I have to confess that I am still puzzling what was the real intention of the authors in submitting this long and, to some extent, verbose report for publication to AMT.

Although I appreciate the effort in contributing to simplify the exchange of L2 data and explain their error characteristics, in its present version the paper seems just an occasion for the many authors to recount and self-reference what they did in the area of inverse/retrieval algorithms for the sounding of atmospheric parameters.

The title seems to open to a wide tutorial, however at the end of the abstract they say the goal of the paper is just to provide a list of *recommendations which shall help to unify retrieval error reporting*. In section 3, it seems that the authors want to redefine terminology about errors. Do we have to call the root mean square error, simply uncertainty? And the variance, precision? Or whatsoever? Do we have to stick to new definitions issued by JCGM and BIPM? Is it a problem of terminology or contents? Or simply, do authors want to set up a sort of protocol for exchanging L2 products? By the way, in the end, I count 6 CoAs and 18 (with subpoints) recommendations, for a total of 24 and more. To me, more than 3 recommendations are effective as no recommendations at all. In effect, 24 recommendations are normally much more than the degrees of freedom or pieces of information conveyed by common retrievals.

Looking deep inside the paper, I can see interesting aspects about trying to define a common paradigm to interpret data coming from a large variety of satellite data processors. However, this objective is somewhat lost among unnecessary details of retrieval schemes, methodological issues and what I could call a silent but insistent criticism to Optimal Estimation. Furthermore, I think that the format of the present study is much more adequate for a report.

Concerning *retrieval error reporting*, the canonical Theory of Statistics has been teaching us (e.g. Kendall and Stuart Vol I, II, III, *The Advanced Theory of Statistics, Fourth Edition, 1979*) for so many years that the performance of a given statistic or estimator, say $\hat{x}$, is measured in terms of its mean square error or deviation from the *true value*, which can be decomposed in variance and bias, namely

$$E[(\hat{x}-x)^2] = E[(\hat{x}-\bar{x})^2] + E[(x-\bar{x})^2]$$

For the assessment of the root mean square error and its reporting, the consolidated usage is today to share and/or distribute.

1. Estimated state (of course) and related retrieval covariance matrix
2. Background (state and covariance)
3. Averaging Kernels

Based on the above items, the performance of any estimator (bias and variance) can be unambiguously quantified. From what I can see, in the end, the above three ingredients are what authors agree with to be the basic items to share. In this respect, a potential list of recommendations, included that of authors, could be made and explained in one-two pages.

I have also to say, that authors' recommendation list itself is largely independent of the bulk of the present paper.

General Remarks

The paper is lacking a correct definition and assessment of bias. Authors seem to identify the random component of the root mean square error as the error or uncertainty of a given retrieval system. What about the bias? What's the strategy they want to set up to estimate it and eventually share with end users?

I have found a bit confusing the question about grid-independent retrievals, which for me is a non-sense, since normally one works with a discretized state vector. Apart from forward model (FM), the bias depends on the given constraints, which are normally grid-dependent, in the sense that their definition and use is contingent to the way the state vector has been discretized. In effect, for a regularized estimator the bias depends solely on the constraints (again apart from FM biases). This basic aspect has been largely overlooked in the paper, and in fact their recommendations are not consistent with a correct sharing of the root mean square error.

On the same line, their CoA2 is inconsistent with the idea of root mean square error. Furthermore, I am not sure if it can be implemented, in practice. To streamline my personal thinking, let's suppose W is a suitable interpolation/extrapolation operator, which transforms a given estimator $\hat{x}_{n1}$, defined on a grid with n_1 layers, into a new one, say $\hat{x}_{n2}$, defined on a grid with n_2 layers, we have

$$\hat{x}_{n2} = W\hat{x}_{n1},$$

with W a matrix of size $n_2 \times n_1$. CoA2 requires that, using authors' language,

$$WS_{x,noise,n1}W^T = S_{x,noise,n2},$$

where, $S_{x,noise,XX}$ is the error covariance directly retrieved on the grid with XX layers. However, I cannot see how the above condition can be met for any choice of W and $n_1 \geq n_2$or $n_1 \leq n_2$. Atmospheric state vectors are not band-limited signals, therefore a mere extrapolation/interpolation of a given retrieval from a coarser to a finer grid will not show finer structures of the underlying state. Hence, the above condition would normally not be met.

Condition CoA2 seems to have been set up just to criticize the concept of smoothing error, which is the way Rodgers considers for the bias. Since the bias of the individual, single, retrieval depends on the true value, which is normally not know, Rodgers considers the variability of the true value (variance-covariance) in order to have at least an estimates of the interval in which the bias is expected to range. However, the variability and/or stochastic behaviour of the state vector, which is correctly considered in OE, is overlooked by authors. They say, "natural variability is not a genuine retrieval error". It seems to me that authors purposely mislead statistical error with mistake. Natural variability is what makes our weather to be forecastable, but not exactly predictable. This is why we need statistics to address natural variability.

Taking into account the natural variability of the state vector, it is possible to perform an assessment of the estimator's bias, e.g., through the (unfortunately named) *smoothing error*, whose meaning has been, in fact, completely mislead by authors (see also later when dealing with the smoothing error).

Finally, because of the many issues addressed in the paper, in the end it looks like a confusing revision of Rodgers 2000; a sort of *pout-pourri* of about everything is known today on atmospheric inverse problems: Twomey, Tikhonov, Rodgers, LS, ML. Furthermore, the estimator described in Eq. (4) in the text is not rigorously derived from any basic principle of statistics, it is just copied from OE and rewritten by substituting S_a^{-1} with R.

Specific Comments

Pag. 3. At best, CoA2 is only consistent with the variance component of the estimated error. What they want to do with the bias is not clear. Stand as is, I have doubt CoA2 is effective and can really work.

