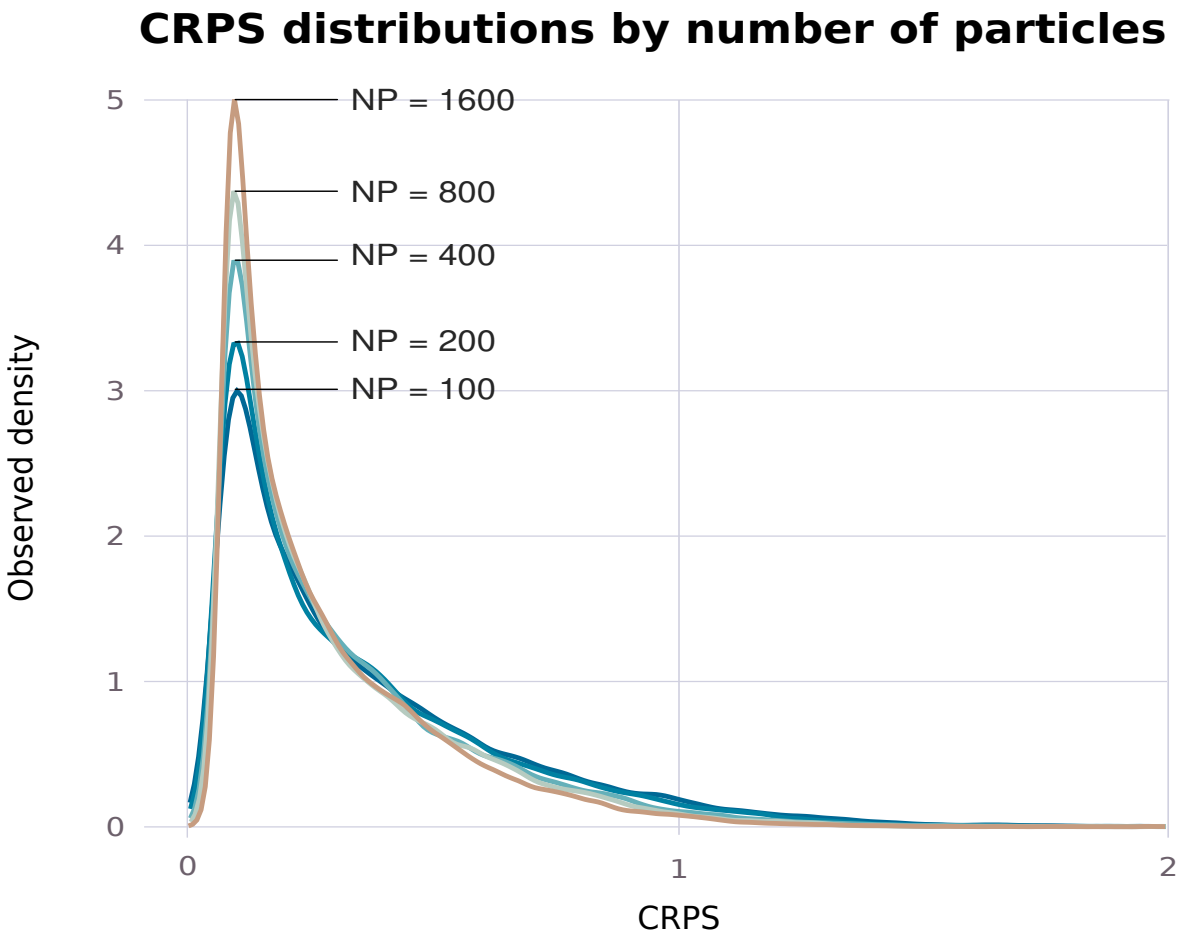


Figure 2.6: Kernel density estimates (KDE) of the CRPS observed for different numbers of particles demonstrate the concentration of probability as the number of particles increases while $\ell^2 = 0.30$ and $N_y = 64$ are held fixed, for a fixed simulation and fixed observations thereof. Each KDE is built from the CRPS computed for each of 2048 grid cells and all 100 timesteps. The slow convergence in the number of particles is one of the reasons it is attractive to seek other means of making the particle filter more effective in sampling high-dimensional distributions.

Chapter 3

A tunable smoother for scattered measurements with application to particle filtering

3.1 Introduction

We present a smoothing[1] algorithm that works for data measured at scattered points in $\mathbb{R}^d$, for arbitrary dimension $d > 0$, and that permits choice in the shape of its spectrum. This smoothing algorithm is equivalent to solving a positive definite self-adjoint elliptic partial differential equation.

Such a smoother is motivated by an effort to mitigate the dimensional curse of particle filtering, which we will refer to as sequential importance sampling with resampling (SIR) hereinafter.[2] This introduction will emphasize the derivation of the smoother from considerations about SIR, in terms of an observation covariance matrix, and the smoother itself is described in section 3.2.

SIR is susceptible to a phenomenon called **collapse**, characterized by essentially all the ensemble weight accumulating on a single ensemble member that is closest to the observations, causing the filter to catastrophically underestimate posterior dispersion. The number of ensemble members required to avoid SIR collapse depends on system covariance and observation error covariance, scaling exponentially in an effective system dimension.

Specific estimates of ensemble size required to avoid collapse, provided in Snyder et al. (2008,

[1] This paper uses the term **smoothing** as it appears in the statistics and image processing literature, to denote the process of attenuating noise while preserving important patterns. In more precise signal processing vernacular, we introduce a spatial-domain filter with a gradual low-pass effect. This is not to be confused with the term smoothing as in the Kalman smoother, though the concepts are related.

[2] Sequential importance sampling (SIS) is also known as **particle filtering**. In this paper we specialize on SIS with resampling (SIR), the most famous variety of particle filter, though the dimensional curse and presumably our applications also extend to other particle filter varieties.

2015), suggest one can reduce the required ensemble size by increasing eigenvalues of the observation error covariance.[3] Doing so carefully can also improve uncertainty quantification for a fixed number of ensemble members. Ref. Robinson et al. (2018) suggests inflating the observation error variance at small scales, letting variance grow in wavenumber, since small scales have very limited predictability in geophysical fluids (Judt, 2018, Lorenz, 1969, Rotunno and Snyder, 2008).

To be more precise, consider an observing system

$$y(\boldsymbol{q}) = \mathscr{H}\{x\}(\boldsymbol{q}) + r^{1/2}(\boldsymbol{q})\ \epsilon(\boldsymbol{q}), \tag{3.1}$$

where $y(\boldsymbol{q}) \in \mathbb{R}$ is the observation at location $\boldsymbol{q} \in \mathbb{R}^d$, $\mathscr{H}$ is a function-valued observation operator acting on x, which describes the scalar system state as a function of location, $r^{1/2}(\boldsymbol{q})$ can be imagined as the standard deviation of the observation error at $\boldsymbol{q}$, and $r^{1/2}(\boldsymbol{q})\ \epsilon(\boldsymbol{q})$ is the random observation error.

It is natural to think of ϵ as a random field. But letting the spectrum grow in wavenumber precludes pointwise definition of ϵ, with probability 1, so it is not a random field in the traditional sense. The idea of imposing a correlation structure with a growing spectrum can instead be understood in the framework of generalized random fields. In this case ϵ can be treated as a random process with realizations taking the form of tempered distributions, i.e. elements of the topological dual to a Schwartz space of rapidly decaying functions on $\mathbb{R}^d$.

Since realizations of ϵ with a growing spectrum cannot be described pointwise, instead interpret the spatially parametrized terms in eq. (3.1) as averages with respect to a Schwartz function ν that is closely concentrated near $\boldsymbol{q}$. For example,

$$\epsilon(\boldsymbol{q}) \equiv \int_{\mathbb{R}^d} \epsilon\ \nu\ \mathrm{d}x \Big/ \int_{\mathbb{R}^d} \nu\ \mathrm{d}x\ . \tag{3.2}$$

Narrowing our attention within the scope of generalized random fields, let ϵ be a mean-zero stationary Gaussian generalized random field (GGRF). Then the vector $\big(\epsilon(\boldsymbol{q}_1), \cdots, \epsilon(\boldsymbol{q}_{N_y})\big)$ is a multivariate normal random variable with zero mean and a covariance matrix $\boldsymbol{C}$ with entries C_{ij} that

[3] Decreasing eigenvalues of the system covariance has the same effect, for the same reasons, but it is potentially harder to justify and implement manipulations to the dynamical model.

depend only on $\|\boldsymbol{q}_i - \boldsymbol{q}_j\|$. Hence the vector of observations $\boldsymbol{y} \equiv \big(y(\boldsymbol{q}_1), \cdots, y(\boldsymbol{q}_{N_y})\big)$ conditioned on x is a multivariate normal random variable with mean $H(\boldsymbol{x})$ and covariance

$$\mathbf{R} = \mathbf{R}_0^{1/2} \mathbf{C} \mathbf{R}_0^{1/2},$$

where $\boldsymbol{R}_{\mathbf{0}}^{1/2}$ is a diagonal matrix of the discrete observation standard deviations $r^{1/2}(\boldsymbol{q}_i)$ that can be treated as instrument errors and $H(\cdot) : \mathbb{R}^{N_y} \to \mathbb{R}^{N_y}$ is an observation operator acting on the discrete vector $\mathbf{x}$ that characterizes the underlying system state. The discrete observing system can be summarized in the form

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \boldsymbol{R}_0^{1/2} \boldsymbol{\epsilon} \in \mathbb{R}^{N_y}. \tag{3.3}$$

We specifically treat the continuous field of observation error as $\epsilon = \mathscr{D}\mathcal{W}$, where

$$\mathscr{D} = \big(1 - \ell^2 \Delta\big)^{\beta} \tag{3.4}$$

is the fractional bound-state Helmholtz operator acting on a spatial white noise $\mathcal{W}$ with mean zero and unit pointwise variance, Δ is the formal Laplacian operator, $\ell > 0$ is a tuning parameter with dimensions of length, and $\beta > 0$ is a dimensionless tuning parameter that controls the rate of growth of eigenvalues. Eigenfunctions of $\mathscr{D}$ are Fourier modes of wavenumber k and corresponding eigenvalues $(1 + \ell^2 |k|^2)^{\beta}$. The characteristic scale of this operator is $\ell / (2\pi \sqrt{2^{1/\beta} - 1})$, in the sense that eigenfunctions with length scales longer than this have corresponding eigenvectors close to 1. Modeling ϵ in this manner therefore ascribes a variance to large scales that is commensurate with instrument error, but it also progressively and unboundedly inflates variance for small scales at a rate controlled by β. The GGRF description of observation error is thus a kind of surrogate model for the assumption of uncorrelated observations at large scales, but with inflated variance at small scales that are of lesser concern in geophysical forecasting.

We will use the fact that preferentially inflating observation variance at small scales is equivalent to treating smoothed innovations as uncorrelated.[4] To see this equivalence, observe how the

[4] The **innovation** of an ensemble member is the difference between observation and forecast.

correlation matrix features in the Gaussian likelihood, the logarithm of which is proportional to

$$(\mathbf{y}-\mathbf{H}\mathbf{x})^T\mathbf{R}_0^{-1/2}\mathbf{C}^{-1}\mathbf{R}_0^{-1/2}(\mathbf{y}-\mathbf{H}\mathbf{x}). \tag{3.5}$$

Then consider preprocessing the standardized innovations with a linear operation

$$\mathbf{R}_0^{-1/2}(\mathbf{y}-\mathbf{H}\mathbf{x}) \quad \longmapsto \quad \mathbf{S}\mathbf{R}_0^{-1/2}(\mathbf{y}-\mathbf{H}\mathbf{x}). \tag{3.6}$$

If these smoothed observations are now assimilated under the assumption that the errors in the smoothed field are standard normal, then the log-likelihood is proportional to

$$(\mathbf{y}-\mathbf{H}\mathbf{x})^T\mathbf{R}_0^{-1/2}\mathbf{S}^T\mathbf{S}\mathbf{R}_0^{-1/2}(\mathbf{y}-\mathbf{H}\mathbf{x}). \tag{3.7}$$

If $\mathbf{S}$ is a positive definite smoothing operator – i.e. a positive definite operator with a decaying spectrum toward small scales – then $\mathbf{C}=(\mathbf{S}^T\mathbf{S})^{-1}$ is a positive definite operator with a spectrum that grows toward small scales.

Regularly-spaced data on a periodic domain would enable straightforward application of Fourier methods to implement a smoother that obeys a desired spectrum. But high-dimensional data assimilation problems in geophysics often involve measurements made at irregularly scattered locations in a spatially-extended domain, which is the purpose for seeking a smoother that does not require a regular grid and that provides a freedom to shape the degree of smoothing at different length scales.

Our approach to smoothing irregularly scattered data $\mathbf{z}$ is to construct a Gaussian radial basis function (RBF) interpolant of the data, and then apply $\mathscr{D}^{-1}$ to this interpolant. Evaluating this smoothed interpolant at the data locations yields the vector $\mathbf{S}\mathbf{z}$. In the spirit of Lindgren et al. (2011), Rue and Held (2005), the connection between elliptic stochastic partial differential equations and random fields enables us to make use of fast algorithms for PDEs in the context of GGRFs, rather than naively developing a dense approximation of $\mathbf{C}$ and then solving the associated linear system.

The paper is organized as follows. Our method for smoothing data measured on a spatially-extended domain is described in section 3.2, an illustrative example is shown in section 3.3, another

example using real meteorological data is in section 3.4 to show the smoother has the desired effect on SIR, discussion about algorithmic complexity and generalizations toward practical application of our method is in section 3.5, and conclusions follow in section 3.6.

3.2 Method

Let $\mathbf{z} \in \mathbb{R}^{N_z}$ be a vector of standardized innovations at locations $\mathbf{q}_i \in \mathbb{R}^d$ for each $i \in (1, \cdots, N_z)$, where N_z is the number of observations. Define $\mathscr{D} = (1 - \ell^2\Delta)^\beta$ with the same parameters as described of eq. (3.4). The proposed smoother $\mathbf{S}$ works by solving for a discrete approximation of $\mathscr{D}^{-1}\zeta$, where $\zeta : \mathbb{R}^d \to \mathbb{R}$ is a continuous-domain interpolant of the data $\mathbf{z}$ expressed as a sum of Gaussians. This obtains by approximating the Green's function g of $\mathscr{D}$ as a multiresolution sum of Gaussians, computing the convolution of that approximation with ζ, and evaluating the result at the locations $\{\mathbf{q}_i\}$.

A drawback of this approach is that convolution with g attenuates all but constant functions on $\mathbb{R}^d$, so even a constant data vector $\boldsymbol{z}$ will be attenuated to some degree. We will discuss a way to mitigate this effect in section 3.3.

To represent the data in a continuous form that allows convolution with g, we choose radial basis function (RBF) interpolation (Fornberg and Flyer, 2015). RBF interpolation of the observations requires us to choose a kernel $\psi : \mathbb{R}^+ \to \mathbb{R}$ that is used as the **radial basis** in which the interpolant will represent the data. The interpolant takes the form

$$\zeta(\cdot) = \sum_{j=1}^{N_z} b_j \psi(\| \cdot - \mathbf{q}_j \|), \tag{3.8}$$

where (b_j) are interpolation weights such that

$$\zeta(\mathbf{q}_i) = \sum_{j=1}^{N_z} b_j \psi(\|\mathbf{q}_i - \mathbf{q}_j\|) = z_i. \tag{3.9}$$

We write this linear system in matrix form as

$$\mathbf{B}\mathbf{b} = \mathbf{z}. \tag{3.10}$$

We take the RBF kernel to be $\psi(\|\cdot\|) = \phi(\cdot; 0, \xi\mathbf{I})$, where $\phi(\cdot; \mu, \mathbf{\Sigma})$ is the density of a d-variate Gaussian random variable with mean μ and covariance $\mathbf{\Sigma}$

$$\phi(\cdot; \boldsymbol{\mu}, \mathbf{\Sigma}) = (2\pi \det \mathbf{\Sigma})^{-d/2} \exp\left((\cdot - \boldsymbol{\mu})^T \mathbf{\Sigma}^{-1} (\cdot - \boldsymbol{\mu})\right). \tag{3.11}$$

This notation is used as a convenient description of Gaussian functions even though we will not use them to describe any random variables.

One may worry that this method is hardly fast, despite using a fast PDE solver, since a naive approach to solving for $\mathbf{b}$ requires $\mathcal{O}(N_y^3)$ operations. Computational complexity of our algorithm, including faster alternatives to solving for $\boldsymbol{b}$, is discussed in section 3.5.

The multiresolution Gaussian approximation of the Green's function begins by writing the Fourier transform of the inverse of eq. (3.4) as an inverse-power function $\hat{g}(t) = t^{-\beta}$ where $t = 1 + \ell^2|k|^2$. Exponential approximations of inverse power functions like this are studied in Beylkin and Monzón (2010), McLean (2018). The approach therein is to write a finite trapezoid-type discretization of an integral representation of $\hat{g}$. With the change of variables introduced by McLean, the integral representation to be discretized is

$$\frac{1}{t^\beta} = \frac{1}{\Gamma(\beta)} \int_{-\infty}^{\infty} \exp\left(-\varphi(x,t)\right)\left(1 + e^{-x}\right)\, dx, \tag{3.12}$$

where

$$\varphi(x,t) = t\exp(x - e^{-x}) - \beta(x - e^{-x}). \tag{3.13}$$

The finite trapezoid rule discretizes this into

$$\frac{1}{t^\beta} \approx \frac{1}{\Gamma(\beta)} \sum_{n=-M_-}^{M_+} v_n e^{-a_n t}, \tag{3.14}$$

with

$$a_n = \exp(nh - e^{-nh}), \tag{3.15}$$

$$v_n = h(1 + e^{-nh}) \exp\left(\beta(nh - e^{-nh})\right). \tag{3.16}$$

Ref. McLean (2018) Lemma 4 shows that the total required number of terms $M_- + M_+ + 1$ scales as $(\ln E)^2$ to achieve uniform relative error bounded by $E > 0$ in the limit $E \to 0$.

The approximation in eq. (3.14) can now be rewritten in terms of normalized multivariate isotropic Gaussian functions of k. Given weights v_n and exponential rates a_n from the exponential approximation above, we can derive the multiplicative factors required of this equivalent formulation:

$$v_n e^{-a_n(1+\ell^2 k^2)} = v_n e^{-a_n} e^{-a_n \ell^2 k^2} \tag{3.17}$$

$$= v_n e^{-a_n} (2\pi/2\ell^2 a_n)^{d/2} \left((2\pi/2\ell^2 a_n)^{-d/2} e^{-2\ell a_n^2 k^2/2} \right) \tag{3.18}$$

$$= v_n e^{-a_n} (\pi/\ell^2 a_n)^{d/2} \phi(k; 0, 1/2\ell^2 a_n). \tag{3.19}$$

The second line obtains from simultaneously multiplying and dividing by the constant required to normalize the Gaussian term in large parentheses, which is written in that manner to ease visual comparison to the standard form of an isotropic d-variate Gaussian probability density function of mean 0 and variance $1/2\ell^2 a_n$. Combining eq. (3.14)-eq. (3.19) yields

$$\frac{1}{(1+\ell^2|k|^2)^\beta} \approx \frac{1}{\Gamma(\beta)} \sum_{n=-M_-}^{M_+} v_n e^{-a_n} (\pi/\ell^2 a_n)^{d/2} \phi(k; 0, 1/2\ell^2 a_n). \tag{3.20}$$

A plot of the relative error committed by this approximation is shown in fig. 3.1.

