The ensemble Kalman filter regularized with non-parametric non-stationary spatial convolutions

M Tsyrulnikov and A Sotskiy
HydroMetCenter of Russia

March 11, 2022

1 Introduction

Modern data assimilation increasingly relies on the ensemble technique, in which the prior probability distribution of the truth is represented by a finite sample (ensemble) of pseudo-random realizations (called ensemble members). In practical applications, there are two most widely used approaches: (1) Ensemble Kalman Filter (EnKF) and (2) Ensemble Variational schemes (EnVar). In the latter, the ensemble statistics is accommodated within the variational analysis. The principal problem of the ensemble approach is that running many ensemble members is computationally expensive. In real-world very-high-dimensional problems, this means that only very small ensembles (normally, we tens of members) are affordable. As a result, such ensembles can provide the analysis (i.e., the observation update at each assimilation cycle) with only scarce information on the true prior distribution. In this situation, the sample covariance matrix is a poor estimate of the true background-error covariance matrix and thus requires a kind of *regularization* (i.e., the introduction of additional information on the prior covariances).

1.1 Ensemble Kalman filter

Bayesian data assimilation.

Ensemble (Monte-Carlo) approach. The ensemble members aim to represent the uncertainty of the estimates of the true system state.

At the forecast step EnKF, generates an ensemble of perturbed forecast realizations (ensemble members) $\mathbf{x}^{\mathrm{fc},\mu}$ (where μ labels the ensemble member).

EnKF.

Sample covariance matrix $\mathbf{S}$. For small sample (ensemble) size, $\mathbf{S}$ is a very poor estimate of the true covariance matrix $\mathbf{B}^{\mathrm{true}}$. $\mathbf{S}$ is heavily contaminated by the sampling noise and it is rank deficient. A kind of regularization of the EnKF analysis problem is needed to address these deficiencies. Normally this is done by applying an ad-hoc device

(various kinds of so-called localization are most popular). We propose to use methods of spatial statistics to regularize the problem.

...

1.2 Covariance regularization in EnKF

There exist the following main practical approaches to covariance matrix regularization.

1. The most popular approach is covariance localization (tapering), (e.g. Houtekamer and Mitchell, 1998; Furrer and Bengtsson, 2007), which reduces spurious long-distance correlations through element-wise multiplication of the sample covariance matrix by an ad-hoc analytical localization covariance matrix. This technique efficiently removes a lot of noise in the sample covariance matrix but it cannot cope with the noise at small distances. The multiplication by an ad-hoc localization function also reduces the length scale and, as a result of this, can destroy balances between different fields (**?**).

2. A similar approach is smoothing and reducing (shrinking) the Kalman gain matrix (Sætrom and Omre, 2013). This technique filters out the sampling noise by spatially smoothing the weights with which observations are impact the resulting analysis field.

3. Blending (more precisely, computing a linear combination of) sample covariances and static (time-mean) covariances helps reduce the sampling noise and is now widely used in meteorological ensemble-variational schemes (Buehner et al., 2013; Lorenc et al., 2014). In statistical literature, similar techniques are known as shrinkage estimators (Ledoit and Wolf, 2004) [1]. Sample covariances are noisy but containing useful flow-dependent "signal". Static covariances are noise-free but can be irrelevant for current weather situation. Mixing the two kinds of covariances proved to be useful (see the above references) but it is not selective: the noise in the sample covariances is reduced to same extent as the flow-dependent non-stationary signal.

4. Another approach is the spatial averaging of the covariances (that is, blending with neighboring in space covariances) (e.g. Berre and Desroziers, 2010). The technique damps the noise in sample covariances due to an increase in the effective ensemble size, but at the expense of somewhat distorting the covariances due to their spatial smoothing. The optimal spatial filtering of the covariances (Ménétrier et al., 2015) further develops this idea.

[1] The term "shrinkage" means that such estimators decrease the range of the covariance matrix eigenvalues. This is meaningful because the eigenvalues of the sample covariance matrix are known to be too dispersed, with the largest eigenvalue being too large whilst the smallest eigenvalue too small (e.g. Ledoit and Wolf, 2004, section 2.2).

5. Similar to the previous approach is the *temporal* averaging of the covariances (i.e., blending with recent past covariances). Berre et al. (2015); Bonavita et al. (2016) use ensemble members from several previous days to increase the ensemble size and Lorenc (2017) found that using time-shifted perturbations increases the effective ensemble size. Tsyrulnikov and Rakitko (2017) theoretically arrived at this technique by assuming that the true covariance matrix is an unknown random matrix with and introducing a secondary filter in which the covariances are updated. In the (Bayesian) update of the covariance matrix, the hyperprior probability distribution of the covariance matrix is inverse Wishart. Its posterior (hyperposterior) distribution is obtained by treating ensemble members are used as generalized observations on the covariance matrix.

   Tsyrulnikov and Rakitko (2019) compared the above three covariance blending techniques (that is, mixing with climatological, neighboring in space, and neighboring in time covariances) and found that their usefulness crucially depends on the degree of the spatiotemporal non-stationarity (inhomogeneity) of background errors. Time mean (static) covariances are useful under low non-stationarity, whereas the spatial and temporal covariance blending are more useful when non-stationarity is stronger. They also found (using their doubly-stochastic advection-diffusion-decay model) that the temporal covariance blending is systematically more beneficial than spatial covariance blending.

6. (Ueno and Tsuchiya, 2009) proposed to regularize the sample covariance matrix by imposing a *sparse* structure in the inverse covariance (precision) matrix. A similar approach is taken in Hou et al. (2021).

7. One more option is to adopt a parametric background-error covariance model and estimate parameters of the model from the forecast ensemble.

   This class of covariance regularization techniques includes, first, wavelet based models. In high-dimensional problems, most popular (and affordable) is the so-called wavelet-diagonal approach, in which the wavelet coefficients are postulated to be independent and variances are estimated from the ensemble, see (Fisher, 2003; Berre et al., 2015; Kasanickỳ et al., 2015). (Theoretically, an unpleasant feature of the wavelet-diagonal approach with overlapping spectral bands is its inability to represent a stationary process. The overlapping bands are needed to achieve spatial localization, see, e.g., chapter 10 in Marinucci and Peccati (2011)).)

   Second, physical-space parametric covariance models were used by Skauvold and Eidsvik (2019), who found that simple models were more useful than sophisticated models.

   The approach we propose here belongs to this category of covariance regularization techniques with the caveat that our model is nonparametric.

1.3 Our contribution

2 Spatial model

Following (Higdon et al., 1999), we rely on the *process convolution* model. In contrast to most applications of the process convolution model we do not specify a parametric model for the spatial kernel. This is motivated by the desire to allow for variable shapes of spatial covariances for a non-stationary spatial process (field). The spatial kernel is estimated "online" in a non-parametric way. To facilitate the estimation, a number of constraints on the kernel are imposed.

We consider processes defined on the two-dimensional sphere and on the circle. The spherical case is more practically relevant whereas the circular case is technically simpler. On the sphere, we use the terms stationarity and isotropy interchangeably.

2.1 General process convolution model

Let ξ be a general zero-mean linear process, that is, the process whose values are linear combinations of the white Gaussian noise $\alpha(y)$:

$$\xi(x) = \int_D w(x,y)\,\alpha(y)\,\mathrm{d}y \equiv \int_D w(x,y)\,Z(\mathrm{d}y), \tag{1}$$

where $D = \mathbb{S}^1$ or $\mathbb{S}^2$ is the domain of interest, Z is the spatial Gaussian orthogonal stochastic measure (such that the expectation $\mathbb{E}\,Z(\mathrm{d}A) = 0$, $\mathbb{E}(Z(\mathrm{d}A))^2 = |\mathrm{d}A|$, and $\mathbb{E}\,Z(\mathrm{d}A)Z(\mathrm{d}B) = 0$ whenever $\mathrm{d}A \cap \mathrm{d}B = \emptyset$), $\mathrm{d}A$ is an area element, $|\mathrm{d}A|$ its surface area, and $w(x,y)$ is a real function (called the convolution kernel or the weighting function). In theoretical statistics, processes defined by Eq.(1) are sometimes called of Karhunen class (Kakihara, 1997).

For $\mathsf{Var}\,\xi(x)$ to be finite, the kernel $w(x,y)$ is required to be square integrable w.r.t. its second argument

$$\int w(x,y)^2\,\mathrm{d}y < \infty, \tag{2}$$

Besides the technical constraint Eq.(2), we impose below four fundamental constraints on the weighting function w that will make it unique and identifiable from a realistic-size ensemble.

2.2 Space discrete process convolution model

In data assimilation, a space discrete representation of the process in question is needed. Discretizing Eq.(1) yields the spatial *moving average* model:

$$\boldsymbol{\xi} = \mathbf{W}\boldsymbol{\alpha}, \tag{3}$$

where $\mathbf{W}$ is an $n_\mathrm{x} \times n_\mathrm{x}$ matrix and the entries of the white noise vector $\boldsymbol{\alpha}$ are independent $N(0,1)$ random variables. Equation (3) implies that the covariance matrix of $\boldsymbol{\xi}$ satisfies

the "square-root" decomposition

$$\mathbf{B} = \mathbf{W}\,\mathbf{W}^{\top}. \tag{4}$$

The model Eq.(3) is capable of representing *any* covariance matrix because there is always the positive definite square root of $\mathbf{B}$, which satisfies Eq.(3). The representation Eq.(3) is, actually, "too general" as there are infinitely many such representations. Indeed, for the non-degenerate $\mathbf{B}$, any matrix $\mathbf{W}' = \mathbf{W}\mathbf{Q}$, where $\mathbf{Q}$ is an orthogonal matrix, also satisfies Eq.(3). Our goal (dictated by computational considerations) is to select a *sparse* weighting matrix $\mathbf{W}$. The general strategy is to *constrain* the space continuous model and then discretize it in space.

2.3 Convolution model with locally isotropic kernel

Given the redundancy of the class of space discrete moving average models, we aim at reducing the number of degrees of freedom of the model. This will isolate a single model among those which satisfy Eq.(3).

We begin with the space continuous model, constraining $w(x, y)$ in Eq.(1) to be of the locally isotropic form

$$w(x, y) = u(x, \rho(x, y)), \tag{5}$$

where u is the real valued function and $\rho(x, y)$ is the great-circle distance between the points x and y. Note that the restriction Eq.(5) is the **first constraint** we impose on the general process convolution model.

Substituting Eq.(5) into Eq.(1) we obtain

$$\xi(x) = \int u(x, \rho(x, y))\, \alpha(y)\, \mathrm{d}y \equiv \int u(x, \rho(x, y))\, Z(\mathrm{d}y). \tag{6}$$

Next, we develop a spectral representation of the the process in question and of its spatial covariances.

On $\mathbb{S}^2$, we employ, first, the spectral representation of the real valued band limited white noise:

$$\alpha(x) = \sum_{\ell=0}^{L} \sum_{m=-\ell}^{\ell} \widetilde{\alpha}_{\ell m}\, Y_{\ell m}(x). \tag{7}$$

Here L is the maximal total wavenumber, x stands for the pair (θ, ϕ) with θ being the co-latitude and ϕ the longitude, and $Y_{\ell m}$ is the spherical harmonic (normalized such that $\int_{\mathbb{S}^2} |Y_{\ell m}(x)|^2\, \mathrm{d}x = 1$). It can be seen that $\widetilde{\alpha}_{\ell m}$ are mutually uncorrelated complex-valued random Fourier coefficients with $\mathbb{E}\,\widetilde{\alpha}_{\ell m} = 0$ and $\mathsf{Var}\,\widetilde{\alpha}_{\ell m} = 1$. More specifically, $\widetilde{\alpha}_{l0}$ are real valued and all the other $\widetilde{\alpha}_{\ell m}$ are complex circularly symmetric random variables (e.g. Searle and Khuri, 2017, section 9.5) such that (since $\alpha(x)$ is real valued) $\widetilde{\alpha}_{l,-m} = \widetilde{\alpha}^{*}_{\ell m}$ (where * denotes complex conjugation).

We perform the spectral (Fourier-Legendre) expansion of $u(x, \rho)$ with x being fixed:

$$u(x, \rho) = \frac{1}{4\pi} \sum_{\ell=0}^{L} (2\ell + 1) \sigma_{\ell}(x) P_{\ell}(\cos \rho), \tag{8}$$

where $P_\ell(\cos\rho)$ is the Legendre polynomial (normalized such that $P_\ell(1) = 1$). Substituting $\rho = \rho(x, y)$ into Eq.(8) and applying the addition theorem for spherical harmonics, we obtain

$$u(x, \rho(x, y)) = \sum_{\ell=0}^{L} \sigma_\ell(x) \sum_{m=-\ell}^{\ell} Y_{\ell m}(x)\, Y^*_{\ell m}(y), \tag{9}$$

Finally, we substitute Eqs.(7) and (9) into Eq.(6). Utilizing orthonormality of spherical harmonics, we obtain:

$$\xi(x) = \sum_{\ell=0}^{L} \sum_{m=-\ell}^{\ell} \sigma_\ell(x)\, \widetilde{\alpha}_{\ell m}\, Y_{\ell m}(x). \tag{10}$$

Note that from Eq.(8) it follows that $\sigma_\ell(x)$ are real valued. We call $\sigma_\ell(x)$ the spectral functions and $f_\ell(x) := \sigma^2_\ell(x)$ the *local spatial spectrum*.

