Efficient Estimation of Non-stationary Spatial Covariance Functions with Application to Climate Model Emulation

Yuxiao Li and Ying Sun

King Abdullah University of Science and Technology

Supplementary Material

This file contains the details of computations, additional simulation studies and figures, as well as the source of R code and data for reproducible research.

S1 Details of Computations

S1.1 The Bandwidth Selection

The weight function depends on the choice of bandwidth, h. Typically, h is selected by cross-validation (CV) or generalized cross-validation (GCV). The options are various and details can be found in Wand and Jones (1994). For example, we can either leave one pair of locations or one subregion out at a time; we can also minimize the sum of squared prediction errors (SSPE) for either spatially varying parameters or for covariance functions. Here, we prefer the option employed by Kleiber and Nychka (2012). They imple-

mented a leave-one-out cross-validation (LOOCV) process for each location pair by minimizing the SSPE for the covariance functions in the multivariate case. The univariate SSPE is $\sum_{k,l=1}^{n}\left\{\hat{C}^{\mathrm{NS}}_{l,k}(\mathbf{s}_k,\mathbf{s}_l)-\hat{C}^{\mathrm{NS}}_{-l,-k}(\mathbf{s}_k,\mathbf{s}_l)\right\}^2$, where $\hat{C}^{\mathrm{NS}}_{l,k}(\mathbf{s}_k,\mathbf{s}_l)$ is the estimated covariance function of $C^{\mathrm{NS}}(\mathbf{s}_k,\mathbf{s}_l)$ and $\hat{C}^{\mathrm{NS}}_{-l,-k}(\mathbf{s}_k,\mathbf{s}_l)$ is the predicted estimate at the location pair $(\mathbf{s}_k,\mathbf{s}_l)$ based on the remaining locations, $\mathbf{s}\neq\mathbf{s}_k,\mathbf{s}_l$.

This option is preferable since it contains more location pairs and is not affected by the choice of spatially varying parameters. However, this procedure could be quite time consuming and tends to underestimate the bandwidth. In Gaussian kernels, h is the variance of the Gaussian distribution. We therefore choose the bandwidth to be the square of half of the distance between the two closest anchor locations, i.e., $h=\{\|\mathbf{s}_1-\mathbf{s}_2\|/2\}^2$, so that the local estimates depend only on the data in the subregions with range $2\sqrt{h}$, indicating the range of high density points in a Gaussian distribution.

S1.2 Sequentially Conditional Simulation

We are generating data from a multivariate Gaussian distribution with covariance matrix of size $N\times N$. It generally requires a Cholesky decomposition of the covariance matrix, which takes $O(N^3)$ running time, and it takes

two hours to simulate one spatial field with 13,000 locations on a personal laptop. Therefore, we propose to use a more scalable method that simulates observations from the Gaussian random fields sequentially. The sequential Gaussian simulation has been widely used in geostatistical simulations Gotway and Rutherford (1994); Fredericks and Newman (1998). The idea is to simulate realizations only on a subset at a time. In our application, we simulate one dataset for a subregion, then the next subset are simulated conditional only on all or a part of the previous simulated one.

Specifically, we propose a computationally efficient algorithm with sequentially conditional simulation methods. Let X be the $l_1 \times l_2$ data matrix observed on a grid, where $N = l_1 \times l_2$. First, we separate the region into J blocks of equal size by columns such that $X = (\mathbf{X}_1, \ldots, \mathbf{X}_J)$ is ordered from left to right. Since $\mathbf{X}_1$ can be simulated unconditionally, for $k = 2, \ldots, J$, $\mathbf{X}_k|\mathbf{X'}_{k-1} \sim N_{n_k}(\Sigma_{k-1,k}^T \Sigma_{k-1,k-1}^{-1} \mathbf{X'}_{k-1}, \Sigma_{k,k} - \Sigma_{k-1,k}^T \Sigma_{k-1,k-1}^{-1} \Sigma_{k-1,k})$, where $\mathbf{X'}_{k-1}$ is partial or all columns of $\mathbf{X}_{k-1}$ closest to $\mathbf{X}_k$ and $\Sigma_{k,l}$ is the covariance matrix between two random matrices.

