Supplemental Material for "Mission CO_2ntrol: A statistical scientist's role in remote sensing of atmospheric carbon dioxide"

Noel Cressie
National Institute for Applied Statistics Research Australia
School of Mathematics and Applied Statistics
University of Wollongong
Wollongong NSW 2522
Australia
(ncressie@uow.edu.au)

December 13, 2017

Abstract

Supplemental Material for the article, "Mission CO_2ntrol: A statistical scientist's role in remote sensing of atmospheric carbon dioxide," by Noel Cressie, forthcoming in the *Journal of the American Statistical Association.*

Video: Fixed Rank Kriging (FRK) of daily XCO2 data from OCO-2 in 2015, where the model is spatio-temporal (defined on geoid×time, in days). Click here to view [Also available on YouTube at `https://www.youtube.com/watch?v=KXId_dBuHoU`]

Figure 1: "Blue Marble" (Credit: Astronaut Photograph, NASA Johnson Space Center, 7 December 1972)

(a) Launch of OCO-2 satellite on 2 July 2014 (Credit: Photo from NASA)

(b) OCO-2 satellite in orbit (Credit: Illustration from NASA)

Figure 2: OCO-2 launch and NASA illustration of OCO-2 satellite in orbit

Details of the Main Paper's Section 2.1: Optimal retrievals for a linear forward model

In order to understand the so-called Optimal Estimation retrieval (Rodgers, 2000; Miller et al., 2007; Connor et al., 2008; Bréon and Ciais, 2010; Bösch et al., 2011; O'Dell et al., 2012) used by OCO-2, it helps to consider initially what a retrieval would look like if the forward model were linear:

$$\mathbf{F}^L(\mathbf{X}) \equiv \mathbf{c} + \mathbf{K}\mathbf{X}\,,$$

where $\mathbf{c}$ is n_ε-dimensional and $\mathbf{K}$ is an $n_\varepsilon \times n_\alpha$ matrix. Clearly, $\mathbf{F}^L(\mathbf{X})$ could be thought of as an approximation to a nonlinear vector-valued function $\mathbf{F}(\mathbf{X})$, subjected to a Taylor-series expansion about a known vector $\mathbf{X}^0$:

$$\begin{aligned}\mathbf{F}(\mathbf{X}) &= \mathbf{F}(\mathbf{X}^0) + \left.\frac{\partial \mathbf{F}(\mathbf{x})}{\partial \mathbf{x}}\right|_{\mathbf{x}=\mathbf{X}^0} \times (\mathbf{X} - \mathbf{X}^0) + \boldsymbol{\lambda} \\ &\equiv \mathbf{c} + \mathbf{K}\mathbf{X} + \boldsymbol{\lambda}\,,\end{aligned}$$

where $\boldsymbol{\lambda}$ models the lack of fit of $\mathbf{F}^L$ to (the non-linear) $\mathbf{F}$, and the choice of linearization point $\mathbf{X}^0$ is important. Thus, if $\mathbf{F}$ is non-linear and the linear forward model, $\mathbf{F}^L(\mathbf{X}) = \mathbf{c} + \mathbf{K}\mathbf{X}$, were used, then the forward-model error $\boldsymbol{\varepsilon}$ would contain yet another component of variability.

In this subsection, I assume a linear forward model,

$$\mathbf{Y} = \mathbf{c} + \mathbf{K}\mathbf{X} + \boldsymbol{\varepsilon}\,,$$

and I describe Optimal Estimation (Rodgers, 2000) in terms of it. Section 2.2 of the Main Paper considers a more realistic non-linear forward model. Instead of working with the marginal distribution of $\mathbf{Y}$ in the forward model, a strategically clever alternative is to describe the probability structure (from which the uncertainty is quantified) through conditional distributions. Let $[\mathbf{Y}|\mathbf{X}]$ denote the density of $\mathbf{Y}$ given $\mathbf{X}$. Then

$$-2\ln[\mathbf{Y}|\mathbf{X}] = (\mathbf{Y} - \mathbf{c} - \mathbf{K}\mathbf{X})'\mathbf{S}_\varepsilon^{-1}(\mathbf{Y} - \mathbf{c} - \mathbf{K}\mathbf{X}) + c_1\,.$$

This distribution corresponds to the so-called *measurement equation* in a state-space model or, equivalently, corresponds to the *data model* in a hierarchical statistical model (and it is called a forward model in Optimal Estimation).

The so-called *state equation* (equivalently a *process model* in a hierarchical statistical model, or a *prior distribution* in Optimal Estimation) is,

$$\mathbf{X} = \mathbf{X}_\alpha + \boldsymbol{\alpha}\,,$$

where $\boldsymbol{\alpha} \sim \text{Gau}(\mathbf{0}, \mathbf{S}_\alpha)$ independent of $\boldsymbol{\varepsilon}$, and $\mathbf{X}$, $\mathbf{X}_\alpha$, and $\boldsymbol{\alpha}$ are n_α-dimensional vectors. Notice that biases, or systematic errors, can be incorporated into this model, along with random errors whose variances characteristically decrease with averaging. Hence, the density $[\mathbf{X}]$ of the atmospheric state is given by:

$$-2\ln[\mathbf{X}] = (\mathbf{X} - \mathbf{X}_\alpha)'\mathbf{S}_\alpha^{-1}(\mathbf{X} - \mathbf{X}_\alpha) + c_2\,.$$

The implication of this is that the "true" state $\mathbf{X}$ is random, although some geophysicists find this a leap too far. While the true state might be independent of the scientist modeling it, the lack of complete knowledge of the state is inside the head of the modeler (perhaps representing a consensus model after much discussion with fellow scientists). Thus, the source of randomness here is the scientist's uncertainty about the true state.

Now assume that all parameters necessary for retrieval (e.g., $\mathbf{K}$, $\mathbf{S}_\varepsilon$, $\mathbf{X}_\alpha$, $\mathbf{S}_\alpha$) are known. Inference on the n_α-dimensional state vector $\mathbf{X}$ is based on its *predictive distribution* (called a *posterior distribution* in Optimal Estimation) of $\mathbf{X}$ given $\mathbf{Y}$, namely $[\mathbf{X}|\mathbf{Y}]$. From Bayes' Rule:

$$\begin{aligned} -2\ln[\mathbf{X}|\mathbf{Y}] &= (\mathbf{Y} - \mathbf{c} - \mathbf{K}\mathbf{X})'\mathbf{S}_\varepsilon^{-1}(\mathbf{Y} - \mathbf{c} - \mathbf{K}\mathbf{X}) + (\mathbf{X} - \mathbf{X}_\alpha)'\mathbf{S}_\alpha^{-1}(\mathbf{X} - \mathbf{X}_\alpha) + c_3 \\ &= (\mathbf{X} - \hat{\mathbf{X}})'\hat{\mathbf{S}}^{-1}(\mathbf{X} - \hat{\mathbf{X}}) + c_4\,, \end{aligned}$$

