Spatial Statistical Data Fusion for Remote Sensing Applications

Hai Nguyen (398L), Peter Kalmus (398K), Amy Braverman (398L)

Jet Propulsion Laboratory
California Institute of Technology

Outline

- Motivation- Fusion of t and q from AIRS and CrIS
- Spatial Statistical Data Fusion
- Bias Estimation
- Results

Motivation

What is the benefit of data fusion?

- Data collection is often incomplete, sparse, and yields spatially incompatible results. Our goal is to infer the true process from all available data sources.
- Data fusion can capitalize on complementary strengths to minimize prediction errors.

Difficulties encountered when fusing remote sensing datasets:

- massive size,
- change of support,
- isotropy and stationarity,
- bias correction.

Examples

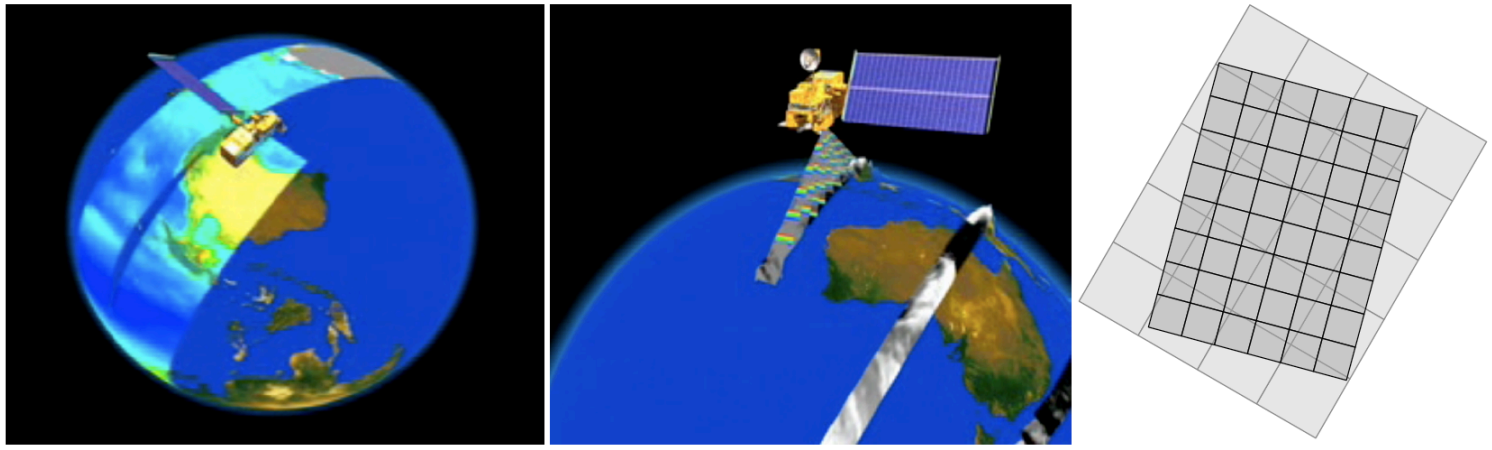


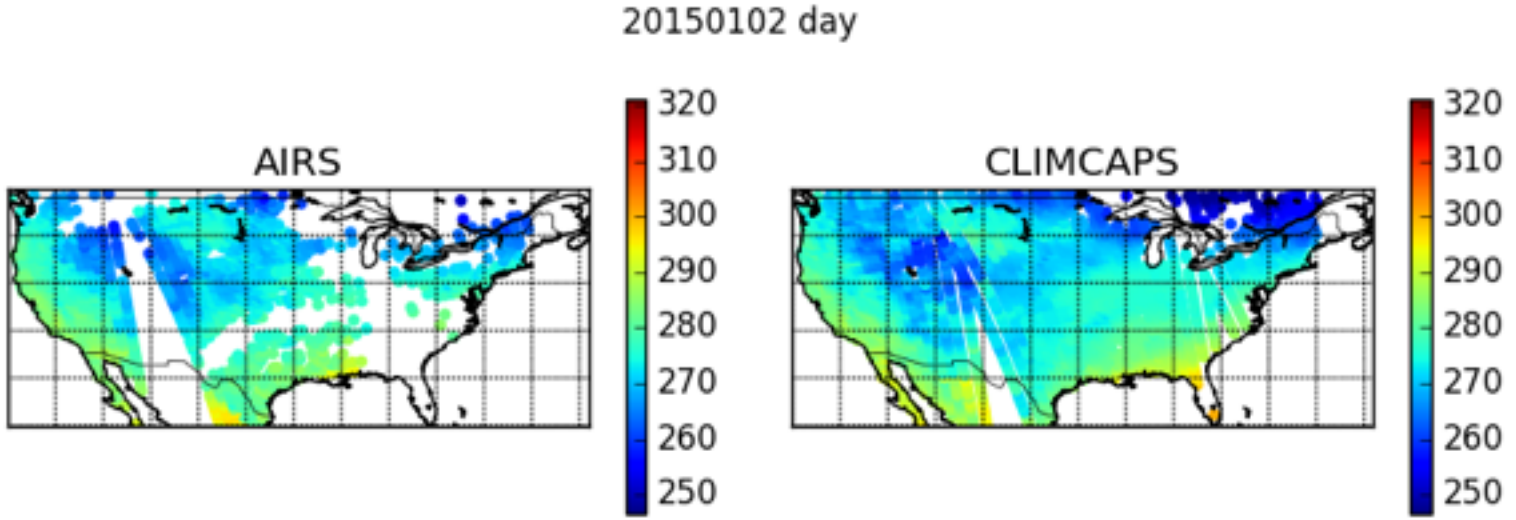
Figure 1: Example satellite swatch from two different instruments and their overlapping footprints (rightmost panel)

