Supplemental Information

S1 Real data case study

This section of the SI describes the real data case study. This case study covers the same time period as the summer case study in the main article (July and August 2015), but we use real OCO-2 observations (version 9) instead of synthetic observations. The results provide a check to ensure that the synthetic case study has similar features to real world conditions.

The setup for the real data case study is identical to the synthetic case studies with a few additional steps. In the real data case study, we require a CO_2 boundary condition, an estimate of CO_2 mixing ratios in air masses before they enter the model domain. We generate this boundary condition using an empirical estimate of the CO_2 in air masses over the Pacific and Atlantic oceans adjacent to North America. Specifically, we utilize a boundary curtain developed for CarbonTracker-Lagrange that is generated by smoothing and interpolating aircraft and marine boundary layer CO_2 observations from NOAA's Global Greenhouse Gas Reference Network (e.g., as in Gourdji et al., 2012; Shiga et al., 2018; Hu et al., 2019). This boundary curtain varies by latitude, altitude, and time. We tabulate the ending latitude, altitude, and time of each particle trajectory in a given STILT simulation and find the nearest neighbor boundary curtain value for each of those endpoints. These values are then averaged and the pressure weighting function applied to generate a final CO_2 boundary condition estimate for a specific OCO-2 observation.

The results of the real data case study have features that are similar to the synthetic case study described in the main article. Figure S1 displays the estimated correlation lengths for the real data case study, analogous to Fig. 1 in the main article, and Fig. S2 displays the reduced datasets calculated using these correlation lengths. The estimated correlation lengths in the real data study have a similar magnitude and similar spatial variability as the synthetic study. Note, however, that the estimated correlation lengths sometimes differ between the real and synthetic studies. These differences may occur for two reasons. First, the synthetic case study includes randomly-generated model and measurement errors. These errors will not look identical to the real modeling and observation errors, in part because the synthetic errors are random, and that difference may yield slightly different correlation length estimates in some locations. Second, real-world CO_2 fluxes may not match the spatial and/or temporal patterns in the CarbonTracker CO_2 flux estimate, the estimate used to generate the synthetic observations. Any differences between real-world, unknown CO_2 fluxes and CarbonTracker fluxes may yield slightly different estimates for the correlation length.

The estimated fluxes in the synthetic (Fig. 3) and real data (Fig. S3) case studies also respond similarly as the XCO_2 data is reduced further and further. For example, the patterns in the flux maps begin to degrade at a level of data compression greater than $0.15l$ to $0.2l$. Note that the patterns in the estimated fluxes between the real and synthetic case studies will not look identical; real world fluxes may not precisely match the spatiotemporal patterns in CarbonTracker, and there may be biases in the OCO-2 observations or the estimated boundary condition. For example, Fig. S3a shows a large CO_2 sink in northwest British Columbia that is not present in the synthetic case studies (Fig. 3). There are few OCO-2 observations in that region of northern Canada, and OCO-2 observations at high latitudes often exhibit high noise relative to lower latitudes in comparisons with Total Carbon Column Observing Network (TCCON) observations (e.g., O'Dell et al., 2018). Interestingly, the fluxes estimated using data reduction do not produce a similar sink in northwest British Columbia. In this particular case, data reduction likely reduces the influence of anomalous observations on the estimated CO_2 sink in the region; the kriging step of the data reduction algorithm will smooth out data points that are not consistent with other observations in the region (i.e., have a magnitude that is not consistent with surrounding observations, the known spatial properties of the observations, and known error characteristics of the observations).

We also compare the accuracy of the fluxes estimated using the proposed approach to data reduction against fluxes estimated using data reduced through binning and averaging (Fig. S4). Like Fig. 5 in the main article, we compute the root mean squared error (RMSE) of the grid-scale fluxes estimated with a reduced dataset against fluxes estimated using the full dataset. The results show similar patterns to Fig. 5. Specifically, the approach proposed here yields a lower RMSE relative to fluxes that have been estimated from data reduced with binning and averaging. Like the synthetic case studies, the RMSE for the geostatistical approach shows a clear inflection point, after which the RMSE begins to increase much more rapidly. As in the synthetic case studies (Figs. 3 and 5), the spatial definition of the flux maps (Fig. S3) begins to degrade before the RMSE reaches an inflection point (Fig. S4). In the main manuscript, we point out that different metrics (e.g., flux maps and RMSE) provide similar albeit somewhat different information, and we argue that a modeler may need to consider multiple different metrics or criteria when deciding on an appropriate level of data reduction.

