Fusion of AIRS and OCO-2 carbon dioxide data for mapping lower-atmospheric CO2

Amy Braverman[1], Noel Cressie[1,2], Emily Kang[3], Matthias Katzfuss[4], Pulong Ma[3], Anna Michalak[5], Hai Nguyen[1], Tim Stough[1], and Vineet Yadav[1]

[1]Jet Propulsion Laboratory, California Institute of Technology
[2]NIASRA, University of Wollongong
[3]Department of Mathematical Sciences, University of Cincinnati
[4]Department of Statistics, Texas A&M University
[5]Carnegie Institution of Washington

April 21, 2015

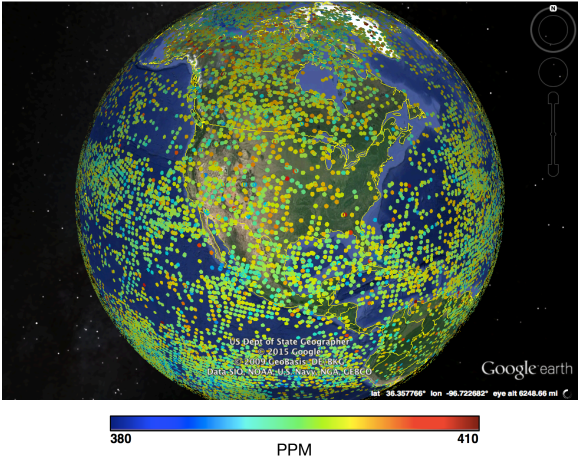
Outline

- Data fusion is spatial/spatio-temporal inference
- OCO-2 and AIRS data
- Exploiting synergy
- Example
- Data fusion strategy
- Validation
- Take home

Introductory comments

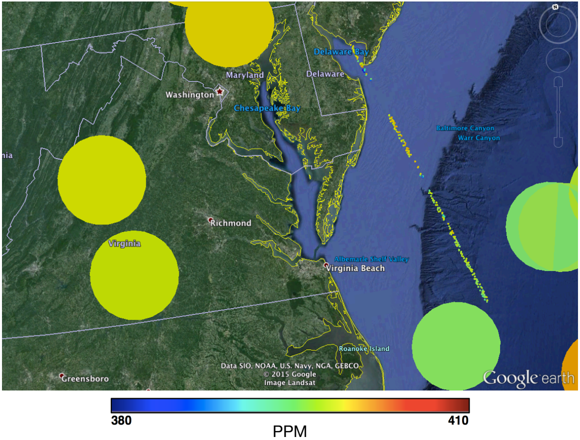
- We want to estimate a complete geophysical field from massive, heterogeneous, observational data.
- The result is input to further science investigations and applications, so uncertainties must be propagated rigorously.
- Uncertainties should also be minimized so that conclusions, and decisions based on them, are as robust as possible. Need to leverage spatial and temporal dependencies.
- Challenge: Accomplish this in the face of massive data volumes from multiple instruments and complex calculations required.

OCO-2 and AIRS data



AIRS mid-tropospheric CO_2, October 30 through November 2, 2014.

OCO-2 and AIRS data



AIRS mid-tropospheric and OCO-2 total column CO_2, October 30 through November 2, 2014.

OCO-2 footprints actual size.

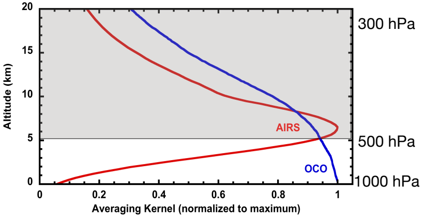
Exploiting synergy

- Instrument sensitivities are similar at and above the mid-troposphere, but not below: OCO-2 is sensitive down to the surface, but AIRS is not.

- To the extent that CO2 mole-fraction near the surface and in the mid-troposphere are correlated, we should be able to improve estimates of both by exploiting this correlation.

- We should also be able to
  - exploit the coverage of AIRS and the accuracy of OCO-2
  - exploit spatial and temporal correlations within and between data sets.