Taking the inverse Fourier transform and combining terms finally yields the desired approximation of the Green's function in physical space in terms of normalized Gaussians:

$$g(\cdot) \approx \sum_{n=-M_-}^{M_+} c_n \phi(\cdot; 0, \rho_n), \tag{3.21}$$

$$\rho_n = 2\ell^2 a_n, \tag{3.22}$$

$$c_n = \frac{v_n (\pi/\ell^2 a_n)^{d/2-1}}{\Gamma(\beta) e^{a_n}}. \tag{3.23}$$

With approximations of the data and the Green's function now constructed in terms of d-variate isotropic normalized Gaussians, convolution of which is trivial, applying a discrete version

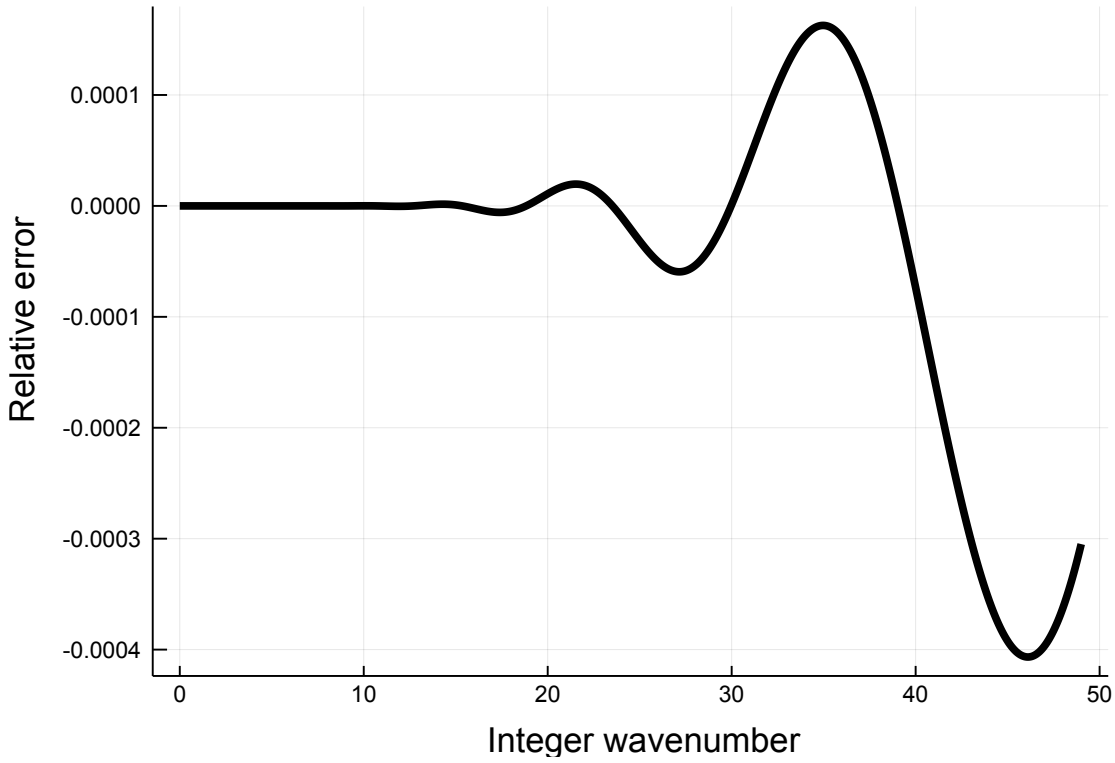


Figure 3.1: Relative error of a Gaussian approximation in the form eq. (3.19) of the Fourier-space Green's function $(1 - \ell^2 k^2)^{-1/\beta}$ with parameters $\ell = 1.0$, $\beta = 0.5$, $h = 0.2$, $M = 32$, and $N = 28$.

of the integral operator that inverts $\mathscr{D}$ is just:

$$\mathscr{D}^{-1}y \approx \left(\sum_{i=-M_-}^{M_+} c_i\phi(\cdot;0,\rho_i I)\right) * \left(\sum_{j=1}^{N_y} b_j\phi(\cdot;\mu_j,\xi I)\right) \tag{3.24}$$

$$= \sum_{i=-M_-}^{M_+}\sum_{j=1}^{N_y} c_i b_j \phi(\cdot;0,\rho_i I) * \phi(\cdot;\mu_j,\xi I) \tag{3.25}$$

$$= \sum_{j=1}^{N_y} b_j \left(\sum_{i=-M_-}^{M_+} c_i\phi(\cdot;\mu_j,(\rho_i+\xi_j)I)\right). \tag{3.26}$$

Evaluating this quantity at the observation locations yields the output of our smoother. The last line in the manipulation above shows that the continuous function we evaluate to arrive at outputs can be interpreted as an RBF interpolant of the smoothed data in terms of a new smoothed basis function $\tilde{\psi}$, given by

$$\tilde{\psi}(\cdot) = \sum_{i=-M_-}^{M_+} c_i\phi\left(\cdot;0,(\rho_i+\xi_j)I\right). \tag{3.27}$$

The weights of the smoothed interpolant are identical to the weights of the input interpolant, which will be valuable later in this section.

We now describe our smoothing algorithm more concretely. Algorithm 2.1 ties together pieces of the Green's function approximation specified in eq. (3.12)-eq. (3.23). Algorithm 2.2 combines RBF interpolation of the data with the output of Algorithm 2.1 with a convolution and evaluates the result at the data locations. These algorithms, used together, are a complete description of our smoother.

Algorithm 2.2 defines a linear operator $\mathbf{S}$ on $\mathbb{R}^{N_z}$. As described in section 3.1, the application we propose is to smooth standardized innovations with $\mathbf{S}$ in such a way that $\mathbf{S}^T\mathbf{S}$ is a covariance for a discretized GGRF. To be a valid covariance matrix, $\mathbf{S}^T\mathbf{S}$ must be symmetric and positive definite. These conditions directly follow if $\boldsymbol{S}$ is positive definite.

We will prove that $\mathbf{S}$ is positive definite in theorem 1. The theorem is more general than the specific algorithm so far presented, which will set the stage for potential variants to be described in section 3.5. To ease into the theorem, we will summarize the preceding development of $\mathbf{S}$ and connect it to a briefer alternative formulation that is easier to treat analytically.

Algorithm 1 Gaussian approximation of fractional bound-state Helmholtz kernel

Input

Smoothing scale parameter $\ell > 0$.
Smoothing shape parameter $\beta > 0$.
Integration mesh size $h > 0$.
Number of negative integration steps $M_- \in \mathbb{N}$.
Number of positive integration steps $M_+ \in \mathbb{N}$.
Dimension of each measurement location $d \in \mathbb{N}$.

Output

Vector of positive weights $\boldsymbol{c} \in \mathbb{R}^{M_-+M_++1}$.
Vector of variances $\boldsymbol{\rho} \in \mathbb{R}^{M_-+M_++1}$.

function GaussianBSH(ℓ, β, h, M_-, M_+, d)
  for $n = -M_- \to M_+$ **do**
    $\hat{a} \leftarrow \exp(nh - \exp(-nh))$.
    $\hat{w} \leftarrow h\,(1 + \exp(-nh)) \exp(\beta\,(nh - \exp(-nh)))$.
    $\hat{w}' \leftarrow \hat{w} \exp(-\hat{a}) \left(\pi/\ell^2 \hat{a}\right)^{d/2} / \Gamma(\beta)$.
    $\rho_n \leftarrow 2\ell^2 \hat{a}$.
    $c_n \leftarrow \hat{w}' \rho / 2\pi$.
  end for
  return $\boldsymbol{c}, \boldsymbol{\rho}$.
end function

Algorithm 2 Smoother

Input

Vector of measured data $\mathbf{z} \in \mathbb{R}^{N_z}$.
Array of data location vectors $\mathbf{q}_i \in \mathbb{R}^d$, $i \in (1, \ldots, N_z)$.
RBF scale parameter $\xi > 0$.
Vector of positive weights $\boldsymbol{\omega} \in \mathbb{R}^{M_-+M_++1}$.
Vector of variances $\boldsymbol{\rho} \in \mathbb{R}^{M_-+M_++1}$.

Output

Array of smoothed data $\tilde{\mathbf{z}}_i \in \mathbb{R}$, $i \in (1, \ldots, N_z)$

function Smooth($\mathbf{z}$, $\mathbf{q}$, ξ, $\boldsymbol{\omega}$, $\boldsymbol{\rho}$)
  Let $\phi(\|\cdot\|; 0, \xi\mathbf{I})$ be a unit-mass Gaussian to use as the RBF kernel.
  Generate RBF weight matrix $\mathbf{B}$ with elements:
  for $i = 1 \to N_z, j = 1 \to N_z$ **do**
    $B_{ij} \leftarrow \phi(\|\mathbf{z}_i - \mathbf{q}_j\|; 0, \xi\mathbf{I})$.
  end for
  $\mathbf{b} \leftarrow \mathbf{B}^{-1}\mathbf{z}$.
  for $i = 1 \to N_y$ **do**
    $\tilde{z}_i \leftarrow \sum_{i',n} (b_i \cdot \omega_n) \phi(\|y_i - \boldsymbol{q}'_i\|; 0, (\xi + \rho_n) I)$.
  end for
  return $\tilde{\mathbf{z}}$.
end function

We described a sequence of mappings between vector spaces, with smoothing taking place most explicitly in the function space $L^2(\mathbb{R}^{N_y})$ of interpolants by way of the convolution $\psi_i \mapsto \mathscr{G}\psi_i$, where $\mathscr{G}$ denotes an operator that performs convolution with the Gaussian approximation of g, and ψ_i is defined as the interpolation basis function $\psi(\|\cdot - \mathbf{q}_i\|)$ centered at location $\mathbf{q}_i$. Taken literally, that conceptual development prescribes the following composition of linear operations:

$$Y \xrightarrow{\mathbf{B}^{-1}} W \xrightarrow{\mathscr{F}} X \xrightarrow{\mathscr{G}} \tilde{X} \xrightarrow{\tilde{\mathscr{F}}^{-1}} \tilde{W} \xrightarrow{\tilde{\mathbf{B}}} \tilde{Y}. \tag{3.28}$$

Nodes in this diagram represents the various vector spaces found along the way of describing our smoothing algorithm:

- Y is the space of standardized innovations,
- W is the space of interpolant weights in the basis $\{\psi_i\}$,
- $X = \text{span}\{\psi_i\} \subset L^2(\mathbb{R}^n)$ is the space of interpolants,
- $\tilde{X} = \text{span}\{\mathscr{G}\psi_i\} \subset L^2(\mathbb{R}^n)$ is the space of smoothed interpolants,
- $\tilde{W}$ is the space of smoothed interpolant weights in the basis $\{\mathscr{G}\psi_i\}$, and
- $\tilde{Y}$ is the space of smoothed standardized innovations.

Arrows in the diagram represent the action of the operators superscribed on them:

- $\mathbf{B}^{-1}$ maps standardized innovations to RBF weights,
- $\mathscr{F}$ maps RBF weights to interpolated functions,
- $\mathscr{G}$ maps interpolated functions to smoothed functions,
- $\tilde{\mathscr{F}}^{-1}$ maps smoothed functions to weights in a smoothed RBF basis, and
- $\tilde{\mathbf{B}}$ maps smoothed weights to smoothed standardized innovations.

The complicated sequence of steps above can simplify greatly; observe in eq. (3.26) that the weights in the smoothed basis $\{\mathscr{G}\psi_i\}$ are always identical to the weights in the unsmoothed basis

$\{\psi_i\}$. Therefore $\tilde{\mathscr{F}}^{-1}\mathscr{G}\mathscr{F} = \mathbf{I}$, leaving just $\mathbf{S} = \tilde{\mathbf{B}}\mathbf{B}^{-1}$ where $\tilde{\mathbf{B}}$ is the RBF matrix in the smoothed RBF basis:

$$\tilde{B}_{i,j} = \tilde{\psi}(\|\mathbf{q}_i - \mathbf{q}_j\|). \tag{3.29}$$

This alternative perspective demonstrates that $\mathbf{S}$ is equivalent to finding weights $\mathbf{b}$ for an RBF interpolant of the unsmoothed data using a basis $\{\psi_i\}$, and then evaluating an RBF interpolant of the **smoothed** data using the same weights $\boldsymbol{b}$ that now act as coefficients on a smoothed basis $\{\mathscr{G}\psi_i\}$. The following theorem is stated in terms of this simplified perspective.

Theorem 1. *Let $\psi : \mathbb{R}^+ \to \mathbb{R}$ be an interpolating radial basis function and let $g : \mathbb{R}^+ \to \mathbb{R}$ be a convolution kernel. Suppose ψ and g each have positive Fourier transforms, and define $\tilde{\psi} = g * \psi$. Then the matrices $\mathbf{B}$ with entries $B_{ij} = \psi(\|\mathbf{q}_i - \mathbf{q}_j\|)$ and $\tilde{\mathbf{B}}$ with entries $\tilde{B}_{ij} = \tilde{\psi}(\|\mathbf{q}_i - \mathbf{q}_j\|)$ are symmetric positive definite, and the product $\mathbf{S} = \tilde{\mathbf{B}}\mathbf{B}^{-1}$ is positive definite.*

Proof. A standard theorem of RBF interpolation (e.g. Section 3 of Fornberg and Flyer (2015)) states that $\mathbf{B}$ is positive definite under the assumption that the Fourier transform of ψ is positive.

The Convolution Theorem guarantees that $\tilde{\psi} = g * \psi$ has a positive Fourier transform if g and ψ both have positive Fourier transforms, so $\tilde{\mathbf{B}}$ is positive definite for the same reason that $\mathbf{B}$ is positive definite.

Observe that $\mathbf{B}$ and $\tilde{\mathbf{B}}$ are symmetric by construction, and that $\mathbf{B}^{-1}$ is symmetric positive definite since it is the inverse of a symmetric positive definite matrix. Theorem 7.6.3 in Horn and Johnson (1990) states that the product of a positive definite matrix $\mathbf{P}$ and a Hermitian matrix $\mathbf{Q}$ is a matrix with the same number of negative, zero, and positive eigenvalues as $\mathbf{Q}$. It follows that the product of two Hermitian positive definite matrices is also positive definite. Therefore, since $\tilde{\mathbf{B}}$ and $\mathbf{B}^{-1}$ are Hermitian positive definite matrices, $\tilde{\mathbf{B}}\mathbf{B}^{-1}$ is positive definite. □

The coefficients in the multiresolution approximation eq. (3.21) are all positive, and the Fourier transform of a positive Gaussian is also a positive Gaussian. Therefore $\mathbf{S}$, as defined by Algorithm 2.2 together with Algorithm 2.1, is positive definite as a corollary of theorem 1.

Although theorem 1 shows that $\mathbf{S}$ is positive definite, it is typically **not** symmetric, and the symmetric part of $\mathbf{S}$ is not necessarily positive definite. For this reason we must treat $\mathbf{R}_0^{-1/2}\mathbf{S}^T\mathbf{S}\mathbf{R}_0^{-1/2}$ as a covariance matrix, rather than $\mathbf{R}_0^{-1/2}\mathbf{S}\mathbf{R}_0^{-1/2}$. We design $\mathbf{S}$ to approximate $\mathscr{D}^{-1}$.

3.3 Example 1: circular measurement locations embedded in a 2-plane

We want to verify that the smoother's effect resembles what we would expect of a discrete approximation to $\mathscr{D}^{-1}$. To that end, we smooth equally-spaced circular data embedded in $\mathbb{R}^2$ to provide insight into the spectral properties of the smoother in practice. This example will also describe heuristics in choosing parameters ℓ, β, and ξ. In the course of this example we will also demonstrate an undesirable phenomenon whereby $\mathbf{S}$ attenuates even the largest scales, and suggest a workaround.

Locations were chosen to encircle the origin in $\mathbb{R}^2$ with $N_y = 100$ distinct locations separated by unit distance from nearest neighbors, i.e.

$$\mathbf{q}_i = \left| e^{2in\pi/100} - 1 \right|^{-1} \begin{bmatrix} \cos(2n\pi/100) \\ \sin(2n\pi/100) \end{bmatrix}.$$

The interpolation kernel was chosen to be the isotropic Gaussian PDF with standard deviation $\xi^{1/2} = 2.5$. The convolution kernel is the multiresolution Gaussian approximation eq. (3.21) to the fractional bound-state Helmholtz kernel with $\ell = 1$ and $\beta = 1/2$, using approximation parameters $h = 0.2$, $M = 32$, and $N = 28$. These parameters yield an approximation of $\hat{g}$ with $< 0.05\%$ relative error up to $k_{max} = 49$, the Nyquist number for the one-dimensional problem that this example simulates embedded in two dimensions. Recall that fig. 3.1 shows this relative error as a function of k.

The smoother thus constructed defines a linear operator $\mathbf{S}$ on $\mathbb{R}^{N_y}$, and a precision matrix $\mathbf{S}^T\mathbf{S}$. For the purpose of inspecting the effect of smoothing at different scales for the purpose of improving SIR performance, we numerically constructed a matrix representation of $\boldsymbol{S}^T\boldsymbol{S}$, though for practical application of the smoothing algorithm it is inadvisable to actually construct $\mathbf{S}$. Due to the rotational symmetry of observation locations, the matrices $\boldsymbol{B}^{-1}$ and $\tilde{\boldsymbol{B}}$ are circulant. The class

of circulant matrices is stable under inversion, transposition, and matrix multiplication, so $\boldsymbol{S}^T\boldsymbol{S} = (\tilde{\boldsymbol{B}}\boldsymbol{B}^{-1})^T\tilde{\boldsymbol{B}}\boldsymbol{B}^{-1}$ is also circulant. Therefore eigenvectors of $\boldsymbol{S}^T\boldsymbol{S}$ are discrete Fourier vectors. This connection provides a rationale for comparing eigenvalues of $\boldsymbol{S}^T\boldsymbol{S}$ to the spectrum of $\mathscr{D}^{-2}$, whose eigenfunctions are Fourier modes. However, it is important to recognize that the comparison is imprecise because the interpolant eq. (3.9) represents these discrete Fourier eigenvectors as functions that differ from Fourier modes on $\mathbb{R}^2$.

Eigenvalues of $\boldsymbol{S}^T\boldsymbol{S}$ are plotted in fig. 3.2 as circles, and some examples of eigenfunctions of $\boldsymbol{S}^T\boldsymbol{S}$ are visualized beneath the plot for $k \in (1, 2, 25, 49)$. **Eigenfunctions of $\boldsymbol{S}^T\boldsymbol{S}$** are defined here as continuous interpolants of the matrix's eigenvectors found by the RBF interpolation scheme utilized in the smoother. This figure also shows a solid trace labelled "Fourier" that plots $(1 + \ell^2k^2)^{-2\beta}$, which is the spectrum of $\mathscr{D}^{-2}$ that corresponds to $\mathbb{R}^2$ Fourier modes. Since eigenvectors of $\boldsymbol{S}^T\boldsymbol{S}$ do not correspond to $\mathbb{R}^2$ Fourier modes, this trace of the Fourier spectrum is only a rough comparison rather than an analytical prediction that we are trying to match. The observed spectrum of $\mathbf{S}^T\mathbf{S}$ behaves as expected, with gradual smoothing of small scale features.

