Data Quality Control and Observation Error Estimation

Martin S Lohmann

Outline

- Quality control (QC) and error estimation.
  - Introduction and principles
  - QC techniques

- Radio Occultation QC (general CDAAC approach).

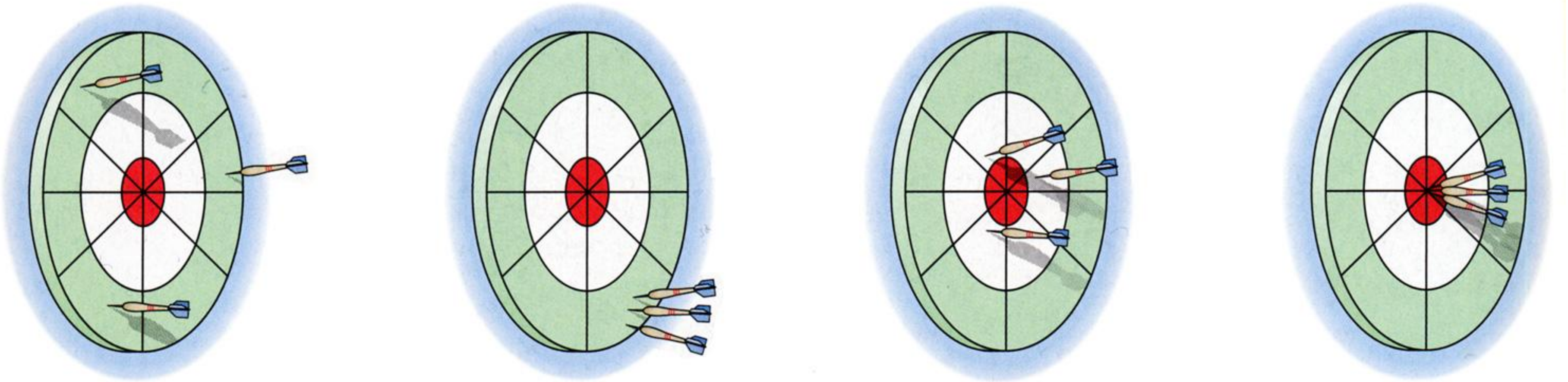
- RO error estimation.
  - Error estimation strategies
  - Dynamic error estimation
  - Statistical optimization

What is good data?

DEFINITION:

Accuracy: The measure of how close the result of the experiment comes to the “true” values

Precision: The measure of how exactly the result is determined without reference to any “true” value



(a) Low accuracy
Low precision

(b) Low accuracy
High precision

(c) High accuracy
Low precision

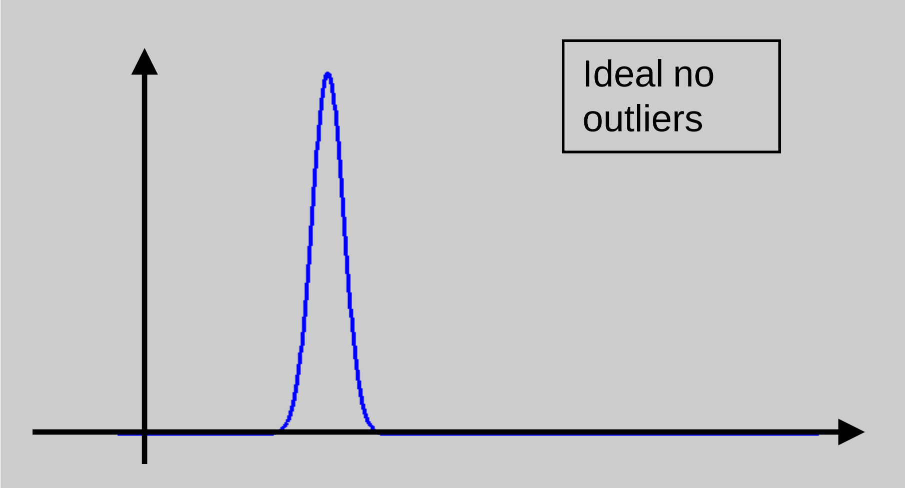
(d) High accuracy
High precision

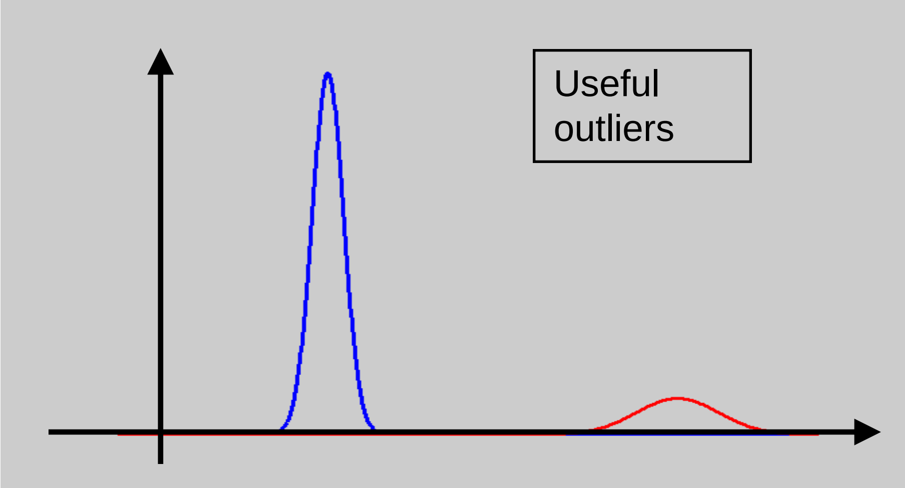
J. Anderson

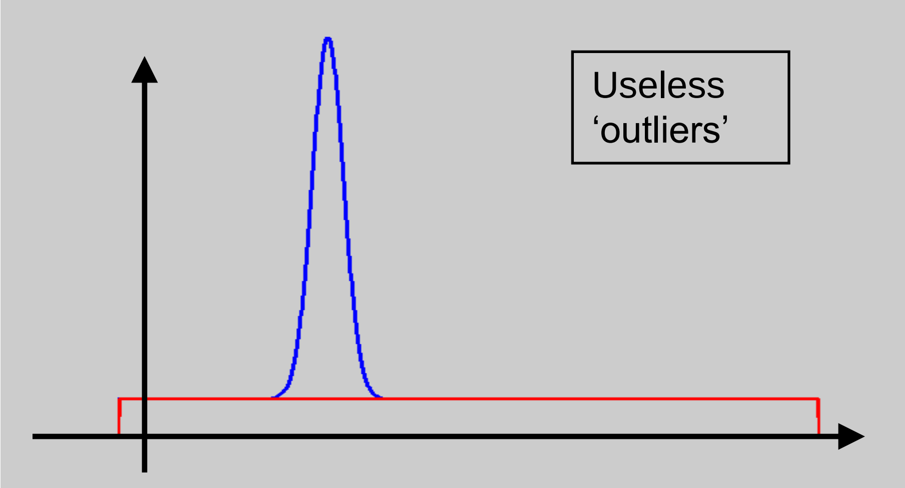
Introduction

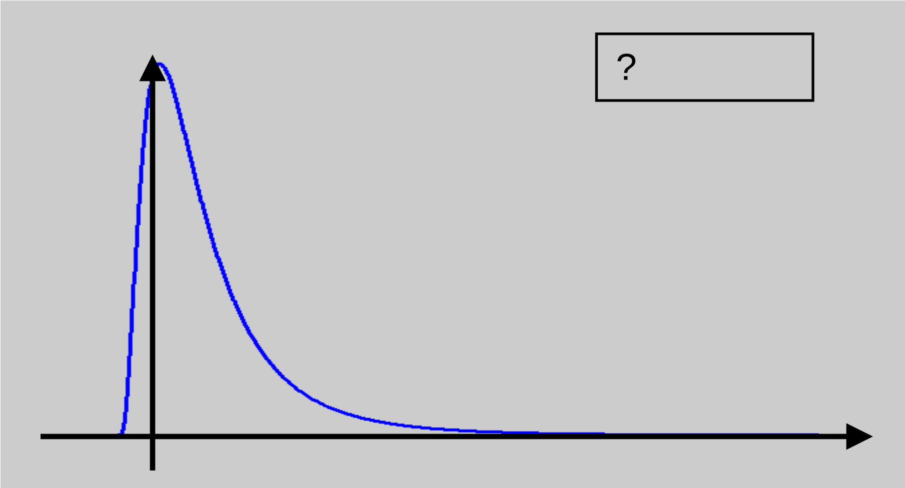
- Why do we care?
- An observation is only useful if the minimum precision and accuracy can be assessed (worst case performance must be known).
- Any data provider must provide reliable error estimates to the users if he wants anybody to use his data.
- Any instrument will under some non-ideal conditions provide data with an accuracy and/or precision which is considerably lower than under ideal conditions (uncorrelated with the truth).
- The purpose of QC is to remove these data.
- Sometimes data obtained under non ideal conditions can still be used if the precision and accuracy of these observations can be estimated.

Example of distribution with outliers









Ideal QC and error estimation

- The QC should only remove data that do not convey any information about the property we are trying to measure. These observation are referred to as ‘bad’.
- Remaining data should be assigned appropriate errors characteristics which may depend on the measurement conditions. These observation are referred to as ‘good’.
- This allows the user to do their own quantity quality trade-off.
- Problems:
  - How do we know if an observations is ‘bad’?
  - How do we determine the mean and standard deviation?
  - When are outliers useful?

When are outliers useful?

- Different users have different requirements.
- Climatologists are mostly concerned about accuracy and particularly long term drift.
- Operational centers are concerned about erroneous observations and observations for which the standard deviation are under estimated.
- Operational centers prefer to do their own complementary QC to decide what is useful or not (advanced user). (see lecture ‘Practical issues related to the assimilation of GPS radio occultation data’ by Sean Healy.
- Individual researchers prefer to get data they can trust (‘simple’ user).

How can error estimates be used in NWP?

- Basically error estimates are used in data assimilation to weight the observations relative to the background.
- Too small error estimates may degrade a forecast, therefore error estimates should rather be to big than too small.
- Important to assign different errors for different observation conditions otherwise the centers will use worst case error estimates which will give a smaller impact.
- In practice, if we are within 50% we are doing good (John Derber Personal communication).

For details and more information see lecture:

‘Introduction to atmospheric data assimilation: Basic concepts and Methodologies’ by Xiaolei Zou.

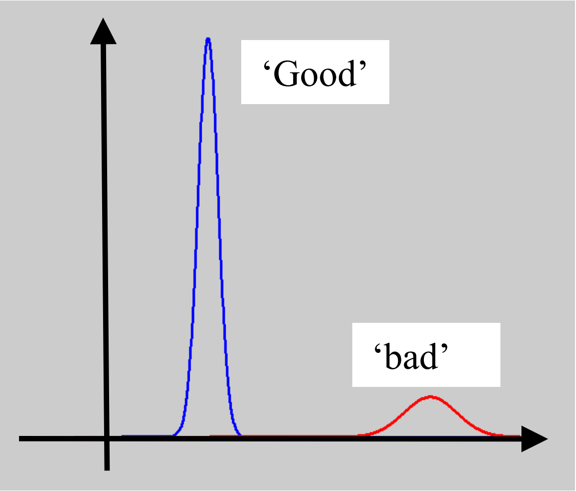
How can we know if an observation is ‘bad’?

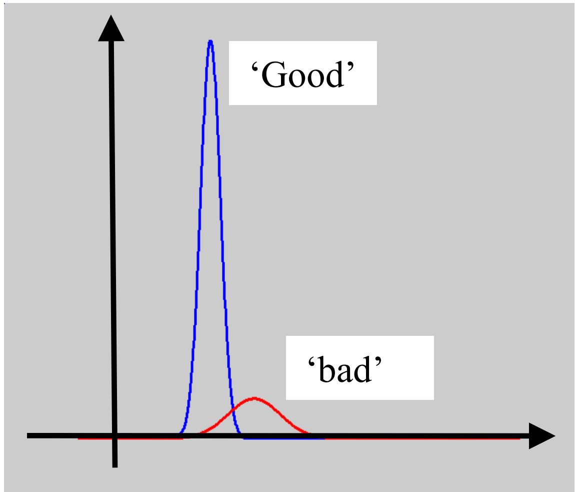
1. Look for ensemble outliers (remove the tail of the distribution).
2. Compare to other data sources e.g. a climatology or reanalysis.
3. Data based QC.
4. Manual inspection.
5. Each technique has it own advantages disadvantages.

