Space-time Data Fusion for Remote Sensing Applications

A. Braverman[a], H. Nguyen[a], and N. Cressie[b]

[a] Jet Propulsion Laboratory, California Institute of Technology, Mail Stop 306-463, 4800 Oak Grove Dr., Pasadena, CA 91109, USA – (Amy.Braverman, Hai.Nguyen)@jpl.nasa.gov

[b] Department of Statistics, The Ohio State University, 1958 Neil Ave., Room 404, Columbus, OH 43210, USA – ncressie@stat.osu.edu

Abstract- NASA has been collecting massive amounts of remote sensing data about Earth's systems for more than a decade. Missions are selected to be complementary in quantities measured, retrieval techniques, and sampling characteristics, so these datasets are highly synergistic. To fully exploit this, a rigorous methodology for combining data with heterogeneous sampling characteristics is required. For scientific purposes, the methodology must also provide quantitative measures of uncertainty that propagate input-data uncertainty appropriately. We view this as a statistical inference problem. The true but not-directly-observed quantities form a vector-valued field continuous in space and time. Our goal is to infer those true values or some function of them, and provide to uncertainty quantification for those inferences. We use a spatio-temporal statistical model that relates the unobserved quantities of interest at point-level to the spatially aggregated, observed data. We describe and illustrate our method using CO2 data from two NASA data sets.

Keywords: data fusion, spatio-temporal statistics, uncertainty quantification, massive datasets, carbon dioxide.

1. INTRODUCTION

The motivation for this work is the need to combine data from two remote sensing instruments to paint a complete and quantitative picture of the distribution of carbon dioxide (CO2) in the lower part of Earth's atmosphere. CO2 enters and leaves the atmosphere only near the surface, and so monitoring changes in this important greenhouse gas near the surface may shed light on sources and sinks of CO2 in the Earth's system. However, no instrument observes everywhere all the time so the best way to get a complete global picture is to combine information from multiple sources. In addition, different satellite instruments use different technologies and have different strengths and weaknesses. Combining their data provides the added advantage of capitalizing on complementary strengths. Finally, if the combined data are to be useful for scientific analyses and policy making, quantitative uncertainty measures must be provided.

In this article, we show how mid-tropospheric CO2 data from NASA's Atmospheric Infrared Sounder (AIRS) can be combined with total column CO2 data from NASA's Atmospheric Carbon Dioxide Observations from Space (ACOS; based on data from Japan's Greenhouse Gases Observing Satellite (GOSAT)) to estimate CO2 in the lower atmosphere. In Section 2, we describe our methodology, and in Section 3 we provide estimates and their uncertainties of lower atmosphere CO2 over the continental US for 15 days in June 2009. We conclude with a discussion in Section 4.

2. SPATIO-TEMPORAL DATA FUSION

Consider two different remote sensing instruments' views of a spatial field at a single instant in time as shown in the top panel of Figure 1. The two instruments discretize the scene differently and add measurement errors with different biases and variances to the discretized pixel values. The instruments will also typically have different patterns of missingness, as shown by the black pixels. Presented with only the middle images in Figure 1, could we infer the true fields shown on the left? Could we infer the true fields from the images on the right alone? The answer is yes in both cases if we can rigorously account for: 1) the fact that pixel values are spatial averages, 2) the fact that there is measurement error associated with pixel values, and 3) the fact that there is spatial correlation in the true field. We could make even better inferences if we could exploit the middle and right images simultaneously, and do better yet if could make use information across time periods.

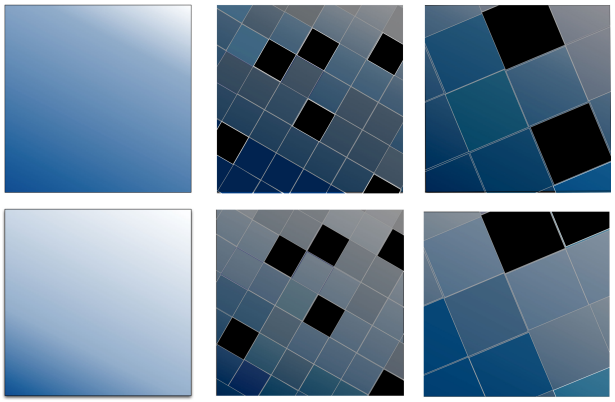

Figure 1. Left panels: examples of a true, spatially continuous geophysical field at two successive time points. Middle panels: the field as viewed by a remote sensing instrument. Right panels: the field as viewed by another remote sensing instrument.

