Jet Propulsion Laboratory
California Institute of Technology

Update on AIRS+CrIS data fusion of near-surface T, RH, and VPD

Peter Kalmus, Hai Nguyen, Amy Braverman

Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA

AIRS Science Team Meeting

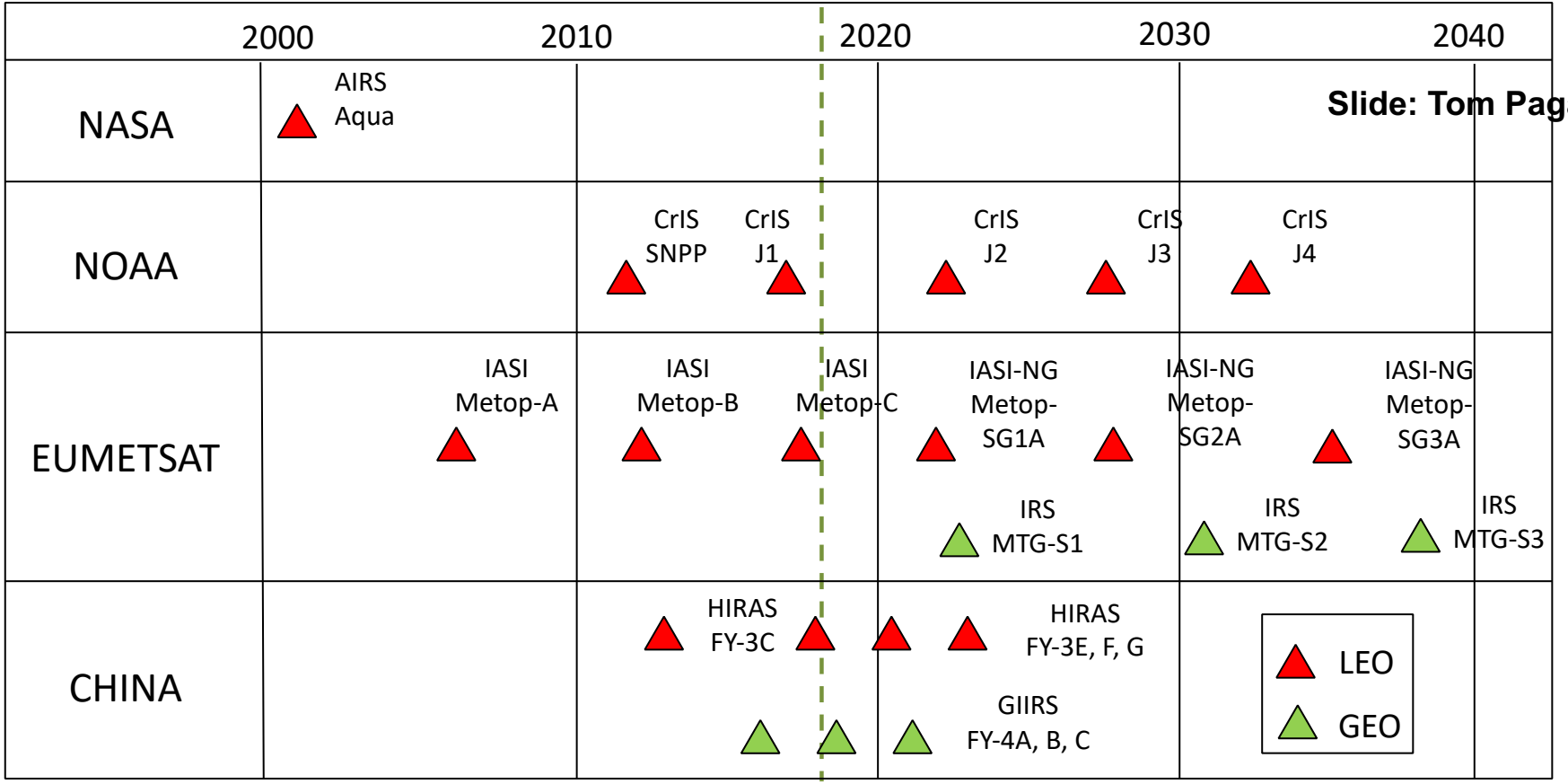
April 5, 2019

Copyright 2019. California Institute of Technology.
Government sponsorship acknowledged.

NOTES FROM THE MEETING

Time to improve bias estimation, possibly with UQ, to improve variance.
Fix CONUS blue dot bug.
Fix RH bug.
Fix long tail bug (note: Berry did not see it, so maybe it’s in my validation?)
CLIMCAPS! (when?)
Berry should do a VPD validation.
Need to meet up with Alireza to get this product into applications.
Should we use more ground truth besides ISD?
How to do a sine wave model and pull in IASI?
Go global

AIRS Success Led to Numerous Operational Sounders Worldwide



Slide: Tom Pagano

AIRS IR radiance product accuracy estimates provided to the Global Space-based Inter-Calibration of Sensors (GSICS). AIRS is an in-orbit intercalibration reference sensor for GSICS.

How to optimally utilize this information?

- Start seeing these instruments as a network of information providers
- We need good uncertainty estimates to do good science!

- Spatial statistical data fusion (SSDF, Nguyen et. al., 2012)
  - can fuse *M* estimates of the same process (sparse, incomplete OK)
  - Produces estimates in data gaps
  - Requires estimates of bias and variance for each data set
  - Quantifies uncertainty in the fused product

Fusion of AIRS + CrIMSS CHART near-surface temps

Eastern US: delivered

8 months: [Jan, Apr, Jul, Oct] x [2013, 2015]

Fusion done on a daily basis, separately for day & night, 0.5° output grid

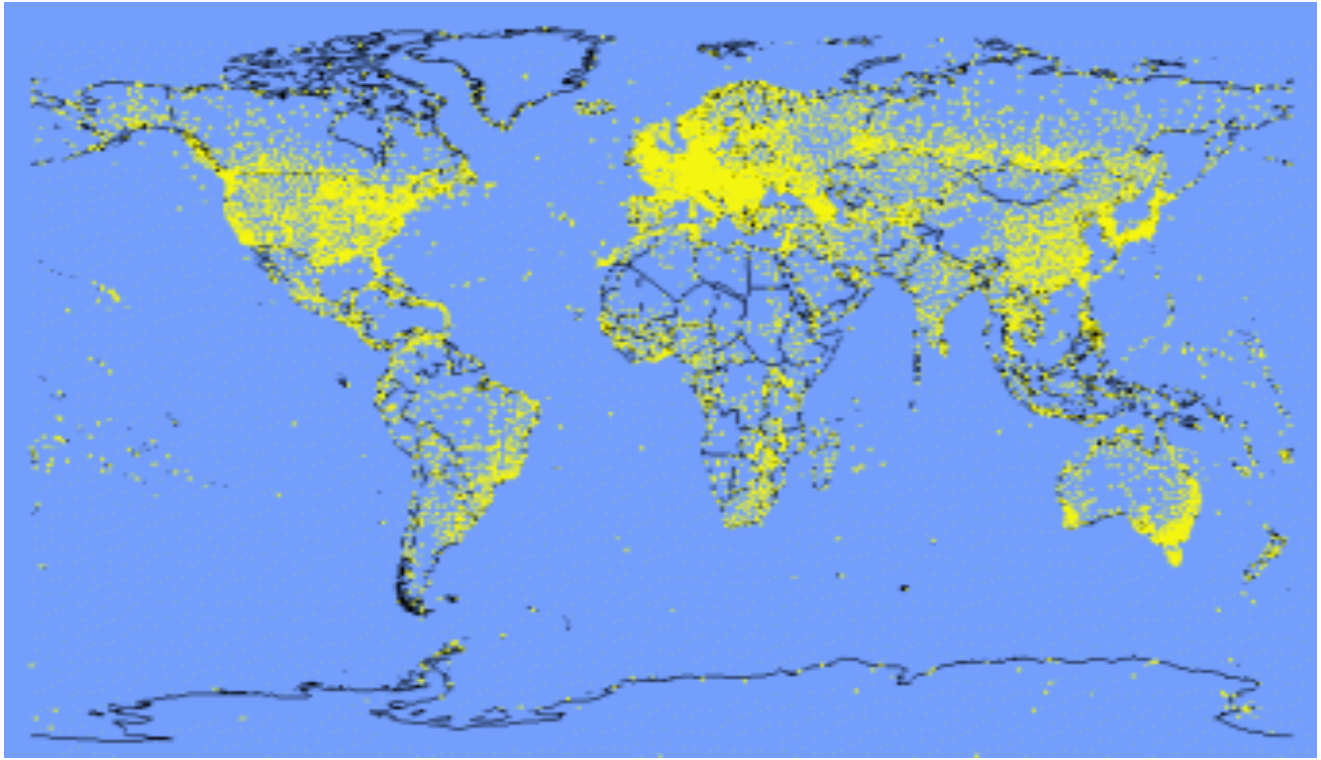
- AIRS v6 IR+MW support product
- CrIMSS CHART IR+MW
- DQ 0, 1

