Objective Determination of the Number of Principal Components to Use in Data Reconstruction

Dave Turner, Bob Knuteson,
Hank Revercomb, and Ralph Dedecker

SPACE SCIENCE AND ENGINEERING CENTER
University of Wisconsin-Madison

Advanced High Spectral Resolution Infrared Observations Workshop
Madison, Wisconsin
26 – 28 April 2006

Office of Science
U.S. DEPARTMENT OF ENERGY

ARM Atmospheric Radiation Measurement Program

DOE Atmospheric Radiation Measurement (ARM) Program

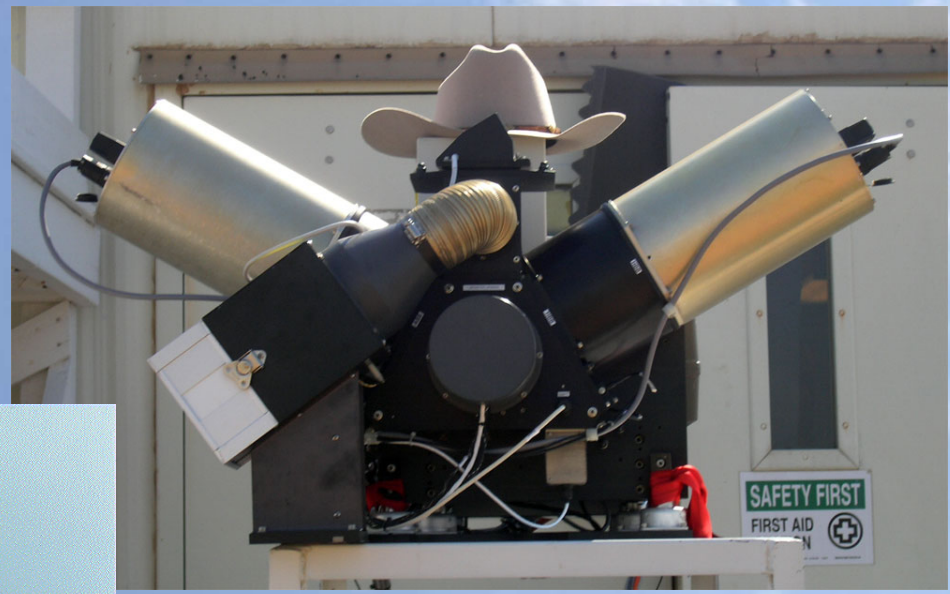
- Objective: to improve the treatment of clouds and radiation in global climate models
- Approach:
  - Collect a long-term dataset of atmospheric state, cloud and aerosol properties, and radiative fluxes at a variety of climatologically interesting sites
  - Confront the models with this data to improve the parameterizations used within the models

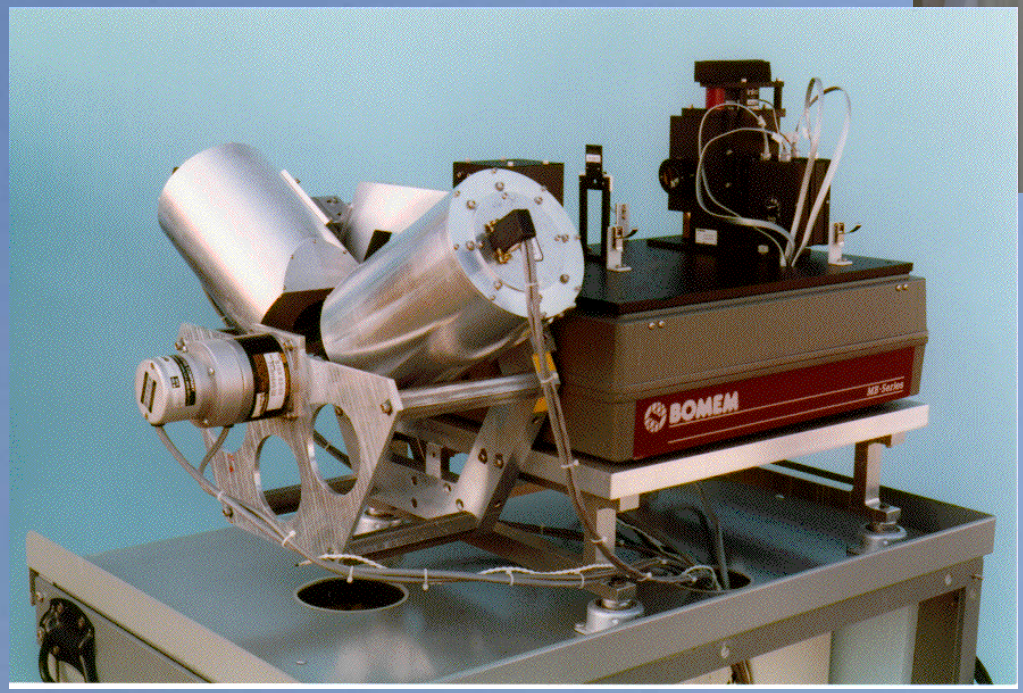
Location of ARM Sites

Atmospheric Emitted Radiance Interferometer (AERI)

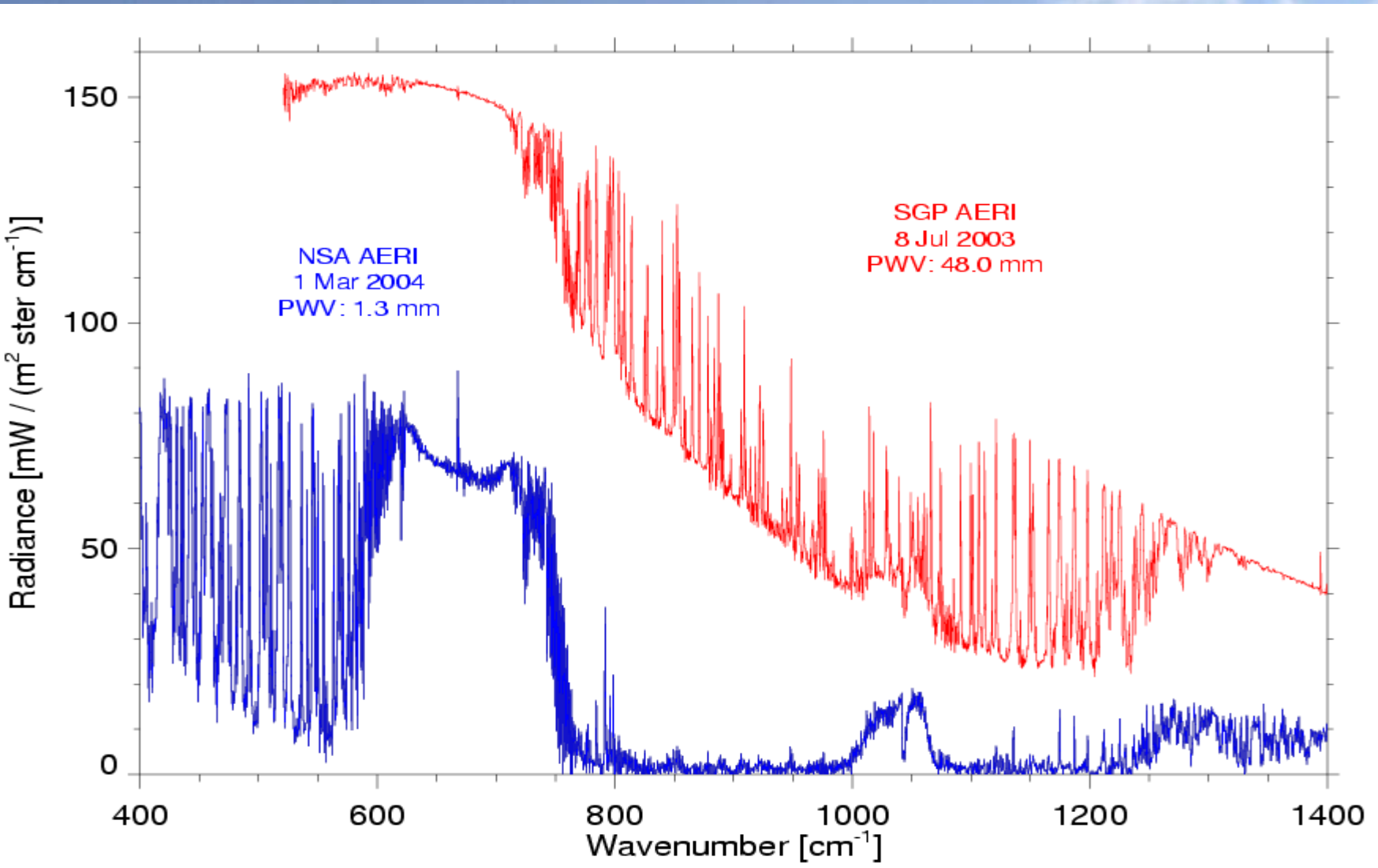
- Hardened, automated longwave interferometer
  - Nominal spectral range: 3-19 μm
  - Range extend to 3-25 μm for deployment in the Arctic
  - Spectral resolution: ~1 cm^{-1}
- Absolute calibration better than 1% of the ambient radiance (3-sigma)
- Provides ground-truth for evaluation of infrared radiative transfer models
- Data used in retrievals of temperature and water vapor profiles, cloud properties, aerosol properties, trace gases, etc.

