Data reduction for inverse modeling: an adaptive approach v1.0

Xiaoling Liu[1], August L. Weinbren[1], He Chang[1], Jovan Tadić[2], Marikate E. Mountain[3], Michael E. Trudeau[4], Arlyn E. Andrews[4], Zichong Chen[1], and Scot M. Miller[1]

[1]Department of Environmental Health and Engineering, Johns Hopkins University, Baltimore, MD, USA
[2]Lawrence Berkeley National Laboratory, Berkeley, CA, USA
[3]Atmospheric and Environmental Research, Inc., Lexington, MA, USA
[4]Global Monitoring Laboratory, National Oceanic and Atmospheric Administration, Boulder, CO, USA

Correspondence: Scot M. Miller (smill191@jhu.edu)

Abstract. The number of greenhouse gas (GHG) observing satellites has greatly expanded in recent years, and these new datasets provide an unprecedented constraint on global GHG sources and sinks. However, a continuing challenge for inverse models that are used to estimate these sources and sinks is the sheer number of satellite observations, sometimes in the millions per day. These massive datasets often make it prohibitive to implement inverse modeling calculations and/or assimilate the observations using many types of atmospheric models. Although these satellite datasets are very large, the information content of any single observation is often modest and non-exclusive due to redundancy with neighboring observations and due to measurement noise. In this study, we develop an adaptive approach to reduce the size of satellite datasets using geostatistics. A guiding principle is to reduce the data more in regions with little variability in the observations and less in regions with high variability. We subsequently tune and evaluate the approach using synthetic and real data case studies for North America from NASA's Orbiting Carbon Observatory-2 (OCO-2) satellite. The proposed approach to data reduction yields more accurate CO_2 flux estimates than the commonly-used method of binning and averaging the satellite data. We further develop a metric for choosing a level of data reduction; we can reduce the satellite dataset to an average of one observation per $\sim 80 - 140$ km for the specific case studies here without substantially compromising the flux estimate, but we find that reducing the data further quickly degrades the accuracy of the estimated fluxes. Overall, the approach developed here could be applied to a range of inverse problems that use very large trace gas datasets.

1 Introduction

Satellite observations of greenhouse gases (GHGs) have dramatically expanded over the past decade. New satellites with smaller footprints, wider viewing angles, and efficient scanning can collect millions of observations per day at high density and with broad spatial coverage. Remote sensing of carbon dioxide (CO_2) is a prime example. The Greenhouse Gases Observing Satellite (GOSAT), launched in early 2009, is the first satellite dedicated to observing CO_2 and methane (CH_4) from space. GOSAT collects a modest $\sim 1 \times 10^3$ cloud-free soundings or observations per day. The Orbiting Carbon Observatory 2 (OCO-2) launched five years later in late 2014 and is NASA's first satellite dedicated to observing CO_2. It collects far more cloud-free soundings than GOSAT – on the order of 1×10^5 (Crisp, 2015; Eldering et al., 2017). By contrast, NASA's forthcoming

Geostationary Carbon Observatory (GeoCarb), planned for launch in the early 2020s, is slated to collect $\sim 1 \times 10^7$ soundings each day (Buis, 2018). A substantial fraction of these soundings will be unusable due to cloud contamination, but GeoCarb can reduce contamination by scanning cloud-free regions.

These satellites observe average CO_2 mixing ratios across a vertical column of the atmosphere (XCO_2), and these XCO_2 measurements can be used to estimate surface CO_2 fluxes using inverse modeling. Specifically, an inverse model will combine satellite observations (z, dimensions $n \times 1$) with an estimate of atmospheric transport ($\mathbf{H}$, $n \times m$) to estimate surface fluxes (s, $m \times 1$):

$$z = \mathbf{H}s + \epsilon, \tag{1}$$

where ϵ ($n \times 1$) is a vector of errors in the measurements and atmospheric modeling system. The objective of the inverse model is to estimate s given z and $\mathbf{H}$. Most existing inverse models also require an estimate of the statistical properties of ϵ to ensure that the solution does not under- or over-fit the atmospheric observations (z). There are many different strategies for estimating the fluxes (s); most existing studies implement an inverse model that is based upon Bayesian statistics. Refer to Rodgers (2000), Michalak et al. (2004), and Brasseur and Jacob (2017) for an overview of commonly-used strategies for inverse modeling.

Large satellite datasets often pose computational problems for inverse modeling, specifically for calculations that involve the atmospheric model, $\mathbf{H}$. The associated challenges vary depending upon the the type of atmospheric model. For example, one approach to inverse modeling is to estimate $\mathbf{H}$ using a Lagrangian model, also known as a back-trajectory model. These models estimate how fluxes or emissions from different regions would impact mixing ratios at a downwind observation site. Commonly-used models include the Stochastic Time-Inverted Lagrangian Transport (STILT) model (Lin et al., 2003; Nehrkorn et al., 2010), the FLEXible PARTicle dispersion model (FLEXPART) (Pisso et al., 2019), and the Hybrid Single-Particle Lagrangian Integrated Trajectory model (HYSPLIT) (Stein et al., 2016). One must run simulations of the back-trajectory model for each of n observations used in the inverse model, and each simulation becomes a different row of $\mathbf{H}$. As a result of this setup, the computational cost of the back-trajectory model scales with the number of observations, the size of the modeling domain, and the resolution of the model. This approach is commonly employed for ground- and aircraft-based atmospheric observations but can quickly become computationally challenging for large satellite datasets (e.g., Wu et al., 2018).

Another common approach to atmospheric modeling is to use the a gridded atmospheric model, also known as an Eulerian model. Common models include the Goddard Earth Observing System – Chemistry model (GEOS-Chem) (e.g., Henze et al., 2007; Liu et al., 2017) and TM5 (e.g., Krol et al., 2005; Bergamaschi et al., 2005). These models are not typically used to explicitly calculate $\mathbf{H}$. Rather, these models are often used to calculate the product of $\mathbf{H}$ or $\mathbf{H}^T$ and a vector~~(e. g., a vector of estimated CO_2 fluxes).~~ One can then estimate s by iterating toward the minimum of an objective function (e.g., Eq. S1) using a series of matrix-vector products that involve $\mathbf{H}$ and $\mathbf{H}^T$ (e.g., Brasseur and Jacob, 2017). Large satellite datasets can also create computational challenges for inverse models that follow this approach. The model output must be interpolated to the locations and times of the observations ~~, often referred to as the observation operator. The observation operator must be~~ and multiplied by satellite parameters like the pressure weighting function and averaging kernel. These calculations are often

applied repeatedly during the course of ~~these~~ iterative inverse modeling algorithms, and the computational cost of ~~the operator~~ these calculations will increase with more observations. In addition, file input/output (I/O) can be a bottleneck for some types
of atmospheric models, and this cost increases as the number of observations increase. The GEOS-Chem model provides an illustrative example. In test simulations, the current GEOS-Chem forward/adjoint models for CO_2 (v35, at the time of writing) required $\sim$30 days of wall clock time to calculate the objective function and its gradient (i.e., first derivative) using a year of OCO-2 observations from the "lite" file (i.e., using 3.1×10^7 total observations for year 2016; computed on the Maryland Advanced Research Computing Center cluster using a global model spatial resolution of $2°$ latitude by $2.5°$ longitude). Most iterative inverse modeling algorithms require calculating this ~~cost~~ objective function and its gradient multiple times – at each iteration of the algorithm. By contrast, these same calculations required $\sim$0.5 days of wall clock time using 10-second averages of the OCO-2 observations (9×10^4 total observations).

The most common solution to date for these computational problems is to reduce the size of the satellite dataset. One approach is to bin and average the data across a set interval and/or run the atmospheric model at a set interval along the satellite flight track. For example, recent inverse models for OCO-2 use data that has been binned and averaged every 10-seconds along the satellite flight track (e.g., Crowell et al., 2019). This approach yields approximately one observation per 70 km, far fewer observations than the original OCO-2 dataset.

Relatedly, scientists that use a back-trajectory model for atmospheric simulations will often run the model at a set interval along the flight track. For example, scientists at NOAA have generated trajectory simulations for OCO-2 data over North America using STILT as part of the CarbonTracker-Lagrange project (e.g., NOAA Global Monitoring Laboratory, 2020a; Miller et al., 2020). These runs have been generated for a single location every 2-seconds along each satellite flight track, thereby reducing the number of model simulations required. This 2-second interval yields just under one simulation per 10 km, the spatial resolution of the meteorology fields used in the trajectory model simulations. The total computational cost of these simulations is substantial, even using a dataset reduced to 2-second intervals; each STILT simulation (i.e., for a single observation location) requires $\sim$5 hours of computing time on a single core of the the NASA Pleiades supercomputer. NOAA scientists generated $\sim 9.88 \times 10^4$ STILT simulations for year 2015.

These existing strategies for data reduction present several challenges. First, one must decide how frequently to average the data (i.e., across how many seconds or kilometers) or how frequently to generate atmospheric simulations along a satellite flight track. It is not alway practical to re-run the atmospheric model and inverse model with different levels of data reduction to decide on an optimal approach – due to the computing time involved. Instead, this decision is often based upon the spatial resolution of the atmospheric transport model (e.g., NOAA Global Monitoring Laboratory, 2020a) or the anticipated spatial resolution of the flux estimate (e.g., Crowell et al., 2019). Second, one level of averaging or data reduction (e.g., 2-second versus 10-second) may work better for one inverse modeling setup or one satellite dataset than another. Lastly, satellite observations are typically non-stationary: they exhibit different spatial and temporal variability in different locations and in different seasons (e.g., Katzfuss and Cressie, 2011; Hammerling et al., 2012a). These differences may be important to account for when reducing the size of the satellite dataset. For example, OCO-2 observations collected over the remote ocean have a lower variance and are correlated across longer distances than observations collected over terrestrial regions with heterogeneous surface sources

and sinks (e.g., Eldering et al., 2017). A one-size-fits-all approach to data reduction may not be ideal in this circumstance. Instead, it may be advantageous to reduce the size of the dataset more in regions with little variability and less in regions with greater variability.