where the posterior mean is:

$$E(\mathbf{X}|\mathbf{Y}) \equiv \hat{\mathbf{X}} = \mathbf{X}_\alpha + \mathbf{G}(\mathbf{Y} - \mathbf{c} - \mathbf{K}\mathbf{X}_\alpha)\,;$$

the posterior covariance matrix is:

$$\text{cov}(\mathbf{X}|\mathbf{Y}) \equiv \hat{\mathbf{S}} = \{\mathbf{S}_\alpha^{-1} + \mathbf{K}'\mathbf{S}_\varepsilon^{-1}\mathbf{K}\}^{-1}\,;$$

the so-called "gain matrix" $\mathbf{G}$ is:

$$\mathbf{G} = \{\mathbf{S}_\alpha^{-1} + \mathbf{K}'\mathbf{S}_\varepsilon^{-1}\mathbf{K}\}^{-1}\mathbf{K}'\mathbf{S}_\varepsilon^{-1}\,;$$

and the so-called "averaging kernel matrix" is:

$$\mathbf{A} = \mathbf{G}\mathbf{K} = \{\mathbf{S}_\alpha^{-1} + \mathbf{K}'\mathbf{S}_\varepsilon^{-1}\mathbf{K}\}^{-1}\mathbf{K}'\mathbf{S}_\varepsilon^{-1}\mathbf{K}\,.$$

Hence, the retrieved state is:

$$\hat{\mathbf{X}} = \mathbf{X}_\alpha + \{\mathbf{S}_\alpha^{-1} + \mathbf{K}'\mathbf{S}_\varepsilon^{-1}\mathbf{K}\}^{-1}\mathbf{K}'\mathbf{S}_\varepsilon^{-1}(\mathbf{Y} - \mathbf{c} - \mathbf{K}\mathbf{X}_\alpha)\,,$$

and no iterative scheme is needed to obtain $\hat{\mathbf{X}}$. Provided the forward model $\mathbf{F}$ is linear, the predictive distribution, $[\mathbf{X}|\mathbf{Y}]$, is Gaussian; specifically,

$$\mathbf{X}|\mathbf{Y} \sim \text{Gau}(\hat{\mathbf{X}}, \hat{\mathbf{S}})\,.$$

Finally, it is straightforward to see that the prediction error is:

$$\hat{\mathbf{X}} - \mathbf{X} = (\mathbf{A} - \mathbf{I})\boldsymbol{\alpha} + \mathbf{G}\boldsymbol{\varepsilon}\,,$$

where recall that $\boldsymbol{\varepsilon}$ is the error in the measurement equation, which is independent of $\boldsymbol{\alpha}$, the error in the state equation. This relationship becomes a first-order approximation when the forward model is non-linear, as do many other relationships given in this subsection; see Section 2.2 of the Main Paper.

Details of the Main Paper's Section 3.1: Optimal spatial prediction (kriging)

In what follows, I consider the generic spatial prediction problem of predicting the underlying univariate spatial process $X(\cdot)$ defined by the vector,

$$\mathbf{X}_p \equiv (X(\mathbf{u}_1), \ldots, X(\mathbf{u}_N))'\,, \tag{1}$$

whose elements are indexed by the centroids $\{\mathbf{u}_1, \ldots, \mathbf{u}_N\}$ of the N BAUs that tessellate D. The application in Subsection 3.3 of the Main Paper is to the underlying random field $X(\cdot) = XCO2(\cdot)$, the true column-averaged CO_2 values. Kriging is a statistical method of spatial prediction originally proposed by Matheron (1963). It is linear, unbiased, and amongst all such predictors minimizes the mean squared prediction error; see (9) and (10) below. With a slight abuse of notation, after tessellation, I re-define the spatial domain of interest to be

$$D \equiv \{\mathbf{u}_1, \ldots, \mathbf{u}_N\}, \tag{2}$$

recognizing that $\mathbf{u}_j$ represents the j-th BAU, for $j = 1, \ldots, N$. The prediction of $\mathbf{X}_p$ is obtained from the L2 retrievals $\mathbf{Y}_d$, given by (15) in the Main Paper, and I shall now review briefly how to do this optimally using kriging (i.e., so that the uncertainty of the predictor is minimized). Visualization in the form of a map of the L3 data and a map of their uncertainties is a powerful way to generate hypotheses about the behavior of $\mathbf{X}_p$.

Data are incomplete and noisy; now model the data $\mathbf{Y}_d = (Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n))'$ with a *data model*,

$$Y(\mathbf{s}_i) = X(\mathbf{s}_i) + \xi(\mathbf{s}_i); \quad \mathbf{s}_i \in D^O, \tag{3}$$

where $\xi(\cdot)$ is defined on $D^O = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ and represents mean-zero independent measurement errors that are also independent of $X(\cdot)$. Then $\text{var}(Y(\mathbf{s}_i)|X(\mathbf{s}_i)) = \text{var}(\xi(\mathbf{s}_i)) \equiv \sigma^2_\xi$,

which is the measurement-error variance. The BAUs are defined at a very fine resolution, so in practice locations in D^O are moved slightly in order that $D^O \subset D$ (Cressie and Kornak, 2003). Often $n = |D^O| << |D| = N$. In what follows, it is useful to write

$$\mathbf{X}_{d:p} \equiv (X(\mathbf{s}_1), \ldots, X(\mathbf{s}_n))', \tag{4}$$

which is the hidden process $X(\cdot)$ restricted to the data locations, D^O.

Now model the true process $X(\cdot)$ with a *process model*,

$$X(\mathbf{s}) = \mu(\mathbf{s}) + \delta(\mathbf{s}); \quad \mathbf{s} \in D, \tag{5}$$

where $\mu(\cdot)$ represents large-scale spatial variation (deterministic trend) that is often assumed to be a regression, $\mu(\cdot) \equiv \mathbf{f}(\cdot)'\boldsymbol{\beta}$, for covariates $\mathbf{f}(\mathbf{s}) \equiv (f_1(\mathbf{s}), \ldots, f_p(\mathbf{s}))'$ known at all $\mathbf{s} \in D$; and $\delta(\cdot)$ represents small-scale spatial variation modeled as a mean-zero stochastic process with second moment,

$$\text{cov}(\delta(\mathbf{s}), \delta(\mathbf{u})) \equiv C(\mathbf{s}, \mathbf{u}; \boldsymbol{\phi}). \tag{6}$$

When the process $X(\cdot)$ is Gaussian, it is usually called a Gaussian Field (GF); see, for example, Banerjee et al. (2015). While the ensuing development does not require an assumption of a GF, relying instead on optimality of linear predictors, henceforth I assume that both measurement error in (3) and the underlying spatial process in (5) are Gaussian. Some discussion of how to handle non-Gaussian L2 retrieval data is given in Section 5 of the Main Paper.