Mug shots of the AERI





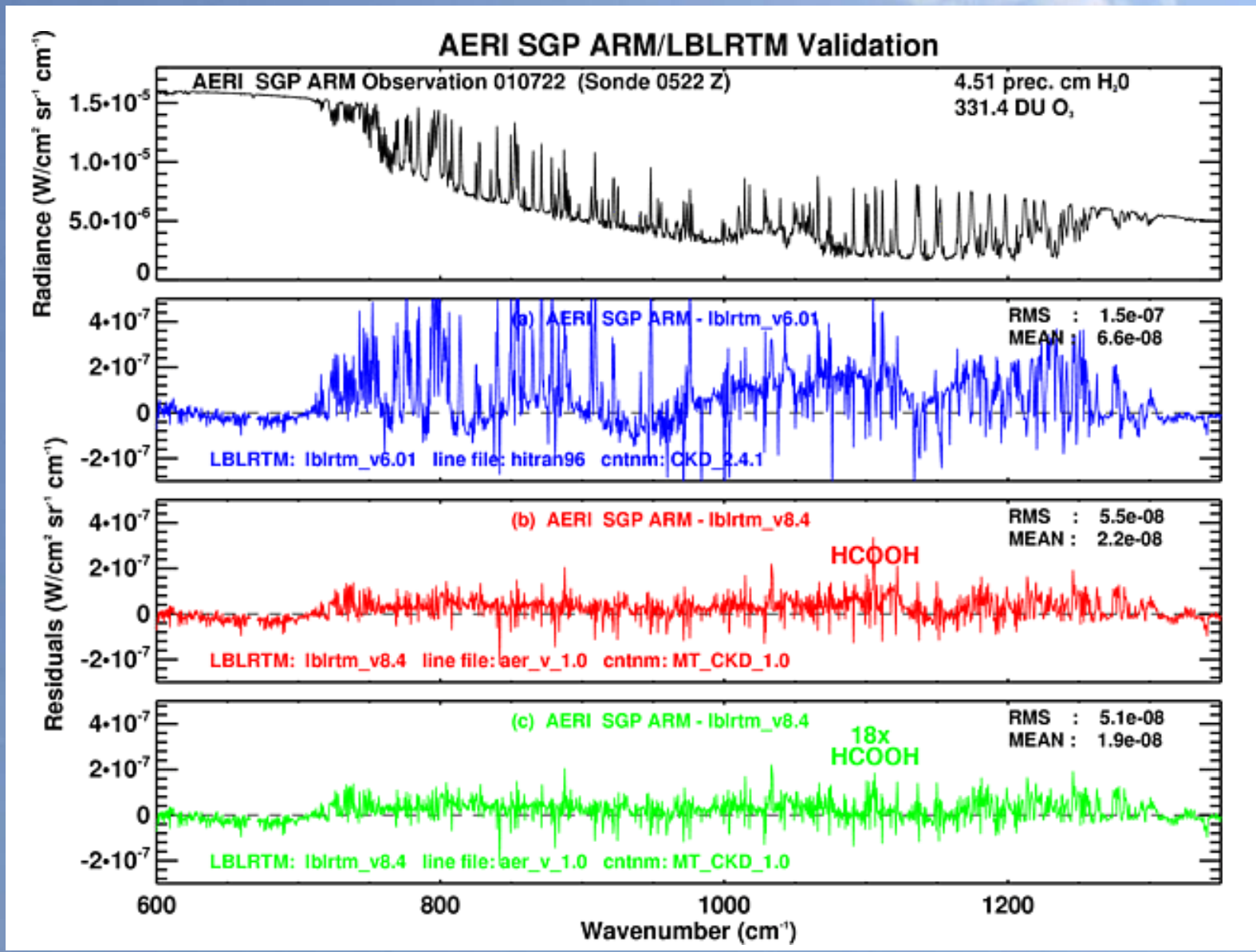
Clear Sky AERI Spectra



Early Use of AERI in ARM:
Radiative transfer model evaluation

- AERI has spectral resolution to validate line-by-line radiative transfer (RT) models
  - These models, once validated, can be used to build faster RT models
- Established a "Quality Measurement Experiment" (QME) to evaluate:
  - AERI observations
  - Physics in the LBLRTM, especially the WV continuum
  - Atmospheric state used in the model (esp. the WV profile)
- Temporal resolution of AERI set to 3-min sky averages every 8 min; ideal for this clear-sky problem
- Temporal resolution also ideal for thermodynamic profiling
- Good progress made in this arena

AERI – LBLRTM Residuals



From AER website (http://rtweb.aer.com)

Moving towards clouds…

- Infrared radiance provides an excellent constraint when retrieving cloud properties
- Algorithms have been developed that use AERI data in combination with lidar and/or radar
- 3-min sky average every 8 min was unsatisfactory, given cloud conditions can vary markedly in this period
- Prototyped a “rapid-sample” mode in July 2002, which allows 12-s sky averages to be collected every 20-30 s
- ARM is upgrading all AERIs to run in rapid-sample mode
- Decreasing the sky averaging interval increases the random error in the data…

PCA noise filter

- PCA provides a mechanism to reduce the uncorrelated random error in the observations
- Challenge: find an objective way to determine the "proper" number of PCs to use in the reconstruction
  - Important as ARM's data system requires all processes be automated; human interaction not really allowed in the process
  - Many AERIs to process (ARM currently has 6 AERIs in the field)
  - Ideally, we hope to avoid repeated reconstructions to find the proper number of PCs to use

PCA noise filter method

1. Normalize each observed spectrum with the estimated NESR
2. Compute covariance matrix $\mathbf{C} = \mathbf{M}^{T}\,\mathbf{M}$
3. Derive PCs (eigenvectors) and eigenvalues of **C**
4. Determine the number *k* of PCs to use in the reconstruction
5. Project each spectrum onto the vector space spanned by these *k* PCs
6. Reconstruct the data using the projection coefficients and the *k* PCs
7. Multiply each spectrum by the NESR used in step 1

Some errata

$$\mathbf{C} = \mathbf{U}\,\mathbf{D}\,\mathbf{V}^T$$

Singular value decomposition of the covariance matrix **C**

$$\mathbf{D}_{ii} = \lambda_i, \quad \mathbf{D}_{ij} = 0 \; for \; i \neq j$$

D is a diagonal matrix of eigenvalues λ

$$\mathbf{U} = \mathbf{V}$$

Columns are real-valued eigenvectors (PCs) since **C** is real-valued and symmetric. Eigenvectors of covariance matrix **C** are same as for data matrix **M**!

$$\sum_{i=1}^{n} \lambda_i$$

Sum of the eigenvalues is the total variance in the data

$$\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots \geq \lambda_n$$

Typically arrange eigenvalues into descending order (**U**, **V** accordingly)

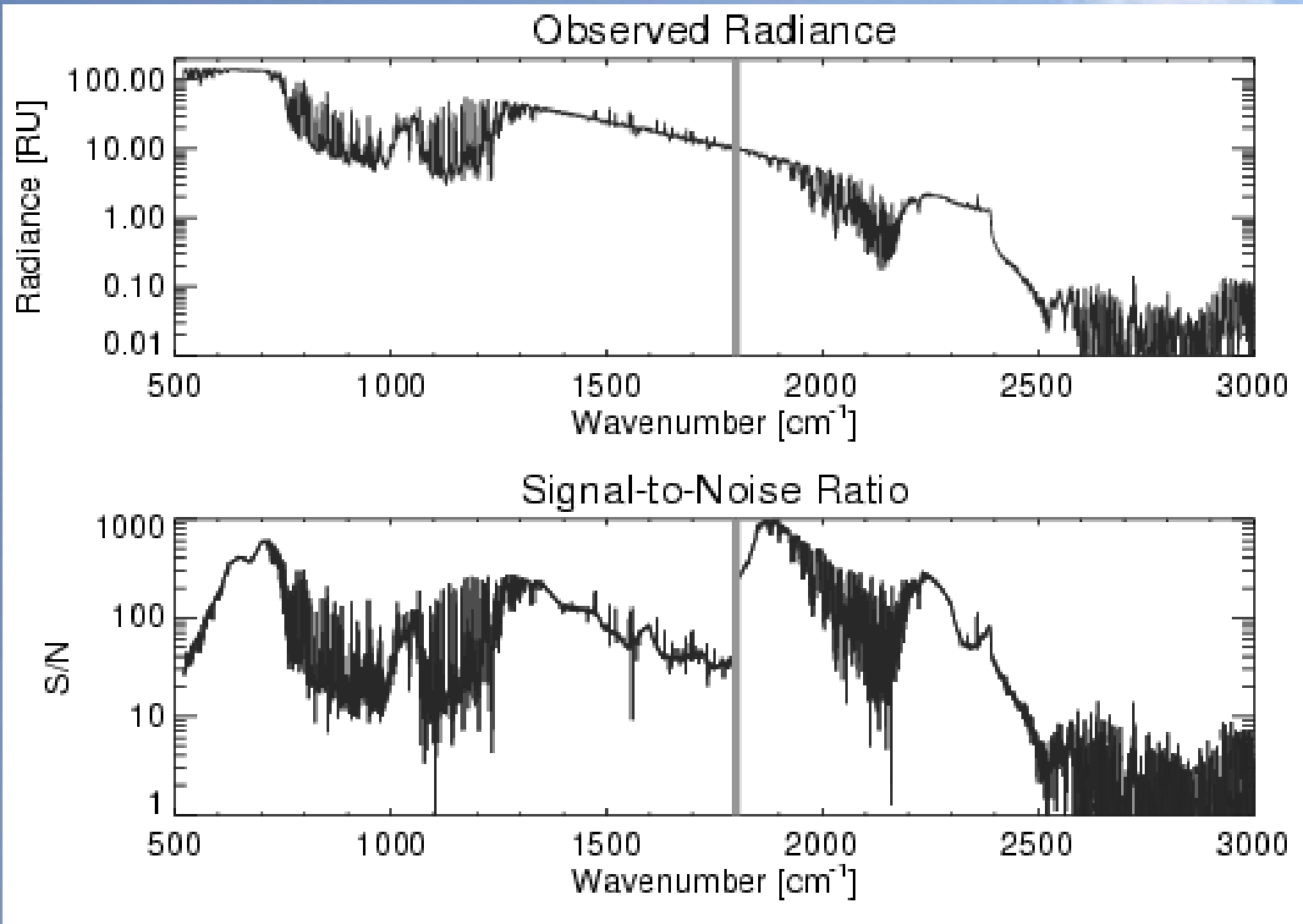
$$\hat{\mathbf{C}}_{n,n} = \mathbf{U}_{n,k}\,\mathbf{D}_{k,k}\left(\mathbf{V}_{n,k}\right)^T$$

