The ensemble Kalman filter regularized with nonstationary nonparametric convolutions

Working paper

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

July 7, 2021

1 Introduction

2 Notation

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

On $\mathbb{S}^1$, we represent the *background error* field, ξ, on the regular grid with $n_{\rm x}$ points.

On $\mathbb{S}^2$, we represent ξ 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 $n_{\rm e}$.

$n_{\rm e}$ stands for the ensemble size.

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.

On $\mathbb{S}^1$, the forward spectral (discrete Fourier) transform of a function, $f(x)$, is denoted by $\mathcal{F}_{x\to l} : f(x) \mapsto \widetilde{f}_l$. The inverse transform is denoted by $\mathcal{F}_{l\to y} : \widetilde{f}_l \mapsto f(x)$.

On $\mathbb{S}^2$, the forward spectral (spherical harmonic) transform of a function, $f(x)$, we denote by $\mathcal{S}_{x\to lm} : f(x) \mapsto \widetilde{f}_{lm}$. The inverse transform is denoted by $\mathcal{S}_{lm\to x} : \widetilde{f}_{lm} \mapsto f(x)$. 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.

By default, we derive equations for the spherical case and then give their analogs for the case on the circle.

Background error *biases* are beyond te scope of this study, so we assume that $\xi(x)$ is a mean zero process.

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 $w(x,y)$ is a real function (called the convolution kernel or the weighting function). $\forall x \in D$ ($D = \mathbb{S}^1$ or $\mathbb{S}^2$ is the domain on which ξ is defined), $w(x,y)$ 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 $G = \{\mathbf{s}_i\}_{j=1}^{n_\mathrm{x}}$, where n_x is the number of grid points and $\mathbf{s}_i$ 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 $\mathbf{W}' = \mathbf{W}\mathbf{Q}$, 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 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) 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 at all, 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 real valued white noise $\alpha(x)$ on $\mathbb{S}^2$:

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

where L 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 $\mathsf{Var}\,\widetilde{\alpha}_{lm} = 1$. More specifically, $\widetilde{\alpha}_{l0}$ are real valued and all the other $\widetilde{\alpha}_{lm}$

are complex circularly symmetric random variables such that (since $\alpha(x)$ is real valued) $\widetilde{\alpha}_{l,-m} = \widetilde{\alpha}^*_{lm}$ (where * denotes complex conjugation).

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

$$u(x, \rho) = \frac{1}{4\pi} \sum_{l=0}^{L} (2l + 1)\sigma_l(x) P_l(\cos \rho). \tag{8}$$

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

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

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

$$\boxed{\xi(x) = \sum_{l=0}^{L} \sum_{m=-l}^{l} \sigma_l(x)\, \widetilde{\alpha}_{lm}\, Y_{lm}(x).} \tag{10}$$

Note that from Eq.(8) it follows that $\sigma_l(x)$ are real valued.

On $\mathbb{S}^1$, the analogs of Eqs.(8) and (9) are

$$u(x, \Delta) = \frac{1}{\sqrt{2\pi}} \sum_{l=-L+1}^{L} \sigma_l(x)\, \mathrm{e}^{\mathrm{i}l\Delta} \tag{11}$$

and

$$\xi(x) = \sum_{l=-L+1}^{L} \sigma_l(x)\, \widetilde{\alpha}_l\, \mathrm{e}^{\mathrm{i}lx}. \tag{12}$$

Here all the $\widetilde{\alpha}_l$ are mutually independent complex-valued random Fourier coefficients with mean zero and variance one. More specifically, $\widetilde{\alpha}_0$ and $\widetilde{\alpha}_L$ are real valued, whilst all the others are complex circularly symmetric random variables (e.g. Searle and Khuri, 2017, section 9.5). With the real valued $\xi(x)$, it holds that $\widetilde{\alpha}_{-l} = \widetilde{\alpha}^*_l$. Finally, for Eq.(5) to hold in on $\mathbb{S}^1$, $u(x, \Delta)$ should be an even function of its second argument.

Note that the restriction Eq.(5) is the **first constraint** we impose on the general convolution model. It implies, in particular, that $\sigma_l(x)$ are real valued in both the spherical and the circular case. On $\mathbb{S}^1$ we have in addition that $\sigma_{-l}(x) = \sigma_l(x)$ (because the kernel $u(x, \Delta)$ is real valued by construction).

The space discrete equivalent of Eq.(6) is

$$\xi_i = \sum_j u(x_i, \rho(x_i, y_j))\, Z(\Delta y_j) = \sum_j u(x_i, \rho(x_i, y_j))\, \sqrt{\Delta y_j}\alpha_j \equiv \sum_j w_{ij}\alpha_j, \tag{13}$$

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

$$w_{ij} = u(x_i, \rho(x_i, y_j))\, \sqrt{\Delta y_j}. \tag{14}$$

3.4 Random field'd local isotropy

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:

$$\mathsf{Var}\,(\zeta(x; x_0) - \xi(x)) < C\rho^2(x, x_0), \tag{15}$$

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 $\sigma_l(x)$ vary slowly (smoothly) with x. Indeed, for any fixed x_0, the process

$$\zeta(x; x_0) = \sum_{l,m} \sigma_l(x)\, \widetilde{\alpha}_{lm}\, Y_{lm}(x) \tag{16}$$

is stationary and

$$\mathsf{Var}\,(\zeta(x; x_0) - \xi(x)) = \sum_{l=0}^{L} (\sigma_l(x_0) - \sigma_l(x))^2 \sum_{m=-l}^{l} Y_{lm}(x) Y^*_{lm}(x_0) \le C\,\rho^2(x, x_0) \tag{17}$$

if $\sigma_l(x)$ are Lipschitz continuous, uniformly in l. We ensure the uniform Lipschitz continuity by limiting the support of the spectrum of $\sigma_l(x)$ as a function of x, see Eq.(23), and by requiring that $\mathsf{Var}\,\xi(x)$ is uniformly bounded (see Appendix **??** for details). The inequality in Eq.(17) is justified by the boundedness of the spectrum $\mathcal{S}_{x\to mn}\sigma_l(x)$, the addition theorem for spherical harmonics, and the bounds $|P_l(\cos\rho)| \le 1$.

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

3.5 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.(111) we find a significant similarity. Indeed, the only difference is $\sigma_l(x)$ in Eq.(10) vs. σ_l in Eq.(111). Thus, the difference of the process convolution model from the isotropic model is in the dependency, in the latter, of the spectral standard deviations σ_l on x. Since in the isotropic model $b_l = \sigma_l^2$ is called the (modal) spectrum, we call $b_l(x) := \sigma_l^2(x)$ the *local (modal) spectrum* (first introduced by (Priestley, 1965) and called "evolutionary spectrum" in the time series context). We call the model, Eq.(10), the Local Spectrum Model (LSM).

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:

$$B(x, x') := \mathbb{E}\,\xi(x)\xi(x') = \frac{1}{4\pi} \sum_{l=0}^{L} (2l+1)\,\sigma_l(x)\,\sigma_l(x')\,P_l(\cos\rho(x, x')). \tag{18}$$

In particular,

$$\mathsf{Var}\,\xi(x) = \frac{1}{4\pi}\sum_{l=0}^{L}(2l+1)\,b_l(x) = \sum_{l=0}^{L} f_l(x), \tag{19}$$

where

$$f_l(x) = \frac{1}{4\pi}(2l+1)b_l(x) \tag{20}$$

is the local *variance* spectrum. On $\mathbb{S}^1$,

$$\mathsf{Var}\,\xi(x) = \sum_{l=-L+1}^{L} b_l(x) \tag{21}$$

where $b_l(x) = \sigma_l(x)^2 = f_l$.

We also notice that Eq.(8), which relates the kernel $u(x,\rho)$ to the local spectrum $b_l = \sigma_l^2(x)$ constitutes the inverse Fourier-Legendre transform $\mathcal{S}_{l\to\rho} : \sigma_l(x) \mapsto u(x,\rho)$. The respective forward transform $\mathcal{S}_{\rho\to l}$ reads

$$\sigma_l(x) = 2\pi\int_{-1}^{1} u(x,\rho)\,P_l(\cos\rho)\sin\rho\,\mathrm{d}\rho. \tag{22}$$

This equation implies that $\sigma_l(x)$ are non-negative if, for any x, $u(x,\rho)$ is a non-negative definite function of the angular distance ρ. We will make this latter assumption, $\sigma_l(x) \geq 0$ as the our **second constraint** imposed on the general process convolution model (in both spherical and circular cases).

