NIASRA

NATIONAL INSTITUTE FOR APPLIED STATISTICS RESEARCH AUSTRALIA

National Institute for Applied Statistics Research Australia

University of Wollongong, Australia

Working Paper

13-16

Multivariate Spatial Data Fusion for Very Large Remote Sensing Datasets

Hai Nguyen, Noel Cressie, and Amy Braverman

National Institute for Applied Statistics Research Australia, University of Wollongong, Wollongong NSW 2522, Australia Phone +61 2 4221 5435, Fax +61 2 4221 4998.
Email: karink@uow.edu.au

Multivariate Spatial Data Fusion for Very Large Remote Sensing Datasets

Hai Nguyen[1], Noel Cressie[2,1], and Amy Braverman[1]

[1]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA
[2]National Institute for Applied Statistics Research Australia, University of Wollongong, Australia

August 22, 2016

Abstract

Global maps of carbon dioxide (CO_2) mole fraction (in units of parts per million) in the lower atmosphere are important tools for climate research since they can help identify sources and sinks of CO_2. No satellite instrument currently provides estimates of the lower-atmosphere CO_2, though inferences are possible using data from existing instruments. Two remote sensing instruments, the Orbiting Carbon Observatory 2 (OCO-2) and the Greenhouse gases Observing SATellite (GOSAT), both observe column-averaged CO_2. These data are then used as inputs into flux inversion, which combines a transport model, *a priori* atmospheric information, and satellite-derived column-averaged CO_2 to produce estimates of sources and sinks.

Here, we demonstrate a method for improving inferences for column-averaged CO_2 using OCO-2 and GOSAT. Both instruments produce estimates of CO_2 concentration, called profiles, at 20 different pressure levels. Operationally, each profile estimate is then convolved into a single estimate of column-averaged CO_2 using a pressure weighting function. However, CO_2 may be more efficiently estimated by making optimal estimates of the vector-valued CO_2 profiles and applying the pressure weighting function afterwards. These estimates will be more efficient if there is multivariate dependence between CO_2 values in the profile. In this article, we describe a methodology that uses a modified Spatial Random Effects model to account for the multivariate nature of the data fusion of OCO-2 and GOSAT. We show that multivariate fusion of the profiles has improved mean squared error relative to scalar fusion of the column-averaged CO_2 values from OCO-2 and GOSAT. The computations scale linearly with the number of data points, making it suitable for the typically massive remote sensing datasets. Furthermore, the methodology properly accounts for differences in instrument footprint, measurement-error characteristics, and data coverages.

Keywords: data fusion, EM algorithm, multivariate geostatistics, Spatial Random Effects model.

1 Introduction

The monitoring of the global carbon cycle is an important component of atmospheric science due to carbon dioxide's impact on Earth's climate and biology. Carbon dioxide (CO_2) is one of the key inputs in photosynthesis, the process in which plants, algea, and cyanobacteria convert sunlight into chemical energy. It is also one of the by-products of respiration, the reverse process. In its atmospheric form, CO_2 acts as a greenhouse gas, and its increase since the late 19th century is believed to be playing an important role in global warming [Houghton et al., 2001]. In water, CO_2 dissolves to form carbonic acid, which contributes to ocean acidification and poses a threat to food chains connected to the oceans.

The exchange of CO_2 between the atmosphere and Earth's surface is a critical part of the global carbon cycle and an important determinant of future climate [Gruber et al., 2009]. Of particular interest to climate scientists is the global distribution of CO_2 flux, or the net amount of CO_2 exchanged between the atmosphere and the terrestrial biomes (plants, oceans, etc.) per unit of time. Regions that release CO_2 into the atmosphere (positive net flux) are called carbon *sources*, and examples of these include plant respiration, land-clearing for agriculture, and forest burning. Regions that sequester CO_2 from the atmosphere (e.g., high-photosynthesis-activity forests or grasslands) are called carbon *sinks*. Accurately identifying the location of sources and sinks over all of Earth's surface is an important research topic because of its implications for political, social, and scientific decision-making. For example, mitigation strategies can be implemented by a country with this knowledge, compliance with treaties can be monitored, and feedback can be obtained on the efficacy of policy decisions. Importantly, knowledge about sources and sinks is used in comprehensive ocean-atmosphere general circulation models (OAGCMs or GCMs), which approximate the atmosphere-ocean circulation based on equations describing motion of fluids and the input of thermodynamic energy sources such as solar radiation and latent heat [e.g., McGuffie and Henderson-Sellers, 1997].

A proxy for determining the CO_2 flux is the average CO_2 mole fraction (in units of parts per million or ppm) between the surface of Earth and the planetary boundary layer. This measure, which we call the lower-atmosphere CO_2 mole fraction, is not available from

a single remote sensing satellite instrument. However, it is possible to derive estimates of lower-atmosphere CO_2 from CO_2 measurements made by remote sensing instruments such as the Orbiting Carbon Observatory 2 (OCO-2) and the Greenhouse gases Observing SATellite (GOSAT); see Nguyen et al. [2014]. In general, space-based fleets of satellites provide an unprecedented opportunity to leverage massive, global, high-resolution, observational datasets in environmental studies.

The goal of this paper is to promote remote sensing analyses that go beyond descriptive data analyses and incorporate inferential statistical methodologies that allow rigorous quantification of uncertainty and probabilistic assessment of scientific hypotheses. We demonstrate this through the estimation of lower-atmosphere CO_2. This can be done through flux inversion, which uses *a priori* knowledge of sources and sinks, a chemistry and transport model, and satellite and ground-based column-averaged CO_2 observations [e.g., Chevallier et al., 2005].

Nguyen et al. [2012] show that it is possible to obtain better inferences on geophysical processes by combining complementary and reinforcing data from multiple satellite instruments and Nguyen et al. [2014] use this idea to obtain an estimate of the lower-atmosphere CO_2. In principle, a multivariate approach can improve upon the inferences for column-averaged CO_2 by working more closely with the retrieval processes for these two instruments.

GOSAT and OCO-2 instruments consist of spectrometers that measure that number of photons reaching the top of the atmosphere in several spectral bands. Since the amount of photons within a spectrum is differentially absorbed by the atmosphere and its CO_2 concentration, the amount of CO_2 in the atmosphere can be inferred from photon counts across multiple spectra through a process called inverse modeling or *optimal estimation* [Rodgers, 2000]. The GOSAT and OCO-2 optimal-estimation algorithms produce, at each observation location, CO_2 concentrations at 20 different atmospheric pressure levels, which they combine in a weighted average to form the column-averaged CO_2 estimate [Crisp et al., 2010]. It is this column-averaged CO_2 that is used in both flux inversions and statistical data fusion as described in Nguyen et al. [2014]. In this paper, our new approach is to reverse the order and first perform data fusion on the CO_2 profiles before combining them to form an

estimate of the column-averaged CO_2. This accounts for the statistical dependence among the profile heights and should produce more efficient statistical inferences.

In what follows, we leverage the computational efficiency provided by the Spatial Random Effects model [Cressie and Johannesson, 2008; Nguyen et al., 2012] to solve a data-fusion problem where the data sources are massive and multivariate. In Section 2, we review the data sources used in this article. Section 3 gives the statistical methodology underlying Multivariate Spatial Data Fusion (MSDF). In Section 4, we apply this methodology to the GOSAT and OCO-2 Level 2 data from calendar-year 2015 and compare the predictions to independent data from the ground-based Total Carbon Column Observing Network (TCCON). Section 5 contains discussion and conclusions, including the extension of MSDF to spatio-temporal settings.

2 Remote Sensing Data Sources for Atmospheric CO_2

The Greenhouse gases Observing Satellite (GOSAT) was launched by Japan on January 23, 2009 as a joint venture by Japan's National Institute for Environmental Studies (NIES), the Japanese Space Agency (JAXA), and the Ministry of the Environment (MOE). It is a polar-orbiting satellite dedicated to the observation of column-averaged CO_2 and CH_4, both major greenhouse gases, from space using spectra of reflected sunlight [Hamazaki et al., 2005]. GOSAT flies at approximately 665 kilometers (km) altitude, and it completes an orbit every 100 minutes. The satellite returns to the same observation location every three days [Morino et al., 2011]. There are several 'retrievals' of the GOSAT raw-radiance data. In this paper, we make use of the NASA version, which was produced by the Atmospheric CO_2 Observations from Space (ACOS) team at the Jet Propulsion Laboratory. Hereafter, we will refer to this GOSAT dataset as ACOS data.

The Orbiting Carbon Observatory-2 (OCO-2) is NASA's first Earth remote sensing instrument dedicated to studying carbon dioxide's global distribution. It was launched on July 2, 2014, and it uses three high-resolution grating spectrometers to acquire observations of the atmosphere in three observation modes: nadir, glint, and target. In nadir mode, the instrument points to the local nadir to collect data directly below the spacecraft. Nadir mode does not provide adequate signal-to-noise ratio over the dark ocean surface,

and thus over ocean OCO-2 uses glint mode. In that mode, OCO-2 points its mirrors at bright glint spots where the solar radiation is specularly reflected from the surface. Finally, in target mode the instruments locks its view onto specific surface locations (usually a ground-based TCCON station or observational tower) while flying overhead. OCO-2 has a repeat cycle of sixteen days and a sampling rate of about one million observations per day, making it a high-density and high-resolution complement to GOSAT. The CO_2 concentrations in an atmospheric column are inferred from the observed spectra through optimal estimation [Crisp et al., 2010]. The outputs are available as 20-dimensional CO_2 profiles and column-averaged CO_2 concentrations. The latter is derived from the former using a pressure weighting function, which is a 20-dimensional vector of weights derived from local atmospheric conditions. A pressure weighting function is convolved with the 20-dimensional CO_2 vector in a linear combination to form the column-averaged estimate [O'Dell et al., 2012].

However, in principle, it is possible to fuse the 20-dimensional CO_2 data vectors from GOSAT and OCO-2 directly to obtain an optimal estimate of this vector profile. Then the column-averaged CO_2 value can be obtained by applying the pressure weighting function to the fused vector of CO_2 values. This multivariate-data-fusion apporach has a distinct advantage when there is dependence down the profile, as is expected for physical reasons due to atmospheric transport.