Note that the model Eq.(10) becomes the stationary (isotropic) random field model if $u(x, \rho) = u(\rho)$ or, equivalently, all spectral functions $\sigma_\ell(x)$ do not depend on x, see Eq.(59).

The LSM's spatial covariances can be readily obtained from Eq.(10) by taking into account that all $\widetilde{\alpha}_{\ell m}$ are mutually uncorrelated and applying the addition theorem for spherical harmonics:

$$B(x, x') := \mathbb{E}\, \xi(x)\xi(x') = \frac{1}{4\pi} \sum_{\ell=0}^{L} (2\ell + 1)\, \sigma_\ell(x)\, \sigma_\ell(x')\, P_\ell(\cos\rho(x, x')). \tag{11}$$

In particular,

$$\mathsf{Var}\, \xi(x) = \frac{1}{4\pi} \sum_{\ell} (2\ell + 1)\sigma^2_\ell(x) \equiv \frac{1}{4\pi} \sum_{\ell=0}^{L} \sum_{m=-\ell}^{\ell} f_\ell(x) \tag{12}$$

so that the pointwise process variance $\mathsf{Var}\, \xi(x)$ is proportional to the sum of the spectral variances $f_\ell(x)$ — exactly as in the stationary case — this is why we call $f_\ell(x)$ the local spectrum and the model Eq.(10) the Local Spectrum Model (LSM). The formulation of the LSM on the circle is outlined in appendix A.

2.4 Non-negative local spectrum constraint

In Appendix B we consider the stationary process ξ_{statio} defined by Eq.(59). We examine a kernel $u(\rho)$ that yields the fixed spatial covariance function of ξ_{statio}. This kernel is not unique. In an attempt to select a unique kernel we impose a *spatial localization* requirement. We show that there are three solutions to this optimization problem, all of them having the same shape of the kernel. Among these three equivalent solutions we select the non-negative definite one. A similar result is valid for the process defined on the circle. So, for a stationary process both on the sphere and on the circle, defined by the convolution $\xi = u * \alpha$ (where α the white noise) and having a given spatial covariance function, the most spatially localized kernel can be considered a non-negative definite function of distance.

Motivated by this result and acknowledging that spatial localization is essential for fast computations, we postulate that in the non-stationary case, for any x, the kernel $u(x, \rho)$ is a *non-negative* definite function of the distance ρ. As a consequence, $\sigma_\ell(x) \geq 0$ both in the spherical and in the circular case. This constitutes our **second constraint** imposed on the general process convolution model.

2.5 Local stationarity

The notion of local stationarity has been defined differently by different authors (most often for processes on the real line). The general idea is that a locally stationary process can be approximated by a stationary process locally, i.e., in a vicinity of any point in time (Mallat et al., 1998). Starting from Dahlhaus (1997), the common approach is to use the "infill" asymptotics, which considers the process $\xi(t)$ in rescaled time t/T with $T \to \infty$, e.g., Dahlhaus (2012). With this approach, effectively, just one segment of a non-stationary process is studied in an increasingly sharper detail. On a compact manifold G like the circle or the two-dimensional sphere, this rescaling cannot be used because of their compactness.

To overcome this obstacle, we assume that the process in question, $\xi(x)$, indexed on the sphere or on the circle and satisfying Eqs.(10) or (57), respectively, is conditionally Gaussian given the set of spectral functions $\sigma_\ell(x)$, *which themselves are random processes*. Note that randomness of $\sigma_\ell(x)$ is equivalent to the kernel $u(x, \rho)$ being a function-space valued random process as a function of its first argument x. We are interested in the distribution of the process $\xi(x)$ given the set of processes $\{\sigma_\ell(x)\}_{\ell=0}^{L}$. We denote this conditioning as $\xi \,|\, \sigma$. The processes $\sigma_\ell(x)$ can be non-stationary as functions of x but here we assume that any $\sigma_\ell(x)$ is *stationary*. We also postulate that the processes $\sigma_\ell(x)$ are smooth: mean-square differentiable with differentiable sample paths.

We will need length scales Λ_ℓ of the processes $\sigma_\ell(x)$, which we define as micro-scales from

$$\mathbb{E}\left(\frac{\partial \sigma_\ell(x)}{\partial x}\right)^2 =: \frac{\mathsf{Var}\,\sigma_\ell}{\Lambda_\ell^2}, \tag{13}$$

where $\frac{\partial}{\partial x}$ is the gradient. We call $\Lambda := \inf \Lambda_\ell$ the *non-stationarity length scale*. We also define a measure, Σ, of spatial variability in the variances of the processes $\sigma_\ell(x)$ as follows

$$\Sigma^2 := \sum \frac{2\ell + 1}{4\pi} \mathsf{Var}\,\sigma_\ell. \tag{14}$$

We call Σ the *non-stationarity strength*.

Now we are in a position to consider the *limit of weak non-stationarity*:

$$\Lambda \to \infty, \tag{15}$$

$$\frac{\Sigma}{\Lambda} \to 0, \tag{16}$$

$$\mathbb{E}\,\sigma_\ell^2 = \mathsf{const}. \tag{17}$$

Here the last equation implies that the mean (i.e., averaged over the randomness of the processes $\sigma_\ell(x)$) spectrum of ξ, and thus its mean variance and mean spatial correlation function are kept constant. It is the distribution of the processes $\sigma_\ell(x)$ that is being changed in this limit: their spectra become increasingly localized at small wavenumbers and their correlation functions become increasingly flat.

We show that the LSM process $\xi(x)$ is *locally stationary* in the sense that for any spatial point x_0, there is a *stationary* process $\zeta(x; x_0)$ such that $\mathbb{E}\,\mathbb{E}[(\xi(x) - \zeta(x; x_0))^2 \,|\, \sigma] \to 0$ uniformly in x, where the limit is understood in the above sense.

With the process in question, $\xi(x)$ defined by Eq.(10), for any x_0, we define the stationary process ζ as

$$\zeta(x; x_0) := \sum_{\ell=0}^{L} \sum_{m=-\ell}^{\ell} \sigma_\ell(x_0)\, \widetilde{\alpha}_{\ell m}\, Y_{\ell m}(x), \tag{18}$$

where $\widetilde{\alpha}_{\ell m}$ is the same realization of the driving white noise as in Eq.(10). Then

$$\mathbb{E}[(\xi(x) - \zeta(x; x_0))^2 \,|\, \sigma] = \sum \frac{2\ell + 1}{4\pi} (\sigma_\ell(x) - \sigma_\ell(x_0))^2. \tag{19}$$

By the mean value theorem, there is a point $\bar{x}$ on the great circle connecting the two points, x and x_0, such that $\sigma_\ell(x) - \sigma_\ell(x_0) = \frac{\mathrm{d}\sigma_\ell}{\mathrm{d}\rho}\, \rho$, where ρ is the distance between x and x_0 and the derivative is along the great circle and evaluated at $\bar{x}$. Since $|\frac{\mathrm{d}\sigma_\ell}{\mathrm{d}\rho}| \le |\frac{\partial \sigma_\ell(\bar{x})}{\partial x}|$, we obtain from Eq.(19) using Eq.(13) and the inequality $\Lambda_\ell \ge \Lambda$:

$$\mathbb{E}\,\mathbb{E}[(\xi(x) - \zeta(x; x_0))^2 \,|\, \sigma] \le \sum \frac{2\ell + 1}{4\pi}\, \mathbb{E}\left(\frac{\partial \sigma_\ell(\bar{x})}{\partial x}\right)^2 \rho^2 \le$$
$$\rho^2 \sum \frac{2\ell + 1}{4\pi} \frac{\mathsf{Var}\,\sigma_\ell}{\Lambda_\ell^2} \le \frac{\rho^2}{\Lambda^2} \sum \frac{2\ell + 1}{4\pi} \mathsf{Var}\,\sigma_\ell = \left(\frac{\rho\Sigma}{\Lambda}\right)^2 \to 0. \tag{20}$$

The limit here is due to the assumptions Eqs.(15) and (16). So, indeed, the LSM process $\xi(x)$ is locally stationary. In practical situations, we assume that a process can be modeled by the Locally Stationary Model if its non-stationarity length scale is significantly greater than its typical length scale and if the variability in the local process variance is smaller than or comparable to the typical process variance.

The assumption of smooth and uniformly slow variation of the processes $\sigma_\ell(x)$, and hence of the local spectra $f_\ell(x)$, with x is our **third constraint**. It implies that the kernel $u(x, \rho)$ slowly varies with location x (as compared with its variation with the distance ρ) and, actually, justifies the term "local spectrum" introduced by Priestley (1965, 1988), who called it evolutionary spectrum in the time series context.

2.6 Smoothness of local spectra

Studies of real-world spatio-temporal processes showed that spatial (and temporal) spectra in geosciences, say in meteorology, are often quite smooth, exhibiting typically a power-law behavior at large wavenumbers, e.g., Gage and Nastrom (1986), Trenberth and Solomon

(1993). For this reason and with the intention to regularize the LSM by further reducing its effective number of degrees of freedom, we postulate that the spatial spectra $f_\ell(x)$ are smooth (and, tentatively, monotone) functions of the wavenumber ℓ — this is our **fourth constraint**.

2.7 Summary of constraints

We have introduced four constraints:

1. The convolution kernel has the form $w(x, y) = u(x, \rho(x, y))$.
2. The locally isotropic kernel $u(x, \rho)$ is a non-negative definite function of the distance ρ.
3. The kernel $u(x, \rho)$ is a smooth function of location x. More specifically, the spectral transform of $u(x, \rho)$ w.r.t. the distance ρ, that is, the spectral functions $\sigma_\ell(x)$ are smooth functions of the spatial location x.
4. The spectral functions $\sigma_\ell(x)$ or, equivalently, the local spatial spectra $f_\ell(x)$ are smooth functions of the wavenumber ℓ. Requiring the spectra to be monotone may further regularize the problem.

The specific formulation of constraints 3 and 4 in this list will be given below when we introduce an estimator of the LSM.

2.8 Identifiability

The question here is whether the kernel $u(x, \rho)$ that satisfies the above four constraints can be uniquely determined from the output of the LSM, that is, from its non-stationary covariances $B(x, x')$? In Appendix C we prove that, with some technical assumptions, the answer is yes.

3 Locally stationary ensemble filter (LSEF)

3.1 State variables

In the time discrete cycling scheme, LSEF provides estimates of the main state vector $\mathbf{x}$ as well as the set of *local spatial log-spectra*, $\varphi_{\ell i} = \log(\sigma_\ell(x_i))$ collected in the matrix $\mathbf{\Phi}$. The reason of treating the local spectra on the log scale is that the spectra are highly non-Gaussian distributed. We assume that their logarithm can be modeled as Gaussian. One assimilation cycle consists of the forecast step that advances current information on the state variables forward in time and the analysis step that updates the state variables using current data.

The forecast-step variables are denoted by the superscript $^{\mathrm{fc}}$, the analysis-step variables by the superscript $^{\mathrm{an}}$.

3.2 Forecast step

At the forecast step, LSEF, like EnKF, generates an ensemble of perturbed forecast realizations (ensemble members) $\mathbf{x}^{\mathrm{fc},\mu}$ (where $\mu = 1, \ldots, M$ labels the ensemble member). These ensemble members are produced by a stochastic model of truth, which starts from the previous-cycle analysis ensemble members. We define the *forecast ensemble perturbations* as $\mathbf{e}^{\mu} = \mathbf{x}^{\mathrm{fc},\mu} - \mathbf{x}^{\mathrm{fc}}$, where $\mathbf{x}^{\mathrm{fc}}$ is a central forecast. The latter is set to be the deterministic (non-perturbed) forecast if the forecast model is linear and the ensemble mean forecast if the forecast model is nonlinear. In either case, $\mathbf{e}^{\mu}$ are assumed to be mutually independent equally distributed zero mean Gaussian random vectors.

