National Aeronautics and
Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

A Statistical Perspective on Climate Informatics

Amy Braverman[1,2]

[1]Jet Propulsion Laboratory,
California Institute of Technology

[2]Department of Statistics, UCLA

September 20, 2012

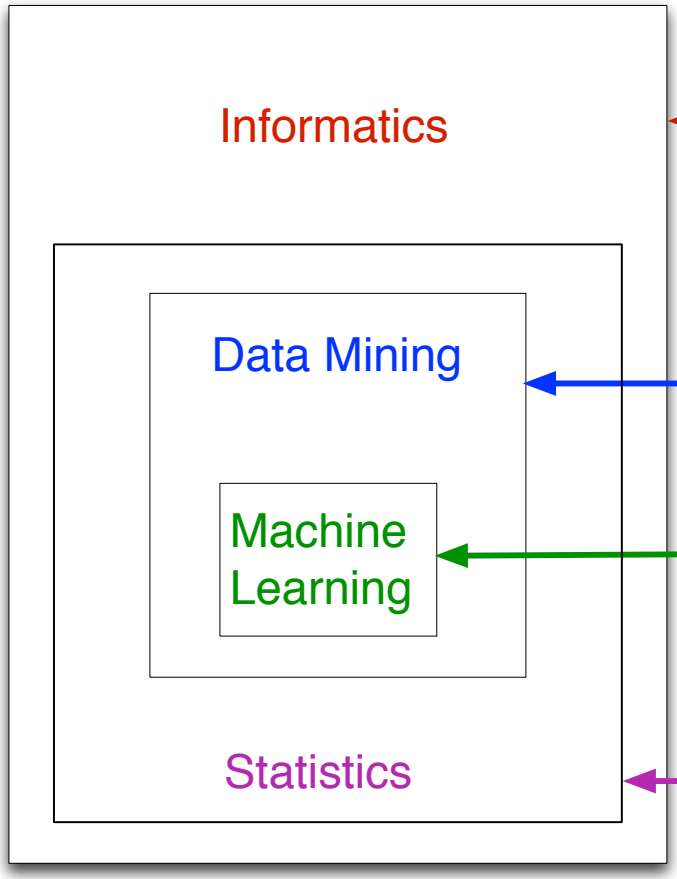
Outline

- What is climate informatics and why do we care?
- Machine Learning, Data Mining, Statistics, and Earth Science: the Second NASA Data Mining Workshop.
  - What was it?
  - What did we learn?
  - How have things changed since then?
- An example addressing one of today's informatics challenges at NASA.
  - A statistical model for remote sensing data.
  - A family of inference problems.
  - Massiveness and change of support.
  - Exploiting spatial and temporal dependence.
  - Data "fusion": an example.
- Some concluding thoughts.

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

What is climate informatics?



``... the science of information, the practice of information processing, and the engineering of information systems." (Wikipedia)

``... the process that attempts to discover patterns in large data sets." (Wikipedia)

``Machine learning focuses on prediction, based on *known* properties learned from the training data." (Wikipedia)

``... the study of the collection, organization, analysis, interpretation, and presentation of data." (Wikipedia)

What is climate informatics?

"Data Mining for Insight into Processes":

- "The rate of data acquisition via the satellite network and the re-analyses projects is very rapid. Similarly the amount of model output is equally fast growing. Model-observation comparisons based on processes (ie. the multi-variate changes that occur in a single event (or collection of events)... have the potential to provide very useful information on model credibility, physics and new directions for parameterisation improvements."
- "However, data services usually deliver data in single variable, spatially fixed, time varying formats that make it very onerous to apply space and time filters to the collection of data to extract generic instances of the process in question. This is surely a task that computer science should be able to improve."

From the Climate Informatics Wiki, Scientific Problems in Climate Informatics

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

The Second NASA Data Mining Workshop

Second NASA Data Mining Workshop: Issues and Applications in Earth Science

Home | Agenda | Instructions for Presenters | Program Committee | Attendees | Logistics

May 23-24, 2006, Pasadena, CA

Call for Papers

Instructions for Presenters

Agenda
Updated 7-6-06

Attendees
Updated 8-2-06

Hotel Information

Final Report (PDF)

This workshop is sponsored by NASA.

Data from Earth-orbiting satellites have been accumulating at a very high rate for several years now. In combination with in-situ observations and physical model output, this enormous, distributed repository holds the answers to important questions about our planet's past, present and future. However, the information is only accessible if effective analysis capabilities can be brought to bear. Data mining has the potential to provide these capabilities, and, if employed in close coordination with Earth science research, could increase the science return from NASA's vast Earth science data collection.