Attempts to apply spatial inferences on these datasets would have to deal with both the change-of-support issue and the massive data sizes in addition to the vector-valued nature of the observations. The first issue concerns the problem of inferring a spatial process at one resolution using data that are obtained at another resolution. Geophysical processes of interest are often assumed to be continuous and smooth, but most remote sensing satellites observe and record the relevant processes as pixel values, where each pixel corresponds to some area in the domain. These pixels are also called footprints, and in remote sensing the value observed over a footprint is assumed to be a spatial average of the true process over the area of the footprint plus a measurement error. When estimating the underlying processes, we need to properly account for the differences in the footprints' sizes, shapes, and orientations. Inferences that do not deal with this change-of-support

issue appropriately are susceptible to a so-called "ecological fallacy," namely erroneous conclusions can occur when inferences drawn from aggregated data are assumed to apply to individual units [e.g., Cressie, 1996].

3 Multivariate Spatial Data Fusion

Spatial interpolation and statistical inference for massive data is an active area of research. Scalable spatial and spatio-temporal approaches in the recent literature include Berliner et al. [1999; hierarchical Bayesian spatio-temporal model with multiresolution wavelet basis functions and two data sources of different support], Wikle et al. [2001; more general than Berliner et al., 1999, with science-based orthogonal eigenfunctions and multiresolution basis functions to capture residual dependencies], Nychka et al. [2002; modelling nonstationary covariance functions with multiresolutional wavelet models], Hooten et al. [2003; hierarchical Bayesian model with FFT representation of spatial random effects], Royle and Wikle [2005; spectral parameterization of the spatial Poisson process], Banerjee et al. [2008; approximate optimal prediction with dimension reduction through conditioning on a small set of space-filling locations], Calder [2008; bivariate dynamic process convolution model], Cressie and Johannesson [2008; Fixed Rank Kriging based on the Spatial Random Effects model], Stein and Jun [2008; modelling nonstationary covariance models using the discrete Fourier transform], Cressie et al. [2010; Fixed Rank Filtering and Fixed Ranked Smoothing based on the Kalman filter and the Spatio-Temporal Random Effects model], and Lindgren et al. [2011; linking Gaussian fields and Gaussian Markov random fields using stochastic partial differential equations].

To our knowledge, there has been no attempt in the literature to combine multivariate profiles of considerable length (here, 20 atmospheric levels) from massive spatial datasets. The spatio-temporal methods in Nguyen et al. [2014] are in principle generalizable to multiple profile levels, but a key parameter, the spatial covariance matrix of the Spatial Random Effects vector, increases quadratically in size with respect to the number of levels, thereby making it quickly infeasible. In this article, we modify the approach in Nguyen et al. [2014] to use a three-dimensional (surface $\times$ height) spatial covariance model whose computational complexity does not depend on the number of atmospheric levels.

In this section, we introduce the data model and the process model of a hierarchical statistical model. The motivation and partial derivations for Multivariate Spatial Data Fusion (MSDF) are presented in Section 3.2. Multivariate spatial basis functions are constructed in Section 3.3. The model's parameters are estimated using the EM algorithm, which we describe in Section 3.4.

3.1 Data model and properties

Let $Y(\mathbf{s}, h)$ represent the CO_2 concentration at location $\mathbf{s} \in D$ and and physical or geopotential height $h \in H$, where D and H represent the domain in the horizontal and vertical directions, respectively. Let the horizontal domain of interest be defined as $\cup\{A_l \subset \mathbb{S} : l = 1, \ldots, N_D\}$, which is made up of N_D fine-scale, non-overlapping, Basic Areal Units (BAUs) $\{A_l\}$ with locations $D \equiv \{\mathbf{p}_l \in A_l : l = 1, \ldots, N_D\}$, and $\mathbb{S}$ is the surface that a sphere that approximates Earth. Similarly, the vertical domain of interest is $\cup\{V_m \subset \mathbb{R}^+ : m = 1, \ldots, N_H\}$, which is a collection of non-overlapping Basic Vertical Units (BVUs) with $H = \{q_m \in V_m : m = 1, \ldots, N_H\}$. The BAUs and BVUs represent the smallest resolution at which we will make predictions with our model.

Suppose we have data at K different heights $\{h_1, h_2, \ldots, h_K\} \subset H$. For convenience, we will refer to these heights by their index k rather than their height h_k (e.g., when referring to a dataset at height h_2, we will simplify the notation and refer to the dataset with a superscript $k = 2$). We define the observations from instrument i at height k as an $N_{k,i}$-dimensional vector,

$$\mathbf{Z}^{(k,i)} \equiv \left(Z^{(k,i)}(A_{i,1}), \ldots, Z^{(k,i)}(A_{i,N_{k,i}})\right)',$$

where $A_{i,j}$ is the j-th footprint of the i-th instrument, which is made up of an appropriate subset of BAUs $\{A_l\}$. Here, for ease of exposition, we assume that there are two separate instruments to be fused (i.e., $i = 1, 2$), although in principle that number could be larger than 2.

We assume that the observation at footprint $A_{i,j}$ and height h_k is a spatial average of the true unobserved process $Y(\cdot, h_k)$ over the footprint plus an instrument-and-height-specific

Gaussian measurement-error process. That is,

$$Z^{(k,i)}(A_{i,j}) = \frac{1}{|D \cap A_{i,j}|} \left\{ \sum_{\mathbf{u} \in D \cap A_{i,j}} Y(\mathbf{u}, h_k) \right\} + \epsilon^{(i,k)}(A_{i,j}); \quad i = 1, 2,\ k = 1, \ldots, K, \tag{1}$$

where $\mathbf{s}$ is a BAU in D, and $\epsilon^{(i,k)}(\cdot)$ is a measurement-error process that is potentially a function of the location, size, shape, and orientation of the footprint $A_{i,j}$. Here, we assume that for a given i and k the measurement errors are independently and identically distributed as a Gaussian process with standard deviation $\sigma_\epsilon^{(i,k)}$, which is independent of $Y(\cdot)$. This measurement-error component in (1) may have a height-dependent non-zero mean that captures the instrument bias. Because many remote sensing instruments have multiplicative bias [Nguyen et al., 2014], we assume that the measurement-error process $\epsilon^{(i,k)}(A_{i,j})$ satisfies $E(\epsilon^{(i,k)}(A_{i,j})) = c^{(i,k)} E(Y(A_{i,j}, h_k))$, where $c^{(i,k)}$ is a known multiplicative bias constant for the i-th instrument at the k-th height, and the case of zero bias is captured by $c^{(i,k)} = 0$. We further assume that for $i_1 \neq i_2$, $\epsilon^{(i_1,k_1)}(\cdot)$ is indepedent of $\epsilon^{(i_2,k_2)}(\cdot)$. The underlying true process at height h is assumed to follow the form,

$$Y(\mathbf{s}, h) = \mu(\mathbf{s}, h) + \nu(\mathbf{s}, h) + \xi(\mathbf{s}, h); \quad \mathbf{s} \in D,\ \ h \in H, \tag{2}$$

where at height h, $\mu(\cdot, h)$ is the large-scale trend process, $\nu(\cdot, h)$ is the small-scale spatial-variability process, and $\xi(\cdot, h)$ is the fine-scale spatial-variability process. Equivalently, we can group $Y(\cdot, h)$ in (2) over all the BAUs in D into vector form as follows,

$$\mathbf{Y}(h) = \boldsymbol{\mu}(h) + \boldsymbol{\nu}(h) + \boldsymbol{\xi}(h), h \in H,$$

where $\boldsymbol{\mu}(h), \boldsymbol{\nu}(h)$, and $\boldsymbol{\xi}(h)$ are all N_D-dimensional vectors, $\boldsymbol{\nu}(h)$ and $\boldsymbol{\xi}(h)$ are mean-zero random vectors, and $\boldsymbol{\nu}(h)$ is statistically independent of $\boldsymbol{\xi}(h)$.

We assume that the trend has the form $\mu(\mathbf{s}, h) = \mathbf{t}(\mathbf{s}, h)'\boldsymbol{\alpha}(h)$, and that the small-scale spatial variability term has the form $\nu(\mathbf{s}, h) = \mathbf{S}(\mathbf{s}, h)'\boldsymbol{\eta}$, which is the Spatial Random Effects (SRE) model [Cressie and Johannesson, 2008]. Then (2) can be written as,

$$Y(\mathbf{s}, h) = \mathbf{t}(\mathbf{s}, h)'\boldsymbol{\alpha}(h) + \mathbf{S}(\mathbf{s}, h)'\boldsymbol{\eta} + \xi(\mathbf{s}, h); \quad \mathbf{s} \in D,\ \ h \in H, \tag{3}$$

where at height h, $\mathbf{t}(\mathbf{s}, h)$ is a p-dimensional vector of known covariates at location $\mathbf{s}$, $\boldsymbol{\alpha}(h)$ is a p-dimensional vector of unknown regression coefficients, $\mathbf{S}(\mathbf{s}, h)$ is a q-dimensional vector

of given spatial basis functions (of location $\mathbf{s}$ and height h), and $\boldsymbol{\eta}$ is a q-dimensional Gaussian random vector with mean $\mathbf{0}$ and unknown variance-covariance matrix $\mathbf{K}$. Similarly, equation (3) can be stacked over the N_D BAUs into vector form as follows:

$$\mathbf{Y}(h) = \mathbf{T}(h)\boldsymbol{\alpha}(h) + \mathbf{S}(h)\boldsymbol{\eta} + \boldsymbol{\xi}(h),$$

where $\mathbf{T}(h)$ and $\mathbf{S}(h)$ are the matrices $[\mathbf{t}(\mathbf{s}_1, h), \cdots, \mathbf{t}(\mathbf{s}_{N_D}, h)]'$ and $[\mathbf{S}(\mathbf{s}_1, h), \cdots, \mathbf{S}(\mathbf{s}_{N_D}, h)]'$ of dimension $N_D \times p$ and $N_D \times q$, respectively.

