The authors thank the reviewers for their comments.

Review #1:

(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. One of its purposes is to lay down a formal framework in which the work of the TUNER project takes place. This, we hope, justifies several sections which might not be necessary in an original research paper, and may make some of the criticism obsolete.

Action: We now mention 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.

Action: Instead of splitting the paper we have tried 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 re-
porting section, additional section references have been included.

On page 3, l1, we have added: “[We then systemize and discuss the various sources of retrieval error] and, if applicable, their dependence on the retrieval scheme chosen] (Section 6)”

On page 18, Section 6.1.1., l 15, we have added: “[The situation is different in a decomposed retrieval] (Section 5.4)”

On page 18, Section 6.1.1., l22, we have included 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 have added “[by the radiative transfer model \( \hat{F} \) in use] (Section 4)”

On page 24, Section 6.3, l4-5, we have added “[... is already known] (see, Section 5.4)”.

On page 24, Section 6.3.1, l23, we have added ”[In onion peeling] (Section 5.4.5) [the ray...].

On page 25, Section 6.4: There has already been 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 have added 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.

Action: The links between these sections and the recommendations have been
made clearer. The amendment on the altitude resolution has been added to (old) Recommendation 16 (now, after reordering Recommendation 13).

(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 possibly 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 directly 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.

Action: We have extended the recommendations a little bit. After each recommendation where appropriate we could add a short discussion where we add examples and provide the links to the various retrieval approaches. Further, we make links to the sections under dispute to make clear how they are related to the recommendations.

(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.

Action: The link between the recommendations and the theoretical parts has been 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.

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.

Action: The introduction has been rewritten and the new version includes examples of the requested type.

(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.”

Action: We have expanded 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 broadly agree. Recommendations 1–7, 10 and 13–15 (new order) apply both to profile and column retrievals. Recommendations 8 and 9 have been amended to cover also column retrievals. Recommendations 11,12, 16, 17
are only applicable to profile retrievals

**Action:** Recommendations 8 and 9 have been amended to cover also column retrievals. Recommendations 11, 12, 16 and 17 have no column counterpart. Her occasionally the wording has been changed to make clear that these recommendations refer to profile retrievals only.

(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.

**Action:** We now mention 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.

**Action:** We have added 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:** Validation is not the main content of the paper. Our work and validation studies are linked in that validation verifies or falsifies the error estimates. This is clearly stated in CoA5 and R18. Besides the retrieval errors, different a
priori and resolution as well as natural variability contributes to the differences between measurements. That is where we mention validation or data comparison. We do not have to say much on the other aspects of validation.

**Action:** None. We have not succeeded to include more on validation without disrupting the logic of this paper.

(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.

**Action:** We now mention in Section 5.2 and in the conclusion 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 (R18).

**Action:** We have added after the list of recommendations “On the face of it, the list of recommendations appears quite weak, leaving a lot of freedom to the data provider. This is, however, not the case. Recommendation , that the error budget should be as complete as possible, along with Recommendation , which gives a criterion for the completeness of the error budget quickly make
the apparent freedom disappear.”

(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.

**Action:** We have tried to 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’.

**Action:** We have added a statement saying that most of the pre-existing literature is not specific enough for indirect measurements.

(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.

**Action:** In the paper we have made the rationale outlined above explicit.
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.

Action: We have added 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 has become obsolete because major parts of the introduction have been rewritten anyway.

(1.21) Review: l16: “The project ... is a consortium of”. Please modify

Reply: Thanks for spotting.

Action: This has been reworded.
(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”.

**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).

**Action:** Particularly in the recommendations section more weight is now given to column retrievals.

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

**Reply:** Thanks for spotting!

**Planned Action:** This has been 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.

Action: We have added some sentences on this.
(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.

**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, the large rank of the matrix to be inverted, which will impose some practical limitations.

**Action:** We have added “[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.

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

**Action:** We have added “$\tilde{F}$ is a vector-valued non-linear function and deviates from $\tilde{f}$ in that it involves numerical approximations and may not include the full physics of radiative transfer.” Further, we have corrected $F$ to $\tilde{F}$ 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 Hadamard, 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 of the equations is not in conflict with overdeterminedness.

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 on purpose. Otherwise the reader might think that we claim that other error sources are omitted also in the error propagation.

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.

Action: None.

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

Reply: Some information has been added.

Action: This has been reworded: “With the $(n - 1) \times n$ first order differences
matrix $L_1$ and $\gamma$ a scaling parameter to control the strength of the regularization, the choice of
\[ R = \gamma L_1 L_1^T, \] (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”.

**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.

**Action:** The introduction has been 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.

**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. It is the decomposition which gives rise to parameter errors.

Action: We have edited the text to make the logical flow better visible.

(1.43) Review: Sec 6.1.1, l37: “cheerful” ...

Reply: That’s what we felt...

Action: None by now.

(1.44) Review: Sec 6.1.3, l33: “measurments”

Reply: Thanks for spotting.

Action: 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 have decided 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. Contrary to what we have said in our initial reply, we find now that this paragraph was more on mitigating uncertainties than on characterizing them. Thus we think that this paragraph does not belong here.

Action: This part has been deleted.
(1.47) Review: l24: “retrieved”

Reply: Thanks for spotting.

Action: 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 \( \hat{x} \) with respect to I-A explicitly. In the paper there is a reference to the book and several papers by Rodgers, where the use of I-A is introduced.

Action: None.

(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.

Action: We have included “Depending on the source of the information on the parameter vector – climatology, preceding retrieval step, independent measurements, or whatsoever, the parameter errors can be correlated or uncorrelated in space and time.”.

(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 was misleading in the way it was written.

Action: The first lines of Section 6.3. have been reworded: “We define parameter errors as those errors caused by the decomposition of the full retrieval
problem such that a part of the atmospheric state is assumed to be already known and thus not included in the retrieval vector \( \hat{x} \) (see, Section 5). The assumed values can derive from either a preceding retrieval step or from climatologies or any other source of prior information."

(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.

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.

Action: We have added: “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.
Action: 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 original 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.

Action: We have added to R13 (old R16, before re-ordering) 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: The room for different approaches is intentional. We will accept any method as long as the resulting errors explain differences encountered between independent measurement systems. Together with this latter condition, we do not find R2 weak.

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 ...”. The recommendations, however, does not explicitly require that all related single-component covariances shall be reported.

Planned Action: We have added an example that shows that (and why) even scalar component-wise uncertainty information can be useful.”

(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.

Action: We have included a statement into the intro on our rationale not to stipulate conventions.
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.

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

Review: What does “indicate” mean in practice? Please be more specific.

Reply: We have kept the verb ‘indicate’ but we have reworded the sentence to make the recommendation more specific.

Action: Reworked: “[... ] should also indicate which of them contribute primarily to the random error and which contribute primarily to the systematic error.”.

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.

Action: Illustrative examples have been included here and in other recommendations to make clear what we mean.

Review: What is the difference between an error source and an error component?

Reply: E.g., 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.
**Action:** we have added in Section 4: “[is the measurement noise mapped into the retrieved atmospheric state.] In other words, $S_{x, \text{noise}}$ is the error component in $\tilde{x}$ due to the error source $S_{y, \text{noise}}$.

(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 traceable 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.

**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?

**Action:** The text has been reworded in a way that it should now be 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.

**Action:** “or to perform similar operations” has been 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.

Action: We have changed 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.

Action: The text has been 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.

Action: We have added: “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 have added “E.g., in data assimilation, where the averaging kernel is part of the observation operator, it would be redundant and lead to incorrect double counting to include the smoothing error in the observation error covariance matrix.”

(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.

Action: We have included a reference to this paper and mention the relevance of this method to data assimilation.

(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 reviewer, it must be reported that errors are autocorrelated in the time domain.

Action: We have added this example to R7 (now R5).
Review #2:

(2.1) Review: 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.

Reply: The intention of this paper is to summarize retrieval approaches actually in use in satellite remote sensing, to systematize them in a common framework and notation, to discuss the implications of related choices on error propagation and to infer related recommendations on unified error reporting.

Action: The introduction has been rewritten to make our intention clear.

(2.2) Review: 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 selfreference what they did in the area of inverse/retrieval algorithms for the sounding of atmospheric parameters.

Reply: The purpose of the article is to cover all the satellites for remote sensing of atmospheric compositions over the past 20 years from the all frequency ranges, microwave, infrared, NIR and UV/VIS, as a review paper. This implies numerous references, and since the list of authors includes many scientists from many different groups working in this field, it appears quite natural to us that self-referencing is inavoidable.

Planned Action: We have added more references from scientists who are not involved in this paper in order to make the article more balanced.

(2.3) Review: 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.

Reply: We are afraid that this is misreading; we do not say that the goal of the paper is just to provide a list of recommendations. We do mention that we provide some recommendations, but the rest of the abstract summarizes the problem areas tackled in this overview paper. Only in the section on conditions of adequacy we indeed mention the “ultimate goal of presenting a list of recommendations”. The attribute “ultimate” makes clear that this is by no means the only goal.

