Review #1:

The authors thank the reviewer for their careful and thorough reading of our manuscript and for their constructive comments.

(1.0) **Review: In their paper Thomas von Clarmann and co-authors provide a list of recommendations on how to report on errors, based on the activities of the TUNER project. To my opinion this is a very important and timely topic, and I acknowledge the effort made by the author team to write a dedicated paper discussing this point. I found the paper interesting, but also had two major reservations and a substantial number of comments, as detailed below, which require a major revision of the paper before publication.**

Reply: We thank the reviewer for this encouraging overall appreciation. However, we suspect that it remained somehow unnoticed that the paper was submitted as a review paper. This, we hope, justifies several sections which might not be necessary in an original research paper, and may make some of the criticism obsolete.

Planned Action: We will metion in the introduction that this paper is a review paper.

(1.2) **Review: Major general comments:**
My first major reservation: the purpose of the paper is the formulation of the list of recommendations for a more uniform and complete error reporting in level-2 satellite data products (see the last line of the abstract and section 7). However, the bulk of the material presented in the paper is basically a review of real-world implementations of optimal-estimation based (and related) profile retrievals. As such the authors could consider to split the document into two papers, a review of profile retrievals and a shorter more focussed paper about unified error reporting.

Reply: Retrieval and error estimation are intertwined. The relevance of error sources can depend on the retrieval scheme chosen. We have thus intended to lay down the entire framework in which the error estimation takes place. It is often the real-world implementation which gives rise to some errors in the retrieval. Thus we think that the detailed description of the real-world implementations and the error reporting issue should not be torn apart. The consideration of various specific retrieval implementations makes the difference between our paper and existing literature. We concede that the link between these sections and the recommendations should be made clearer.

Planned Action: Instead of splitting the paper we will try to make the logical connection between the sections of the paper clearer.

In order to strengthen the link between the retrieval section and the error reporting section, additional section references will be included.
On page 3, l1, we will add: “[We then systemize and discuss the various sources of retrieval error] **and, if applicable, their dependence on the retreval scheme chosen**[(Section 6)]”
On page 18,Section 6.1.1., l. 15, we will add: “[The situation is different in a decomposed retrieval] **(Section 5.4)**”
On page 18, Section 6.1.1., l22, we will make a direct link to Section 5.5: “[In the context of error propagation in the the Levenberg-Marquardt algorithm] **(Section 5.5)**[, it is important to distinguish two different applications]”
On page 21, Section 6.2.1, l30, we will add “[by the radiative transfer model $\vec{F}$ in use] **(Section 4)**”
On page 24, Section 6.3, l4-5, we will add “[... is already known] **(see, Section 5.4)**”.
On page 24, Section 6.3.1, l23, we will add ”[In onion peeling] **(Section 5.4.5)** [the ray...].
On page 25, Section 6.4: There is already a link to Eq. (4).

(1.3) **Review: Several sections of the paper are providing useful functional background information for section 7. But for quite some sections I could not find the link with the final recommendations. Examples are section 5.4 and also parts of 4, 5.1 and 6.4 (e.g. 6.4.3 to 6.4.6), 6.7. Because of these sections the paper is very long.**

Reply: Section 4: Eqs 1-4 lay down the notational basis for the rest of the paper. The text on maximum likelihood approaches seems important to us because the current literature is strongly biased towards Bayesian methods. Our intention has been to get the community together and to form a common framework.
Section 5.1 Discretization: Here we introduce concepts where the result is represented in other ways than value over vertical coordinate. Without introducing this, Recommendation 11 would be incomprehensible.
Section 5.4 Decomposition: Without decomposition of the retrieval problem there would be no parameter errors at all, and different decomposition approaches require different ways to estimate the retrieval errors. Thus we consider this section as essential.
Section 6.4.3 Altitude resolution: We consider to add an amendent to Rec. 16: “If averaging kernels are only provided for a few representative cases, one might still consider to show the vertical resolution profiles for each profile.”
Section 6.4.6 Regularization crosstalk: This is an often overlooked error component and may be essential for a complete error budget.
Section 6.7 Drifts: This section is indeed not referred to from anywhere else in the paper, but without it, we are pretty sure, we will be blamed to have forgotten this issue.

Planned Action: The links between these sections and the recommendations will be made clearer.

(1.4) **Review: Section 5.4 is a review of (profile) retrieval approaches, but contains a lot of material which is not directly relevant for the paper. Personally, I would propose to shorten this section, keeping possiby the tables (and references) and keeping those remarks which are important in the context of error reporting. This review-like section also leads to a very long list of references. It would be good to mention only those references that bring new information to the discussion how to present the retrieval errors.**

Reply: Well, this paper is indeed meant as a review paper. It is for this reason that we touch more topics than directly necessary for the recommendations and aim for a comprehensive list of references. Even if not all the sections are necessary to infer the recommendations, the information they provide is often still necessary to comply with the recommendations, i.e., to provide the data characterization as requested in the recommendations.