Then the process can be further stacked over all BVU heights $q_m \in H$, for $m \in 1, \ldots, N_H$, as follows:

$$\begin{pmatrix} \mathbf{Y}(q_1) \\ \vdots \\ \mathbf{Y}(q_{N_H}) \end{pmatrix} = \begin{pmatrix} \mathbf{T}(q_1) & \ldots & \mathbf{0} \\ \vdots & \ddots & \\ \mathbf{0} & & \mathbf{T}(q_{N_H}) \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha}(q_1) \\ \vdots \\ \boldsymbol{\alpha}(N_H) \end{pmatrix} + \begin{pmatrix} \mathbf{S}(q_1) \\ \vdots \\ \mathbf{S}(q_{N_H}) \end{pmatrix} \boldsymbol{\eta} + \begin{pmatrix} \boldsymbol{\xi}(q_1) \\ \vdots \\ \boldsymbol{\xi}(q_{N_H}) \end{pmatrix},$$

where the process vectors are $N_D N_H$-dimensional, the regression-coefficient vector is $N_D\, p$-dimensional, and importantly the basis-function coefficient vector remains q-dimensional. Recall that for the $q \times q$ variance-covariance matrix $\mathbf{K}$, the basis-function coefficient vector is distributed as,

$$\boldsymbol{\eta} \;\sim\; \text{Gau}\left(\mathbf{0}, \mathbf{K}\right),$$

independently of the $N_D N_H$-dimensional fine-scale variation vector,

$$\begin{pmatrix} \boldsymbol{\xi}(q_1) \\ \vdots \\ \boldsymbol{\xi}(q_{N_H}) \end{pmatrix} \;\sim\; \text{Gau}\left(\mathbf{0}, \begin{pmatrix} (\sigma_\xi^{(q_1)})^2\mathbf{I} & \ldots & \mathbf{0} \\ \vdots & \ddots & \\ \mathbf{0} & & (\sigma_\xi^{(q_{N_H})})^2\mathbf{I} \end{pmatrix}\right).$$

3.2 Multivariate Spatial Data Fusion (MSDF)

Having described the process model and its representation in terms of spatial basis functions, we now describe how to optimally combine (i.e., fuse) information when we have K-variate spatial data for two instruments, written here as $\{\mathbf{Z}^{(k,i)} : k = 1, \ldots, K;\ \ i = 1, 2\}$. Notice that we have chosen to keep the discussion quite general in terms of fusing K-variate

data from different instruments; for our application, $K = 20$, but fusing sub-vectors where $K < 20$ is also possible.

We first concatenate all datasets at the *same* height and denote that dataset as the $N^{(i,k)}$-dimensional vector $\mathbf{Z}^{(k)}$ for $k \in 1, \ldots, K$. The full data model for all heights $h_1, \ldots, h_K$ can be expressed as

$$\begin{pmatrix} \mathbf{Z}^{(1)} \\ \vdots \\ \mathbf{Z}^{(K)} \end{pmatrix} = \begin{pmatrix} \tilde{\mathbf{T}}^{(1)} & \ldots & \mathbf{0} \\ \vdots & \ddots & \\ \mathbf{0} & & \tilde{\mathbf{T}}^{(K)} \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha}^{(1)} \\ \vdots \\ \boldsymbol{\alpha}^{(K)} \end{pmatrix} + \begin{pmatrix} \tilde{\mathbf{S}}^{(1)} \\ \vdots \\ \tilde{\mathbf{S}}^{(K)} \end{pmatrix} \boldsymbol{\eta} + \begin{pmatrix} \tilde{\boldsymbol{\xi}}^{(1)} \\ \vdots \\ \tilde{\boldsymbol{\xi}}^{(K)} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\epsilon}^{(1)} \\ \vdots \\ \boldsymbol{\epsilon}^{(K)} \end{pmatrix}, \quad (4)$$

or equivalently,

$$\mathbf{Z} = \tilde{\mathbf{T}}\boldsymbol{\alpha} + \tilde{\mathbf{S}}\boldsymbol{\eta} + \tilde{\boldsymbol{\xi}} + \boldsymbol{\epsilon}, \quad (5)$$

where the tilde "$\sim$" on the components of (5) indicates that the corresponding term or process is aggregated over BAUs within the observed data footprints; see (1). For instance, the aggregated matrix $\tilde{\mathbf{S}}$ is made up of terms,

$$\tilde{\mathbf{S}}^{(k,i)}(A_{i,j}) \equiv \frac{1}{|D \cap A_{i,j}|} \sum_{\mathbf{u} \in D \cap A_{i,j}} \mathbf{S}^{(k,i)}(\mathbf{u}),$$

where recall that $A_{i,j}$ is the j-th footprint of the i-th instrument. This procedure very effectively accounts for the change-of-support issue between the two instruments, taking advantage of the linearity of the SRE model. For more details, see Nguyen et al. [2012].

Given the formulation above, we can carry out spatial prediction of the process Y at all (BAU, BVU) combinations using a linear combination of the data $\mathbf{Z} \equiv \{\mathbf{Z}^{(k)} : k = 1, \ldots, K\}$. That is, our estimate of $Y(\mathbf{s}, h)$, the true process at location $\mathbf{s}$ and height h, is,

$$\hat{Y}(\mathbf{s}, h) = \mathbf{a}'\mathbf{Z}, \quad (6)$$

where $\mathbf{a}$ implicitly depends on $\mathbf{s}$ and h. Subject to an unbiasedness constraint, we minimize,

$$\begin{aligned} E(Y(\mathbf{s}, h) - \hat{Y}(\mathbf{s}, h))^2 &= \mathrm{var}(Y(\mathbf{s}, h) - \mathbf{a}'\,\mathbf{Z}) \\ &= \mathrm{var}(Y(\mathbf{s}, h)) - 2\mathbf{a}'\,\mathrm{cov}(\mathbf{Z}, Y(\mathbf{s}, h)) + \mathbf{a}'\,\mathrm{var}(\mathbf{Z})\,\mathbf{a}, \end{aligned} \quad (7)$$

with respect to $\mathbf{a}$, where the unbiasedness constraint is p-dimensional:

$$\mathbf{0} = \mathbf{a}'\mathbf{B}\mathbf{T} - \mathbf{t}(\mathbf{s}, h)', \quad (8)$$

where

$$\mathbf{T} = \begin{pmatrix} \mathbf{T}^{(1)} & \dots & \mathbf{0} \\ \vdots & \ddots & \\ \mathbf{0} & & \mathbf{T}^{(K)} \end{pmatrix}, \text{ for } \mathbf{T}^{(k)} \equiv \mathbf{T}(h_k),$$

and $\mathbf{B}$ is a matrix with the multiplicative bias coefficients down the diagonals. That is,

$$\mathbf{B} = \begin{pmatrix} \mathbf{B}^{(1)} & \dots & \mathbf{0} \\ \vdots & \ddots & \\ \mathbf{0} & & \mathbf{B}^{(K)} \end{pmatrix}, \text{ for } \mathbf{B}^{(k)} \equiv \begin{pmatrix} (1+c^{(k,1)})\mathbf{I}_{N^{(k,1)}} & \mathbf{0} \\ \mathbf{0} & (1+c^{(k,2)})\mathbf{I}_{N^{(k,2)}} \end{pmatrix}.$$

The minimization of (7) with respect to $\mathbf{a}$ can be solved using the method of Lagrange multipliers. Let $\mathbf{\Sigma} \equiv \text{var}(\mathbf{Z})$ and $\mathbf{c} \equiv \text{cov}(\mathbf{Z}, Y(\mathbf{s},h))$. Then, the solution for $\mathbf{a}$ is

$$\hat{\mathbf{a}}' = \left(\mathbf{c}' + (\mathbf{t}(\mathbf{s},h)' - \mathbf{c}'\,\mathbf{\Sigma}^{-1}\mathbf{B}\mathbf{T})(\mathbf{T}'\mathbf{B}\mathbf{\Sigma}^{-1}\mathbf{B}\mathbf{T})^{-1}\mathbf{T}'\mathbf{B}\right)\mathbf{\Sigma}^{-1}. \quad (9)$$

Having derived the data-fusion coefficients $\mathbf{a}$, we can produce the fused prediction and its prediction standard error at $\mathbf{s} \in D$ and $h \in H$, as follows,

$$\hat{Y}(\mathbf{s},h) \equiv \hat{\mathbf{a}}'\mathbf{Z} \quad (10)$$

$$\hat{\sigma}(\mathbf{s},h) \equiv \left(\text{var}(Y(\mathbf{s},h)) - 2\hat{\mathbf{a}}'\,\mathbf{c} + \hat{\mathbf{a}}'\,\mathbf{\Sigma}\,\hat{\mathbf{a}}\right)^{\frac{1}{2}}, \quad (11)$$

where $\hat{\mathbf{a}}$ is given by (9).

Computation of (9) and hence of (10) and (11) requires inversion of $\mathbf{\Sigma}$, which is typically enormous. However, because of the SRE parameterization in (3), inversion of $\mathbf{\Sigma}$ can be computed exactly with linear computational complexity using the Sherman-Morrison-Woodbury formula [Henderson and Searle, 1981]. Due to the data model given by (5), the covariance matrix of the dataset has the following form:

$$\mathbf{\Sigma} = \mathbf{S}\mathbf{K}\mathbf{S}' + \mathbf{C}\mathbf{E} + \mathbf{V},$$

where $\mathbf{V} \equiv \text{cov}(\boldsymbol{\epsilon})$ and $\text{cov}(\boldsymbol{\xi}) = \mathbf{C}\mathbf{E}$. The component matrices of the fine-scale covariance matrix are defined as follows,

$$\mathbf{C} = \begin{pmatrix} (\sigma_\xi^{(1)})^2\mathbf{I}_{N^{(1)}} & \dots & \mathbf{0} \\ \vdots & \ddots & \\ \mathbf{0} & & (\sigma_\xi^{(K)})^2\mathbf{I}_{N^{(K)}} \end{pmatrix}, \text{ and } \mathbf{E} = \begin{pmatrix} \mathbf{E}^{(1)} & \dots & \mathbf{0} \\ \vdots & \ddots & \\ \mathbf{0} & & \mathbf{E}^{(K)} \end{pmatrix},$$

where

$$\mathbf{E}^{(k)} \equiv \left[\frac{|D \cap A^{(k)}_{i_1,j_1} \cap A^{(k)}_{i_2,j_2}|}{|D \cap A^{(k)}_{i_1,j_1}||D \cap A^{(k)}_{i_2,j_2}|}; i_1, i_2 = 1, 2 \text{ and } j_i = 1, \ldots, N^{(k,i)} \right],$$

is a matrix constructed from footprint overlaps. Under this parameterization, the inversion can be computed exactly using the Sherman-Morrison-Woodbury formula,

$$\mathbf{\Sigma}^{-1} = \mathbf{D}^{-1} - \mathbf{D}^{-1}\mathbf{S}\left[\mathbf{K}^{-1} + \mathbf{S}'\mathbf{D}^{-1}\mathbf{S}\right]^{-1}\mathbf{S}'\mathbf{D}^{-1}, \tag{12}$$

where $\mathbf{D} \equiv \mathbf{CE} + \mathbf{V}$. The procedure requires inversion of the $N \times N$ matrix $\mathbf{D}$, which is typically very sparse, and inversion of $\mathbf{K}$ and $(\mathbf{K}^{-1}+\mathbf{S}'\mathbf{D}^{-1}\mathbf{S})$, both of which are $q \times q$. Since q is the number of spatial basis functions and is chosen by the user, it is typically much smaller than N, which results in very substantial speed-ups. The overall computational complexity for MSDF is $O(Nq^2)$, where N, the total number of observations, is much larger than the number of basis functions q. Furthermore, the model in (3) does not assume isotropy or stationarity in the data. Together, the scalability and flexibility of MSDF makes it well-suited for the typically massive and heterogeneous data found in remote sensing applications.