The objectives of this second NASA Data Mining Workshop are to bring together Earth scientists and data miners to match the needs of the scientific community to existing capabilities provided by computer scientists and statisticians, and suggest future research directions they may pursue to help advance Earth science research. In particular, we seek to facilitate formation of collaborative relationships between Earth and data scientists, and identify specific problems those collaborations can address. We are soliciting *short papers* (no more than four pages) focusing on the use of data mining techniques in Earth science research Please refer to the Call For Papers.

The workshop is now full and registration is closed.

Please SEE AGENDA PAGE FOR ROOM CHANGES

Important Dates:
Paper submission deadline: 11:00 pm (PST), January 31, 2006
Notification of acceptance: March 17, 2006
Workshop: May 23-24, 2006, Pasadena, CA

- Goal: to bring together Earth scientists and data miners to match the needs of the scientific community to existing capabilities provided by computer scientists and statisticians, and suggest future research directions they may pursue to help advance Earth science research.
- See http://datamining.itsc.uah.edu/meeting06/ to obtain the Final Report.

Among the key findings:

- One obstacle to infusion of modern data analysis methods in Earth science is the disconnect between “modeling the data" and relating it back to underlying physical processes. (“Science is hypothesis driven, but data mining is data driven.")
- A conceptual framework is needed to articulate the roles that statistics and data mining can play in advancing Earth science research. Such a framework should:
  - link questions about Earth system processes to questions about data,
  - provide an infrastructure for making inferences from the data back to the underlying state of the Earth system, and translating those inferences into physically meaningful conclusions.

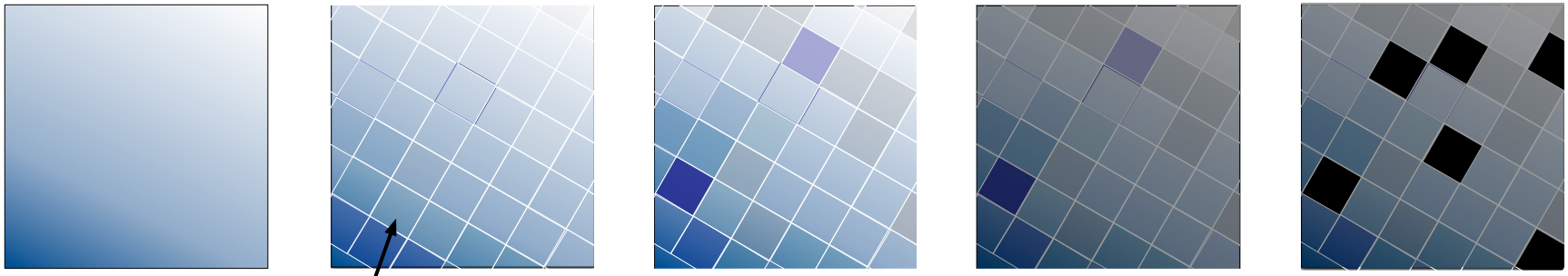
The Second NASA Data Mining Workshop

What's changed in the last six years?

- Large-scale, climate modeling experiments (CMIP3, CMIP5) produce massive climate simulation datasets.
- Simulations instantiate hypotheses about the behavior, causes, and effects of processes making up the climate system. We are now in a position to test these hypotheses by comparing simulations to observations, but we need uncertainties...
- We have many diverse observational datasets that report the same/similar/related quantities (e.g., CO2 concentrations) with different sampling and measurement error characteristics. Can we estimate the underlying true fields that are the targets of these (noisy) observations?
- Can we use multiple data sources simultaneously to get better estimates (reduce uncertainies)?

National Aeronautics and Space Administration
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

A statistical model for remote sensing data



$Y(\mathbf{s})$ B_m $\epsilon(B_m) \sim N(c, \sigma_\epsilon^2)$ $N =$ number non-missing pixels

$$Y(B_m) = \frac{1}{|B_m|} \int_{\mathbf{s} \in B_m} Y(\mathbf{s})\, d\mathbf{s} \qquad Z(B_m) = Y(B_m) + \epsilon(B_m) \qquad \mathbf{Z} = (Z(B_1), \ldots, Z(B_N))'$$

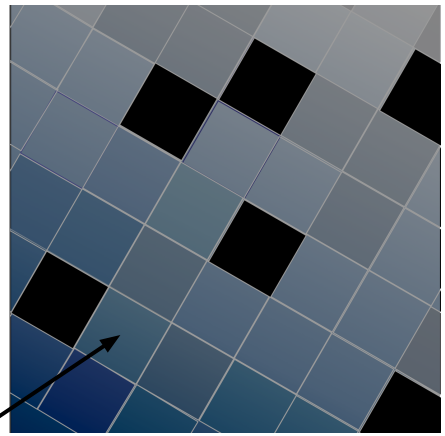
- Remote sensing data are noisy spatial aggregates of the true field.
- Goal: infer $Y(\mathbf{s})$ from $\mathbf{Z}$ (optimal spatial prediction).

