Spatial Prediction of Column-Averaged Carbon Dioxide Over the Globe

Josh Jacobson

PhD Candidate

Advisors:

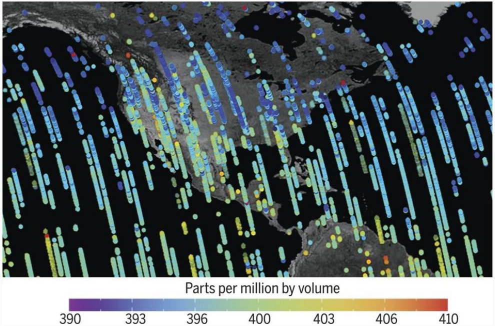
Distinguished Professor Noel Cressie

Dr. Andrew Zammit Mangion

July 6, 2021

NASA's Orbiting Carbon Observatory-2 (OCO-2) seeks to quantify the global geographic distribution of carbon dioxide (CO_2) (Eldering et al., 2017):

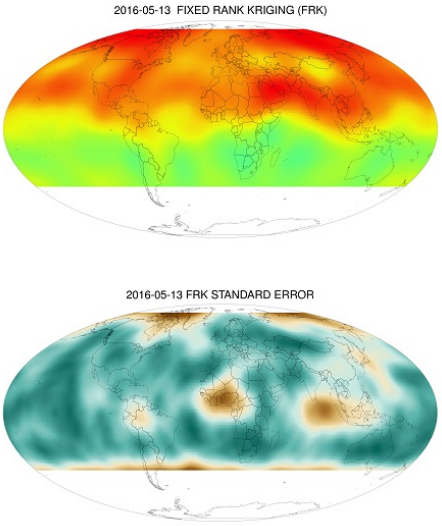
- Primary "Level 2" data product is column-averaged CO_2 dry-air mole fraction (XCO_2) at orbit locations/times.
- Spatial resolution of $\sim$3 km^2 with a 16-day repeat cycle; 2014 – present.
- De-noised and gap-filled "Level 3" data products are needed for analyses of atmospheric carbon.



Credit: Eldering et al. (2017)

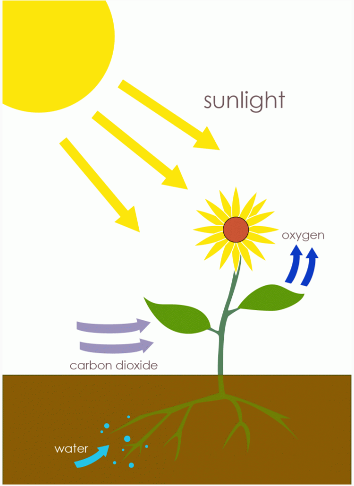
Background / Spatial statistical models provide predictions and uncertainty

- Statistical methods like kriging (i.e., optimal spatial prediction) leverage the spatial dependence in these observations to produce de-noised and gap-filled estimates along with their statistical uncertainty (e.g., Cressie, 1993).
- This spatial inference can be made more efficient/accurate if cross-correlations with other observed variables are identified.



Credit: Zammit-Mangion et al. (2018)

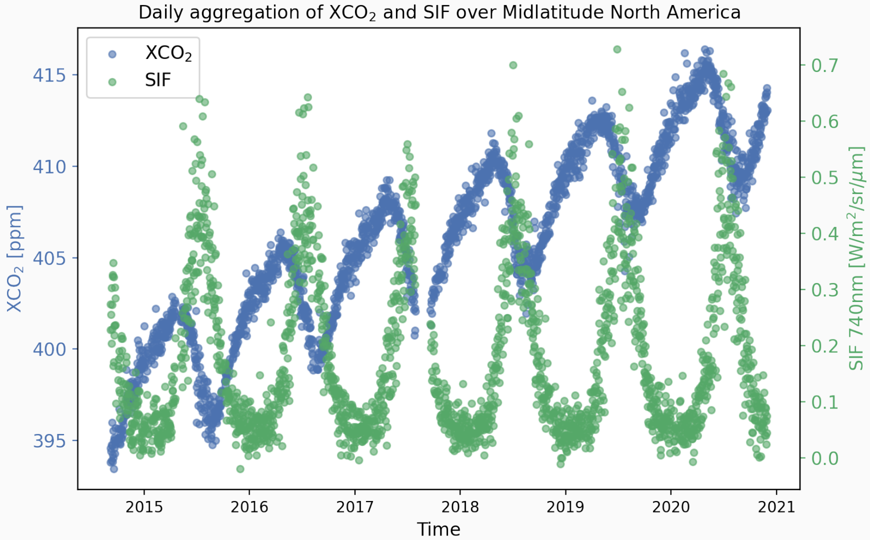
Background / A bivariate process: carbon dioxide and chlorophyll fluorescence



