Review of "Estimating and Reporting Uncertainties in Remotely Sensed Atmospheric Composition and Temperature" by Thomas von Clarmann et al.

On reading this first time I got the impression that the authors were biassed towards one kind of remote sounding - spectrally resolved limb sounding. It is not until the end of the Recommendations that I found the reason: "These recommendations have been developed from the perspective of mainly satellite-borne limb sounding and occultation observations but some of these concepts are equally applicable to other types of remote sensing missions." This limitation should be stated right at the start, at least in the introduction, if not in the abstract or even the title. Better still they should attempt to write a paper about generic remote sounding.

Another impression I got is that in places there is a bias against the use of regularisation, without mentioning the disadvantages of not using it. This seems to be irrelevant to the apparent aim of the paper, namely objectively developing a unified scheme for reporting the characteristics of remotely sensed data.

Done properly, the description of any information-conserving retrieval in terms of a linearisation point, an averaging kernel and an error covariance matrix contains all of the information of the measurement in a unified and standard way, such that the user does not need to know much about the instrument. It is simply a given linear function of the atmospheric state, about a given linearisation point, together with a given error covariance, valid in so far as the forward function is linear within the error bounds. As such it could be used as data for any kind of further retrieval or transformation that the user may wish to do. (It may be sensible to divide the error covariance into different independent sources in cases where there is temporal correlation between different measurements.) It is not important whether the retrieval is ML or MAP based.

The paper is much too long, though I am at a loss to recommend in detail how it might be shortened. A lot of attention to removing unnecessary detail would help. A couple of sections - Onion Peeling and Chahine's Method seem a bit obsolete. Are they really needed?

Specific comments

Page 3 line 8: In what way are the characterisation schemes described in Rodgers 2000 not of general applicability?

Page 3 line 27: It seems to me that this requirement is not possible. The requirement should be that means should be provided for transforming error estimated errors from one grid to another.

Page 4 after section 2: Should there be a requirement that spatial and temporal correlations in errors between retrievals should be described? This is mentioned in the recommendations at the end but I feel it should be in the conditions of adequacy.

Page 4 line 24: sentence starting “The interested reader” and ending on line 29. I got lost in the sentence. Use a list papers and then say a few words about each reference.

Section 3.1: This whole section reminds me of angels dancing on pinheads. Various committees are producing different varieties of camels, to mix metaphors. (This is a complaint about committees, not about this paper)

Page 6 line 3: "two different ways to evaluate this quantity." It seems to me that one of these ways is a way of validating the other.

Page 7 line 7: Insert "an explicit" after “because"

Page 9 line 12: "holds" is a bit vague. How about something like "has a Bayesian interpretation”

Page 10 section 5.1: You should remind the reader here that discretisation itself is a kind of regularisation, and if the discretisation is too fine, further regularisation will needed to deal with the ill-posedness of the inversion.

To say that in a maximum likelihood retrieval, the grid width is identical to the spatial resolution of the retrieval may be formally correct but can be misleading. A fine structure whose amplitude is less than the retrieval noise is not usefully resolved.

Page 12 line 18: "problems" problems is not quite the right word perhaps “considerations” would be more appropriate. These are not problems unless they are ignored, dealing with them is part of the process.

Page 15 line 2: Using layer values instead of level values should not make much difference unless the layers are made thicker. It is simply a different kind of representation with the same number of unknowns.

Page 18 line 7: Marquardt was aware that using $\lambda\mathbf{I}$ lead to various kinds of problems. He suggested using a scaling matrix $\mathbf{D}$ with different elements down the diagonal. Many variants of Levenberg-Marquardt are available in the literature.

Page 19 line 8: should “inaccurate instrument model” be included in this list or is it implied by item 1?

Page 20 bottom paragraph: Another problem with residual-based noise characterisation is that systematic spectroscopic errors will lead to systematic retrieval errors, whereas random noise will lead to random retrieval errors. The user needs to be able to distinguish these. Equation (19) applies separately to each independent source of error which can be described by a covariance matrix. Examination of the spectral residual is an important part of validation of the instrument error model.

Page 21 rest of 6.1.1: Again, problem (line 2) is not quite the right word. Propagated noise from preceding retrieval steps may be formally dealt with internally as a parameter error, but as far as the data user is concerned it is still noise and should be presented as such.

You need a clear definition *from the users point of view* of what is meant by noise and what is meant by parameter error. It seems to me that anything which is uncorrelated between successive measurements is noise (primarily as a result of detector noise), and anything which is correlated, or even constant, between successive measurements should be classified as a parameter error. From the user's point of view internal details of how you do the retrieval should not matter, but the experimenter needs a formal way of calculating it correctly, the analysis should treat the retrieval as a whole, however it works internally.

My overall impression of section 6 is that it is something of a ragbag, going into too much detail in the case of some kinds of instrument while ignoring other kinds of instrument almost completely. E.g. nadir sounders, radiometers, GPS occultation.

