Geostatistical modeling using non-Gaussian Matérn fields

David Bolin
Chalmers University of Technology

joint work with Jonas Wallin

Búzios, RJ, Brazil,
June 25, 2014

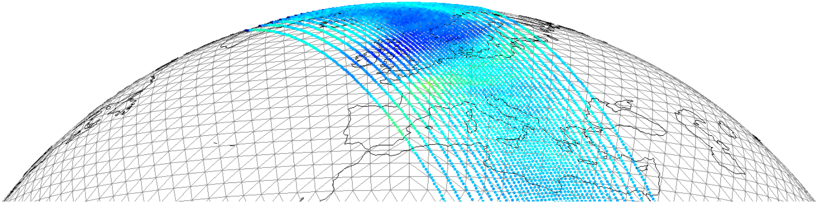
Modeling spatial data

- A typical geostatistical model:

$$Y(s) = X(s) + \mathcal{E}(s)$$

  X is Gaussian with some covariance function $r(s,t)$ and some mean, and $\mathcal{E}$ is Gaussian white noise.

- Modeling spatial environmental data is a challenging problem:
  - Non-stationary covariance models are often needed.
  - Spatially irregular data on other domains than $\mathbb{R}^d$.
  - Large datasets.
  - Gaussianity can not always be assumed.



So what do we do for non-Gaussian data?

- The standard approach is to try to find some non-linear transformation that enables the use of Gaussian models.
- This is commonly referred to as trans-Gaussian Kriging.
- Common transformations include the square root transform and the log transform.
- For example, consider a standard square root transformed latent Gaussian model

$$\sqrt{y_i} = Z(\mathbf{s}_i) + \epsilon_i$$
$$Z(\mathbf{s}) = \mathbf{B}(\mathbf{s})\boldsymbol{\beta} + X(\mathbf{s})$$

where

- y_i are the observations,
- $\epsilon_i \sim N(0, \sigma_\epsilon^2)$ is measurement noise,
- $X(\mathbf{s})$ is a mean-zero Gaussian field with a stationary covariance function.
- The mean $\mathbf{B}(\mathbf{s})\boldsymbol{\beta} = \sum_{k=1}^K B_i(\mathbf{s})\beta_k$ is modeled using covariates $\mathbf{B}(\mathbf{s})$.

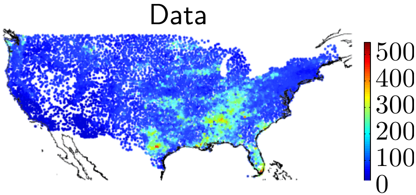
What does this actually mean?

- According to this model, the mean and covariance of the data in the original scale is given by

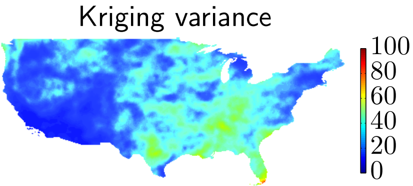
$$\begin{aligned}\mathsf{E}[y_i] &= \mathbb{C}_X(0) + 2(\mathbf{B}(\mathbf{s}_i)\boldsymbol{\beta})^2,\\ \mathbb{C}[y_i, y_j] &= 2\,\mathbb{C}_X(\mathbf{s}_i - \mathbf{s}_j)^2 + 4(\mathbf{B}(\mathbf{s}_i)\boldsymbol{\beta})(\mathbf{B}(\mathbf{s}_j)\boldsymbol{\beta})\,\mathbb{C}_X(\mathbf{s}_i - \mathbf{s}_j),\end{aligned}$$

  where $\mathbb{C}_X$ is the stationary covariance function of $X(\mathbf{s})$ with the measurement variance σ_ϵ^2 added at $\mathbf{0}$.
- It is not obvious how to interpret the effect of the measurement error.
- The usage of covariates for the mean induces a non-stationary covariance function for the data.

How does it affect kriging predictions?



Data



Kriging variance

- The posterior variance of the process in the same scale as the data is given by

$$\mathsf{V}[X(\mathbf{s})^2|\mathbf{y}] = 2\mathsf{V}[X(\mathbf{s})|\mathbf{y}]^2 + 4\mathsf{E}[X(\mathbf{s})|\mathbf{y}]^2\mathsf{V}[X(\mathbf{s})|\mathbf{y}].$$

- Hence, the observations $\mathbf{y}$ and the mean field affects the kriging variance for the transformed Gaussian model.