National Aeronautics and Space Administration
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

A statistical model for remote sensing data

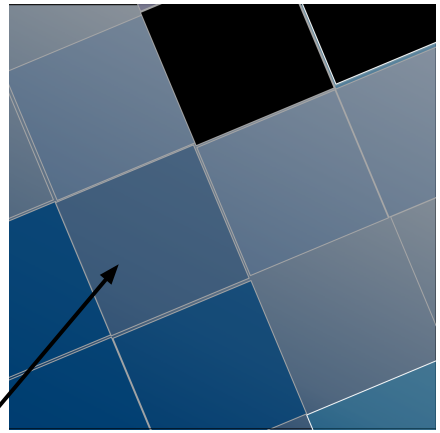
True

Instrument 1



B_{1m}

Instrument 2



B_{2m}

$$Y(\mathbf{s})$$

$$\epsilon_1(B_{1m}) \sim N(c_1, \sigma^2_{\epsilon,1})$$

$$\epsilon_2(B_{2m}) \sim N(c_2, \sigma^2_{\epsilon,2})$$

$$Y_1(B_{1m}) = \frac{1}{|B_{1m}|}\int_{s\in B_{1m}} Y(\mathbf{s})\ d\mathbf{s}$$

$$Z_1(B_{1m}) = Y_1(B_{1m}) + \epsilon_1(B_{1m})$$

$$\mathbf{Z}_1 = (Z_1(B_{11}), \ldots, Z_1(B_{1N_1}))'$$

$$Y_2(B_{2m}) = \frac{1}{|B_{2m}|}\int_{s\in B_{2m}} Y(\mathbf{s})\ d\mathbf{s}$$

$$Z_2(B_{2m}) = Y_2(B_{2m}) + \epsilon_2(B_{2m})$$

$$\mathbf{Z}_2 = (Z_2(B_{21}), \ldots, Z_2(B_{2N_2}))'$$

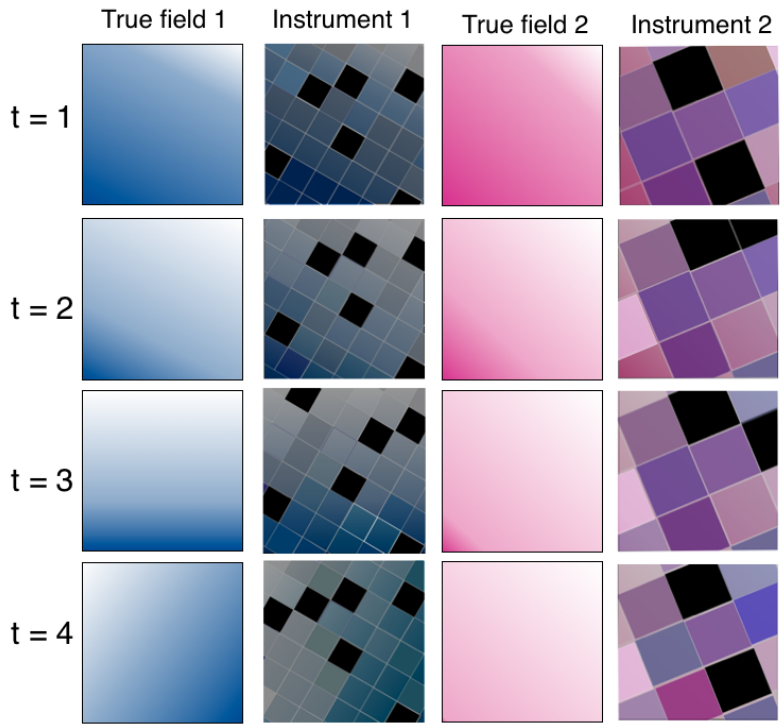
- Better yet: infer $Y(\mathbf{s})$ from $\mathbf{Z}_1$ and $\mathbf{Z}_2$ (data fusion).

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

A family of inference problems

Exploit spatial correlations	Exploit spatial and temporal correlations
Infer the true field (single quantity) from one remote sensing image of it at a single time point. (Fixed Rank kriging)	Infer the true field (single quantity) from one remote sensing image of it at multiple time points. (Fixed Rank filtering)
Infer the true field from *two* different remote sensing images of it at a single time. (Single process, multiple source spatial data fusion)	Infer the true field from *two* different remote sensing images of it at multiple time points. (Single process, multiple source spatio-temporal data fusion)
Infer true values of *two* fields from *two* different remote sensing images at a single time. (Muliple process, multiple source spatial data fusion)	Infer true values of *two* fields from *two* different remote sensing images at multiple time points. (Muliple process, multiple source spatio-temporal data fusion)



Joint work with Noel Cressie (U. of Wollongong and Ohio State), Matthias Katzfuss (U. of Heidelberg), Emily Kang (U. of Cincinnati), Anna Michalak (Stanford U.) and Hai Nguyen (JPL).