Figure S5 provides an additional metric to help guide the choice of data reduction; it displays a measure of the variance in the OCO-2 observations that is lost through the process of data reduction, analogous to Fig. 6. The shape of the curve in Fig. S5, calculated for the real data case study, is very similar to the shape of the curve computed for the synthetic data case study (Fig. 6). In each figure, the best fit line shows a clear inflection point between 1000 and 2000 observations, and the lost variance increases more quickly at greater levels of data reduction. Note that the y-axes in Figs. S5 and 6a have slightly different magnitudes, and this difference is due to the way we generate the synthetic observations. In the synthetic data case study, we add randomly-generated error to the observations to not only represent observational errors but also to mimic errors in the atmospheric modeling system. The addition of simulated modeling errors to the synthetic data also increases the variance of that data. As a result, the synthetic observations have a higher variance than the real data observations – because simulated atmospheric modeling errors have been added to the synthetic observations.

S2 Example variograms

Variogram fitting is a key aspect of the proposed approach to data reduction and yields an estimate of the decorelation length in the observations. Figure S6 shows two example variograms from the synthetic OCO-2 observations for summer 2015. Both variograms are from western Canada. The empirical variogram is a measure of the differences among pairs of observations at different distances from one another (e.g., Kitanidis, 1997; Wackernagel, 2003). Each dot in Fig. S6 represents the average value for many pairs of observations. The lines in Fig. S6 displays the variogram model, fitted to the gray points using a least squares fit. The fitted parameters of the model provide an estimate of the spatial properties of the observations, including the decorrelation length. In each panel, the semi-variance, a measure of the differences among observations, is smallest among pairs of observations that are located near one another. The semi-variance increases at greater distances and then levels off. The decorrelation length is defined as the distance at which the variogram model levels off; observations at that distance and greater distances show no spatial correlation.

S3 Additional detail on the inverse modeling setup

We estimate CO_2 fluxes in the case studies using a geostatistical inverse model. The case studies here are similar to those described in Miller et al. (2020). That study provides additional details on the inverse modeling setup for the case studies, but we also summarize the inverse modeling setup here. Specifically, we estimate CO_2 fluxes ($\boldsymbol{s}$, dimensions $m \times 1$) by solving a linear system of equations that minimizes the following cost function (e.g., Kitanidis and Vomvoris, 1983; Michalak et al., 2004):

$$L(\boldsymbol{s}, \boldsymbol{\beta}) \;=\; \tfrac{1}{2}(\boldsymbol{z} - \mathbf{H}\boldsymbol{s})^T \mathbf{R}^{-1} (\boldsymbol{z} - \mathbf{H}\boldsymbol{s}) + \tfrac{1}{2}(\boldsymbol{s} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{Q}^{-1} (\boldsymbol{s} - \mathbf{X}\boldsymbol{\beta}) \qquad \text{(S1)}$$

where $\boldsymbol{z}$ ($n \times 1$) are the synthetic or real data observations from OCO-2, $\mathbf{H}$ ($n \times m$) is the atmospheric transport model (in this case, footprints from WRF-STILT), $\mathbf{R}$ ($n \times n$) is a covariance matrix that describes errors in the observations and atmospheric model, and $\mathbf{Q}$ ($m \times m$) is a covariance matrix that describe the variance, spatial covariances, and temporal covariances in the fluxes. In addition, $\mathbf{X}$ ($m \times p$) contains different covariates that help describe patterns in the unknown fluxes ($\boldsymbol{s}$), and $\boldsymbol{\beta}$ ($p \times 1$) are coefficients that scale the magnitude of the columns in $\mathbf{X}$. These coefficients are estimated as part of the inverse model, along with the fluxes ($\boldsymbol{s}$). In the particular setup here, we use a non-informative prior, so $\mathbf{X}$ contains columns of ones. As a result, any spatial patterns in the fluxes are solely the result of information in the synthetic or real atmospheric observations, not the results of any prior flux estimate. Specifically, $\mathbf{X}$ has dimensions $m \times 8$ for each case study. Each column of $\mathbf{X}$ contains ones and zeros that correspond to each 3-hour time period of the day (for a total of 8 columns) (similar to Gourdji et al., 2012). The different columns of $\mathbf{X}$ account for the fact that the fluxes have different overall magnitudes at different times of day.

