MEASURES 2017 Data Fusion (v2)

Algorithm Theoretical Basis Document

November 23, 2020

ATBD version 0.95

Jet Propulsion Laboratory

California Institute of Technology [1]

[1]A portion of this research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration (80NM0018F0527).

1 Introduction

This Algorithm Theoretical Basis Document (ATBD) describes the theoretical basis for the algorithm used to generate the v2 fused data products for the Making Earth Science Data Records for Use in Research Environments project (MEaSUREs '17: Records of Fused and Assimilated Satellite Carbon Dioxide Observations and Fluxes from Multiple Instruments). This ATBD is divided into three parts. In Section 2, we describe the data sources (here the Orbiting Carbon Observatory 2 and the Greenhouse Gases Observing Satellite) along with some detail on data quality filtering. In Section 3, we present a theoretical justification for our fusion approach. Finally, in Section 4 we review the motivation and fundamentals of our methodology (kriging) and provide the implementation details.

2 Data Sources for Fusion

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).

GOSAT is a polar-orbiting satellite dedicated to the observation of carbon dioxide and methane, both major greenhouse gases, from space. It 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). NASA's Atmospheric CO_2 Observations from Space (ACOS) team uses the raw-radiance data from GOSAT to estimate the column-average CO_2 mole fraction in ppm, extending from the surface to the satellite over a base area corresponding to the instrument's footprint. In this article, we will be using GOSAT retrievals that are processed by the ACOS team to yield Level 2 column-average CO_2 data (see Crisp et al., 2012, for more details), which were available to us through NASA's Goddard Earth Sciences Data and Information Services Center. Hereafter, we refer to these as ACOS data. Since the ACOS product is produced at the Jet Propulsion

Laboratory by the same team behind the OCO-2 instruments, much of the retrieval characterization (e.g., priors, choice of pressure levels, forward models, etc.) are the same between the two products.

2.1 Data version and quality filter

For our fusion products, we use ACOS Version 9 data, which are produced by the Jet Propulsion Lab at NASA. These data are available at `https://oco2.gesdisc.eosdis.nasa.gov/data/GOSAT_TANSO_Level2/ACOS_L2_Lite_FP.9r/`. For the OCO-2 Level 2 data, we use the Version 10 data, which are available at `https://disc.gsfc.nasa.gov/datasets/OCO2_L2_Lite_FP_10r/summary`. The User Data Guide for ACOS V9 can be found at `https://docserver.gesdisc.eosdis.nasa.gov/public/project/OCO/ACOS_v9_DataUsersGuide.pdf`, and the Data User Guide for OCO-2 Version 10 can be found at `https://docserver.gesdisc.eosdis.nasa.gov/public/project/OCO/OCO2_OCO3_B10_DUG.pdf`.

Typically, OCO-2 and ACOS L2 data vary in retrieval quality due to different atmospheric conditions (e.g., contamination of the radiance by clouds or uncertainties in the atmospheric aerosols). Hence, the OCO-2 team recommends that the Level XCO2 data be filtered to eliminate potential 'bad' data. Here, we make use of the 'xco2_quality_flag' quality flag from the Lite products. From the OCO-2 Level 2 Data Quality Guide:

> "xco2_quality_flag [...] is simply a byte array of 0s and 1s. This filter has been derived by comparing retrieved XCO2 for a subset of the data to various truth

proxies, and identifying thresholds for different variables that correlate with poor data quality. It applies a number of quality filters based on retrieved or auxiliary variables that correlate with excessive XCO2 scatter or bias."

For the fusion product, we filter both ACOS and OCO-2 L2 product by selecting only values for which xco2_quality_flag == 0. Both data products employ a bias correction process, which is a post-processing algorithm that applies a small offset to each retrieved XCO2 value to correct for instrument biases. For our fusion, we make use of the bias-corrected XCO2 values from both ACOS and OCO-2 products.

