Non-stationary nested SPDE models applied to global ozone data

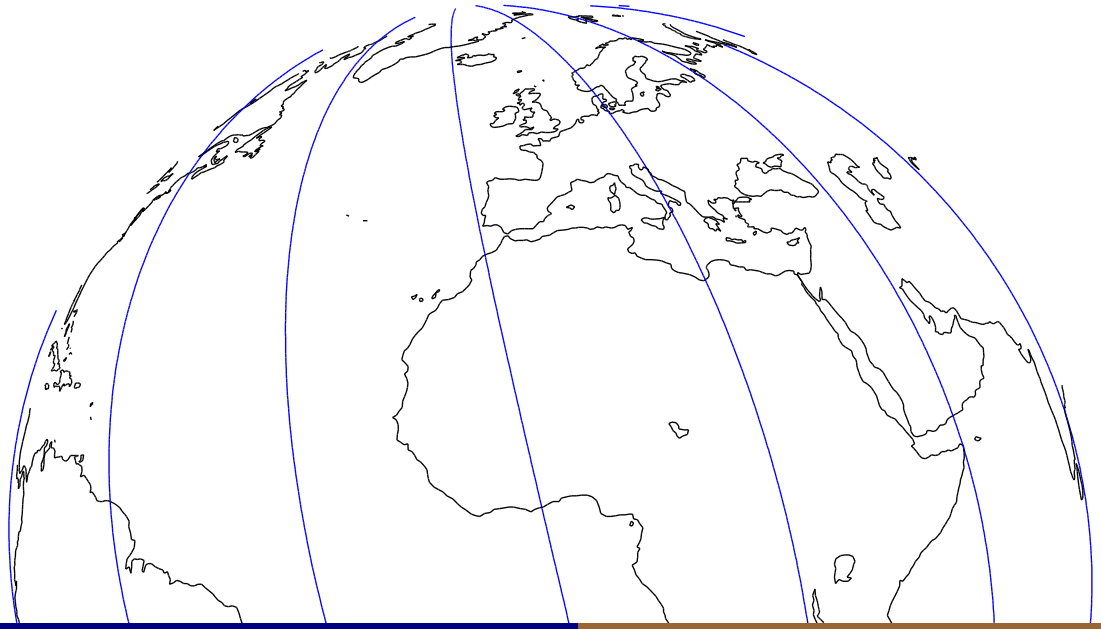
David Bolin, Finn Lindgren

Centre for Mathematical Sciences
Lund University

Smögen, August 16, 2010

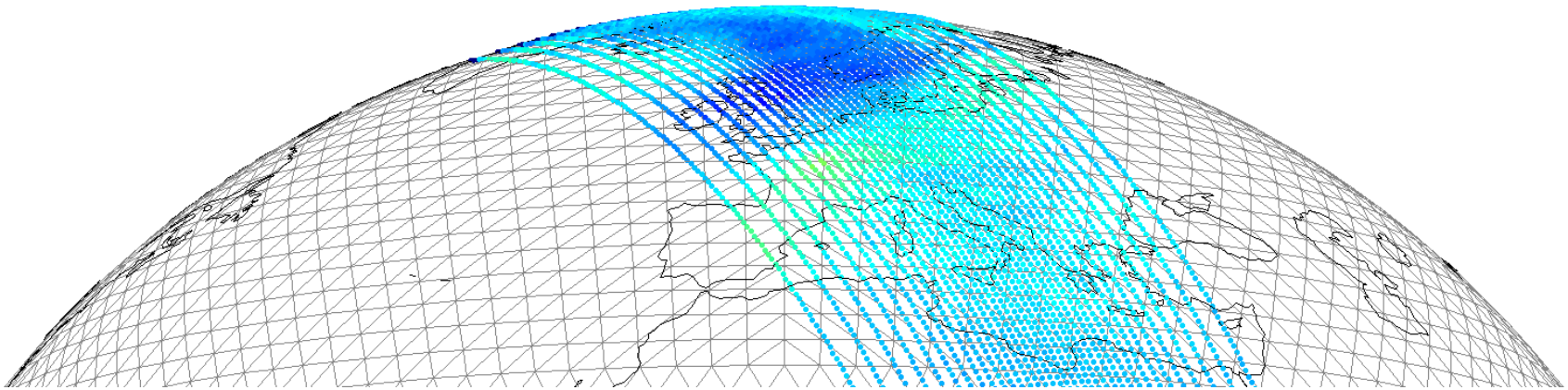
Modeling spatial data

- Modeling spatial environmental data is a challenging problem:
  - Non-stationary covariance models are often needed.
  - Computational efficiency is an increasingly important property.
- One such data set is the the Total Ozone Mapping Spectrometer (TOMS) atmospheric ozone data.
- Collected by a TOMS instrument onboard the the near-polar, Sun-synchronous orbiting satellite Nimbus-7.

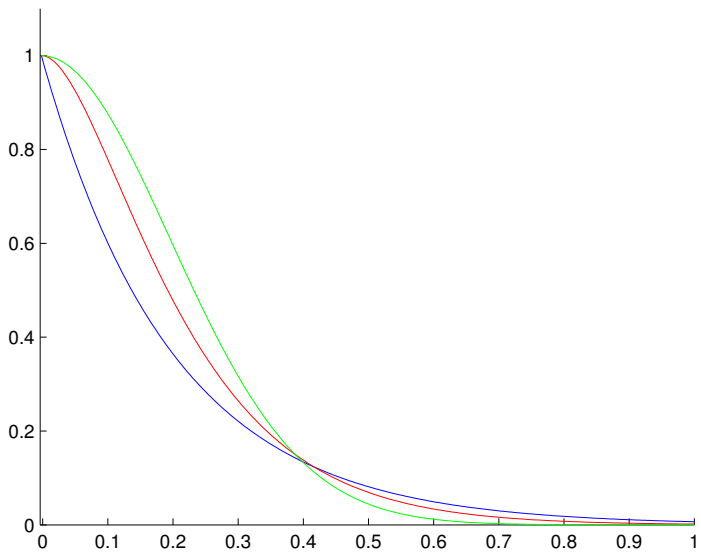


TOMS ozone data

- During the sunlit portions of the satellite's orbit, the instrument collected data in scans perpendicular to the orbital plane.
- The data is spatially and temporally irregular, requires non-stationary covariance structures, and efficient computational techniques due to the massive amounts of data.



