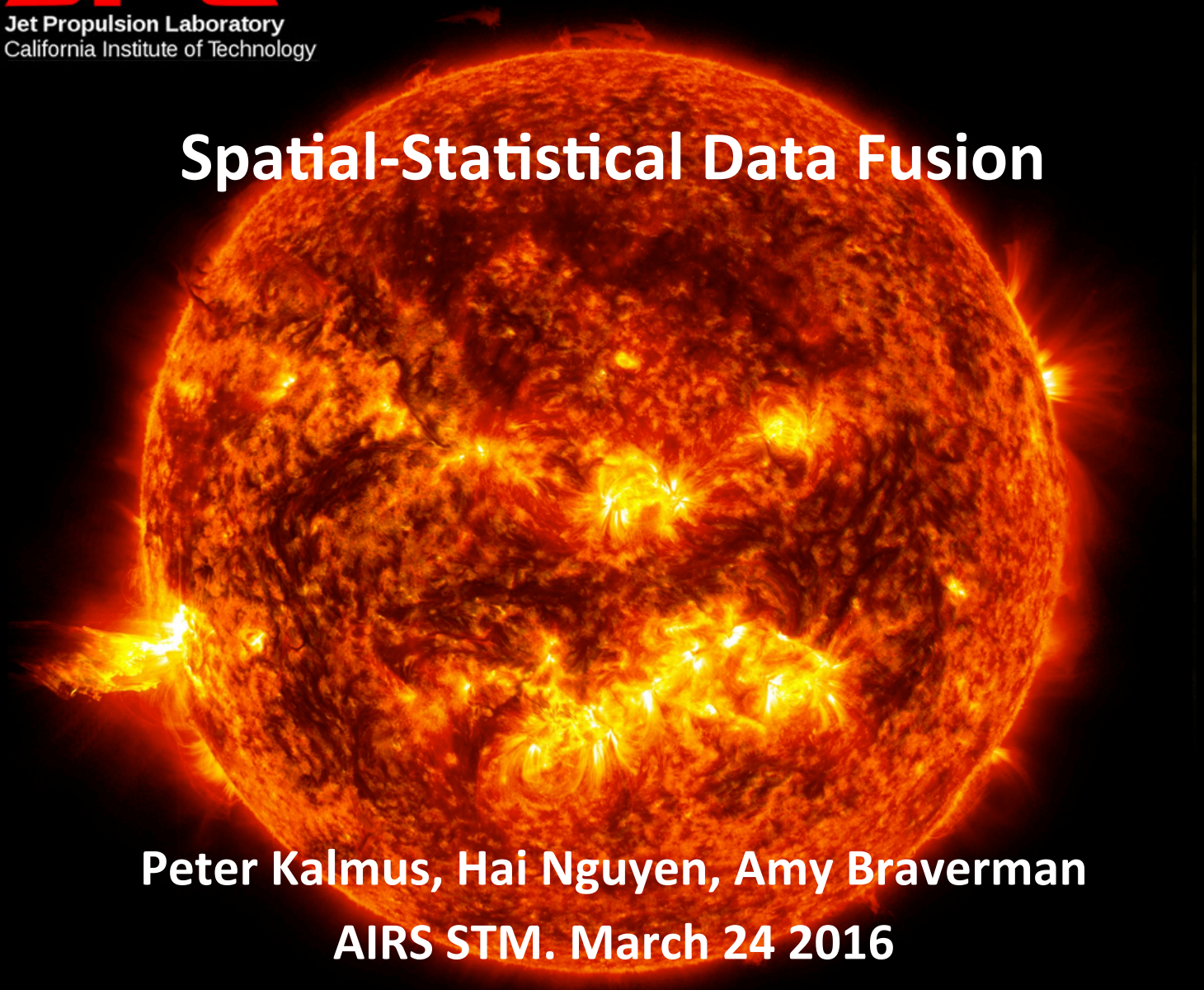
Jet Propulsion Laboratory
California Institute of Technology

Spatial-Statistical Data Fusion



Peter Kalmus, Hai Nguyen, Amy Braverman

AIRS STM. March 24 2016

Motivation

Suppose there are *K* global data sets estimating the same field.

- Different spatio-temporal coverages, strengths, weaknesses.
- Example: satellite temperature soundings.

Users are forced to choose. By choosing a single data set:

- Users throw away information.
- Users forced to resolve spatio-temporal disparities between fields.
- Users forced to accept unnecessarily large amounts of missing data.
- Users aren't necessarily well-positioned to choose.

Why not fuse the *K* data sets into one optimal estimate?

- And also provide robust uncertainty estimate to users?

Spatial-Statistical Data Fusion

Data fusion produces optimal estimates of the true field (minimum uncertainty and unbiased) on a grid of our choosing.

It also yields uncertainties.

Start with 2+ noisy, incomplete data sets representing some physical field (e.g. T or q).