Note that the covariance matrices must be estimated by the modeler prior to estimating the fluxes ($\boldsymbol{s}$). We populate the covariance matrices with different values for the summer and winter case studies, but we use the same values for both the summer synthetic and summer real data case studies. The covariance matrix $\mathbf{R}$ describes errors in the model–data system ($\boldsymbol{\epsilon}$ in Sect. 3), and Sect. 3 of the main article lists the estimated values of these errors. By contrast, we use restricted maximum likelihood (RML) estimation to estimate the variance, decorrelation time, and decorrelation length in 3-hourly CarbonTracker 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 CarbonTracker (e.g., Corbeil and Searle, 1976; Kitanidis, 1986; Mueller et al., 2008). We specifically estimate these properties by minimizing a cost function that describes the likelihood of the data (in this case, the CarbonTracker fluxes) given some guess for the variance, decorrelation time, and decorrelation length. 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 July 5 will covary with fluxes from noon to 15:00 UTC on July 4 and July 6 but will not covary with fluxes from 9am to noon or 15:00 to 18:00 on July 5. This setup parallels that of Gourdji et al. (2010) and Gourdji et al. (2012). In the summer case studies, we estimate a variance of (10 μmol m^{-2} s^{-1})^2, decorrelation length of 555 km, and decorrelation time of 9.9 days. For the winter case study, we estimate a variance of (3.1 μmol m^{-2} s^{-1})^2, decorrelation length of 647 km, and decorrelation time of 14.7 days.

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., 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.

Hu, L., Andrews, A. E., Thoning, K. W., Sweeney, C., Miller, J. B., Michalak, A. M., Dlugokencky, E., Tans, P. P., Shiga, Y. P., Mountain, M., Nehrkorn, T., Montzka, S. A., McKain, K., Kofler, J., Trudeau, M., Michel, S. E., Biraud, S. C., Fischer, M. L., Worthy, D. E. J., Vaughn, B. H., White, J. W. C., Yadav, V., Basu, S., and van der Velde, I. R.: Enhanced North American carbon uptake associated with El Niño, Science Advances, 5, https://doi.org/10.1126/sciadv.aaw0076, 2019.

Kitanidis, P.: Introduction to Geostatistics: Applications in Hydrogeology, Stanford-Cambridge program, Cambridge University Press, Cambridge, 1997.

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. and Vomvoris, E. G.: A geostatistical approach to the inverse problem in groundwater modeling (steady state) and one-dimensional simulations, Water Resour. Res., 19, 677–690, https://doi.org/10.1029/WR019i003p00677, 1983.

Michalak, A. M., Bruhwiler, L., and Tans, P. P.: A geostatistical approach to surface flux estimation of atmospheric trace gases, J. Geophys. Res.-Atmos., 109, D14 109, https://doi.org/10.1029/2003JD004422, 2004.

Miller, S. M., Saibaba, A. K., Trudeau, M. E., Mountain, M. E., and Andrews, A. E.: Geostatistical inverse modeling with very large datasets: an example from the Orbiting Carbon Observatory 2 (OCO-2) satellite, Geosci. Model Dev., 13, 1771–1785, https://doi.org/10.5194/gmd-13-1771-2020, 2020.

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, D21 114, https://doi.org/10.1029/2007JD009734, 2008.

O'Dell, C. W., Eldering, A., Wennberg, P. O., Crisp, D., Gunson, M. R., Fisher, B., Frankenberg, C., Kiel, M., Lindqvist, H., Mandrake, L., Merrelli, A., Natraj, V., Nelson, R. R., Osterman, G. B., Payne, V. H., Taylor, T. E., Wunch, D., Drouin, B. J., Oyafuso, F., Chang, A., McDuffie, J., Smyth, M., Baker, D. F., Basu, S., Chevallier, F., Crowell, S. M. R., Feng, L., Palmer, P. I., Dubey, M., García, O. E., Griffith, D. W. T., Hase, F., Iraci, L. T., Kivi, R., Morino, I., Notholt, J., Ohyama, H., Petri, C., Roehl, C. M., Sha, M. K., Strong, K., Sussmann, R., Te, Y., Uchino, O., and Velazco, V. A.: Improved retrievals of carbon dioxide from Orbiting Carbon Observatory-2 with the version 8 ACOS algorithm, Atmos. Meas. Tech., 11, 6539–6576, https://doi.org/10.5194/amt-11-6539-2018, 2018.

