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

M Tsyrulnikov and A Sotskiy
HydroMetCenter of Russia

September 9, 2022

1 Introduction

Modern data assimilation increasingly relies on the ensemble technique, in which the prior probability distribution of the truth is represented by a finite sample (ensemble) of pseudo-random realizations (called ensemble members). This allows for realistic spatially varying, flow-dependent prior (forecast-error) covariances in the analysis (i.e., in the observation update step of the data assimilation cycle). In practical applications, the most widely used approach is the Ensemble Kalman Filter (EnKF), which makes use of ensemble (sample) covariances. The principal problem of the ensemble approach is that running many ensemble members is computationally expensive. In real-world very-high-dimensional problems, this means that only very small ensembles (normally, comprising just tens of members) are affordable. As a result, such ensembles can provide the analysis with only scarce information on the true prior distribution. The sample covariance matrix (a sufficient statistic for the true covariance matrix) is a poor estimate if the sample size is much lower than the dimensionality if the state space. As a result, a kind of *regularization* (i.e., the introduction of additional information on the prior covariances) is required.

1.1 Covariance regularization in EnKF

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

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

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

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

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

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

   Tsyrulnikov and Rakitko (2019) compared the above three covariance blending techniques (that is, mixing with climatological, neighboring in space, and neighboring in time covariances) and found that their usefulness crucially depends on the degree of the spatiotemporal non-stationarity (inhomogeneity) of background errors. Time mean (static) covariances are useful under low non-stationarity, whereas the

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

spatial and temporal covariance blending are more useful when non-stationarity is stronger. They also found (using their doubly-stochastic advection-diffusion-decay model) that the temporal covariance blending is systematically more beneficial than spatial covariance blending.

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

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

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

   Second, physical-space parametric covariance models were used by Skauvold and Eidsvik (2019).

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

1.2 Our contribution

We define local stationarity on the sphere (and on the circle) by introducing a "spectral rescaling" asymptotics. We introduce a non-parametric Locally Stationary (process) Convolution Model on the sphere and on the circle. The Locally Stationary Convolution Model is characterized by local spectra, which determine the spatially varying kernel. We design a computationally efficient estimator of the local spectra from an ensemble of random field's realizations. The estimator involves a spatial multi-scale bandpass filter and a neural network trained on synthetic data. We design an ensemble filter that involves online estimation of the spatial model, which is then used to compute the gain matrix and solve the analysis (observation update) equation. The new filter termed Locally Stationary Ensemble Filter (LSEF) is tested in numerical experiments with a toy model on the circle. Besides, the analysis step of the filter is tested both on the circle and on the sphere with pre-specified "true" non-stationary covariances. In all numerical experiments the developed technique outperformed three alternative methods of specifying the forecast-error covariances: (i) as localized sample covariances (the baseline stochastic

EnKF approach), (ii) as the static (time mean) covariances, and (iii) as a convex linear combination of localized sample covariances and static covariances.

2 Spatial model

As noted in the Introduction, we employ a non-parametric spatial model for the forecast-error random field to regularize the data assimilation problem. Following (Higdon et al., 1999), we rely on the *process convolution* model. In contrast to most applications of this approach, which postulate a parametric model for the spatial kernel (e.g., Lemos and Sansó (2009); Bhat et al. (2012); Heaton et al. (2014); Li and Zhu (2020)), we choose a non-parametric approach to allow for variable shapes of spatial covariances. An attractive approach is Bayesian kernel learning proposed for stationary process convolutions by Tobar et al. (2015), see also Bruinsma et al. (2022). However, this latter hierarchical approach assumes the kernel to be a random function and involves placement of a non-stationary Gaussian process prior on it, which leads to intensive computations. Aiming at high-dimensional applications, we developed a simpler and faster estimator of the non-stationary spatial kernel using bandpass filtering and machine learning. The spatially variable kernel is estimated "online" from an ensemble of pseudo-random realizations of the forecast-error field generated in sequential data assimilation.

We consider processes defined on the two-dimensional unit sphere $\mathbb{S}^2$ and on the unit circle $\mathbb{S}^1$ (referred to as the sphere and the circle in the sequel). The spherical case is more practically relevant (and so is the default case) whereas the circular case is technically simpler. On the sphere, we use the terms stationarity and isotropy interchangeably. We confine ourselves to band-limited functions because in applications, functions need to be represented on a (limited resolution) grid.

2.1 General process convolution model

Let $\xi(x)$ (where x is the spatial coordinate) be a general real-valued space-continuous zero-mean linear Gaussian process, that is, the process whose values are linear combinations of the real valued white Gaussian noise $\alpha(y)$:

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

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

The (non-stationary in general) covariance function of the process $\xi(x)$ defined by Eq.(1) is $B(x,x') = \mathbb{E}\,\xi(x)\xi(x') = \int w(x,y)w(x',y)\,\mathrm{d}y$. Technically, this equation implies

that for $\mathsf{Var}\,\xi(x)$ to be finite, the kernel $w(x,y)$ needs to be square integrable w.r.t. its second argument for any x: $\int w^2(x,y)\,\mathrm{d}y < \infty$.

Rather than using the kernel w that depends on two points in the domain, (x,y), we wish to work with a kernel, v, that depends on the point x *and* on the *relative location*, let it be called z, of the point y w.r.t. x. This is most easily done in the circular case, where we define $z := y - x$ ("distance with the sign") and $v(x,z) := w(x, x+z)$. As a result, we can rewrite Eq.(1) as

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

The new kernel $v(x,z)$ can be viewed as a spatially varying (with x) convolution kernel w.r.t. z. If $v(x,z)$ is independent of x, Eq.(2) becomes the ordinary convolution and ξ becomes a stationary random process.

On the sphere, the definition of $v(x,z)$ is a bit more complicated because we cannot add and subtract points on the sphere. To proceed, we note that, still on the circle, Eq.(2) can be rewritten using *rotations* of the coordinate system. Specifically, for any x, let us rotate (counter-clockwise) the coordinate system by the angle x so that in the unrotated system, the angular coordinate of a coordinate vector changes from e to $R_x e = e + x$. Then, the point whose angle coordinate was y in the unrotated system will have coordinate $z = y - x = R_x^{-1} y$ in the rotated system. Thus we can rewrite Eq.(2) as

$$\xi(x) = \int v(x,z)\,\alpha(R_x z)\,\mathrm{d}z. \tag{3}$$

Now, we note that this equation is applicable to the spherical domain as well. To demonstrate this, we consider the coordinate system's rotation that takes the North Pole to the point x, say, in the following way. First, we rotate around the unit vector e_3 by the angle ϕ_x (longitude of the point x in the unrotated system) and then around the new e_2 unit vector by the angle θ_x (co-latitude of the point x in the unrotated system). That is, the Euler angles of the rotation are $(\phi_x, \theta_x, 0)$ (Varshalovich et al., 1988, sec. 1.4.1). Denoting the respective (orthogonal) change-of-basis matrix by R_x, we realize that the point whose coordinates were y in the unrotated system will have coordinates $z = R_x^{-1} y$ in the rotated system. Therefore, denoting

$$v(x,z) := w(x, R_x z) \tag{4}$$

and changing variables in the integral $\int w(x,y)\,\alpha(y)\,\mathrm{d}y$, we obtain Eq.(3).

2.2 Constraining the model: strategy

In the formulations Eq.(1) and (3), however, the model is not identifiable, that is, the function $w(x,y)$ is not unique given the process covariance function $B(x,x')$. Indeed, consider an isotropic kernel $\psi(\rho(y,z))$ (where $\rho(y,z)$ stands for the great-circle distance between the points y and z) whose Fourier transform (on $\mathbb{S}^1$) Fourier-Legendre transform

(on $\mathbb{S}^2$) is equal to one in modulus (see Appendix A for the definition of the Fourier-Legendre transform we use). Then, it is straightforward to see that the covariance function remains unchanged if $w(x, y)$ is convolved with ψ. (The convolution with such a kernel is analogous to the multiplication by an orthogonal matrix in the space discrete case.)

Our strategy is to make the model identifiable by imposing a few constraints on the kernel $v(x, y)$.

2.3 Convolution model with locally isotropic kernel

The **first constraint** we impose on the kernel requires that it is of the *locally isotropic* form in which for any x, the function $v(x, z)$ depends only on the co-latitude of z. Recall that $z = R_x^{-1}y$ represents y in the rotated coordinate system in which x is at the North Pole. Then the co-latitude of z is nothing other than the great-circle distance between the points x and y. So, there is a real valued function $u(x, \rho(x, y))$ such that

$$w(x, y) \equiv v(x, R_x^{-1}y) = u(x, \rho(x, y)). \tag{5}$$

Substituting Eq.(5) into Eq.(1) yields

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

In the remainder of this subsection, we develop spectral representations of the process defined by Eq.(6) and of its spatial covariances.

