Data Compression – Data User Perspective

Andrew Collard

IMSG @ NOAA/NCEP/EMC

Introduction

- **This talk is:**
  - **An attempt to put what is to come into context:**
    - **So I will not go into too much technical detail**
  - **Wrong (I hope)…**
    - **… in that some of the issues highlighted will have been resolved … and the following talks will tell us how.**

Talk Overview

- Why do we want to compress hyperspectral data?
- What properties would we prefer?
- Overview of hyperspectral spectrum
- Possibilities for compressing data for assimilation
- Spatial Compression
- Discussion

Talk Overview

- **Why do we want to compress hyperspectral data?**
- **What properties would we prefer?**
- **Overview of hyperspectral spectrum**
- **Possibilities for compressing data for assimilation**
- **Spatial Compression**
- **Discussion**

Why do we want to compress hyperspectral data?

- **Hyperspectral infrared satellite sounder observations are comprised of many thousands of channels but information content studies show that they contain only a few tens of pieces of independent information.**
- **There is therefore significant redundancy in these measurements.**
- **We have an interest in compressing these data for two main reasons:**
  - **Efficiency in distributing data**
    - **We would prefer this to be lossless**
    - **See talks by Atkinson, Hultberg**
  - **Efficiency in assimilating data**
    - **This is my focus today**

Talk Overview

- Why do we want to compress hyperspectral data?
- **What properties would we prefer?**
- Overview of hyperspectral spectrum
- Possibilities for compressing data for assimilation
- Spatial Compression
- Discussion

Desired properties

- **Observation Error Characteristics**
  - **We need to be able to derive an observation error covariance matrix that will allow close to optimal assimilation of the measurements.**
  - **Uncertainties in the definition of this matrix should not affect the quality of the analysis**
  - **We should be aware that the total error budget will include forward model error, representivity error, errors due to imperfect bias correction or quality control and non-linearity error.**
    - **Reducing the instrument noise down to very low values will have very little benefit if these other errors dominate.**

Desired properties

- **Forward Modelling**
  - We need to have an efficient and sufficiently accurate forward model plus its adjoint/Jacobians
- **Quality Control**
  - Can the compressed observations be quality controlled sufficiently. In particular can clouds be accurately detected, compensated for or assimilated?
- **Bias Correction**
  - Do we need to develop new bias correction schemes for the observations?
- **Monitoring**
  - In some cases, monitored quantities will be less directly related to meteorological or instrument properties.
- **Robustness**
  - If instrument characteristics change will the compressed data need to be retuned?