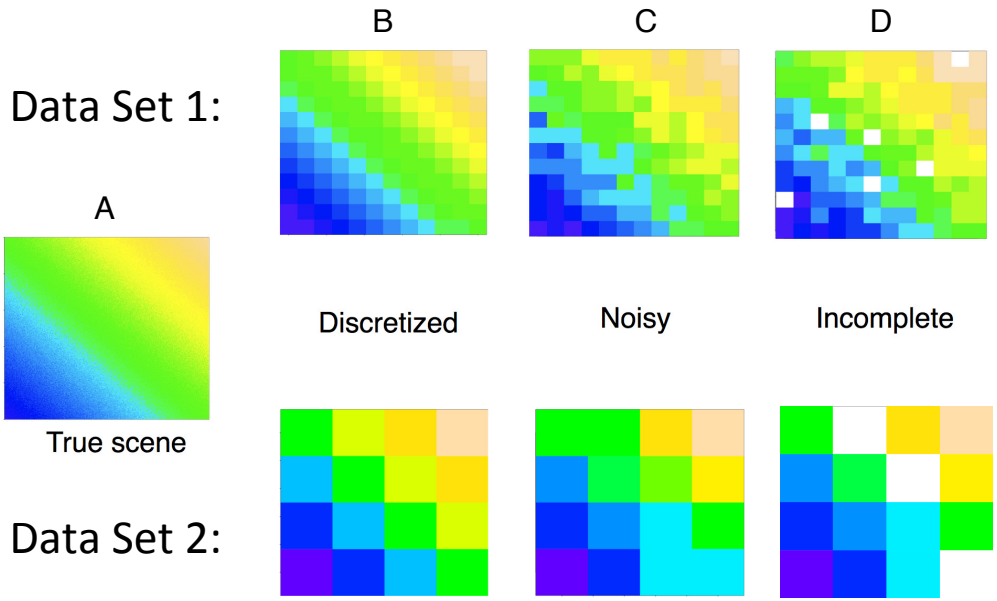
Each data set must come with robust bias and variance estimates. Spatial footprints can be anything.

We model the underlying (true) physical field as a spatial Gaussian process.

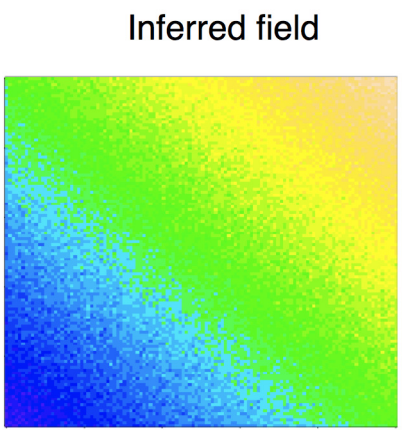
We model the observations as functions of these random variables (e.g., spatial aggregates plus measurement error).

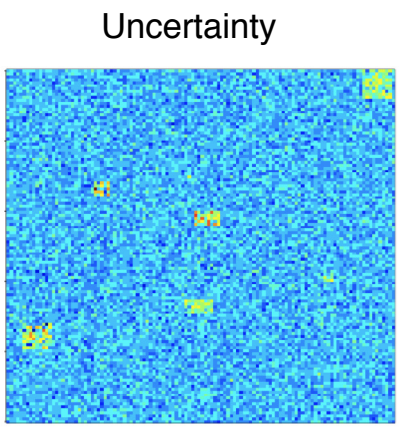
From the observations, we estimate the spatial covariance of the underlying field and infer the true field, accounting for spatial dependence and error of each input data set.

see Nguyen, Cressie, and Braverman 2012



Fusion of column D yields:





Evolution of fusion methods

Two major challenges:

- different spatial footprints of input data sets

  Bayesian melding (e.g. Fuentes & Raftery 2005)

- massive data sets of total size N

  Bayesian melding has computational complexity $O(N^3)$

  Fixed-Rank Kriging (FRK), single dataset spatial interpolation $O(N)$

  (Cressie & Johannesson 2008)

SSDF generalizes FRK to multiple data sets

Different areal footprints, measurement errors

Performance comparable to Bayesian melding

Much faster!