Scientists in other academic disciplines have also grappled with many of these challenges, albeit in the context of very different scientific applications. For example, data reduction has become common in computer graphics and data visualization because many remote sensing and/or medical images are too large to render and display at native resolution (e.g., Li et al., 2018). Numerous studies reduce the size of the image through a process known as mesh reduction; these algorithms reduce the mesh more in regions of the image with little variability and less in locations with high variability (e.g., Schroeder et al., 1992; Garland and Heckbert, 1997; Brodsky and Watson, 2000; Li et al., 2018). The algorithms are therefore also adaptable to different images.

Data reduction has also become common in weather data assimilation, where the reduced datasets are typically referred to as "superobservations" or "superobs". In most existing meteorology studies, the data is divided into different grid boxes and averaged, analogous to the approach used in recent GHG studies (e.g., Lorenc, 1981; Miyoshi and Kunii, 2012). More recently, however, several studies have proposed "adaptive" or "intelligent" approaches to data reduction (e.g., Ochotta et al., 2005; Ramachandran et al., 2005; Lazarus et al., 2010; Richman et al., 2015). These studies preferentially reduce or thin the data more in regions where the observations have little variability or provide redundant information. Existing studies have used different algorithms to attain this goal, including mesh reduction (Ramachandran et al., 2005), data clustering (e.g., Ochotta et al., 2005), and machine learning (Richman et al., 2015). Compared to these meteorology studies, data reduction in atmospheric inverse modeling presents unique challenges. In weather data assimilation, observations are used to directly nudge or adjust a weather model in adjacent grid boxes. In inverse modeling, by contrast, the atmospheric observations and unknown GHG fluxes are fundamentally different quantities with complex relationships determined by atmospheric winds.

In the present study, we develop an approach to data reduction for inverse modeling of GHG observations. This approach follows the principles of adaptive reduction: we reduce the XCO_2 data more in regions with little variability in the observations and less in regions with high variability. The goals of this approach are two-fold. First, improve the computational feasibility of inverse modeling using satellite data while preserving the accuracy of the estimated fluxes. Second, develop an objective means to decide on the optimal level of data reduction for a given satellite dataset and a given inverse modeling problem. We subsequently tune and evaluate this approach using several case studies from the OCO-2 satellite – case studies that use synthetic and real data and case studies from different seasons of the year. We then compare CO_2 fluxes estimated using the proposed approach against fluxes estimated using a satellite dataset that has been averaged to reduce its size. This comparison provides a lens to evaluate the costs and benefits of the proposed approach to data reduction versus the commonly-used approach of averaging the data. The approach described here is designed not only for OCO-2 but could be applied to current and future observations of CO_2 (e.g., from GeoCarb) and observations of CH_4 (e.g., from the TROPOspheric Monitoring Instrument (TROPOMI) and GeoCarb).

2   Approach to data reduction

We develop an approach to data reduction for inverse modeling that leverages tools from geostatistics. Geostatistical tools, like variogram modeling and kriging, have become widespread in spatial data analysis (e.g., Kitanidis, 1997; Wackernagel, 2003), and these tools are often straightforward to implement using software packages in R, Matlab, Python, and other scientific programming languages. Furthermore, geostatistical tools are already used throughout inverse modeling and therefore offer an appealing framework for data reduction.

The overall strategy developed here is to first characterize the spatial properties of the observations using variogram analysis and second, use kriging to interpolate the satellite observations to a number of locations that is smaller than the original dataset. The choice of locations is informed by the variogram analysis: we retain fewer locations in regions where the observations are correlated over longer distances and more locations in regions with a shorter decorrelation length.

2.1   Step 1: Evaluate the spatial properties of the satellite data

We estimate the degree of spatial correlation in the satellite observations using a variogram analysis (e.g., Kitanidis, 1997). This analysis yields an estimate of the decorrelation length – the distance at which the correlation between any two observations is effectively zero.

In this study, we estimate the decorrelation length by creating a variogram of the satellite observations (Fig. S6). A variogram is a geostatistical tool that is used to quantify the differences among observations as a function of distance. The variogram of the observations is known as an empirical variogram, and we then fit a model to this empirical variogram using a least squares fit to estimate the decorrelation length (e.g., Kitanidis, 1997; Wackernagel, 2003). There are many possible choices for a variogram model, and we choose an exponential model with a nugget because it has been used in several existing studies of satellite-based XCO_2 observations (e.g., Hammerling et al., 2012a; Zeng et al., 2014; Guo et al., 2015; Tadić et al., 2015, 2017). The covariances between observations in this model decay exponentially as a function of distance. Furthermore, the nugget component of the variogram model accounts for fine-scale variability and errors in the observations – specifically errors that are spatially uncorrelated. Note that the exponential model yields an estimate of the e-folding distance, the distance at which covariances decay by a factor of e. In this study, we report the decorrelation length (l) or three times the e-folding distance; this is the distance at which the covariances effectively decay to zero. Refer to Kitanidis (1997) or Wackernagel (2003) for a review of different variogram models and model fitting.

We specifically estimate the decorrelation length along individual satellite flight tracks and estimate different lengths at different locations along each track. The spatial properties of the satellite observations often differ in different regions of the globe, and these differences are important to account for. In this particular study, we include all observations within 2000 km when making the estimate at each location along a flight track (as in Hammerling et al., 2012a, b). Note that we do not quantify correlations or covariances among different flight tracks or different days for the case studies using OCO-2 (Sect. 3); that satellite has a narrow swath of ~ 10 km and a 16 day revisit time, so the individual flight tracks on a given day or week are spaced relatively far apart (e.g., Crisp, 2015; Eldering et al., 2017). For new and forthcoming satellites with a wider swath

and/or more frequent revisit time, one could quantify zonal, meridional, and/or temporal decorrelation lengths, depending upon the characteristics of the satellite in question.

2.2 Step 2: Reduce the data using kriging

We subsequently reduce the satellite dataset by estimating atmospheric CO_2 at one location per fraction of a correlation length along each satellite flight track. For example, a modeler could reduce the dataset to one observation per $0.1l$ or $1.0l$. The latter choice would reduce the size of the dataset to a much greater degree but increase the risk of losing information that would ultimately inform the inverse model. Sect. 2.3 discusses strategies for deciding on an optimal level of data reduction.

The correlation length will differ in different locations, and this procedure will therefore yield a different density of points in different regions. For example, the proposed approach will result in a greater density of points in regions where XCO_2 varies across small spatial scales and a lower density of points in regions where XCO_2 is correlated across long distances.

This approach is conceptually similar to several adaptive strategies for data reduction in other scientific fields. Many existing studies either remove, merge, or cluster data points based on spatial variability. In computer visualization, mesh reduction studies merge or remove vertices from the image based upon the curvature or flatness of the original image, and different studies use various metrics to quantify this curvature and flatness (e.g., Schroeder et al., 1992; Garland and Heckbert, 1997; Brodsky and Watson, 2000; Li et al., 2018). Studies in meteorology use similar algorithms. For example, Ochotta et al. (2005) developed a metric to cluster observations based upon both the squared distance between observations and the squared difference in the observation values. Similarly, Ramachandran et al. (2005) reduced the data based upon the variance of the data in each locale. In this particular study, we use the decorrelation length, a common tool in geostatistics, to quantify the variability of the original data and guide the data reduction.

At each chosen location along the flight track, we subsequently interpolate the observations using ordinary kriging (e.g., Kitanidis, 1997). Numerous existing studies have applied various forms of kriging to interpolate satellite observations of CO_2 (e.g., Katzfuss and Cressie, 2011; Hammerling et al., 2012a, b; Tadić et al., 2015). Kriging accounts for the spatial and/or temporal properties of the quantity of interest, yielding a more accurate estimate. Kriging also yields an estimate of uncertainty in the estimated quantity, in this case uncertainties in the reduced XCO_2 dataset. These uncertainties account for the variability (or lack therefore) of the OCO-2 data in the vicinity of each location, the density or sparsity of the original OCO-2 dataset near each location, and random noise in the OCO-2 observations, among other error sources. Both the best estimate of XCO_2 and the corresponding uncertainties can be calculated using a simple linear system of equations.

We specifically implement ordinary kriging using a moving neighborhood (e.g., Kitanidis, 1997; Hammerling et al., 2012a, b); the quantity of interest is only estimated at a single location or a subset of locations at one time using nearby observations. This approach contrasts with other variants of kriging that incorporate all observations to estimate all unknown locations simultaneously. Ordinary kriging with a moving neighborhood is particularly useful when the observations are non-stationary and exhibit different spatial properties and/or error characteristics in different regions, as is often the case with XCO_2 (e.g., Hammerling et al., 2012a).

Ordinary kriging with a moving neighborhood requires two steps. First, a modeler must estimate the spatial properties of the observations in the vicinity of the estimation location. We estimate these properties as part of the analysis in Sect. 2.1 and use that estimate as an input in ordinary kriging. Second, we estimate XCO_2 at each location of interest by solving a system of linear equations. Kitanidis (1997) describes this approach in detail, and Hammerling et al. (2012a), Hammerling et al. (2012b) describe the application of moving neighborhood kriging to observations of XCO_2. The XCO_2 estimates from kriging can then be incorporated as observations in inverse modeling.

Note that traditional kriging models are designed to interpolate the quantity of interest to the same spatial support as the observations. In other words, the footprint size of the kriging estimate will be the same as that of the observations. This setup works well for the case study presented here; the atmospheric simulations in this study are generated using a back-trajectory model, and each simulation corresponds to a specific point location and time along an OCO-2 flight track. By contrast, a variant of kriging known as block kriging can be used to estimate a representative or average value for an entire grid box (e.g., Wackernagel, 2003; Tadić et al., 2015, 2017). This approach may be desirable when generating atmospheric simulations using an Eulerian model where the outputs represent grid averages. Tadić et al. (2015) and Tadić et al. (2017) describe this approach in detail, including applications to interpolating XCO_2.