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

$$u(x, \rho) = \sum_{\ell=0}^{\ell_{\max}} \frac{2\ell + 1}{4\pi} \sigma_\ell(x) P_\ell(\cos \rho), \tag{7}$$

where P_ℓ is the Legendre polynomial. Substituting $\rho = \rho(x, y)$ into Eq.(7) and applying the addition theorem for spherical harmonics (see again Appendix A), we obtain

$$u(x, \rho(x, y)) = \sum_{\ell=0}^{\ell_{\max}} \sigma_\ell(x) \sum_{m=-\ell}^{\ell} Y_{\ell m}(x)\, Y_{\ell m}^*(y). \tag{8}$$

Then, we write down the spectral expansion of the band limited Gaussian white noise:

$$\alpha(y) = \sum_{\ell=0}^{\ell_{\max}} \sum_{m=-\ell}^{\ell} \widetilde{\alpha}_{\ell m}\, Y_{\ell m}(y). \tag{9}$$

Here $\widetilde{\alpha}_{\ell m}$ are mutually uncorrelated complex-valued random Fourier coefficients with $\mathbb{E}\,\widetilde{\alpha}_{\ell m} = 0$ and $\mathsf{Var}\,\widetilde{\alpha}_{\ell m} = 1$. More specifically, $\widetilde{\alpha}_{l0}$ are real valued and all the other $\widetilde{\alpha}_{\ell m}$ are complex circularly symmetric random variables (e.g. Searle and Khuri, 2017, section 9.5) such that $\widetilde{\alpha}_{l,-m} = \widetilde{\alpha}_{\ell m}^*$.

Finally, we substitute Eqs.(9) and (8) into Eq.(6). Utilizing orthonormality of spherical harmonics, we obtain the basic spectral representation of the random process that satisfies Eq.(6):

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

Note that from Eq.(7) it follows that $\sigma_\ell(x)$ (we call them the spectral functions) are real valued. The spatial covariances can be readily obtained from Eq.(10) by taking into account that all $\widetilde{\alpha}_{\ell m}$ are mutually uncorrelated and again applying the addition theorem for spherical harmonics:

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

If $u(x, \rho) = u(\rho)$ or, equivalently, if all spectral functions $\sigma_\ell(x)$ do not depend on x, then the model Eq.(6) or Eq.(10) becomes a stationary (isotropic) random field model. Equation (11) shows that the process variance $\mathsf{Var}\, \xi(x)$ can be obtained by summing $\frac{2\ell+1}{4\pi} f_\ell(x)$ known in the stationary case as the variance spectrum $\frac{2\ell+1}{4\pi} f_\ell$ (energy per degree ℓ). Again in the stationary case, f_ℓ are often called the *modal* spectrum (energy per 'mode", i.e., per pair (ℓ, m)). In the non-stationary case, in which f_ℓ depend on x, we call $\frac{2\ell+1}{4\pi} f_\ell(x)$ the local variance spectrum and $f_\ell(x)$ the local modal spectrum or just the *local spectrum*.

The formulation of the spatial model Eq.(10) on the circle is outlined in Appendix B.

2.4 Non-negative local spectrum constraint

In Appendix C we consider the process ξ_{statio} defined by the process convolution model with the *stationary* kernel $u(\rho)$, see Eq.(39). As we discussed above, the kernel is not unique given the covariance function of the process. In an attempt to select a unique kernel we impose a computationally motivated *spatial localization* requirement, seeking $u(\rho)$ that has the smallest spatial scale. We show that there are three solutions to this optimization problem, all of them having the same shape of the kernel. Among these three equivalent solutions we select the *non-negative* definite one. A similar result is valid for the process defined on the circle. So, with a stationary process defined by the convolution $\xi = u * \alpha$ (where α the white noise) and having a given spatial covariance function, the most spatially localized kernel u can be considered to be a non-negative definite function of distance.

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

2.5 Local stationarity

Our **third constraint** is *local stationarity* defined below.

The model formulations based on the kernel $v(x, R_x^{-1}y)$ or on the locally isotropic kernel $u(x, \rho(x, y))$ rather than on the kernel $w(x, y)$, see Eqs.(3) and (5), enable us to introduce *two spatial scales*. The first one is the scale at which the kernel $v(x, z)$ varies as a function of z (or $u(x, \rho)$ varies as a function of ρ). It determines a *local length scale* of the process $\xi(x)$. The other one is the scale $v(x, z)$ at which the kernel $v(x, z)$ (or $u(x, \rho)$) varies as a function of x. We call the latter the *non-stationarity length scale* (because, as we noted, the process becomes stationary if $v(x, z)$ is independent of x).

Informally, we say that the process $\xi(x)$ is locally non-stationary if the local length scale of the process is everywhere much less than the non-stationarity length scale. This statement is formalized in Appendix D, where we propose a "spectral rescaling" asymptotics in which the dependence of the kernel $u(x, \rho)$ (and of the processes $\sigma_\ell(x)$ on x) becomes arbitrarily weak.

Note that the assumed slow variation of the kernel $u(x, \rho)$ with the location x (as compared to its variation with the distance ρ), actually, justifies the above term "local spectrum" introduced first by Priestley (1965, 1988), who called it evolutionary spectrum in the time series context.

2.6 Smoothness of local spectra

Studies of real-world spatio-temporal processes showed that spatial (and temporal) spectra in geofluids, say in meteorology, are often quite smooth, exhibiting typically a power-law behavior at large wavenumbers, e.g., Gage and Nastrom (1986), Trenberth and Solomon (1993). For this reason and with the intention to regularize the spatial model by further reducing its effective number of degrees of freedom, we postulate that the spatial spectra $f_\ell(x)$ are smooth functions of the wavenumber ℓ — this is our **fourth constraint**.

2.7 Summary of constraints

1. The convolution kernel has the locally isotropic form $w(x, y) = u(x, \rho(x, y))$.
2. The kernel $u(x, \rho)$ is a non-negative definite function of the distance ρ for any location x.
3. The kernel $u(x, \rho)$ is a smooth function of location x.
4. The local spatial spectra $f_\ell(x)$ are smooth functions of the wavenumber ℓ.

Below we refer to the above spatial model that obeys these four constraints as the Locally Stationary Convolution Model.

2.8 Identifiability

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

3 Locally stationary ensemble filter (LSEF)

The main idea of the new filter is to use the above spatial model (Locally Stationary Convolution Model) in the *analysis* (the observation update step in the filtering process) with the intention to regularize the problem of specifying the prior covariance matrix from the inevitably small ensemble of forecast realizations. In this section we propose an estimator of the Locally Stationary Mode from the forecast ensemble. We also outline other aspects of the filter, which we take from classical stochastic EnKF in order to isolate the impact of our regularization approach.

3.1 Estimating local spectra

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

3.1.1 Multiscale bandpass filter

In the stationary case, it is straightforward to derive from Eq.(37) that the likelihood of the spectrum f_ℓ depends on the sample (the ensemble of M independent realizations of $\xi^\mu_{\text{statio}}(x)$, where $\mu = 1, \dots, M$) trough $\widehat{v}_\ell := \frac{1}{M} \sum_{\mu=1}^{M} \sum_{m=-\ell}^{\ell} |\widetilde{\xi}_{\ell m}|^2$. (Note that with $M = 1$, $\widehat{v}_\ell$ is the spherical analog of what is known in the time series context as the *periodogram*.) Therefore, the vector of the sample mean spectral variances $\widehat{v}_\ell$ (with $\ell = 0, 1, \dots, \ell_{\max}$) is a sufficient statistics for the spectrum f_ℓ. This implies that no information on the true spectrum is lost if we switch from the raw ensemble to the set of $\widehat{v}_\ell$.

In the locally stationary case, however, relying on the spectral variances for individual wavenumbers is not a good idea because the spatial filter that isolates individual spectral components has non-local response functions. Say, on $\mathbb{S}^1$, the spectral transfer function that is equal to one at the wavenumbers $\pm\ell_0$ and zero otherwise corresponds to the impulse response function $2\cos(\ell_0 x)$. This is clearly not acceptable in the non-stationary case. To localize the response function and thus make the technique applicable to the locally stationary case, we specify broader filter spectral transfer functions, which have narrower (i.e., localized) response functions, and use ensemble variances of the outputs of these *bandpass filters* to estimate the local spectra. Broadening the filter's transfer functions reduces the spectral resolution of the estimator, but according to our fourth

constraint (section 2.6), the spatial spectra are smooth, so a reduced spectral resolution should not be a problem.

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

We call the set of J bandpass filters the *multiscale bandpass filter*. Its application to the prior ensemble yields *band variances*, which are then used to estimate the local spectra at each spatial grid point independently.