Exploiting synergy



- If we knew the “true" values of total-column and mid-tropospheric mole-fraction at a location **s** =lat,lon, $Y_1(\mathbf{s})$ and $Y_2(\mathbf{s})$, then we could compute

$$Y_{LA}(\mathbf{s}) = \frac{(1000-300)\,Y_1(\mathbf{s}) - (500-300)\,Y_2(\mathbf{s})}{1000-500}.$$

- Can we get estimates, with uncertainties, of (total-column, mid-trop) pairs at reasonable resolution so we can compute this?

Example

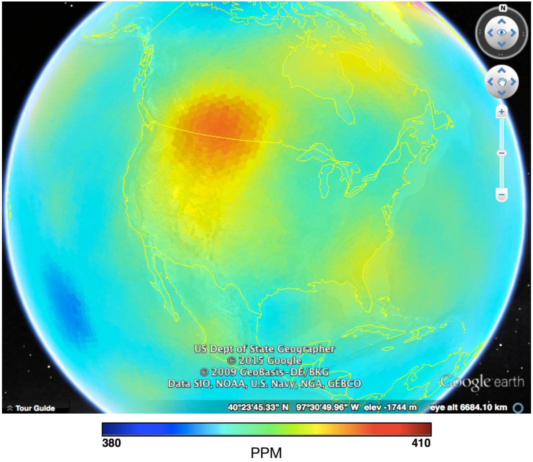
- Accumulate 12 days of AIRS and OCO-2 data into three, four-day blocks: Oct 30 - Nov 2, Nov 3 - 6, Nov 7 - 10.
- Run Spatio-Temporal Data Fusion algorithm (STDF) on the three blocks, producing three output data sets, one for each block. (See Nguyen, Katzfuss, Cressie, and Braverman (2014) for details.)
- STDF accounts for spatial correlations among footprints for both instruments (including corrections for different sizes and orientations) and for temporal correlations from time block to time block.
- Timing: 90 minutes to process the three blocks on a single, Intel Xeon 2.0 GHz processor.
- Crucial assumptions: uncertainty on AIRS datum is 1.5 ppm, and uncertainty on OCO-2 datum is 2 ppm.

National Aeronautics and
Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Example

Fused estimate of lower-atmosphere CO_2, Oct 30 - Nov 2, 2014:



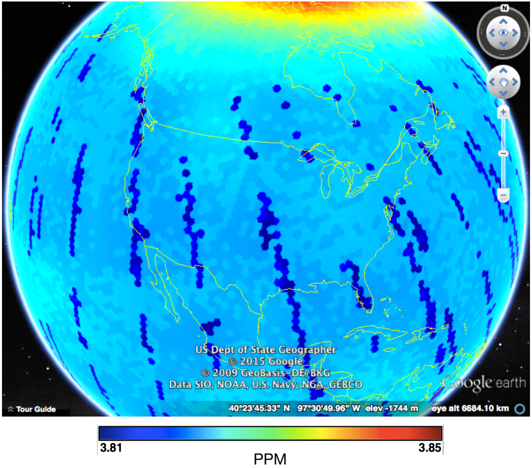
Produced using STDF with analysis resolution $\approx$ 30 km.

Visualization resolution $\approx$ 120 km.

How to validate estimates?

Example

Uncertainties of fused uncertainties, Oct 30 - Nov 2, 2014:

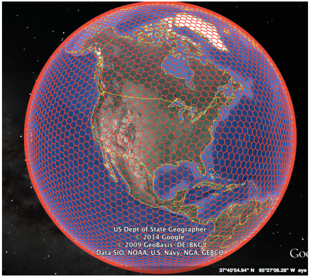


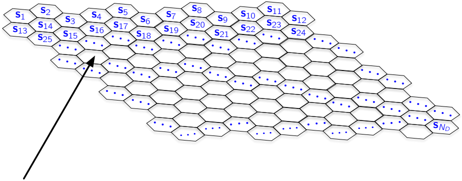
Lower uncertainties coincide with OCO-2 tracks.