Planned Action: We will try to find a place in the paper where examples can be included which demonstrate the relevance of the material presented.

(1.5) **Review: In general, there is quite a big conceptual step (gap) between sections 3-6 and the summarising recommendations in section 7. Ideally all recommendations should be complemented by motivations, examples and explanatory information in the sections preceding section 7. The link between the recommendations and the rest of the text (which reads as a review of retrieval methods and theory) is often unclear to me.**

Reply: As stated above, the paper has indeed been submitted as a review paper. We think that retrieval theory and error estimation are fairly intertwined issues. Any change in the retrieval setup needs to be reflected by the error estimation approach.

Planned Action: The link between the recommendations and the theoretical parts will be made clearer.

(1.6) **Review: I would suggest that the authors go through sections 3 to 6 (see list of subsections mentioned above) and remove discussions which are not functional to motivate the requirements presented in section 7.**

Reply: We think that a section is justified not only if it motivates the recommendations but also if it provides the technical information needed to understand and to provide what the recommendations require.

Planned Action: Again: The link between the recommendations and the theoretical parts will be made clearer and more specific.

(1.7) **Review: After reading the first sections of the paper it was not fully clear to me what is really the problem which is addressed? In what sense are retrieval products not comparable? Please provide (generic) examples of retrieval products which miss information which makes a direct comparison between retrievals, or comparisons with independent data difficult or impossible.**

Reply: We agree that generic examples will be useful.

Planned Action: Examples will be included.

(1.8) **Review: In what sense is there a need for a new set of recommendations, e.g. what is missing after the work of e.g. QA4EO or the GUM?**

Reply: Regarding the treatment of uncertainty, QA4EO basically recommends to apply the GUM recommendations. GUM does not take constrained retrievals using prior information into account. It does not take the problems into account which are caused by the real-world retrieval schemes. Broadly speaking, it is not specific enough for our purpose. limitations of the GUM for remote sensing application were identified already in, e.g., Povey and Grainger, AMT, 2015, section 2.3, where they state "These conventions [from GUM] apply equally to satellite remote sensing data but represent an impractical ideal that does not help an analyst fully represent their understanding of the uncertainty in their data. This is due to the simplistic treatment of systematic errors."

Planned Action: We will expand on this in the paper.

(1.9) **Review: The final set of recommendations are focussing on profile retrievals. But the tables include also total column retrieval examples (e.g. DOAS). I think this is a missed opportunity, and I would encourage the authors to formulate explicitly what their recommendations are for column retrieval products (some recommendations are generic, but several parts of section 7 explicitly refer to profiles)**.

Reply: We agree.

Planned Action: For those recommendations which refer to profiles only, a respectice recommendation for column retrievals will be included.

(1.10) **Review: Arguably the atmospheric composition data assimilation community is the main user of satellite retrieval products. This**

community and their needs are basically not discussed in the paper.

Reply: The quantities needed by the data assimilation community are basically those we recommend to provide: Covariance matrices of the data uncertainties and averaging kernels. The former are needed for their observation error covariance matrix; the latter are needed for their observation operator.

Planned Action: We will metion the need of correct diagnostics for data assimilation.

(1.11) **Review: More generally the users of the data do not receive much attention, and the requirements are discussed from a L2 data provider point of view.**

Reply: This paper is indeed addressed to data providers. A tutorial paper addressed to data users is in preparation. There the correct use of the data characterization will be demonstrated. We agree, however, that the relevance of the correct use of the error estimates, averaging kernels, etc. in quantitative applications like data comparison, time series and trends, data assimilation etc. deserve to be mentioned.

Planned Action: We will add to the intro: "This review paper, the first 'foundational' paper from the TUNER team, is mainly addressed to the providers of remotely sensed data. A paper addressed to the data users, guiding them through the correct use of the uncertainty information, is currently being written (Livesey et al., in preparation)" In addition, we will mention some typical applications of satellite data along with their specific needs with respective to data characterization.

(1.12) **Review: This is my second major reservation. Some parts of the text refer to the validation activities, but this is not presented in a very structured way. The needs and feedback from the validation and assimilation communities on existing L2 satellite products would be an important starting point to discuss requirements for satellite data products.**

Reply: While validation and assimilation are not meant to be the main content of the paper, we agree that the relevance of correct data characterization is essential for these purposes.

Planned Action: We will mention the relevance of correct data characterization for validation and assimilation, and we will highlight the relevance of validation studies for the assessment of the adequacy of error estimates.