Credit: Wikipedia Commons, Author At09kg

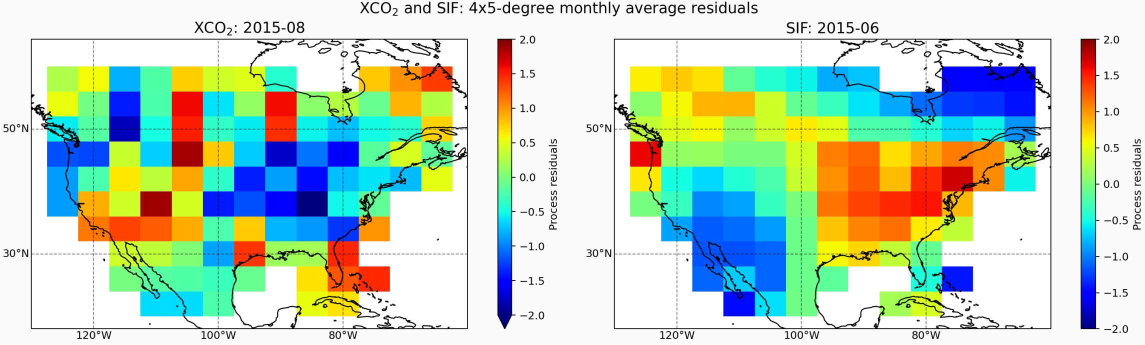
- Solar-induced chlorophyll fluorescence (SIF) is another primary data product of the OCO-2 mission.
- SIF is a small amount of light emitted during photosynthesis that can be detected in remote sensing measurements of radiance within solar Fraunhofer lines.
- As an indicator of photosynthetic activity, an inverse relationship between SIF and XCO_2 is expected and observed.
- Incorporation of SIF into a bivariate spatial model will help improve prediction of XCO_2 (and vice versa).

Inverse relationship strongest at 1-2 month temporal lag



Process residuals over midlatitude North America

Monthly data aggregated to a resolution of 4-degrees latitude by 5-degrees longitude for compatibility with recent work in CO_2 flux inversion (Liu et al., 2021).



Each dataset is pre-processed to remove large-scale variability and achieve a standard scale.

At the 4×5-degree resolution, each set of pre-processed residuals $\mathbf{Z}_i \equiv (Z_{i1}, \ldots, Z_{in_i})^\top$ with corresponding spatial locations $\{\mathbf{s}_{i1}, \ldots, \mathbf{s}_{in_i}\} \in D \subset \mathbb{S}^Z$, are modelled as realisations of a mean-zero Gaussian process $Y_i(\cdot) \equiv \{Y_i(\mathbf{s}) : \mathbf{s} \in D\}$.

In a bivariate setting, the within-process spatial dependence for $Y_i(\cdot)$ is captured by the covariance function of $Y_i(\cdot)$,

$$C_{ii}(\mathbf{s}, \mathbf{u}) \equiv \mathrm{cov}(Y_i(\mathbf{s}), Y_i(\mathbf{u})); \quad i = 1, 2; \ \mathbf{s}, \mathbf{u} \in \mathbb{R}^d,$$

and the between-process spatial dependence is captured by the cross-covariance function:

$$C_{ij}(\mathbf{s}, \mathbf{u}) \equiv \mathrm{cov}(Y_i(\mathbf{s}), Y_j(\mathbf{u})); \quad i, j = 1, 2; \ \mathbf{s}, \mathbf{u} \in \mathbb{R}^d.$$

Importantly, not all functions are valid (cross-) covariance functions.