Page 23 line 29: I suggest you replace "but are not part of" by “but are not usually thought of as part of"

Page 23 bottom line: I do not see a list of "the most prominent auxiliary data uncertainties", but just a discussion of pointing.

Page 24 lines 20 and 21: this only applies for some kinds of instrument.

Page 28 lines 24-26: "one": there will normally be an infinite set of atmospheric states which are consistent with the pure measurement information, not just one. This is true regardless of the number of levels used in the retrieval representation. The atmosphere is effectively a continuum. The purpose of the grid is to reduce the dimensionality of the problem, so as to search for a solution in a finite a dimensional sub-space. Thus a grid is just another form of regularisation. Furthermore, there is no particular reason for choosing a sub-space defined by a grid, any nonsingular representation could be chosen. Even if a grid is chosen, the implied interpolation rule must be specified. A linear interpolation, for example, introduces non-physical gradient discontinuities.

Any proper retrieval method with its correct characterisation will conserve information, as long as the problem is not grossly nonlinear, hence there is no “degradation” of the data, and the grid id fine enough to represent all singular vectors of the weighting functions for which there is non-neglible signal to noise. The retrieval can always be adjusted to allow for a required change in (or removal of) prior data, if used. Think of the characterisation as the forward model for a new retrieval.

A retrieval method that does not use a priori can be distorted by large errors and degraded by resulting in retrieval noise so large that the error analysis no longer remains in the near-

linear region. It also discards information available in the measurement by using a too coarse grid.

Page 28 line 25: The meaning of resolution “with respect to the true state of the atmosphere” is unclear. We do not know what the true state of the atmosphere is. Resolution is a property of an instrument, including its noise characteristics, conceptually it is the spacing of two delta functions that can be distinguished from each other, and that depends critically on both the size of the delta functions and the noise of the instrument. Hence we need to know something about the variability of the atmosphere before we can make sensible judgements about resolution.

Page 29 line 29: you should define precisely what you mean by biased. There will only be bias if the prior state and covariance matrix, if that is what is used, are inadequate. Any bias would be towards the prior state only in regions where there is not much real information from the measurement.

It might be worth including somewhere the usual warning about averaging retrievals which include a priori information.

Page 30 line 1: These so-called distortions occur generally in areas where there is no real information in the measurement. The user should be aware of this.

Page 30 line 12: I am unhappy about the use of the word “criticised” in this context. The use of (28) is quite appropriate in its own context. This section is not “criticism” as you usually used in English but rather a discussion of the appropriate context. A discussion of retrieval grids and interpolation in the context of prior data can be found in Rodgers (2000) section 10.3.1. (Note the obvious error in the heading for section 10.3.1.3.)

Page 30 paragraph starting line 19: the interpretation of the matrix $\mathbf{S}_a$ is not relevant to error analysis. The retrieval is some more or less known function of the true state. All that is required for error analysis are the derivatives of this function with respect to quantities that might have error components.

Page 30 line 24-26: I agree completely.

Page 31 lines 1-2: For a priori you have to use what information you have - that’s what prior information means. Like the measurement, it needs properly assessing and validating. The better your a priori the better your result will be.

Page 31 line 8: “true": perhaps "appropriate" would be a better word.

Page 31 lines 9-11: If they sample different parts of the atmosphere, you are not going to compare them.

Page 32 line 19 and 20: If profile retrievals are to be assimilated the proper way to use them is the characterisation of the retrieval as a smoothed version of the profile, with the averaging kernel and the appropriate error covariance. This includes all the information in the measurement, and nothing else. The a priori in that case is simply a linearisation point. However modern data assimilation systems use radiances not retrievals. It makes validation and allowance for temporally correlated errors easier.

Page 33 line 10: Please state what you mean by the vertical resolution of the instrument itself. If you mean the vertical spacing of observations for limb sounders then please remember that not all instruments are limb sounders.

Page 33 lines 11-15: The vertical resolution is limited primarily by the physics of the measurement, as expressed by the continuous weighting functions, together with the instrumental noise. A profile representation such as a vertical grid should be chosen not to limit the resolution of the measurement.

It is easy to be misled by the nature of a limb sounder to think that the resolution must be something to do with the measurement spacing, and hence grid spacing (I think this is what you mean, but it isn't clear). Different points in the measured spectrum will have different weighting functions, and consequently it is quite possible for there to be information on a finer vertical scale. Noise matters too, because it can hide the structure you are trying to resolve. To understand the useful resolution of a system you need to have some idea of the size of the structures you want to see.

Alternatively if you mean that you choose a grid spacing so the the problem is formally over-constrained, then you have to choose so that the problem is not unstable, and hence the choice of grid size becomes your method of regularisation. Noise and resolution are intimately linked by a trade-off.

Page 34 line 5: The Backus-Gilbert spread was developed in order to design a retrieval method which optimises resolution; this is why it was designed to suppress sidelobes. I do not understand why you think that a retrieval on a finer grid will produce more pronounced sidelobes.

Page 34 line 18: You can say this more simply by stating that the averaging kernel characterises the vertical resolution of the difference between the retrieved profile and the a priori.