Does it really matter?

- The dependence between mean and covariance is often not unreasonable for real data.
- However, as the models grow more complex, for example by introducing
  - non-stationary covariance functions,
  - spatially varying measurement errors,
  - or covariates for the mean,

  the effects of the transformation methods become less transparent and more stale.
- In these situations, one would like to use latent non-Gaussian models without resorting to transformations.
- By doing this, we will be able to separate non-stationarity in the mean from non-stationarity in covariance.

non-Gaussian Matérn fields

A Matérn field is a solution to the 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 Laplacian (Lindgren et al 2011).

- Goal: Formulate a model with Matérn covariances and non-Gaussian marginal distributions.
- Idea: Replace the Gaussian noise $\sigma \mathcal{W}$ with a non-Gaussian process $\dot{M}$:

$$(\kappa^2 - \Delta)^{\frac{\alpha}{2}} X = \dot{M}$$

- We need to make sure that this makes sense, and we need to know if we can obtain anything useful (i.e. computationally feasible) from it.

Yes, we can solve this and it makes sense.

Theorem (Bolin, 2013)

Assume that M is an independently scattered L_2-valued random measure with $\mathsf{E}(|M(\mathrm{d}\mathbf{x})|^2) = C\,\mathrm{d}\mathbf{x}$. Then for $\kappa > 0$, $\alpha > 0$, there exists a random functional $X : H_n \times \Omega \to \mathbb{R}$ such that for a certain set Ω_0, $P(\Omega_0) = 1$ and for all $\omega \in \Omega_0$ and all $\varphi \in H_n$

$$X(\varphi, \omega) = \int G^\alpha \varphi(\mathbf{x}) M(\mathrm{d}\mathbf{x}, \omega),$$

where $G^\alpha \varphi(\mathbf{x}) = \int G_\alpha(\mathbf{s}, \mathbf{x})\varphi(\mathbf{s})\,\mathrm{d}\mathbf{s}$ and G_α is given by

$$G_\alpha(\mathbf{s}, \mathbf{t}) = \frac{2^{1-\frac{\alpha-d}{2}}}{(4\pi)^{\frac{d}{2}}\Gamma(\frac{\alpha}{2})\kappa^{\alpha-d}}(\kappa\|\mathbf{s}-\mathbf{t}\|)^{\frac{\alpha-d}{2}} K_{\frac{\alpha-d}{2}}(\kappa\|\mathbf{s}-\mathbf{t}\|).$$

This is the unique H_n-solution to $(\kappa^2 - \Delta)^{\frac{\alpha}{2}} X = \dot{M}$ if $n > d/2$, and moreover we have $X \in H_m$ almost surely for $m < \alpha - d/2$.

Just a bit more math before we do something useful

- The solution X is in general a random linear functional.
- However, it can be identified with a random function if $\alpha > d/2$ since $X \in H_m$ almost surely for $m < \alpha - d/2$.
- Recall that the Sobolev embedding theorem shows that the Sobolev space H_n can be embedded in the Hölder space $C_k^r(\mathbb{R}^d)$ where $n - (r + k) = d/2$ and $r \in (0, 1)$.
- If $\nu > d/2$, one has that $X \in C_k^r(\mathbb{R}^d)$ almost surely, where k is the integer part of $\nu - d/2$ and $r = \nu - d/2 - k$.
- One should note here that only $\nu > 0$ is required for continuity in the Gaussian case.
- Thus, the non-Gaussian Matérn fields are in general less smooth than the Gaussian Matérn fields, for the same smoothness parameter ν.

Computationally efficient representations

- For Gaussian fields, Lindgren et al used a finite element method to represent a solution to the SPDE.
- Similarly, we use a finite element matrix transfer technique to obtain a discretized approximation of the solution.
- This technique yields computationally efficient representations if the driving noise is a type-G Lévy process.
- The most well known subclass of the type-G processes are the the generalized hyperbolic processes.
- We need the distribution to be closed under convolution.
- Only two special cases have this property (Podgórski and Wallin, 2013)
    - Generalized asymmetric Laplace (GAL) fields
    - Normal inverse Gaussian (NIG) fields

Computationally efficient representations (cont.)

- We represent $X(s)$ using a (high-rank) basis expansion:

$$X(s) = \sum_{i=1}^{n} \varphi_i(s) w_i$$

where $\{\varphi_i\}$ are usual piecewise linear FEM basis functions.