Besides, LSEF also propagates the analysis estimate of the local spatial log-spectra $\boldsymbol{\Phi}$. For the auxiliary state variable $\boldsymbol{\Phi}$, LSEF does not generate any ensemble. Neither it attempts to represent the uncertainty in this time propagation of $\boldsymbol{\Phi}$, is postulated to be constant in time. Instead, the analysis point estimate (deterministic analysis) $\boldsymbol{\Phi}^{\mathrm{an}}_{k-1}$ at cycle $k-1$ is just transferred to the next cycle and combined with the time mean log-spectra $\bar{\boldsymbol{\Phi}}$ in specifying a kind of regression-to-the-mean forecast log-spectra $\boldsymbol{\Phi}^{\mathrm{fc}}_{k} = w\boldsymbol{\Phi}^{\mathrm{an}}_{k-1} + (1-w)\bar{\boldsymbol{\Phi}}$ (where $0 \le w \le 1$ is the memory parameter). These approximations in treating $\boldsymbol{\Phi}^{\mathrm{fc}}$ at the LSEF forecast step are motivated by the absence of a forecast model for the local spectra and by the *simplicity requirement*: the technique should be applicable to real-world multivariate ultrahigh-dimensional problems.

3.3 Analysis step

The goal of the analysis is to update the joint distribution of the state and the local spectra computing the posterior $p_{\mathrm{post}}(\mathbf{x}, \boldsymbol{\Phi})$.

The central assumption in the LSEF analysis is that the LSM spatial model adequately represents the prior distribution of the state $\mathbf{x}$ given the forecast $\mathbf{x}^{\mathrm{fc}}$ and the local spectra $\boldsymbol{\Phi}$, which themselves are random and subject to update using the prior ensemble.

3.3.1 Prior

The prior density is $p(\mathbf{x}, \boldsymbol{\Phi}) = p(\boldsymbol{\Phi})\, p(\mathbf{x} \,|\, \boldsymbol{\Phi})$.

To make the numerical scheme tractable in a high-dimensional setting, we we confine ourselves to updating the local spectra pointwise in physical space assuming that $p(\boldsymbol{\Phi}) = \prod_x p(\boldsymbol{\varphi}(x))$, where x runs over the spatial grid points. The marginal prior density of the log-spectrum $p(\boldsymbol{\varphi}(x))$ at a given spatial grid point x, is multivariate Gaussian with the mean $\boldsymbol{\varphi}^{\mathrm{fc}}$ and covariance matrix $\sigma^2_{\varphi}\mathbf{I}$ (where σ^2_{φ} is a hyperparameter and $\mathbf{I}$ is the identity matrix):

$$p(\boldsymbol{\varphi}(x)) \propto \exp\left(-\frac{(\boldsymbol{\varphi}(x) - \boldsymbol{\varphi}^{\mathrm{fc}}(x))^2}{2\sigma^2_{\varphi}}\right). \tag{21}$$

The conditional prior density of the state given the local spectra, $p(\mathbf{x} \,|\, \boldsymbol{\Phi})$, is also multi-

variate Gaussian:

$$p(\mathbf{x} \,|\, \mathbf{\Phi}) \propto \frac{1}{|\det \mathbf{B}|} \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{x}^{\mathrm{fc}})^\top \mathbf{B}^{-1}(\mathbf{x}-\mathbf{x}^{\mathrm{fc}})\right), \tag{22}$$

where the covariance matrix $\mathbf{B} = \mathbf{B}(\mathbf{\Phi})$ obeys the LSM, i.e., its entries are determined by the local spectra as in Eqs.(11) or (58).

3.3.2 Likelihoods

The *data* available to update the prior distribution are: (i) the prior ensemble perturbations (deviations from the central forecast) $\mathbf{e}^\mu$ ($\mu = 1, \ldots, M$ and M is the ensemble size) collected as columns in the matrix $\mathbf{E}$ and (ii) the observations $\mathbf{y}$.

The *ensemble likelihood* $p(\mathbf{E}|\mathbf{\Phi})$ readily follows from the assumption that $\mathbf{e}^\mu$ are draws from the same distribution as the forecast error $\boldsymbol{\xi} = \mathbf{x} - \mathbf{x}^{\mathrm{fc}}$:

$$p(\mathbf{E}|\mathbf{\Phi}) \propto \frac{1}{|\det \mathbf{B}|^M} \prod_\mu \exp\left(-\frac{1}{2}(\mathbf{e}^\mu)^\top \mathbf{B}^{-1}\mathbf{e}^\mu\right), \tag{23}$$

where, again, $\mathbf{B} = \mathbf{B}(\mathbf{\Phi})$.

The *observation likelihood* $p(\mathbf{y}|\mathbf{x})$ follows from the *observation equation*

$$\mathbf{y} = \mathbf{H}^{\mathrm{obs}}\mathbf{x} + \boldsymbol{\eta}, \tag{24}$$

where $\mathbf{H}^{\mathrm{obs}}$ is the observation operator (assumed, for simplicity, to be linear) and $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$ is a Gaussian vector of observation errors (with $\mathbf{R}$ being its covariance matrix):

$$p(y \,|\, \mathbf{x}) \propto \exp\left(-\frac{1}{2}(\mathbf{y}-\mathbf{H}^{\mathrm{obs}}\mathbf{x})^\top \mathbf{R}^{-1}(\mathbf{y}-\mathbf{H}^{\mathrm{obs}}\mathbf{x})\right). \tag{25}$$

3.3.3 Posterior

As shown in Appendix D, the computation of the posterior (the analysis) can be split into two sub-analyses: first, the local spatial log-spectra $\mathbf{\Phi}$ are updated (section 4) and second, the state is updated in the traditional Kalman-filter analysis using the updated spatial spectra (section 5).

Analysis ensemble. Discuss Katzfuss et al. (2020): "Tsyrulnikov and Rakitko (2017) seem to ignore the fact that forecast independence does not hold in their model."

"Maybe, you mean that we implicitly used the forecast independence at some point, but I cannot find out where.."

"Regarding the implicit assumption of forecast independence, we were thinking about what is hinted at in the second-to-last paragraph in our Section 4.1.1. In your setting, to get the filtering distribution of the state at time t, I believe one would need to integrate over the distribution of the covariance matrix at time t-1 given the data up to time t (not just t-1), for example."

4 Updating local spectra

The goal is to estimate the set of local spectra $f_\ell(x)$ (or, equivalently, $\sigma_\ell(x)$) at all spatial grid points x.

4.1 Approach

In the stationary case, it is straightforward to derive from Eq.(59) that the likelihood of the spectrum f_ℓ depends on the sample (the ensemble of M independent realizations of $\xi^\mu_{\text{statio}}(x)$, where $\mu = 1, \ldots, M$) trough $\widehat{v}_\ell := \frac{1}{M} \sum_{\mu=1}^{M} \sum_{m=-\ell}^{\ell} |\widetilde{\xi}_{\ell m}|^2$. (Note that with $M = 1$, $\widehat{v}_\ell$ is a spherical analog of what is known in the time series context as the *periodogram*.) Therefore, the vector of the sample mean spectral variances $\widehat{v}_\ell$ (with $\ell = 0, 1, \ldots, L$) is a sufficient statistics for the spectrum f_ℓ. This implies that no information on the spectrum is lost if we switch from the raw ensemble to the set of $\widehat{v}_\ell$.

In the locally stationary case, however, relying on the spectral variances for individual wavenumbers is not a good idea because the spatial filter that isolates individual spectral components has non-local response functions. Say, on $\mathbb{S}^1$, the spectral transfer function that is equal to one at the wavenumbers $\pm\ell_0$ and zero otherwise corresponds to the impulse response function $2\cos(\ell_0 x)$. This is clearly not acceptable in the non-stationary case. To localize the response function and thus make the technique applicable to the locally stationary case, we specify broader filter spectral transfer functions, which have narrower (i.e., localized) response functions, and use ensemble variances of the outputs of these *bandpass filters* to estimate the local spectra. Broadening the filter's transfer functions reduces the spectral resolution of the estimator, but according to our fourth constraint (section 2.6), the spatial spectra are smooth, so a reduced spectral resolution should not be a problem. On the other hand, with broader spectral transfer functions, the approximate filtering Eq.(26) we use to estimate the spectrum becomes more accurate (see Appendix E).

Technically, we perform bandpass filtering of the non-stationary process that satisfies Eq.(10) using J filters $\mathcal{H}_j$, where $j = 1, \ldots, J$. The filters are isotropic, with overlapping spectral transfer functions $H_j(\ell)$. To address the non-stationarity of the process in question, the filters' impulse response functions $h_j(x)$, that is, the inverse Fourier-Legendre transforms of $H_j(\ell)$, must be localized in space (note that the filtered processes are convolutions of $\xi(x)$ with $h_j(x)$). To ensure this, we require $H_j(\ell)$ to be smooth functions of the wavenumber ℓ. The bandpass filtered processes $\xi_{(j)}(x)$ (at any spatial grid point x separately) are used to update the local spectra $f_\ell(x)$.

Our approach can be regarded as a modification of the technique by Priestley (1965), who applied a lowpass filter.... Marinucci.

4.2 Bandpass filters

As shown in Appendix E, applying the filter $\mathcal{H}_j$ to the field $\xi(\theta,\phi)$ that obeys the LSM, Eq.(10) yields, approximately

$$\xi_{(j)}(x) \approx \sum_{\ell=0}^{L} H_j(\ell)\,\widetilde{\sigma}_\ell(x) \sum_{m=-\ell}^{\ell} \widetilde{\alpha}_{\ell m}\, Y_{\ell m}(x) \tag{26}$$

so that the variances $v_{(j)}(x)$ of the bandpass filtered processes $\xi_{(j)}(x)$ are related to the local variance spectrum $f_\ell(x)$ as follows:

$$v_{(j)}(x) \approx \frac{1}{4\pi} \sum_{\ell=0}^{L} (2\ell+1) H_j^2(\ell)\, f_\ell(x). \tag{27}$$

On the other hand, having an ensemble (i.e., a sample) of M independent fields (ensemble members) taken from the same probability distribution as the field in question $\xi(x)$, we estimate the variances of the processes $\xi_{(j)}(x)$ as their sample (ensemble) variances at each x independently:

$$\widehat{v}_{(j)}(x) := \widehat{\mathsf{Var}}\,\xi_{(j)}(x), \tag{28}$$

where $\widehat{\mathsf{Var}}$ stands for the sample variance operator. As discussed above, the vector $\widehat{\mathbf{v}}(x) = (\widehat{v}_{(1)}(x), \ldots, \widehat{v}_{(1)}(x))^\top$ can be regarded as an approximately sufficient statistic. Accepting this assumption, confining, for computational reasons, to a pointwise estimation formulation, and switching to log-spectra $\varphi_\ell(x) = \log(\sigma_\ell(x))$, we write a Bayesian estimator of the local spectrum φ_ℓ at the grid point x as a maximizer of the posterior density

$$p(\varphi_\ell \,|\, \widehat{\mathbf{v}}) \propto p(\varphi_\ell)\, p(\widehat{\mathbf{v}} \,|\, \varphi_\ell). \tag{29}$$

Technically, we used the filters' spectral transfer functions of the form

$$H_j(\ell) = \exp\left(- \left| \frac{\ell - \ell_j^{\mathrm{c}}}{d_j} \right|^q \right), \tag{30}$$

where ℓ_j^{c} is the central wavenumber of the jth waveband, d_j is its half-width, and q is the shape parameter. We took q in the range from 2 to 3. Both ℓ_j^{c} and d_j were taken linearly growing with the log wavenumber.

4.3 Preliminary estimation of local spectrum

Since $\widehat{v}_{(j)}(x)$ is an unbiased estimate of $v_{(j)}(x)$, we obtain from Eqs.(27) and (28):

$$\widehat{v}_{(j)} = \frac{1}{4\pi} \sum_{\ell=0}^{L} (2\ell+1) H_j^2(\ell) f_\ell + \text{error}. \tag{31}$$

Here error stands for both methodological error (involved in Eq.(27) and discussed in Appendix E) and the sampling error. We have dropped the dependencies on the spatial

grid point x because the estimation of the local spectrum is performed independently for different x.

Denoting $\frac{1}{4\pi}(2\ell+1)H_j^2(\ell) =: \omega_{jl}$, we rewrite Eq.(31) in the vector-matrix form as

$$\mathbf{\Omega}\,\mathbf{f} = \widehat{\mathbf{v}}. \tag{32}$$

Here $\widehat{\mathbf{v}}$ is a length-J vector, $\mathbf{f}$ is an N-vector ($N \equiv L+1$ on the sphere and $N \equiv 2L$ on the circle), and $\mathbf{\Omega}$ is a $J \times N$ matrix.

Equation (32) constitutes the standard linear inverse problem, in which $\widehat{\mathbf{v}}$ is the data we have at our disposal (from the ensemble) and $\mathbf{\Omega}$ is the known matrix that relates the unknown vector of spectral variances $\mathbf{f}$ to the data $\widehat{\mathbf{v}}$. A reasonable way to solve a linear inverse problem such as Eq.(32) is to seek the *minimal-norm least-squares* solution, that is, to use pseudo inversion:

$$\mathbf{f}^+ = \mathbf{\Omega}^+\,\widehat{\mathbf{v}}, \tag{33}$$

where $\mathbf{\Omega}^+$ is the Moore-Penrose pseudo-inverse matrix. The pseudo inverse solution effectively dampens noise in the solution because the minimal-norm solution has zero projection on the null space of $\mathbf{\Omega}$ (where a solution would contain only noise and no signal). However, it does not respect other constraints we wish to impose on the solution: first of all, (i) non-negativity and (ii) prior information. Besides, the pseudo inverse solution disregards (iii) monotonicity, (iv) resemblance to the typical shape of the spectrum, and (v) smoothness of the spectrum. The simplest way to impose these four constraints is to fit a parametric model to $\{f_\ell^+\}_{\ell=0}^L \equiv \mathbf{f}^+$.