Page 34 line 25: You could append "but can result in noise which makes the retrieval useless"

Page 35 line 1: Rodgers (2000) uses the term degrees of freedom **for** signal. This is associated with the degrees of freedom for noise, and is not particularly to do with retrieval. The sum of the two is the total number of degrees of freedom for the measurements, m.

Degrees of freedom for signal is a good guide to a suitable number of levels to use in a grid, or number of elements in some other representation.

Page 35 line 16: Please define what do you mean by bias and distort.

What normally happens is that in a part of the profile where the ML retrieval has low noise, an MAP retrieval on the same grid gives the same value and the same averaging kernel. Where the noise is large, the MAP retrieval moves smoothly to the a priori, while the ML gives only noise. Which you prefer depends on your application. You can have “bias” or noise.

Page 35 line 23:"Systematically": I presume here you are referring to ensembles of retrievals rather than a single retrieval. Perhaps you should give the usual warning about averaging retrievals containing a priori. I.e. the a priori component should be removed before averaging.

Page 35 line 25: This concept goes back to Backus and Gilbert who use the constraint that the sum of the rows of the average kernel matrix should be unity in developing their retrieval method.

Page 36 line 17: "fake": only if the retrieval method has been badly designed, for example by using an unsuitable prior covariance matrix.

Page 36 line 26: What is the case under discussion?

Page 36 line 29: This line appears to mean “in order to avoid problems due to regularisation, regularisation by means of a coarse discretisation is used”.

Page 37 line 7: It is not clear how this would be done.

Page 37 line 15: The English is a bit peculiar here. How about "Even if the prior information can be conceived as the….” See also comment on Page 31 lines 1-2. Any deviation of your estimate of instrument noise covariance from reality is likewise a source of error.

Page 39 last sentence: I strongly agree; should this have been in the conditions of adequacy?

Page 42 line 16: Sensitivity studies don't necessarily have to be based on delta functions, this may sometimes lead to numerical problems. Any kind of suitable complete set of orthogonal functions will do.

Page 42 line 20: I suggest that "ideally the data provider" be replaced by "the data provider must”.

Page 42 line 23: I suggest adding somewhere "For retrievals given on a grid, the implied interpolation scheme must be specified”.

Page 42 line 31 onwards: If smoothing error is reported, the a priori covariance on which it is based should be given. However I am inclined to think that you should recommend that smoothing error should not be reported, for the reasons you give.

Page 43 line 13: Precisely what is meant by a “profile representation” must be defined somewhere. It not only includes values on a grid, it also includes an implied interpolation between the grid points. This interpolation must be specified.

Page 44 recommendation 18: This does not sound like a recommendation. Either construct a recommendation from this discussion, or separate it as an explicit discussion section.

Page 45 line 5: This mention of satellite limb soundings and occultation observations is important it should not be left as a throwaway comment at the end of the paper, it should be mentioned right at the start. Preferably the recommendations should have been developed from the perspective of remote sounding generically.

Minor points

Page 2 line 3 “shall” should be “should”.

Page 3 line 20: The sentence starting “We refer to diagnostic metadata” is unclear. I assume it means something like "By diagnostic metadata we mean”

Page 4 line 3 Insert "proper" before combination.

Page 4 line 8: QA4EO should be spelled-out at its first use.

Page 5 line 32: This sentence is unclear, I assume it means "no particular terminology" rather than "not having a terminology".

Page 7 line 21: “$m>n$” does not necessarily imply over-determined. It is quite possible for the rank of the Jacobian to be less than n if weighting functions are not linearly independent. (I’m nitpicking)

Page 12 line 30: Limb sounding was first used to my knowledge for Mariner Mars in the 1965 in the form of radio occultation sounding

Page 24 line 20: Replace "makes the" by "computes a”.

Page 24 line 22: Knowledge **of** radiative transfer

Page 24 section 6.2.1: Do we need a single subsection within a section? This comment also applies to section 6.4.1.

Page 25 line 23: Replace "along" by "according to” or “by”.

Page 34 line 15: Delete comma after retrieval.

Page 37 line 30: Replace "exclude that they all are" by ‘assume that they are not"

Review of “Estimating and Reporting Uncertainties in Remotely Sensed Atmospheric Composition and Temperature" by Thomas von Clarmann et al.

Addendum

Strictly, the averaging kernel and the weighting functions are continuous functions. The discrete versions, with an interpolation rule, are approximations.

In the near-linear case, the continuous averaging kernel is a linear combination of the continuous weighting functions. A fine enough grid should be chosen to approximate it well - the singular vectors of the weighting functions are a good guide. It can be approximated onto a coarse grid such as one used for a ML retrieval, but it cannot be restored by interpolation from that grid, as its fine structure is lost. For example, the fine grid averaging kernel for a coarse grid retrieval with linear interpolation is definitely not triangular, as the interpolation rule might imply.

Data users should be able to access the averaging kernel computed on a fine grid so they can evaluate smoothing error on whatever grid they need. This applies to any retrieval method, not just ML.