Recall from section 3.2 that this method has a drawback of attenuating large scales. That behavior is evident in the eigenvalues plotted in fig. 3.2, which are all less than 1. Eigenvalues less than 1 correspond to attenuation, so the smoother is attenuating even the largest scale (eigenmode index 0). This over-attenuation occurs because convolution with g attenuates every Fourier eigenmode of $\mathscr{D}$ except constant functions in $\mathbb{R}^d$. Since a finite Gaussian approximation can never fully describe a nonzero spatial constant, even the largest-scale function in the space of possible RBF interpolants will be attenuated by our smoother. The largest-scale eigenfunctions in this example are thin in the direction transverse to the circle, causing those modes to be smoothed more than continuous Fourier modes in $\mathbb{R}^2$ with the same wavenumber (i.e. Fourier modes with **planar** length scale equal to the **circumferential** length scale of $\boldsymbol{S}^T\boldsymbol{S}$ eigenfunctions)

For a similar reason that large scales are attenuated too much, the smoother does not suppress the smallest-scale eigenmodes as much as $\mathscr{G}$ would suppress a true $\mathbb{R}^2$ Fourier mode of the same wavenumber. This is because the RBF interpolants of the most highly-oscillatory eigenvectors in

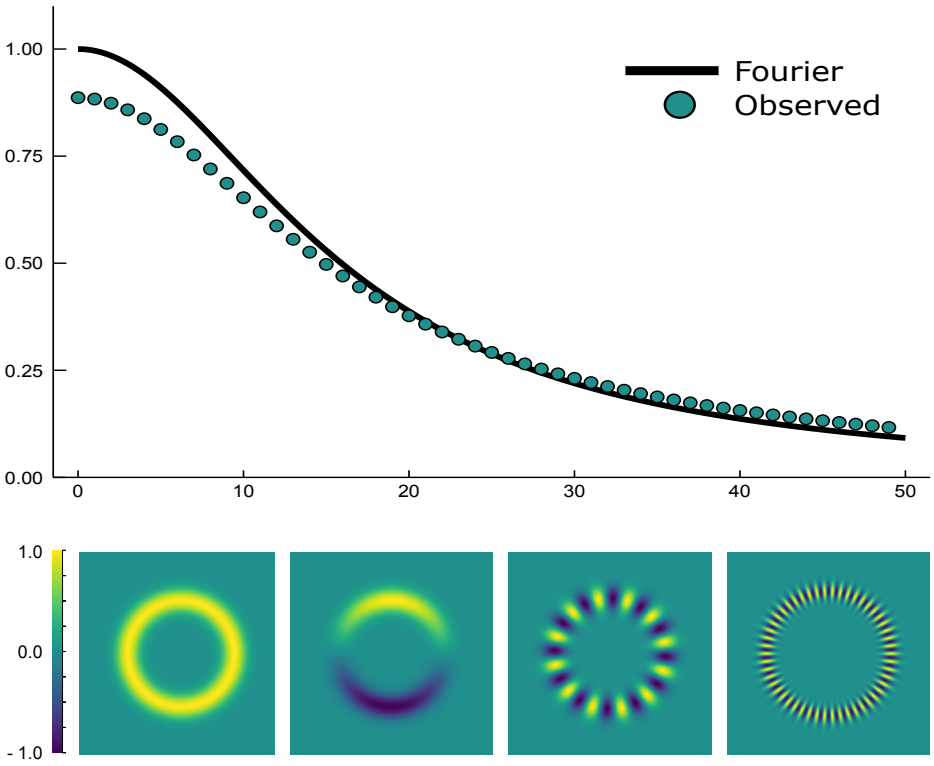


Figure 3.2: Top: points indicate eigenvalues of the corvariance matrix $\mathbf{S}^T\mathbf{S}$ where $\mathbf{S}$ comprises interpolation by Gaussian radial basis functions of standard deviation $\xi^{1/2} = 2.5$, followed by convolution with a Gaussian approximation of the Green's function for the bound-state fractional Helmholtz kernel of $\mathscr{D} = (1 - \Delta)^\beta$ with $\ell = 1.0$ and $\beta = 1/2$, acting on 100 equally spaced points around the origin in $\mathbb{R}^2$ with unit nearest-neighbor distance. The solid trace shows the spectrum $(1 + k^2)^{-2\beta}$ of $\mathscr{D}^{-2}$, eigenfunctions of which are Fourier modes; this serves to highlight the similarity between the spectra of $\boldsymbol{S}^T\boldsymbol{S}$ and of $\mathscr{D}^{-2}$, but is not an analytical solution to match since interpolating the eigenvectors of $\boldsymbol{S}^T\boldsymbol{S}$ does not produce Fourier modes in the plane. Bottom: some example eigenfunctions, defined as interpolants given by eq. (3.9) of the eigenvectors of $\boldsymbol{S}^T\boldsymbol{S}$, for $k \in \{1, 2, 25, 49\}$. Duplicate eigenpairs that arise due to symmetry are suppressed in this figure.

this example have more large-scale content than $\mathbb{R}^2$ Fourier modes with the same length scale.

Over-attenuation can be mitigated, so that the largest scales are closer to unity, by rescaling the operator by replacing $\mathbf{S}^T\mathbf{S} \mapsto \mathbf{S}^T\mathbf{S}/\|\mathbf{1}^T\mathbf{S}^T\mathbf{S}\mathbf{1}\|$, where $\mathbf{1}$ is a unit-norm vector with all entries identical. The eigenvectors of $\mathbf{S}^T\mathbf{S}$ are usually **not** discrete Fourier vectors like they are in this symmetric example, so the largest-scale eigenvector is not necessarily $\mathbf{1}$. Therefore this mitigation technique is only a heuristic, which derives from the idea that an input with identical entries contains little small-scale information.

Choosing the RBF standard deviation parameter $\xi^{1/2}$ is not to be taken lightly. We recommend choosing it to be roughly on the order of the nearest-neighbor distance between measurements. A value too small prevents the RBF interpolation step from resolving gradual transitions from location to location, causing the interpolant to appear as a rugged set of "spikes" that are overly suppressed by the convolution step on account of their inappropriately small scale. Choosing an interpolation kernel that is too large, however, can cause numerical problems related to ill-conditioning of the linear system we must solve to arrive at RBF coefficients. Choosing $\xi^{1/2}$ to be as large as possible, while avoiding insurmountable instability due to ill-conditioning, is considered a best practice in RBF literature (Fornberg and Flyer, 2015).

Choosing the parameters ℓ and β in eq. (3.4) is a problem-specific process that depends on the scales of interest for the data assimilation task at hand, as well as the density of measurements. These considerations are explored for ℓ in Robinson et al. (2018), which found that smoothing too aggressively is detrimental when observation locations are sparse but also that aggressive smoothing can still be beneficial when observations are very dense.

One can expect optimal choice of the shape parameter β also to depend on the dynamics and observing system. A large value of β corresponds to more aggressive smoothing of small scales but also to flatter, more permissive response for scales less than $\ell/2\pi\sqrt{2^{1/\beta}-1}$. Therefore how β affects the tractability of a filtering problem using SIR will depend not only on density of observations and what scales are of interest but also the particular covariance spectrum required to resolve the dynamics of the physical system under observation.

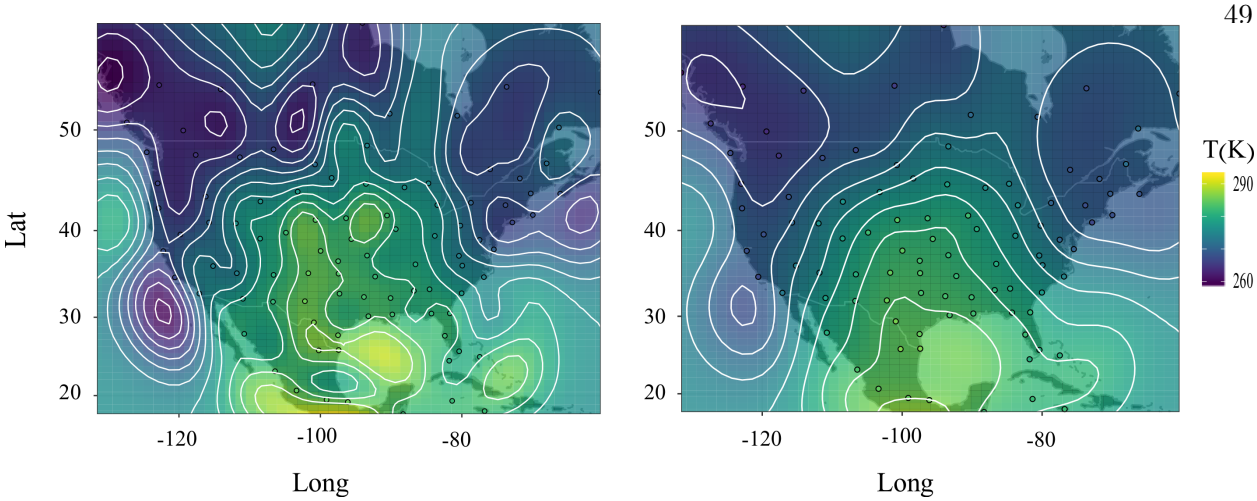


Figure 3.3: Left: Temperature fields interpolated from radiosonde data measured on May 15, 2017 at a pressure level of 70kPa. Right: the result of applying our smoother with parameters $\xi^{1/2} = 5°$, $\beta = 1$, and $\ell = 4°$. The shape of the North America is underlaid to give a sense of scale, and small circles indicate measurement locations.

3.4 Example 2: radiosonde data

To demonstrate the behavior of our smoothing algorithm on scattered data and its impact on SIR weights, we make use of data from the U.S. National Center for Atmospheric Research (NCAR) Convection Allowing Ensemble (Schwartz et al., 2015, 2019). The NCAR ensemble produced real-time 48 hour forecasts over the conterminous United States (CONUS) from April 7, 2015 to December 30, 2017. The ensemble forecasting system consisted of two components: an 80 member ensemble assimilation system operating at 15 km resolution and a 10 member ensemble forecast system operating at 3 km resolution. We make use of the 80 member ensemble data. The assimilation system used the Advanced Research version of the Weather Research and Forecasting (WRF) model; observations were assimilated in a 6 hour cycle via the Ensemble Adjustment Kalman Filter (Anderson, 2001) implemented in the Data Assimilation Research Testbed software suite (Anderson et al., 2009). Every assimilation cycle processed between 66,000 and 70,000 observations from a variety of sources including radiosondes, aircraft measurements, satellite wind measurements, and Global Positioning System radio occultation data, among others. Further details are provided in Schwartz et al. (2015).

To verify that our smoothing algorithm performs as expected on scattered data, we apply it to radiosonde temperature measurements at a single pressure level. Every 12 hours, i.e. every other assimilation window, there are between 90 and 97 radiosonde measurements scattered across North and Central America and the Caribbean available at various pressure levels. An example of the locations of these observations at a pressure level of 70 kPa on May 15, 2017 is shown in Fig. 3.3. The left panel shows the locations of the measurements along with an interpolated temperature field obtained using Gaussian RBFs with standard deviation $\xi^{1/2} = 5°$. The right panel shows the result of applying our smoothing algorithm with smoothing exponent $\beta = 1/2$ and smoothing length scale $\ell = 4°$. Figure 3.3 provides visual evidence that our algorithm indeed smooths scattered data. Interpolating the raw data would leave strange regions of approximately zero Kelvins in the interpolants depicted. So for the purpose of visualization, we subtract the mean before applying the smoother, and then add the mean back to the smoothed data.

We next verify that the algorithm has a controllable degree of smoothing with the desired effect on SIR weights, viz. that the effective sample size increases as the smoothing length scale ℓ increases. To that end we use the 80 member ensemble forecast for temperature at the locations of the radiosonde temperature observations, assume that the forecast weights are all equal to 1/80, and update the weights based on mismatch to the observations using the standard SIR update formula.

To be precise, let $\mathbf{y}$ be the vector of radiosonde temperature observations at a given time, let $\mathbf{H}\mathbf{x}^{(i)}$ be the vector of forecast temperatures at the same time and locations for ensemble member i, and let $\mathbf{R}_0^{1/2}$ be a diagonal matrix whose diagonal contains the standard deviations of the observation errors. The un-normalized weight for the i^{th} ensemble member is

$$\overline{w}_i = \exp\left\{-\frac{1}{2\sigma}\left(\mathbf{y} - \mathbf{H}\mathbf{x}^{(i)}\right)^T \mathbf{R}_0^{-1/2}\mathbf{S}^T\mathbf{S}\mathbf{R}_0^{-1/2}\left(\mathbf{y} - \mathbf{H}\mathbf{x}^{(i)}\right)\right\} \tag{3.30}$$

where $\mathbf{S}$ is the matrix corresponding to the smoothing operator and $\sigma = \|\mathbf{1}^T\boldsymbol{S}^T\boldsymbol{S}\mathbf{1}\|$ is a rescaling factor with $\mathbf{1}$ a unit vector of identical entries. The normalized weights are

$$w_i = \frac{\tilde{w}_i}{\sum_{j=1}^N \tilde{w}_j} \tag{3.31}$$

and the effective sample size (ESS) is

$$\text{ESS} = \frac{1}{\sum_{i=1}^{N} w_i^2}. \tag{3.32}$$

Figure 3.4 shows violin plots of the ESS computed twice daily over the entire month of May 2017 using radiosonde temperature data at a pressure level of 50 kPa. The standard particle filter, without smoothing, exhibits very poor performance with ESS only rarely rising much beyond 1.

As the smoothing length scale ℓ increases from 0, the distribution of ESS also increases. With $\ell = 4°$, the median ESS is approximately 3 and ESS occasionally rises beyond 5. We emphasize that these values of ESS are still quite small for an 80 member ensemble, but the goal here has not been to demonstrate the performance of the particle filter per se, but of the smoothing algorithm for scattered data. That said, even a tiny increase of the ESS beyond its minimum possible value of 1 is promising because it offers hope that uncertainty quantification will improve faster upon increasing the ensemble size.

3.5 Discussion

The first step of our smoother involves solving a dense linear system eq. (3.9) for interpolation weights $\boldsymbol{b}$. A naive approach to doing so would require $\mathcal{O}(N_y^2)$ storage and $\mathcal{O}(N_y^3)$ operations. This would be highly undesirable, especially if the interpolation matrix $\boldsymbol{B}$ changes between assimilation cycles due to changing observation locations. Happily, there exist algorithms to solve a Gaussian RBF problem much faster. These notably include PetRBF, which is based on GMRES iteration with a restricted additive Schwarz method preconditioner. PetRBF requires $\mathcal{O}(N_y)$ storage and $\mathcal{O}(N_y)$ operations in arbitrary dimension d, and it scalable to many cores as implemented in PETSc (Yokota et al., 2010).

Another potential bottleneck is evaluating the sum of $N_y M$ Gaussians at N_y target locations, with $M = M_- + M_+ + 1$ terms in the approximation of g. A direct approach to evaluating this sum, as is written in Algorithm 2.2, has time complexity $\mathcal{O}(N_y^2 M)$. This Gaussian sum approximation can be reduced to $\mathcal{O}(N_y + N_y M)$ with the Fast Gauss Transform (FGT) (Greengard and Strain,

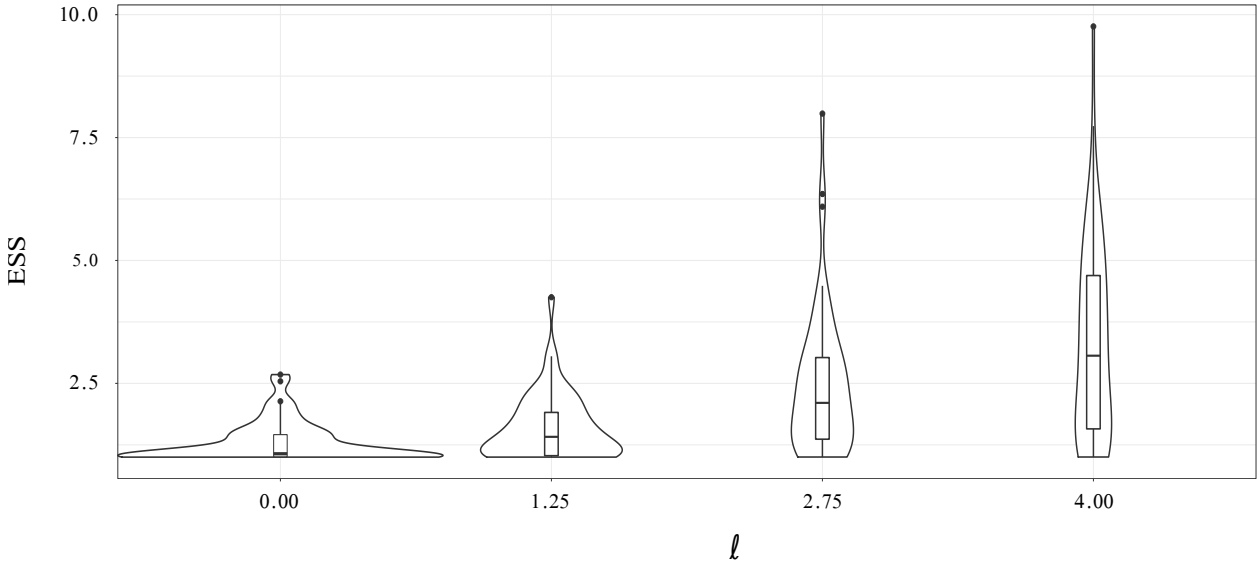


Figure 3.4: Distributions of effective sample sizes, for different values of smoothing length scale ℓ, observed for the posterior ensembles on radiosonde temperature. The posterior ensembles are obtained by performing an importance sampling update on the WRF forecast as a prior ensemble, using actual radiosonde temperature measurements at an atmospheric pressure level of 50 kPa. The distributions in this plot depict ESS values for SIR weights computed twice daily over the entire month of May 2017.

1991). The FGT has exponential time complexity in location dimensionality d, so it often runs slower than direct evaluation when d is greater than 2. The Improved Fast Gauss Transform (IFGT) exists to eliminate that exponential scaling that hinders the original FGT for large d (Yang et al., 2003). The IFGT can be challenging to use in practice, but there exist approaches to assist in automatic tuning such as Morariu et al. (2009).

Our algorithm uses an interpolation basis function that remains fixed throughout the domain. This may be problematic when the density of observation locations is highly heterogenous, since an RBF standard deviation ξ large enough to resolve smooth features in a sparsely-sampled region may be large enough that it causes numerical problems in densely-sampled regions. Those numerical problems may arise from ill-conditioning of $\boldsymbol{B}$ or from insufficient data locality expected of some divide-and-conquer solvers like PetRBF. There is some extant literature on the use of nonuniform RBF width parameters to address this situation (Fornberg and Zuev, 2007), so that the size of the basis function can adapt to the density of observations. Using adaptive width involves interpolation with a basis $\{\psi_i\}$ that is allowed to vary with i. Adaptive width can be incorporated into our smoother just by modifying eq. (3.9) and eq. (3.24) to let ξ vary with i. Using nonuniform width parameters no longer comes with guaranteed nonsingularity, but Fornberg and Zuev (2007) suggests that singularity is more of an exception than a rule.