2.2 Data fusion output modes

The OCO-2 instrument have three primary observation modes: glint, nadir, and target. The nadir mode consists of observations where the surface solar zenith angle is less than 85 degrees, and the glint mode consist of observation at latitudes where the solar zenith angle of the glint spot is less than 75 degrees. Finally, target mode consists of very localized observations are conducted over selected OCO-2 validation sites. The three modes differ in their quality and biases. They also differ in their spatial coverage. Nadir mode, for instance, is only collected over land, while glint mode can collect observations over both land and ocean.

It has been shown that the bias correction process for ACOS and OCO-2 still demonstrate residual bias, which depends on surface type, latitude, and scattering by aerosol Wunch et al. (2017). One significant factor in determining the residual bias is whether the surface is land or ocean. Therefore, many flux inversion studies opt for

assimilate the XCO2 data separately for land and ocean. Consequently, we stratify our fusion products into 4 different products, as seen in the table below: In the fusion

Table 1: Fusion output modes

Product	Description
Land Only	Uses only Land observations from ACOS and OCO-2 (Land Nadir + Land Glint)
Ocean Only	Uses only Ocean observations from ACOS and OCO-2 (Ocean Glint)
Land and Ocean	Uses all ACOS observations and OCO-2 Glint and Nadir modes
Target	Uses only Target observations from OCO-2

outputs, these different modes can be identified by the variable 'source_data_mode', which is an integer ranging from 1 to 4, where 'Land Only' = 1, 'Ocean Only' = 2, 'Land and Ocean' = 3, and 'Target' = 4.

3 Fusion approach

The data fusion approach based on kriging is well developed in the literature, specifically for Level 3 XCO2 generation (e.g. Nguyen et al., 2012). However, this MEaSUREs project is specifically geared towards producing data that could be incorporated into flux inversion studies, and hence the approach needs to be modified. The key difference between Level 3 XCO2 generation and flux inversion is the number of variables required for the outputs. The variables that the flux modelers require include the following: longitude, latitude, pressure levels (which varies at each foot-

print), pressure weighting functions, XCO2, time in UTC , prior mean, and column averaging kernel. Using the naming convention of the OCO-2 Lite files and the fusion output files, these variables are described in the table below:

Table 2: Variables required for flux inversion

Name	Dimension	Description
longitude	1x1	The longitude at the center of the sounding field-of-view
latitude	1x1	The latitude at the center of the sounding field-of-view
xco2	1x1	The bias-corrected XCO2 (in units of ppm)
time	1x1	The time of the sounding in seconds since 1970-01-01
co2_profile_apriori	20x1	The prior mean profile of CO2 in ppm
xco2_averaging_kernel	20x1	The normalized column averaging kernel for the retrieved XCO2
pressure_levels	20x1	The retrieval pressure level grid for each sounding in hPa
pressure_weight	20x1	The pressure weighting function on levels used in the retrieval

The data fusion approach in Nguyen et al. (2012) provides a framework for fusing scalar quantities such as XCO2 or aerosols, however, the presence of multivariate profiles (e.g., co2_profile_apriori, xco2_averaging_kernel) requires extra care. In principle, one approach would be to apply the scalar fusion method to all fields in Table 1 individually for each pressure level. However, we note that certain variables such as time and co2_profile_apriori are deterministic functions, and as such do not conform to assumptions of a statistical spatial dependence model (i.e., a semi-variogram model

such as that described in 4.1).

Our approach in this section is a generalization of a common technique called the '10 second average' that is used in many flux inversions (e.g., Basu et al., 2018). There, the approach is to simply take the average of all the fields in Table 1 in 10 seconds intervals. Here, we generalize the approach by using a spatial statistics approach (specifically local kriging) to compute a weighted vector of coefficients, which we then apply to the fields in Table 1 to get the fused outputs.

The Bayesian Optimal Estimation framework, as formalized in Rodgers (2000), is popular for inverse problems in remote sensing and it is the method of choice for OCO-2 retrievals (Crisp et al., 2010). In this section, we will review the background of Optimal Estimation (OE), and then discuss how our fusion approach is consistent with the OE formulation. For ease of exposition, we will consider the inverse problem where the forward model is linear.

