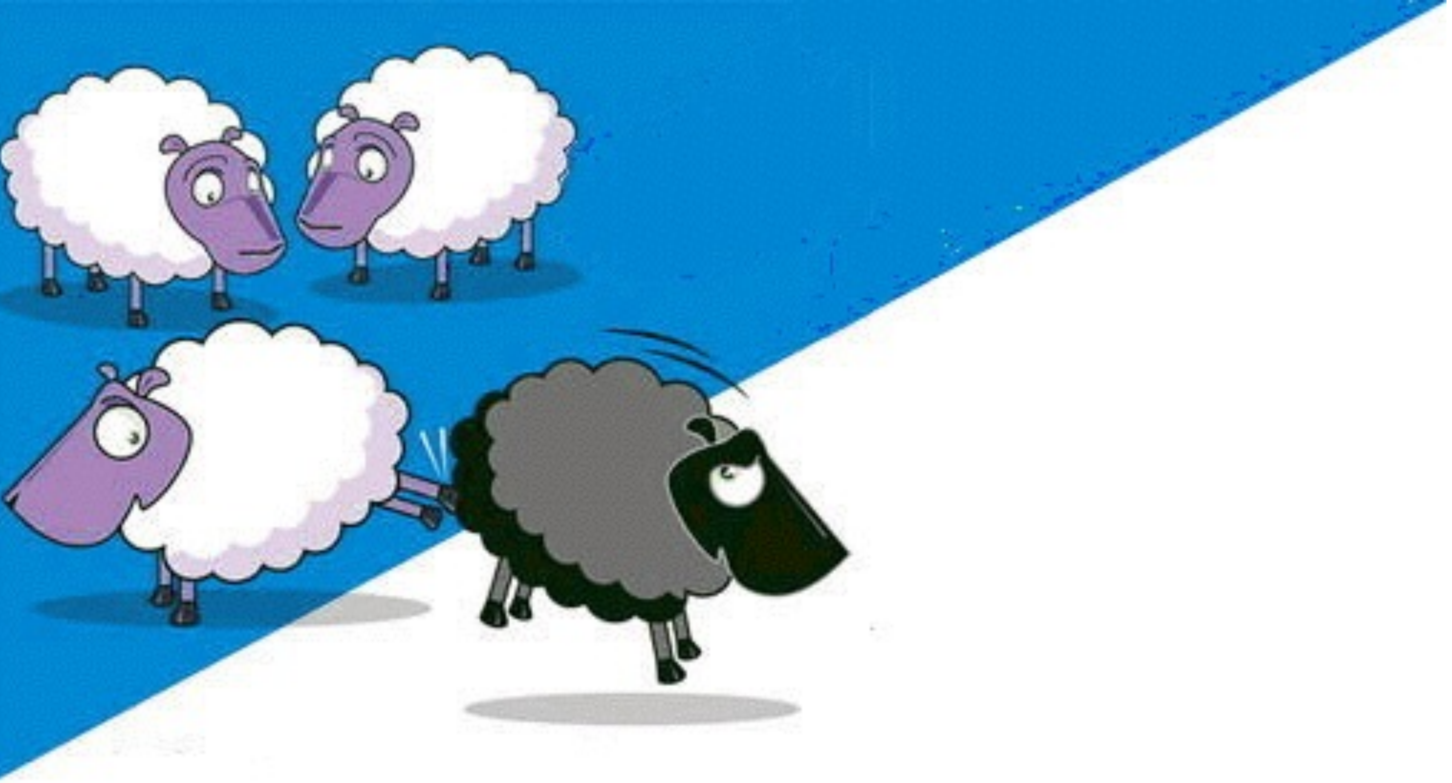


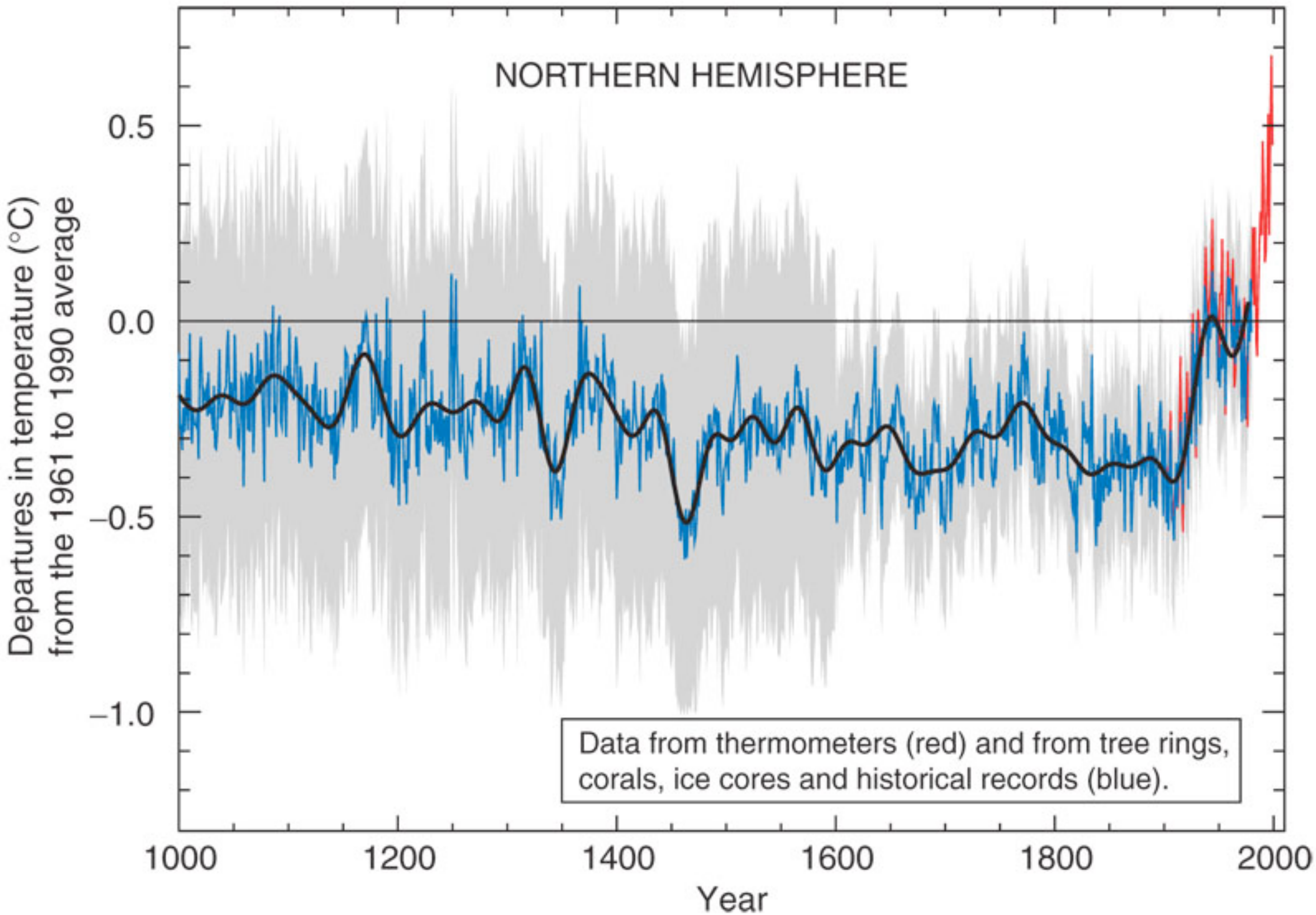
All sunspots are equal, but some are more equal than others

T. Dudok de Wit
LPC2E, University of Orléans, France

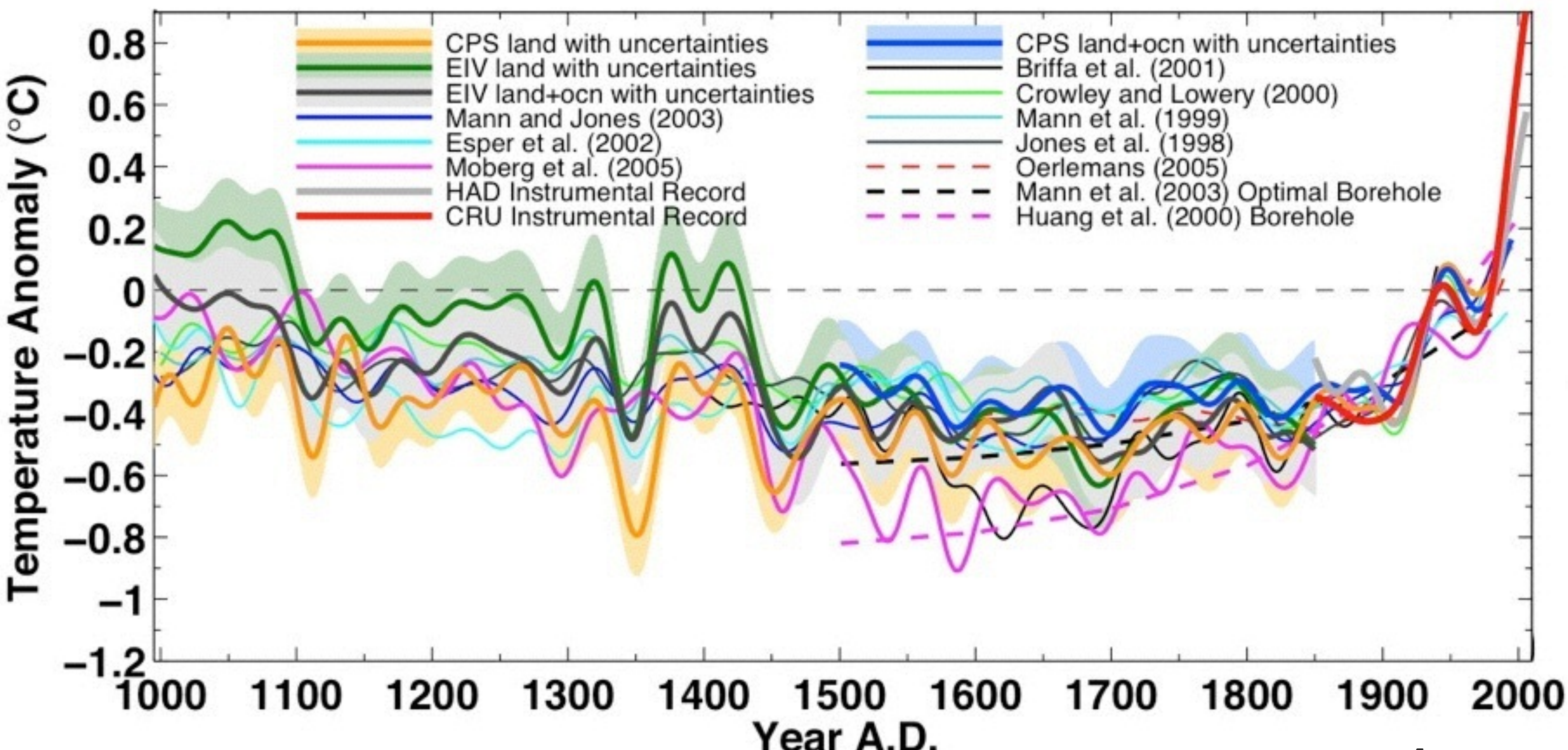
With thanks to Rodney Howe, Laure Lefèvre, Frédéric Clette, Marty Snow, Greg Kopp, Claus Fröhlich, and many more

