The ensemble Kalman filter regularized with nonstationary nonparametric convolutions

Working paper

M Tsyrulnikov, A Sotskiy, and D Gayfulin
HydroMetCenter of Russia

August 12, 2020

1 Introduction

2 Notation

We use the terms "process" and "field" interchangeably.

On $\mathbb{S}^2$, we represent the fields on a regular longitude-latitude grid with $n_{\rm lon}$ and $n_{\rm lat}$ grid points over longitude and latitude, respectively. The dimensionality of state space is denoted by $n_{\rm x} = n_{\rm lon} \cdot n_{\rm lat}$. The ensemble size is denoted by $n_{\rm e}$.

Vectors of length $n_{\rm x}$ are written in bold case, e.g., $\boldsymbol{\xi}$. Vectors of length $n_{\rm x} \times n_{\rm e}$ are written in bold case with an arrow, e.g., $\vec{\boldsymbol{\xi}}$.

Matrices of size $n_{\rm x} \times n_{\rm e}$ are denoted by bold capital letters, e.g., $\mathbf{X}$.

Points on the sphere are denoted either by the pair (θ, φ) (where θ is the co-latitude and φ is longitude) or sometimes simply by a lowercase letter s, x, y, etc. By $\rho(r, s)$ we denote the great-circle (angular) distance between the two points r and s on the sphere.

The forward spectral (spherical harmonic) transform of a function, $f(x)$, on the sphere we denote by $\mathcal{S}_{y \to mn} : f(x) \mapsto \widetilde{f}_{lm}$. The inverse transform is denoted by $\mathcal{S}_{mn \to y} : \widetilde{f}_{lm} \mapsto f(x)$. The spectrum of a function, f, that depends on the great-circle distance ρ on the sphere, is provided by the Fourier-Legendre transform denoted $\mathcal{S}_{\rho \to n} : f(\rho) \mapsto \widetilde{f}_l$.

Stationarity = isotropy on the sphere.

3 Process convolution 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 nonstationary spatial process. To constrain the non-parametric convolution model and make it identifiable we require that the spatial kernel is locally isotropic. We show that this means that the model has a *local spectrum* so we call it a Local Spectrum Model (LSM).

The general intention is to introduce a model that can be made more or less "tight" dependent on the ensemble size and the non-stationarity of the problem at hand.

3.1 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 $\alpha(y)$:

$$\xi(x) = \int_{\mathbb{S}^2} w(x,y)\,\alpha(y)\,\mathrm{d}y \equiv \int_{\mathbb{S}^2} w(x,y)\,Z(\mathrm{d}y), \tag{1}$$

where Z is the spatial orthogonal stochastic measure (such that $\mathbb{E}\,Z(\mathrm{d}A) = 0$, $\mathbb{E}(Z(\mathrm{d}A))^2 = |\mathrm{d}A|$, 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 on the sphere, $|\mathrm{d}A|$ its surface area, and w is a real function (called the convolution kernel or the weighting function) which, $\forall x \in \mathbb{S}^2$, is required to be square integrable w.r.t. its second argument,

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

in order for $\mathsf{Var}\,\xi(x)$ to be finite. In what follows, we impose a number of additional constraints, besides Eq.(2), on the weighting function w.

3.2 Space discrete process convolution model

In data assimilation, the process in question is always represented by a vector, $\boldsymbol{\xi}$, on a spatial grid $G = \{\mathbf{s}_j\}_{j=1}^{n_\mathrm{x}}$, where n_x is the number of grid points and $\mathbf{s}_j$ are the grid point locations. If the space discrete process is an approximation to a space continuous process, the values of the latter are, normally, smoothed or averaged to get the values of the former (to avoid aliasing). Another possibility is that the process is *defined* to be space discrete without explicitly having its space continuous "parent".

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 symmetric 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 $\mathbf{W}' = \mathbf{WQ}$, where $\mathbf{Q}$ is an orthogonal matrix, also satisfies Eq.(3). Our goal will be to select the *sparsest* or the *most localized* weighting matrix $\mathbf{W}$. The strategy will be to *constrain* the space continuous model and then to discretize it in space.

3.3 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) and facilitate its estimation from an ensemble of process realizations, We begin with the space continuous model, constraining $w(x, y)$ in Eq.(1) to be of the form

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

where the dependence of $u(x, \rho)$ on its first argument x is much weaker than on its second argument ρ, so that the kernel $u(x, \rho)$ can be called *locally isotropic*. In the limit of no dependence of $u(x, \rho)$ on its first argument, we obtain an isotropic kernel.

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

$$\xi(x) = \int_{\mathbb{S}^2} u(x, \rho(x, y))\, \alpha(y)\, \mathrm{d}y \equiv \int_{\mathbb{S}^2} u(x, \rho(x, y))\, Z(\mathrm{d}y). \tag{6}$$

Next, we develop a spectral representation of the the process in question and of its spatial covariances. To this end, first, we employ the spectral representation of the white noise $\alpha(x)$:

$$\alpha(y) = \sum_{l=0}^{l_{\max}} \sum_{m=-l}^{l} \widetilde{\alpha}_{lm}\, Y_{lm}(y), \tag{7}$$

where $l_{\max}$ is the maximal total wavenumber resolvable on the spatial grid. It can be seen that $\widetilde{\alpha}_{lm}$ are mutually independent complex-valued random Fourier coefficients with $\mathbb{E}\,\widetilde{\alpha}_{lm} = 0$ and $\mathsf{Var}\,\widetilde{\alpha}_{lm} = 1$. Second, we perform the spectral (Fourier-Legendre) expansion of $u(x, \rho)$ with x being fixed:

$$u(x, \rho) = \frac{1}{4\pi} \sum_{l} (2l + 1) \widetilde{u}_l(x) P_l(\cos \rho). \tag{8}$$

In this equation, substituting $\rho = \rho(x, y)$ and applying the addition theorem for spherical harmonics, we obtain

$$u(x, \rho(x, y)) = \sum_{l=0}^{l_{\max}} \widetilde{u}_l(x) \sum_{m=-l}^{l} Y_{lm}(x)\, \overline{Y_{lm}(y)}, \tag{9}$$

Finally, we substitute Eqs.(61) and (9) into Eq.(6). Utilizing orthonormality of spherical harmonics, we obtain:

$$\xi(x) = \sum_{l=0}^{l_{\max}} \sum_{m=-l}^{l} \widetilde{u}_l(x)\, \widetilde{\alpha}_{lm}\, Y_{lm}(x). \tag{10}$$

Note that the restriction Eq.(5), which postulates local isotropy of the convolution kernel $w(x, y)$, is the **first constraint** we impose on the general convolution model.