3.6 Physical-space smoothness of the local spectrum

Of course, the term "local spectrum" is justified only if $\sigma_l(x)$ 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 $\sigma_l(x)$ w.r.t. x is confined to the wavenumber range $|n| \leq M \ll L$:

$$\sigma_l(x) = \sum_{p=0}^{M}\sum_{q=-p}^{p} \widetilde{\sigma}_l^{pq}\,Y_{pq}(x) \tag{23}$$

The limitation $M \ll L$ is our **third 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 $\mathbb{S}^2$).

3.7 Implications for data assimilation

Equation (13) implies that the nonstationary space-discrete random vector $\boldsymbol{\xi} = (\xi_1,\ldots,\xi_{n_\mathrm{x}})$ satisfies

$$\boldsymbol{\xi} = \mathbf{W}\boldsymbol{\alpha}, \tag{24}$$

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

$$(\mathbf{W})_{ij} := w_{ij} \tag{25}$$

are defined in Eq.(14). Then, the covariance matrix $\mathbf{B}$ of the random vector $\boldsymbol{\xi}$ (whose entries are grid-point values of $\xi(x)$) becomes, obviously,

$$\mathbf{B} = \mathbf{W}\,\mathbf{W}^{\top}. \tag{26}$$

The representation of the background-error covariance matrix $\mathbf{B}$ in the "square-root" form, Eq.(26) 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 **fourth constraint**.

3. The computation of rows of matrix $\mathbf{W}$ from the (estimated online) local spectra $\sigma_l(x)$ 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 $\sigma_l(x)$ 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 $f_l(x)$ 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: outline

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

We use f_l generically for both b_l on the circle and the varoiance spectrum e_l on the sphere.

The goal is to estimate the local spectrum $f_l(x)$ for any x we wish.

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.(10) for the wavenumber bands $j = 1, \dots, J$, getting the respective bandpass filtered processes $\xi_{(j)}(x)$.

2. For any point of interest x, estimate the waveband *variances* $\mathsf{Var}\,\xi_{(j)}(x)$ (for $j = 1, \dots, J$) from the ensemble, relate them to $\{f_l(x)\}$ and produce the estimator, $\widehat{f_l}(x)$.

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).

4.3 Spectral bandpass filters

We introduce J filters $\mathcal{H}_j$, where $j = 1, \dots, J$. The j-th filter is characterized by its real valued spectral *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 applied to an *isotropic* random field (on the sphere), the filtered field will be *isotropic* as well.

The bands' transfer functions $H_j(l)$ can overlap and cover the whole wavenumber range resolved by the analysis grid.

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

$$\xi(\theta,\phi) = \sum_{l=0}^{L} \sum_{m=-l}^{l} \widetilde{\xi}_{lm}\, Y_{lm}(\theta,\phi) \tag{27}$$

is

$$\boxed{(\mathcal{H}\xi)(\theta,\phi) = \sum_{l=0}^{L} H(l) \sum_{m=-l}^{l} \widetilde{\xi}_{lm}\, Y_{lm}(\theta,\phi).} \tag{28}$$

On $\mathbb{S}^1$, we have

$$(\mathcal{H}\xi)(x) = \sum_{l=-L+1}^{L} H(l)\, \widetilde{\xi}_l\, \mathrm{e}^{\mathrm{i}lx}. \tag{29}$$

An important technical difference between Eqs.(28) and (29) is that on $\mathbb{S}^2$, the filtered process $\mathcal{H}\xi$ is always real valued, whereas on $\mathbb{S}^1$, the filtered process is complex valued unless the filter's spectral transfer function $H(l)$ is an even function.

4.4 Preliminary parametric estimation of local spectrum

Applying the filter $\mathcal{H}_j$ $(j = 1, \ldots, J)$ to the field $\xi(\theta,\phi)$ yields the jth band-pass-filtered field:

$$\xi_{(j)}(x) \approx \sum_{l=0}^{L} \sum_{m=-l}^{l} H_j(l)\, \widetilde{\xi}_{lm}\, Y_{lm}(x). \tag{30}$$

As shown in Appendix **??**, if ξ obeys the LSM, Eq.(10), then the variances of the bandpass filtered processes $\xi_{(j)}(x)$ are related to the local variance spectrum $f_l(x)$ as follows:

$$\mathsf{Var}\, \xi_{(j)}(x) \approx \sum_{l=0}^{L} H_j^2(l)\, f_l(x). \tag{31}$$

Note that on the circle, the lower summation limit in this equation is $-L+1$ (and the same is true for Eq.(33) below). Note also that $\mathsf{Var}\, \xi_{(j)}(x)$ are (by definition of variance) real valued on both the sphere (where the filtered processes are real valued) and on the circle (where the filtered processes can be complex valued).

On the other hand, having an ensemble (i.e., a sample of size n_e) of 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,

$$d_j(x) := \widehat{\mathsf{Var}}\, \xi_{(j)}(x), \tag{32}$$

(d stands for "data" we have from the ensemble) at each x independently.

Equating $\mathsf{Var}\, \xi_{(j)}(x)$ to $d_j(x)$ (see Eqs.(31)–(32)) provides us with the "observation equation" w.r.t. the variance spectrum f_l:

$$d_j = \sum_{l=0}^{L} H_j^2(l) f_l + \text{error}. \tag{33}$$

Here error stands for both methodological error (involved in Eq.(31)) and sampling error (involved in Eq.(32)). We have dropped the dependencies on the spatial grid point x because the estimation of the local spectrum is going to be performed independently for different x (for computational reasons in a real-world very high-dimensional problem).

Denoting $H_j^2(l) =: \omega_{jl}$, we rewrite Eq.(33) in the vector-matrix form as

$$\mathbf{\Omega}\,\mathbf{f} = \mathbf{d}. \tag{34}$$

Here $\mathbf{d}$ 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 (34) constitutes the standard linear inverse problem, in which $\mathbf{d}$ 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 $\mathbf{d}$. A reasonable way to solve a linear inverse problem such as Eq.(34) is to seek the *minimal-norm least-squares* solution, that is, to use pseudo inversion:

$$\mathbf{f}^+ = \mathbf{\Omega}^+\,\mathbf{d}, \tag{35}$$

where $\mathbf{\Omega}^+$ is the Moore-Penrose pseudo-inverse matrix. It is well known that with the singular value decomposition $\mathbf{\Omega} = \mathbf{U}_\Omega \mathbf{\Sigma}_\Omega \mathbf{V}_\Omega^\top$, the pseudo-inverse matrix is

$$\mathbf{\Omega}^+ = \mathbf{U}_\Omega \mathbf{\Sigma}_\Omega^{-1} \mathbf{V}_\Omega^\top. \tag{36}$$

Here $\mathbf{U}_\Omega$ is a $J \times r$ matrix (where r is the rank of $\mathbf{\Omega}$) with orthonormal columns, $\mathbf{\Sigma}_\Omega$ is an $r \times r$ diagonal matrix with positive diagonal entries (singular values placed in order of decreasing magnitude), and $\mathbf{V}_\Omega$ is an $N \times r$ matrix with orthonormal columns. Note that, with the reasonably selected bands, $r = J$ so that $\mathbf{U}_\Omega, \mathbf{\Sigma}_\Omega, \mathbf{V}_\Omega$ are, normally, full-rank matrices.

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 also (ii) monotonicity, (iii) resemblance to the typical *shape* of the spectrum, and even (iv) smoothness of the spectrum. The simplest way to impose these four constraints is to fit a parametric model to $\{f_l^+\}_{l=0}^L \equiv \mathbf{f}^+$.

We chose to fit a two-parameter scale-magnitude model: $g_l \approx A \cdot g(l/a)$, 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.

4.5 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 (independently at each x).

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.

