Supplement to Geostatistical inverse modeling with very large datasets: an example from the OCO-2 satellite

Scot M. Miller[1], Arvind K. Saibaba[2], Michael E. Trudeau[3], Marikate E. Mountain[4], and Arlyn E. Andrews[3]

[1]Department of Environmental Health and Engineering, Johns Hopkins University, Baltimore, MD, USA
[2]Department of Mathematics, North Carolina State University, Raleigh, NC, USA
[3]Global Monitoring Division, National Oceanic and Atmospheric Administration, Boulder, CO, USA
[4]Atmospheric and Environmental Research, Inc., Lexington, MA, USA

Correspondence: Scot M. Miller (smill191@jhu.edu, scot.m.miller@gmail.com)

S1 Efficient calculation of the GIM cost function and gradient

In the main manuscript, we describe an iterative approach to estimate the fluxes that requires calculating the cost function and gradient using the L-BFGS algorithm paired with a variable transformation (Sect. 4.1). These calculations require computing the product $\mathbf{G}^*\boldsymbol{s}^*$ (dimensions $m\times 1$) where $\mathbf{G}^* = \mathbf{I}-\mathbf{Q}^{-\frac{1}{2}}\mathbf{X}(\mathbf{X}^T\mathbf{Q}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Q}^{-\frac{1}{2}}$. The matrix $\mathbf{G}^*$ has dimensions $m\times m$ and is often prohibitively large to compute explicitly or store in memory.

Instead, these calculations can be broken down into smaller components that are far more computationally efficient and require comparatively less memory:

$$\mathbf{G}^*\boldsymbol{s}^* = \underbrace{\boldsymbol{s}^*}_{m\times 1} - \underbrace{\mathbf{A}}_{m\times p}\underbrace{(\mathbf{X}^T\mathbf{B})^{-1}}_{p\times p}\underbrace{(\mathbf{A}^T\boldsymbol{s}^*)}_{p\times 1} \tag{S1}$$

$$\mathbf{A} = \mathbf{Q}^{-\frac{1}{2}}\mathbf{X} \tag{S2}$$

$$\mathbf{B} = \mathbf{Q}^{-1}\mathbf{X} \tag{S3}$$

The calculations in Eq. S1 require multiplying several matrices with p rows and/or columns, where $p << m$. Furthermore, we can compute the matrices $\mathbf{A}$ and $\mathbf{B}$ efficiently in a manner that does not require explicitly formulating or manipulating $\mathbf{Q}$ in its entirety. To do so, we use a Kronecker product and reformulate the approach described in Yadav and Michalak (2013):

$$\mathbf{A} = \mathbf{Q}^{-\frac{1}{2}}\mathbf{X} = \begin{bmatrix} \mathbf{E}^{-\frac{1}{2}}\sum_{i=1}^{q}\frac{1}{\sigma_Q}D^{-\frac{1}{2}}(i,1)\mathbf{X}_{i,*} \\ \mathbf{E}^{-\frac{1}{2}}\sum_{i=1}^{q}\frac{1}{\sigma_Q}D^{-\frac{1}{2}}(i,2)\mathbf{X}_{i,*} \\ \vdots \\ \mathbf{E}^{-\frac{1}{2}}\sum_{i=1}^{q}\frac{1}{\sigma_Q}D^{-\frac{1}{2}}(i,q)\mathbf{X}_{i,*} \end{bmatrix} \tag{S4}$$

$$\mathbf{B} = \mathbf{Q}^{-1}\mathbf{X} = \begin{bmatrix} \mathbf{E}^{-1}\sum_{i=1}^{q}\frac{1}{\sigma_Q^2}D^{-1}(i,1)\mathbf{X}_{i,*} \\ \mathbf{E}^{-1}\sum_{i=1}^{q}\frac{1}{\sigma_Q^2}D^{-1}(i,2)\mathbf{X}_{i,*} \\ \vdots \\ \mathbf{E}^{-1}\sum_{i=1}^{q}\frac{1}{\sigma_Q^2}D^{-1}(i,q)\mathbf{X}_{i,*} \end{bmatrix} \tag{S5}$$