Thus, we can simulate $\mathbf{X}_2$ conditionally on $\mathbf{X'}_1$. For $k > 2$, to simulate $\mathbf{X}_k$, instead of conditioning on all $\mathbf{X}_1, \ldots, \mathbf{X}_{k-1}$, we condition only on $\mathbf{X'}_{k-1}$. Here, we only need to invert $\Sigma_{k-1,k-1}$, it indicates that the size of the conditioning set determines the computational cost. By choosing $\mathbf{X'}_{k-1}$

with a small size, this simulation method is nearly as fast as generating $\mathbf{X}_k$ unconditionally from Σ_{kk}. In Vecchia's method (Vecchia, 1988; Katzfuss and Guinness, 2017), the conditioning set is chosen to be the nearest neighbors. In our application, we consider 20% (2%/10%) of the neighbors, on the boundary. We can use a larger conditioning set, however, the computation will become more expensive.

S2 Further Simulation Studies

S2.1 Computation Time

In this subsection, we would like to compare the exact computation time between various 1) non-stationary models, 2) numbers of subregions, and 3) numbers of non-stationary parameters. all comparisons are made based on the same scenario as the 1D simulation study in Section 5, except for the test variable(s).

1. It is not easy to compare different methods if different techniques are employed. First, few of them mention the exact time of the estimations and the data size varies. Second, there are many different fundamental theories, it is not fair to compare them since existing nonstationary models are not all about nonstationary Matérn and all the methods

require the choice of certain tuning parameters.

However, it is possible to make comparisons among the approaches based on non-stationary Matérn covariance function (Paciorek and Schervish, 2006). In the main manuscript, our model is compared with two rough approximation methods (S0 and WS0) and conventional moving window method (moving window S0), which is mentioned in the last two paragraphs of Section 2.2. The following table shows the running time of those methods, based on the same scenario in the 1D simulation study in Section 5. The comparison is relatively fair since the region division and the number of non-stationary parameters are the same. As we can see, for the same number of subregions, since our model has more parameters to estimate, it requires more computational time than the other two rough approximation methods, but much faster than the conventional moving window method when the window size is the same as the size of the subregion.

Model	S0	WS0	NS1	Moving-window S0
Running time (sec)	1.122517	1.1969	2.054463	51.42817

2. Since the smoothing step is very fast, the total computation time is nearly equal to $T = \sum_{i=1}^{m} t_i$, where m is the number of subregions

and t_i is the computational time for each subregion. When m increases, t_i will decrease since the sample size or the number of locations gets smaller in each subregion. Typically, more subregions result in shorter computational time T, as the reduced sample size in each subregion saves the computational cost more significantly. However, the estimation will become less accurate. Based on the same setting in the 1D simulation, The following table shows the running time for $m = 2, 4, 8, 10$, and the results support our argument.

# of subregions	2	4	8	10
Running time (sec)	3.571554	2.054463	1.656103	1.629547

3. Increasing the number of non-stationary parameters implies that more parameters need to be estimated, which will affect both the accuracy and efficiency. Again, using the same setting as in the 1D simulation, the following table shows the running time when we also consider different numbers of nonstationary parameters: 1) σ only, 2) σ and λ, 3) σ, λ and η. It can be seen that when the number of spatially varying parameters increases, the estimation becomes slower. Besides, when too many parameters are required to be estimated, the optimization often fails to achieve the global maximum.

# of nonstationary parameter	1	2	3
Running time (sec)	3.571554	7.516602	9.151497

Similarly, the degree of nonstationarity refers to the order. Higher order of the nonstationarity increases the number of parameters dramatically, and thus increases both the computation and uncertainty. Actually, the comparisons between WS0 (zero-order) and NS1 (first-order) can be viewed as the comparisons for different degrees of nonstationarity.

S2.2 Sensitivity to Region Division

Using the same setting as 1D simulation, we show the estimation results for $m = 2, 4, 8, 10$, and compare with WS0 and S0 models. From the Response Fig. 1-4, our model is less sensitive than WS0 and S0 models. However, a good estimation still heavily depends on the correct division. When $m = 2$, indicating an inappropriate division, all of three methods are not well fitted. When $m = 4, 8, 10$, the true curve always situates inside the 95% CI of our estimation, whereas WS0 and S0 models can obtain a good estimation only after $m = 8$. However, WS0 and S0 have smaller uncertainties than NS1 model, especially when m is large. The reason is that our NS1 model has more parameters, so that our estimation requires more data in the local

fitting.

In general, too few subregions lead to a poor estimation while too many will increase the estimation uncertainty. For example, in the 1D simulation, two subregions ($m = 2$) are not enough, while $m = 4, 8, 10$ are helpful. However, the uncertainty increases too much when $m = 8, 10$, since there are not enough data in local fitting.