2.3 Step 3: Decide on an optimal level of data reduction

A modeler must decide on an optimal level of data reduction. That decision is often based on multiple considerations – the native spatial resolution of the atmospheric model, the computational demands of the inverse model, and the accuracy of the resulting flux estimate. For example, the resolution of the atmospheric model may help dictate a level of data reduction. An atmospheric model will not be able to resolve patterns in the fluxes at spatial scales smaller than the model resolution, so there may be little need to assimilate CO_2 observations at finer density than the model resolution.

For the specific algorithm described here, one must decide on a fraction of a correlation length and reduce the dataset accordingly. A modeler could decide on the optimal level of data reduction using a brute force approach: create numerous datasets with different levels of data reduction, run the inverse model on each, and decide on a level of data reduction based upon a comparison of the estimated fluxes. In practice, this approach is time consuming.

Instead, we propose a criterion for choosing a level of data reduction based upon the variance of the satellite data. We first select all data points in the original CO_2 dataset that fall between two specific kriged points. We then calculate the variance of those selected points using the *var()* function in R. We repeat this procedure for each pair of kriged points in the model domain. We finally average these variances calculated across each pair of kriged points. Some of this variance will undoubtedly be due to measurement error, but some of this variance will likely be due to real variability in atmospheric CO_2.

This variance represents the variability in the data that is lost through the process of data reduction, and it provides a metric for choosing a level of data reduction that does not require re-running the inverse model. This number is smallest when the data reduction is minimal and increases for greater levels of data reduction. For the case studies in Sect. 3, this variance is often a non-linear function of the level of data reduction; it increases slowly if the data reduction is minimal, reaches an inflection point, and then increases more quickly at greater levels of data reduction. A modeler can then choose a level of data reduction that is

preferably below the inflection point and therefore reduces the potential for information loss while balancing the computational requirements of the inverse model.

We evaluate this proposed approach for deciding on a level of data reduction through several case studies based upon the OCO-2 satellite, described in detail in the next section (Sect. 3).

3 Description of the case studies

We evaluate the data reduction algorithm using three case studies based on the OCO-2 satellite. In each case, we estimate CO_2 fluxes across North America for six weeks at a 3-hourly temporal resolution and a $1° \times 1°$ latitude-longitude spatial resolution. Note that this setup targets a particular application of OCO-2 observations to inverse modeling across a continent. One could apply data reduction to inverse models that target urban areas or the entire globe, but the algorithm tuning (e.g., Sect. 2.3) and inverse modeling results will depend upon the particular application involved.

We specifically estimate fluxes using synthetic observations from July and early August 2015, using synthetic observations from March and early April 2015, and using real observations from July and early August 2015. Synthetic observations make it possible to compare the results against a known solution; they are therefore particularly useful for the evaluating the data reduction algorithm proposed here. We further evaluate the algorithm in a real data simulation that mirrors real-world inverse modeling applications. Note that we present the details of the summer real data case study in the SI and focus on the synthetic case studies in the main text. The results of the real data case study are qualitatively very similar to the synthetic case studies, so we include that information in the SI to avoid duplicating similar information in the main text.

We further estimate the CO_2 fluxes using geostatistical inverse model (e.g., Kitanidis and Vomvoris, 1983; Michalak et al., 2004; Miller et al., 2020). The inverse model used here also has a non-informative prior. In other words, the prior has no spatial variability (e.g., Michalak et al., 2004; Mueller et al., 2008). As a result, any patterns in the estimated fluxes reflect the information content of the observations, not any prior information. This setup is identical to the case studies in Miller et al. (2020), and the reader is referred to both the SI and that study for additional detail.

The case studies here also use atmospheric transport simulations from NOAA's CarbonTracker-Lagrange project (e.g., Hu et al., 2019; NOAA Global Monitoring Laboratory, 2020a). These simulations are generated using the Weather Research and Forecasting (WRF) model coupled with the STILT model (e.g., Lin et al., 2003; Nehrkorn et al., 2010). The simulations have a spatial resolution of 10 km over most of the Continental US and a resolution of 30 km across other regions of North America. Miller et al. (2020) provides additional detail on the specific setup of the WRF-STILT runs used here. Note that the STILT simulations for CarbonTracker-Lagrange were generated every two seconds along the OCO-2 flight track and not at every individual OCO-2 observation due to the large number of observations and due to computational constraints. Hence, we only evaluate data reduction that yields fewer than one observation every two seconds for the case studies here.

We further create the synthetic data for each case study using WRF-STILT and CO_2 fluxes from NOAA's CarbonTracker (CT2017) product (Peters et al., 2007; NOAA Global Monitoring Laboratory, 2020b). The synthetic CO_2 fluxes not only include biospheric fluxes but also anthropogenic and biomass burning emissions. The synthetic observations also include noise

(ϵ) that is added to simulate measurement and atmospheric modeling errors. For the summer case studies here, these errors have a variance of $(2\text{ ppm})^2$ (as in Miller et al., 2020). We also include error covariances to account for spatial correlation among these errors. We use decorrelation length from Kulawik et al. (2019), who quantified errors in OCO-2 observations and estimated a decorrelation parameter of $0.3°$ using an exponential variogram model. In the winter case study, we use a slightly smaller error variance of $(1.5\text{ ppm})^2$ because there is less regional variability in atmospheric CO_2 in winter. Note that we only include land nadir and land glint observations in the case studies and exclude ocean glint observations because those observations have known biases (O'Dell et al., 2018).

4 Results and discussion

4.1 Spatial properties of the OCO-2 observations

We estimate correlation lengths that are generally longer in winter when biospheric fluxes are small than in summer when there are large spatial and temporal variations in biospheric fluxes. Figure 1 displays the estimated correlation lengths along the OCO-2 flight tracks for the summer (a) and winter (b) case studies. Most of the estimated correlation lengths range from $\sim$250 km to 1000 km. Note that there are likely multiple different scales of variability in the OCO-2 observations: fine-scale variability due to retrieval errors (e.g., Kulawik et al., 2019; O'Dell et al., 2018), small-scale variability due to variations in mesoscale meteorology (e.g., Torres et al., 2019), and broad variability due to synoptic meteorology and regional patterns in CO_2 fluxes. We specifically focus on quantifying synoptic scale variability in Fig. 1 – because the objective of the case studies is to estimate broad, regional patterns in CO_2 fluxes across an entire continent.

The analysis in Fig. 1 also indicates substantial heterogeneity in the correlation lengths. Correlation lengths are often similar along a single flight track but vary among different tracks. These differences between flight tracks are most likely due to a combination of variations in synoptic meteorology and variability in the underlying CO_2 fluxes. Indeed, several studies have shown that meteorological variability can explain a substantial fraction of variability in XCO_2 across different spatial scales (e.g., Parazoo et al., 2008; Keppel-Aleks et al., 2011; Torres et al., 2019).

Two flight tracks that cross California, Oregon, and Washington illustrate the likely impacts of fluxes and meteorology on heterogeneity in the synthetic OCO-2 observations. One track, on July 16, 2015, exhibits relatively little variability in XCO_2, and we estimate an average correlation length of 794 km along the track with a standard deviation of 266 km. By contrast, another nearby track from July 21 exhibits far more XCO_2 variability, and we estimate a shorter mean correlation length of 221 km with a standard deviation of 53 km along the track. These large differences likely reflect differences in the underlying CO_2 fluxes and in meteorology on the respective days. The July 16 track passes through eastern California, Nevada, Eastern Oregon, and Eastern Washington, desert regions with little heterogeneity in CO_2 fluxes. By contrast, the track from July 21 passes over the Sierra Nevada mountains and over multiple heterogeneous biome types (e.g., desert and temperate rainforest). Furthermore, weather maps indicate that a cold front passed through the Pacific Northwest on July 21 with variable winds on either side of the front (NOAA National Centers for Environmental Prediction Weather Prediction Center, 2020). These differences in transport and surface fluxes likely yield very different patterns of variability in two satellite flight tracks that are geographically close to

one another. These results also imply that it is important to calculate correlation lengths for each individual flight track and not apply correlation length estimates from one track to another track or from one month/year to another month/year. Meteorology can easily vary from one day to another. Hence, we advise against pre-computing the correlation lengths for an individual track or an individual year and applying them to other tracks or other years.

4.2   Estimated CO_2 fluxes using the reduced OCO-2 dataset

The data reduction approach proposed here yields flux estimates that faithfully reproduce patterns in the synthetic CO_2 observations. With that said, data reduction is always a compromise between the accuracy of the flux estimate and the computational requirements of the inverse model. As such, the accuracy of the flux estimate begins to degrade at high levels of data reduction. Figures 3 and 4 summarizes many of these features and displays maps of the time-averaged fluxes from the summer and winter case studies, respectively. The first panel (a) in each figure contains the CarbonTracker fluxes that were used to generate the synthetic OCO-2 observations. The second panel (b) shows the fluxes estimated using the full, synthetic OCO-2 dataset with no reduction; the estimated CO_2 fluxes shown in these panels do not have the same level of spatial definition as the original CarbonTracker fluxes (panel a), but the estimates broadly reproduce the spatial patterns in CarbonTracker. The inverse model in this study uses a non-informative prior, so any patterns in panel (b) are solely informed by the observations and not the result of prior flux information. The patterns in panel (b) indicate that the synthetic OCO-2 observations can be used to recover continental-scale spatial features in the fluxes, but the observations and inverse model do not have the sensitivity or information content to recover more detailed features. Subsequent panels (c) display the fluxes estimated using observations that have been reduced to a modest level – one observation per $0.2l$ or an average of one observation per 100 km for the summer case study and 140 km for the winter case study. The final panel in each figure displays a severe level of reduction – one observation per $0.75l$, an average of one observation per 400 km for the summer case study and 540 km for the summer case study. In both the summer and winter case studies, the modest level of data reduction (one observation per $0.2l$) reduces the total number of observations by $\sim 70\%$ and the severe data reduction (one observation per $0.75l$) by $\sim 90\%$. A $70+\%$ data reduction and corresponding reduction in the number of STILT simulations would yield substantial computational savings, given the large computational cost required associated with STILT (Sect. 1).