4.5.1 Prior

As it is common in spectral analysis (.), we place a prior on the *log-spectrum* $\lambda_l = \log f_l$. We postulate that λ is a stationary Gaussian process of the log-wavenumber variable $s_l = \log(l + l_0)$ (where l_0 is introduced to permit treatment of $l = 0$) so that

$$\lambda(s) \sim GP(s; \bar{\lambda}(s), K). \tag{37}$$

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

As for the kernel K, we specify it implicitly by, first, penalizing deviations from $\bar{\lambda}(s)$, 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,

$$\mathcal{L}^{\text{prior}}(\boldsymbol{\lambda}) = \mathcal{L}^{\text{prior}}_{\text{clim}}(\boldsymbol{\lambda}) + \mathcal{L}^{\text{prior}}_{\text{smoo}}(\boldsymbol{\lambda}), \tag{38}$$

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

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

$$\mathcal{L}^{\text{prior}}_{\text{clim}}(\boldsymbol{\lambda}) = \frac{w_{\text{c}}}{2} (\boldsymbol{\lambda} - \bar{\boldsymbol{\lambda}}, \boldsymbol{\lambda} - \bar{\boldsymbol{\lambda}}), \tag{39}$$

where w_{c} is the weight (a tuning parameter) and the inner product $(.,.)$ is defined as

$$(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2) = \sum_l \lambda_1(s_l)\lambda_2(s_l)\, \Delta^{\text{c}}_l \tag{40}$$

(with Δ^{c}_l denoting the grid-cell size: $\Delta^{\text{c}}_l = (s_{l-1} + s_{l+1})/2$, where $s_{-1} = s_0$ and $s_{L+1} = s_L$) to be consistent with the continuous-s inner product $\int \lambda_1(s)\lambda_2(s)\, \mathrm{d}s$.

Thus,

$$\boxed{\mathcal{L}^{\text{prior}}_{\text{clim}}(\boldsymbol{\lambda}) = \frac{w_{\text{c}}}{2} \sum_{l=0}^{L} \Delta^{\text{c}}_l (\lambda_l - \bar{\lambda}_l)^2.} \tag{41}$$

The smoothness constraint is defined as a discrete analog of the non-negative definite functional:

$$\frac{w_{\text{s}}}{2} \int (\mathcal{D}\lambda(s))^2\, \mathrm{d}s \equiv \frac{w_{\text{s}}}{2} (\mathcal{D}\lambda(s), \mathcal{D}\lambda(s)), \tag{42}$$

where w_s is the tunable weight and $\mathcal{D}$ is the first-order differentiation operator. We specify

$$\mathcal{L}^{\mathrm{prior}}_{\mathrm{smoo}}(\boldsymbol{\lambda}) = \frac{w_\mathrm{s}}{2}\,(\mathbf{D}\boldsymbol{\lambda}, \mathbf{D}\boldsymbol{\lambda}) \equiv \frac{w_\mathrm{s}}{2}\,\sum_{l=0}^{L-1} (\mathbf{D}\boldsymbol{\lambda})_l^2\,\Delta s_l, \tag{43}$$

where $\mathbf{D}$ is a finite difference analog of $\mathcal{D}$ and $\Delta s_l = s_{l+1} - s_l$. With $(\mathbf{D}\boldsymbol{\lambda})_l = \frac{\lambda_{l+1}-\lambda_l}{\Delta s_l}$, we have

$$\boxed{\mathcal{L}^{\mathrm{prior}}_{\mathrm{smoo}}(\boldsymbol{\lambda}) = \frac{w_\mathrm{s}}{2}\,\sum_{l=0}^{L-1} \frac{(\lambda_{l+1}-\lambda_l)^2}{\Delta s_l}.} \tag{44}$$

Remark 1. It can be seen that the relation between the weights w_c and w_s determines the effective length scale of the Gaussian prior process $\lambda(s)$.

Remark 2. It might appear more useful to penalize the second derivative in the smoothness constrain. It is not hard to see that this would lead to a smoother behavior of the kernel K near the origin.

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

In the spherical case, the set of passband filtered processes $\xi_{(j)}(x)$ with $j = 1, \ldots, J$ taken at the same x is a zero mean real valued Gaussian vector fully characterized by their covariances (derived from Eq.(30) using the addition theorem for spherical harmonics)

$$(\xi_{(j)}(x), \xi_{(j')}(x)) \approx \sum_{l=0}^{L} H_j(l)\,H'_j(l)\,f_l(x). \tag{45}$$

In the matrix-vector form, this equation reads

$$\boldsymbol{\Gamma} = \mathbb{E}\,\boldsymbol{\xi}_{()}\boldsymbol{\xi}_{()}^\top \approx \mathbf{H}\mathbf{F}\mathbf{H}^\top = \mathbf{H}\mathrm{e}^{\boldsymbol{\Lambda}}\mathbf{H}^\top, \tag{46}$$

where $\boldsymbol{\xi}_{()} = (\varphi_1(x), \ldots, \varphi_J(x))^\top$, $\mathbf{F} = \mathsf{diag}(\mathbf{f}) \equiv \mathsf{diag}(\mathrm{e}^{\boldsymbol{\lambda}}) \equiv \mathrm{e}^{\Lambda}$, and $\boldsymbol{\Lambda} = \mathsf{diag}(\boldsymbol{\lambda})$. Note that $\boldsymbol{\Gamma}$ is a full-rank $J \times J$ matrix. In this equation and till the end of this section the dependencies on x are dropped.

We consider $\boldsymbol{\xi}_{()}$ at the given x as the *data* or *observations* we have from the ensemble *on* the spectrum we seek to estimate. Since $\xi(x)$ is a zero-mean Gaussian random process, $\boldsymbol{\xi}_{()}$ is multivariate Gaussian distributed. Therefore, we can easily write down the *likelihood* of the log-spectrum, $\boldsymbol{\lambda}$, given the ensemble, i.e., the probability density of the set of n_e fields $\boldsymbol{\xi}^{[\mu]}_{()}$, where $\mu = 1, \ldots, n_\mathrm{e}$ stands for the ensemble member:

$$\mathrm{lik}(\boldsymbol{\lambda}|\boldsymbol{\xi}_{()}) \propto \prod_{\mu} \frac{1}{|\boldsymbol{\Gamma}|^{1/2}} \mathrm{e}^{-\frac{1}{2}\boldsymbol{\xi}^{[\mu]\top}_{()}\boldsymbol{\Gamma}^{-1}\boldsymbol{\xi}^{[\mu]}_{()}}, \tag{47}$$

with $|.|$ denoting the matrix determinant. The minus log-likelihood, which we denote by $\mathcal{L}^{\mathrm{lik}}$, reads:

$$\mathcal{L}^{\mathrm{lik}}(\boldsymbol{\lambda}) = \mathcal{L}^{\mathrm{lik}}_{\mathrm{det}}(\boldsymbol{\lambda}) + \mathcal{L}^{\mathrm{lik}}_{\mathrm{tr}}(\boldsymbol{\lambda}), \tag{48}$$

where

$$\boxed{\mathcal{L}^{\text{lik}}_{\text{det}}(\boldsymbol{\lambda}) = \frac{n_{\text{e}}}{2}\,\log|\boldsymbol{\Gamma}|} \tag{49}$$

and

$$\mathcal{L}^{\text{lik}}_{\text{tr}}(\boldsymbol{\lambda}) = \sum_{\mu} \boldsymbol{\xi}_{()}^{[\mu]\top}\boldsymbol{\Gamma}^{-1}\boldsymbol{\xi}_{()}^{[\mu]} \equiv \frac{1}{2}\,\text{tr}\left(\sum_{\mu}\boldsymbol{\xi}_{()}\boldsymbol{\xi}_{()}^{\top}\boldsymbol{\Gamma}^{-1}\right) = \frac{n_{\text{e}}}{2}\,\text{tr}(\mathbf{S}\boldsymbol{\Gamma}^{-1}), \tag{50}$$

where $\mathbf{S} = \frac{1}{n_{\text{e}}}\sum \boldsymbol{\xi}_{()}^{[\mu]}\boldsymbol{\xi}_{()}^{[\mu]\top}$ is the non-centered (since we have assumed that ξ is unbiased) sample covariance matrix. In practice, when the $\mathbb{E}\,\boldsymbol{\xi}_{()}$ is uncertain, a more common unbiased estimate for the sample covariance matrix can be used:

$$\mathbf{S}_{\varphi} = \frac{1}{n_{\text{e}}-1}\sum_{\mu=1}^{n_{\text{e}}}\boldsymbol{\varphi}^{[\mu]}\boldsymbol{\varphi}^{[\mu]\top}, \tag{51}$$

where $\boldsymbol{\varphi}^{[\mu]} = \boldsymbol{\xi}_{()}^{[\mu]} - \bar{\boldsymbol{\xi}}_{()}$ are the centered ensemble members (a.k.a. ensemble perturbations) in "band space", with $\bar{\boldsymbol{\xi}}_{()}$ being the ensemble mean. As a result, we may rewrite Eq.(50) as

$$\boxed{\mathcal{L}^{\text{lik}}_{\text{tr}}(\boldsymbol{\lambda}) = \frac{n_{\text{e}}}{2(n_{\text{e}}-1)}\,\text{tr}(\boldsymbol{\Phi}\boldsymbol{\Phi}^{\top}\boldsymbol{\Gamma}^{-1}),} \tag{52}$$

where $\boldsymbol{\Phi}$ is a $J \times n_{\text{e}}$ matrix whose columns are the centered bandpass filtered ensemble members $\boldsymbol{\varphi}^{[\mu]}$.

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

In the circular case, the passband filtered processes φ_j are complex valued, so we decompose them into the real and imaginary parts: $\varphi_j(x) = \varphi'_j + \mathsf{i}\varphi''_j$. Then, the LSM implies that φ'_j is uncorrelated with $\varphi''_{j'}$ (for all wavenumber-band pairs (j, j') - check again), so that the likelihood $p(\boldsymbol{\varphi}|\boldsymbol{\lambda})$ is the product of the two likelihoods: $p(\boldsymbol{\varphi}'|\boldsymbol{\lambda})$ and $p(\boldsymbol{\varphi}''|\boldsymbol{\lambda})$. The covariance matrices of $\boldsymbol{\varphi}'$ and $\boldsymbol{\varphi}''$ are, respectively,

$$\boldsymbol{\Gamma}' \approx \mathbf{H}'\mathbf{F}\mathbf{H}'^{\top} \qquad \text{and} \qquad \boldsymbol{\Gamma}'' \approx \mathbf{H}''\mathbf{F}_{-2}\mathbf{H}''^{\top}, \tag{53}$$

where $\mathbf{F}_{-2} = \mathsf{diag}(f_1, \ldots, f_{L-1})$, i.e., the entries f_0 and f_L are omitted in the diagonal of $\mathbf{F}_{-2}$ (because with the discrete Fourier transform, these two spectral components have zero imaginary parts). In other words, the only non-zero elements of $\mathbf{F}_{-2}$ are $(\mathbf{F}_{-2})_{kk} = f_{k+1}$ with $k = 1, ..., L-2$. Correspondingly, $\mathbf{H}''$ has two columns less than $\mathbf{H}'$, i.e., $L-1$ columns.

$\mathbf{H}'$ is a $J \times (L+1)$ matrix defined as follows. $(\mathbf{H}')_{j1} = H_j(0)$, $(\mathbf{H}')_{j,L+1} = H_j(L)$, and for $l \in [1, L-1]$ we have

$$(\mathbf{H}')_{j,l+1} = \frac{1}{\sqrt{2}}(H_j(l) + H_j(-l)). \tag{54}$$

With the $\mathbf{H}''$ matrix, we note that it not only has two columns less than $\mathbf{H}'$ as discussed above, it also normally has two rows less than $\mathbf{H}'$. The reason is that we specify the first and the last band's transfer functions to be *even* functions of the wavenumber in the sense

that for these two bands $H(l) = H(n_\mathrm{x} - l)$. With these transfer function, the imaginary parts of the two passband filtered processes vanish, therefore we omit the first and the last entries from the vector φ''_j and omit the first and the last rows from the matrix $\mathbf{H}''$. As a result, $\mathbf{H}''$ is a $(J-2) \times (L-1)$ matrix defined for $l \in [1, L-1]$ as

$$(\mathbf{H}'')_{j,l} = \frac{1}{\sqrt{2}}(H_j(l) - H_j(-l)). \tag{55}$$

As a result, we obtain,

$$\boxed{\mathcal{L}^{\mathrm{lik}}_{\mathrm{det}}(\boldsymbol{\lambda}) = \frac{n_\mathrm{e}}{2}(\log|\boldsymbol{\Gamma}'| + \log|\boldsymbol{\Gamma}''|)} \tag{56}$$

and

$$\boxed{\mathcal{L}^{\mathrm{lik}}_{\mathrm{tr}}(\boldsymbol{\lambda}) = \frac{n_\mathrm{e}}{2(n_\mathrm{e}-1)}\left[\mathrm{tr}(\boldsymbol{\Phi}'\boldsymbol{\Phi}'^{\top}\boldsymbol{\Gamma}'^{-1}) + \mathrm{tr}(\boldsymbol{\Phi}''\boldsymbol{\Phi}''^{\top}\boldsymbol{\Gamma}''^{-1})\right].} \tag{57}$$

4.5.4 Posterior

The minus log-posterior (cost function) $\mathcal{L}(\boldsymbol{\lambda})$ is, obviously, the sum of the above four components, two from the prior, $\mathcal{L}^{\mathrm{prior}}_{\mathrm{clim}}(\boldsymbol{\lambda})$ and $\mathcal{L}^{\mathrm{prior}}_{\mathrm{smoo}}(\boldsymbol{\lambda})$, and two from the likelihood, $\mathcal{L}^{\mathrm{lik}}_{\mathrm{det}}(\boldsymbol{\lambda})$ and $\mathcal{L}^{\mathrm{lik}}_{\mathrm{tr}}(\boldsymbol{\lambda})$. 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

$$\mathcal{L}(\boldsymbol{\lambda}) = \mathcal{L}^{\mathrm{prior}}_{\mathrm{clim}}(\boldsymbol{\lambda}) + \mathcal{L}^{\mathrm{prior}}_{\mathrm{smoo}}(\boldsymbol{\lambda}) + \mathcal{L}^{\mathrm{lik}}_{\mathrm{det}}(\boldsymbol{\lambda}) + \mathcal{L}^{\mathrm{lik}}_{\mathrm{tr}}(\boldsymbol{\lambda}) \to \min. \tag{58}$$

4.5.5 Numerical solution

To facilitate the numerical minimization in Eq.(58) we derive gradients and Hessians of all the four components of the cost function as follows. (We treat here only the spherical case, for modifications in the circular case are straightforward, see section 4.5.3.)

1. Differentiating the "climatological" part of the cost function, Eq.(41), is easy:

$$\boxed{\frac{\partial \mathcal{L}^{\mathrm{prior}}_{\mathrm{clim}}}{\partial \boldsymbol{\lambda}} = w_\mathrm{c}\,\boldsymbol{\Delta}^\mathrm{c} \circ (\boldsymbol{\lambda} - \bar{\boldsymbol{\lambda}}),} \tag{59}$$

   where $\boldsymbol{\Delta}^\mathrm{c} = (\Delta^\mathrm{c}_0, \ldots, \Delta^\mathrm{c}_L)$ and the sign $\circ$ stands for component-wise multiplication. The matrix of second derivatives is diagonal here:

$$\boxed{\frac{\partial^2 \mathcal{L}^{\mathrm{prior}}_{\mathrm{clim}}}{\partial \boldsymbol{\lambda}^2} = w_\mathrm{c}\,\mathsf{diag}(\boldsymbol{\Delta}^\mathrm{c})} \tag{60}$$