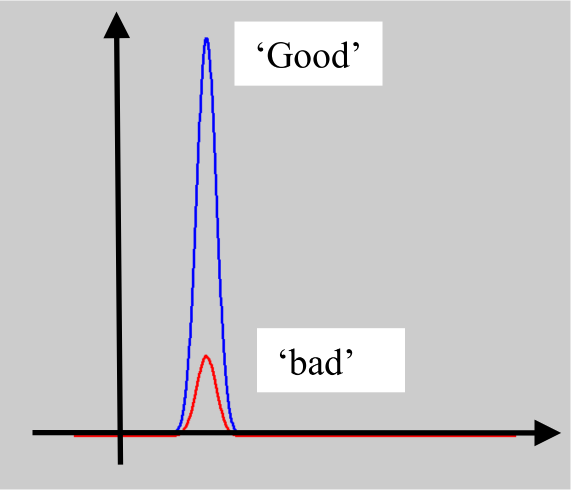
Removal of ensemble outliers (1)

- Basic assumption: The ‘good’ observation belong to some ‘well’ behaved pdf. distribution which do not overlap with the pdf. distribution of the ‘bad’ observations.

- Even if there is some overlap between the two distribution the largest outliers can still be removed.







Removal of ensemble outliers (2)

Procedure (From *Kuo et al*. 2004).

1. Take one sample and treat the remaining samples as a mother group.
2. Compute mean value and standard deviation of mother group.
3. Assuming that the deviations follow a Gaussian distribution we can determine if the sample belongs to the mother group or not.
4. If the observation falls outside some predefined significance level the particular sample is considered and outlier and is excluded from the data set.
5. Repeat this procedure for all samples.
6. Repeat this procedure until no additional outliers is detected.

See also lecture:
‘Recent result on quality control, simulation, and assimilation of GPS RO data’
By Xiaolei Zou.

Removal of ensemble outliers (3)

Advantages:

- A priori data is not required.
- Processing independent.
- Works well for big outliers.
- Fairly straightforward.
- Convenient for 'simple' users.

Disadvantages:

- Do not remove smaller outliers.
- This approach should not be used by. data providers as it is basically a sanity. check - operational centers prefer to do their own sanity check.
- Observations of extreme weather may me tagged as 'bad'.

Compare to other data sources (1)

Advantages:

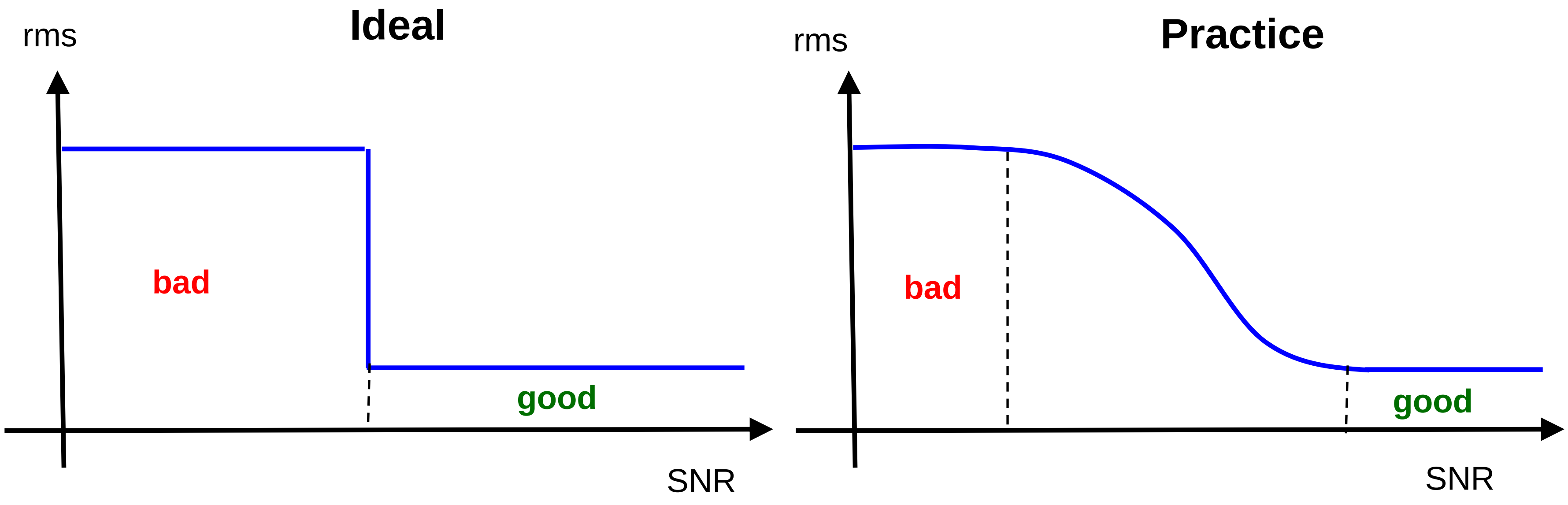
- Processing independent.
- Works well for big and smaller outliers.
- Fairly straightforward.
- Convenient for ‘simple’ users.
- Very useful for improving other QC approaches.

Disadvantages:

- Performance depends on quality of the ancillary data.
- This approach should not be used by data providers as it is basically a sanity. check - operational centers prefer to do their own sanity check. (May be used by data providers to remove observations which are obviously not correct.)
- Observations of extreme weather may me tagged as ‘bad’.

Data based QC (1).

Basic assumption: ‘bad’ data can be identified by some extreme property of the raw data (e.g. SNR).



Multiple QC parameters are normally used to better pinpoint ‘bad’ data.

Data based QC (2).

Advantages:

- Does not require ancillary data.
- Works for extreme events.
- Independent of the magnitude of the errors.
- Well suited for data providers.

Disadvantages:

- Requires detailed knowledge of the processing techniques.
- May allow big outliers to be tagged as ‘good’.

Manual inspection

Advantages:

- Can be very reliable.
- Useful for case studies of extreme events.
- Very useful for improving other QC approaches.

Disadvantages:

- Can only be performed by a processing expert.
- Extremely time consuming.

Radio Occultation data based QC - detection of tracking errors (General CDAAC QC)

- Mean and standard dev. of Ionospheric free bending angle from climatology at altitudes 60-80 km - indication of bias errors, tracking problems, and large measurement errors.

$$\sigma_{obs} = \sqrt{\left\langle \ \alpha_{obs} - \alpha_{clim} \ {}^{2} \right\rangle}, \quad \delta\alpha = \left\langle \alpha_{obs} - \alpha_{clim} \right\rangle$$

- $(\alpha_{12})_{max:}$:maximum differences between L1 and L2 bending angles – indication of tracking problems.

- S4: Normalized standard deviations of L1 signal amplitude at 40-80 altitudes - large values indicate strong ionospheric scintillations and high probability of tracking errors at all altitudes.

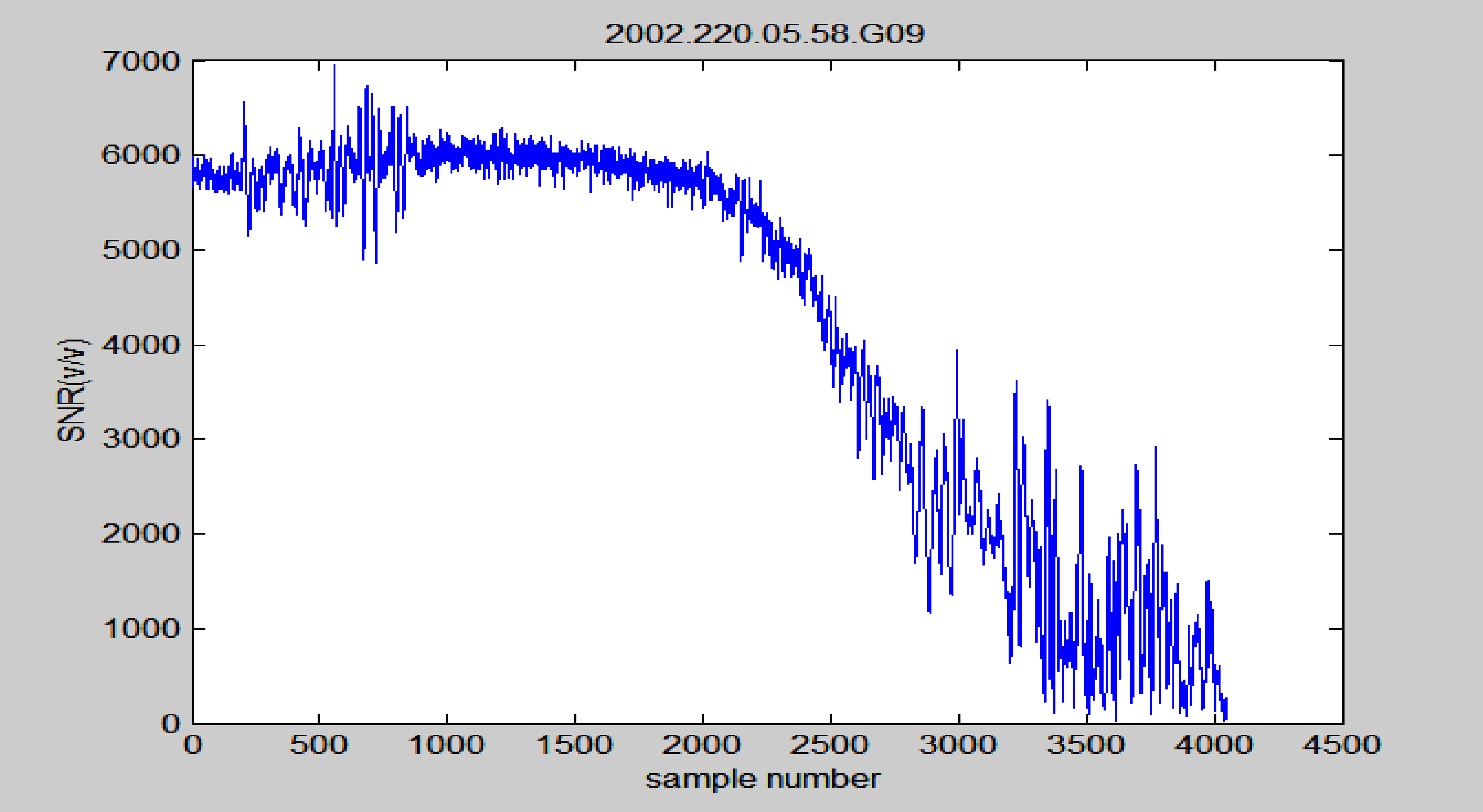
- Max. fractional deviation from climatology – indication of severe measurement/tracking errors.

- Specific thresholds are based on comparison with reanalysis (ECMWF) and manual data inspection.

- Fore more details see (*Kuo et al.* 2004).

Radio Occultation QC – signal truncation (1) (General CDAAC QC)

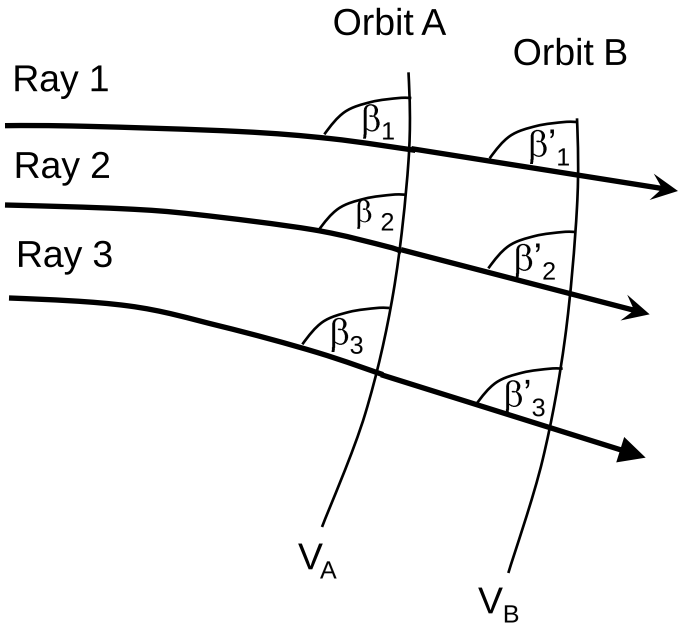
Where should the signal be truncated?