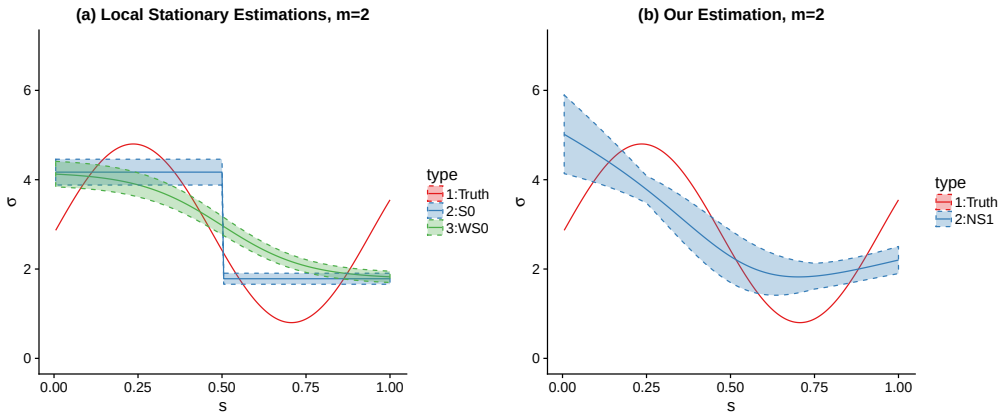

Figure S1: Estimations of $\sigma(s)$ with 95% confidence intervals for $m = 2$ using (a) local stationary model (S0) and weighted local stationary (WS0) model and (b) our first-order non-stationary (NS1) model

As we briefly mentioned in Section 7 and in the Supplement (Section S2.2), this problem is not well solved neither in our paper nor in the related works. To make it consistent, we equally divide the region by four in our paper. In practice, the decision should be made by some exploratory analyses. However, it is hard to visualize the covariance behavior from the raw

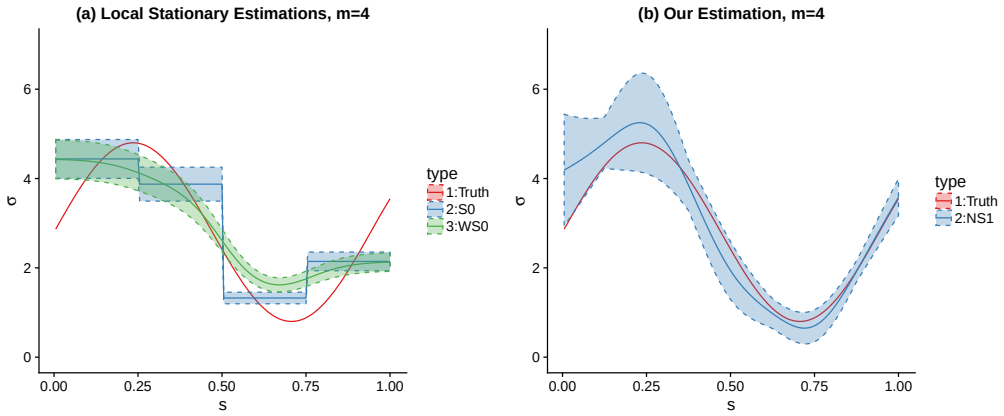

Figure S2: Estimations of $\sigma(s) = 2\sin(s/0.015) + 2.8$ with 95% confidence intervals for $m = 4$ using (a) local stationary model (S0) and weighted local stationary (WS0) model and (b) our first-order non-stationary (NS1) model