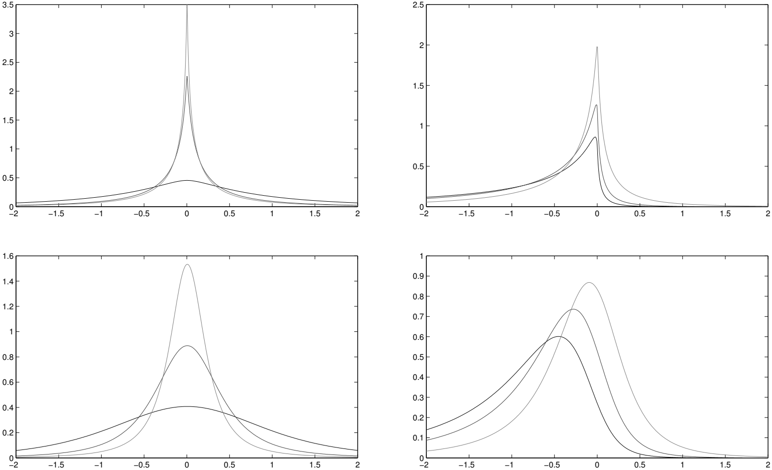
- Let $\mathbf{K} = \mathbf{G} + \kappa^2 \mathbf{C}$, where $C_{ij} = \langle \varphi_i, \varphi_j \rangle$ and $G_{ij} = \langle \nabla\varphi_i, \nabla\varphi_j \rangle$
- For NIG and GAL noise, we then have

$$\mathbf{Kw} \sim \mathsf{N}(\gamma\tau\mathbf{a} + \mu\mathbf{V}, \mathrm{diag}(\mathbf{V}))$$

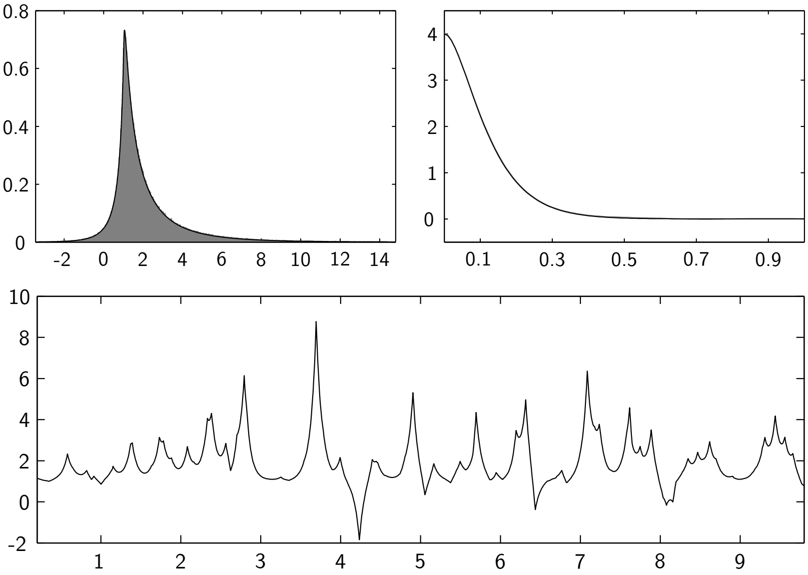
where $V_i \sim \Gamma(a_i\tau, 1)$ for GAL and $V_i \sim IG(\nu^2 a_i, 2)$ for NIG.

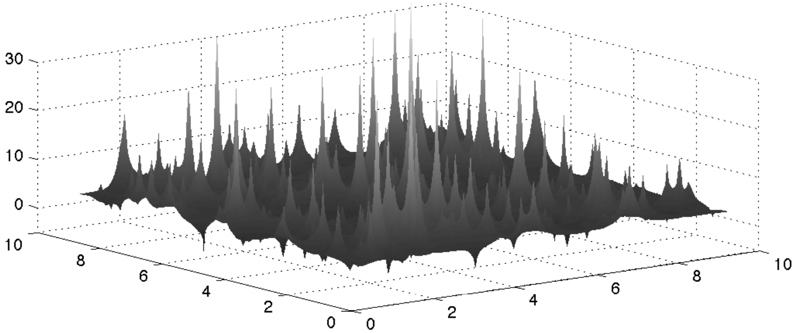
- The complicated SPDE with Sobolev space solutions has now transformed into a very simple linear equation system.