(1.13) **Review: Some assimilation users would prefer to work directly with the level-1b data, an option which is also not discussed.**

Reply: The problem of the assimilation of L1B data is that all errors (parameter errors etc.) have to be mapped into the radiance space to get the observation error covariance matrix right. In spectral measurements this will typically lead to fully correlated non-sparse covariance matrices, which are, to our knowledge, not favoured by the assimilation community. While direct radiance assimilation is successfully applied to nadir sounders, we are not aware of any application where, say, tangent altitude offsets are correctly dealt with in high-resolution limb sounding data. There is certainly a lot to say with respect to this issue, but we think that a deeper discussion of related issues is beyond the scope of this paper, which is already quite long.

Planned Action: We will mention in Section 5.2 that in direct radiance assimilation it is important to consider a measurement covariance matrix which does not only contain noise but the mapping into the radiance space of all uncertainties of parameters which are not assimilated.

(1.14) **Review: The recommendations in section 7 are not always formulated as a recommendation, but leave room for interpretation and implementation. I sometimes found the CoA points in section 2 even more clear and explicit than the recommendation points. It may be useful to split the list in section 7 in actual (strong) recommendations and related discussion points. Sometimes it is not so clear what is actually recommended by the authors, e.g. due to a trade-off between completeness and data volume, or aspects are left to the retrieval teams to decide (e.g. point 1, 2, 3, 4, 16, 18).**

Reply: The general problem is that the specific decision depends on the instrument and the retrieval approach chosen. The decision is under responsibility of the retrieval scientist. However, we provide criteria to judge if the decision was correct.

Planned Action: We will make clear that the choices are not *ad libitum* choices but must comply with our general rationale that the error estimates must explain observed inter-instrument differences.

(1.15) **Review: I was expecting recommendations also regarding the naming (see section 3). The authors discuss in particular "error" versus "uncertainty", but do not really provide a clear guidance on what to use..**

Reply: Whatever naming we would suggest, it would always be in conflict with a part of the community. We consider it as a major progress if awareness of the ambiguities of language is created, and if authors clearly define the language they use. Usually no language is *per se* better than another language, and we do not want to stipulate conventions. We want to restrict the recommendations to the objectively necessary.

Planned Action: We will make clearer in the paper that no language is *per se* better than another language, but authors should clearly define the language they use.

(1.16) **Review: Also, the consistency or inconsistency with the GUM activity are not clear to me after reading the section. The reader is referred to a paper in preparation.**

Reply: The criticism of GUM is controversial among the TUNER community. Thus we found it appropriate to defer these issues to a paper which is authored only by those who wholeheartedly endorse this criticism. We agree with most of the technical recommendations in GUM, although we think that these are not specific enough for our purpose. We do not all agree with their construal of the concept of 'error', and we do not all agree that we can dispense with the concept of the 'true value'.

Planned Action: We will check if we can make the related paragraph clearer without becoming biased towards or against some of the concepts under dispute.

(1.17) **Review: Retrieval datafiles contain parameters labelled as "precision", "accuracy", "trueness" etc. and different guidelines exist from different space agencies and for different application areas. It would be useful if the authors can discuss naming conventions also in this paper and express a clear opinion/recommendation.**

Reply: We have intentionally avoided this. One naming convention might not be more adequate than another one. We do not want to make prescriptions with respect to conventions but we have tried to restrict our recommendations to those that can be inferred from the conditions of adequacy.

Planned Action: In the paper we will make the rationale outlined above explicit.

(1.18) **Review: Machine learning approaches are getting more and more popular and deserve some special attention. Several machine learning implementations for retrievals are limited on the error information they provide. It would be useful to have some targeted recommendations for these approaches as well.**

Reply: There are indeed some relevant issues here. If in machine learning the machine is trained with retrieved data (from a conventional retrieval) then all uncertainties of the latter propagate on the regression parameters created by the machine-learning scheme. A variant of machine learning is supervised learning with neural networks. There are two distinct approaches to the use of neural networks in remote sensing. They can be used for the forward modelling of radiative transfer. In this case all the error analysis we describe is still feasible and valid. Or they can be used directly for the retrieval. Then the error estimation schemes presented in our paper are not easily applicable.

Planned Action: We will try to add a short paragraph on these issues but we do not want to add too much length to the paper.

(1.19) **Review: Detailed comments:**
Abstract: The abstract reads like an introduction. I would encourage the authors to summarise (shorten) the first part and expand on the last sentence with a summary of the content and main results of the paper.

Reply: We think that the abstract summarizes the main parts of the paper and the general ideas behind them.