3.4 Space discrete locally isotropic kernel

The space discrete equivalent of Eq.(6) is

$$\xi_i = \sum_j u(x_i, \rho(x_i, y_j))\, Z(\Delta y_j) = \sum_j u(x_i, \rho(x_i, y_j))\, \sqrt{\Delta y_j}\alpha_j \equiv \sum_j w_{ij}\alpha_j, \tag{11}$$

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

$$w_{ij} = u(x_i, \rho(x_i, y_j))\, \sqrt{\Delta y_j}. \tag{12}$$

3.5 Locally isotropic convolution model

Extending the definition by Vogt et al. (2012) to the sphere, we say that $\xi(x)$ is *locally stationary (isotropic)* on the sphere if, for any location x_0, there is a *stationary* random process $\zeta(x; x_0)$ such that in the vicinity of any fixed x_0, the stationary process $\zeta(x; x_0)$ is close the nonstationary process in question $\xi(x)$ in the following sense:

$$\mathsf{Var}\,(\zeta(x; x_0) - \xi(x)) < C\rho^2(x, x_0), \tag{13}$$

where C is the constant independent of x and x_0.

It can be seen that the process defined by Eq.(10) is locally stationary in the sense of this definition if $\widetilde{u}_l(x)$ vary slowly (smoothly) with x. Indeed, for any fixed x_0, the process

$$\zeta(x; x_0) = \sum_{l,m} \widetilde{u}_l(x)\, \widetilde{\alpha}_{lm}\, Y_{lm}(x) \tag{14}$$

is stationary and

$$\mathsf{Var}\,(\zeta(x; x_0) - \xi(x)) = \sum_{l=0}^{l_{\max}} (\widetilde{u}_l(x_0) - \widetilde{u}_l(x))^2 \sum_{m=-l}^{l} Y_{lm}(x)\overline{Y_{lm}(x_0)} \le C\, \rho^2(x, x_0) \tag{15}$$

if $\widetilde{u}_l(x)$ are Lipschitz continuous, uniformly in l. We ensure the uniform Lipschitz continuity by limiting the support of the spectrum of $\widetilde{u}_l(x)$ as a function of x, see Eq.(21), and by requiring that $\mathsf{Var}\,\xi(x)$ is uniformly bounded (see Appendix **??** for details). The inequality in Eq.(15) is justified by the boundedness of the spectrum $\mathcal{S}_{x\to mn}\widetilde{u}_l(x)$, the addition theorem for spherical harmonics, and the bounds $|P_l(\cos\rho)| \le 1$.

Thus, we have shown that the convolution model with a locally isotropic convolution kernel defines a locally isotropic process $\xi(x)$.

3.6 Locally isotropic convolution model as a local-spectrum model (LSM)

If we compare the spectral Eq.(10) with the spectral expansion of an *isotropic* random field on the sphere, Eq.(59) we find a significant similarity. Indeed, the only difference is $\widetilde{u}_l(x)$ in Eq.(10) vs. $\widetilde{u}_l$ in Eq.(59). Thus, the difference of the process convolution model from the isotropic model is in the dependency, in the latter, of the spectral standard deviations $\widetilde{u}_l$ on x. Since in the isotropic model $b_l = \widetilde{u}_l^2$ is called the spectrum, we call

$$b_l(x) := \widetilde{u}_l^2(x) \tag{16}$$

the *local spectrum* and the resulting model the Local Spectrum Model (LSM). For the spectral standard deviations $\widetilde{u}_l(x)$ we will also use the notation

$$\sigma_l(x) \equiv \widetilde{u}_l(x). \tag{17}$$

So, we may rewrite Eq.(10) as

$$\boxed{\xi(x) = \sum_{l=0}^{l_{\max}} \sum_{m=-l}^{l} \sigma_l(x)\, \widetilde{\alpha}_{lm}\, Y_{lm}(x)} \tag{18}$$

Using Eq.(10), taking into account that all $\widetilde{\alpha}_{lm}$ are mutually uncorrelated, and applying the addition theorem for spherical harmonics, we obtain the covariance between the points x and x' on the sphere:

$$B(x,x') := \mathbb{E}\,\xi(x)\xi(x') = \frac{1}{4\pi} \sum_{l=0}^{l_{\max}} (2l+1)\, \sigma_l(x)\, \sigma_l(x')\, P_l(\cos\rho(x,x')). \tag{19}$$

We also notice that Eq.(8), which relates the kernel $u(x,\rho)$ to the local spectrum B_l constitutes the inverse Fourier-Legendre transform $\mathcal{S}_{l\to\rho} : \sigma_l(x) \mapsto u(x,\rho)$. The respective forward transform reads

$$\sigma_l(x) = 2\pi \int_{-1}^{1} u(x,\rho)\, P_l(\cos\rho) \sin\rho \,\mathrm{d}\rho. \tag{20}$$

This equation implies that with the real valued function $u(x,\rho)$, $\sigma_l(x)$ are necessarily real. They will be positive if, for any x, $u(x,\rho)$ is a positive definite function of the angular distance ρ.

3.7 Physical-space smoothness constraint

Of course, the term "local spectrum" is justified only if $\widetilde{u}_l(x)$ varies slowly with x (Priestley, 1965, 1988). We adopt this assumption because it further constrains the LSM. Specifically, we postulate that the spectral expansion of $\widetilde{u}_l(x)$ w.r.t. x is confined to the wavenumber range $|n| \le M \ll l_{\max}$:

$$\widetilde{u}_l(x) = \sum_{p=0}^{M} \sum_{q=-p}^{p} \widetilde{\sigma}_l^{pq}\, Y_{pq}(x) \tag{21}$$

The limitation $M \ll l_{\max}$ is our **second constraint** imposed on the general process convolution model. One can show that this constraint ensures that the LSM is *unique* given its spatial covariances. For the sake of simplicity, we prove this statement in Appendix ... for the process indexed on $\mathbb{S}^1$ (instead of $\mathbb{S}^2$).