Reconstruction reduces dimensionality of problem if $k < n$

Example data used here…

- Operational ARM AERI (the "AERI-01") at ARM SGP central facility
- Deployed a second AERI system at SGP site in Oct-Nov 2003
  - Univ of Wisconsin – Madison in "Bago"
  - This unit was in rapid-sample (RS) mode
- Noise filter applied to RS data; comparisons with AERI-01 used to help evaluate the filter

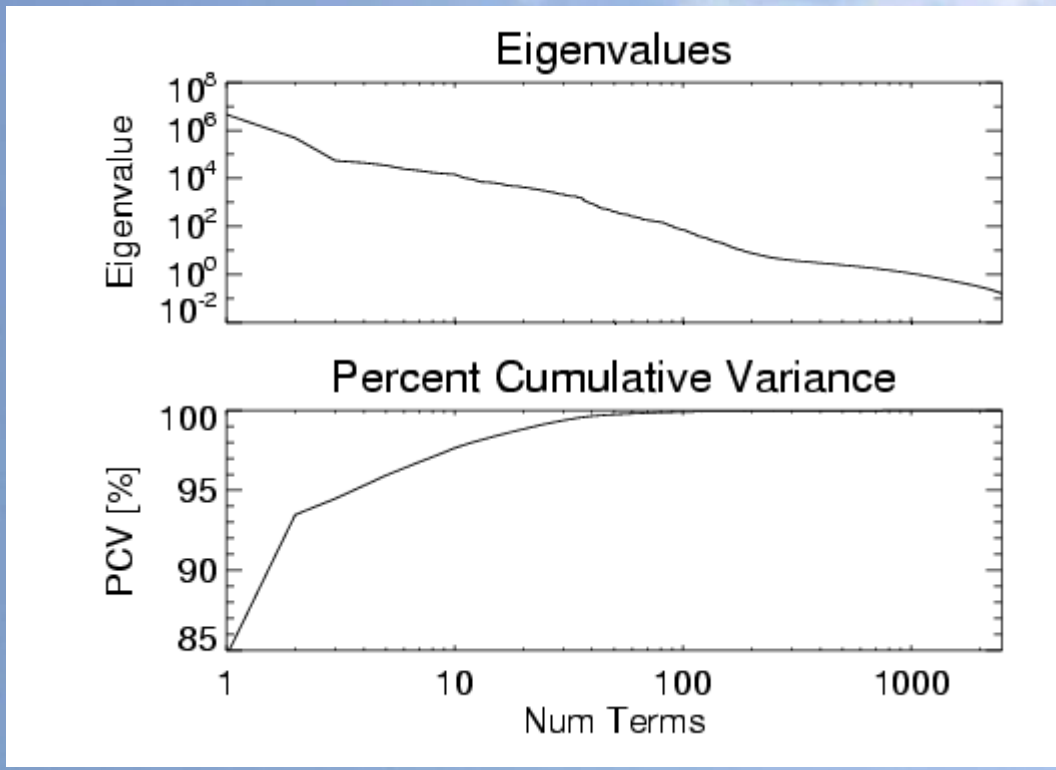
Example AERI obs and NESR



Determining the "proper" *k* (1)

- Many subjective methods being used
  - E.g., **P**ercent **C**umulative **V**ariance below some threshold

$$PCV(k) = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{n} \lambda_i}$$



  - Fast, as it works directly from the eigenvalues
  - Problem: how to set the threshold?

Determining the "proper" *k* (2)

- More computational expensive methods
  - E.g., **R**econstruction **Sc**ore below 1

$$RSc(k) = \left[\frac{1}{n} \sum_{v} \left(I_v - R_v(k) \right)^2 \right]^{1/2}$$

  - Expensive, as one must reconstruct data for all *k*

Determining the "proper" *k* (3)

$$\sum_{i=1}^{n} \lambda_i = \sum_{i=1}^{k} \lambda_i^x + \sum_{i=k+1}^{n} \lambda_i^o$$

λ^x associated with PCs that contain mixture of true signal and random error
λ^o associated with PCs that contain uncorrelated random error only

- Chemical analysis community has been using PCA for decades
- Edmund Malinowski (in a series of papers) showed:

$$(RE)^2 = (IE)^2 + (XE)^2$$

Real "error"

Imbedded "error"

Extracted "error"

E.R. Malinowski's Formulations

$$RE(k) = \left[\frac{\sum_{i=k+1}^{n} \lambda_i^0}{t(n-k)} \right]^{1/2}$$

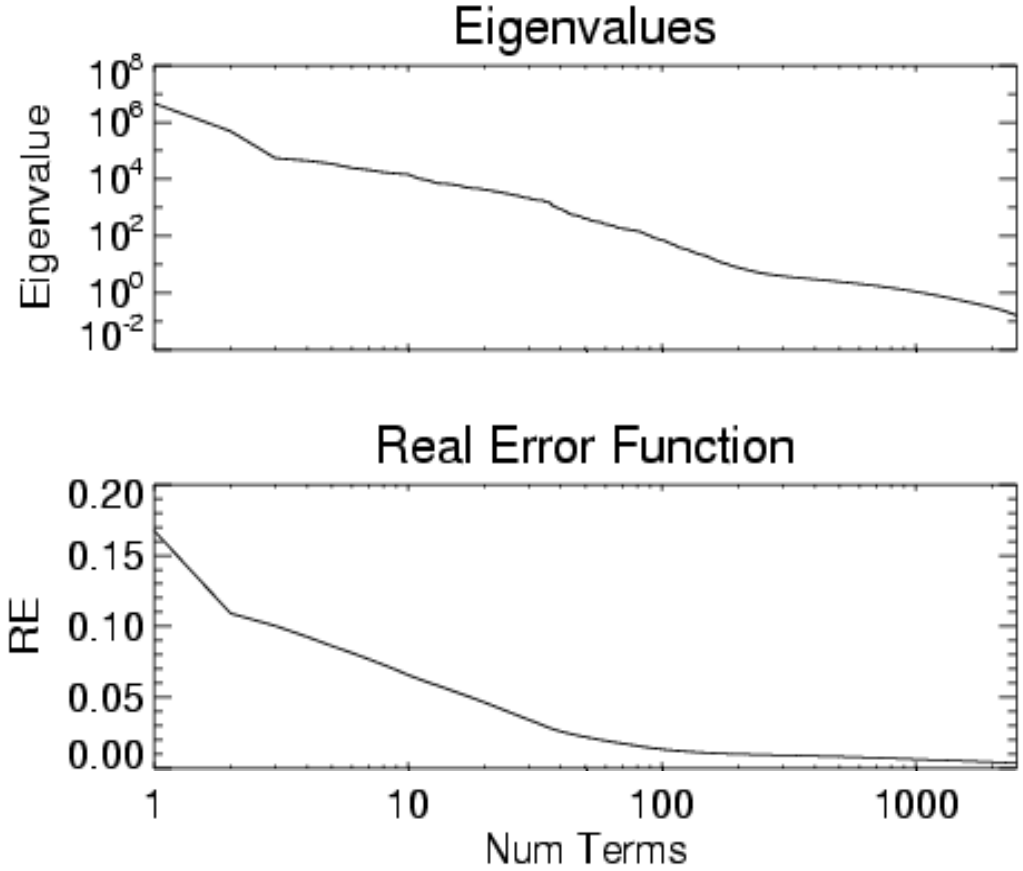
$$IE(k) = \left[\frac{k \sum_{i=k+1}^{n} \lambda_i^0}{t\,n(n-k)} \right]^{1/2}$$

$$XE(k) = \left[\frac{\sum_{i=k+1}^{n} \lambda_i^0}{t\,n} \right]^{1/2}$$

t ≡ number of temporal samples
n ≡ number of spectral elements
Assumes that t > n

- Important: Noise must be random (gaussian) and uniform (i.e., constant standard deviation for all times and spectral elements)