2. The smoothness part of the cost function, Eq.(44), can be represented as

$$\mathcal{L}^{\mathrm{prior}}_{\mathrm{smoo}}(\boldsymbol{\lambda}) = \frac{w_\mathrm{s}}{2}\,(\mathbf{T}\boldsymbol{\lambda}, \boldsymbol{\lambda}), \tag{61}$$

where $\mathbf{T}$ is the tri-diagonal non-negative definite symmetric matrix with the entries $-(\Delta s_0)^{-1}, -(\Delta s_1)^{-1}, \ldots, -(\Delta s_{L-1})^{-1}$ on both the first subdiagonal and first super-diagonal, and with the sum of entries in each row equal to zero. Differentiating this expression is, again, trivial:

$$\boxed{\frac{\partial \mathcal{L}_{\text{smoo}}^{\text{prior}}}{\partial \boldsymbol{\lambda}} = w_{\text{s}} \mathbf{T} \boldsymbol{\lambda}} \tag{62}$$

and

$$\boxed{\frac{\partial^2 \mathcal{L}_{\text{smoo}}^{\text{prior}}}{\partial \boldsymbol{\lambda}^2} = w_{\text{s}} \mathbf{T}} \tag{63}$$

3. Differentiating $\mathcal{L}_{\text{det}}^{\text{lik}} = \frac{n_{\text{e}}}{2} \log |\boldsymbol{\Gamma}|$ is slightly more involved. From (e.g. Searle and Khuri, 2017, section 9.5), we have that for any matrix $\mathbf{X}$,

$$\frac{\partial \log |\mathbf{X}|}{\partial X_{ij}} = (\mathbf{X}^{-1})_{ji}. \tag{64}$$

Therefore, since $\boldsymbol{\Gamma} = \mathbf{H} \mathrm{e}^{\boldsymbol{\Lambda}} \mathbf{H}^\top$ (see Eq.(46)), we obtain the differential

$$\mathrm{d} \log |\boldsymbol{\Gamma}| = \operatorname{tr} \left(\frac{\partial \log |\boldsymbol{\Gamma}|}{\partial \boldsymbol{\Gamma}} \mathrm{d}\boldsymbol{\Gamma} \right) = \operatorname{tr} \left(\boldsymbol{\Gamma}^{-1} \mathbf{H} \mathrm{e}^{\boldsymbol{\Lambda}} \mathrm{d}\boldsymbol{\Lambda} \mathbf{H}^\top \right) = \operatorname{tr} \left(\mathbf{N} \mathrm{e}^{\boldsymbol{\Lambda}} \mathrm{d}\boldsymbol{\Lambda} \right), \tag{65}$$

where

$$\mathbf{N} = \mathbf{H}^\top \boldsymbol{\Gamma}^{-1} \mathbf{H} \tag{66}$$

(an $L \times L$ matrix). From Eq.(65) it follows that $\partial \log |\boldsymbol{\Gamma}| / \partial \lambda_l = (\mathbf{N})_{ll} \, e^{\lambda_l}$, so that

$$\boxed{\frac{\partial \mathcal{L}_{\text{det}}^{\text{lik}}}{\partial \lambda_l} = \frac{n_{\text{e}}}{2} (\mathbf{N})_{ll} \, e^{\lambda_l} \equiv \frac{n_{\text{e}}}{2} (\mathbf{N})_{ll} f_l.} \tag{67}$$

To compute the Hessian matrix, we derive the second differential of $\log |\boldsymbol{\Gamma}|$ From Eq.(65) we have

$$\mathrm{d}^2 \log |\boldsymbol{\Gamma}| = \mathrm{d} \operatorname{tr} \left(\mathbf{N} \mathrm{e}^{\boldsymbol{\Lambda}} \mathrm{d}\boldsymbol{\Lambda} \right) = \operatorname{tr} \left(\mathrm{d}\mathbf{N} \, \mathrm{e}^{\boldsymbol{\Lambda}} \mathrm{d}\boldsymbol{\Lambda} \right) + \operatorname{tr} \left(\mathbf{N} \mathrm{e}^{\boldsymbol{\Lambda}} \mathrm{d}\boldsymbol{\Lambda}^2 \right). \tag{68}$$

Here we express $\mathrm{d}\mathbf{N}$ using the equality for the differential of any (square) matrix:

$$\mathrm{d}\mathbf{X}^{-1} = -\mathbf{X}^{-1} \, \mathrm{d}\mathbf{X} \, \mathbf{X}^{-1} \tag{69}$$

(again, e.g. Searle and Khuri, 2017, section 9.5). With this equality, we have

$$\mathrm{d}\boldsymbol{\Gamma}^{-1} = -\boldsymbol{\Gamma}^{-1} \mathbf{H} \mathrm{e}^{\boldsymbol{\Lambda}} \, \mathrm{d}\boldsymbol{\Lambda} \, \mathbf{H}^\top \boldsymbol{\Gamma}^{-1}. \tag{70}$$

and therefore

$$\mathrm{d}\mathbf{N} = \mathbf{H}^\top \mathrm{d}\boldsymbol{\Gamma}^{-1} \mathbf{H} = -\mathbf{N} \mathrm{e}^{\boldsymbol{\Lambda}} \mathrm{d}\boldsymbol{\Lambda} \mathbf{N}. \tag{71}$$

Substituting this equation into Eq.(68) yields

$$\mathrm{d}^2 \log |\boldsymbol{\Gamma}| = \operatorname{tr} \left(\mathbf{N} \mathrm{e}^{\boldsymbol{\Lambda}} \mathrm{d}\boldsymbol{\Lambda} \mathbf{N} \mathrm{e}^{\boldsymbol{\Lambda}} \mathrm{d}\boldsymbol{\Lambda} \right) + \operatorname{tr} \left(\mathbf{N} \mathrm{e}^{\boldsymbol{\Lambda}} \mathrm{d}\boldsymbol{\Lambda}^2 \right). \tag{72}$$

Symmetry of $\mathbf{N}$ allows us to rewrite this equation in the component-wise form as

$$\mathrm{d}^2 \log|\mathbf{\Gamma}| = \sum_{ll'} (\mathbf{N})_{ll'}^2 \, \mathrm{e}^{\lambda_l + \lambda_{l'}} \, d\lambda_l \mathrm{d}\lambda_{l'} + \sum_{l} (\mathbf{N})_{ll} \, \mathrm{e}^{\lambda_l} \, d\lambda_l^2. \tag{73}$$

This equation implies that the Hessian matrix is

$$\boxed{\frac{\partial^2 \mathcal{L}_{\mathrm{det}}^{\mathrm{lik}}}{\partial \boldsymbol{\lambda}^2} = \frac{n_{\mathrm{e}}}{2} \left(\mathbf{F}\mathbf{N}^2\mathbf{F} + \mathsf{diag}(\mathsf{diag}(\mathbf{N}) \circ \mathbf{f}) \right).} \tag{74}$$

Here, the operation diag, when applied to a matrix is defined to return its main diagonal, and, when applied to a vector is defined to return the diagonal matrix with the vector on its main diagonal.

4. To derive $\nabla \mathcal{L}_{\mathrm{tr}}^{\mathrm{lik}}$ and $\nabla^2 \mathcal{L}_{\mathrm{tr}}^{\mathrm{lik}}$, we compute the first and second differential of the temporary variable $\tau = \mathrm{tr}(\mathbf{\Phi}\mathbf{\Phi}^\top \mathbf{\Gamma}^{-1})$ such that $\mathcal{L}_{\mathrm{tr}}^{\mathrm{lik}} = \tau \cdot n_{\mathrm{e}}/(2(n_{\mathrm{e}} - 1))$. From Eq.(52), we have

$$\mathrm{d}\tau = \mathrm{tr}(\mathbf{\Phi}\mathbf{\Phi}^\top \, \mathrm{d}\mathbf{\Gamma}^{-1}) = -\,\mathrm{tr}(\mathbf{\Phi}\mathbf{\Phi}^\top \, \mathbf{\Gamma}^{-1}\mathbf{H}\mathrm{e}^{\mathbf{\Lambda}} \, \mathrm{d}\mathbf{\Lambda} \, \mathbf{H}^\top \mathbf{\Gamma}^{-1}) \equiv -\,\mathrm{tr}(\mathbf{Z}\mathbf{Z}^\top \mathrm{e}^{\mathbf{\Lambda}} \, \mathrm{d}\mathbf{\Lambda}), \tag{75}$$

where

$$\mathbf{Z} = \mathbf{H}^\top \mathbf{\Gamma}^{-1} \mathbf{\Phi} \tag{76}$$

(an $L \times n_{\mathrm{e}}$ matrix). From Eq.(75),

$$\boxed{\frac{\partial \mathcal{L}_{\mathrm{tr}}^{\mathrm{lik}}}{\partial \boldsymbol{\lambda}} = -\frac{1}{2} \frac{n_{\mathrm{e}}}{n_{\mathrm{e}} - 1} \, \mathbf{f} \circ \mathsf{diag}(\mathbf{Z}\mathbf{Z}^\top)} \tag{77}$$