The multivariate Matérn model (e.g., Gneiting et al., 2010) in $\mathbb{R}^d$ is popular for its flexible parameterisation of spatial smoothness. For $\mathbf{h} = \mathbf{s} - \mathbf{u}$,

$$C^{\circ}_{ij}(\mathbf{h}) = C^{\circ}_{ji}(\mathbf{h}) = \begin{cases} \sigma_i^2 M(\mathbf{h} \mid \nu_i, \ell_i); & i = j, \\ \rho_{ij}\sigma_i\sigma_j M(\mathbf{h} \mid \nu_{ij}, \ell_{ij}); & i \neq j, \end{cases}$$

where

$$M(\mathbf{h} \mid \nu, \ell) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}}{\ell} ||\mathbf{h}|| \right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu}}{\ell} ||\mathbf{h}|| \right); \quad \mathbf{h} \in \mathbb{R}^d,$$

is the Matérn correlation function, which is an isotropic and stationary characterisation of spatial dependence.

The model can be quite flexible but is limited to situations where stationary and symmetric covariance structures are realistic (such as in smaller spatial domains).

An empirical measure of spatial dependence

The (cross-) semivariogram function is $\gamma_{ij}(\mathbf{s},\mathbf{u}) \equiv \frac{1}{2}\mathrm{var}(Y_i(\mathbf{s}) - Y_j(\mathbf{u}))$. Under stationarity assumptions, the bivariate Matérn (cross-) semivariogram model is given as

$$\gamma^{\circ}_{ij}(\mathbf{h} \mid \theta_{ij}) = \begin{cases} \sigma_i^2(1 - M(\mathbf{h} \mid \nu_i, \ell_i)) + \tau_i^2; & i = j, \\ \frac{1}{2}(\sigma_i^2 + \sigma_j^2 + \tau_i^2 + \tau_j^2) - \rho_{ij}\sigma_i\sigma_j M(\mathbf{h} \mid \nu_{ij}, \ell_{ij}); & i \neq j, \end{cases}$$

where $\theta_{ii} = \{\sigma_i, \nu_i, \ell_i, \tau_i\}$ and $\theta_{ij} = \{\rho_{ij}, \nu_{ij}, \ell_{ij}\}$, for $i, j = 1, 2$.

The unbiased estimator is known as the empirical semivariogram when $i = j$, and as the empirical cross-semivariogram otherwise:

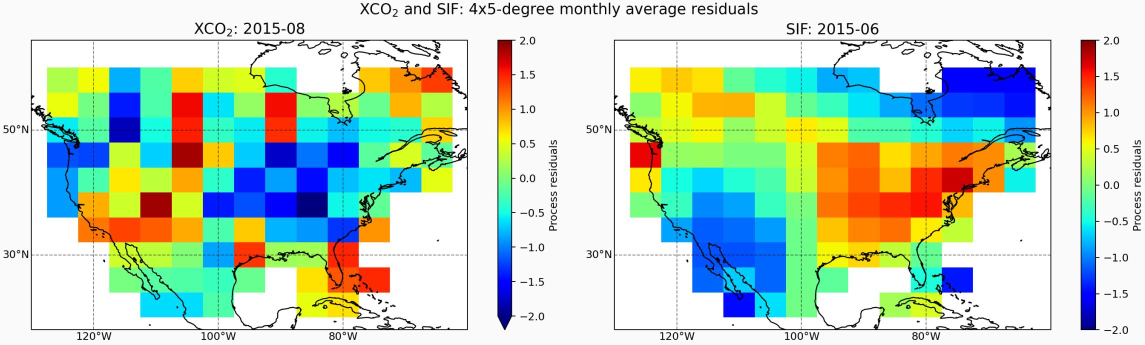
$$\widehat{\gamma}^{\circ}_{ij}(\mathbf{h}) = \frac{1}{2|N(\mathbf{h})|} \sum_{\mathbf{s}_{ik}, \mathbf{s}_{jl} \in N(\mathbf{h})} ((Z_{ik} - \hat{\mu}_i) - (Z_{jl} - \hat{\mu}_j))^2; \quad \mathbf{h} \in \mathbb{R}^d.$$

A popular approach for fitting semivariograms is via weighted least squares (WLS; Cressie, 1985). Here, the approach is extended to the bivariate case via a *composite* WLS. For a fixed set of lags $\mathbf{h}_1, \dots, \mathbf{h}_K$, model parameters $\boldsymbol{\theta} \equiv \cup\{\theta_{11}, \theta_{12}, \theta_{22}\}$ are estimated simultaneously as

$$\hat{\boldsymbol{\theta}}^{\mathrm{WLS}} \equiv \underset{\boldsymbol{\theta}\in\Theta}{\arg\min} \left\{ \sum_{i=1}^{2} \sum_{j=i}^{2} \sum_{k=1}^{K} |N(\mathbf{h}_k)| \left(\frac{\widehat{\gamma}^{\circ}_{ij}(\mathbf{h}_k) - \gamma^{\circ}_{ij}(\mathbf{h}_k \mid \theta_{ij})}{\gamma^{\circ}_{ij}(\mathbf{h}_k \mid \theta_{ij})} \right)^2 \right\}.$$