The hockey-stick curve

IPCC 3rd AR, 2001



and what it is made of



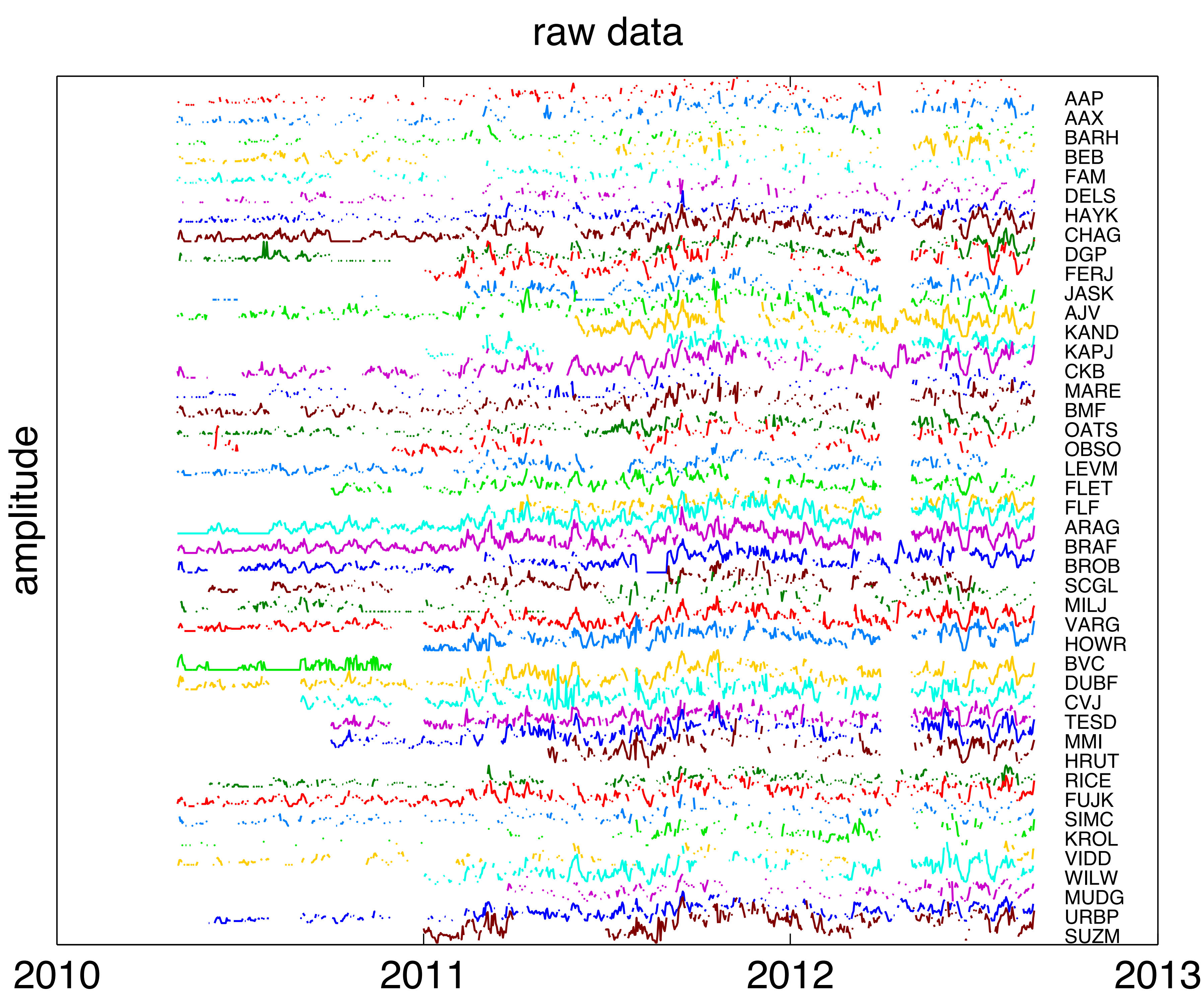
Mann [PNAS, 2008]

Key question

How do I combine incomplete and partly-overlapping records of noisy data into one single composite ?

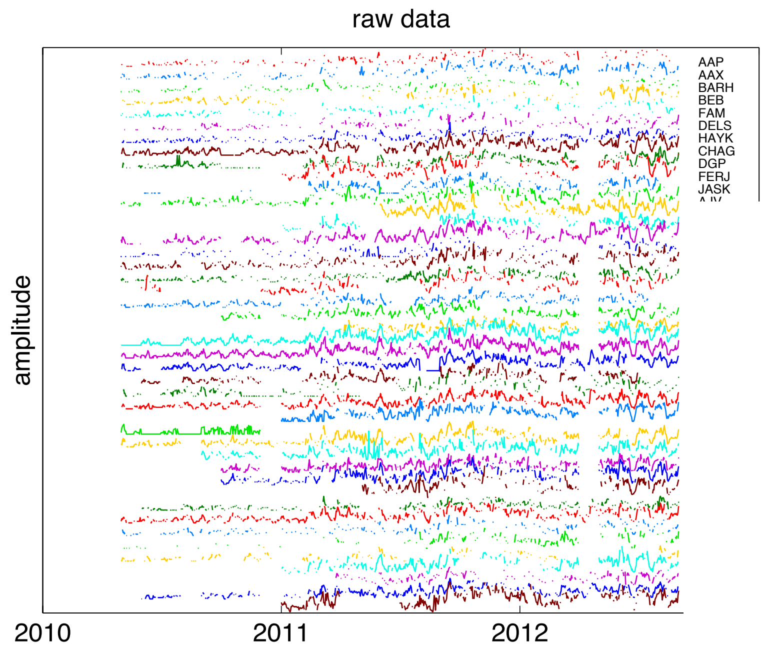
This is a **multi-sensor data fusion** problem.

Data fusion with sunspot records

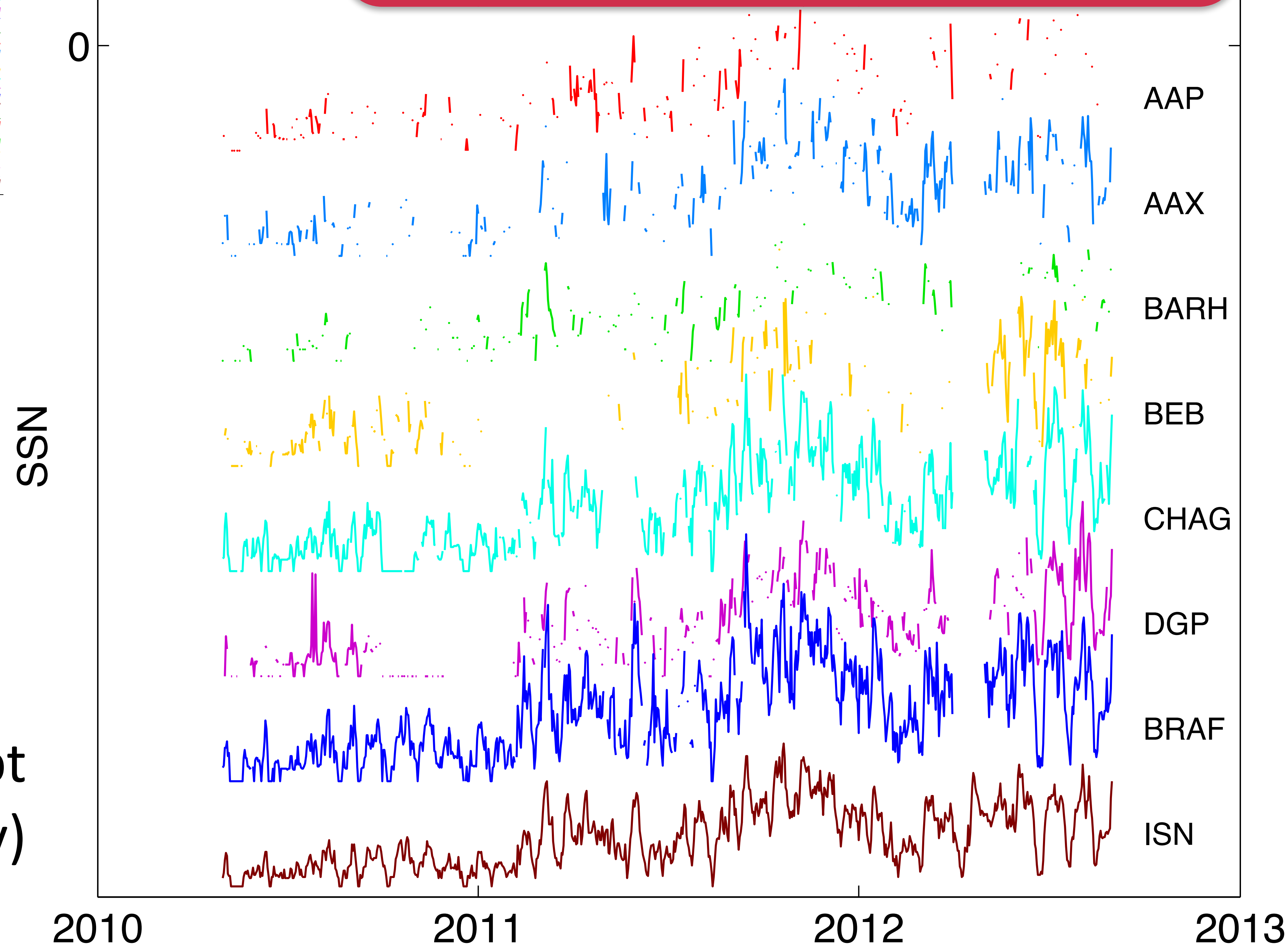


2.5 years of data, 44 observers in the US

Data fusion with sunspot records



What is the best way of making a composite ?

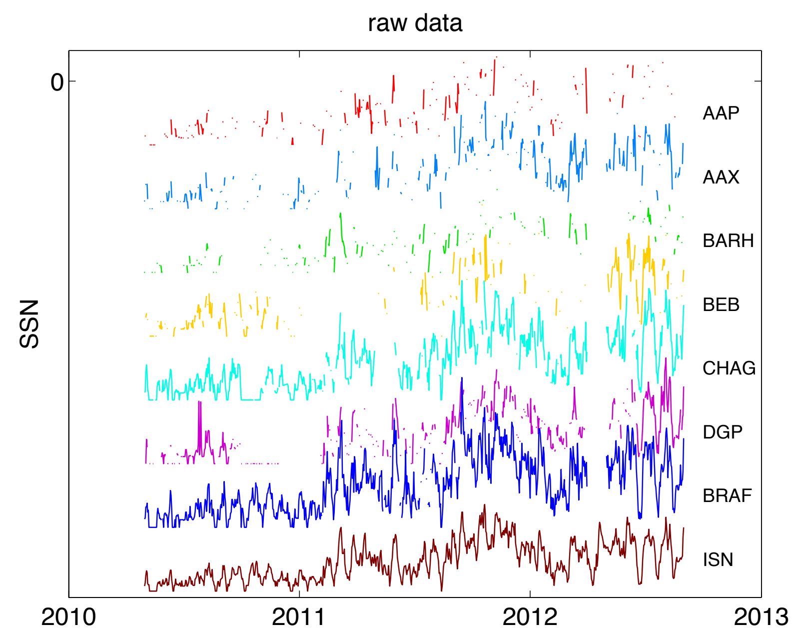


excerpt
(7 observers only)