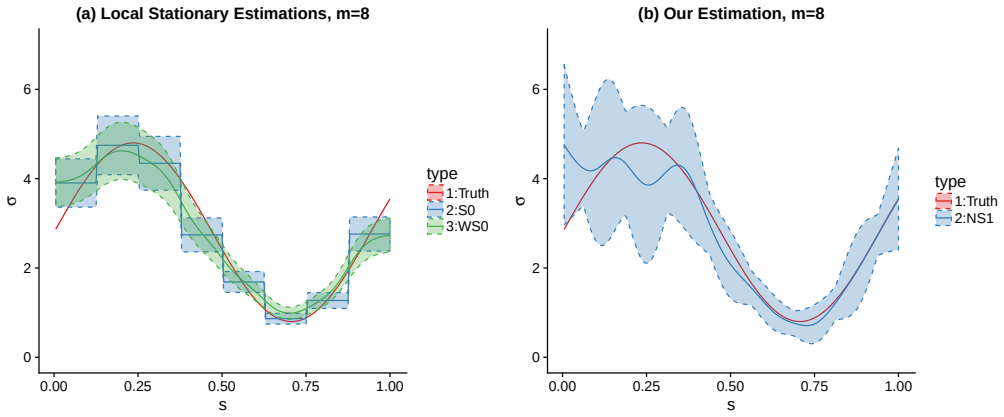

Figure S3: Estimations of $\sigma(s) = 2\sin(s/0.015) + 2.8$ with 95% confidence intervals for $m = 8$ using (a) local stationary model (S0) and weighted local stationary (WS0) model and (b) our first-order non-stationary (NS1) model

data, since it measures the second-order property and there are different types of stationarity. Therefore, other exogenous information regarding the

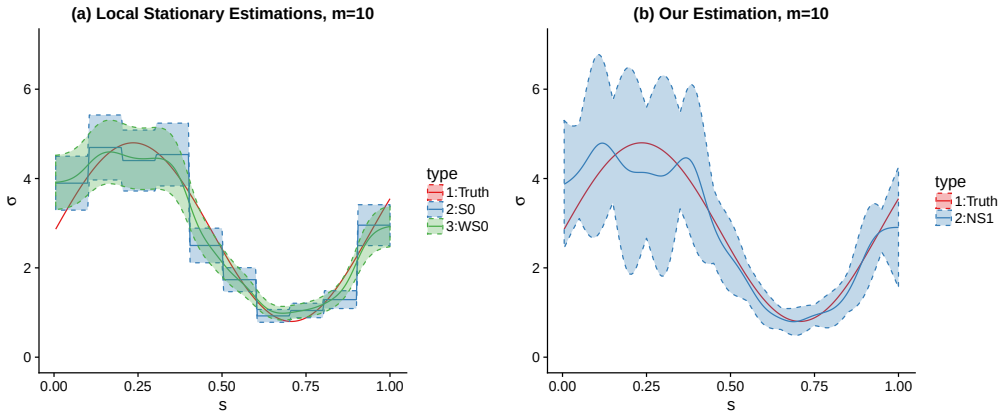

Figure S4: Estimations of $\sigma(s) = 2\sin(s/0.015) + 2.8$ with 95% confidence intervals for $m = 10$ using (a) local stationary model (S0) and weighted local stationary (WS0) model and (b) our first-order non-stationary (NS1) model

heterogeneity of the spatial might be helpful for determining the partition. In principle, each subregion should be small enough to be homogeneous but large enough to have enough data for estimation.

In fact, based on the Taylor theorem, the spatially varying parameter as a function over space can be expanded as a polynomial function up to an infinitely small error. If we can correctly estimate the coefficients for all polynomial terms, the region division does not matter. However, we only recommend use linear approximation without going for higher order polynomials but with a reasonable number of subregions. This is because there are computational issues due to estimating too many parameters, and the uncertainty also increases dramatically when considering higher order

polynomials.

Future research is needed for developing automatic and adaptive partitioning criteria. It could make use of some validation tools or the information from covariates (see Risser et al. 2016).

S2.3 Estimation of Non-stationary Kernel Matrix

Our second example is to estimate the non-stationary kernel matrix $\Sigma(\mathbf{s})$, characterized by $\{\lambda_1(\mathbf{s}), \lambda_2(\mathbf{s}), \phi(\mathbf{s})\}$, $\mathbf{s} \in [0,1]^2$. Here, $\mathbf{s}$ is gridded and the true spatially varying parameters are chosen to be the transformed quadratic functions, where the transformations depend on the constraints of the parameters. Other parameters are set to be constant, i.e., $(\sigma^2, \tau^2, \nu) = (1, 0.1, 2)$. Figure S6(a) shows the contour-image plots of the spatially varying parameters and the corresponding covariance matrix.

Specifically, $\lambda_1(\mathbf{s}) = \exp\{-3 - 6(s_1 - 0.5)^2 - 7(s_2 - 0.5)^2\}$, $\lambda_2(\mathbf{s}) = \exp\{-5+6(s_1-0.5)^2-4(s_2-0.5)^2\}$, $\phi(\mathbf{s}) = \frac{\pi}{2}\left[\frac{\exp\{(s_1-0.5)-2(s_2-0.5)+(s_2-0.5)^2\}}{1+\exp\{(s_1-0.5)-2(s_2-0.5)+(s_2-0.5)^2\}}\right]$. Since λ_1 and λ_2 represent the squared spatial ranges, we choose $\lambda_1(\mathbf{s_0}) = -3$ and $\lambda_2(\mathbf{s_0}) = -5$ at the center, $\mathbf{s_0} = (0.5, 0.5)$, to allow the effective range (the distance at which the correlation reduce to 0.05) to be reasonable.