3.1.2 Band variances

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

$$\xi_{(j)}(x) \approx \sum_{\ell=0}^{\ell_{\max}} H_j(\ell)\, \widetilde{\sigma}_\ell(x) \sum_{m=-\ell}^{\ell} \widetilde{\alpha}_{\ell m}\, Y_{\ell m}(x) \tag{12}$$

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

$$v_{(j)}(x) \approx \sum_{\ell=0}^{\ell_{\max}} \frac{2\ell + 1}{4\pi} H_j^2(\ell)\, f_\ell(x). \tag{13}$$

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

$$\widehat{v}_{(j)}(x) := \widehat{\mathsf{Var}}\, \xi_{(j)}(x), \tag{14}$$

where $\widehat{\mathsf{Var}}$ stands for the sample variance operator. Technically, we used the filters' spectral transfer functions of the form

$$H_j(\ell) = \exp\left(- \left| \frac{\ell - \ell_j^{\mathrm{c}}}{d_j} \right|^q \right), \tag{15}$$

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

3.1.3 Neural network estimator of local spectrum from band variances

Equating the sample band variances $\widehat{v}_{(j)}(x)$ for $j = 1, \ldots, J$ to the respective right-hand sides of Eq.(13) yields a linear inverse problem to be solved w.r.t. the local spectrum $f_\ell(x)$

at each spatial grid point x independently. We applied a standard feed-forward neural network to solve this inverse problem (using pytorch, **?**).

The size of the input layer was J (the number of spectral bands). The size of the output layer was $\ell_{\max}+1$ (the number of local spectral variances). There were two hidden layers with ... neurons each QQ.

In the hidden layers, the activation function was RelU (Goodfellow et al., 2016). We tried a few other activation functions (leaky ReLU, tanh, and sigmoid) and found little difference in the performance of the estimator (not shown). In the output layer, the quadratic activation function was chosen to ensure that the estimated spectrum is non-negative QQ. The exponential activation function led to similar performance (not shown).

The essential component of the neural network's design appeared to be the *loss function*. We started with standard loss functions that penalize the quadratic norm of the function and, possibly, of its first and second derivative. But this choice gave mediocre results. The key point was the introduction of the loss function which is equal to the analysis error variance for a hypothetical idealized analysis, see Appendix F for details.

As for the learning process, we used the Adam optimizer and tried different learning hyperparameters such as learning rate, momentum, and batch size. The results were quite robust (not shown). The number of epochs QQ needed to get stable results was selected by trial and error QQ. Learning data were generated by the models of synthetic pseudo-random fields on $\mathbb{S}^1$ and $\mathbb{S}^2$ described in section 4.

END OF POLISHED TEXT
APPENDIXES ARE MORE OR LESS OK
9 SEP 2022

3.2 Analysis of the state

Having estimated the local spectra $f_\ell(\mathbf{r}_i)$ (where i labels the spatial grid point), we then compute the kernels $u(\mathbf{r}_i, \rho(\mathbf{r}_i, \mathbf{r}_j))$ using Eq.(7) and build the $\mathbf{W}$ matrix in the space discrete random vector representation:

$$\boldsymbol{\xi} = \mathbf{W}\boldsymbol{\alpha}, \tag{16}$$

where $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_{n_\mathbf{x}})$ and $\boldsymbol{\alpha} \sim \mathbf{N}(\mathbf{0}, \mathbf{I})$. The entries of the weighting matrix $\mathbf{W}$ are are defined by the space discrete equivalent of Eq.(6):

$$w_{ij} = u(\mathbf{r}_i, \rho(\mathbf{r}_i, y_j))\,\sqrt{\Delta y_j}, \tag{17}$$

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

$$\mathbf{B} = \mathbf{W}\,\mathbf{W}^\top. \tag{18}$$

The representation of the background-error covariance matrix $\mathbf{B}$ in the "square-root" form, Eq.(18), is common in data assimilation practice because, first, it provides efficient

preconditioning of the analysis equations. Second, it allows for *thresholding* of the $\mathbf{W}$ matrix, i.e., nullifying its small in modulus entries in order to make the matrix sparse and facilitate fast computations (we nullify all w_{ij} whose absolute value is less than a threshold, ϑ_{W}).

In the analysis, we are given, first, the forecast vector $\mathbf{x}^{\mathrm{fc}}$ of length $n_{\mathbf{x}}$ and second, the vector of observations $\mathbf{x}^{\mathrm{o}}$ of length n_{obs}. The observations are assumed to satisfy the observation equation

$$\mathbf{x}^{\mathrm{o}} = \mathbf{H}\mathbf{x}^{\mathrm{true}} + \boldsymbol{\eta}, \tag{19}$$

where $\boldsymbol{\eta}$ is the unknown observation-error vector with known covariance matrix $\mathbf{R}$, $\mathbf{x}^{\mathrm{true}}$ is the unknown true system state, and $\mathbf{H} : \mathbb{R}^{n_{\mathrm{obs}}} \to \mathbb{R}^{n_{\mathbf{x}}}$ is the linear observation operator (an $n_{\mathbf{x}} \times n_{\mathrm{obs}}$ matrix). Then, the optimal analysis (the best linear estimate of $\mathbf{x}^{\mathrm{true}}$ given $\mathbf{x}^{\mathrm{fc}}$ and $\mathbf{x}^{\mathrm{obs}}$) is

$$\mathbf{x}^{\mathrm{an}} = \mathbf{x}^{\mathrm{fc}} + \mathbf{K}(\mathbf{x}^{\mathrm{obs}} - \mathbf{H}\mathbf{x}^{\mathrm{fc}}), \tag{20}$$

where

$$\mathbf{K} = (\mathbf{B}^{-1} + \mathbf{H}^{\top}\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^{\top}\mathbf{R}^{-1} \tag{21}$$

(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.(18). We proceed as follows:

$$\mathbf{K} = (\mathbf{W}^{-\top}\mathbf{W}^{-1} + \mathbf{H}^{\top}\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^{\top}\mathbf{R}^{-1} = \mathbf{W}(\mathbf{I} + \mathbf{W}^{\top}\mathbf{H}^{\top}\mathbf{R}^{-1}\mathbf{H}\mathbf{W})^{-1}\mathbf{W}^{\top}\mathbf{H}^{\top}\mathbf{R}^{-1}. \tag{22}$$

Now the matrix to be inverted is, clearly, well conditioned. (For Eq.(22) to be valid, $\mathbf{W}$ need not, actually, be invertible and even square. This can be proved by changing the control variable from $\mathbf{x}$ to $\boldsymbol{\chi}$, where $\mathbf{x} = \mathbf{W}\boldsymbol{\chi}$, see, e.g., Lorenc et al. (2000).)

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

3.3 Cycling

Stochastic EnKF.

4 Numerical experiments with synthetic fields

In this section we explore the above LSEF analysis technique in the setting with known "truth" that obeys the Locally Stationary Convolution Model. We compare the LSEF analysis with the classical stochastic-EnKF analysis.

4.1 Generating the "true" local spectrum

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

4.1.1 The lowest level in the hierarchy

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

$$f_\ell(x) = \frac{c(x)}{1 + [\lambda(x)(\ell + \ell_0)]^{\gamma(x)}}. \tag{23}$$

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

4.1.2 Parameter fields

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

$$\begin{aligned} S(x) &:= S_{\text{add}} + S_{\text{mult}} \cdot g(\log \varkappa_S \cdot \chi_S(x)), \\ \lambda(x) &:= \lambda_{\text{add}} + \lambda_{\text{mult}} \cdot g(\log \varkappa_\lambda \cdot \chi_\lambda(x)), \\ \gamma(x) &:= \gamma_{\text{add}} + \gamma_{\text{mult}} \cdot g(\log \varkappa_\gamma \cdot \chi_\gamma(x)). \end{aligned} \tag{24}$$

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

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

$$g(z) := \frac{1 + \mathrm{e}^b}{1 + \mathrm{e}^{b-z}}, \tag{25}$$

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

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

With $\varkappa_S = \varkappa_\lambda = \varkappa_\gamma = 1$, the respective spectra are constant in space. The higher $\varkappa_S, \varkappa_\lambda, \varkappa_\gamma$, the more variable in space becomes the respective parameter: $S(x)$ (the standard deviation of the process at the given x), $\lambda(x)$ (the spatially variable length scale of

the process), and $\gamma(x)$ (the spatially variable shape of the local correlations). We specify $\varkappa_\bullet$ to lie between 1 (stationarity) and 4 ("wild" non-stationarity), with 2 being the default value.

4.1.3 Pre-transform fields

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

$$f_\ell^\chi \propto \frac{1}{1 + (\Lambda_{\text{NSL}} \cdot \ell)^\Gamma}, \tag{26}$$

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

4.1.4 Local spectra and W

After the processes $S(x)$, $\lambda(x)$, and $\gamma(x)$ are computed at each analysis grid point, $c(x)$ is adjusted pointwise so that $\mathsf{Var}\,\xi(x) = S^2(x)$. With $c(x)$, $\lambda(x)$, and $\gamma(x)$ in hand, we compute the true spectrum $f_\ell(x)$ using Eq.(23) and $\sigma_\ell(x) = \sqrt{f_\ell(x)}$. Next, we make use of Eq.(7) to compute $u(x, \rho)$. After that, we build $\mathbf{W}^{\text{true}}$ using Eq.(17). The $\mathbf{W}^{\text{true}}$ matrix is then used both to generate the non-stationary random field ξ (and the ensemble members) using Eq.(16) and to compute the best possible analysis (following Eq.(22)).