Planned Action: We have addressed this more clearly in the introduction, and in the beginning of the recommendations section.
(2.4) Review: 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?

Reply: We want to avoid quibbling about words. The reviewer is free to call the quantities mentioned as they like, as long as the terms are clearly defined somewhere. However, the terminology we use is applicable also to single measurements, while we have problems to assign a meaning to the terms ‘bias’ or the ‘root mean square error’ in the case of a single measurement.

Planned Action: None.

(2.5) Review: 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.

Reply: We do not understand how it is logically justified to calculate the sum of the CoAs and the recommendations. We thought that sums can only be calculated of items of the same category. We also do not understand what the logical link between the number of recommendations and the degrees of freedom of a retrieval is. We do not see how these quantities connected. To us, these quantities seem incommensurable.

We would have preferred less recommendations but condensing them makes them less specific, and finally we would end up with some vague truisms which would not be helpful at all.

Planned Action: We have added examples to the recommendations to make clear why they are important.

(2.6) Review: 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 [...] 

Reply: The retrieval schemes and methodical issues belong to the core content of this review paper. Without understanding the underlying simplifying assumptions of a retrieval scheme, it is difficult, perhaps even impossible, to provide a reliable error estimate.
Action: The introduction has been rewritten to make the purpose of the paper clearer.

(2.7) Review: [...] and what I could call a silent but insistent criticism to Optimal Estimation.

Reply: We neither endorse nor dispraise any particular method but we describe the methods which are currently in use, or whose data products are currently in use. For each method we discuss the underlying assumptions.

Action: We now make an explicit statement that the superiority of either maximum-likelihood based or optimal-estimation based retrievals cannot be decided on scientific grounds but is a purely philosophical question.

(2.8) Review: Furthermore, I think that the format of the present study is much more adequate for a report.

Reply: Reports typically report technical information related to one instrument, processor, etc. We present, in a unified notation, an overview of all methods we are aware of. Thus we think that this paper serves well as an overview paper for the TUNER special issue, because it provides a framework the other papers of the special issue can refer to.

Action: None.

(2.9) Review: 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 onetwo pages.
Reply: First a side remark: The fact that canonical theory of statistics relates the performance of a statistic estimator to the true value strengthens our position against GUM. We do agree that the errors of an ensemble of retrievals can be decomposed into the mean square error and the bias, and we use this concept ourselves in order to validate the error estimates. We concede that our list of recommendations is three pages, but it covers issues not mentioned by the reviewer (correlations in other domains; data traffic, and others).

Action: None.

(2.10) Review: I have also to say, that authors’ recommendation list itself is largely independent of the bulk of the present paper.

Reply: We admit that not all parts of the paper are needed to derive the recommendations, but the information contained in the bulk of the paper is needed to provide the quantities requested by the recommendations.

Action: The relation between the recommendations and the rest of the paper has been made clearer, e.g. by cross-references.

(2.11) Review: 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?

Reply: It is not true that we identify the random component of the root mean square error as the error or uncertainty of a given retrieval system. We conceive the error or uncertainty as a quantity which is composed of a random part (corresponding to what the reviewer calls root mean square error) and a systematic part (corresponding to what the reviewer calls bias). We state explicitly that the systematic error estimates can be tested using the bias between collocated measurements of independent measurement systems, and that the random part can be tested using the standard deviation of the difference between collocated data from different measurement systems. The bias is commonly defined as a mean difference between the measured value and the true value unless explicitly specified differently. In our paper, when we use the term ‘bias’ with any other meaning than the mean difference between the measured and the true value, we state explicitly what the mean difference refers to.

Planned Action: In the former Section “Measurement response” (now retitled “Regularization bias and measurement Response”, more room is given to the discussion of the bias.

(2.12) Review: I have found a bit confusing the question about gridin-
dependent retrievals, which for me is a nonsense, since normally one works with a discretized state vector. Apart from forward model (FM), the bias depends on the given constraints, which are normally griddependent, 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).[...]

Reply: Another contribution to the bias can be calibration issues. The role of the constraint is discussed in Section 6.

Action: We now mention that the choice of a prior which is not the expectation of an ensemble the actual measurement is taken from will cause a bias.

(2.13) Review: 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.

Reply: We recommend that the averaging kernels and priors used shall be communicated to the users. The users can then evaluate the smoothing error on the final grid they use, after evaluating the additional averaging kernel component entailed by the interpolation. Sharing the total error will cause inadequate error estimates after resampling and respective generalized Gaussian error propagation.

Action: As said above, the discussion of biases has been expanded.

(2.14) Review: On the same line, their CoA2 is inconsistent with the idea of root mean square error.

Reply: The intention is to avoid that data users interpolate the smoothing error on a finer grid. Instead they should be provided with all information they need to directly evaluate it on the grid of their choice. Any possible inconsistency with the root mean square error comes only from conceiving the retrieved state as a smoothed estimate of the truth, a conception we do not explicitly endorse. Conceiving the retrieved state as an estimate of the smoothed truth removes this inconsistency.

Planned Action: None.

(2.15) Review: Furthermore, I am not sure if it can be implemented, in practice.

Reply: For noise alone, CoA2 can be implemented. It is only the combined noise and smoothing error which causes the problem.
(2.16) Review: 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 \( n_1 \times n_1 \). CoA2 requires that, using authors’ language,

\[
WS_{x,\text{noise},n1}W^T = S_{x,\text{noise},n2},
\]

where, \( S_{x,\text{noise},XX} \) 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 \( n_1 \leq n_2 \) or \( W \) and \( n_1 \geq n_2 \). Atmospheric state vectors are not bandlimited 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.

Reply: We agree with everything above except that the additional error is not part of the noise but of the smoothing error, which, we suggest, should be evaluated newly on the finer grid. The example presented by the reviewer shows perfectly why we insist that noise and smoothing error should be reported separately. For noise alone, CoA2 is fulfilled by using generalized Gaussian error propagation. And again, conceiving the retrieval as an estimate of the smoothed truth removes this inconsistency.

Planned Action: None.

(2.17) Review: 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.

Reply: We do not agree. We do not criticize the concept of the smoothing error in general (except for the ambiguity of the underlying interpretations of probability, which we criticize in a very careful and moderate wording). The central point of our criticism is the inclusion of the smoothing error in the total error, which will lead to inconsistent results after resampling of profiles.

Planned Action: None.
Planned Action: None.

(2.18) Review: 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).

Reply: To us, natural variability explains that the atmospheric state at one time and one place is different from the state at another place and another time. Due to this natural variability we cannot expect that two instruments measuring at different places and/or times will render the same result. Detected differences thus do not hint at any malfunction of one of the instruments or retrieval and thus are not genuine measurement errors. Still, these differences have to be considered in comparisons. The reviewer has torn this quotation out of a very different context in our paper. From the context of Section 6.6, where the quoted statement comes from, it should be very clear what we mean. We do not understand how the reviewer can, on the basis of this text, accuse us to “purposely mislead statistical error with mistake”.

Action: We have added “[...natural variability] in a sense that the atmospheric state at place $s_1$ and time $t_1$ differs from the one at $s_2$ and $t_2$. “ And we have given more weight in the text to the regularization bias.

(2.19) Review: 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 poutpourri of about everything is known today on atmospheric inverse problems: Twomey, Tikhonov, Rodgers, LS, ML.

Reply: Our intention is to cover all relevant (in the sense that data retrieved with these methods are still around) methods within a consistent framework and a common notation. This is a precondition for unified error reporting. While the book of Rodgers (2000) provides an excellent theoretical basis, we apply this theory (and other variants) to the real-world retrieval schemes and investigate which uncertainties are caused by the assumptions and approximations in place. We understand this as a systematic compilation rather than a potpourri. We first lay down the basic theory. Then we discuss how retrieval schemes used in the real world deviate from the idealized theory. Then we discuss all error sources and their relevance. We find that the content is clearly structured, and goes beyond the content of the available literature in that it treats also the relevant real-world problems.
**Action:** The introduction of the paper has been rewritten to make the purpose of the paper more evident.

(2.20) **Review:** 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 $R$

**Reply:** From our introduction it should be clear that we do not only consider methods which have a probabilistic interpretation. T. von Clarmann and U. Grabowski, Atmos. Chem. Phys. 7:397-408 (2007), their appendix, have shown that there is even a probabilistic interpretation of Eq 4 with R defined as shown in Eq 5. We do not see what is wrong with putting a method in a more general context.

**Planned Action:** None.

(2.21) **Review:** 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.

**Reply:** Resampling of profiles and associated error propagation works well for all error components (noise, instrumental calibration biases, forward model biases...) except those which depend on the sampling of the $S_a$ matrix. Thus we insist that the latter should be evaluated on the final grid, using the respective sufficiently resolved covariance matrix.

**Planned Action:** None.

(2.22) **Review:** 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.

**Reply:** We do not say that ‘uncertainty’ shall not be used. We say that the claimed difference between ‘uncertainty’ and ‘error’ is controversial. And according to GUM (and with respect to this issue we agree with GUM), ‘uncertainty’ does NOT refer to the root mean square error of an estimator but includes also systematic effects.