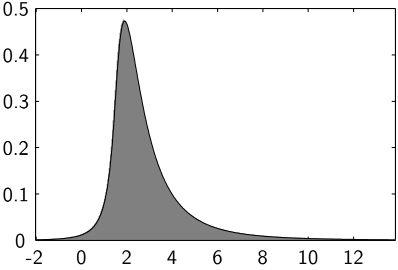
Examples of marginal distributions

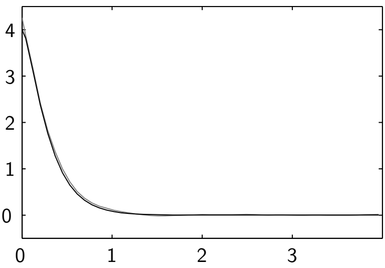


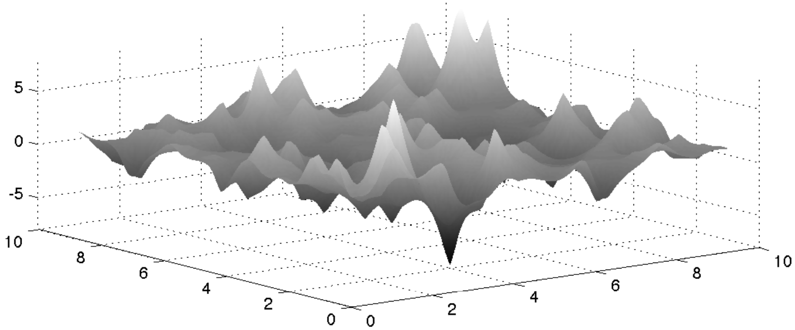
Simulated examples

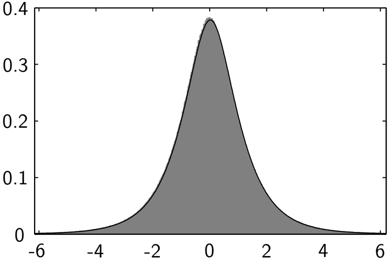


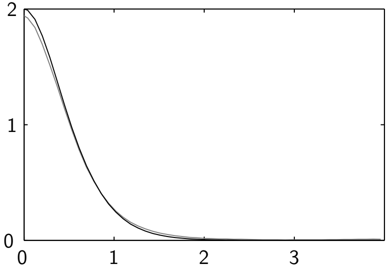

30
20
10
0
10
8
6
4
2
0
0
2
4
6
8
10


0.5
0.4
0.3
0.2
0.1
0
-2
0
2
4
6
8
10
12


4
3
2
1
0
0
1
2
3


5
0
-5
10
8
6
4
2
0
0
2
4
6
8
10


0.4
0.3
0.2
0.1
0
-6
-4
-2
0
2
4
6


2
1
0
0
1
2
3

Latent non-Gaussian models

- We can use Matérn fields driven by noise processes of this type in a hierarchical model:

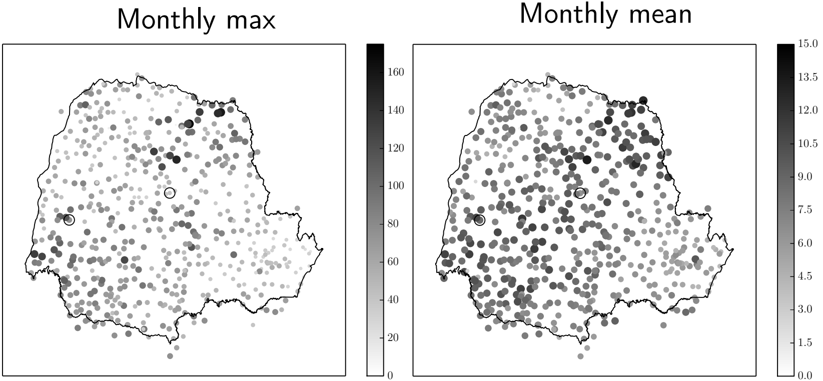
$$Y_i = Z(s_i) + \mathcal{E}_i$$

$$Z(s) = \sum_{i=1}^{p} B_i(s)\beta_i + X(s)$$

where $X(s)$ now is a non-Gaussian Matérn field.