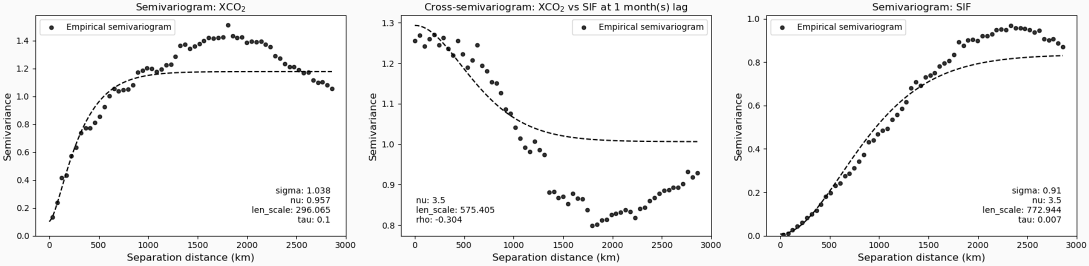
Main advantage: WLS automatically gives the most weight to lags where the spatial (cross-) dependence is strongest and down-weights those lags associated with the fewest spatial pairs.

A fitted bivariate Matérn model (semivariogram scale)

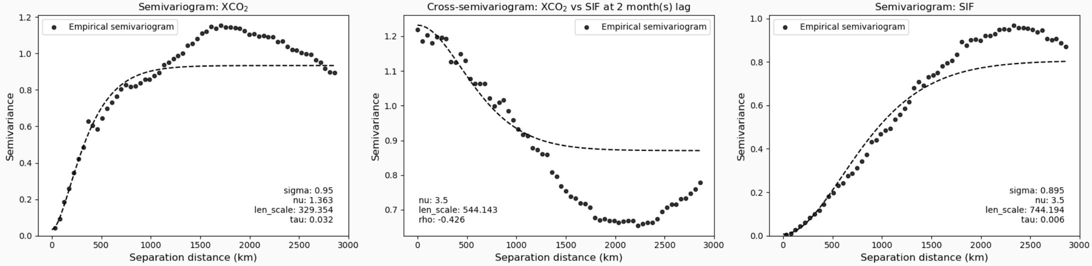


In this analysis, the domain D is midlatitude North America (Liu et al., 2021) with a spatial support of 4-degrees latitude by 5-degrees longitude.

July 2015, 1 month lag



August 2015, 2 month lag



- Bivariate spatial dependence between XCO_2 and SIF can be exploited to obtain better predictions than using either process alone.
- Updated parameter estimates will be plugged into bivariate prediction equations (cokriging), and prediction results will be analysed over midlatitude North America and other regions around the globe.
- In future work, it will be necessary to develop a multivariate model capable of handling non-stationarity for application over larger domains.

Credit: NASA on unsplash.com

Cressie, N. (1985). Fitting variogram models by weighted least squares. *Journal of the International Association for Mathematical Geology*, 17:563–586.

Cressie, N. (1993). *Statistics for Spatial Data*. Wiley, Hoboken, NJ, revised edition.

Eldering, A., Wennberg, P. O., Crisp, D., Schimel, D. S., Gunson, M. R., Chatterjee, A., Liu, J., Schwandner, F. M., Sun, Y., O'Dell, C. W., Frankenberg, C., Taylor, T., Fisher, B., Osterman, G. B., Wunch, D., Hakkarainen, J., Tamminen, J., and Weir, B. (2017). The Orbiting Carbon Observatory-2 early science investigations of regional carbon dioxide fluxes. *Science*, 358:eaam5745.

Gneiting, T., Kleiber, W., and Schlather, M. (2010). Matérn cross-covariance functions for multivariate random fields. *Journal of the American Statistical Association*, 105:1167–1177.