**Action:** None.

(2.23) **Review:** 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 be-
cause 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.

**Reply:** The term ‘maximum likelihood’ in this context is used by Rodgers (2000) for a solution which is free of formal prior information. And this terminology is consistent with that of Fisher, who coined that term. If we search for a solution of which the probability that it reproduces the (noisy) measurement is largest, we get, by definition, the maximum likelihood solution. If we apply this principle to Gaussian noise, the maximum likelihood solution happens to be the least squares solution. We do not use the maximum likelihood solution because it yields LS under normality but we use least squares because ML plus normally distributed yields least squares. It is agreed - and even conceded by Fisher - that ML does not yield the solution of maximum posterior probability. But what is untenable about it? We do never claim that we consider only methods which have a probabilistic interpretation. And more generally speaking: We do not particularly endorse any of the methods we describe. In this paper, we just describe and characterize them.

**Action:** None.

(2.24) **Review:** Why do authors not qualify the bias and variance of the estimator?

**Reply:** Because we have organized the paper such that first the methods are presented, and in Section 6 error estimation is discussed. This seems justified to us, because a lot of the error propagation stuff can be treated in parallel for all the estimators, and touching this issue here would lead to redundancies which would make the paper even longer. Both bias and variance of the estimates depend on many more choices than the estimator alone.

**Action:** None.

(2.25) **Review:** Why the reader has to wait until section 6, just to see the variance alone of the estimator.

**Reply:** An estimator does not have a variance, only the estimate has one. There are a lot more sources which contribute to the variance of the estimate than measurement noise. Making an exception for this particular source of variance does not seem adequate to us.

**Action:** None.

(2.26) **Review:** 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 adhoc substitution, authors claim that the estimator (4) becomes more flexible and powerful than the OE shown in Eq. (6).

**Reply:** We do not make such a statement.

**Action:** None.

(2.27) **Review:** 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).

**Reply:** We find it quite natural to first present the methods and then discuss the error sources. This seems particularly adequate to us since Eq. 18 represents only one component (often not even the leading one!) of the random error.

**Action:** None.

(2.28) **Review:** In addition, (a) The bias of the estimator is not qualified/assessed/quantified in any part of the document.

**Reply:** The bias caused by the regularization is only one component of the total bias. We do not see any good reason to give it an extra treatment by discussing it in Section 4 while all other bias-generating errors are discussed in Section 6. This would disrupt the logical structure of the paper and may even lead the readers astray because they may think that the bias caused by the regularization term is always the most important one.

**Action:** We have rewritten Section 6.4.5 and discuss the bias-generating properties of the retrieval approaches there. Some further thoughts on this issue are now included in the Conclusions.

(2.29) **Review:** b. What is the reason to change \( S_a^{-1} \) with \( R \)? What are the expected improvements?

**Reply:** We do not claim in this paper that there are improvements. We simply want to systematize existing retrieval methods by presenting them in a common framework and notation.

**Action:** None.

(2.30) **Review:** c. Why has the Tikhonov/Twomey regularization \( \gamma \)-parameter disappeared? That is why not \( \gamma R \)?

**Reply:** Thanks for spotting!
Action: The equation has been corrected. The text above the equation has been changed to: ‘[... first order differences matrix,] and $\gamma$ a scaling parameter to control the strength of the regularization’.

(2.31) Review: d. What’s the role of $x_a$, and why not $x_0$ as in Eq. (3)?

Reply: It makes a difference with respect to what the solution is smoothed. The solution of Eq (3) (in the linear case or after iteration in a well-behaved case) does not depend on $x_0$. Thus $x_0$ can be freely chosen. The solution of Eq (4) does depend on $x_a$, because the smoothing operator will not smooth the profile but the difference between the profile and the a priori. Thus, $x_a$ cannot be freely chosen.

Action: The text has been changed to make this clear.

(2.32) Review: e. With $R$ set to any of the suggested matrices, 012 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?

Reply: First of all, we do not propose anything, but we describe methods which are actually in use. And back to the question: By an appropriate definition of $\gamma$ (which has admittedly been missing in the discussion paper) dimensions can easily be included.

Action: We have defined $\gamma$ in the text and mention that it is in addition supposed to balance the units.

(2.33) Review: 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 satelliteborne 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.

Reply: Eq (4) is the algebraic generalization. Both Tikhonov smoothing and optimal estimation are particular instanciations of it. We think that this is a fairly neutral way to present these methods. It shows how the methods are related.

Action: None.
Review: 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.

Reply: A smoothed profile is a profile where the altitude-to-altitude differences of the profile values are reduced. With a reasonably smooth a priori profile, also the difference profile (result minus a priori) will be smooth. The question why it is a good property is answered in the second part of the criticized sentence: “thus avoiding unphysical oscillations...”

Action: We have modified the sentence as follows: “[...smoothed] in the sense of reduced altitude-to-altitude differences...”.

Review: 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?

Reply: If the retrieval is conceived as an estimate of the smoothed truth as discussed in Section 6.4.2 and if the measurement response as discussed in Section 6.4.5 is unity (as it typically is with first order differences Tikhonov regularization) then estimator (4) is bias-free. If $S_n$ does not equal the (typically unknown) $\langle \tilde{x}_{true} \rangle$, optimal estimation will have a bias. Thus, things are not as simple as they seem to be. We thus think that the bias discussion should not be touched upon passing in Section 3 but should be deferred to Section 6.4.5, where we have the content of Section 6.4.2 available, and where we can discuss the bias issue at more depth.

Action: Section 6.4.5 has been rewritten to include the bias issue.

Review: Page 8. Eq (6). Now that the authors have invented $R$, they can say our estimator retrieves the OE estimate if we put $R = S_n^{-1}$, unbelievable!

Reply: We find it quite natural that, when we generalize over formalisms and then specify again, we get the original specification back. We do not see what is wrong about this.

Planned Action: None.

Review: By the way, to me, to $R = S_n^{-1}$, is the only possible choice, if we want to reach dimensional consistency.

Reply: We disagree. With the correct units (which can be imported via $\gamma$), any $R$ will be dimensionally consistent, regardless if it has a probabilistic interpretation or not.
Planned Action: None.

(2.38) Review: Page 8, paragraph beginning at line 8. This comment seems to stay here just to add some references.

Reply: The fact that in the case of logarithmic retrievals the data characterization also refers to the logarithm of the state value is often overlooked and has already caused some confusion among data users. Thus, we find it appropriate to mention this issue.

Planned Action: None.

(2.39) Review: 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.

Reply: We disagree. The caveat regarding the Gaussian probability density function is relevant only if the estimate is given a probabilistic interpretation, i.e., in the context of Eq. (6). And the suggested method using $\tilde{x}$ as a retrieval variable does not solve the problem that, for a variable which mostly has small values but a large natural variability (i.e. large $s_a$), the wings of a Gaussian penetrate wide into the negative. That is to say, optimal estimation assigns positive probability densities to negative temperatures or mixing ratios.

Action: We have better highlighted the problem of positive probability densities to negative temperatures or mixing ratios.

(2.40) Review: 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?

Reply: Here we neither discuss the variance nor the bias. Both variance and bias include more than only noise and regularization, respectively. Thus, the discussion of both is deferred to Section 6.

Action: The bias will is now discussed in Section 6.4.5.

(2.41) Review: Section 5. All is said in this section is today overcome by Simultaneous Retrieval. Section 5 is outdated and should be to-
tally removed.

Reply: A scientist trying to figure out what the total error budget of HALOE or SOFIE data is, is not much helped by this statement. And for, e.g., infrared spectroscopic instruments with 30-40 data products, represented at tens of altitudes each, and – depending on the instrument type – more than 1000 profiles per day with overlapping lines of sight, simultaneous retrieval of everything is still beyond reach. And if, e.g., spectroscopic data of one species are inconsistent in different parts of the spectrum, simultaneous retrieval can even be worse than a sequential approach.

Action: None.

(2.42) Review: Section 5.4.5 Still Onion Peeling?

Reply: As said above: the users of, say, HALOE or SOFIE data are not helped very much by saying “the data providers should have used another retrieval method.”

Action: None.

(2.43) Review: Section 5.4.6 See point above. I recommend a CoA0: Please forget about adhoc and nonoptimal methods!