The vector $\boldsymbol{X}_{i,*}$ denotes the rows of $\mathbf{X}$ corresponding to time period i. The variable $D^{-\frac{1}{2}}(i,1)$ corresponds to the i^{th} row and first column of the matrix $\mathbf{D}^{-\frac{1}{2}}$. Note that each line or block in the equations above has dimensions $r \times p$. There are q blocks, and collectively the matrix $\mathbf{A}$ and $\mathbf{B}$ have dimensions $rq \times p$ or, equivalently, $m \times p$. Note that $\mathbf{A}$ and $\mathbf{B}$ need only be calculated once and can subsequently be used in every iteration of the iterative solver.

One can use an analogous approach to calculate $\mathbf{Q}(\mathbf{H}^T\boldsymbol{\xi})$ in the minimum residual approach to the solution (Eq. 7–8 and Sect. 4.2) and $\mathbf{Q}^{\frac{1}{2}}\boldsymbol{s}^*$ in Eqs. 14, 16, and 17 (Sect. 4.1).

S2 Generation of conditional realizations using L-BFGS and a variable transformation

A conditional realization or simulation of the fluxes (denoted $\boldsymbol{s_c}$, dimensions $m \times 1$) is an estimate of the fluxes that randomly samples from the posterior distribution. A large number of conditional realizations can be used to represent the posterior uncertainties in the fluxes. Several existing studies detail how to create conditional realizations using the GIM system of linear equations (Eq. 7–8, Kitanidis, 1996; Saibaba and Kitanidis, 2012). Furthermore, Kitanidis (1995) and Snodgrass and Kitanidis (1997) describe how to generate conditional realizations using the cost function and gradient without a variable transformation (e.g., using the L-BFGS algorithm, Sect. 4.1). Here, we detail how to create conditional realizations using L-BFGS with a variable transformation (Eq. 11–16).

We modify the cost function and gradient functions slightly to obtain the equations required for generating conditional realizations ($\boldsymbol{s_c}$). In this study, we estimate $\boldsymbol{s_c^*}$ using the cost function and gradient paired with a L-BFGS minimum-finding algorithm. Subsequently, we back-transform $\boldsymbol{s_c^*}$ to obtain $\boldsymbol{s_c}$:

$$L(\boldsymbol{s_c^*}) = \tfrac{1}{2}(\boldsymbol{z} + \boldsymbol{\epsilon}_c - \mathbf{H}\mathbf{Q}^{\frac{1}{2}}\boldsymbol{s_c^*})^T\mathbf{R}^{-1}(\boldsymbol{z} + \boldsymbol{\epsilon}_c - \mathbf{H}\mathbf{Q}^{\frac{1}{2}}\boldsymbol{s_c^*}) + \tfrac{1}{2}(\boldsymbol{s_c^*} - \boldsymbol{u})^T\mathbf{G}^*(\boldsymbol{s_c^*} - \boldsymbol{u}) \tag{S6}$$

$$\nabla L(\boldsymbol{s_c^*}) = -\tfrac{1}{2}\mathbf{Q}^{\frac{1}{2}}\mathbf{H}^T\mathbf{R}^{-1}(\boldsymbol{z} + \boldsymbol{\epsilon}_c - \mathbf{H}\mathbf{Q}^{\frac{1}{2}}\boldsymbol{s_c^*}) + \tfrac{1}{2}\mathbf{G}^*(\boldsymbol{s_c^*} - \boldsymbol{u}) \tag{S7}$$

$$\boldsymbol{s_c} = \mathbf{Q}^{\frac{1}{2}}\boldsymbol{s_c^*} \tag{S8}$$

where $\boldsymbol{\epsilon}_c$ is a random $n \times 1$ vector with $\boldsymbol{\epsilon}_c \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$. If $\mathbf{R}$ is a diagonal matrix with diagonal elements $\sigma_{\mathbf{R}}^2$, then $\boldsymbol{\epsilon}_c$ is a random vector with a mean of zero and standard deviation of $\sigma_{\mathbf{R}}$. The vector $\boldsymbol{u}$ has dimensions $m \times 1$ and is randomly drawn from a standard normal distribution. Both vectors $\boldsymbol{\epsilon}_c$ and $\boldsymbol{u}$ must be re-generated anew when estimating each new conditional realization.