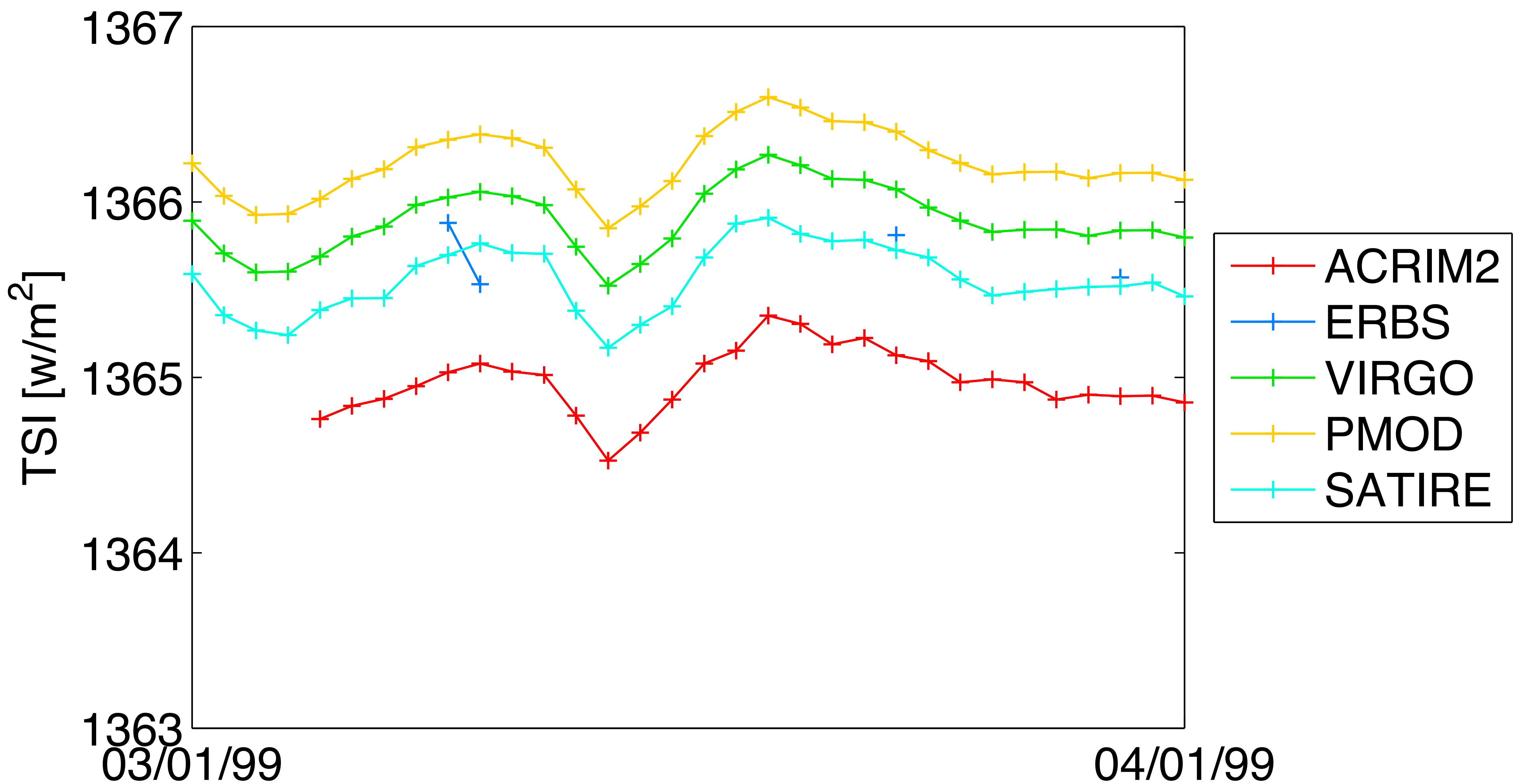
We need some best 'average' of all

Various strategies

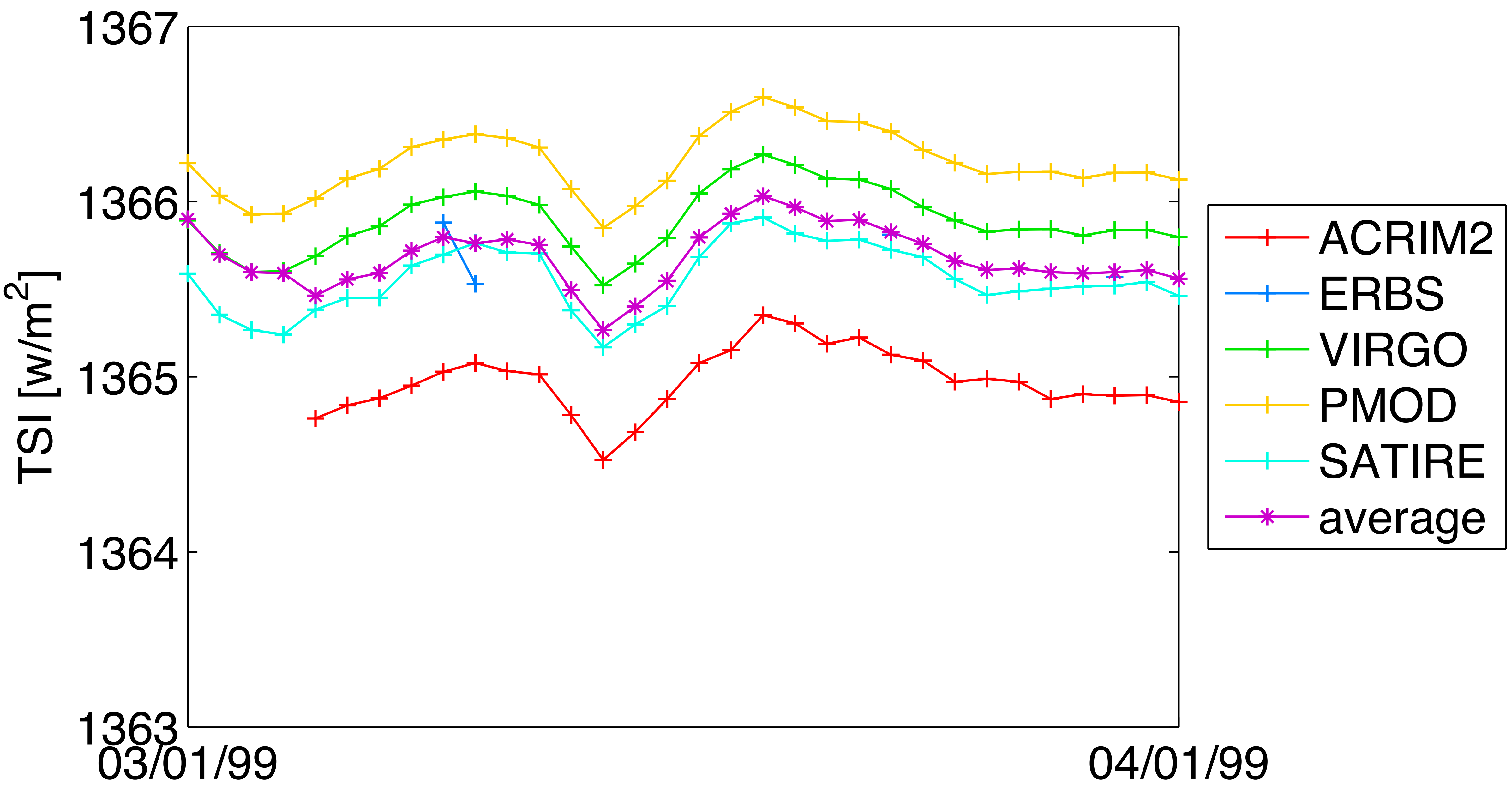
- Average (easy, but outcome depends on sample size)
- Median (less sensitive to outliers)
- TYFO (take your favorite observer)
- BOS (best of Swiss)
- ...
- Maximum a posteriori
  (Bayes estimator - probabilistic)