The optimal spatial predictor of $X(\mathbf{s})$ (for squared error loss) is $E(X(\mathbf{s})|\mathbf{Y}_d)$, for all $\mathbf{s} \in D$, which is the mean of the predictive distribution. Note that it is implicit in what follows that all parameters are fixed, so the statistical analyses I give are not fully Bayesian (Section 5 of the Main Paper). In practice, the parameters are estimated from $\mathbf{Y}_d$ and "plugged in" to the spatial predictor. This is called a "Case 1" state-of-knowledge in Section 5 of the Main Paper, and the resulting inference on $X(\cdot)$ has been called empirical BLUP (Best Linear Unbiased Prediction); in the context of small area estimation, bias-corrected inference for this problem has been considered by Prasad and Rao (1990).

The predictive mean is linear in $\mathbf{Y}_d$ if all processes are GFs, which is henceforth assumed. Consider the class of linear predictors,

$$X^*(\mathbf{s}) = \boldsymbol{\lambda}(\mathbf{s})'\mathbf{Y}_d + \kappa(\mathbf{s}); \quad \mathbf{s} \in D,$$

where $\boldsymbol{\lambda}(\mathbf{s}) \equiv (\lambda_1(\mathbf{s}), \ldots, \lambda_n(\mathbf{s}))'$ is an n-dimensional vector of real numbers and $\kappa(\mathbf{s})$ is a scalar. Under squared-error loss and assuming $E(X^*(\mathbf{s})) = E(X(\mathbf{s}))$ (unbiasedness),

the optimal (i.e., minimum mean squared prediction error) linear predictor is a (simple) *kriging predictor* (e.g., Cressie, 1993, pp. 109-110). Because I assume known parameters for this derivation, the subtle differences between simple, ordinary, and universal kriging are avoided. The kriging coefficients, $\boldsymbol{\lambda}(\mathbf{s})$ and $\kappa(\mathbf{s})$, are obtained by minimizing the mean squared prediction error (MSPE),

$$E(X^*(\mathbf{s}) - X(\mathbf{s}))^2\,, \tag{7}$$

with respect to $\lambda_1(\mathbf{s}), \ldots, \lambda_n(\mathbf{s})$, and $\kappa(\mathbf{s})$, subject to unbiasedness, namely

$$E(X^*(\mathbf{s})) = E(X(\mathbf{s})). \tag{8}$$

Assuming $\mu(\cdot) = \mathbf{f}(\cdot)'\boldsymbol{\beta}$ in (5) and using the method of Lagrange multipliers, the optimal coefficients $\tilde{\boldsymbol{\lambda}}(\mathbf{s})$ and $\tilde{\kappa}(\mathbf{s})$ are (e.g., Cressie, 1993, p. 110),

$$\tilde{\boldsymbol{\lambda}}(\mathbf{s})' = \mathbf{c}_X(\mathbf{s})'\boldsymbol{\Sigma}_{Y_d}^{-1} \text{ and } \tilde{\kappa}(\mathbf{s}) = \{\mathbf{f}(\mathbf{s})' - \mathbf{c}_X(\mathbf{s})'\boldsymbol{\Sigma}_{Y_d}^{-1}\mathbf{F}\}\boldsymbol{\beta},$$

where $\mathbf{c}_X(\mathbf{s})' \equiv \text{cov}(X(\mathbf{s}), \mathbf{X}_{d:p})$, recall that $\mathbf{X}_{d:p}$ is given by (4), $\boldsymbol{\Sigma}_{Y_d} \equiv \text{cov}(\mathbf{Y}_d)$, and $\mathbf{F} \equiv (\mathbf{f}(\mathbf{s}_1), \ldots, \mathbf{f}(\mathbf{s}_n))'$. Finally, the kriging predictor of $X(\mathbf{s})$ is:

$$\tilde{X}(\mathbf{s}) \equiv \tilde{\boldsymbol{\lambda}}(\mathbf{s})'\mathbf{Y}_d + \tilde{\kappa}(\mathbf{s}); \quad \mathbf{s} \in D\,, \tag{9}$$

and its MSPE (or kriging variance) is:

$$E(\tilde{X}(\mathbf{s}) - X(\mathbf{s}))^2 = C(\mathbf{s}, \mathbf{s}; \boldsymbol{\phi}) - \mathbf{c}_X(\mathbf{s})'\boldsymbol{\Sigma}_{Y_d}^{-1}\mathbf{c}_X(\mathbf{s})\,. \tag{10}$$

It is straightforward to show (e.g., Cressie and Wikle, 2011, p. 141) that the predictor $E(X(\mathbf{s})|\mathbf{Y}_d)$ has MSPE given by $E(\text{var}(X(\mathbf{s})|\mathbf{Y}_d))$. Under the Gaussian assumptions detailed above, $E(X(\mathbf{s})|\mathbf{Y}_d) = \tilde{X}(\mathbf{s})$, the kriging predictor, and hence

$$E(\tilde{X}(\mathbf{s}) - X(\mathbf{s}))^2 = E(\text{var}(X(\mathbf{s})|\mathbf{Y}_d)) = \text{var}(X(\mathbf{s})|\mathbf{Y}_d)\,. \tag{11}$$

Computing $\boldsymbol{\Sigma}_{Y_d}^{-1}$ is generally $O(n^3)$, which means that kriging given by (9) and (10) is not generally scalable. In Subsection 3.2 of the Main Paper, I present a spatial random effects (SRE) model for $C(\mathbf{s}, \mathbf{u}; \boldsymbol{\phi})$ defined by (6) that results in scalable kriging. Another way to reduce the computational burden of kriging is to use a moving window around the prediction location that limits the dimension of the data vector $\mathbf{Y}_d$ and hence of the matrix $\boldsymbol{\Sigma}_d$ (Haas, 1995; Hammerling et al., 2012; Tadić et al., 2017). Zammit-Mangion et al. (2017) discuss the pros and cons of both approaches and point out that local kriging's underlying

probability model is not coherent, which will lead to difficulty when making change-of-support calculations or when attempting to carry out flux inversion from the L3 maps.

Details of the Main Paper's Section 3.2: The spatial random effects (SRE) model

There are a number of ways to make spatial predictions scalable: Reduced-rank methods are reviewed by Wikle (2010), and the use of sparse precision matrices have been proposed by Lindgren et al. (2011) and Nychka et al. (2015). In what follows, I shall use one of the reduced-rank methods known as Fixed Rank Kriging (FRK), which is derived from the spatial random effects model.