1. Estimate bias and variance of AIRS and CrIMSS data, vs. NOAA ISD truth
   a. Matchup criteria: 45 km, ±30 minutes (ISD is typically every 15 minutes)
   b. Bias, variance estimates made in 2-degree hexagonal bins, 3-day averages
2. Perform data fusion with **pySSDF**
3. Validate fused result

NOAA ISD
Integrated Surface Database

14k active stations globally

https://www.ncdc.noaa.gov/isd

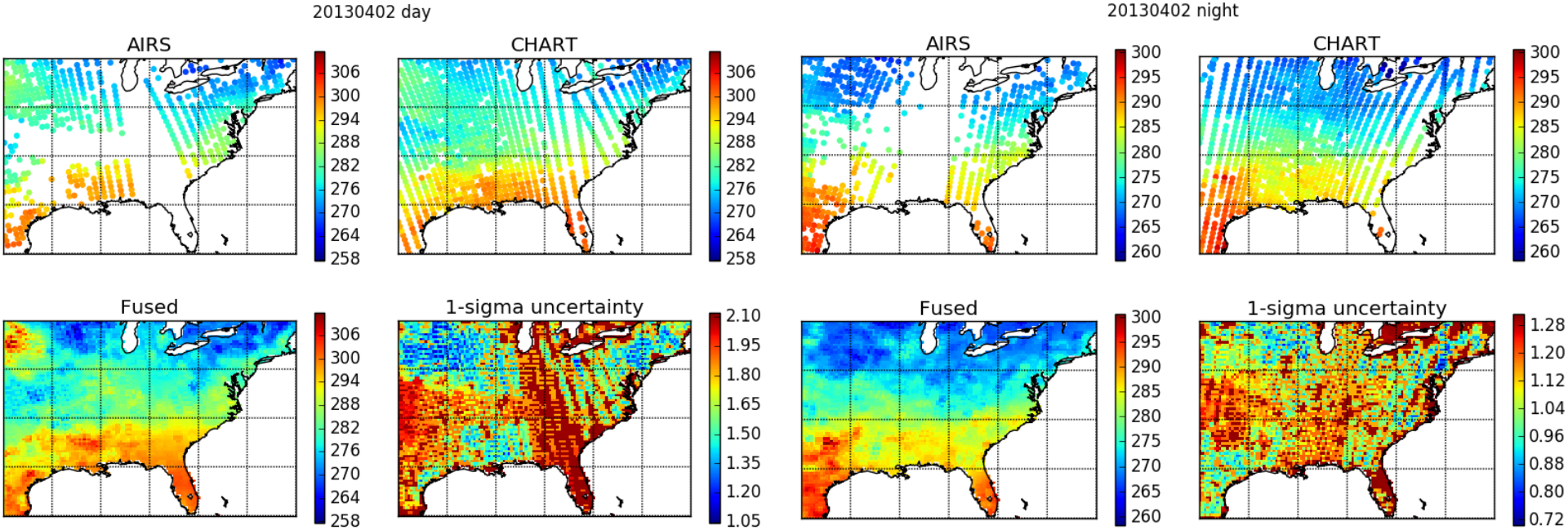


SSDF – spatial statistical data fusion

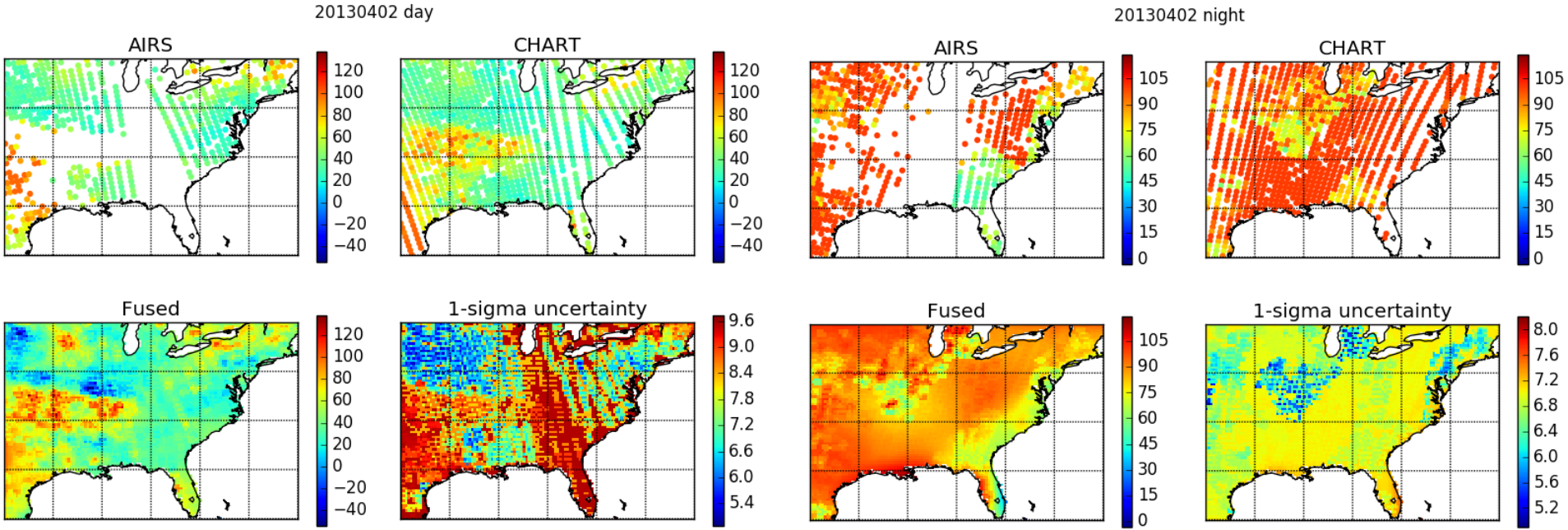
- Generalization of optimal interp. (kriging). Nguyen et al. 2012, 2014, 2017
- SSDF accounts for heterogeneities in input data with a spatial statistical model relating the true quantity of interest at a particular location, and all the observations at all locations from all data sources
  - estimate bias and variance of each input data set relative to ground truth
  - subtract bias from each data set
  - make optimal (minimum RMSE) estimates of a process of interest (i.e. temperature) from spatial data by accounting for spatial dependence structure
    - *best linear unbiased estimates:* standard errors are guaranteed to be smallest possible
    - estimate at new location a function of data covariance matrix (correlation between observed data points) and prediction covariance vector (correlation between observed data points and new location)
    - variance informs the weighting of the various input data sets
    - converge on optimal estimate using expectation-maximization algorithm