Planned Action: none

(1.20) **Review: Introduction l6: "reduction"? Should this be "deduction"?**

Reply: We think "reduction" is correct here (in the sense of the technical term 'data reduction').

Planned Action: This issue may become obsolete because major parts of the introduction will be rewritten anyway.

(1.21) **Review: l16: "The project ... is a consortium of". Please modify**

Reply: Thanks for spotting.

Planned Action: The entire introduction will be rewritten.

(1.22) **Review: l24: "atmospheric composition and temperature profiles". What about other profiles, e.g. water vapour?**

Reply: We consider water vapour as a constituent of the atmosphere and think that it is thus covered by "composition".

Planned Action: none

(1.23) **Review: Is the paper limited to profiles, or are single property (column) retrievals also included?**

Reply: We do not see a fundamental difference. A column amount can be conceived as a profile containing one element. Some parts of the paper refer directly to column retrievals, e.g., the column averaging kernel (Eq. 26).

Planned Action: Particularly in the recommendations section more weight will be given to column retrievals.

(1.24) **Review: l37: "are do not need to be", please correct.**

Reply: Thanks for spotting!

Planned Action: This will be corrected.

(1.25) **Review: Section 2:**
l82, CoA 1: "and/or error estimation schemes". Would it not be better to say "and/or retrieval schemes"?

Reply: yes, indeed.

Planned Action: This will be reworded: "[The error estimates should be intercomparable among different instruments], retrieval schemes, [and/or error estimation schemes.]"

(1.26) **Review: CoA 2: "independent of the vertical grid". But I assume at this point that error covariances are specified on a specific grid used in the retrieval ?!**

Reply: Yes, they are. But generalized Gaussian error estimation applied to the resampling on other grids will produce the correct covariance matrices also on other grids. This is not usually true for the smoothing error, thus our critical position with respect to the latter. Needless to say that interpolation to a finer grid causes a sort of smoothing error but this is not what we understood should be included in the noise error. It is another category of error. Thus it cannot be expected to be rendered by the propagation of noise to a finer grid.

Planned Action: none

(1.27) **Review: CoA 5: "and different amounts of prior information". Do you mean "and different sources of prior information"?**

Reply: Actually we need both: amounts (weight) of the prior and the values of the prior itself.

Planned Action: We will reword this : "[...different amounts of] possibly different [prior information]".

(1.28) **Review: p3, l10: "but we consider it unrealistic to assign quality indicators for 'fitness for purpose' for all conceivable applications." This is an interesting remark. It would be useful to expand on this: explain how it is discussed by QA4EO and which parts are unrealistic.**

Reply: There is an almost infinite number of possible applications and purposes. Thus it is impossible to provide fitness-for-purpose indicators for all, and we consider it not useful to select a few of them on an ad hoc basis.

Planned Action: none

(1.29) **Review: Sec 3.1: please introduce the acronym "Joint Committee for Guides in Metrology (JCGM)" just once, and use only the acronym "JCGM" in the rest of the paper.**

Reply: Agreed.

Planned Action: This will be corrected.

(1.30) **Review: sec 3.1, l34: "actually claimed that there are conceptual differences between error analysis and uncertainty estimation." For readers who did not follow this debate it is hard to follow this section. It would be helpful to add a few sentences to list the claimed conceptual differences between these two terms.**

Reply: In wide parts of the literature - from Gauss to Rodgers and beyond - 'error' is used also as a statistical estimate of the absolute difference between the estimate and the true value. There are concepts which allow to estimate this quantity without knowing the true value (we say 'estimate', not 'know'!). GUM seems to ignore this connotation and it appears (however, this is not quite clear in their documents) that they refer to error only as the actual difference between the estimate and the true value.

Planned Action: We will try to find a clearer wording for this part.

(1.31) **Review: Section 4: I find it useful to include a section with the theoretical background and notation. In fact, using this notation could be a recommendation (Section 7, point 1).**

Reply: We are happy that our section on theory and notation is appreciated. In our recommendation #1 we write " We hope that this paper serves that [clearly defined language and notation] purpose". But since no notation is *per se* better than any other one, we do not feel to be in a position to dictate which convention others should use, as long as everything is clearly defined.

Planned Action: We will make this recommendation clearer by adding "[that purpose] and that the terminology and notation introduced here will be found useful."

(1.32) **Review: eq. 2: "can only be approximated" What does this refer to? The ill-posed or underdetermined nature of many inverse problems?**

Reply: yes, exactly. And beyond this, large rank of the matrix to be inverted, which will impose some practical limitations.

Planned Action: We will add "[only be approximated] due to the over- or underdetermined or otherwise ill-posed nature of the problem and the large rank of the matrix to be inverted."

(1.33) **Review: l70: "macrorcopic"**

Reply: Thanks for spotting.

