Correlated Nugget Effects in Multivariate Spatial Models: Enhancing Ocean Data Predictions

Damilya Saduakhas[1], Alexandre B. Simas[1], Jonas Wallin[2], David Bolin[1]

1 - Statistics Program, CEMSE, King Abdullah University of Science and Technology,
2 - Department of Statistics, Lund University

Summary

- **Observation level.** Each Argo float measures temperature (T) and salinity (S) with the same sensor $\Rightarrow$ their measurement errors (the nugget) are typically correlated.
- **Model level.** Classical bivariate spatial models force the nugget matrix to be diagonal. As a result, latent T–S dependence is overestimated.
- **Our extension.** Add a nugget-correlation parameter ρ_ε and fit with Gaussian or Normal-Inverse-Gaussian (NIG) driving noise. Predictions combine the Matérn-SPDE with a moving-window scheme.
- **Tools.** Implemented in the open-source `ngme2` R package; moving-window SPDE estimation scales to ∼100k profiles.

Random fields with Matérn covariance

A class of commonly used isotropic covariance functions for geostatistical applications is the stationary Matérn covariance family:

$$c(\boldsymbol{s}, \boldsymbol{s}') = \frac{\sigma^2}{\Gamma(\nu)2^{\nu-1}} (\kappa \|\boldsymbol{s} - \boldsymbol{s}'\|)^\nu K_\nu (\kappa \|\boldsymbol{s} - \boldsymbol{s}'\|)$$

where $K_\nu(\cdot)$ is the modified Bessel function of the second kind, $\nu > 0$ is the shape parameter, $\kappa > 0$ is the spatial scale parameter, and σ^2 is the variance of the covariance function.

SPDE Approach

A Gaussian process $X(\boldsymbol{s})$ with Matérn covariance function solves the stochastic partial differential equation (SPDE) (Whittle, 1963)

$$\left(\kappa^2 - \Delta\right)^{\frac{\alpha}{2}} X(\boldsymbol{s}) = \dot{\mathcal{W}}, \quad \text{where } \boldsymbol{s} \in \mathcal{D} := \mathbb{R}^d$$

where $\alpha = \nu + d/2$ and Δ is the Laplacian, and $\dot{\mathcal{W}}$ is Gaussian white noise on a general domain $\mathcal{D}$.

- Using the connection between SPDE and Gaussian processes with Matérn covariance functions, described in (Lindgren et al., 2011), we can use computationally efficient approximation of $X(\boldsymbol{s})$ on bounded domain $\mathcal{D} \subset \mathbb{R}^d$.
- We will consider the following extension introduced in (Bolin, 2014):

$$\left(\kappa^2 - \Delta\right)^{\frac{\alpha}{2}} X(\boldsymbol{s}) = \dot{\mathcal{M}}, \quad \text{where } \boldsymbol{s} \in \mathcal{D} := \mathbb{R}^d$$

  where $\dot{\mathcal{M}}$ is non-Gaussian white-noise, specifically, we assume $\mathcal{M}$ to be a type-G Lévy process.
- A Lévy process is of type-G if its increments can be represented as location-scale mixtures:

$$\gamma + v\mu + \sqrt{v}z,$$

  where $\boldsymbol{\gamma} \in \mathbb{R}^p$ and $\boldsymbol{\mu} \in \mathbb{R}^p$ are parameters, $z \sim \mathcal{N}(0, 1)$, and v is a non-negative random variable. For the model in use, we assume $\boldsymbol{\mu} = -\boldsymbol{\gamma}$.

Bivariate Matérn-SPDE formulation

The general parametrization, introduced in (Bolin and Wallin, 2020) allows us to separate control of variances, cross-correlations, and higher moments. For bivariate model, the parametrization will be as follows:

$$\mathbf{D}(\theta, \rho) \begin{bmatrix} c_1 \left(\kappa_1^2 - \Delta\right)^{\alpha_1/2} & \\ & c_2 \left(\kappa_2^2 - \Delta\right)^{\alpha_2/2} \end{bmatrix} \begin{bmatrix} X_1(\mathbf{s}) \\ X_2(\mathbf{s}) \end{bmatrix} = \begin{bmatrix} \dot{\mathcal{M}}_1 \\ \dot{\mathcal{M}}_2 \end{bmatrix}$$