How to validate uncertainties?

Data fusion strategy

- In order to do this calculation, we need to infer the true mole-fractions of (total-column, mid-trop) pairs on a fine grid of locations.
- We define that grid by partitioning the world into very small hexagonal tiles called basic areal units (BAU's) Notionally, each BAU contains a pair.





$$Y(\mathbf{s}) = (Y_1(\mathbf{s}), Y_2(\mathbf{s}))$$

Data fusion strategy

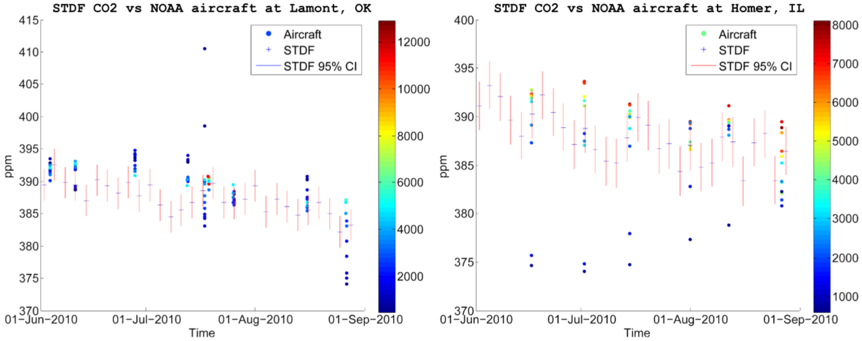
- Since this bivariate field is unknown, we model it with a random vector that behaves according to a probability distribution.
- We use Bayes' Theorem: before acknowledging the observations, we assume a "prior" distribution.
- After seeing the data, we update that distribution and call it the "posterior".
- We report the mean vector and covariance matrix of the posterior distribution as our inference.
- Key innovations: capture spatial dependence with a hidden, low-dimensional variable; capture temporal dependence from time block to time block with a Kalman smoother on the hidden variable.

Example

Caveats:

- OCO-2 data are very preliminary: just a placeholder here to show data fusion machinery.
- The formula for computing lower-atmosphere mole-fraction is unrealistically crude.
- Uncertainties on the input data are unrealistic (but the best we've got right now). This is a *major* issue.
- We have built a simulation system for characterizing the performance of STDF on synthetic “truth" data, and are in the process of assessing how various design choices affect our results.

Validation example



- Validation of STDF estimates of lower-atmosphere CO_2 based on AIRS and Japan's Greenhouse Gases Observing Satellite (NASA retrievals). See Nguyen et al. (2014) for details.

Take home

- Data fusion is necessary to realize benefits of synergy among NASA missions.

- What is new about this data fusion technology:
  - based on uncertainty quantification and minimization
  - uses a formal probabilistic framework that is coherent
  - exploits spatial and temporal correlations to drive uncertainties down
  - corrects for heterogeneous footprints
  - feasible for massive data sets and operational implementation.

- Better results are possible if mission provide formal uncertainty estimates for their retrieivals.

References

The methodology used here is described in detail in

Nguyen, H., Katzfuss, M., Cressie, N., and Braverman, A. (2014). Spatio-Temporal Data Fusion for Very Large Remote Sensing Datasets, *Technometrics*, 56, pp. 174-185.

The spatial-only methodology is described in

Nguyen, H., Cressie, N., and Braverman, A. (2012). Spatial Statistical Data Fusion for Remote-Sensing Applications, *Journal of the American Statistical Association*, 107, pp. 1004-1018.

This work was partially performed at the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration. Government sponsorship acknowledged.

This research was funded by NASA's Earth Science Technology Office through its Advanced Information Systems Technology program (AIST-08 and AIST-11).

We would like to thank the OCO-2 and AIRS Projects for their intellectual and practical contributions to this work.

Contact information: `Amy.Braverman@jpl.nasa.gov`