3.1 Background

Consider the case where an N-dimensional radiance vector $\mathbf{y}$ is related to the r-dimensional (hidden) true state $\mathbf{x}$ by the following data model:

$$\mathbf{y} \;=\; \mathbf{F}(\mathbf{x}) + \boldsymbol{\epsilon}, \tag{1}$$

where $\mathbf{F}(\,\cdot\,)$ is the forward model, $\mathbf{x}$ is the r-dimensional true state with true mean $\mathbf{x}_T$ and true covariance matrix $\mathbf{S}_T$, and $\boldsymbol{\epsilon}$ is the N-dimensional measurement-error vector with mean $\mathbf{0}$ and covariance matrix $\mathbf{S}_\epsilon$. That is, $\mathbf{x} \sim N_r(\mathbf{x}_T, \mathbf{S}_T)$ and $\boldsymbol{\epsilon} \sim N_N(\mathbf{0}, \mathbf{S}_\epsilon)$.

The true mean $\mathbf{x}_T$ is defined at a set of r pressure levels $\mathbf{p}$ that will change from observation to observation. Since we assume that the forward model is linear, the general data model in (1) becomes,

$$\mathbf{y} \;=\; \mathbf{c} + \mathbf{K}\mathbf{x} + \boldsymbol{\epsilon},$$

where $\mathbf{K}$ is the Jacobian of the forward model, and $\mathbf{c}$ is an N-dimensional constant vector.

Without lack of generality, we can assume that $\mathbf{c} = \mathbf{0}$ (since $\mathbf{c}$ is known and hence in principle could be subtracted from $\mathbf{y}$). Our data model then becomes,

$$\mathbf{y} \;=\; \mathbf{K}\mathbf{x} + \boldsymbol{\epsilon}. \tag{2}$$

Rodgers (2000) proposed a loss function that is the negative logarithm of the posterior distribution of $\mathbf{x}$ given $\mathbf{y}$, dropping constant terms,

$$L(\mathbf{x}) \equiv -2\log P(\mathbf{x}|\mathbf{y}) = (\mathbf{y} - \mathbf{K}\mathbf{x})' \, \mathbf{S}_\epsilon^{-1} \, (\mathbf{y} - \mathbf{K}\mathbf{x}) - (\mathbf{x} - \mathbf{x}_T)' \, \mathbf{S}_T^{-1} \, (\mathbf{x} - \mathbf{x}_T). \tag{3}$$

The maximum a posteriori solution (also the posterior mean in our linear forward model case) is then given by,

$$\hat{\mathbf{x}} = \mathbf{x}_T + \mathbf{G}_T(\mathbf{y} - \mathbf{K}\mathbf{x}_T), \tag{4}$$

where $\mathbf{G}_T$ is called the gain matrix and is given by $\mathbf{G}_T = (\mathbf{S}_T^{-1} + \mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K})^{-1}\mathbf{K}'\mathbf{S}_\epsilon^{-1}$.

The uncertainty on $\hat{\mathbf{x}}$ is then given by,

$$\mathbf{\Sigma}_T \equiv \mathrm{Var}(\hat{\mathbf{x}} - \mathbf{x}) = (\mathbf{S}_T^{-1} + \mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K})^{-1}. \tag{5}$$

The relationship in (4) is sometimes expressed as a relationship between the true state, the retrieved state, and the prior mean state as follows,

$$\hat{\mathbf{x}} = \mathbf{x}_T + \mathbf{A}_T(\mathbf{x} - \mathbf{x}_T) + \epsilon_x, \tag{6}$$

where $\mathbf{A}_T \equiv (\mathbf{S}_T^{-1} + \mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K})^{-1}\mathbf{K}'\mathbf{S}_\epsilon^{-1}\mathbf{K}$ is a $r \times r$ matrix called the averaging kernel. Typically, the column CO_2 amount is not used in flux inversion, and a linear combination is applied to this state vector to compute what is called the total-column CO_2 (XCO_2). That is, $x_{xco2} = \mathbf{h}'\hat{\mathbf{x}}$, where $\mathbf{h}$ is the pressure weighting vector. Note that the averaging kernel can be interpreted as a derivative with respective to the true state $\mathbf{x}$ as follows,

$$\mathbf{A}_T = \frac{d\hat{\mathbf{x}}}{d\mathbf{x}} \tag{7}$$

Finally, flux inversion also requires a quantity called column averaging kernel $\mathbf{c}$, which is given as

$$\mathbf{c} = (\mathbf{h}'\mathbf{A}_T) \oslash \mathbf{h}', \tag{8}$$

where $\oslash$ denotes element-wise division.