3.8 Implications for data assimilation

Equation (11) implies that the nonstationary space-discrete random vector $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_{n_x})$ satisfies

$$\boldsymbol{\xi} = \mathbf{W}\boldsymbol{\alpha}, \tag{22}$$

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

$$(\mathbf{W})_{ij} := w_{ij} \tag{23}$$

are defined in Eq.(12). Then, the covariance matrix $\mathbf{B}$ of the random vector $\boldsymbol{\xi}$ (whose entries are grid-point values of $\xi(x)$) becomes, obviously,

$$\mathbf{B} = \mathbf{W}\,\mathbf{W}^\top. \tag{24}$$

The representation of the background-error covariance matrix $\mathbf{B}$ in the "square-root" form, Eq.(24) is common in data assimilation practice and provides the following benefits for a data assimilation (analysis) scheme:

1. The decomposition $\mathbf{B} = \mathbf{W}\,\mathbf{W}^\top$ allows *preconditioning* of the analysis equations (as it is common in variational schemes). This ensures fast convergence of a variational-analysis solver.

2. If ξ has *short-range* correlations (which is often the case in practice), that is, if $B(\rho(x_i, x_j))$ rapidly decays as x_j moves away from x_i, so will the function $u(x_i, \rho(x_i, x_j))$. Restricting the support of the function $u(x, \rho(x, y)$ with respect to y introduces a *sparsity* pattern in the matrix $\mathbf{W}$ and provides a kind of *localization*, which is key to fast computations.

   Limiting the number of entries in a row of the $\mathbf{W}$ matrix which are allowed to be non-zero or nullifying its small enough entries further constrains LSM, constituting our **third constraint**.

3. The computation of rows of matrix $\mathbf{W}$ from the (estimated online) local spectra $\widetilde{u}_l(x)$ can be done perfectly in parallel, which implies fast computations on current and future massively parallel machines.

4. As the spatial covariances are assumed to vary *smoothly* (at a spatial scale significantly larger than the length scale of the process ξ itself), the local spectrum $\widetilde{u}_l(x)$ can be evaluated on a *coarse* spatial grid. This may also contribute to computational efficiency of an LSM based covariance model and analysis scheme.

4 Estimation of W from the ensemble

We proceed in two steps: (i) estimate the local spectra $\sigma_l^2(x)$ and (ii) compute the weighting matrix $\mathbf{W}$.

4.1 Existing approaches

1. Original technique (Priestley, 1988). Narrow-band filtering.

2. Local periodogram (windowed Fourier transform, short-time Fourier transform, estimation in segments). (Dahlhaus, 1997), (Wieczorek and Simons, 2005)

   Rosen: partition/segmentation of the interval of time.

3. Wavelet spectra

   (e.g. Spanos et al., 2005): perform a discrete wavelet transform and estimate the variances of the wavelet coefficients.

   (Nason et al., 2000)

   (Berre et al., 2015)

4.2 The proposed estimator

We propose a modification of the original technique by Priestley (1965). The modification is needed because Priestley (1965) worked with a single realization of the random process in the time domain, whereas we have multiple realizations (an ensemble) defined in a spatial domain.

The goal is to estimate the local spectrum $\sigma_l^2(x)$ for any x we wish. Recall that x stands for (θ, φ) everywhere.

We propose a technique that resembles what is called in signal processing "complex demodulation", (e.g Webb, 1979). Specifically,

1. Perform a bandpass filtering of the nonstationary process that satisfies Eq.(18) for the wavenumber bands $j = 1, \ldots, n_{\text{band}}$, getting the respective bandpass filtered processes $\xi_{(j)}(x)$.

2. For any point of interest x, estimate the waveband *variances* $V_{(j)}(x) := \mathsf{Var}\, \xi_{(j)}(x)$ (for $j = 1, \ldots, n_{\text{band}}$) from the ensemble, getting $\widehat{V}_{(j)}(x)$ Relate $\widehat{V}_{(j)}(x)$ to $\sigma_l^2(x)$ and produce the estimator $\widehat{\sigma}_l^2(x)$.

We assume that the available *data* are *band-pass filtered ensemble members for a number of spectral bands* (wavebands) as explained below. The reason for using spectral-bands data is threefold: (1) to reduce sampling noise (as the sample size is small), (2) to address horizontal non-stationarity, and (3) to impose smoothness of the spectrum (note the a small number of wavebands implies a low resolution in spectral space).

4.2.1 Spectral bands

We introduce n_{band} filters $\mathcal{H}_j$, where $j = 1, \ldots, n_{\mathrm{band}}$. The j-th band is characterized by its *transfer function* $H_j(l)$. Note that we postulate H_j to depend only on the *total* wavenumber l so that if the filter is fed by an *isotropic* random field (on the sphere), the output will be an *isotropic* random field as well.

The filter $\mathcal{H}$ is a *strictly bandpass* filter if its transfer function is the indicator function supported on the wavenumber segment $[\underline{l}, \bar{l}]$. Besides such strictly bandpass filters, we will also employ filters whose transfer function is a smoothed version of the indicator function. We will call them also bandpass filters since they favor a limited band of wavenumbers.

The bands' transfer functions can overlap and cover the whole wavenumber range represented on the analysis grid.

The action of the filter $\mathcal{H}$, which has the transfer function $H(l)$, on the input field

$$\xi(\theta, \varphi) = \sum_{l=0}^{l_{\max}} \sum_{m=-l}^{l} \widetilde{\xi}_{lm} \, Y_{lm}(\theta, \varphi), \tag{25}$$

is

$$\boxed{(\mathcal{H}\xi)(\theta, \varphi) = \sum_{l=0}^{l_{\max}} \sum_{m=-l}^{l} H(l) \widetilde{\xi}_{lm} \, Y_{lm}(\theta, \varphi).} \tag{26}$$

4.2.2 Variances of bandpass filtered processes

Applying the filter $\mathcal{H}_j$ to the field $\xi(\theta, \varphi)$ yields the jth band-pass-filtered field:

$$\xi_{(j)}(\theta, \varphi) = \sum_{l=0}^{l_{\max}} \sum_{m=-l}^{l} H_j(l) \widetilde{\xi}_{lm} \, Y_{lm}(\theta, \varphi).), \tag{27}$$

with $j = 1, \ldots, n_{\mathrm{band}}$.