Planned Action: This will be corrected.

(1.34) **Review: l77: What is the approximation which turns "f" into "F". Are these real-life uncertainties in f? Is F now a matrix or still a non-linear function?**

Reply: $\vec{f}$ represents the (unknown) true radiative transfer function, and $\vec{F}$ the model we use. $\vec{F}$ typically is a nonlinear vectorial function.

Planned Action: We will add: "$\vec{F}$ is a vector-valued non-linear function and deviates from $\vec{f}$ in that it involves numerical approximations and may not include the full physics of radiative transfer." We will correct **F** to $\vec{F}$ in the text.

(1.35) **Review: l87: "overdetermined case (m ¿ n)". Whether or not the inverse problem is overdetermined also depends on F, and not only on the size of the vectors. Add "and not ill-posed". (This is discussed on next page)**

Reply: We are afraid that here different terminologies clash. According to the conditions of well-posedness by Hamadard, every inhomogeneous over-determined problem without collinear equations is by definition ill-posed because it does not have an exact solution. Thus 'overdetermined and not ill-posed' is usually an unsatisfiable condition. We use the convention endorsed, e.g., by James E. Gentle, Numerical Linear Algebra for Applications in Statistics, DOI https://doi.org/10.1007/978-1-4612-0623-1, Springer-Verlag New York, Inc. 1998, Print ISBN 978-1-4612-6842-0, Online ISBN 978-1-4612-0623-1, Series Print ISSN 1431-8784, who states in Chapter 3, page 94: "However, many of the linear systems that occur in scientific applications are overdetermined; that is, there are more equations than there are variables, resulting in a nonsquare coefficient matrix." The Wikipedia article entitled "Overdetermined System" (retrieved 16 Jan 2020) they even state implicitly that collinearity ot the equations is not in conflict with overdeterminedness.

Planned Action: none

(1.36) **Review: p5, l7: "In most real-world applications, only measurement noise is considered here, while other measurement uncertainties like calibration errors are neglected at this stage." Remove "here" and "at this stage"**

Reply: We have added these words with intention. Otherwise the reader might think that we claim that other error sources are omitted also in the error propagation.

Planned Action: none

(1.37) **Review: p5, l21-44: This is an interesting historical note, but not essential for this paper and may be removed.**

Reply: Since this paper is intended to be a review paper, we think that it is appropriate to put the methods in their historical context. Furthermore, there seems to be quite some confusion about what maximum likelihood is, how it can be justified, and which dubitable assumptions it avoids, and the pro/contra likelihood discussion seems often to be based on half-truths. Thus we think it would be useful to guide the interested reader to the original literature.

Planned Action: none

(1.38) **Review: eq. 5: What is L1? What are its properties?**

Reply: Some information will be added.

Planned Action: This will be reworded: "With the $(n-1) \times n$ first order differences matrix $\mathbf{L}_1$ and γ a scaling parameter to control the strength of the regularization, the choice of

$$\mathbf{R} = \gamma \mathbf{L}_1 \mathbf{L}_1^T, \tag{1}$$

renders fields of profiles..."

(1.39) **Review: Sec.5: l22: mention the loss of information**

Reply: We think that the loss of information is included in "and limits the spatial resolution of the solution"?

Planned Action: none

(1.40) **Review: Sec 5.4. This section is basically a review of retrieval approaches: why is it relevant for this paper to include such a review? See my general remark above.**

Reply: We consider it as relevant, because any error propagation scheme can only be understood in the context of the related retrieval scheme. What in one kind of decomposition is accounted for by the error propagation of noise needs explicit evaluation of the related parameter error in another kind of decomposition.
Further, we recall that this paper has indeed been submitted as a review paper and we suspect that this information has been lost somewhere in the system.

Planned Action: The introduction will be rewritten to make the purpose of the paper clearer.

(1.41) **Review: Sec 6, p10, point 2: Model errors: It would seem logical to me to split this into RT model errors and inputs used by the forward and inverse models, e.g. influence of atmospheric aspects like surface characterisation, aerosols and clouds, other meteorological variables (humidity, temperature).**

Reply: According to our systematics, the latter are not model errors but parameter errors. Otherwise the reviewer's suggestion and the way we organize this section are very close. "Incomplete Models" and "Numerical Issues" together cover the RT model errors, and the inputs used by the RT model (as far as they are not parameter errors) are the uncertainties in the "model constants". If we combined both RT subsubsections into one subsection, we would need the paragraph caption which is not allowed according to AMT formatting standards. Thus we are forced to keep the hierarchy of sections flat.

