TUNER
TOWARDS UNIFIED ERROR REPORTING

Minutes of the 1st TUNER Meeting, June 15, Univ. of Saskatchewan, Saskatoon

by Thomas von Clarmann, based on notes by Nathaniel Livesey

Agenda:

1. Status of project
2. Analysis of the questionnaire
3. Deductive error analysis
4. Inductive Error Analysis
5. What do Data Users Need?
6. Other Business

1. **New and old news** (Thomas von Clarmann)

   - TUNER has been approved as an emerging SPARC activity at the last SPARC SSG meeting;
   - TUNER has been selected as an ISSI International Team;
   - TUNER has no funding except some travel funds. Perhaps the status as SPARC/ISSI activity might encourage further funding, but currently no source of funding of specific work has been identified.
   - The agenda was approved.
   - The total duration of the TUNER-SPARC activity is foreseen to be 3 years; in the first year it has the status of an "emerging" activity.

2. **Analysis of questionnaire** (Thomas von Clarmann)

   A questionnaire was circulated among the retrieval scientists in order to obtain an overview over the characteristics of the various retrieval and error analysis schemes in use by the various satellite groups. Responses were obtained for 12 limb and 1 nadir mission. Limb missions include limb emission, limb scattering, and occultation. Measurements in the following frequency ranges were represented: MW, FIR, IR, NIR, vis, UV.

Two groups retrieve full 2D distributions per inversion, one group retrieves 1D vertical profiles along with a horizontal gradient. For 7 instruments a full 1D profile is retrieved in one step and for two instruments one value per inversion is retrieved in an onion peeling mode. All retrievals are based on a matrix formalism with or without regularization, the latter being either optimal estimation or Tikhonov-type. Some groups provide their data on the native retrieval grid, while others interpolate their data to a regular grid after the retrieval. In the latter case care has to be taken to also transform the diagnostic data onto the new grid.

Noise is evaluated by most groups using $\mathbf{S}_{noise} = \mathbf{G}^T \mathbf{S}_y \mathbf{G}$ for each profile separately, and this is considered adequate. Typically the measurement error covariance matrix $\mathbf{S}_y$ is diagonal unless spectra are apodized. 9 instruments analyze noise estimates for each single profile while 3 instruments analyze noise for selected examples. Estimating error propagation through the layers for onion peeling retrievals may need some further thoughts. In summary, there is good agreement on the treatment of retrieval noise and only a few open issues have been identified.

Much less agreement has been found for the treatment of **parameter errors**. Two schemes in use are perturbation studies and linear error propagation using the full parameter uncertainty covariance matrix. Assumptions on the input parameter errors are considered critical. Bernd Funke stated that regarding this unification is needed, and that ideally each instrument should use the same input parameter errors. Thomas von Clarmann replied that different instruments have different sensitivities to the various error sources, and that a different background statistic may be applicable to each instrument. Gabi Stiller mentioned that different error assumptions are needed for different instruments because some instruments use a priori parameter information while others retrieve the parameters in a preceding step. This leads to different parameter uncertainties. Another issue is correlations between input parameter errors. Correlations of spectroscopic data uncertainties are a particular problem. For example, if a retrieval uses multiple lines in multiple bands, it is typically not clear to which degree band/line intensity errors of different bands/lines cancel out. Further issues with parameter errors have been identified: The impact of hydrostatics on profile retrieval can be huge, and it is not sure if this is always done in a consistent way (Bernd Funke). Often parameter errors are not assessed for each single profile but only for selected cases. Here it is not clear if these examples are really representative. It has also been highlighted that parameter error correlations in the x-space are important for certain applications. For example, when computing a column total while neglecting the (typically negative) correlations in the systematic errors, then they will grossly overestimate the uncertainty in the column amount.

Forward model errors depend largely on the particular instrument. The hottest candidates were found to be line shape issues, horizontal inhomogeneities, non-LTE, self-absorption (when transparent atmosphere approximation is used) and instrument specific issues. During the discussion also the issue of aerosol and cloud effects was mentioned. There seems to be a good awareness of the issues to be considered but quantification is not always easy. Thomas von Clarmann suggested that issues here are best left to the experts of the particular instrument.