Figure S5(a) shows how these parameters bring non-stationarity to the corresponding covariance function. We select five reference locations on

the four sides and the center. For each chosen location, we calculate the covariance between the location with all others, and the contour-image plots show the covariances. If a covariance function is stationary, the contours at reference locations will present an homogeneous pattern, where the four contours on the sides are able to combine into the one in the center. In our example, however, the contours are apparently non-homogeneous and thus the generated random process is non-stationary.

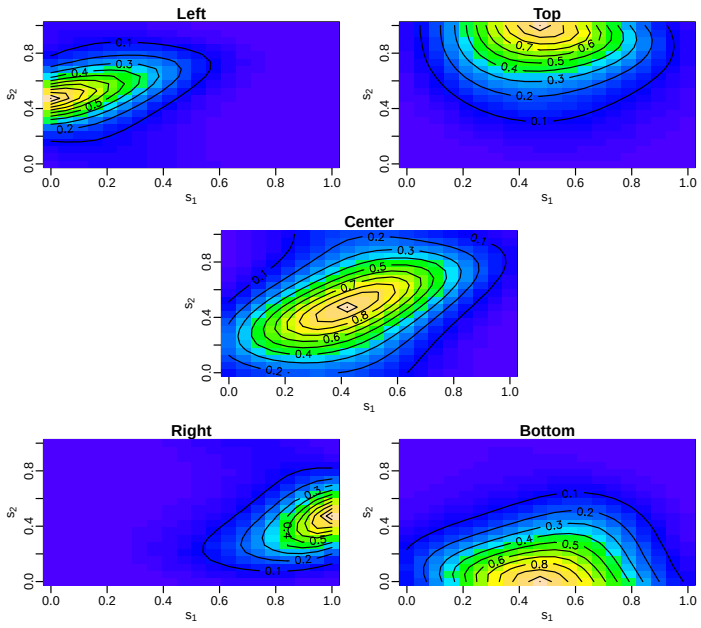

Figure S5: Contour-image plots of the covariance between the contour center and the other points, where the contour center is selected at five reference locations.

We simulate 8000 observations from zero-mean GRFs at 400 locations with 20 independent replicates. Our NS1 estimates are obtained following

the algorithm in Section 4, whereas the S0 and WS0 estimates are implemented using the `convoSPAT` package. With four subregions, i.e. $m = 4$, we have $3 \times 3 \times 4 + 3 = 39$ parameters to estimate in our model. However, since we separate the optimization as mentioned in Section 4.1, at most six parameters are optimized simultaneously.

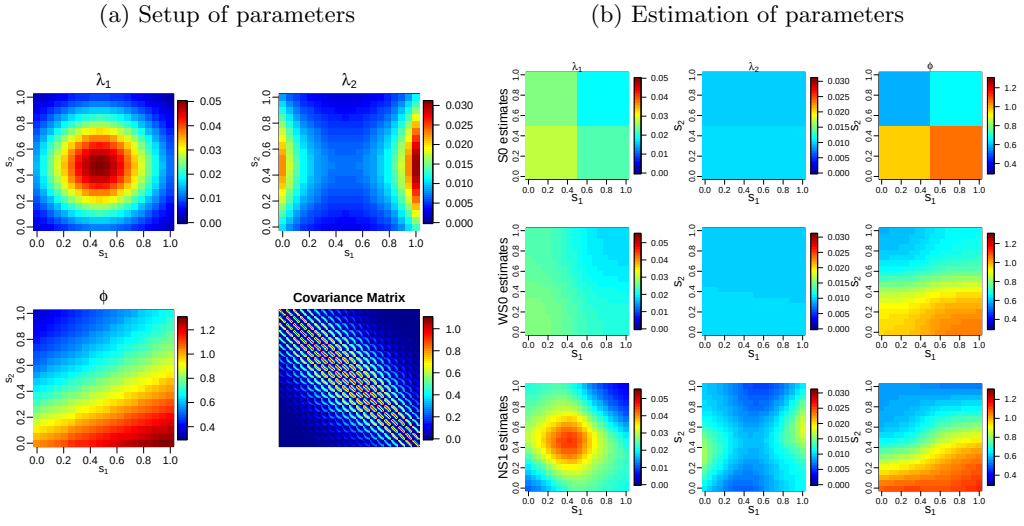

Figure S6: (a): The contour-image plots of three parameters, $(\lambda_1, \lambda_2, \phi)$, of the kernel matrix and the corresponding covariance matrix. (b): The contour-image plots of the corresponding estimations based on three models.