Massiveness and change of support

- Bayesian hierarchical model: $[Y|Z] \propto [Z|Y][Y]$.

$$Z(B) = \frac{1}{|B|}\int_{\mathbf{u}\in B} Y(\mathbf{u})\,d\mathbf{u} + \epsilon(B) \qquad \text{(data model*)},$$

$$Y(\mathbf{s}) = \mu(\mathbf{s}) + \nu(\mathbf{s}) + \xi(\mathbf{s}) = \mu(\mathbf{s}) + \mathbf{S}(\mathbf{s})'\boldsymbol{\eta} + \xi(\mathbf{s}) \qquad \text{(process model)},$$

where $\mu(\mathbf{s})$ is "trend" (assumed known or estimated), $\nu(\mathbf{s})$ is the spatial covariance term, $\xi(\mathbf{s})$ is "fine-scale" variation (a residual), and $\epsilon(B)$ is measurement error. $\xi(\mathbf{s}) \sim N(0, \sigma^2_\xi)$, $\epsilon(B) \sim N(c\mu(B), \sigma^2_\epsilon)$. ($c$ is a multiplicative bias, $\mu(B)$ defined below.)

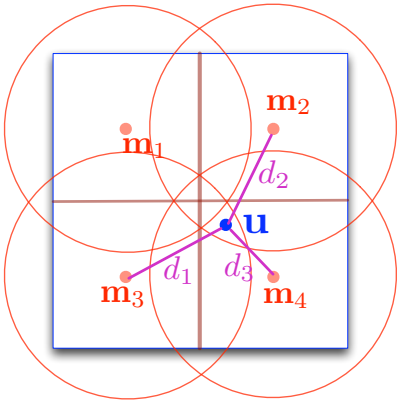
- $\mu(\mathbf{s}) = \mathbf{t}(\mathbf{s})'\boldsymbol{\alpha}$, $\mathbf{t}(\mathbf{s}) = (lat, lon)'$, for example.
- $\mathbf{S}(\mathbf{s})$ is a low-dimensional ($r \times 1$) basis vector that encodes the location of $\mathbf{s}$ relative to a set of multi-resolution basis centers.
- $\boldsymbol{\eta}$ is a low-dimensional ($r \times 1$) hidden structure variable to be estimated.

* *abuse of notation!*

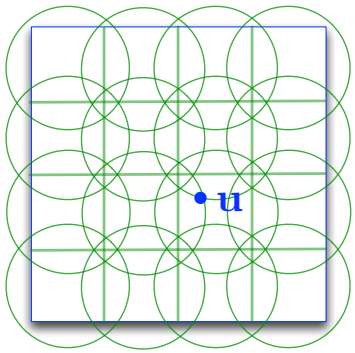
National Aeronautics and Space Administration
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Massiveness and change of support

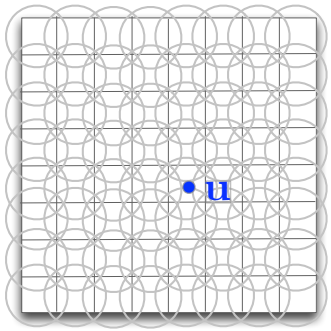
Multi-resolution basis functions:



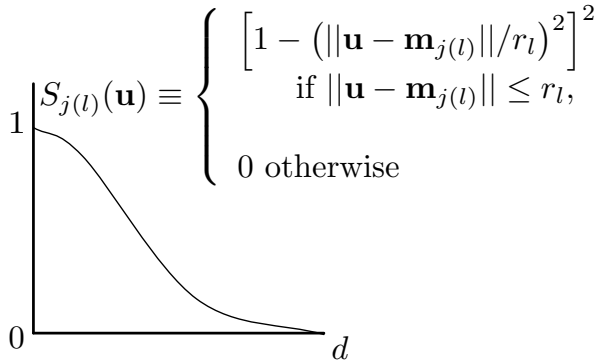
Resolution 1



Resolution 2



Resolution 3



Local Bisquare Functions

S(**u**): location **u** is encoded by distances to multiresolution basis centers.

Change of support property:

$$\nu(B) = \frac{1}{|B|}\int_{u\in B}\nu(\mathbf{u})\,d\mathbf{u} = \left[\frac{1}{|B|}\int_{u\in B}\mathbf{S}(\mathbf{u})'\eta\,d\mathbf{u}\right] = \left[\frac{1}{|B|}\int_{u\in B}\mathbf{S}(\mathbf{u})\,d\mathbf{u}\right]'\eta = \mathbf{S}(B)'\eta.$$

Massiveness and change of support

Other change of support properties:

- $Y(B) = \frac{1}{|B|} \int_{\mathbf{u} \in B} Y(\mathbf{u})\, d\mathbf{u}$.
- $\mathbf{t}(B) = \frac{1}{|B|} \int_{\mathbf{u} \in B} \mathbf{t}(\mathbf{u})\, d\mathbf{u}$.
- $\mu(B) = \frac{1}{|B|} \int_{\mathbf{u} \in B} \mathbf{t}(\mathbf{u})' \boldsymbol{\alpha}\, d\mathbf{u} = \left[\frac{1}{|B|} \int_{\mathbf{u} \in B} \mathbf{t}(\mathbf{u})\, d\mathbf{u} \right]' \boldsymbol{\alpha} = \mathbf{t}(B)' \boldsymbol{\alpha}$.
- $\nu(B) = \frac{1}{|B|} \int_{\mathbf{u} \in B} \nu(\mathbf{u})\, d\mathbf{u}$.
- $\xi(B) = \frac{1}{|B|} \int_{\mathbf{u} \in B} \xi(\mathbf{u})'\, d\mathbf{u}$.
- $\mathbf{S}(B) = \frac{1}{|B|} \int_{\mathbf{u} \in B} \mathbf{S}(\mathbf{u})\, d\mathbf{u}$.

Exploiting spatial and temporal dependence

- Optimal (minimum variance, unbiased) estimates:

$$\hat{Y}(\mathbf{s}) = \mathbf{t}(\mathbf{s})'\boldsymbol{\alpha} + \mathbf{S}(\mathbf{s})'\hat{\boldsymbol{\eta}} + \hat{\xi}(\mathbf{s}),$$
$$\hat{Y}(B) = \mathbf{t}(B)'\boldsymbol{\alpha} + \mathbf{S}(B)'\hat{\boldsymbol{\eta}} + \hat{\xi}(B).$$

- $\boldsymbol{\eta} \sim N_r(\mathbf{0}, \mathbf{K})$ a priori ($\mathbf{K}$ estimated off-line), $\hat{\boldsymbol{\eta}} = \mathrm{E}(\boldsymbol{\eta}|\mathbf{Z})$.
- $\xi(\mathbf{s}) \sim N(0, \sigma^2_\xi)$ a priori (σ^2_ξ estimated off-line), $\hat{\xi}(\mathbf{s}) = \mathrm{E}(\xi(\mathbf{s})|\mathbf{Z})$.
- Uncertainties of $\hat{\boldsymbol{\eta}}$, $\hat{\xi}(\mathbf{s})$ and $\hat{\xi}(B)$ are posterior variances. Propagate through to yield uncertainties of $\hat{Y}(\mathbf{s})$ and $\hat{Y}(B)$ relative to true, but not directly observed $Y(\mathbf{s})$ and $Y(B)$.
- Computationally tractable thanks to linearity and the parameterization of $\nu(\cdot)$ in terms of basis functions. See Cressie and Johannesson (2008), Nguyen (2009), Nguyen, Cressie, and Braverman (2012).

Exploiting spatial and temporal dependence

What about time?

- Incorporate time by letting $\boldsymbol{\eta}$ evolve according to a first-order autoregressive model:

$$\boldsymbol{\eta}_t | \boldsymbol{\eta}_{t-1}, \ldots, \boldsymbol{\eta}_0 \sim N_r(\mathbf{H}_t \boldsymbol{\eta}_{t-1}, \mathbf{U}_t), \;\; t = 1, 2, \ldots, T,$$

  where subscript t indicates time period, $\mathbf{H}_t$ and $\mathbf{U}_t$ are the propagator and innovation matrices, respectively, and the initial state is $\boldsymbol{\eta}_0 \sim N_r(\mathbf{0}, \mathbf{K}_0)$.

- Assuming the parameters $\boldsymbol{\alpha}$, $\mathbf{K}_0$, σ^2_ϵ, σ^2_ξ, $\mathbf{H}_t$, and $\mathbf{U}_t$ are known/given/estimated, we can optimally estimate the posterior expectations and covariances of $\{\boldsymbol{\eta}_t\}$ and of $\{\xi_t(\cdot)\}$ at all prediction locations using a Kalman smoother.

- See Katzfuss and Cressie (2010), Cressie, Shi, and Kang (2010), Kang, Cressie, and Shi (2010), Katzfuss and Cressie (2011), Nguyen, Katzfuss, Cressie, and Braverman (2012).