Real Error (RE) example

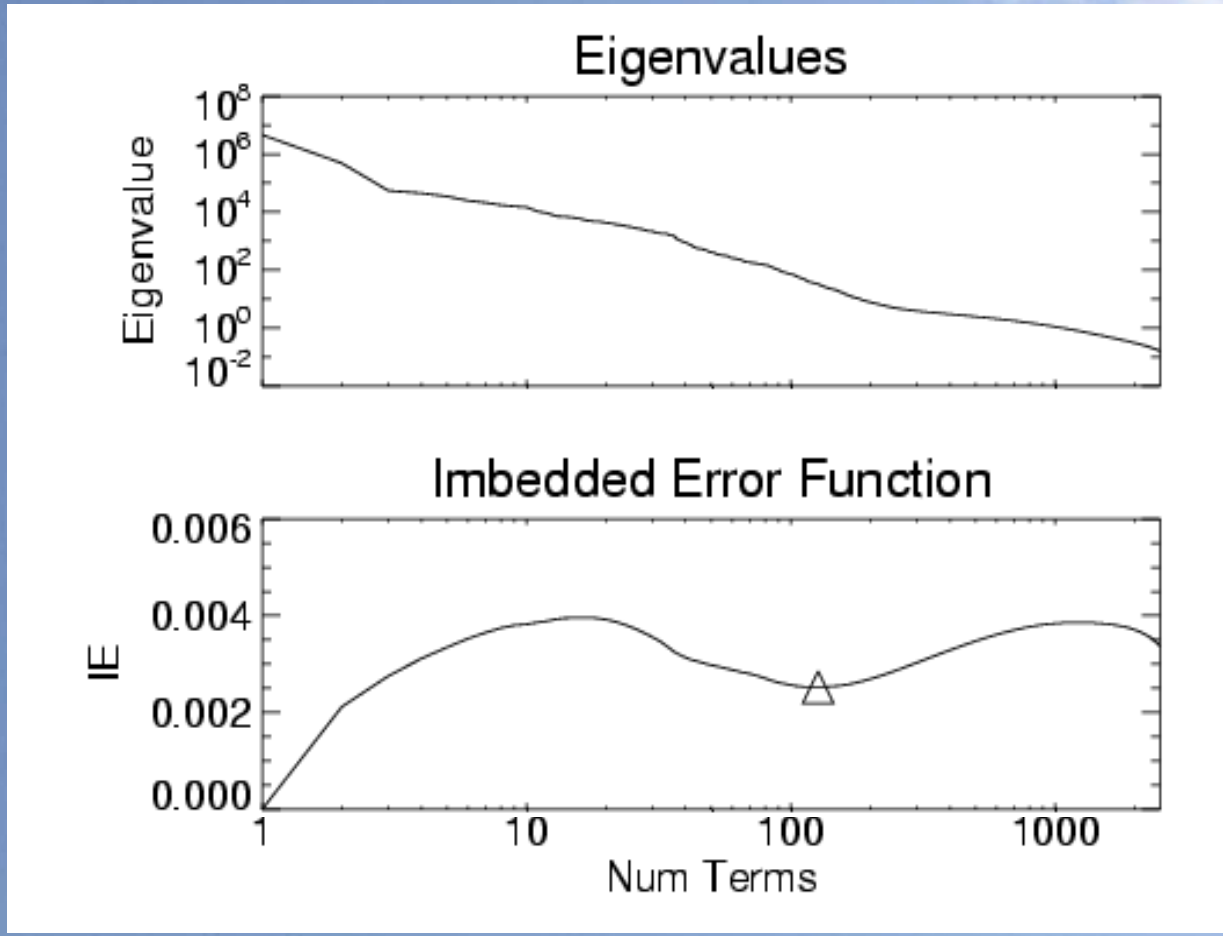


As *k* increases, this implies that more PCs contain a mixture of signal and noise (i.e., more λ^x)

So the RE in the data is smaller because the "noise sphere" has a smaller radius

$$(RE)^2 = (IE)^2 + (XE)^2$$

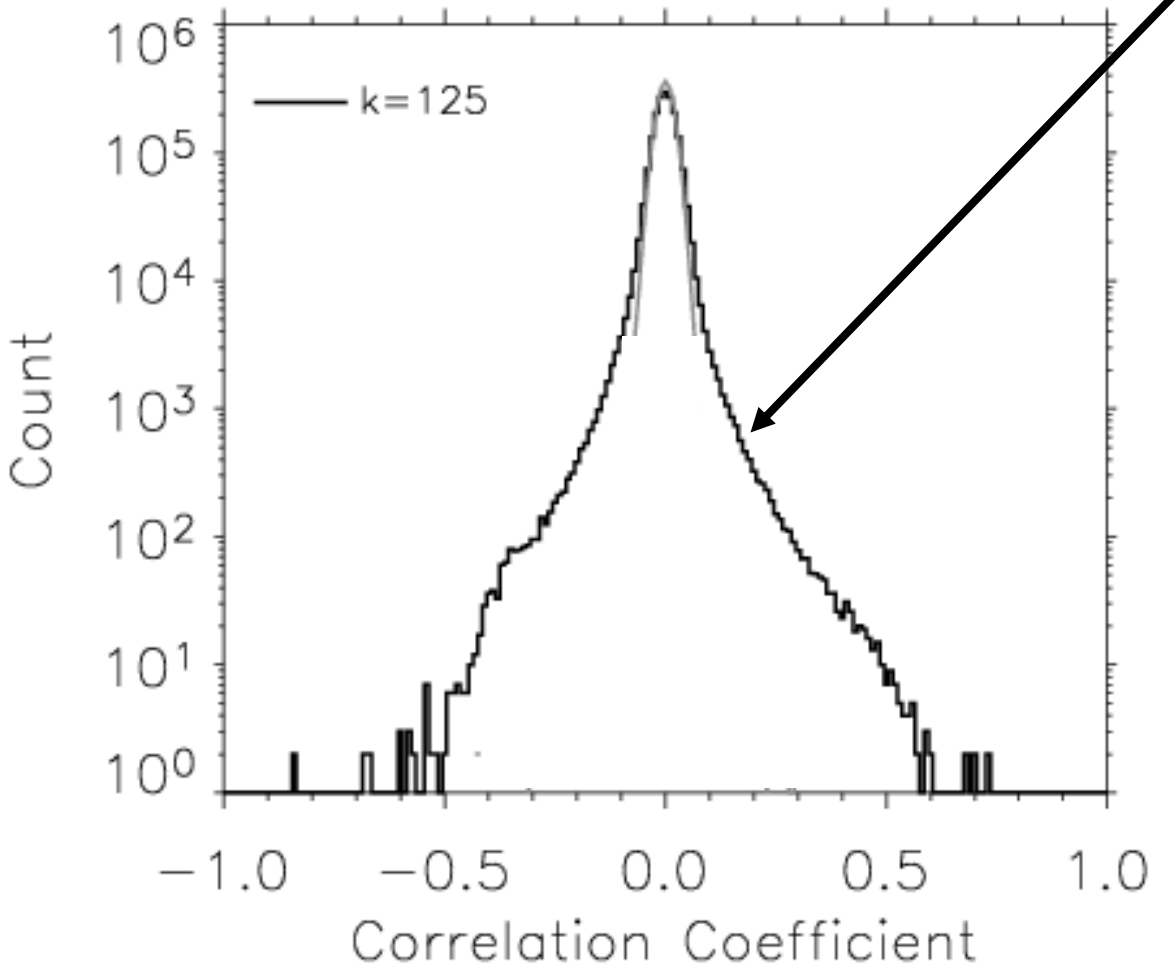
Imbedded Error (IE) example



- Well defined minima at $k = 125$ (this example)
- Is this the ideal value for k ?