SSDF uncertainty estimate

Assume that that we have the following data model

$$Z = Y + \epsilon$$

where Z, Y, and ϵ are N-dimensional vectors of the observations, the true process, and the measurement errors, respectively.
The kriging estimators at a location s are assumed to the linear combination of the data Z that minimizes the root mean squared error as follows

$$(\hat{Y}(s)) = aZ \text{ where } a \text{ miminizes } (aZ - Y(s))^2.$$

The coefficients a could be found by differentiating rmse expression with respective to a, setting it to zero, and solving for a. Once found, the error could be computed as,

$$\begin{aligned} \mathrm{Var}(aZ - Y(s)) &= \mathrm{Var}(aZ) + \mathrm{Var}(Y(s)) - 2\mathrm{Cov}(aZ, Y(s)) \\ &= a'\mathrm{Var}(Z)a + \mathrm{Var}(Y(s)) - 2a\mathrm{Cov}(Z, Y(s)). \end{aligned}$$

Data Fusion Proof-of-Concept

Fusion of 3 data sets: AIRS+ECMWF+MERRA2

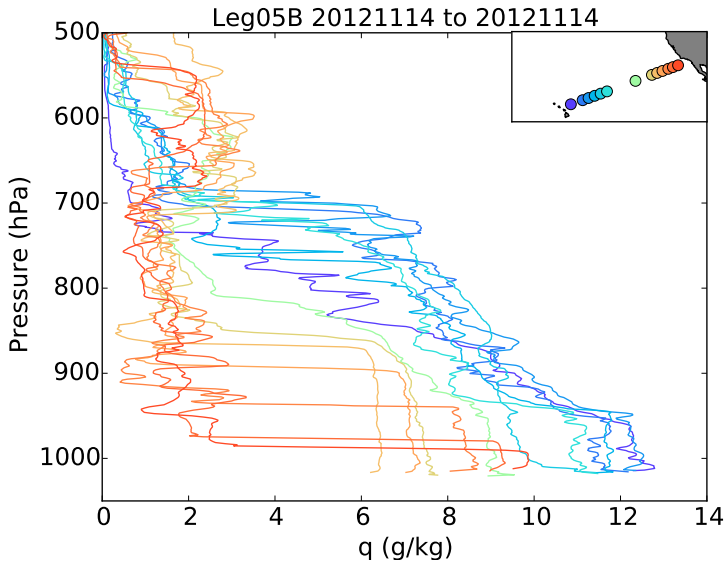
- q and T on 14 vertical levels
- levels fused separately
- T, q fused separately (for now...)
- **July 2013**, 10N—40N, 130W—160W

MAGIC radiosondes:

- to estimate bias, variance of input data sets (over entire MAGIC campaign)
- to evaluate results (July only)
- matchups: 200km, ±6 hours

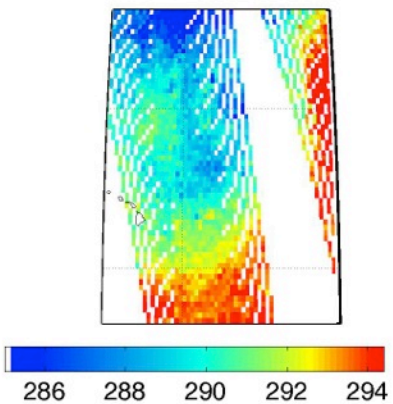
Fused data output:

- 6-hourly, 0.5 degree grid (ECMWF grid)
- 14 levels: 1000, 925, 850, 700, 600, 500, 400, 300, 250, 200, 150, 100, 70, 50 hPa