4.2 Experimental setup

The grid:

$n_{\mathbf{x}} = 60$

The ensemble:

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

The DLSM:

$\overline{S} = 1$

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

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

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

$\gamma_{med} = 2.5$

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

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

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

The bands:

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

4.3 Accuracy of the estimator of band variances

Here we experimentally evaluate the error in the approximate Eq.(13) to confirm the theoretical result obtained in Appendix G.

4.4 Accuracy of the estimator of the local spectra

4.5 Accuracy of the analysis

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

Observations.

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

5 Numerical experiments with cyclic LSEF

5.1 Model

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

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

We preferred DSADM over popular nonlinear models like Lorenz-96 (Lorenz and Emanuel, 1998) 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

6 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.

6.1 Comparison with wavelet-diagonal approach

The Locally Stationary Convolution Model 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).

6.2 Kernel convolution modeling

The kernel convolution approach in modeling stationary random processes has a limitation. Banerjee et al. (2015, section 12.3) note, however, that some correlation models, e.g., the popular exponential correlation function, cannot be reproduced with the kernel convolution approach. The reason is that the spectrum of the exponential correlation function, f_ℓ, decays too slowly as $n \to \infty$ so that the Fourier transform of the kernel, $\sigma_\ell = \sqrt{(f_\ell)}$ does not converge at all. But if we confine ourselves to band-limited functions (evaluated on a spatial grid), then the approximate convolution square root of the exponential correlation function $B(\rho)$ does exist and the approximation error is very small for a reasonable band width (we have checked that but omit those results here).

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

6.3 Application area

Loc statio

Smooth spectra, no lines in spectrum.

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

6.4 Wavelet based filtering

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

6.5 Computations

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

6.6 Extensions

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

$$\widehat{\mathbf{v}}_{\mathbf{j}}(x) = \sum_{\ell} \frac{2\ell+1}{4\pi} H_j^2(\ell)\mathbf{f_l}(x) + \boldsymbol{\zeta} \tag{27}$$

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

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.

7 Conclusions

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

Positive-spectrum requirement fulfilled automatically. ...

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

The traditional covariance localization is *not* capable of suppressing noise at small distances (near the diagonal of the sample covariance matrix), where it is the largest. Our the Locally Stationary Convolution Model 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 for a realistic small-size ensemble, the sample covariance matrix is low-rank and largely uncertain. The regularizing information comes by means of the following assumptions made about the Locally Stationary Convolution Model.

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 Locally Stationary Convolution Model estimator based on spatial band-pass filtering of ensemble members to be consistent (i.e., to give useful results). Assumption 3 is needed for the analysis technique to be computationally efficient.

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

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

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

Appendices

A Spherical harmonics and Fourier-Legendre transform

Here we list some conventions and useful properties of spherical harmonics and Legendre polynomials used throughout the article.

We adopt the following normalization of spherical harmonics:

$$\int_{\mathbb{S}^2} |Y_{\ell m}(x)|^2 \,\mathrm{d}x = 1 \tag{28}$$

where x stands for a point on the unit sphere with the spherical coordinates θ (co-latitude or polar angle) and ϕ (longitude). Spherical harmonics are orthonormal w.r.t. the inner product $(\varphi, \psi) := \int_{\mathbb{S}^2} \varphi(x)\psi^*(x)\,\mathrm{d}x$. We call ℓ the total wavenumber.

The normalization of Legendre polynomials is $\int_{-1}^{1} P_\ell^2(t)\,\mathrm{d}t = \frac{2}{2\ell+1}$ so that $P_\ell(1) = 1$. Legendre polynomials are orthogonal w.r.t. the inner product $(\Phi, \Psi) := \int_{-1}^{1} \Phi(t)\Psi(t)\,\mathrm{d}t$.

The addition theorem for spherical harmonics reads

$$\sum_{m=-\ell}^{\ell} Y_{\ell m}(x) Y_{\ell m}^*(y) = \frac{2\ell+1}{4\pi} P_\ell(\cos\rho(x,y)), \tag{29}$$

where x and y are two points on the unit sphere, $\rho(x,y)$ is the great-circle distance between them, and the asterisk denotes complex conjugation.

By the forward Fourier-Legendre transform of a function, $F(\rho)$, we mean the set of real numbers $\widetilde{F}_\ell$ (where $\ell = 0, 1, \ldots, \ell_{\max}$ and $\ell_{\max}$ is the maximal total wavenumber) defined by the formula

$$\widetilde{F}_\ell = 2\pi \int_0^\pi F(\rho) P_\ell(\cos\rho) \sin\rho \, \mathrm{d}\rho, \tag{30}$$

The backward Fourier-Legendre transform is then

$$F(\rho) = \sum_{\ell=0}^{\ell_{\max}} \frac{2\ell+1}{4\pi} \widetilde{F}_\ell \, P_\ell(\cos\rho). \tag{31}$$

B Locally Stationary Convolution Model on the unit circle

On $\mathbb{S}^1$, the model Eq.(6) specializes to

$$\xi(x) = \int_{\mathbb{S}^1} u(x, x-y) \, \alpha(y) \, \mathrm{d}y, \tag{32}$$

where the real valued kernel $u(x,z)$ is an even function of the second argument. Let us employ the truncated Fourier series expansion of the kernel in the form

$$u(x,z) = \frac{1}{\sqrt{2\pi}} \sum_{\ell=-\ell_{\max}+1}^{\ell_{\max}} \sigma_\ell(x) \, \mathrm{e}^{\mathrm{i}\ell z}. \tag{33}$$

We wish the band limited functions to be representable on the grid, therefore the grid spacing is selected to be $\Delta x = \pi/\ell_{\max}$. Note that the summation limits are such that, being restricted to the grid, the basis functions $\mathrm{e}^{\mathrm{i}\ell z}$ are linearly independent. Since $u(x,z)$ is an even real valued function of z, all $\sigma_\ell(x)$ are real valued functions such that $\sigma_{-\ell}(x) = \sigma_\ell(x)$.

The spectral expansion of the band limited Gaussian white noise reads

$$\alpha(y) = \sum_{\ell=-\ell_{\max}+1}^{\ell_{\max}} \widetilde{\alpha}_\ell \, \mathrm{e}^{\mathrm{i}\ell y}, \tag{34}$$

where all the $\widetilde{\alpha}_\ell$ are mutually uncorrelated random Fourier coefficients with mean zero and variance $\frac{1}{2\pi}$. $\widetilde{\alpha}_0$ and $\widetilde{\alpha}_{\ell_{\max}}$ are real valued while all the others are complex circularly symmetric random variables.

Substituting Eqs.(33) and (34) into Eq.(32) yields the Locally Stationary Convolution Model equation on the circle:

$$\xi(x) = \sum_{\ell=-\ell_{\max}+1}^{\ell_{\max}} \sigma_\ell(x) \, \widetilde{\alpha}_l \, \mathrm{e}^{\mathrm{i}\ell x}, \tag{35}$$

for which the local spectrum is defined as $f_\ell(x) := \sigma_\ell^2(x)$. From Eq.(35) it follows that the Locally Stationary Convolution Model's covariances on $\mathbb{S}^1$ are

$$B(x, x+s) := \mathbb{E}\, \xi(x)\xi(x') = \sum \sigma_\ell(x) \, \sigma_\ell(x+s) \, \mathrm{e}^{\mathrm{i}\ell s}. \tag{36}$$

C The most localized kernel in the stationary case

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

Consider a *stationary* (isotropic) process on the sphere defined by its spectral representation

$$\xi_{\text{statio}}(x) = \sum_{\ell=0}^{\ell_{\max}} \sum_{m=-\ell}^{\ell} \widetilde{\xi}_{\ell m}\, Y_{\ell m}(x) \equiv \sum_{\ell=0}^{\ell_{\max}} \sum_{m=-\ell}^{\ell} \sigma_\ell\, \widetilde{\alpha}_{\ell m}\, Y_{\ell m}(x), \tag{37}$$

where $\mathbb{E}\,\widetilde{\xi}_{\ell m}\widetilde{\xi}^*_{\ell' m'} = \sigma_\ell^2\, \delta_\ell^{\ell'} \delta_m^{m'}$ e.g., Yadrenko (1983, section 5.1). The process $\xi_{\text{statio}}(x)$ can be modeled as the convolution of the kernel

$$u(\rho) = \sum \frac{2\ell+1}{4\pi} \sigma_\ell P_\ell(\cos\rho) \tag{38}$$

with the white noise α:

$$\xi(x) = \int u(\rho(x,y))\, \alpha(y)\, \mathrm{d}y. \tag{39}$$

Multiple kernels $u(\rho)$ give rise to the same spectrum $f_\ell = \sigma_\ell^2$ of the process $\xi_{\text{statio}}(x)$, differing one from another in the signs of σ_ℓ (note that with the real valued kernel $u(\rho)$, all σ_ℓ are real valued). To isolate a unique kernel, we require it to be *most spatially localized* in the sense that it has a minimal width. We define the width as a kind of macro scale from

$$R_{\text{u}}^2 = \frac{\int_{\mathbb{S}^2} u^2(\rho(x,y))\, \mathrm{d}y}{(\max|u(\rho)|)^2}. \tag{40}$$