Similarly, we repeat the estimation 100 times. In Table 1, we show the the mean, standard error, and the nonstationarity index, D_1, of 100 estimations for each non-stationary parameter. In addition, we also calculate the MSEs based on the mean of 100 estimations. To visualize it, Figure

S6(b) shows the mean estimates of each spatially varying parameter for the three methods. It can be seen that, similar to the 1D simulation study, our method gives more accurate estimation in terms of MSEs but with larger uncertainty for the estimators. The difference is less significant than 1D case since more parameters are estimated together.

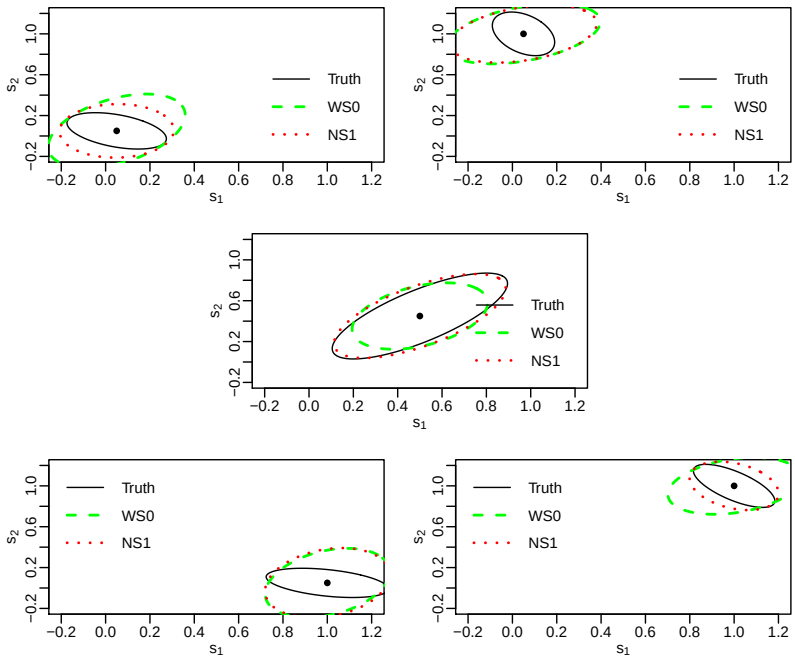

Figure S7: the ellipse-contour plots of the true and estimated kernel matrix at five reference locations.

Furthermore, to validate the estimations of the corresponding kernel matrices of size 2×2, we choose five reference locations and draw the contour ellipses as shown in Figure S7. In general, our model has a better fit with

the kernel matrix corresponding to the location at the center, whereas the borders are more difficult to estimate. Even so, wherever the location is, our NS1 model never performs worse than the WS0 model with regard to the contour ellipses. Therefore, for spatial problem with overall or local trend-non-stationarity, our model is a better solution.

Table 1: Summary of results from 100 estimations. Definitions of columns: (1)-(2) the mean (standard deviation) of the S0 estimators $\hat{\beta}_0$ and slope estimators $\hat{\beta}_1$ and $\hat{\beta}_2$, respectively; (3) the empirical trend-nonstationarity index in the k-th subregion $\hat{D}_{1k}$; (4) the mean squared error MSE_{NS1} of $\hat{\lambda_1}(s)$,$\hat{\lambda_2}(s)$, and $\hat{\phi}(s)$, calculated from the mean of $\hat{\beta}_0$s, $\hat{\beta}_1$s, and $\hat{\beta}_2$s from the NS1 model; (5) $\text{MSE}_{\text{S0}}/\text{MSE}_{\text{NS1}}$; (6) $\text{MSE}_{\text{WS0}}/\text{MSE}_{\text{NS1}}$.