In each case, the fluxes using the $0.2l$ data (c) reproduce spatial patterns in the fluxes estimated with no reduction (b). By contrast, the fluxes estimated using the $0.75l$ data lack spatial definition, and are therefore not an ideal estimate of the synthetic fluxes (a). Note that we also conducted a real data case study for summer 2015. Those results have broadly similar characteristics to the synthetic data case study and are discussed in detail in the SI.

In both of the case studies, the data reduction approach proposed here yields more accurate flux estimates than binning and averaging the observations. We reduce the data using both averaging and the geostatistical approach proposed in this study. We subsequently estimate fluxes using the reduced datasets and compare the results against the fluxes estimated using the full dataset without any data reduction. Figure 5 displays the results of this analysis – the root mean squared error (RMSE) of the grid-scale, 3-hourly estimated fluxes relative to fluxes calculated from the full dataset. In both the winter and summer case studies and at almost all levels of data reduction, the geostatistical approach produces fluxes with a lower RMSE.

Note that all results in Fig. 5 display a clear inflection point: the RMSE is relatively low at low levels of data reduction and increases rapidly at high levels of reduction. The chosen level of data reduction should be at or below this inflection point, or the inverse model will yield an inaccurate flux estimate. The inflection point for the geostatistical approach occurs at a higher level of data reduction than for data averaging. In other words, the RMSE for the geostatistical approach remains low at a greater degree of data reduction than for averaging.

The RMSE, however, may not be the only criteria to consider when deciding on a level of data reduction. Specifically, the spatial patterns in the fluxes begin to degrade at a lower level of data reduction than the RMSE. For example, in both the winter, summer, and real data case studies, the monthly-averaged fluxes begin to lose spatial definition at data reduction levels greater than $0.15l$ to $0.2l$ (an average of one observation per 80-100 km for the summer case study and one observation per 100-140 km for the winter case study). Hence, it may be advisable to balance multiple criteria when deciding on an optimal level of data reduction, depending upon the goals of the inverse modeling study.

4.3 Determining an optimal level of data reduction

In the previous section (Sect. 4.2), we evaluate the data reduction by comparing the resulting estimates of CO_2 fluxes. However, this approach may not work well if the inverse model is time-consuming and/or computationally intensive, as is often the case for satellite-based inverse modeling. For example, in Fig. 5 we run the inverse model twenty times for each case study to estimate the fluxes using different levels of data reduction and evaluate the results against fluxes estimated using the original, synthetic OCO-2 data. An inverse model can take days to run using large satellite datasets, so it may not be feasible or desirable to run the inverse model numerous times. Furthermore, one may want to decide on an optimal level of data reduction before running back-trajectory simulations using a model like STILT. We therefore propose an approach to evaluate the data reduction in a way that does not require running the inverse model or an atmospheric model like STILT (Sect. 2.3). This approach is based upon the variance of the original data between each of the reduced data points; this variance is a measure of the variability in the data that is lost through the process of data reduction.

In each of the OCO-2 case studies, this metric provides a reasonable and informative means to decide on an appropriate level of data reduction. Figure 6 displays this metric calculated for the summer (a) and winter (b) case studies at multiple different levels of data reduction. The variance lost through the process of data reduction is lowest at small levels of data reduction and increases non-linearly at higher levels of data reduction. The summer case study (a) is highly non-linear and reaches a very clear inflection point. By contrast, the winter case study (b) does not have as clear of an inflection point, but the variance does increase more quickly at higher levels of data reduction.

This metric also mirrors many of the patterns in the flux maps (Figs. 3 and 4) and RMSE calculations (Fig. 5). For example, fluxes in the summer case study loose spatial definition at data reduction levels greater than $0.2l$ (equivalent to 2263 data points). Indeed, the variances in Fig. 6a begin to increase more rapidly after at data reduction levels greater than $0.2l$. By contrast, the fluxes in the winter case study progressively lose spatial definition, but that loss is particularly notable at high levels of reduction. That pattern is similar to the pattern in the variances in Fig. 6b. Furthermore, the patterns in Fig. 6 also mirror many of the patterns in the RMSE (Fig. 5). Both the RMSE and variances for the summer case study reach an inflection

point between 2000 and 1000 observations, at which point both begin to increase rapidly. The RMSE and variance plots for the winter case study do not look identical (Figs. 5b and 6b). With that said, the pattern in the winter case study looks qualitatively more akin to the degradation of spatial patterns in the plotted fluxes than it does to the RMSE in Fig. 5b.

At the end of the day, it is arguably difficult to identify a single metric for deciding on an appropriate level of data reduction, and the right metric may depend upon the goal of the inverse model (e.g., identifying spatial patterns, temporal patterns, and/or flux totals). With that said, the metric proposed in this section is a computationally efficient option that summarizes many of the features of data reduction described in the previous sections (e.g., Sect. 4.2).

4.4 Computational costs

We find that the computational costs of the data reduction algorithm are reasonable for the case studies explored here and are far less than the computing time associated with generating atmospheric model simulations. Both the variogram analysis (Sect. 2.1) and kriging (Sect. 2.2) are implemented using a moving neighborhood, thereby limiting the number of observations included in any given variogram or kriging calculation and reducing computing time. For example, we use a moving neighborhood with a radius of 2000 km for the case studies (as in Hammerling et al. (2012a)), approximately half the width of the continental United States. Each empirical variogram required an average of 0.05 s to calculate using the R programming language, and each kriging estimate required an average of 0.02 s. By contrast, a single STILT model simulation corresponding to a single OCO-2 observation required far more computing time (see Sect. 1).

Furthermore, one can distribute the variogram and kriging calculations across multiple computing cores and nodes, reducing the required computing time. Specifically, the calculation of each decorrelation length and each kriging estimate is independent of every other calculation or location, so these individual calculations can be spread across as many cores or nodes as desired. There is also some flexibility in the implementation of this algorithm and therefore in its computational cost. We calculate the variogram every two seconds along the OCO-2 flight track, but one could calculate the variogram at less frequent intervals. For example, Hammerling et al. (2012a) implement moving neighborhood kriging for synthetic OCO-2 observations and calculated variogram parameters for each location on a 1° latitude by 1.25° grid. One can also define the moving neighborhood differently with computational considerations in mind. For example, Tadić et al. (2015) and Tadić et al. (2017) limit the moving neighborhood to 500 XCO_2 observations. Instead of including all observations within a given radius, they choose observations to include in the moving neighborhood using a randomized algorithm, and the algorithm preferentially chooses observations that are closer to the estimation location over observations that are far away. That strategy yields accurate variogram parameters and kriging estimates while ensuring that the number of observations within a moving neighborhood is not so large as to pose a computational burden.

5 Conclusions

In many instances, new satellite datasets are simply too large to assimilate in an inverse model given the current computational limitations of existing atmospheric models. In an ideal world, it would be possible to assimilate all available GHG observations

to exploit the full information content of these massive new satellite datasets. However, that ideal is not computationally feasible in many instances, and modelers often need a strategy to reduce the size of these datasets. At minimum, this strategy should reduce the computational demands of inverse modeling while yielding flux estimates that accurately reproduce key information on the magnitude and distribution of surface fluxes. A complicating factor is that satellite observations often exhibit very different variability in different regions and/or on different days – depending on factors like regional variability in GHG fluxes and variations in meteorology. In this work, we argue that a data reduction strategy should account for this variability and that doing so typically yields a more accurate flux estimate.

One could develop a strategy for data reduction using many different statistical and mathematical tools, and we specifically develop a strategy using geostatistics because it provides a convenient way to quantify and account for the spatial variability of the satellite observations. In the case studies presented here based on NASA's OCO-2 satellite, this strategy outperforms data averaging, a common and straightforward approach to data reduction but one that does not account for the variable spatial properties of the observations. The specific implementation of this strategy will likely vary depending upon the satellite dataset in question and the specifics of the atmospheric model. To that end, we also develop and evaluate a computationally efficient metric to help choose an appropriate level of data reduction – a metric that does not require re-running the inverse model numerous times.

Future computational improvements to atmospheric models and increased access to high performance computing resources will hopefully make it possible to implement inverse modeling with larger and larger atmospheric datasets while minimizing the need for data reduction. With that said, forthcoming satellites like NASA's GeoCarb mission promise to collect unprecedented numbers of atmospheric GHG observations, and these new missions may make data reduction more necessary than ever.

Code and data availability. The code used in this study for data reduction is available on Github at https://doi.org/10.5281/zenodo.3899307. The inverse modeling code used in the case studies is also available at https://doi.org/10.5281/zenodo.3241524, and the model simulations used to construct the summer case study are available on Zenodo at https://doi.org/10.5281/zenodo.3241466.

Author contributions. X.L. A.L.W. J.T. and S.M.M. designed the study. X.L. A.L.W. H.C., Z.C. and S.M.M. conducted the analysis. M.M., M.T. and A.A. developed the strategy for simulating OCO-2 soundings using WRF-STILT and provided the footprints. S.M.M. drafted the manuscript, and all authors helped edit the manuscript.

Competing interests. The authors declare they have no competing interests.

Acknowledgements. We thank Thomas Nehrkorn (Atmospheric and Environmental Research (AER), Inc.) for help with the atmospheric modeling simulations used in this study. We also thank Amy Braverman (NASA) for her advice and input on the project.