Here, the numerator is equal to $\sum \frac{2\ell+1}{4\pi} f_\ell$ and thus is fixed given the set of f_ℓ. So, R_{u} is minimized when $\max|u(\rho)|$ is maximal among all kernels with the fixed $|\sigma_\ell|$. That is, we seek to maximize $|u(\rho)|$ over both ρ and the signs of σ_ℓ. From Eq.(38), we have

$$|u(\rho)| = \left| \sum_{\ell=0}^{\ell_{\max}} \frac{2\ell+1}{4\pi} \sigma_\ell P_\ell(\cos\rho) \right| \le \sum_{\ell=0}^{\ell_{\max}} \frac{2\ell+1}{4\pi} |\sigma_\ell| \tag{41}$$

because $|P_\ell(t)| \le 1$ for $-1 \le t \le 1$ (Szegő, 1939, section 7.21). Since $|P_\ell(t)| = 1$ if and only if $t = \pm 1$, the upper bound (i.e., the maximum of $|u(\rho)|$) indicated in Eq.(41) is reached if all $|P_\ell(\cos\rho)| = 1$ (which implies that $\rho = 0$ or $\rho = \pi$) and all products $\sigma_\ell P_\ell(\cos\rho)$ have the same sign. This happens in three cases so that there are three solutions to the optimization problem $R_{\text{u}} \to \min$.

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

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

The third solution is $\sigma_\ell = (-1)^\ell |\sigma_\ell|$. The corresponding kernel $u_3(\rho)$ is non-definite and attains its maximum at $\rho = \pi$ (note that $P_\ell(-1) = (-1)^\ell$). Importantly, it has

exactly the same shape as the first two kernels in the sense that $u_3(\rho) = u_1(\pi - \rho)$ (this follows from the identity $P_\ell(-t) = (-1)^\ell P_\ell(t)$).

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

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

D Definition of local stationarity

Here we give a definition of local stationarity applicable to processes on the sphere and on the circle. To do so, we define a "spectral rescaling" asymptotics.

D.1 Background

The notion of local stationarity has been defined differently by different authors (most often for processes on the real line). The general idea is that a locally stationary process can be approximated by a stationary process locally, i.e., in a vicinity of any point in time (Mallat et al., 1998). Starting from Dahlhaus (1997), the common approach is to use the "infill" asymptotics, which considers the process $\xi(t)$ in rescaled time t/T with $T \to \infty$, e.g., Dahlhaus (2012). With this approach, effectively, just one segment of a non-stationary process is studied in an increasingly sharper detail. The prototypical example is the autoregressive process

$$\xi_{t+1} = a_t\, \xi_t + \varepsilon_t \equiv a(t/T)\xi_t + \varepsilon_t, \tag{42}$$

where $t = 0, 1, \ldots, T$, the function $a(u) : [0, 1] \to (-1, 1)$ defines the non-stationarity pattern, $u := t/T$ is the rescaled time, and $\varepsilon_t \sim N(0, 1)$. In the non-rescaled time t, the non-stationarity pattern is infinitely stretched as $T \to \infty$. Indeed, let variations in $a(u)$ that define the non-stationarity pattern occur at the rescaled time scale L_a (no matter how this time scale is defined). Then, in the non-rescaled time $t = Tu$, this time scale will obviously be stretched to $\Lambda := TL_a$ and thus will tend to infinity as $T \to \infty$. On the other hand, for a fixed u, in a vicinity of the point $t = uT$ in the non-rescaled time t, the process ξ_t will be governed by the auto-regression with asymptotically constant (equal to $a(u)$) coefficients and thus will have an asymptotically constant local length scale $L(u)$ (agian, no matter how it is defined). Therefore, the infill asymptotics is the one in which the ratios $\Lambda/L(u)$ uniformly go to infinity as $T \to \infty$.

On compact manifolds like the circle or the two-dimensional sphere, this rescaling cannot be used because of their compactness. To overcome this obstacle, we define local stationarity using "spectral rescaling" as follows.

D.2 Process on the circle

To define the spectral rescaling asymptotics, we make the kernel $v(x,z)$ depend also on a *spatial scale parameter*, T: $v(x,z;T)$ in such a way that $v(x,z;T)$ becomes only weakly dependent on x for large T. This makes $\xi(x;T)$ "almost stationary" in a vicinity of any point x_0.

Specifically, we expand $v(x,z;T) = \sum \widetilde{v}^*_\ell(x;T)\mathrm{e}^{-\mathrm{i}\ell z}$, $\alpha(x+z) = \sum \widetilde{\alpha}_{\ell'}\mathrm{e}^{\mathrm{i}\ell'(x+z)}$, substitute these expansions into Eq.(2) getting

$$\xi(x;T) = 2\pi \sum_\ell \widetilde{v}^*_\ell(x;T)\,\widetilde{\alpha}_l\,\mathrm{e}^{\mathrm{i}\ell x}. \tag{43}$$

Then, we expand

$$\widetilde{v}_\ell(x;T) = \sum_q \widetilde{v}_{\ell q}(T)\,\mathrm{e}^{\mathrm{i}qx}. \tag{44}$$

Finally, we consider the spectral rescaling asymptotics in which

$$\widetilde{v}_{\ell q}(T) = \phi_\ell(qT), \tag{45}$$

where $T\to\infty$, and for any $\ell \in [-\ell_{\max}+1, \ell_{\max}]$,

$$\phi_\ell(t) \to 0 \quad \text{as} \quad t\to\infty. \tag{46}$$

(For band unlimited functions we would require that Eq.(46) holds uniformly in ℓ but for band limited functions we are interested in this study this is unnecessary.)

In the limit $T\to\infty$, the Fourier transform of $\widetilde{v}_\ell(x;T)$ becomes more and more localized near $q=0$, implying that for any ℓ, $\widetilde{v}_\ell(x;T)$ increasingly flattens as a function of x, tending to a constant. With $T=\infty$, we obtain the stationary model whilst at a finite but large T we obtain *weak non-stationarity*.

Now, we are in a position to prove that the *weakly non-stationary random process* $\xi(x;T)$ defined by Eqs.(43)–(46) is *locally stationary* in the following sense: for any T and any x_0, there exists an approximating stationary process, $\zeta(x;x_0;T)$, such that $\mathbb{E}(\xi(x;T)-\zeta(x;x_0;T))^2 \to 0$ uniformly in x as $T\to\infty$.

Indeed, let the approximating process be defined as

$$\zeta(x;x_0;T) := \sum \widetilde{v}^*_\ell(x_0;T)\,\widetilde{\alpha}_l\,\mathrm{e}^{\mathrm{i}\ell x} \tag{47}$$

where $\widetilde{\alpha}_\ell$ correspond to the same realization of the driving white noise as in Eq.(43) and $\widetilde{v}^*_\ell$ are defined in Eqs.(44)–(46). Then

$$\mathbb{E}(\xi(x;T)-\zeta(x;x_0;T))^2 = \sum (v^*_\ell(x;T)-v^*_\ell(x_0;T))^2 = \sum_{\ell=-\ell_{\max}+1}^{\ell_{\max}} |\,\mathrm{e}^{\mathrm{i}\ell x}-\mathrm{e}^{\mathrm{i}\ell x_0}|^2 \sum_q |\phi_\ell(qT)|^2. \tag{48}$$

Here, in the rightmost double sum, the terms with $\ell=0$ all vanish so that

$$\mathbb{E}(\xi(x;T)-\zeta(x;x_0;T))^2 \le 4\sum_{\ell\neq 0}\sum_q |\phi_\ell(qT)|^2 \to 0, \tag{49}$$

where the limit is due to Eq.(46). So, indeed, the weakly stationary process $\xi(x;T)$ defined by Eqs.(43)–(46) is locally stationary.

D.3 Process on the sphere

We make the kernel $v(x,z)$ depend also on a *non-stationarity scale parameter*, T: $v(x,z;T)$ so that $v(x,z;T)$ becomes only weakly dependent on x for large T. As a result, $\xi(x;T)$ that satisfies Eq.(3) becomes "almost stationary" in a vicinity of any point x_0.