Now, we estimate the variances of the processes $\xi_{(j)}(\theta, \varphi)$,

$$\widehat{V}_{(j)}(\theta, \varphi) := \widehat{\mathsf{Var}}\, \xi_{(j)}(\theta, \varphi), \tag{28}$$

from an ensemble (of size n_{e}) of independent fields (ensemble members) taken from the same probability distribution as the field in question $\xi(\theta, \varphi)$.

4.2.3 Estimating equations

As shown in Appendix **??**, the variances $V_{(j)}(\theta, \varphi)$ of the bandpass filtered processes $\xi_{(j)}(\theta, \varphi)$ are related to the local variance spectrum $v_l(\theta, \varphi)$ as follows:

$$V_{(j)}(\theta, \varphi) = \sum_{l=0}^{l_{\max}} H_j^2(l) v_l(\theta, \varphi) + \zeta, \tag{29}$$

where ζ is the error. This equation provides us with an "observation equation", where $V_{(j)}(\theta, \varphi)$ is the "observation", the set of $v_l(\theta, \varphi)$ for $l \in \mathcal{B}_j$ is the vector of the unknown

"truth", and the summation is the "observation operator". In these terms, the "observation error" ζ is a kind of "representativeness error": the error due to the neglect of the above mentioned "end effects" (which arises if the effective spectral width M of $\widetilde{u}_l(\theta,\varphi)$ is not negligible as compared with the waveband width $|\mathcal{B}_j|$). Note that the role of the "end effects" (ans thus the magnitude of the "representativeness error") should decrease with the growing width of the wavebands. This is why we prefer *broad* wavebands over narrow bands advocated by Priestley (1965).

Of course, the reduction of the "representativeness error" comes at a price: with *broad* bands, the spectral resolution is lower, which, in turn, implies that resulting spectrum is smoother. This smoothness of the spectrum might be a distortion of the real spectrum but it also has a positive effect. Namely, a smooth spectrum implies *fast decay* of the spatial correlations and of the above function $u(x,u)$ w.r.t. its second argument. This promotes sparseness of the the weighting matrix $\mathbf{W}$, which is highly desirable from the computational point of view.

In reality, we have only estimates of the band variances, so we rewrite Eq.(29) as

$$\boxed{\widehat{V}_{(j)}(\theta,\varphi)=\sum_{l=0}^{l_{\max}} H_j^2(l)\, v_l(\theta,\varphi)+\zeta+\varepsilon,} \tag{30}$$

where ε is the random (sampling) error due to the finite size of the ensemble.

4.3 Estimation of the spectral variances $v_l(x)$ from the set of the waveband variances $\widehat{V}_{(j)}(x)$

Having n_{band} "observations" $\widehat{V}_{(j)}(x)$ at any point x, is not enough to restore $l_{\max}+1$ spectral variances $b_l(x)\equiv\sigma_l^2(x)$ because the number of the wavebands is much smaller than the number of the unknowns. Additional information is needed to regularize the problem. We provide this additional informationby assuming that $b_l(x)$ *smoothly* change as functions of l Specifically, we formulate a variational problem for the spectra at each x independently. For this reason, we omit the dependence on x till the end of this section.

Technically, we prefer here to work with the *variance spectrum* $v_l=\frac{1}{4\pi}(2l+1)b_l$. We require that for any waveband j, the set of v_l (over $l\in[0,l_{\max}]$) is such that the following requirements are simultaneously met:

1. $\sum_{l\in\mathcal{B}_j} v_l$ is close to $\widehat{V}_{(j)}$, see Eq.(29).

2. The spectrum v_l is a *smooth* function of l. This is reasonable (i) because real-world spectra tend to be smooth, (ii) because a spectral-space smoothing is standard in building a consistent estimate of the spectral density (see, e.g., Yaglom, 1987, section 18), and (iii) because, as we noted above, a smooth spectrum implies that physical-space covariances *decay rapidly*, leading to a *sparse* weighting matrix and thus fast computations.

This is achieved by minimizing the following *loss function* for all x separately:

$$\mathcal{L}(\mathbf{v}) := \mathcal{L}_{\mathrm{e}}(\mathbf{v}) + \mathcal{L}_{\mathrm{s}}(\mathbf{v}) \to \min, \tag{31}$$

where $\mathbf{v} = (v_0, v_1, \ldots, v_{l_{\max}})$.

Now we elaborate on the two terms on the r.h.s. of Eq.(31).

1. In Eq.(31) the first term is the model misfit (i.e., a discrepancy between the local spectrum and the available data, the waveband variances $\widehat{V}_{(j)}$, see Eq.(29)):

$$\mathcal{L}_{\mathrm{e}}(\mathbf{b}) := \sum_{j=1}^{n_{\mathrm{band}}} \left(\widehat{V}_{(j)} - \sum_{l \in \mathcal{B}_j} v_l \right)^2 . \tag{32}$$

2. The second term in Eq.(31) measures the "smoothness" of the solution (the smoothness constraint. We choose $\mathcal{L}_{\mathrm{s}}(\mathbf{v})$ to penalize the second finite difference of the *log-variance-spectrum* $\log v_l$, that is the deviation of $\log v_l$ from the linear function:

$$\mathcal{L}_{\mathrm{s}}(\mathbf{b}) := w_{\mathrm{s}} \sum_{l=1}^{l_{\max}-1} \left(\log v_{l-1} - 2 \log v_l + \log v_{l+1} \right)^2 , \tag{33}$$

   where w_{s} is the tuning parameter.

   Note that the smoothness constraint Eq.(33) is the **fourth constraint** on LSM we introduce.

The loss function defined in Eqs.(31)–(33) is *non*-quadratic with respect to the vector of unknowns $\mathbf{v}$ so its minimum must be found using an iterative numerical optimization technique. The additional constraint in the optimization problem Eqs.(31)–(33) is that all spectral variances v_l need to be non-negative.

...

5 Numerical experiments with synthetic nonstationary covariances

5.1 True covariances

Here we took nonstationary 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 background spectrum b^{f}_l, we took "climatology": the time and space averaged spatial field covariances produced by DSADM.

5.2 True models

Here we describe how the two true models (a stationary model and a doubly stochastic nonstationary model, DLSM) are specified. The common external parameter is the grid size, n_x equal 360 if not otherwise stated.