420 *Financial support.* This work is supported by NASA ROSES grant no. 80NSSC18K0976. CarbonTracker-Lagrange footprint production was supported by NASA the NASA Carbon Monitoring System via interagency agreement NNH14AY37I.

References

Bergamaschi, P., Krol, M., Dentener, F., Vermeulen, A., Meinhardt, F., Graul, R., Ramonet, M., Peters, W., and Dlugokencky, E. J.: Inverse modelling of national and European CH_4 emissions using the atmospheric zoom model TM5, Atmos. Chem. Phys., 5, 2431–2460, https://doi.org/10.5194/acp-5-2431-2005, https://www.atmos-chem-phys.net/5/2431/2005/, 2005.

Brasseur, G. and Jacob, D.: Modeling of Atmospheric Chemistry, Cambridge University Press, Cambridge, 2017.

Brodsky, D. and Watson, B.: Model simplification through refinement, in: Graphics Interface, vol. 2000, pp. 221–228, 2000.

Buis, A.: GeoCarb: A New View of Carbon Over the Americas, ExploreEarth, https://www.nasa.gov/feature/jpl/geocarb-a-new-view-of-carbon-over-the-americas, 2018, last access: 17 Jul 2020.

Crisp, D.: Measuring atmospheric carbon dioxide from space with the Orbiting Carbon Observatory-2 (OCO-2), in: Earth Observing Systems XX, edited by Butler, J. J., Xiong, X. J., and Gu, X., vol. 9607, pp. 1 – 7, International Society for Optics and Photonics, SPIE, https://doi.org/10.1117/12.2187291, 2015.

Crowell, S., Baker, D., Schuh, A., Basu, S., Jacobson, A. R., Chevallier, F., Liu, J., Deng, F., Feng, L., McKain, K., Chatterjee, A., Miller, J. B., Stephens, B. B., Eldering, A., Crisp, D., Schimel, D., Nassar, R., O'Dell, C. W., Oda, T., Sweeney, C., Palmer, P. I., and Jones, D. B. A.: The 2015–2016 carbon cycle as seen from OCO-2 and the global in situ network, Atmos. Chem. Phys., 19, 9797–9831, https://doi.org/10.5194/acp-19-9797-2019, 2019.

Eldering, A., Wennberg, P., Crisp, D., Schimel, D., Gunson, M., Chatterjee, A., Liu, J., Schwandner, F., Sun, Y., O'Dell, C., et al.: The Orbiting Carbon Observatory-2 early science investigations of regional carbon dioxide fluxes, Science, 358, eaam5745, https://doi.org/10.1126/science.aam5745, 2017.

Garland, M. and Heckbert, P. S.: Surface simplification using quadric error metrics, in: Proceedings of the 24th annual conference on Computer graphics and interactive techniques, pp. 209–216, 1997.

Guo, L., Lei, L., Zeng, Z., Zou, P., Liu, D., and Zhang, B.: Evaluation of Spatio-Temporal Variogram Models for Mapping XCO_2 Using Satellite Observations: A Case Study in China, IEEE J. Sel. Top. Appl., 8, 376–385, https://doi.org/10.1109/JSTARS.2014.2363019, 2015.

Hammerling, D. M., Michalak, A. M., and Kawa, S. R.: Mapping of CO_2 at high spatiotemporal resolution using satellite observations: Global distributions from OCO-2, J. Geophys. Res.-Atmos., 117, D06 306, https://doi.org/10.1029/2011JD017015, 2012a.

Hammerling, D. M., Michalak, A. M., O'Dell, C., and Kawa, S. R.: Global CO_2 distributions over land from the Greenhouse Gases Observing Satellite (GOSAT), Geophys. Res. Lett., 39, L08 804, https://doi.org/10.1029/2012GL051203, 2012b.

Henze, D. K., Hakami, A., and Seinfeld, J. H.: Development of the adjoint of GEOS-Chem, Atmos. Chem. Phys., 7, 2413–2433, https://doi.org/10.5194/acp-7-2413-2007, 2007.

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.

Katzfuss, M. and Cressie, N.: Tutorial on fixed rank kriging (FRK) of CO_2 data, Tech. Rep. 858, Department of Statistics, The Ohio State University, Columbus, Ohio, https://niasra.uow.edu.au/content/groups/public/@web/@inf/@math/documents/mm/uow175999.pdf, 2011, last access: 17 Jul 2020.

Keppel-Aleks, G., Wennberg, P. O., and Schneider, T.: Sources of variations in total column carbon dioxide, Atmos. Chem. Phys., 11, 3581–3593, https://doi.org/10.5194/acp-11-3581-2011, 2011.

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

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.

Krol, M., Houweling, S., Bregman, B., van den Broek, M., Segers, A., van Velthoven, P., Peters, W., Dentener, F., and Bergamaschi,

465 P.: The two-way nested global chemistry-transport zoom model TM5: algorithm and applications, Atmos. Chem. Phys., 5, 417–432, https://doi.org/10.5194/acp-5-417-2005, 2005.

Kulawik, S. S., Crowell, S., Baker, D., Liu, J., McKain, K., Sweeney, C., Biraud, S. C., Wofsy, S., O'Dell, C. W., Wennberg, P. O., Wunch, D., Roehl, C. M., Deutscher, N. M., Kiel, M., Griffith, D. W. T., Velazco, V. A., Notholt, J., Warneke, T., Petri, C., De Mazière, M., Sha, M. K., Sussmann, R., Rettinger, M., Pollard, D. F., Morino, I., Uchino, O., Hase, F., Feist, D. G., Roche, S., Strong, K., Kivi, R., Iraci,

470 L., Shiomi, K., Dubey, M. K., Sepulveda, E., Rodriguez, O. E. G., Té, Y., Jeseck, P., Heikkinen, P., Dlugokencky, E. J., Gunson, M. R., Eldering, A., Crisp, D., Fisher, B., and Osterman, G. B.: Characterization of OCO-2 and ACOS-GOSAT biases and errors for CO_2 flux estimates, Atmos. Meas. Tech. Discussions, 2019, 1–61, https://doi.org/10.5194/amt-2019-257, 2019.

Lazarus, S. M., Splitt, M. E., Lueken, M. D., Ramachandran, R., Li, X., Movva, S., Graves, S. J., and Zavodsky, B. T.: Evaluation of Data Reduction Algorithms for Real-Time Analysis, Weather Forecast., 25, 837–851, https://doi.org/10.1175/2010WAF2222296.1, 2010.

475 Li, S., Marsaglia, N., Garth, C., Woodring, J., Clyne, J., and Childs, H.: Data Reduction Techniques for Simulation, Visualization and Data Analysis, Comput. Graph. Forum, 37, 422–447, https://doi.org/10.1111/cgf.13336, 2018.

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, 4493, https://doi.org/10.1029/2002JD003161, 2003.

480 Liu, J., Bowman, K. W., Schimel, D. S., Parazoo, N. C., Jiang, Z., Lee, M., Bloom, A. A., Wunch, D., Frankenberg, C., Sun, Y., O'Dell, C. W., Gurney, K. R., Menemenlis, D., Gierach, M., Crisp, D., and Eldering, A.: Contrasting carbon cycle responses of the tropical continents to the 2015–2016 El Niño, Science, 358, https://doi.org/10.1126/science.aam5690, 2017.

Lorenc, A. C.: A Global Three-Dimensional Multivariate Statistical Interpolation Scheme, Mon. Weather Rev., 109, 701–721, https://doi.org/10.1175/1520-0493(1981)109<0701:AGTDMS>2.0.CO;2, 1981.

485 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.

490 Miyoshi, T. and Kunii, M.: Using AIRS retrievals in the WRF-LETKF system to improve regional numerical weather prediction, Tellus A, 64, 18 408, https://doi.org/10.3402/tellusa.v64i0.18408, 2012.

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.

495 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.

NOAA Global Monitoring Laboratory: CarbonTracker – Lagrange, https://www.esrl.noaa.gov/gmd/ccgg/carbontracker-lagrange/, 2020a, last access: 17 Jul 2020.

500 NOAA Global Monitoring Laboratory: CarbonTracker, https://www.esrl.noaa.gov/gmd/ccgg/carbontracker/, 2020b, last access: 17 Jul 2020.

NOAA National Centers for Environmental Prediction Weather Prediction Center: Daily Weather Map, https://www.wpc.ncep.noaa.gov/dailywxmap/, 2020, last access: 17 Jul 2020.

Ochotta, T., Gebhardt, C., Saupe, D., and Wergen, W.: Adaptive thinning of atmospheric observations in data assimilation with vector quantization and filtering methods, Q. J. Roy. Meteor. Soc., 131, 3427–3437, https://doi.org/10.1256/qj.05.94, 2005.

505 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
510 ACOS algorithm, Atmos. Meas. Tech., 11, 6539–6576, https://doi.org/10.5194/amt-11-6539-2018, 2018.

Parazoo, N. C., Denning, A. S., Kawa, S. R., Corbin, K. D., Lokupitiya, R. S., and Baker, I. T.: Mechanisms for synoptic variations of atmospheric CO_2 in North America, South America and Europe, Atmos. Chem. Phys., 8, 7239–7254, https://doi.org/10.5194/acp-8-7239-2008, 2008.

Peters, W., Jacobson, A. R., Sweeney, C., Andrews, A. E., Conway, T. J., Masarie, K., Miller, J. B., Bruhwiler, L. M. P., Pétron,
515 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.

Pisso, I., Sollum, E., Grythe, H., Kristiansen, N. I., Cassiani, M., Eckhardt, S., Arnold, D., Morton, D., Thompson, R. L., Groot Zwaaftink, C. D., Evangeliou, N., Sodemann, H., Haimberger, L., Henne, S., Brunner, D., Burkhart, J. F., Fouilloux, A., Brioude, J., Philipp, A.,
520 Seibert, P., and Stohl, A.: The Lagrangian particle dispersion model FLEXPART version 10.4, Geosci. Model Dev., 12, 4955–4997, https://doi.org/10.5194/gmd-12-4955-2019, 2019.