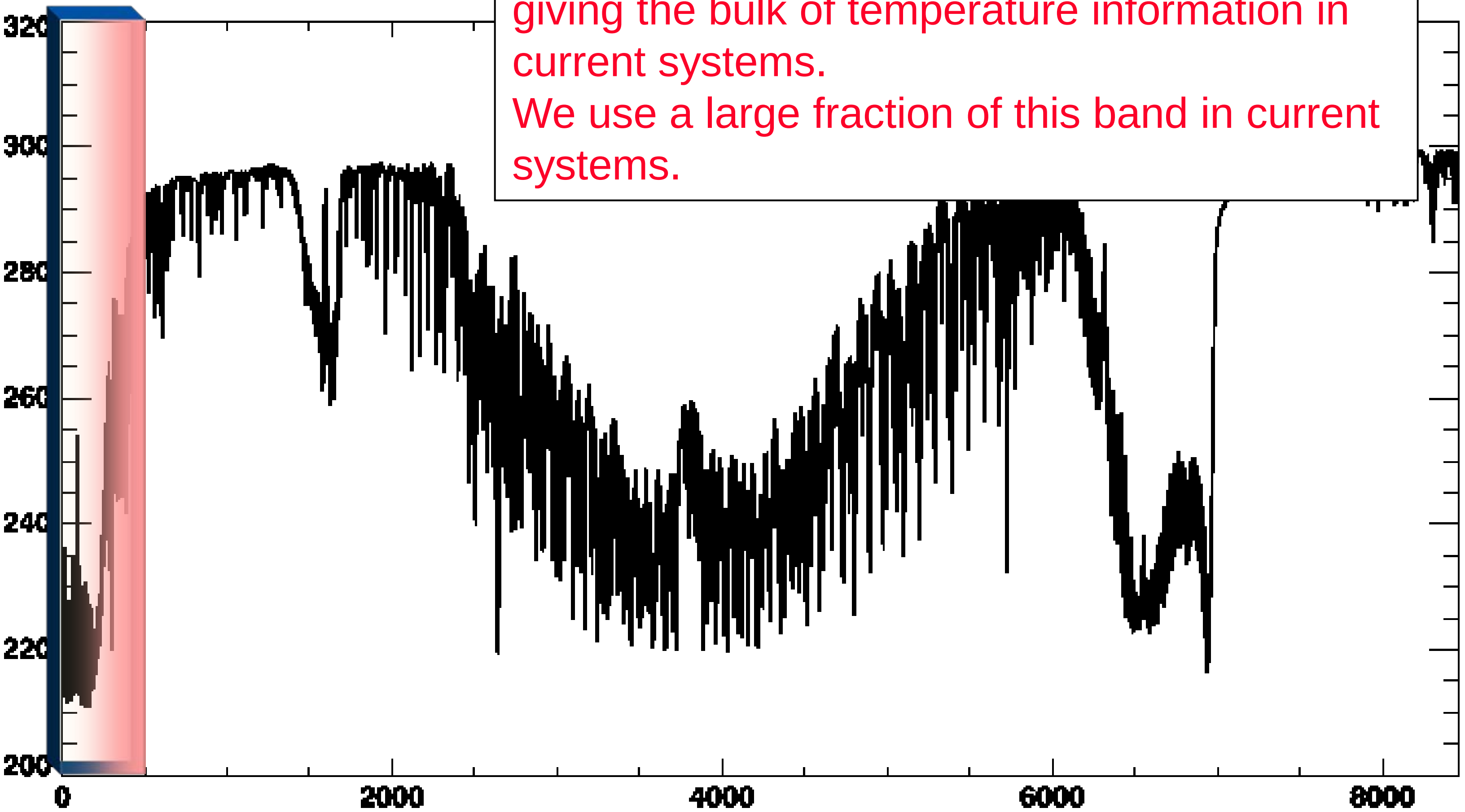
Talk Overview

- Why do we want to compress hyperspectral data?
- What properties would we prefer?
- **Overview of hyperspectral spectrum**
- Possibilities for compressing data for assimilation
- Spatial Compression
- Discussion

Using the IASI Spectrum

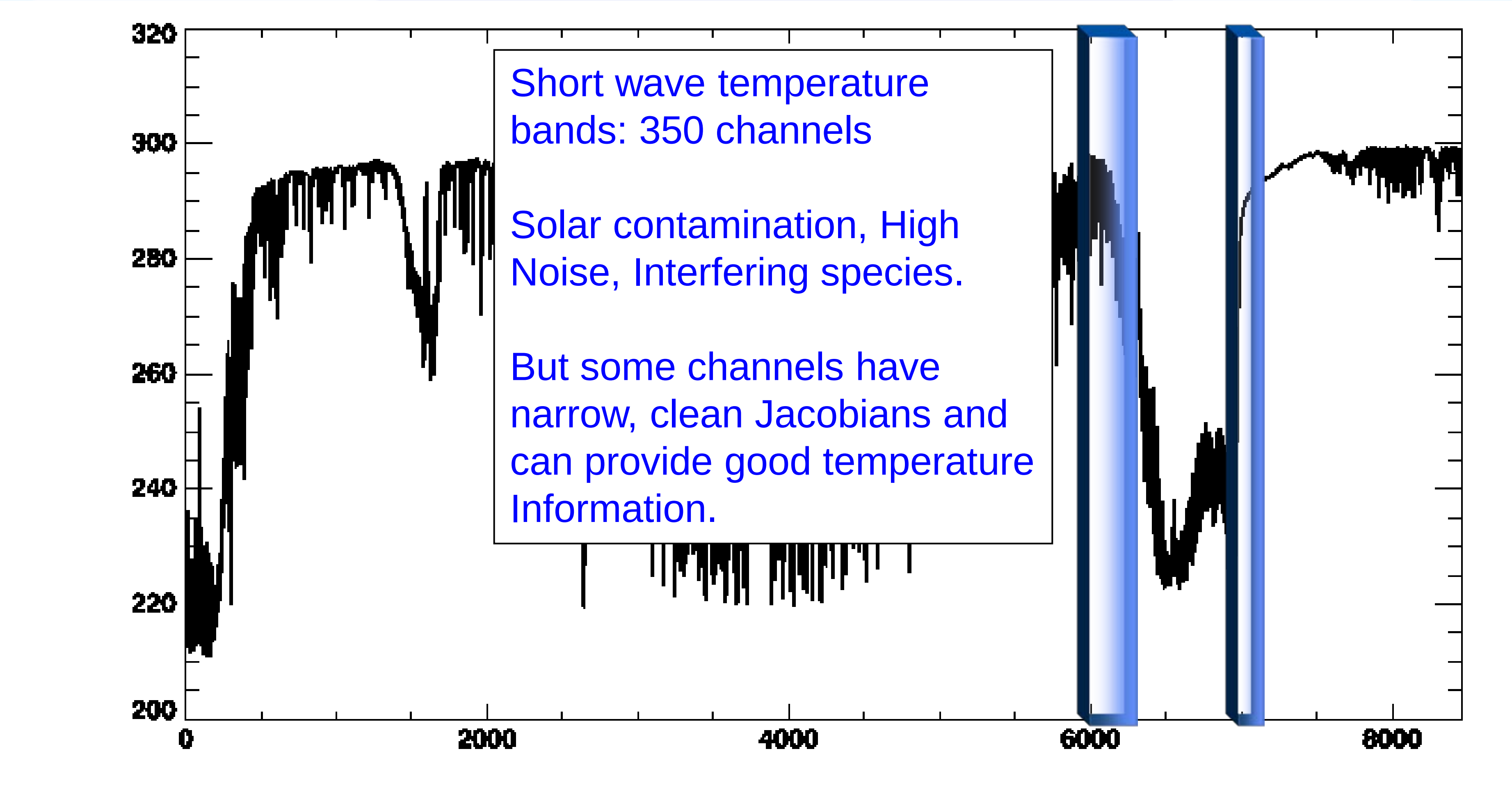
Longwave CO_2 Band

1st 500 channels cover the 15µm CO_2 band giving the bulk of temperature information in current systems.
We use a large fraction of this band in current systems.