Reply:
1. It is the purpose of this paper to get error estimation for existing data sets under control. We are not proposing a data analysis scheme for a future instrument. The reviewer seems to have misunderstood the conditions of adequacy. They are not about retrieval schemes, but for error propagation schemes for given (not necessarily favoured or endorsed) retrieval schemes.
2. Optimal methods are optimal only if a real $x_0$ and a real $S_0$ are available. These are often not available, and many “optimal estimation” retrievals are non-optimal retrievals in disguise. Some of the instruments covered by our study have made measurements of some species for the first time. Where to get the prior and its statistics from in this case? And finally: Who says that the prior which was valid until yesterday is still valid today? Remember the turkey that came to the gate of the enclosure every day at 9:00 expecting to be fed. This went well until Thanksgiving. But according to the rules of inductive inference, on which optimal estimation is based, the turkey behaved in a fully rational way!
3. Forgetting methods not favoured by the reviewer clashes with comment 2.33, where we are requested to be neutral. (cf 2.33)
4. We understand that science is the generation and aggregation of knowledge. Based on this assumption it is unclear to us how forgetting anything should advance science.
(2.44) **Review:** Sections 6.1 to 6.3 can be summarized under a very short section entitled “Instrument Noise and Forward Model bias”

**Reply:** First, we have organized Section 6 by causes of the errors and not by random versus systematic errors. This is because the same source of error can show up as the one or the other, according to the retrieval scheme. And second, we do not see how this reorganization should make the section shorter.

**Action:** None.

(2.45) **Review:** 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.

**Reply:** The bias caused by the regularization is dealt with in Section 6.4.5.

**Action:** Section 6.4.5 has been expanded and restructured.

(2.46) **Review:** 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, smooth, 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?

**Reply:** We do not criticize the smoothing error as such but we criticise that it may be included in the total error and will thus be inappropriately propagated for resampled profiles. We find the claim that interpolation is unnecessary somewhat odd. Science does not only consist of plotting data. Some more quantitative approaches are required. Time series at one altitude, when the original data have a varying altitude grid, quantitative profile compari-
son as suggested by Rodgers and Connor (J. Geophys. Res. 108(D3):4116, doi:10.1029/2002JD002299, 2003) and many more scientific applications need interpolation of the data to a common grid.

**Action:** None.

(2.47) **Review:** Pag. 27 and 28. Eqs (28) and (29) can be left to more elaborated comparisons. There is no need to cover this aspect in the present paper.

**Reply:** Here we agree with reviewer #1 who finds this section particularly important. Furthermore, these sections are important to understand when regularization can cause a bias and when not. Equations 28 and 29 are essential for these sections.

**Action:** None

(2.48) **Review:** 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.

**Reply:** This statement refers to situations only where the contrast in the resolution is large. Thus, this statement does not depend on the particular definition of vertical resolution, any of the resolution concepts introduced in Section 6.4.3 will do. We had planned to move Section 6.4.3 before Section 6.4.1. With this we would have had the definitions of altitude resolution available and could have made reference to these definitions. This, however, disrupted the logical flow of Section 3, and we have dismissed this idea.

**Action:** We make a reference to Section 6.4.3, where different concepts of altitude resolution will be introduced.

(2.49) **Review:** Pag. 6.4.3 From section 6.4.3 on, until section 7, the paper appears to be unnecessary long.

**Reply:** At many instances the reviewer criticizes that the various consequences of regularization are not sufficiently discussed, and here, where these issues are dealt with, the paper is criticized to be too long. We are confused.

**Planned Action:** None.

(2.50) **Review:** Section 7. As said at the beginning 18 recommendations are too many to be useful.
Reply: As said before, we would have preferred less recommendations but condensing them makes them less specific, and finally we would end up with some vague truisms which would not be helpful at all.

Planned Action: We have included examples which show why the recommendations are important.

(2.51) Review: 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

Reply: “official L2 providers” and “agencies” are no terms of scientific relevance. And including a “list of the many scientists dealing with Satellite Data Processors” would not be useful either. Our criterion is: we included data processors of which the data are distributed to the scientific community. We think that this is a sensible criterion.

Planned Action: We have included another IASI processor.
Estimating and Reporting Uncertainties in Remotely Sensed Atmospheric Composition and Temperature

Thomas von Clarmann¹, Douglas A. Degenstein², Nathaniel J. Livesey³, Stefan Bender⁴, Amy Braverman³, André Butz⁵, Steven Compernolle⁶, Robert Damadeo⁷, Seth Dueck², Patrick Eriksson⁸, Bernd Funke⁹, Margaret C. Johnson³, Yasuko Kasai¹⁰, Arno Keppens⁶, Anne Kleinert¹, Natalya A. Kramarova¹¹, Alexandra Laeng¹, Vivienne H. Payne³, Alexei Rozanov¹², Tomohiro O. Sato¹⁰, Matthias Schneider¹, Patrick Sheese¹³, Viktoria Sofieva¹⁴, Gabriele P. Stiller¹, Christian von Savigny¹⁵, and Daniel Zawada²

¹Karlsruhe Institute of Technology, Institute of Meteorology and Climate Research, Karlsruhe, Germany
²University of Saskatchewan, Saskatoon, Canada
³NASA Jet Propulsion Laboratory and California Institute of Technology, Pasadena, CA, USA
⁴Department of Physics, Norwegian University of Science and Technology (NTNU), NO-7491 Trondheim, Norway
⁵Institut für Umweltphysik, Department of Physics and Astronomy, Universität Heidelberg
⁶Department of Atmospheric Composition, Royal Belgian Institute for Space Aeronomy (BIRA-IASB), 1180 Brussels, Belgium
⁷NASA Langley Research Center, Hampton, VA
⁸Chalmers University of Technology, SE-412 96 Gothenburg, Sweden
⁹Instituto de Astrofísica de Andalucía, CSIC
¹⁰National Institute of Information and Communications Technology (NICT), 4-2-1 Nukui-kita, Koganei, Tokyo 184-8795, JAPAN.
¹¹NASA Goddard Space Flight Center, Greenbelt, MD, USA
¹²Institute of Environmental Physics (IUP), University of Bremen, Germany
¹³Department of Physics, University of Toronto
¹⁴Finnish Meteorological Institute, Helsinki, Finland
¹⁵Greifswald University

Correspondence: T. von Clarmann (thomas.clarmann@kit.edu)

Abstract. Remote sensing of atmospheric state variables typically relies on the inverse solution of the radiative transfer equation. An adequately characterized retrieval provides information on the uncertainties of the estimated state variables as well as on how any constraint or a priori assumption affects the estimate. Reported characterization data should be intercomparable between different instruments, empirically validatable, grid-independent, usable without detailed knowledge of the instrument or retrieval technique, traceable, and still have reasonable data volume. The latter condition of adequacy may force one to work with representative rather than individual characterization data. The main sources of uncertainty are measurement noise, calibration errors, simplifications and idealizations in the radiative transfer model and retrieval scheme, auxiliary data errors, and uncertainties in atmospheric or instrumental parameters. Some of these errors affect the result in a random way, while others chiefly cause a bias or are of mixed character. Beyond this, it is of utmost importance to know the influence of any constraint
and prior information on the solution. While different instruments or retrieval schemes may require different error estimation schemes, we provide a list of recommendations which shall help to unify retrieval error reporting.

1 Introduction

Atmospheric science studies often rely on measurements Observations from remote sensing instruments. The ability to draw robust scientific conclusions from these observations are central to many studies in atmospheric science. The robustness of the conclusions drawn in these studies is critically dependent on adequate characterization of the data and all aspects of their reduction--the characteristics of the reported data, including their uncertainty, resolution, and dependence on any a priori information. Adequate communication of these data characteristics is therefore essential. Further, when, as is increasingly the case, observations from multiple sensors are considered, it is important that these characteristics be described in a manner that allows for appropriate intercomparison of those characteristics and the observations they describe. In the satellite community, however, the definition of adequate what constitutes "adequate communication" is far from uniform. Currently, multiple retrieval methods are used by different remote sounding instrument groups, and various approaches to error or uncertainty estimation are applied. Furthermore, reported uncertainties are not always intercomparable, e.g., because they include different error components. Some kinds of uncertainties are sometimes not reported at all. The different spatial resolutions and the different content of prior information in the data products are a particular problem, readily intercomparable. For example, the metrics used as uncertainty values for a data set might not be properly defined (as, say, 1σ or 2σ values, or as an appropriate confidence interval), uncertainty values might not be adequately described as "random" or "systematic" in nature (let alone any more nuanced description of inter-error correlations), spatial resolution information or the influence of a priori content might not be given, etc. This paper discusses these issues and proposes a common framework for the appropriate communication of uncertainty and other measurement characteristics.

The project--This review has been undertaken under the aegis of the ‘Towards Unified Error Reporting’ (TUNER) is a consortium of project and was carried out by retrieval experts from the atmospheric remote sensing community (including active participation from eight different instrument science teams) who have come together to tackle the (arguably daunting) goal of establishing a consensus approach towards reporting errorsto enable for reporting errors, hopefully enabling more robust scientific studies using the retrieved data. To achieve this, TUNER largely aims at promoting the adoption of consistent and intercomparable error estimates for remote measurements of atmospheric composition and temperature profiles. Such an adoption also requires a consistent and intercomparable characterization of spatial resolution and content of a priori information of the geophysical data products. This review paper, the first ‘foundational’ paper from the TUNER team, is mainly addressed to the providers of remotely sensed data. This is a summary paper of sorts, which methodically reviews background material before making suggestions as to how unified error reporting can be achieved. 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).