Here we suggest a formal statistical framework for inferring the true value of the geophysical field at any point location, **s**, using remote sensing observations from two instruments at multiple time points. We begin by focusing on a single time point.

The relationships between the true field at a single point in time and the corresponding instrument images in the middle and right panels of Figure 1 can be formalized. Let $Y(\mathbf{s})$ be a random variable representing the true, underlying geophysical phenomenon of interest. The top-left panel of Figure 1 is an image of $Y(\mathbf{s})$ for all locations in the domain. Let $B_1(\mathbf{s})$ be a pixel, centered at $\mathbf{s}$, observed by instrument 1, and let $B_2(\mathbf{s})$ be a pixel, centered at $\mathbf{s}$, observed by instrument 2. In general, the pixels for two instruments can be different. Let

$$Z_1(B_1(\mathbf{s})) = \frac{1}{|B_1(\mathbf{s})|}\int_{\mathbf{s}\in B_1(\mathbf{s})} Y(\mathbf{s})d\mathbf{s} + \varepsilon_1(B_1(\mathbf{s})), \quad (1)$$

and

$$Z_2(B_2(\mathbf{s})) = \frac{1}{|B_2(\mathbf{s})|}\int_{\mathbf{s}\in B_2(\mathbf{s})} Y(\mathbf{s})d\mathbf{s} + \varepsilon_2(B_2(\mathbf{s})), \quad (2)$$

where $Z_j(B_j(\mathbf{s}))$ is the value observed by instrument j in the pixel $B_j(\mathbf{s})$, $|B_j(\mathbf{s})|$ is the area of the pixel, and $\varepsilon_j(B_j(\mathbf{s}))$ is the measurement error. Denote the set of observations from instrument j by $\mathbf{Z}_j = (Z_j(B_j(\mathbf{s}_{j1})),\ldots,Z_j(B_j(\mathbf{s}_{jN_j})))'$, where prime indicates vector transpose, and N_j is the number of observations. This is a column vector formed by concatenating all non-missing observations for the instrument. For any location $\mathbf{s}$ in the domain, $Y(\mathbf{s})$ can be optimally estimated as a linear combination of the available observations from instrument j by $\tilde{Y}_j(\mathbf{s}) = \mathbf{a}'_{j\mathbf{s}}\mathbf{Z}_j$. Optimal estimates are obtained by solving for the coefficient vectors, $\mathbf{a}_{j\mathbf{s}}$, that minimize

$$MSE(Y(\mathbf{s}),\tilde{Y}_j(\mathbf{s})) = E\left\|Y(\mathbf{s}) - \mathbf{a}'_{j\mathbf{s}}\mathbf{Z}_j\right\|^2 \quad (3)$$

subject to the unbiasedness constraint,

$$E(Y(\mathbf{s})) = E(\tilde{Y}_j(\mathbf{s})) = E(\mathbf{a}'_{j\mathbf{s}}\mathbf{Z}_j), \quad (4)$$

where $E(\cdot)$ is the statistical expectation operator. The two estimates $\tilde{Y}_1(\mathbf{s})$ and $\tilde{Y}_2(\mathbf{s})$ are unbiased and optimal in the sense of having minimum mean squared error given their respective input data, but they will not be identical. Their mean squared errors given in (3) will also be different and the one with the lower value is preferred.