We chose to fit a two-parameter scale-magnitude model: $g_\ell \approx A \cdot g(\ell/a)$, where a is the scale parameter (a scalar), A is the magnitude parameter (a scalar), and g is a function estimated from an archive of ensembles as the time mean stationary (isotropic) spectrum. The fitting was done using the method of moments. Specifically equating the first and second moments of the parametric model to their empirical counterparts (computed using $\mathbf{f}^+$) and replacing sums over l with integrals we easily obtain two easily solvable linear algebraic equations for A and a.

The procedure presented in this section is very fast and effective but it can struggle with situations where the shape of the local spectrum significantly differs from the time-mean spectrum. To cope with this problem, a more general approach is described next.

4.4 Non-parametric Bayesian solution

We regard the local spectrum $\mathbf{f}(x) = (f_0(x), f_1(x), \ldots, f_L(x))$ as a random vector (at each grid point x), specify its prior distribution, formulate its likelihood given the data (the ensemble of bandpass filtered fields), and describe a numerical scheme aimed at the maximization of the posterior density.

On the sphere, the filtered processes are real valued and therefore fully characterized by their covariance matrix. On the circle, on the contrary, the filtered processes are, in

general, complex valued so that a different (and slightly more complex) treatment of the likelihood is needed there, see below.

4.4.1 Prior

We place a prior on the *log-spectrum* $\varphi_\ell = \log f_\ell$. We postulate that φ_ℓ is a stationary Gaussian process of the log-wavenumber variable $s_\ell = \log(l + l_0)$ (where l_0 is introduced to permit the treatment of $l = 0$) so that

$$\varphi(s) \sim GP(s; \bar{\varphi}(s), K). \tag{34}$$

Here $\bar{\varphi}(s)$ is the mean function and K is a covariance kernel. We assume that $\bar{\varphi}(s)$ is known, say, from fitting a stationary model to an archive of data. Knowing $\bar{\varphi}(s)$, we switch from $\varphi(s)$ to the zero mean stationary process

$$\delta(s) := \varphi(s) - \bar{\varphi}(s).$$

As for the kernel K, we specify it implicitly by, first, penalizing deviations from $\bar{\varphi}(s)$, and second, promoting *smoothness* of the spectrum. The resulting minus-log-prior is the (quadratic) function $\mathcal{L}^{\text{prior}}$ that consists of the two terms,

$$\mathcal{L}^{\text{prior}}(\boldsymbol{\delta}) = \mathcal{L}^{\text{prior}}_{\text{bckg}}(\boldsymbol{\delta}) + \mathcal{L}^{\text{prior}}_{\text{smoo}}(\boldsymbol{\delta}), \tag{35}$$

where $\mathcal{L}^{\text{prior}}_{\text{bckg}}(\boldsymbol{\delta})$ is the "background constraint" and $\mathcal{L}^{\text{prior}}_{\text{smoo}}(\boldsymbol{\delta})$ is the smoothness constraint, both defined just below.

We define the "background constraint" in the following simplest way:

$$\mathcal{L}^{\text{prior}}_{\text{bckg}}(\boldsymbol{\delta}) = \frac{1}{2}\,(\boldsymbol{\delta}, \boldsymbol{\delta}), \tag{36}$$

where the inner product $(.,.)$ is defined as

$$(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2) = \sum_\ell \delta_1(s_\ell)\delta_2(s_\ell)\,\Delta^{\text{c}}_\ell \tag{37}$$

(with Δ^{c}_ℓ denoting the grid-cell size: $\Delta^{\text{c}}_\ell = (s_{l+1} - s_{l-1})/2$, where $s_{-1} = s_0$ and $s_{L+1} = s_\ell$) to be consistent with the continuous-s inner product $\int \delta_1(s)\delta_2(s)\,\mathrm{d}s$.

Thus,

$$\boxed{\mathcal{L}^{\text{prior}}_{\text{bckg}}(\boldsymbol{\delta}) = \frac{1}{2}\sum_{\ell=0}^{L} \Delta^{\text{c}}_\ell(\delta_\ell - \bar{\delta}_\ell)^2.} \tag{38}$$

The smoothness constraint is defined as a discrete analog of the positive definite functional that penalizes the magnitudes of both first and second derivatives:

$$\frac{w_{\text{s1}}}{2}\int (\mathcal{D}\delta(s))^2\,\mathrm{d}s + \frac{w_{\text{s2}}}{2}\int (\mathcal{D}^2\delta(s))^2\,\mathrm{d}s \equiv \frac{w_{\text{s1}}}{2}\,(\mathcal{D}\delta(s), \mathcal{D}\delta(s)) + \frac{w_{\text{s2}}}{2}\,(\mathcal{D}^2\delta(s), \mathcal{D}^2\delta(s)), \tag{39}$$

where $w_{\rm s1}$ and $w_{\rm s2}$ are the tunable weights and $\mathcal{D}$ is the first-order differentiation operator. We specify

$$\mathcal{L}^{\rm prior}_{\rm smoo}(\boldsymbol{\delta}) = \frac{w_{\rm s1}}{2}\,(\mathbf{D}_1\boldsymbol{\delta}, \mathbf{D}_1\boldsymbol{\delta}) + \frac{w_{\rm s2}}{2}\,(\mathbf{D}_2\boldsymbol{\delta}, \mathbf{D}_2\boldsymbol{\delta}), \tag{40}$$

where $\mathbf{D}_1$ and $\mathbf{D}_2$ are finite difference analogs of $\mathcal{D}$ and $\mathcal{D}^2$, respectively. With $(\mathbf{D}\boldsymbol{\delta})_\ell = \frac{\delta_{l+1}-\delta_{l-1}}{2\Delta^{\rm c}_\ell}$ and $(\mathbf{D}^2\boldsymbol{\delta})_\ell = \frac{\delta_{l-1}-2\delta_l+\delta_{l+1}}{(\Delta^{\rm c}_\ell)^2}$ and using Eq.(37), we have

$$\boxed{\mathcal{L}^{\rm prior}_{\rm smoo}(\boldsymbol{\delta}) = \frac{w_{\rm s1}}{2}\sum_{\ell=0}^{L-1}\frac{(\delta_{l+1}-\delta_{l-1})^2}{2\Delta^{\rm c}_\ell} + \frac{w_{\rm s2}}{2}\sum_{\ell=0}^{L-1}\frac{(\delta_{l-1}-2\delta_l+\delta_{l+1})^2}{(\Delta^{\rm c}_\ell)^3}.} \tag{41}$$

Remark 1. It can be seen that the relations between $w_{\rm s1}$, $w_{\rm s2}$, and unity determine the effective length scale of the prior Gaussian process $\delta(s)$.

Remark 2. It is not hard to see that more a stronger penalty on the second derivative, i.e., a larger $w_{\rm s2}$ leads to a smoother behavior of the kernel K near the origin.

Remark 3. The largest scales appear to be the most noisy in the ensemble statistics, so, technically, to avoid unrealistic behaviour of the estimated spectra at low wavenumbers, it might be reasonable to set the constant ℓ_0 significantly larger than 1 (say, 5–10).

4.4.2 Likelihood: $\mathbb{S}^2$

4.4.3 Likelihood: $\mathbb{S}^1$

4.4.4 Posterior

The minus log-posterior (loss function) $\mathcal{L}(\boldsymbol{\delta})$ is, obviously, the sum of the above four components, two from the prior, $\mathcal{L}^{\rm prior}_{\rm bckg}(\boldsymbol{\delta})$ and $\mathcal{L}^{\rm prior}_{\rm smoo}(\boldsymbol{\delta})$, and two from the likelihood, $\mathcal{L}^{\rm lik}_{\rm det}(\boldsymbol{\delta})$ and $\mathcal{L}^{\rm lik}_{\rm tr}(\boldsymbol{\delta})$. In this application, we choose to seek the *mode* of the posterior density. Maximizing the posterior is equivalent to minimizing the minus log-posterior, so we have to solve the optimization problem

$$\mathcal{L}(\boldsymbol{\delta}) = \mathcal{L}^{\rm prior}_{\rm bckg}(\boldsymbol{\delta}) + \mathcal{L}^{\rm prior}_{\rm smoo}(\boldsymbol{\delta}) + \mathcal{L}^{\rm lik}_{\rm det}(\boldsymbol{\delta}) + \mathcal{L}^{\rm lik}_{\rm tr}(\boldsymbol{\delta}) \to \min\,. \tag{42}$$

4.4.5 Numerical solution

4.5 Neural network-based solution

4.5.1 Approach

The estimator is applied pointwise on the physical domain (that is, for any spatial grid point x independently), so that on input we have the band variances $\widehat{v}_{(j)} = \mathsf{Var}\,\xi_{(j)}(x)$ (for $j = 1, \ldots, J$) and on output we have the local spectrum $f_\ell(x)$ (for $\ell = 0, \ldots, L$ on the sphere and $\ell = -L, \ldots, L$ on the circle).

4.5.2 Architecture of the neural network

Feed-forward.

Depth, width.

Activation function. ...

4.5.3 Learning

Learning data were generated by the models on $\mathbb{S}^1$ and $\mathbb{S}^2$ described in section 6.

4.5.4 Loss function

5 Analysis of the state

The goal

We are interested in cyclic (intermittent) data assimilation with alternating time update (forecast) and observation update (analysis) steps. Let the data assimilation technique involve maintaining an *ensemble* that represents the uncertainties in resulting state estimates. Then , at each analysis step, we, first, estimate $\mathbf{W}$ using the ensemble, and second, estimate the system state using current observations.

According to Equation (45) implies that the non-stationary space-discrete random vector $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_{n_\mathrm{x}})$ satisfies

$$\boldsymbol{\xi} = \mathbf{W}\boldsymbol{\alpha}, \tag{43}$$

where $\boldsymbol{\alpha} \sim \mathbf{N}(\mathbf{0}, \mathbf{I})$ and the entries of the weighting matrix $\mathbf{W}$ are

$$(\mathbf{W})_{ij} := w_{ij} \tag{44}$$

are defined by the space discrete equivalent of Eq.(6),

$$\xi_i = \sum_j u(x_i, \rho(x_i, y_j))\, Z(\Delta y_j) = \sum_j u(x_i, \rho(x_i, y_j))\, \sqrt{\Delta y_j}\alpha_j \equiv \sum_j w_{ij}\alpha_j, \tag{45}$$

where Δx_j is the area of jth grid cell, $\alpha_j \sim \mathcal{N}(0, 1)$, and the last equality defines the weights

$$w_{ij} = u(x_i, \rho(x_i, y_j))\, \sqrt{\Delta y_j}. \tag{46}$$

Sparsity of $\mathbf{W}$ is enforced by nullifying all $\mathbf{w}_{ij}$ whose absolute value is less than a threshold, θ_W.

The covariance matrix $\mathbf{B}$ of the random vector $\boldsymbol{\xi}$ (whose entries are grid-point values of $\xi(x)$) becomes, obviously,

$$\mathbf{B} = \mathbf{W}\,\mathbf{W}^\top. \tag{47}$$

The representation of the background-error covariance matrix $\mathbf{B}$ in the "square-root" form, Eq.(47), is common in data assimilation practice because, first, it provides efficient *preconditioning* of the analysis equations. Second, it allows for *thresholding* of the $\mathbf{W}$

matrix, i.e., nullifying its small in modulus entries in order to make the matrix sparse and facilitate fast computations.

Given the forecast vector $\mathbf{x}^{\mathrm{fc}}$ of length n_{x}, the vector of observations $\mathbf{x}^{\mathrm{o}}$ of length n_{o}, the observation operator $\mathbf{H} : \mathbb{R}^{n_{\mathrm{o}}} \to \mathbb{R}^{n_{\mathrm{x}}}$ (an $n_{\mathrm{x}} \times n_{\mathrm{o}}$ matrix), the optimal analysis is

$$\mathbf{x}^{\mathrm{an}} = \mathbf{x}^{\mathrm{fc}} + \mathbf{K}(\mathbf{x}^{\mathrm{obs}} - \mathbf{H}\mathbf{x}^{\mathrm{fc}}), \tag{48}$$

where

$$\mathbf{K} = (\mathbf{B}^{-1} + \mathbf{H}^{\top}\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^{\top}\mathbf{R}^{-1} \tag{49}$$

(the so-called gain matrix). The matrix to be inverted in this last equation is normally ill conditioned. The standard way to improve its conditioning is to use matrix factorization of the type Eq.(47). We proceed as follows:

$$\mathbf{K} = (\mathbf{W}^{-\top}\mathbf{W}^{-1} + \mathbf{H}^{\top}\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^{\top}\mathbf{R}^{-1} = \mathbf{W}(\mathbf{I} + \mathbf{W}^{\top}\mathbf{H}^{\top}\mathbf{R}^{-1}\mathbf{H}\mathbf{W})^{-1}\mathbf{W}^{\top}\mathbf{H}^{\top}\mathbf{R}^{-1}. \tag{50}$$

Now the matrix to be inverted is, clearly, well conditioned. (For Eq.(50) to be valid, $\mathbf{W}$ need not, actually, be invertible and even square. This can be proved by changing the control variable from $\mathbf{x}$ to $\boldsymbol{\chi}$, where $\mathbf{x} = \mathbf{W}\boldsymbol{\chi}$, see, e.g., Lorenc et al. (2000).)