Radio Occultation QC – signal truncation (2) (General CDAAC QC)

- To determine the cut-off point the following principle can be applied (*Sokolovskiy* 2001)*:*

  For a large enough distance from the Earth's limb to the receiver, the fractional variability of the Doppler frequency shift of the RO signal is much smaller than the corresponding variability of the refractivity in the atmosphere.



Radio Occultation QC – signal truncation (3) (General CDAAC QC)

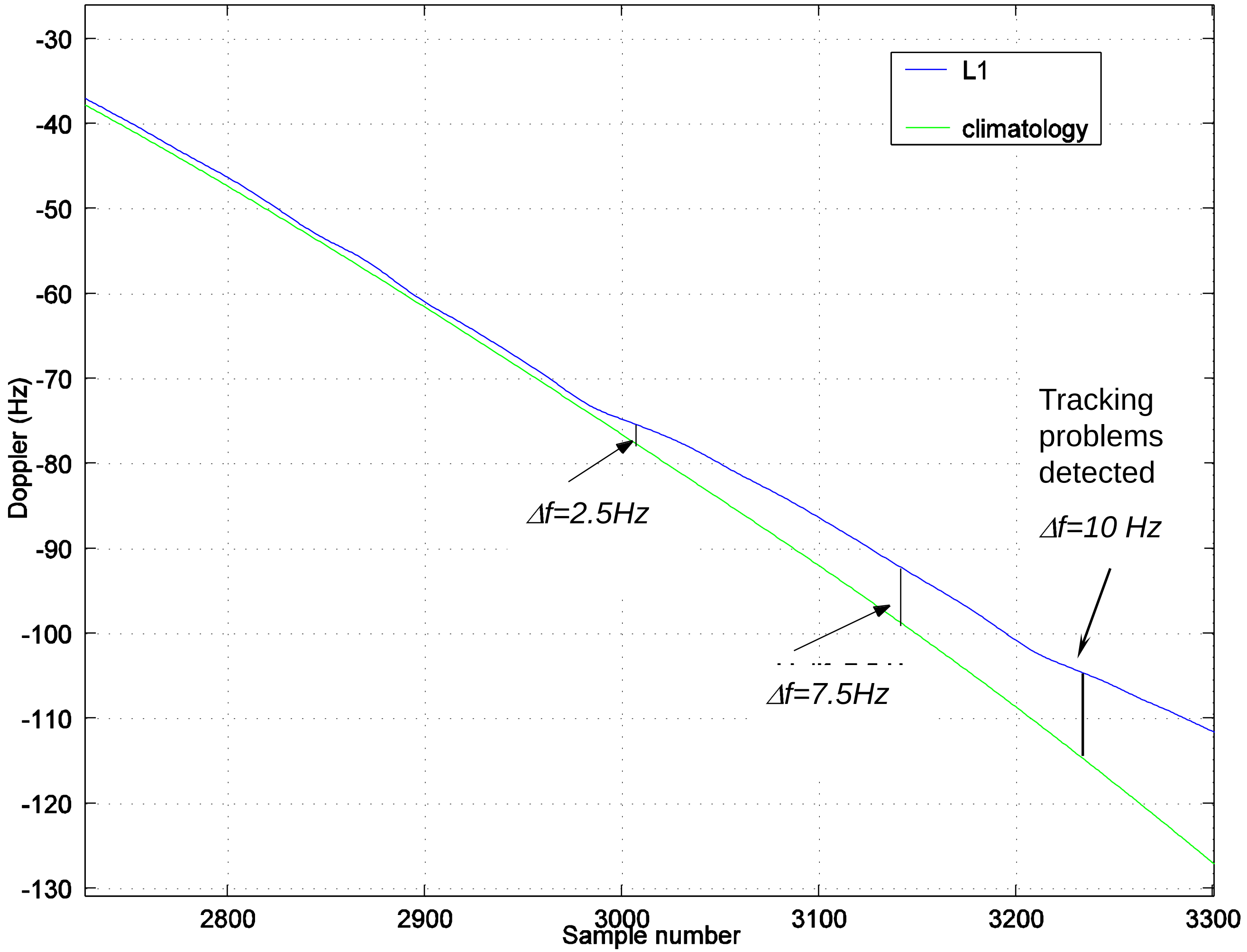
- As the distance from the limb increases the total Doppler and Doppler spread decreases and the duration of the signal increases.

- For GPS-LEO radio occultations the distance to Earth's limp is approximately 3000 km and it can be shown that Doppler shift can be predicted by a climatology with and accuracy of +/- 15-20 Hz (*Sokolovskiy* 2001).

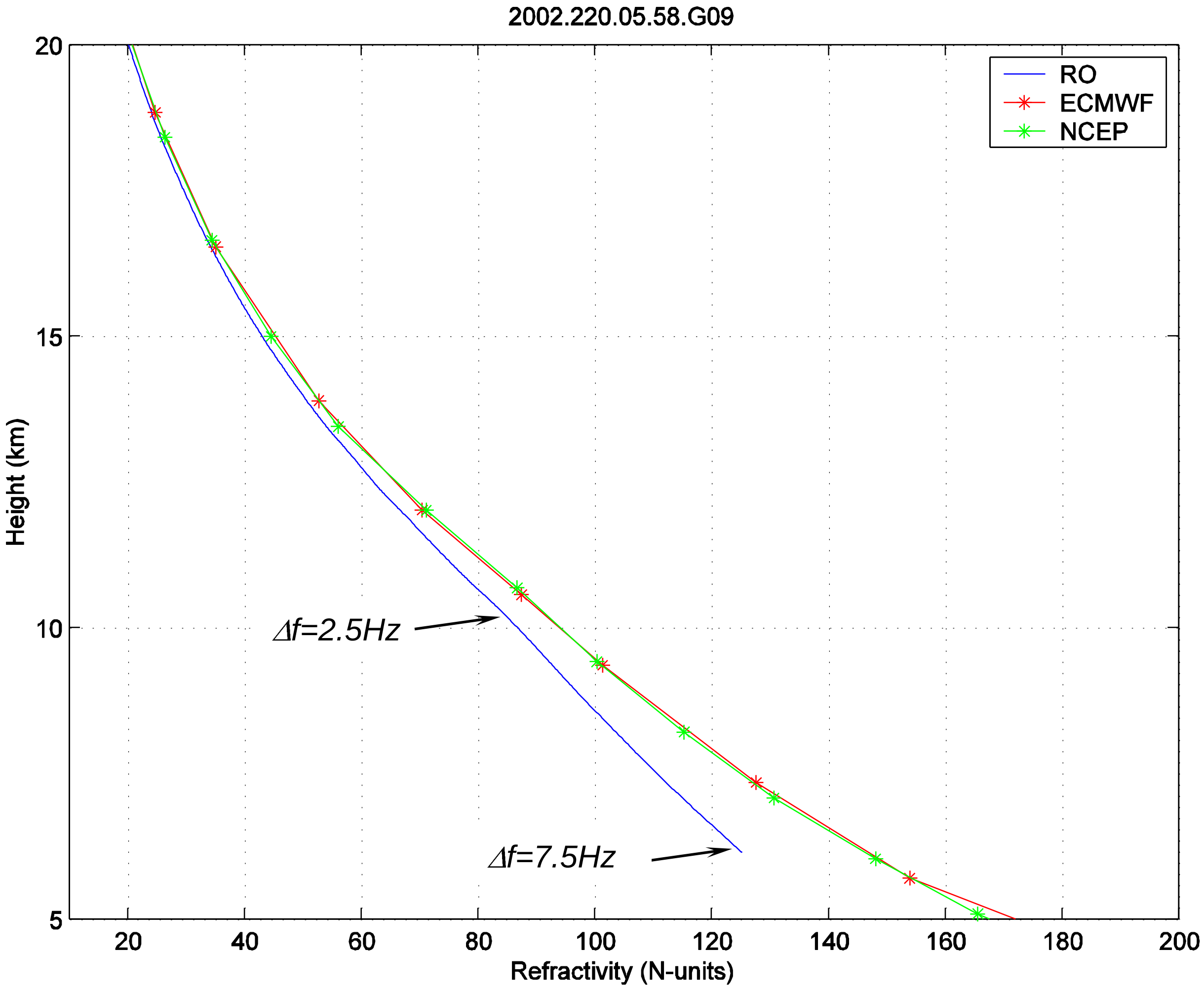
- This important results can be used for QC (and for signal tracking in open loop mode).

- The +/- 15-20 Hz criteria can be used to tell us that the instrument is not tracking correctly but it cannot tell us when the problems started.

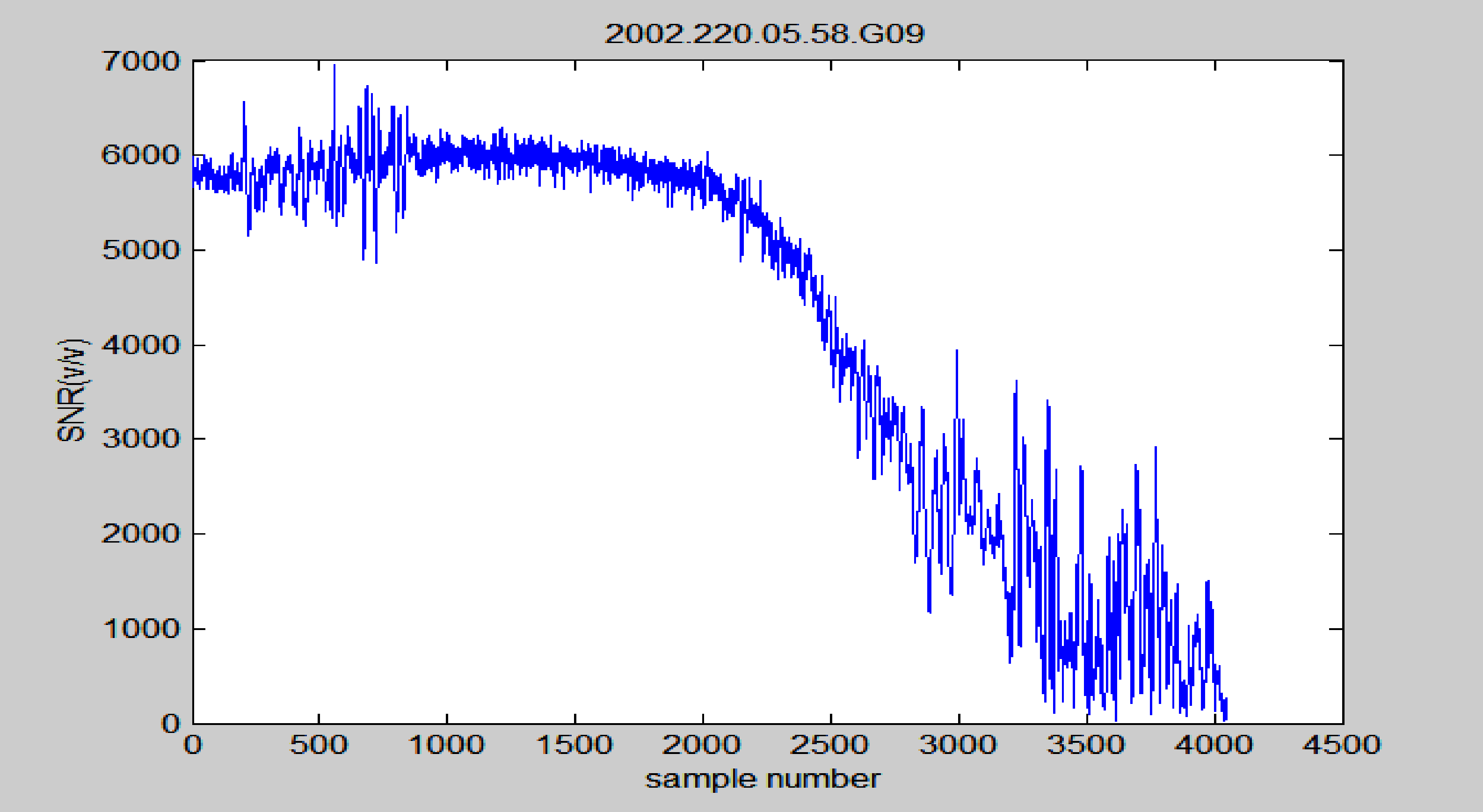
Radio Occultation QC – signal truncation (4) (General CDAAC QC)