3.2 Fusion approach

Let's consider the case of OCO-2 and GOSAT, both of which provide all the information required above. The main difficulty is that the data fusion methodology we develop should also produce *fused* quantities for all the variables above. We have explored methodologies for fusing scalars (e.g., Nguyen et al., 2012), but some of these variables (e.g., pressure levels, prior means, and column averaging kernel) are r-dimensional vectors.

We take a similar approach as the 10-seconds average (Basu et al., 2018) using a variation of local kriging based on the work of Hammerling et al. (2012). Since the fused estimate is a linear combination of the two individual datasets, we can perform a scalar data fusion on XCO_2, and then use the linear coefficients therefrom to form linear combination of the remaining quantities. For instance, let the XCO2 data vector and time vector be indicated by $\mathbf{Z}_i$ and $\mathbf{T}_i$ where $i = 1$ for OCO-2 and $i = 2$ for ACOS. Our fused estimate for XCO_2 is $Z_F = \mathbf{a}_1'\mathbf{Z}_1 + \mathbf{a}_2'\mathbf{Z}_2$, and the fused UTC time would then be $T_F = \mathbf{a}_1'\mathbf{T}_1 + \mathbf{a}_2'\mathbf{T}_2$, where $\mathbf{a}_1'$ and $\mathbf{a}_2'$ are derived from the XCO_2 fusion.

This formulation is consistent with the derivation of averaging kernels above. That is, if we consider the original space of the state vector $\mathbf{x}$,

$$\begin{aligned}\mathbf{x}_F &= a_{11}\hat{\mathbf{x}}_{11} + a_{12}\hat{\mathbf{x}}_{12} + \ldots + a_{1N_1}\hat{\mathbf{x}}_{1N_1} + \ldots \\ &\quad + a_{21}\hat{\mathbf{x}}_{21} + a_{22}\hat{\mathbf{x}}_{22} + \ldots + a_{2N_2}\hat{\mathbf{x}}_{2N_2},\end{aligned}$$

Then the averaging kernel of this fused estimate is given as

$$\begin{aligned}
\mathbf{A}_F &= \frac{d\hat{\mathbf{x}}_F}{d\mathbf{x}} \\
&= a_{11}\frac{d\hat{\mathbf{x}}_{11}}{d\mathbf{x}} + a_{12}\frac{d\hat{\mathbf{x}}_{12}}{d\mathbf{x}} + \ldots + a_{1N_1}\frac{d\hat{\mathbf{x}}_{1N_1}}{d\mathbf{x}} + \ldots \\
&\quad + a_{21}\frac{d\hat{\mathbf{x}}_{21}}{d\mathbf{x}} + a_{22}\frac{d\hat{\mathbf{x}}_{22}}{d\mathbf{x}} + \ldots + a_{2N_2}\frac{d\hat{\mathbf{x}}_{2N_2}}{d\mathbf{x}}, \\
&= a_{11}\mathbf{A}_{11} + a_{12}\mathbf{A}_{12} + \ldots + a_{1N_1}\mathbf{A}_{1N_1} + \ldots \\
&\quad + a_{21}\mathbf{A}_{21} + a_{22}\mathbf{A}_{22} + \ldots + a_{2N_2}\mathbf{A}_{2N_2}. \qquad (9)
\end{aligned}$$

The result in (9) indicates that our generalization of the '10-seconds-average' approach is consistent with the Optimal Estimation interpretation of the fields in Table 2. This simplifies the fusion problem to that of optimally fusing the XCO2 fields. The theoretical basis for fusing XCO2 based on their geospatial dependence is well-explored in the literature (e.g., Hammerling et al., 2012; Nguyen et al., 2012). In the next section we will describe the motivation and implementation details for the fusion of XCO2 from OCO-2 and ACOS.