Ramachandran, R., Li, X., Movva, S., Graves, S., Greco, S., Emmitt, D., Terry, J., and Atlas, R.: Intelligent data thinning algorithm for earth system numerical model research and application, in: Proc. 21st Intl. Conf. on IIPS, 2005.

Richman, M. B., Leslie, L. M., Trafalis, T. B., and Mansouri, H.: Data selection using support vector regression, Adv. Atmos. Sci., 32,
525 277–286, https://doi.org/10.1007/s00376-014-4072-9, 2015.

Rodgers, C. D.: 4, Series On Atmospheric, Oceanic And Planetary Physics, World Scientific Publishing Company, London, 2000.

Schroeder, W. J., Zarge, J. A., and Lorensen, W. E.: Decimation of triangle meshes, in: Proceedings of the 19th annual conference on Computer graphics and interactive techniques, pp. 65–70, 1992.

Stein, A. F., Draxler, R. R., Rolph, G. D., Stunder, B. J. B., Cohen, M. D., and Ngan, F.: NOAA's HYSPLIT Atmospheric Transport and
530 Dispersion Modeling System, B. Am. Meteorol. Soc., 96, 2059–2077, https://doi.org/10.1175/BAMS-D-14-00110.1, 2016.

Tadić, J. M., Qiu, X., Yadav, V., and Michalak, A. M.: Mapping of satellite Earth observations using moving window block kriging, Geosci. Model Dev., 8, 3311–3319, https://doi.org/10.5194/gmd-8-3311-2015, 2015.

Tadić, J. M., Qiu, X., Miller, S., and Michalak, A. M.: Spatio-temporal approach to moving window block kriging of satellite data v1.0, Geosci. Model Dev., 10, 709–720, https://doi.org/10.5194/gmd-10-709-2017, 2017.

535   Torres, A. D., Keppel-Aleks, G., Doney, S. C., Fendrock, M., Luis, K., De Maziere, M., Hase, F., Petri, C., Pollard, D. F., Roehl, C. M., Sussmann, R., Velazco, V. A., Warneke, T., and Wunch, D.: A Geostatistical Framework for Quantifying the Imprint of Mesoscale Atmospheric Transport on Satellite Trace Gas Retrievals, J. Geophys. Res.-Atmos., 124, 9773–9795, https://doi.org/10.1029/2018JD029933, 2019.

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

540   Wu, D., Lin, J. C., Fasoli, B., Oda, T., Ye, X., Lauvaux, T., Yang, E. G., and Kort, E. A.: A Lagrangian approach towards extracting signals of urban CO_2 emissions from satellite observations of atmospheric column CO_2 (XCO_2): X-Stochastic Time-Inverted Lagrangian Transport model ("X-STILT v1"), Geosci. Model Dev., 11, 4843–4871, https://doi.org/10.5194/gmd-11-4843-2018, 2018.

Zeng, Z., Lei, L., Hou, S., Ru, F., Guan, X., and Zhang, B.: A Regional Gap-Filling Method Based on Spatiotemporal Variogram Model of CO_2 Columns, IEEE T. Geosci. Remote., 52, 3594–3603, https://doi.org/10.1109/TGRS.2013.2273807l, 2014.
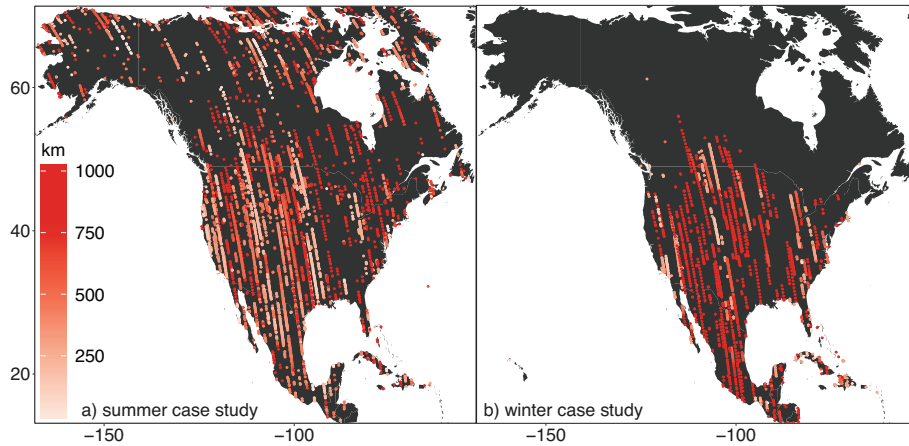
Figure 1. Correlation lengths estimated along OCO-2 flight tracks for (a) the summer synthetic case study and (b) the winter synthetic case study. The estimated correlation lengths are typically shorter in summer when biospheric fluxes exhibit high spatiotemporal variability and longer in winter when biospheric fluxes are small.
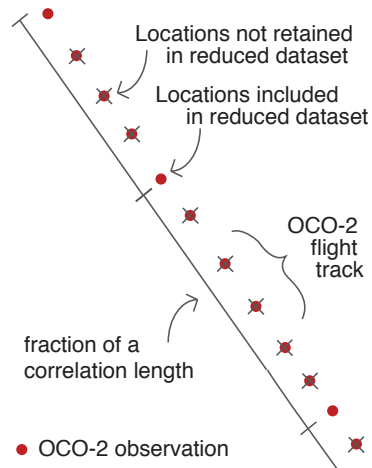
Figure 2. A schematic of the approach to data reduction proposed here. We estimate XCO_2 at one location per fraction of a correlation length along the satellite flight track, where the specific fraction must be chosen by the user. We subsequently estimate XCO_2 at each chosen location using ordinary kriging with a moving neighborhood.
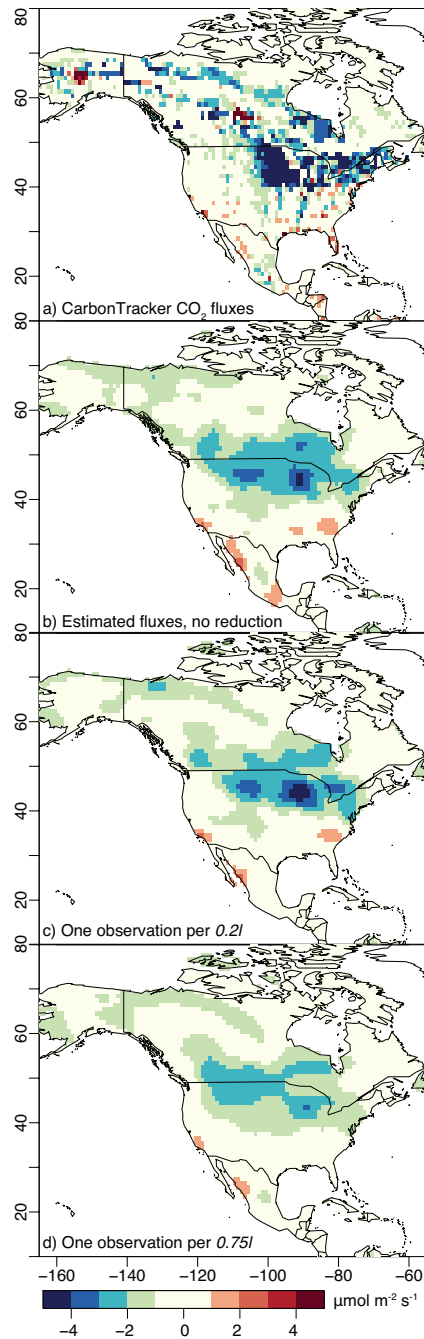
Figure 3. CO_2 fluxes estimated for the summer 2015 synthetic case study, averaged across the 6-week study window: (a) the synthetic CO_2 fluxes from NOAA's CarbonTracker estimate, (b) fluxes estimated from XCO_2 data with no reduction (6799 data points), (c) fluxes estimated from data reduced to one point per $0.2l$ (2263 data points), and (d) fluxes estimated from data reduced to one point per $0.75l$ (755 data points). The estimate with no reduction (b) and a reduction of $0.2l$ (c) reproduce broad, continental-scale spatial patterns in the synthetic fluxes (a), while the estimate with $0.75l$ reduction has lost spatial definition. Note that the inverse model here uses a non-informative prior, so any patterns in the flux estimates are informed by the observations, and not by any prior flux information.
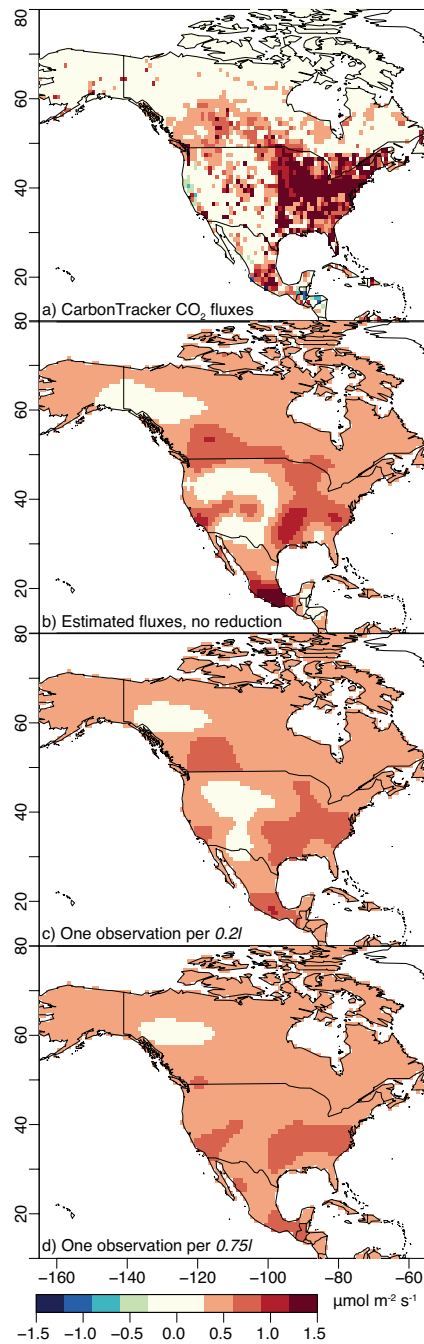
Figure 4. CO$_2$ fluxes estimated for the winter case study, analagous to Fig. 3. The fluxes estimated with modest data reduction (c, 1098 data points) reproduce the patterns in the flux estimate with no reduction (b, 4183 data points). By contrast, large data reduction (d, 251 data points) yields fluxes with little spatial variability.
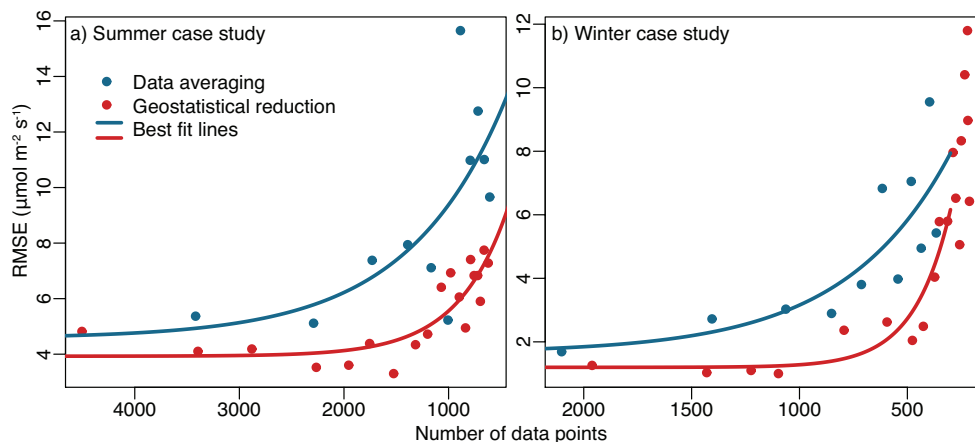
Figure 5. Root mean squared error (RMSE) of the fluxes estimated using data reduction relative to the fluxes estimated without data reduction. The figure displays results for the summer (a) and winter (b) case studies. Fluxes estimated using the data reduction approach proposed here have a lower RMSE relative to those estimated using the binning and averaging approach to data reduction.
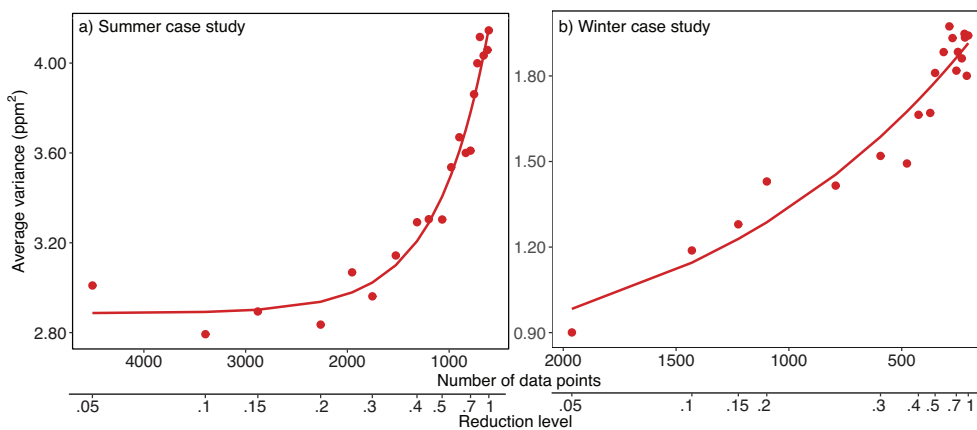
Figure 6. The amount of variance in the data that is lost through the process of data reduction for the summer (a) and winter (b) case studies. These plots provide a metric to help decide on an appropriate level of reduction and do not require costly runs of the inverse model to generate.