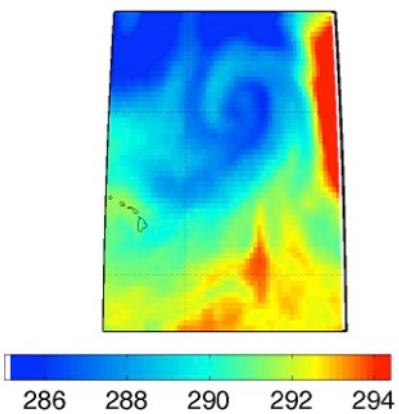


Input and fused maps of q

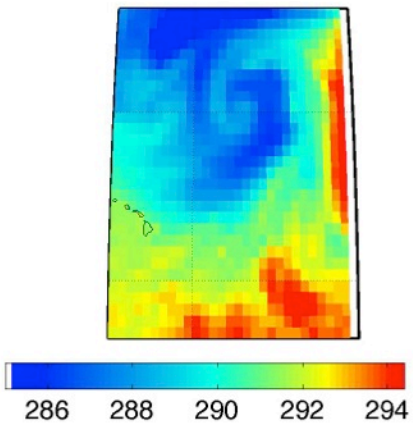
AIRS



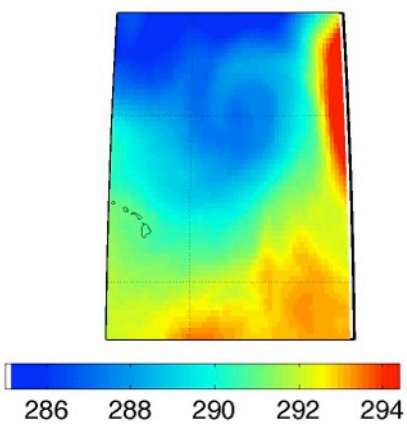
ECMWF ERA-Interim



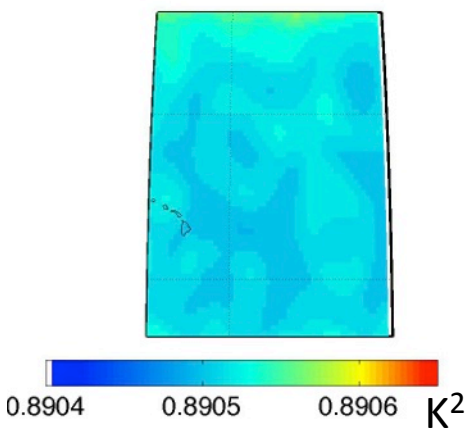
MERRA2



Fused



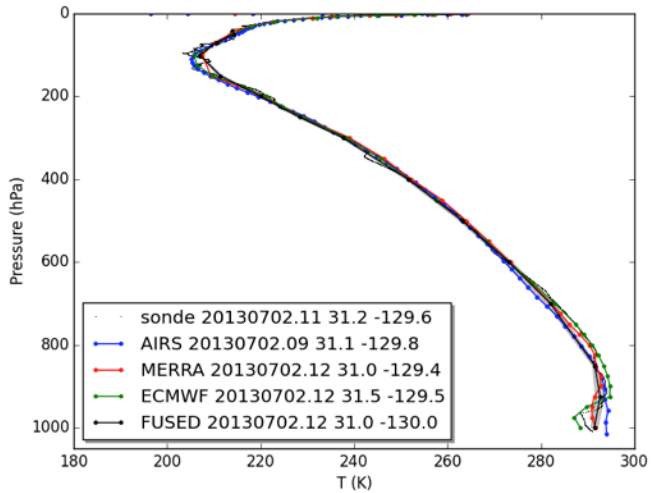
Variance of fused

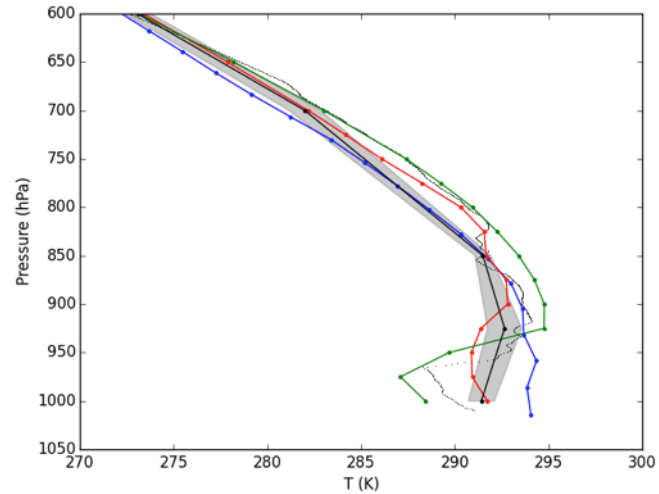


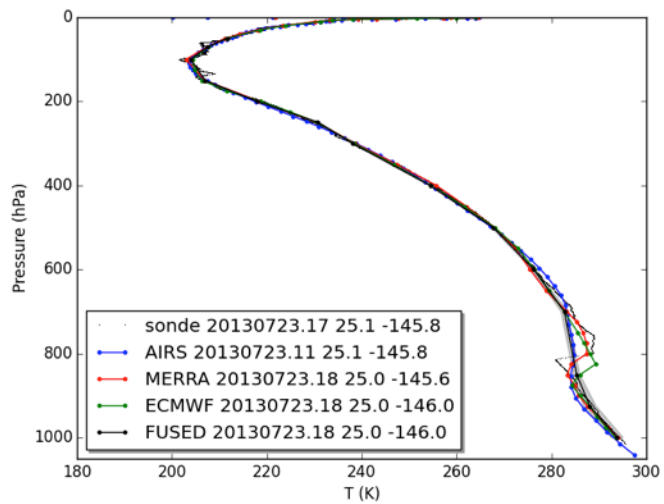
Example: 925 hPa temperature (K)
July 2 2013, 1800 UTC

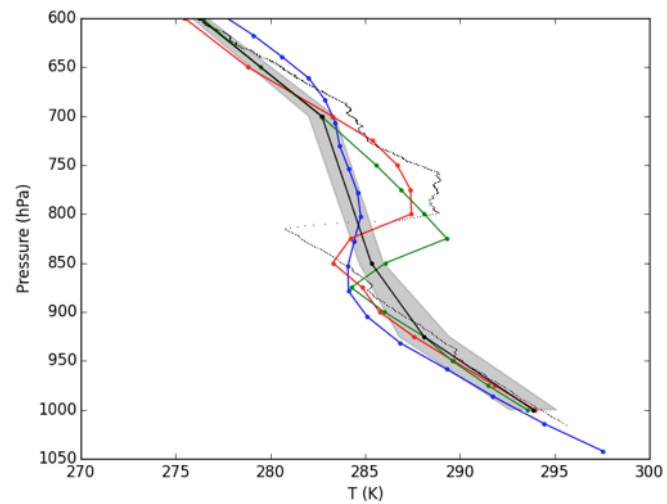
Fusing 3 sparse data sets would further highlight the usefulness of fusion.