Sample of delivered Eastern CONUS results: NST

fast: 2 hours to fuse all 8 months of NST on one core on `weather`



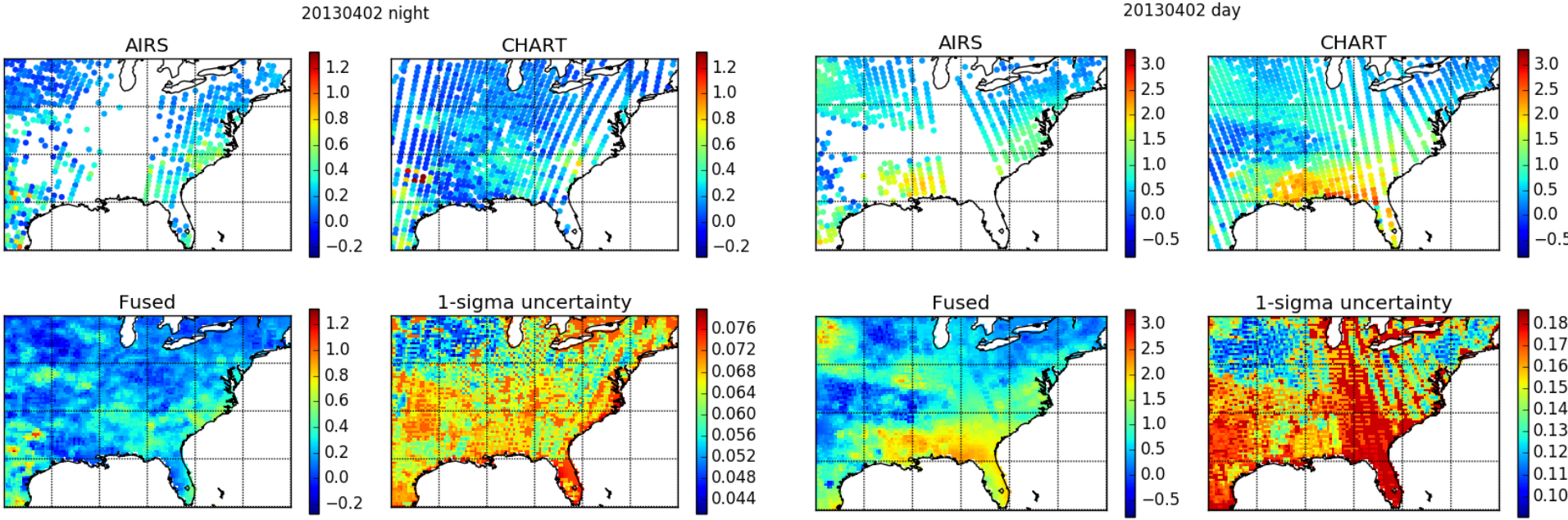
Sample of delivered Eastern CONUS results: NSRH



Sample of delivered Eastern CONUS results: VPD

VPD = humidity.esat(NST+273.15)/1000.*(1.-NSRH/100.) in kPa

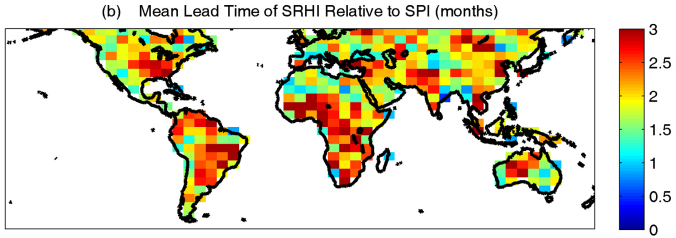
VPD of 0.5-1.2 kPa “OK” Too low→mold Too high→water stress



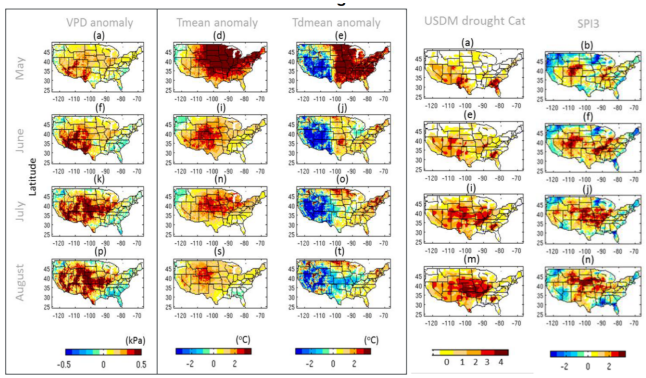
Drought Early Detection

- Relative Humidity, Temperature and Vapor Pressure Deficit can potentially detect drought onset ***earlier*** than other indicators for up to two months

Also:
Fire risk assessment
Agricultural applications



RH: Farahmand et al 2015

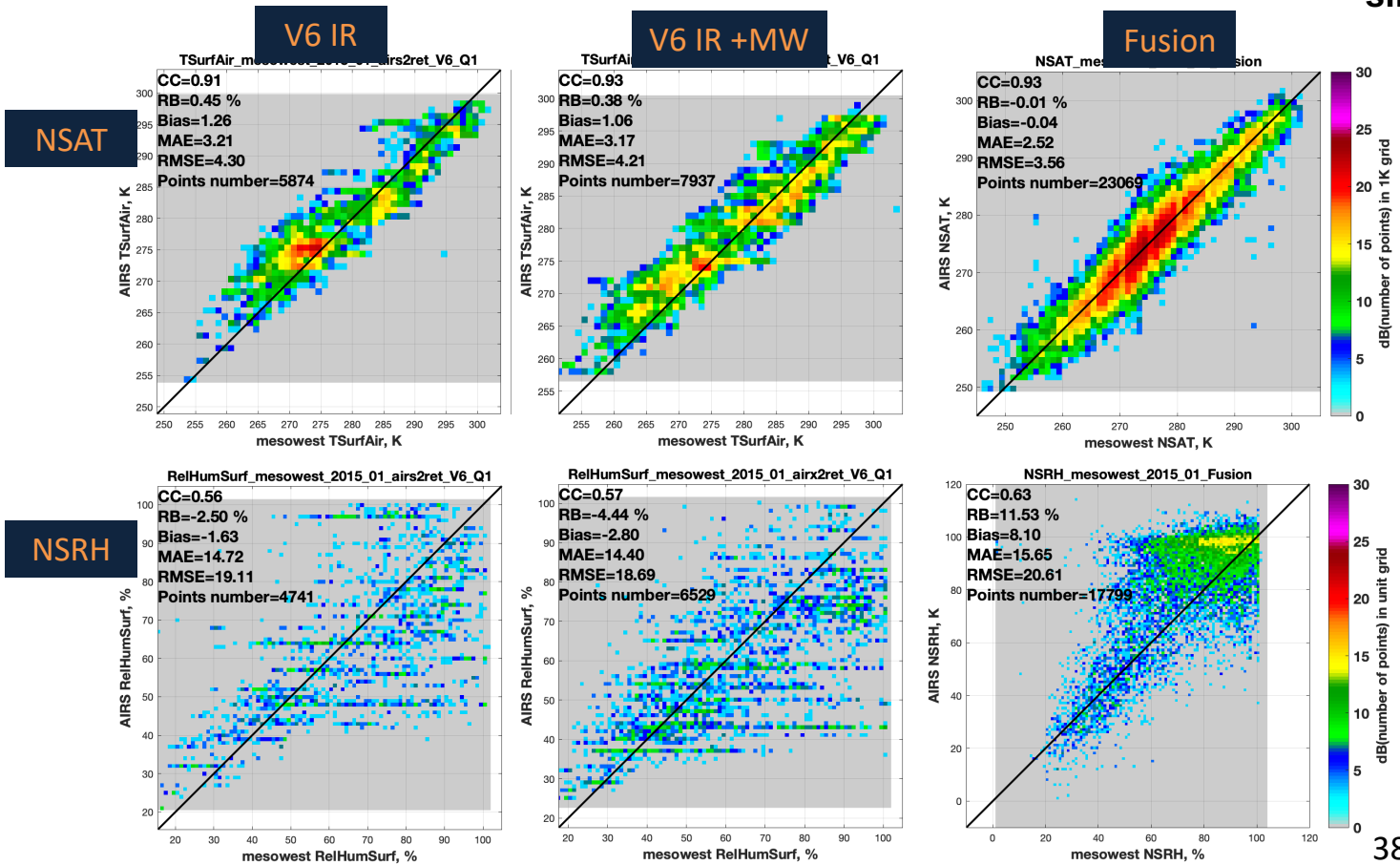


VPD and T: Behrangi et al 2015

Slide: Alireza Farahmand

Preliminary results: 2015 Jan, Q==0

slide: Berry Wen



Not bad, but room for improvement

Mean values over all 8 months

monthly

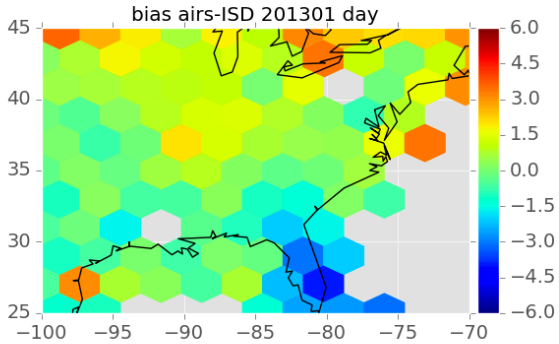
bias airs:	0.62
bias chart:	0.58
bias fused:	**0.14**
std airs:	2.55
std chart :	2.93
std fused:	**2.56**