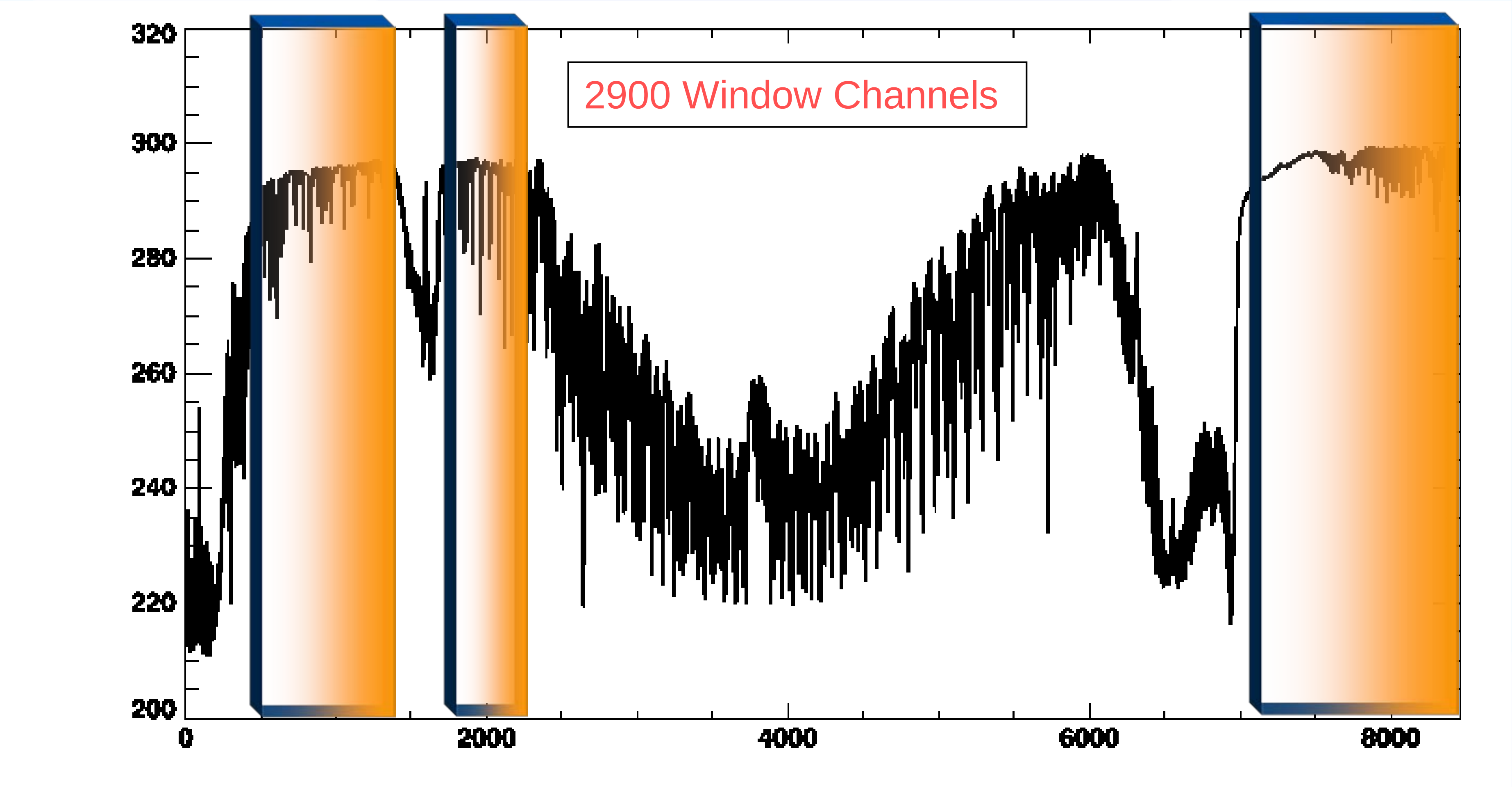
Using the IASI Spectrum

Shortwave CO_2 Band



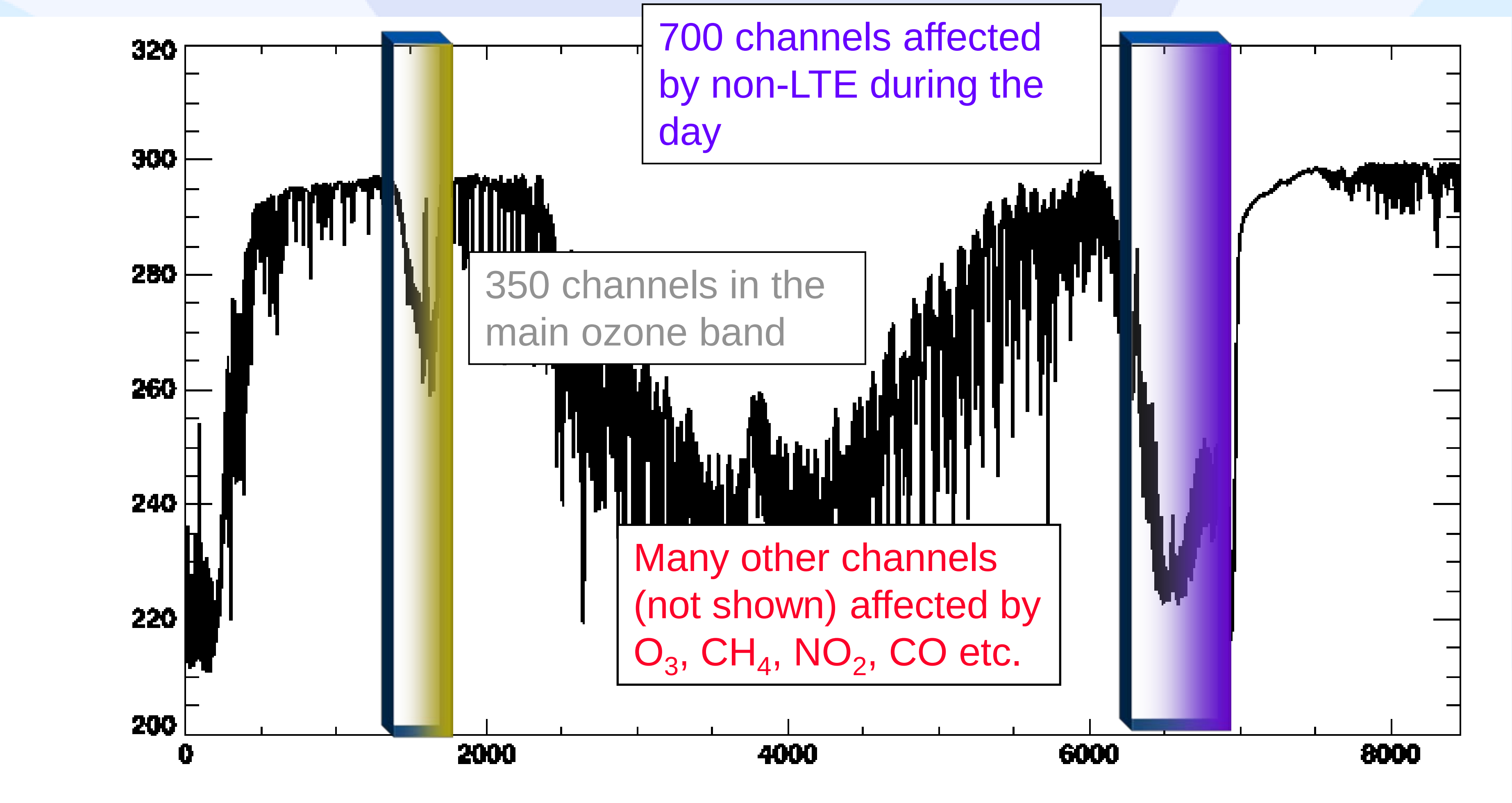
Using the IASI Spectrum

Channels Primarily Sensitive to the Surface



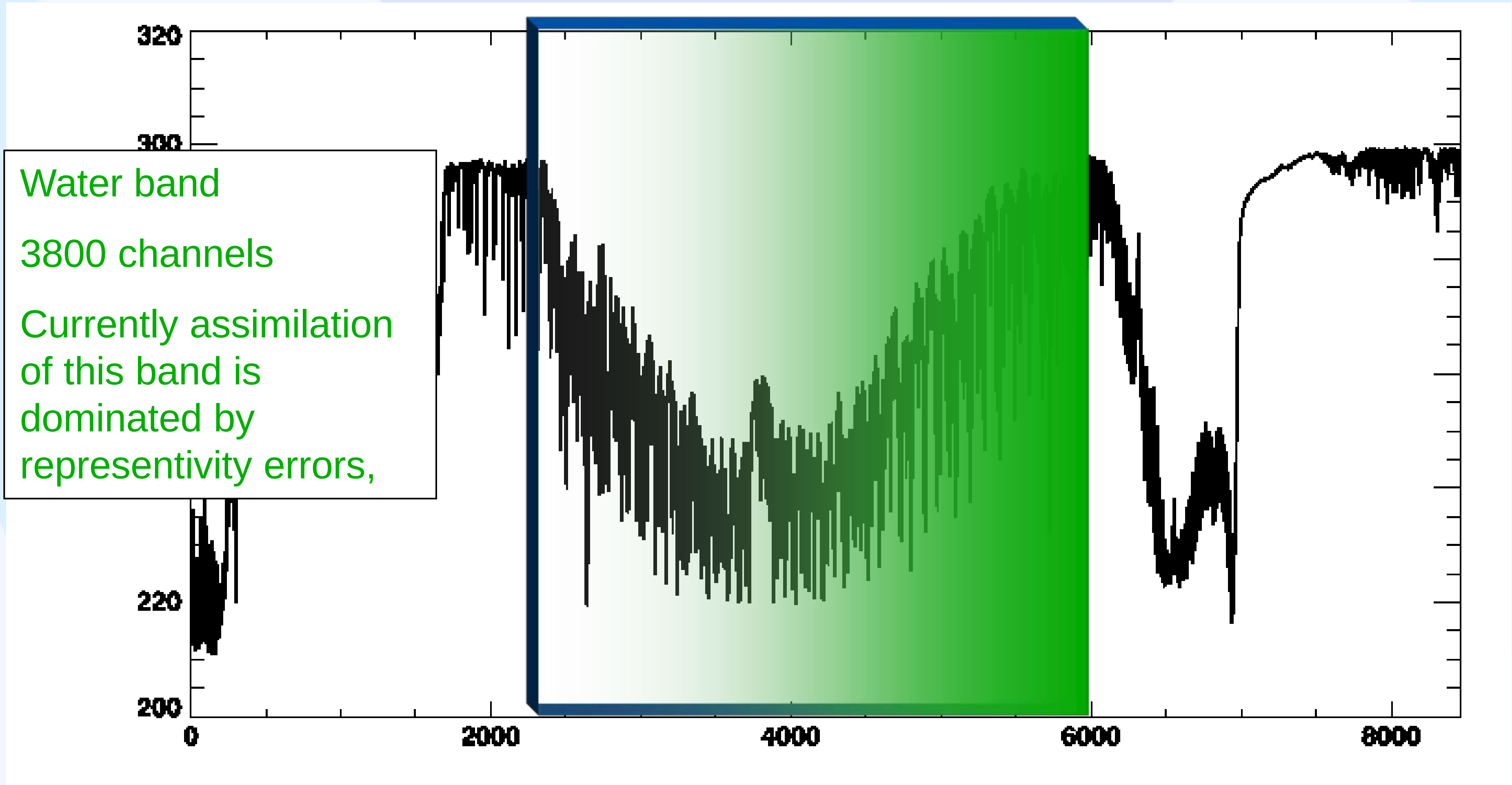
Using the IASI Spectrum

Trace Gases and RT Challenges



Using the IASI Spectrum

The 6.3μm Water Band



Water band

3800 channels

Currently assimilation of this band is dominated by representivity errors,

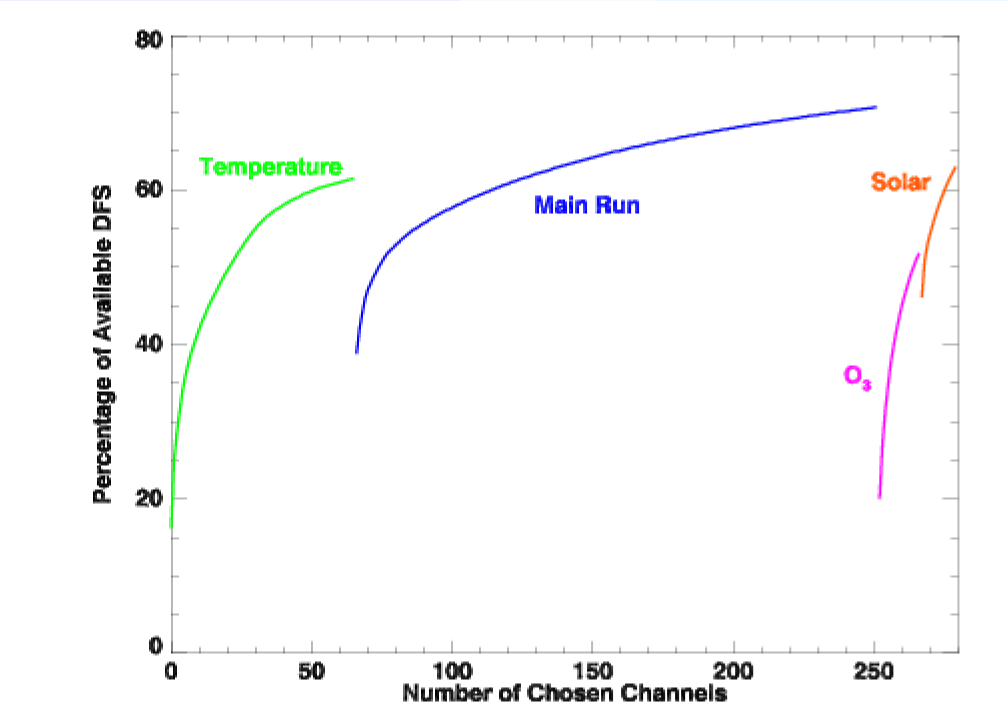
Talk Overview

- Why do we want to compress hyperspectral data?
- What properties would we prefer?
- Overview of hyperspectral spectrum
- **Possibilities for compressing data for assimilation**
- Spatial Compression
- Discussion

Possibilities for compressing data for assimilation

Channel Selection

- **The bulk of the information in the spectrum can be represented by a subset of a few hundred channels.**
- **Studies include work by Rodgers (2000), Rabier et al. (2002), Collard(2007), Ventress and Dudhia (2013)**
- **Lossy**



Possibilities for compressing data for assimilation: Spectral data compression with PCA*

The complete spectrum can be compressed using a truncated principal component analysis (e.g. 200PCAs v 2300 rads)

Leading eigenvectors (200,say) of covariance of spectra from (large) training set

Mean spectrum

$$\mathbf{p} = \mathbf{V}^{\mathrm{T}}(\mathbf{y} - \overline{\mathbf{y}})$$

Coefficients

Original Spectrum

- To use PCs in assimilation requires an efficient RT model to calculate PCs directly
- PCs are more difficult to interpret physically than radiances