Most concepts presented in this paper rely on the assumption that it is enough to provide providing the user with the result of the retrieval, a measure of estimated error or uncertainty along with correlation information, and sensitivity to possible a
priori information used. That is to say, is sufficient for most scientific uses. In other words, there is no need for more detailed discussion of the expected distribution functions of the retrieved values around a true value (or around the expectation value of the retrievals are do not need) to be provided, although. That said, we recognize that they might be useful for other quantitative application some specialized quantitative applications.

The well-informed reader will already be acquainted with most of the material in this paper, while although those less familiar with retrieval algorithms may find this it a useful introduction. In either case, the following outline will hopefully assist the reader in navigating the wealth of information discussed here. Firstly we list conditions of adequacy of the reporting of error and uncertainty (desiderata desiderata), which summarize the information that should be provided to the data user (Section 2). Next, before diving into headlong into the technical details, Section 3 attempts to offer some necessary clarification with some of various terminological issues. In Section 4 we lay down the formal background. In particular, we discuss the retrieval equation and try to provide unambiguous interpretations of all involved terms. Here, enabling the informed reader may have a chance to calibrate their to map their own notation and terminology to ours that discussed herein. In our discussions of retrieval theory we will not reinvent the wheel but build heavily on the framework laid out by Rodgers (1976, 1990, 2000). Importantly, however, our characterization is applicable beyond the retrieval schemes endorsed therein, including many in every day use among remote sounding teams. Section 5 discusses how the theory translates into real world problems. In particular, we discuss, centering on how the full retrieval problem is decomposed into sub-problems. Following this, we turn towards error estimation and uncertainty assessment. We then systemize and discuss the various sources of retrieval error (Section 6) and, if applicable, their dependence on the retrieval scheme chosen. We identify data characterization methods currently in use and relate them to the theoretical concepts presented. Finally recommendations Recommendations on unified error reporting for space-borne atmospheric temperature and composition measurements are given in Section 7. In these recommendations we refrain from stipulating conventions and confine ourselves to recommendations that can directly be inferred from the conditions of adequacy. Finally, we identify unsolved problems and applications which might not be fully covered by our framework (Section 8).

2 Conditions of Adequacy adequacy for Diagnostic Metadata diagnostic metadata

With the ultimate goal of presenting a list of recommendations to the community of data providers, we must first discuss a list of desired properties of diagnostic metadata from the point-of-view of a data user. We refer to diagnostic metadata as error or uncertainty estimates and all information on the content of a priori data, spatial resolution, and the like. The list of possible metadata to characterize retrievals of atmospheric state variables is huge, but some of them are more useful than others. Here we define conditions of adequacy (CoA) for error and uncertainty reporting. These conditions will be used as criteria when recommendations are developed on which metadata are indeed essential and should thus find their way into the recommendations.

CoA 1. The error estimates should be intercomparable among different instruments, retrieval schemes, and/or error estimation schemes.
CoA 2. The estimated errors should be independent of the vertical grid such that correct propagation of the errors to a different grid yields the same errors as would the direct evaluation for a retrieval on the new grid.

CoA 3. The error budget shall be useable without detailed technical knowledge of the instrument or retrieval technique. This enables the data user to correctly apply error propagation laws and calculate uncertainty in higher level data products.

CoA 4. The error analysis shall be traceable in a sense that all relevant underlying assumptions are documented.

CoA 5. In principle the error estimates should be empirically validatable. Empirical validation is achieved via comparison between independent measurements because the true values of the atmospheric state are unknowable. We consider error estimates as empirically adequate if differences between independent measurements can be fully explained by the combination of their error bars, natural variability in the case of less than perfect collocations, different resolution in time and space, and different amounts of possibly different prior information.

CoA 6. The data volumes associated with this reporting should be reasonable. This is particularly important because involved matrices (e.g., covariances and averaging kernels) exceed the data volume of the data themselves by orders of magnitude.

These conditions of adequacy comply in part with the principles issued by the QA4EO task team (2010). That document requests traceability and fitness for purpose. We endorse traceability of the uncertainty estimates but we consider it unrealistic to assign quality indicators for ‘fitness for purpose’ for all conceivable applications.

3 Terminological Issues

Unification of error reporting is only achievable if at least a minimum agreement on terminology and the underlying concepts is achieved. Most of the terms used are largely self-explanatory and are introduced in the following sections. There are, however, two troublesome terminological issues. One consists of the dispute as to whether ‘estimated error’ and ‘uncertainty’ relate to the same concept and, if not, which concept is appropriate. The other is related to the exact connotation of these terms with respect to the underlying methodology. In the following, both issues will be briefly discussed.

3.1 Error versus Uncertainty

A particularly troublesome terminological issue is the use of the term ‘error’ and the concept behind it. Given that the Joint Committee for Guides in Metrology (JCGM) and the Bureau International des Poids et Mesures (BIPM) aim to replace the concept of error analysis by the concept of uncertainty analysis (Guide to the expression of uncertainty in measurement (GUM), 2008a), some conceptual and terminological remarks are in order. While on the face of it, this is quibbling about words, it is actually claimed in these documents that there are conceptual differences between error analysis and uncertainty estimation. A deeper discussion of this issue is beyond the scope of this paper. The interested reader is referred to, e.g., Thomas von Clarmann (paper in preparation), who challenges the principal difference between the error concept and the uncertainty concept; Bich (2012), who, although a Working Group leader of the Joint Committee for Guides in Metrology (JCGM), claims inconsistencies
between the GUM document and its supplements; Grégis (2015), who challenges the position that one can dispense with the notion of ‘true value’ in metrology as suggested in GUM; or Elster et al. (2013) and European Centre for Mathematics and Statistics in Metrology (2019), where the applicability of the GUM concept to inverse problems is critically discussed. Conversely, QA4EO task team (2010), Merchant et al. (2017), and Povey and Grainger (2015), e.g., largely endorse the GUM-based uncertainty concept. The latter authors, however, state that the GUM conventions “[...] 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.” The application of GUM to remote sensing of the atmosphere are hampered by the fact that GUM does not explicitly take indirect measurements into account, that GUM assumes a well defined measurand while the atmosphere is characterized by statistical variables which do not relate to a canonical ensemble, and that the problem of a priori information is not considered. For our purposes it is sufficient to say that the claim of the conceptual difference is still under debate, and that we have not fully adopted the terminology stipulated by the Joint Committee for Guides in Metrology (JCGM). Instead, we invoke the statement in Joint Committee for Guides in Metrology (JCGM) (2008a) that the error concept and the uncertainty concept are ‘not inconsistent’; we understand this in a sense that the underlying methodology and mathematical tools are the same, and that the differences are restricted to the interpretation of the terms under dispute.

The GUM-stipulated framework, however, does present a dilemma when seeking to unify terminology in the TUNER arena. On the one hand, we are not in favour of brushing away the common interpretation whereby the term ‘estimated error’ is used for a statistical quantity that reflects the difference between the true value and the value inferred from the measurement. It remains to be seen whether the new terminology stipulated by the Joint Committee for Guides in Metrology (JCGM, 2008a and 2008b) will be widely accepted. Accordingly, given the significant heritage within the atmospheric remote sensing community, renaming long-established concepts would not promote our goal of ‘unification’. In recent scientific literature, terms like ‘estimated measurement error’, ‘error analysis’, ‘error covariance matrix’ or ‘standard error of the mean’ are still widely in use, and replacement terms like ‘standard uncertainty of the mean’ etc., are rarely invoked. On the other hand, we recognize that explicitly breaking with the official stipulations of the JCGM does not advance the overall goal of ‘unification’ either.

For the purposes of the following discussion we define ‘error’ to be the difference between an unknown truth and a value inferred from measurements. ‘Uncertainty’ describes the distribution of an error. This can be summarized with metrics such as the total squared error, which can be decomposed into systematic and random components that are reflected by bias and variance. We will often use the word ‘error’ as a part of composite terms, (e.g., ‘parameter error’, ‘noise error’, ‘retrieval error’, ‘estimated error’, etc.). When we use a composite containing the term ‘error’, this does not imply that the uncertainty interpretation is excluded, and conversely, when we use a composite term containing the term ‘uncertainty’, this does not imply that the error interpretation is excluded. The use of the term ‘error’ as a generic term in the sense of ‘measurement noise causes an error in the inferred quantity’ is probably uncontroversial and can be accepted both by adherents of the error concept and adherents of the uncertainty concept.
We think that no terminology is *per se* better than another one, as long as it is clearly defined. Instead of further fueling the terminological conflict, we try to concentrate on the content and to lay down an error reporting framework custom-tailored to remote measurements of atmospheric temperature and constituents that is more detailed and specific than most of the previous literature.