Shiga, Y. P., Michalak, A. M., Fang, Y., Schaefer, K., Andrews, A. E., Huntzinger, D. H., Schwalm, C. R., Thoning, K., and Wei, Y.: Forests dominate the interannual variability of the North American carbon sink, Environ. Res. Lett., 13, 084 015, https://doi.org/10.1088/1748-9326/aad505, 2018.

Wackernagel, H.: Multivariate Geostatistics: An Introduction with Applications, Springer, Berlin, 2003.

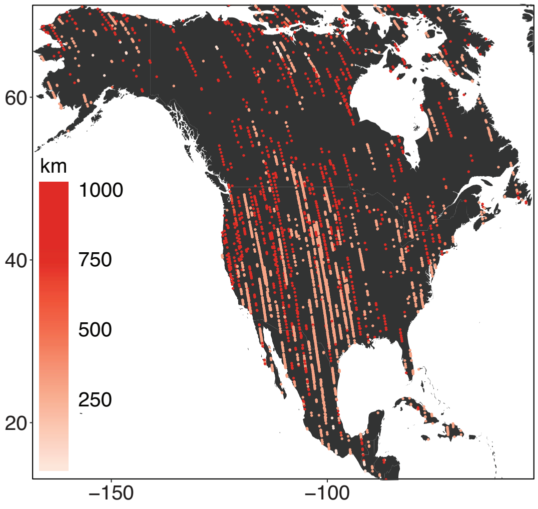

Figure S1: Correlation lengths estimated along OCO-2 flight tracks for the summer real data case study. The estimated correlation lengths are similar in magnitude and exhibit similar spatial heterogeneity to those in the synthetic case study (Fig. 1).

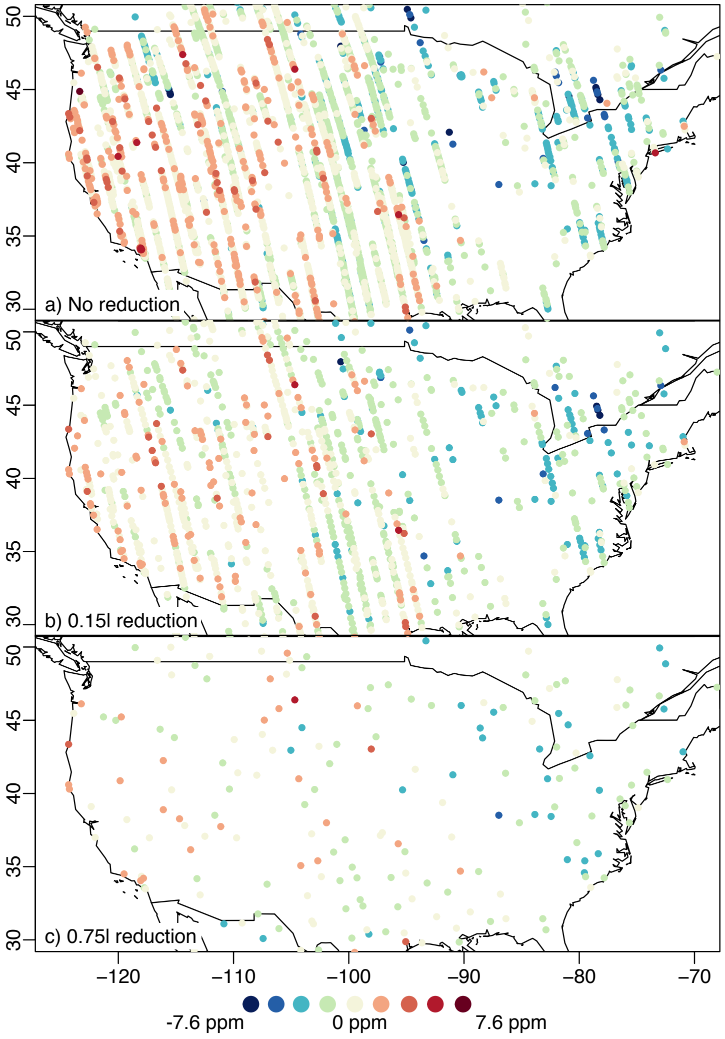

Figure S2: The original XCO_2 dataset after subtracting the CO_2 background or boundary condition (a), the dataset reduced to one observation per $0.15l$ (b), and the dataset reduced to one observation per $0.75l$ (c). Panels (a) and (b) display similar spatial patterns, but the data in panel (c) is very sparse. This figure is zoomed in over the United States to better show spatial patterns in the observations.