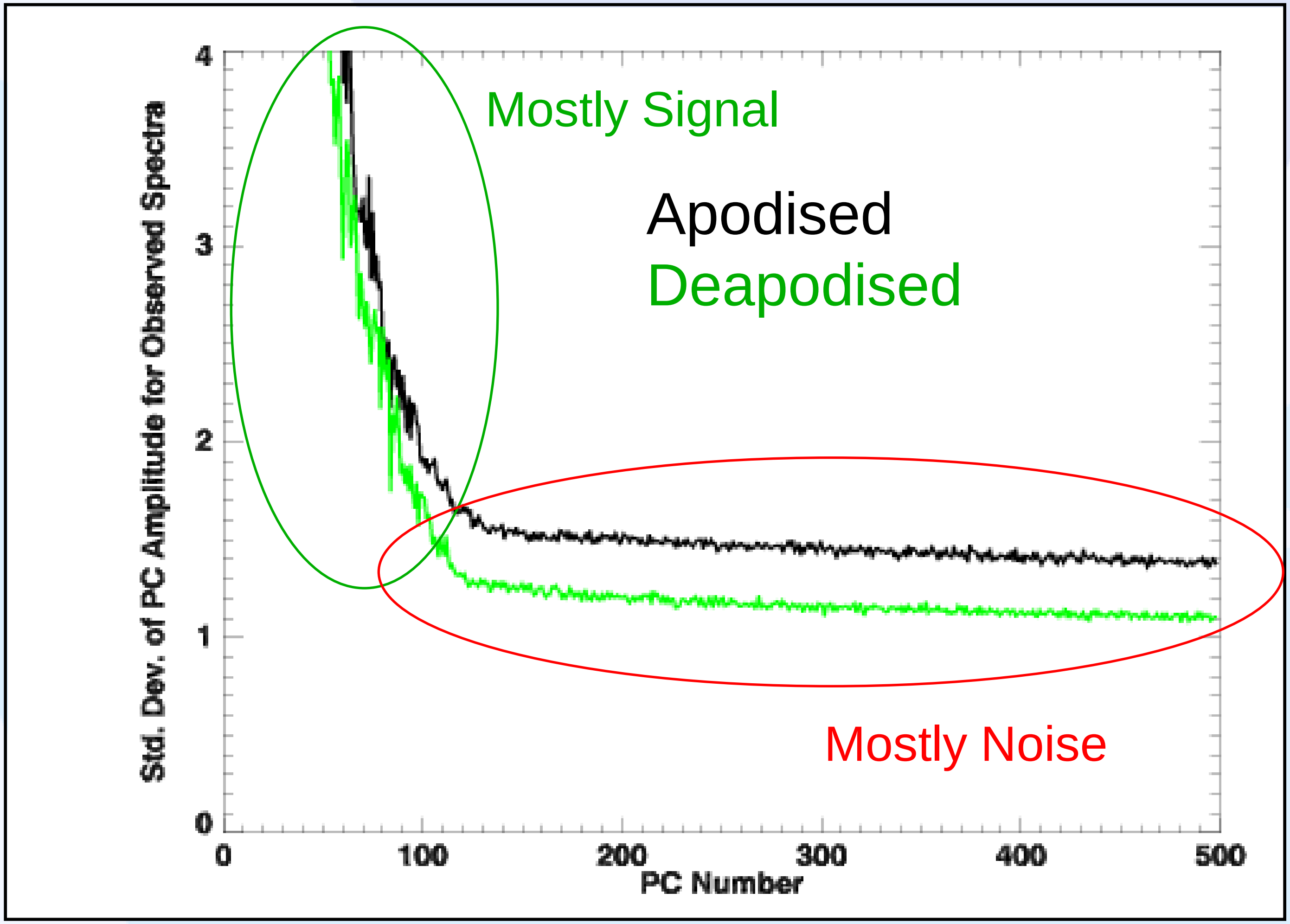
N.B. This is usually performed in noise-normalised radiance space

This allows data to be transported efficiently

***Principal Component Analysis**

See talks from Matricardi, McNally and Bormann

PC amplitudes



Std.Dev. of PC amplitudes for real observations

EOFs from Nigel Atkinson

Loop of Jacobians of PCs

Advantages with direct assimilation of PC amplitudes

- **Components containing signal may be preferentially chosen, eliminating noise up to 90% of the noise.**
  - **In theory optimal assimilation of all channels will provide more information that the leading PCs, but requires an accurate error model for the noise contained in the rejected PCs.#**
- **Instrument noise model is relatively simple.**
  - **In theory it is the identity matrix if the initial radiances are correctly noise-normalised.**

Issues with direct assimilation of PC amplitudes

- **Jacobians of principal components are more non-localised than raw channels**
  - **Signals are less separable (in the vertical and between temperature and humidity etc.)**
  - **It is harder to have a cloud detection scheme where only “channels” above the cloud/surface are used.**
  - **We are therefore limited to “hole hunting” or to doing assimilation of cloudy radiances**
- **the appropriate selection of sub-bands for the PC calculation**
- **Requires new RT modelsIt is possible that the degree of non-locality can be reduced with , but these have been developed for some years now**

- **Need to develop new quality control, bias correction and monitoring methods**

Possibilities for compressing data for assimilation: Spectral data compression and de-noising

The complete spectrum can be compressed using a truncated principal component analysis (e.g. 200PCAs v 2300 rads)

Leading eigenvectors (200,say) of covariance of spectra from (large) training set