The Matérn covariance function

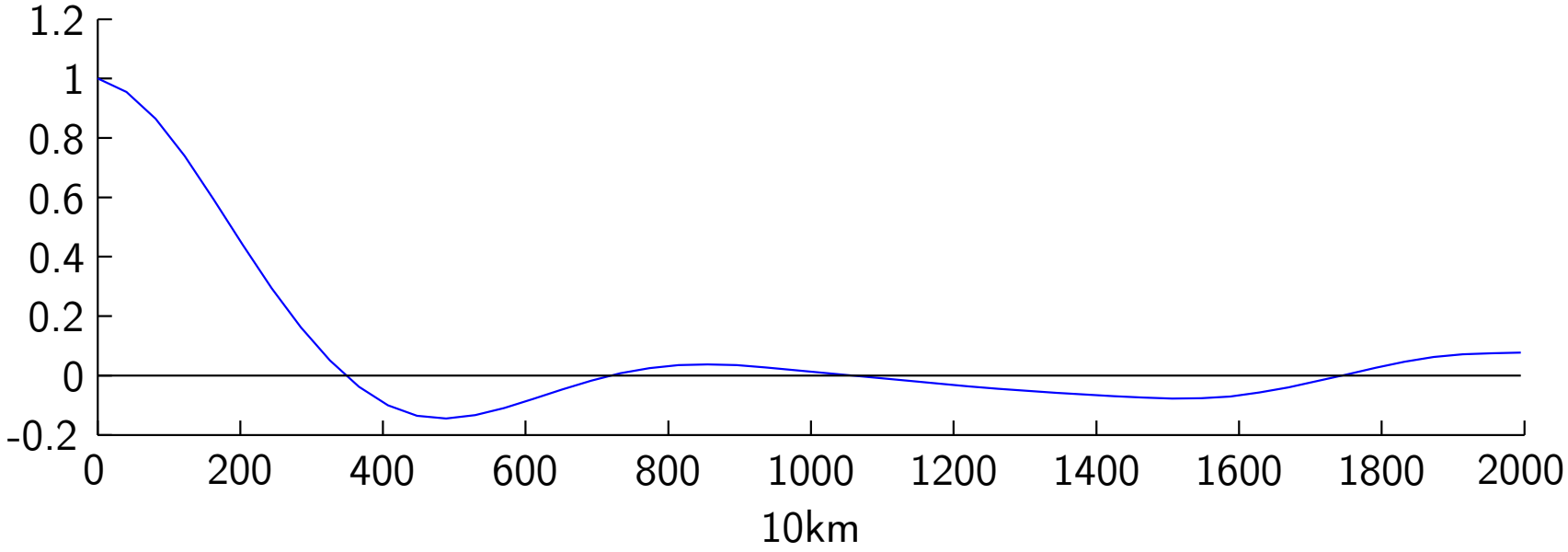


- One of the most popular covariance functions for spatial data.
- Has three parameters, κ, ν and σ, and can be parametrized as

$$C(\mathbf{h}) = \frac{2^{1-\nu}\sigma^2}{(4\pi)^{\frac{d}{2}}\Gamma(\nu + \frac{d}{2})}(\kappa\|\mathbf{h}\|)^{\nu} K_{\nu}(\kappa\|\mathbf{h}\|), \qquad \mathbf{h} \in \mathbb{R}^d,$$

where K_{ν} is a modified Bessel function of the second kind of order $\nu > 0$.

The Matérn covariance family



- There are no oscillating covariance functions in the Matérn family.
- Computational cost for spatial prediction of one location given m observations is $\mathcal{O}(m^3)$.

Nested SPDE models

- A Matérn field is the solution to a stochastic partial differential equation (SPDE)

$$(\kappa^2 - \Delta)^{\frac{\alpha}{2}} X(\mathbf{s}) = \sigma \mathcal{W}(\mathbf{s})$$

  where $\alpha = \nu + d/2$, $\mathcal{W}(\mathbf{s})$ is Gaussian white noise and Δ is the Laplace operator.
- An easy generalization of the generating SPDE is

$$(\kappa^2 - \Delta)^{\frac{\alpha}{2}} X(s) = (b + \mathbf{B}^\top \nabla) \mathcal{W}(s) \tag{1}$$

  where ∇ is the gradient and $\mathbf{B}$ is a vector.
- Since the two operators are commutative, (1) is equivalent to the following system of nested SPDEs

$$(\kappa^2 - \Delta)^{\frac{\alpha}{2}} X_0(s) = \mathcal{W}(s),$$
$$X(s) = (b + \mathbf{B}^\top \nabla) X_0(s).$$

Nested SPDE models

- To get a larger class of random fields, the orders of the operators can be increased.
- The class of nested SPDE models is defined as solutions to

$$\left(\prod_{i=1}^{n_1}(\kappa_i^2-\Delta)^{\frac{\alpha_i}{2}}\right)X(s)=\left(\prod_{i=1}^{n_2}(b_i+\mathbf{B}_i^\top\nabla)\right)\mathcal{W}(s)$$

for $\kappa_i>0, b_i\in\mathbb{R}, \alpha_i\in\mathbb{N}$, and $\mathbf{B}_i\in\mathbb{R}^d$.

- Differentiations in several directions $\mathbf{B}_i$.
- If the solution to this system, X, should be at least as "well behaved" as white noise, one must have $\sum_{i=1}^{n_1}\alpha_i\geq n_2$.