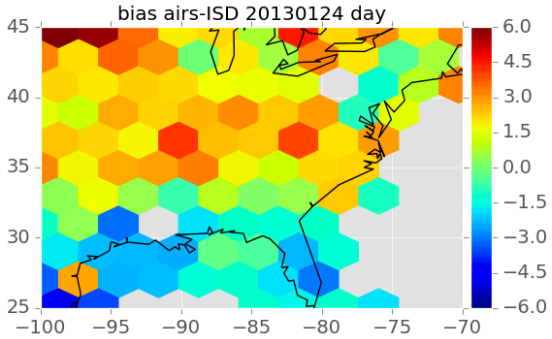
7-day

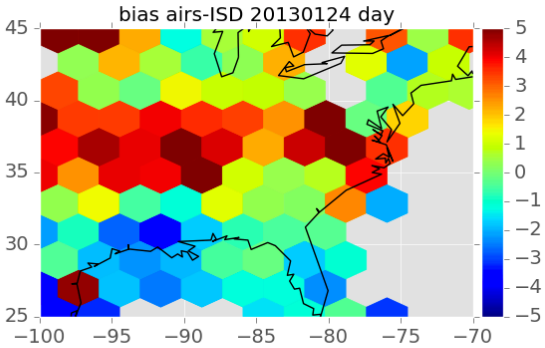
bias airs:	0.63
bias chart :	0.57
bias fused:	**0.12**
std airs:	2.56
std chart :	2.94
std fused:	**2.45**

3-day

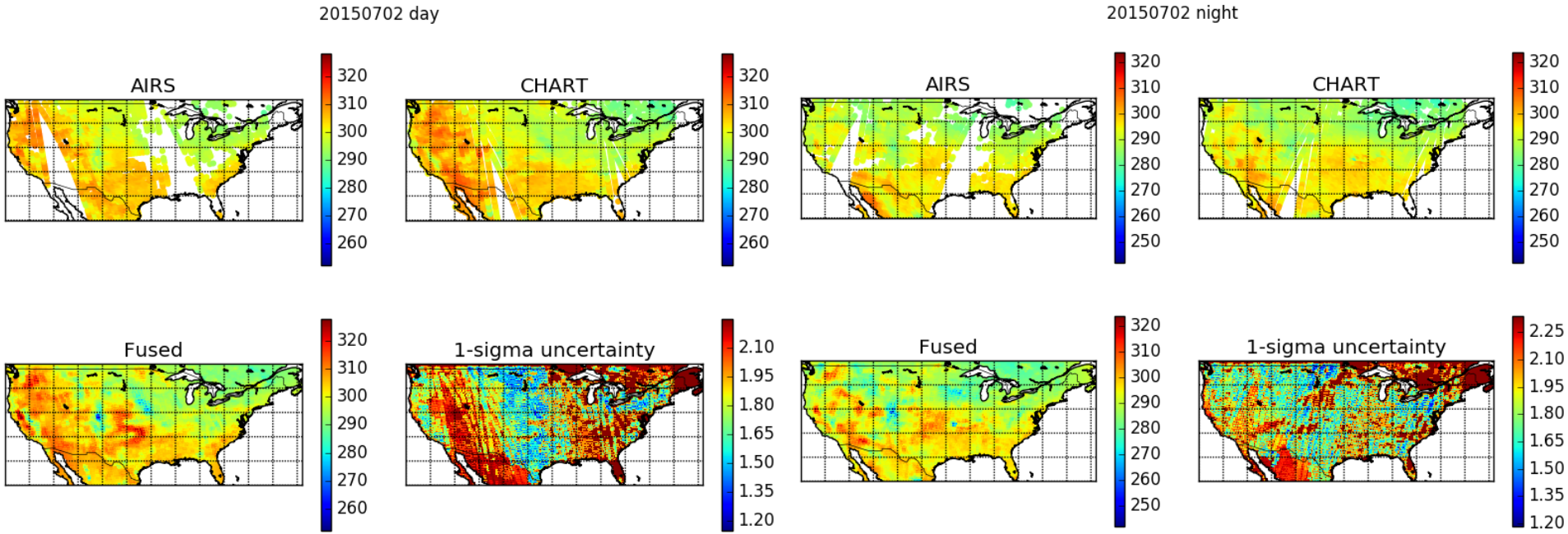
bias airs:	0.62
bias chart :	0.58
bias fused:	**0.13**
std airs:	2.56
std chart :	2.94
std fused:	**2.25**