- In practise, we need to estimate the following model parameters: $\boldsymbol{\beta}_x, \kappa, \sigma_\varepsilon, \gamma, \mu, \sigma$ and τ (GAL) or ν (NIG).
- We can do the parameter estimation in a likelihood framework using an MCEM algorithm.
- A better alternative is to use a gradient-based approach.
- For details on the estimation procedure, see the references.

Parana data



Monthly max

Monthly mean

Precipitation data for October 2012 over Parana, Brazil.

We want to create a high-resolution map of precipitation.

Models for the data

We compare four different models, for all models we use a latent field $Z(s) = \sum_{i=1}^{p} B_i(s)\beta_i + X(s)$ with $p = 3$ three covariates for the mean: $B_1 = 1$ (intercept), $B_2 =$ longitude, $B_3 =$ latitude.

For the first three models, we assume

$$Y_i = Z(s_i) + \mathcal{E}_i$$

and assume that $X(s)$ is a Matérn field driven by

1. Gaussian noise.
2. GAL noise.
3. NIG noise.

For the final model, we assume

$$\sqrt{Y_i} = Z(s_i) + \mathcal{E}_i$$

and assume that $X(s)$ is a Gaussian Matérn field.

Parameter estimates

	max				mean			
	Gauss	tGauss	NIG	GAL	Gauss	tGauss	NIG	GAL
κ	2.6	1.9	5.8	5.9	1.40	1	2.0	2.0
ϕ	11.6	2.5	-	-	2.75	1	-	-
σ_ε	16.3	1.1	14 .4	13.5	1.3	0.3	1.3	1.3
$\boldsymbol{\beta}$	$\begin{bmatrix}-79\\-3\\0.3\end{bmatrix}$	$\begin{bmatrix}7.8\\-0.3\\0.0\end{bmatrix}$	$\begin{bmatrix}36\\-6\\-6\end{bmatrix}$	$\begin{bmatrix}28\\-5\\-3\end{bmatrix}$	$\begin{bmatrix}6.0\\-0.1\\-0.16\end{bmatrix}$	$\begin{bmatrix}2.7\\-0.0\\0.0\end{bmatrix}$	$\begin{bmatrix}8\\-0.36\\0.36\end{bmatrix}$	$\begin{bmatrix}11\\-0.2\\0.5\end{bmatrix}$
$\boldsymbol{\mu}$	-	-	312	74	-	-	-1.8	-1.0
σ	-	-	0.0	0.0	-	-	8.3	2.3
ν^2	-	-	0.7	-	-	-	0.2	-
τ	-	-	-	17	-	-	-	15

Parameter estimates for the different models for the precipitation max and mean data. Note that the tGauss parameters should not be compared directly with the others since they are for transformed data.

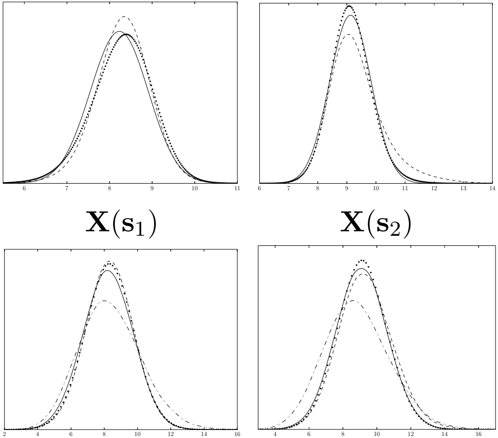
Model comparisons

	max				mean			
	Gauss	tGauss	NIG	GAL	Gauss	tGauss	NIG	GAL
V($\mathbf{r}_s$)	0.99	0.89	0.99	1.00	0.99	0.67	1.05	1.02
V($\mathbf{r}$)	327	334	330	295	2.05	2.12	2.12	2.06
ES	301	304	310	287	24.0	24.5	24.2	24.0
CRPS	9.8	9.7	9.7	9.3	0.76	0.79	0.77	0.76

Crossvalidation results for the different models. Here,

- V($\mathbf{r}$) denotes the variance of crossvalidated kriging residuals.
- V($\mathbf{r}_s$) denotes the variance of crossvalidated kriging residuals. standardized by the estimated kriging variances.
- $CRPS$ denotes the continuous ranked probability score of $\mathbf{r}$.
- ES denotes the energy score.