Assume in (5) that

$$\delta(\mathbf{s}) = \boldsymbol{\Phi}(\mathbf{s})'\boldsymbol{\eta} + \nu(\mathbf{s}); \quad \mathbf{s} \in D, \tag{12}$$

where $\boldsymbol{\eta}$ is an r-dimensional random vector with mean $\mathbf{0}$ and $\text{var}(\boldsymbol{\eta}) = \mathbf{K}_\eta$; $\boldsymbol{\Phi}(\mathbf{s})$ is an r-dimensional vector of spatial basis functions of $\mathbf{s} \in D$, and write $\boldsymbol{\Phi}(\cdot) = (\Phi_1(\cdot), \ldots, \Phi_r(\cdot))'$; r is fixed $<< n$; and $\nu(\cdot)$ is a stochastic process defined on D that represents fine-scale variation. One could think of $\nu(\cdot)$ as having (equivalent) range less than the smallest distance between any pair of data locations or with support below the spatial resolution of the data. Model (12) has been called a spatial random effects (SRE) model (Cressie and Johannesson, 2006, 2008), where the k-th spatial random effect is η_k with contribution $\eta_k\Phi_k(\mathbf{s})$ to the error (12); $k = 1, \ldots, r$.

Recall that $\mu(\cdot)$ in (5) is given by

$$\mu(\cdot) = \mathbf{f}(\cdot)'\boldsymbol{\beta} = \sum_{j=1}^{p} \beta_j f_j(\cdot)\,, \tag{13}$$

which is a linear combination of p spatial covariates that defines a deterministic spatial trend. Because this term represents deterministic large-scale variation, the covariates' role in describing spatial variation is different from the role of the spatial basis functions. It is important to choose the fixed effects, $\{f_j(\cdot) : j = 1, \ldots, p\}$, and the random effects, $\{\Phi_k(\cdot) : k = 1, \ldots, r\}$, wisely. Intuitively, dependence of $\mathbf{Y}_d$ on a basis function $\Phi_k(\cdot)$ could be positive or negative for data collected under similar circumstances. The same intuition tells us that dependence of $\mathbf{Y}_d$ on a covariate $f_j(\cdot)$ should always be the same sign for data collected under similar circumstances. Hughes and Haran (2013) give ways to avoid confounding fixed effects (covariates) and random effects (basis functions).

When there is no physical reason for using a specific collection of basis functions, multi-resolutional classes could be used, such as Fourier functions, wavelets, or bisquares. They

do not have to be orthogonal, but generally they are multiresolutional with centers that are space-filling as resolutions get finer and finer, and their apertures are the same for a given resolution (Cressie and Johannesson, 2008). For example, the bisquare spatial basis function $\phi(\cdot)$ centered at $\mathbf{c}$ with aperture w is defined by,

$$\phi(\mathbf{s}; \mathbf{c}, w) \equiv (1 - \|\mathbf{s} - \mathbf{c}\|^2/w^2)^2 I(\|\mathbf{s} - \mathbf{c}\| \leq w); \quad \mathbf{s} \in \mathbb{R}^d,$$

where $I(\cdot)$ denotes the indicator function. Figure 3(a) of the Supplemental Material shows a bisquare basis function in $\mathbb{R}^2$, and Figure 3(b) of the Supplemental Material is an example of basis-function centers at three resolutions, distributed regularly across the plane. Then $\Phi_k(\mathbf{s}) = \phi(\mathbf{s}; \mathbf{c}_k, w_k)$ for some choice of center $\mathbf{c}_k$ and aperture $w_k; k = 1, \ldots, r$.

The SRE model given by (12) has covariance function,

$$\text{cov}(\delta(\mathbf{s}), \delta(\mathbf{u})) = \boldsymbol{\Phi}(\mathbf{s})' \mathbf{K}_\eta \boldsymbol{\Phi}(\mathbf{u}) + C_\nu(\mathbf{s}, \mathbf{u}); \quad \mathbf{s}, \mathbf{u} \in \mathbb{R}^d, \tag{14}$$

where $\mathbf{K}_\eta \equiv \text{cov}(\boldsymbol{\eta})$ is an $r \times r$ positive-definite matrix with r fixed $<< n$; and C_ν is a covariance function that is often represented as white noise:

$$C_\nu(\mathbf{s}, \mathbf{u}) = \sigma^2_\nu I(\mathbf{s} = \mathbf{u}); \quad \mathbf{s}, \mathbf{u} \in \mathbb{R}^d. \tag{15}$$

That is, the spatial covariance function, $C(\mathbf{s}, \mathbf{u}; \boldsymbol{\phi})$ defined in (6), is given by (14) and (15) for the SRE model, where $\boldsymbol{\phi} \equiv \{\mathbf{K}, \sigma^2_\nu\}$ denotes the SRE model's parameters. It should be noted that (14) is spatially non-stationary, and it is a valid covariance function for $\mathbf{s}$ and $\mathbf{u}$ defined on the surface of the sphere (or any other manifold). Other covariance models on the sphere embed it in $\mathbb{R}^3$ and assume stationarity, isotropy, and a chordal distance; the use of the SRE model avoids this non-physical feint to ensure that all covariance functions are non-negative definite.

The covariance matrix of the data, $\mathbf{Y}_d$, is of particular interest because its inverse, $\boldsymbol{\Sigma}^{-1}_{Y_d}$, appears prominently in the kriging predictor (9) and its MSPE (10). Now under the SRE model,

$$\boldsymbol{\Sigma}_{Y_d} = \text{var}(\mathbf{Y}_d) = \boldsymbol{\Phi}' \mathbf{K}_\eta \boldsymbol{\Phi} + \sigma^2_\nu \mathbf{E} + \sigma^2_\xi \mathbf{V}, \tag{16}$$

where $\boldsymbol{\Phi} \equiv (\boldsymbol{\Phi}(\mathbf{s}_1), \ldots, \boldsymbol{\Phi}(\mathbf{s}_n))'$ is an $n \times r$ matrix of basis functions evaluated at data locations D^O; recall that $\mathbf{K}_\eta = \text{var}(\boldsymbol{\eta})$ and σ^2_ν is given in (15); σ^2_ξ is the measurement-error-variance parameter of the measurement-error term $\xi(\cdot)$ in (3); and $\mathbf{E}$ and $\mathbf{V}$ are known diagonal matrices.

Inversion of $\boldsymbol{\Sigma}_{Y_d}$ is generally $O(n^3)$ in computational complexity. However, for the SRE model (12), it is $O(nr^2) = O(n)$, for r fixed (Cressie and Johannesson, 2008). The

inverse of the data covariance matrix using the Sherman-Morrison-Woodbury formula (e.g., Henderson and Searle, 1981), is

$$\boldsymbol{\Sigma}_{Y_d}^{-1} = \mathbf{U}^{-1} - \mathbf{U}^{-1}\boldsymbol{\Phi}'(\mathbf{K}_\eta^{-1} + \boldsymbol{\Phi}\mathbf{U}^{-1}\boldsymbol{\Phi}')^{-1}\boldsymbol{\Phi}\mathbf{U}, \tag{17}$$

where $\mathbf{U} \equiv \sigma_\nu^2\mathbf{E} + \sigma_\xi^2\mathbf{V}$ is a diagonal $n \times n$ matrix. Notice that the only non-trivial inverse in (17) is that of $(\mathbf{K}_\eta^{-1} + \boldsymbol{\Phi}\mathbf{U}^{-1}\boldsymbol{\Phi}')$, which is of fixed dimension $r \times r$.