Radio Occultation QC – signal truncation (5) (General CDAAC QC)

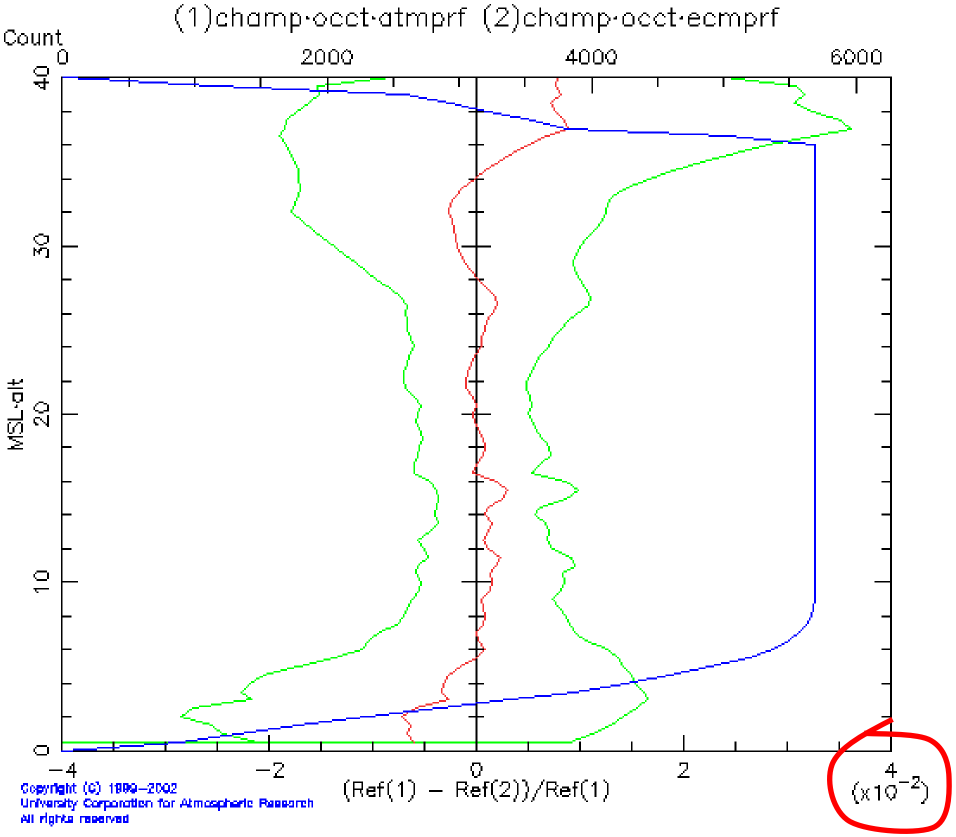


Radio Occultation QC – signal truncation (5) (General CDAAC QC)

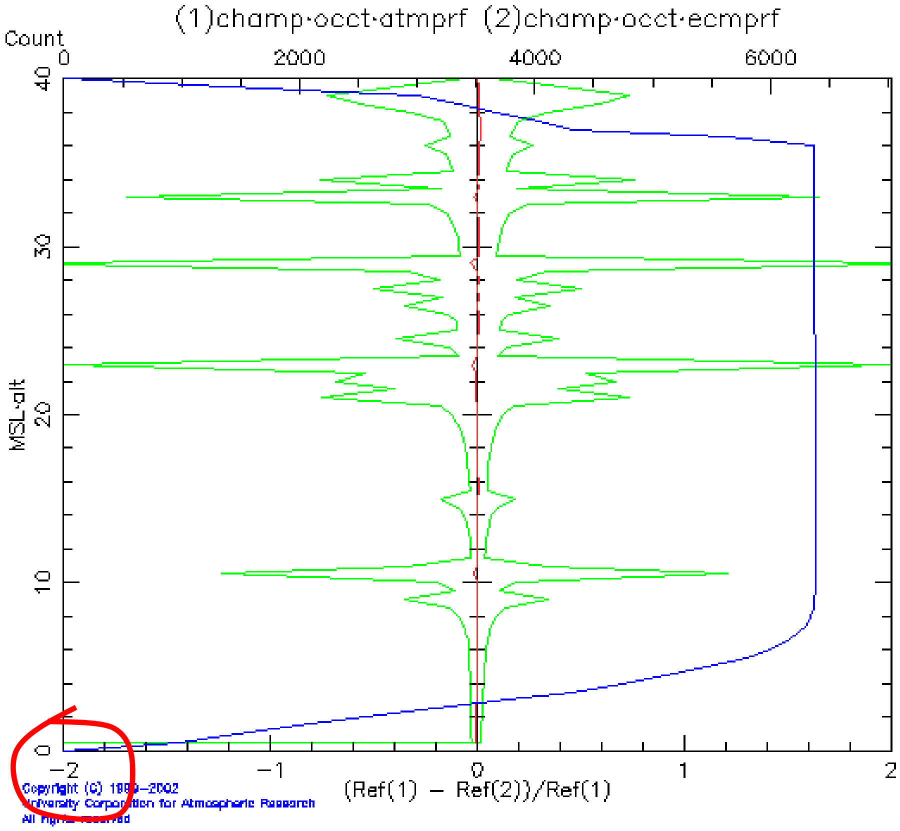


Comparison with ECMWF - QC vs. no QC

Processing with QC



Processing without QC



How can we determine error statistics?

1. Compare retrieved data to the true atmospheric state and compute statistics – not possible! Generally error estimation is an underdetermined problem.
2. Error budget - estimate contributions from individual error sources.
3. Compare retrieved data to other data sources e.g. the atmospheric state taken from a model or other observations.
4. Simulations.
5. Dynamic error estimation - estimate measurement errors and propagate these errors to geophysical parameters.

Observation errors vs. measurement errors (Definitions from *Sokolovskiy et al.*, 2005)

- Observation error: the apparent difference between a physical quantity (observable) obtained from observations and modeled by use of an atmospheric model.

- The observation error include:
  1. The modeling error
  2. The representativeness error
  3. Measurement error

- Modeling error: error related to approximations applied for modeling an observable (e.g. treating Abel retrieved refractivity as a local measurement).

- The representativeness error: the error related to the discrete representation of continuous meteorological fields by an atmospheric model and the use of discrete representation in the model of an observable.

- Measurement error: error related to the physical process that effect observations, but are not (or may be not) modeled in the deterministic sense (e.g. receiver noise, ionospheric, irregularities, turbulence).

- Only the measurement error should be supplied by a data provider. Why?

Radio Occultation Measurement Error Sources

Instrumental Errors

- **Local multi path**
- **Thermal noise**

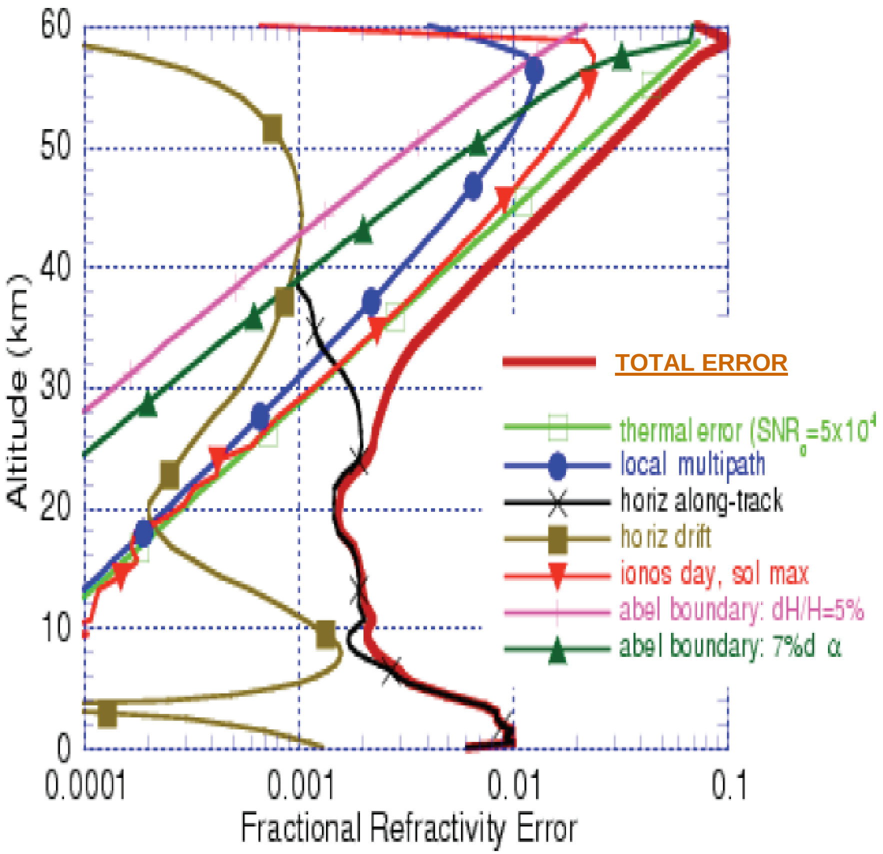
Atmospheric Effects

- **Ionosphere**
- Diffraction
- Scintillations/turbulence
- Scattering

Retrieval Errors

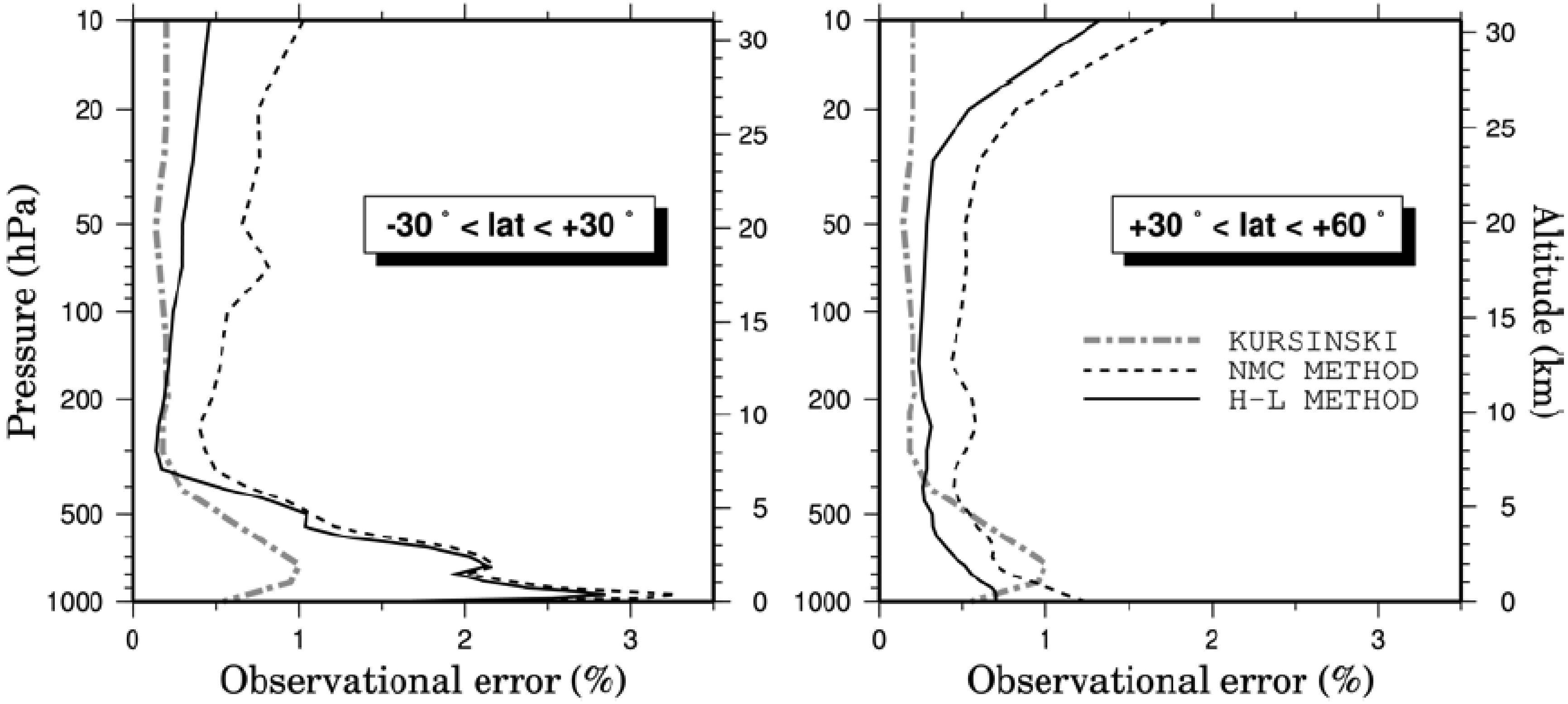
- Integral initialisation (**Climatology**, hydrostatic integral)