Example fusion applications

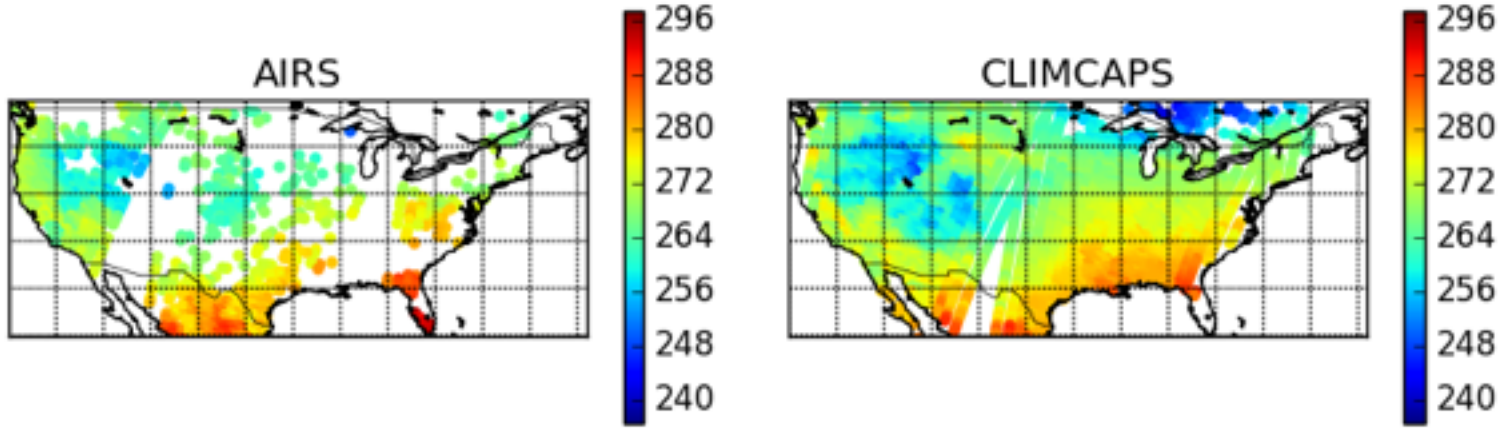
- XCO2 from Orbiting Carbon Observatory-2 (OCO-2) and Greenhouse gases Observing SATellite (GOSAT)
- Temperature and water vapor from Atmospheric Infrared Sounder (AIRS) and Cross-track Infrared Sounder (CrIS)
- Sea Surface Temperature from AQUA and TERRA satellites
- Aerosol optical depth from the Multi-angle Imaging SpectroRadiometer (MISR) and the Moderate Resolution Imaging Spectroradiometer (MODIS) instruments
- In an era of multiple observations of the same variable in Earth's system, data fusion makes sense both logistically and scientifically.

AIRS and CrIS (CLIMCAPS) near surface temperature

Daytime

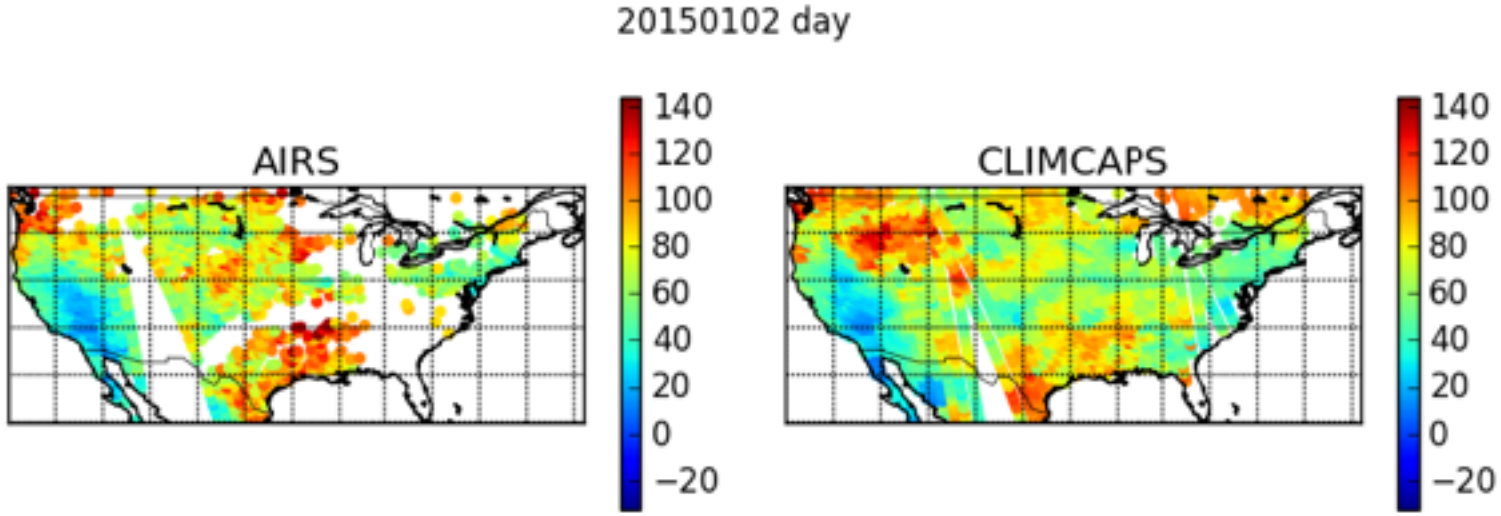


Nighttime

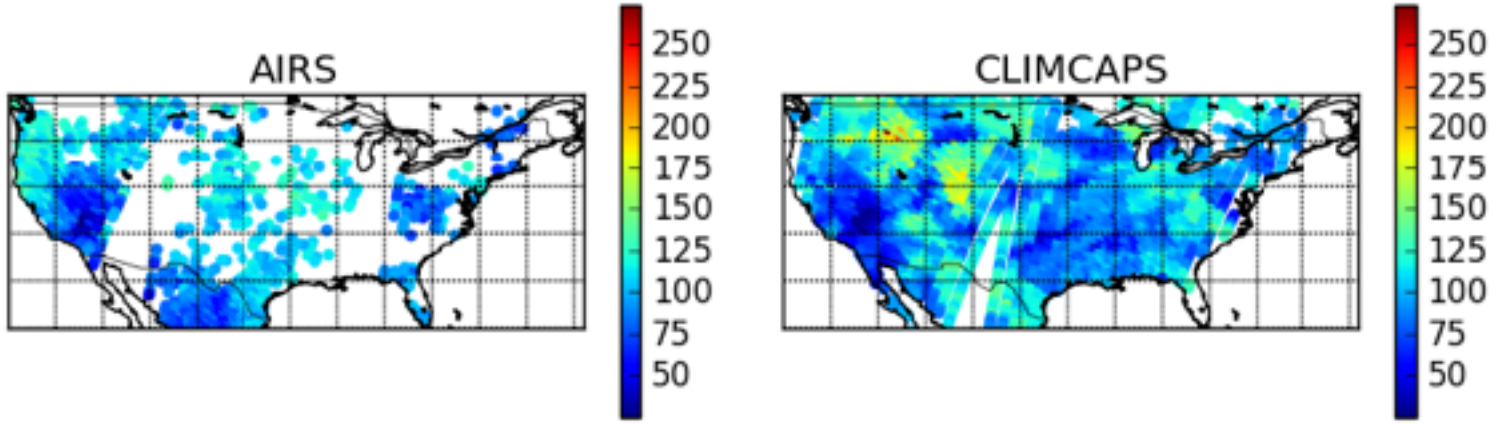


AIRS and CrIS water vapor

Daytime



Nighttime



AIRS and CrIS fusion

- AIRS records temperature and water vapor from 2002-present. The Cross-track Infrared Sounder (CrIS) instrument is a follow-up mission that was launched in 2011.