The parameters $\mathbf{K}$ and $\{(\sigma^{(k)}_\xi)^2 : k = 1, \ldots, K\}$ are typically unknown and have to be estimated from the data. The maximum likelihood estimates of these parameters are analytically intractable, so the preferred method is to obtain them iteratively using the EM algorithm [see Katzfuss and Cressie, 2011; Nguyen et al., 2014, for more details]. We present the details of the EM algorithm for MSDF in Section 3.4.

3.3 Constructing the spatial basis function $\mathbf{S}(\mathbf{s}, h)$

In the previous section we defined the term $\mathbf{S}(\mathbf{s}, h)$ as a q-dimensional vector of basis functions at location $\mathbf{s}$ and height h. The model in (3) allows scalable computations as a consequence of this formulation, but in practice specifying and estimating a consistent three-dimensional (varying in space and height) covariance function is difficult. This problem is somewhat related to that of estimating the joint covariance between mid-tropospheric and column-averaged CO_2 mole fractions as described in Nguyen et al. [2014]. There, they specified a covariance model in which the number of basis centers q varied linearly with respect to the number of processes P they considered. That is, $q = c \cdot P$, where $c \approx 300$

and $P = 2$ in that case. However, $P = 20$ in our CO_2-profile application, which for $c = 300$ yields $q = 6000$, a number that is too large for fast computations.

We take advantage of the fact that CO_2 concentration is a function of height and assume that the vector of spatio-temporal basis functions $\mathbf{S}(\mathbf{s}, h)$ can be written as,

$$\mathbf{S}(\mathbf{s}, h) = \boldsymbol{\tau}(\mathbf{s}, h) \otimes \mathbf{B}(\mathbf{s}),$$

where $\mathbf{B}(\mathbf{s})$ is a b-dimensional multi-resolutional "horizontal" basis expansion over latitudes and longitudes (e.g., bi-square basis functions, wavelets, etc.), $\boldsymbol{\tau}(\mathbf{s}, h)$ is an r-dimensional "vertical" basis expansion, and $\otimes$ denotes the Kronecker product. The dimension b typical ranges between 100-300, while r is much smaller and ranges between 4 and 10. Note that the horizontal basis functions in $\mathbf{B}(\mathbf{s})$ are independent of height, while the vertical basis functions change depending on the location $\mathbf{s}$ and the height h. There is a good physical reason for this, as we explain below.

For $\mathbf{B}(\mathbf{s})$, we choose to use three resolutions of bisquare basis functions based on the Discrete Global Grid [for more detail, see Carr et al., 1998]. For $\boldsymbol{\tau}(\mathbf{s}, h)$, we choose to use cubic B-splines with exterior knots placed at the surface and the top of the atmosphere [for more details, see Bartels et al., 1998]. The interior knots should naturally be placed at the boundary of the atmospheric layers (planetary boundary layers, troposphere, mesospheres, etc.) where a change in the behavior of CO_2 is expected as the height transitions between different atmospheric regimes.

OCO-2 and GOSAT have 20 different levels $\{h_k : k = 1, \ldots, 20\}$, which range from the surface (0 km) to the middle of the troposphere (approximately 32 km). Consequently, when combining profiles from these instruments, we use two geophysical transition points (planetary boundary layer height and tropopause height) between 0 km and 32 km as the interior knots.

The tropopause is the height at which the atmospheric temperature transitions from decreasing with altitude to increasing with altitude. Below the troposphere is the planetary boundary layer (PBL), which is another important transition point; it is the upper boundary of the the lowest layer of the troposphere where wind is influenced by friction. The meteorological and atmospheric regime changes substantially at these two vertical transitions, hence our decision to place the two interior cubic-spline knots there.

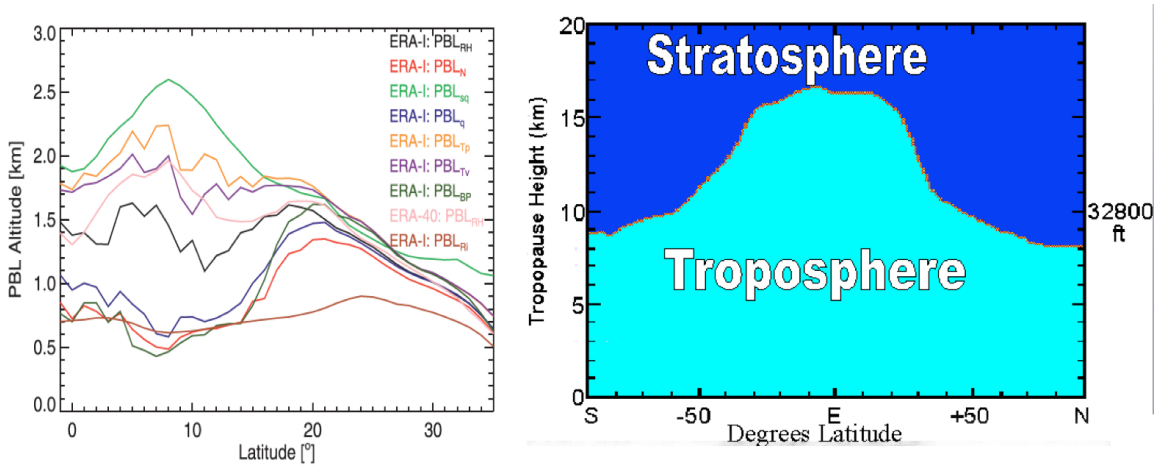

Figure 1: Left: Planetary boundary layer (PBL) height as a function of latitude from the European Centre for Medium-Range Weather Forecasts using different PBL definitions. Right: tropopause height as a function of latitude from the Atmospheric Infrared Sounder.

PBL heights are available through European Centre for Medium-Range Weather Forecasts (ECMWF) Reanalysis Data. However, there are differences in the estimated heights depending on the working definition for "planetary boundary layer" (see left panel of Figure 1). In our application, we make the simplifying assumption that the PBL height is constant at 1 km across all latitudes.

We use tropopause height data from the AIRS instrument. The tropopause height typically depends on latitude, longitude, and time, but we make the simplifying assumption that the tropopause height depends only on the latitude as shown in the right panel of Figure 1.

Since we have two exterior knots and two interior knots, the resulting cubic splines has dimension 5. More specifically, given a four-dimensional knot vector $\mathbf{t}(\mathbf{s}) = (t_1(\mathbf{s}), t_2(\mathbf{s}), t_3(\mathbf{s}), t_4(\mathbf{s}))'$, which are the surface, PBL, tropopause, and top-of-atmosphere heights, respectively, the vertical basis functions are $\boldsymbol{\tau}(\mathbf{s}, h) \equiv (B_{1,3,\mathbf{s}}(h), \ldots, B_{5,3,\mathbf{s}}(h))$, where

$$B_{l,d,\mathbf{s}}(h) = \frac{h - t_l(\mathbf{s})}{t_{l+d}(\mathbf{s}) - t_l(\mathbf{s})} B_{l,d-1,\mathbf{s}}(h) + \frac{t_{l+1+d}(\mathbf{s}) - h}{t_{l+1+d}(\mathbf{s}) - t_{l+1}(\mathbf{s})} B_{l+1,d-1,\mathbf{s}}(h),$$

with

$$B_{l,0,\mathbf{s}}(h) = \begin{cases} 1, & \text{if } t_l(\mathbf{s}) \leq h < t_{j+1}(\mathbf{s}); \\ 0, & \text{otherwise.} \end{cases}$$

. The horizontal basis functions $\mathbf{B}(\mathbf{s})$ have approximately 300 basis functions, so q $\approx 300 \times 5 = 1500$, and the resulting variance-covariance matrix $\mathbf{K}$ of the SRE model is approximately 1500×1500. Hence, the dimension of the problem has been reduced even further from around 6000 to 1500, which is now computationally feasible.

3.4 EM algorithm for parameter estimation

To apply MSDF to actual data, we need to estimate the parameters $\boldsymbol{\theta} \equiv \{\boldsymbol{\alpha}, \mathbf{K}, (\sigma_\xi^{(k)})^2; k = 1, \ldots, K\}$. We estimate them using the EM algorithm in a like manner to the estimation found in Nguyen et al. [2014]. There, we define $\boldsymbol{\eta}$ and $\tilde{\boldsymbol{\xi}}$ as "missing data." Let $\boldsymbol{\theta}^{[b]}$ denote the vector of parameter values at the b-th iteration. Using the current value of the parameter vector, $\boldsymbol{\theta} = \boldsymbol{\theta}^{[b]}$, Katzfuss and Cressie [2011] give the following conditional expectations and covariance matrices for the missing data:

$$\begin{aligned}
\boldsymbol{\eta}^{[b]} &\equiv \mathrm{E}_{\boldsymbol{\theta}^{[b]}}(\boldsymbol{\eta}|\mathbf{Z}) = \mathbf{K}^{[b]}\tilde{\mathbf{S}}'(\boldsymbol{\Sigma}^{[b]})^{-1}(\mathbf{Z} - \mathbf{B}\tilde{\mathbf{T}}\boldsymbol{\alpha}^{[b]}) && (13)\\
\tilde{\boldsymbol{\xi}}^{[b]} &\equiv \mathrm{E}_{\boldsymbol{\theta}^{[b]}}(\tilde{\boldsymbol{\xi}}|\mathbf{Z}) = \mathbf{C}^{[b]}\mathbf{E}(\boldsymbol{\Sigma}^{[b]})^{-1}(\mathbf{Z} - \mathbf{B}\tilde{\mathbf{T}}\boldsymbol{\alpha}^{[b]}) && (14)\\
\mathbf{P}^{[b]} &\equiv \mathrm{cov}_{\boldsymbol{\theta}^{[b]}}(\boldsymbol{\eta}|\mathbf{Z}) = \mathbf{K}^{[b]} - \mathbf{K}^{[b]}\tilde{\mathbf{S}}'(\boldsymbol{\Sigma}^{[b]})^{-1}\tilde{\mathbf{S}}\mathbf{K}^{[b]} && (15)\\
\mathbf{R}^{[b]} &\equiv \mathrm{cov}_{\boldsymbol{\theta}^{[b]}}(\tilde{\boldsymbol{\xi}}|\mathbf{Z}) = \mathbf{C}^{[b]}\mathbf{E} - \mathbf{C}^{[b]}\mathbf{E}(\boldsymbol{\Sigma}^{[b]})^{-1}\mathbf{E}\mathbf{C}^{[b]}, && (16)
\end{aligned}$$

where

$$\mathbf{C}^{[b]} \equiv \begin{pmatrix} (\sigma_\xi^{(1)})^{2\,[b]}\,\mathbf{I}_{N^{(1)}} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & (\sigma_\xi^{(K)})^{2\,[b]}\,\mathbf{I}_{N^{(2)}} \end{pmatrix}, \quad (17)$$

and $\boldsymbol{\Sigma}^{[b]} \equiv \tilde{\mathbf{S}}\mathbf{K}^{[b]}\tilde{\mathbf{S}}' + \mathbf{D}^{[b]}$.