### 3.2 Ex ante versus ex post error estimates

Regardless of whether one prefers to call the estimated retrieval error ‘uncertainty’ or the uncertainty of the measurement ‘estimated error’, there are still two different ways to evaluate this quantity. One relies on generalized Gaussian error propagation or, particularly in grossly nonlinear problems, on sensitivity studies, either as case studies or in a Monte Carlo sense. Uncertainties of input quantities are propagated through the data analysis system to yield the uncertainties of the target quantities. The other way relies on a statistical analysis of the results, e.g., by comparison to other observations. Many different terms are commonly used to distinguish between these different approaches. In Joint Committee for Guides in Metrology (JCGM) (2008a), the first fall into their ‘category B’, while the second are ‘category A’. Von Clarman (2006) distinguishes between *ex ante* and *ex post* error estimates, reflecting the fact that error propagation can be calculated even before the measurement has been made, while the statistical analysis of the measurements requires the availability of actual measurements. Along the same line of thought, one could also talk about error prediction *versus* evidence of errors. Since error estimation is deterministic with respect to the estimated variances (but certainly not with respect to the actual realizations of the measurement error), and since statistical analysis of any evidence follows the laws of inductive logic (Carnap and Stegmüller, 1959), one could also distinguish between deductive and inductive error estimation. Others prefer to use the terms ‘bottom up’ and ‘top down’ for this dichotomy. This study focuses chiefly on *ex ante* error estimation. To validate these estimates, *ex post* error estimation is relevant, as expounded, e.g., by Keppens et al. (2019).

### 4 Retrieval Theory and Notation

Measurements – also most so-called direct measurements – invoke inverse methods. The only exception is a direct comparison where the measurand is directly accessible via human sensation, like length measurement by comparison with a yardstick or determination of colour by comparison with a colour table. The inverse nature of most measurements is due to the fact that the measurand \( x \) is the cause and the measured signal \( y \) is the effect. These are connected via a natural regularity which is formalized via a function

\[
f : \mathbb{R}^n \to \mathbb{R}^m : x \mapsto y = f(x),
\]

which maps the discretized measurand onto the respective observable signal, and where \( m \) and \( n \) designate the number of measured data points and the number of state values, respectively.

In the macroscopic world, exempt from quantum processes, the measured effect is thus, for given conditions, a deterministic unambiguous function of the measurand. While microscopic processes can admittedly be indeterministic, their statistical...
treatment for canonical ensembles ensembles of sufficient size leads to deterministic laws. Irreducibly non-deterministic components contribute to the measurement noise. In contrast, the conclusion from the measured signal \( y \) to the measurand \( x \) is not always unambiguous because in many cases the inverse function

\[
g : \mathbb{R}^m \rightarrow \mathbb{R}^n : y \mapsto x = f^{-1}(y)
\]

can 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.

In some cases, the inverse process can be quite trivial, e.g., in the case of a temperature measurement with a mercury thermometer. The causal process is the thermal expansion of mercury, and the inverse conclusion goes from the volume of the mercury to the ambient temperature. The scale of the mercury thermometer is simply a pretabulated solution of the inverse process for various temperatures. In other applications, such as remote sensing of the atmosphere from space, the inverse process is slightly more complicated because \( f^{-1}(y) \) does not usually exist. Related workarounds to solve this problem are discussed below.

Remote sensing of the atmospheric state from space relies in one form or another on the radiative transfer equation (Chandrasekhar, 1960). This equation is deterministic in a sense that its formulation \( f \) simulates the measured signal via causal processes. The deterministic characteristic of \( f \) in the macroscopic world is achieved via a statistical treatment of the underlying microscopic processes. While its forward solution allows the calculation of the radiance received by the instrument, its inverse solution allows for the determination of the state of the atmosphere from a known radiance signal.

Roughly following the notation of Rodgers (2000), we define \( \mathbf{F}^{-1} \) to be the radiative transfer model which approximates \( f \); \( \mathbf{F} \) is a vector-valued non-linear function and deviates from \( f \) in that it is discrete in space and frequency, involves numerical approximations and may not include the full physics of radiative transfer. \( x \in \mathbb{R}^n \) is the vector representing the atmospheric state, and \( y \in \mathbb{R}^m \) the vector containing the measured radiance signal. The elements of \( x \) contain both the ‘target variables’ and ‘joint-fit’ variables. Target variables are those variables we are actually interested in. Conversely, the joint-fit variables are variables needed by \( \mathbf{F} \) that, while not the focus of our interest, have to be sought in the inversion if not accurately known in order to minimize error propagation because they may not be accurately known and their uncertainties would otherwise make an unacceptably large contribution to the total error budget.

Typically \( m \neq n \), i.e., the dimension of \( x \) does not equal the dimension of \( y \). For the overdetermined case \( (m > n) \), Gauss (1809)\(^1\) suggested an approximate inversion obtained by minimizing the sum of squares of the residual \( \mathbf{F}(x) - y \). If we assume, for now, that \( \mathbf{F} \) is linear and that Gaussian distributions of are adequate to characterize the measurement (see Section 5.5 for related problems) the “maximum likelihood” unconstrained \(^2\) solution of the inverse problem is

\[
\hat{x}_{\text{ML}} = x_0 + \left[ K^T S_{y,\text{total}}^{-1} K \right]^{-1} K^T S_{y,\text{total}}^{-1} \left( y - \mathbf{F}(x_0) \right).
\]

\(^1\)The first publication of a least squares method was actually by Legendre (1805), but Gauss is said to have had this idea about ten years before. Obviously unaware of Legendre’s work, also Robert Adrain (1808) proposed the least squares method as the most advantageous solution in this context. See Merriman (1877), Sprott (1978), or Stahl (2006) for a deeper discussion of the priority regarding this method.

\(^2\)See below for a deeper discussion of this term.
\( K \) is the Jacobian matrix with the elements \( K_{ij} = \frac{\partial y_i}{\partial x_j} \), \( x_0 \) represents an initial guess of the atmospheric state, and \( S_{y,\text{total}} \) is the covariance matrix characterizing the total measurement error. Here the ' symbol indicates that, due to measurement noise mentioned above, and other uncertainties and ambiguities which will be discussed below, the result of the inversion is only an estimate of the measurand \( x \). In most real-world applications, only measurement noise is considered here, while other measurement uncertainties like calibration errors are neglected at this stage. Since the solution provided by Eq. 3 does not consider any prior information, it is a “maximum likelihood” solution in the sense of Fisher (1922, 1925).^3

One major difference between our notation and Rodgers’ notation refers to the error covariance matrices \( S \). We use two subscripts. The first indicates if the uncertainties refer to the retrieved quantities \( x \) or to the ingoing quantities. The second subscript specifies the source of the uncertainty. For example, \( S_{y,\text{noise}} \) is noise in the measurement data, while \( S_{x,\text{noise}} \) is the measurement noise mapped into the retrieved atmospheric state. In other words, \( S_{x,\text{noise}} \) is the error component in \( x \) due to the error source \( S_{y,\text{noise}} \). In some cases, e.g., if any ambiguity can be excluded or if the sources of the error are not known, the second subscript can be missing.

By explicitly assuming equally distributed, i.e., uniform prior, state values Gauss (1809, p. 211) gave this solution a probabilistic interpretation without clashing with the Bayes (1763) theorem. In a linear context and for measurement errors following a normal\(^4\) distribution around the true value, the Gaussian least squares solution corresponds formally — but certainly not in terms of its interpretation — to a maximum likelihood solution in the terminology of Fisher (1922, 1925) (thus the index ML in \( \hat{x}_{\text{ML}} \)). An interesting overview on the history of maximum likelihood estimates is given by Hald (1999), while the justification of this method is critically discussed in Aldrich (1997). For instructive discussions of the relevance of the Bayes theorem in inductive statistics, see, e.g., Bar-Hillel (1980) and Thompson and Shuman (1987). The original Gaussian least squares method was valid for independent measurement errors only. The introduction of the correlation coefficient by Galton (1888) and Pearson (1896) paved the way towards a wider range of applications. The matrix formulation as used today, where correlated measurement errors are represented in the measurement error covariance matrix \( S_y \), owes much to Yule (1907), Fisher (1925), and Aitken (1935). A reconstruction of the historical development of this technique was performed by Aldrich (1998).

If the inverse problem is underdetermined \((m < n)\) or ill-posed in a sense that the \[ K^T S_{y,\text{total}}^{-1} K \] matrix is singular or has a high condition number, then a constraint has to be used. Even in formally well-conditioned problems but large measurement noise, the use of a constraint can be helpful. With a prior assumption on the atmospheric state \( x_a \) and a regularization matrix \( R \) we can modify Eq. (3) in a way that the matrix inversion can be accomplished. This so-called regularized solution is (von Clarmann et al. 2003, building upon Rodgers 2000; Phillips 1962; Tikhonov 1963; Twomey 1963; Steck and von Clarmann 2001)

\[
\hat{x}_{\text{reg}} = x_a + \left[ K^T S_{y,\text{total}}^{-1} K + R \right]^{-1} K^T S_{y,\text{total}}^{-1} [y - F(x_a)].
\]

\(^3\)See below for a deeper discussion of this term.