S3 Additional detail on the OCO-2 case study setup

The inverse model requires an estimate of atmospheric transport, and we use the Weather Research and Forecasting (WRF) model paired with the Stochastic Time-Inverted Lagrangian Transport Model (STILT) for all atmospheric modeling in this study. We generate WRF-STILT simulations for North America and for adjacent regions of the Pacific and Atlantic Oceans (10° to 80° latitude and -10° to -180° longitude).

WRF is a meteorological model, and the simulations used here have been specifically generated to be compatible with STILT (e.g.. Nehrkorn et al., 2010). STILT, by contrast, is a particle trajectory model, described in detail by Lin et al. (2003). STILT leverages a meteorological model like WRF to estimate how CO_2 fluxes in different locations and at different times would affect atmospheric CO_2 levels at the measurement locations. Specifically, we use STILT to generate a set of gridded influence footprints with units of ppm per unit flux. Each STILT footprint subsequently becomes one of n rows in $\mathbf{H}$ (e.g., Eq. 1), a key input of the inverse model. The WRF model outputs used here have a resolution of 10km over the continental US and 30km for other regions of North America. By contrast, we compute the STILT footprints at a spatial resolution of 1° by 1° and a 3-hourly temporal resolution. Note that we do not run WRF-STILT for every OCO-2 observation due to the computational cost. Instead, we generate STILT footprints corresponding to every 2 seconds along the OCO-2 flight track, yielding a total of 9.88×10^4 footprints for the one-year study period (Sept. 2014 – Aug. 2015).

All of the experiments in this study are synthetic; we generate a set of OCO-2-like observations using a known CO_2 flux estimate, and we use these synthetic observations in all of the inverse modeling experiments. This setup makes it possible to compare the CO_2 fluxes estimated using the inverse model against a known set of true CO_2 fluxes. We further add randomly-generated errors to the synthetic observations – errors with a standard deviation of 2ppm, a number similar to that estimated in a recent top-down study using OCO-2 observations (Miller et al., 2018).

Note that we only consider biospheric CO_2 fluxes in the synthetic data setup. Many inverse modeling studies of CO_2 fix anthropogenic emissions at a specific value and/or pre-subtract the modeled contribution of anthropogenic emissions from the observations. Hence, we do not consider anthropogenic emissions in the synthetic case study here.

The covariance matrices also play a key role in the inverse modeling setup. The covariance matrix $\mathbf{R}$ describes errors in the inverse model that are unrelated to uncertain fluxes – errors due to the satellite retrieval, modeled atmospheric transport, and errors due to the finite resolution of the inverse model. For the setup here, $\mathbf{R}$ is a diagonal matrix with a variance of (2 ppm)^2 on the diagonals. By contrast, the covariance matrix $\mathbf{Q}$ describes the spatial and temporal properties of the CO_2 fluxes; this matrix helps to guide the structure of the fluxes estimated by the inverse model. We use restricted maximum likelihood (RML) estimation to estimate the variance, decorrelation time, and decorrelation length in 3-hourly CT2017 fluxes, and we use these values to populate $\mathbf{Q}$. RML is a statistical technique that can be used to estimate the spatial and temporal properties that are most likely given a dataset, in this case CO_2 fluxes from CT2017 (e.g., Corbeil and Searle, 1976; Kitanidis, 1986; Mueller et al., 2008). In this application, RML requires minimizing a cost function that describes the likelihood of the data (i.e., CT2017 fluxes) given some guess for the variance, decorrelation time, and decorrelation length. For the case studies here, we use unique values of the variance for each month, and we use a single estimate for the decorrelation length and time that