Unfortunately, many fast solvers for the RBF problem are incompatible with basis functions that vary by location. One possibility to reduce the cost of solving for interpolation weights in this case is to choose compactly-supported basis functions ψ_i so that $\boldsymbol{B}$ is sparse. We are unaware of any compactly-supported radial basis functions with positive Fourier transforms that are simple to convolve with a Gaussian, particularly for arbitrary d. But performing the interpolation in terms of compactly-supported bases ψ_i can be made compatible with the rest of our smoothing method, simply by approximating each ψ_i with a sum of Gaussians. The resulting Gaussian approximation of the data will not be an interpolant, but careful construction can make it accurate. Therefore theorem 1 does not apply, but we can still expect this substitution to yield a good approximation of the convolution acting on the original interpolant.

It is similarly possible to choose a different convolution kernel g to approximate with a sum of Gaussians. This idea can be used to implement a smoother of the form presented here with a wider variety of characteristics, such as a non-monotonic response in length scale. If the Guassian approximation kernel possesses a positive Fourier transform, and the RBF interpolation employs a uniform basis function, then theorem 1 still applies to guarantee that $\mathbf{S}^T\mathbf{S}$ is a valid covariance matrix.

To reduce the M prefactor in the convolution step, we can apply a reduction algorithm based on Prony's method with the suboptimal approximation eq. (3.14) as a starting point (Beylkin and Monzón, 2010). Doing so yields an optimal multiresolution approximation of the integral kernel for given uniform relative error bounds, which may require substantially fewer terms to attain the same relative accuracy. This reduction method may be particularly helpful for different forms of $\mathscr{D}$ (ergo g) that do not yield such a rapidly-convergent approximation as eq. (3.21).

3.6 Conclusions

We have a described a method to smooth data measured at N_y locations that are arbitrarily scattered in $\mathbb{R}^d$, for arbitrary d, by applying a discrete approximation to the integral equation that inverts the fractional bound state Helmholtz operator $(1 - \ell^2\Delta)^\beta$. The degree of attenuation for different length scales can be tuned by adjusting the parameters $\ell > 0$ and $\beta > 0$; large scales are attenuated little, but length scales shorter than $\ell/(2\pi\sqrt{2^{1/\beta} - 1})$ are rapidly suppressed with a rate determined by β.

The discrete approximation results from a multiresolution Gaussian approximation to the differential operator's Green's function. This readily permits convolution with a sum of Gaussians that approximate the data; we take the sum of Gaussians approximation to be a radial basis function (RBF) interpolant with a Gaussian kernel. The smoother is shown to be a positive definite linear operator on $\mathbb{R}^{N_y}$ in a more general context where the interpolation basis and the convolution kernel have positive Fourier transforms.

Our smoother is developed with application to Sequential Importance Sampling with Resam-

pling (SIR) particle filters in mind. Smoothing observations with $\mathbf{S}$ before assimilating them as if they have uncorrelated errors is equivalent to assuming that the observation errors have covariance $(\mathbf{S}^T\mathbf{S})^{-1}$, which gives observation errors the correlation structure of a stationary generalized Gaussian random field. Relative to an uncorrelated model, an observation error model of this type decreases the number of ensemble members required to achieve good uncertainty quantification from SIR for spatially-extended dynamical systems (Robinson et al., 2018).

Spectral properties of our smoother are examined with an example shown in section 3.3. This example provides evidence that the algorithm operates as expected: attenuation gradually increases in wavenumber, roughly approximating the differential operator's inverse spectrum, with a caveat that our smoother attenuates even the largest scale. It is particularly important to preserve large scales for the application to SIR; for that we propose dividing $\boldsymbol{S}^T\boldsymbol{S}$ by $\mathbf{1}^T\boldsymbol{S}^T\boldsymbol{S}\mathbf{1}$, where $\mathbf{1}$ is a unit vector with all entries identical.

Section 3.4 demonstrates that this smoother has the desired effect of helping balance SIR weights in an example with real meteorological data, which improves uncertainty quantification by reducing the tendency of SIR to produce underdispersed posterior distributions in high dimensions. This example is chosen to be provocative of potential future applications to geophysical fluids, but it is worth characterizing traits of applications that would be more appropriate. The extratropical temperature field in section 3.4 probably features little dynamical nonlinearity at large scales, and its measurements are linear and Gaussian, so this corpus of data is an excellent candidate for assimilation with any one of the many variants of the Ensemble Kalman Filter. A more appropriate application of smoothed SIR would feature substantially non-Gaussian behavior at large scales. That can arise due to nonlinear dynamics of large scales or due to large dispersion of a non-negative state variable relative to its mean, or due to a nonlinear observation operator inducing a non-Gaussian posterior distribution. Moist convective systems, for example, have nonlinear dynamics and substantially skewed sign-definite variables. Examples of nonlinear observation operators that could be similar motivation for SIR include satellite radiance and precipitation measurements. Any of these features could provide motivation for accepting the computational challenge of SIR

in exchange for provable consistency.

A naive implementation of Algorithm 2.1 requires $\mathcal{O}(N_y^2)$ memory and $\mathcal{O}(N_y^3)$ operations to solve for interpolant weights. However section 3.5 describes how specialized kernel matrix solvers and the Fast Gauss Transform can reduce the asymptotic complexity of our algorithm to $\mathcal{O}(N_y)$.

Chapter 4

A smoothed-observation SIR-ESRF hybrid filter for data assimilation scenarios with 'medium' non-Gaussianity

4.1 Introduction

Data assimilation of high-dimensional dynamical systems routinely falls to ensemble approximations of the Square Root Filter, including various types of Ensemble Kalman Filter (EnKF) and the Ensemble Square Root Filter (ESRF), despite known inconsistency when the dynamics or the observing system yield non-Gaussian probability distributions (Katzfuss et al., 2016). In contrast, Sequential Importance Sampling with Resampling (SIR a.k.a. Particle Filtering) is proven to weakly converge to the correct posterior in the large-ensemble limit — with remarkably mild constraints on the dynamics, prior, and observing system (Crisan and Doucet, 2002, Gordon et al., 1993). This flexibility makes SIR superficially attractive for nonlinear fluids problems such as atmospheric and oceanographic forecast. Unfortunately, however, SIR suffers a severe curse of dimensionality that has prevented its practical application to state estimation and forecast of fluids.

Previous work (Robinson et al., 2018) demonstrated that smoothing innovations can improve how SIR scales in the number of observations. That improvement obtains from the way SIR's effective dimension depends inversely on observation error variances. Supposing that smoothed data have uncorrelated errors is equivalent to supposing that observations at small scales have high error variance. Doing so reduces the effective dimensionality of the particle filtering problem, which reduces the incidence of collapse for a fixed ensemble size, thereby improving uncertainty quantification of large-scale physical features. Preferring accurate uncertainty quantification of large

scales, at the expense of accurately reproducing small scales, is informed by a practical distinction: small scales mix so rapidly that they are essentially unpredictable.

Although smoothing observations does substantially reduce the computational burden of SIR for assimilating data observed of spatially-extended high dimensional systems, is not a silver bullet. The number of dynamically important and forecastable large scales in a problem, observations of which must be preserved to avoid filter divergence, is often large enough to make the particle filter infeasible. Snyder et al. (2015) estimate that global numerical weather prediction, broken into 500 independent tiles over the globe, would require e^{5000} ensemble members. If we can reduce the dimensionality by a factor of 100 using smoothed observations, the exponent would decrease from 5000 to 50 — but an ensembe size of $e^{50} \approx 10^{22}$ is still far too large.

In the spirit of work by Frei, Künsch, et al., we seek to extend this work by interleaving smoothed SIR with ESRF (Chustagulprom et al., 2016, Frei and Künsch, 2013). This approach, termed 'bridging', aims to address the respective shortcomings of SIR and ensemble implementations of square root filters. At each assimilation cycle, part of the observational information is incorporated by means of an SIR step, and then the remaining observational information is incorporated with the ESRF. On one hand, the aim is to mitigate ESRF's bias by feeding it with a prior ensemble that is more Gaussian than the original forecast ensemble. On the other hand, the ESRF step moves ensemble members closer to observations in a manner that can mitigate SIR's tendency to become degenerate.

4.2 The hybrid algorithm

4.2.1 SSIR

Standard sequential importance resampling particle filters work as follows (Doucet et al., 2001, Gordon et al., 1993). Each ensemble member $\boldsymbol{x}_0^{(i)}$ (or 'particle') starts with equal weight $w_0^{(i)} = 1/N$, where N is the ensemble size. Each ensemble member is forecast until the next

assimilation cycle. At assimilation cycle j the weights are updated using the likelihood $L(\boldsymbol{x})$

$$w_j^{(i)} = w_{j-1}^{(i)} \frac{L\left(\boldsymbol{x}_j^{(i)}\right)}{Z_j} \tag{4.1}$$

where Z_j is a normalization constant to ensure that the weights sum to one. A resampling is then applied whereby particles with high weights are replicated and particles with low weights are eliminated. After resampling, all weights are equal. There are a variety of resampling algorithms; we here use systematic resampling (Kitagawa, 1996).

It is well known that the weights of a particle filter tend to collapse in high dimensions, i.e. a small number of particles receive a weight near one while all others receive a weight near zero. After resampling, only the high-weight particles are left. If an optimal-transport based alternative to resampling is used (Acevedo et al., 2017, Reich, 2013), then all particles are transported to a very small vicinity of the high-weight particles. In both cases, the posterior distribution is poorly estimated. The number of particles with a substantial portion of the weight can be approximated by the effective sample size

$$\text{ESS} = \frac{1}{\sum_{i=1}^{N} (w_j^{(i)})^2}. \tag{4.2}$$

The ESS takes values between 1 and N, with large ESS indicating approximately equal weights and small ESS indicating that the weights have collapsed.

The development of particle filters that avoid or reduce the incidence of collapse is an active area of research, with most methods focused on alternative methods of forecasting particles. These methods use observational data to guide the forecast (Ades and Van Leeuwen, 2015, 2013, Chorin et al., 2010, Chorin and Tu, 2009, 2012, Morzfeld et al., 2012, Pulido and van Leeuwen, 2019, van Leeuwen, 2010). The authors recently proposed an alternative that uses the same forecast as the standard particle filter, but imposes a generalized random field model of observation errors (Robinson et al., 2018). When the observation errors are Gaussian, the likelihood takes the form

$$L(\boldsymbol{x}) \propto \exp\left\{-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{H}(\boldsymbol{x}))^T \mathbf{R}^{-1} (\boldsymbol{y} - \boldsymbol{H}(\boldsymbol{x}))\right\}. \tag{4.3}$$

Here $\boldsymbol{y}$ is the observation vector, $\boldsymbol{H}$ is the observation (or 'forward') operator, and $\mathbf{R}$ is the

observation error covariance matrix. In the smoothed particle filter of (Robinson et al., 2018), the observation error covariance matrix $\mathbf{R}$ is replaced by a covariance matrix that has increasing variance at small spatial scales. In practice this is implemented by smoothing the innovations $\boldsymbol{y}-\boldsymbol{H}(\boldsymbol{x})$. The authors recently developed a fast algorithm for smoothing scattered data in arbitrary dimensions for this purpose (Robinson and Grooms, 2019).

In the numerical experiments presented here, the spatial domain is a circle and Fourier methods are used to apply the smoothing. The true observation error covariance matrix is $\gamma^2\mathbf{I}$. In the smoothed particle filter this is replaced by $\gamma^2\left(\mathbf{S}^T\mathbf{S}\right)^{-1}$, where the matrix $\mathbf{S}$ corresponds to an operator that attenuates the Fourier coefficients according to the spectrum

$$\frac{1}{(1+(\ell k)^2)^\beta}, \tag{4.4}$$

where β and ℓ are tunable parameters and k is the Fourier wavenumber. More general smoothing spectra are trivial to implement in our experiments, but the above smoothing corresponds to the spectrum of the fast smoother for scattered data developed in (Robinson and Grooms, 2019).

Replacing the true likelihood (or a more common approximation of the true likelihood, such as an uncorrelated Gaussian likelihood) by a likelihood that is associated with spatial smoothing means that the particle filter is approximating a distribution other than the true Bayesian posterior. The effect of this smoothing is to make the likelihood less informative at small scales, so that the posterior reverts to the prior at small scales. At large scales the smoothing likelihood closely resembles the true likelihood, so the posterior obtaining from this model is close to the true posterior for large scales. Smoothing reduces the effective dimension of the problem by confining the action of the importance sampling to large scales. This has the effect of reducing the minimum ensemble size needed to avoid collapse, and can improve uncertainty quantification of large scales for a fixed ensemble size.

4.2.2 ESRF

There are many ensemble Kalman filters, any of which could be hybridized with the smoothed particle filter. We focus here on an ensemble square root filter (ESRF) developed in (Whitaker and Hamill, 2002) for sequential assimilation of observations possessing uncorrelated errors. At a single assimilation cycle the ensemble is denoted $\{\boldsymbol{x}^{(i)}\}_{i=1}^{N}$. The ensemble mean is denoted $\bar{\boldsymbol{x}}$, and the scaled ensemble perturbation matrix is denoted

$$\mathbf{A} = \frac{1}{\sqrt{N-1}}\left[\boldsymbol{x}^{(1)} - \bar{\boldsymbol{x}}, \ldots, \boldsymbol{x}^{(N)} - \bar{\boldsymbol{x}}\right]. \tag{4.5}$$

The ensemble covariance matrix is thus $\mathbf{A}\mathbf{A}^T$. Covariance inflation is applied by replacing $\mathbf{A}$ with $(1+r)\mathbf{A}$, where r is a tunable inflation factor.

Observations are linear, and for a single scalar observation the observations take the form

$$y = \mathbf{H}\boldsymbol{x} + \epsilon. \tag{4.6}$$

Here the observation error ϵ is a zero-mean normal with variance γ^2. It is convenient to define the row vector $\mathbf{V} = \mathbf{H}\mathbf{A}$. With this notation, the ESRF from Whitaker and Hamill (2002) corresponds to the following update of the ensemble mean

$$\bar{\boldsymbol{x}}_+ = \bar{\boldsymbol{x}} + \frac{(y - \mathbf{H}\bar{\boldsymbol{x}})}{\sigma^2 + \gamma^2}\mathbf{A}\mathbf{V}^T \tag{4.7}$$

and the following update of the scaled ensemble perturbation matrix

$$\mathbf{A}_+ = \mathbf{A} - b\mathbf{A}\mathbf{V}^T\mathbf{V}, \tag{4.8}$$

$$b = \frac{1}{\sigma^2 + \gamma^2 + \gamma\sqrt{\sigma^2 + \gamma^2}} \tag{4.9}$$

where $\sigma^2 = \mathbf{V}\mathbf{V}^T$.

Localization is applied by multiplying the increments elementwise by a localization vector $\boldsymbol{\rho}$. The elements of $\boldsymbol{\rho}$ are $e^{-(d/L)^2/2}$, where d is the distance from $\boldsymbol{x}_i$ to y and L is a tunable localization radius. This amounts to updating eq. (4.7) and eq. (4.8) to

$$\bar{\boldsymbol{x}}_+ = \bar{\boldsymbol{x}} + \frac{(y - \mathbf{H}\bar{\boldsymbol{x}})}{\sigma^2 + \gamma^2}\boldsymbol{\rho} \circ \mathbf{A}\mathbf{V}^T \tag{4.10}$$

and

$$\mathbf{A}_+ = \mathbf{A} - b\left(\boldsymbol{\rho} \circ \mathbf{A}\mathbf{V}^T\right)\mathbf{V} \tag{4.11}$$

where $\circ$ denotes an elementwise product.

Evensen was the first to suggest resampling the posterior within the context of an ensemble square root filter by multiplying $\mathbf{A}$ from the right by a random orthogonal matrix after assimilating but before reconstructing the ensemble (Evensen, 2004). Since the ensemble covariance matrix is $\mathbf{A}\mathbf{A}^T$, this kind of resampling does not change the ensemble covariance matrix. Sakov & Oke pointed out that the random orthogonal matrix should have $\mathbf{1}$ as an eigenvector in order for the resampling to preserve the ensemble mean (Sakov and Oke, 2008). To that end, we construct a new scaled ensemble perturbation matrix $\mathbf{A}$ by multiplying $\mathbf{A}$ from the right by a random orthogonal matrix $\mathbf{Q}$ that has $\mathbf{1}$ as an eigenvector. The matrix $\mathbf{Q}$ is constructed as follows (Sakov and Oke, 2008)

$$\mathbf{Q} = \mathbf{U}\left[\begin{array}{c|c} 1 & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{P} \end{array}\right]\mathbf{U}^T. \tag{4.12}$$

The matrix $\mathbf{U}$ is a time-independent orthogonal matrix whose first column is proportional to $\mathbf{1}$, formed by analytically solving the Gram-Schmidt orthogonalization of the matrix whose nonzero entries are ones in the first column and the diagonal. The matrix $\mathbf{P}$ is a random orthogonal matrix of size $N-1 \times N-1$. With a large ensemble size it can become costly to sample a new $\mathbf{P}$ at each assimilation cycle. In principle the random matrix $\mathbf{Q}$ could be constructed once and used repeatedly, but in our numerical experiments $\mathbf{P}$ is resampled at each assimilation cycle.

Using this method, a single assimilation cycle proceeds as follows

- Form the ensemble mean $\bar{\boldsymbol{x}}$ and scaled ensemble perturbation matrix $\mathbf{A}$.
- Inflate the scaled ensemble perturbation matrix: $\mathbf{A} \leftarrow (1+r)\mathbf{A}$
- For each observation, update $\bar{\boldsymbol{x}}$ and $\mathbf{A}$ using eq. (4.10) and eq. (4.11).
- Resample the posterior ensemble by replacing $\mathbf{A}$ with $\mathbf{A}\mathbf{Q}$.

- Reconstitute the ensemble according to $\boldsymbol{x}^{(i)} = \bar{\boldsymbol{x}} + \sqrt{N-1}\mathbf{A}_i$ where $\mathbf{A}_i$ is the i^{th} column of $\mathbf{A}$.

4.2.3 SSIR-ESRF hybrid

The smoothed SIR/ensemble square root filter (SSIR-ESRF) hybrid developed here adheres closely to the bridging method of Frei and Künsch (Frei and Künsch, 2013). The likelihood $L(\boldsymbol{x})$ is factored into a product $(L(\boldsymbol{x}))^{1-\alpha} \cdot (L(\boldsymbol{x}))^{\alpha}$ where $\alpha \in [0, 1]$ is the 'splitting factor'. The hybrid proceeds by having the smoothed SIR particle filter assimilate using the likelihood $(L(\boldsymbol{x}))^{1-\alpha}$, followed by an ESRF assimilation using the likelihood $(L(\boldsymbol{x}))^{\alpha}$. In principle, the methods can be applied in either order, but the method is intended for situations where the prior is non-Gaussian but the posterior is nearly Gaussian ('medium' nonlinearity according to (Morzfeld and Hodyss, 2019)). In such cases the intermediate posterior produced after the first assimilation with the particle filter should be closer to Gaussian than the prior. The ESRF subsequently performs an assimilation on a problem that more closely conforms to its Gaussian assumptions.

Following Frei and Künsch (2013) we choose the splitting factor α to ensure that the effective sample size is within 10 of a tunable theshold. This is achieved with a rootfinding method. A small α generally leads to large ESS. If $\alpha = 1$, then the hybrid reverts to a pure ESRF because all the particle filter weights become equal.

The resampling step of the SSIR particle filter leads to a degeneracy where there are multiple copies of some ensemble members. In our numerical experiments we use a deterministic system of ordinary differential equations, so the dynamics do not break the degeneracy. Resampling could be replaced by an optimal transport method (Acevedo et al., 2017, Reich, 2013) that does not lead to degeneracy. Alternatively, degeneracy could be broken by using a perturbed-observation EnKF instead of an ESRF, but we opt instead to follow the ESRF assimilation with a mean-preserving random rotation that resamples the ensemble within the Gaussian posterior, as described in the previous section.