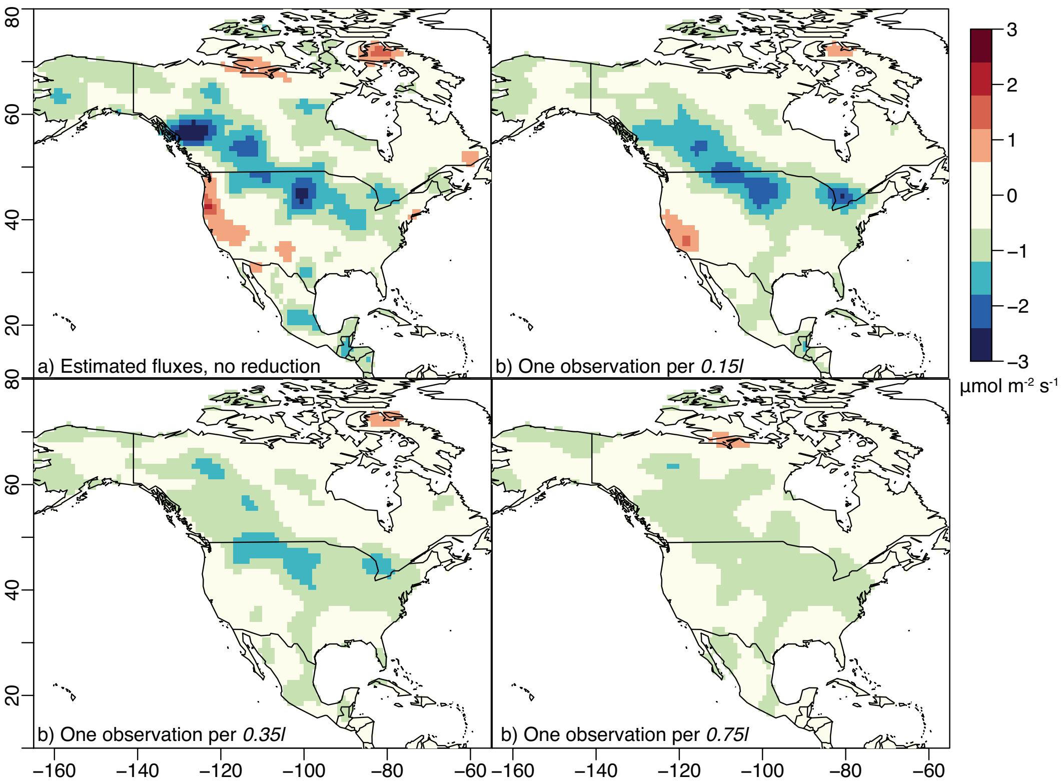
80
60
40
20
80
60
40
20
a) Estimated fluxes, no reduction
b) One observation per 0.15l
b) One observation per 0.35l
b) One observation per 0.75l
−160
−140
−120
−100
−80
−60
−160
−140
−120
−100
−80
−60
3
2
1
0
−1
−2
−3
µmol m-2 s-1

Figure S3: CO_2 fluxes estimated for the summer 2015 real data case study, averaged across the 6-week study window: (a) fluxes estimated from OCO-2 data with no reduction (5032 data points), (b) fluxes estimated from data reduced to one point per $0.15l$ (2156 data points), and (c) fluxes estimated from data reduced to one point per $0.75l$ (565 data points). The estimate with a reduction of $0.15l$ (b) reproduces most of the spatial patterns in panel (a), while the estimate with $0.75l$ reduction has lost spatial definition (c).

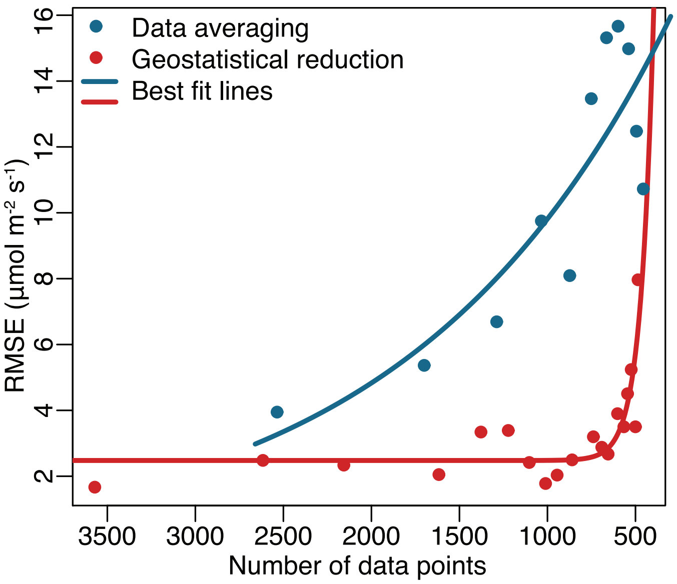

Figure S4: Root mean squared error (RMSE) of the fluxes estimated using data reduction relative to the fluxes estimated without data reduction (analogous to Fig. 5). The RMSE for the geostatistical approach proposed here is less than the RMSE for data reduced using binning and averaging.