Note that for computational reasons, spatial covariances are never used *per se* in a high-dimensional analysis.

6 Numerical experiments with synthetic non-stationary covariances

In this section we explore the above LSEF analysis technique in the setting with known "truth" that obeys the LSM model.

6.1 Generating the "true" local spectrum

The spatially variable true spectrum $f_\ell(x)$ is specified in a hierarchical way.

6.1.1 The lowest level in the hierarchy

$f_\ell(x)$ is on the lowest level in the hierarchy. It is computed as at each spatial point x as

$$f_\ell(x) = \frac{c(x)}{1 + [\lambda(x)(\ell + \ell_0)]^{\gamma(x)}}. \tag{51}$$

Here ℓ_0 is a non-negative number. $\lambda(x), \gamma(x)$ are two *parameter* random fields. $\lambda(x)$ is the local length scale of the process in question ξ. $\gamma(x)$ is the shape parameter of the local spectrum. $c(x)$ is the normalizing variable that ensures that the field variance $\mathsf{Var}\,\xi(x) = \frac{1}{4\pi}\sum_\ell (2\ell + 1) f_\ell(x)$ equals $S^2(x)$, where $S(x)$ is the third parameter random field.

6.1.2 Parameter fields

The *parameter fields* $S(x), \lambda(x), \gamma(x)$ are on the second level in the hierarchy. These three fields are set to be non-Gaussian *stationary* random fields computed as follows.

$$\begin{aligned} S(x) &:= S_{\text{add}} + S_{\text{mult}} \cdot g(\log \varkappa_S \cdot \chi_S(x)), \\ \lambda(x) &:= \lambda_{\text{add}} + \lambda_{\text{mult}} \cdot g(\log \varkappa_\lambda \cdot \chi_\lambda(x)), \\ \gamma(x) &:= \gamma_{\text{add}} + \gamma_{\text{mult}} \cdot g(\log \varkappa_\gamma \cdot \chi_\gamma(x)). \end{aligned} \tag{52}$$

Here g is the nonlinear transformation function defined in the next paragraph, $\chi_S, \chi_\lambda, \chi_\gamma$ are the three independent zero mean and unit variance *pre-transform* stationary Gaussian fields (defined in the next subsection) and the coefficients $\varkappa_S, \varkappa_\lambda, \varkappa_\gamma$, along with the parameters with subscripts $_{\text{add}}$ and $_{\text{mult}}$, determine the strength of the spatial non-stationarity.

The transformation function is selected, following (Tsyrulnikov and Rakitko, 2019), to be a scaled and shifted logistic function:

$$g(z) := \frac{1 + \mathrm{e}^b}{1 + \mathrm{e}^{b-z}}, \tag{53}$$

where b is the constant (with the default value of 1). The function $g(z)$ behaves like the ordinary exponential function everywhere except for $z \gg b$, where the exponential growth is tempered. The reason to replace $\exp(z)$ with $g(z)$ is the desire to avoid too large values in the parameter fields, which can give rise to unrealistically large spikes in ξ.

Due to nonlinearity of the transformation function g, the above parameter fields $S(x), \lambda(x), \gamma(x)$ are non-Gaussian. Their pointwise distributions are known as logit-normal or logit-Gaussian. As $g(z)$ is a "tempered" exponential function, it is worth measuring the standard deviation of the pre-transform fields on the log scale: say, $\text{SD}(\log(\lambda - \lambda_{\text{add}}))$ is approximately proportional to $\log \varkappa$, so that the typical deviation of the transformed field from its unperturbed value is about $\varkappa$ *times*.

With $\varkappa_S = \varkappa_\lambda = \varkappa_\gamma = 1$, the respective spectra are constant in space. The higher $\varkappa_S, \varkappa_\lambda, \varkappa_\gamma$, the more variable in space becomes the respective parameter: $S(x)$ (the standard deviation of the process at the given x), $\lambda(x)$ (the spatially variable length scale of the process), and $\gamma(x)$ (the spatially variable shape of the local correlations). We specify $\varkappa_\bullet$ to lie between 1 (stationarity) and 4 ("wild" non-stationarity), with 2 being the default value.

6.1.3 Pre-transform fields

On the highest level in the hierarchy are the *pre-transform* stationary Gaussian processes $\chi_S, \chi_\lambda, \chi_\gamma$. These are mutually independent zero mean, unit variance stationary processes whose common spatial spectrum is

$$f_\ell^\chi \propto \frac{1}{1 + (\Lambda_{\text{NSL}} \cdot \ell)^\Gamma}, \tag{54}$$

where $\Gamma := \gamma_{\text{add}} + \gamma_{\text{mult}}$, $\Lambda_{\text{NSL}} := (\lambda_{\text{add}} + \lambda_{\text{mult}}) \cdot \mu_{\text{NSL}}$ is the non-stationarity length scale of $\xi(x)$, and $\mu_{\text{NSL}} > 1$ is the non-stationarity length scale parameter. We specified μ_{NSL} in the range from 1 ("wildly" non-stationary) to 10 (almost stationary), with 3 being the default value.

6.1.4 Local spectrum and W

After the processes $S(x)$, $\lambda(x)$, and $\gamma(x)$ are computed at each analysis grid point, $c(x)$ is adjusted pointwise so that $\mathsf{Var}\,\xi(x) = S^2(x)$. With $c(x)$, $\lambda(x)$, and $\gamma(x)$ in hand, we compute the true spectrum $f_\ell(x)$ using Eq.(51) and $\sigma_\ell(x) = \sqrt{f_\ell(x)}$. Next, we make use of Eq.(8) to compute $u(x, \rho)$. After that, we build $\mathbf{W}^{\text{true}}$ using Eq.(46). The $\mathbf{W}^{\text{true}}$ matrix is then used both to generate the non-stationary random field ξ (and the ensemble members) using Eq.(43) and to compute the best possible analysis (following Eq.(50)).

6.2 Experimental setup

The grid:

$n_x = 60$

The ensemble:

$M = 20(5...100)$

The DLSM:

$\overline{S} = 1$

$W = 4(1...10)$.

$\overline{\lambda} = 250(125...500)$ km (?)

$\lambda_{\min} = \Delta x$

$\gamma_{med} = 2.5$

$\gamma_{mult} = \gamma_{med} * 5/6$

$\gamma_{add} = \gamma_{med} * 1/6$

$\varkappa_\bullet = 2(1...4)$.

The bands:

$J = 3...4$ (?)

6.3 Accuracy of the estimator of band variances

Here we experimentally evaluate the error in the approximate Eq.(27) to confirm the theoretical result obtained in Appendix E.

6.4 Accuracy of the estimator of the local spectra

6.5 Accuracy of the analysis

Can LSM improve the ensemble *sample variances* $(\mathbf{B})_{ii} = ((\mathbf{W})_{i,:}, (\mathbf{W})_{j,:})$ (which cannot be denoised by covariance localization!)?

Observations.

Point-support obs randomly located at the circle/sphere.

7 Numerical experiments with LSEF

7.1 Model

Here we took non-stationary covariances produced by the Doubly Stochastic Advection-Diffusion-decay Model (DSADM, Tsyrulnikov and Rakitko (2019)). Specifically, we tried to fit LSM to spatial covariance matrices of a field (on the 60-point 1D grid on the circle) simulated by DSADM. We had 5000 60*60 covariance matrices $\mathbf{\Gamma}_k$ computed for $k = 1, 2, \ldots, 5000$ consecutive cycles with field correlations between adjacent cycles resembling 1-day lag correlations of meteorological fields in the mid-latitude troposphere.

As the "shape" spectrum $G(.)$, we took "climatology": the time and space averaged spatial field covariances produced by DSADM.

We preferred DSADM over popular nonlinear models like Lorenz-96 (**?**) because it is the spatial covariance estimation problem that we addressed within EnKF, which .. and avoid possible side-effects due to nonlinearity of the forecast model.. cleaner setup.. model error

8 Discussion

W is a random matrix. Bayesian estimation. Hyperprior: Inverse Wishart. HBEF, DSADM: mixing with time-mean and recent past **W** yields apx-ly the posterior mode of **W**|**E** (scnd flt). We use it in the primary filter.

8.1 Kernel convolution modeling

The kernel convolution approach in modeling stationary random processes has a limitation. Banerjee et al. (2015, section 12.3) note, however, that some correlation models, e.g., the popular exponential correlation function, cannot be reproduced with the kernel convolution approach. We argue that this latter statement is true only if $L = \infty$ in the above equations. The reason is that the spectrum of the exponential correlation function, b_ℓ, may decay too slowly as $n \to \infty$ for the series in Eq.(**??**) to converge at $\rho = 0$. But if we truncate the series in Eq.(**??**) and confine ourselves to band-limited functions (evaluated on a spatial grid), then the convolution square root of the exponential correlation function $B(\rho)$ does exist.

?Instead of stoch dfr eqs (...) we use stoch integrals.

8.2 Comparison with wavelet-diagonal approach

LSM contains the stationary model as a special case, whereas a wavelet-diagonal model cannot represent a stationary field since it requires that the bands have to intersect (which creates cross-covariances, at least between adjacent bands).

8.3 Application area

Loc statio

Smooth spectra, no lines in spectrum.

Using the Loc Spec Mdl is an approach of the bias-variance-tradeoff kind: the mdl does introduce a bias but it reduces the sampling noise considerably. The approach is expected to be beneficial whenever the reduction in the sampling noise is greater than the methodological error introduced by the model.

8.4 Wavelet based filtering

The technique we have proposed in this article relies on a multi-scale bandpass filter. We used a spectral-space filter because it is easy to implement on "global" domains like the circle or the sphere. On other domains such as a limited area domain or a domain with complex boundaries (like an ocean or sea) on the sphere, the spectral-space formulation can be changed to a physical-space formulation by using wavelet filters. Indeed, applying a bandpass filter with the spectral transfer function H_ℓ is equivalent to convolving the signal with the impulse response function of filter, that is, the inverse spectral transform of the transfer function.

8.5 Computations

The computation of rows of matrix $\mathbf{W}$ from the (estimated online) local spectra $\sigma_\ell(x)$ can be done perfectly in parallel.

8.6 Extensions

Multivar, multi-level – with the bandpass filters, we can estm the "vertical" covariance matrices $\mathbf{B_l}(x)$:

$$\widehat{\mathbf{v}}_\mathbf{j}(x) = \frac{1}{4\pi} \sum_\ell H_j^2(\ell)\,(2\ell+1)\mathbf{B_l}(x) + \boldsymbol{\zeta} \tag{55}$$

Then recover $\mathbf{B_l}(x)$.

2D - isotropic. Intro anisotropy by applying directionally dependent filters (for a parametric version of the resulting model, see Heaton et al. (2014)).

Spatial *auto-regressive* models: simultaneous and conditional (MRF).

Multigrid representations to cope with a wide range of scales in a computationally efficient way.

9 Conclusions

As a result, the much desired scale dependent mixing of "background" and local spectra.

Positive-spectrum requirement fulfilled automatically. ...

The four constraints on the general process convolution model: ... Thus, the model we have proposed can be tightened or relaxed — depending on the problem in question (the prior uncertainty in the spatial covariances) and the available data (the ensemble size and the quality of the ensemble).

The traditional covariance localization is *not* capable of suppressing noise at small distances (near the diagonal of the sample covariance matrix), where it is the largest. Our LSM based technique has this capability. More generally, it regularizes the analysis problem by supplying additional information about the true covariance matrix. This additional information is inevitable because the sample covariance matrix is low-rank and thus largely uncertain. The regularizing information comes by means of the following assumptions made about the LSM.

1. The local spatial spectrum is assumed to *vary smoothly in physical space*.
2. The local spatial spectrum is assumed to be *smooth in spectral space*.
3. The local spectra are smooth enough at the origin for the entries of the weighting matrix $\mathbf{W}$ to decay quickly away from the diagonal so that their *thresholding* (i.e., nullifying small entries below a threshold) is acceptable.
4. The local spectra are monotonically decreasing.
5. The *shape* of local spectra are required to be "not too far" from the shape of the mean spectrum.