Mean spectrum

Reconstructed spectrum

$$\mathbf{p} = \mathbf{V}^{\mathbf{T}}(\mathbf{y} - \bar{\mathbf{y}})$$

$$\mathbf{y}_{\mathbf{R}} = \bar{\mathbf{y}} + \mathbf{V}\mathbf{p}$$

Coefficients

Original Spectrum

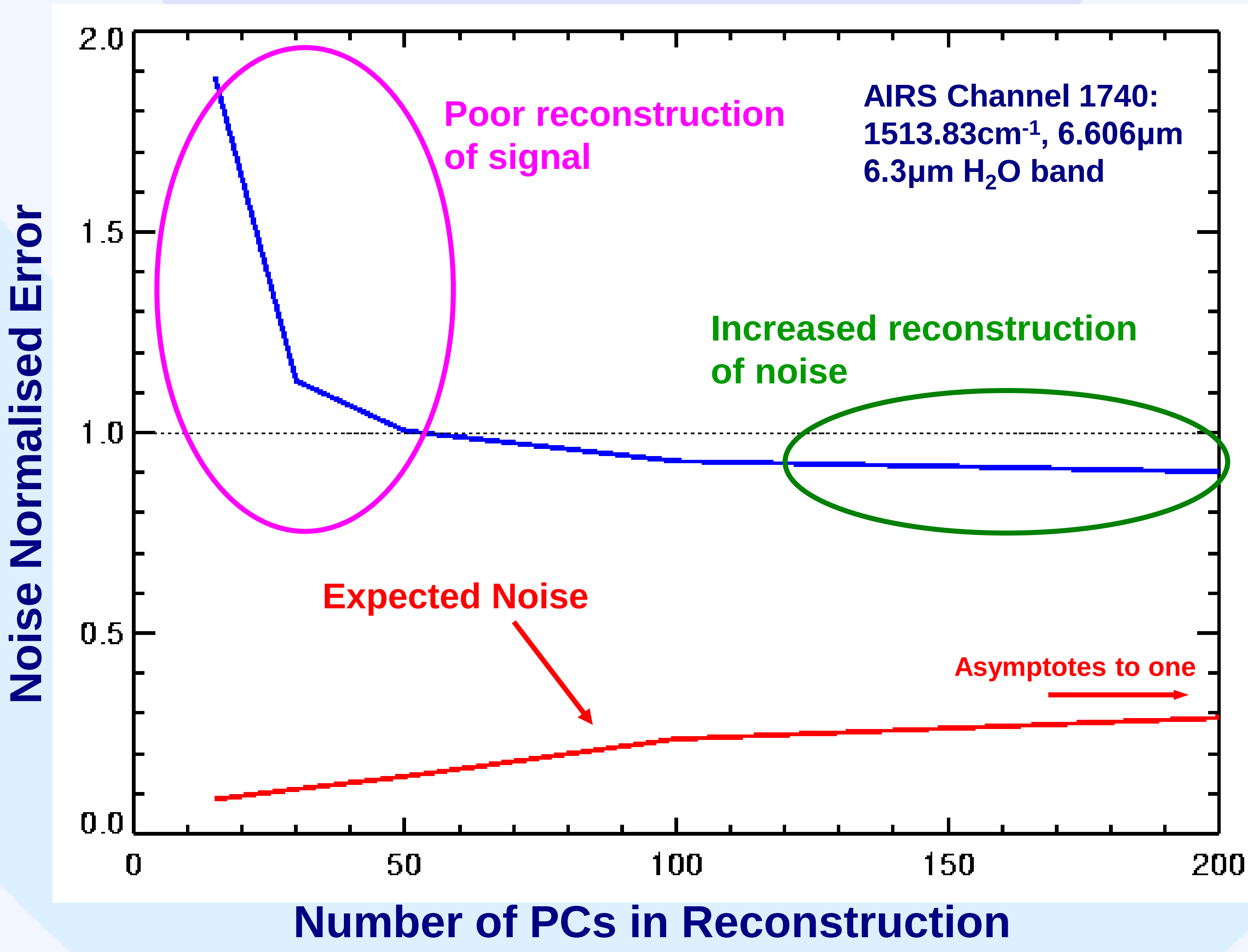
N.B. This is usually performed in noise-normalised radiance space

Each reconstructed channel is a linear combination of all the original channels and the data is significantly de-noised.

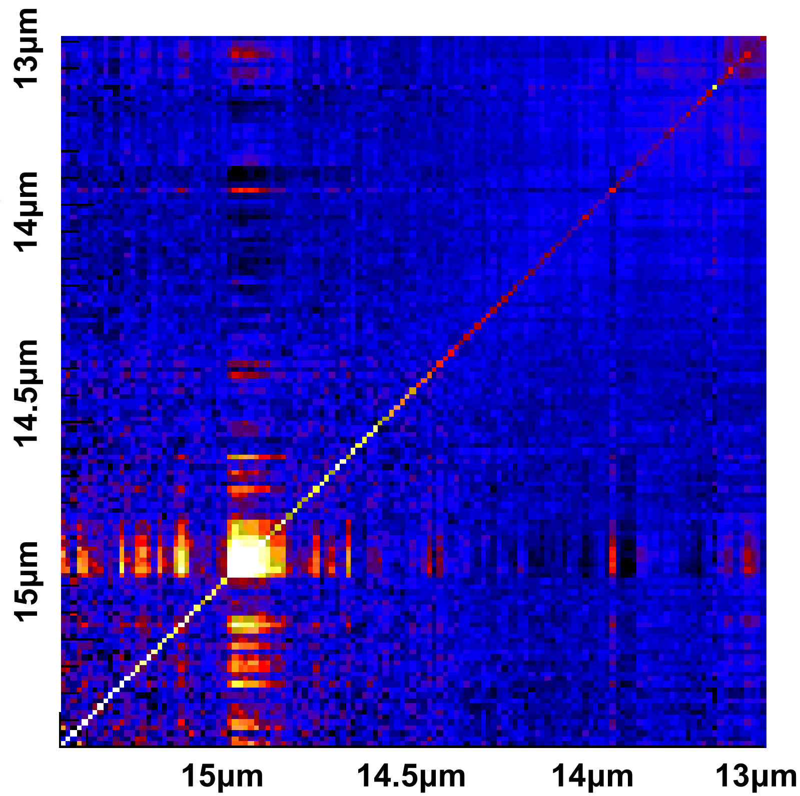
If *N* PCs are used all the information is contained in *N* reconstructed channels (theoretically)