Averaging kernels are provided for all instruments where regularization is used, but there is no full agreement what is the altitude resolution of non-regularized profiles is. This has been deferred to later. Another open issue is if exemplar averaging kernels are sufficient or if an averaging kernel matrix is needed for each single profile. In some cases, sensitivity is state dependent, so profile by profile kernels can be needed. The applicability of "averaged averaging kernels" is not clear. The possible need for those arises because, for example, data users often want things like monthly zonal means. A presentation on this issue was given later in the meeting. Another issue where the data user may need advice from the experts may be how averaging kernels are to be interpolated to another vertical grid.

Combination of errors can be a task for the data providers, who then provide the total error estimates to the data users; alternatively, the error components can be distributed to the data users, who combine the errors themselves. Both approaches are used by the participating data providers, and both have their pros and cons.

Validation papers are available for most of the participating instruments (but not necessarily for all, or most recent, data products). Within TUNER no validation studies will be made, but it will be heavily drawn upon existing validation studies. These are considered particularly useful to judge which error estimation schemes are adequate.

Many instruments have **drift** issues. Difficulties to determine the drift of one instrument in absolute terms (rather than in comparison to another instrument) were discussed. There seemed to be agreement that drift issues should not be covered by TUNER, because drifts are systematically assessed by Daan Hubert in the framework of other project and duplication of work should be avoided.

3. **Deductive Error Analysis**

   This is propagation of ingoing uncertainties through the retrieval system. Several talks were given

   OMPS error estimation (Natalya Kramarova)

   The main sources of uncertainties in limb scattering observations are sensor pointing errors, instrumental effects (systematic errors in measurements), background aerosols, inhomogeneity along the line of sight, the forward radiative transfer model and the inverse model. To the sensor pointing, a correction of ∼1.1 — 1.5km, varying from slit to slit is applicable. The overall uncertainty in these corrections is about 200m. First a static corrections was applied for each slit but tangent height changes over the orbit were observed. As a possible explanation, the flexure of the s/c platform due to graduate heating along the orbit is considered. Since LP is mounted at the end of the s/c platform the thermal flexure can cause an error ∼20 arc-sec in determining s/c pitch angle at the LP location. Almost linear dependence from south pole to north pole was found. Static and (daily varying) orbit corrections were already implemented in Level 1.Systematic patterns in measured radiances were found in spectral and spatial domains (in both UV and VIS spectra). These structures in measured radiances propagate into the retrieved ozone profiles. The following candidate explanations are under consideration: Overly simplified band shape assumed by L2 (error in gridding process - mapping radiances on the regular spectral and vertical grids); band center mismatch between radiance and irradiance spectra; change in radiometric response associated with thermal-induce spectral shifts (seasonal and intra-orbital). The role of background aerosol as an error source as well as horizontal inhomogeneities with respect to aerosol, temperature and ozone were discussed. The forward model has been validated by Loughman et al., (2004, 2015). Error budgets, averaging kernels and gain matrices were presented. The discussion of this paper focused on the calculation of error bars for zonal, monthly, and monthly zonal means and the problem that error terms can be systematic in one domain and random in another.

The ACE-FTS Error Budget (Patrick Sheese)

This presentation started with a caveat that results shown are both preliminary and very old because computer problems have caused considerable delay in this work. For ACE-FTS error propagation, for a sample of occultations different variables are perturbed by their expected uncertainty, and the errors are allowed to propagate through the retrieval. For ACE v3.5 retrievals, preliminary results for O_3, H_2O, NO_2, CH_4, using 100 sample occultations. Also numeric averaging kernels have been inferred by perturbation studies. Resulting averaging kernel matrices were not unity as one might expect for non-regularized retrievals. This issue had led to some discussion. Error sources considered were measurement error in terms of the inverse instrument SNR at each wavenumber, spectroscopic error, line strength and position uncertainties (from HITRAN 2004), tangent height error (assumed max of $\pm$0.5km, which might be too large, the influence of the initial guess profiles, the neglect of the consideration of the finite field of view (single ray versus 7 rays for integration over the field of view; the difference in the results was found to be about 5%.) The dependence of the results on the initial guess profiles is generally small. Propagation of pressure, temperature, instrument line shape errors still have to be assessed.