4.3 Numerical experiment

4.3.1 A two-scale Lorenz-'96 Model

The experiments make use of a model inspired by the Lorenz-'96 model (Lorenz, 1996) and developed in (Grooms and Lee, 2015). The standard two-scale (or 'two-layer') Lorenz-'96 model includes two sets of variables, X_k and $Y_{j,k}$. There are fewer X_k variables, and they evolve more slowly than the $Y_{j,k}$ variables, so the X_k variables are typically viewed as 'large-scale' while the $Y_{j,k}$ variables are viewed as 'small-scale.' The difficulty with this model is that a quantity like temperature or velocity contains both large and small scales, and is not conveniently presented to the practitioner in a form partitioned by spatial scale. Grooms and Lee (2015) developed a model inspired by the Lorenz-'96 models that has a single set of variables x_i that have distinct large-scale and small-scale parts. The model is governed by a system of ordinary differential equations of the form

$$\dot{\boldsymbol{x}} = h\boldsymbol{N}_S(\boldsymbol{x}) + J\mathbf{T}^T\boldsymbol{N}_L(\mathbf{T}\boldsymbol{x}) - \boldsymbol{x} + F\mathbf{1} \tag{4.13}$$

where $h, F \in \mathbb{R}$, $J \in \mathbb{N}$, $\mathbf{1}$ is a vector of ones, and

$$(\boldsymbol{N}_S(\boldsymbol{x}))_i = -x_{i+1}(x_{i+2} - x_{i-1}) \tag{4.14}$$

$$(\boldsymbol{N}_L(\boldsymbol{X}))_k = -X_{k-1}(X_{k-2} - X_{k+1}). \tag{4.15}$$

The number of state variables in $\boldsymbol{x}$ is $41J$; here $J = 128$ for a total system dimension of 5248. As in the Lorenz-'96 model, the indices extend periodically. The matrix $\mathbf{T}$ projects onto the 41 largest-scale discrete Fourier modes and then evaluates that projection at 41 equally-spaced points on the grid of state variables. The matrix $J\mathbf{T}^T$ interpolates a vector of length 41 back to the full dimension of $\boldsymbol{x}$.

The large-scale part of the model dynamics is obtained by applying $\mathbf{T}$ to $\boldsymbol{x}$. The result is identical to large-scale dynamics of the standard Lorenz-'96 model. Large scales are coupled to small scales via the term $h\mathbf{T}\boldsymbol{N}_S(\boldsymbol{x})$. While the standard Lorenz-'96 model has 40 large-scale variables, Grooms and Lee (2015) used 41 variables so that the 20^{th} Fourier mode is not split between large

and small scales. At small scales, the dynamics are same as those of original Lorenz-'96 model but with the direction of indexing reversed.

The experiments presented here use $h = 0.38$ and $F = 8$. With these parameters the large-scale dynamics are very similar to the standard Lorenz-'96 model, with fairly weak coupling to the small scales. The exception is when the large-scale Lorenz-'96 component reaches large values (e.g. amplitudes ≥ 10). This occurrence excites a fast small-scale instability, causing the small scales also to reach similarly large amplitudes that feed back onto the large-scale dynamics. A Hovmöller plot of these dynamics appears in fig. 4.1.

4.3.2 Data assimilation system configuration

Reference solutions are generated by drawing initial conditions from uncorrelated standard normal random variables and propagating them by 9.0 time units by numerical intergration of the dynamical model, at which point a statistically-steady state is reached. Once that statistically steady state is reached, 16 reference solutions are produced and their states are stored at 1500 time intervals separated by 1.2 time units. In the usual interpretation of the standard Lorenz-'96 model, this time interval corresponds to 6 days. That is quite long. At shorter time intervals the model exhibits only mild nonlinearity (Morzfeld and Hodyss, 2019), where the forecast distribution is still very nearly Gaussian even though the dynamics are nonlinear. At 6 days the forecast distributions are certifiably non-Gaussian, as shown in fig. 4.2. This figure was produced by projecting a forecast ensemble of 1200 members onto the three leading singular vectors of the ensemble's empirical covariance matrix. The forecast distribution is dramatically non-Gaussian within this subspace — therefore the EnKF assumption of a Gaussian prior is invalid.

Our hybrid is intended for situations with medium nonlinearity, where the prior is not Gaussian but the posterior is nearly Gaussian (Morzfeld and Hodyss, 2019). To achieve an approximately Gaussian posterior in the face of a non-Gaussian prior requires a large number of sufficiently-accurate observations. Reference observation timeseries are generated from the 16 reference solutions by taking observations at every fourth grid point, with observation error variance $1/2$. This

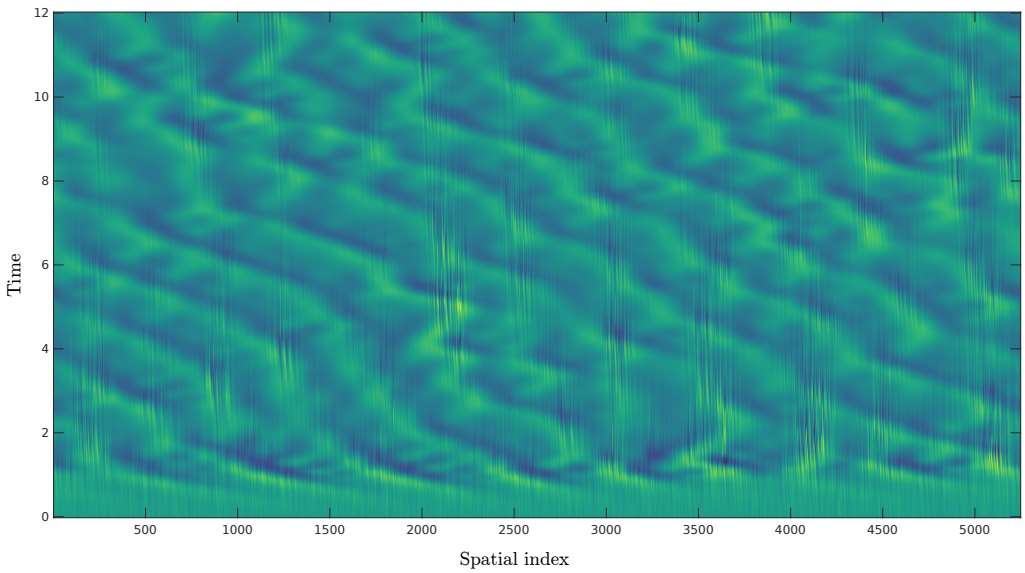


Figure 4.1: Hovmöller diagram of modified Lorenz-'96 dynamics starting from uncorrelated initial data drawn from a standard normal distribution. The ordinate is time, the abscissa is spatial index, and color represents the dynamical state as a function of those space and time variables.

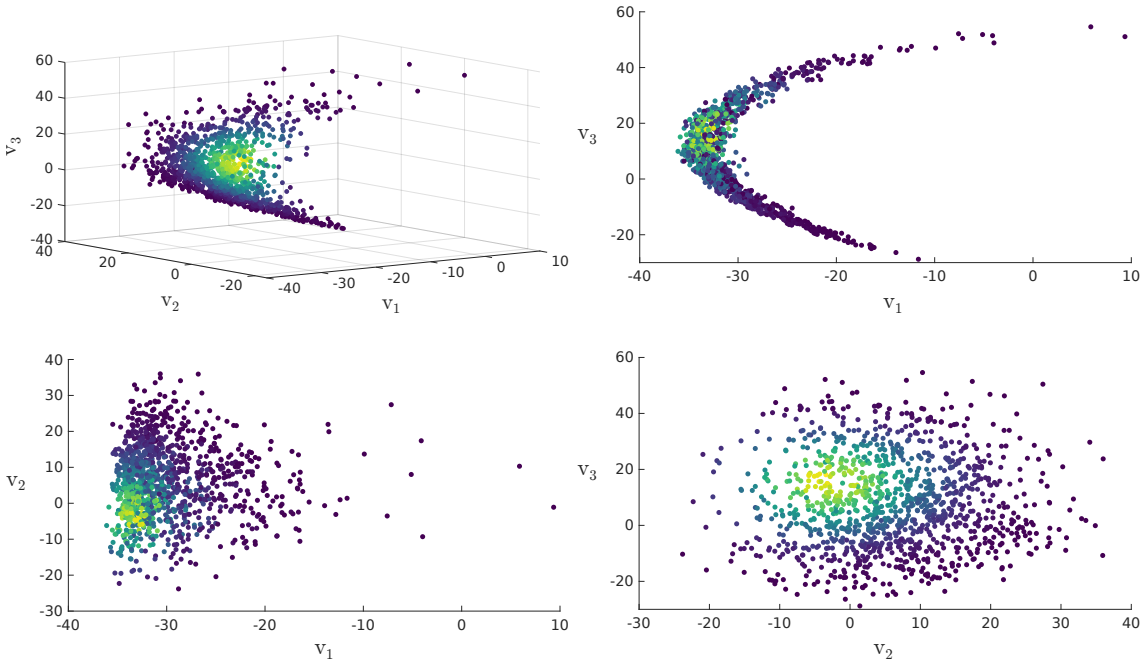


Figure 4.2: (a) An example forecast ensemble with $N_e = 1200$ projected onto three leading singular vectors of the ensemble's empirical covariance matrix. Points are shaded according to the importance sampling weights using a likelihood splitting factor α that produces an effective sample size of 640.5. (b-d) Two-dimensional projections of the same ensemble, identically shaded, provide detail of the marginal distributions on the three principal components that are jointly visualized in (a).

density and accuracy of observations is sufficient to produce a nearly-Gaussian posterior without rendering the data assimilation procedure superfluous. (If the observations were dense enough and accurate enough, then the filter would add essentially no information to the observations.)

Ensemble members are initialized by propagating a sample from the uncorrelated multivariate standard normal distribution by 9.0 time units to arrive at an ensemble of substantially disparate states near the dynamic's attractor. Because this initial forecast ensemble is fairly uninformative of the true state, there is a transient in filter performance while the filter burns in, approaching its asymptotic optimal performance. The results of the first 100 assimilation cycles are ignored in computations of filter performance statistics, so that the results presented are reflective of the statistical steady state of the filter. The data assimilation system was run for 1500 cycles, i.e. nearly 25 years, for each trial in the experiment.

4.3.3 Parameter Optimization

The ESRF used here has two primary tunable parameters: the inflation factor r and the localization radius L. The SSIR-ESRF hybrid filter has an additional tunable parameter, the splitting factor α. Since the system is configured to choose the splitting factor that leads to a pre-specified effective sample size (ESS), that pre-specified ESS threshold should be viewed as the third tunable parameter for the adaptive-α hybrid we investigate in this article. The smoothed particle filter also has tunable parameters ℓ and β related to the smoothing, but these should not be viewed as a primary means of optimizing performance. We expect the hybrid to outperform the pure ESRF using only reasonable smoothing parameters chosen a priori on the basis of approximate considerations about the dynamics and observing system. The demarcation between large and small scales occurs at Fourier wavenumber 20 for the Lorenz '96 variant considered here as a test problem, so the smoothing is chosen to have a Fourier spectrum

$$\frac{1}{\left(1+\left(\frac{k}{20}\right)^2\right)^2}.$$

To help substantiate a comparison between our SSIR-ESRF-MPRR hybrid approach against

the classic ESRF-MPRR filter, we independently tuned the respective filter parameters. This began by generating parameter configurations, described hereafter as 'arms', from a Sobol sequence of low-discrepancy quasirandom numbers (Owen, 1998) in a bounding box that we chose as a search space. The term 'arm' comes from the literature on multi-armed bandits and denotes a particular configuration to be tested.

For each arm, we ran the filter on at least four of the precomputed timeseries and their paired initialization ensembles. For each assimilation cycle we computed the resulting root mean squared error (RMSE) and continuous ranked probability score (Gneiting and Raftery, 2007, Hersbach, 2000, CRPS) for both the forecast and analysis. RMSE is a scalar quantity at each timestep, but CRPS was computed for each state variable at each timestep. We then aggregated these quantities by computing a mean over all state variables, timesteps, and assimilation trials — excluding the first 100 timesteps to allow for filter burn-in.

We elected to optimize for mean analysis CRPS because it quantifies the accuracy of the entire distributional estimate, whereas RMSE only describes accuracy of the ensemble mean point estimate. The ensemble spread (Fortin et al., 2014) would also provide an estimate of the distributional accuracy, but CRPS is preferable in its ability to quantify the accuracy of non-Gaussian distributional estimates. The median was excluded as an aggregation function to optimize because we found it to be insufficiently sensitive to situations in which the filter exhibits large intermittent excursions from the true state.

After exploring broad patterns with a Sobol sequence, we switched to a Bayesian optimization method for choosing new arms to evaluate. Using a Bayesian optimization method substantially accelerated convergence to optimal filter parameters relative to the quasirandom search. In short, this involved fitting a Gaussian process surrogate model to the mean CRPS observations as a function on the parameter search space, and then choosing new arms that maximize a utility function under that surrogate model. We chose a utility function that estimates improvement from previously observed arms expected under the surrogate model. The arms are then evaluated in parallel, by running the filter on a subset of the reference simulations using those arms' filtering

parameters. Those results are then incorporated with previous results to fit a new Guassian process surrogate model used in the next iteration of the Bayesian optimization loop. The remainder of this section is dedicated to the technical details of the optimization strategy we used, implemented in Ax/BoTorch (Balandat et al., 2019).

Let $c_i(p)$ denote the observed mean CRPS for trial i with parameters p. It is reasonable to expect that the mean CRPS is a piecewise-continuous latent function f of the filter parameters, with small discontinuities due only to resampling discrepancies, for fixed values of observed data, initial ensembles, random resamplings, and random rotations. But since these values stochastically vary in practice, we can view each such f as a realization of a random field $\mathscr{F}$ of functions $f : P \to \mathbb{R}$ where $P \subset \mathbb{R}^m$ is the search space of m filter parameters. In this view, the quantities $c_i(p)$ are noisy observations of the random field's true mean $\overline{\mathscr{F}}$. Our Bayesian optimizer seeks the minimizer of $\overline{\mathscr{F}}$ using these noisy observations.

Let $\overline{c}_p$ be the mean CRPS observed over all assimilation trials that were run with parameters p. The quantity $\overline{c}_p$ is, in other words, a mean of all the mean CRPS values observed for the arm p. Then let ξ_p be the standard error of that mean over trials for the arm p. We approximate this standard error of the mean as the sample standard deviation, of the set of trial means $\{\overline{c}_i(p)\}_{i\in(1,\ldots,N_{trials})}$, divided by the square root of the number of trials.

For convenience in setting hyperparameters of the Guassian process model, we scale the search space to the unit cube and standardize the observations. The raw search spaces are hyperrectangles, so they are scaled in each coordinate in the obvious manner to arrive at a unit cube. To standardize the observations, we subtract the mean of the set $\{\overline{c}_p\}$, for all parameter sets p previously evaluated in the experiment, and divide the result by the sample standard deviation $\sigma_{\overline{c}}$ of the same set. We do not introduce new notation for these values transformed to the unit cube; the remainder of this section will treat c in the standardized output space and will treat values of p in the scaled parameter space. The raw standard errors ξ_p are simultaneously divided by $\sigma_{\overline{c}}$ to preserve their validity in this standardized output space; define $\boldsymbol{\Xi}$ to be a diagonal matrix whose nonzero entries are these values $\xi_p/\sigma_{\overline{c}}$.

In these scaled and standardized variables, we form a surrogate model supposing that $\overline{\mathscr{F}}$ depends on p as a Gaussian process

$$\mathcal{GP} \sim \mathcal{N}(0, k(p, p')). \tag{4.16}$$

We take $k(p, p')$ to be the Matérn covariance kernel

$$k(p_i, p_j) = \frac{\Theta_s \; 2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu}\; d(p_i, p_j)\right) K_\nu \left(\sqrt{2\nu}\; d(p_i, p_j)\right), \tag{4.17}$$

where K_ν is the modified Bessel function of the second kind, and

$$d(\mathbf{p}_i, \mathbf{p}_j) = (\mathbf{p}_i - \mathbf{p}_j)^\top \mathbf{\Theta}_d^{-1} (\mathbf{p}_i - \mathbf{p}_j). \tag{4.18}$$

Here $\mathbf{\Theta}_d$ is a diagonal matrix of length scale hyperparameters. Each of the scalars on the diagonal of $\mathbf{\Theta}_d$ corresponds to a length scale of a feature in the space of scaled filter parameters, and we endow each with a Gamma-distributed prior $\Gamma(\lambda_d, r_d)$ with shape $\lambda_d = 6$ and rate $r_d = 3$. The factor Θ_s is another hyperparameter that controls the covariance function's overall scale, on which we also impose a Gamma distribution prior $\Gamma(\lambda_S, r_S)$ with shape $\lambda_S = 2$ and rate $r_S = 0.15$.

Whereas each f can only be expected to be piecewise-continuous, this Gaussian process regression aims to fit the true mean $\overline{\mathscr{F}}$. We can expect $\overline{\mathscr{F}}$ to be substantially smoother than individual realizations f, and we set the smoothness parameter $\nu = 5/2$ so that realizations are almost surely twice-differentiable. Let us summarize for clarity: we are modeling the mean of one discontinuous random field as a realization of a separate random field, of the Matérn class, that has twice-differentiable realizations almost surely.

Another popular choice of Gaussian process covariance kernel is the squared exponential (i.e. Gaussian) function, which yields infinitely differentiable realizations almost surely. Without having seen how smooth our results might be, the Matérn kernel was the more conservative choice so that we could resolve a potentially rugged CRPS landscape. Given the smoothness of the results, however, a kernel such as a squared exponential may have been a better choice. It was overly optimistic to assume that observations would be heteroskedastic, and the regression quality suffered due to that misspecification, but the resulting regression was still good enough to substantially

accelerate the search for optimal parameters and provide confidence that we are making a fair comparison between the three assimilation methods investigated.

Setting $\mathbf{c}$ to be the vector of mean CRPS values $\bar{c}_{p_i}$ observed for each attempted parameter configuration p_i and marginalizing over the latent function f yields a log-likelihood

$$\ln P\left(\mathbf{c}|\{\mathbf{p}_i\},\Theta\right) = -\frac{1}{2}\mathbf{c}^\top(\mathbf{K}+\boldsymbol{\Xi})^{-1}\mathbf{c} - \frac{1}{2}\ln|\mathbf{K}+\boldsymbol{\Xi}| - \frac{N_p}{2}\ln(2\pi) \tag{4.19}$$

$$+\sum_{j=1}^{N_p}\left[(\lambda_L-1)\ln\left(\Theta_{L,j}\right) - r_L\Theta_{L,j} + \lambda_L\ln r_L - \ln\Gamma(\lambda_L)\right] \tag{4.20}$$

$$+(\lambda_S-1)\ln\left(\Theta_S\right) - r_S\Theta_{S,j} + \lambda_S\ln r_S - \ln\Gamma(\lambda_S), \tag{4.21}$$

where $K_{ij} = k(p_i, p_j)$. The equation above obtains by adding the log-likelihood of our hyperparameter priors to Equation 2.30 of Williams and Rasmussen (2006). The Gaussian process surrogate is then fit to the rescaled data by maximizing the log-density eq. (4.19) using the L-BFGS-B method (Byrd et al., 1995) to arrive at a maximum marginal likelihood estimate for the hyperparameters. With the hyperparameters now fixed, we then have a well-defined Gaussian process predictor for f. Finally, a batch of candidate arms is generated that approximately optimizes the batched noisy expected improvement acquisition function (Letham et al., 2019) acting on that predictor. Batch sizes varied between 1 and 32 depending on computational resources available at the time.