The updating equations for the parameters are:

$$\begin{aligned}
\boldsymbol{\alpha}^{[b+1]} &= (\tilde{\mathbf{T}}'\mathbf{B}\mathbf{V}^{-1}\mathbf{B}\tilde{\mathbf{T}})^{-1}\tilde{\mathbf{T}}'\mathbf{B}\mathbf{V}^{-1}\left[\mathbf{Z} - \tilde{\mathbf{S}}\boldsymbol{\eta}^{[b]} - \tilde{\boldsymbol{\xi}}^{[b]}\right] && (18)\\
\mathbf{K}^{[b+1]} &= \mathbf{P}^{[b]} + \boldsymbol{\eta}^{[b]}(\boldsymbol{\eta}^{[b]})' && (19)\\
(\sigma_\xi^{(k)})^{2\,[b+1]} &= \frac{1}{N^{(k)}}\mathrm{trace}\left(\left(\mathbf{E}^{-1}\left[\mathbf{R}^{[b]} + \tilde{\boldsymbol{\xi}}^{[b]}\tilde{\boldsymbol{\xi}}^{[b]'}\right]\right)_{[d_1(k),d_2(k)]}\right), && (20)
\end{aligned}$$

where $d_1(k) = \sum_{m=0}^{k-1} N_m + 1$; $d_2(k) = \sum_{m=0}^{k} N_m$; $(\mathbf{A})_{[i,j]}; j \geq i$ is the sub-block of the square matrix $\mathbf{A}$ consisting of all elements of $\mathbf{A}$ whose row and column indices both belong to the set given by the sequence of successive integers $\{i, \ldots, j\}$; and k =1, . . . , K.

4 Application to CO_2 data from ACOS and OCO-2

In this section we present a demonstration of our MSDF methodology for fusion the ACOS (GOSAT) and OCO-2 profile data of CO_2 and compare it to independent validation data. Unfortunately, ground-based CO_2 profile data are mostly limited to aircraft flights or balloon campaigns [e.g., Karion et al., 2010], which tend to be sparse in both spatial and temporal coverage. Because of this difficulty, we make use of column-averaged CO_2 data from the Total Carbon Column Observing Network (TCCON) instead. These are a set of about 30 globally distributed ground-based stations that record daily observations of total-column (i.e., column-averaged) CO_2. While this dataset does not provide a CO_2 profile, we can convert our profile estimates at these TCCON locations into the corresponding column-averaged CO_2 values for comparison. To assess the impact of accounting for the vertical dependence in the data-fusion algorithm, we also interpolate the already-column-averaged univariate CO_2 data from ACOS and OCO-2 to the TCCON sites using a form of kriging modified to accommodate massive datasets [Spatial Statistical Data Fusion or SSDF; Nguyen et al., 2012] and compare column-averaged MSDF to SSDF relative to the "ground-truth" TCCON values.

4.1 Overview of ACOS, OCO-2, and TCCON data

The Greenhouse gases Observing Satellite (GOSAT) was launched on January 23, 2009 and the OCO-2 satellite was launched on July 2, 2014. As of August 2016, both instruments are still operational. For the period of comparison, we make use of OCO-2 and ACOS (i.e., GOSAT) Level 2 data between January 1 and December 31, 2015. We use v3.5 release of the ACOS data, which are available from the Goddard Data and Information Services Center. Following recommendations from the ACOS Data User's Guide, we apply the recommended data-screening procedure to eliminate potentially bad retrievals [Osterman et al., 2016b]. For the OCO-2 data, we use version 7.0, which is also available from the Goddard Data and Information Services Center . The OCO-2 data have a set of data-quality flags called Warn Levels [Osterman et al., 2016a], and for this exercise we only use data with Warn Levels less than 15. This particular filter eliminated about 25% of the converged OCO-2 data. For more information on both datasets, see 'ACOS Data Access'

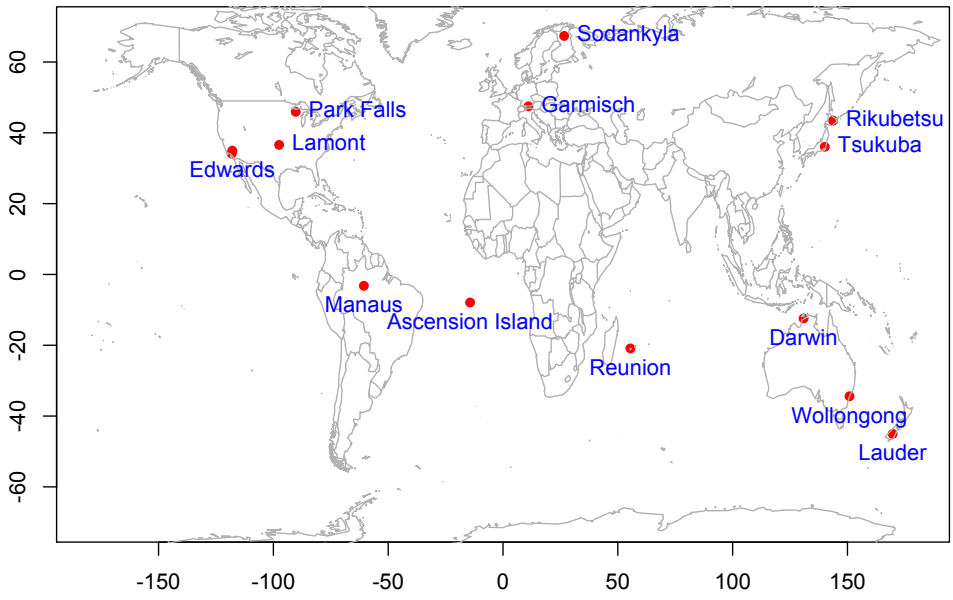

Figure 2: Locations of the 14 TCCON sites that were used to compare to column-averaged, data-fused CO_2 values.

and 'OCO2 Data Access' in our reference list.

The Total Carbon Column Observing Network (TCCON) consists of ground-based Fourier Transform Spectrometers that record direct solar spectra in the near-infrared. These spectra are then used to retrieve column-averaged abundances of atmospheric constituents including CO_2, CH_4, N_2O, HF, CO, and H_2O, which are directly comparable with the near-infrared column-averaged measurements from space-based instruments [Wunch et al., 2011]. Whereas ACOS and OCO-2 retrievals are susceptible to variability resulting from contamination by optically-thick clouds and aerosols that were missed by the cloud screening process [O'Dell et al., 2012], TCCON makes direct observation of the solar disk and hence is less sensitive to errors from scattered light [Crisp et al., 2012].

The TCCON sites sample in a diverse range of atmospheric states, which include tropical and polar, continental and maritime, polluted and clean, providing a valuable validation link between the space-based measurements and an extensive ground-based *in situ* network

[Wunch et al., 2011]. TCCON's column-averaged CO_2 data are in turn validated against integrated aircraft profiles [Washenfelder et al., 2006; Deutscher et al., 2010; Messerschmidt et al., 2011; Wunch et al., 2010] and have a precision and accuracy of ∼0.8 ppm [Wunch et al., 2010]. We obtained TCCON data at all sites that were operational in 2015 via the TCCON Data Archive (see TCCON Data Access). There were 27 of these sites, but only 14 produced data on days when both GOSAT and OCO-2 were operational in 2015. Their locations are presented in Figure 2. Fortunately, these 14 sites are fairly well distributed globally and include both continental and maritime regions as well as spanning a range of diverse geographical and ecological regimes.

For our demonstration, we obtained daily global ACOS and OCO-2 Level 2 data (using both the profile and column-averaged CO_2 estimates) and fused them to produce MSDF and SSDF estimates of column-averaged CO_2 at the TCCON locations. The scalar column-averaged CO_2 concentrations from ACOS and OCO-2 were fused together using SSDF, while the 20-dimensional CO_2 profiles for ACOS and OCO-2 were combined using MSDF. The MSDF profile estimate is convolved into a column-averaged CO_2 value using a linear combination pressure weighting function that is a function of the local specific humidity and the local acceleration due to gravity [see Appendix A of O'Dell et al., 2012]. We did not account for temporal dependence in the observed data, treating all the data within the same day as if they were observed at a fixed time point.

These two estimates were then compared to the daily median CO_2 values from TCCON. Since both the scalar (SSDF) and profile (MSDF) fusions are spatial only, our research illustrates most clearly whether the extra effort involved in doing a *multivariate* data fusion is worth the effort.