Average Averaging Kernels (Thomas von Clarmann)

Some users prefer to use averaged satellite data, e.g., zonal monthly means, as produced within the framework of the SPARC Data Initiative. For some applications averaging kernels might be required. The naive solution to this problem might be to generate zonal mean averaging kernels and to distribute them along with the data. This also avoids a lot of data traffic compared to the distribution of averaging kernels for each single profile. The problem, however is that, if the averaging kernel depends on the profile, the covariance between these has to be considered.

$$< \mathbf{Ax} >=< \mathbf{A} >< \mathbf{x} > -\mathrm{cov}(\mathbf{A}, \mathbf{x})$$

It was suggested to provide climatological estimates of $\mathrm{cov}(\mathbf{A}, \mathbf{x})$. By providing this term, the use of average averaging kernels can be justified, which would be user-friendly. Consideration of this term is particularly an issue when radiative transfer is nonlinear or when the retrieval is performed in the logarithmic domain. During the discussion the issue of the dimensionality of $\mathrm{cov}(\mathbf{A}, \mathbf{x})$ was raised. (Note by TvC added after the meeting: There is no conflict, because

$$\mathrm{cov}(\mathbf{A}, \mathbf{x}) = \frac{1}{\mathrm{n}-1} \sum (\mathbf{A}_{\mathrm{i}} - < \mathbf{A} >)(\mathbf{x}_{\mathrm{i}} - < \mathbf{x} >)$$

which has the same dimension as $\mathbf{Ax}$ and $\mathbf{x}$.) Further, the distribution of the smoothing error was critically discussed.

Machine Learning Methods (Stefan Bender)

Inspired by the book by Rasmussen and Williams "Gaussian processes for machine learning", (2006), machine learning methods were assessed for their applicability to error estimation. A lot of these methodes share their mathematical structure with those of retrieval or error propagation. The main problem seems to be that machine learning aims at prediction rather than the "internal state". The advantage seems to be that machine learning methods might help to construct a suitable a prior covariance matrix for an optimal estimation like retrievals.

General Discussion of Deductive Error Analysis (All)

There was agreement that within the TUNER project, we should come up with a recommendation as to how deductive error analysis is best performed such that error estimates from multiple instruments are comparable. Deductive error analysis was found to be best be the domain of the ISSI-TUNER project, because this includes the satellite data scientists.

4. Inductive error analysis

Inductive error analysis is understood to be error analysis based on the observations and is thus closely related to validation. Inductive error analysis can help to judge if the choices and approximations made in deductive error analysis are adequate. A number of presentations were given.

Natural variability: The Universal excuse in the validation (Arne Babenhauserheide, Quentin Errera and Thomas von Clarmann, presented by TvC)

When validation studies show that the standard deviation of the differences between two datasets of collocated measurements is larger than the combined error bars of the two data sets. This discrepancies are often attributed to less than perfect collocations and natural variability. The purpose of this study is that to provide a tool to test if this universal excuse is valid. To obtain statistics of the natural variability as a function of temporal and spatial mismatch, model calculations have been performed with the BASCOE (BIRA) model. The horizontal resolution was 1x1 degrees; the vertical resolution was that of ERA interim. ERA interim analyses were also used to drive the model as a CTM. For one week at the end of January 2009 hourly fields of a lot of species have been made available and resampled on regular z and p coordinates. A lot of variability is seen even in region where a quite calm atmosphere is expected. It is planned to use these fields to develop parametrizations of the mean difference of the state variable under investigation as a function of spatial and temporal mismatch. This parametrization can then be used to estimate which fraction of the hitherto unexplained discrepancy between two datasets can be attributed to natural variabilty. Horizontal and vertical resolution of the datasets have to be considered. The main issues raised in the discussion were if one week in January 2009 is sufficiently representative and if it would be worthwhile to use potential vorticity as a third independent variable of the parametrization besides spatial and temporal distance.

Recalibrating precision (Thomas von Clarmann)

Typically the standard deviation of the differences $\sigma_{\rm diff}$ between two collocated data products is larger than the estimated error of the differences, sometimes even after consideration of natural variability $\sigma_{\rm nat}$ (see above). This then hints at problems with at least on of the precision estimates. When only two data sets are compared, there is no way to identify which data set's precision bars are too optimistic. For three instruments, however, correction factors c_1, c_2, c_3 can unambiguously be obtained by solving the following system of equations:

$$\begin{aligned}
c_1\sigma_1^2 + c_2\sigma_2^2 + \sigma_{\rm nat;1,2}^2 &= \sigma_{diff;1,2}^2 \\
c_1\sigma_1^2 + c_3\sigma_3^2 + \sigma_{\rm nat;1,3}^2 &= \sigma_{diff;1,3}^2 \\
c_2\sigma_2^2 + c_3\sigma_3^2 + \sigma_{\rm nat;2,3}^2 &= \sigma_{diff;2,3}^2
\end{aligned}$$

For four and more data sets, the system is even over-determined and the solution is hoped to be more robust. In the discussion it was mentioned that this approach is clearly predicated on accurate estimates of the natural variability terms.