Correlation in the extracted data

$(\upsilon_i, \upsilon_j), \forall\ i \neq j$



$k_{IE} = 125$

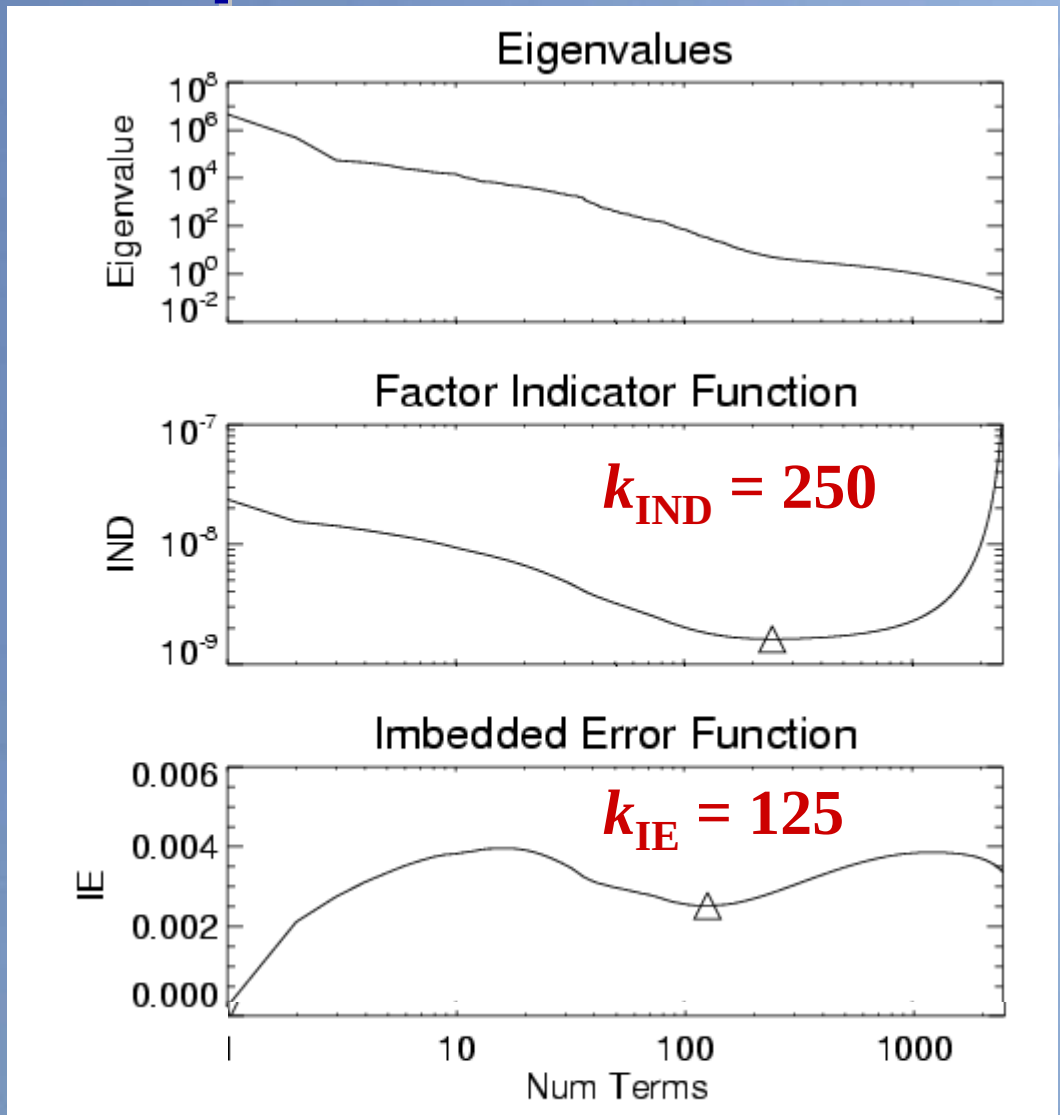
Significant number of spectral pairs have fair correlation when $k = 125$, which implies that spectrally correlated information was lost

Factor Indicator Function (IND)

- Problems with using IE to determine k
  - Occassionally, IE does not have well defined minimum, as IE places too much emphasis on outliers
  - Significant correlation among different pair of spectral elements can remain in XE, implying signal was lost
- Malinowski developed another empirical function, called factor indicator function (IND), which overcomes the limitations of the IE function

$$IND(k) = \frac{RE(k)}{(n-k)^2}$$

Factor Indicator Function (IND) example

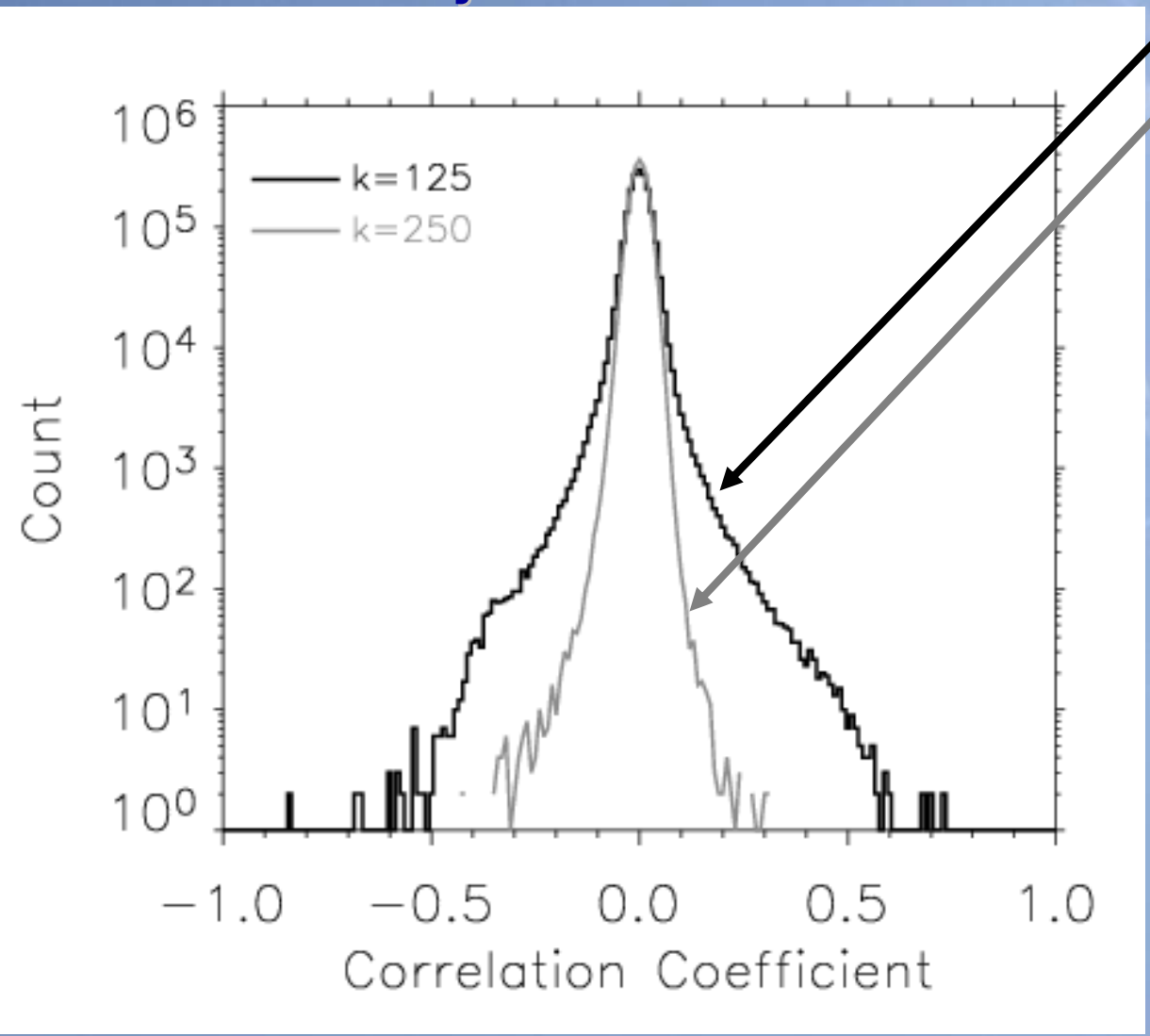


My experience:

$k_{IND} \geq k_{IE}$

Correlation in the extracted data

$(\upsilon_i, \upsilon_j), \forall\ i \neq j$

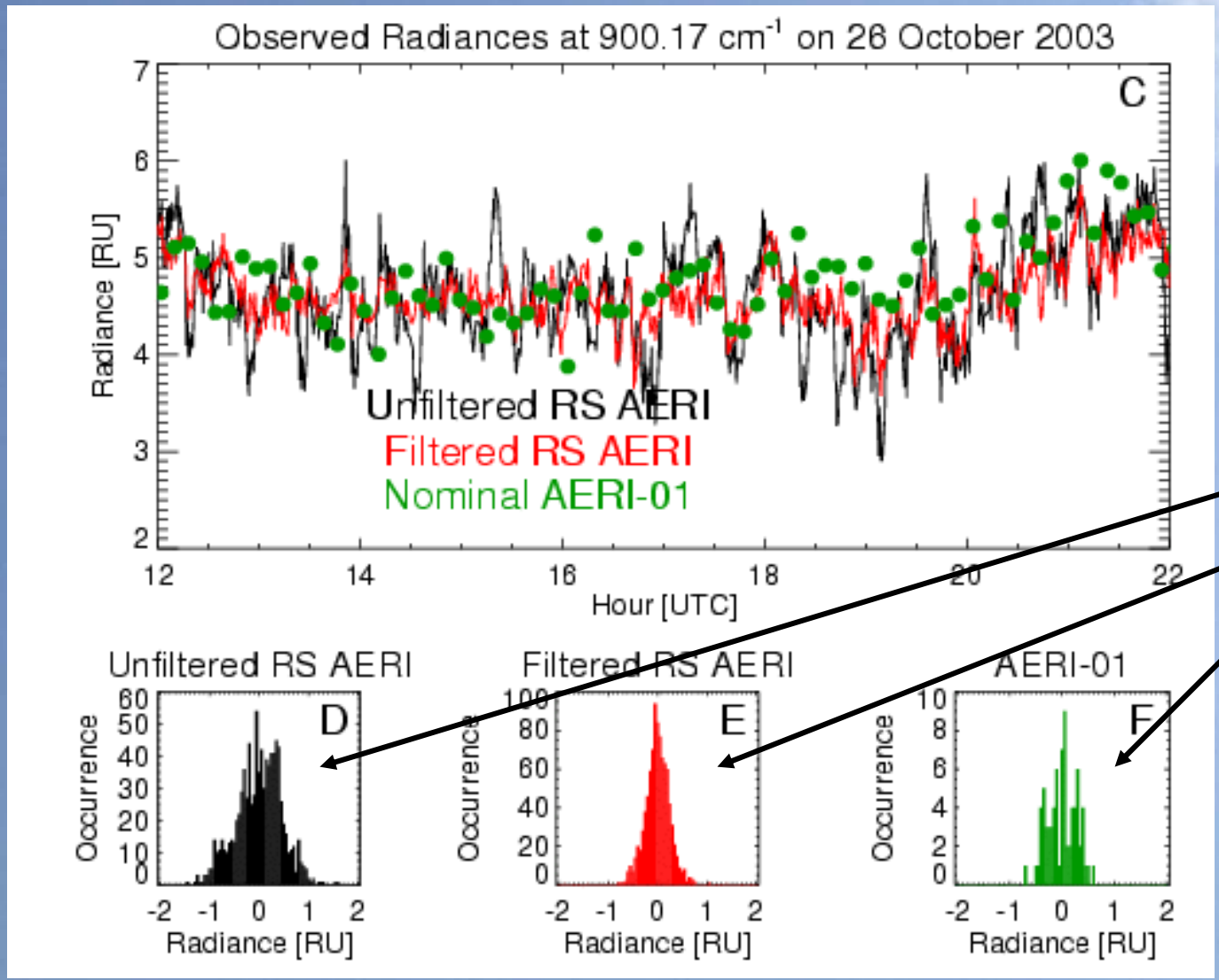


$k_{IE} = 125$

$k_{IND} = 250$

Significantly smaller correlation between pairs of spectral elements when using IND-determined *k* for reconstruction

Impact of noise filter: reduction in clear sky variance



St. Dev.