Example T profiles

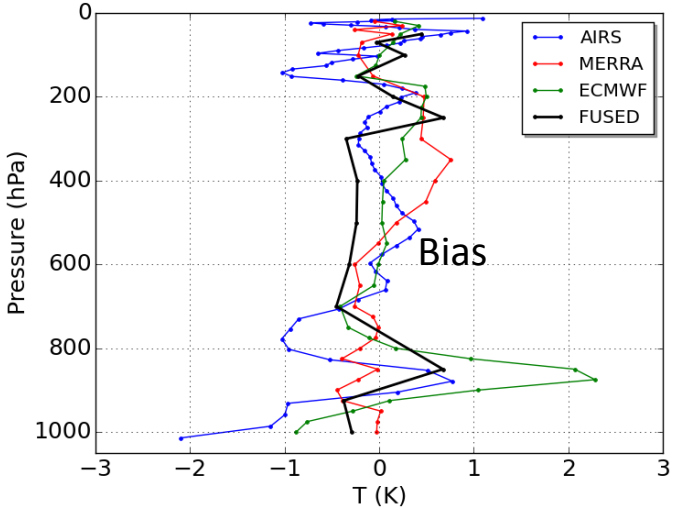


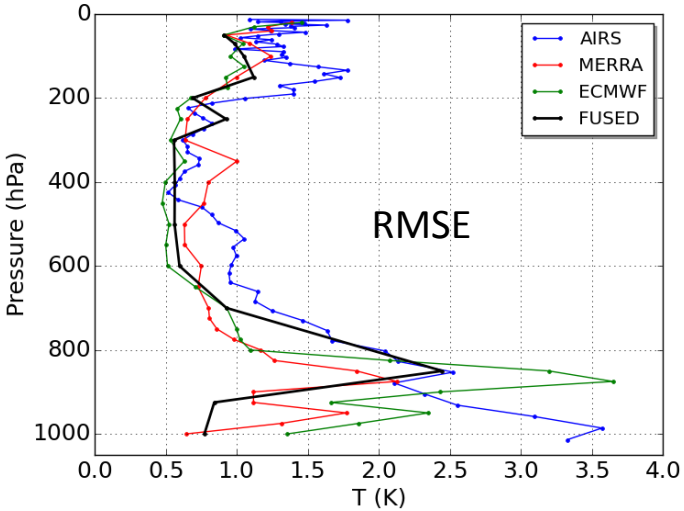




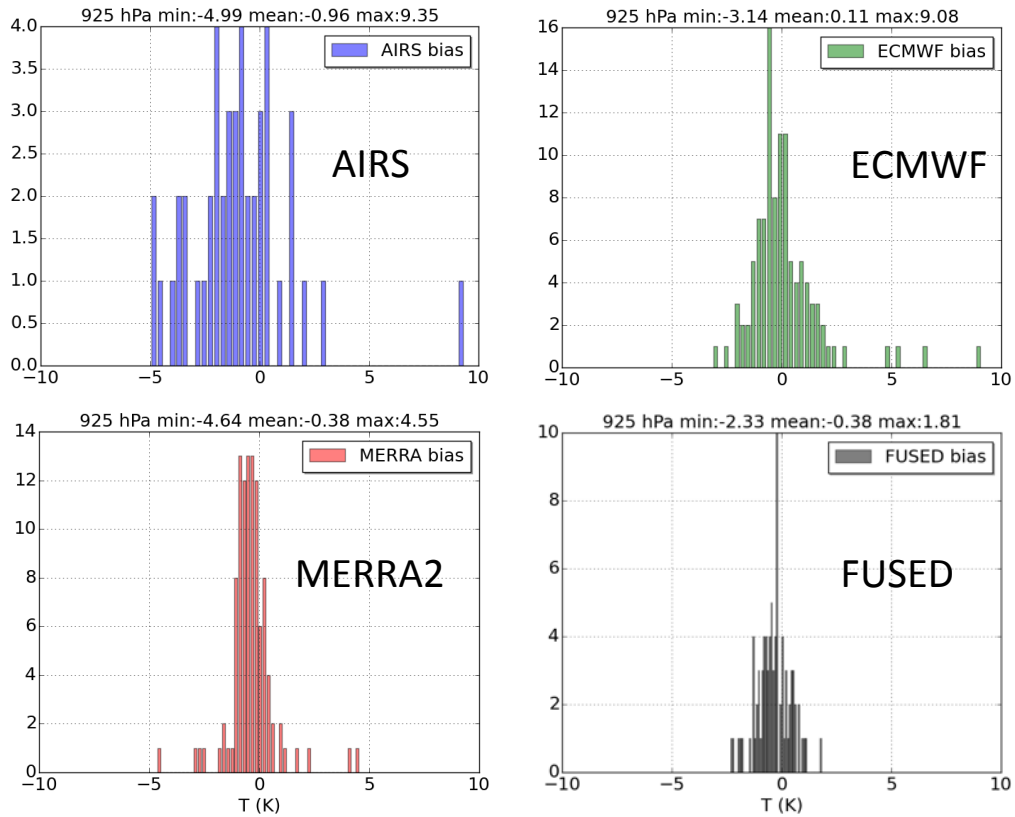


Data fusion: early results: T





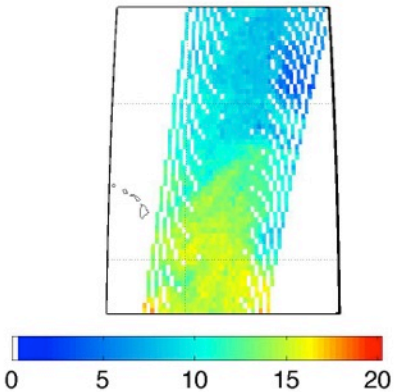
Above: Profiles of T mean bias, RMSE