- The goal of fusion between AIRS and CrIS is to produce a long climate record that span the life-time of both missions.

- Our fusion methodology (Nguyen et al., 2012) is based on kriging, which is a best linear unbiased predictor and which produces estimates of uncertainties.

Overview of kriging

We assume the data are generated according to the following model:

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

where

- $\mathbf{s}_i$ is the ith footprint ,
- $\mathbf{Z}$ is the vector of response variable,
- $Y(\cdot)$ is the true process,
- $\epsilon(\cdot)$ is the error process.

Overview of kriging

Under this formulation, the (linear unbiased) optimal interpolation can be written as

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

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

Overview of kriging

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

$$\begin{aligned} E(Y(\mathbf{s}) - \hat{Y}(\mathbf{s}))^2 &= \mathrm{var}(Y(\mathbf{s}) - \mathbf{a}'\mathbf{Z}), \\ &= \mathrm{var}(Y(\mathbf{s})) - 2\mathbf{a}'\,\mathrm{cov}(\mathbf{Z}, Y(\mathbf{s})) + \mathbf{a}'\,\mathrm{var}(\mathbf{Z})\,\mathbf{a}, \qquad (1) \end{aligned}$$

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

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

Overview of kriging

We can solve (1) for the optimal $\mathbf{a}$ using the method of Lagrange multiplier. The equation for the optimal kriging coefficients is,

$$\begin{pmatrix} \mathbf{\Sigma} & \mathbf{1} \\ \mathbf{1}' & 0 \end{pmatrix} \begin{pmatrix} \mathbf{a} \\ \lambda \end{pmatrix} = \begin{pmatrix} \mathbf{c}(\mathbf{s}) \\ 1 \end{pmatrix} \tag{2}$$

where

- $\mathbf{\Sigma} \equiv \mathrm{var}(\mathbf{Z})$
- $\mathbf{c}(\mathbf{s}) \equiv \mathrm{cov}(\mathbf{Z}, Y(\mathbf{s}))$
- λ is the Lagrange multiplier

Spatial Random Effects model

We assume that the spatial process $Y(\cdot)$ has the following model,

$$Y(\mathbf{s}) = \mathbf{S}(\mathbf{s})'\boldsymbol{\eta},$$

which leads to the following covariance model,

$$\boldsymbol{\Sigma} \equiv \mathrm{var}(\mathbf{Z}) = \mathbf{S}'\mathbf{K}\mathbf{S} + \mathbf{D},$$

where

- $\mathbf{S}(\mathbf{s})$ is an r-dimensional basis expansion of $\mathbf{s}$, and $r << N$,
- $\mathbf{S} \equiv (\mathbf{S}(\mathbf{s}_1), \ldots, \mathbf{S}(\mathbf{s}_N))'$,
- $\mathbf{K} = \mathrm{var}(\boldsymbol{\eta})$: fixed dimension $r \times r$,
- $\mathbf{D}$ is the variance-covariance matrix of the measurement errors.

Inversion of Σ

Since $\mathbf{\Sigma}$ has the convenient form

$$\mathbf{\Sigma} \equiv \mathrm{var}(\mathbf{Z}) = \mathbf{S}'\mathbf{K}\mathbf{S} + \mathbf{D},$$

We can quickly invert $\mathbf{\Sigma}$ using the Sherman-Morrison-Woodbury formula

$$\mathbf{\Sigma}^{-1} = \mathbf{D}^{-1} - \mathbf{D}^{-1}\mathbf{S}'\left(\mathbf{K}^{-1} + \mathbf{S}\mathbf{D}^{-1}\mathbf{S}'\right)^{-1}\mathbf{S}\mathbf{D}^{-1}. \qquad (3)$$

Fusion of two instruments

We can rewrite the data models in vector forms as

$$\begin{aligned} \mathbf{Z}_1 &= \mathbf{S}_1'\boldsymbol{\eta} + \boldsymbol{\epsilon}_1 \\ \mathbf{Z}_2 &= \mathbf{S}_2'\boldsymbol{\eta} + \boldsymbol{\epsilon}_2 \end{aligned}$$

Given the data vectors $\mathbf{Z}_1$ and $\mathbf{Z}_2$, we can simplify the problem by stacking them to form the following model,

$$\begin{pmatrix} \mathbf{Z}_1 \\ \mathbf{Z}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{S}_1' \\ \mathbf{S}_2' \end{pmatrix} \boldsymbol{\eta} + \begin{pmatrix} \boldsymbol{\epsilon}_1 \\ \boldsymbol{\epsilon}_2 \end{pmatrix},$$

or equivalently,

$$\mathbf{Z}_F = \mathbf{S}_F'\boldsymbol{\eta} + \boldsymbol{\epsilon}_F.$$

Application to AIRS and CrIS

- We reviewed Spatial Statistical Data Fusion, which can handle massive datasets through the Spatial Random Effects model (along with change-of-support and stationarity)
- We need to, however, estimate and remove instrument biases from AIRS and CrIS before applying data fusion
- We choose to use Integrated Surface Database (ISD) as a data source for validation.