A leave-one-out cross validation is performed to provide some confidence that it is reasonable to compare optimal values observed: we forecast the mean CRPS for each arm based on the Gaussian process regressor developed from the remaining mean CRPS observations. The forecasted values can then be compared to their corresponding actual observed values. When the forecasts closely match true observations, that helps establish that the Gaussian process model is a fair predictor of the mean CRPS that we seek to optimize. Since the Bayesian optimizer uses that Gaussian process model to select candidate arms to evaluate, such concordance of prediction and observation provides some evidence that the Bayesian optimizer has hope to do its job of finding optimal values.

4.4 Results

4.4.1 Large ensemble

Both the smoothed and unsmoothed hybrid filter outperformed the pure ESRF with an ensemble size $N_e = 1200$. The SSIR-ESRF hybrid attained a mean analysis CRPS that is 15% less than the minimum mean analysis CRPS observed of the ESRF. The same filtering parameters also resulted in a 10% reduction in mean analysis RMSE, a 14% reduction in forecast CRPS, and a 10% reduction in forecast RMSE. These data and corresponding RMSE for trials that optimized mean CRPS are summarized in table 4.1.

In all of the optimal cases observed, the analysis RMSE is smaller than the observation error standard deviation. This means that the filters are making meaningful use of the forecast, rather than merely trusting observations, despite the long assimilation timestep and small observation error.

The unhybridized ESRF achieved a minimum mean CRPS of 0.0972 ± 0.0013 with inflation factor $r = 0.027$ and localization radius $L = 474$. The unsmoothed hybrid filter achieved a minimum mean CRPS of 0.0931 ± 0.0014 with parameters $r = 0.04$, $L = 237$, and target effective sample size $ESS_0 = 727$. The smoothed hybrid filter achieved a minimum mean CRPS of 0.0823 ± 0.0008 with parameters $r = 0.04$, $L = 529$, and $ESS_0 = 761$. Recall that there are 128 spatial variables within each of the 41 large-scale Lorenz-96 modes, so L=474 corresponds to a localization radius of about 3.7 in the standard Lorenz-96 model.

Results from the cross validation for the unhybridized ESRF, plotting the predicted value versus actual observations, are shown in fig. 4.3. It is only imperative that the optimizer's surrogate model accurately represents a region near the optimum, in a case like this where we have reasonable cause to believe that the optimum is a well-behaved well (c.f. contours below), so we also performed a 'zoomed-in' version of the cross validation. This used the same training data but withheld points for cross-validation only from the subset of arms for which the observed mean CRPS is less than 0.12. On this restricted test dataset, the mean absolute predictive error is 0.04, the predictive

Pearson correlation is $\rho_p = 0.49$, and the predictive Spearman correlation is $\rho_s = 0.67$. The predictive error is fairly large, but the fair correlation statistics provide some credibility to the surrogate model. More so, it is promising to see the good alignment of the predicted and observed means within the cross validation test dataset. This suggests that the Guassian process surrogate model is fairly representing the CRPS landscape as a function of the filtering parameters, lending a degree of confidence in the optimizer having a chance to effectively find the minimum.

Results from the cross validation for the unsmoothed SIR-ESRF hybrid are shown in fig. 4.4. For the restricted case using a training set having observed CRPS mean less than 0.2, the mean absolute predictive error is 0.19, the predictive Pearson correlation is $\rho_p = 0.53$, and the predictive Spearman correlation is $\rho_s = 0.54$. These suggest that the Guassian process surrogate model is a poor representation of the CRPS landscape as a function of the filtering parameters within the selected region, so it is plausible that further iterations will result in measurably better performance. That said, the unrestricted cross validation shown in section 4.4.1 shows that the optimizer has successfully focused computational effort in regions with fairly good performance, and the optimum appears to be fairly shallow and well-behaved. This provides some confidence that the optimal parameter set, while probably a bit better, is not dramatically dissimilar from that which we have already found.

4.4.2 Small ensemble

For $N_e = 400$, there were no significant differences between optimal performance metrics observed of the three cases we examined. For the unhybridized ESRF-MPRR filter, the best result identified had a mean analysis CRPS of 0.1022 ± 0.0015 with parameters $r = 0.026$ and $L = 209$. The best arm identified of the unsmoothed hybrid SIR-ESRF-MPRR filter resulted in a mean analysis CRPS of 0.0996 ± 0.0016 with parameters $r = 0.067$, $L = 236$, and $ESS_0 = 239$. The smoothed SIR-ESRF-MPRR filter, for smoothing parameters $\ell = 0.5$ and $\beta = 2$ fixed in the search for optimal values of the remaining filtering parameters, the best result identified had an analysis CRPS of 0.0993 ± 0.0014 with parameters $r = 0.055$, $L = 239$, and $ESS_0 = 328$. Forecast CRPS,

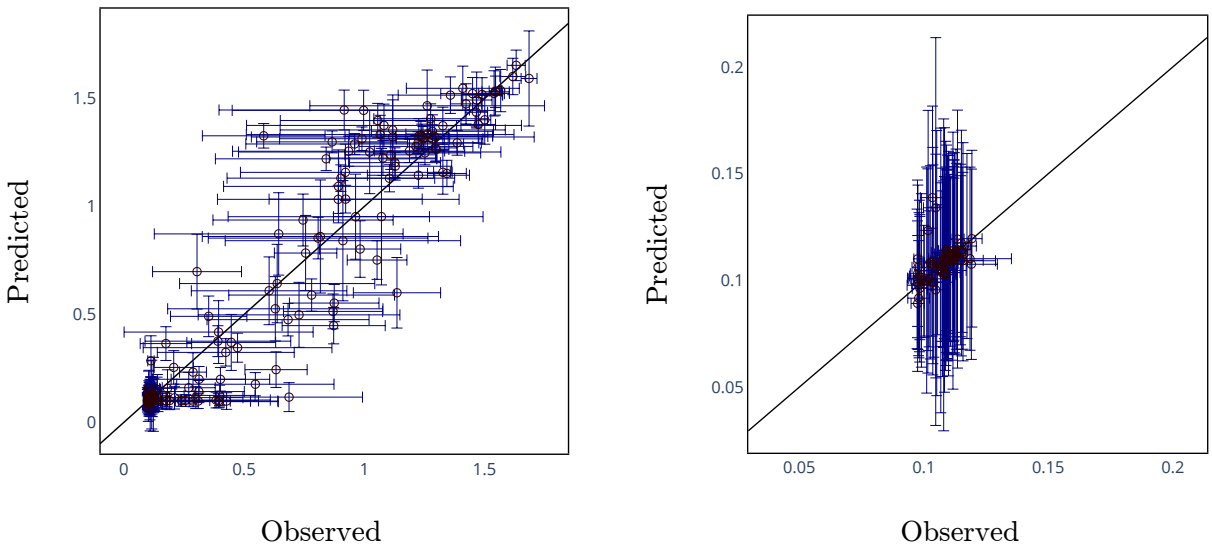


Figure 4.3: (a) Observed analysis CRPS versus the Gaussian surrogate model's prediction for a leave-one-out cross validation of all observed arms in the unhybridized ESRF with $N_e = 1200$. (b) Likewise, but with the validation set restricted to those outcomes having observation mean less than 0.12. For the restricted case, the mean absolute predictive error is 0.04, the predictive Pearson correlation is $\rho_p = 0.49$, and the predictive Spearman correlation is $\rho_s = 0.67$. These provide weak evidence that the Guassian process surrogate model reasonably represents the CRPS landscape as a function of the filtering parameters.

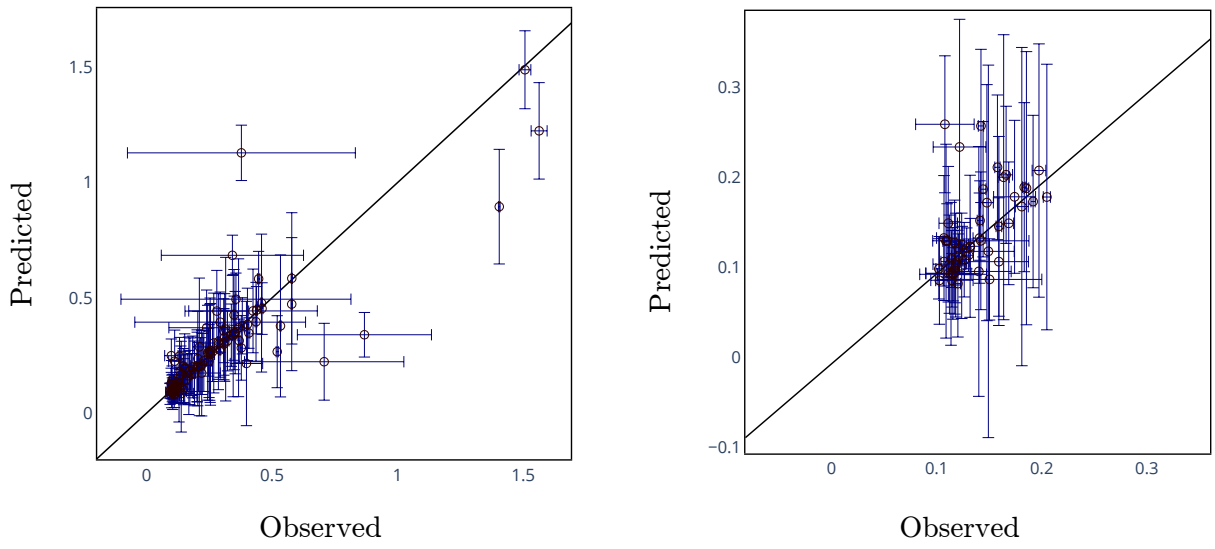


Figure 4.4: (a) Observed analysis CRPS versus the Gaussian surrogate model's prediction for a leave-one-out cross validation of all observed arms in the unhybridized ESRF with $N_e = 1200$. (b) Likewise, but with the validation set restricted to those outcomes having observation mean less than 0.12. For the restricted case, the mean absolute predictive error is 0.19, the predictive Pearson correlation is $\rho_p = 0.53$, and the predictive Spearman correlation is $\rho_s = 0.54$. These suggest that the Guassian process surrogate model is a poor representation of the CRPS landscape as a function of the filtering parameters within the selected region, so it is not implausible that further iterations will result in better performance.

as well as RMSE for both the analysis forecast, are also indicated in table 4.1.

Results from a cross validation of Bayesian optimizer's surrogate model are plotted in fig. 4.5 for the unhybridized ESRF case. As before, we also performed a cross validation of the surrogate model on a subset of test data comprising those observations with mean observed CRPS less than 0.2. the mean absolute predictive error is 0.025, the predictive Pearson correlation is $\rho_p = 0.97$, and the predictive Spearman correlation is $\rho_s = 0.96$. These provide some evidence that the Guassian process surrogate model accurately represents the CRPS landscape as a function of the filtering parameters, lending a degree of confidence in the optimizer having arrived at a minimum that is visually apparent in fig. 4.7a. On this restricted test dataset, the mean absolute predictive error is 0.025, the predictive Pearson correlation is $\rho_p = 0.97$, and the predictive Spearman correlation is $\rho_s = 0.96$.

Cross validation results for the unsmoothed hybrid are plotted in fig. 4.6. Within a test set having observed CRPS < 0.12, the mean absolute predictive error is 0.044, the predictive Pearson correlation is $\rho_p = 0.48$, and the predictive Spearman correlation is $\rho_s = 0.65$. These provide weak evidence that the Guassian process surrogate model represents the CRPS landscape within reason, though they provide reason to believe that further iterations of the optimizer would yield some benefit. That said, on account of how shallow the minimum appears to be (c.f. contours), we do not anticipate dramatic improvement from additional iterations.

4.4.3 Exploring the parameter space

Gaussian process surrogate models for analysis CRPS and analysis RMSE, in the case of an unhybridized ESRF with $N_e = 400$, are shown in fig. 4.7a and fig. 4.7b. Corresponding plots for forecast CRPS and RMSE are shown in fig. 4.8a and fig. 4.8b. These depict the Gaussian distribution predicted by the surrogate model, as a function of the parameter values of a single unobserved arm. They demonstrate that the optimum is quite shallow — a wide range of parameters yields similar performance within a region near the optimum. The figures have been cropped from a larger search space to show more detail in the optimal region, outside of which performance

		Analysis		Forecast	
N_e	Configuration	CRPS	RMSE	CRPS	RMSE
400					
	smoothed	0.0993(14)	0.2998(74)	0.4887(50)	1.178(12)
	unsmoothed	0.0999(27)	0.3168(102)	0.4832(94)	1.188(26)
	unhybridized	0.1022(15)	0.3033(92)	0.5070(39)	1.233(10)
800					
	smoothed	0.0929(23)	0.2891(112)	0.4529(71)	1.106(19)
	unsmoothed				
	unhybridized				
1200					
	smoothed	0.0823(8)	0.2536(36)	0.4173(33)	1.028(8)
	unsmoothed	0.0931(14)	0.2694(71)	0.4673(70)	1.129(18)
	unhybridized	0.0973(13)	0.2814(74)	0.4847(28)	1.146(7)

Table 4.1: RMSE and CRPS for optimal configurations of the ESRF, unsmoothed hybrid filter, and smoothed-observation hybrid filter. Numbers in parentheses indicate uncertainty in the least-significant digits, computed as one standard error of the mean.

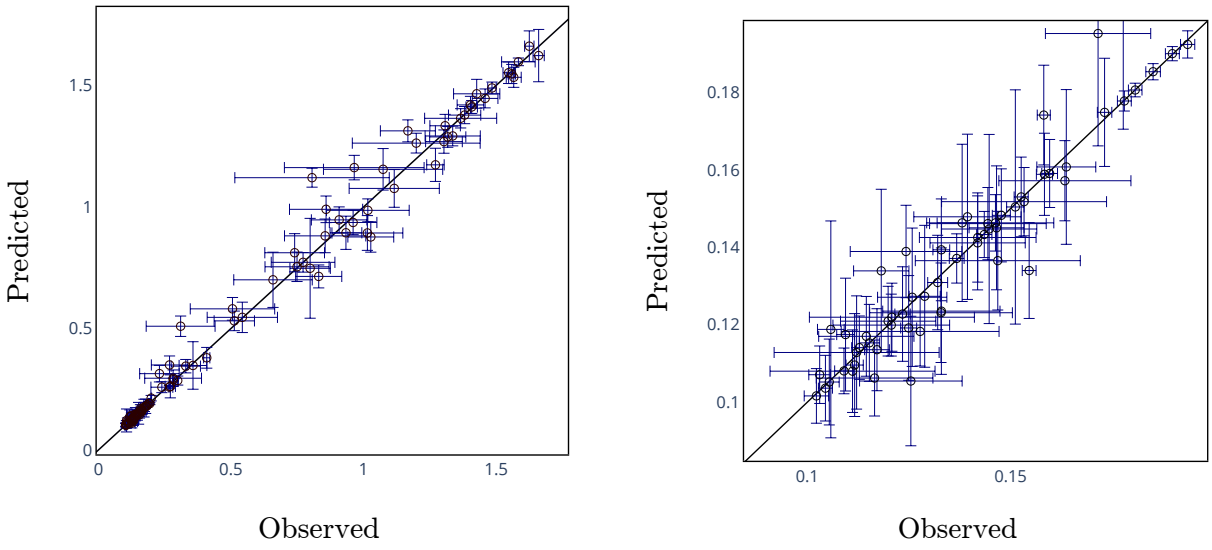


Figure 4.5: (a) Observed analysis CRPS versus the Gaussian surrogate model's prediction for a leave-one-out cross validation of all observed arms in the unhybridized ESRF with $N_e = 400$. (b) Likewise, but with the validation set restricted to those outcomes having observation mean less than 0.2. For the restricted case, the mean absolute predictive error is 0.025, the predictive Pearson correlation is $\rho_p = 0.97$, and the predictive Spearman correlation is $\rho_s = 0.96$. These provide some evidence that the Guassian process surrogate model accurately represents the CRPS landscape as a function of the filtering parameters.

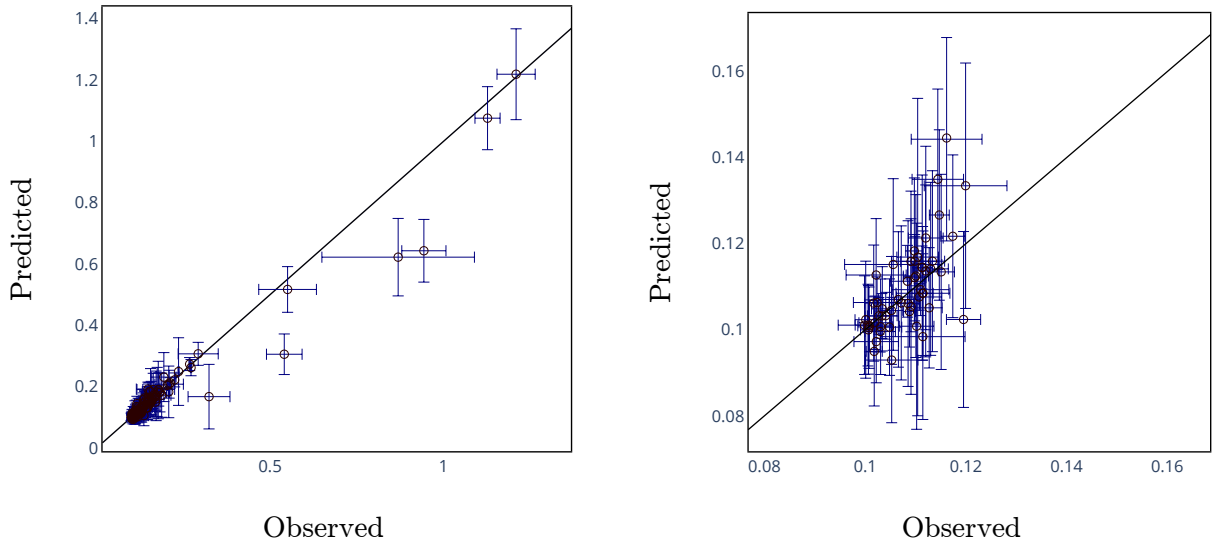


Figure 4.6: (a) Observed analysis CRPS versus the Gaussian surrogate model's prediction for a leave-one-out cross validation of all observed arms in the unsmoothed SIR-ESRF hybrid with $N_e = 400$. (b) Likewise, but with the validation set restricted to those outcomes having observation mean less than 0.12. For the restricted case, the mean absolute predictive error is 0.044, the predictive Pearson correlation is $\rho_p = 0.48$, and the predictive Spearman correlation is $\rho_s = 0.65$. These provide weak evidence that the Guassian process surrogate model represents the CRPS landscape within reason, though they provide reason to believe that further iterations of the optimizer would yield some benefit. That said, on account of how shallow the minimum appears to be (c.f. contours), we do not anticipate dramatic improvement from additional iterations.