Above: 925 hPa T bias histograms

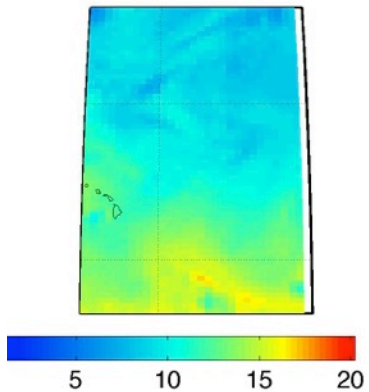
Input and fused maps of q

AIRS

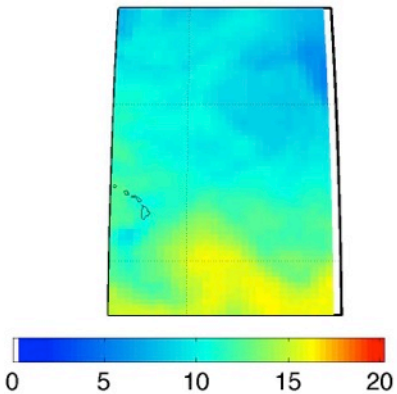


g/kg

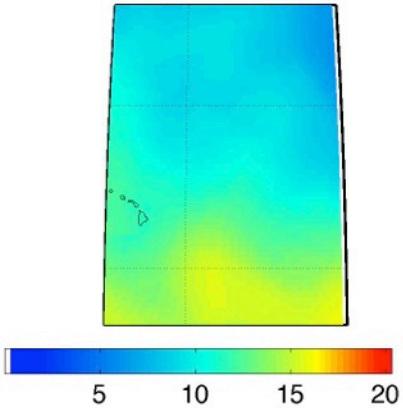
ECMWF ERA-Interim



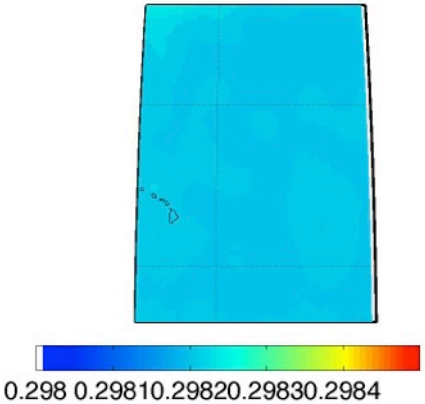