5.2.1 Stationary model

The stationary model is Eq.(59) with independent complex random numbers ν_{lm} whose variances (modal spectrum) are $b_l = \widetilde{u}_l^2$, which are specified as follows:

$$b_l = \frac{c}{1 + (\lambda l)^\gamma}. \tag{34}$$

Here $\gamma > 1$ defines the *shape* of the spectrum (and, correspondingly, the shape of the covariance function), λ controls the length scale of the process, and c is the normalizing constant needed to ensure the desired variance of the random field in question. Thus, the model of the spectrum, Eq.(34), has the three *internal* parameters, c, λ, γ, to be computed from the three *external* parameters we are interested in: the variance V, the macro-scale L, and the smoothness of the process (quantified as the number r of mean-square derivatives of the underlying process $\xi(x)$). For the process covariances to be reasonably resolvable on the grid, we make sure that L is at least several times as large as the mesh size in latitude $\Delta\theta$ so that $L/\Delta\theta \geq 5$.

We now describe the specification of the internal parameters γ, λ, c (in this order).

1. γ. The random field is mean-square differentiable iff its gradient $\nabla\xi$ has finite variance. With the stationary random field, its gradient (and Laplacian) is stationary as well, so that the computation of the variance of the gradient can be computed using the Green's formula as follows:

$$\mathsf{Var}\,\nabla\xi(\theta,\varphi) = \mathbb{E}\,\frac{1}{4\pi}\int_{\mathbb{S}^2} \nabla\xi\cdot\nabla\xi\,\mathrm{d}s = -\,\mathbb{E}\,\frac{1}{4\pi}\int_{\mathbb{S}^2} \xi\nabla^2\xi\,\mathrm{d}s = -\,\mathbb{E}\,\xi\nabla^2\xi. \tag{35}$$

   Substituting ξ from Eq.(59) and using the spherical-harmonics identity $\nabla^2 Y_{lm} = -l(l+1)Y_{lm}$ into Eq.(35) yields

$$\mathsf{Var}\,\nabla\xi(\theta,\varphi) = \frac{1}{4\pi}\sum_{l=0}^{\infty} l(l+1)(2l+1)b_l. \tag{36}$$

   With b_l defined in Eq.(34), the series in Eq.(36) converges if (and only if) $\gamma > 4$. More generally, ξ is r times mean-square differentiable iff

$$\gamma > 2r + 2. \tag{37}$$

   This equation allows us to specify the internal parameter γ (from the external parameter r). Assuming that realistic fields we are dealing with in meteorology and related fields are moderately smooth, say, having $r = 1$ or 2, we specify γ in the

range from 4 to 6. As Eq.(37) provides us with a direct relationship between the smoothness of $\xi(x)$ and the parameter γ, we decide to specify γ immediately, thus replacing the (integer) external parameter r by the the more convenient continuous parameter γ. The greater γ, the smoother the process $\xi(x)$ (the less small-scale "noise" it has).

2. λ. Having specified γ and knowing the externally specified macro-scale L, we wish here to compute λ. At the moment, we define L as $L_{0.5}$, the minimal angular distance at which the covariance function (computed using Eq.(54)) implied by the spectrum Eq.(34) falls to 0.5: $B(L_{0.5}) = 0.5$. We manually tune λ to obtain the specified $L_{0.5}$. Then, we work directly with λ.

3. c. With γ and λ in hand, we finally calculate, using Eq.(56), the normalizing coefficient c from the requirement that the resulting variance be equal to V:

$$c = \frac{4\pi V}{\sum_{l=0}^{l_{\max}} \frac{2l+1}{1+(\lambda l)^{\gamma}}}. \tag{38}$$

5.2.2 A doubly stochastic locally stationary model (DLSM)

The LSM is defined by Eq.(18). In that equation, ν_{lm} are the standard Gaussian (real or complex) random numbers, so to set up the model all we need to do is to specify the local spectra $b_l(x) = \sigma_l^2(x)$ as functions of x.

To define the specific "model of truth" to be used in the below experiments, the doubly stochastic LSM (DLSM), we postulate that $b_l(x)$ are specified in the same way as in section 5.2.1 but with the three parameters V, λ, γ being functions of x. We call them the *parameter* fields. First, we simulate the parameter fields $V(x)$, $\lambda(x)$, and $\gamma(x)$ as detailed below. Second, for each grid point x, we compute $c(x), \lambda(x)$, and $\gamma(x)$ as described in section 5.2.1 above. Finally, we set

$$b_l(x) = \frac{c(x)}{1 + (\lambda(x) l)^{\gamma(x)}}. \tag{39}$$

The parameter processes $V(x)$, $\lambda(x)$, and $\gamma(x)$ are defined as transformed Gaussian processes generically written as $g(GP(x))$, where g is the transformation function and $GP(x)$ stands for a stationary Gaussian process. As the simplest choice, we postulate the "pre-transform" Gaussian processes $GP(x)$ to be proportional to band-limited Gaussian white noises as detailed below.

We define the band-limited Gaussian white noise as the random process

$$\chi(x, W) := \sum_{l=0}^{W} \sum_{m=-l}^{l} \widetilde{\chi}_{lm} \, Y_{lm}(x) \tag{40}$$

where $W \ll l_{\max}$ is the *non-stationarity width* (an external parameter). The modal spectral variances of $\chi(x, W)$, that is, $b_l^{\chi} = \mathbb{E}\, \widetilde{\chi}_{lm}$, are specified to be same for all l (note

that the term "white" refers to equal variances of all the spectral components) and are specified from the normalization requirement: $\mathsf{Var}\,\chi = \frac{1}{4\pi}\sum_{l=0}^{W}(2l+1)b_l^{\chi} = 1$.

Then, we postulate that $GP(x) = \log\varkappa \cdot \chi(x, W)$, where $\varkappa$ is a hyperparameter (see below for the justification of the use of logarithm here). Specifically, we have

$$V(x) := V_0 \cdot g(\log\varkappa_V \cdot \chi_V(x, W)), \tag{41}$$

$$\lambda(x) := \lambda_{\min} + \lambda_0 \cdot g(\log\varkappa_\lambda \cdot \chi_\lambda(x, W)), \tag{42}$$

and

$$\gamma(x) := \gamma_{\min} + \gamma_0 \cdot g(\log\varkappa_\gamma \cdot \chi_\gamma(x, W)), \tag{43}$$

where g is the transformation function such that $g(0) = 1$ (defined below), $\chi_V, \chi_\lambda, \chi_\gamma$ are the three independent white noises, the variables with the subscript 0 control the magnitude of the spatial variation of the respective parameter field. We take V_0, λ_0, γ_0 equal to the respective values of the above stationary model. $\lambda_{\min}$ is the minimal allowed length scale (set to be equal to the grid spacing Δx), $\gamma_{\min}$ is the minimal shape parameter γ we allow in order to avoid unrealistic degenerate spectra (selected to be equal to 1.5), and the coefficients $\varkappa_V, \varkappa_\lambda, \varkappa_\gamma$ determine the strength of the spatial non-stationarity. γ_0 is chosen to be in the range 2–3, which leads to a reasonable degree on variability in the shape of the local spectrum (and local correlations). With too large $\gamma > 10$, oscillations in the correlations decay unrealistically slowly. With $\varkappa_\bullet = 1$, the respective spectrum does not depend on x: $b_l(x) = b_l$. The higher $\varkappa_\bullet$, the more variable in space becomes the respective parameter: $V(x)$ (the variance 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).