We expand $v(x,z;T)$ in the spherical harmonics basis w.r.t. z:

$$v(x,z;T) = \sum_{\ell=0}^{\ell_{\max}} \sum_{m=-\ell}^{\ell} \widetilde{v}_{\ell m}(x;T)\, Y_{\ell m}(z) \tag{50}$$

(for notation, see Appendix A). Then, we make use of the spectral expansion of the band limited Gaussian white noise, Eq.(9). Substituting Eqs.(50) and (9) (with $y = R_x z$) into Eq.(3), we obtain

$$\xi(x;T) = \sum \widetilde{v}_{\ell' m'}(x;T)\, \widetilde{\alpha}^*_{\ell m} \int Y_{\ell' m'}(z)\, Y^*_{\ell m}(R_x z)\, \mathrm{d}z. \tag{51}$$

Next, we note that the rotation acts on spherical harmonics as follows (Varshalovich et al., 1988, sec. 5.5.1):

$$Y^*_{\ell m}(R_x z) = \sum_{m''} D^{\ell}_{m m''}(\phi_x, \theta_x, 0)\, Y^*_{\ell m''}(z), \tag{52}$$

where $D^{\ell}_{m m''}(\alpha,\beta,\gamma)$ is the Wigner D-function that corresponds to the rotation operator with the Euler angles (α,β,γ). Substituting Eq.(52) into Eq.(51) and taking into account orthonormality of spherical harmonics, we rewrite Eq.(3) as

$$\xi(x;T) = \sum_{\ell m m'} \widetilde{v}_{\ell m'}(x;T)\, D^{\ell}_{m m'}(\phi_x, \theta_x, 0)\, \widetilde{\alpha}^*_{\ell m}. \tag{53}$$

After that, we expand

$$\widetilde{v}_{\ell m}(x;T) = \sum_{pq} \widetilde{v}^{pq}_{\ell m}(T)\, Y_{pq}(x), \tag{54}$$

let

$$\widetilde{v}^{pq}_{\ell m}(T) := \phi_{\ell m}(pT, qT), \tag{55}$$

where $T \to \infty$ and postulate that for all indexes within the resolvable range indicated in Eq.(50),

$$\phi_{\ell m}(t,t') \to 0 \quad \text{as} \quad t \to \infty \quad \text{and} \quad t' \to \infty. \tag{56}$$

Equations (55)–(56) imply that for any pair ℓ, m, the coefficients $\widetilde{v}^{pq}_{\ell m}(T) \to 0$ as $T \to \infty$ unless $p = q = 0$, so that $\widetilde{v}_{\ell m}(x;T)$ tends to a constant as a function of x. This means that $\widetilde{v}_{\ell m}(x;T)$ become increasingly smooth functions of the spatial coordinate x as $T \to \infty$ and for any admissible x, x_0, ℓ, m,

$$\widetilde{v}_{\ell m}(x;T) - \widetilde{v}_{\ell m}(x_0;T) \to 0. \tag{57}$$

Similarly to the circular case above, we define the approximating process to be

$$\zeta(x;x_0;T) = \sum \widetilde{v}_{\ell m'}(x_0;T)\, D^{\ell}_{m m'}(\phi_x, \theta_x, 0)\, \widetilde{\alpha}^*_{\ell m}, \tag{58}$$

where the only difference from Eq.(53) is the fixed spatial argument x_0 in $\widetilde{v}_{\ell m'}(x_0;T)$, in particular, $\widetilde{\alpha}_{\ell m}$ are same as in Eq.(53). Then

$$\mathbb{E}(\xi(x;T)-\zeta(x;x_0;T))^2 = \sum_{\ell m m' m''}(v_{\ell m'}(x;T)-v_{\ell m'}(x_0;T))(v_{\ell m''}(x;T)-v_{\ell m''}(x_0;T))^* D^{\ell}_{mm'} D^{\ell}_{mm''} \to 0, \quad (59)$$

where the limit follows from Eq.(57) and from the boundedness of the Wigner D-functions for a limited range of wavenumbers (Varshalovich et al., 1988, sec. 4.3).

So, indeed, the weakly stationary process $\xi(x;T)$ defined by Eqs.(3), (50), (54), (55), and (56) is locally stationary.

E Locally Stationary Convolution Model with positive definite kernel is identifiable

E.1 Circular case

Here we consider the LSM on $\mathbb{S}^1$ defined by Eq.(35), in which $\sigma_\ell(x) \geq 0$ (equivalently, $u(x,\rho)$ is a non-negative definite function of ρ for any x), and prove that if the non-stationary covariances $B(x,x')$ are produced by an LSM, then this LSM is unique, i.e., the spectral functions $\sigma_\ell(x)$ are uniquely determined by $B(x,x')$.

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

$$\sigma_\ell(x) = \sum_{q=-Q}^{Q} \widetilde{\sigma}_{\ell q}\, \mathrm{e}^{\mathrm{i}qx}, \quad (60)$$

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

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

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

Let us substitute the expansion Eq.(60) into Eq.(35) and change the summation variable q to $n = \ell + q$:

$$\xi(x) = \sum_{\ell=-\ell_{\max}}^{\ell_{\max}} \widetilde{\alpha}_\ell\, \mathrm{e}^{\mathrm{i}\ell x} \sum_{q=-Q}^{Q} \widetilde{\sigma}_{\ell q}\, \mathrm{e}^{\mathrm{i}qx} = \sum_{n=-N}^{N} \widetilde{\xi}_n\, \mathrm{e}^{\mathrm{i}nx}, \quad (61)$$

where $N = \ell_{\max} + Q$ and

$$\widetilde{\xi}_n = \sum_{\substack{\ell=n-Q \\ |\ell|\leq \ell_{\max}}}^{n+Q} \widetilde{\alpha}_\ell\, \widetilde{\sigma}_{\ell,n-\ell}, \quad (62)$$

Since the bandwidth for $\xi(x)$ is finite, $\xi(x)$ is uniquely represented by the set of its spectral coefficients $\{\widetilde{\xi}_n\}_{-N}^{N}$. Therefore the covariances $B(x,x')$ are uniquely represented by the

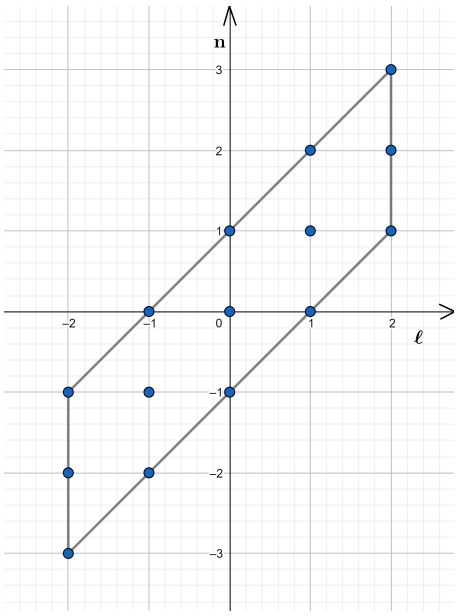

Figure 1: Circular case. Summation area for $\ell_{\max} = 2$ and $Q = 1$

set of the spectral covariances $\widetilde{B}_{n\bar{n}} = \mathbb{E}\,\widetilde{\xi}_n \widetilde{\xi}^*_{\bar{n}}$. Now, we derive $\widetilde{B}_{n\bar{n}}$ from Eq.(62) taking into account that $\widetilde{\alpha}_\ell$ are all mutually uncorrelated zero mean unity variance random variables:

$$\widetilde{B}_{n\bar{n}} = \mathbb{E} \sum_{\substack{\ell=n-Q \\ |\ell|\le\ell_{\max}}}^{n+Q} \widetilde{\alpha}_\ell \, \widetilde{\sigma}_{\ell,n-\ell} \sum_{\substack{\bar{\ell}=\bar{n}-Q \\ |\bar{\ell}|\le\ell_{\max}}}^{\bar{n}+Q} \widetilde{\alpha}^*_{\bar{\ell}} \, \widetilde{\sigma}^*_{\bar{\ell},\bar{n}-\bar{\ell}} = \sum_{\substack{\ell=n-Q \\ |\ell|\le\ell_{\max}}}^{n+Q} \widetilde{\sigma}_{\ell,n-\ell} \, \widetilde{\sigma}^*_{\ell,\bar{n}-\ell} \tag{63}$$

and show that all $\widetilde{\sigma}_{\ell q}$ are uniquely determined by Eq.(63).

The summation area in Eq.(63) for all possible n and $\bar{n}$ is shown in Fig.1 for $\ell_{\max} = 2$ and $Q = 1$.