RO occultation error characteristics.
Error budget (From *Kursinski et al.*, 1997)



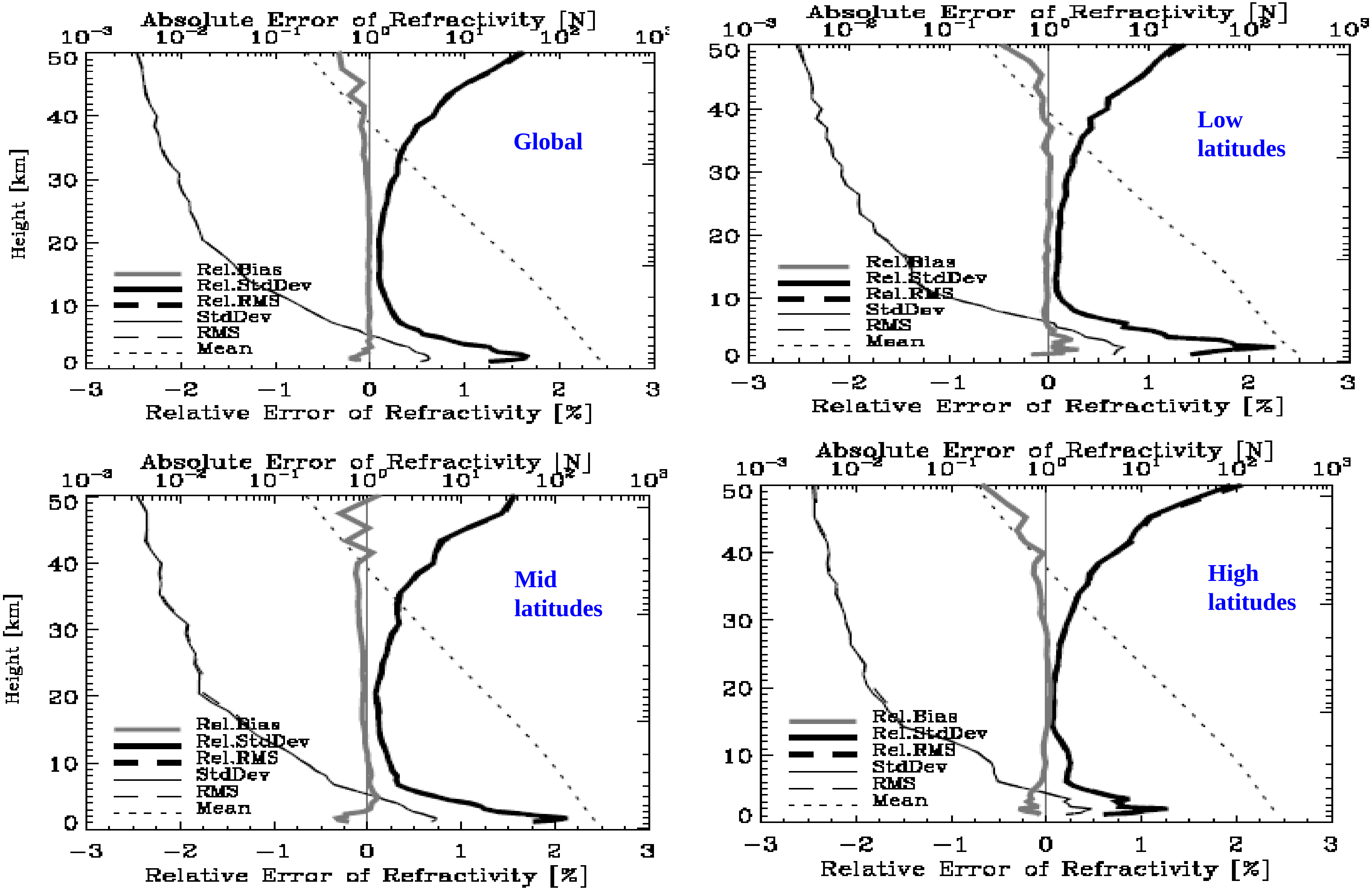
1. To first order high altitude errors are constant while the atmospheric effect is growing exponentially with decreasing altitude => Fractional error increases exponentially with altitude.
2. Above approx. 25 km the dominant error sources are ionospheric noise, thermal noise, and local multipath – measurement errors.
3. Below approx. 25 km the dominant errors are due to non-spherical symmetry – a model error.

RO occultation error characteristics
Comparison between CHAMP RO and 12 hour forecast (From *Kuo et al.* 2004)



RO observational errors (in terms of fractional differences in refractivity) based on the NMC method (dashed line) and the Hollingsworth and Lonnberg (1986) method (solid line). The RO measurement errors as estimated by Kursinski et al. (1997) are shown as dot-dashed lines.

RO occultation error characteristics
Simulations (From *Steiner and Kirchengast* 2004)



Dynamic error estimation

1. What are the dominant errors?

2. How can we be estimate these errors?

3. How do these errors propagate through the retrieval chain?

Statistical optimization (1)

- In RO, the atmospheric refractivity is computed from observed bending angles through the Abel transform:

$$n(r_0) = \exp\left(\frac{1}{\pi} \int_{a_0}^{\infty} \frac{\alpha(a')}{\sqrt{a'^2 - a_0^2}} da' \right)$$

- To prevent noise from propagating to lower altitudes, measured bending angles can be combined with a First Guess bending angle profile in a statistical optimal way by minimizing the cost function:

$$J(\alpha) = (\alpha-\alpha_{bgn})^{\mathrm{T}} \mathbf{B}_{bgn}^{-1} (\alpha-\alpha_{bgn}) + (\alpha-\alpha_{guess})^{\mathrm{T}} \mathbf{B}_{guess}^{-1} (\alpha-\alpha_{guess}) = \min$$

$$\Downarrow$$

$$\alpha_{opt} = (\mathbf{B}_{bgn}^{-1} + \mathbf{B}_{guess}^{-1})^{-1} (\mathbf{B}_{bgn}^{-1} \alpha_{bgn} + \mathbf{B}_{guess}^{-1} \alpha_{guess})$$

- If vertical correlations are neglected this equation simplifies too:

$$\alpha_{opt}(a) = \frac{\sigma_{guess}^2(a)\alpha_{bgn}(a) + \sigma_{bgn}^2(a)\alpha_{guess}(a)}{\sigma_{guess}^2(a) + \sigma_{bgn}^2(a)} = (1 - w(a))\alpha_{bgn} + w(a)\alpha_{guess}(a)$$

Error estimation

To perform statistical optimization and error estimation we need to:

1. Estimate the correlation length and standard deviation of the background noise (ionospheric noise, thermal noise, local multi path).

2. Estimate the correlation length and standard deviation of the First Guess.

3. Propagate these errors to refractivity.

Estimation of background error covariance:

- The background noise error covariance can be estimated in the height range 60-80 km from differences between observations and First Guess bending angles.

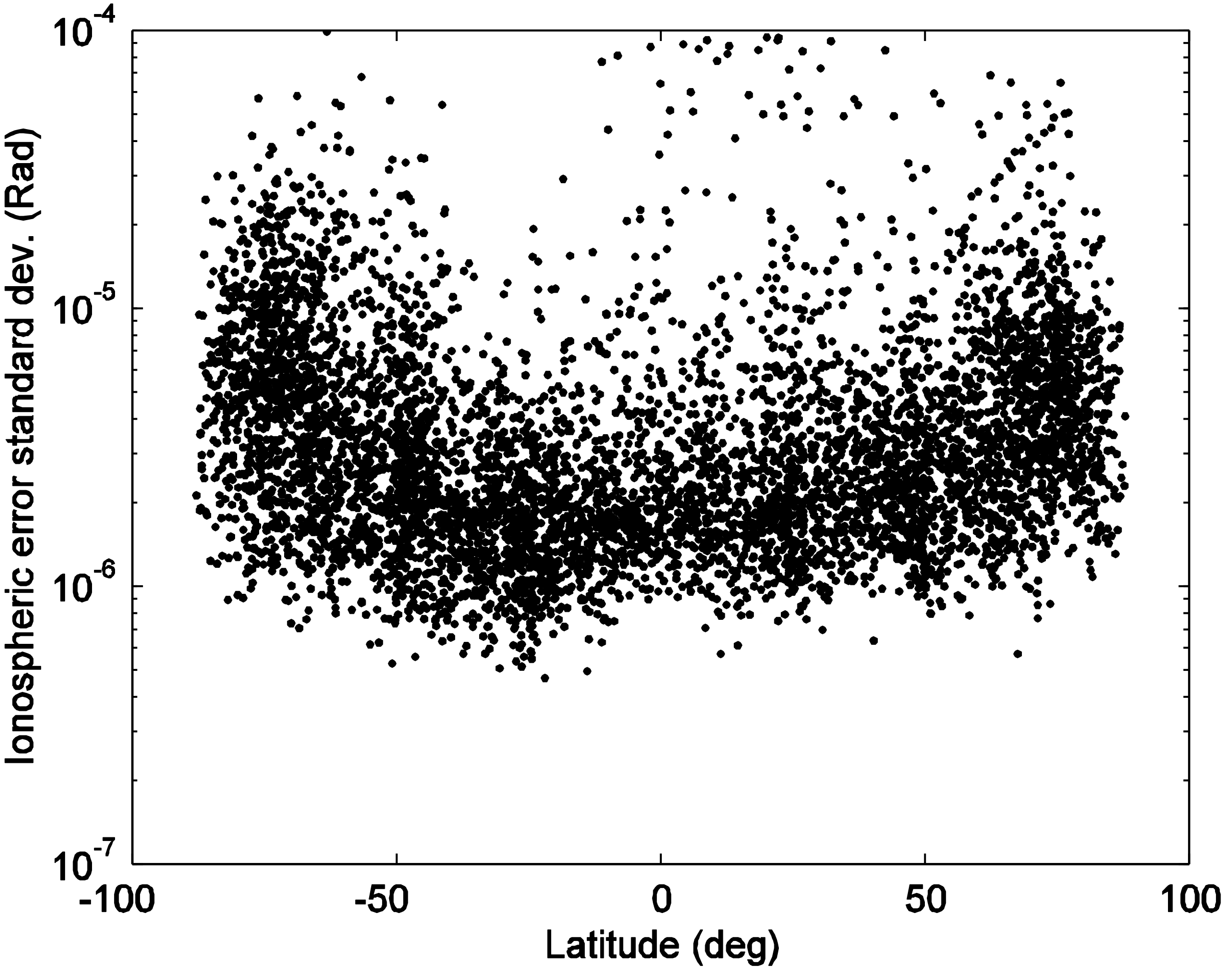
$$r_{obs}\ \ \tau\ \ = \left\langle \Delta\alpha\ \ a\ \ \Delta\alpha\ \ a+\tau\ \right\rangle \approx \frac{1}{M_{\tau}}\sum_{i}\Delta\alpha\ \ a_i\ \ \Delta\alpha\ \ a_i+\tau$$