Several comments should be made about the SRE model: It looks like a truncated Karhunen-Loéve expansion (e.g., Papoulis, 1965), except that for the SRE model the basis functions do not have to be orthogonal and $\mathbf{K}_\eta$ does not have to be diagonal. It defines a spatial process that is not stationary and hence not isotropic, and its covariance parameters are given by the $r \times r$ positive-definite matrix $\mathbf{K}_\eta$ and the fine-scale variance $\sigma_\nu^2 \geq 0$; note that it is possible to parameterize $\mathbf{K}_\eta$, although when datasets are large, estimation of $\mathbf{K}_\eta$ and σ_ν^2 can proceed directly via the method of moments (Cressie and Johannesson, 2008), the EM algorithm (Katzfuss and Cressie, 2009), or maximum likelihood estimation (Tzeng and Huang, 2017). From a rich class of multi-resolution spatial basis functions (e.g., wavelets), one can approximate stationary covariance models (e.g., Matérn) with an SRE model as described in Kang and Cressie (2011). The SRE model also handles spatial change-of-support seamlessly: On spatial support $B \subset D$,

$$\delta(B) \equiv \frac{1}{|B|}\int_B \delta(\mathbf{s})\,d\mathbf{s} = \boldsymbol{\Phi}(B)'\boldsymbol{\eta} + \nu(B),$$

where $\boldsymbol{\Phi}(B) \equiv \frac{1}{|B|}\int_B \boldsymbol{\Phi}(\mathbf{s})\,d\mathbf{s}$ and $\nu(B) \equiv \frac{1}{|B|}\int_B \nu(\mathbf{s})\,d\mathbf{s}$. Upon comparing this expression to (12), it is clear that the process δ with spatial support B also follows an SRE model with the same parameter $\mathbf{K}_\eta$ and basis functions $\boldsymbol{\Phi}(B)$ that can be obtained by (off-line) integration of $\boldsymbol{\Phi}(\cdot)$.

The model has been criticized as giving realizations that are too smooth (Stein, 2014), but that is relative to what "too" means. Bradley et al. (2015) and Zammit-Mangion et al. (2017) address this by establishing that less-smooth predictors lack the full optimality obtained from kriging. Further, since the variance of the fine-scale variation, σ_ν^2, is estimated and not specified, the SRE model is adaptive to the smoothness of the process; further details are given in Zammit-Mangion and Cressie (2017).

The more important generalization is to move away from fine-scale variation $\nu(\cdot)$ having white-noise covariance given by (15). That is, instead of using a nugget effect (Matheron, 1963) for the fine-scale variation, one could use a model that incorporates local spatial dependence while maintaining fast computation for the inverse of the no-longer-diagonal

matrix $\mathbf{U}$ in (17). A nearest-neighbor conditional autoregressive (CAR) process at the BAU scale is one possible choice for $\nu(\cdot)$; then the matrix $\mathbf{E}$ in (16) (and hence $\mathbf{U}$ in (17)) is not diagonal, but it is very sparse and its inverse can be obtained quickly (Ma and Kang, 2017).

Details of the Main Paper's Section 3.4: Equal area hexagonal grids

The latitude-longitude (lat-lon) coordinate system has served navigation on the geoid well, but for the purpose of mapping geophysical variables aggregated onto equal fractions of longitude, every change in latitudinal zone results in a change of the cells' areas, with the biggest areal difference occurring between cells in the equatorial zones and cells in the polar zones. The geophysical variable $X(B)$ on spatial support B is:

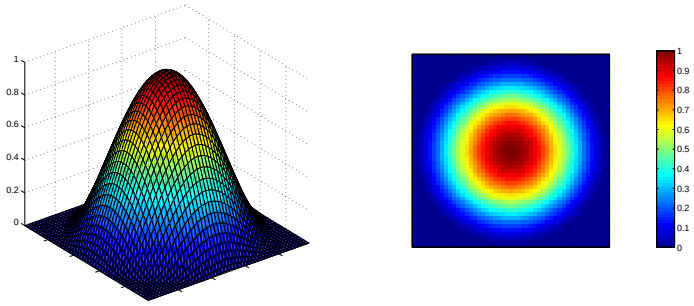
$$X(B) = \frac{1}{|B|} \int_B X(\mathbf{s})\, d\mathbf{s},$$

and the variability of $X(B)$ is greater for smaller B. For this reason, global gridding systems that have (mostly) equal area have been developed (e.g., Sahr et al., 2003).

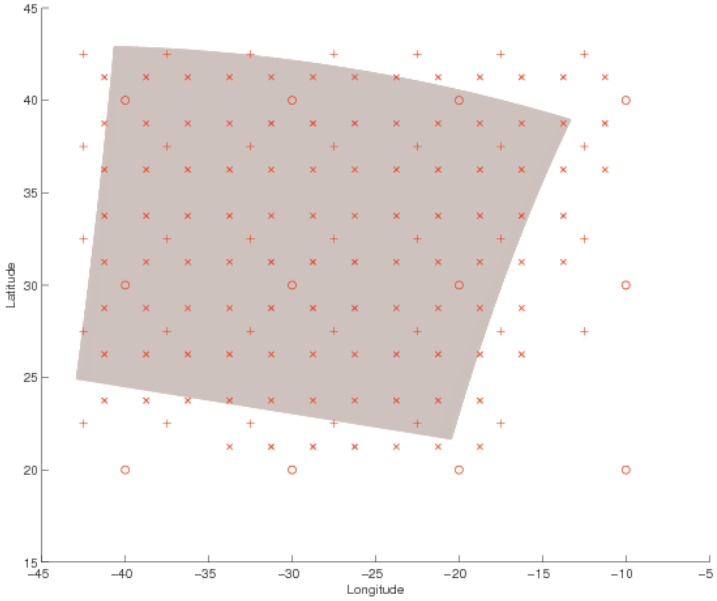
The Icosahedral Snyder Equal Area (ISEA) grid tessellates the sphere and offers hexagonal Basic Areal Units (BAUs) of equal area, apart from a small number of pentagons (Sahr et al., 2003). The ISEA grid is based on a hexagonal tessellation of the sphere with 12 pentagons at corners of the icosahedron. All calculations are done on the flattened icosahedron superimposed on a hexagonal grid; see Figure 5. Whether a latitude-longitude grid or a hexagonal grid is used, the per-grid-cell computation time for FRK is the same.

The hexagons are wrapped back onto the sphere and different views of Earth show a map of the geophysical variable. Figure 6 shows two successive resolutions of an ISEA grid, a map of the geophysical variable XCO2 from the Parametrized Chemistry Transport Model (or PCTM; see Kawa et al., 2004) on the flattened icosahedron, and an Earth view of the same map over North America (Stough et al., 2014).