Posterior marginals for monthly mean



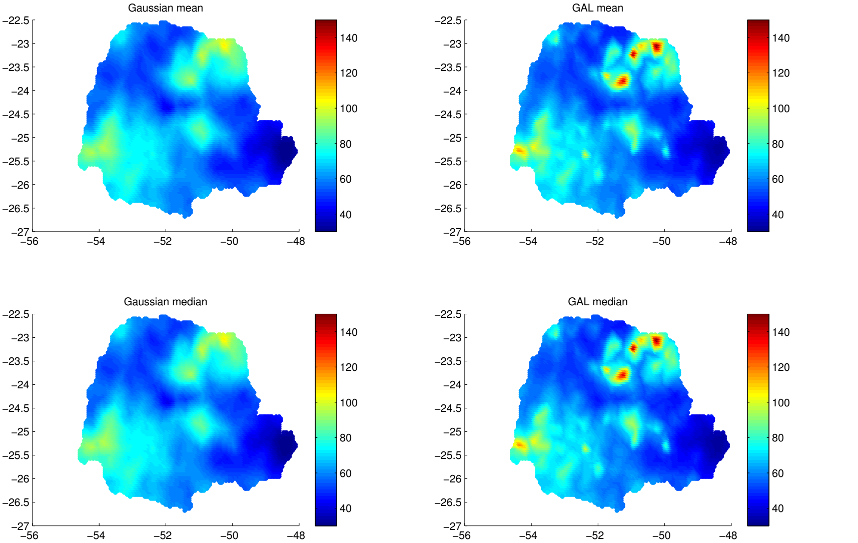
$\mathbf{X}(\mathbf{s}_1)$ $\mathbf{X}(\mathbf{s}_2)$

$\mathbf{X}(\mathbf{s}_1) + \boldsymbol{\epsilon}(\mathbf{s}_1)$ $\mathbf{X}(\mathbf{s}_2) + \boldsymbol{\epsilon}(\mathbf{s}_2)$

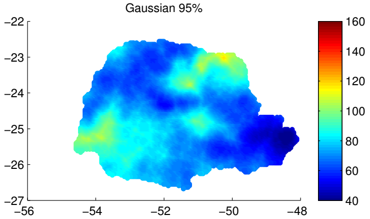
Gaussian model (solid), NIG model (dotted), GAL model (dashed), and transformed Gaussian model (dash-dotted).

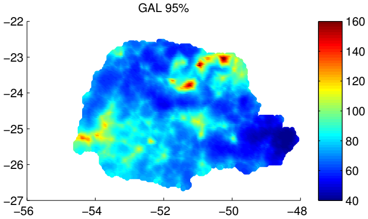
CHALMERS

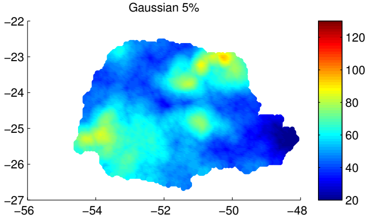
Kriging surfaces for the monthly maximum

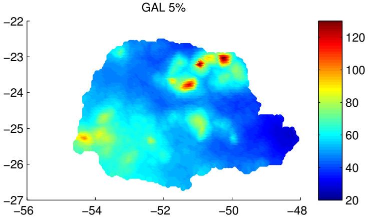


Quantiles

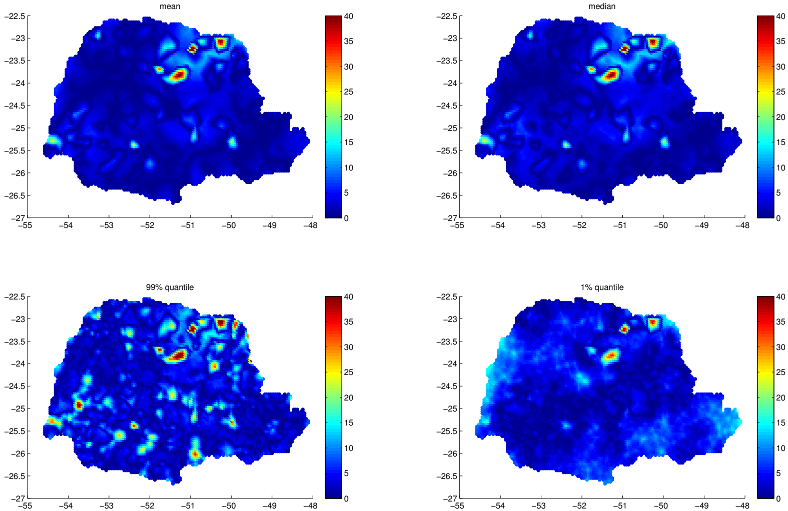

Gaussian 95%
GAL 95%
-22
-23
-24
-25
-26
-27
-56
-54
-52
-50
-48
160
140
120
100
80
60
40
Gaussian 5%
GAL 5%
120
100
80
60
40
20