- The correlation length, l_{bgn}, is computed from a Gaussian fit to the first part of the correlation function ($0.1 < \rho(\tau) \ \leq 1$):

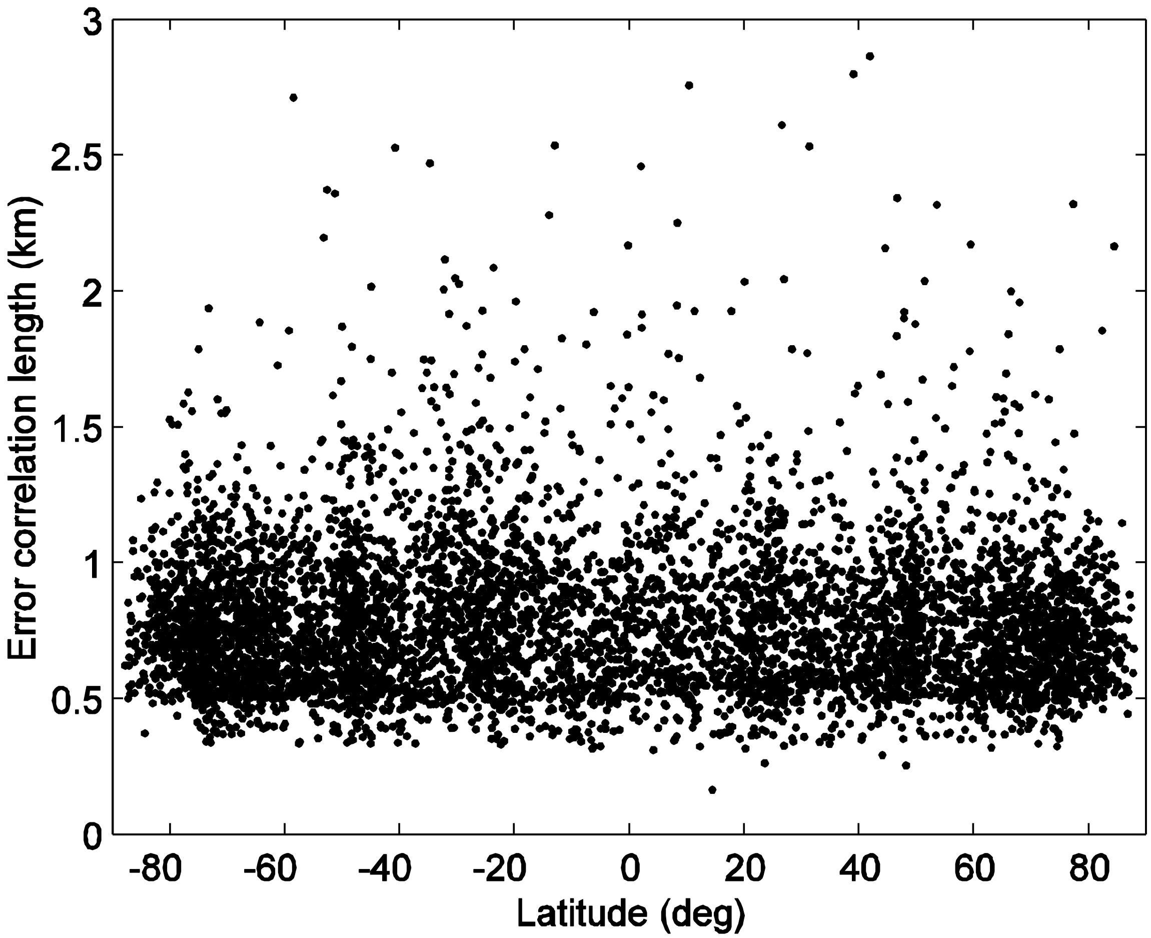
$$\rho_{bgn}\ \ \tau\ \ \approx\sim \exp\left(-\left(\frac{\tau}{l_{bgn}}\right)^{2}\right)$$

- As the background errors are constant with height these error estimate can be applied through out the profile.

Standard deviation of bending angle (60-80 km) background noise (from CHAMP August 02)



Estimated observation error correlation lengths (from CHAMP August 02)



Estimation of relative First Guess errors for each occultation

- First Guess standard deviation is assumed to correspond to a fixed ratio, *K,* of the First Guess bending angle profile.

- This ratio is estimated for each occultation from the differences between observations and First Guess bending angles, $\Delta\alpha$, in the height range 20-60 km.

$$\left\langle \alpha_{bgn}(a)-\alpha_{guess}(a)\ {}^{2}\right\rangle=\left\langle \Delta\alpha(a)\ {}^{2}\right\rangle=\sigma_{bgn}^{2}+\ K\alpha_{guess}(a)\ {}^{2}$$

$$\Downarrow$$

$$K^{2}\approx\frac{\dfrac{1}{N}\sum_{i}\ \Delta\alpha(a_{i})\ {}^{2}-\sigma_{bgn}^{2}}{\dfrac{1}{N}\sum_{i}\alpha_{guess}^{2}(a_{i})}$$

Estimation of First Guess error correlation function

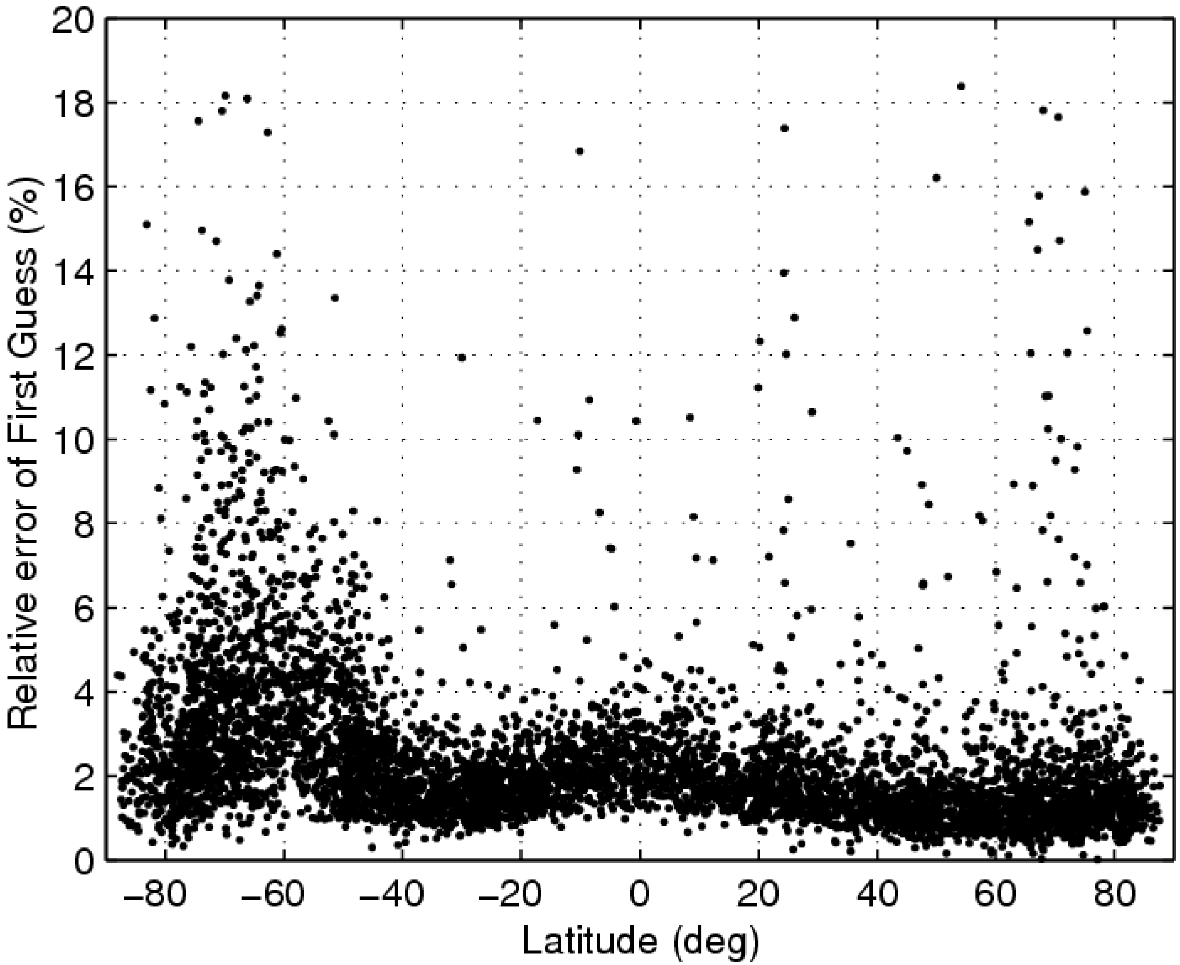
- The First Guess error correlation can be expressed as:

$$\rho_{guess}(\tau) = \frac{\frac{1}{N_\tau}\sum_i \Delta\alpha(z_i)\Delta\alpha(a_i+\tau) - r_{obs}(\tau)}{\frac{K^2}{N_\tau}\sum_i \alpha_{guess}(a_i)\alpha_{guess}(z_i+\tau)}$$

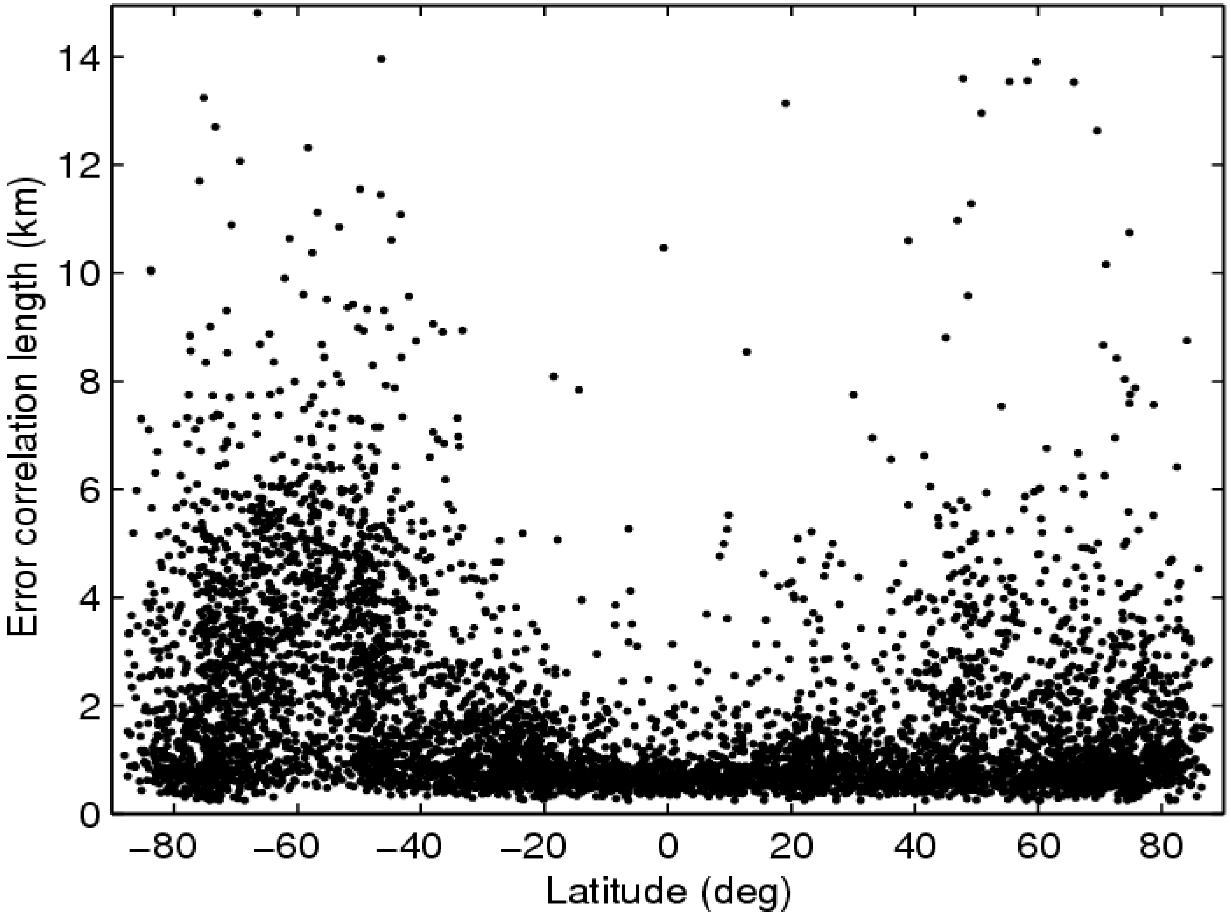
- The correlation length, l_{guess}, is computed from a Gaussian fit to the first part of the correlation function $(0.1 < \rho(\tau) \leq 1)$:

For more details on dynamic error estimation see *Lohmann* 2005.