Page 4. Section 3.1 This is confusing. Please state exactly why uncertainty cannot be used or why it sounds ambiguous if referred to the root mean square error of an estimator.

Page 6, Eq (3), I cannot see any point why the unconstrained Least Squares solution should be called "Maximum Likelihood". This is a misconception. The assumption of Normal pdf is what really qualify the estimator (3). The reason of using ML because it yields LS under normality is untenable; it is like saying that a meteorologist is using Einstein General Gravity (EGG) theory when forecasting the atmosphere with the Newton dynamical equations, because EGG retrieves Newton in the limit of low velocity. Why do authors not qualify the bias and variance of the estimator? Why the reader has to wait until section 6, just to see the variance alone of the estimator.

Page 7, Eq. (4). This is the worst part of the paper. Equation (4) is the OE estimator where S_a^{-1} has been substituted with R. In force of this unjustified and ad-hoc substitution, authors claim that the estimator (4) becomes more flexible and powerful than the OE shown in Eq. (6). Also, in this case the variance of the estimator has been presented to the reader in instalments; first Eq. (7) and then an incredible jump to go to Eq. (18). In addition,

a. The bias of the estimator is not qualified/assessed/quantified in any part of the document
b. What is the reason to change S_a^{-1} with R? What are the expected improvements?
c. Why has the Tikhonov-Twomey regularization γ-parameter disappeared? That is why not γR?
d. What's the role of x_a, and why not x_0 as in Eq. (3)?
e. With R set to any of the suggested matrices, 0-1-2 order difference matrices, Eq. (4) is dimensionally inconsistent. The authors seek a protocol-independent of constraints and other assumptions, but they propose to use an estimator, which is dimensionally inconsistent and depending on the units used for the state vector. In which way do they achieve dimensional consistency between the two terms in the squared brackets?

It would be much fairer to say "Equation (4), as well as Eq. (3) (e.g. global fit), has been normally in use for the retrievals from satellite-borne limb sounding and occultation observations. It is here considered because still now many satellite processors rely on it. Or something similar. The description of the various estimators, LS, TT, OE should be as much as neutral and respond to the need to just explain their error characteristics.

Page 7, line 30. What do you exactly mean with *smoothed?* What is a *smoothed profile?* How smoothing is quantified, and why this is a good property. In comparison to estimator (3), estimate (4) is biased and the bias structure is determined by R, which is grid dependent. So, how the estimated errors can be propagated according to CoA2? What is the solution proposed by authors: just forget about bias?

Page 8. Eq (6). Now that the authors have invented R, they can say *our estimator* retrieves the OE estimate if we put $R = S_a^{-1}$, unbelievable! By the way, to me, to $R = S_a^{-1}$, is the only possible choice, if we want to reach dimensional consistency.

Page 8, paragraph beginning at line 8. This comment seems to stay here just to add some references. By the way, it is not appropriate for Eq. (6). This is a comment to be added soon after Eq. (5). It does not apply to Eq. (6), in fact, OE elegantly solve the problem of high dynamic range of the state vector, because S_a has the right dimension to properly scale the state vector. As shown in many papers, OE can be solved for the scaled variable $\tilde{x} = S_a^{-1/2}x$, which is equalized to a standardized variate, at each layer.

Page 8, Eq. (7) and discussion after. Here it seems that an *essential role in error estimation* is played by the variance of the estimator alone, and the bias? Once again, how the bias of estimator (4) is qualified/assessed/quantified?

Section 5. All is said in this section is today overcome by *Simultaneous Retrieval.* Section 5 is out-of-date and should be totally removed.

Section 5.4.5 Still Onion Peeling?

Section 5.4.6 See point above. I recommend a CoA0: Please forget about ad-hoc and non-optimal methods!

Sections 6.1 to 6.3 can be summarized under a very short section entitled “Instrument Noise and Forward Model bias”

Section 6.4. Authors here simply miss the important point that the Averaging Kernels matrix, A, qualifies and serves to assess the bias error, at least the part coming from the background constraint. In fact, if we take expectations on both side of Eq. (25) all random components associated to the instruments are averaged to zero, and we remain with the expectation value, $E(\hat{x})$. Systematic component, originating from the forward model, can be dealt with appropriate transforms of the radiance vector, e.g., *random projections*.

Section 6.4.1. All the verbose premise of the paper points straight to this criticism of the smoothing error. However, the only thing which is fairly criticisable here is the word *smoothing. In fact, s*mooth, smoother and similar terms should be banned from the context of error assessment and analysis. If Rodgers had said the retrieval can be regarded as a **biased estimate** of the true state, then everything would have gone to the right place. In effect, the smoothing error is the missing bias term to be added to the variance in order to have an estimate of the root mean square error, $E[(\hat{x}-x)^2]$. In principle, there is no need to interpolate/extrapolate to different grids a given state vector for the purpose of comparison. For visual inspection, one can just plot the given estimators and confidence intervals on the same plot, using the proper pressure-altitude grid. Why the quest of plotting differences?

Pag. 27 and 28. Eq. s (28) and (29) can be left to more elaborated comparisons. There is no need to cover this aspect in the present paper.

Pag. 28. Eq. (30). What do you mean “better resolved”? Please, quantify. The paper is aiming at providing recommendation, this cannot be given in terms of ambiguous qualitative terms.

Pag. 6.4.3 From section 6.4.3 on, until section 7, the paper appears to be unnecessary long.

Section 7. As said at the beginning 18 recommendations are too many to be useful.

Table 1 and Table 2. I do not understand the scope of these two tables. If authors want to provide a list of official L2 data providers, the list is too long since it should show only Agencies. If the authors want to provide a list of the many scientists dealing with Satellite Data Processors, it is too short.