Assumptions 1 and 2 are needed for the LSM estimator based on spatial band-pass filtering of ensemble members to be consistent (i.e., to give useful results). Assumption 3 is needed for the analysis technique to be computationally efficient.

Our approach is fundamentally different from the *wavelet diagonal* approach (as in ECMWF). In the latter the coefficients of the wavelet expansion are assumed *uncorrelated*. In our approach this assumption is not introduced, which allows the model to cover the stationary case (which is not possibly with the wavelet diagonal approach since the wavelet spectral transfer functions overlap).

If, in a practical application, the $\mathbf{W}$ matrix appears to be not sparse enough, then it can be redefined for a number of spatial scales, so that large scales are represented on a sparse spatial grid whereas smaller scales are represented on denser grids. As a result, the number of non-zero entries in each row of each scale-dependent $\mathbf{W}$ will be small.

In a practical problem, at each assimilation cycle, an advantage of our approach is that the (online) estimation of LSM can be done *before observations are collected* (only background ensemble members are needed for this task).

Appendices

A LSM on the circle

On $\mathbb{S}^1$, the analogs of Eqs.(8) and (10) are

$$u(x,\rho) = \frac{1}{\sqrt{2\pi}} \sum_{\ell=-L}^{L} \sigma_\ell(x)\, \mathrm{e}^{\mathrm{i}\ell\rho}, \tag{56}$$

where $u(x,\rho)$, as a function of the (non-negative) distance ρ, is extended to negative ρ by defining it to be an even function of ρ, and

$$\xi(x) = \sum_{\ell=-L}^{L} \sigma_\ell(x)\, \widetilde{\alpha}_l\, \mathrm{e}^{\mathrm{i}\ell x}. \tag{57}$$

Here all the $\widetilde{\alpha}_\ell$ are mutually uncorrelated random Fourier coefficients with mean zero and variance one. $\widetilde{\alpha}_0$ is real valued, whilst all the others are complex circularly symmetric random variables. The kernel $u(x,\rho)$ is a real valued even function of ρ. Therefore, $\sigma_\ell(x)$ are real valued functions such that $\sigma_{-\ell}(x) = \sigma_\ell(x)$. The variance spectrum is defined as $f_\ell(x) := \sigma_\ell^2(x)$. From Eq.(57) it follows that the LSM covariances are

$$B(x, x+s) := \mathbb{E}\,\xi(x)\xi(x') = \sum_{\ell} \sigma_\ell(x)\, \sigma_\ell(x+s)\, \mathrm{e}^{\mathrm{i}\ell s}. \tag{58}$$

B The most localized kernel in the stationary case

Here we show that an isotropic kernel whose convolution with the white noise is a stationary process with the given spatial spectrum (equivalently, with the given covariance function) is not unique. Uniqueness can be achieved by imposing the requirement that the kernel be most spatially localized.

Consider a *stationary* (isotropic) process on the sphere,

$$\xi_{\text{statio}}(x) = \sum_{\ell=0}^{L} \sum_{m=-\ell}^{\ell} \widetilde{\xi}_{\ell m}\, Y_{\ell m}(x) \equiv \sum_{\ell=0}^{L} \sum_{m=-\ell}^{\ell} \sigma_\ell\, \widetilde{\alpha}_{\ell m}\, Y_{\ell m}(x), \tag{59}$$

where $\mathbb{E}\,\widetilde{\xi}_{\ell m}\widetilde{\xi}^*_{\ell' m'} = \sigma_\ell^2\, \delta_\ell^{\ell'}\delta_m^{m'}$ e.g., Yadrenko (1983, section 5.1). The process $\xi_{\text{statio}}(x)$ can be modeled as the convolution of the kernel $u(\rho) = \frac{1}{4\pi}\sum(2\ell+1)\sigma_\ell P_\ell(\cos\rho)$ with the white noise $\alpha(x)$. Multiple kernels give rise to the same spectrum $f_\ell = \sigma_\ell^2$ of the process $\xi_{\text{statio}}(x)$, differing one from another in the signs of σ_ℓ (note that on the sphere σ_ℓ are real valued). To isolate a unique kernel, we require it to be *most spatially localized* in the sense that it has the minimal width. We define the width as a kind of macro scale from

$$R^2 = \frac{\int_{\mathbb{S}^2} u^2(\rho(x,y))\, \mathrm{d}y}{(\max|u(\rho)|)^2}. \tag{60}$$

Here, the numerator is equal to $\sum f_\ell$ and thus is fixed given the set of f_ℓ. So, R is minimized when $\max|u(\rho)|$ is maximal among all kernels with the fixed $|\sigma_\ell|$. That is,

we seek to maximize $|u(\rho)|$ over both ρ and the signs of σ_ℓ. Writing down the Fourier-Legendre transform of $u(\rho)$, we observe that

$$|u(\rho)| = \left| \frac{1}{4\pi} \sum_{\ell=0}^{L} (2\ell+1)\sigma_\ell P_\ell(\cos\rho) \right| \le \frac{1}{4\pi} \sum_{\ell=0}^{L} (2\ell+1)|\sigma_\ell| \tag{61}$$

and the equality sign here holds in three cases, that is, there are three solutions to the above optimization problem.

The first solution is $\sigma_\ell \ge 0$ for all ℓ. The corresponding kernel $u_1(\rho)$ is non-negative definite and attains its maximum at $\rho = 0$.

The second solution is $\sigma_\ell \le 0$ for all ℓ. The corresponding kernel $u_2(\rho)$ is non-positive definite and attains its minimum at $\rho = 0$ so that $u_2(\rho) = -u_1(\rho)$.

The third solution is $\sigma_\ell = (-1)^\ell |\sigma_\ell|$. The corresponding kernel $u_3(\rho)$ is non-definite and attains its maximum at $\rho = \pi$. Importantly, it has exactly the same shape as the first two kernels in the sense that $u_3(\rho) = u_1(\pi - \rho)$.

These conclusions are straightforward consequences of the following facts: $P_\ell(1) = 1$, $P_\ell(-t) = (-1)^\ell P_\ell(t)$ for $-1 \le t \le 1$, and $|P_\ell(t)| < 1$ for $-1 < t < 1$ (Szegő, 1939, section 7.21).

As the above three kernels that minimize the macro scale R have exactly the same shape and thus are equivalent, we select the first solution: the non-negative definite kernel $u(\rho)$.

On the circle, similar arguments lead to virtually the same conclusion: the most localized kernel $u(x)$ is a non-negative definite function or its negated or translated version $\pm u(x-h)$ (the proof is omitted).

C LSM with positive definite kernel is identifiable

C.1 Circular case

Here we consider the LSM on $\mathbb{S}^1$ defined by Eq.(57), in which $\sigma_\ell(x) \ge 0$ (equivalently, $u(x,\rho)$ is a non-negative definite function of ρ for any x), and prove that if the non-stationary covariances $B(x,x')$ are produced by an LSM, then this LSM is unique, i.e., the spectral functions $\sigma_\ell(x)$ are uniquely determined by $B(x,x')$.

To show this, we expand $\sigma_\ell(x)$ into the truncated Fourier series

$$\sigma_\ell(x) = \sum_{q=-Q}^{Q} \widetilde{\sigma}_{\ell q}\, \mathrm{e}^{\mathrm{i}qx}, \tag{62}$$

where Q is the bandwidth of the processes $\sigma_\ell(x)$. Since $\sigma_\ell(x)$ are smooth functions, Q is normally smaller than L.

We assume that the bandwidth $[-Q, Q]$ is tight in the sense that $\widetilde{\sigma}_{\ell Q} \ne 0$ for all ℓ.

We note also that $\widetilde{\sigma}_{-\ell,q} = \widetilde{\sigma}_{\ell q}$ and $\widetilde{\sigma}_{\ell,-q} = \widetilde{\sigma}^*_{\ell q}$ — this follows from both $\xi(x)$ and $\sigma_\ell(x)$ being real valued.

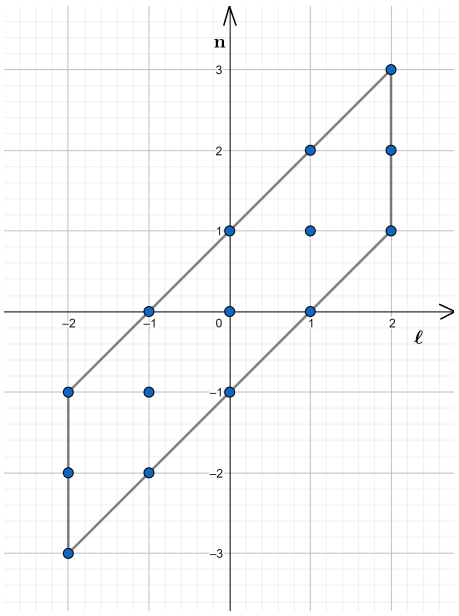

Figure 1: Circular case. Summation area for $L = 2$ and $Q = 1$

Let us substitute the expansion Eq.(62) into Eq.(57) and change the summation variable q to $n = \ell + q$:

$$\xi(x) = \sum_{\ell=-L}^{L} \widetilde{\alpha}_\ell \, \mathrm{e}^{\mathrm{i}\ell x} \sum_{q=-Q}^{Q} \widetilde{\sigma}_{\ell q} \, \mathrm{e}^{\mathrm{i}qx} = \sum_{n=-N}^{N} \widetilde{\xi}_n \, \mathrm{e}^{\mathrm{i}nx}, \tag{63}$$

where $N = L + Q$ and

$$\widetilde{\xi}_n = \sum_{\substack{\ell=n-Q \\ |\ell|\le L}}^{n+Q} \widetilde{\alpha}_\ell \, \widetilde{\sigma}_{\ell,n-\ell}, \tag{64}$$

Since the bandwidth for $\xi(x)$ is finite, $\xi(x)$ is uniquely represented by the set of its spectral coefficients $\{\widetilde{\xi}_n\}_{-N}^{N}$. Therefore the covariances $B(x, x')$ are uniquely represented by the set of the spectral covariances $\widetilde{B}_{n\bar{n}} = \mathbb{E}\,\widetilde{\xi}_n \widetilde{\xi}^*_{\bar{n}}$. Now, we derive $\widetilde{B}_{n\bar{n}}$ from Eq.(64) taking into account that $\widetilde{\alpha}_\ell$ are all mutually uncorrelated zero mean unity variance random variables:

$$\widetilde{B}_{n\bar{n}} = \mathbb{E} \sum_{\substack{\ell=n-Q \\ |\ell|\le L}}^{n+Q} \widetilde{\alpha}_\ell \, \widetilde{\sigma}_{\ell,n-\ell} \sum_{\substack{\bar{\ell}=\bar{n}-Q \\ |\bar{\ell}|\le L}}^{\bar{n}+Q} \widetilde{\alpha}^*_{\bar{\ell}} \, \widetilde{\sigma}^*_{\bar{\ell},\bar{n}-\bar{\ell}} = \sum_{\substack{\ell=n-Q \\ |\ell|\le L}}^{n+Q} \widetilde{\sigma}_{\ell,n-\ell} \, \widetilde{\sigma}^*_{\ell,\bar{n}-\ell} \tag{65}$$

and show that all $\widetilde{\sigma}_{\ell q}$ are uniquely determined by Eq.(65).

The summation area in Eq.(65) for all possible n and $\bar{n}$ is shown in Fig.1 for $L = 2$ and $Q = 1$.

We start with $n = N$, for which the sum in Eq.(64) reduces to the single term, $\widetilde{\xi}_N = \widetilde{\alpha}_L \, \widetilde{\sigma}_{LL}$ (the upper right corner of the summation area in Fig.1). This means that the covariance of $\widetilde{B}_{Nn} = \mathbb{E}\,\widetilde{\xi}_N \, \widetilde{\xi}^*_n$ for any $n \in [L-Q, L+Q]$ will also contain just one term, $\widetilde{B}_{Nn} = \widetilde{\sigma}_{LL} \, \widetilde{\sigma}_{L,n-L}$ so that we can restore $\widetilde{\sigma}_{Lq}$ for all q. First, we note that $\widetilde{B}_{NN} = |\widetilde{\sigma}_{LQ}|^2$. Next, we turn to $\widetilde{B}_{N,L-Q} = \widetilde{\sigma}_{LQ} \, \widetilde{\sigma}^*_{L,-Q}$ (the lower right corner of the summation area in

Fig.1). Recalling that $\widetilde{\sigma}_{\ell,-q} = \widetilde{\sigma}^*_{\ell q}$, we obtain $\widetilde{B}_{N,L-Q} = \widetilde{\sigma}^2_{LQ}$. So, from $\widetilde{B}_{NN} \equiv \widetilde{B}_{N,L+Q}$ and $\widetilde{B}_{N,L-Q}$ we know both the modulus and the square of the complex number $\widetilde{\sigma}_{LQ}$, hence it is uniquely determined. By the tight bandwidth assumption (see above in this subsection), $\widetilde{\sigma}_{LQ} \neq 0$, therefore we easily recover $\widetilde{\sigma}_{Lq}$ for all $|q| < Q$ from $\widetilde{B}_{N,L+q} = \widetilde{\sigma}_{LQ}\,\widetilde{\sigma}^*_{Lq}$ (the right edge of the summation parallelogram in Fig.1).