\(^4\)Normal distribution and Gaussian distribution are the same. The term ‘normal distribution’ was probably coined by Karl E. Pearson in 1893. While this term evades the question of priority in its discovery, it “has the disadvantage of leading people to believe that all other distributions of frequency are in one sense or another abnormal”, as Pearson (1920) self-critically states.
Many choices of the regularization matrix \( R \) are possible. With the choice of \( R \) as a squared \((n-1) \times n\) first order differences matrix \( \Delta \)

\[
R = L_1 L_1^T,
\]

the resulting \( L_1 \) and \( \gamma \) a scaling parameter to control the strength of the regularization and balancing the units, the choice of

\[
R = \gamma L_1 L_1^T,
\]

renders fields of profiles of atmospheric state variables are smoothed in the sense of reduced altitude-to-altitude differences of the \( \hat{x}_{\text{reg}} - x_a \) profile, thus avoiding unphysical oscillations that typically result from instabilities associated with ill-posed inverse problems.

If we represent the best known a priori statistics about the targeted atmospheric state as \( x_a \), its covariance matrix as \( S_a \), the inverse of this matrix as \( R \), and continue to assume Gaussian error distributions, then we get a Bayesian solution that is usually referred to as ‘optimal estimate’ (Rodgers, 1976) or ‘maximum a posteriori (MAP) solution’ (Rodgers, 2000) and is fully compatible with the Bayes (1763) theorem and information theory by Shannon (1948), and thus gives the solution a probabilistic interpretation in the sense of the maximum a posteriori probability:

\[
\hat{x}_{\text{MAP}} = x_a + \left[ K^T S_{y,\text{total}}^{-1} K + S_a^{-1} \right]^{-1} K^T S_{y,\text{total}}^{-1} \left[ y - F(x_a) \right].
\]

The formalism of Eq. 6 can also be used without committing oneself to a probabilistic interpretation of \( S_a \). For example, \( S_a \) can be rescaled to give less weight to the priori information.

This equation, however, holds only if the variability of the atmospheric state is fairly well covered by a Gaussian probability density function. To characterize the variability of highly variable trace gases, a log-normal probability density function can be more adequate. It avoids, e.g., that non-zero a priori probability densities are assigned to negative mixing ratios. Technically, this is achieved by using Eq. (6) but re-interpreting \( x \) as the logarithm of the concentrations and \( S_a \) as the covariance matrix of these logarithms. This is, for instance, important for tropospheric water vapour (e.g., Hase et al., 2004 or Schneider et al., 2006). However, there is a price to be paid, in that this then casts the measurement error in terms of a log-normal distribution also. The positive bias caused by the retrieval of logarithms of concentrations in the case of measurement noise oscillating around zero signal has been investigated by Funke and von Clarmann (2012).

For brevity, we define the gain matrix (Rodgers, 2000)

\[
G = \frac{\partial \hat{x}}{\partial y} = \left( K^T S_{y,\text{total}}^{-1} K + R \right)^{-1} K^T S_{y,\text{total}}^{-1},
\]

which will play an essential role in error estimation. The remainder of this paper broadly identifies all relevant sources of uncertainties including measurement noise, approximations, idealizations, and assumptions.

Tables 1 and 2 summarize the retrieval schemes used by a number of satellite data processors.
5 Retrieval in the Real World

Application of Eqs. (3–6) usually involves many approximations and idealizations including discretization, decomposition of the argument of the radiative transfer function into variables and parameters, spatial decomposition, and non-linearity issues, just to name a few. Since all these approximations give rise to retrieval errors, a full understanding of them is of utmost importance when quantifying the error budget of a measurement.

5.1 Discretization

At least on macroscopic scales, atmospheric state variables are construed as continuously varying in space and time. In the retrieval equations they are, however, represented by vectors with a finite number of elements, each representing the atmospheric state at a gridpoint. If the discretization is too fine, a stronger regularization is needed to fight ill-posedness of the inversion, while a too coarse discretization can cause errors in the radiative transfer modelling and limits the spatial resolution of the solution. In a maximum likelihood retrieval, the grid-width is identical to the spatial resolution of the retrieval.

In this context we note that the atmospheric state does not necessarily need to be represented as vertical profiles where each element of $\mathbf{x}$ is a state variable at a certain altitude or location, or represents an atmospheric layer. Alternative representations include, e.g., principal components/empirical orthogonal functions (see, e.g., Boukabara et al. 2011; Munchak et al. 2016; Duncan and Kummerow 2016). These can be inferred from an ensemble of spatially highly resolved prior measurements. The unknowns in the retrieval are the weights of the principal components. Complete or partial neglect of higher principal components will regularize the retrieval. Such an approach is under consideration for the Atmospheric Limb Tracker for Investigation of the Upcoming Stratosphere (ALTIUS) mission (Fussen et al., 2016). A similar approach was tried for the Scanning Imaging Absorption Spectrometer for Atmospheric Chartography (SCIAMACHY) (Doicu et al., 2007), for the multi-channel infrared radiometer on the Geostationary Operational Environmental Satellite (GOES-13) and infrared sounder on the Indian National Satellite (INSAT-3D) by Jindal et al. (2016). The retrieval of vertical column amounts by simply scaling the initial guess profile reduces the profile retrieval to a single degree of freedom. Alternatively, the altitude axis of the profile can be stretched or compressed using the so-called ‘downwelling factor’ as suggested by Toon et al. (1992). These approaches are often used for analysis of measurements which do not provide direct information on the vertical distribution of the target species. Particularly in the greenhouse gas monitoring community, retrieved column amounts of target species are divided by the molecular oxygen column amount retrieved with the same instrument. Rescaling of the quotient by the 0.20946 gives the column-averaged dry-air mole fraction XCO$_2$ or XCH$_4$. The benefit of this approach is a cancellation of multiplicative systematic error components (see, e.g., Wallace and Livingston, 1990; Yang et al., 2002; Wunch et al., 2010; Reuter et al., 2011). Similar arguments hold for isotopic ratios (e.g., Piccolo et al., 2009; Schneider et al., 2016) or ratios between trace gas profiles (e.g. García et al., 2018).
5.2 The Measurement Error Covariance Matrix

Typically in real-world applications, the measurement error $S_y$ in Eqs. (3–6) contains only measurement noise, while other sources of measurement error are often ignored during the retrieval (see Section 6.1 for details) and typically analyzed after performing the retrieval. Since this treatment deprives any solution from its claimed optimality (Cressie, 2018), in some cases the measurement noise is artificially “inflated” to account for potential calibration uncertainties. A method for how to include multiple types of uncertainties can be included in the measurement error covariance matrix is discussed in Marks and Rodgers (1993), Tarantola and Valette (1982), Eriksson (2000), and von Clarmann et al. (2001). These authors discuss the possibility of mapping all relevant error contributions into the measurement space and include them in the $S_y$ matrix. Rodgers (2000, Section 4.1.2) views this problem from a different perspective but the suggested solution is mathematically equivalent to the approach suggested above.

5.3 Variables and Parameters

While the measurement typically depends on a large number of atmospheric state variables, only a few of them are actually dealt with as unknowns. The other variables are assumed to be known and are dealt with as constant parameters. For example, in an ozone profile retrieval the atmospheric temperature profile may be assumed to be known and thus not be included in the retrieval vector $x$. With this, the forward problem can be formalized as

$$y = F(x; b),$$

where $b$ is the vector of parameters, which are separated in the argument of function $f$ by the semicolon. The respective inverse solution reads in its general form

$$\hat{x}_{\text{reg}} = x_a + \left(K^T S^{-1}_{y,\text{total}} K + R\right)^{-1} K^T S^{-1}_{y,\text{total}} \left[y - F\left(x_a; b\right)\right].$$

The uncertainties of parameters $b$ affect the estimate $\hat{x}$ and thus a parameter error term has to be included in the error budget.

5.4 Decomposition of the inverse problem

Practical reasons typically force one to decompose the inverse problem, e.g., to reduce the size of the problem in order to achieve numerical efficiency. Often a part of the measurement is virtually insensitive to some of the atmospheric state variables. The general idea of decomposition is to isolate subsets of the entire set of measurements that are mainly sensitive to only a subset of the unknown variables. This decomposition can be made according to spectral or geometrical criteria (see below).

Decomposition of the inverse problem can be done either in an “optimal” or in a “non-optimal” way. The optimal decomposition solves the inverse problem sequentially, where at each step the retrieval is made for the full $x$-vector but based only on a subset of the measurements, whereby each measurement is used only once during the entire process (see, e.g., Rodgers,

---

5 This issue seems to be of particular importance when observation error covariance matrices are built in contexts where a data assimilation scheme uses radiance measurements instead of retrieved state variables, as suggested by Andersson et al. (1994).
2000, Ch. 5.8.1.3; his requirement of a diagonal measurement covariance matrix can be replaced by the weaker requirement of a block-diagonal covariance matrix if the algebra is adjusted accordingly). Initially, the retrieval, which typically is patently under-determined because of the temporarily ignored measurements, is constrained by an initial $S_a$ matrix. For each subsequent step, the $S_a$ matrix is replaced by the so-called ‘retrieval covariance matrix’

$$S_x = \left( K^T S_{y,\text{total}}^{-1} K + S_a^{-1} \right)^{-1}$$

of the preceding step. Within linear theory, the solution of such sequential methods is equivalent to the direct solution of the full inverse problem.