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 averagedand~~ _averaged_. _Then_ the pressure weighting function _and averaging kernel are_ 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 reduction 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 decorrelation 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~~ objective function (e.g., Kitanidis and Vomvoris, 1983; Michalak et al., 2004):

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

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.

We minimize Eq. S1 and estimate the fluxes ($\boldsymbol{s}$) by solving the following linear system of equations (e.g. Kitanidis, 1996; Saibaba and Kitanidis, 2012):

$$\begin{bmatrix} \mathbf{HQH}^T + \mathbf{R} & \mathbf{HX} \\ (\mathbf{HX})^T & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{\xi} \\ \boldsymbol{\beta} \end{bmatrix} = \begin{bmatrix} \boldsymbol{z} \\ \mathbf{0} \end{bmatrix} \tag{S2}$$

We solve the system in Eq. S2 to estimate the weights ($\boldsymbol{\xi}$, $n \times 1$) and coefficients ($\boldsymbol{\beta}$, $p \times 1$). The fluxes ($\boldsymbol{s}$) can then be estimated using the following equation:

$$\hat{\boldsymbol{s}} \;=\; \mathbf{X}\boldsymbol{\beta} + \mathbf{QH}^T\boldsymbol{\xi} \tag{S3}$$

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). In the synthetic case studies, we randomly generate these errors ($\boldsymbol{\epsilon}$) and add them to the synthetic observations (Sect. 3 of the main article). We use these values in $\mathbf{R}$ for the inverse modeling simulations without data reduction – $(2\ \mathrm{ppm})^2$ for the summer case study and $(1.5\ \mathrm{ppm})^2$ for the summer case study. We also use $(2\ \mathrm{ppm})^2$ for the real data case study described in Sect. S1.

For the inverse modeling simulations using the reduced datasets, we use the uncertainties computed for the kriging estimates as the values for $\mathbf{R}$. These values are slightly lower than the errors ($\boldsymbol{\epsilon}$) described in the previous paragraph; this result is expected because ordinary kriging assimilates many observations to make the best possible XCO_2 estimate at each location (in this case at the observation locations). For example, in the summer case study, we estimate average kriging uncertainties of $(1.5\ \mathrm{ppm})^2$. For the winter case study, we estimate average

kriging uncertainties of $(1.44 \text{ ppm})^2$. As a result, the model-data mismatch errors used in the inverse modeling simulations with the reduced dataset are slightly smaller than those used in the inverse modeling simulations with the original CO_2 dataset. Note that in the inverse modeling simulations using averaged data (e.g., the blue line in Fig. S4), we use the same model-data mismatch errors as in the inverse modeling simulations using the original XCO_2 dataset.

We further 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~~ an objective 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 $(10 \ \mu\text{mol m}^{-2} \text{ s}^{-1})^2$, decorrelation length of 555 km, and decorrelation time of 9.9 days. For the winter case study, we estimate a variance of $(3.1 \ \mu\text{mol m}^{-2} \text{ s}^{-1})^2$, 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.

175 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.

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.,
180 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.: Geo-
185 statistical 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 mea-
190 surements, 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,
195 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.
200 Tech., 11, 6539–6576, https://doi.org/10.5194/amt-11-6539-2018, 2018.

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.