We start with $n = N$, for which the sum in Eq.(62) reduces to the single term, $\widetilde{\xi}_N = \widetilde{\alpha}_{\ell_{\max}} \widetilde{\sigma}_{LL}$ (the upper right corner of the summation area in Fig.1). This means that the covariance of $\widetilde{B}_{Nn} = \mathbb{E}\,\widetilde{\xi}_N \widetilde{\xi}^*_n$ for any $n \in [\ell_{\max} - Q, \ell_{\max} + Q]$ will also contain just one term, $\widetilde{B}_{Nn} = \widetilde{\sigma}_{LL} \widetilde{\sigma}_{\ell_{\max},n-\ell_{\max}}$ so that we can restore $\widetilde{\sigma}_{Lq}$ for all q. First, we note that $\widetilde{B}_{NN} = |\widetilde{\sigma}_{LQ}|^2$. Next, we turn to $\widetilde{B}_{N,\ell_{\max}-Q} = \widetilde{\sigma}_{LQ} \widetilde{\sigma}^*_{\ell_{\max},-Q}$ (the lower right corner of the summation area in Fig.1). Recalling that $\widetilde{\sigma}_{\ell,-q} = \widetilde{\sigma}^*_{\ell q}$, we obtain $\widetilde{B}_{N,\ell_{\max}-Q} = \widetilde{\sigma}^2_{LQ}$. So, from $\widetilde{B}_{NN} \equiv \widetilde{B}_{N,\ell_{\max}+Q}$ and $\widetilde{B}_{N,\ell_{\max}-Q}$ we know both the modulus and the square of the complex number $\widetilde{\sigma}_{LQ}$, hence it is uniquely determined. By the tight bandwidth assumption (see above in this subsection), $\widetilde{\sigma}_{LQ} \neq 0$, therefore we easily recover $\widetilde{\sigma}_{Lq}$ for all $|q| < Q$ from $\widetilde{B}_{N,\ell_{\max}+q} = \widetilde{\sigma}_{LQ} \widetilde{\sigma}^*_{Lq}$ (the right edge of the summation parallelogram in Fig.1).

Then, we consider $\widetilde{\xi}_{N-1}$ and realize that it, again, has just one term, $\widetilde{\alpha}_{\ell_{\max}-1} \widetilde{\sigma}_{\ell_{\max}-1,\ell_{\max}-1+Q}$ besides the term that contains $\widetilde{\sigma}_{\ell_{\max},\ell_{\max}+Q-1}$, which has already been recovered. This allows us to repeat the above process and uniquely recover $\widetilde{\sigma}_{\ell_{\max}-1,\ell_{\max}-1+q}$ for all $q \in [-Q, Q]$. And so on, we recover all non-zero $\widetilde{\sigma}_{\ell q}$, and therefore

all spectral functions $\sigma_\ell(x)$, from the set of the spectral covariances $\widetilde{B}_{n\bar{n}}$. This completes the proof of uniqueness of the LSM on the circle provided that $\sigma_\ell(x) \geq 0$ and the half-bandwidth Q of $\sigma_\ell(x)$ is such that $\widetilde{\sigma}_{\ell Q} \neq 0$ for any ℓ.

E.2 Spherical case

On $\mathbb{S}^2$, the same reasoning is applicable. We replace the Fourier series in Eq.(60) by the spherical harmonic expansion (Laplace series)

$$\sigma_\ell(x) = \sum_{q=0}^{Q} \sum_{q'=-q}^{q} \widetilde{\sigma}_\ell^{qq'}\, Y_{qq'}(x), \tag{64}$$

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

$$\xi(x) = \sum_{\ell=0}^{\ell_{\max}} \sum_{\ell'=-\ell}^{\ell} \sum_{q=0}^{Q} \sum_{q'=-q}^{q} \widetilde{\sigma}_\ell^{qq'}\, \widetilde{\alpha}_{\ell\ell'}\, Y_{\ell\ell'}(x) Y_{qq'}(x) \tag{65}$$

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

$$Y_{\ell\ell'}(x) Y_{qq'}(x) = \sum_{n=0}^{N} \sum_{n'=-n}^{n} C_{\ell qn}^{\ell' q' n'} Y_{nn'}(x), \tag{66}$$

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

Substituting Eq.(66) into Eq.(65) and utilizing orthogonality of spherical harmonics, we write down the expansion $\xi(x) = \sum_{n=0}^{N} \sum_{n'=-n}^{n} \widetilde{\xi}_{nn'}\, Y_{nn'}(x)$, where

$$\widetilde{\xi}_{nn'} = \sum C_{\ell qn}^{\ell' q' n'}\, \widetilde{\alpha}_{\ell\ell'}\, \widetilde{\sigma}_\ell^{qq'}. \tag{67}$$

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

Then, like in the circular case, we start from $\widetilde{\xi}_{NN}$ and realize that the respective sum in Eq.(67) contains only one term: $C_{LQN}^{LQN}\, \widetilde{\alpha}_{LL}\, \widetilde{\sigma}_L^{QQ}$. We derive $\mathsf{Var}\, \widetilde{\xi}_{NN}$ and $\mathsf{Cov}\,(\widetilde{\xi}_{NN}, \widetilde{\xi}_{N,\ell_{\max}-Q})$, which allows us to recover $\widetilde{\sigma}_L^{QQ}$. As in the circular case, we assume $\widetilde{\sigma}_L^{QQ} \neq 0$. This allows us to recover all $\widetilde{\sigma}_L^{Qq'}$ by computing $\mathsf{Cov}\,(\widetilde{\xi}_{NN}, \widetilde{\xi}_{N,\ell_{\max}+q'})$ for all $-Q \leq q' \leq Q$.

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

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

F Spectral loss function

Here we derive an *analysis-error related* loss function to train the neural network that extracts the local spectrum from the local band variances. Devising such a targeted loss function was crucial for the success of the neural network based approach.

We set up a hypothetical simplified linear analysis and look at the analysis error variance. We consider, first, the (mean-square) optimal analysis, which has access to the *true background-error* (forecast-error) covariances. Second, we consider the same analysis but with *misspecified background error* covariances (the suboptimal analysis). Both analyses have access to the true *observation-error* covariances. We define the loss function to be the *excess of the suboptimal-analysis error variance over that of the optimal analysis*.

Specifically, on the unit sphere, let the forecast-error field, $\xi(x)$, be zero-mean, stationary, band-limited (with the maximum wavenumber $\ell_{\max}$ so that it can be represented on an analysis grid), and have the spectrum f_ℓ. Let the observations X_i^{obs} be located at every analysis grid point (so that X^{obs} can be viewed as a spatial field) and have independent zero-mean errors η_i all with the same variance. Finally, assume, for simplicity, that the analysis background is zero.

Since both ξ and η are stationary (isotropic on the sphere), their spectral-space covariance matrices are diagonal. Therefore, the simplified analysis decouples into a series of partial analyses performed for each wavenumber pair (ℓ, m) separately and independently.

For each partial analysis, by $X_{\ell m}^{\text{true}} = -\xi_{\ell m}$ denote the projection of the truth $X^{\text{true}}(x)$ on $Y_{\ell m}$ (recall that the background is postulated to be equal to zero) and by $X_{\ell m}^{\text{obs}} = X_{\ell m}^{\text{true}} + \eta_{\ell m} = \eta_{\ell m} - \xi_{\ell m}$ the respective projection of the observation field. The optimal estimate of the truth (i.e., the optimal analysis) in spectral space will then be

$$X_{\ell m}^{\text{a}} = K_\ell X_{\ell m}^{\text{obs}} = K_\ell(\eta_{\ell m} - \xi_{\ell m}), \tag{68}$$

where $K_\ell = \frac{f_\ell}{f_\ell + r}$ is the optimal-analysis gain and r is the spectral-space observation error variance (which is constant since $\eta(x)$ is the band-limited white noise by construction). With the misspecified analysis, we assume that the gain is defined with the misspecified f'_ℓ instead of the true f_ℓ, that is, $K'_\ell = \frac{f'_\ell}{f'_\ell + r}$.

Then, for the sub-optimal K'_ℓ, the analysis error is $\delta X_{\ell m}^{\text{a}} = X_{\ell m}^{\text{a}} + \xi_{\ell m} = (1 - K'_\ell)\xi_{\ell m} + K'_\ell \eta_{\ell m}$. Its variance is

$$A_\ell := \mathsf{Var}\, \delta X_{\ell m}^{\text{a}} = \frac{r^2 f_\ell}{(f'_\ell + r)^2} + \frac{(f'_\ell)^2 r}{(f'_\ell + r)^2}. \tag{69}$$

Its excess over the optimal-analysis error variance is then

$$\Delta A_\ell = \frac{r^2 f_\ell}{(f'_\ell + r)^2} + \frac{(f'_\ell)^2 r}{(f'_\ell + r)^2} - \frac{r^2 f_\ell}{(f_\ell + r)^2} - \frac{(f_\ell)^2 r}{(f_\ell + r)^2}. \tag{70}$$

Summing all spectral-space contributions, we obtain the loss function

$$\mathcal{L}(\mathbf{f}, \mathbf{f}') := \sum_{\ell=0}^{\ell_{\max}} \frac{2\ell+1}{4\pi} \Delta A_\ell = r^2 \sum_{\ell=0}^{\ell_{\max}} \frac{2\ell+1}{4\pi} \frac{(f'_\ell - f_\ell)^2}{(f_\ell + r)(f'_\ell + r)^2}. \tag{71}$$

Here $\mathbf{f}, \mathbf{f}'$ are the vectors comprised by all spectral variances f_ℓ and f'_ℓ, respectively. Note that the resulting loss function Eq.(71) is not a norm, it is rather a *deviance*, that is, $\mathcal{L}(\mathbf{f}, \mathbf{f}') \geq 0$ and $\mathcal{L}(\mathbf{f}, \mathbf{f}') = 0$ if and only if $\mathbf{f} = \mathbf{f}'$.