Now suppose we form an estimator that uses both instruments' data simultaneously:

$$\hat{Y}(\mathbf{s}) = \mathbf{b}'_{1\mathbf{s}}\mathbf{Z}_1 + \mathbf{b}'_{2\mathbf{s}}\mathbf{Z}_2. \quad (5)$$

The coefficients $\mathbf{b}_{j\mathbf{s}}$ are obtained by minimizing

$$MSE(Y(\mathbf{s}),\hat{Y}(\mathbf{s})) = E\left\|Y(\mathbf{s}) - (\mathbf{b}'_{1\mathbf{s}}\mathbf{Z}_1 + \mathbf{b}'_{2\mathbf{s}}\mathbf{Z}_2)\right\|^2 \quad (6)$$

subject to,

$$E(Y(\mathbf{s})) = E(\hat{Y}(\mathbf{s})) = E(\mathbf{b}'_{1\mathbf{s}}\mathbf{Z}_1 + \mathbf{b}'_{2\mathbf{s}}\mathbf{Z}_2). \quad (7)$$

Now, $MSE(Y(\mathbf{s}),\hat{Y}(\mathbf{s})) \le \min_j\left[MSE(Y(\mathbf{s}),\tilde{Y}_j(\mathbf{s}))\right]$ because if either instrument's data were to contain no "useful" information, the corresponding $\mathbf{b}_{j\mathbf{s}}$ would be the zero vector. We call the estimates $\tilde{Y}_1(\mathbf{s})$ and $\tilde{Y}_2(\mathbf{s})$ kriging estimators and $\hat{Y}(\mathbf{s})$ the statistical data fusion estimator.

Solving the constrained minimization problems in (3), (4), (6), and (7) requires expanding the expressions for mean squared error and substituting in the definitions (1) and (2). This results in a set of terms that depend on various parameters of the joint (spatial) distribution of the true field, such as $Cov(Y(\mathbf{s}_i),Y(\mathbf{s}_j))$, for all pairs of locations $(\mathbf{s}_i,\mathbf{s}_j)$ in the domain.

Estimating these distributional parameters requires some additional modeling assumptions. In particular, we assume $Y(\mathbf{s})$ behaves according to a spatial mixed effects model,

$$Y(\mathbf{s}) = \mathbf{t}(\mathbf{s})'\alpha + \nu(\mathbf{s}), \quad \nu(\mathbf{s}) = \mathbf{S}(\mathbf{s})'\eta + \xi(\mathbf{s}). \quad (8)$$

The term $\mathbf{t}(\mathbf{s})'\alpha$ is the spatial trend and captures the effect of simple explanatory variables. For example, $\mathbf{t}(\mathbf{s})$ may be the latitude and longitude of $\mathbf{s}$. The trend term reflects a modeling assumption that, to a coarse approximation, the value of $Y(\mathbf{s})$ can be "explained" by its latitude and longitude. The term $\nu(\mathbf{s})$ explains additional variation in $Y(\mathbf{s})$ not captured by the trend. This additional variation has spatial structure: $Y(\mathbf{s}_i)$ may be correlated with $Y(\mathbf{s}_j)$ where $\mathbf{s}_i$ and $\mathbf{s}_j$ are two different spatial locations. The term $\nu(\mathbf{s})$ is further broken down into $\mathbf{S}(\mathbf{s})'\eta$ and $\xi(\mathbf{s})$, where η is a hidden (unobserved) vector-valued random variable that captures key features of the spatial-dependence structure in the domain. The coefficient vector $\mathbf{S}(\mathbf{s})$ provides location-specific weights for combining the elements of η to produce a contribution to $\nu(\mathbf{s})$ for each specific location. Finally, $\xi(\mathbf{s})$, called fine-scale variation, is a residual term to account for variation in $\nu(\mathbf{s})$ not accounted for by $\mathbf{S}(\mathbf{s})'\eta$.

Nguyen, Cressie and Braverman (2010) use (1) - (8) to formulate and implement Spatial Statistical Data Fusion (SSDF), a methodology for optimally estimating $Y(\mathbf{s})$ from two remote sensing data sets with different statistical characteristics (at a single time point). For SSDF it is not necessary to explicitly estimate η. The derivations of the data fusion coefficients, the optimal estimate $\hat{Y}(\mathbf{s})$, and its

uncertainty depend on η only through its covariance matrix, so this covariance matrix is estimated directly.