Planned Action: none

(1.42) **Review: Sec 6, p10, point 3: "errors caused by decomposing the inverse problem". Does this deserve a separate section?**

Reply: We think so, because it has major implication on error estimation and reporting. In a joint retrieval of species A and B, there is no parameter error due to uncertainties in A, but in a sequential retrieval (first A then B), there is.

Planned Action: We will edit the text to make the logical flow better visible.

(1.43) **Review: Sec 6.1.1, l37: "cheerful" ...**

Reply: That's what we felt...

Planned Action: None by now.

(1.44) **Review: Sec 6.1.3, l33: "measurments"**

Reply: Thanks for spotting.

Planned Action: This will be corrected.

(1.45) **Review: Sec 6.2.1: "If a complete model is available but not used .., the effect of the missing processes can be assessed via sensitivity analyses based on the complete model ...". This sounds like a recommendation (could be part of section 7).**

Reply: We think that this is implicitly included in the completeness requirement in the recommendations. Our recommendations only say what we want but not how it should be achieved. It is the content of the preceding sections to present and discuss methods how this can be achieved.

Planned Action: While we are reluctant to add further recommendations, we consider to mention this as an example along with the respective recommendation.

(1.46) **Review: p15, l3-7: "The OCO-2 team is currently working on ..". I could not understand this paragraph. I suggest to either explain the approach in more detail or omit.**

Reply: We agree that this is hard to understand.

Planned Action: This part will be rewritten.

(1.47) **Review: l24: "retrived"**

Reply: Thanks for spotting.

Planned Action: This will be corrected.

(1.48) **Review: p16, l40: "the derivative". I do not understand how to take such a derivative.**

Reply: **A** is the derivative of the retrieved state with respect to the true state. With a *tertium non datur* assumption (this is here that we disregard the dependence of the solution with respect to noise and other uncertainties, which are addressed elsewhere), then **I-A** is the derivative of the retrieved state with respect to the prior information. Thus it is not necessary to differentiate $\hat{\vec{x}}$ with respect to **I-A** explicitly.

Planned Action: We are somewhat reluctant to add much length to the paper with respect to this but we will make reference to the literature where the use of **I-A** is introduced. Probably the Rodgers book, his Section 3.2.

(1.49) **Review: Sec 6.3: The parameter errors are often very relevant and could be discussed more extensively. For these parameters often simplifying assumptions are made (e.g. climatologies) or they are taken from elsewhere (e.g. actual weather model output) or they may be derived in the retrieval itself (or previous step in the retrieval). All these choices will lead to different characteristics for the related errors, often introducing quasi-systematic error correlations.**

Reply: We agree.
Planned Action: We will expand on this.

(1.50) **Review: 6.3: Why is this section called "parameter errors" instead of something like "Inverse model decomposition errors"**

Reply: Because these errors are not always caused by the decomposition. Sometimes just prior assumptions or external information are used. We call them parameter errors because they are related to the parameters of the forward model as discussed in Section 5.3. We concede that the first line of 6.3 is misleading in the way it is written.

Planned Action: We will reword the first lines of Section 6.3.

(1.51) **Review: Sec 6.4.1 and 6.4.2: I'm happy that the authors include these two "interpretations". This is a subtle point, often not understood by satellite data users.**

Reply: We are glad for appreciation, particularly because reviewer #2 finds this unnecessary.

Planned Action: none

(1.52) **Review: p16, l87: "the undesirable effect that a smoothing error evaluated on a coarse grid will be smaller than a smoothing error evaluated on a fine grid." I do not really understand why this is undesirable. This property seems to make sense to me: more layers allow more detail to be resolved (and smoothed away by the retrieval process).**

Reply: Yes, but if there is a profile on a fine grid, there should be no difference in the error estimate between (case 1) the profile has been retrieved directly on the fine grid, with a certain constraint limiting the vertical resolution to a certain value and (case 2) the profile has been retrieved on a coarser grid, and the resampled to the fine grid. In both cases the difference from the truth can be the same but the (propagated) smoothing errors will be different. This seems absurd to us.

Planned Action: none, since this is discussed in detail in the literature referenced.

(1.53) **Review: p18, l18: "also commonly applied when measurements are compared to model data". It would be good to mention explicitly the data assimilation application here.**

Reply: Agreed.

Planned Action: We will add: "In data assimilation the averaging kernel has to be included in the observation operator."

(1.54) **Review: p19, l10: "reasons reasons"**

Reply: Thanks for spotting.

Planned Action: This will be corrected.

(1.55) **Review: Sec 6.4.3: I was wondering if this section (on altitude resolution) is needed as background to section 7.**

Reply: We agree that the old version of the recommendations does not make any use of the concept of altitude resolution; we think, however, that at least the altitude resolution should be reported for each single profile if only representative averaging kernels are provided.