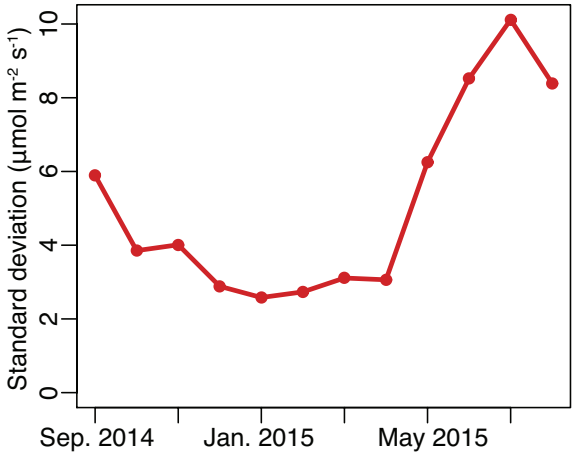

Figure S1. The standard deviation of CO_2 fluxes from CT2017 at long spatial and temporal distances, separated by month. The estimates here are used to construct the $\mathbf{Q}$ covariance matrix. The estimated values show a distinct seasonal cycle that peaks during summer months.

has been averaged across all months. Furthermore, we set up $\mathbf{Q}$ such that fluxes from one three-hour time window covary with fluxes from the same three hour window on adjacent days. However, fluxes from one three-hour time window will not covary with fluxes from other time windows on the same day. For example, fluxes from noon to 15:00 UTC on May 5 will covary with fluxes from noon to 15:00 UTC on May 4 and May 6 but will not covary with fluxes from 9am to noon or 15:00 to 18:00 on May 5. This setup follows that of Gourdji et al. (2010) and Gourdji et al. (2012).

The estimated covariance matrix parameters show a distinct seasonality. For the six week case study, we estimate a standard deviation in the fluxes at long spatial and temporal distances of 10.0 μmol m^{-2} s^{-1} (i.e., the square root of the diagonal elements of $\mathbf{Q}$), a decorrelation length of 555 km (assuming a spherical covariance model), and a correlation time of 9.8 days (also assuming a spherical model). Figure S1 shows the standard deviation of the fluxes at long spatial and temporal distances by month for the one year case study. This figure has a distinct sinusoidal shape, and the standard deviations are highest during summer months. In that case study, we also use a decorrelation length of 586 km and a decorrelation time of 12.4 days. These values are the average of the values estimated for each month of the year.

The matrix $\mathbf{X}$ is also a key input to the GIM. In many studies, the matrix $\mathbf{X}$ can include predictor variables that may help describe the spatial and/or temporal distribution of the fluxes. For the setup here, $\mathbf{X}$ is non-informative and consists of columns of ones. As a result of this setup, the fluxes estimated by the GIM will only reflect the information in the atmospheric observations and will not reflect any prior information about the fluxes. In the 6-week case study for June and July 2015, $\mathbf{X}$ has dimensions $m \times 8$. Each column of $\mathbf{X}$ corresponds to a different 3-hour time period of the day (for a total of 8 columns). For the annual case, we include one column for each month. Both of these setups for $\mathbf{X}$ have been used in past GIM studies (e.g., Gourdji et al., 2008, 2012). The different columns of $\mathbf{X}$ account for the fact that the fluxes have different overall magnitudes at different times of day and/or during different months of the year. The setup here also avoids including too many columns in $\mathbf{X}$; the coefficients ($\boldsymbol{\beta}$) estimated by the GIM will scale each column in $\mathbf{X}$ to fit the observations. The estimated values of these

coefficients are not guided by any prior estimate. As a result, there is nothing to regularize the coefficient ($\boldsymbol{\beta}$) estimates, and those coefficients are only estimated using the atmospheric observations. Hence, we do not want to include too many columns in $\mathbf{X}$ to avoid overfitting the atmospheric observations and/or obtaining unrealistic estimates for the coefficients ($\boldsymbol{\beta}$).

References

Corbeil, R. R. and Searle, S. R.: Restricted Maximum Likelihood (REML) Estimation of Variance Components in the Mixed Model, Technometrics, 18, 31–38, https://doi.org/10.1080/00401706.1976.10489397, 1976.

Gourdji, S. M., Mueller, K. L., Schaefer, K., and Michalak, A. M.: Global monthly averaged CO_2 fluxes recovered using a geostatistical inverse modeling approach: 2. Results including auxiliary environmental data, J. Geophys. Res.-Atmos., 113, https://doi.org/10.1029/2007JD009733, d21115, 2008.