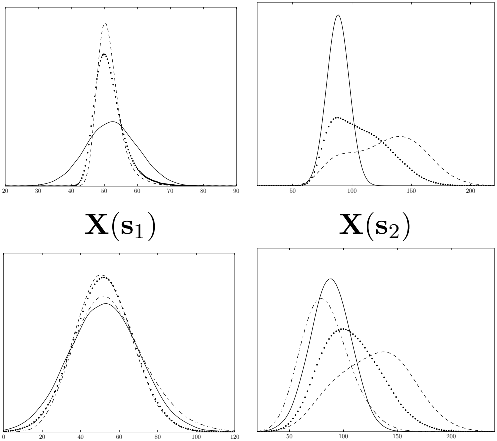




Absolute differences between GAL and Gaussian



Posterior marginals for monthly maximum

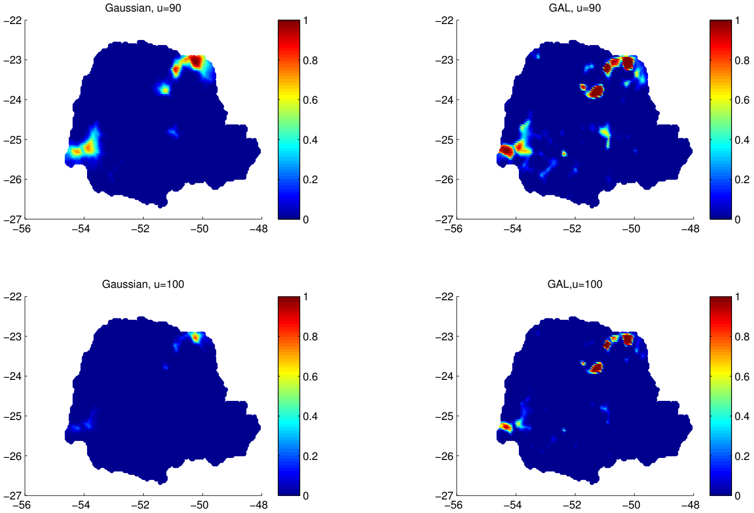


$\mathbf{X}(\mathbf{s}_1)$ $\mathbf{X}(\mathbf{s}_2)$

$\mathbf{X}(\mathbf{s}_1) + \boldsymbol{\epsilon}(\mathbf{s}_1)$ $\mathbf{X}(\mathbf{s}_2) + \boldsymbol{\epsilon}(\mathbf{s}_2)$

Gaussian model (solid), NIG model (dotted), GAL model (dashed), and transformed Gaussian model (dash-dotted).

Marginal excursion probabilities, $\mathsf{P}(X(s) > u|Y)$



Conclusions

- One should be aware of the assumptions that are implied by using transformed Gaussian models.
- Non-Gaussian Matérn fields are interesting alternatives to transformed Gaussian models for non-Gaussian data.
- By using the SPDE representation we can obtain a computationally feasible representation.
- We can handle latent non-Gaussian models with measurement noise and partial observations for large datasets.
- Estimation for similar models has previously been done using the method of moments, we have constructed a likelihood-based estimation procedure.
- Currently, we have Matlab and C code for performing the analysis. We are planning on making an R package.

References

- Bolin, Spatial Matérn fields driven by non-Gaussian noise, *Scandinavian Journal of Statistics*, 2013
- Wallin and Bolin, Non-Gaussian Matérn fields with an application to precipitation modeling, *ArXiv preprint*, 2013.
- Lindgren, Rue, and Lindström, An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach, *Journal of the Royal Statistical Society: Series B*, 2011
- Podgórski and Wallin, Convolution invariant generalized hyperbolic sub-classes. *Communications in Statistics - Theory and Methods*, in press, 2014
- Åberg and Podgórski, A class of non-Gaussian second order random fields. *Extremes* 14, 187-222, 2011