When areas of spatial support are equal, the comparative interpretation of grid values is uncomplicated, as is the computation of kriging standard errors. In the case of OCO-2, the high-latitude regions generally have no OCO-2 data, and kriging is done in the convex hull of the retrieval locations, D^O. Consequently, the difference between kriging on latitude-longitude grids and kriging on hexagonal grids will be small for L3 maps based on OCO-2 data. For other geophysical variables, like stratospheric ozone, the polar regions are the most important, and L3 maps on equal area grids like ISEA have a distinct advantage.



(a) Generic bisquare basis function in $\mathbb{R}^2$



(b) An example of multi-resolution centers of a collection of spatial basis functions on a sub-geoid flattened onto $\mathbb{R}^2$. The symbols 'o', '+', and 'x' are used to distinguish coarse, medium, and fine resolutions, respectively

Figure 3: Multi-resolution bisquare spatial basis functions in $\mathbb{R}^2$ (Credit: Sengupta et al., 2016)

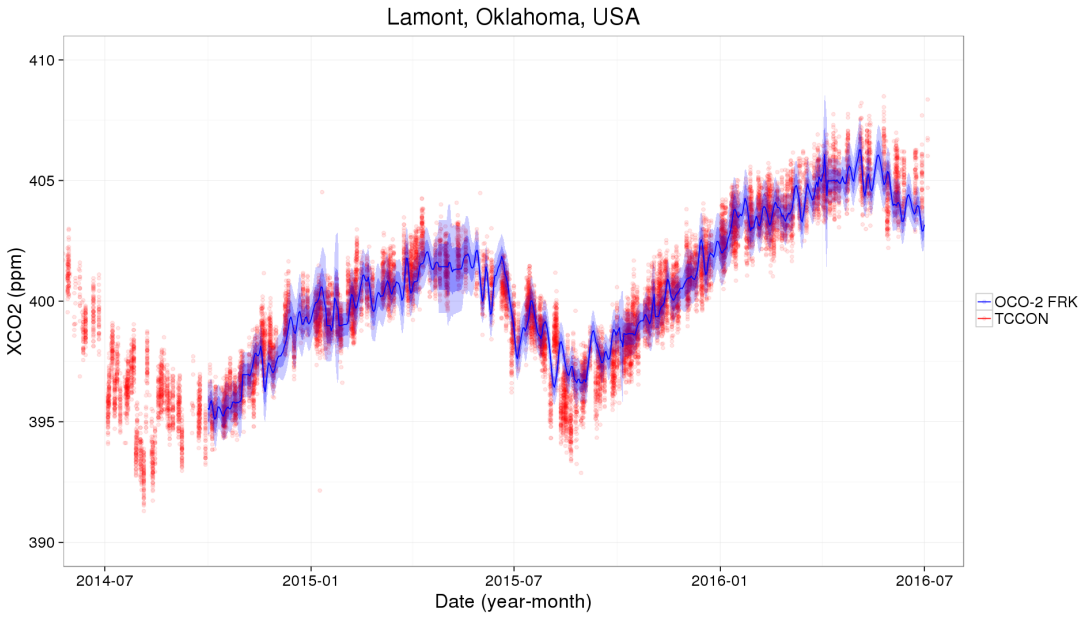


Figure 4: Comparison of XCO2 values for FRK predictors from OCO-2 data and Total Carbon Column Observing Network (TCCON) averages of measurements in a 60-minute window around the satellite's local crossing time at the Lamont, OK TCCON site. The dark-blue shading gives a prediction interval between $\pm$ FRK standard error, and the light-blue shading gives a prediction interval between $\pm$ 2 $\times$ FRK standard error. The OCO-2 data are sparse in the middle of the period plotted, and the two prediction intervals are most easily distinguished there

In the FRK software described in Zammit-Mangion and Cressie (2017), there is an option to use either the lat-lon grid or the Icosahedral Snyder Equal Area (ISEA) grid based on a hexagonal tessellation. The per-grid-cell computation time is the same for either grid.

References

Banerjee, S., B. Carlin, and A. Gelfand (2015). *Hierarchical Modeling and Analysis for Spatial Data, 2nd edn.* Chapman and Hall/CRC, Boca Raton, FL.

Bösch, H., D. Baker, B. Connor, D. Crisp, and C. Miller (2011). Global characterization of CO_2 column retrievals from shortwave-infrared satellite observations of the Orbiting Carbon Observatory-2 mission. *Remote Sensing 3*(2), 270–304.

Bradley, J. R., N. Cressie, and T. Shi (2015). Rejoinder on discussion of: Comparing and selecting spatial predictors using local criteria. *Test 24*, 54–60.

Bréon, F.-M. and P. Ciais (2010). Spaceborne remote sensing of greenhouse gas concentrations. *Comptes Rendus Geoscience 342*(4), 412–424.

Connor, B., H. Bösch, G. Toon, B. Sen, C. Miller, and D. Crisp (2008). Orbiting Carbon Observatory: Inverse method and prospective error analysis. *Journal of Geophysical Research: Atmospheres 113*, D05305.

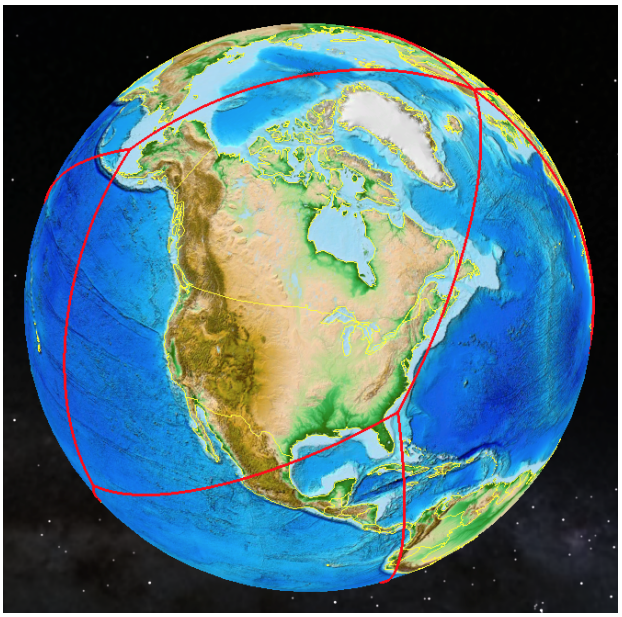
Cressie, N. (1993). *Statistics for Spatial Data, rev. edn.* Wiley, New York, NY.

Cressie, N. and G. Johannesson (2006). Spatial prediction for massive datasets. *Mastering the Data Explosion in the Earth and Environmental Sciences*, Australian Academy of Sciences, Canberra, Australia, 1-11.