MERRA2



Fused



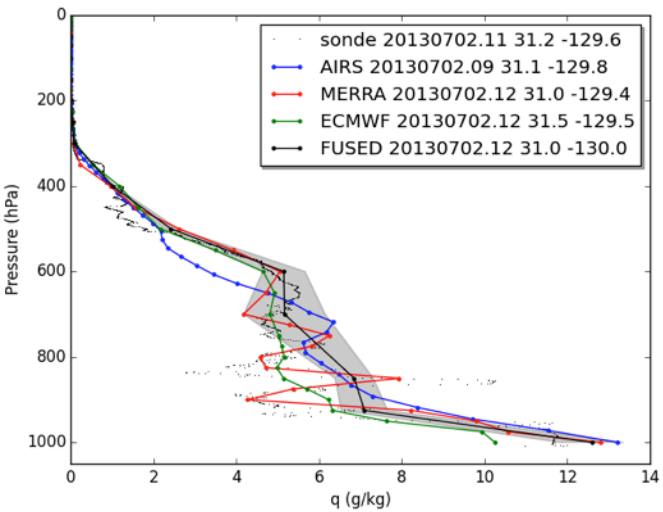
Variance of fused

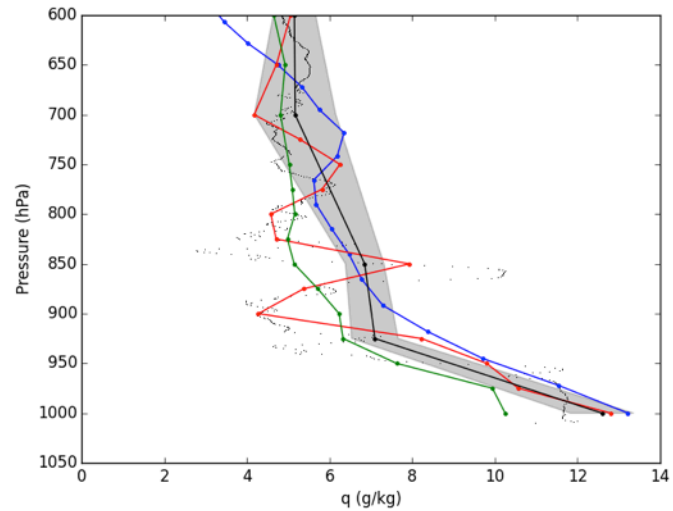


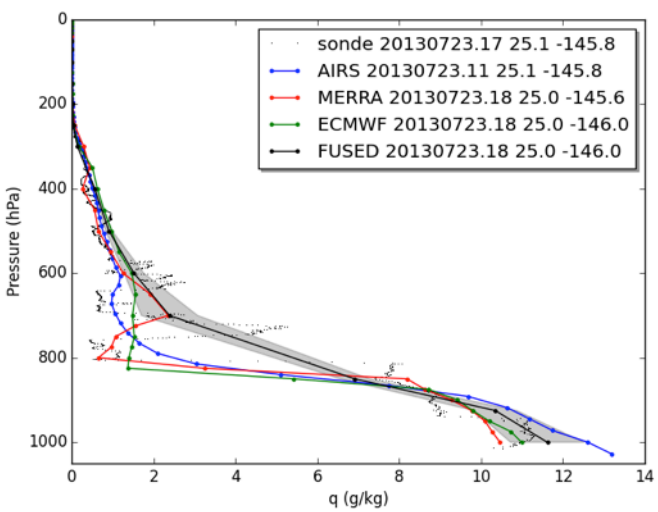
$(g/kg)^2$

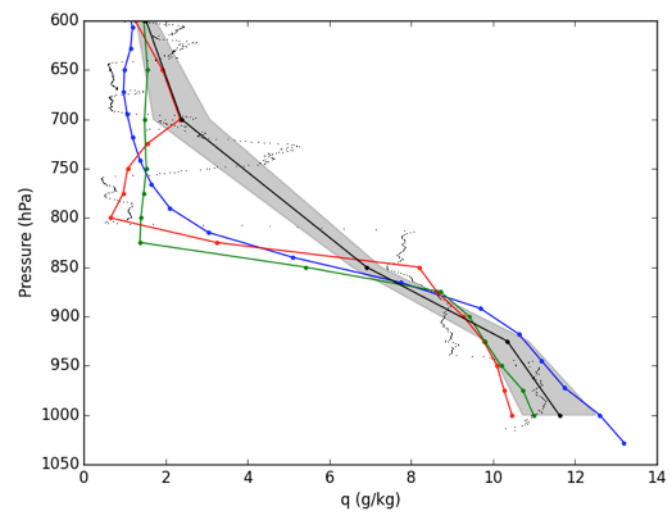
Example: 925 hPa q (g/kg)
July 19 2013, 0600 UTC