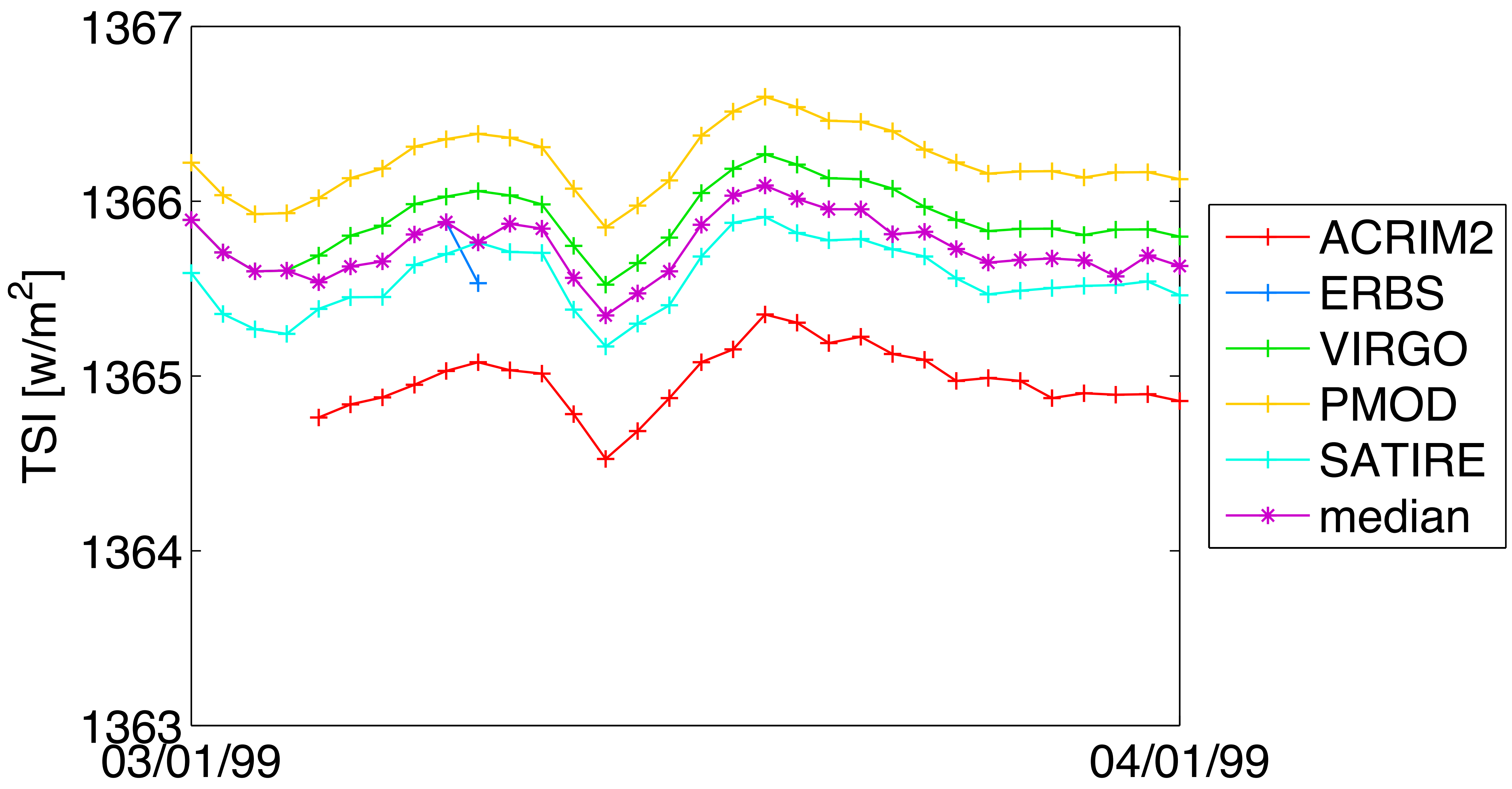
Example based on the TSI



Example based on the TSI

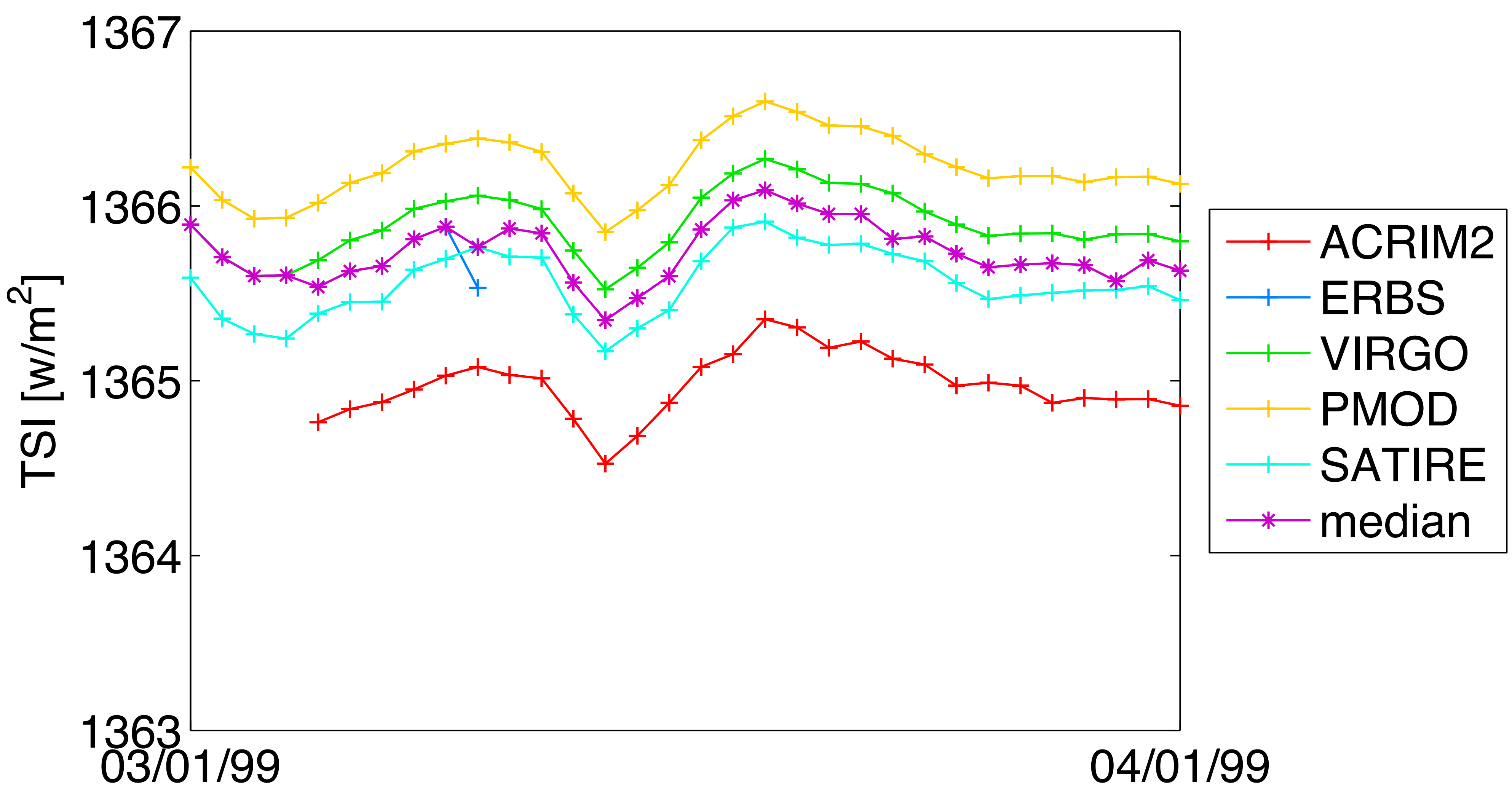


Example based on the TSI



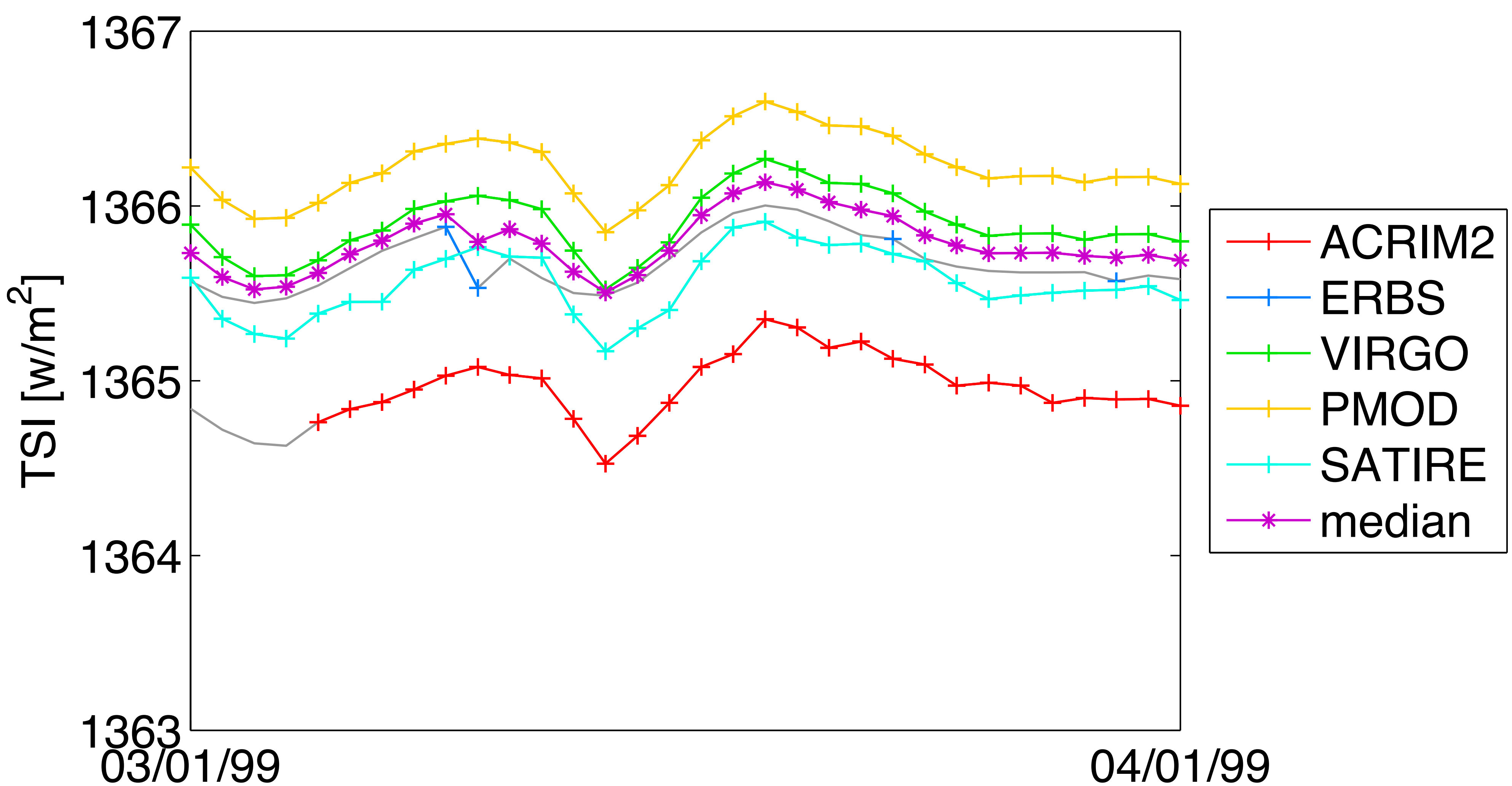
Example based on the TSI

- Some of these problems can be alleviated by filling in all data gaps → much easier to work on

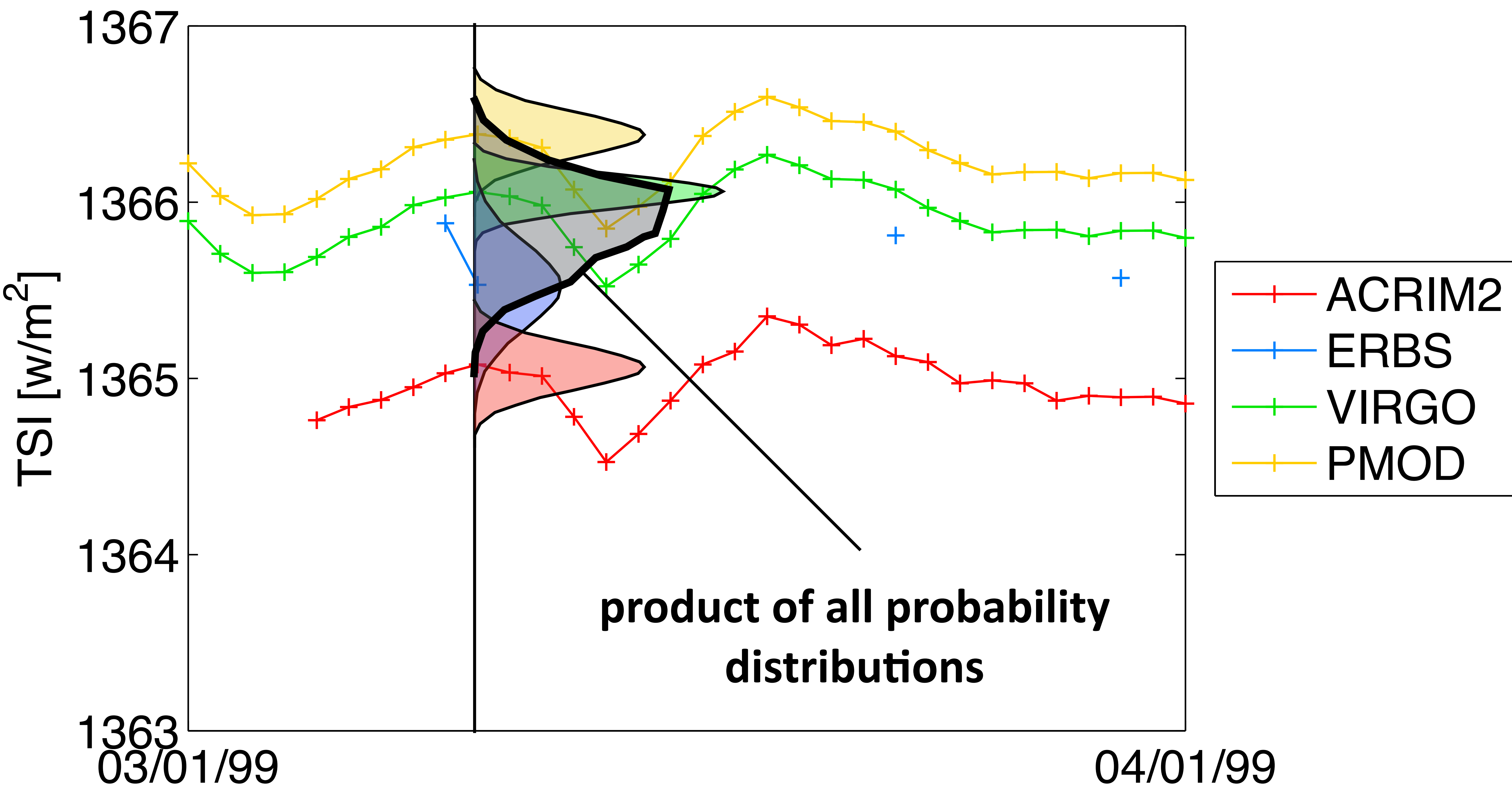


Example based on the TSI

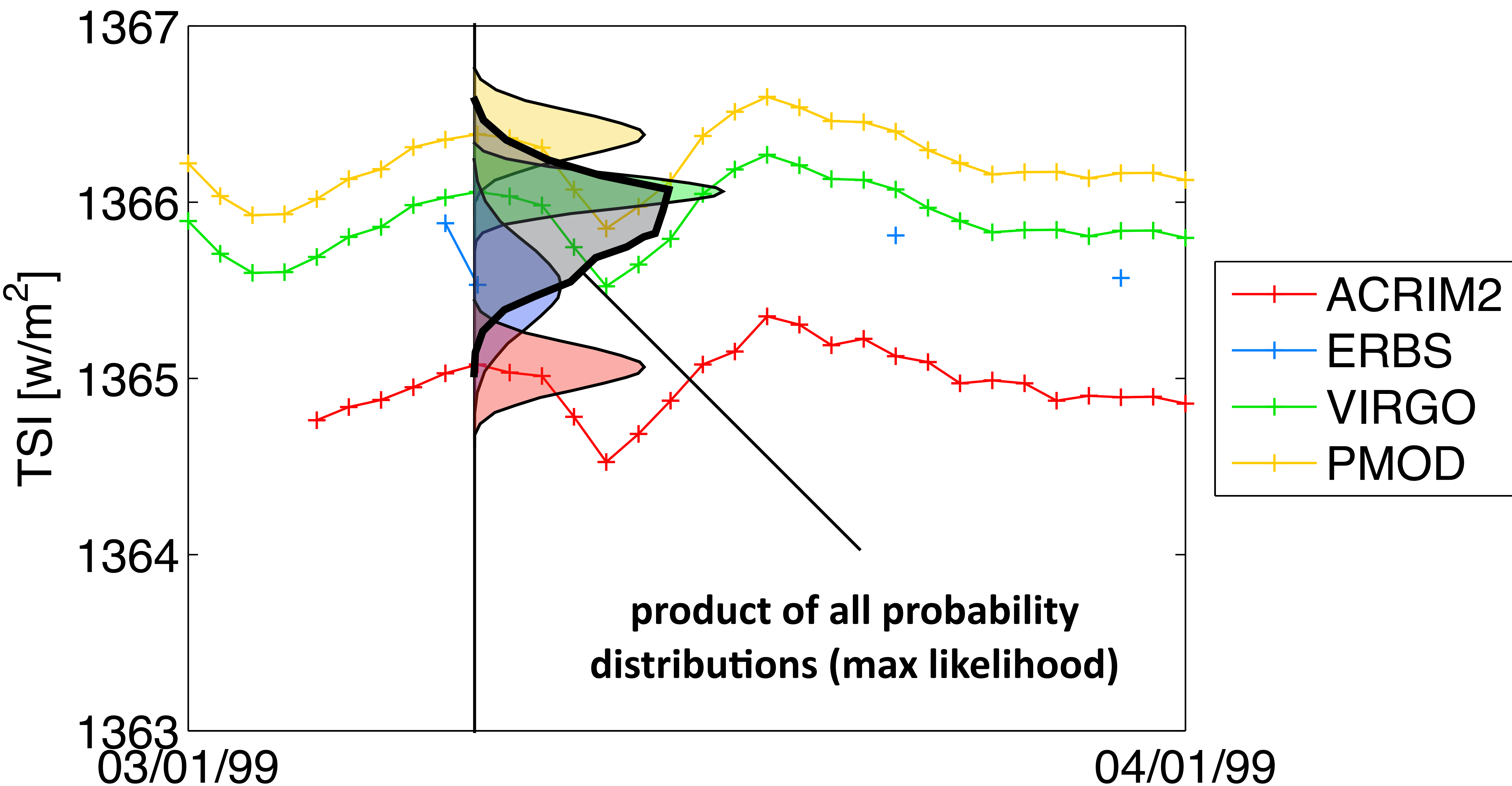
- Some of these problems can be alleviated by filling in all data gaps → much easier to work on