Next, we compute the second differential

$$\mathrm{d}^2\tau = -\,\mathrm{tr}(\mathbf{Z}\mathbf{Z}^\top \mathrm{e}^{\mathbf{\Lambda}} \, \mathrm{d}\mathbf{\Lambda}^2) + 2\,\mathrm{tr}(\mathbf{Z}^\top \mathrm{e}^{\mathbf{\Lambda}} \, \mathrm{d}\mathbf{\Lambda} \mathrm{d}\mathbf{Z}). \tag{78}$$

Substituting

$$\mathrm{d}\mathbf{Z} = \mathbf{H}^\top \mathrm{d}\mathbf{\Gamma}^{-1} \mathbf{\Phi} = -\mathbf{H}^\top \mathbf{\Gamma}^{-1} \mathbf{H} \mathrm{e}^{\mathbf{\Lambda}} \, \mathrm{d}\mathbf{\Lambda} \, \mathbf{H}^\top \mathbf{\Gamma}^{-1} \mathbf{\Phi} \tag{79}$$

into Eq.(78) while using Eq.(66) yields

$$\mathrm{d}^2\tau = -\,\mathrm{tr}(\mathbf{Z}\mathbf{Z}^\top \mathrm{e}^{\mathbf{\Lambda}} \, \mathrm{d}\mathbf{\Lambda}^2) + 2\,\mathrm{tr}(\mathbf{Z}\mathbf{Z}^\top \mathrm{e}^{\mathbf{\Lambda}} \, \mathrm{d}\mathbf{\Lambda}\mathbf{N}\mathrm{e}^{\mathbf{\Lambda}} \, \mathrm{d}\mathbf{\Lambda}). \tag{80}$$

Rewriting this equation in the component-wise form allows us to simplify it:

$$\mathrm{d}^2\tau = 2 \sum_{ll'} (\mathbf{Z}\mathbf{Z}^\top)_{ll'} (\mathbf{N})_{ll'} \mathrm{e}^{\lambda_l + \lambda_{l'}} \, \mathrm{d}\lambda_l \, \mathrm{d}\lambda_{l'} - \sum_{l} (\mathbf{Z}\mathbf{Z}^\top)_{ll} \mathrm{e}^{\lambda_l}. \tag{81}$$

This equation implies that the Hessian matrix is

$$\boxed{\frac{\partial^2 \mathcal{L}_{\mathrm{tr}}^{\mathrm{lik}}}{\partial \boldsymbol{\lambda}^2} = \frac{n_{\mathrm{e}}}{2(n_{\mathrm{e}} - 1)} \left(\mathbf{F}[\mathbf{N} \circ (\mathbf{Z}\mathbf{Z}^\top)]\mathbf{F} - \mathsf{diag}(\mathsf{diag}(\mathbf{Z}\mathbf{Z}^\top) \circ \mathbf{f}) \right).} \tag{82}$$

With the gradient and the Hessian of the cost function $\mathcal{L}(\boldsymbol{\lambda})$ in hand, we perform a few Newton-Raphson iterations in search of the *maximum a posteriori* estimate of the log-spectrum, $\widehat{\boldsymbol{\lambda}}$. The starting point for the iterations is the parametric solution described in section 4.4.

4.6 Parametric Bayesian solution

$\log g_l = \log A + \log g(l/a)$

$\mathcal{A} = \log A$

$\alpha = \log a$

$$p(\alpha, \mathcal{A}|\boldsymbol{\varphi}) \propto p(\alpha) \cdot p(\mathcal{A}|\alpha) \cdot p(\boldsymbol{\varphi}|\alpha, \mathcal{A}) \cdot$$

Log-normal priors:

$$\alpha \sim \mathbf{N}(0, \sigma_\alpha)$$

$$\mathcal{A}|\alpha \sim \mathbf{N}(\log V_{\text{clim}} - \log G_0 - \alpha,\, \sigma_{\mathcal{A}}) \equiv \mathbf{N}(c_{\mathcal{A}} - \alpha,\, \sigma_{\mathcal{A}})$$

$$p(\alpha) \propto \mathrm{e}^{-\frac{1}{2}\frac{\alpha^2}{\sigma_\alpha^2}}$$

$$\mathcal{L}_{\text{prior}}(\alpha) = -\frac{1}{2}\frac{\alpha^2}{\sigma_\alpha^2} + \mathsf{const}$$

$$\mathcal{L}_{\text{prior}}(\mathcal{A}|\alpha) = -\frac{1}{2}\frac{(\mathcal{A} + \alpha - c_{\mathcal{A}})^2}{\sigma_{\mathcal{A}}^2} + \mathsf{const}$$

$$\mathcal{L}_{\text{det}}^{\text{lik}}(.) = \frac{n_{\text{e}}}{2}(\log|\boldsymbol{\Gamma}'| + \log|\boldsymbol{\Gamma}''|) \tag{83}$$

and

$$\mathcal{L}_{\text{tr}}^{\text{lik}}(.) = \frac{n_{\text{e}}}{2(n_{\text{e}}-1)}\left[\text{tr}(\boldsymbol{\Phi}'\boldsymbol{\Phi}'^{\top}\boldsymbol{\Gamma}'^{-1}) + \text{tr}(\boldsymbol{\Phi}''\boldsymbol{\Phi}''^{\top}\boldsymbol{\Gamma}''^{-1})\right]. \tag{84}$$

$$\boldsymbol{\Gamma}' = \mathbf{H}'\mathbf{F}\mathbf{H}'^{\top} = A \cdot \mathbf{H}'\mathbf{G}\mathbf{H}'^{\top} \qquad \text{and} \qquad \boldsymbol{\Gamma}'' = A \cdot \mathbf{H}''\mathbf{G}_{-2}\mathbf{H}''^{\top}, \tag{85}$$

$$\det\boldsymbol{\Gamma}' = \det(A \cdot \check{\boldsymbol{\Gamma}}') = A^J \det(\check{\boldsymbol{\Gamma}}')$$

$$\log\det\boldsymbol{\Gamma}' = J\log A + \det\check{\boldsymbol{\Gamma}}' \equiv J\mathcal{A} + \det\check{\boldsymbol{\Gamma}}'$$

$$\log\det\boldsymbol{\Gamma}'' = (J-2)\mathcal{A} + \det\check{\boldsymbol{\Gamma}}''$$

$$\mathcal{L}_{\text{det}}^{\text{lik}}(\boldsymbol{\lambda}) = \frac{n_{\text{e}}}{2}(\log|\boldsymbol{\Gamma}'| + \log|\boldsymbol{\Gamma}''|) \tag{86}$$

and

$$\mathcal{L}_{\text{tr}}^{\text{lik}}(\boldsymbol{\lambda}) = \frac{n_{\text{e}}}{2(n_{\text{e}}-1)}\left[\text{tr}(\boldsymbol{\Phi}'\boldsymbol{\Phi}'^{\top}\boldsymbol{\Gamma}'^{-1}) + \text{tr}(\boldsymbol{\Phi}''\boldsymbol{\Phi}''^{\top}\boldsymbol{\Gamma}''^{-1})\right]. \tag{87}$$

$$\text{tr}(\boldsymbol{\Phi}'\boldsymbol{\Phi}'^{\top}\boldsymbol{\Gamma}'^{-1}) = \mathrm{e}^{-\mathcal{A}}\,\text{tr}(\boldsymbol{\Phi}'\boldsymbol{\Phi}'^{\top}\check{\boldsymbol{\Gamma}}'^{-1})$$

All combined,

$$\mathcal{L}^{\text{post}} = \frac{1}{2}\frac{\alpha^2}{\sigma_\alpha^2} + \frac{1}{2}\frac{(\mathcal{A}+\alpha-c_\mathcal{A})^2}{\sigma_\mathcal{A}^2} + \frac{n_\text{e}}{2}\left[(2J-2)\mathcal{A} + \det\check{\mathbf{\Gamma}}' + \det\check{\mathbf{\Gamma}}''\right] + \frac{n_\text{e}}{2\mathcal{A}(n_\text{e}-1)}\left[\text{tr}(\mathbf{\Phi}'\mathbf{\Phi}'^\top\check{\mathbf{\Gamma}}'^{-1}) + \text{tr}(\mathbf{\Phi}''\mathbf{\Phi}''^\top\check{\mathbf{\Gamma}}''^{-1})\right] \quad (88)$$

5 Numerical experiments with synthetic nonstationary covariances

5.1 True models

Here we describe how the two true models (a stationary model and a doubly stochastic nonstationary model, DLSM) are specified.