Example q profiles

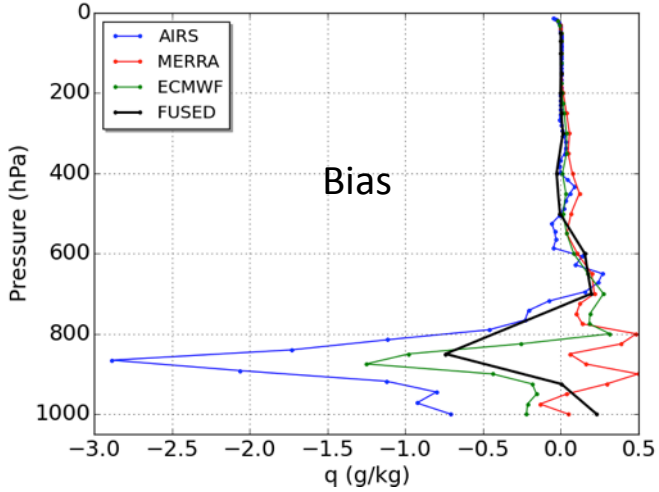


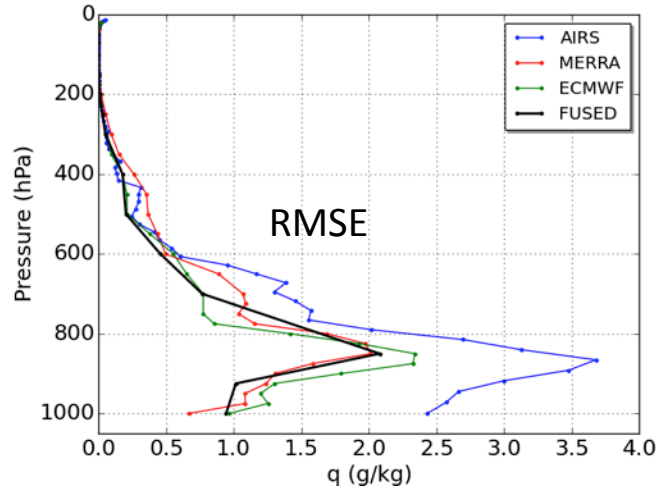




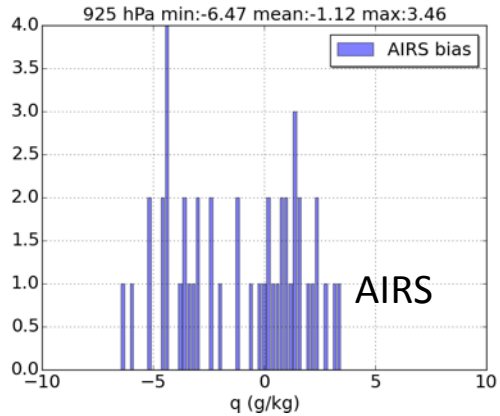


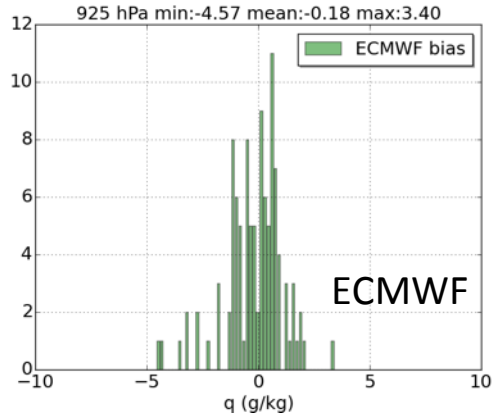
Data fusion: early results: q

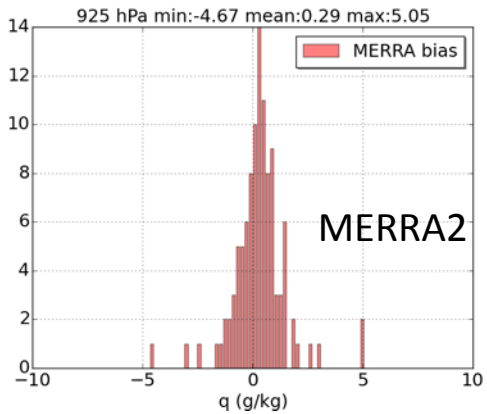


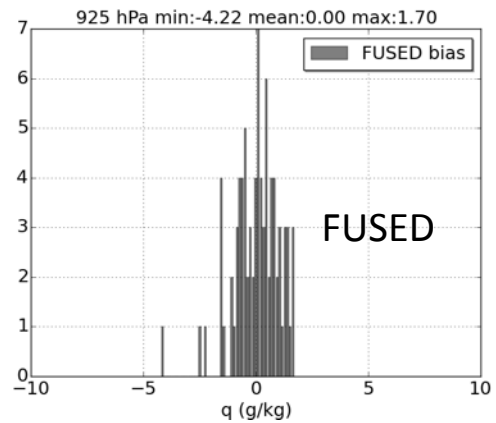


Above: Profiles of q mean bias, RMSE









Above: 925 hPa q bias histograms

Uncertainty estimates & synthetic data

Fused product is only as good as input uncertainty estimates.

We expect the biggest barrier will be obtaining robust global bias and variance estimates for the input data sets.

One possibility: characterize uncertainty in retrievals with simulations using synthetic “truth” data. Follow protocol similar to that used for OCO2 uncertainty quantification.

Conclusions

Fused data sets would be valuable to the community.
SSDF is a promising method for fusing data sets.

Next steps:

- simultaneous T, q fusion (AIRS+ECMWF+MERRA2)
  - quantify improvement
- AIRS+CrIS+IASI T, q fusion in "MAGIC sandbox" (NE Pacific 30°x30°)
  - Assess usefulness
- optimize calculation of uncertainty estimate
  - currently: takes 50—90% of CPU time
  - currently: 1 yr of global AIRS+ECMWF+MERRA would take 60 hours on 5000 CPUs
  - we may also face memory challenges
- global estimates of AIRS T, q bias and variance using synthetic data
- global proof-of-concept for input data sets
- global AIRS+CrIS+IASI T,q fusion

SDO/AIA 304 2015-02-10 17:54:56 UT