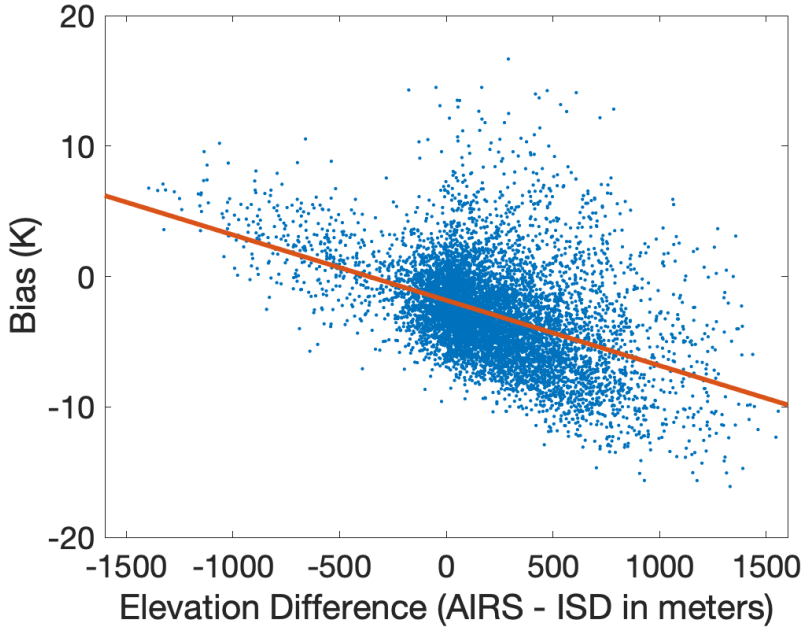
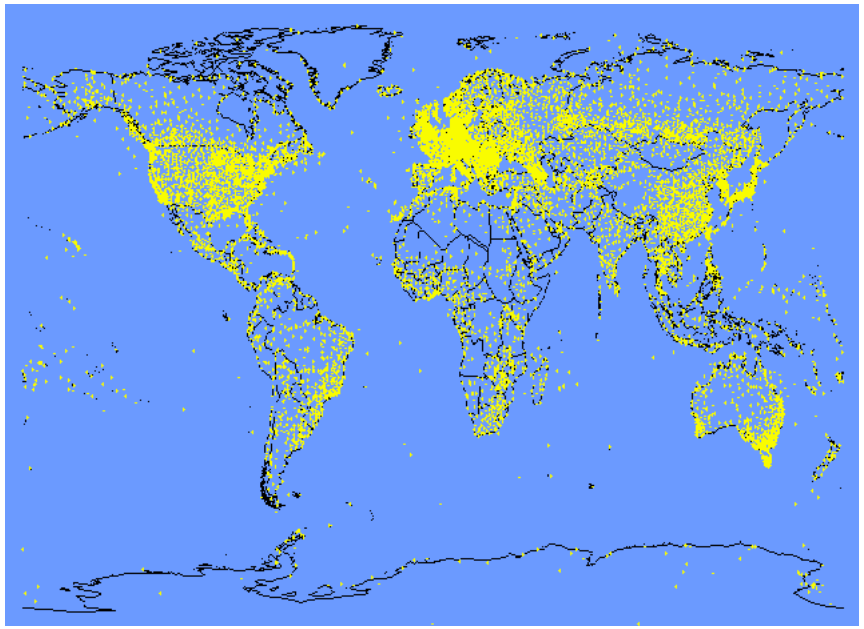


Figure: Bias between AIRS and ISD as a function of elevation difference

Integrated Surface Database (ISD)

- ISD consists of global hourly and synoptic observations compiled from numerous sources
- Here, we matched ISD temperature and water vapor to AIRS and CrIS, respectively, to compute their biases
- These biases are computed based on a sliding temporal (30 minutes) and spatial window (45 km)
- We did find an major contributing factor to the biases is the elevation difference, which we model and remove using random forest.



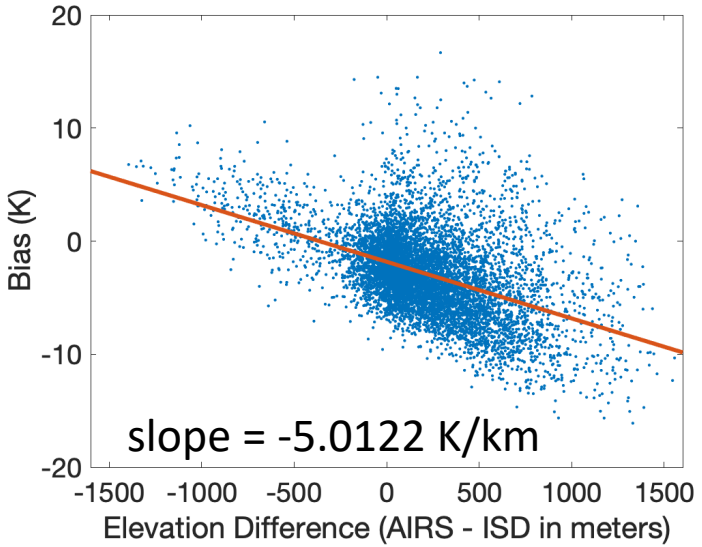
Map of Station Locations for ISD

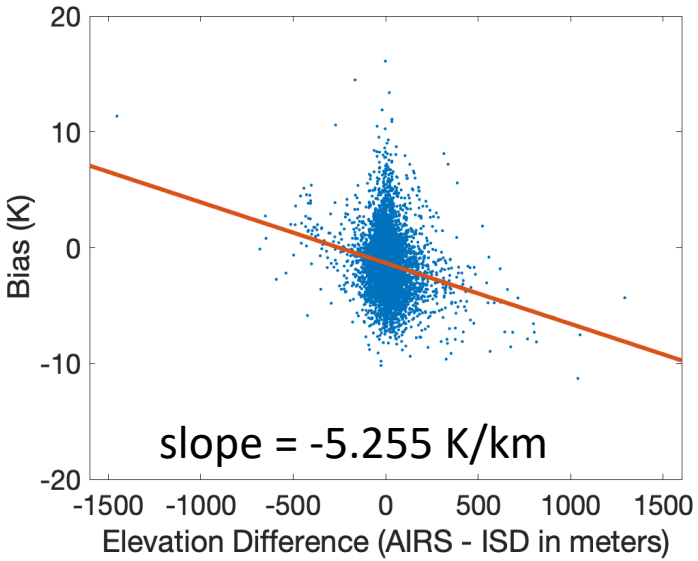
Elevation difference versus bias

West (lon < -105)