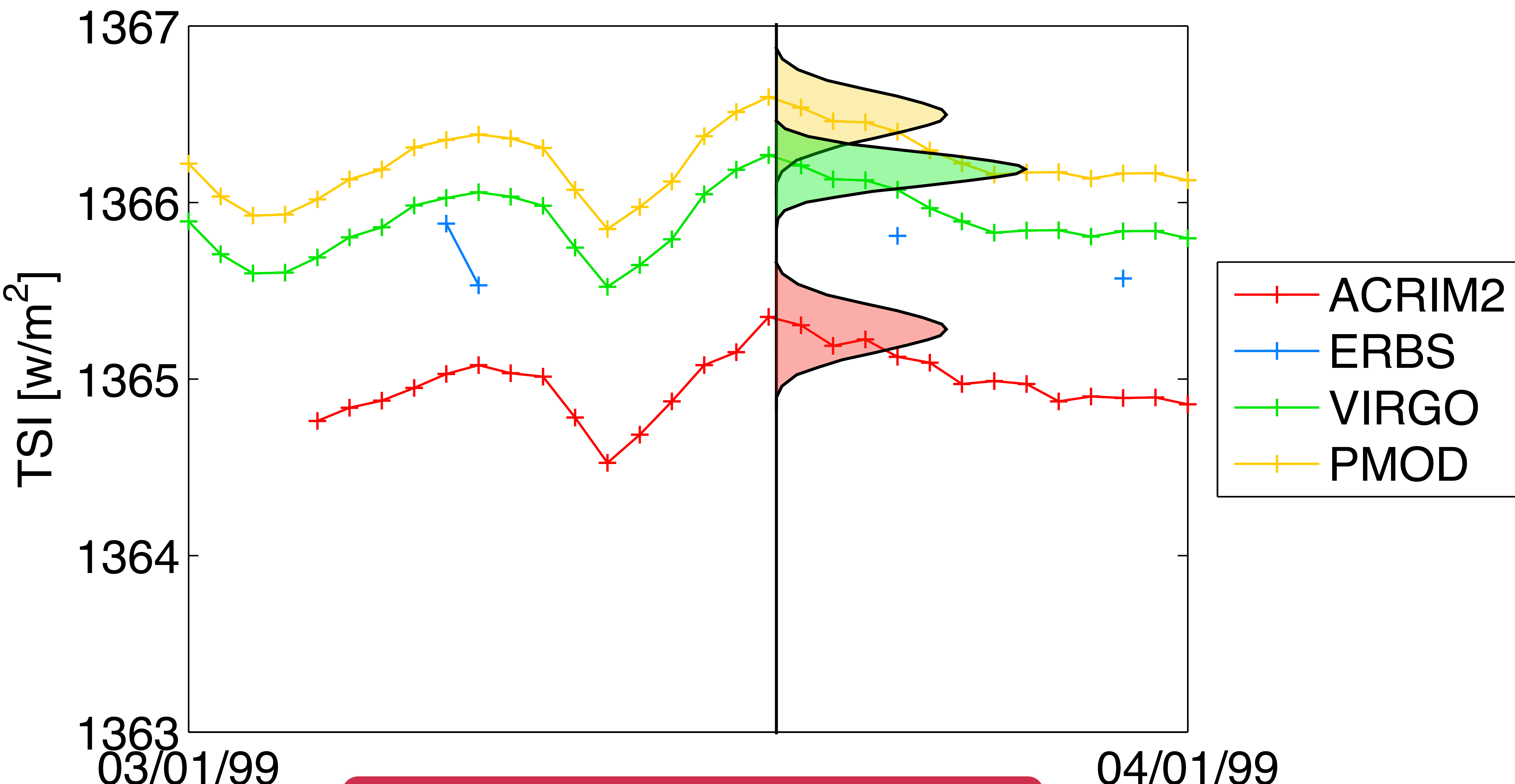
A probabilistic approach



A probabilistic approach



A probabilistic approach



Solution may not always exist

Main ingredients

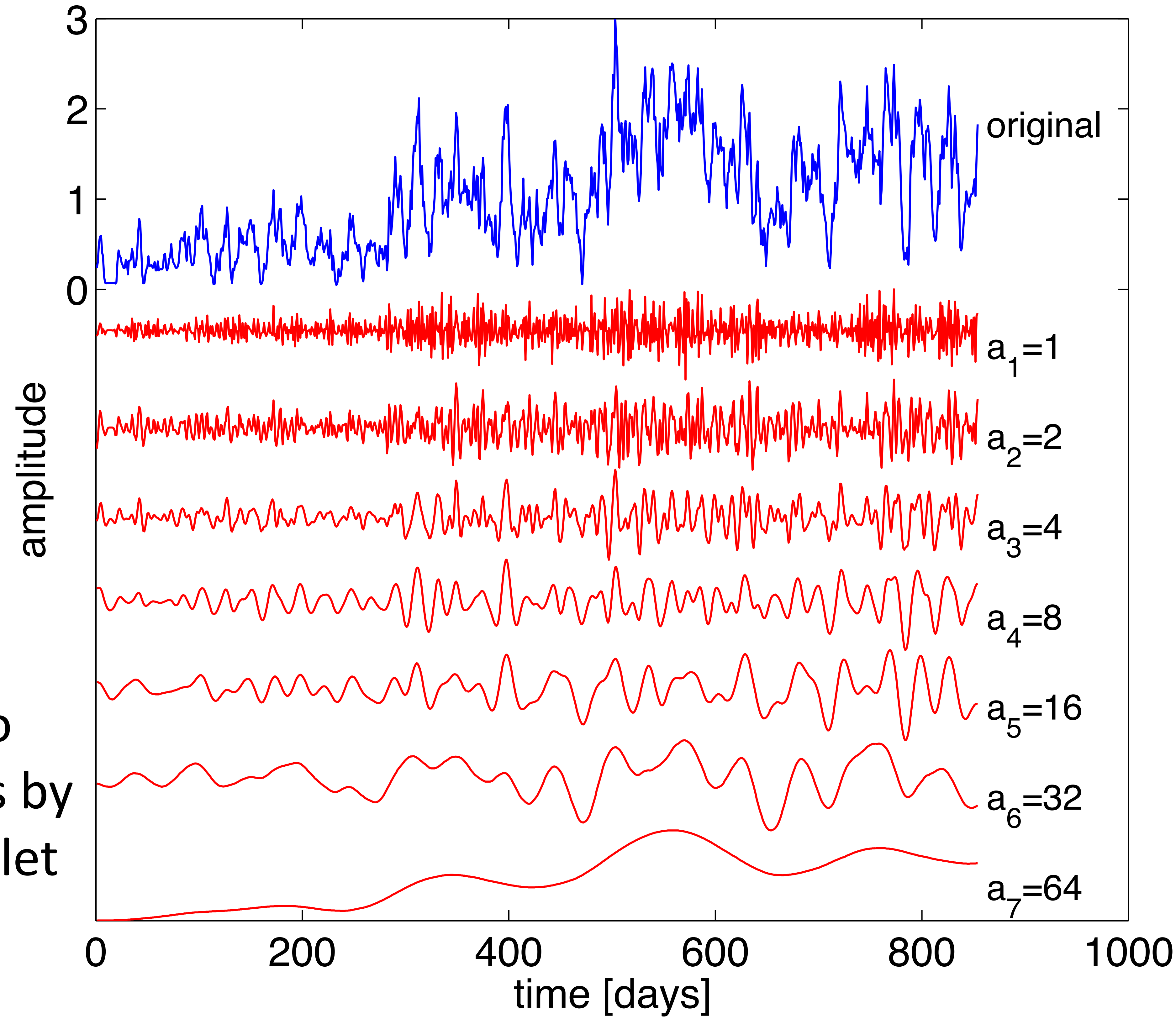
- **Keep your trash**: no not throw away (or average) any data

- **Working hypothesis**: all observations proceed from the same unknown "true" sunspot number

$$SSN_{observed}(t) = k\ SSN_{true}(t) + error(t)$$

- **Process the data scale by scale** (multiscale decomposition, aka wavelet decomposition)

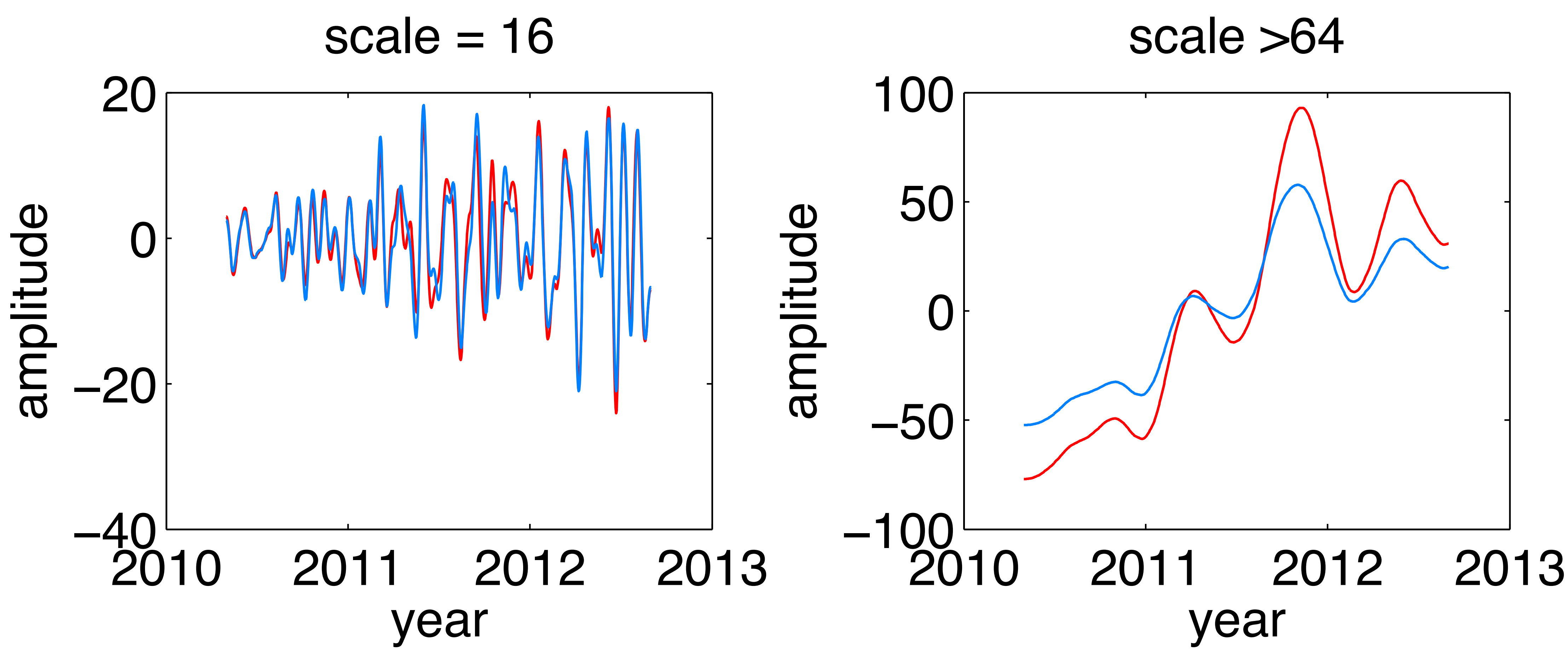
Multiscale decomposition



Each record is decomposed into multiple time scales by undecimated wavelet decomposition