Properties in $\mathbb{R}^d$

- Spectral density

  Proposition 2.1 The spectral density for X is given by

$$S(\mathbf{k}) = \frac{\phi^2}{(2\pi)^d} \frac{\prod_{j=1}^{n_2}(b_j^2 + \mathbf{k}^\top \mathbf{B}_j \mathbf{B}_j^\top \mathbf{k})}{\prod_{j=1}^{n_1}(\kappa_j^2 + \|\mathbf{k}\|^2)^{\alpha_j}}$$

- Sample path regularity

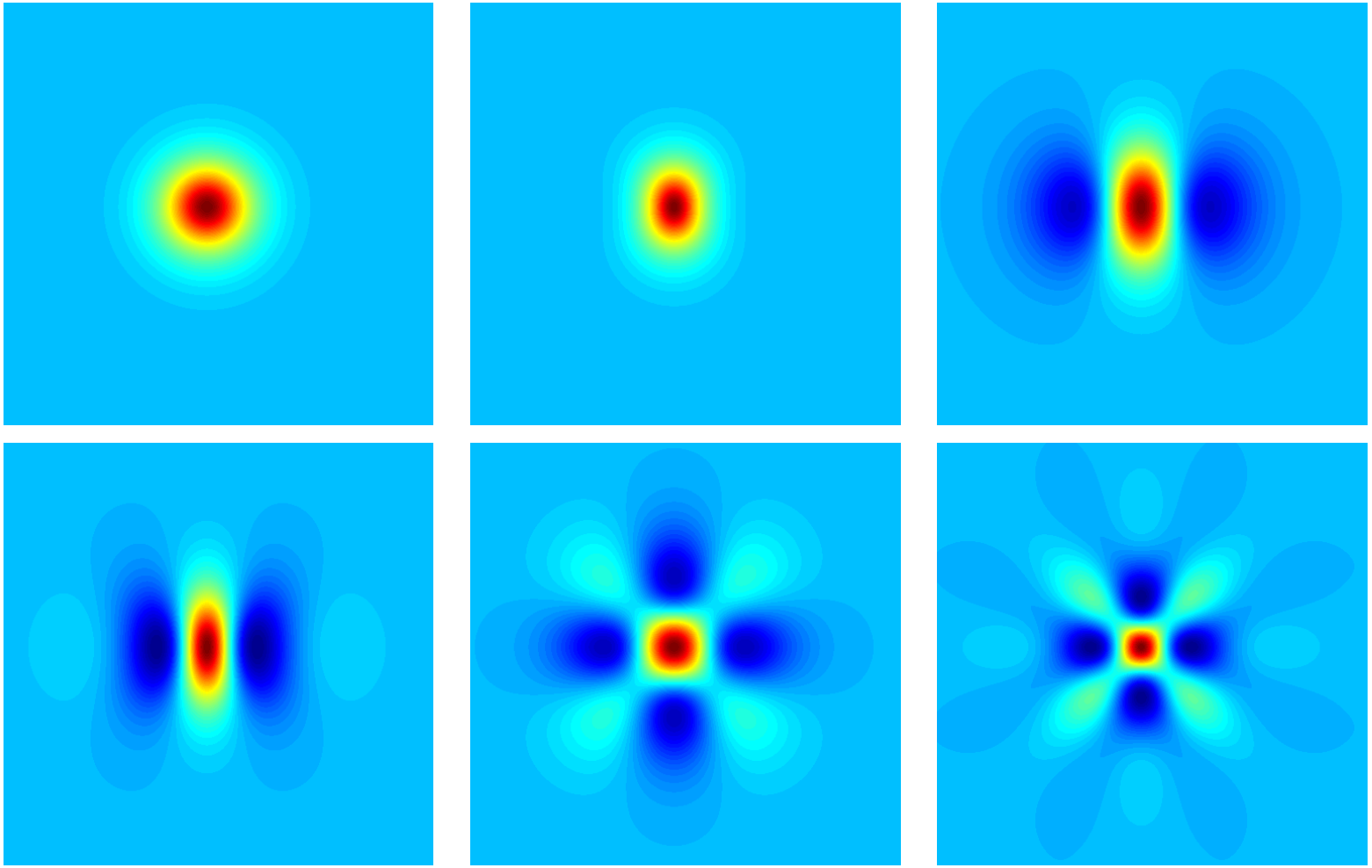
  Proposition 2.2 X has almost surely continuous sample functions if $2\sum_{i=1}^{n_1} \alpha_i - 2n_2 > d$.

- Sample path differentiability

  Corollary 2.3 The m:th order directional derivative of X has almost surely continuous sample functions given that $2\sum_{i=1}^{n_1} \alpha_i - 2n_2 - d > m$.