4 Kriging equations

Here, we will take the approach of Hammerling et al. (2012) and fuse the XCO2 field ('/xco2' from the OCO-2 Lite product and the ACOS product) using a form of local kriging based on the exponential semivariogram. For ease of reference, we provide a review of kriging below.

Assume that we have observed CO2 data in the following form:

$$\begin{aligned} \mathbf{Z} &= (Z(\mathbf{s}_1), Z(\mathbf{s}_2), \ldots, Z(\mathbf{s}_N))', \\ Z(\mathbf{s}_i) &= Y(\mathbf{s}_i) + \epsilon(\mathbf{s}_i), \end{aligned}$$

where, for simplicity of notation, we assume that the data vector $\mathbf{Z}$ consist of both the OCO-2 and ACOS data concatenated together. Under this formulation, the measurement error process $\epsilon(\mathbf{s}_i)$ would sample from different distributions depending on whether $\mathbf{s}_i$ is an ACOS or OCO-2 observation. The (linear unbiased) optimal interpolation can be written as

$$\hat{Y}(\mathbf{s}_0) = \mathbf{a}_0'\mathbf{Z}$$

where $\mathbf{a}_0$ is a N-dimensional vector of kriging coefficients at location $\mathbf{s}$.

We wish to find the vector $\mathbf{a}$ that minimizes,

$$\begin{aligned} E(Y(\mathbf{s}) - \hat{Y}(\mathbf{s}))^2 &= \mathrm{Var}(Y(\mathbf{s}) - \mathbf{a}_0'\,\mathbf{Z}), \\ &= \mathrm{Var}(Y(\mathbf{s})) - 2\mathbf{a}_0'\,\mathrm{Cov}(\mathbf{Z}, Y(\mathbf{s})) + \mathbf{a}_0'\,\mathrm{Var}(\mathbf{Z})\,\mathbf{a}_0, \qquad (10) \end{aligned}$$

with respect to $\mathbf{a}_0$, subject to the unbiasedness constraint,

$$1 = \mathbf{a}_0'\mathbf{1},$$

Note that this vector of kriging coefficient $\mathbf{a}_0$ is precisely the one required for forming

linear combinations of the fields in Table 1 (also see Section 3.2). We can solve the minimization problem above for the optimal $\mathbf{a}_0$ using the method of Lagrange multiplier, which gives the following matrix equation:

$$\begin{pmatrix} C_{11} & C_{12} & \dots & C_{1N} & 1 \\ C_{21} & C_{22} & \dots & C_{1N} & 1 \\ \vdots & \vdots & \ddots & \vdots & 1 \\ C_{N1} & C_{N2} & \dots & C_{NN} & 1 \\ 1 & 1 & \dots & 1 & 0 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_N \\ \lambda \end{pmatrix} = \begin{pmatrix} C_{10} \\ C_{20} \\ \vdots \\ C_{N0} \\ 1 \end{pmatrix} \tag{11}$$

where $C_{ij} = \mathrm{Cov}(\mathbf{s}_i, \mathbf{s}_j)$, $\mathbf{a}_0 = (a_1, \dots, a_N)$, and λ is the Langrange multiplier. In our implementation of the data fusion, we prefer to use an alternative measure of spatial dependence called semi-variogram, which is defined between any two location $\mathbf{s}_i$ and $\mathbf{s}_j$ as

$$\begin{aligned} \gamma(\mathbf{s}_1, \mathbf{s}_2) &= \frac{1}{2} E(Z(\mathbf{s}_i) - Z(\mathbf{s}_j))^2 \\ &= \frac{1}{2} \mathrm{Var}(Z(\mathbf{s}_i) - Z(\mathbf{s}_j)). \end{aligned}$$

It is easy to see that the semi-variogram is related to the covariance by the following:

$$\begin{aligned} \gamma(\mathbf{s}_1, \mathbf{s}_2) &= \frac{1}{2} \mathrm{Var}(Z(\mathbf{s}_i) - Z(\mathbf{s}_j)) \\ &= \frac{1}{2}(C_{ii} + C_{jj}) - C_{ij}. \end{aligned} \tag{12}$$