Planned Action: We consider to make an amendment to R16 or R17 that, if AKs are presented only for individual cases, a vertical resolution profile for each single profile still is useful.

(1.56) **Review: Section 7 Point 2: A bit weak, it leaves a lot of room for different approaches.**

Reply: This is intentional. We will accept any method as long as the resulting errors explain differences encountered between independent measurement systems.

Planned Action: none

(1.57) **Review: Point 3: Does this have repercussions for the data volume? Especially when each component has its own covariance matrix?**

Reply: Yes, it certainly has. This is why we write "The ideal approach ...".

Planned Action: We will add: "It is the responsibility of the data provider to judge to which degree simplifications are justified."

(1.58) **Review: Point 4: Again leaves much freedom. What about proposing a 1 sigma as default?**

Reply: This is controversial. This default is not better than any other default, and in order not to upset researchers following other conventions, we refrain as much as possible from stipulating conventions and limit ourselves to recommendations which can in some way be inferred from agreed conditions of adequacy.

Planned Action: We will add a statement to the intro on our rationale not to stipulate conventions.

(1.59) **Review: Point 6: "error components available, they should also indicate how they contribute to the random and/or systematic error" What about the total error: should this consist of a random and a systematic part?**

Reply: Yes, the total error consists of both components. We insist on reporting them separately, because, depending on the application, only one or the other component may be important. E.g., for time series or trend analysis only the random error component is important, and additive systematic errors are irrelevant. Conversely, in monthly zonal means of densely sampled data the standard error of the mean, representing the random error, goes down to almost zero, and the systematic part of the error will survive.

Planned Action: We will add that, if a total error is reported, it should include both the random and systematic error components.

(1.60) **Review: What does "indicate" mean in practice? Please be more specific.**

Reply: We mean "report" or "describe".

Planned Action: We will consider to reword this.

(1.61) **Review: Point 7: It is difficult to understand what is meant here. What is the domain of a subset of a component of a source of error? It would be good to provide an example.**

Reply: We agree. The recommendations are very generic. This is intentional, since they should be applicable to different measurement systems and retrieval schemes. Examples can help to clarify what we mean in practice.

Planned Action: Illustrative examples will be added to the recommendations.

(1.62) **Review: What is the difference between an error source and an error component?**

Reply: In an ozone retrieval, the retrieved ozone mixing ratio may depend on temperature. The temperature uncertainty is the error source and that part of the ozone error which is caused by the temperature uncertainty is the error component. We agree that this terminology needs to be defined somewhere but we find that Section 4 "Retrieval Theory and Notation" is the better place to do this.

Planned Action: we will add in Section 4: "[is the measurement noise mapped into the retrieved atmospheric state.] **In other words, $\mathbf{S}_{\mathrm{x,noise}}$ is the error component in $\vec{x}$ due to the error source $\mathbf{S}_{\mathrm{y,noise}}$.**

(1.63) **Review: Point 9: "assumed ingoing uncertainties shall be reported". What is meant by "reported"? Does this refer to the ATBD, a journal paper or to the L2 datafiles themselves?**

Reply: At the place where the respective resulting errors are reported. There are many ways to make error analysis traceble in this respect. The important thing is that the information can easily be found. If the source of these data is accessible, a link may do.

Planned Action: [For all error components, the assumed ingoing uncertainties shall be reported] **in the relevant documentation**[, otherwise error propagation would not be traceable.]

(1.64) **Review: Point 10: Sometimes (I-A)x$_a$ = 0 even though the retrieval still needs/depends on a-priori information. Should the a-priori be reported also in this case?**

Reply: We do not quite understand how the retrieval can depend on a priori information when (I-A)x$_a$ = 0. Does the reviewer mean cases when the quantity the a priori refers to is not an element of x? According to our terminology this would not be "a priori" in the narrow sense but parameter. Or does the reviewer talk about implicit a priori information imposed by a coarse grid?

Planned Action: The text will be reworded in a way that it will become clear that a priori information is meant only in a narrow technical sense and does refer only to elements of the state vector actually retrieved.

(1.65) **Review: Point 10: What are "similar operations"? Please be more explicit.**

Reply: we mean variants of Eq. 30 which may be formally different but follow the same rationale.