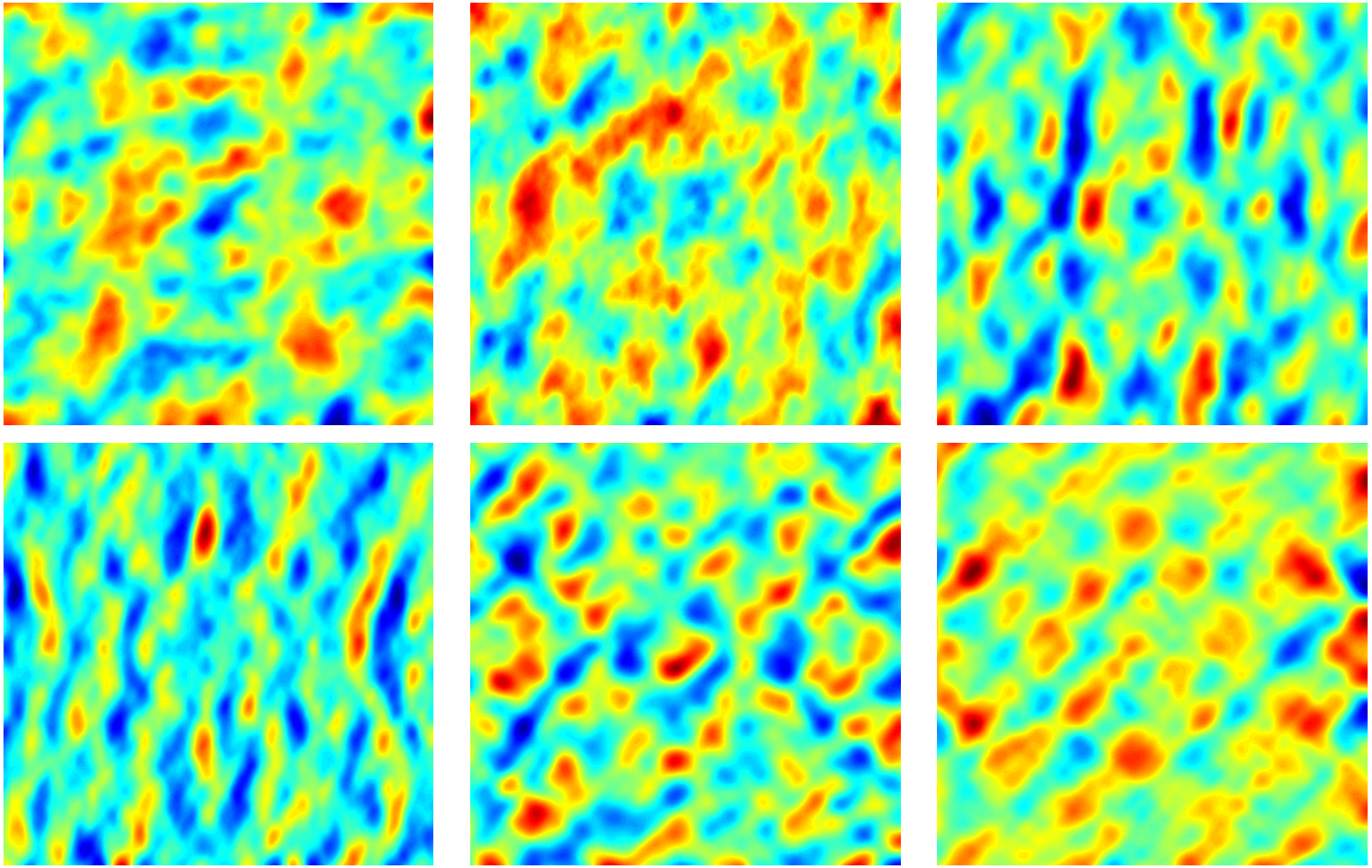
Covariance functions

- Wide class of explicit covariance functions.



Realizations

- Corresponding realizations.



Hilbert space approximations

- For large data sets, a computationally efficient model representation is needed.
- For Matérn fields, Lindgren, Rue and Lindström (2010) derived a Hilbert space approximation technique for obtaining *high-rank* basis expansions $X(s) \approx \sum_{i=1}^{n} \varphi_i(s) v_i$.
- The method uses the stochastic weak formulation of the SPDE

$$(\kappa^2 - \Delta)^{\frac{\alpha}{2}} X(\mathbf{s}) = \sigma \mathcal{W}(\mathbf{s}),$$

  with respect to some test functions to calculate the weights $\mathbf{v}$.
- For certain bases, the weights form a Gaussian Markov Random Field:

$$\mathbf{v} \sim \mathrm{N}(0, \mathbf{Q}^{-1}),$$

- The precision matrix, $\mathbf{Q}$, is sparse.

Hilbert space approximations

$$\left(\prod_{i=1}^{n_1}(\kappa_i^2 - \Delta)^{\frac{\alpha_i}{2}}\right) X(s) = \left(\prod_{i=1}^{n_2}(b_i + \mathbf{B}_i^\top \nabla)\right) \mathcal{W}(s)$$

- The method can be extended to give a similar basis expansion as for Matérn fields: $X(s) \approx \sum_{i=1}^{n} \varphi_i(s) w_i$.
- $\mathbf{w}$ will not be a GMRF, but we have

$$\mathbf{w} \sim \mathrm{N}(0, \mathbf{H}\mathbf{Q}^{-1}\mathbf{H}^\top)$$

  where $\mathbf{H}$ and $\mathbf{Q}$ are sparse matrices.
- We have lost the Markov structure, but the computational benefits can be preserved using the fact that $\mathbf{w} = \mathbf{H}\mathbf{v}$, where $\mathbf{v}$ is a GMRF.
- The computational cost for Kriging in $\mathbb{R}^2$ is $\mathcal{O}(n^{\frac{3}{2}})$ instead of $\mathcal{O}(m^3)$ for covariance based models.

A non-stationary nested SPDE model

- We analyze the data from October 1, 1988.
- This data set contains approximately 180 000 measurements.
- Assume that the data are measurements of the latent ozone field $X(s)$ under Gaussian measurement noise.

$$Y(s) = X(s) + \epsilon(s), \quad \epsilon(s) \sim \mathrm{N}(0, \sigma^2)$$

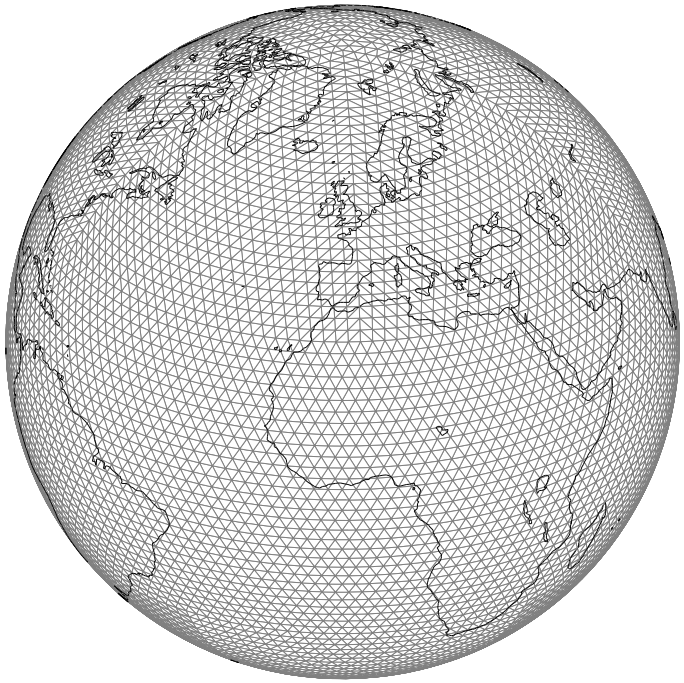
- We let $X(s)$ have some mean value $\mu(s)$, and, inspired by Jun and Stein (2008) assume that $Z(s) = X(s) - \mu(s)$ follows the simplest nested SPDE model:[1]