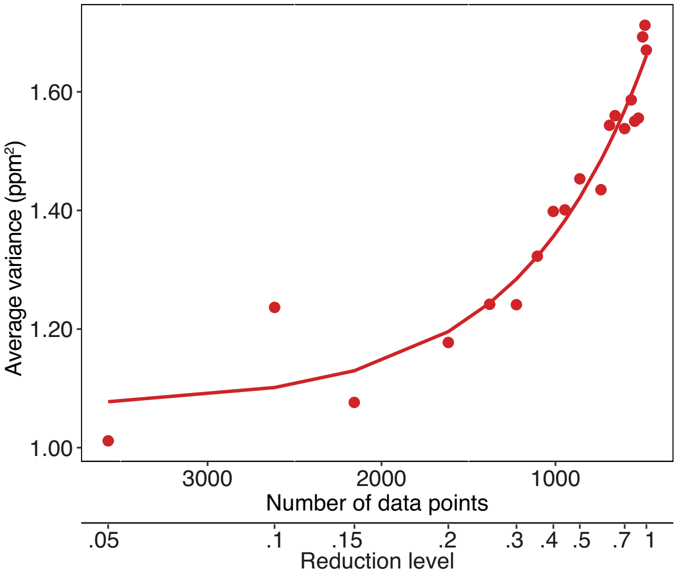

Figure S5: The amount of variance in the data that is lost through the process of data reduction. The patterns in this figure from the real data case study are similar to those from the summer synthetic case study (Fig. 6).

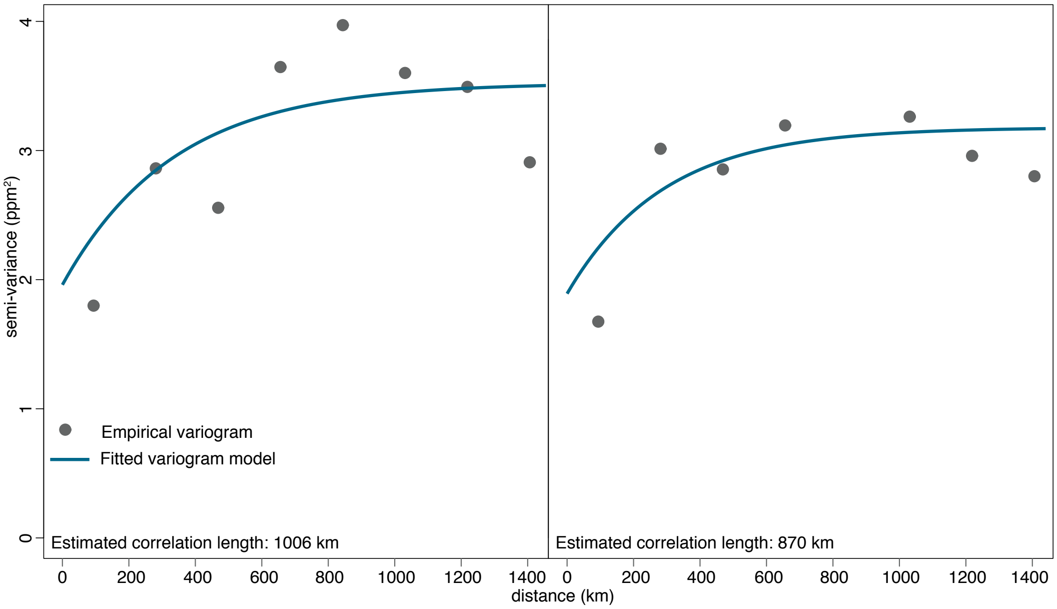

Figure S6: Two example variograms from the summer synthetic case study. The emperical variogram (grey) is calculated from the synthetic observations, and the variogram model (blue) is fitted to these points using least squares.