Substituting (12) into (13), we get the following expression in terms of the semi-

variograms:

$$\begin{pmatrix} \gamma(\mathbf{s}_1,\mathbf{s}_1) & \gamma(\mathbf{s}_1,\mathbf{s}_2) & \dots & \gamma(\mathbf{s}_1,\mathbf{s}_N) & 1 \\ \gamma(\mathbf{s}_2,\mathbf{s}_1) & \gamma(\mathbf{s}_2,\mathbf{s}_2) & \dots & \gamma(\mathbf{s}_2,\mathbf{s}_N) & 1 \\ \vdots & \vdots & \ddots & \vdots & 1 \\ \gamma(\mathbf{s}_N,\mathbf{s}_1) & \gamma(\mathbf{s}_N,\mathbf{s}_2) & \dots & \gamma(\mathbf{s}_N,\mathbf{s}_N) & 1 \\ 1 & 1 & \dots & 1 & 0 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_N \\ \lambda \end{pmatrix} = \begin{pmatrix} \gamma(\mathbf{s}_1,\mathbf{s}_0) \\ \gamma(\mathbf{s}_2,\mathbf{s}_0) \\ \vdots \\ \gamma(\mathbf{s}_N,\mathbf{s}_0) \\ 1 \end{pmatrix} \tag{13}$$

The solution to the minimization problem in (10) can easily be found by solving (13) using matrix inversion. This provides the vector of kriging coefficient $\mathbf{a}_0$ required by Section 3.2. In this section we have discussed single-instrument interpolation, and the extension to multi-instrument fusion is straightforward by combining data from multiple instruments into a single meta-dataset (e.g., Nguyen et al., 2012).

4.1 Implementation

As described in the previous section, we use local kriging to obtain kriging coefficients for the XCO2 field, which we then apply to all remaining fields (e.g., longitude, latitude, pressure levels (which varies at each footprint), pressure weighting functions, XCO2, time in UTC , prior mean, and column averaging kernel) to obtain the fused outputs. The local neighborhood that we consider is a circular region of radius of 300 km. That is, for every fusion location $\mathbf{s}_0$, we search for all available ACOS and OCO-2 data within 300 km in the same day and use that as our fusion data. The

semi-variogram model that we use is the exponential model, which has the form

$$\gamma(\mathbf{s}_i, \mathbf{s}_j) = (s - n_{ij}) \left(1 - e^{-|\mathbf{s}_i - \mathbf{s}_j|/(r)}\right) + I(\mathbf{s}_i \neq \mathbf{s}_j) n_{ij} \tag{14}$$

where I($\mathbf{s}_i \neq \mathbf{s}_j$) is an indicator function that returns 1 if $\mathbf{s}_i \neq \mathbf{s}_j$, and 0 otherwise; $n_{ij} = (\beta(\mathbf{s}_i) + \beta(\mathbf{s}_j))/2$, and that

$$\beta(\mathbf{s}_i) = \begin{cases} n_A & \text{if } \mathbf{s}_i \text{ is an ACOS observation} \\ n_O & \text{otherwise} \end{cases} \tag{15}$$

where n_A and n_O are the ACOS and OCO-2 nugget terms; and r and s are the sill and range, respectively. Note here that we are employing an exponential variogram function, so the r term here does not fit the typical understanding of the range as the distance at which the variogram becomes level. However, conventional wisdom indicate that we can multiply the term r by 3 to get an approximation of the distance at which observations are considered independent. Since both instruments are observing total column dioxide (XCO2), we assume that the range and sill terms are identical for both instruments. Current literature indicate that OCO-2 and GOSAT have different instrument bias and measurement error (e.g., Wunch et al., 2017; Inoue et al., 2013). In our implementation we assume that the instrument-dependent bias has been removed by the bias-correction process (see OCO-2 and ACOS Data User Guide). The existing literature indicates that GOSAT typically has higher measurement-error variability, and hence here we model that by assuming that the nugget term for ACOS is 1.5 times that of OCO-2. That is, we assume that $n_A = 1.5\, n_O$.