Space-time Data Fusion (STDF) builds on SSDF by assuming that the random vector η evolves in time according to a lag-1 auto-regressive process (AR(1)). That is, at each time step, one can exploit not only spatial dependence in the domain at that time, but also the temporal dependence with the previous time step via the relationship between η_t and η_{t-1}. Our STDF methodology is motivated by Cressie, Shi, and Kang (2010) who introduced Fixed Rank Filtering (FRF), a framework for capturing temporal dependence in the context of optimal estimation from a single data set.

Suppose now that we have access to data at more than one time point, say $t-1$ (top panel in Figure 1) and t (bottom panel in Figure 1). The evolution of the spatial dependence structure is described by a first-order autoregressive relationship between η_{t-1}, and η_t. The space-time data fusion estimator of Y at location $\mathbf{s}$ and time t, $\hat{Y}(\mathbf{s},t)$, is based on the optimal estimation of η_t through a Kalman Filter.

Using the model in (8), the data fusion estimator and its mean squared error are,

$$\hat{Y}(\mathbf{s},t) = \mathbf{t}(\mathbf{s})'\hat{\alpha}_t + \mathbf{S}(\mathbf{s})'\hat{\eta}_{t|t} + \hat{\xi}_{t|t}(\mathbf{s}), \quad (9)$$

$$MSE(Y(\mathbf{s},t),\hat{Y}(\mathbf{s},t)) = E(Y(\mathbf{s},t) - \hat{Y}(\mathbf{s},t))^2, \quad (10)$$

where

$$\hat{\eta}_{t|t} = E(\eta_t \,|\, \mathbf{Z}_1(1),\mathsf{K}\,,\mathbf{Z}_1(t),\mathbf{Z}_2(1),\mathsf{K}\,,\mathbf{Z}_2(t)), \quad (11)$$

$$\hat{\xi}_{t|t}(\mathbf{s}) = E(\xi(\mathbf{s},t) \,|\, \mathbf{Z}_1(1),\mathsf{K}\,,\mathbf{Z}_1(t),\mathbf{Z}_2(1),\mathsf{K}\,,\mathbf{Z}_2(t)), \quad (12)$$

and $E(\cdot\,|\,\cdot)$ is the statistical expectation of the quantity on the left of the bar, given the quantity on the right. All parameters or their estimates that vary in time are subscripted by t. The subscript $t\,|\,t$ indicates that the subscripted variable is statistically conditional on all information up through and including time t, and $\mathbf{Z}_j(t)$ is the vector of observations from instrument j at time t. Note that we do not explicitly derive the data fusion coefficients because they are not of interest in and of themselves. The formulas required to compute (9) – (12) can be obtained by concatenating the data vectors $\mathbf{Z}_1(t)$ and $\mathbf{Z}_2(t)$ into a supervector, making commensurate adjustment to covariance matrices and other quantities, and applying the fixed rank filtering formulas given by Cressie, Shi, and Kang (2010). Interested readers can find details and a thorough discussion there.

3. ESTIMATING CO2 IN THE LOWER ATMOSPHERE FROM AIRS AND ACOS

Our goal is to estimate the amount of CO2 in the lower part of the atmosphere using data from AIRS and ACOS. AIRS retrieves mid-tropospheric CO2 on 90 km footprints with near-global coverage every three days. ACOS retrieves total column CO2 on 10 km footprints spaced about 150 km apart, in a narrow swath that repeats every three days.

The theory in the previous section assumes $Y(\mathbf{s},t)$ is a scalar, and instruments 1 and 2 both measure this quantity with different resolutions and other sampling characteristics. AIRS and ACOS do not measure the thing: AIRS measures the amount of CO2 in the mid-troposphere and above, and ACOS measures the amount of CO2 in the total column. The sensitivities of the two to different vertical levels in the atmosphere are depicted in the lower-left panel of Figure 2. To estimate CO2 in the lower atmosphere, we need to estimate the difference between ACOS and AIRS CO2 values. Fortunately, the theory in Section 2 easily accommodates this.