Subregion	1	2	3	4
		$\log\{\lambda_1(\mathbf{s})\}$		
$\hat{\beta}_0$	-3.59 (0.31)	-3.82 (0.44)	-3.70 (0.35)	-4.01 (0.35)
$\hat{\beta}_1$	1.95 (0.64)	-0.16 (0.65)	0.36 (0.50)	-2.47 (0.85)
$\hat{\beta}_2$	2.53 (0.73)	0.16 (0.64)	-0.38 (0.53)	-2.68 (0.76)
$\hat{D}_1$	2.24	0.16	0.37	2.57
MSE_{NS1}		0.360		
$\text{MSE}_{\text{S0}}/\text{MSE}_{\text{NS1}}$		1.585		
$\text{MSE}_{\text{WS0}}/\text{MSE}_{\text{NS1}}$		1.570		
		$\log\{\lambda_2(\mathbf{s})\}$		
$\hat{\beta}_0$	-4.57 (0.32)	-4.58 (0.43)	-4.63 (0.35)	-4.64 (0.33)
$\hat{\beta}_1$	-2.06 (0.71)	0.29 (0.69)	-0.09 (0.54)	2.37 (0.61)
$\hat{\beta}_2$	1.75 (0.70)	-0.15 (0.68)	-0.02 (0.56)	-1.85 (0.71)
$\hat{D}_1$	1.91	0.22	0.05	2.11
MSE_{NS1}		0.148		
$\text{MSE}_{\text{S0}}/\text{MSE}_{\text{NS1}}$		2.331		
$\text{MSE}_{\text{WS0}}/\text{MSE}_{\text{NS1}}$		2.900		
		$\text{logit}\{\phi(\mathbf{s})\}$		
$\hat{\beta}_0$	0.46 (0.17)	0.73 (0.25)	-0.52 (0.21)	-0.30 (0.35)
$\hat{\beta}_1$	0.40 (0.96)	0.50 (0.52)	0.03 (0.30)	0.05 (1.43)
$\hat{\beta}_2$	-2.20 (0.84)	-0.52 (0.47)	-0.10 (0.26)	-1.43 (1.31)
$\hat{D}_1$	1.30	0.51	0.06	0.74
MSE_{NS1}		0.042		
$\text{MSE}_{\text{S0}}/\text{MSE}_{\text{NS1}}$		529.105		
$\text{MSE}_{\text{WS0}}/\text{MSE}_{\text{NS1}}$		11.363		

S3 Additional Figures

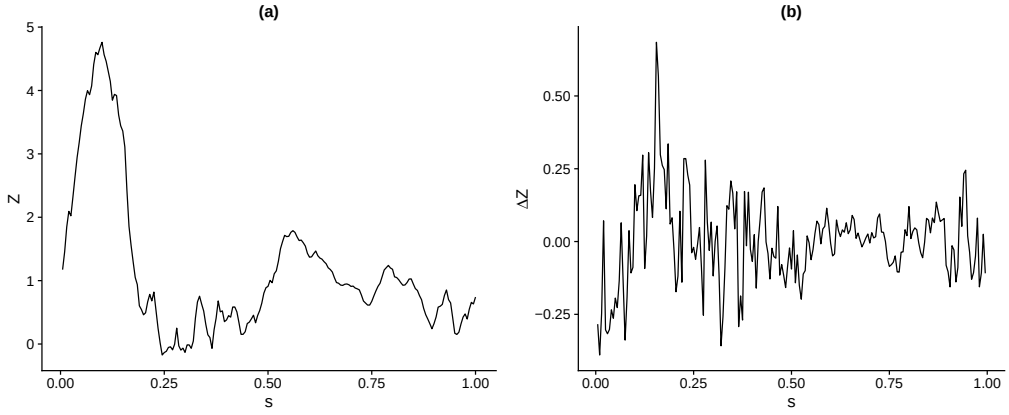

Figure S8: (a) A realization and (b) its increment of a GRF, $Z(s) = \sigma(s)W(s), s \in [0, 1]$, where $W(s)$ is a zero-mean stationary GRF with a Matérn covariance function with parameters $(\tau, \nu, \lambda) = (0, 1, 0.2)$ and $\sigma(s) = 2\sin(s/0.15) + 2.8$.