Both the GOSAT and OCO-2 instruments have long periods of observation punctuated with brief periods of non-operation for reasons of maintenance, thermal control, or spacecraft maneuvers. For each day in 2015, we performed data fusion and estimated column-averaged CO_2 at the TCCON sites if and only if both instruments produced observations on that day. There were 258 days in 2015 in which both instruments were operationally making observations. Typically, there are about 70,000 observations from OCO-2 per day, while GOSAT has about 2,000 observations per day. Figure 3 illustrates

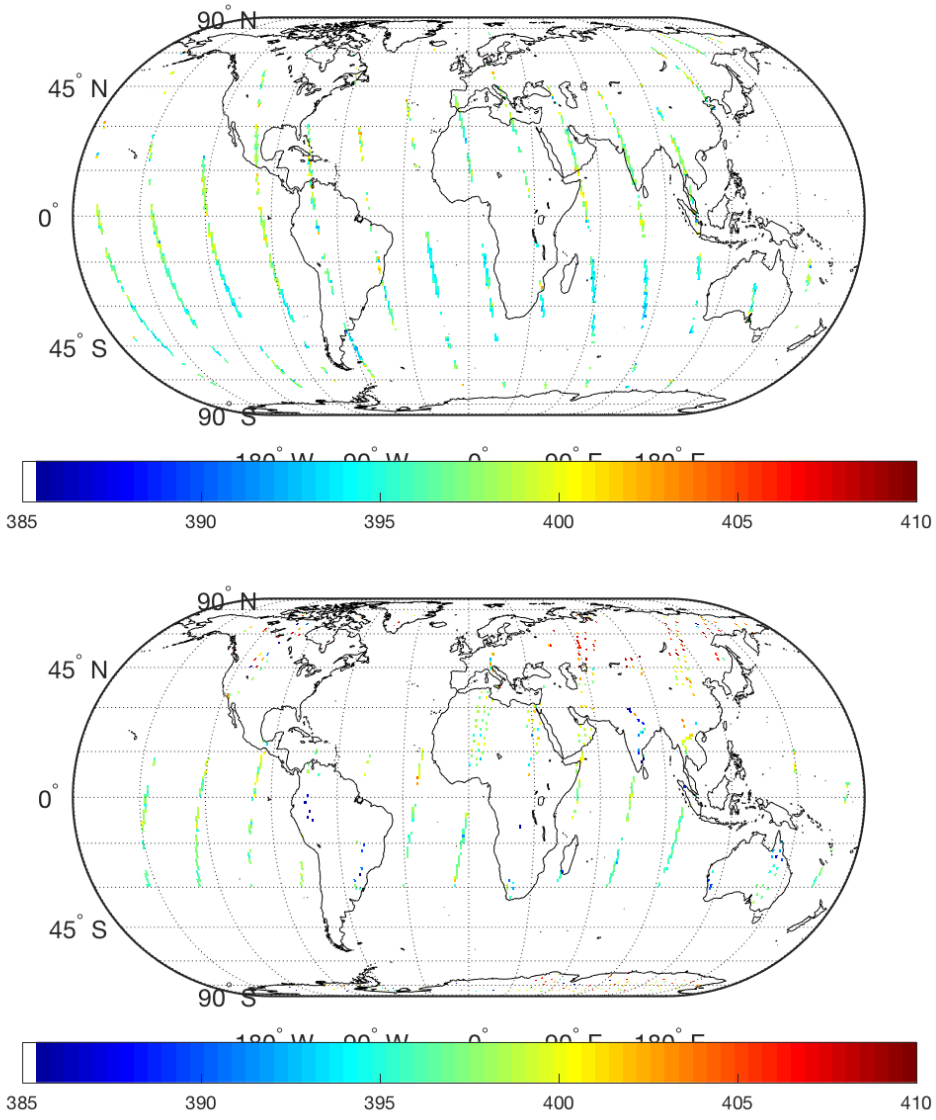

Figure 3: Retrieved column-averaged CO_2 data from OCO-2 (top panel) and ACOS (GOSAT) (bottom panel) on March 3, 2015.

the spatial patterns of each of the instruments on a particular day, March 3, 2015. Both maps show thin stripes of data running in a north-south direction, which are indicative of their polar, sun-synchronous orbits. Both instruments notably have sections of their observational swath missing, which is likely due to failure in the retrieval process because of contamination from clouds. Note that OCO-2's and GOSAT's observational patterns are both reinforcing and complementary. In regions such as the southern oceans, their overlapping coverage can contribute to lower uncertainties in the estimates. In other places such as the high-latitude Siberian tundra, one instrument (in this case, GOSAT) is able to make observations while the other cannot. However, in the North American continent, the opposite is true.

The measurement-error standard deviation of column-averaged CO2 from OCO-2 is between 1.5-3.5 ppm on land and 1.5-2.5 over ocean [Conner et al, 2016]. For simplicity, we use a constant value of 3 ppm. There is no corresponding validation study for the ACOS v3.5 data due to its recent production, so we assume that its measurement error standard deviation is also 3 ppm. These are the values that we use in the scalar SSDF of column-averaged CO_2.

To the best of our knowledge, there has been no attempt in the literature to quantify the biases and measurement-error variance of the individual height-specific CO_2 concentrations for either OCO-2 or ACOS. To obtain rough estimates of the height-specific measurement-error variance, $\{(\sigma_\epsilon^{(i,k)})^2\}$, we first compute the empirical variance of the CO_2-concentration data at each height. That is, we obtain an instrument-specific vector of empirical variance $\mathbf{v}^{(i)} \equiv (v^{(i,1)}, \ldots, v^{(i,K)})'$, where $v^{(i,k)}$ is the empirical variance of the elements in $\mathbf{Z}^{(i,k)}$. Similarly, we also compute a vector of the average pressure weighting function over the entire dataset, and we call it $\hat{\mathbf{h}}^{(i)}$. We then specify that the measurement error $\{(\sigma_\epsilon^{(i,k)})^2\}$ is approximately equal to $\beta^{(i)}\,\mathbf{v}^{(i)}$, where $\beta^{(i)}$ is a constant for the i-th instrument and is chosen such that $\beta^{(i)}\ \hat{\mathbf{h}}^{(i)\prime} \cdot \mathbf{v}^{(i)} = 3^2$. For GOSAT and ACOS, this coefficient is .35 and .31, respectively. Erring on the conservative side, we choose .35 as the coefficient for both instruments. Essentially, this procedure assigns about 35% of the marginal variability of CO_2 at each height as the corresponding measurement-error variance.

Regarding the bias, the OCO-2 instrument team uses a linear bias correction scheme

that estimates a bias in column-averaged CO_2 at each location as a function of the footprint index, the difference between the retrieved and the *a priori* surface pressure, the relative abundance of coarse aerosols, and the variation in the retrieved profile from that assumed in the prior [Osterman et al., 2016a]. The coefficients of this bias correction model are estimated by comparison to TCCON data and by using the southern hemisphere approximation, which assumes that the entire region from -25 to -60 latitude has minimal and negligible variation in the signal for column-averaged CO_2. ACOS v3.5 has a similar linear bias correction scheme for column-averaged CO_2 [Osterman et al., 2016b]. These bias correction schemes make it straightforward to remove the bias from the GOSAT and OCO-2 column-averaged CO_2 values before fusing them with SSDF. However, at present it is not clear how to translate this linear bias-correction scheme in column-averaged CO_2 into biases at each of the 20-dimensional CO_2 concentrations, as is required for applying MSDF. Since we are principally concerned with understanding the effect of "vertical" CO_2 dependence on inferences (i.e., MSDF versys SSDF), we opted not to apply bias correction on ACOS and OCO-2 data in either methods. We see later that we can detect this bias in the predictions, which is consistent for both MSDF and SSDF.

For the horizontal basis functions, we make use of multi-resolutional bisquare basis functions. Specifically, we use three resolutions whose centers are the centers of Level 1, 2, and 3, respectively, of an Icosahedral Snyder Equal Area Aperture 3 (ISEA3) grid, which is generated from the Discrete Global Grid software [Carr et al., 1998]. The three resolutions have 32, 92, and 272 basis centers, respectively, for a total of 396 functions. These basis centers are used for both MSDF and SSDF; and for the former a set of vertical basis functions based on cubic B-splines are also used, as discussed in Section 2.3. At each height, the EM algorithm's initial values for the fine-scale and the small-scale parameters are set as 10% and 90%, respectively, of the difference between the empirical variance minus the measurement-error variance [see Supplementary Materials of Nguyen et al., 2014, for more details].

Having performed the data fusions, we compare the SSDF and MSDF predictions to the TCCON daily median column-averaged CO_2 values, of which there are 1365 separate values. Table 1 displays the MSE, Mean, and Variance for the prediction errors. Looking at

Table 1: Table of MSE, Mean, and Variance of prediction errors

	SSDF	MSDF
MSE	18.07909	10.59946
Mean	1.263243	1.120176
Variance	16.48331	9.344668

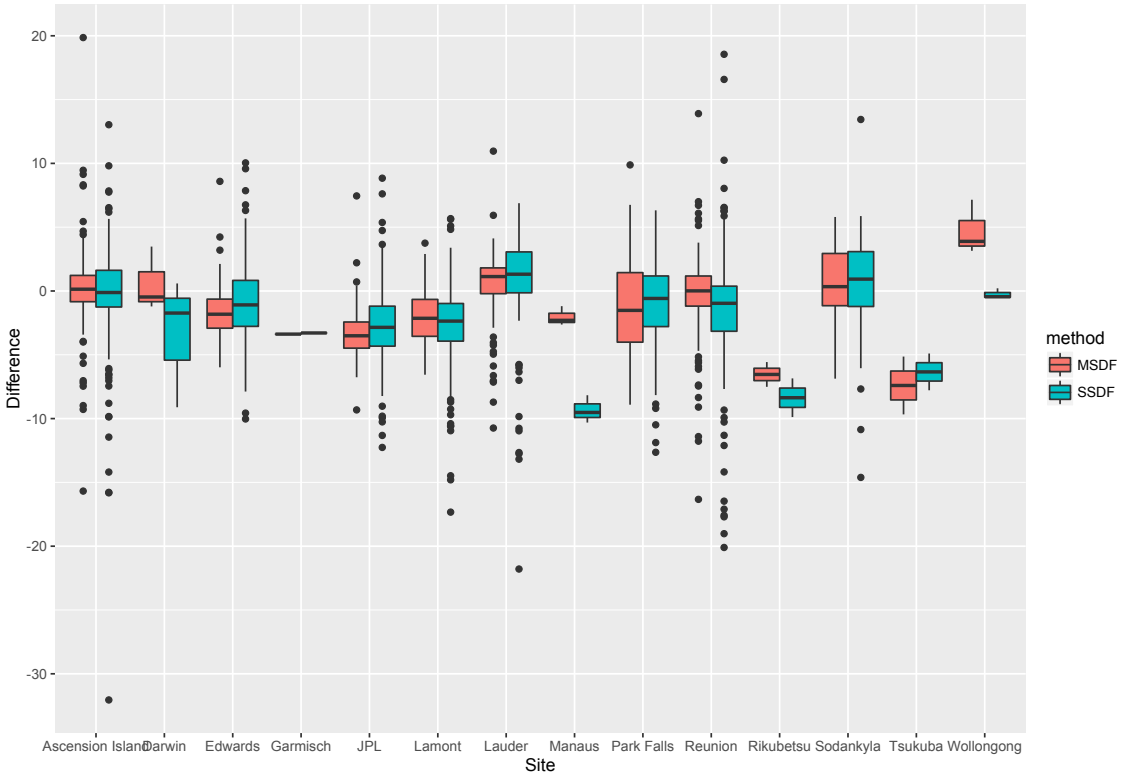

Figure 4: Boxplots of the "predicted minus TCCON" column-averaged CO_2, where "predicted" is given by SSDF (teal) and MSDF (red).

the MSE, we see that MSDF's value of 9.4 ppm^2 is substantially smaller than SSDF's value of 16.48 ppm^2, indicating that MSDF is able to take advantage of the vertical dependence to reduce the prediction MSE. This alone is a good metric for performance but, for better understanding of the improvement, we decomposed the MSE into the Mean and Variance in the bottom rows of Table 1. Both SSDF and MSDF have about the same average mean, also called bias, of about 1.2 ppm. This is not surprising, as we did not correct for the bias before carrying out data fusion. MSDF has a slightly smaller bias, but this is probably due to noise not signal.