- where $\kappa_i > 0$ and $\alpha_i > d/2$, $c_i = \sqrt{\sigma_i^{-2}(4\pi)^{-d/2}\kappa_i^{-2\nu_i}\Gamma(\nu_i)/\Gamma(\alpha_i)}$ for $i = 1, 2$.

$$\mathbf{D}(\theta, \rho) = \begin{pmatrix} \cos(\theta) + \rho\sin(\theta) & -\sin(\theta)\sqrt{1+\rho^2} \\ \sin(\theta) - \rho\cos(\theta) & \cos(\theta)\sqrt{1+\rho^2} \end{pmatrix}$$

- ρ controls the correlation between X_1 and X_2 and θ the higher moments for non-Gaussian models
- Here $\dot{\boldsymbol{\mathcal{M}}}$ is L_2-valued independently scattered random measure, whose components are mutually uncorrelated. It includes non-Gaussian processes, as well as Gaussian noise.
- For the model in use, the Normal-Inverse Gaussian (NIG) driving noise will be used and α is fixed to 2.

Simulation Study

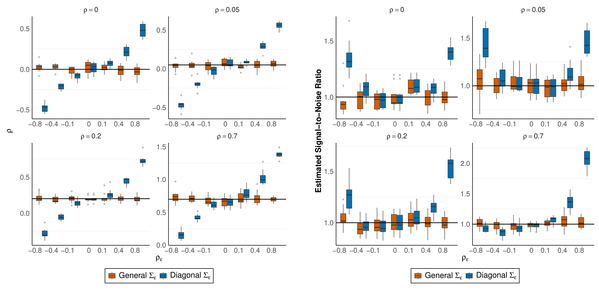

Figure: Simulation study results. (a) We compare the two Gaussian bivariate SPDE models based on the measurement-noise covariance $\boldsymbol{\Sigma}_\varepsilon$. (b) Estimated signal-to-noise ratio, $\sigma_1^2/\sigma_{\varepsilon,1}^2$, from the simulation results. The horizontal line shows the true value.

The mean value function $\mathbf{m}(\mathbf{s})$ is specified as (Roemmich and Gilson, 2009)

$$m_i(\mathbf{s}) = \beta_{i,0} + \beta_{i,x}x_c + \beta_{i,y}y_c + \beta_{i,xy}x_c y_c + \beta_{i,x^2}x_c^2 + \beta_{i,y^2}y_c^2 + \sum_{k=1}^{K}\left[\beta_{i,c_k}\cos\left(\frac{2\pi kt}{365}\right) + \beta_{i,s_k}\sin\left(\frac{2\pi kt}{365}\right)\right], \quad (1)$$

$\mathbf{s} = (x_i, y_i)$ (with x and y corresponding to longitude and latitude, respectively), $x_c := x - x^*$ and $y_c := y - y^*$ are spatial coordinates centered around x^* and y^*, and K is a predefined maximum number of harmonics $K = 6$.

Application: Argo float data

- Our analysis focused on Argo data collected annually from 2007 to 2020.
- First, we preprocessed the data and selected 1,349,863 profiles. We then computed the monthly mean using the yearly data and equation (1). Subtracting the monthly mean from the raw data, we then obtained monthly residuals.
- Our final model fits the data from January over 14 years and treats them as independent replicates.

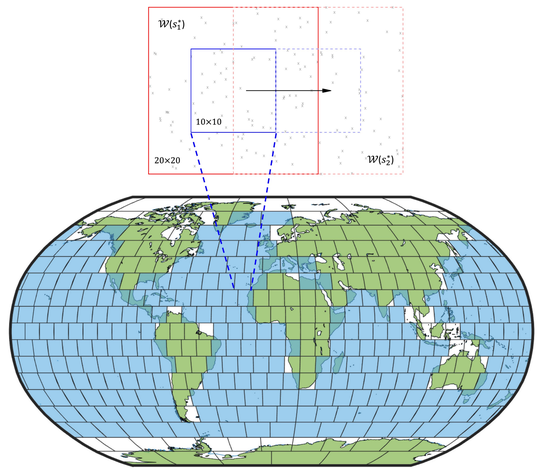

Figure: Each 10°×10° cell (solid blue) is estimated with data inside its 20°×20° window (dashed). Overlap ensures smooth parameter fields. Grid used for model fitting. Each grid box has an approximately equal surface area. For the pressure level 300 dbar, there are 410 grid boxes; 106 are omitted based on insufficient data. The remaining boxes are shown in blue.