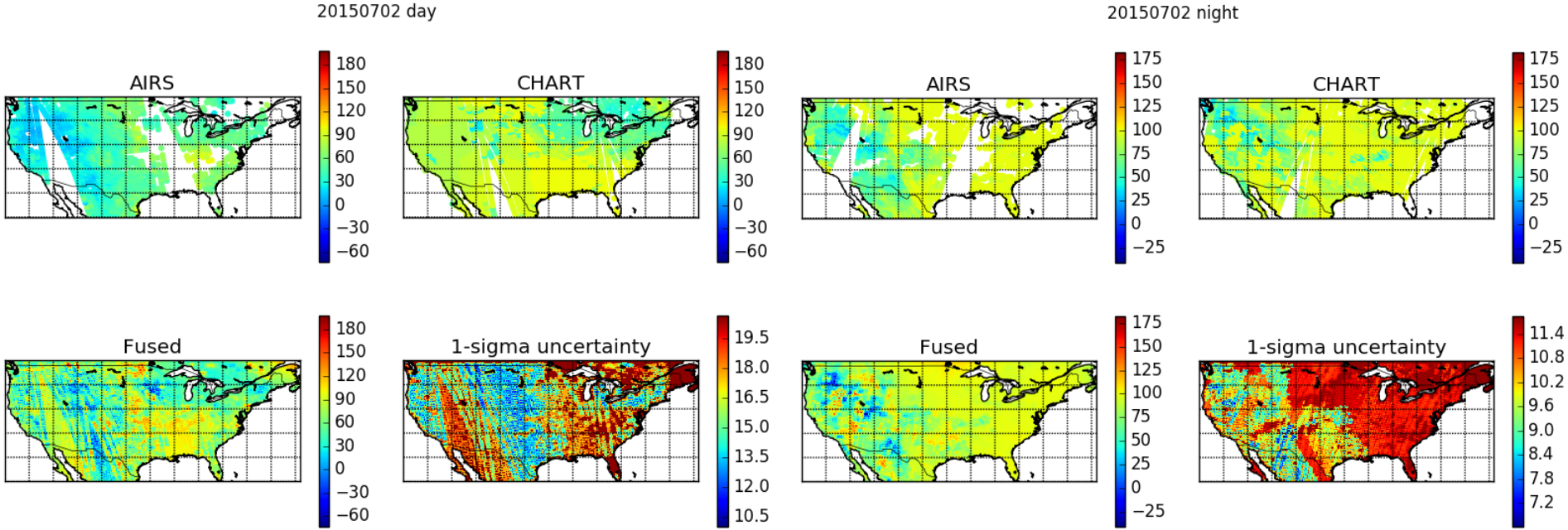




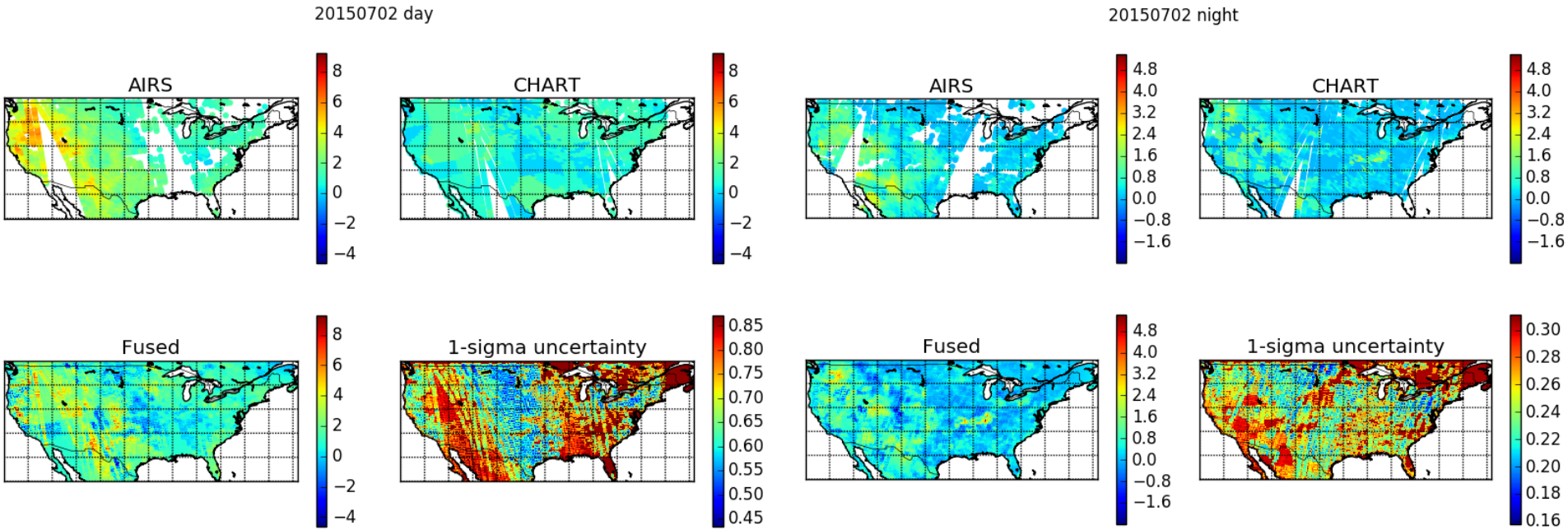
Preliminary sample of full CONUS: NST



Preliminary sample of full CONUS: NSRH

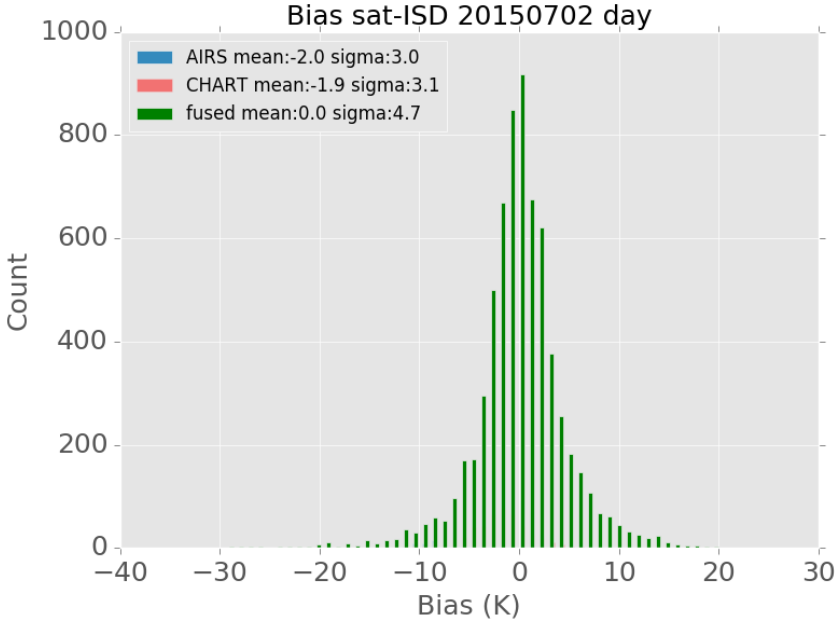


Preliminary sample of full CONUS: VPD

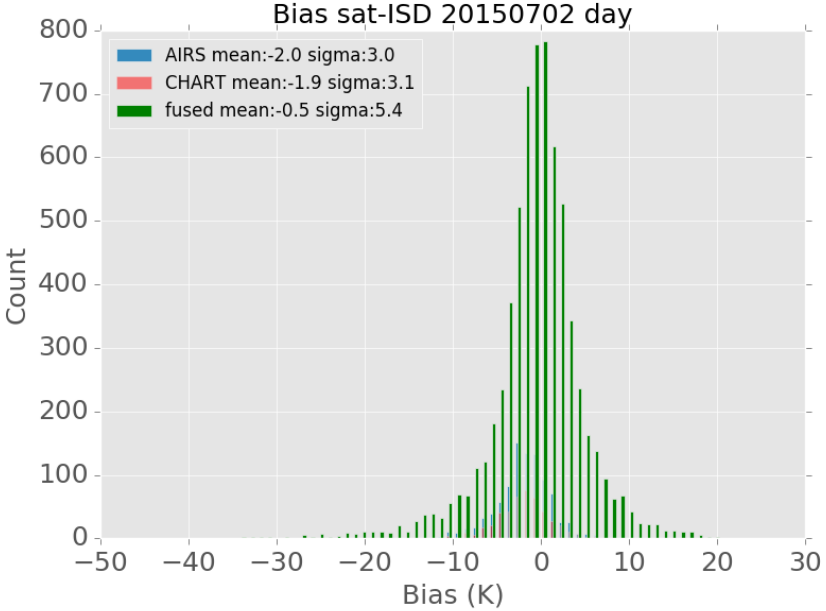


Initial attempt to elevation-adjust NST matchups failed

regular



ISD shifted along dry adiabat according to elev. differences with sat. matchups



Summary

Eastern CONUS AIRS+CHART NST, NSRH, VPD delivered.

- Significantly reduces bias; variance reduced but room for improvement
- Produces estimates even over data gaps
- Produces uncertainty estimates.
- Fusion calculation is fast.

Full CONUS products are under development.

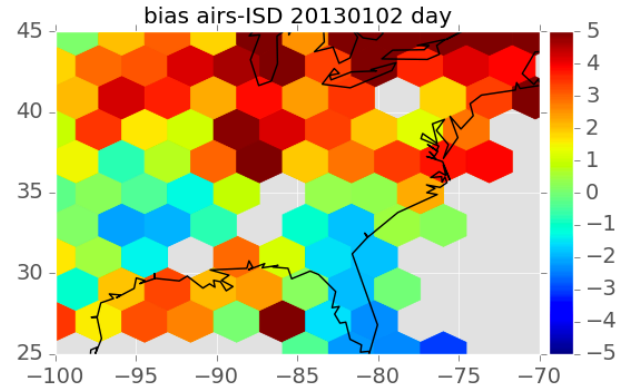
- Applications (drought, fire, agriculture, vector-borne disease)
- Explore potential for kriging or UQ to improve bias and variance inputs

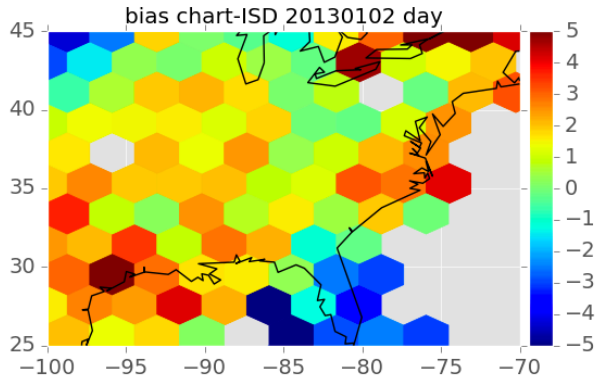
Switch to CLIMCAPS! And make this a real product.