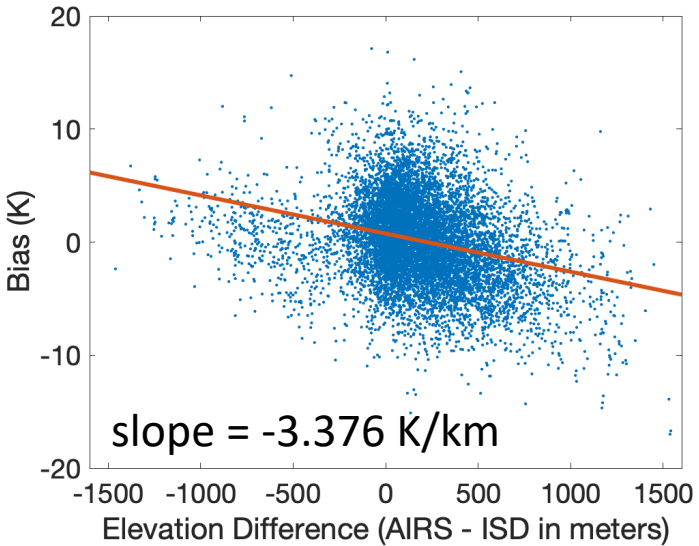
East (lon > -105)