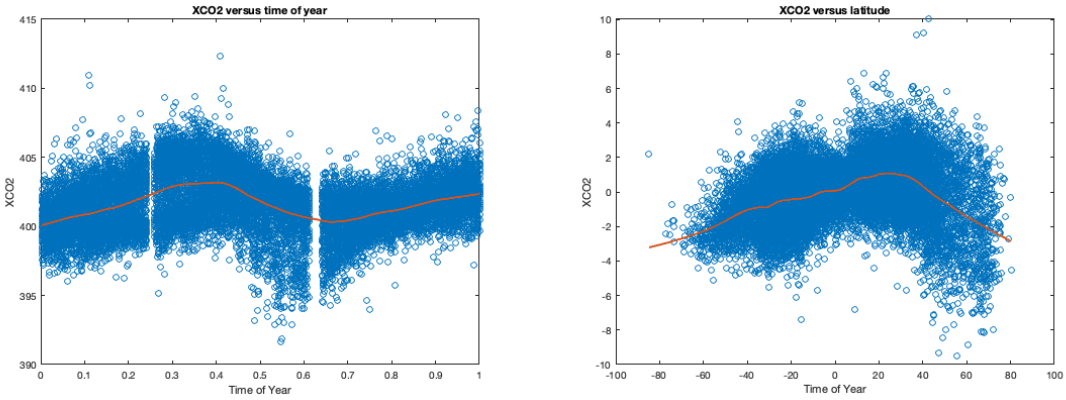

Figure 1: Left: a scatterplot of XCO2 versus time of year for 2016. Right: a scatterplot of the time-detrended residuals versus latitude. The red lines indicate a smoothed fit using locally-weight regression with span = 10%.

Solving for the weighting coefficients in (13) for any arbitrary fusion location requires the variogram parameters $\{n, r, n_O\}$. Here, we opt to construct a 'climatology' of $\{n, r, n_O\}$ as function latitude and longitude. To do this, we gathered all XCO2 data from the years 2015-2017. Simple scatterplots of the XCO2 against time of year and latitude indicate that there are time and latitude-dependent non-linear trends that need to be removed before the variogram estimation process. Therefore, we removed these trends using a step-wise process where we first fit a locally-weighted regression line (lowess; Cleveland and Devlin, 1988) against time, and then removing the smoothed value from every observed value. This forms a set of time-detrended residuals, on which we repeat the procedure to obtain another lowess fit versus latitude. Once again, this latitude-dependent trend is removed to produce a detrended dataset for variogram estimation. The scatterplots and lowess fits against time of year and latitude are shown in Figure 1.

Having detrended the XCO2 data, for each grid point we gather all data within 400 km and computed the variogram parameters using weighted least-squares (for details, see Cressie, 1985). In general, the global climatology tends to have a nugget within the range [.5 and 1], the effective range (defined as $3r$) within the range [100 km, 300 km], and the sill between [.5, 4].

4.2 Workflow

Having constructed the variogram parameter climatology, the workflow for generating the fused product is as below:

1. For every day, divide the globe into a $.5^\circ \times .5^\circ$ output grid
2. Filter the OCO-2 and ACOS data as described in Section 2.1 and subset them by one of four modes in Table 1.
3. For every fusion location $\mathbf{s}_0$, search for all ACOS and OCO-2 data within the same day and within 300 km radius.
    - If there are less than 5 retrieved data points from Step 3, proceed to the next fusion location.
4. Set the variogram parameters from the climatology (Section 4.1) using climatology in Section 4.1.
5. Compute the optimal fusion coefficient vector $\mathbf{a}_0$ by solving (13).
6. Compute linear combinations of the fields in Table 2 using coefficient vector $\mathbf{a}_0$.