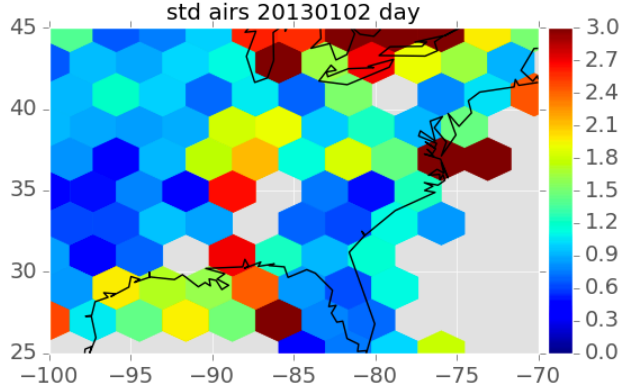
extra slides

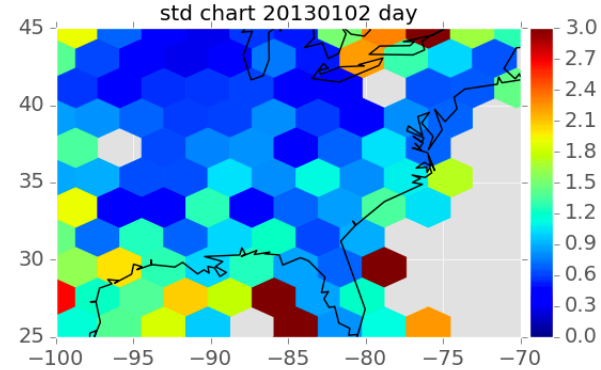
Sample biases and variances

3-day averages (day / night separate), 2° hex grid

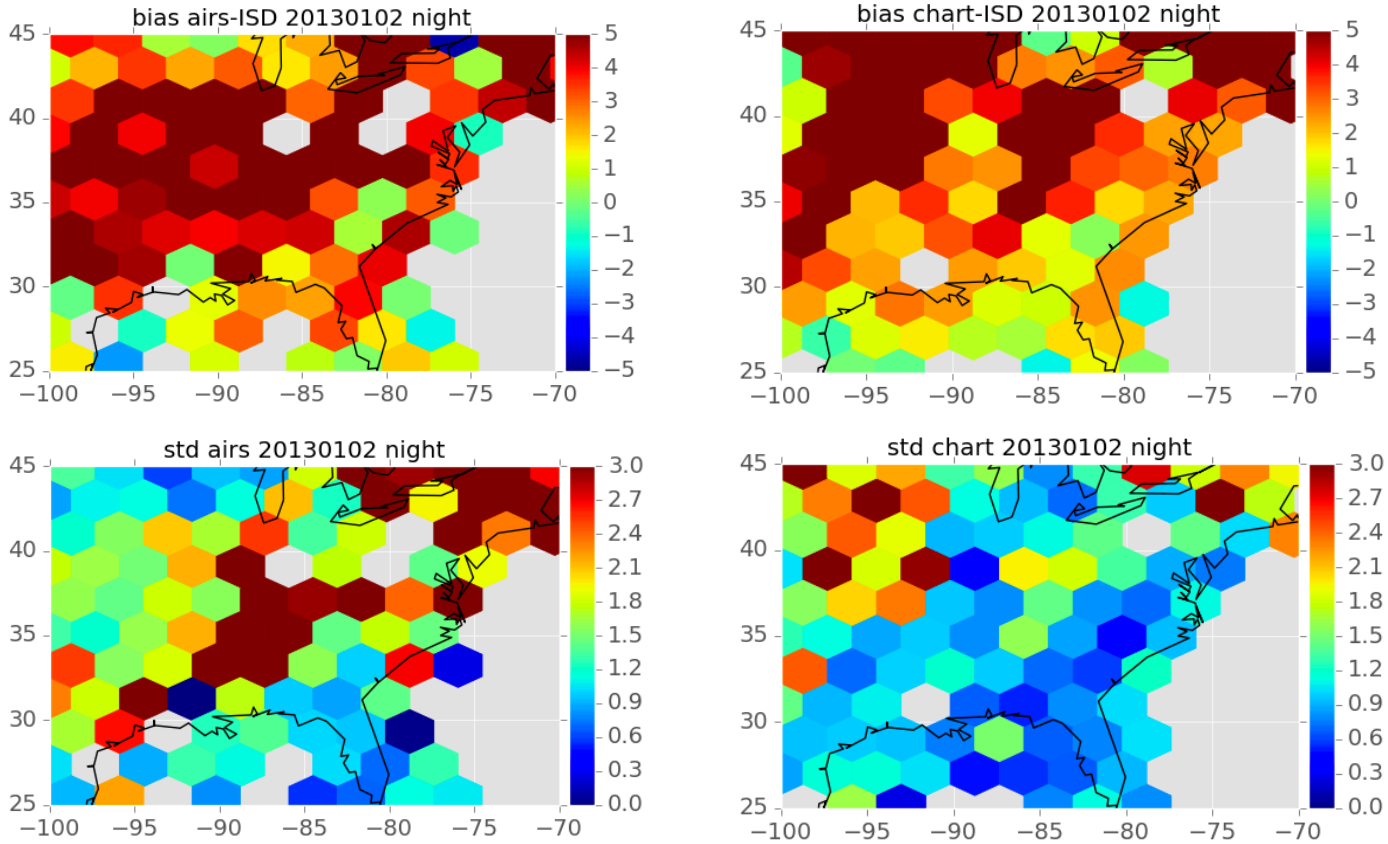








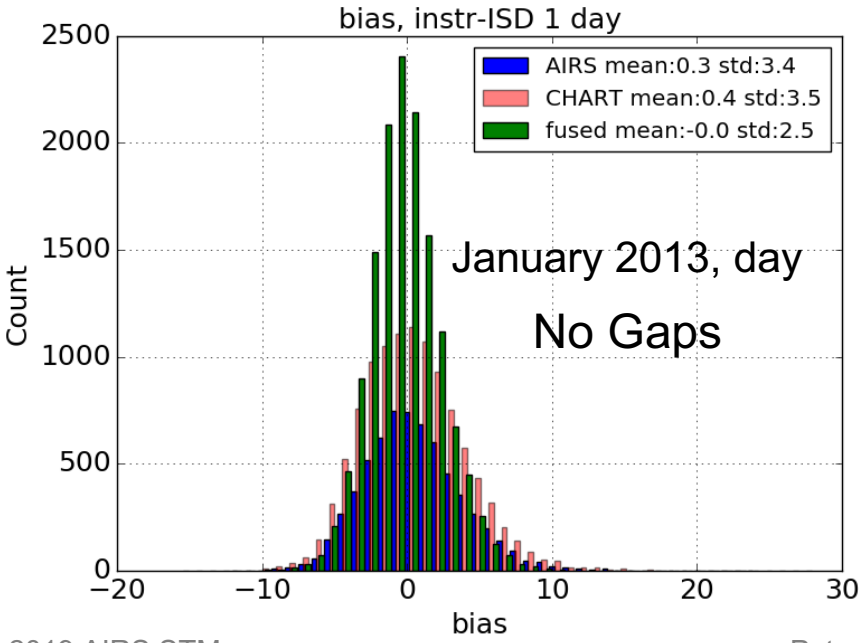
Sample biases and variances

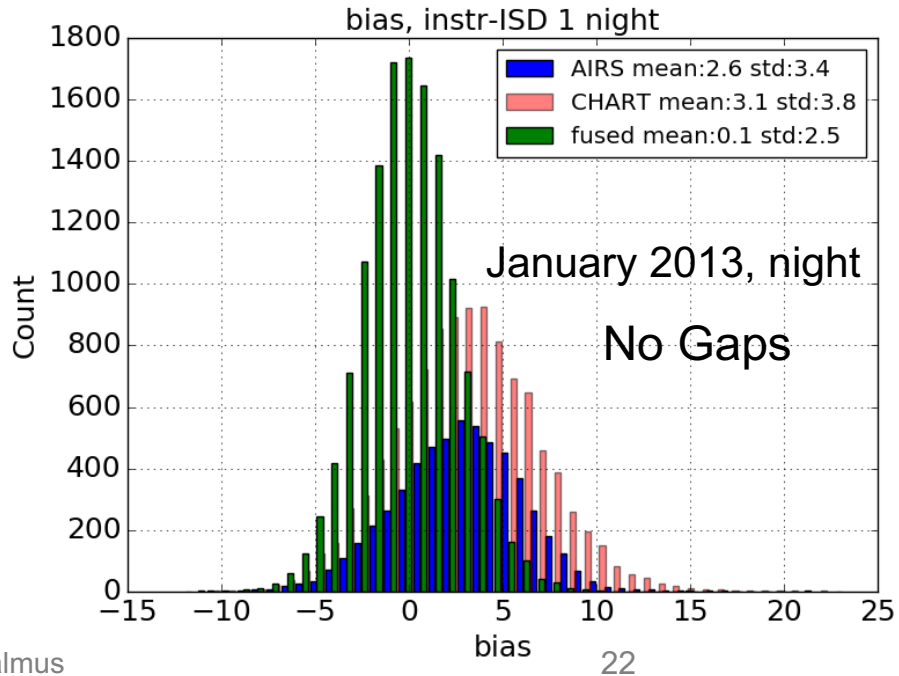


Validation

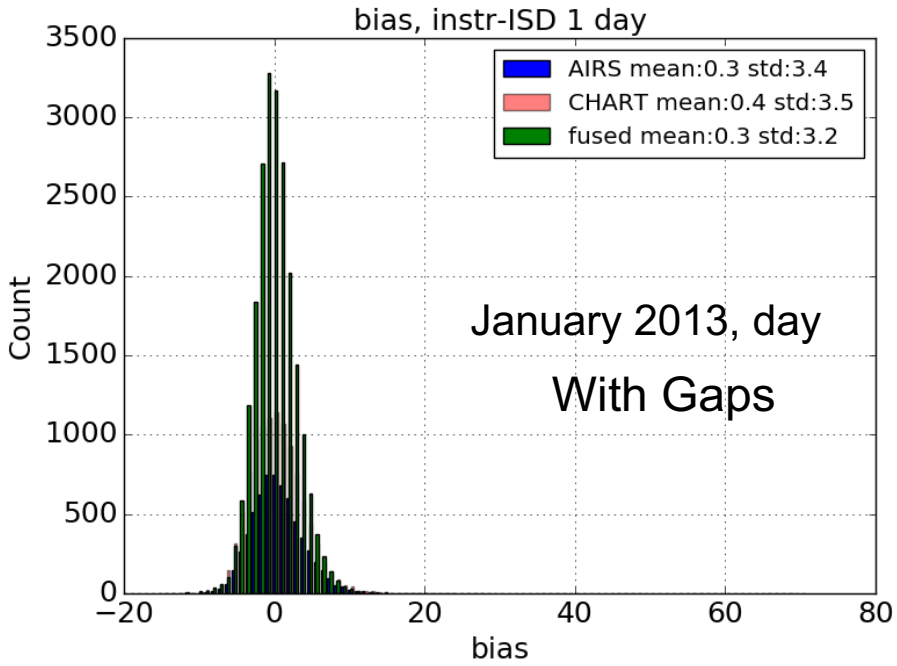