Shiga, Y. P., Michalak, A. M., Fang, Y., Schaefer, K., Andrews, A. E., Huntzinger, D. H.,
205 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.
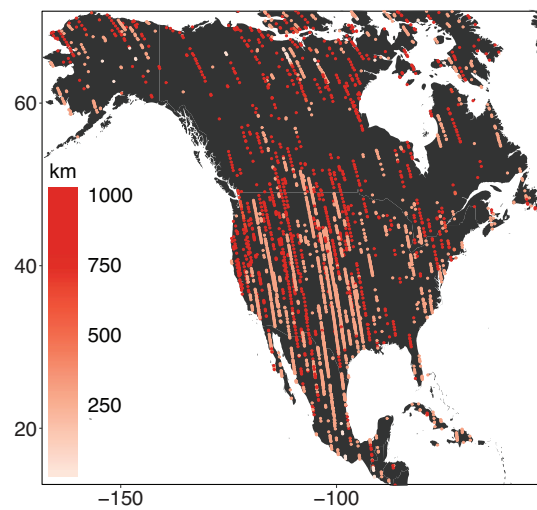
Figure S1: Correlation lengths estimated along OCO-2 flight tracks for the summer real data case study. The estimated correlation lengths are similar in magnitude and exhibit similar spatial heterogeneity to those in the synthetic case study (Fig. 1).
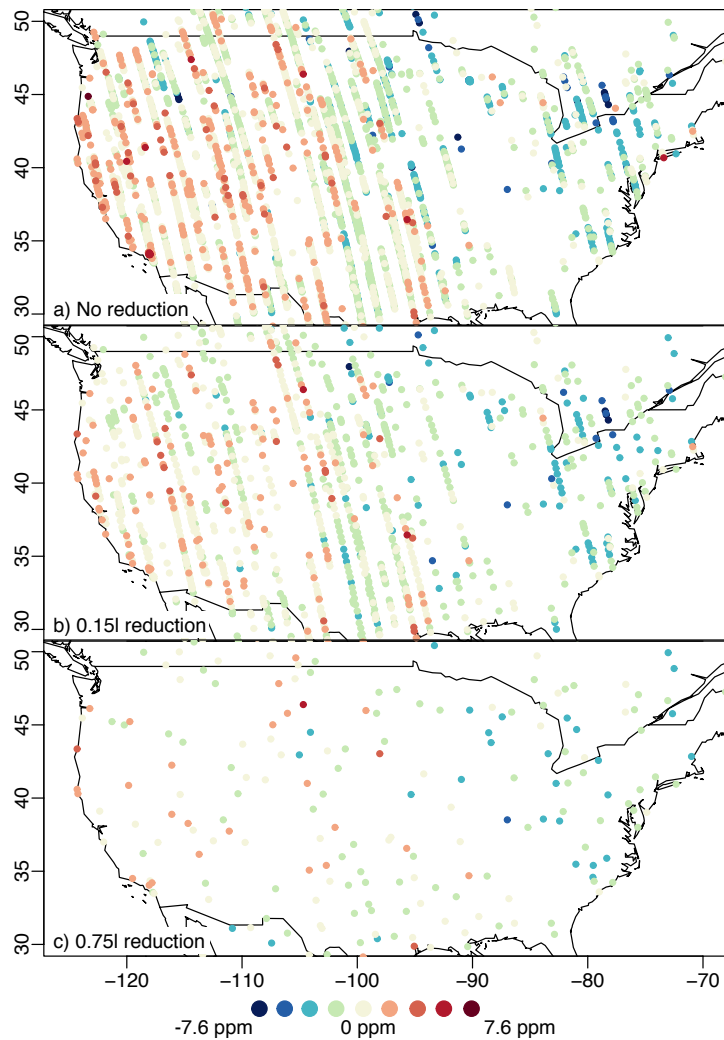
Figure S2: The original XCO_2 dataset after subtracting the CO_2 background or boundary condition (a), the dataset reduced to one observation per $0.15l$ (b), and the dataset reduced to one observation per $0.75l$ (c). Panels (a) and (b) display similar spatial patterns, but the data in panel (c) is very sparse. This figure is zoomed in over the United States to better show spatial patterns in the observations.
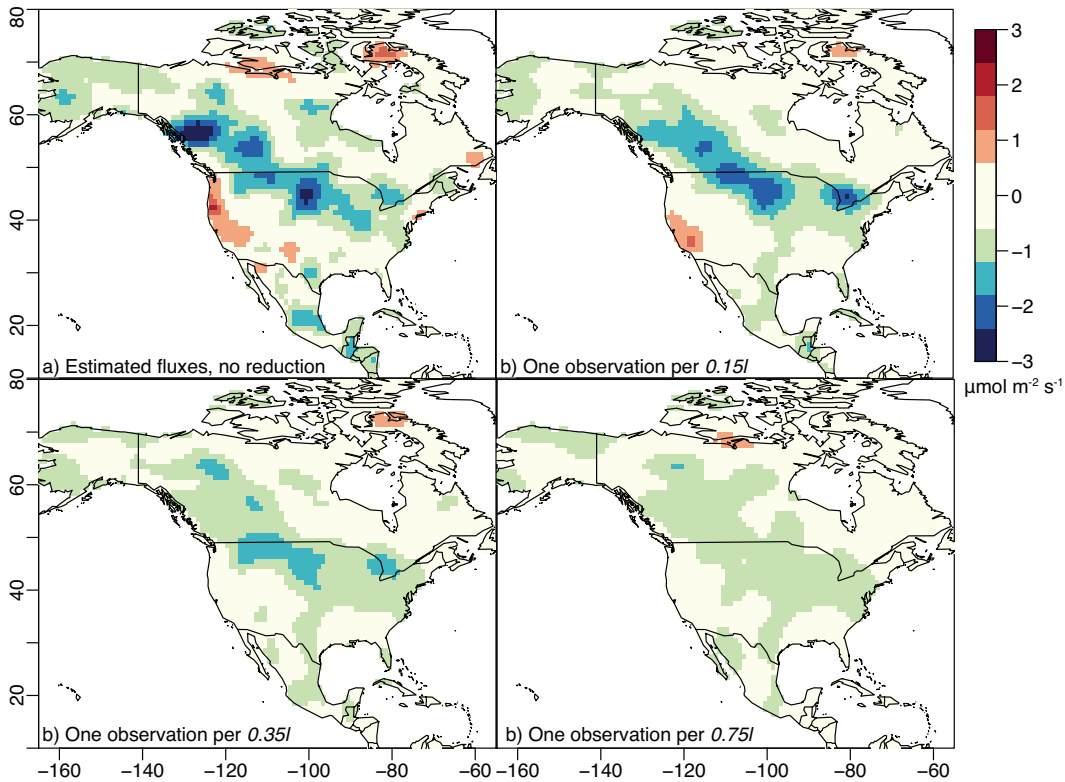
Figure S3: CO_2 fluxes estimated for the summer 2015 real data case study, averaged across the 6-week study window: (a) fluxes estimated from OCO-2 data with no reduction (5032 data points), (b) fluxes estimated from data reduced to one point per $0.15l$ (2156 data points), and (c) fluxes estimated from 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).
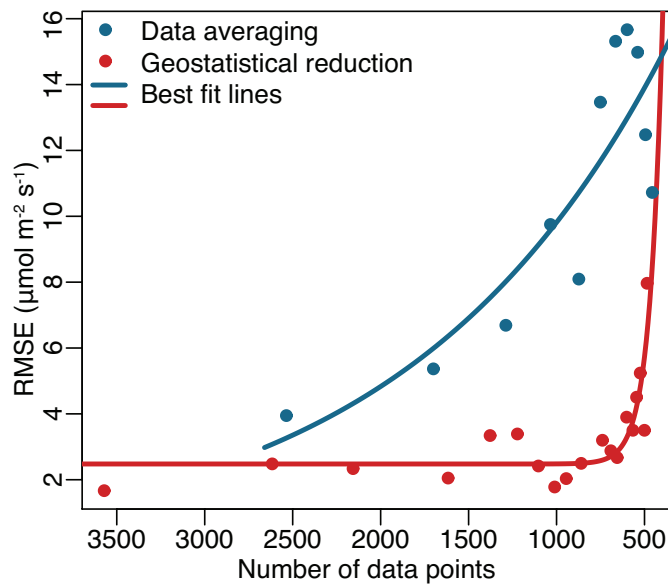
Figure S4: Root mean squared error (RMSE) of the fluxes estimated using data reduction relative to the fluxes estimated without data reduction (analogous to Fig. 5). The RMSE for the geostatistical approach proposed here is less than the RMSE for data reduced using binning and averaging.
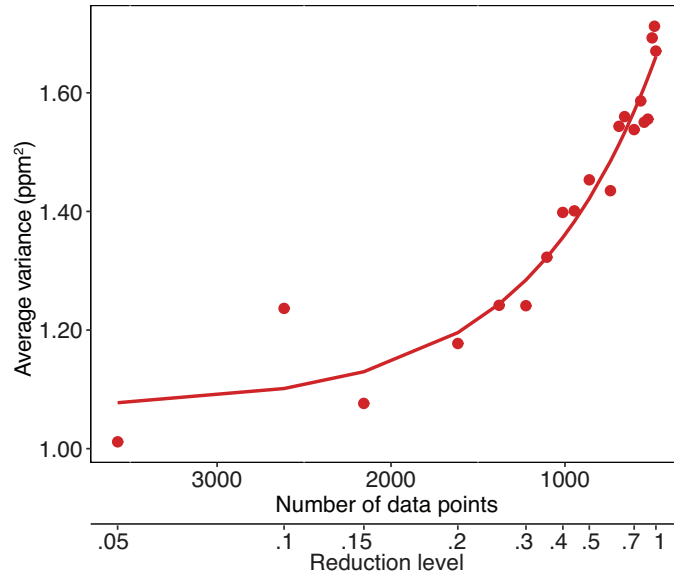
Figure S5: The amount of variance in the data that is lost through the process of data reduction. The patterns in this figure from the real data case study are similar to those from the summer synthetic case study (Fig. 6).
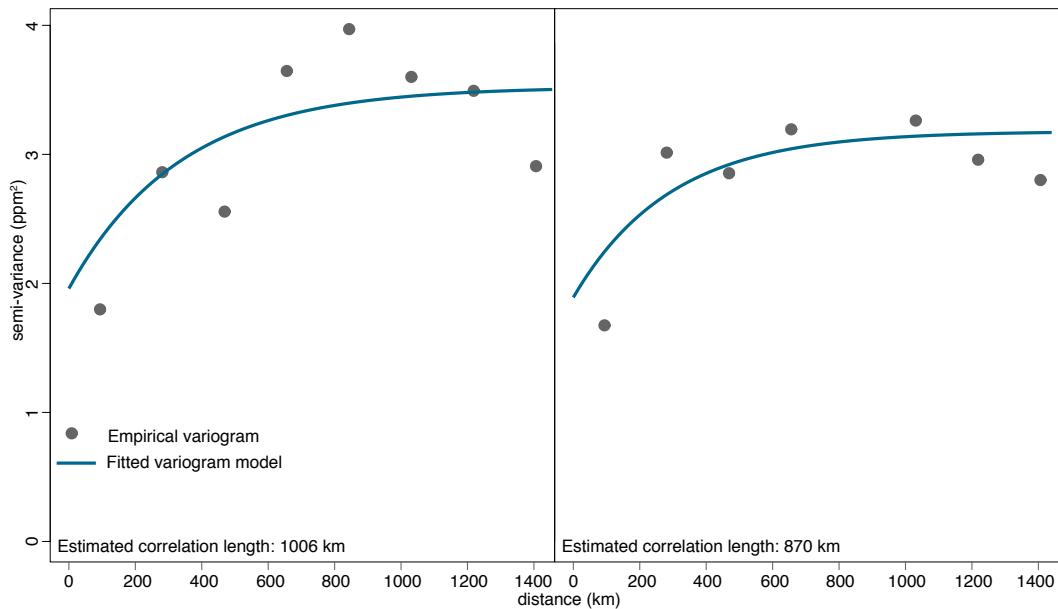
Figure S6: Two example variograms from the summer synthetic case study. The emperical variogram (grey) is calculated from the synthetic observations, and the variogram model (blue) is fitted to these points using least squares.