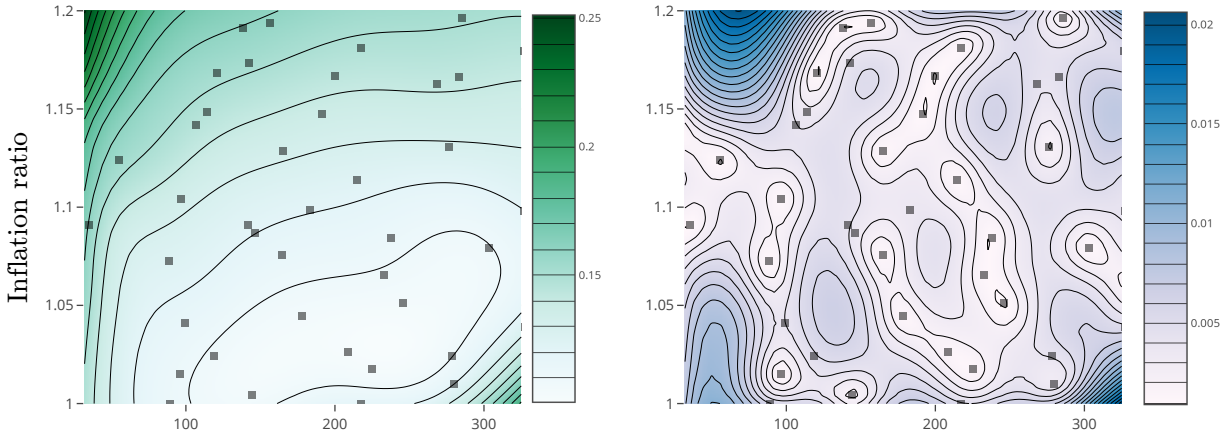
substantially degrades. Analysis CRPS strongly correlates with forecast CRPS, and with RMSE both in the analysis and forecast. Therefore the other metrics yield visually similar results, as in the visual similarity of these figures, so additional plots of this format will show only analysis CRPS.

Contour plots for the Gaussian process surrogate model for analysis CRPS, in the case of the unsmoothed hybrid filter, are shown in fig. 4.9a, fig. 4.10a, and fig. 4.9b. As in the unhybridized case, these demonstrate a fairly shallow minimum. Results from the cross validation of the unsmoothed case are shown in fig. 4.6, with a closeup cross validation plotted in the adjacent panel.

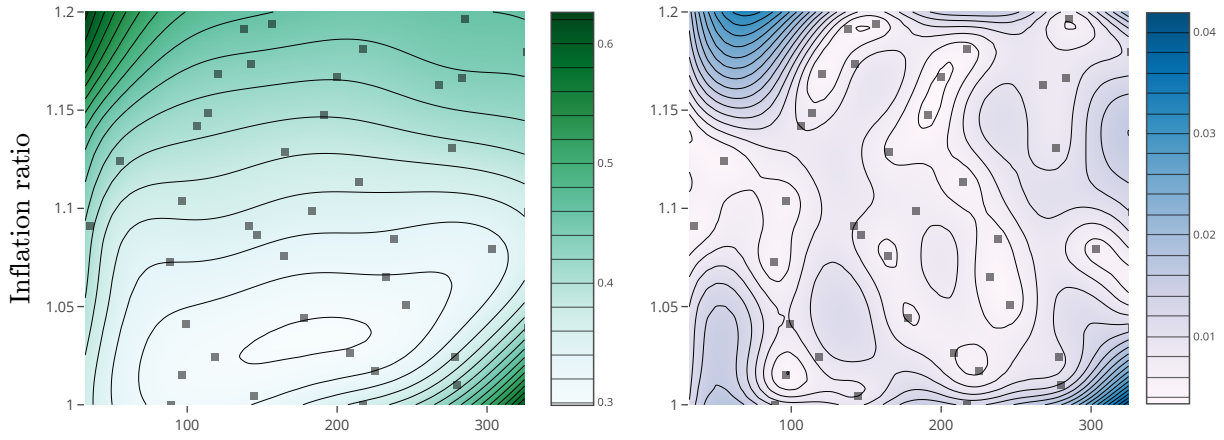
The search space, having an extra parameter, is substantially larger than that of the pure ESRF. Substantial computational effort went into searching for the minimum values, and we are satisfied that it is unlikely for a point to exist that is dramatically superior to those already sampled, even though the cross validation results described above make a less compelling case for that (likely as a result of this larger search space). Those results, while unimpressive, still provide weak evidence that the Guassian process surrogate model represents the CRPS landscape within reason, though they provide reason to believe that further iterations of the optimizer would yield some benefit. That said, it is qualitatively evident how shallow the minimum is (c.f. contour plots). This is why we do not anticipate dramatic improvement from additional iterations.

4.5 Discussion

The aforementioned differences are statistically large for an ensemble size $N_e = 1200$. The difference in analysis CRPS between the smoothed hybrid and ESRF filters when $N_e = 1200$ corresponds to a Cohen's d-statistic > 13, the smoothed hybrid versus unsmoothed hybrid filters differ in CRPS with $d > 9$, and the unsmoothed versus ESRF filters differ in CRPS with $d > 3$. However, this should be interpreted with a healthy measure of caution since the comparison is based on fixed filtering parameters. We have so far been unable to achieve strong evidence to rule out the possibility that there exists another set of parameters for the unsmoothed hybrid filter or the ESRF that would improve upon those we tested. The cross validation study shows that

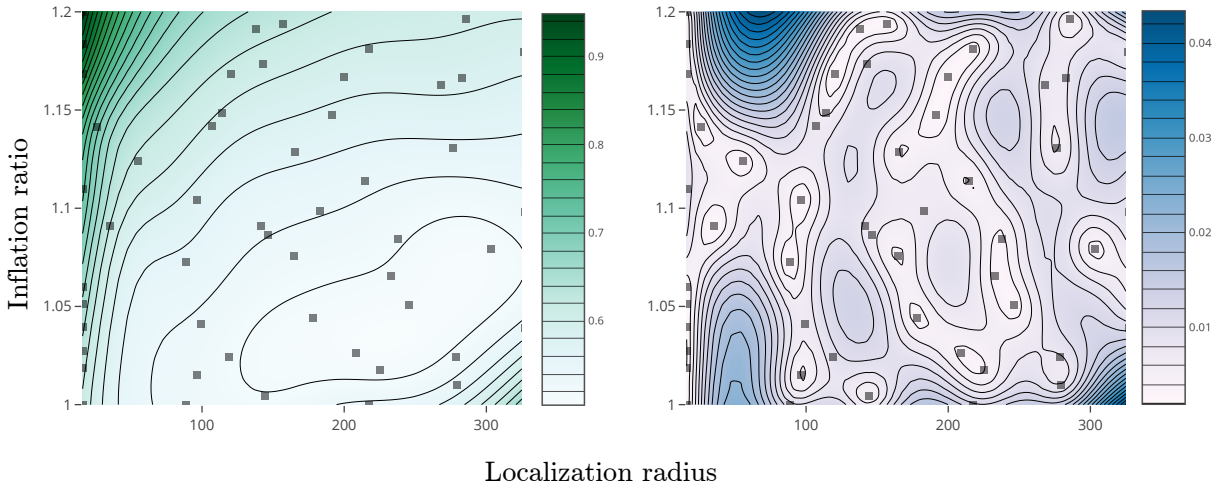


(a) Mean analysis CRPS

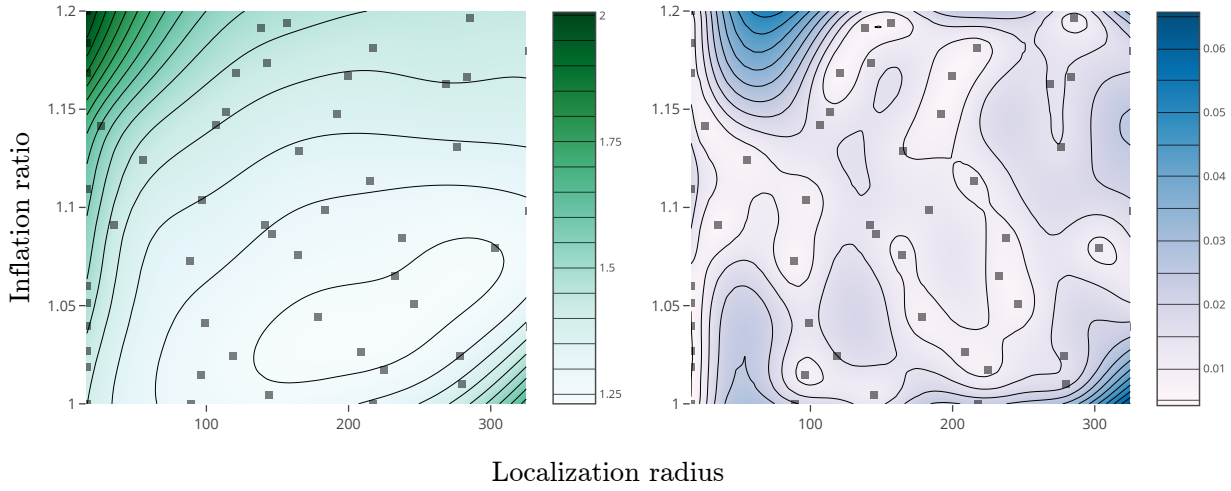


(b) Mean analysis RMSE

Figure 4.7: Filter performance in the analysis ensemble as a function of filtering parameters, for the case of a pure ESRF using $N_e = 400$ ensemble members. (a) Left: contours and shading represent the mean prediction for analysis CRPS as a function of filtering parameters under the Gaussian process surrogate model. Right: contours and shading represent standard deviation of the prediction under the same surrogate model. Squares represent observations. (b) As in (a) but for analysis RMSE.

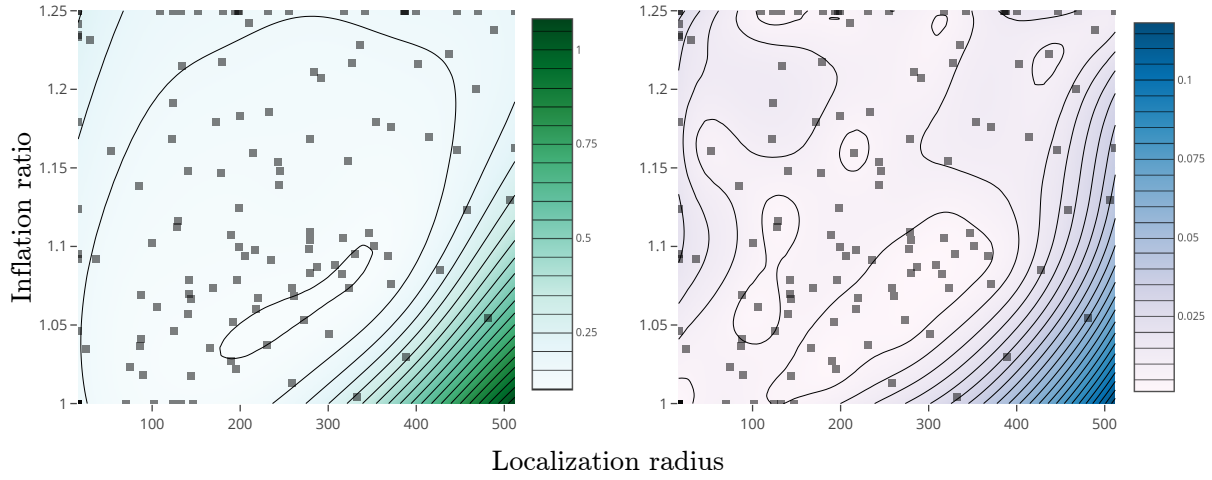


(a) Mean forecast CRPS

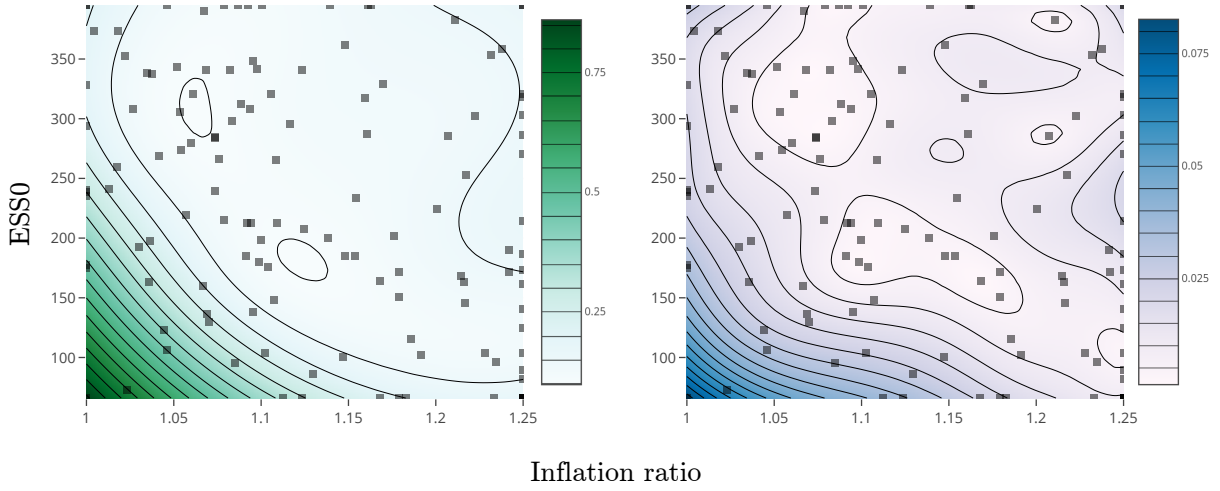


(b) Mean forecast RMSE

Figure 4.8: Variation of filter performance in the analysis ensemble as a function of filtering parameters, for the case of a pure ESRF using $N_e = 400$ ensemble members. (a) Left: contours and shading represent the mean prediction for analysis CRPS as a function of filtering parameters under the Gaussian process surrogate model. Right: contours and shading represent standard deviation of the prediction under the same surrogate model. (b) As in (a) but for analysis RMSE.

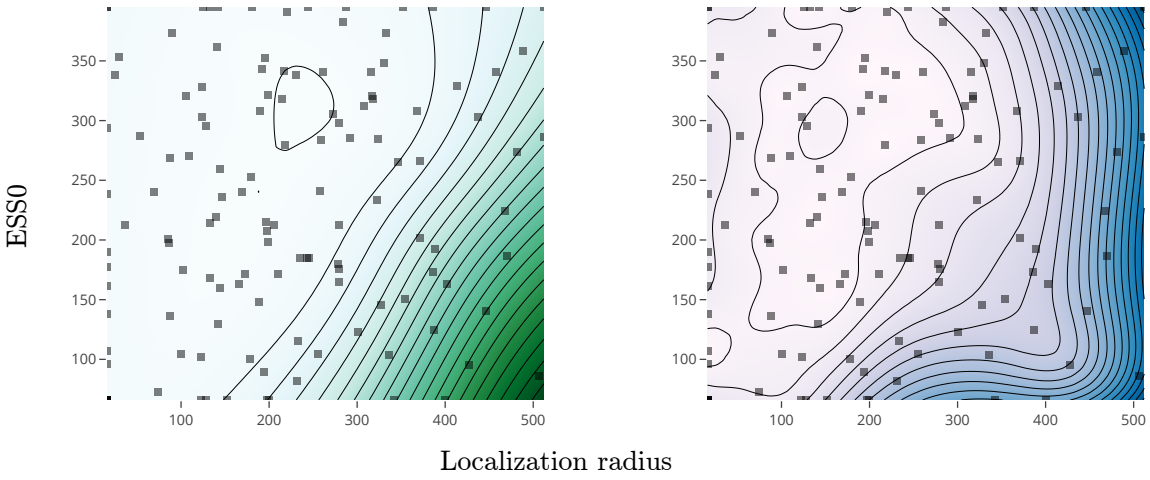


(a) Mean analysis CRPS cross-section, $ESS_0 = 306$



(b) Mean analysis CRPS cross-section, $L = 273$

Figure 4.9: Filter performance in the analysis ensemble as a function of filtering parameters, for the case of an unsmoothed hybrid SIR-ESRF using $N_e = 400$ ensemble members. (a) Left: contours and shading represent the mean prediction for analysis CRPS as a function of filtering parameters r and L under the Gaussian process surrogate model, keeping $ESS_0 = 306$ fixed. Right: contours and shading represent standard deviation of the prediction under the same surrogate model. Squares represent observations. (b) Likewise, but with varying r and ESS_0 while keeping $L = 273$ fixed.



(a) Mean analysis CRPS cross-section, $r = 0.053$

Figure 4.10: Filter performance in the analysis ensemble as a function of filtering parameters, for the case of an unsmoothed hybrid SIR-ESRF using $N_e = 400$ ensemble members. (a) Left: contours and shading represent the mean prediction for analysis CRPS as a function of filtering parameters L and ESS_0 under the Gaussian process surrogate model, keeping $r = 0.053$ fixed. Right: contours and shading represent standard deviation of the prediction under the same surrogate model. Squares represent observations.

the Guassian process surrogate model utilized by the Bayesian optimizer still displays significant uncertainty and bias. That said, we have dedicated substantial computational resources, with the same optimization methodology, in a good-faith effort to tune the parameters both of the ESRF status-quo and of the unsmoothed hybrid filter.

Suffice it to say that we did not bother including cross validation plots for the smoothed hybrid because they are terrible. The cross validation results that are included here are superior to those found of the smoothed hybrid filter because we have dedicated extra effort toward tuning the competition, in order that we can be more confident in the smoothed hybrid's dominance. The poor cross validation results for the smoothed hybrid gives just as much (if not more) reason to believe that we can further tune the parameters of the smoothed hybrid. This also suggests that our smoothed hybrid algorithm is easier to tune, even if it does prove to be inferior to the ESRF for some yet undiscovered configuration.

The ease with which we tuned the smoothed hybrid filter may relate to how shallow its optimum is. All methods investigated here exhibit shallow CRPS landscapes, as a function of their free parameters, in the vicinity of their minima. This feature means they are robust to fairly large misspecification of filtering parameters around their optimum.

4.6 Conclusion

Ensemble approximations to the class of square root filters are highly effective in solving data assimilation problems that are weakly nonlinear and non-Gaussian. On the other hand, they are biased for problems that possess non-Gaussian prior or posterior filtering distributions. Problems that possess 'strong nonlinearity,' wherein both the prior and posterior are substantially non-Gaussian, suggest the use of sequential Monte Carlo methods (particle filters) that weakly converge to the correct posterior distribution even in the context of severely non-Gaussian behavior.

A subtler class of data assimilation problems possess 'medium nonlinearity', in which the filtering posterior is nearly Guassian but the filtering prior is substantially non-Gaussian. For this class of problems, Morzfeld and Hodyss (2019) recommend variational approaches that make a

Gaussianity assumption in the posterior, but variational data assimilation methods usually do not provide uncertainty estimates. Alternatively, medium nonlinearity is a promising application of 'bridging' (Chustagulprom et al., 2016, Frei and Künsch, 2013) between the sequential importance resampling particle filter and an ensemble implementation of a square root filter.

We presented an approach to adaptively bridge between sequential importance sampling (SIR) and the ensemble square root filter (ESRF), in order to glean some of each of their complementary benefits. This derives from the hope that the SIR step provides ESRF with a prior that more closely conforms to ESRF's Guassianity assumptions, and that ESRF helps shepherd ensemble members toward observations in order to present SIR with a more efficient prior ensemble in the next assimilation cycle. We pair this bridging approach with observation error variance inflation in the SIR step, selectively applied to small scales. Selectively inflating observation variance at small scales, is equivalent to smoothing innovations and assuming that the smoothed innovations have spatially uncorrelated observation error. This improves the performance of the SIR step by reducing its notorious tendency to produce an underdispersed posterior ensemble.