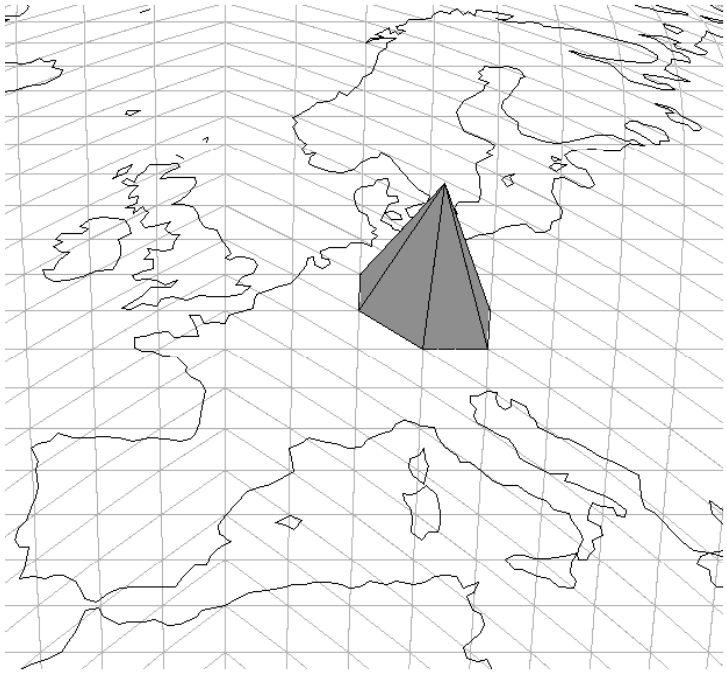
$$(\kappa^2(s) - \Delta) Z_0(s) = \mathcal{W}(s)$$
$$Z(s) = (b(s) + B(s)^\top \nabla) Z_0(s)$$

[1] Jun, M. and Stein M.L., Nonstationary covariance models for global data, *The Annals of Applied Statistics*, 2(4) (2008) 1271-1289

Hilbert space approximations

- The Hilbert space approximation technique is extended and used for the class of non-stationary nested SPDE models.
- A basis of 9002 piecewise linear functions induced by a triangulation of the earth is used.





Parameter estimation

- We assume a constant but unknown mean $\mu(s) = \mu$ and use the following regression models for the varying parameters.

$$b(s) = \exp\left(\sum_{k,m} b_{k,m} Y_{k,m}(s)\right), \quad B(s) = \sum_k b_k F_k(s)$$

$$\kappa^2(s) = \exp\left(\sum_{k,m} \kappa_{k,m} Y_{k,m}(s)\right),$$

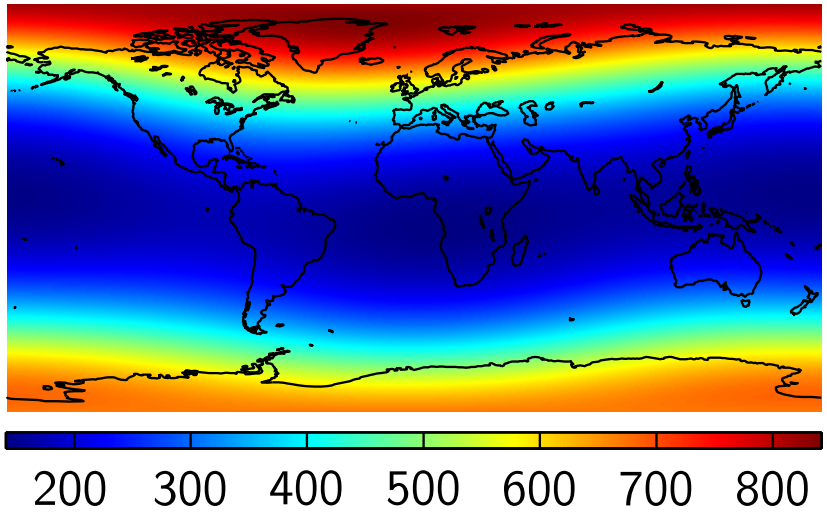
  where $Y_{k,m}$ is the spherical harmonic of order k and mode m and $\{F_k\}$ is a basis for vector fields on the sphere.

- The regression parameters are estimated by numerical optimization of the marginal posterior $\pi(\boldsymbol{\kappa}, \boldsymbol{\phi}, \mathbf{b}, \sigma^2 | Y)$.

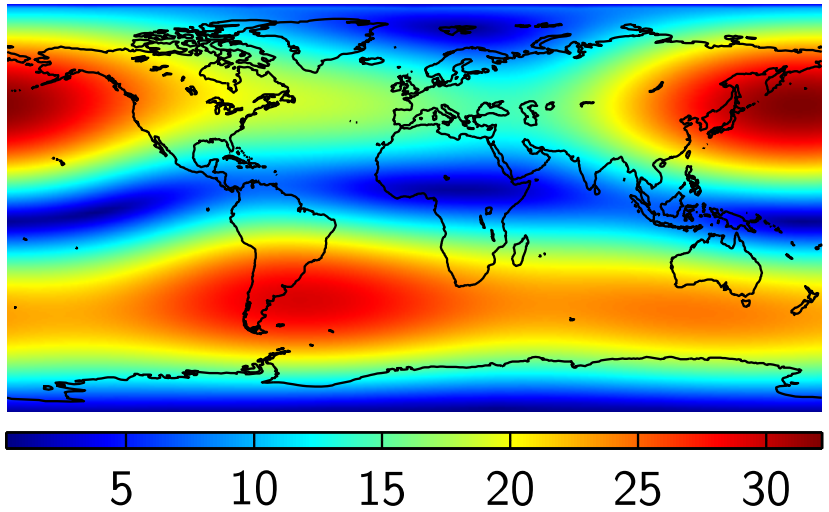
Parameter estimates

- The number of basis functions are chosen using AIC and BIC.
- Best model has constant range parameter $\kappa(\mathbf{s})$ and spatially varying vector field $\mathbf{B}(\mathbf{s})$ and variance scaling $b(\mathbf{s})$.