Multiscale decomposition

- Agreement between SSN records often better on rotational time scales, than on other ones



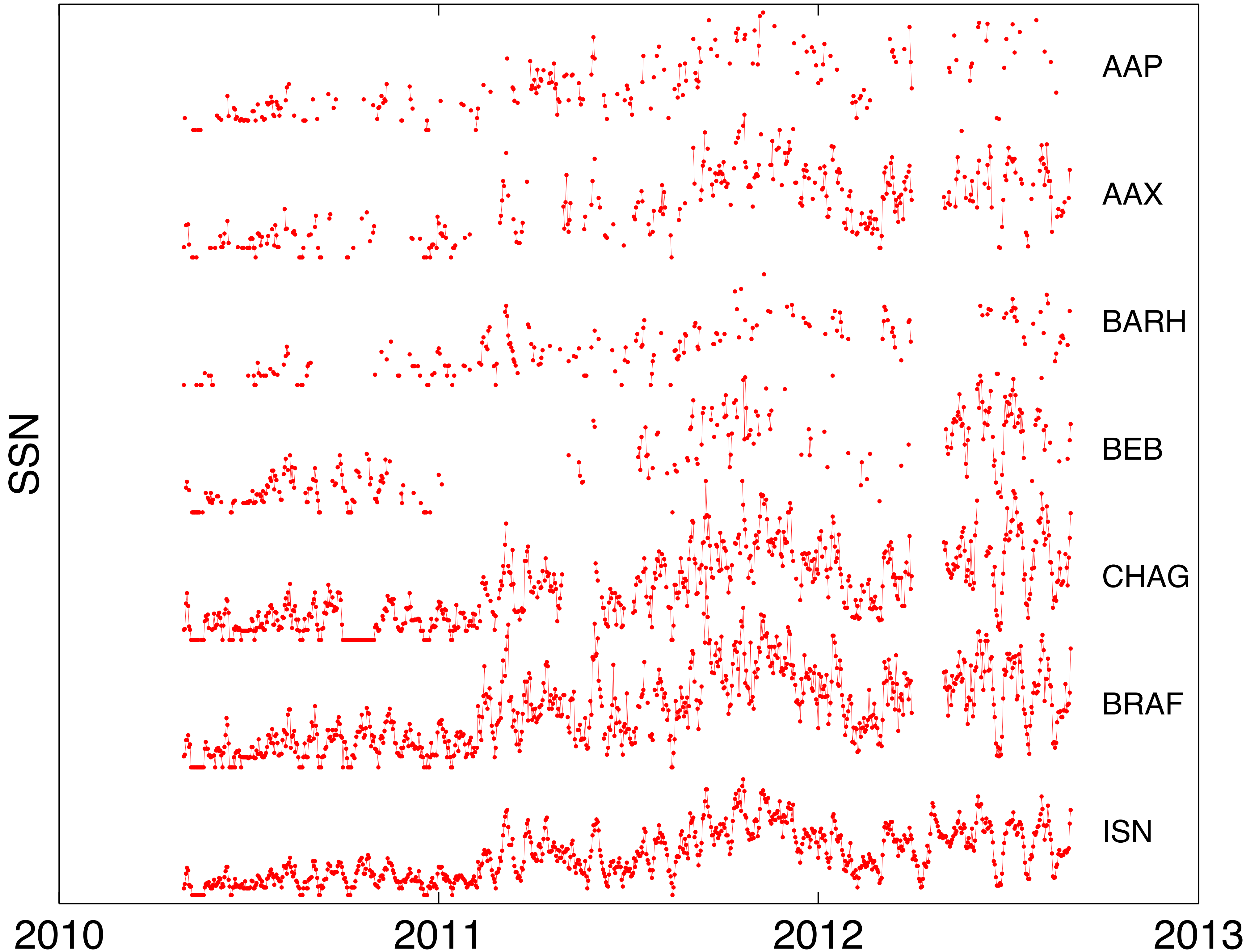
Agreement between observers is time-scale dependent

Interpolation

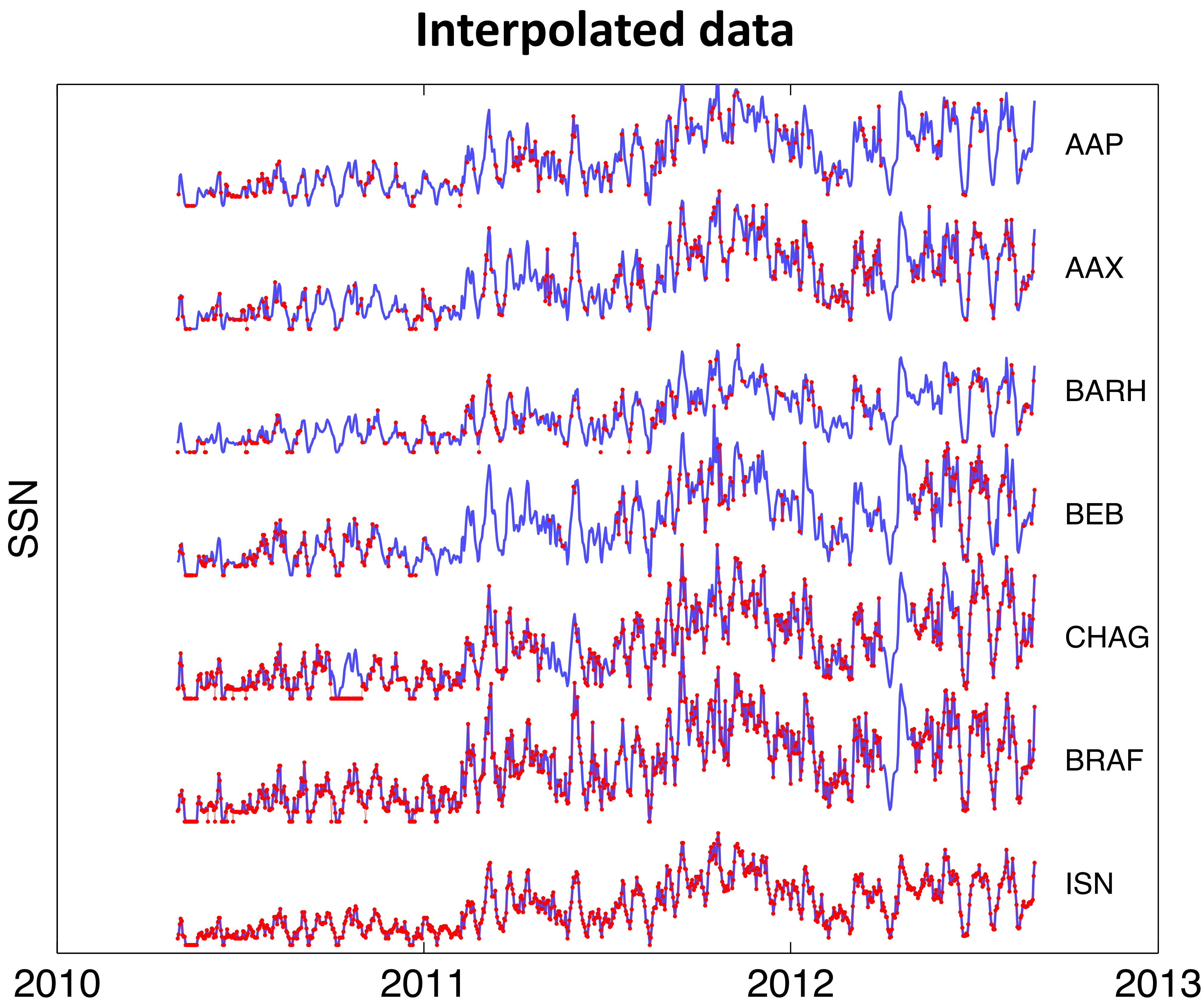
- Before decomposing the data, all gaps must be filled in
- We use **expectation-maximization**, which is well suited for highly correlated data with large gaps.
  - this method also provides confidence intervals for the interpolated data
- The interpolated values are flagged and **NOT** taken into account the subsequent analysis

Interpolation

Original data



Interpolation



One example (among many)

- Many powerful interpolation techniques exist for reconstructing missing data in correlated records

  Example: inpainting

Criminisi et al., 2004

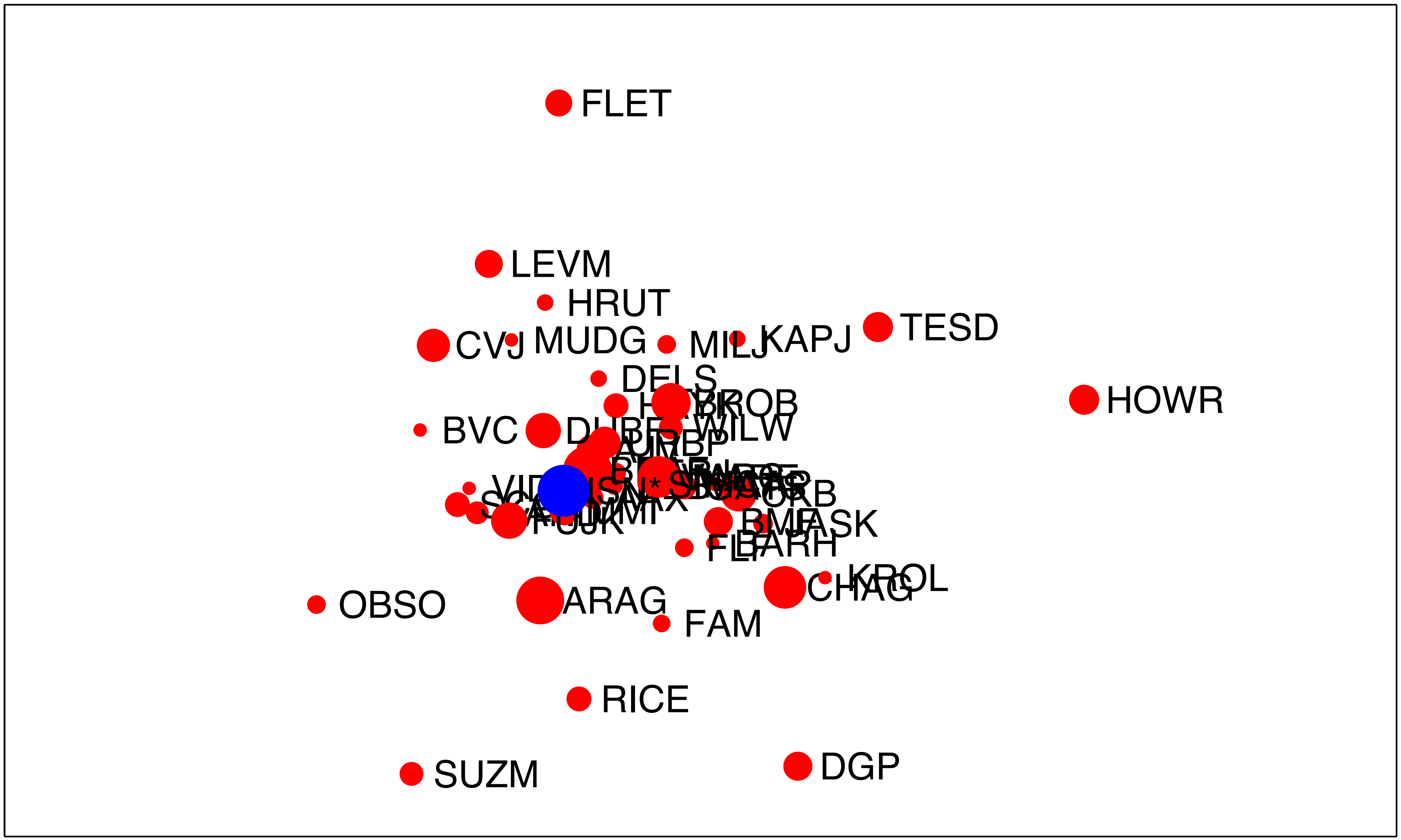
Intermezzo
Are all observers equally good ?