Relative errors of First Guess



Error correlation lengths of the First Guess errors



Error propagation

- To propagate errors from bending angle to refractivity we linearize and discretize the Abel transform (*Syndergaard* 1999):

$$\mathbf{N}=\mathbf{A}\boldsymbol{\alpha}_{opt}$$

- At high altitudes (above approx. 20 km) refractivity error covariance can be expressed as:

$$\mathbf{C}_N= \mathbf{A}\mathbf{C}_\alpha\mathbf{A}^\mathrm{T}$$

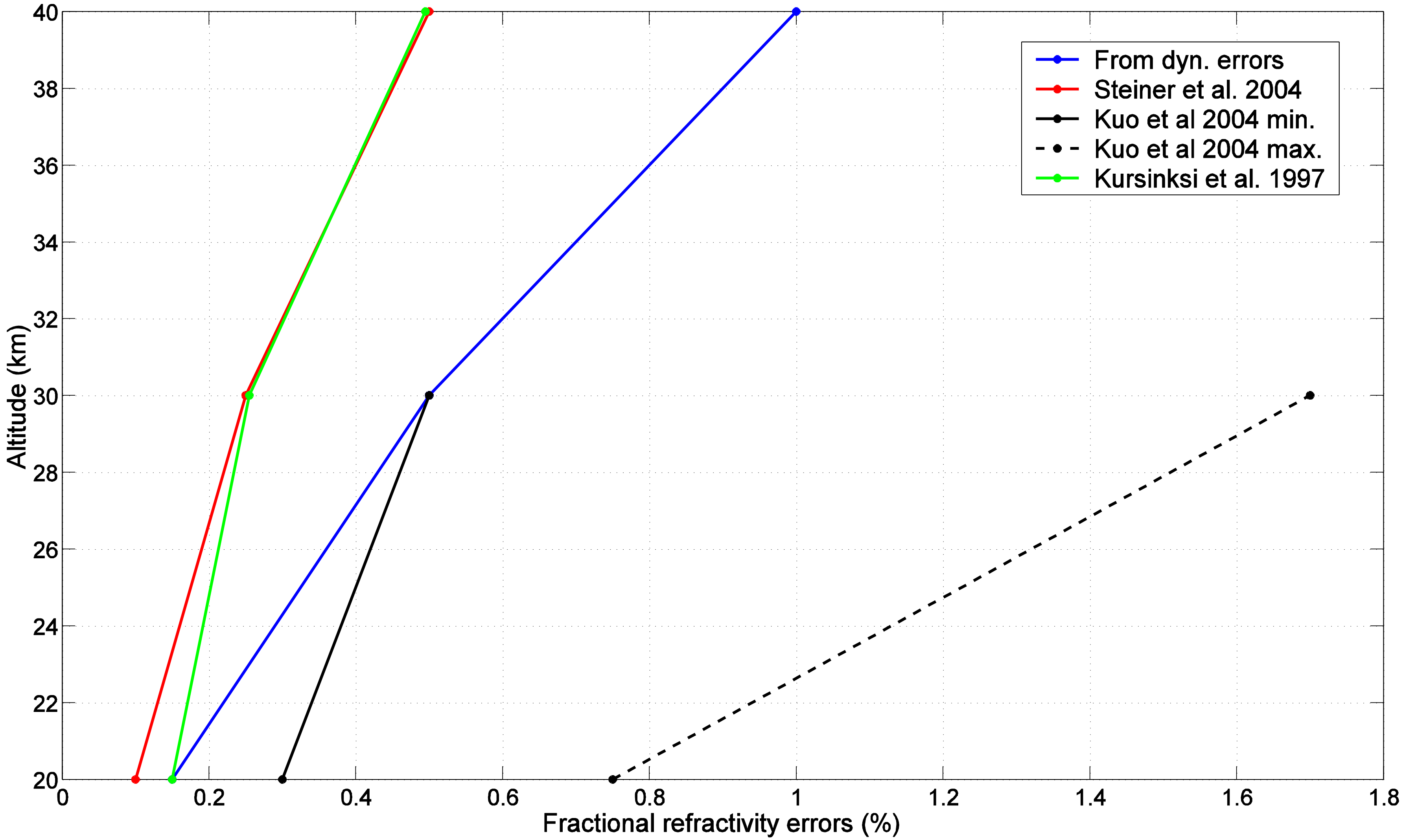
- At lower altitudes we must account for errors in height as well since $r_0=n/a_0$ (see *Syndergaard* 1999).

- The bending angle error covariance can be estimated as:

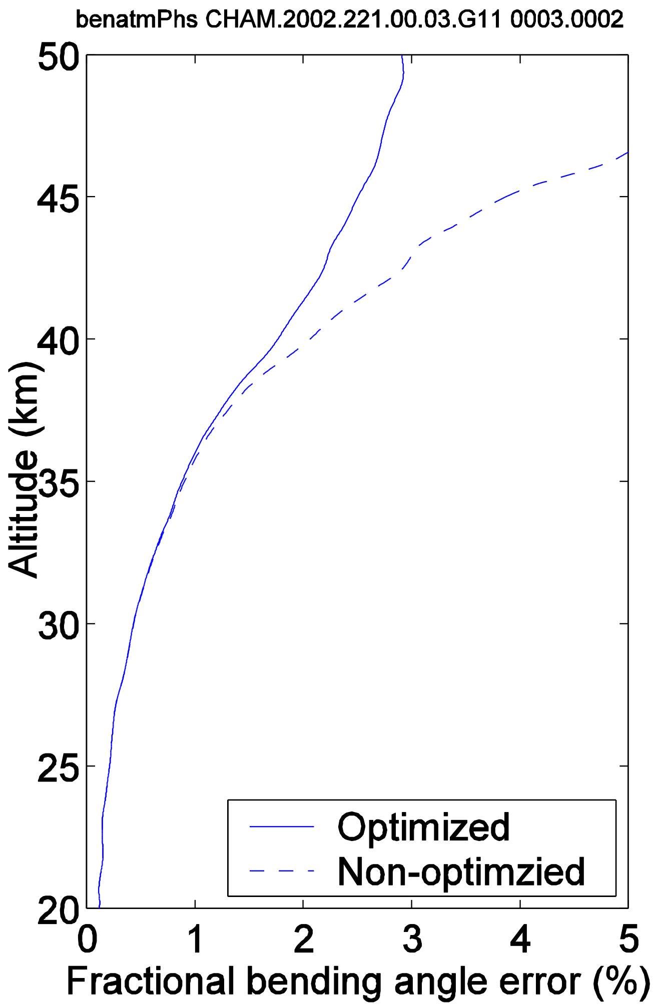
$$\mathrm{C}_\alpha(i,j)=w\;\;a_i\;\;w\;\;a_j\;\;K^2\alpha_{\mathrm{guess}}(a_j)\alpha_{\mathrm{guess}}(a_i)\mathrm{exp}\left(\left(\frac{a_i - a_j}{\ell_{guess}}\right)^2\right)+\;1-W\;\;a_i\quad 1-w\;\;a_j\quad\sigma^2_{bgn}\mathrm{exp}\left(\left(\frac{a_i - a_j}{\ell_{bgn}}\right)^2\right)$$

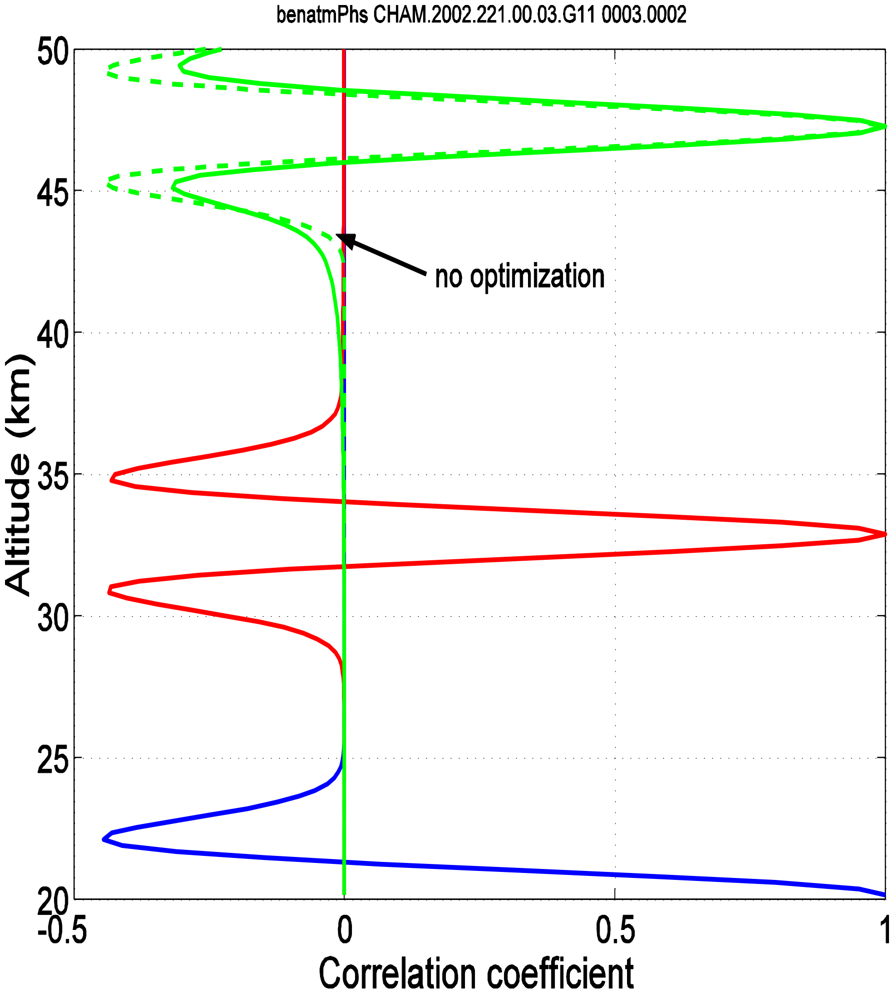
HOW DOES THIS WORK?