Planned Action: "or to perform similar operations" will be replaced by "or variants of it"

(1.66) **Review: Point 11: I do not understand why it is crucial to have the results as vertical profiles (as opposed to desirable). The vertical profile retrievals are linked to the real physical world through the averaging kernels, as specified in Eq. 30. Ignoring this link leads to all the smoothing error considerations (and problems) as discussed extensively by the authors. Especially when the kernel is very different from the unity matrix I, the interpretation of the retrieved profile as a real profile becomes troublesome. The retrieval at a given altitude then contains physical information from (depends on concentrations in) many other layers, as specified by the averaging kernel matrix. The kernels will always have altitude on one axis, even if presented in eigenvalue space, and relate the retrieval to real physical profiles. Please explain why this strong statement ("should be presented") is made.**

Reply: We do not mean "exclusively represented as vertical profile". There is nothing wrong with presenting the data in a different way, and we agree that for certain applications this can even be advantageous. But when data from different sources are combined in one study, this happens almost always on the basis of a vertical profile representation. We just want to make sure that in this case the data characterization is available. We do not want to allow the data provider the excuse "I do not need averaging kernels because in my representation they are irrelevant and everything else is the business of the data user". We think that the data provider is in a better position than the data user to provide the averaging kernels in the profile space.

Planned Action: We will change the text to: "[...then the final result should] in addition [be presented as vertical profiles and also all diagnostic data (error estimates, averaging kernels) should be transformed to an altitude-dependent...]".

(1.67) **Review: Point 12: "Ideally the data provider calculates the averaging kernels on the final grid". What is proposed here? It sounds**

like a commitment of the retrieval team (data provider) to provide support to all users with a grid which differs from the retrieval grid. This would imply a major commitment. Or would this imply that each retrieval product should be accompanied by software to do the interpolation (extrapolation is also very likely!) to different grids.

Reply: This is not meant. We are talking about the final grid on which the data producer distributes the data. Sometimes data are retrieved on some specific grid (e.g., related to the tangent altitudes) and then resampled on a uniform grid. In this case the user is not helped much with an averaging kernel which refers to the original retrieval grid.

Planned Action: The text will be clarified: "[...on the final grid] on which the data are provided to the user."

(1.68) **Review: Point 13: "This is particularly important when data are reported in a form that differs from that of the retrieval state vector". This may not be very clear to the reader. Please provide an example. Why is it important in this case?**

Reply: E.g. the application of log averaging kernels to vmr profiles gives a mess.

Planned Action: We will add: "E.g., the averaging kernels resulting from a retrieval of the logarithms of mixing ratios must not be applied to mixing ratios. It is thus of utmost importance to communicate to the data user to which quantities the averaging kernels refer."

(1.69) **Review: Point 15: "If the data are understood to be a representation of the smoothed state of the atmosphere, the smoothing error is not needed and averaging kernels along with the prior information are sufficient". I suggest that the authors explicitly mention applications here, e.g. model-satellite comparisons and data assimilation.**

Reply: We agree.

Planned Action: We will explicitly mention applications.

(1.70) **Review: Point 16: "Communication of a complete error budget ... is not always technically feasible and often creates unnecessary data traffic." I would suggest that the authors include a reference to the work of S. Migliorini, DOI: 10.1175/2007MWR2236.1. This paper describes how the data volume can be reduced drastically (explicit a-priori profile and error covariance no longer needed) while preserving the full error information, to support data assimilation applications. Do the authors consider this a possible alternative for storing the retrieved profiles, see e.g. point 10, 11?**

Reply: We will mention this possibility. However, we find it hard enough to prevent data users from simply ignoring the diagnostic data because they seem to be too complicated, and a data-reduced representation of the matrices may make the problem even worse.

Planned Action: We will reference this paper and mention that it might provide (at least partly) a solution to the data traffic problem.

(1.71) **Review: Point 18: This important point distinguishes random and systematic errors, related to real-world validation activities. I agree that this is the ultimate test for the errors provided. In practice there will be a difficult to quantify group of contributions to the error budget which are quasi-random, quasi-systematic. Error terms related to input parameters (climatologies, estimates of auxiliary information on the surface, clouds, aerosols impact on trace gas retrievals, temperature/humidity profile information, measurements from other space instruments, the a-priori and other model information) may average out over long time periods (e.g. a year) but are typically (strongly) correlated in space and/or time. Are there any general recommendations that can be made for this group of error contributions? Sometimes such contributions are presented to users as "random" and sometimes as "systematic" by the retrieval teams. It would be good if the authors discuss this random/systematic distinction in more detail and, where possible, provide clear recommendations how to deal with this.**

Reply: Again, it is the responsibility of the data providers to make a sensible distinction here. The final criterion is that the error estimates can be confirmed by the validation of the standard deviation of the differences and the bias. In particular situations it may even be appropriate to split the contribution of one error source into a systematic and a random component. Errors which are random in longer time-scales but systematic in shorter timescales are exactly what we mean with 'errors correlated in certain domains'. In cases like those mentioned by the reviewers, it must be reported that errors are autocorrelated in the time domain.

Planned Action: We will add this example to R7.