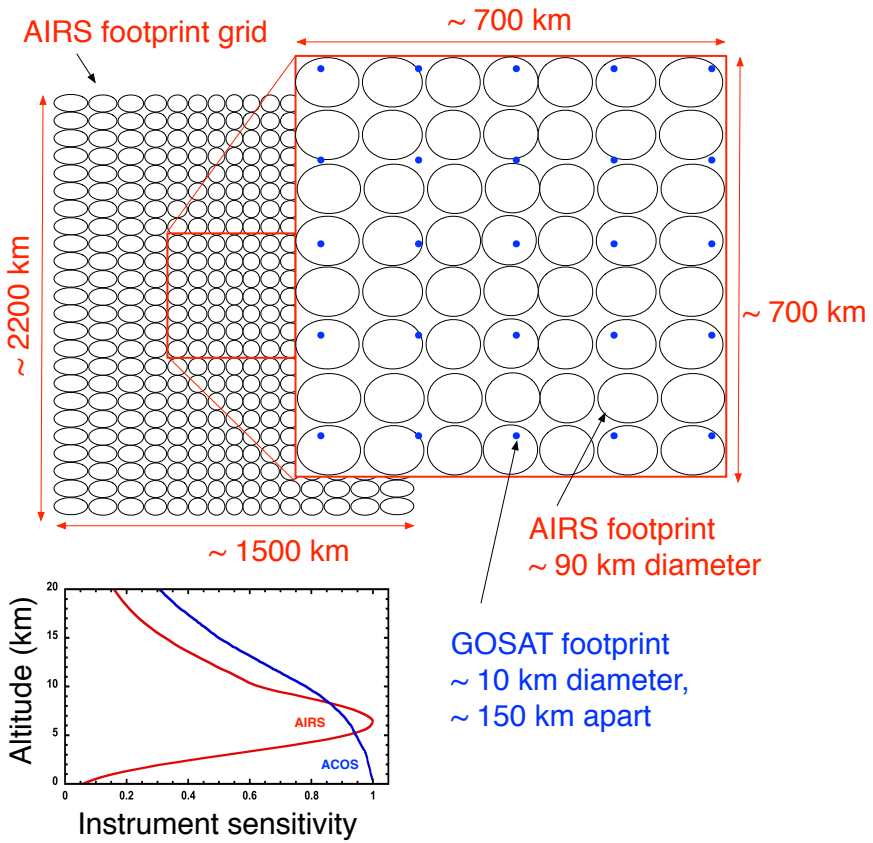
Data fusion example

Data fusion: optimal spatio-temporal inference on multiple processes from multiple, heterogenous datasets.

- $Y_{1t}(\cdot)$ = total column CO2 concentration at time t.
- $Y_{2t}(\cdot)$ = mid-tropospheric CO2 concentration at time t.
- $Y_t(\cdot) = (Y_{1t}(\cdot), Y_{2t}(\cdot))'$.
- $\mathbf{Y}_t = (Y_t(A_1), Y_t(A_2), \ldots, Y_t(A_P),)'$ is the vector of true process (bivariate) values at given support, A, at P locations.
- Goal: produce $\hat{\mathbf{Y}}_t = \left(\hat{Y}_t(A_1), \hat{Y}_t(A_2), \ldots, \hat{Y}_t(A_P), \right)'$ and associated mean squared prediction error matrices.

Data fusion example

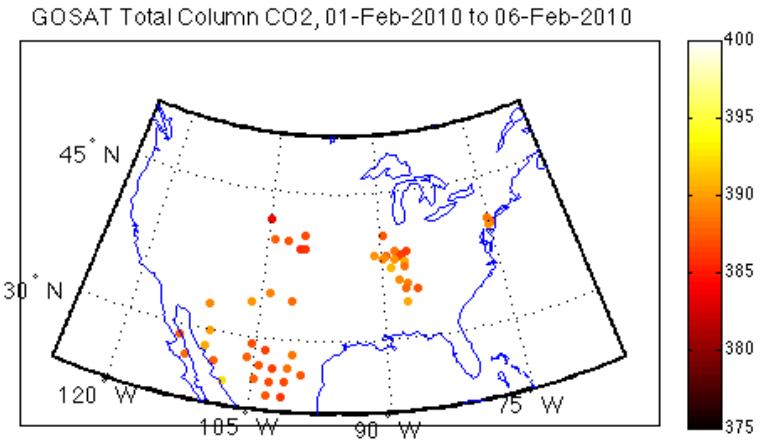
Data fusion: optimal spatio-temporal inference on multiple processes from multiple, heterogenous datasets.

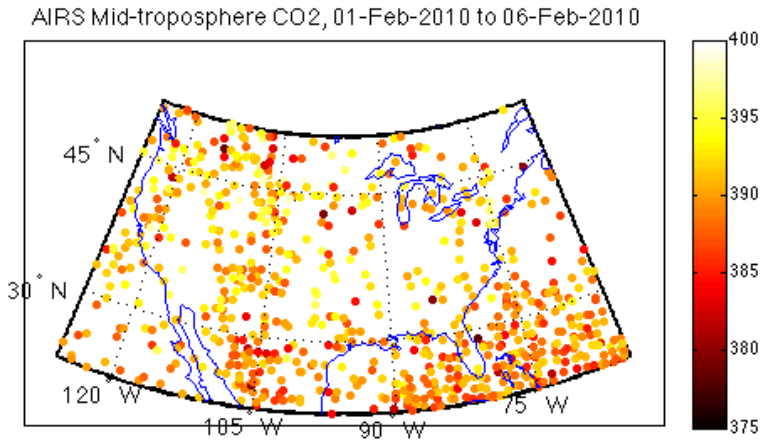


- $\mathbf{Z}_{1t}$ = vector of total column CO2 concentrations observed by JAXA's Greenhouse Gases Observing Satellite (GOSAT) at time t.
- $\mathbf{Z}_{2t}$ = vector of mid-tropospheric CO2 concentrations observed by NASA's Atmospheric Infrared Sounder (AIRS) at time t.
- $\mathbf{Z}_t = (\mathbf{Z}_{1t}', \mathbf{Z}_{2t}')'$.