The internal model parameter W is specified through the external parameter $L_{\text{nstat}} = R/W$, the *length scale of the non-stationarity*. The smaller L_{nstat}, the more rapidly changes the local spectrum in space.

Finally, we define the transformation function $g(z)$. Following (Tsyrulnikov and Rakitko, 2019), we selected the scaled and shifted logistic function (also known as the sigmoid function in machine learning):

$$g(z) := \frac{1 + \mathrm{e}^b}{1 + \mathrm{e}^{b-z}}, \tag{44}$$

where b is the constant. The function $g(z)$ has the following property: it behaves like the ordinary exponential function everywhere except for $z \gg b$, where the exponential growth is tempered (moderated). Indeed, it exponentially decays as $z \to -\infty$. Like $\exp(z)$, it is equal to 1 at $z = 0$. With $b > 0$, $g(z)$ saturates as $z \to \infty$ at the level $1 + \mathrm{e}^b$; this is the main difference of g from the exponential function and the reason why we replace $\exp(z)$ by $g(z)$: to avoid too large values in $\sigma(t, s)$, which can give rise to unrealistically large spikes in ξ. We will refer to b as the g-function's saturation hyperparameter. For $b = 1$, the function $g(z)$ in plotted in Fig.1 alongside the exponential function.

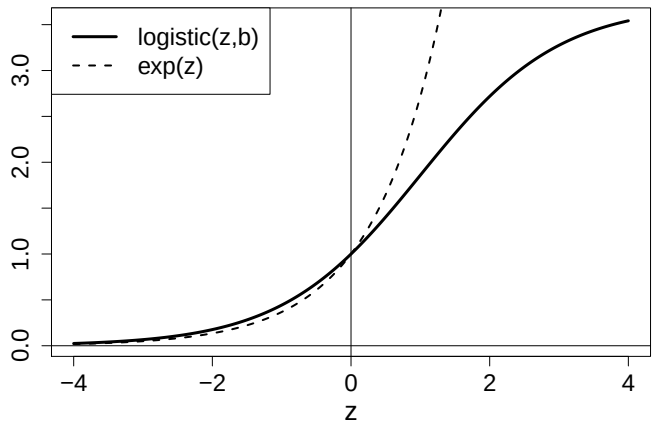

Figure 1: Logistic function $g(z)$ with $b = 1$ and exponential function

Due to nonlinearity of the transformation function g, the above transformed Gaussian white noises are non-Gaussian. Their pointwise distribution is known as logit-normal or logit-Gaussian.

As $g(z)$ (defined in Eq.(44) and shown in Fig.1) is a "tempered" exponential function, it is worth measuring the standard deviation of, the pre-transform fields on the log scale: $\mathrm{SD}(GP) = \log \varkappa$, so that the typical deviation of the transformed field from its unperturbed value is $\varkappa$ *times*.

After the processes $V(x)$, $\lambda(x)$, and $\gamma(x)$ are computed at each analysis grid point, Eq.(38) is used to find $c(x)$. With $c(x)$, $\lambda(x)$, and $\gamma(x)$ in hand, we finally compute the "true" modal spectrum $b_l(x)$ using Eq.(39). Then, following Eqs.(16) and (17), we compute

$$\widetilde{u}_l(x) \equiv \sigma_l(x) = \sqrt{b_l(x)}. \tag{45}$$

Next, we make use of Eq.(8) to apply the inverse Fourier-Legendre transform and get the function $u(x, \rho)$ (at each grid point x independently). After that, we build the $\mathbf{W}$ matrix using Eq.(12). The $\mathbf{W}$ matrix is then used both to generate the nonstationary random field using Eq.(22), to compute the covariance matrix (only if absolutely necessary) using Eq.(24), and in the analysis algorithm following Eq.(50).

Testing. In order to test the computation of the $\mathbf{W}$ matrix, we compute $\mathbf{B} = \mathbf{W}\mathbf{W}^\top$ and compare it with $\mathbf{B}$ computed in two different ways:

1. First, we set up the stationary mode by specifying all $\varkappa_\bullet = 1$. In this regime, $\mathbf{B}$ is computed using the stationary (isotropic) covariance function $B(\rho)$, see Eq.(54).

2. In the nonstationary regime, the alternative way of computing $\mathbf{B}$ is given by the nonstationary covariance function, see Eq.(19).

In addition, we visually inspect the nonstationary random field generated using Eq.(22). We check if the generated field is, indeed, (i) larger in areas where $V(x)$ is large, (ii) has

larger spatial scale in areas where $\lambda(x)$ is large, and (iii) has smaller spatial scale in areas where $\lambda(x)$ is small.

5.3 Experimental setup

The grid:

$n_x = 60$

The ensemble:

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

The DLSM:

$\overline{V} = 1$

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

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

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

$\overline{\gamma} = 5(3...6)$ (?)

$\gamma_{\min} = 1$ (?)

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

The bands:

$n_{\text{band}} = 3...4$ (?)

5.4 Accuracy of the estimating equation

The estimating Eq.(29) is exact for the stationary process $\xi(x)$ and becomes approximate when $\xi(x)$ is nonstationary. So, the accuracy of Eq.(29) certainly depends on the strength of the non-stationarity. With the DLSM, this implies that the magnitude of the error ζ in Eq.(29) depends on W and $\varkappa_\bullet$. Besides, as it follows from the analysis in Appendix **??**, ζ depends on the width of the band (because the error is due to the "end-effects", whose influence is expected to grow with the shrinking waveband $\mathcal{B}_j$).