5. **What do the Data Users Need?**

   There is a general following conflict: On the one hand, this study should be "user-driven", which is that TUNER shall provide the data users with the information they need. On the other hand, it is common experience that data users often are not aware which diagnostics are relevant for which purpose. Particularly averaging kernels and error correlations leave a lot of data users clueless, and these data users would not request these diagnostics because they are not aware of their importance. In this case to educate the data users such that they finally want the right things would be more adequate than fulfilling their immediate wishes.

   The solution could be to ask the data users not about the diagnostic data they want but what they typically intend to do with the satellite data. The retrieval scientist should then decide which diagnostic data need to be considered. For typical applications the correct use of diagnostic data could be demonstrated in a tutorial paper.

   A questionnaire could be used to find out which are the most common applications of the data. This questionnaire could be sent with highest priority to the liaison scientists of other SPARC activities and to scientists involved in data merging activities.

6. **Other Business**

   (a) **The TUNER ISSI Project**

   TUNER was proposed as an ISSI project with 12 people allowed in previous year but was not selected. Resubmission to ISSI in 2017 has been successful. While TUNER-SPARC is larger and involves more people, TUNER-ISSI is limited to 12 scientists. A challenge has been how to make a fair decision as to who from TUNER SPARC should be in TUNER ISSI. For several reasons the same participants were included which had been involved in the TUNER ISSI 2016 proposal. There is still the possibility to invite additional experts, depending on what the focus of the first ISSI TUNER meeting can be. Possible dates for the first TUNER ISSI meeting in Berne will be determined and a doodle poll will be set up, with priority around end of November 2017. The agenda could foresee one day of presentations and then three days of workshop work where deductive error analysis is thoroughly discussed. A website for ISSI TUNER will be made.

(b) **The AMT Special Issue**

TUNER results will best be reported in journal papers, and this way of reporting seems to be agreeable for the SPARC officials. Positive signals were received from AMT executive editors that they would support an AMT special issue. This special issue could be open for multiple years and include articles on all the work performed under the umbrella of TUNER (over-arching and instrument specific) as well as external topically related work. There was agreement that the AMT Special Issue could be proposed immediatey now.

(c) **The SPARC SSG-Meeting in Seoul**

The SPARC Scientific Steering Group Meeting will take place in Seoul in 16-20 October 2017. Nathaniel Livesey will represent TUNER there.

(d) **Next Meetings**

One TUNER meeting could be held as a side meeting of the next Atmospheric Limb Conference in Greifswald in 2019. An earlier TUNER SPARC meeting would be useful. One option would be to collocate it with the SPARC General Assembly in October 2018 in Kyoto. No decision has been made yet.

7. **Action Items**

Not all the action items were explicitly mentioned during the meeting because the self-evidence of some of them was only recognized when the minutes were written.

(a) Push on instrument-specific work on error analysis (all)

(b) Push on work related to "Average Averaging Kernels" (TvC)

(c) Continue work on "Natural Variability" (AB)

(d) Continue work on "Recalibrating Precision" (TvC)

(e) Design data-user questionnaire (not assigned. Volunteers?)

(f) Set up doodle poll for TUNER ISSI meeting (TvC)

(g) Prepare TUNER ISSI meeting (TvC)

(h) Design TUNER ISSI website (TvC; delegated to Andrea Linden)

(i) Design project logos for TUNER SPARC and TUNER ISSI (NL and PS)

(j) Propose AMT Special Issue (TvC)

(k) Prepare TUNER presentation for SPARC SSG-mtg (TvC, NL, DD)

(l) Represent TUNER at SPARC SSG Mtg (NL)

(m) Find possibilities for the next TUNER SPARC meeting (Tvc, further suggestions welcome)