Multidimensional scaling

- We use **multidimensional scaling** to compare observers

1. Compute similarity between each pair of observers (e.g. Pearson correlation coefficient)
2. Distribute all observers on a 2D map in such a way that they pairwise distance reflects their lack of similarity
3. Check that this can indeed be represented in 2D

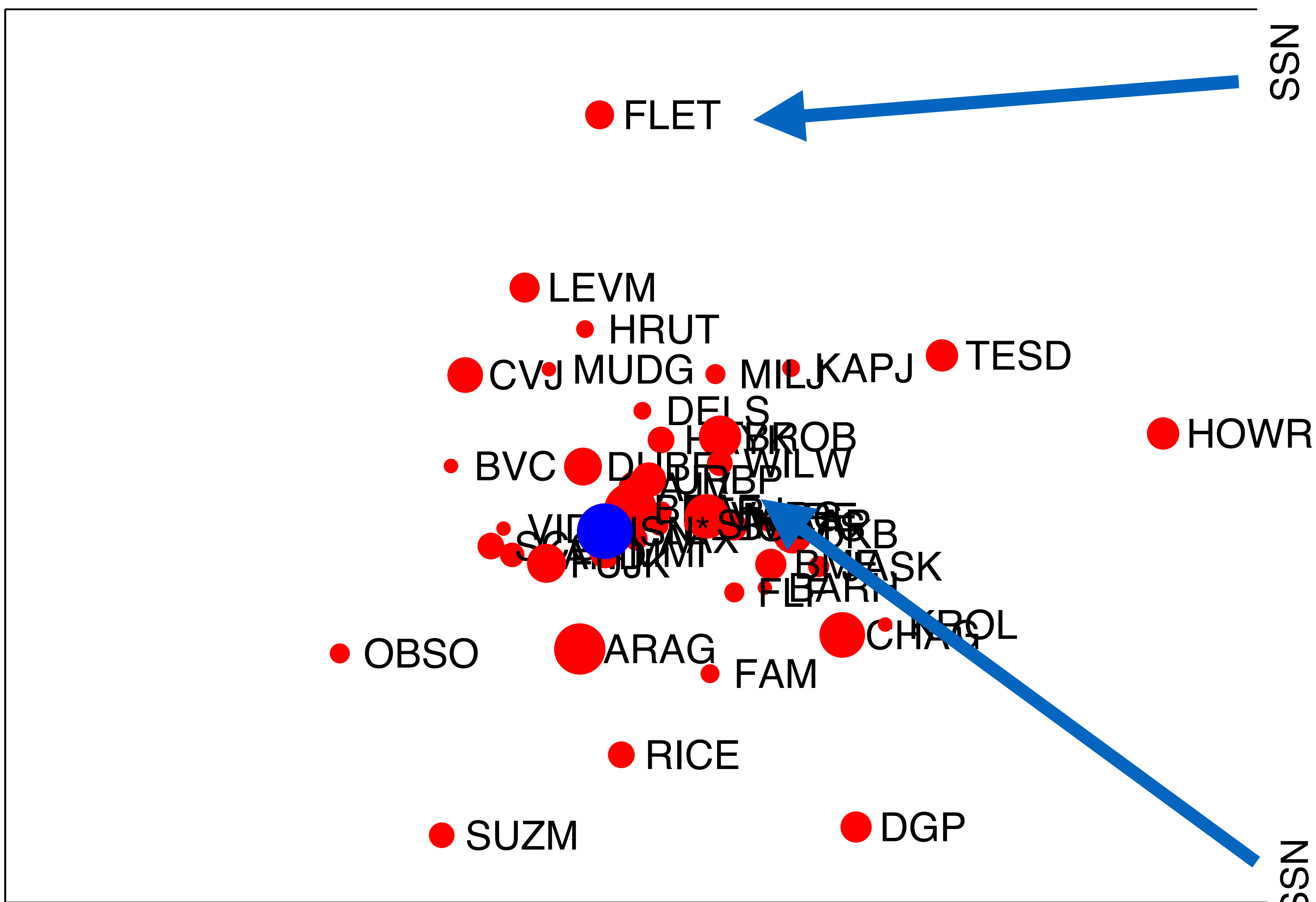
Multidimensional scaling

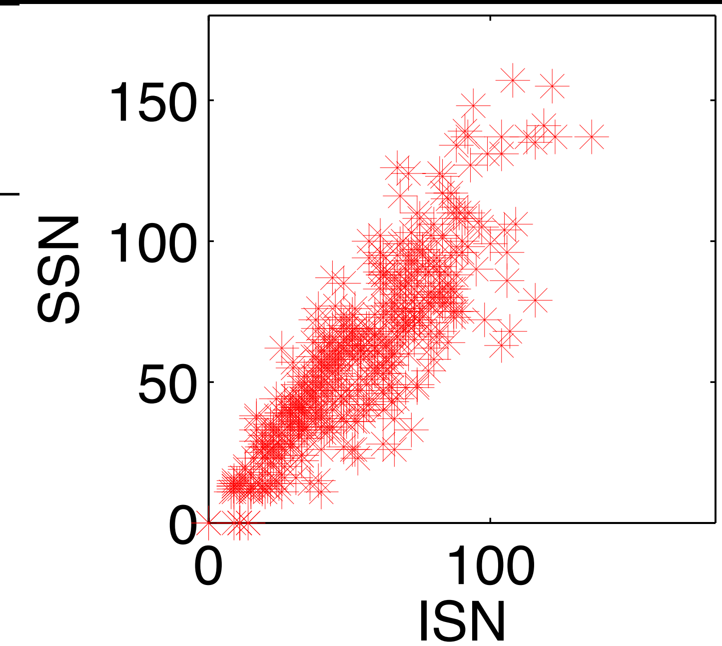


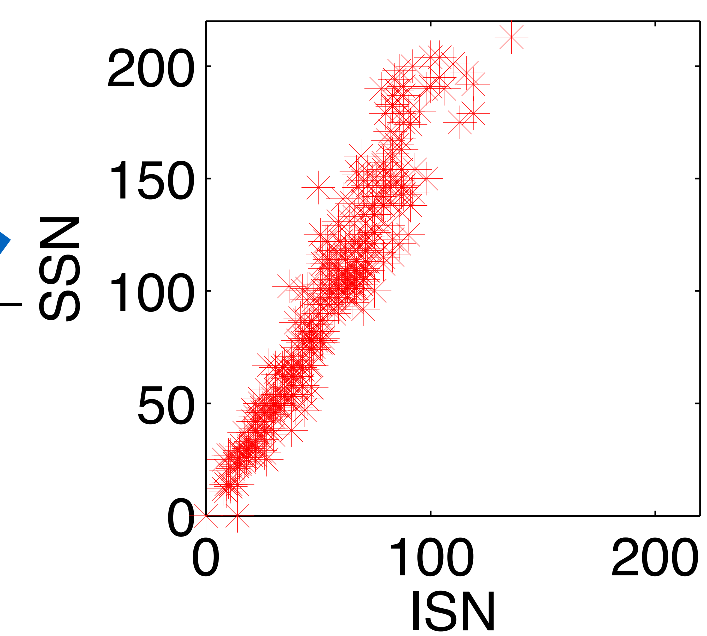
dot size reflects number of observations