While both methods do not differ very much with regard to the bias, they diverge greatly for the prediction-error variance. SSDF has a variance of 16.48 ppm^2 (standard deviation = 4.06 ppm), while MSDF comes with at a much improved variance of 9.34 ppm^2 (standard deviation = 3.06 ppm), or about a 43% reduction in prediction-error variance. We note that MSDF's prediction-error standard deviation of 3.06 ppm is consistent with validation studies showing biases of 1.5-3.5 ppm over land and 1.5-2.5 over ocean [Conner et al, 2016].

These metrics are averaged across the 14 TCCON sites, and in Figure 4 we display boxplots of the prediction errors at the individual sites. From this figure, it seems that the distribution of the errors within any particular TCCON site has generally the same mean for both methods, which is consistent with the results of Table 1. The only exceptions to this are Manaus, Brazil and Wollongong, Australia where SSDF and MSDF have large difference in their biases. These anomalies are likely due to the types of observations obtained near the two sites; Manaus is located within the Amazon, where the consistent cloudy conditions pose recurring retrieval problems for GOSAT and OCO-2. Wollongong, on the other hand, is at the southern edge of the Australian continent and has escarpment on one side and ocean on the other. Another remarkable feature from Figure 4 is that MSDF has consistently fewer outliers than SSDF, indicated on the boxplots by dots extending far past the "whiskers." Overall, Figure 4 shows what statistical theory suggests. Data fusion of column-averaged CO_2 based on MSDF does not improve upon the bias of an estimate, but it does substantially reduce its variability.

The improvement in prediction-error variance is most likely due to the fact that we are able to dampen more efficiently the large variability of the column-averaged estimates.

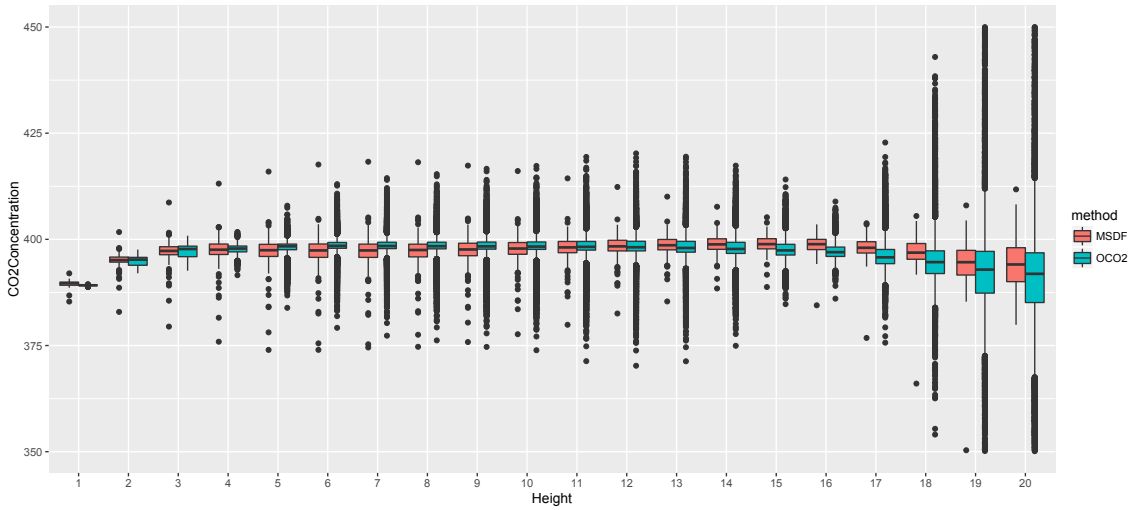

Figure 5: Boxplots of CO_2 concentrations for MSDF predictions and colocated OCO-2 retrieved profiles at Darwin. Heights on the horizontal axis are the index k from height $h_k; k = 1, \ldots, 20$.

Retrieving CO_2 concentrations at 20 different altitudes from photon counts is a difficult process and, in particular, the OCO-2 retrieval algorithm that yields the profile vetor of CO_2 concentrations has large variability in its estimates of the individual components. While the empirical measurement-error standard deviation of the column-averaged CO_2, compared to validation data, is about 3 ppm, the standard deviations of the individual components, particularly those at lower altitudes, can be much larger. We demonstrate this by collecting all retrieved profiles from OCO-2 within 100 km of the Darwin, Australia TCCON site and compare them to the fused estimates from MSDF at the same location. In Figure 5 we give the boxplots of CO_2 concentration as a function of height for both the OCO-2 and the MSDF predicted values. As can be seen, the MSDF predictions have much smaller variability compared to the OCO-2 profile data, indicating that our methodology is able to take advantage of the vertical dependence in addition to the horizontal dependence to dampen down the inherently large "measurement error" in the retrieved profile vectors.

5 Conclusions and discussion

This article examines a methodology for estimating column-averaged CO_2 from two remote sensing instruments– GOSAT and OCO-2. Both instruments produce multivariate profile CO_2 values and a column-averaged CO_2 value that is a linear combination of the profile with coefficients given by a deterministic pressure weighting function. Derivative uses of the CO_2 data, primarily within the flux-inversion community, tend to make use of the column-averaged CO_2 field only. We show here that there is significant information in the profile that is potentially lost when the two profiles are first convolved with the pressure weighting function. In the application section, we show that performing data fusion on the profile first, and then convolving the fused profile prediction into a column-averaged CO_2 value afterwards, has smaller mean square error.

The datasets in this application are fairly large, totaling about 70,000 observations per day. This is not a problem for MSDF, which has computational complexity that is linear with respect to the data size. Furthermore, MSDF does not assume isotropy or stationary in the data, which makes it appropriate for a wide range of applications. Note that having large amounts of data yields robust estimates of the spatial-dependence parameters through the EM algorithm. Hence, the methodology is well suited for data-rich applications, where the dependence on massive amounts of data is a strength rather than a weakness. This is particularly relevant to remote sensing, where a typical instrument returns thousands to hundreds of thousand of observations per day., and future instruments will be able to return millions of observations per day.

One direction for further research is incorporating temporal dependence into the model. We could allow the parameter $\boldsymbol{\eta}$ to evolve in discrete time steps according to a first-order autoregressive process as demonstrated in Cressie et al. [2010] and Nguyen et al. [2014]. In principle, that approach can be applied to the MSDF algorithm, but we anticipate that the main challenge would be the need to estimate the significantly larger propagation and innovation matrices. Further research is needed to balance the computational needs of a spatio-temporal MSDF against robustness of the estimation.

Finally, we note that we are not required to collapse the predicted profiles into a scalar column-averaged CO_2 value. Currently, this column-averaged CO_2 is used in conjunction

with atmospheric circulation models and ground-based column-averaged CO_2 to derive lower atmosphere CO_2 via data assimilation. This approach imposes physics-based atmospheric transport models, however the profile data fusions from MSDF could be used to obtain lower-atmosphere CO_2 directly by averaging over only the higher pressure levels. This could serve as a useful data-driven estimate of lower-atmosphere CO_2 that is complementary to that derived from data assimilation.

Acknowledgement

The research described in this paper was carried out in part by the Jet Propulsion Laboratory, California Institute of Technology, under a contract with NASA. It is supported by NASA's Earth Science Technology Office through its Advanced Information Systems Technology program. Noel Cressie's research is was partially supported by a 2015-2017 Discovery Grant, DP150104576, from Australian Research Council, and by NASA grant NNHII-ZDA001N-OCO2. ACOS and OCO-2 data are obtained from Goddard Earth Sciences Data and Information Services Center, operated by NASA, from the website http://daac.gsfc.nasa.gov/. TCCON data were obtained from the TCCON Data Archive, hosted by the Carbon Dioxide Information Analysis Center, from the website http://tccon.ornl.gov/.

References

ACOS Data Access (last access: May 2016). Goddard Earth Sciences Data and Information Services Center, WWW link: http://disc.sci.gsfc.nasa.gov/acdisc/documentation/ACOS.shtml.

Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008). Gaussian prediction process models for large spatial data sets. *Journal of the Royal Statistical Society, Series B*, 70(4):825–848.

Bartels, R. H., Beatty, J. C., and Barsky, B. A. (1998). *Hermite and Cubic Spline Interpolation*. Morgan Kaufmann, San Francisco, CA.

Berliner, L., Wikle, C., and Milliff, R. (1999). Multiresolution wavelet analyses in hierarchical Bayesian turbulence models. In *Bayesian Inference in Wavelet-Based Models*, eds. P. Miller and B. Vidakovic. Springer Lecture Notes in Statistics, No. 141. Springer-Verlag, New York, NY.

Calder, C. (2008). A Bayesian dynamic process convolution approach to modelling the point distribution of PM2.5 and PM10. *Envirometrics*, 19:39–48.

Carr, D., Kahn, R., Sahr, K., and Olsen, T. (1998). ISEA discrete global grids. *Statistical Computing and Statistical Graphics Newsletter*, 8(2/3):31–39.