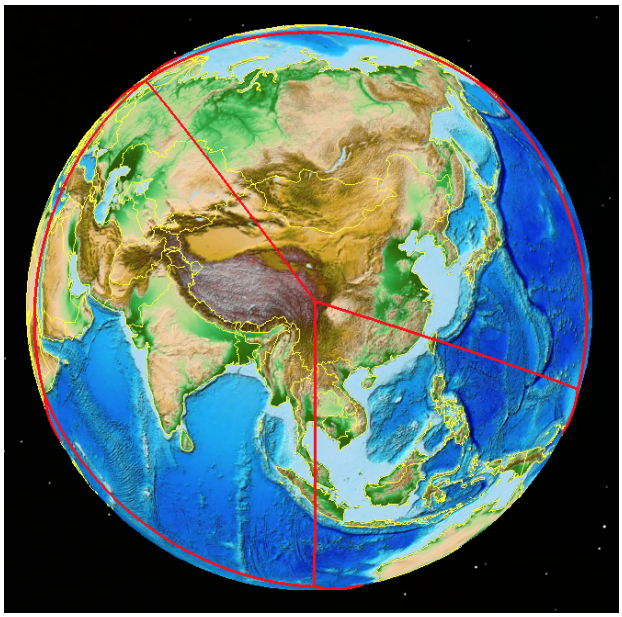
Cressie, N. and G. Johannesson (2008). Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society, Series B 70*, 209–226.

Cressie, N. and J. Kornak (2003). Spatial statistics in the presence of location error with an application to remote sensing of the environment. *Statistical Science 18*, 436–456.

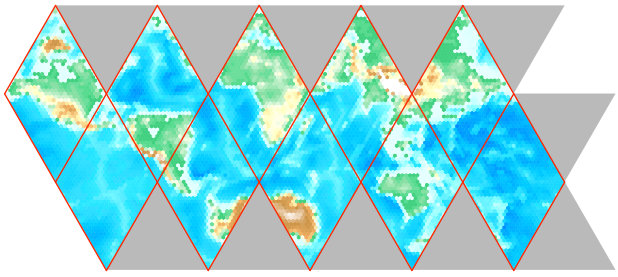
Cressie, N. and C. K. Wikle (2011). *Statistics for Spatio-Temporal Data*. Wiley, Hoboken, NJ.



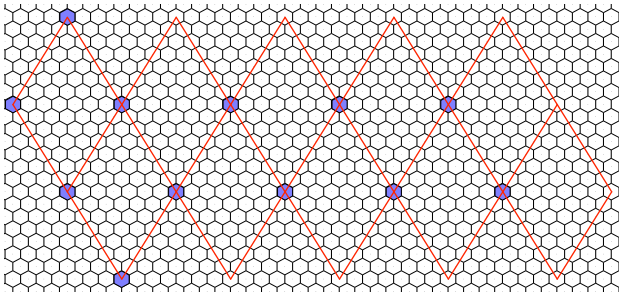
(a) Earth and its icosahedron: North America is featured



(b) Earth and its icosahedron: Asia is featured

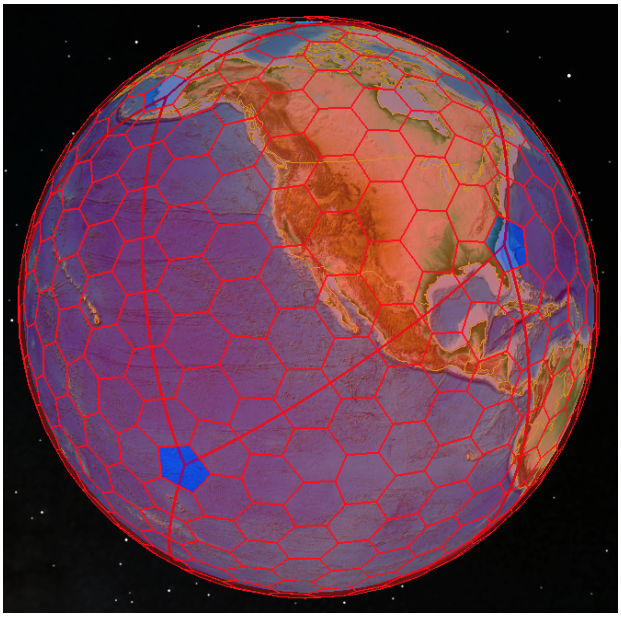


(c) Flattened Icosahedron

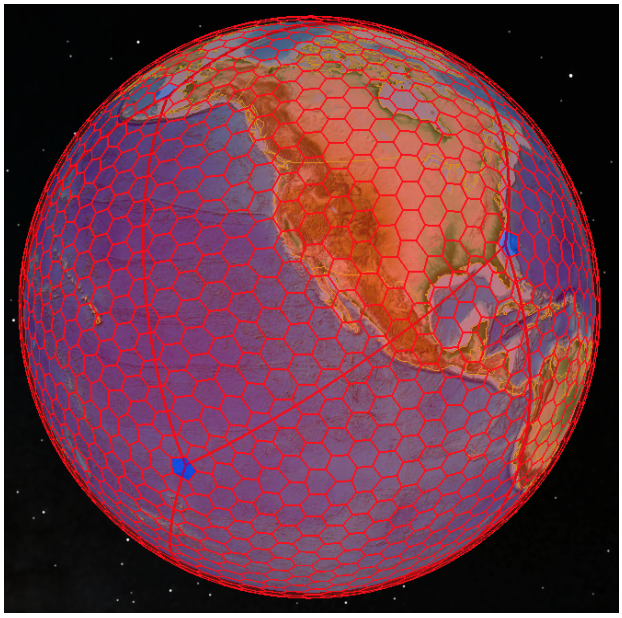


(d) Icosahedron Snyder Equal Area (ISEA) hexagons at Resolution 3 with the flattened icosahedron superimposed. On the sphere, the 12 colored hexagons are represented as pentagons

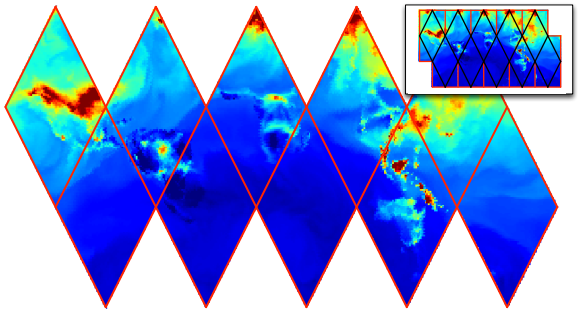
Figure 5: Geoid with icosahedron featured (Credit: Timothy Stough, JPL-Caltech)