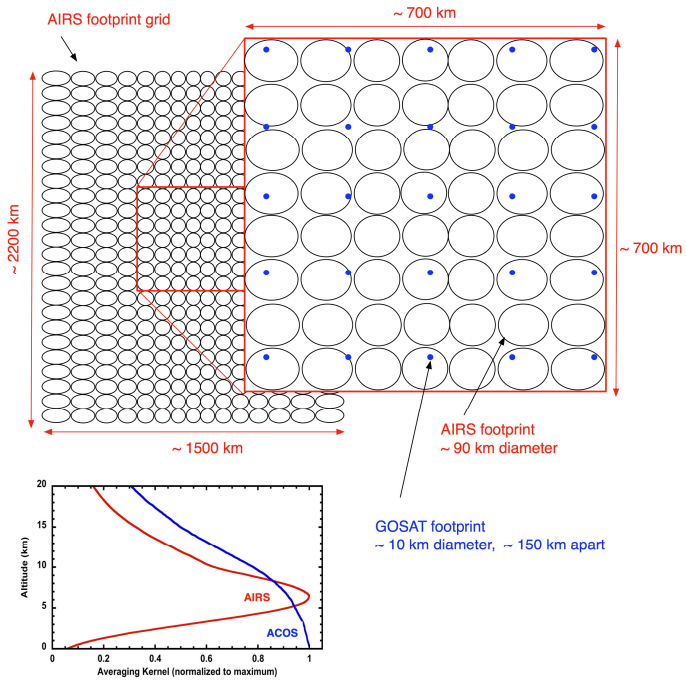

Figure 2. Sampling characteristics of AIRS (red) and ACOS (blue). The lower-left show sensitivities of the two instruments to different atmospheric levels.

Let $Y_1(\mathbf{s},t)$ be the true amount of CO2 in the mid-troposphere and above at location $\mathbf{s}$ and time t, and let $Y_2(\mathbf{s},t)$ be the true amount of CO2 in the total column at location $\mathbf{s}$ and time t. Let $\mathbf{Y}(\mathbf{s},t) = (Y_1(\mathbf{s},t),Y_2(\mathbf{s},t))'$. The entire STDF methodology presented in Section 2 generalizes for this vector-valued case in enable simultaneous inference of the pair $(Y_1(\mathbf{s},t),Y_2(\mathbf{s},t))'$. In fact, there may be additional benefit if the components of

$\mathbf{Y}(\mathbf{s},t)$ are correlated, as they surely are in this case. The methodology automatically exploits such correlations to improve the inferences and reduce uncertainties. The estimate of CO2 in the lower atmosphere is simply $\mathbf{c}'\mathbf{Y}(\mathbf{s},t)$, where $\mathbf{c}'$ is the row vector (-1, 1). The mean squared error of this estimate is $\mathbf{c}'\Sigma(\mathbf{s},t)\mathbf{c}$, where $\Sigma(\mathbf{s},t)$ is the mean-squared-error matrix of $\mathbf{Y}(\mathbf{s},t)$.