Data fusion example

Data: remote sensing observations from GOSAT and AIRS, over the continental US, covering February 2010 through March 2011 in $T = 70$ six-day blocks.





Estimate $\mathbf{Y}_t = (Y_t(A_1), Y_t(A_2), \ldots, Y_t(A_P))'$, $t = 1, \ldots, 70$, at a set of P regularly spaced locations with $1^\circ \times 1^\circ$ block support. Produce covariance (uncertainty) matrix for each $\hat{Y}_t(A_p)$.

Data fusion example

$$\mathbf{Y}_t = \mathbf{T}_t\boldsymbol{\alpha}_t + \mathbf{S}_t\boldsymbol{\eta}_t + \boldsymbol{\xi}_t, \quad \mathbf{Z}_t = \mathbf{Y}_t + \boldsymbol{\epsilon}_t, \quad \Longrightarrow \quad \mathbf{Z}_t = \mathbf{T}_t\boldsymbol{\alpha}_t + \mathbf{S}_t\boldsymbol{\eta}_t + \boldsymbol{\xi}_t + \boldsymbol{\epsilon}_t.$$

- $\mathbf{Z}_t = \begin{pmatrix} \mathbf{Z}_{1t} \\ \mathbf{Z}_{2t} \end{pmatrix}$,
- $\mathbf{T}_t = \begin{pmatrix} \mathbf{T}_{1t} & \mathbf{0} \\ \mathbf{0} & \mathbf{T}_{2t} \end{pmatrix}$,

  $\mathbf{T}_{kt} = (\mathbf{t}(B_{k1t}), \ldots, \mathbf{t}(B_{kN_kt}))'$,
- $\boldsymbol{\alpha}_t = \begin{pmatrix} \boldsymbol{\alpha}_{1t} \\ \boldsymbol{\alpha}_{2t} \end{pmatrix}$,
- $\mathbf{S}_t = \begin{pmatrix} \mathbf{S}_{1t} & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_{2t} \end{pmatrix}$,
- $\boldsymbol{\eta}_t = \begin{pmatrix} \boldsymbol{\eta}_{1t} \\ \boldsymbol{\eta}_{2t} \end{pmatrix}$,
- $\boldsymbol{\xi}_t = \begin{pmatrix} \boldsymbol{\xi}_{1t} \\ \boldsymbol{\xi}_{2t} \end{pmatrix}$,

  $\boldsymbol{\xi}_{kt} = (\xi_{kt}(B_{k1t}), \xi_{kt}(B_{k2t}), \ldots, \xi_{kt}(B_{kN_kt}))'$,
- $\boldsymbol{\epsilon}_t = \begin{pmatrix} \boldsymbol{\epsilon}_{1t} \\ \boldsymbol{\epsilon}_{2t} \end{pmatrix}$,

  $\boldsymbol{\epsilon}_{kt} = (\epsilon_{kt}(B_{k1t}), \epsilon_{kt}(B_{k2t}), \ldots, \epsilon_{kt}(B_{kN_kt}))'$.

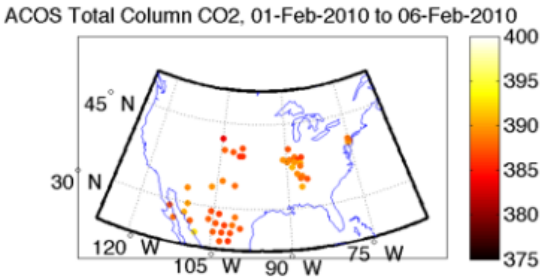
Now use the machinery discussed earlier for the single dataset case...