Comparison between different error estimates



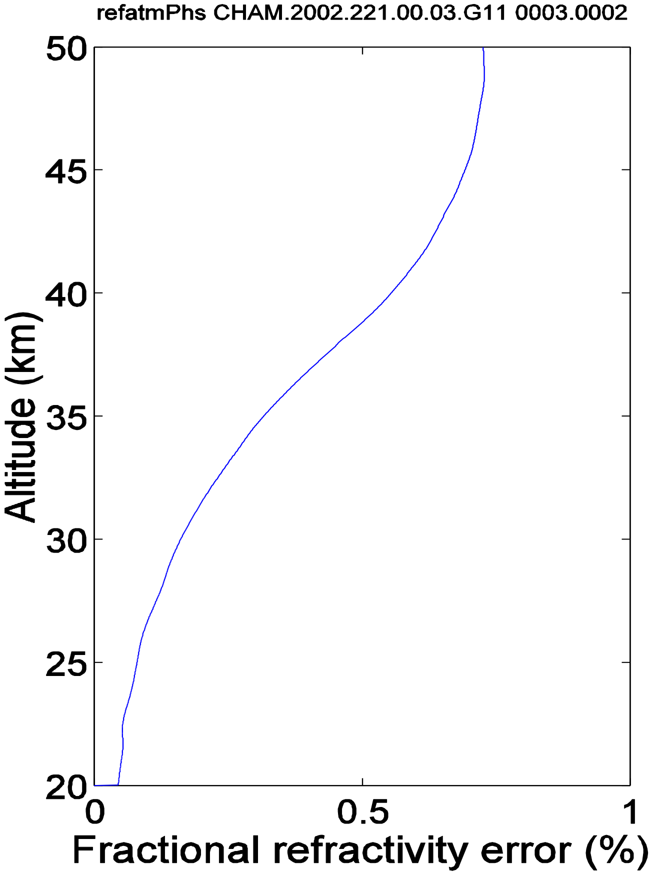
Bending angle errors for a single occultation

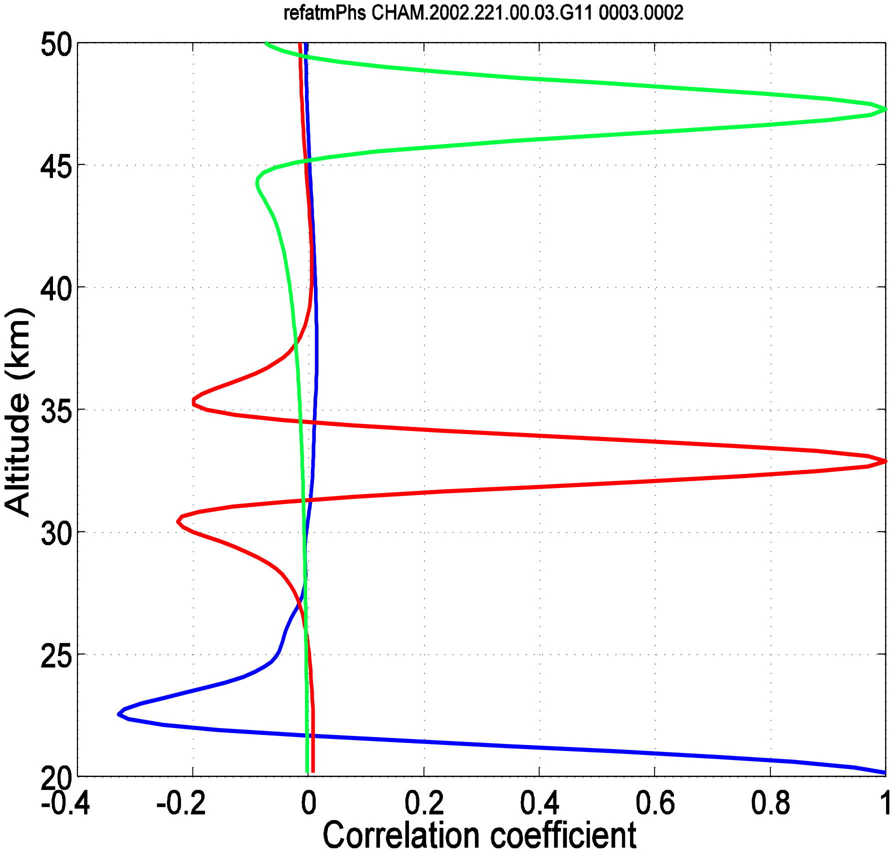




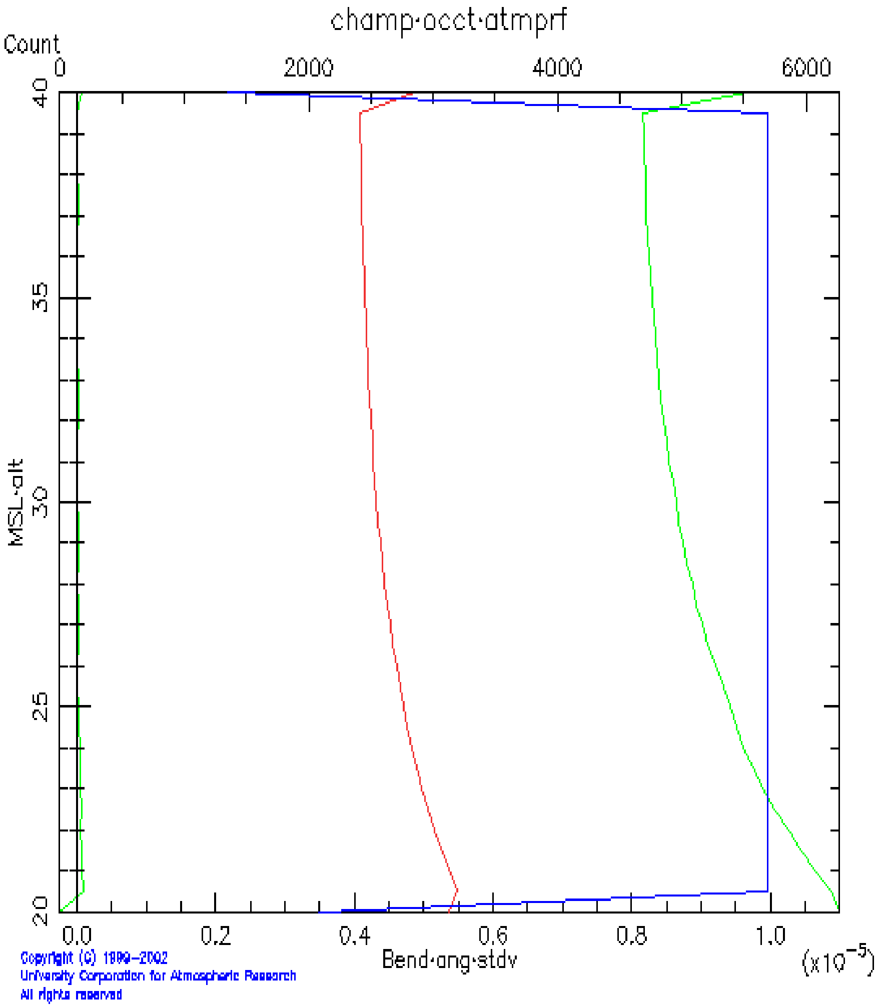
no optimization

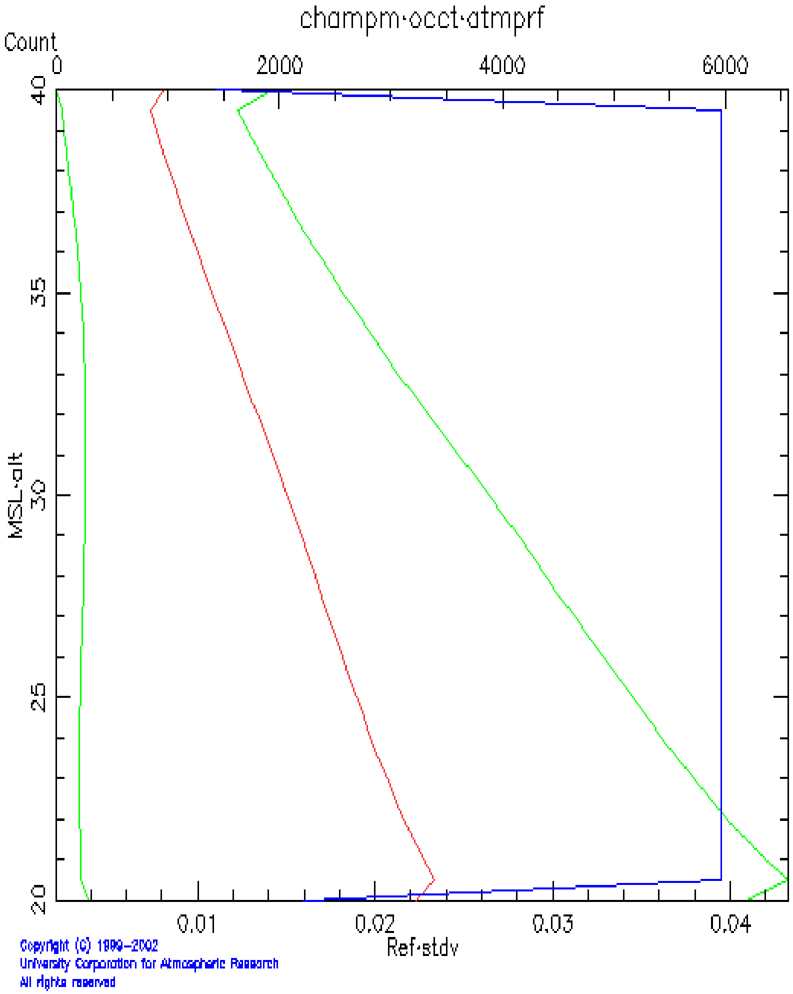
Refractivity errors for a single occultation





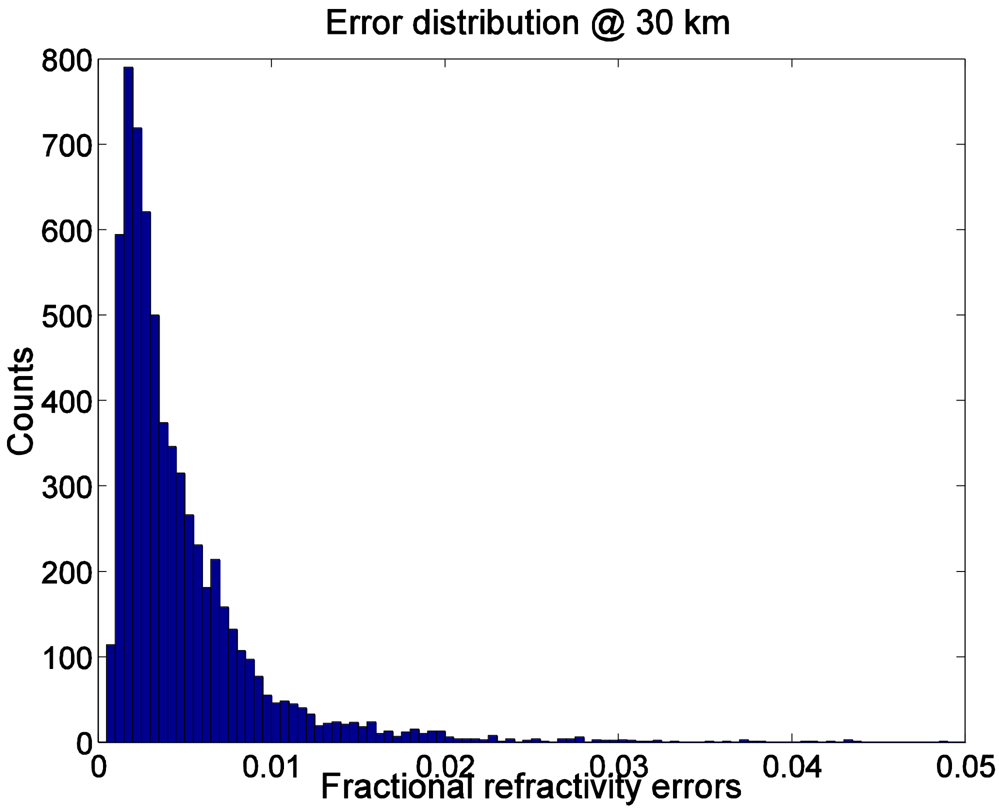
Error profiles for August 2002





Error distributions at 30 km altitude

Refractivity



Summary and conclusions

- QC and error estimations are extremely important.
- The QC applied by a data provider will generally be different from the QC applied by a user - different users may apply different QC strategies.
- Data providers should provide measurement errors – model error and representativeness errors depend on the user.
- RO error estimate shows that RO refractivity errors may reach 3-4 % in the moist troposphere and 0.3 %-0.4% over much of rest of the troposphere and stratosphere.
- Dynamic error estimation allows for assigning different errors to individual occultations.