0.461 RU

0.243 RU

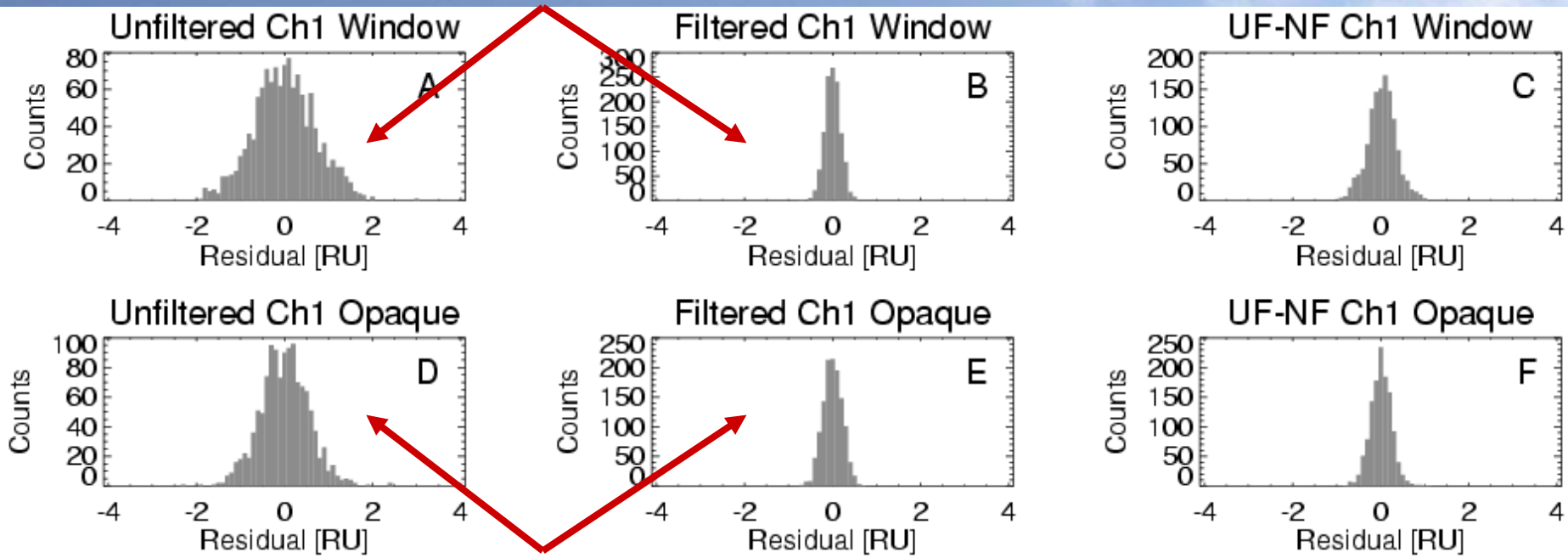
0.278 RU

Reduction of clear sky variance

- NSA AERI system, April 2004

"Window" at 900 cm^{-1}
"Opaque" at 675 cm^{-1}

StDev reduced 4x

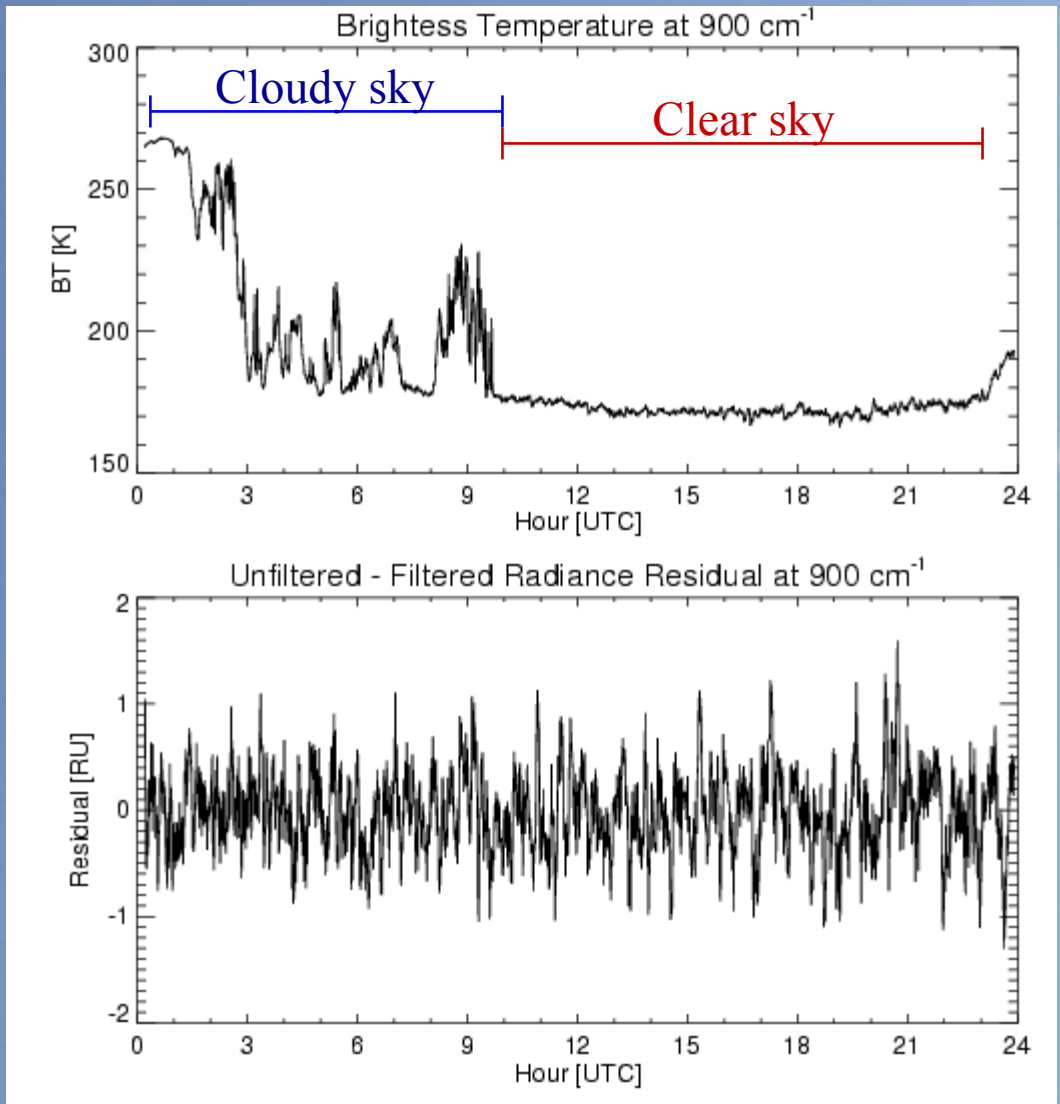


StDev reduced 2.6x

Clear sky, detrended observations

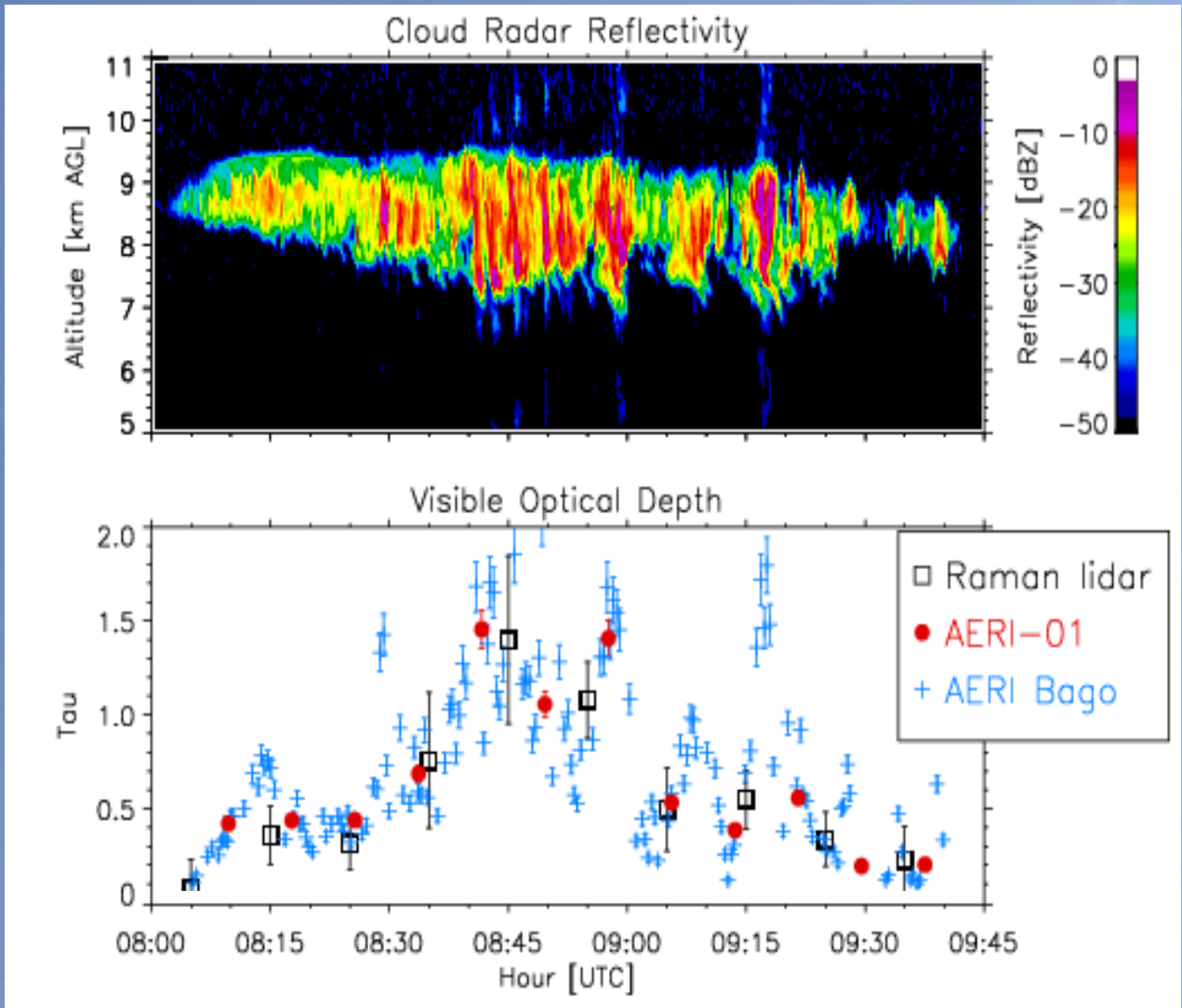
Unfiltered minus filtered residuals

Noise filter in different scenes: an example of clear sky vs. cloudy



No apparent change in the character of the extracted variance (i.e., the unfiltered – filtered radiance)