On the circle, the setup and the derivation are the same (omitted). The expression for the loss function differs from Eq.(71) only in the absence of the factor $\frac{2\ell+1}{4\pi}$ and in the summation range (from $-\ell_{\max} + 1$ to $\ell_{\max}$).

G Consistency of the estimator of band variances

Here we consider bandpass filtering of a locally stationary process and find conditions under which the approximations Eqs.(12) and (13) hold. For simplicity, we examine the circular case.

Let the linear filter $\mathcal{H}$ with the spectral transfer function $H(\ell)$ be applied to the locally stationary process

$$\xi(x) = \sum_{\ell=-\ell_{\max}}^{\ell_{\max}} \sigma_\ell(x)\, \widetilde{\alpha}_l\, \mathrm{e}^{\mathrm{i}\ell x} \tag{72}$$

(we reproduce here our basic Eq.(35) for the reader's convenience). In spectral space, the action of a linear filter on the signal $\xi(x) = \sum_\ell \widetilde{\xi}_\ell\, \mathrm{e}^{\mathrm{i}\ell x}$, amounts to the multiplication of spectral coefficients $\widetilde{\xi}_\ell$ by the respective $H(\ell)$. The spectral-space representation of $\xi(x)$ is given by Eq.(61) so that the filtered process $\xi_\mathcal{H} = \mathcal{H}\xi$ reads

$$\xi_\mathcal{H}(x) = \sum_{\ell=-\ell_{\max}}^{\ell_{\max}} \widetilde{\alpha}_\ell\, \mathrm{e}^{\mathrm{i}\ell x} \sum_{q=-Q}^{Q} H(\ell+q)\, \widetilde{\sigma}_{\ell q}\, \mathrm{e}^{\mathrm{i}q x}. \tag{73}$$

Now we recall that the local stationarity means that the processes $\sigma_\ell(x)$ are slowly changing in space (i.e., with x), which is equivalent to a rapid decrease of the coefficients $\widetilde{\sigma}_{\ell q}$ in their spectral decomposition $\sigma_\ell(x) = \sum_q \widetilde{\sigma}_{\ell q}\, \mathrm{e}^{\mathrm{i}q x}$ (Eq.(60)) for any wavenumber ℓ. If we specify $H(\ell)$ to change with ℓ much more slowly than $\widetilde{\sigma}_{\ell q}$ change with q, then, in the sum over q in Eq.(73), $H(\ell+q)$ can be approximated by $H(\ell)$ leading to the approximation

$$\check{\xi}_\mathcal{H}(x) = \sum \sigma_\ell(x)\, \widetilde{\alpha}_l\, H(l)\, \mathrm{e}^{\mathrm{i}\ell x}. \tag{74}$$

(Its spherical counterpart is given in Eq.(12).) A more rigorous proof of this statement follows. The error in the approximation Eq.(74) is

$$\check{\xi}_\mathcal{H}(x) - \xi_\mathcal{H}(x) = \sum_\ell \widetilde{\alpha}_\ell\, \mathrm{e}^{\mathrm{i}\ell x} \sum_q \left[H(\ell+q) - H(\ell)\right]\, \widetilde{\sigma}_{\ell q}\, \mathrm{e}^{\mathrm{i}q x}, \tag{75}$$

Here we remember that $\mathcal{H}$ is a bandpass filter and assume that its spectral transfer function is generated by a bell-shaped function of continuous argument, $\kappa(z)$, such that $\kappa(0) = 1$ and a half-width of $\kappa(z)$ is also about 1. Specifically, let

$$H(\ell) = \kappa\left(\frac{\ell - \ell^{\mathrm{c}}}{d}\right), \tag{76}$$

where ℓ^{c} is the band's central wavenumber and d is the half-bandwidth (*cf.* Eq.(15)). As we noted above, since $\widetilde{\sigma}_{\ell q}$ rapidly decays with the growing $|q|$ for any ℓ, the use of the first order Taylor series approximation is warranted: $H(\ell + q) \approx H(\ell) + H'(\ell)q = H(\ell) + \kappa'(.)q/d$. Substituting this equation into Eq.(75) yields

$$\check{\xi}_{\mathcal{H}}(x) - \xi_{\mathcal{H}}(x) = \frac{1}{d}\sum_{\ell} \widetilde{\alpha}_{\ell}\, \mathrm{e}^{\mathrm{i}\ell x}\kappa'(.) \sum_{q} q\, \widetilde{\sigma}_{\ell q}\, \mathrm{e}^{\mathrm{i}qx} \equiv \frac{1}{\mathrm{i}d}\sum_{\ell} \widetilde{\alpha}_{\ell}\, \mathrm{e}^{\mathrm{i}\ell x}\kappa'(.)\sigma'_{\ell}(x) \tag{77}$$

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

Next, we compute the mean square value of Eq.(77) and note that the width d of the spectral transfer function $H(\ell)$ equals the inverse width L_{H} of the impulse response function of the filter $h(x)$ (which is the inverse Fourier transform of $H(\ell)$), getting

$$\mathbb{E}\left(\check{\xi}_{\mathcal{H}}(x) - \xi_{\mathcal{H}}(x)\right)^2 = L_{\mathrm{H}}^2 \sum_{\ell} (\kappa'(.)\, \sigma'_{\ell}(x))^2. \tag{78}$$

With

$$\sigma_{\ell}(x) = \sum_{q} \phi_{\ell}(qT)\, \mathrm{e}^{\mathrm{i}qx} \tag{79}$$

(see Eqs.(45) and (60),

$$|\sigma'_{\ell}(x)| \le \sum_{q} |q|\, |\phi_{\ell}(qT)|. \tag{80}$$

Since $\kappa'(.) = O(1)$ and the range of both ℓ in Eq.(78) and q in Eq.(79) are finite, the asymptotics Eq.(46) implies that $\mathbb{E}\left(\check{\xi}_{\mathcal{H}}(x) - \xi_{\mathcal{H}}(x)\right)^2$ as $T \to \infty$.

So, indeed, the mean-square error in the approximation Eq.(74) is asymptotically zero in the limit of local stationarity.

QQ In practical terms, for small approximation error, the width of the filter's impulse response function L_{H} is required to be much shorter than the non-stationarity length scale Λ. This is conceivable because at intervals shorter than Λ, the LSM process behaves like a stationary process.

The above result implies that with the norm defined from $\|.\|^2 := \mathbb{E}(.)^2$, we have $\|\check{\xi}_{\mathcal{H}} - \xi_{\mathcal{H}}\| \to 0$. Therefore, $\|\check{\xi}_{\mathcal{H}}\|^2 \to \|\xi_{\mathcal{H}}\|^2$ so that the approximate band variance estimate $\sum H^2(l)\sigma_{\ell}^2(x)$ is asymptotically unbiased and consistent.

Spherical case...QQ

References

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

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

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

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

K. Bhat, M. Haran, R. Olson, and K. Keller. Inferring likelihoods and climate system characteristics from climate models and multiple tracers. *Environmetrics*, 23(4):345–362, 2012.

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

W. P. Bruinsma, M. Tegnér, and R. E. Turner. Modelling non-smooth signals with complex spectral structure. In *International Conference on Artificial Intelligence and Statistics*, pages 5166–5195. PMLR, 2022.

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

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

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

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

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

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

I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT press, 2016.

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

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

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

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

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

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

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

R. T. Lemos and B. Sansó. A spatio-temporal model for mean, anomaly, and trend fields of north atlantic sea surface temperature. *Journal of the American Statistical Association*, 104(485):5–18, 2009.

Y. Li and Z. Zhu. Spatio-temporal modeling of global ozone data using convolution. *Japanese Journal of Statistics and Data Science*, 3(1):153–166, 2020.

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

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

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

E. N. Lorenz and K. A. Emanuel. Optimal sites for supplementary weather observations: Simulation with a small model. *J. Atmos. Sci.*, 55(3):399–414, 1998.

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

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

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

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

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

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

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

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

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

F. Tobar, T. D. Bui, and R. E. Turner. Learning stationary time series using gaussian processes with nonparametric kernels. *Advances in neural information processing systems*, 28, 2015.

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

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

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

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

D. Varshalovich, A. Moskalev, and V. Khersonskii. *Quantum theory of angular momentum*. World Scientific, 1988.

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