Then, we consider $\widetilde{\xi}_{N-1}$ and realize that it, again, has just one term, $\widetilde{\alpha}_{L-1}\,\widetilde{\sigma}_{L-1,L-1+Q}$ besides the term that contains $\widetilde{\sigma}_{L,L+Q-1}$, which has already been recovered. This allows us to repeat the above process and uniquely recover $\widetilde{\sigma}_{L-1,L-1+q}$ for all $q \in [-Q, Q]$. And so on, we recover all non-zero $\widetilde{\sigma}_{\ell q}$, and therefore all spectral functions $\sigma_\ell(x)$, from the set of the spectral covariances $\widetilde{B}_{n\bar{n}}$. This completes the proof of uniqueness of the LSM on the circle provided that $\sigma_\ell(x) \geq 0$ and the half-bandwidth Q of $\sigma_\ell(x)$ is such that $\widetilde{\sigma}_{\ell Q} \neq 0$ for any ℓ.

C.2 Spherical case

On $\mathbb{S}^2$, the same reasoning is applicable. We replace the Fourier series in Eq.(62) by the spherical harmonic expansion (Laplace series)

$$\sigma_\ell(x) = \sum_{q=0}^{Q} \sum_{q'=-q}^{q} \widetilde{\sigma}^{qq'}_\ell \, Y_{qq'}(x), \tag{66}$$

where $\widetilde{\sigma}^{q,-q'}_\ell = (\widetilde{\sigma}^{qq'}_\ell)^*$. We substitute this expression into Eq.(10) (where the notation m is changed to ℓ'):

$$\xi(x) = \sum_{\ell=0}^{L} \sum_{\ell'=-\ell}^{\ell} \sum_{q=0}^{Q} \sum_{q'=-q}^{q} \widetilde{\sigma}^{qq'}_\ell \, \widetilde{\alpha}_{\ell\ell'} \, Y_{\ell\ell'}(x) Y_{qq'}(x) \tag{67}$$

and project $\xi(x)$ onto $Y_{nn'}(x)$ isolating the spectral component $\widetilde{\xi}_{nn'}$ for all $0 \leq n \leq N = L+Q$ and $-n \leq n' \leq n$. The technical difficulty here is that the product of two spherical harmonics, $Y_{\ell\ell'}(x)Y_{qq'}(x)$, when expanded into the spherical harmonics basis, yields a number of components (not just one component as for the trigonometric series in the circular case above):

$$Y_{\ell\ell'}(x)Y_{qq'}(x) = \sum_{n=0}^{N} \sum_{n'=-n}^{n} C^{\ell' q' n'}_{\ell q n} Y_{nn'}(x), \tag{68}$$

where $C^{\ell' q' n'}_{\ell q n}$ can be expressed using Clebsch-Gordan coefficients and is non-zero if and only if (i) the triple $\ell q n$ satisfies the triangle inequality ($|\ell - q| \leq n \leq \ell + q$), (ii) $n' = \ell' + q'$, and (iii) $\ell + q + n$ is an even number (Arfken, 1985, section 12.9).

Substituting Eq.(68) into Eq.(67) and utilizing orthogonality of spherical harmonics, we write down the expansion $\xi(x) = \sum_{n=0}^{N} \sum_{n'=-n}^{n} \widetilde{\xi}_{nn'} \, Y_{nn'}(x)$, where

$$\widetilde{\xi}_{nn'} = \sum C^{\ell' q' n'}_{\ell q n} \, \widetilde{\alpha}_{\ell\ell'} \, \widetilde{\sigma}^{qq'}_\ell. \tag{69}$$

Here the non-zero terms correspond to the quadruples ℓ, ℓ', q, q' satisfying $0 \le \ell \le L$, $0 \le q \le Q$, $|n-\ell| \le q \le n+\ell$, $\ell+q+n$ is an even number, and $\ell'+q' = n'$.

Then, like in the circular case, we start from $\widetilde{\xi}_{NN}$ and realize that the respective sum in Eq.(69) contains only one term: $C_{LQN}^{LQN}\,\widetilde{\alpha}_{LL}\,\widetilde{\sigma}_L^{QQ}$. We derive $\mathsf{Var}\,\widetilde{\xi}_{NN}$ and $\mathsf{Cov}\,(\widetilde{\xi}_{NN}, \widetilde{\xi}_{N,L-Q})$, which allows us to recover $\widetilde{\sigma}_L^{QQ}$. As in the circular case, we assume $\widetilde{\sigma}_L^{QQ} \neq 0$. This allows us to recover all $\widetilde{\sigma}_L^{Qq'}$ by computing $\mathsf{Cov}\,(\widetilde{\xi}_{NN}, \widetilde{\xi}_{N,L+q'})$ for all $-Q \le q' \le Q$.

After that, we compute $\mathsf{Cov}\,(\widetilde{\xi}_{NN}, \widetilde{\xi}_{N-1,L-1+q'})$ for all $-Q \le q' \le Q$, retrieving $\widetilde{\sigma}_L^{Q-1,q'}$. Proceeding in this way for $n = N-2, N-3, \ldots$, we recover all $\widetilde{\sigma}_L^{qq'}$. Knowing $\widetilde{\sigma}_L^{qq'}$, we can repeat the above process to recover $\widetilde{\sigma}_{L-1}^{qq'}$, and so on, until all $\widetilde{\sigma}_\ell^{qq'}$ are found.

So, we have shown that all $\widetilde{\sigma}_\ell^{qq'}$ and thus all $\sigma_\ell(x)$ can be uniquely determined from the process (spectral) covariances. This proves the uniqueness of the LSM in the spherical case whenever $\sigma_\ell(x) \ge 0$ and the half-bandwidth Q of the process $\sigma_\ell(x)$ is such that $\widetilde{\sigma}_\ell^{QQ} \neq 0$ for any ℓ.

D LSEF analysis: posterior inference

$$p^{\mathrm{an}}(\mathbf{\Phi}, \mathbf{x}) := p(\mathbf{\Phi}, \mathbf{x} \,|\, \mathbf{E}, \mathbf{y}) \propto p(\mathbf{\Phi})\, p(\mathbf{x} \,|\, \mathbf{\Phi})\, p(\mathbf{E} \,|\, \mathbf{\Phi})\, p(\mathbf{y} \,|\, \mathbf{x}). \tag{70}$$

As in Tsyrulnikov and Rakitko (2017), we rewrite this equation as

$$p^{\mathrm{an}}(\mathbf{\Phi}, \mathbf{x}) = p(\mathbf{\Phi})\, p(\mathbf{E} \,|\, \mathbf{\Phi})\, \left[p(\mathbf{x} \,|\, \mathbf{\Phi})\, p(\mathbf{y} \,|\, \mathbf{x})\right]. \tag{71}$$

Here, taking advantage of Gaussianity of the distributions $\mathbf{x} \,|\, \mathbf{\Phi}$ and $\mathbf{y} \,|\, \mathbf{x}$, see Eqs.(22) and (25), we analytically integrate $\mathbf{x}$ out, getting

$$p^{\mathrm{an}}(\mathbf{\Phi}) = p(\mathbf{\Phi})\, p(\mathbf{E} \,|\, \mathbf{\Phi})\, p(\mathbf{y} \,|\, \mathbf{\Phi}), \tag{72}$$

where

$$p(\mathbf{y} \,|\, \mathbf{\Phi}) = \int p(\mathbf{y} \,|\, \mathbf{x})\, p(\mathbf{x} \,|\, \mathbf{\Phi})\, \mathrm{d}\mathbf{x}. \tag{73}$$

Following Tsyrulnikov and Rakitko (2017), we neglect $p(\mathbf{y} \,|\, \mathbf{\Phi})$ in Eq.(72), getting

$$p^{\mathrm{an}}(\mathbf{\Phi}) \approx p(\mathbf{\Phi})\, p(\mathbf{E} \,|\, \mathbf{\Phi}). \tag{74}$$

Dividing Eq.(71) by Eq.(74), we obtain

$$p^{\mathrm{an}}(\mathbf{x} \,|\, \mathbf{\Phi}) \propto p(\mathbf{x} \,|\, \mathbf{\Phi})\, p(\mathbf{y} \,|\, \mathbf{x}). \tag{75}$$

In Tsyrulnikov and Rakitko (2017) we found that with a similar problem, an empirical Bayes solution worked well. Applying that approach, we, first, find a point estimate of the local spectra $\mathbf{\Phi}$ by maximizing the posterior density $p^{\mathrm{an}}(\mathbf{\Phi})$ and, second, plug this estimate, $\mathbf{\Phi}^{\mathrm{an}}$, into Eq.(75). The resulting posterior density $p^{\mathrm{an}}(\mathbf{x} \,|\, \mathbf{\Phi}^{\mathrm{an}})$ appears to be Gaussian so that the classical Kalman filter analysis of $\mathbf{x}$ is applicable. Neglecting the uncertainty of $\mathbf{\Phi}$ in $p^{\mathrm{an}}(\mathbf{x} \,|\, \mathbf{\Phi})$ is an approximation but it works well in low dimensions and

it greatly reduces the cost of the analysis, making it affordable for ultrahigh-dimensional problems we face in geosciences and elsewhere.

Finally, an pseudo-random sample from $p^{\rm an}(\mathbf{x}\,|\,\mathbf{\Phi}^{\rm an})$ is drawn following the stochastic Ensemble Kalman Filter approach (). The strength of that approach is in its reliance on the prior ensemble, which approximately spans the subspace of leading backward Lyapunov vectors (Bocquet). This ensures that the analysis ensemble approximately lies on the model attractor. This is what we always would like to have (but which we have never explicitly encoded in the analysis equations..)

E Consistency of the estimator of band variances

Here we consider bandpass filtering of a locally stationary process and find conditions under which the approximations Eq.(26) and Eq.(27) hold. For simplicity, we examine the circular case.

Let the linear filter $\mathcal{H}$ with the spectral transfer function $H(\ell)$ be applied to the locally stationary process (we reproduce here Eq.(57) for the reader's convenience)

$$\xi(x) = \sum_{\ell=-L}^{L} \sigma_\ell(x)\,\widetilde{\alpha}_l\,\mathrm{e}^{\mathrm{i}\ell x}. \tag{76}$$

In spectral space, the action of a linear filter on a signal is the element-wise multiplication of the signal's spectral representation by $H(\ell)$. The spectral-space representation of $\xi(x)$ is given by Eq.(63) so that the filtered process $\xi_{\mathcal{H}} = \mathcal{H}\xi$ reads

$$\xi_{\mathcal{H}}(x) = \sum_{\ell=-L}^{L} \widetilde{\alpha}_\ell\,\mathrm{e}^{\mathrm{i}\ell x} \sum_{q=-Q}^{Q} H(\ell+q)\,\widetilde{\sigma}_{\ell q}\,\mathrm{e}^{\mathrm{i}q x}. \tag{77}$$

Now we recall that the local stationarity means that the processes $\sigma_\ell(x)$ are slowly changing in space, which is equivalent to a rapid decrease of the coefficients $\widetilde{\sigma}_{\ell q}$ in their spectral decomposition $\sigma_\ell(x) = \sum_q \widetilde{\sigma}_{\ell q}\,\mathrm{e}^{\mathrm{i}qx}$ (Eq.(62)) for any wavenumber ℓ. If we specify $H(\ell)$ to change slowly with ℓ, then, in the sum over q in Eq.(77), $H(\ell+q)$ can be approximated by $H(\ell)$ leading to the approximation

$$\check{\xi}_{\mathcal{H}}(x) = \sum \sigma_\ell(x)\,\widetilde{\alpha}_l\,H(l)\,\mathrm{e}^{\mathrm{i}\ell x}. \tag{78}$$

(Its spherical counterpart is given in Eq.(26).) A more rigorous proof of this statement follows. The error in the approximation Eq.(78) is

$$\check{\xi}_{\mathcal{H}}(x) - \xi_{\mathcal{H}}(x) = \sum_{\ell} \widetilde{\alpha}_\ell\,\mathrm{e}^{\mathrm{i}\ell x} \sum_{q} \left[H(\ell+q) - H(\ell)\right]\,\widetilde{\sigma}_{\ell q}\,\mathrm{e}^{\mathrm{i}q x}, \tag{79}$$

Here we remember that $\mathcal{H}$ is a bandpass filter and assume that its spectral transfer function is generated by a bell-shaped function of continuous argument, $\kappa(z)$, such that

$\kappa(0) = 1$ and a half-width of $\kappa(z)$ is also about 1 (more precisely, we require that $\max|\kappa'(z)| = 1$). Specifically, let

$$H(\ell) = \kappa((\ell - \ell^{\mathrm{c}})/d), \tag{80}$$

where ℓ^{c} is the band's central wavenumber and d is the half-bandwidth (*cf.* Eq.(30)). As we noted above, since $\widetilde{\sigma}_{\ell q}$ rapidly decays with the growing $|q|$ for any ℓ, the use of the first order Taylor series approximation is warranted: $H(\ell + q) \approx H(\ell) + H'(\ell)q = H(\ell) + \kappa'((\ell - \ell^{\mathrm{c}})q/d$. Substituting this equation into Eq.(79) yields

$$\check{\xi}_{\mathcal{H}}(x) - \xi_{\mathcal{H}}(x) = \frac{1}{d}\sum_{\ell}\widetilde{\alpha}_{\ell}\,\mathrm{e}^{\mathrm{i}\ell x}\kappa'((\ell - \ell^{\mathrm{c}})/d)\sum_{q} q\,\widetilde{\sigma}_{\ell q}\,\mathrm{e}^{\mathrm{i}qx} \equiv \frac{1}{\mathrm{i}d}\sum_{\ell}\widetilde{\alpha}_{\ell}\,\mathrm{e}^{\mathrm{i}\ell x}\kappa'((\ell - \ell^{\mathrm{c}})/d)\sigma'_{\ell}(x). \tag{81}$$

Here the second equality is due to $\sigma'_{\ell}(x) = \sum_q \mathrm{i}q\,\widetilde{\sigma}_{\ell q}\,\mathrm{e}^{\mathrm{i}qx}$.