In this section, we work with a single band, so we drop the waveband index j.

5.4.1 Methodology

The methodology here is to

1. Specify the external parameters W, $\overline{\lambda}$, and $|\mathcal{B}|$ within their ranges indicated in section 5.3. Set the other external parameters to be equal to their default values.

2. Generate realizations of the secondary fields $V(x)$, $\lambda(x)$, and $\gamma(x)$. Fix them throughout the following steps within an experiment. Calculate the local spectrum $\widetilde{u}_l(x)^2$. Use the term $\sum_{l\in\mathcal{B}} \sigma_l^2(x)$ in Eq.(29) as the ground truth $t(x)$.

3. Generate an ensemble of processes ${}^{(e)}\xi(x)$ with $e = 1, \dots, n_{\text{e}}$. By averaging over the ensemble (and possibly over x), compute the variance $v_{\mathcal{B}}(x)$ of the bandpass filtered process $\Pi_{\mathcal{B}}\xi(x)$ and use it as the "data" $d(x)$.

4. Compute the error in the "data" as $\zeta(x) = d(x) - t(x)$ (see Eq.(29)). Assess the relative mean and root-mean-square errors (denoted by μ and ρ, respectively) as

$$\mu = \frac{\sum_x \zeta(x)}{\sum_x |t(x)|}, \qquad \rho = \sqrt{\frac{\sum_x \zeta^2(x)}{\sum_x t^2(x)}}, \tag{46}$$

where the sums are over the grid on the circle.

5. Explore the dependencies of ρ on W, $\overline{L}$, $|\mathcal{B}_\bullet|$, $\varkappa_\bullet$, $n_{\rm e}$, and the random seeds on both levels in the DLSM hierarchy.

5.4.2 Results

1) Role of the band width $|\mathcal{B}|$, W, and the location of the band (say, the lower bound n_j).

Experimentally, the relative error in Eq.(29) becomes larger than some 5% when the band width $|\mathcal{B}|$ becomes less than 10. This only weakly depends on W (with the higher W, broader bands are needed). And this is almost independent of the location of the band. ...

2). Spatial resolution.

3) Role of spatial smoothing of $\widehat{V}_{(j)}(x)$.

Savitsky-Golay python smoothing employed.

Greater role for higher wvn.

Stronger optimized degree of smoothing (kernel) for small wvn.

Weaker smoo for higher W.

ker=10-15 for high wvn

30-100 for the lowest wvn

25-30 overall looks acceptable.

4) Optim nu of bands.

8 the same as 4 (?)

5.5 Accuracy of restoring the full spectrum b_l from band variances $\widehat{V}_{(j)}$

This module is debugged, tested, and tuned as follows. With DLSM at a fixed grid point x, we computed: (i) the true spectrum $b_l^{\rm true} = \widetilde{u}_l^2(x)$, (ii) the true band variances $\widehat{V}_{(j)}^{\rm true} = \sum_{l\in\mathcal{B}_j} b_l^{\rm true}$, and (iii) the restored (via solving Eq.(31)) spectrum $b_l^{\rm rstr}$.

The relative error of the restored spectrum w.r.t. the true one is defined as

$$\rho_{\rm rstr} = \frac{\sum_l |b_l^{\rm rstr} - b_l^{\rm true}|}{\sum_l |b_l^{\rm true}|}. \tag{47}$$

Results. In the default setting, $\rho_{\rm rstr}$ averaged over x and over an ensemble of realizations of the DLSM's true spectrum, turned out to be about ???

5.6 Performance of the Estimator

5.7 Efficacy of extraction of nonstationary signal

Compare with B averaged over the diagonals.

5.7.1 Results

2). Spectral resolution.

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

5.8 Analysis algorithm

Given the forecast vector $\mathbf{x}^{\mathrm{f}}$ 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

$$\mathbf{x}^{\mathrm{a}} = \mathbf{x}^{\mathrm{f}} + \mathbf{K}(\mathbf{x}^{\mathrm{o}} - \mathbf{H}\mathbf{x}^{\mathrm{f}}), \tag{48}$$

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.(24). 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 Lorenc et al. (2000).)

In the below experiments we try the following three factorizations of the $\mathbf{B}$ matrix:

1. Using the full $\mathbf{W}$ matrix as defined in Eq.(24).
2. Using a *localized* (thresholded) $\mathbf{W}$ matrix. All $(\mathbf{W})_{ij}$ less than a threshold θ_{W} in modulus are nullified.
3. Using the exact spectral matrix factor $\mathbf{\Phi}$ as defined in Eq.(**??**) (this option is used just for debugging and testing).

5.8.1 Results

Observations.

1) Point-support obs randomly located at the circle.

2) Non-point-support (non-local) obs with an square-exponential aperture function, again, randomly located on the circle (?)

6 Numerical experiments with NCEF

6.1 Model

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..

7 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.

7.1 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).

7.2 Extensions

Multivar, multi-level – with the bandpass filters, we can estm the "vertical" covariance matrices

$$\widehat{\mathbf{V}}_{(\mathbf{j})}(x) = \sum_{l \in \mathcal{B}_j} \mathbf{B_l}(x) + \boldsymbol{\zeta} \tag{51}$$

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.

Specification of a *prior distribution* for $\mathbf{W}$ or $\boldsymbol{\Sigma}$ or $\widetilde{\boldsymbol{\Sigma}}$.

8 Conclusions

...

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.

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.

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 Spectrum of a stationary random process on $\mathbb{S}^2$

A.1 Space-continuous random process

Consider a *stationary* real valued zero-mean random process $\xi(x)$ defined on the unit circle, $\mathbf{x} \in \mathbb{S}^2$. On the circle, isotropy (homogeneity, stationarity) means that the spatial covariances are invariant under rotations:

$$\mathbb{E}\,\xi(\mathbf{x})\,\xi(\mathbf{y}) = \mathbb{E}\,\xi(\mathbf{Q}\mathbf{x})\,\xi(\mathbf{Q}\mathbf{y}), \tag{52}$$

where $\mathbf{x}$ and $\mathbf{y}$ stand for vectors in $\mathbb{R}^3$ (of unit length) that represent the two points on the circle and $\mathbf{Q}$ is any orthogonal matrix.