Benefits of rapid sampling and noise filtering



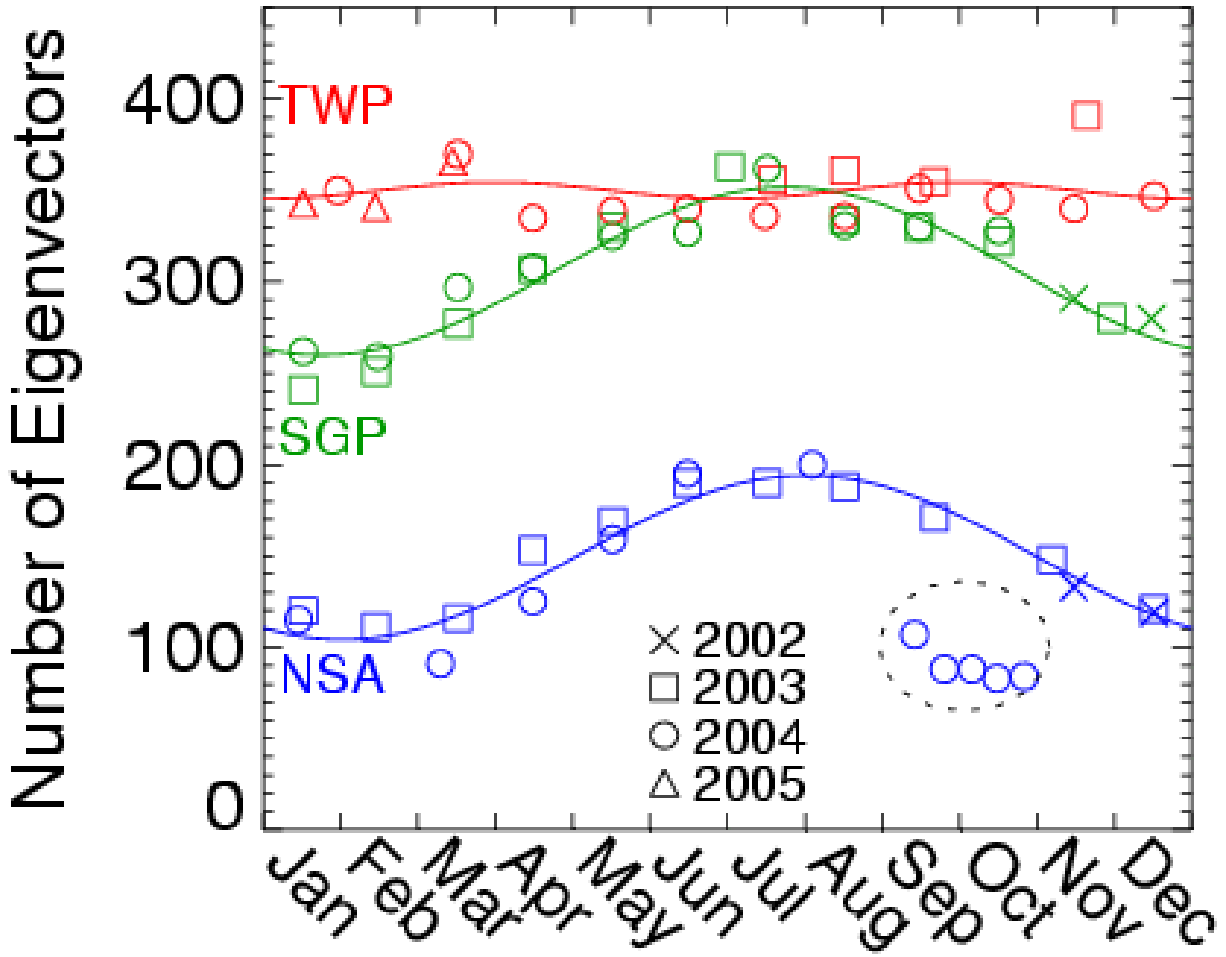
- Able to resolve high temporal features in clouds
- NF reduces uncertainty in retrieved products considerably

Location, seasonal, and instrumental dependence of *k*

- ARM has AERIs in three different environmental regimes (tropics, mid-latitudes, Arctic)
- AERIs collect data continuously
- Two distinct types of AERIs used in ARM

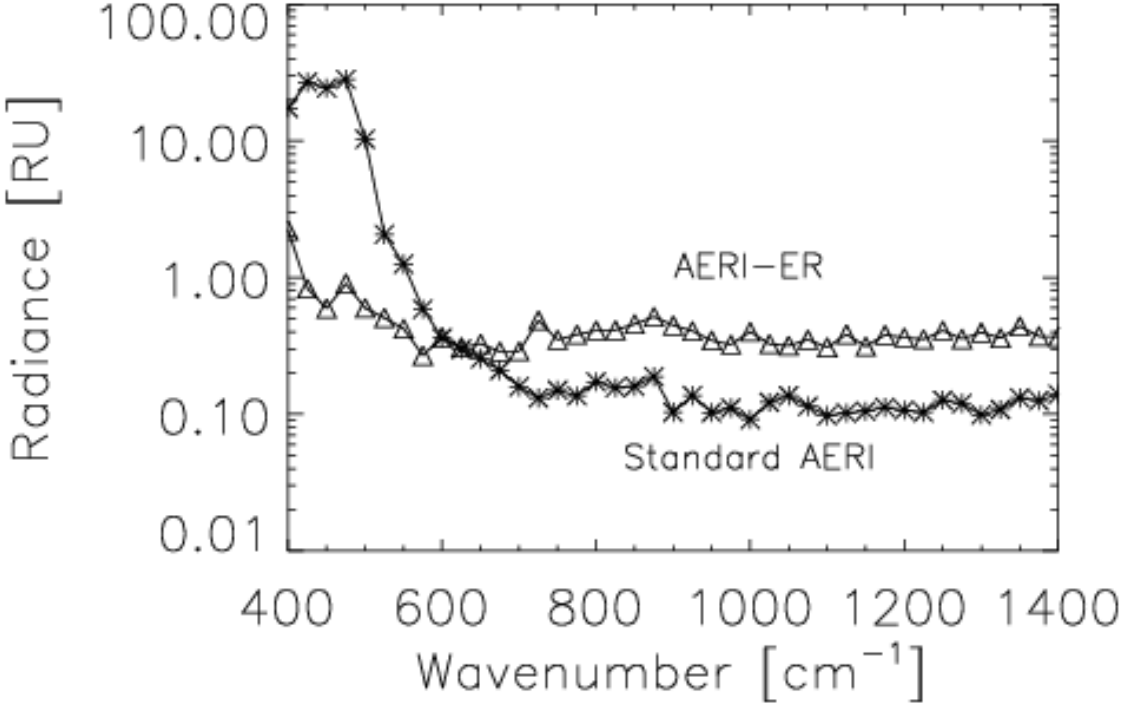
- Questions:
  - Is there any dependence of *k* on the AERI instrument?
  - Is there any seasonal or climatological variability in *k*?