Next, we compute the mean square value of this expression (with all σ_ℓ being fixed) and note that the width d of the spectral transfer function $H(\ell)$ equals the inverse width L_{H} of the impulse response function of the filter $h(x)$ (which is the inverse Fourier transform of $H(\ell)$), getting

$$\mathbb{E}\left((\check{\xi}_{\mathcal{H}}(x) - \xi_{\mathcal{H}}(x))^2 \,|\, \sigma\right) = L_{\mathrm{H}}^2\sum_{\ell}(\kappa'((\ell - \ell^{\mathrm{c}})/d))^2(\sigma'_{\ell}(x))^2. \tag{82}$$

After that, we take the expectation of this equation, make use of Eq.(13), and take into account that $\Lambda_\ell \geq \Lambda$, getting the mean square approximation-error

$$\mathbb{E}\,\mathbb{E}\left((\check{\xi}_{\mathcal{H}}(x) - \xi_{\mathcal{H}}(x))^2 \,|\, \sigma\right) \leq \left(\frac{L_{\mathrm{H}}}{\Lambda}\right)^2\sum(\kappa'((\ell - \ell^{\mathrm{c}})/d))^2\,\mathsf{Var}\,\sigma_{\ell}. \tag{83}$$

At the same time, the mean variance of the approximating process $\check{\xi}_{\mathcal{H}}(x)$ (see Eq.(78)) is

$$\mathbb{E}\,\mathbb{E}\left((\check{\xi}_{\mathcal{H}}(x))^2 \,|\, \sigma\right) = \sum\kappa^2((\ell - \ell^{\mathrm{c}})/d)\;\mathbb{E}\,\sigma_{\ell}^2. \tag{84}$$

Dividing Eq.(83) by Eq.(84) we obtain the relative approximation error

$$\begin{aligned}(\mathrm{rel.err})^2 := \frac{\mathbb{E}\,\mathbb{E}\left((\check{\xi}_{\mathcal{H}}(x) - \xi_{\mathcal{H}}(x))^2 \,|\, \sigma\right)}{\mathbb{E}\,\mathbb{E}\left((\check{\xi}_{\mathcal{H}}(x))^2 \,|\, \sigma\right)} \leq \left(\frac{L_{\mathrm{H}}}{\Lambda}\right)^2\frac{\sum(\kappa'(.))^2\,\mathsf{Var}\,\sigma_{\ell}}{\sum\kappa^2(.)\;\mathbb{E}\,\sigma_{\ell}^2} \equiv \\ \left(\frac{L_{\mathrm{H}}}{\Lambda}\right)^2\frac{\sum(\kappa'(.))^2\,\mathsf{Var}\,\sigma_{\ell}}{\sum\kappa^2(.)\,\mathsf{Var}\,\sigma_{\ell}}\cdot\frac{\sum\kappa^2(.)\,\mathsf{Var}\,\sigma_{\ell}}{\sum\kappa^2(.)\;\mathbb{E}\,\sigma_{\ell}^2} \approx \left(\frac{L_{\mathrm{H}}}{\Lambda}\right)^2\frac{\sum H^2(\ell)\,\mathsf{Var}\,\sigma_{\ell}}{\sum H^2(\ell)\;\mathbb{E}\,\sigma_{\ell}^2}.\end{aligned} \tag{85}$$

Here the last approximate equality is due to the fact that both κ and κ' are smooth functions that have the same effective supports and the same maximum values so that with the smooth spectrum $\mathsf{Var}\,\sigma_\ell$, both $\sum(\kappa'(.))^2\,\mathsf{Var}\,\sigma_\ell$ and $\sum\kappa^2(.)\,\mathsf{Var}\,\sigma_\ell$ are of the same order of magnitude.

Equation (85) implies that if we fix the filter, then the approximation error is asymptotically zero:

$$\mathrm{rel.err} \leq \mathsf{const}\cdot\frac{\Sigma}{\Lambda} \to 0 \tag{86}$$

provided that $\xi(x)$ is locally stationary, see Eq.(16). If, on the other hand, we fix the covariances of $\xi(x)$ (that is, the processes $\sigma_\ell(x)$), then Eq.(85) implies that

$$\text{rel.err} \leq \frac{L_{\text{H}}}{\Lambda}. \tag{87}$$

In words, Eqs.(85)–(87) show that the approximation Eq.(78) of the bandpass filter's output Eq.(77) is accurate whenever the width L_{H} of the filter's impulse response function is much less than the non-stationarity length scale Λ (equivalently, the filter's spectral transfer function is wide enough). If the spatial variability in the processes $\sigma_\ell(x)$ is significantly smaller than the mean variance of the filtered process, then the same level of approximation error can be achieved with larger L_{H} and thus with narrower spectral transfer functions, thus, implying better spectral resolution.

Finally, we note that in the limit we considered in , the above result implies that with the norm defined from $\|.\|^2 := \mathbb{E}\,\mathbb{E}\,((.)^2 \,|\, \sigma)$, we have $\|\check{\xi}_{\mathcal{H}} - \xi_{\mathcal{H}}\| \to 0$. Therefore, $\|\check{\xi}_{\mathcal{H}}\|^2 \to \|\xi_{\mathcal{H}}\|^2$ so that the approximate band variance estimate $\sum H^2(l)\sigma_\ell^2(x)$ (see Eq.(27)), which follows from Eq.(78) is asymptotically unbiased and consistent.

References

G. B. Arfken. *Mathematical methods for physicists*. Academic Press, 1985.

S. Banerjee, B. P. Carlin, and A. E. Gelfand. *Hierarchical modeling and analysis for spatial data*. Taylor and Francis, 2015.

L. Berre and G. Desroziers. Filtering of background error variances and correlations by local spatial averaging: A review. *Mon. Weather Rev.*, 138(10):3693–3720, 2010.

L. Berre, H. Varella, and G. Desroziers. Modelling of flow-dependent ensemble-based background-error correlations using a wavelet formulation in 4D-Var at Météo-France. *Q. J. Roy. Meteorol. Soc.*, 141(692):2803–2812, 2015.

M. Bonavita, E. Hólm, L. Isaksen, and M. Fisher. The evolution of the ECMWF hybrid data assimilation system. *Quart. J. Roy. Meteor. Soc.*, 142(694):287–303, 2016.

M. Buehner, J. Morneau, and C. Charette. Four-dimensional ensemble-variational data assimilation for global deterministic weather prediction. *Nonlin. Process. Geophys.*, 20 (5):669–682, 2013.

R. Dahlhaus. Fitting time series models to nonstationary processes. *Ann. Stat.*, 25(1): 1–37, 1997.

R. Dahlhaus. Locally stationary processes. In *Handbook of statistics*, volume 30, pages 351–413. Elsevier, 2012.

M. Fisher. Background error covariance modelling. *Proc. ECMWF Semin. on recent developments in data assimilation for atmosphere and ocean, 8-12 September 2003*, pages 45–64, 2003.

R. Furrer and T. Bengtsson. Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants. *J. Multivar. Anal.*, 98(2):227–255, 2007.

K. S. Gage and G. D. Nastrom. Theoretical interpretation of atmospheric wavenumber spectra of wind and temperature observed by commercial aircraft during GASP. *Journal of Atmospheric Sciences*, 43(7):729–740, 1986.

M. Heaton, M. Katzfuss, C. Berrett, and D. Nychka. Constructing valid spatial processes on the sphere using kernel convolutions. *Environmetrics*, 25(1):2–15, 2014.

D. Higdon, J. Swall, and J. Kern. Non-stationary spatial modeling. *Bayes. Statist.*, 6(1): 761–768, 1999.

E. Hou, E. Lawrence, and A. O. Hero. Penalized ensemble Kalman filters for high dimensional non-linear systems. *PloS one*, 16(3):e0248046, 2021.

P. L. Houtekamer and H. L. Mitchell. Data assimilation using an ensemble Kalman filter technique. *Mon. Weather Rev.*, 126(3):796–811, 1998.

Y. Kakihara. *Multidimensional second order stochastic processes*. World Scientific, 1997.

I. Kasanickỳ, J. Mandel, and M. Vejmelka. Spectral diagonal ensemble Kalman filters. *Nonlinear Processes in Geophysics*, 22(4):485–497, 2015.

M. Katzfuss, J. R. Stroud, and C. K. Wikle. Ensemble kalman methods for high-dimensional hierarchical dynamic space-time models. *Journal of the American Statistical Association*, 115(530):866–885, 2020.

O. Ledoit and M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivar. Anal.*, 88(2):365–411, 2004.

A. Lorenc, S. Ballard, R. Bell, N. Ingleby, P. Andrews, D. Barker, J. Bray, A. Clayton, T. Dalby, D. Li, et al. The Met. Office global three-dimensional variational data assimilation scheme. *Quarterly Journal of the Royal Meteorological Society*, 126(570): 2991–3012, 2000.

A. C. Lorenc. Improving ensemble covariances in hybrid variational data assimilation without increasing ensemble size. *Quart. J. Roy. Meteor. Soc.*, 143(703):1062–1072, 2017.

A. C. Lorenc, N. E. Bowler, A. M. Clayton, S. R. Pring, and D. Fairbairn. Comparison of hybrid-4DEnVar and hybrid-4DVar data assimilation methods for global NWP. *Mon. Weather Rev.*, 143(2015):212–229, 2014.

S. Mallat, G. Papanicolaou, and Z. Zhang. Adaptive covariance estimation of locally stationary processes. *The annals of Statistics*, 26(1):1–47, 1998.

D. Marinucci and D. Peccati. *Random Fields on the Sphere*. Cambridge University Press, 2011.

B. Ménétrier, T. Montmerle, Y. Michel, and L. Berre. Linear filtering of sample covariances for ensemble-based data assimilation. Part I: optimality criteria and application to variance filtering and covariance localization. *Mon. Weather Rev.*, 143(5):1622–1643, 2015.

M. B. Priestley. Evolutionary spectra and non-stationary processes. *Journal of the Royal Statistical Society. Series B (Methodological)*, 27(2):204–237, 1965.

M. B. Priestley. *Non-linear and non-stationary time series analysis*. 1988.

J. Sætrom and H. Omre. Uncertainty quantification in the ensemble Kalman filter. *Scand. J. Stat.*, 40(4):868–885, 2013.

S. R. Searle and A. I. Khuri. *Matrix algebra useful for statistics*. John Wiley & Sons, 2017.

J. Skauvold and J. Eidsvik. Parametric spatial covariance models in the ensemble Kalman filter. *Spatial statistics*, 29:226–242, 2019.

G. Szegő. *Orthogonal polynomials*. American Mathematical Soc., 1939.

K. E. Trenberth and A. Solomon. Implications of global atmospheric spatial spectra for processing and displaying data. *Journal of climate*, 6(3):531–545, 1993.

M. Tsyrulnikov and A. Rakitko. A hierarchical Bayes ensemble Kalman filter. *Physica D*, 338:1–16, 2017.

M. Tsyrulnikov and A. Rakitko. Impact of non-stationarity on hybrid ensemble filters: A study with a doubly stochastic advection-diffusion-decay model. *Quart. J. Roy. Meteorol. Soc.*, pages 2255–2271, 2019.

G. Ueno and T. Tsuchiya. Covariance regularization in inverse space. *Q. J. Roy. Meteorol. Soc.*, 135(642):1133–1156, 2009.

M. I. Yadrenko. *Spectral theory of random fields*. Optimization Software, 1983.