Liu, J., Baskaran, L., Bowman, K., Schimel, D., Bloom, A. A., Parazoo, N. C., Oda, T., Carroll, D., Menemenlis, D., Joiner, J., Commane, R., Daube, B., Gatti, L. V., McKain, K., Miller, J., Stephens, B. B., Sweeney, C., and Wofsy, S. (2021). Carbon monitoring system flux net biosphere exchange 2020 (CMS-Flux NBE 2020). *Earth System Science Data*, 13:299–330.

Zammit-Mangion, A., Cressie, N., and Shumack, C. (2018). On statistical approaches to generate Level 3 products from satellite remote sensing retrievals. *Remote Sensing*, 10:155.

For prediction location $\mathbf{s}_0 \in D$, the best predictor of $Y_1(\mathbf{s}_0)$ is the conditional mean, $\mathbb{E}(Y_1(\mathbf{s}_0)|\mathbf{Z}_1, \dots, \mathbf{Z}_p)$. Consider the joint distribution,

$$\begin{bmatrix} Y_1(\mathbf{s}_0) \\ \mathbf{Z}^* \end{bmatrix} \sim \mathrm{Gau}\left(\begin{bmatrix} \mathbf{x}_1(\mathbf{s}_0)^\top \boldsymbol{\beta}_1 \\ \mathbf{X}^* \boldsymbol{\beta}^* \end{bmatrix}, \begin{bmatrix} c(\mathbf{s}_0) & \mathbf{c}_0^{*\top} \\ \mathbf{c}_0^* & \mathbf{C}_Z \end{bmatrix} \right),$$

where $c(\mathbf{s}_0) \equiv \mathrm{var}(Y_1(\mathbf{s}_0))$ and $\mathbf{c}_0^{*\top} \equiv \mathrm{cov}(Y_1(\mathbf{s}_0), \mathbf{Z}^*) = \mathrm{cov}(Y_1(\mathbf{s}_0), \mathbf{Y}^*)$. Standard Gaussian identities give:

$$Y_1(\mathbf{s}_0) \mid \mathbf{Z}^* \sim \mathrm{Gau}\left(\mathbf{x}_1(\mathbf{s}_0)^\top \boldsymbol{\beta}_1 + \mathbf{c}_0^{*\top} \mathbf{C}_Z^{-1} (\mathbf{Z}^* - \mathbf{X}^* \boldsymbol{\beta}^*),\ c(\mathbf{s}_0) - \mathbf{c}_0^{*\top} \mathbf{C}_Z^{-1} \mathbf{c}_0^* \right).$$

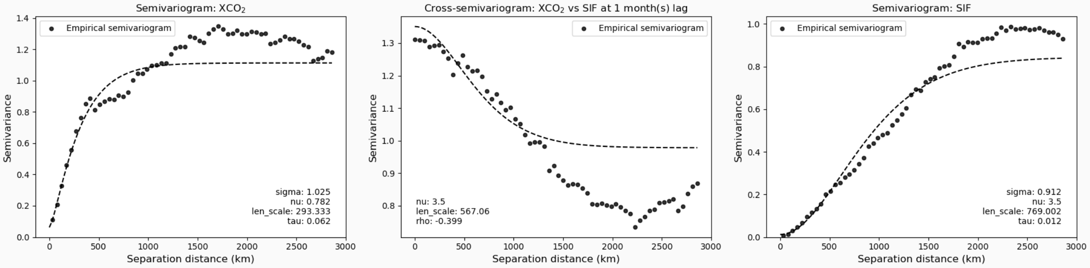
That is, the optimal spatial predictor of $Y_1(\mathbf{s}_0)$ given $\mathbf{Z}_1, \dots, \mathbf{Z}_p$ is

$$\mathbb{E}(Y_1(\mathbf{s}_0) \mid \mathbf{Z}_1, \dots, \mathbf{Z}_p) = \mathbf{x}_1(\mathbf{s}_0)^\top \boldsymbol{\beta}_1 + \mathbf{c}_0^{*\top} \mathbf{C}_Z^{-1} (\mathbf{Z}^* - \mathbf{X}^* \boldsymbol{\beta}^*).$$

The predictive variance, $c(\mathbf{s}_0) - \mathbf{c}_0^{*\top} \mathbf{C}_Z^{-1} \mathbf{c}_0^*$, is a measure of **uncertainty** in the corresponding prediction.

The roles of $Y_1(\cdot)$ and, for example, $Y_2(\cdot)$ can be reversed for optimal spatial prediction of $Y_2(\mathbf{s}_0)$.

July 2019, 1 month lag



August 2019, 2 month lag