axes have NO meaning - what matters is the pairwise distance

Multidimensional scaling





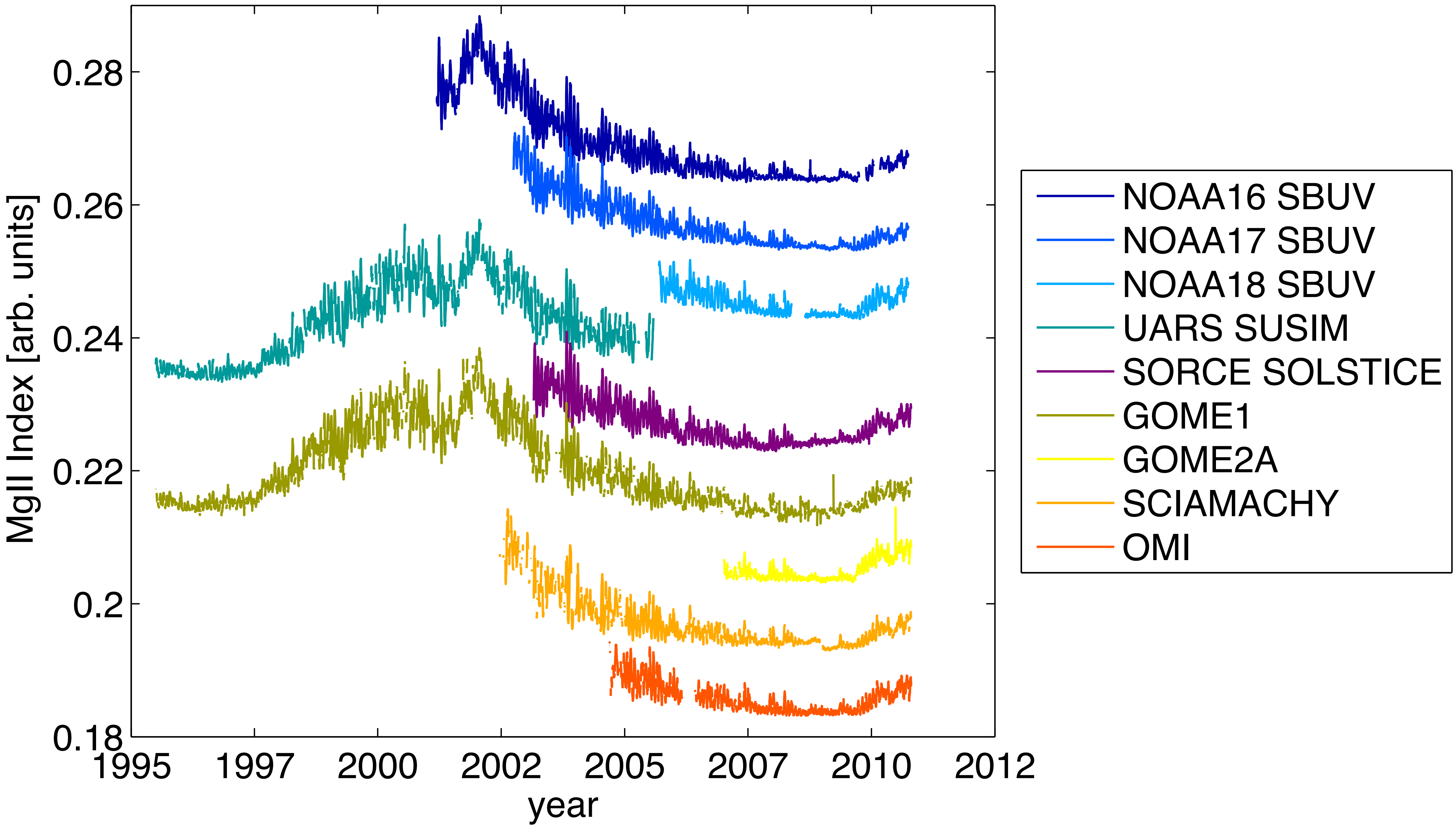


Whover is away from the "consensus cluster" has anomalous observations

Going back to the composite: Application to the MgII index

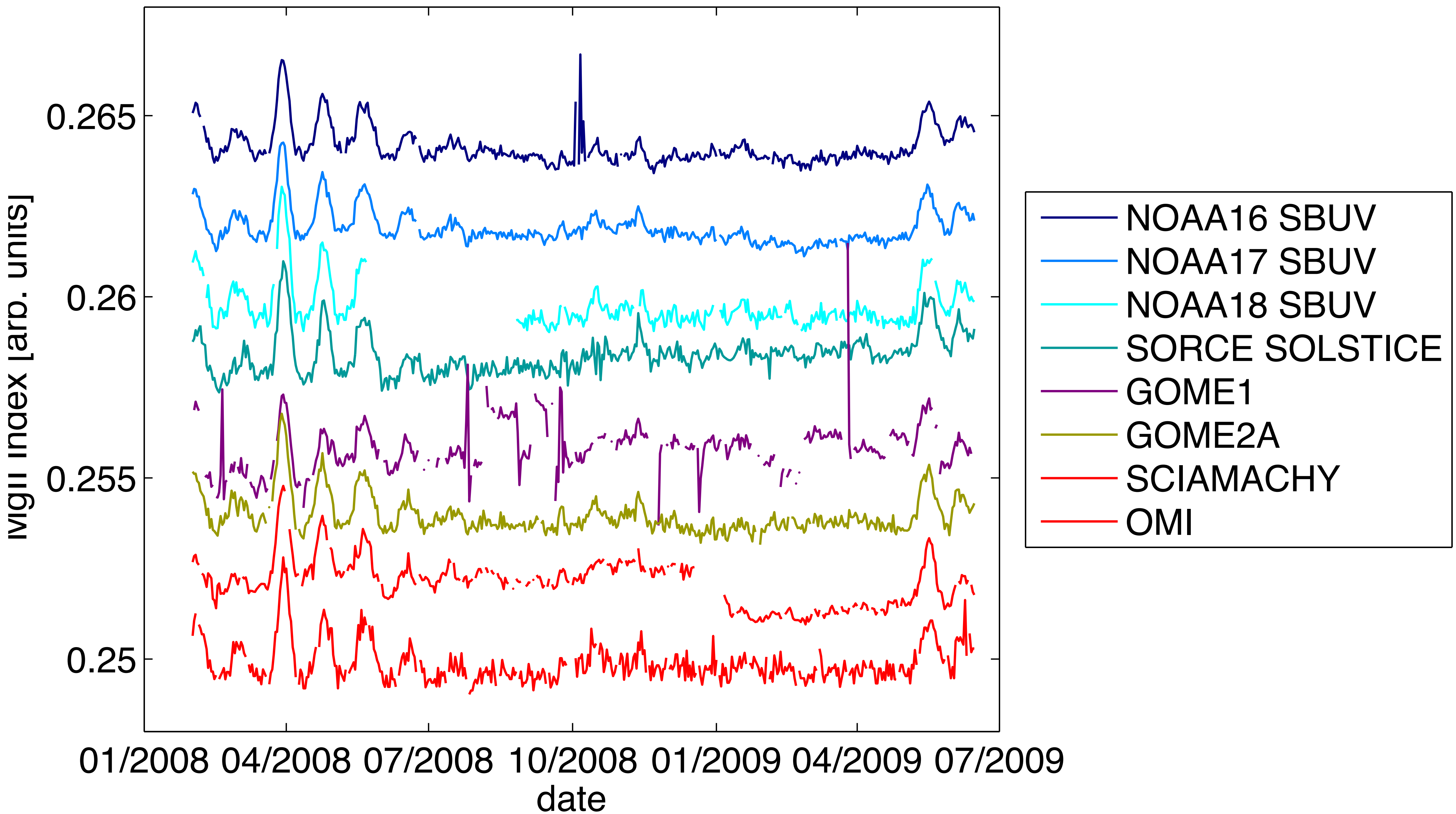
Example: making of a MgII index

9 different measurements of the core-to-wing ratio of the Mg 280nm line (= a UV proxy)



Example: making of a MgII index

- The observations disagree in many ways



Strategy for making the composite

1. Get sunspot numbers AND their confidence intervals
2. Decompose each record into multiple time scales
3. For each time scale
   - compute the maximum a posteriori (MAP) estimate: use for this a Bayesian approach

$$\mathcal{P}(c|\text{data}, I) = \frac{\mathcal{P}(\text{data}|c, I) \cdot \mathcal{P}(c|I)}{\mathcal{P}(\text{data}|I)}$$

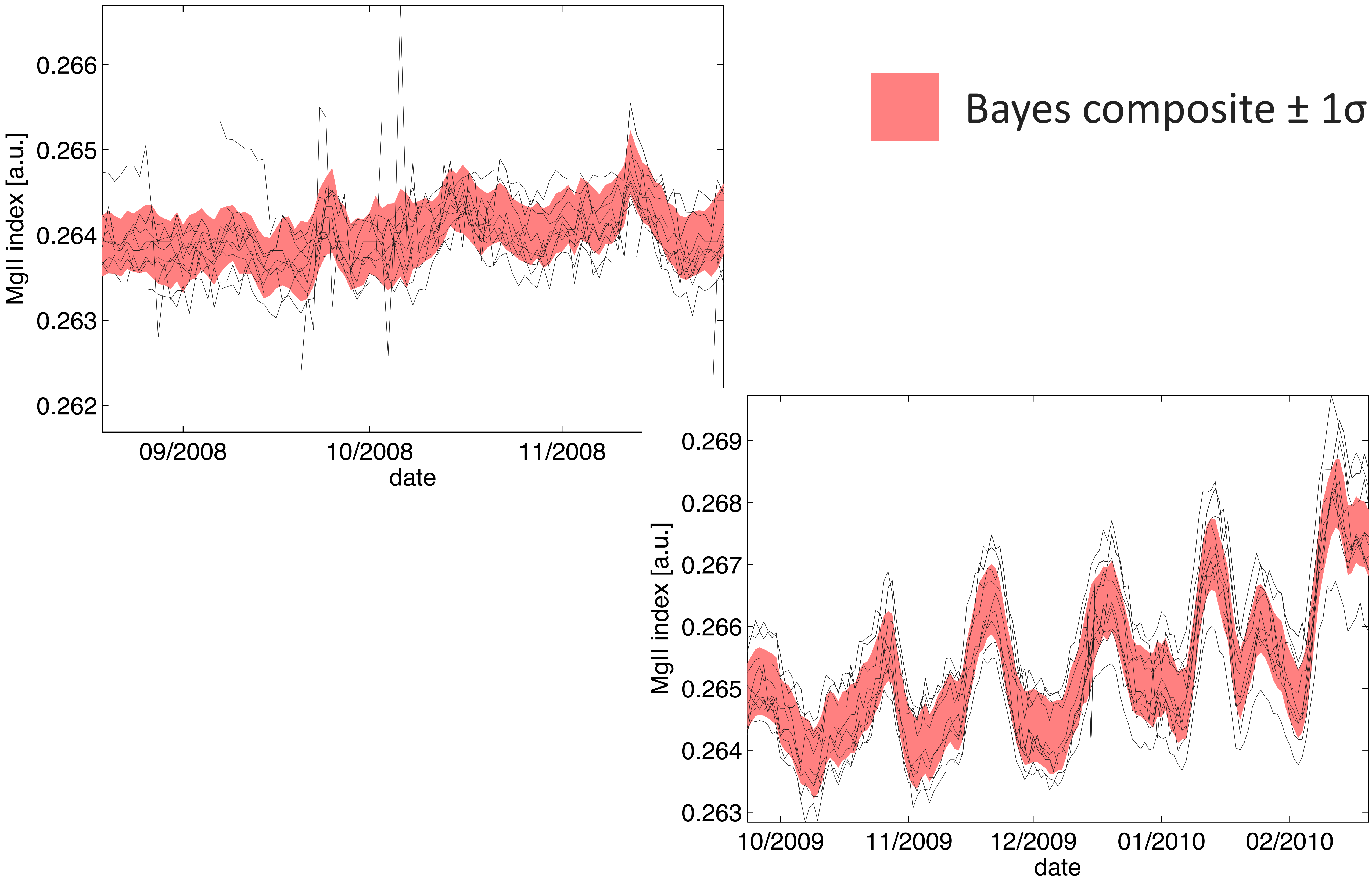
composite

prior information

   - k factors are set automatically: but some reference is needed (i.e. the ones that best matches the group)
4. Build final composite by recombining individual ones

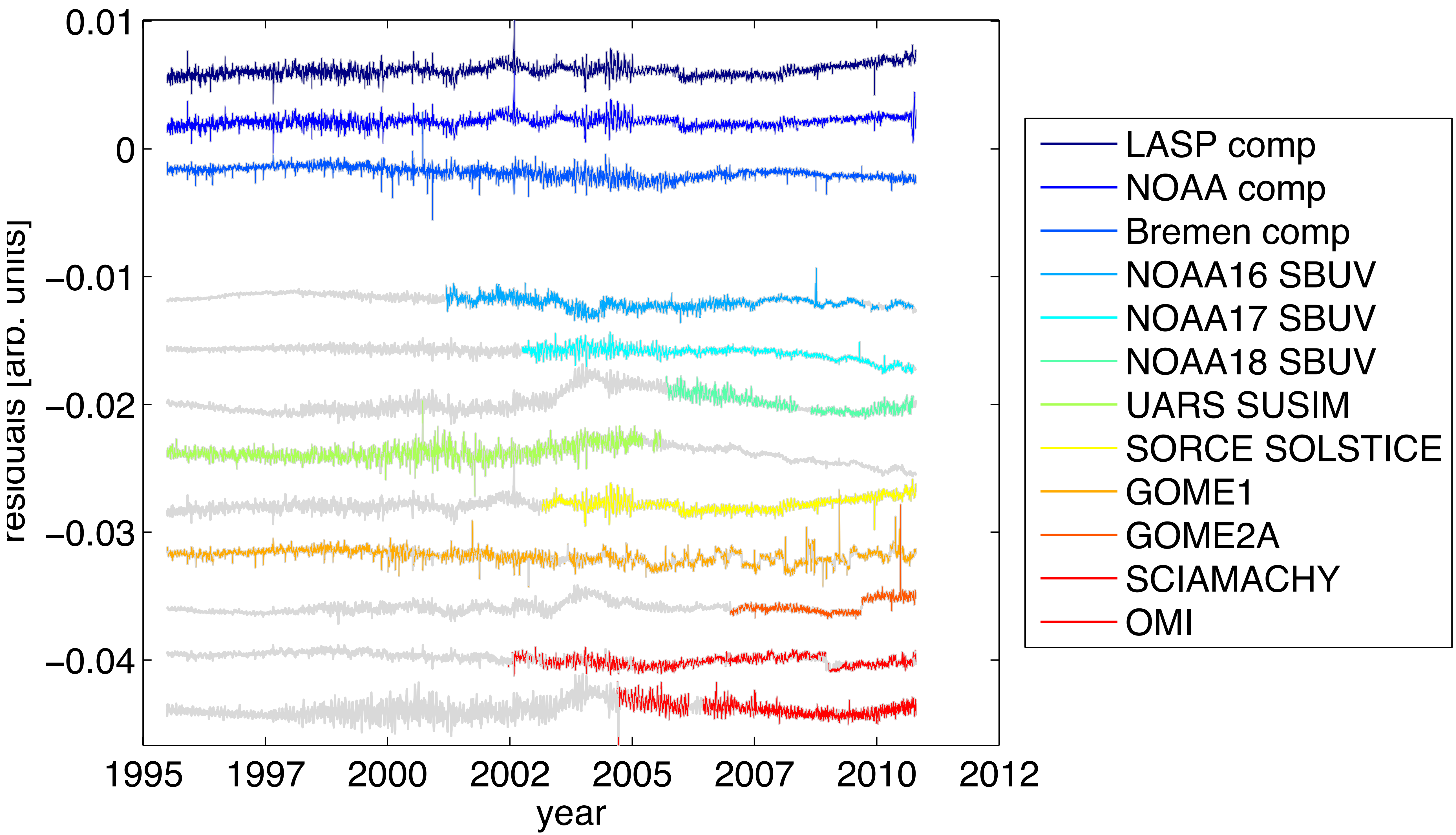
In principle straightforward; computationally expensive

Bayes composite: 2 excerpts



Residuals

Residuals = observations - Bayesian composite



To conclude

Making of composite

- Work still in progress
- Many technical issues have not been addressed here
    - the interpolation can in principle be avoided (tricky)
    - when no confidence intervals are available: give them all the same value
    - constraints can be added, e.g. SSN has to be positive

Conclusions

- The **Bayesian framework** provides a powerful and unbiased way for building a composite
  - no subjective choice of backbones - but strategy for defining the k factor can still be improved
  - working hypotheses must be explicited
  - computationally expensive...
- Decomposing the data into **multiple time scales** is essential for extracting the best out them
- This approach is now being implemented for making a composite TSI, MgII and SSI (solar spectral irradiance)