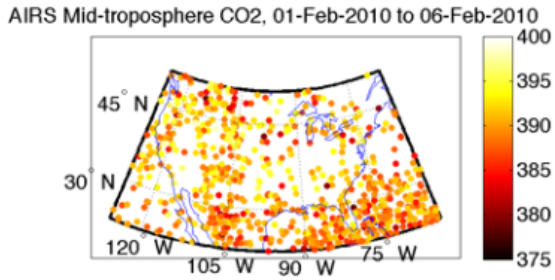
National Aeronautics and
Space Administration

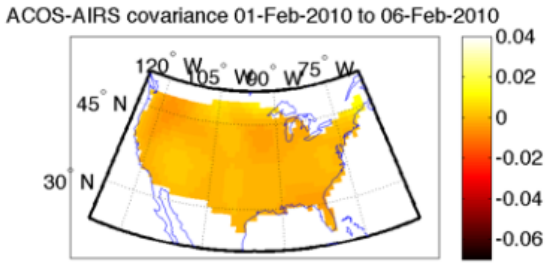
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

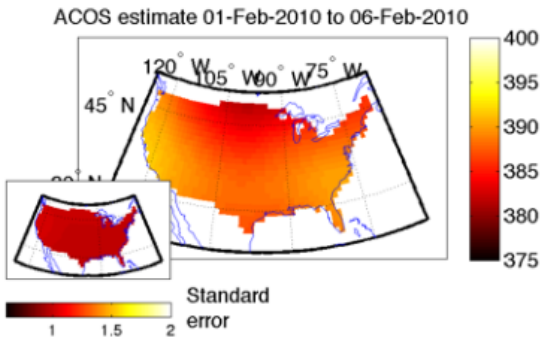
Data fusion example

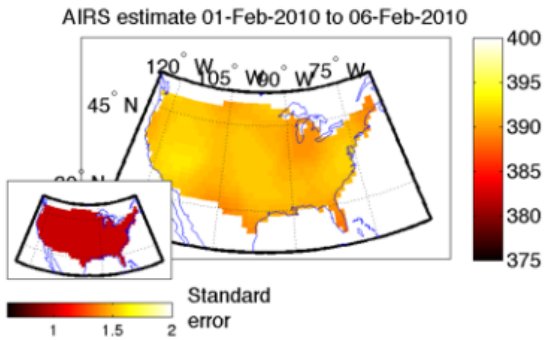
In addition to exploiting spatial and temporal dependence, data fusion also exploits *inter-variable* dependence.

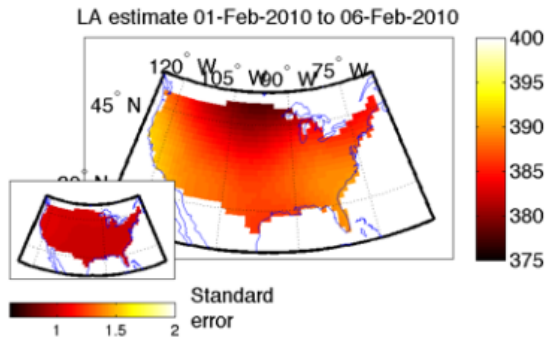












(Analysis took about 42 minutes on a MacBook Pro laptop with ∼140,000 observations.)

Some concluding thoughts

- The advent of massive climate model simulation datasets that instantiate hypotheses about the climate system *changes the landscape for climate informatics*. Provides a framework for hypothesis testing.
- Hypothesis testing requires uncertainties on the observational data sources.
- Remote sensing data are a vast, arguably untapped source of information about the climate system. However, they are themselves inferences ("retrievals"), and subject to uncertainty.
- There's still a lot to learn from them in an "exploratory" mode (finding patterns and empirical relationships).
- *We should begin to think about how to carry these findings forward to inferential conclusions about the climate system.* (Cross-validation error $\neq$ estimation uncertainty: even a massive dataset is a sample...)

References

- Nguyen, H. (2009). Spatial Statistical Data Fusion for Remote Sensing Applications, Ph.D. Dissertation, Department of Statistics, UCLA.
- Katzfuss, M. and Cressie, N. (2010). Spatio-temporal smoothing and EM estimation for massive remote-sensing datasets. *Journal of Time Series Analysis,* Vol. 32, pp. 430-446.
- Cressie, N., Shi, T., and Kang, E.L. (2010). Fixed-rank filtering for spatio-temporal data. *Journal of Computational and Graphical Statistics,* Vol. 19, No. 3, pp. 724-745.
- Kang, E.L., Cressie, N., and Shi, T. (2010). Using temporal variability to improve spatial mapping with application to satellite data. *Canadian Journal of Statistics,* Vol. 38, No. 2, pp. 271-289.
- Katzfuss, M. and Cressie, N. (2011). Bayesian hierarchical spatio-temporal smoothing for massive datasets. *Environmetrics,* Vol. 23, pp. 94-107.
- Nguyen, H., Cressie, N., and Braverman, A. (2012). Spatial Statistical Data Fusion for Remote-Sensing Applications, *Journal of the American Statistical Association,* to appear.
- Nguyen, H., Katzfuss, M., Cressie, N., and Braverman, A. (2012). Spatio-Temporal Data Fusion for Very Large Remote Sensing Datasets, *Technometrics,* in revision.

The end

Questions, comments?

Contact Amy.Braverman@jpl.nasa.gov.

Support for this work is provided by NASA's Earth Science Technology Office under its Advanced Information Systems Technology Program.