7. Repeat Step 2-6 for other output modes.

4.3 Inflation factors

Flux inversion studies often make the assumption that the input data (here, either the Level OCO-2 retrievals, 10-seconds averages, or our fusion product) are statistically independent of one another. One natural consequence of this assumption is that the 'information content' of a product depends on the size of the dataset. We currently are studying various choices of 'inflation factors', which would allow flux modelers to normalize the information content to make it more comparable across different products. In this version, we provide a preliminary metric based on the ratio of the variance of the mean estimates.

Consider the flux inversion cost function, which is given as follow:

$$L_s = (\mathbf{z} - \mathbf{Hs})^T \mathbf{R}^{-1} (\mathbf{z} - \mathbf{Hs}) + (\mathbf{s} - \mathbf{s}_p)^T \mathbf{Q}^{-1} (\mathbf{s} - \mathbf{s}_p)$$

where $\mathbf{z}$ is the retrieval observations, $\mathbf{H}$ is the transport operator, $\mathbf{s}$ is the flux, $\mathbf{R}$ is the covariance matrix for $\mathbf{z}$, and $\mathbf{s}_p$ and $\mathbf{Q}$ are the prior mean vector and the prior covariance matrix for the flux, respectively. $\mathbf{R}$ is typically non-diagonal, but flux inversion often uses only the diagonal components of $\mathbf{R}$ (due to the independence assumption). We denote the diagonal component of $\mathbf{R}$ as $\mathbf{R}_D$. The arithmetic mean of $\mathbf{z}$ is given by

$$m = \mathbf{wz},$$

where $\mathbf{w} = (1, 1, \ldots, 1)/N$, and N is the number of observations in $\mathbf{z}$. The variance

of this estimate is given by

$$\sigma^2 = \mathbf{wRw}',$$

Here, we estimated the inflation factor as the ratio of the variance of the mean estimate (with spatial dependence) to that arising from the independence assumption. The inflation factor is given by

$$c = \frac{\sigma^2}{\sigma_D^2} = \frac{\mathbf{wRw}'}{\mathbf{wR}_D\mathbf{w}'} = \frac{\sum_i \sum_j \mathbf{R}^{ij}}{\sum_i \sum_j \mathbf{R}_D^{ij}},$$

where $\mathbf{R}^{ij}$ denotes the element in the i-th row and j-th column of the matrix $\mathbf{R}$. In our data product, these inflation factors are calculated separately for land and ocean fused estimates, then their values are given in 'land_inflation_factor' and 'ocean_inflation_factor.'

References

Basu, S., Baker, D. F., Chevallier, F., Patra, P. K., Liu, J., and Miller, J. B. (2018). The impact of transport model differences on co 2 surface flux estimates from oco-2 retrievals of column average co 2. *Atmospheric Chemistry & Physics*, 18(10).

Cleveland, W. S. and Devlin, S. J. (1988). Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American statistical association*, 83(403):596–610.

Cressie, N. (1985). Fitting variogram models by weighted least squares. *Journal of the International Association for Mathematical Geology*, 17(5):563–586.

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., Bösch, 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 CO2 retrieval algorithm– Part II: Global XCO2 data characterization. *Atmospheric Measurement Techniques*, 5:687—-707.

Hammerling, D. M., Michalak, A. M., O'Dell, C., and Kawa, S. R. (2012). Global co2 distributions over land from the greenhouse gases observing satellite (gosat). *Geophysical Research Letters*, 39(8).

Inoue, M., Morino, I., Uchino, O., Miyamoto, Y., Yoshida, Y., Yokota, T., Machida, T., Sawa, Y., Matsueda, H., Sweeney, C., et al. (2013). Validation of xco 2 derived from swir spectra of gosat tanso-fts with aircraft measurement data. *Atmospheric Chemistry and Physics*, 13(19):9771–9788.

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.

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., Na-

traj, 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.

Rodgers, C. D. (2000). *Inverse methods for atmospheric sounding: theory and practice*, volume 2. World Scientific.

Wunch, D., Wennberg, P. O., Osterman, G., Fisher, B., Naylor, B., Roehl, C. M., O'Dell, C., Mandrake, L., Viatte, C., Kiel, M., et al. (2017). Comparisons of the orbiting carbon observatory-2 (oco-2) x co2 measurements with tccon.