Equation (52) implies that the covariance function depends, effectively, only on the great-circle distance $\rho(\mathbf{x}, \mathbf{y})$ between the two points:

$$\mathbb{E}\,\xi(\mathbf{x})\,\xi(\mathbf{y}) = B(\rho(\mathbf{x}, \mathbf{y})) \tag{53}$$

We expand $B(\rho)$ in the Fourier-Legendre series (Yadrenko, 1983, section 5.1) as follows

$$B(\rho) = \frac{1}{4\pi} \sum_{l=0}^{\infty} (2l+1) b_l P_l(\cos\rho). \tag{54}$$

Equation (54) is the *inverse* Fourier-Legendre transform. The *forward* Fourier-Legendre transform is then

$$b_l = 2\pi \int_{-1}^{1} B[z] P_l(z)\, \mathrm{d}z \equiv 2\pi \int_{0}^{\pi} B(\rho) P_l(\cos\rho) \sin\rho\, \mathrm{d}\rho, \tag{55}$$

where z stands for $\cos\rho$ and we adopt the notation $B(\rho) \equiv B[\cos\rho]$. We note also that Eq.(54) entails the equation for the process variance:

$$\mathsf{Var}\, \xi = \frac{1}{4\pi} \sum_{l=0}^{\infty} (2l+1) b_l. \tag{56}$$

Using the **Addition theorem** for spherical harmonics,

$$\sum_{m=-l}^{l} Y_{lm}(x) \overline{Y_{lm}}(y) = \frac{1}{4\pi} (2l+1) P_l(\cos\rho(x,y)) \tag{57}$$

and the Karhunen theorem, we can obtain the following spectral expansion of the random field in question (Yadrenko, 1983, section 5.1):

$$\xi(\theta,\varphi) := \sum_{l=0}^{\infty} \sum_{m=-l}^{l} \widetilde{\xi}_{lm}\, Y_{lm}(\theta,\varphi) \tag{58}$$

with ξ_{lm} all mutually uncorrelated complex valued random variables such that $\mathbb{E}\, \widetilde{\xi}_{lm} = 0$ and $\mathsf{Var}\, \widetilde{\xi}_{lm} = b_l$.

Denoting $\widetilde{u}_l = \sqrt{b_l}$, we rewrite Eq.(58) as

$$\xi(\theta,\varphi) := \sum_{l=0}^{\infty} \sum_{m=-l}^{l} \widetilde{u}_l \alpha_{lm}\, Y_{lm}(\theta,\varphi) \tag{59}$$

Here $\alpha_{lm} = \widetilde{\xi}_{lm} / \widetilde{u}_l$ are independent zero-mean and unit-variance Gaussian random variables. For $m = 0$ these are real valued, whereas for $m \neq 0$ complex valued with identically distributed and uncorrelated real and imaginary parts. In other words, $\alpha_l^0 \sim N(0,1)$ and for $m \neq 0$, $\alpha_{lm} \sim CN(0,1)$ (here CN states for the circularly symmetric complex Gaussian (normal) random variable (e.g. Tse and Viswanath, 2005)).

The space discrete (gridded) random field is obtained by limiting the support of $\widetilde{u}_l$, to the range of total wavenumbers from $l = 0$ to $l = l_{\max}$ in Eqs.(58) and (59). To represent these band-limited functions we use the regular latitude-longitude grid with $n_{\mathrm{lat}} = l_{\max} + 1$ points over latitude (including both poles) and $n_{\mathrm{lon}} = 2l_{\max}$ points at each latitude circle.

A.2 Kernel convolution

Equation (59) implies that $\xi(\theta, \varphi)$ can be represented as the convolution of the isotropic kernel

$$u(\rho) = \frac{1}{4\pi} \sum_{l=0}^{l_{\max}} (2l+1) \widetilde{u}_l P_l(\cos \rho) \tag{60}$$

(called the *convolution square root* of $B(\rho)$ since $\widetilde{u}_l = \sqrt{b_l}$) with the white noise process

$$\alpha(\theta, \varphi) := \sum_{l=0}^{l_{\max}} \sum_{m=-l}^{l} \alpha_{lm} Y_{lm}(\theta, \varphi), \tag{61}$$

so that

$$\xi(\mathbf{s}) = \int_{\mathbb{S}^2} u(\rho(\mathbf{s}, \mathbf{s}')) \, \alpha(\mathbf{s}') \, \mathrm{d}\mathbf{s}'. \tag{62}$$

Equation (62) is a very general model: according to Yaglom (1987, ...) it can represent any random field that has spectral density (the spectrum b_l on $\mathbb{S}^2$). Banerjee et al. (2014, section 3.1.4) 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 $l_{\max} = \infty$ 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.(60) to converge at $\rho = 0$. But if we truncate the series in Eq.(60) 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.

Moreover, the band-limited convolution square root $u(\rho)$ exists for any correlation function $B(\rho)$ because the spectrum $\widetilde{u}_l$ is square summable (note that $\frac{1}{4\pi} \sum (2l+1) \widetilde{u}_l^2 = B(0)$) and thus $u(\rho)$ is square integrable... This band-limited convolution square root is well defined in the sense that its self-convolution $u * u$ perfectly reproduces the grid-point values of $B(\rho)$ at any resolution...

References

S. Banerjee, B. P. Carlin, and A. E. Gelfand. *Hierarchical modeling and analysis for spatial data*. CRC press, 2014.

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.

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

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.

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.

G. P. Nason, R. Von Sachs, and G. Kroisandt. Wavelet processes and adaptive estimation of the evolutionary wavelet spectrum. *J. Roy. Statist. Soc.: Ser. B*, 62(2):271–292, 2000.

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.

P. D. Spanos, J. Tezcan, and P. Tratskas. Stochastic processes evolutionary spectrum estimation via harmonic wavelets. *Computer Methods in Applied Mechanics and Engineering*, 194(12):1367–1383, 2005.

D. Tse and P. Viswanath. *Fundamentals of wireless communication*. 2005.

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. doi: 10.1002/QJ.3556.

M. Vogt et al. Nonparametric regression for locally stationary time series. *The Annals of Statistics*, 40(5):2601–2633, 2012.

D. C. Webb. The analysis of non stationary data using complex demodulation. volume 34, pages 131–137, 1979.

M. A. Wieczorek and F. J. Simons. Localized spectral analysis on the sphere. *Geophysical Journal International*, 162(3):655–675, 2005.

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

A. M. Yaglom. *Correlation theory of stationary and related random functions, Volume 1: Basic results*. Springer Verlag, 1987.