Assume we have observations $\boldsymbol{Y}_i = [Y_1(s_i), Y_2(s_i)]^T$ observed at locations $s_1, \ldots s_n$, where Y_i satisfies

$$\boldsymbol{Y}_i = \boldsymbol{X}(\mathbf{s}_i) + \boldsymbol{\epsilon}_i \quad \epsilon_i \sim \mathrm{N}(0, \Sigma_i) \text{ for } i = 1 \ldots n$$

where Σ_i is now with correlation ρ_ϵ:

$$\Sigma_i = \begin{bmatrix} \sigma_{\epsilon,1}^2 & \rho_\epsilon\sigma_{\epsilon,1}\sigma_{\epsilon,2} \\ \rho_\epsilon\sigma_{\epsilon,1}\sigma_{\epsilon,2} & \sigma_{\epsilon,2}^2 \end{bmatrix}$$

Here, $\boldsymbol{X}(\mathbf{s}) = [\boldsymbol{X_1}(\mathbf{s}), \boldsymbol{X_2}(\mathbf{s})]^T$ follows the multivariate Type G model specified earlier.

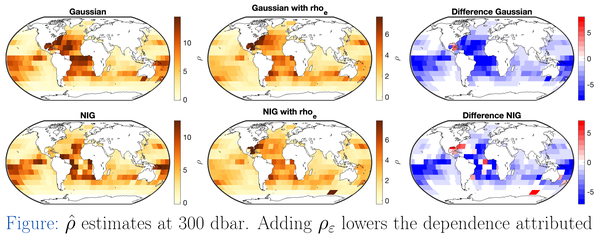

Figure: $\hat{\rho}$ estimates at 300 dbar. Adding ρ_ε lowers the dependence attributed to ρ compared to the independent-nugget model.

Model Evaluation and Results

- Leave-one-out cross-validation inside each 10°×10° window was performed for each grid.
- Predictive distribution approximated with 500 Gibbs samples (Gaussian) and 1000 samples (NIG) per left-out observation.
- Global scores were computed using the weighted mean over 111 000 profiles at 300 dbar (see table).

Model		Temperature				Salinity			
		MAE	MSE	CRPS	SCRPS	MAE	MSE	CRPS	SCRPS
Gaussian	correlated	**0.340**	**0.315**	**0.259**	**0.607**	**0.0410**	**0.0048**	0.0324	−0.454
	independent	0.361	0.354	0.275	0.641	0.0436	0.0054	0.0336	−0.422
NIG	correlated	0.356	0.368	0.268	0.609	0.0428	0.0055	**0.0324**	**−0.462**
	independent	0.380	0.484	0.285	0.645	0.0456	0.0073	0.0345	−0.426

Table: Cross-validation results of the moving-windows model on global Argo data for the pressure level 300 dbar. The lower values indicate a better fit.

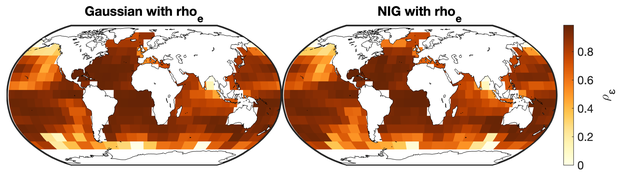

Figure: Measurement noise correlation, ρ_ε for proposed models at 300 dbar pressure level. The models with the additional parameter reveal the hidden dependence between the fields, indicating a possible underestimation of dependence in usual models.

Conclusion

- Allowing $\rho_\varepsilon \neq 0$ sharpens fine-scale structure, and improves the uncertainty quantification for global temperature and salinity predictions.
- After accounting for correlated nugget, Gaussian and NIG fits perform similarly; NIG adds flexibility in skew zones.
- **Open-source:** `github.com/d-saduakhas/Argo-SPDE`

References

Bolin, D. (2014). Spatial matérn fields driven by non-gaussian noise. *Scandinavian Journal of Statistics*, 41:557–579.

Bolin, D. and Wallin, J. (2020). Multivariate type g matérn stochastic partial differential equation random fields. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 82.

Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. *J. R. Statist. Soc. B*, 73:423–498.

Roemmich, D. and Gilson, J. (2009). The 2004–2008 mean and annual cycle of temperature, salinity, and steric height in the global ocean from the Argo Program. *Progress in Oceanography*, 82(2):81–100.

Whittle, P. (1963). Stochastic-processes in several dimensions. *Bulletin of the International Statistical Institute*, 40(2):974–994.