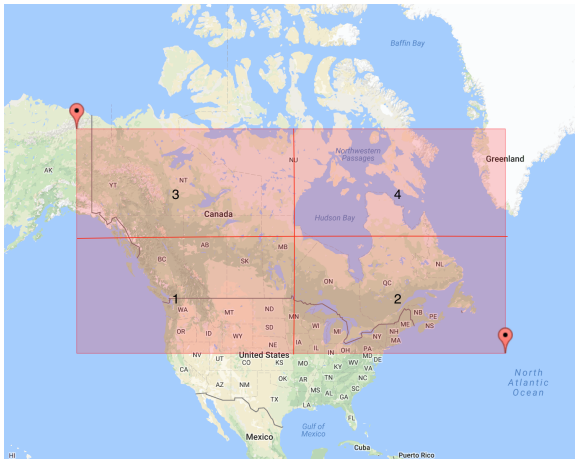

Figure S9: The region of interest with the subregions numbered. The whole region includes most of the North America in a rectangular area from (19.45°N, 157.71°W) to (70.26°N, 35.72°W)

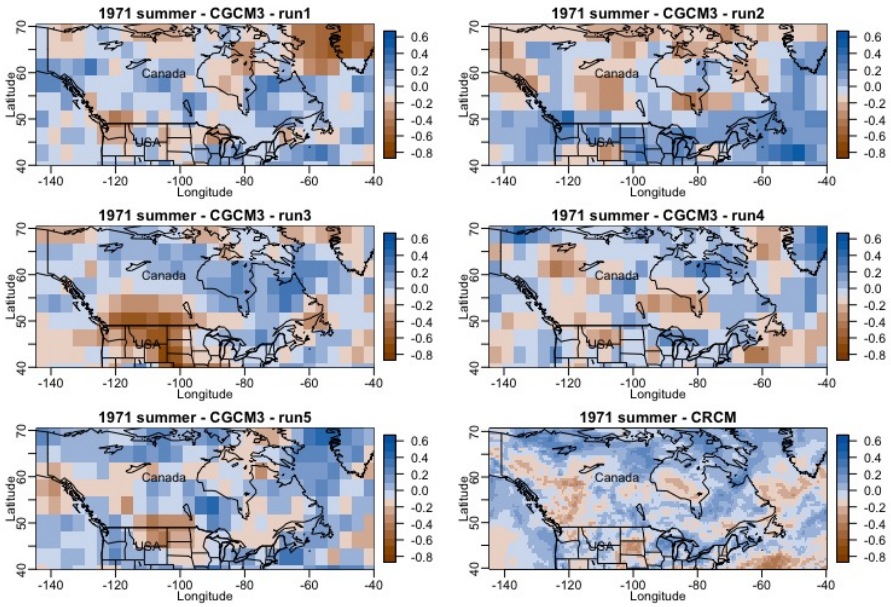

Figure S10: Preprocessed 1971 summer precipitation rate residuals (mm/day): five GCM runs and one RCM run over the region shown in Figure S9.

S4 R Source Codes

The R codes used in this study can be found on the website:

https://github.com/aleksada/Nonstationary-estimation/tree/master.

The R codes of the 1D and 2D simulation studies are stored in `1d.simulation.Rmd` and `2d.simulation.Rmd`, respectively.

The R codes of the application are stored in `application.Rmd`.

S5 GCM and RCM Data and Related Source

The data used in the application can be also found on the website:

https://github.com/aleksada/Nonstationary-estimation/tree/master

The related RCM runs can be downloaded from:

http://www.cccsn.ec.gc.ca/?page=dd-gcm

The related GCM runs can be downloaded from:

http://climate-modelling.canada.ca/data/cgcm3/cgcm3.shtml

Bibliography

Fredericks, A. K. and Newman, K. B. (1998). A comparison of the sequential gaussian and markov-bayes simulation methods for small samples. *Mathematical Geology*, 30(8):1011–1032.

Gotway, C. and Rutherford, B. (1994). Stochastic simultation for imaging spatial uncertainty: Comparison and evaluation of available algorithms. In *Geostatistical simulations*, pages 1–21. Springer.

Katzfuss, M. and Guinness, J. (2017). A general framework for Vecchia approximations of Gaussian processes. *arXiv preprint arXiv:1708.06302*.

Kleiber, W. and Nychka, D. (2012). Nonstationary modeling for multivariate spatial processes. *Journal of Multivariate Analysis*, 112:76–91.

Paciorek, C. J. and Schervish, M. J. (2006). Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics*, 17(5):483–506.

Risser, M. D., Calder, C. A., Berrocal, V. J., and Berrett, C. (2016). Nonstationary spatial process modeling via treed covariate segmentation, with application to soil organic carbon stock assessment. *arXiv preprint arXiv:1608.05655*.

Vecchia, A. V. (1988). Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 297–312.

Wand, M. P. and Jones, M. C. (1994). *Kernel Smoothing*. CRC Press.