We indeed see significant improvement over ESRF from the SIR-ESRF hybrid, but only for a sufficiently large ensemble. The hybrid filter outperformed the ESRF even without smoothing innovations, using 1200 ensemble members, but the hybrid with smoothing significantly outperformed the unsmoothed hybrid too. The ESRF did not perform much better with an ensemble size of 1200 than it did with 400 ensemble members, suggesting that its performance is primarily limited by non-Gaussianity rather than sampling errors even with the smaller ensemble.

With our algorithm using an ensemble size of 1200 we found a 15% improvement in mean analysis CRPS, relative to the plain ESRF using the same ensemble size. This was accompanied by a 14% improvement in forecast CRPS and a 10% improvement in the RMSE of both the analysis and forecast. The unsmoothed version also outperformed ESRF by 4.3% in analysis CRPS, 4.3% in analysis RMSE, 3.4% in forecast CRPS, and 1.5% in forecast RMSE.

Although these are substantial improvements over the ESRF and the unsmoothed hybrid, there is still a good deal of uncertainty in the optimal values of the tunable free parameters. We

made a good-faith effort to tune the parameters of the underperforming models, and we are pleased that our smoothed hybrid filter at least appears easier to tune, but we have not been able to arrive at a statistically compelling case to rule out the existence of a parameter configuration under which the ESRF or the unsmoothed hybrid filter outperform the smoothed hybrid.

In retrospect, given how smoothly the CRPS seem to vary with the filtering parameters, a squared exponential covariance kernel probably would have produced a more efficient parameter search than Matérn kernel did. A Gaussian process with a squared exponential covariance function has realizations that are infinitely differentiable almost surely. We chose the Matérn kernel because we did not have reason to believe that the results would be so gradual, a priori, so the Mateérn kernel with smoothness parameter $\nu = 5/2$ seemed to be a more conservative choice in order to resolve more detail in the event that CRPS was a rugged function of the filtering parameters. But since the CRPS landscape seems not to be so rugged, a kernel that causes the Gaussian process to yield infinitely differentiable realizations may be justified. This could allow the surrogate model to make more confident predictions, accelerating the optimization process.

We also see in retrospect that our search procedure might have proceeded more efficiently with a different choice in how to report the standard error of the mean CRPS when forming the Gaussian process. Whereas we chose to use a standard error based on an estimated variance of an unobserved population of possible realizations of a random field, it may be more efficient to fix a small training set of reference simulations, initial ensembles, and seeds used by the pseudorandom number generator employed in the filter implementations. This would allow one to report the standard error of the mean as zero, i.e. $\boldsymbol{\Xi} = \mathbf{0}$. This changes the optimization problem from minimizing $\overline{\mathscr{F}}$, the mean CRPS in expectation over the population of possible data in the data assimilation problem class, to minimizing the mean CRPS for a specific training set drawn from $\mathscr{F}$. The latter optimization problem would cause the Gaussian process regressor to over-fit the data, relative to the view that we should optimize for expectation over the population, and discontinuities may pose a challenge for optimization routines that require differentiability. That said, a sufficiently large training set ought to be an adequate representation of $\overline{\mathscr{F}}$ to overcome both of these difficulties.

It is unclear whether this approach is more effective than what we chose to try here, in optimization problems of this type, and we hope future research will resolve that question.

Bibliography

Walter Acevedo, Jana de Wiljes, and Sebastian Reich. Second-order accurate ensemble transform particle filters. SIAM J Sci Comput, 39(5):A1834–A1850, 2017.

Mel Ades and Peter J Van Leeuwen. The equivalent-weights particle filter in a high-dimensional system. Quart. J. Roy. Meteor. Soc., 141(687):484–503, 2015.

Melanie Ades and Peter Jan Van Leeuwen. An exploration of the equivalent weights particle filter. Quart. J. Roy. Meteor. Soc., 139(672):820–840, 2013.

S Agapiou, Omiros Papaspiliopoulos, D Sanz-Alonso, and AM Stuart. Importance sampling: Intrinsic dimension and computational cost. Statistical Science, 32(3):405–431, 2017.

Jeffrey Anderson, Tim Hoar, Kevin Raeder, Hui Liu, Nancy Collins, Ryan Torn, and Avelino Avellano. The data assimilation research testbed: A community facility. B Am Metorol Soc, 90 (9):1283–1296, 2009.

Jeffrey L Anderson. An ensemble adjustment Kalman filter for data assimilation. Monthly weather review, 129(12):2884–2903, 2001.

Jeffrey L Anderson and Nancy Collins. Scalable implementations of ensemble filter algorithms for data assimilation. J Atmos Ocean Tech, 24(8):1452–1463, 2007.

Maximilian Balandat, Brian Karrer, Daniel R. Jiang, Samuel Daulton, Benjamin Letham, Andrew Gordon Wilson, and Eytan Bakshy. Botorch: Programmable bayesian optimization in pytorch, 2019.

Thomas Bengtsson, Peter Bickel, and Bo Li. Curse-of-dimensionality revisited: Collapse of the particle filter in very large scale systems. In Deborah Nolan and Terry Speed, editors, Probability and Statistics: Essays in Honor of David A. Freedman, volume Volume 2 of Collections, pages 316–334. Institute of Mathematical Statistics, Beachwood, Ohio, USA, 2008. doi: 10.1214/193940307000000518. URL `http://dx.doi.org/10.1214/193940307000000518`.

Gregory Beylkin and Lucas Monzón. Approximation by exponential sums revisited. Appl Comput Harmon A, 28(2):131–149, 2010.

Craig H Bishop, Brian J Etherton, and Sharanya J Majumdar. Adaptive sampling with the ensemble transform kalman filter. part i: Theoretical aspects. Monthly weather review, 129(3): 420–436, 2001.

Eric Brochu, Vlad M. Cora, and Nando de Freitas. A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. CoRR, abs/1012.2599, 2010. URL http://arxiv.org/abs/1012.2599.

Richard H Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. SIAM Journal on Scientific Computing, 16(5):1190–1208, 1995.

Alexandre Chorin, Matthias Morzfeld, and Xuemin Tu. Implicit particle filters for data assimilation. Comm App Math Com Sc, 5(2):221–240, 2010.

Alexandre J Chorin and Matthias Morzfeld. Conditions for successful data assimilation. J. Geophys. Res., 118(20), 2013.

Alexandre J Chorin and Xuemin Tu. Implicit sampling for particle filters. Proc. Natl. Acad. Sci. (USA), 106(41):17249–17254, 2009.

Alexandre J Chorin and Xuemin Tu. An iterative implementation of the implicit nonlinear filter. ESAIM-Math Model Num, 46(3):535–543, 2012.

CK Chui and G Chen. Kalman Filtering. Springer, 4 edition, 2009.

Nawinda Chustagulprom, Sebastian Reich, and Maria Reinhardt. A hybrid ensemble transform particle filter for nonlinear and spatially extended dynamical systems. SIAM/ASA J. Uncertainty Quantification, 4(1):592–608, 2016.

Dan Crisan and Arnaud Doucet. A survey of convergence results on particle filtering methods for practitioners. IEEE T signal proces, 50(3):736–746, 2002.

A Doucet, Nando De Freitas, and N Gordon. An introduction to sequential Monte Carlo methods. In Sequential Monte Carlo methods in practice, pages 3–14. Springer, 2001.

Arnaud Doucet and Adam M. Johansen. A tutorial on particle filtering and smoothing: Fifteen years later. In in Oxford Handbook of Nonlinear Filtering. University Press, 2009.

Arnaud Doucet, Simon Godsill, and Christophe Andrieu. On sequential Monte Carlo sampling methods for Bayesian filtering. Stat Comput, 10(3):197–208, Jul 2000.

D. M. Endres and J. E. Schindelin. A new metric for probability distributions. IEEE T Inform Theory, 49(7):1858–1860, July 2003.

Geir Evensen. Sampling strategies and square root analysis schemes for the enkf. Ocean dynamics, 54(6):539–560, 2004.

Geir Evensen. Data Assimilation: The Ensemble Kalman Filter. Springer, 2009.

Kira Feldmann, Michael Scheuerer, and Thordis L Thorarinsdottir. Spatial postprocessing of ensemble forecasts for temperature using nonhomogeneous Gaussian regression. Mon. Wea. Rev., 143(3):955–971, 2015.

Bengt Fornberg and Natasha Flyer. A primer on radial basis functions with applications to the geosciences. SIAM, 2015.

Bengt Fornberg and Julia Zuev. The Runge phenomenon and spatially variable shape parameters in RBF interpolation. Computers & Mathematics with Applications, 54(3):379–398, August 2007. ISSN 08981221. doi: 10.1016/j.camwa.2007.01.028. URL https://linkinghub.elsevier.com/retrieve/pii/S0898122107002210.

V Fortin, M Abaza, F Anctil, and R Turcotte. Why should ensemble spread match the rmse of the ensemble mean? Journal of Hydrometeorology, 15(4):1708–1713, 2014.

Marco Frei and Hans R Künsch. Bridging the ensemble Kalman and particle filters. Biometrika, 100(4):781–800, 2013.

IM Gelfand and NY Vilenkin. Generalized functions, volume 4: Applications of Harmonic Analysis. AMS Chelsea Publishing, 1964.

T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. J Am Stat Assoc, 102:359–378, 2007.

I. J. Good. Good Thinking: The Foundations of Probability and Its Applications. University of Minnesota Press, ned - new edition edition, 1983. ISBN 9780816611423. URL http://www.jstor.org/stable/10.5749/j.ctttsn6g.

Neil J Gordon, David J Salmond, and Adrian FM Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. In IEE Proceedings F (Radar and Signal Processing), volume 140, pages 107–113. IET, 1993.

Leslie Greengard and John Strain. The fast gauss transform. SIAM Journal on Scientific and Statistical Computing, 12(1):79–94, 1991.

I Grooms and Y Lee. A framework for variational data assimilation with superparameterization. Nonlinear Proc. Geoph., 22(5):601–611, 2015.

H Hersbach. Decomposition of the continuous ranked probability score for ensemble prediction systems. Weather Forecast, 15:559–570, 2000.

Roger A. Horn and Charles R. Johnson. Matrix analysis. Cambridge Univ. Press, Cambridge, 23. print edition, 1990. ISBN 0-521-38632-2.

Peter L Houtekamer and Herschel L Mitchell. A sequential ensemble kalman filter for atmospheric data assimilation. Monthly Weather Review, 129(1):123–137, 2001.

Falko Judt. Insights into atmospheric predictability through global convection-permitting model simulations. Journal of the Atmospheric Sciences, 75(5):14771497, May 2018. ISSN 1520-0469. doi: 10.1175/jas-d-17-0343.1. URL http://dx.doi.org/10.1175/jas-d-17-0343.1.

E Kalnay. Atmospheric modeling, data assimilation, and predictability. Cambridge University Press, 2002.

Matthias Katzfuss, Jonathan R Stroud, and Christopher K Wikle. Understanding the ensemble kalman filter. The American Statistician, 70(4):350–357, 2016.

Genshiro Kitagawa. Monte carlo filter and smoother for non-gaussian nonlinear state space models. J. Comput. Graph. Stat., 5(1):1–25, 1996.

William Kleiber, Adrian E Raftery, Jeffrey Baars, Tilmann Gneiting, Clifford F Mass, and Eric Grimit. Locally calibrated probabilistic temperature forecasting using geostatistical model averaging and local bayesian model averaging. Mon. Wea. Rev., 139(8):2630–2649, 2011a.

William Kleiber, Adrian E Raftery, and Tilmann Gneiting. Geostatistical model averaging for locally calibrated probabilistic quantitative precipitation forecasting. J Am Stat Assoc, 106 (496):1291–1303, 2011b.

Benjamin Letham, Brian Karrer, Guilherme Ottoni, Eytan Bakshy, et al. Constrained bayesian optimization with noisy experiments. Bayesian Analysis, 14(2):495–519, 2019.

Finn Lindgren, Håvard Rue, and Johan Lindström. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. J Roy Stat Soc B, 73(4):423–498, 2011.

Edward N Lorenz. The predictability of a flow which possesses many scales of motion. Tellus, 21 (3):289–307, 1969.

EN Lorenz. Predictability: A problem partly solved. In T Palmer and R Hagedorn, editors, Proceedings of Seminar on Predicability, volume 1, pages 1–18. ECMWF, Reading, UK, 1996.

Andrew J Majda and John Harlim. Filtering complex turbulent systems. Cambridge University Press, 2012.

William McLean. Exponential sum approximations for t-β. In Contemporary Computational Mathematics-A Celebration of the 80th Birthday of Ian Sloan, pages 911–930. Springer, 2018.

Vlad I Morariu, Balaji V Srinivasan, Vikas C Raykar, Ramani Duraiswami, and Larry S Davis. Automatic online tuning for fast gaussian summation. In Advances in neural information processing systems, pages 1113–1120, 2009.

Matthias Morzfeld and Daniel Hodyss. Gaussian approximations in filters and smoothers for data assimilation. Tellus A, 71(1):1–27, 2019.

Matthias Morzfeld, Xuemin Tu, Ethan Atkins, and Alexandre J Chorin. A random map implementation of implicit filters. J Comput Phys, 231(4):2049–2066, 2012.

Matthias Morzfeld, Daniel Hodyss, and Chris Snyder. What the collapse of the ensemble kalman filter tells us about particle filters. Tellus A, 69(1):1283809, 2017.

Radford M. Neal. The Harmonic Mean of the Likelihood: Worst Monte Carlo Method Ever, August 2008. URL `https://radfordneal.wordpress.com/2008/08/17/the-harmonic-mean-of-the-likelihood-worst-monte-carlo-method-ever/`.

Nicholas A. Nystrom, Michael J. Levine, Ralph Z. Roskies, and J. Ray Scott. Bridges: A uniquely flexible hpc resource for new communities and data analytics. In Proceedings of the 2015 XSEDE Conference: Scientific Advancements Enabled by Enhanced Cyberinfrastructure, XSEDE '15, pages 30:1–30:8, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3720-5. doi: 10.1145/2792745.2792775. URL `http://doi.acm.org/10.1145/2792745.2792775`.

K Okamoto, AP McNally, and W Bell. Progress towards the assimilation of all-sky infrared radiances: an evaluation of cloud effects. Quart. J. Roy. Meteor. Soc., 140(682):1603–1614, 2014.

Bernt Øksendal. Stochastic differential equations. Springer, 6 edition, 2003.

Ferdinand Osterreicher and Igor Vajda. A new class of metric divergences on probability spaces and its applicability in statistics. Annals I Stat Math, 55(3):639–653, Sep 2003.

Art B Owen. Scrambling sobol and niederreiter–xing points. Journal of complexity, 14(4):466–489, 1998.

S. G. Penny and T. Miyoshi. A local particle filter for high-dimensional geophysical systems. Nonlinear Proc Geoph, 23(6):391–405, 2016.

Jonathan Poterjoy. A localized particle filter for high-dimensional nonlinear systems. Mon. Wea. Rev., 144(1):59–76, 2016.

Manuel Pulido and Peter Jan van Leeuwen. Sequential monte carlo with kernel embedded mappings: The mapping particle filter. J. Comput. Phys., 2019.

Patrick Rebeschini and Ramon Van Handel. Can local particle filters beat the curse of dimensionality? Ann Appl Probab, 25(5):2809–2866, 2015.

Sebastian Reich. A nonparametric ensemble transform method for Bayesian inference. SIAM J Sci Comput, 35(4):A2013–A2024, 2013.

Gregor Robinson and Ian Grooms. A tunable multiresolution smoother for scattered measurements with application to particle filters. SIAM J. Sci. Comput., 2019. Submitted.

Gregor Robinson, Ian Grooms, and William Kleiber. Improving particle filter performance by smoothing observations. Monthly Weather Review, 146(2018):2433–2446, 2018.

Richard Rotunno and Chris Snyder. A generalization of lorenzs model for the predictability of flows with many scales of motion. Journal of the Atmospheric Sciences, 65(3):1063–1076, 2008.

H Rue and L Held. Gaussian Markov random fields: theory and applications. CRC press, 2005.

Pavel Sakov and Peter R Oke. Implications of the form of the ensemble transformation in the ensemble square root filters. Mon. Wea. Rev., 136(3):1042–1053, 2008.

Michael Scheuerer and Luca Büermann. Spatially adaptive post-processing of ensemble forecasts for temperature. J Roy Stat Soc C, 63(3):405–422, 2014.

Craig S Schwartz, Glen S Romine, Ryan A Sobash, Kathryn R Fossell, and Morris L Weisman. Ncar's experimental real-time convection-allowing ensemble prediction system. Weather Forecast., 30(6):1645–1654, 2015.

Craig S. Schwartz, Glen S. Romine, Ryan A. Sobash, Kathryn R. Fossell, and Morris L. Weisman. NCAR's real-time convection-allowing ensemble project. B Am Metorol Soc, 100(2):321–343, 2019. doi: 10.1175/BAMS-D-17-0297.1.

Chris Snyder, Thomas Bengtsson, Peter Bickel, and Jeff Anderson. Obstacles to high-dimensional particle filtering. Mon. Wea. Rev., 136(12):4629–4640, 2008.

Chris Snyder, Thomas Bengtsson, and Mathias Morzfeld. Performance bounds for particle filters using the optimal proposal. Mon. Wea. Rev., 143(11):4750–4761, 2015.

J. Towns, T. Cockerill, M. Dahan, I. Foster, K. Gaither, A. Grimshaw, V. Hazlewood, S. Lathrop, D. Lifka, G. D. Peterson, R. Roskies, J. Scott, and N. Wilkins-Diehr. Xsede: Accelerating scientific discovery. Computing in Science & Engineering, 16(05):62–74, sep 2014. ISSN 1521-9615. doi: 10.1109/MCSE.2014.80.

Peter Jan Van Leeuwen. Particle filtering in geophysical systems. Mon. Wea. Rev., 137(12):4089–4114, 2009.

Peter Jan van Leeuwen. Nonlinear data assimilation in geosciences: an extremely efficient particle filter. Quart. J. Roy. Meteor. Soc., 136(653):1991–1999, 2010.

Jeffrey S Whitaker and Thomas M Hamill. Ensemble data assimilation without perturbed observations. Monthly Weather Review, 130(7):1913–1924, 2002.

Christopher KI Williams and Carl Edward Rasmussen. Gaussian processes for machine learning. MIT press Cambridge, MA, 2006.

Akira Moiseevich Yaglom. Correlation theory of stationary and related random functions. Springer, 1987.

Changjiang Yang, Ramani Duraiswami, Nail A Gumerov, and Larry Davis. Improved fast gauss transform and efficient kernel density estimation. In null, page 464. IEEE, 2003.

Rio Yokota, Lorena A Barba, and Matthew G Knepley. Petrbfa parallel o(n) algorithm for radial basis function interpolation with gaussians. Computer Methods in Applied Mechanics and Engineering, 199(25-28):1793–1804, 2010.

Yanqiu Zhu, Emily Liu, Rahul Mahajan, Catherine Thomas, David Groff, Paul Van Delst, Andrew Collard, Daryl Kleist, Russ Treadon, and John C Derber. All-sky microwave radiance assimilation in ncep's gsi analysis system. Mon. Wea. Rev., 144(12):4709–4735, 2016.