5.1.1 Stationary model

The stationary model is Eq.(111) with independent complex random numbers $\widetilde{\alpha}_{lm}$ whose variance spectrum is specified as follows:

$$f_l = \frac{c}{1+(\lambda l)^\gamma}. \quad (89)$$

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 such that, given the parameters γ and λ, the field's standard deviation $\text{SD}\,\xi$ equals the pre-specified value S. From Eq.(110), we find

$$c = \frac{1}{\sum_{l=0}^{L}\frac{1}{1+(\lambda l)^\gamma}}. \quad (90)$$

5.1.2 Doubly stochastic locally stationary model (DLSM)

The LSM is defined by Eq.(10). In that equation, $\widetilde{\alpha}_{lm}$ 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 $f_l(x)$ 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 $f_l(x)$ are specified in the same way as in section 5.1.1 but with the three parameters S, λ, γ being functions of x. We call S, λ, γ the *parameter* fields. First, we simulate the parameter fields $S(x)$, $\lambda(x)$, and $\gamma(x)$ as detailed below in section 5.1.3. Then, for each grid point x, we compute $c(x)$ following Eq.(90) where $S = S(x)$, $\lambda = \lambda(x)$, and $\gamma = \gamma(x)$ and finally set

$$f_l(x) = \frac{c(x)}{1+(\lambda(x) l)^{\gamma(x)}}. \quad (91)$$

5.1.3 Parameter processes

The parameter processes $S(x)$, $\lambda(x)$, and $\gamma(x)$ are defined as transformed Gaussian processes generically written as $g(\chi(x))$, where g is the transformation function and $\chi(x)$ stands for a stationary Gaussian process. We define the "pre-transform" Gaussian processes $\chi(x)$ to have the same *shape* of the modal spectrum as specified for the stationary model by Eq.(89) but with a larger length scale λ than in the model for ξ.

Specifying the same shape of the spectra simplifies the setup and allows an unambiguous comparison of the length scales of the parameter processes $S = S(x)$, $\lambda = \lambda(x)$, and $\gamma = \gamma(x)$ on the one hand and the process ξ in question on the other hand. This latter argument is important because we need to control those length scales as our approach relies on the assumption that the *structure* of the field in question, ξ, changes in space on a significantly larger scale than the length scale of ξ itself. In DLSM, we ensure this by specifying λ for the processes $S = S(x)$, $\lambda = \lambda(x)$, and $\gamma = \gamma(x)$ several times as large as the median λ for ξ as detailed just below.

Specifically, we postulate that

$$S(x) := S_{\text{add}} + S_{\text{mult}} \cdot g(\log \varkappa_S \cdot \chi_S(x, \mu_{\text{NSL}})), \tag{92}$$

$$\lambda(x) := \lambda_{\text{add}} + \lambda_{\text{mult}} \cdot g(\log \varkappa_\lambda \cdot \chi_\lambda(x, \mu_{\text{NSL}})), \tag{93}$$

$$\gamma(x) := \gamma_{\text{add}} + \gamma_{\text{mult}} \cdot g(\log \varkappa_\gamma \cdot \chi_\gamma(x, \mu_{\text{NSL}})), \tag{94}$$

where g is the transformation function such that $g(0) = 1$ (defined below), $\chi_S, \chi_\lambda, \chi_\gamma$ are the three independent pre-transform stationary Gaussian processes (also defined below), 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, and μ_{NSL} is the ratio (common for all three parameter processes) of their length scale Λ to the median length scale λ of ξ.

In more detail, each of the above three pre-transform processes, $\chi_S, \chi_\lambda, \chi_\gamma$ is a realization of the unit-variance stationary process $\chi(x)$ whose variance spectrum is

$$f_l^\chi \propto \frac{1}{1 + (\Lambda l)^\Gamma}, \tag{95}$$

where $\Gamma = \gamma_{\text{add}} + \gamma_{\text{mult}}$ and $\Lambda = (\lambda_{\text{add}} + \lambda_{\text{mult}}) \cdot \mu_{\text{NSL}}$.

With $\varkappa_\bullet = 1$, the respective spectrum does not depend on x: $f_l(x) = f_l$. The higher $\varkappa_\bullet$, 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.

The greater the parameter μ_{NSL}, the smoother in space the parameter processes and, thus, the weaker the spatial non-stationarity of $\xi(x)$. We specify μ_{NSL} in range from 1 to 10, with 3 being the default value.

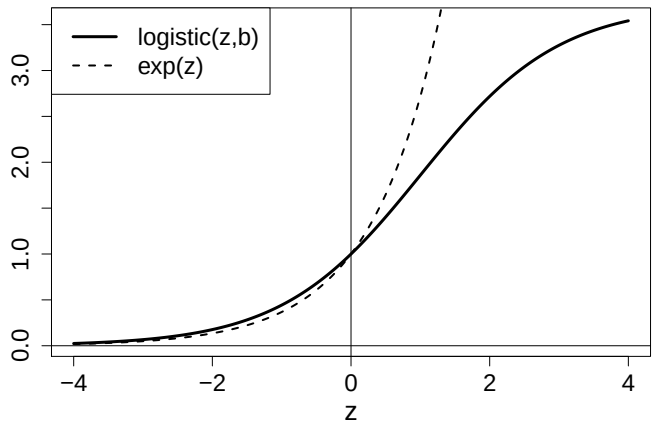

Figure 1: Logistic function $g(z)$ with $b = 1$ and exponential function

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):

$$g(z) := \frac{1 + \mathrm{e}^b}{1 + \mathrm{e}^{b-z}}, \tag{96}$$

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 the parameter fields, 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.

Due to nonlinearity of the transformation function g, the above transformed Gaussian pre-transform fields $\chi_\bullet(x, ...)$ are non-Gaussian. Their pointwise distribution is known as logit-normal or logit-Gaussian.

As $g(z)$ (defined in Eq.(96) 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: $\mathrm{SD}(\chi) = \log \varkappa$, so that the typical deviation of the transformed field from its unperturbed value is $\varkappa$ *times*.

5.1.4 Local spectrum and W

After the processes $S(x)$, $\lambda(x)$, and $\gamma(x)$ are computed at each analysis grid point, Eq.(90) is used to find $c(x)$. With $c(x)$, $\lambda(x)$, and $\gamma(x)$ in hand, we finally compute the true variance spectrum $f_l(x)$ using Eq.(91), then calculate the true modal spectrum $b_l(x)$

using Eq.(20), and finally find

$$\sigma_l(x) = \sqrt{b_l(x)}. \tag{97}$$

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.(14). The $\mathbf{W}$ matrix is then used both to generate the nonstationary random field using Eq.(24), to compute the covariance matrix (only if absolutely necessary) using Eq.(26), and in the analysis algorithm following Eq.(102).

5.1.5 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.(106).

2. In the nonstationary regime, the alternative way of computing $\mathbf{B}$ is given by the nonstationary covariance function, see Eq.(18).

In addition, we visually inspect the nonstationary random field generated using Eq.(24). We check if the generated field is, indeed, (i) larger in areas where $S(x)$ 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.