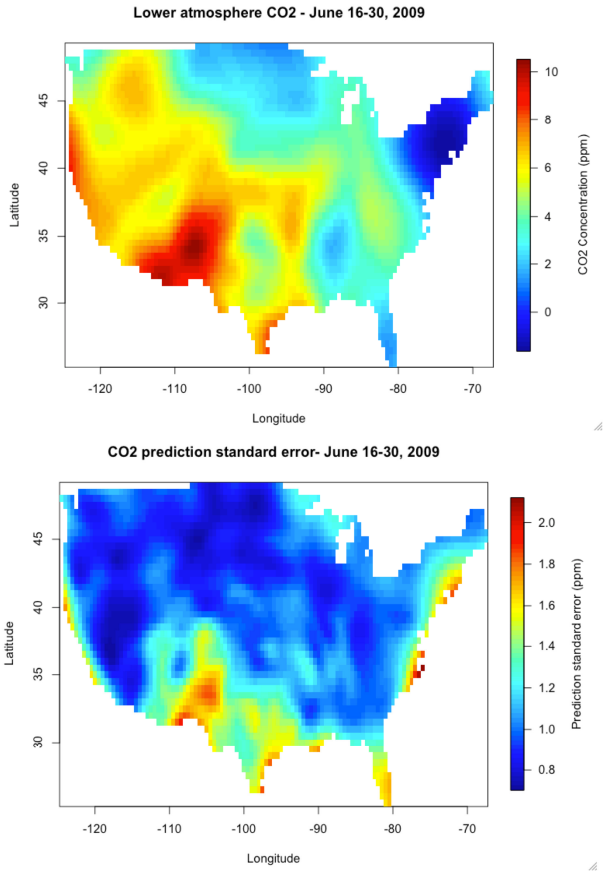

Figure 3. STDF estimate of CO2 in the lower atmosphere derived as the fused difference between ACOS and AIRS measurements (top), and the square-root of the mean squared errors of those estimates (bottom).

We applied STDF to 30 days of AIRS and ACOS data for June 2009 to estimate CO2 in the lower atmosphere over the continental US. Time $t=1$ was designated June 16-30, and time $t=0$ was designated June 1-15. The spatial sampling of AIRS and ACOS data are shown in the top portion of Figure 2. Over the continental US, there were a total of 8407 AIRS data points and 1368 ACOS data points in June 2009. It took about 20 seconds on 2.8 GHz MacBook Pro laptop to produce 3284 estimates, and their mean squared errors, spaced every half-degree. Figure 3 shows the estimates and the corresponding square-root mean squared errors. These extremely fast computations are possible because of the specification of the spatial-dependence structure in (8) using $\mathbf{S}(\mathbf{s})'\eta$, which leads to a fast procedure for estimating and inverting the large matrix $Cov(Y(\mathbf{s}_i),Y(\mathbf{s}_j))$.

4. DISCUSSION

We have demonstrated that STDF can be used to leverage both spatial and temporal dependence to estimate a function of two spatially continuous geophysical fields from noisy observations with different statistical characteristics. The maps in Figure 3 look like they may provide reasonable estimates, but these have yet to be validated against independent in-situ oberservations. The effect of using only land data, however, is evident in the uncertainties in coastal areas. Inland Texas and New Mexico also show elevated uncertainties that must be investigated, especially because of the hot spot in this area in the top panel of Figure 3. It is also worth emphasizing that the validity of both the estimates and uncertainties depends on the means and standard deviations of the measurement error distributions associated with the terms $\varepsilon_1(B_1(\mathbf{s}))$ and $\varepsilon_2(B_2(\mathbf{s}))$ in Equations (1) and (2), and on other modeling choices discussed in Sections 2.1 and 2.2. In this exercise, we used measurement-error statistics based on the judgment and experience of members of the AIRS and ACOS teams. A more rigorous analysis will ultimately be required as will a careful evaluation of the sensitivities of our results to the other modeling assumptions.

Near-term methodological improvements center on reducing the duration of a time step in the STDF analysis. Currently, our method aggregates data over 15 days because the ACOS data are sparse, and estimates of statistical model parameters are unstable with fewer observations. However, CO2 transport occurs on shorter time scales, and the science community would prefer time steps on the order of three days. We have used the method of moments to estimate model parameters here, but we are investigating expectation maximization (EM) as a more stable alternative. We are also beginning the process of validating our lower atmosphere CO2 estimates by comparing them to in-situ observations with the help of the AIRS and ACOS validation teams.

REFERENCES

N. Cressie, T. Shi, and E. Kang, "Fixed Rank Filtering for Spatio-Temporal Data," Journal of Computational and Graphical Statistics, vol. 19, no. 3, pp 724-745, 2010.

H. Nguyen, N. Cressie, and A. Braverman, "Spatial-statistical Data Fusion for Remote Sensing Applications," Journal of the American Statistical Association, submitted, 2010. (Also available as Technical Report No. 849, Department of Statistics, The Ohio State University, November, 2010.)

ACKNOWLEDGEMENTS

The research described in this article was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration. Copyright 2011. All rights reserved. Government sponsorship acknowledged.