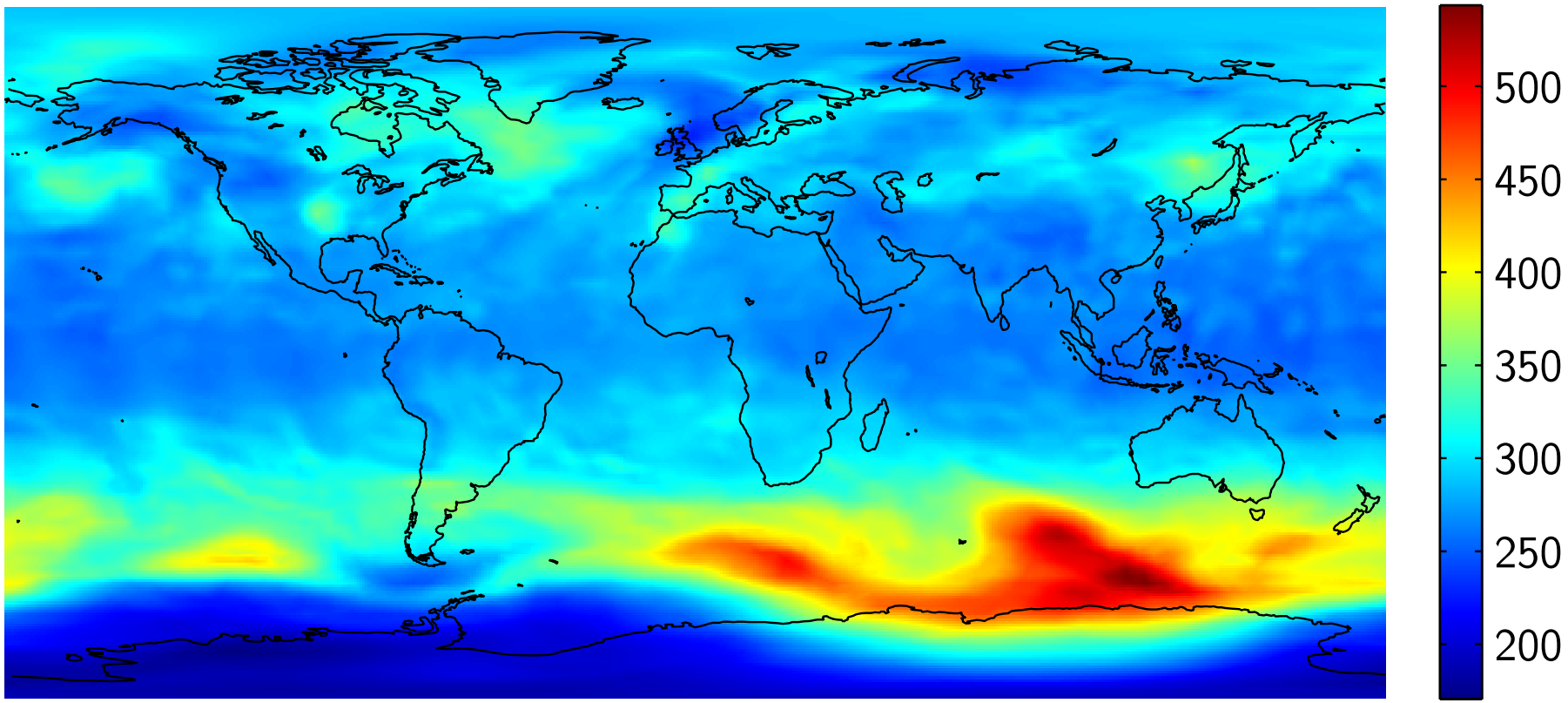
$b(\mathbf{s})$



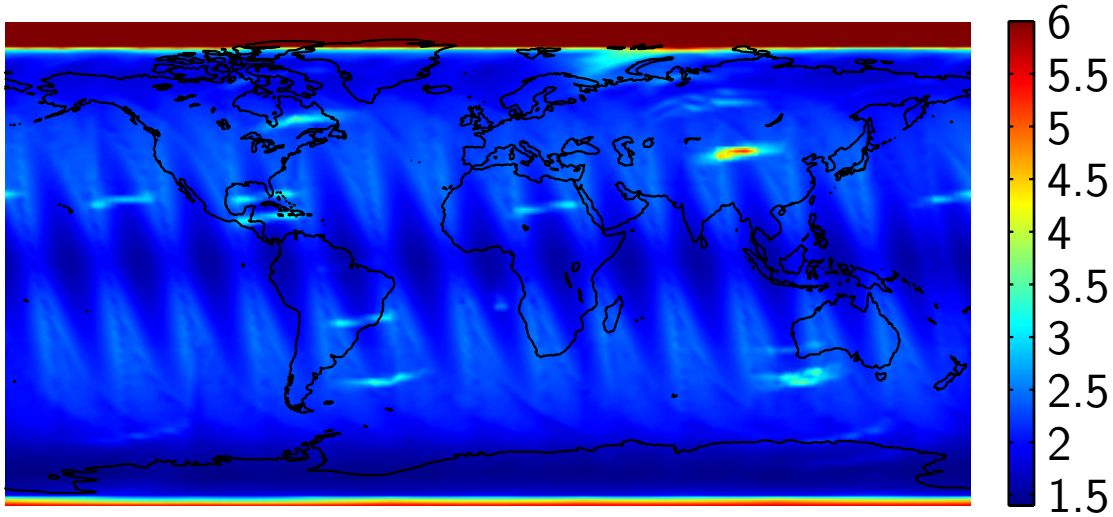
$|\mathbf{B}(\mathbf{s})|$

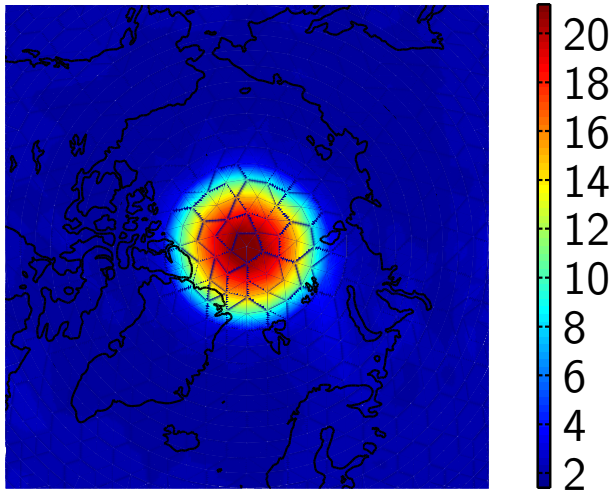


Kriging estimate

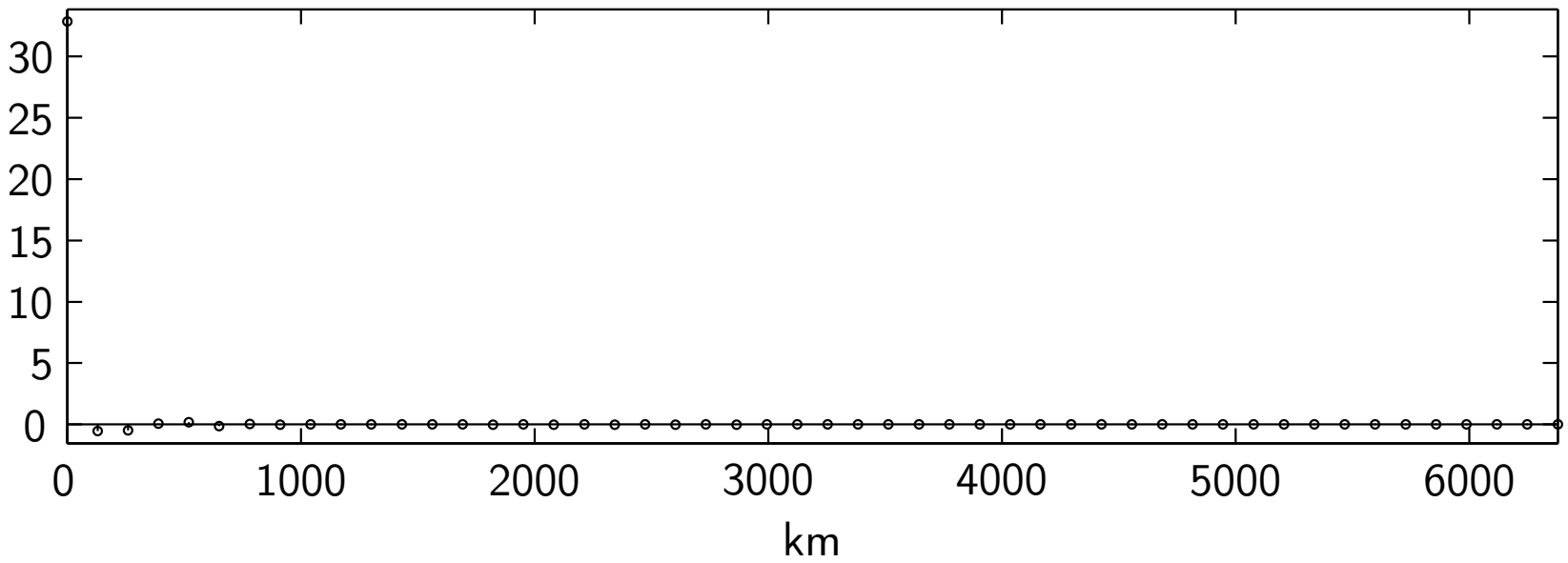


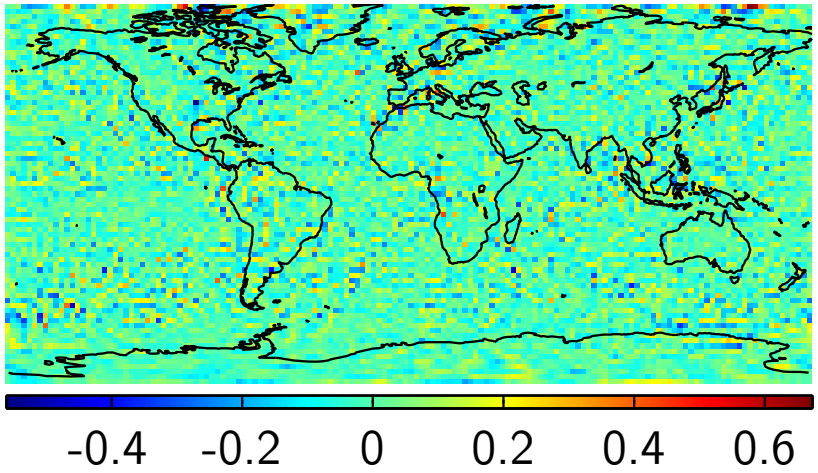
Standard errors

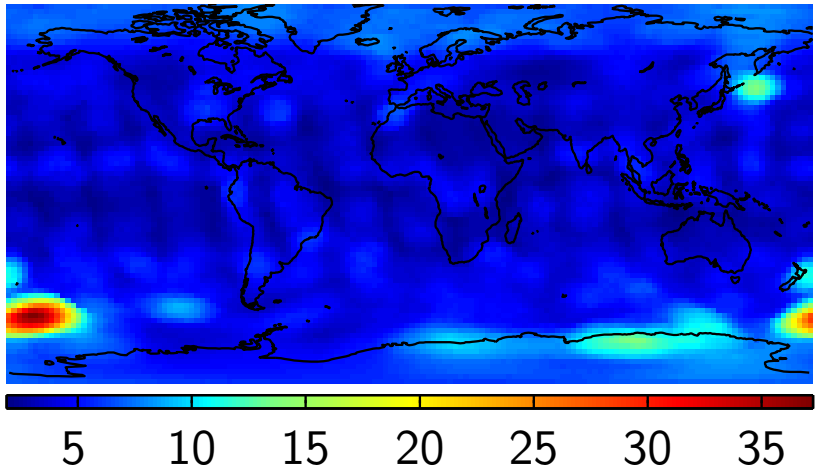




Kriging residuals







Conclusions

- The nested SPDE models is a wide en flexible model class that
  - easily can be made non-stationary.
  - is applicable to data on general smooth manifolds.
  - is computationally efficient.
- Good model fit for the ozone data
  - Non-stationary covariance structures necessary.
  - Likelihood estimation is not computationally feasible using covariance-based modeling.
  - The temporal aspect of the data should be modeled.
- Further work:
  - Spatio-temporal extensions.
  - Non-Gaussian versions.