Location, seasonal, and instrumental dependence of *k* (1)



Instrument or location?

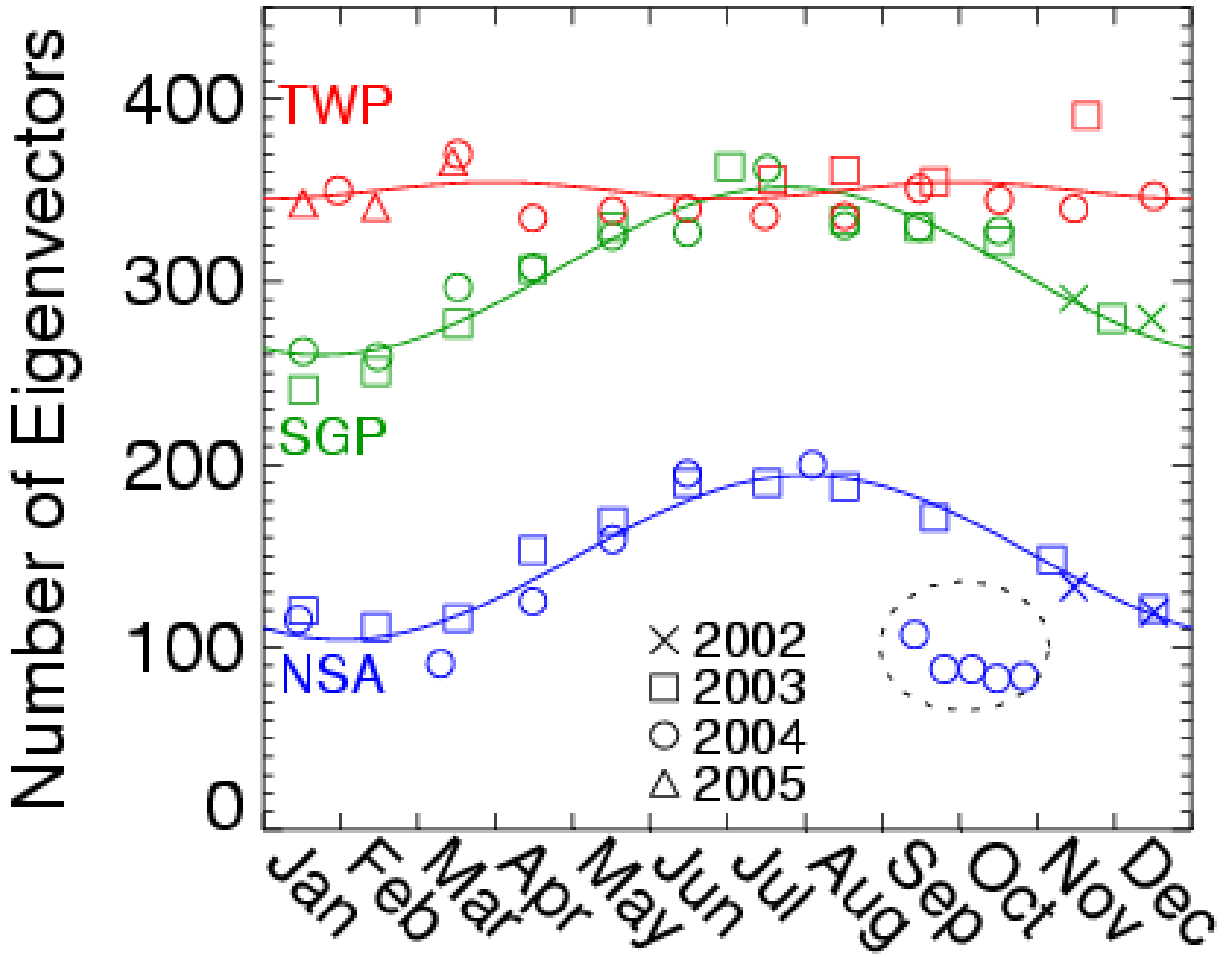
Extended range AERI-ERs at NSA have different noise performance…



- AERI-ER detectors optimized to have improved performance in 15-25 μm region
- Larger "noise sphere"
- Small-scale atmospheric variability may be "lost" in larger noise sphere, resulting in smaller *k*

This likely explains majority of difference between SGP and NSA *k*'s

Location, seasonal, and instrumental dependence of *k* (2)



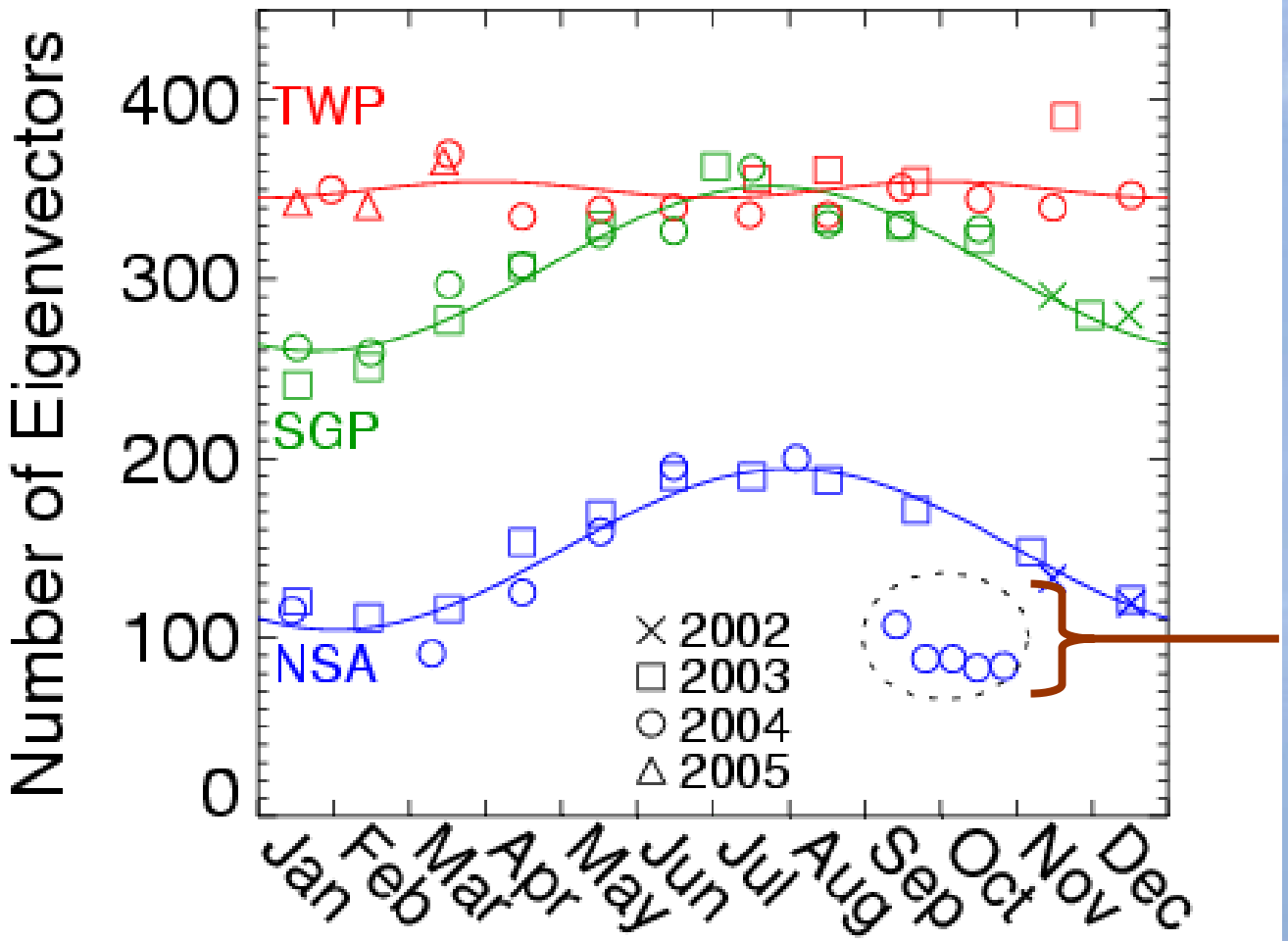
Instrument or real atmospheric seasonal dependence?

Seasonal dependence of *k*: instrument or atmosphere?

- Atmospheric conditions
  - Change from more synoptic conditions in winter to more convective conditions in summer
  - Changes character of clouds, with more broken clouds in summer
  - Increase in *k*
- Calibration equation
  - Summertime has warmer ambient temperatures
  - Delta-T between HBB and ABB decreases, resulting in increased uncertainty due to calibration
  - Increases noise sphere
  - Decrease in *k*

Seasonal dependence of *k*'s likely due to real atmospheric variability

Location, seasonal, and instrumental dependence of *k* (3)



What is this?

Summary

- Factor Indicator Function (IND) method provides a fast and objective way to determine k
  - IND applied to the eigenvalues
- Noise reduction depends on
  - Instrument characteristics
  - Spectral region
  - Instrument location
  - Season
- PCA noise filter reduces random error associated with rapid-sampling, but increased noise level results in real atmospheric variability being “lost”
  - Can recover this signal by averaging raw data in time

Manuscript in press (JTECH)

Noise Reduction of Atmospheric Emitted Radiance Interferometer (AERI) Observations using Principal Component Analysis

D.D. Turner, R.O. Knuteson, H.E. Revercomb,
C. Lo, and R.G. Dedecker

Email: dturner@ssec.wisc.edu