5.2 Experimental setup

The grid:

$n_x = 60$

The ensemble:

$n_\mathrm{e} = 20(5...100)$

The DLSM:

$\overline{S} = 1$

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

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

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

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

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

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

The bands:

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

5.3 Accuracy of the estimating equation

The estimating Eq.(31) is exact for the stationary process $\xi(x)$ and becomes approximate when $\xi(x)$ is nonstationary. So, the accuracy of Eq.(31) certainly depends on the strength of the non-stationarity. With the DLSM, this implies that the magnitude of the error ζ in Eq.(31) 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.

5.3.1 Methodology

The methodology here is to

1. Specify the external parameters W, $\overline{\lambda}$, and $|\mathcal{B}|$ within their ranges indicated in section 5.2. Set the other external parameters to be equal to their default values.

2. Generate realizations of the secondary fields $S(x)$, $\lambda(x)$, and $\gamma(x)$. Fix them throughout the following steps within an experiment. Calculate the local spectrum $\sigma_l(x)^2$. Use the term $\sum_{l\in\mathcal{B}} b_l(x)$ in Eq.(31) as the ground truth $t(x)$.

3. Generate an ensemble of processes ${}^{(e)}\xi(x)$ with $e = 1, \ldots, n_\mathrm{e}$. 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.(31)). Assess the relative mean and root-mean-square errors (denoted by μ and ρ, respectively) as

$$\mu = \frac{\sum_x \zeta(x)}{\sum_x |t(x)|}, \qquad \rho = \sqrt{\frac{\sum_x \zeta^2(x)}{\sum_x t^2(x)}}, \tag{98}$$

   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_e, and the random seeds on both levels in the DLSM hierarchy.

5.3.2 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.(31) 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 (?)

5.4 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 $b_l^{\rm true} = \sigma_l^2(x)$, (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.(**??**)) spectrum $b_l^{\rm rstr}$.

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

$$\rho_{\rm rstr} = \frac{\sum_l |b_l^{\rm rstr} - b_l^{\rm true}|}{\sum_l |b_l^{\rm true}|}. \tag{99}$$

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 ???

5.5 Performance of the Estimator

5.6 Efficacy of extraction of nonstationary signal

Compare with B averaged over the diagonals.

5.6.1 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!)?

5.7 Analysis algorithm

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

$$\mathbf{x}^{\rm a} = \mathbf{x}^{\rm f} + \mathbf{K}(\mathbf{x}^{\rm obs} - \mathbf{H}\mathbf{x}^{\rm f}), \tag{100}$$

where

$$\mathbf{K} = (\mathbf{B}^{-1} + \mathbf{H}^\top \mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^\top \mathbf{R}^{-1} \tag{101}$$

(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.(26). We proceed as follows:

$$\mathbf{K} = (\mathbf{W}^{-\top}\mathbf{W}^{-1} + \mathbf{H}^{\top}\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^{\top}\mathbf{R}^{-1} = \mathbf{W}(\mathbf{I} + \mathbf{W}^{\top}\mathbf{H}^{\top}\mathbf{R}^{-1}\mathbf{H}\mathbf{W})^{-1}\mathbf{W}^{\top}\mathbf{H}^{\top}\mathbf{R}^{-1}. \quad (102)$$

Now the matrix to be inverted is, clearly, well conditioned. (For Eq.(102) 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.(26).
2. Using a *localized* (thresholded) $\mathbf{W}$ matrix. All $(\mathbf{W})_{ij}$ less than a threshold θ_{W} in modulus are nullified.

5.7.1 Results

Observations.

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

6 Numerical experiments with LSEF

6.1 Model

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 $\boldsymbol{\Gamma}_k$ 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

7 Discussion

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

Hou

Eidsvik

Mandel

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 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.

7.3 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.

7.4 Extensions

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

$$\widehat{\mathbf{v}}_\mathbf{j}(x) = \frac{1}{4\pi} \sum_l |H_j(l)|^2 \, (2l+1)\mathbf{B_l}(x) + \boldsymbol{\zeta} \tag{103}$$

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.

8 Conclusions

As a result, the much desired scale dependent mixing of "climatological" 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.

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:

$$\mathbb{E}\,\xi(\mathbf{x})\,\xi(\mathbf{y}) = \mathbb{E}\,\xi(\mathbf{Q}\mathbf{x})\,\xi(\mathbf{Q}\mathbf{y}), \tag{104}$$

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

$$\mathbb{E}\,\xi(\mathbf{x})\,\xi(\mathbf{y}) = B(\rho(\mathbf{x}, \mathbf{y})) \tag{105}$$

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

$$B(\rho) = \frac{1}{4\pi} \sum_{l=0}^{\infty} (2l+1) b_l P_l(\cos\rho). \tag{106}$$

Equation (106) is the *inverse* Fourier-Legendre transform. The *forward* Fourier-Legendre transform is then

$$b_l = 2\pi \int_{-1}^{1} B[z] P_l(z)\,\mathrm{d}z \equiv 2\pi \int_0^{\pi} B(\rho) P_l(\cos\rho) \sin\rho\,\mathrm{d}\rho, \tag{107}$$

where z stands for $\cos\rho$ and we adopt the notation $B(\rho) \equiv B[\cos\rho]$. Using the **Addition theorem** for spherical harmonics,

$$\sum_{m=-l}^{l} Y_{lm}(x) Y_{lm}^*(y) = \frac{1}{4\pi}(2l+1) P_l(\cos\rho(x,y)) \tag{108}$$

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

$$\xi(\theta, \phi) := \sum_{l=0}^{\infty} \sum_{m=-l}^{l} \widetilde{\xi}_{lm}\, Y_{lm}(\theta, \phi) \tag{109}$$

with ξ_{lm} all mutually uncorrelated complex valued random variables such that $\mathbb{E}\,\widetilde{\xi}_{lm} = 0$ and $\widetilde{\xi}_{l,-m} = \widetilde{\xi}_{lm}^*$. In addition, $\mathsf{Var}\,\widetilde{\xi}_{lm} = b_l$, the *modal* spectrum (variances of individual spectral "modes"). We note also that Eq.(106) entails the equation for the process variance:

$$\mathsf{Var}\,\xi = \frac{1}{4\pi} \sum_{l=0}^{\infty} (2l+1) b_l \equiv \sum_{l=0}^{\infty} f_l, \tag{110}$$

where f_l is the *variance* (or power) spectrum. Denoting $\sigma_l = \sqrt{b_l}$, we rewrite Eq.(109) as

$$\xi(\theta, \phi) := \sum_{l=0}^{\infty} \sum_{m=-l}^{l} \sigma_l \, \widetilde{\alpha}_{lm} \, Y_{lm}(\theta, \phi) \tag{111}$$

Here $\widetilde{\alpha}_{lm} = \widetilde{\xi}_{lm}/\sigma_l$ 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 $m \neq 0$, $\widetilde{\alpha}_{lm} \sim CN(0, 1)$ (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 σ_l, to the range of total wavenumbers from $l = 0$ to $l = L$ in Eqs.(109) and (111). To represent these band-limited functions we use the regular latitude-longitude grid with $n_{\text{lat}} = L + 1$ points over latitude (including both poles) and $n_{\text{lon}} = 2L$ points at each latitude circle.

A.2 Kernel convolution

Equation (111) implies that $\xi(\theta, \phi)$ can be represented as the convolution of the isotropic kernel

$$u(\rho) = \frac{1}{4\pi} \sum_{l=0}^{L} (2l + 1) \sigma_l P_l(\cos \rho) \tag{112}$$

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

$$\alpha(\theta, \phi) := \sum_{l=0}^{L} \sum_{m=-l}^{l} \widetilde{\alpha}_{lm} \, Y_{lm}(\theta, \phi), \tag{113}$$

so that

$$\xi(\mathbf{s}) = \int_{\mathbb{S}^2} u(\rho(\mathbf{s}, \mathbf{s}')) \, \alpha(\mathbf{s}') \, \mathrm{d}\mathbf{s}'. \tag{114}$$

Equation (114) 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 $L = \infty$ 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.(112) to converge at $\rho = 0$. But if we truncate the series in Eq.(112) 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 σ_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...

$\binom{n}{x}$

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.

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

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.