Gourdji, S. M., Hirsch, A. I., Mueller, K. L., Yadav, V., Andrews, A. E., and Michalak, A. M.: Regional-scale geostatistical inverse modeling of North American CO_2 fluxes: a synthetic data study, Atmos. Chem. Phys., 10, 6151–6167, https://doi.org/10.5194/acp-10-6151-2010, 2010.

Gourdji, S. M., Mueller, K. L., Yadav, V., Huntzinger, D. N., Andrews, A. E., Trudeau, M., Petron, G., Nehrkorn, T., Eluszkiewicz, J., Henderson, J., Wen, D., Lin, J., Fischer, M., Sweeney, C., and Michalak, A. M.: North American CO_2 exchange: inter-comparison of modeled estimates with results from a fine-scale atmospheric inversion, Biogeosciences, 9, 457–475, https://doi.org/10.5194/bg-9-457-2012, 2012.

Kitanidis, P. K.: Parameter Uncertainty in Estimation of Spatial Functions: Bayesian Analysis, Water Resour. Res., 22, 499–507, https://doi.org/10.1029/WR022i004p00499, 1986.

Kitanidis, P. K.: Quasi-Linear Geostatistical Theory for Inversing, Water Resour. Res., 31, 2411–2419, https://doi.org/10.1029/95WR01945, 1995.

Kitanidis, P. K.: Analytical expressions of conditional mean, covariance, and sample functions in geostatistics, Stoch. Hydrol. Hydraul., 10, 279–294, https://doi.org/10.1007/BF01581870, 1996.

Lin, J. C., Gerbig, C., Wofsy, S. C., Andrews, A. E., Daube, B. C., Davis, K. J., and Grainger, C. A.: A near-field tool for simulating the upstream influence of atmospheric observations: The Stochastic Time-Inverted Lagrangian Transport (STILT) model, J. Geophys. Res.-Atmos., 108, https://doi.org/10.1029/2002JD003161, 2003.

Miller, S. M., Michalak, A. M., Yadav, V., and Tadić, J. M.: Characterizing biospheric carbon balance using CO_2 observations from the OCO-2 satellite, Atmos. Chem. Phys., 18, 6785–6799, https://doi.org/10.5194/acp-18-6785-2018, 2018.

Mueller, K. L., Gourdji, S. M., and Michalak, A. M.: Global monthly averaged CO_2 fluxes recovered using a geostatistical inverse modeling approach: 1. Results using atmospheric measurements, J. Geophys. Res.-Atmos., 113, https://doi.org/10.1029/2007JD009734, d21114, 2008.

Nehrkorn, T., Eluszkiewicz, J., Wofsy, S. C., Lin, J. C., Gerbig, C., Longo, M., and Freitas, S.: Coupled weather research and forecasting–stochastic time-inverted lagrangian transport (WRF–STILT) model, Meteorol. Atmos. Phys., 107, 51–64, https://doi.org/10.1007/s00703-010-0068-x, 2010.

Peters, W., Jacobson, A. R., Sweeney, C., Andrews, A. E., Conway, T. J., Masarie, K., Miller, J. B., Bruhwiler, L. M. P., Pétron, G., Hirsch, A. I., Worthy, D. E. J., van der Werf, G. R., Randerson, J. T., Wennberg, P. O., Krol, M. C., and Tans, P. P.: An atmospheric perspective on North American carbon dioxide exchange: CarbonTracker, P. Natl. Acad. Sci. USA, 104, 18 925–18 930, https://doi.org/10.1073/pnas.0708986104, 2007.

Saibaba, A. K. and Kitanidis, P. K.: Efficient methods for large-scale linear inversion using a geostatistical approach, Water Resour. Res., 48, https://doi.org/10.1029/2011WR011778, 2012.

Snodgrass, M. and Kitanidis, P.: A geostatistical approach to contaminant source identification, Water Resour. Res., 33, 537–546, https://doi.org/10.1029/96WR03753, 1997.

Yadav, V. and Michalak, A. M.: Improving computational efficiency in large linear inverse problems: an example from carbon dioxide flux estimation, Geosci. Model Dev., 6, 583–590, https://doi.org/10.5194/gmd-6-583-2013, 2013.