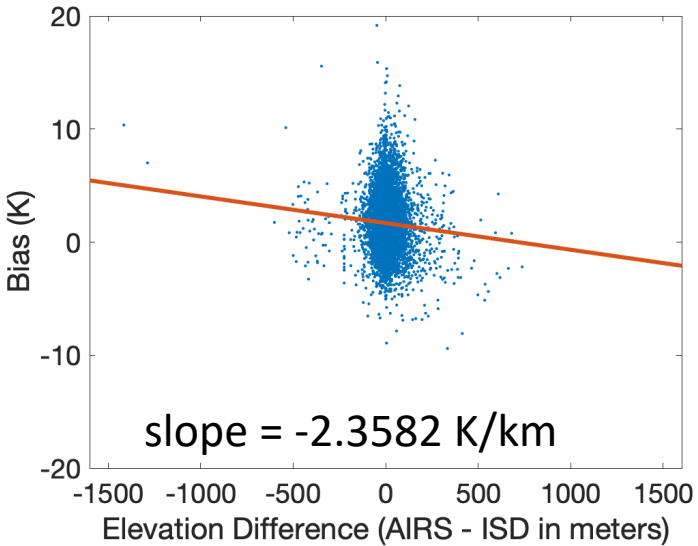
Daytime



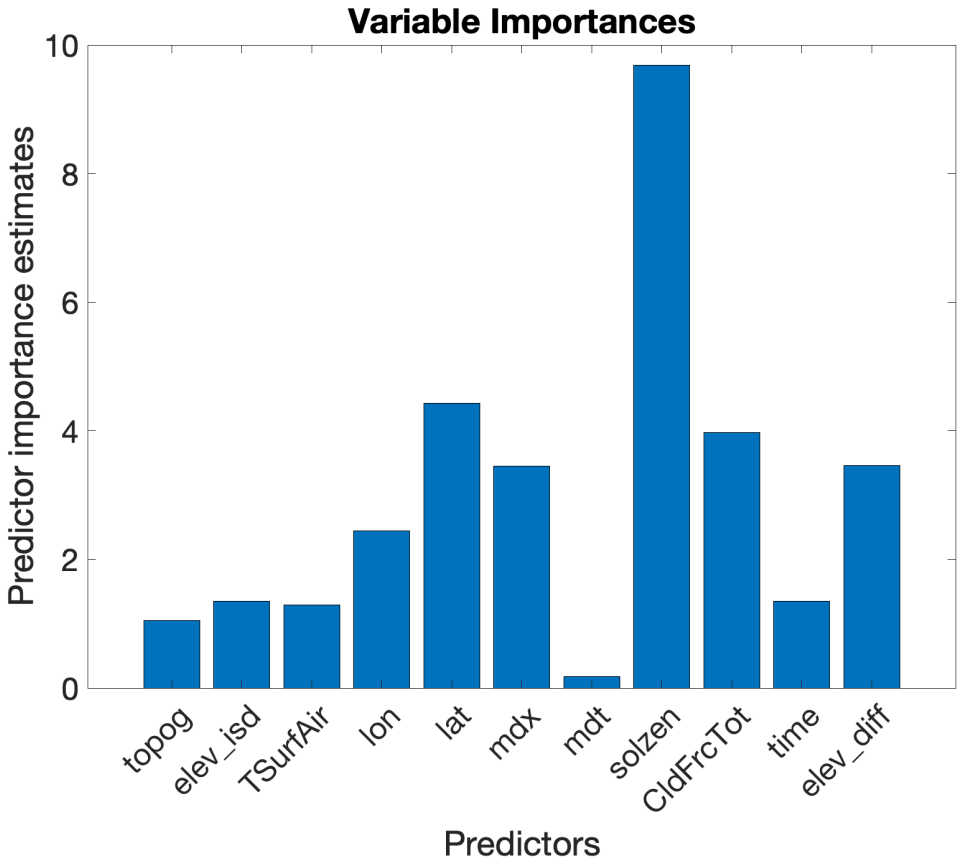


Nighttime

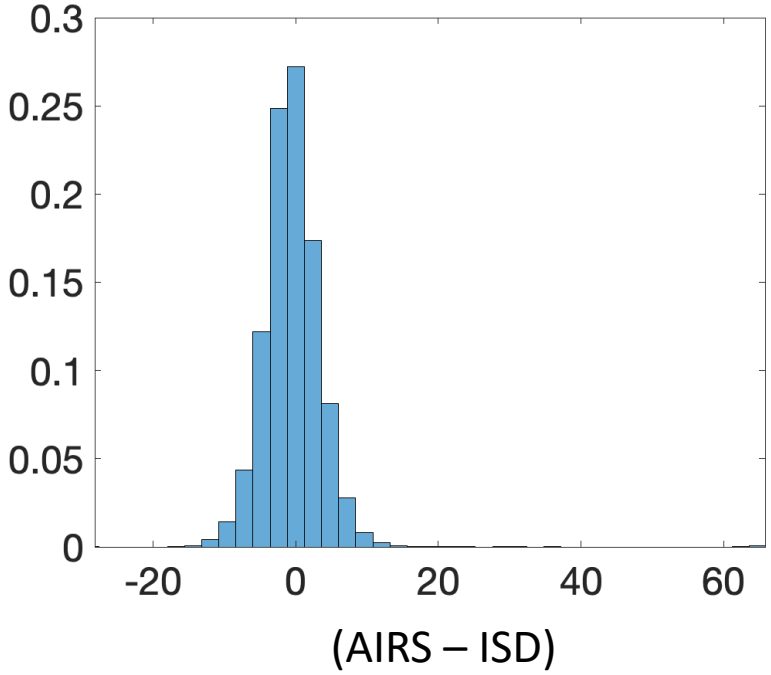




Random Forest bias estimation



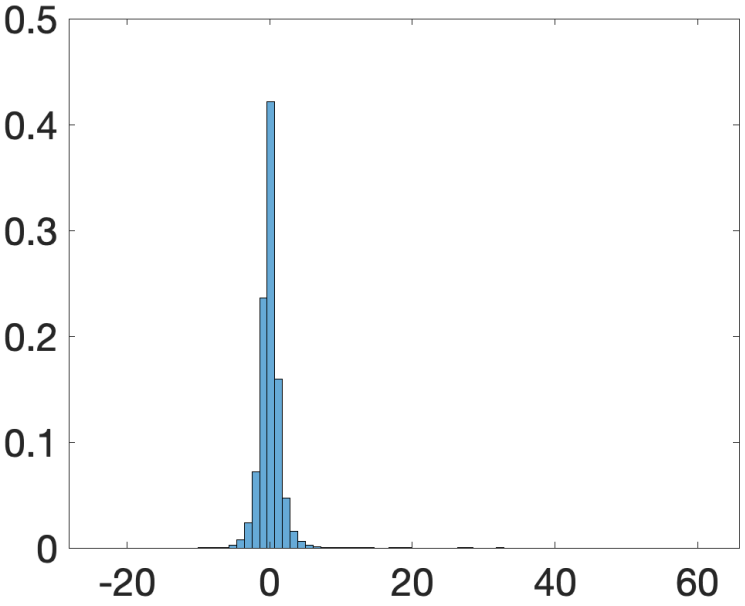
Variable Importance



(AIRS – ISD)

STD = 4.01K

Mean = -.5 K



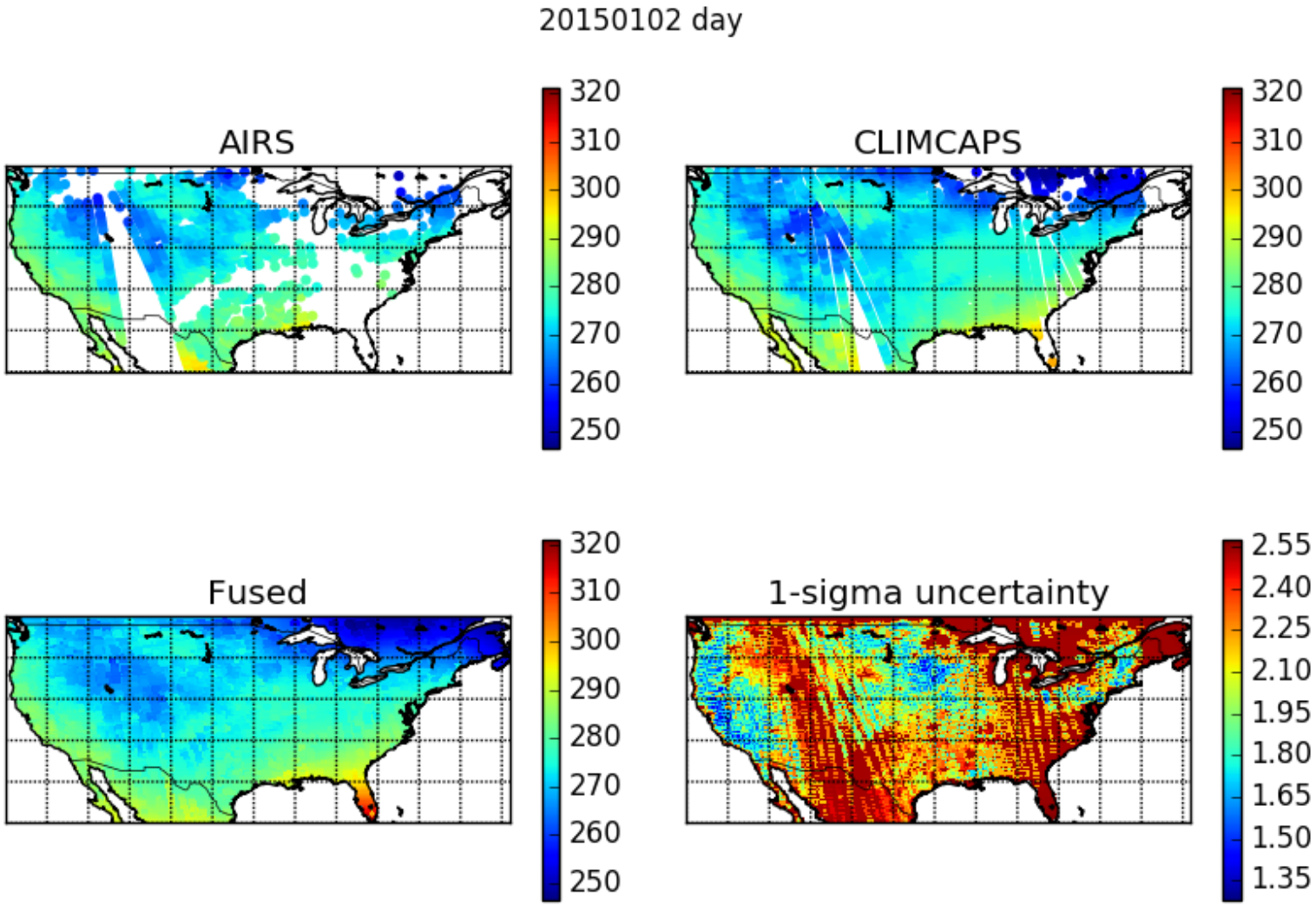
([AIRS - $\overline{bias}$]- ISD)