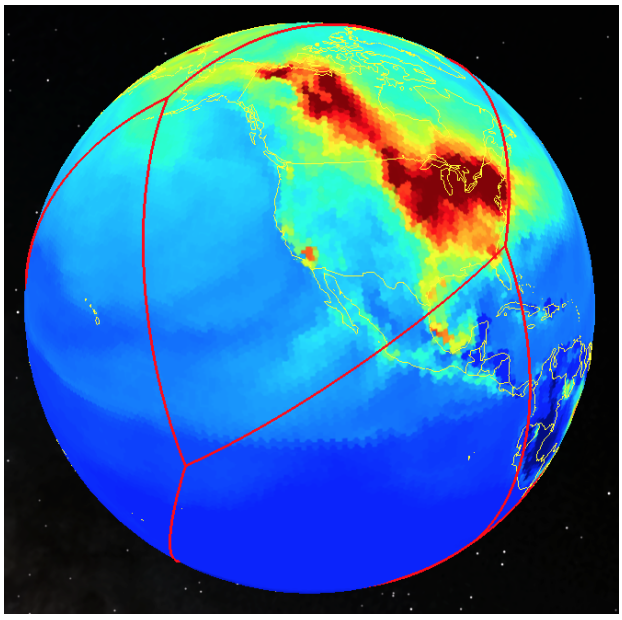
(a) Earth and the ISEA Resolution 3 grid superimposed



(b) Earth and the ISEA Resolution 4 grid superimposed



(c) Flattened icosahedron with model-based XCO2 values shown at Resolution 6. The insert panel shows some repeated triangles for spatial neighborhood calculations



(d) Earth view of (c)

Figure 6: Geoid with ISEA grid featured (Credit: Timothy Stough, JPL-Caltech)

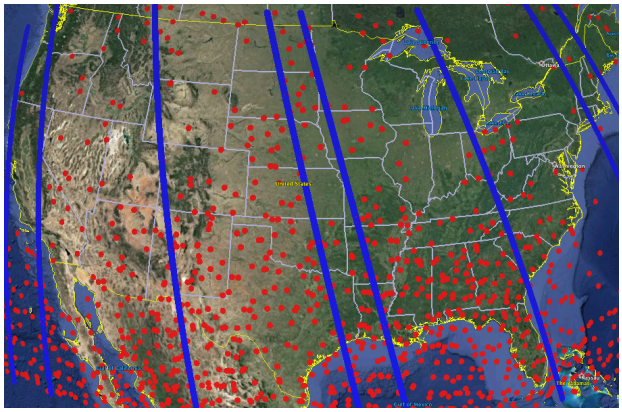


Figure 7: Sounding locations from two remote sensing instruments that retrieve CO_2 measurements: The AIRS soundings are in red, and the OCO-2 soundings are in blue; footprint sizes are not to scale (Credit: Timothy Stough, JPL-Caltech)

Haas, T. C. (1995). Local prediction of a spatio-temporal process with an application to wet sulfate deposition. *Journal of the American Statistical Association 90*, 1189–1199.

Hammerling, D. M., A. M. Michalak, C. O'Dell, and S. R. Kawa (2012). Global CO_2 distributions over land from the Greenhouse Gases Observing Satellite (GOSAT). *Geophysical Research Letters 39*.

Henderson, H. and S. Searle (1981). On deriving the inverse of a sum of matrices. *SIAM Review 23*, 53–60.

Hughes, J. and M. Haran (2013). Dimension reduction and alleviation of confounding for spatial generalized linear mixed models. *Journal of the Royal Statistical Society, Series B 75*(1), 139–159.

Kang, E. L. and N. Cressie (2011). Bayesian inference for the spatial random effects model. *Journal of the American Statistical Association 106*, 972–983.

Katzfuss, M. and N. Cressie (2009). Maximum likelihood estimation of covariance parameters in the spatial-random-effects model. *Proceedings of the Joint Statistical Meetings*, American Statistical Association, Alexandria, VA, 3378-3390.

Kawa, S., D. Erickson, S. Pawson, and Z. Zhu (2004). Global CO_2 transport simulations using meteorological data from the NASA data assimilation system. *Journal of Geophysical Research: Atmospheres 109*, D18.

Lindgren, F., H. Rue, and J. Lindström (2011). 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 73*(4), 423–498.

Ma, P. and E. Kang (2017). Fused Gaussian process for very large spatial data. `arXiv:1702.08797`.

Matheron, G. (1963). Principles of geostatistics. *Economic Geology 58*, 1246–1266.

Miller, C. E., D. Crisp, P. L. DeCola, S. C. Olsen, J. T. Randerson, A. M. Michalak, A. Alkhaled, P. Rayner, D. J. Jacob, P. Suntharalingam, D. B. A. Jones, A. S. Denning, M. E. Nicholls, S. C. Doney, S. Pawson, H. Bösch, B. J. Connor, I. Y. Fung, D. O'Brien, R. J. Salawitch, S. P. Sander, B. Sen, P. Tans, G. C. Toon, P. O. Wennberg, S. C. Wofsy, Y. L. Yung, and R. M. Law (2007). Precision requirements for space-based data. *Journal of Geophysical Research: Atmospheres 112*(D10). D10314.

Nychka, D., S. Bandyopadhyay, D. Hammerling, F. Lindgren, and S. Sain (2015). A multiresolution Gaussian process model for the analysis of large spatial datasets. *Journal of Computational and Graphical Statistics 24*, 579–599.

O'Dell, C., B. Connor, H. Bösch, D. O'Brien, C. Frankenberg, R. Castano, M. Christi, A. Eldering, B. Fisher, M. Gunson, et al. (2012). The ACOS CO_2 retrieval algorithm – Part 1: Description and validation against synthetic observations. *Atmospheric Measurement Techniques 5*, 99–121.

Papoulis, A. (1965). *Probability, Random Variables, and Stochastic Processes*. McGraw Hill, New York, NY.

Prasad, N. and J. N. K. Rao (1990). On the estimation of mean square error of small area predictors. *Journal of the American Statistical Association 85*, 163–171.

Rodgers, C. D. (2000). *Inverse Methods for Atmospheric Sounding*. World Scientific Publishing, Singapore.

Sahr, K., D. White, and A. Kimerling (2003). Geodesic discrete global grid systems. *Cartography and Geographic Information Science 30*, 121–134.

Stein, M. (2014). Limitations on low rank approximations for covariance matrices of spatial data. *Spatial Statistics 8*, 1–19.

Tadić, J. M., X. Qiu, S. Miller, and A. M. Michalak (2017). Spatio-temporal approach to moving window block kriging of satellite data v1.0. *Geoscientific Model Development 10*, 709–720.

Tzeng, S. and C. H. Huang (2017). Resolution adaptive Fixed Rank Kriging. *Technometrics*, forthcoming.

Wikle, C. K. (2010). Low-rank representations for spatial processes. In Gelfand, A.E. et al. (Eds), *Handbook of Spatial Statistics*. Chapman and Hall/CRC Press, Boca Raton, FL, 107-118.

Zammit-Mangion, A. and N. Cressie (2017). FRK: An R package for spatial and spatio-temporal prediction with large datasets. `arXiv:1705.08105`.

Zammit-Mangion, A., N. Cressie, and C. Shumack (2017). On statistical approaches to generate Level 3 products from remote sensing retrievals. `arXiv:1711.07629`.