See talk from Hilton

Reconstruction Errors



A look at Reconstructed Radiances' Errors

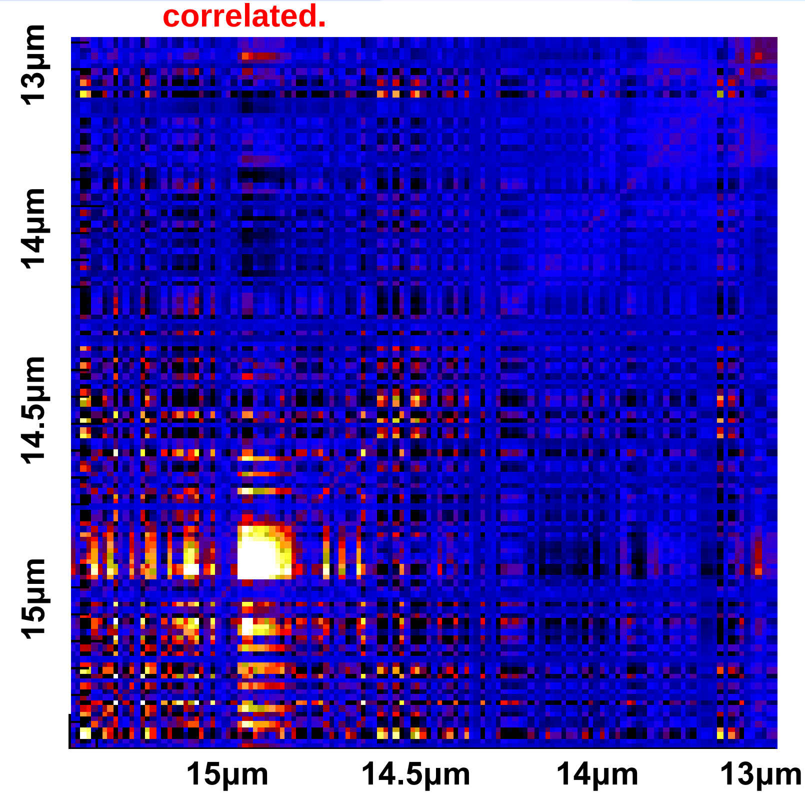


Original Radiances

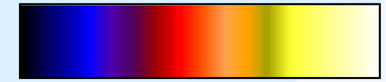
Instrument noise is dominant and diagonal. Correlated noise is from background error

Reconstructed Radiances

Instrument noise is reduced (std. dev. Is approximately halved) but has become correlated.



Covariances of background departures for clear observations in 15µm CO_2 band



Advantages of assimilation of reconstructed radiances

- **Reconstructed radiances look very similar to the original radiances …**
- **… except the noise on each channel is significantly reduced.**

Issues with assimilation of reconstructed radiances

- **The instrument noise is complicated.**
  - **If we want to assimilate all the information in the reconstructed spectrum the observation error covariance matrices may become ill-conditioned**
- **If we treat as a proxy for real radiances we have to allow for an extra "reconstruction error" term …**
- **… but if we correctly forward-model the reconstructed radiances the observation operator is as complicated as for the principal components themselves**

Possibilities for compressing data for assimilation

- **Retrievals**
  - **Compress the spectrum in a retrieved vector of a few tens or hundreds of state-vector elements**
  - **Still requires multiple calls to radiative transfer code …**
    - **… but this can be done outside of the critical timeline**
  - **A priori information is included in the product**
  - **The correct observation operator is the averaging kernel which will vary from observation-to-observation (and should therefore be communicated with the observation)**
  - **The correct observation error will be highly correlated and will vary from observation-to-observation (and should therefore be communicated with the observation)**
  - **See talk from Migliorini**

Spatial Compression

- **Super-Obbing:**
  - **Average background departures within a small spatial area to reduce noise**
    - **This is particularly advantageous when the bulk of the observations are not assimilated because of thinning.**

- **For upcoming geostationary hyperspectral sounders, derived wind products may serve as a spatially and temporally compressed form of radiance data.**

Discussion

- **We need to choose a form for the assimilation of hyperspectral data that preserves as much information as we are reasonably able to use**
- **So we must also consider whether we are able to use the compressed product to its full potential or whether other error sources make this difficult.**
- **These include**
  - **Forward and representivity error (e.g. we can reduce the noise in the 6.3µm water band to <0.2K, but the representivity error is still > 2.0K)**
  - **QC issues**
  - **Forward model complexity**
  - **Non-linearity error**

Thank you