Perform matchups: gridded data fusion output, with ISD

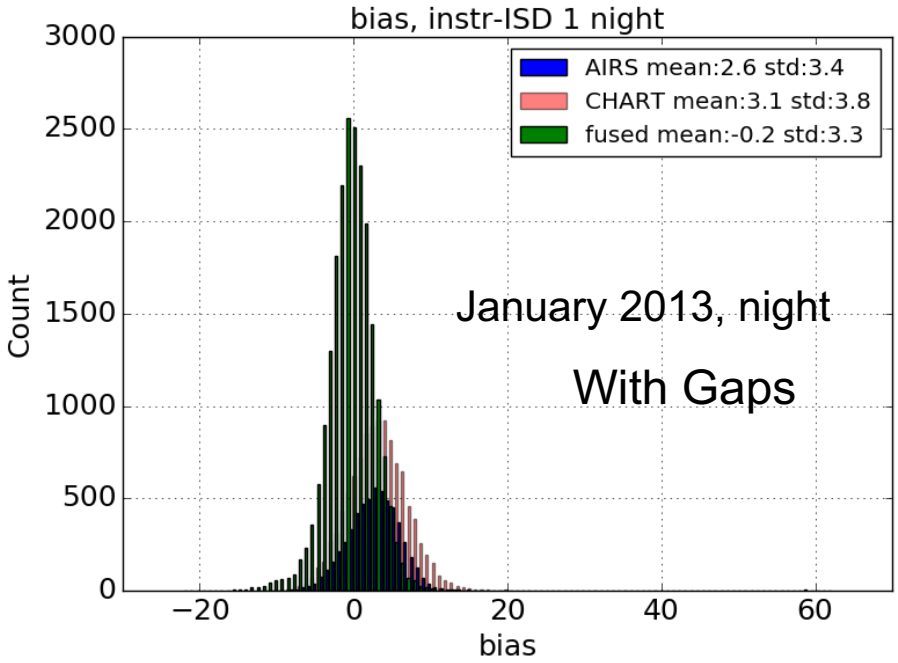
- some fused output matchups also to either AIRS or CrIS datum
- other fused output falls in “gap” areas



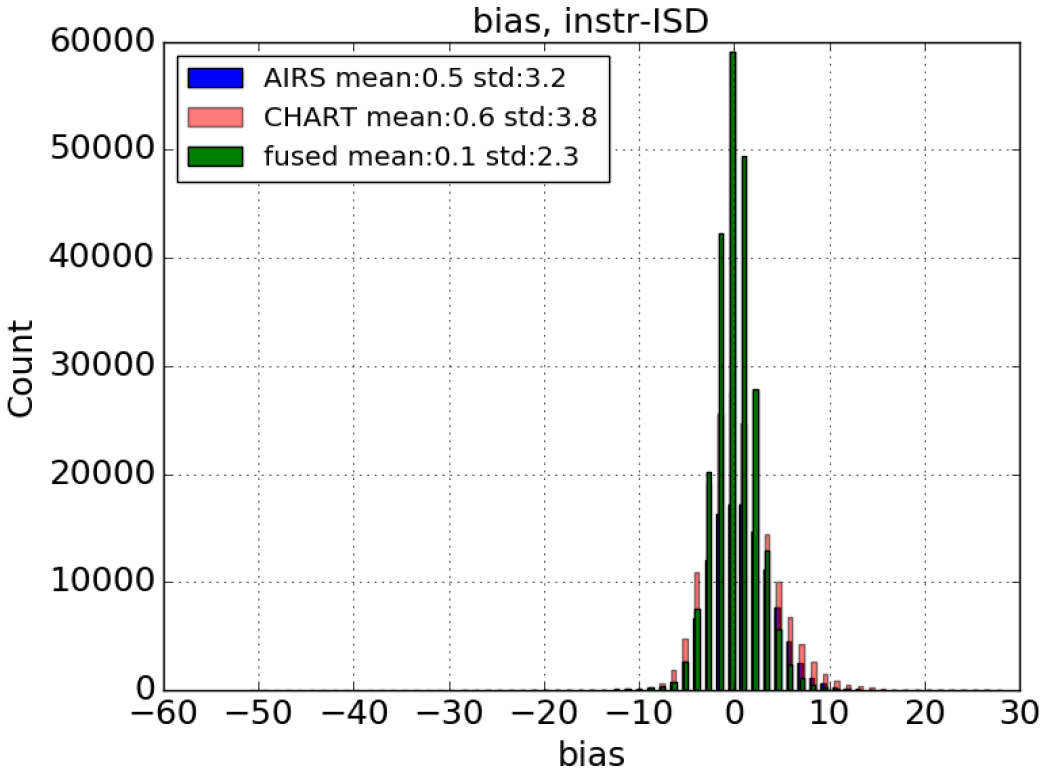


Validation


bias, instr-ISD 1 day
AIRS mean:0.3 std:3.4
CHART mean:0.4 std:3.5
fused mean:0.3 std:3.2
January 2013, day
With Gaps
Count
bias
3500
3000
2500
2000
1500
1000
500
0
−20
0
20
40
60
80