Thanks for listening!

References

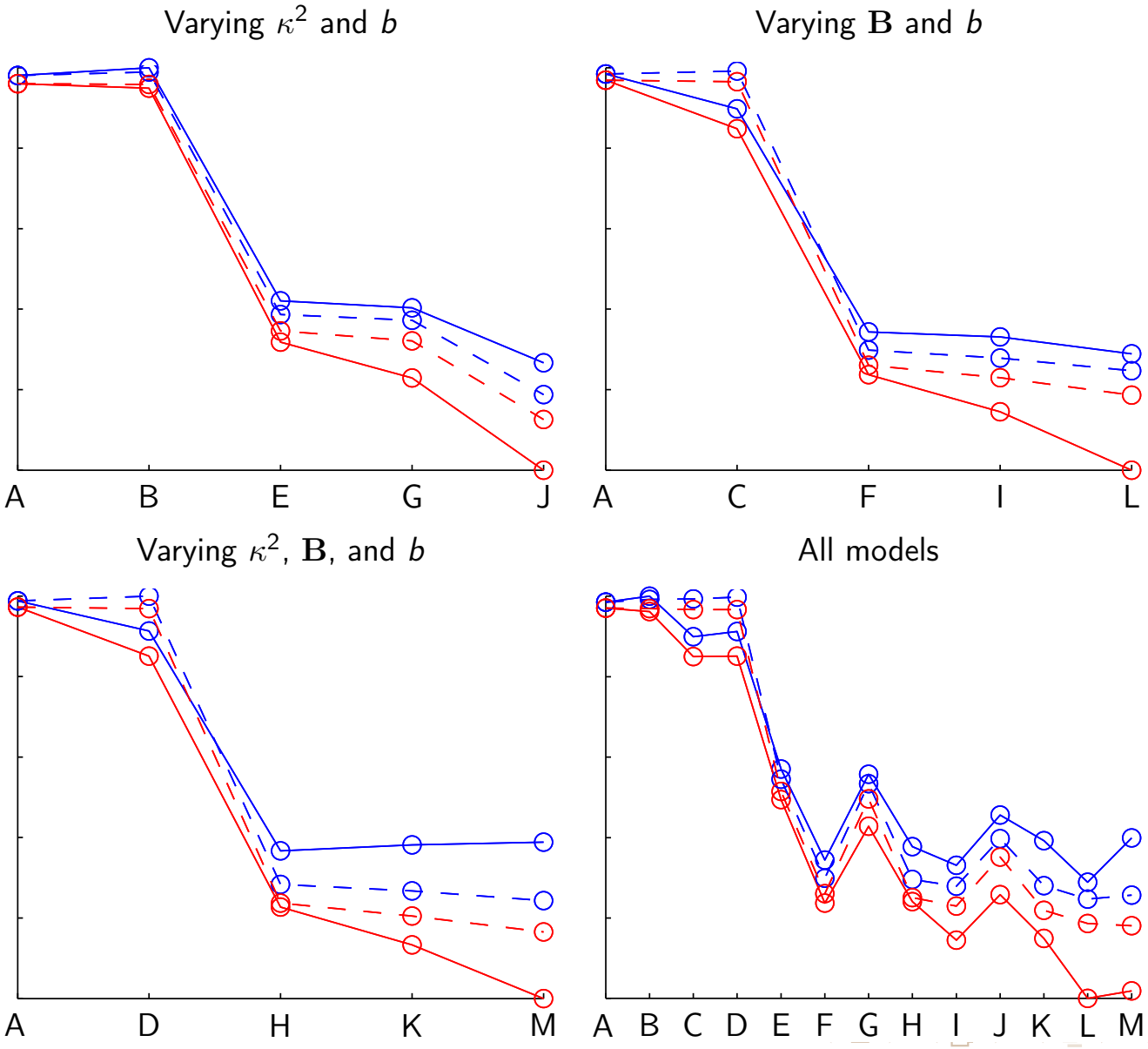
- Bolin, D. and Lindgren, F. Spatial models generated by nested stochastic partial differential equations
  To appear in Annals of Applied Statistics
- Lindgren, F., Lindström, J., and Rue, H., An Explicit link between Gaussian fields and Gaussian Markov random fields (submitted), *Preprints in Mathematical Sciences* (2010) 2010:3, Lund Univ.
- Jun, M. and Stein M.L., Nonstationary covariance models for global data, *The Annals of Applied Statistics*, 2(4) (2008) 1271-1289

Models for ozone data

	A	B	C	D	E	F	G	H	I	J	K	L	M
$\kappa^2(\mathbf{s})$	0	1	0	1	2	0	3	2	0	4	3	0	4
$b(\mathbf{s})$	0	1	1	1	2	2	3	2	3	4	3	4	4
$\mathbf{B}(\mathbf{s})$	0	0	1	1	0	2	0	2	3	0	3	4	4
Total	2	8	11	14	18	26	32	34	47	50	62	75	98

Table: Maximal orders of the spherical harmonics used in the bases for the different parameters and total number of covariance parameters in the different models for $X(\mathbf{s})$. The actual number of basis functions for $\kappa^2(\mathbf{s})$ and $b(\mathbf{s})$ are given by $(ord+1)^2$, and for $\mathbf{B}(\mathbf{s})$, the actual number is $2(ord+1)^2-2$, where ord is the maximal order indicated in the table.

AIC and BIC


Varying κ^2 and b
A B E G J
Varying $\mathbf{B}$ and b
A C F I L
Varying κ^2, $\mathbf{B}$, and b
A D H K M
All models
A B C D E F G H I J K L M