Chevallier, F., Fisher, M., Peylin, P., Serrar, S., Bousquet, P., Bréon, F. M., Chédin, A., and Ciais, P. (2005). Inferring CO_2 sources and sinks from satellite observations: Method and application to TOVS data. *Journal of Geophysical Research: Atmospheres*, 110(D24), doi:10.1029/2005JD006390.

Cressie, N. (1996). Change of support and the modifiable areal unit problem. *Geographical Systems*, 3:159–180.

Cressie, N. and Johannesson, G. (2008). Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society, Series B*, 70(1):209–226.

Cressie, N., Shi, T., and Kang, E. L. (2010). Fixed rank filtering for spatio-temporal data. *Journal of Computational and Graphical Statistics*, 19(3):724–745.

Crisp, D., Boesch, H., Brown, L., Castano, R., Christi, M., Conner, B., Frankenberg, C., McDuffie, J., Miller, C., Natraj, V., O'Dell, C., O'Brien, D., Polonski, I., Oyafuso, F., Thompson, D., Toon, G., and Spurr, R. (2010). OCO (Orbiting Carbon Observatory): Level 2 Full Physics Retrieval Algorithm Theoretical Basis. Version 1.0 Rev. 4, November 10, 2010. JPL, NASA, Pasadena, CA.

Crisp, D., Fisher, B. M., O'Dell, C., Frankenberg, C., Basilio, R., Boesch, H., Brown, L. R., Castano, R., Connor, B., Deutscher, N. M., Eldering, A., Griffith, D., Gunson, M., Kuze, A., Mandrake, L., McDuffie, J., Messerschmidt, J., Miller, C. E., Morino, I., Natraj, V., Notholt, J., O'Brien, D. M., Oyafuso, F., Polonsky, I., Robinson, J., Salawitch, R.,

Sherlock, V., Smyth, M., Suto, H., Taylor, T. E., Thompson, D. R., Wennberg, P. O., Wunch, D., and Yung, Y. L. (2012). The ACOS CO_2 retrieval algorithm– Part II: Global XCO_2 data characterization. *Atmospheric Measurement Techniques*, 5:687–707.

Deutscher, N. M., Griffith, D. W. T., Bryant, G. W., Wennberg, P. O., Toon, G. C., Washenfelder, R. A., Keppel-Aleks, G., Wunch, D., Yavin, Y., Allen, N. T., Blavier, J.-F., Jiménez, R., Daube, B. C., Bright, A. V., Matross, D. M., Wofsy, S. C., and Park, S. (2010). Total column CO_2 measurements at Darwin, Australia; site description and calibration against in situ aircraft profiles. *Atmospheric Measurement Techniques*, 3(4):947–958.

Gruber, N., Gloor, M., Fletcher, S. E. M., Dutkiewicz, S., Follows, M., Doney, S. C., Gerber, M., Jacobson, A. R., Lindsay, K., Menemenlis, D., Mouchet, A., Mueller, S. A., Sarmiento, J. L., and Takahashi, T. (2009). Oceanic sources, sinks, and transport of atmospheric CO_2. *Global Biochemical Cycles*, 23, GB1005, doi:10.1029/2008GB003349.

Hamazaki, T., Kaneko, Y., Kuze, A., and Kondo, K. (2005). Fourier transform spectrometer for Greenhouse Gases Observing Satellite (GOSAT). Komar, G. J., Wang, J., and Kimura, T., editors, *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, 5659:73–80.

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

Hooten, M. B., Larsen, D. R., and Wikle, C. K. (2003). Predicting the spatial distribution of ground flora on large domains using a hierarchical Bayesian model. *Landscape Ecology*, 18:487–502.

Houghton, J. T., Ding, Y., Griggs, D. J., Noguer, M., van der Linden, P. J., Dai, X., Maskell, K., and Johnson, C. A. (2001). *Climate Change 2001: The Scientific Basis. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, UK.

Karion, A., Sweeney, C., Tans, P., and Newberger, T. (2010). Aircore: An innova-

tive atmospheric sampling system. *Journal of Atmospheric and Oceanic Technology*. doi:10.1175/2010JTECHA1448.1.

Katzfuss, M. and Cressie, N. (2011). Spatio-temporal smoothing and EM estimation for massive remote-sensing data sets. *Journal of Time Series Analysis*, 32(4):430–446.

Lindgren, F., Rue, H., and Lindstrom, J. (2011). An explicit link between Guassian fields and Gaussian Markov random fields: The stochastic partial different equation approach. *Journal of Royal Statistical Society, Series B*, 73(4):423–498.

McGuffie, K. and Henderson-Sellers, A. (1997). *A Climate Modelling Primer*. John Wiley and Sons, New York, NY.

Messerschmidt, J., Geibel, M. C., Blumenstock, T., Chen, H., Deutscher, N. M., Engel, A., Feist, D. G., Gerbig, C., Gisi, M., Hase, F., Katrynski, K., Kolle, O., Lavrič, J. V., Notholt, J., Palm, M., Ramonet, M., Rettinger, M., Schmidt, M., Sussmann, R., Toon, G. C., Truong, F., Warneke, T., Wennberg, P. O., Wunch, D., and Xueref-Remy, I. (2011). Calibration of TCCON column-averaged CO_2: the first aircraft campaign over European TCCON sites. *Atmospheric Chemistry and Physics*, 11(21):10765–10777.

Morino, I., Uchino, O., Inoue, M., Yoshida, Y., Yokota, T., Wennberg, P. O., Toon, G. C., Wunch, D., Roehl, C. M., Notholt, J., Warneke, T., Messerschmidt, J., Griffith, D. W. T., Deutscher, N. M., Sherlock, V., Connor, B., Robinson, J., Sussmann, R., and Rettinger, M. (2011). Preliminary validation of column-averaged volume mixing ratios of carbon dioxide and methane retrieved from GOSAT short-wavelength infrared spectra. *Atmospheric Measurement Techniques*, 4(6):1061–1076.

Nguyen, H., Cressie, N., and Braverman, A. (2012). Spatial statistical data fusion for remote sensing applications. *Journal of the American Statistical Association*, 107(499):1004–1018.

Nguyen, H., Katzfuss, M., Cressie, N., and Braverman, A. (2014). Spatio-temporal data fusion for very large remote sensing datasets. *Technometrics*, 56(2):174–185.

Nychka, D., Wikle, C. K., and Royle, J. (2002). Multiresolution models for nonstationary spatial covariance functions. *Statistical Modelling*, 2:315–331.

OCO2 Data Access (last access: May 2016). Goddard Earth Sciences Data and Information Services Center, WWW link: http://disc.sci.gsfc.nasa.gov/OCO-2.

O'Dell, C. W., Connor, B., Bösch, H., O'Brien, D., Frankenberg, C., Castano, R., Christi, M., Eldering, D., Fisher, B., Gunson, M., McDuffie, J., Miller, C. E., Natraj, V., Oyafuso, F., Polonsky, I., Smyth, M., Taylor, T., Toon, G. C., Wennberg, P. O., and Wunch, D. (2012). The ACOS CO2 retrieval algorithm Part 1: Description and validation against synthetic observations. *Atmospheric Measurement Techniques*, 5(1):99–121.

Osterman, G., Eldering, A., Avis, C., Chafin, B., O'Dell, C., Frankenberg, C., Fisher, B., Mandrake, L., Wunch, D., Granat, R., and Crisp, D. (2016a). OCO2 Data Product Users Guide, Operational L1 and L2 Data Versions 7 and 7R. Revision Date: Version G, June 30, 2016. WWW link: http://disc.sci.gsfc.nasa.gov/OCO-2/documentation/oco-2-v7/OCO2_DUG.V7.pdf.

Osterman, G., Eldering, A., Avis, C., O'Dell, C., Martinez, E., Crisp, D., Frankenberg, C., and Fisher, B. (2016b). ACOS Level 2 Standard Product Data Users Guide v3.5. Revision Date: Revision D, March 6, 2016. WWW link: http://disc.sci.gsfc.nasa.gov/OCO-2/documentation/gosat-acos/gosat-acosdoc/ACOS_v3.5_DataUsersGuide.pdf.

Rodgers, C. D. (2000). *Inverse Methods for Atmospheric Sounding : Theory and Practice*. World Scientific, River Edge, N.J.

Royle, J. and Wikle, C. (2005). Efficient statistical mapping of avian count data. *Ecological and Environmental Statistics*, 12:225–243.

Stein, M. L. and Jun, M. (2008). Nonstationary covariance models for global data. *Annals of Applied Statistics*, 2:1271–1289.

TCCON Data Access (last access: June 2013). TCCON Data Archive. WWW link: http://tccon.ornl.gov/.

Washenfelder, R. A., Toon, G. C., Blavier, J.-F., Yang, Z., Allen, N. T., Wennberg, P. O., Vay, S. A., Matross, D. M., and Daube, B. C. (2006). Carbon dioxide column abundances at the Wisconsin Tall Tower site. *Journal of Geophysical Research: Atmospheres*, 111(D22).

Wikle, C. K., Milliff, R. F., Nychka, D., and Berliner, L. M. (2001). Spatio-temporal hierarchical Bayesian modeling: Tropical ocean surface winds. *Journal of the American Statistical Association*, 96:382–397.

Wunch, D., Toon, G., Blavier, J., Washenfelder, R., Notholt, J., Connor, B., Griffith, D., Sherlock, V., and Wennberg, P. (2011). The Total Carbon Column Observing Network. *Philosophical Transactions of the Royal Society A*, 369(1943):2087–2112.

Wunch, D., Toon, G. C., Wennberg, P. O., Wofsy, S. C., Stephens, B. B., Fischer, M. L., Uchino, O., Abshire, J. B., Bernath, P., Biraud, S. C., Blavier, J.-F. L., Boone, C., Bowman, K. P., Browell, E. V., Campos, T., Connor, B. J., Daube, B. C., Deutscher, N. M., Diao, M., Elkins, J. W., Gerbig, C., Gottlieb, E., Griffith, D. W. T., Hurst, D. F., Jiménez, R., Keppel-Aleks, G., Kort, E. A., Macatangay, R., Machida, T., Matsueda, H., Moore, F., Morino, I., Park, S., Robinson, J., Roehl, C. M., Sawa, Y., Sherlock, V., Sweeney, C., Tanaka, T., and Zondlo, M. A. (2010). Calibration of the Total Carbon Column Observing Network using aircraft profile data. *Atmospheric Measurement Techniques*, 3(5):1351–1362.