STD = 1.41 K

Mean = .01 K

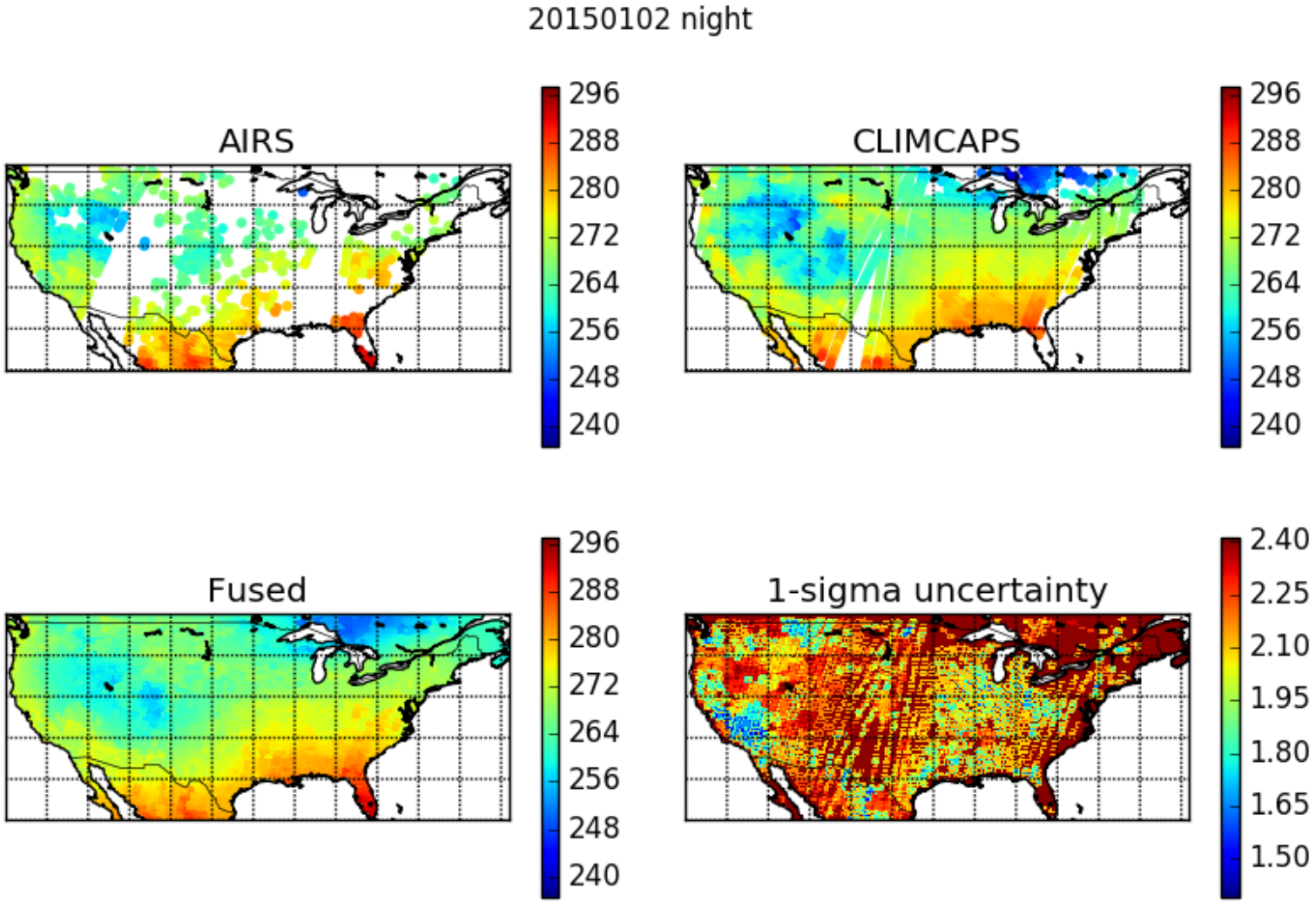
Near Surface Temperature

Near-surface Temperature example



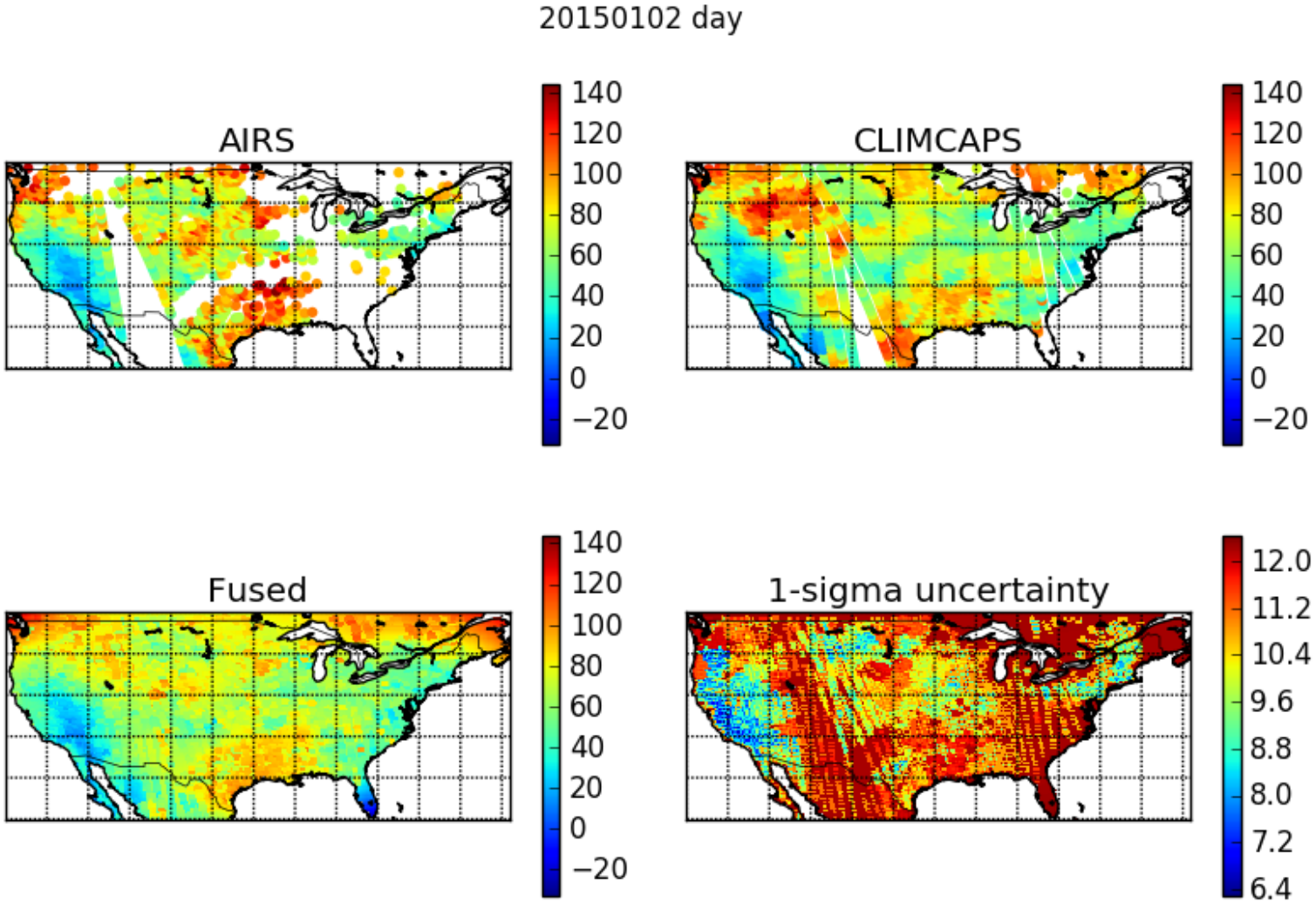
Near Surface Temperature

Near-surface Temperature example



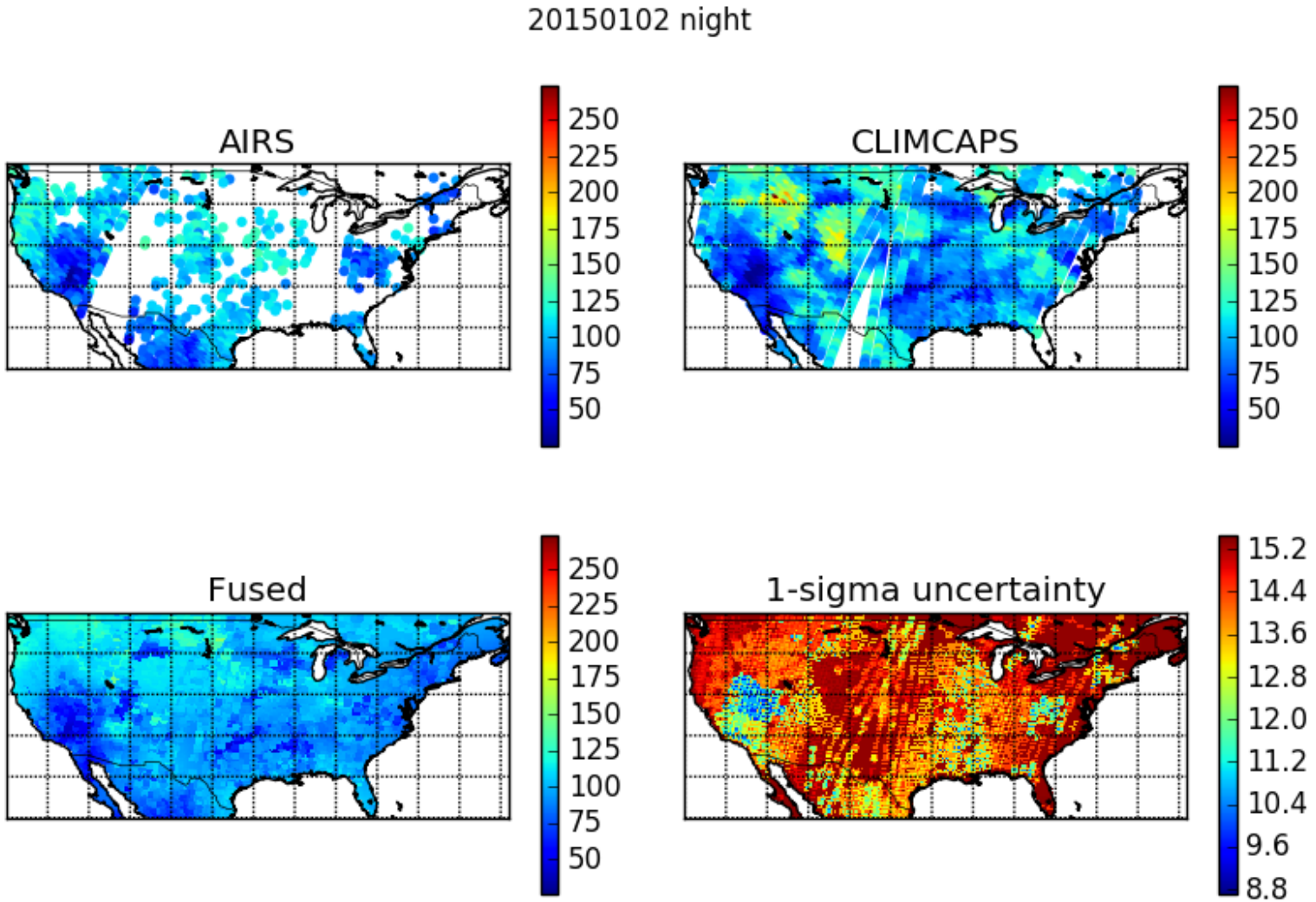
Relative Humidity

Near-surface RH example



Relative Humidity

Near-surface RH example



Summary

- One choice of remote sensing data fusion is kriging, also known as optimal interpolation or Gaussian process regression
- Kriging is typically $O(N^3)$, but here we describe a methodology that is linear $O(Nr^3)$ using the Spatial Random Effects model
- Application to AIRS and CrIS requires a elevation-dependent bias model, which we built using random forest and ISD data.
- For other applications, getting accurate and reliable bias estimates (and uncertainties) are crucial for data fusion, but such validation are often limited by the availability of validation data.

References

- Cressie, Noel, and Gardar Johannesson. "Fixed rank kriging for very large spatial data sets." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70.1 (2008): 209-226.

- Nguyen, Hai, Noel Cressie, and Amy Braverman. "Spatial statistical data fusion for remote sensing applications." *Journal of the American Statistical Association* 107.499 (2012): 1004-1018.

- Nguyen, Hai, et al. "Spatio-temporal data fusion for very large remote sensing datasets." *Technometrics* 56.2 (2014): 174-185.

- Nguyen, Hai, Noel Cressie, and Amy Braverman. "Multivariate spatial data fusion for very large remote sensing datasets." *Remote Sensing* 9.2 (2017): 142.