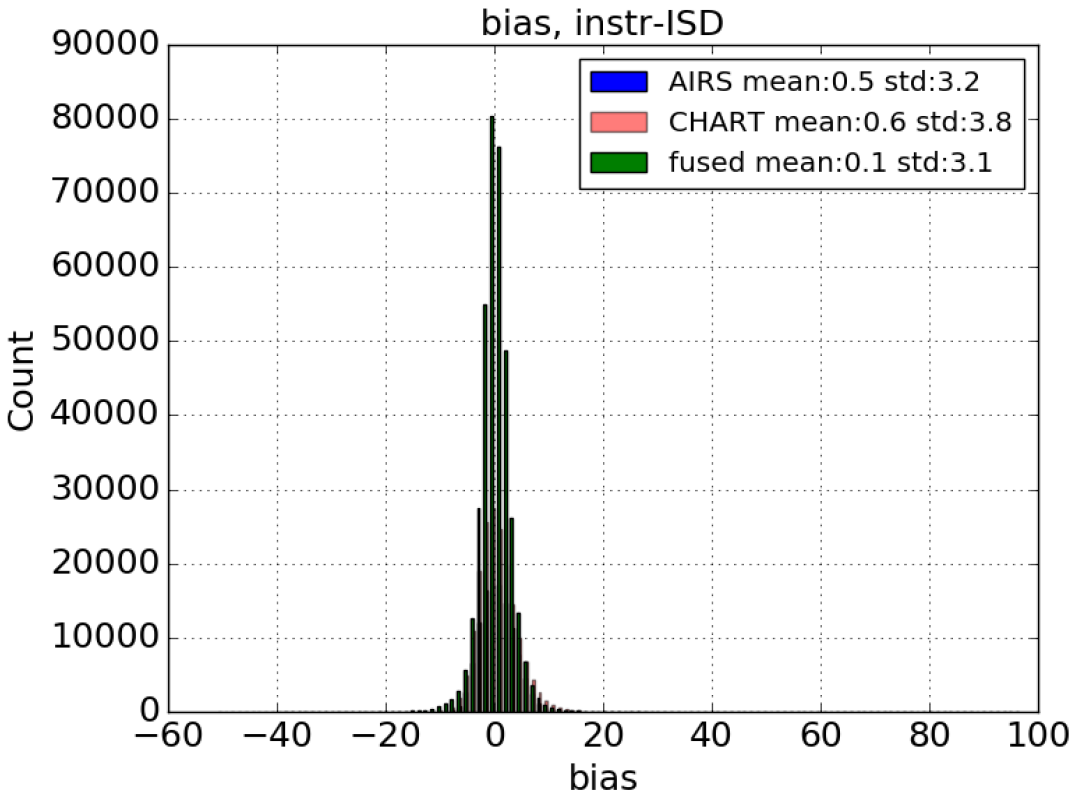

bias, instr-ISD 1 night
AIRS mean:2.6 std:3.4
CHART mean:3.1 std:3.8
fused mean:-0.2 std:3.3
January 2013, night
With Gaps
Count
bias
3000
2500
2000
1500
1000
500
0
−20
0
20
40
60

Bias of all matchups (over the 8 months of data)



excludes input data gaps

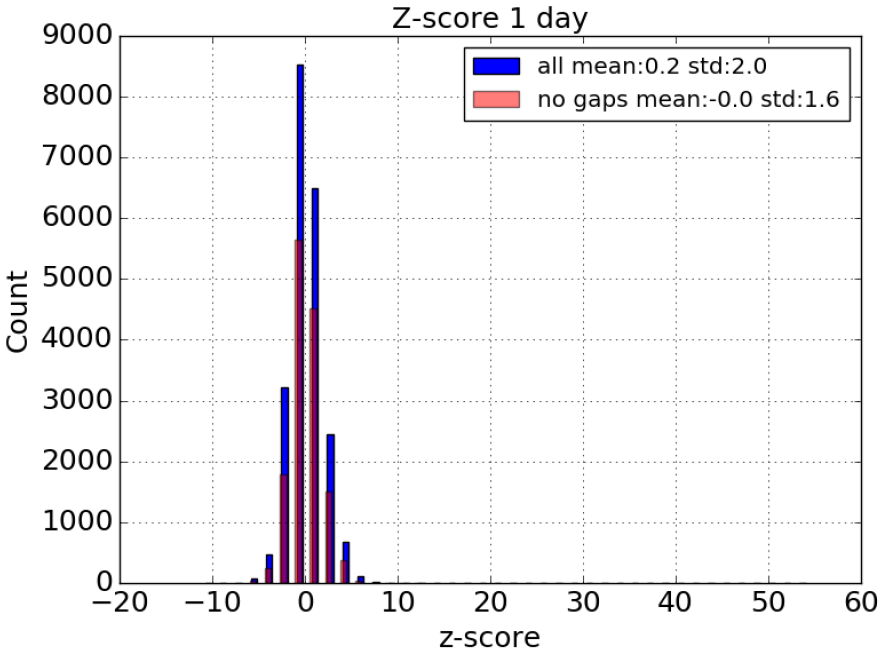
bias of all matchups, including gaps



Validation of uncertainty

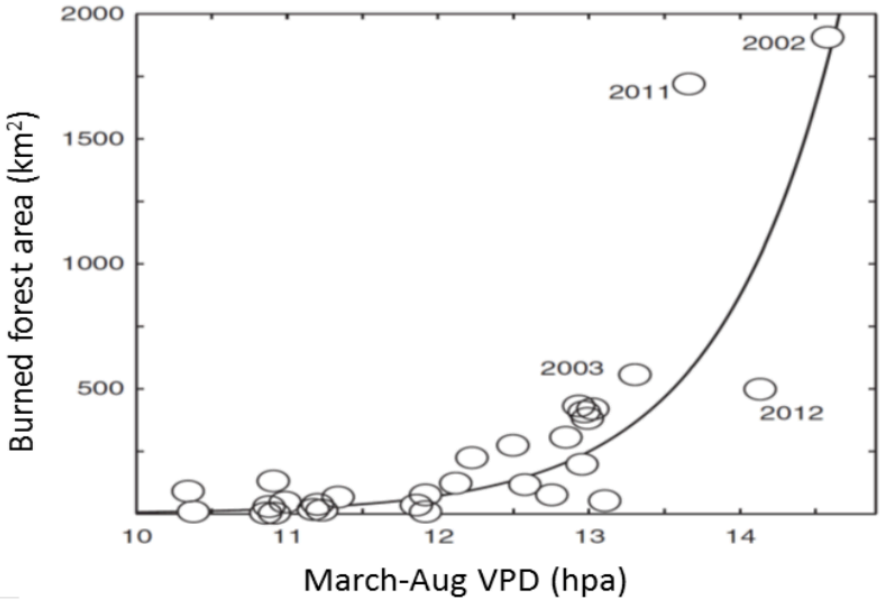
Fusion is capturing ~60% of the uncertainty (1.6^{-1})

Also, there are a few really bad estimates that must be chased down!



January 2013, day

Vapor pressure deficit and fire



Annual burned forest area *v.* mean March–August VPD in SW forest area during 1894–2013. The figure is from Williams et al. (2014a)

Fixed Rank Kriging (FRK) vs. SSDF

Method outline

1. Compute the bias and measurement error of each dataset relative to validation data,
2. Compute the estimates as linear combinations of all the observed data points, subject to unbiasedness and minimum-error constraints,
3. Parametrize the underlying covariance function using the spatial random effects model,
4. Estimate the spatial dependence parameter using the Expectation-Maximization algorithm,
5. Produce the estimates (best linear unbiased predictors) and their associated uncertainty.

Model assumptions

We assume the data from instrument i are generated according to the following model:

$$\begin{aligned} \mathbf{Z}_i &= (Z_i(\mathbf{s}_{i1}), Z_i(\mathbf{s}_{i2}), \ldots, Z_i(\mathbf{s}_{iN_i}))', \\ Z_i(\mathbf{s}_{ij}) &= Y(\mathbf{s}_{ij}) + \epsilon_i(\mathbf{s}_{ij}); \\ Y(\mathbf{s}_{ij}) &= \mathbf{S}(\mathbf{s}_{ij})\boldsymbol{\eta}; \end{aligned}$$

where

- $\mathbf{Z}_i$ is the vector of response variable from dataset i,
- $Y(\cdot)$ is the true process,
- $\epsilon_i(B_{ij})$ is the error process.

Model assumptions

We can rewrite the data models in vector forms as

$$\begin{aligned} \mathbf{Z}_1 &= \mathbf{Y}_1 + \epsilon_1 \\ \mathbf{Z}_2 &= \mathbf{Y}_2 + \epsilon_2, \end{aligned}$$

Or, equivalently, we can simplify the problem by stacking them as follows,

$$\begin{pmatrix} \mathbf{Z}_1 \\ \mathbf{Z}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix},$$

or equivalently,

$$\mathbf{Z} = \mathbf{Y} + \epsilon.$$

Prediction equation

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

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

where **a** is a N-dimensional vector of *kriging coefficients* at location **s**.

Minimization conditions

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

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

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

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

where the spatial dependence between the data points are captured within our parameterization of $\mathrm{cov}(\cdot, \cdot)$.