More frequently used is non-optimal decomposition. Here the relevance of some components of the state vector for the measurements is temporarily ignored, and the retrieval solves the inverse problem only for a part of the state values, using only a subset of the measurements. This approach lends itself to problems where it is adequate to assume that the Jacobian matrix $K$ has an almost block diagonal structure, that is, that there are state variables which have no significant influence on some of the measurements under analysis and vice versa. In the following we discuss spectral and spatial decomposition.

### 5.4.1 Spectral decomposition

Not all spectral gridpoints or channels of a spectrometer or a multi-channel radiometer are equally sensitive to all unknown variables. For example, the subset of the measurements used to retrieve the ozone concentration may be insensitive to the concentration of water vapour (Flittner et al., 2000). In such cases, the abundances of various species can be retrieved in sequence, using dedicated “microwindows” in infrared spectroscopy (see, e.g. von Clarmann and Echle, 1998; Echle et al., 2000; Dudhia et al., 2002), different spectral regions in microwave radiometry (Livesey et al., 2003, 2006) or measurements in the ultraviolet and visible (UV-VIS) spectral range (e.g., Bovensmann and M. Gottwald, 2011). In these cases, a subset of spectral points is selected for analysis. Those unknowns which have sizeable impact on the signal at these spectral points are retrieved. When in later steps other spectral points are analyzed, the results of the first steps can be used and either be treated as known parameters, or as a priori information in an optimal sequential scheme. Uncertainties entailed by this procedure are associated with the following problems: (1) In the first step some of the disregarded variables may still introduce some error; (2) retrieval errors of all kinds resulting from a prior step of the sequential scheme propagate onto the results of later steps; and (3) inconsistencies in spectroscopic parameters between different spectral points can cause a spurious residual signal when, e.g., the concentration of a gas retrieved in one part of the spectrum is used as a known parameter in the analysis of another part of the spectrum.

Spectral decomposition is also often used for the retrieval of a single species. For example, Flittner et al. (2000) or Sonkaew et al. (2009) Kramarova et al. (2018) retrieve ozone sequentially in different spectral bands. An alternative to spectral decomposition is the simultaneous analysis of the full spectrum (e.g., Serio et al., 2016). In cases when spectroscopic data are consistent over the entire spectral range it will best exploit the observational information.
5.4.2 Geometric decomposition

In the case of nadir sounding, lines of sight referring to different ground-pixels cross different parts of the atmosphere and can thus be analyzed independently without sizeable loss of information. In limb sounding, first suggested around the same time by Gille (1968), Blamont and Luton (1972), Hays et al. (1973) and Donehue et al. (1974) for different scientific contexts, the situation is more complicated because the retrieval of a state value at a given altitude depends on the knowledge of the same state value at other altitudes passed by the line of sight.

If the same air parcel is seen under multiple geometries, the measurements have a tomographic nature. Since the simultaneous retrieval of all these intertwined measurements easily exceeds available computational resources, often only a subset of the measurement geometries are analyzed in one step.

More specifically, the algorithm can be constructed such that only a subset of the measurements are needed to retrieve the atmospheric state corresponding to a given subset of the state vector elements that affect signals along the raypath of the considered measurement. The two most prominent examples are single profile retrievals and onion peeling. Typical approaches to decompose the entity of measurements geometrically are listed in Table 1.

In some cases, the geometric profile reconstruction is decoupled from the spectral inversion. In order to gain numerical efficiency, the inversion can be performed in sequential steps. Such an approach is realized for GOMOS two-step inversion, which decompose the retrievals into the spectral inversion followed by the vertical inversion using the concept of effective cross-sections (Kyrölä et al., 1993).

5.4.3 Optimal Decomposition Techniques

Optimal decomposition techniques formally retrieve all relevant variables $x$ in each step but measurement information $y$ of only a subset of the measurement geometries is used. Since, in a maximum likelihood setting, such a retrieval would be hopelessly underdetermined, sequential estimation as described above lends itself to this class of problem. Every state variable can be updated as soon as new information becomes available. In contrast, non-optimal techniques will not update any quantity once retrieved.

5.4.4 Single Profile Retrieval vs. 2D/3D-retrievals

The vast majority of limb sounding retrievals assume local spherical homogeneity of the atmosphere, i.e. considering only vertical variations in the atmospheric state around the line of sight, and neglecting horizontal variability (e.g. Gille, 1968; McKee et al., 1969a, b; House and Ohring, 1969; Carlotti, 1988). Russell III and Drayson (1972) explicitly state the assumption, and only a small number of retrieval schemes relinquish it. In solar occultation observations, where the measurement geometry is determined by the position of the sun and the instrument and where at most one sunset and one sunrise can be observed per orbit, there is not much choice; tomographic multi-limb-scan retrievals are out of reach and the single profile retrieval is the way to go.
For limb measurements, von Clarmann (1993) suggested a non-optimal decomposition similar to “onion peeling” (see below) but in the horizontal domain. This approach, however, was never put into action. Carlotti et al. (2001) proposed to solve the inverse problem for a full satellite orbit instead of for single limb-scans. This tomographic method was published under the name ‘geofit’. Steck et al. (2005) tested an implementation of sequential estimation in the horizontal domain, while the vertical domain was treated in one leap. Livesey and Read (2000); Livesey et al. (2008); Christensen et al. (2015) employ a tomographic approaches, whereby a 2-dimensional along-track curtain of profiles is simultaneously retrieved from multiple sets of limb scans. A similar approach is used for SCIAMACHY retrievals of metals (Scharringhausen et al., 2008; Langowski et al., 2014) and NO (Bender et al., 2013, 2017) and for OMPS-LP ozone (Zawada et al., 2018).

Dudhia and Livesey (1996) and von Clarmann et al. (2009) use prior information on the horizontal variation of state variables in a single limb-scan retrieval. The latter scheme lends itself particularly to reprocessing of data when the initial processing information on the horizontal variability is already available. This approach has been critically analyzed by Castelli et al. (2016). Tomographic approaches and the effect of horizontal gradients were investigated for SCIAMACHY limb measurements by Pukite et al. (2008) and Pukite et al. (2010). A series of OSIRIS orbits allowed the tomographic analysis of polar mesospheric clouds (Hultgren et al., 2013). This application was preceded by theoretical studies on OSIRIS infrared channels tailored for tomography.

Most other limb sounding retrieval schemes use the spherical homogeneity approximation, although this approach can be challenged for limb sounders. For example, Kiefer et al. (2010) provided evidence of biases in trace gas retrievals from MIPAS limb emission spectra due to horizontal temperature gradients. Thus, neglect of the horizontal variation of the atmospheric state needs either to be corrected or to be considered in the error budget.

In the case of nadir sounding, single profile retrievals seem to be the natural thing to do, since a raypath associated with one geolocation intersects each altitude level only once. However, in the UV-VIS spectral range multiple scattering along with strong inhomogeneities in the surface reflection or cloud coverage might cause some interplay between the neighbouring pixels.

**5.4.5 Onion Peeling**

In the “onion peeling” approach, (Gille, 1968; McKee et al., 1969a, b; House and Ohring, 1969; Russell III and Drayson, 1972; Goldman and Saunders, 1979) the collection of limb measurements in a vertical scan is decomposed into a sequence of retrievals, each dealing with one tangent altitude, starting at the top and working down. This method builds upon the fact that the bulk of the information obtained along the horizontal line of sight originates from the vicinity of the tangent point, with limited information from above and essentially none from below. In the first step, the measurement associated with the uppermost tangent altitude is analyzed and the profile above is scaled. Then the second tangent altitude from the top is used and the profile between this tangent altitude and the tangent altitude above is scaled; this is repeated until the lowermost tangent altitude is reached. Often the discretization of the atmospheric state corresponds to the tangent altitude pattern, i.e., there is one profile point per tangent altitude, and the profile shape between the points is determined by interpolation. Gaussian elimination is already provided by the measurement geometry, and the Jacobian $K$ has a quasi-triangular structure. This approach, however,

---

6We are saying ‘quasi-triangular’ here because, due to over-determination at each tangent altitude, $K$ can have more rows than columns.
is prone to instabilities (von Clarmann et al., 1991). The alternative would be that layer values are retrieved instead of level values.

In the early era of limb sounding and solar occultation measurements, onion peeling was the work-horse data analysis algorithm and was used, among others, in the following missions: LIMS (Bailey and Gille, 1986), ATMOS (Norton and Rinsland, 1991), HALOE (Russell III et al., 1993), CRISTA (Offermann et al., 1999). More recently, onion-peeling related algorithms have been used, e.g., for, TIMED-SABER (Russell III et al., 1994), AIM-SOFIE (Gordley et al., 2009a), and SCIAMACHY (Noël et al., 2018). When more computer power along with quasi-analytical algorithms to calculate larger Jacobians became available, onion peeling was often superseded by global-fit-like algorithms (Carlotti, 1988) which solve the inverse problem for the entire limb sequence in one leap.

Approaches related to onion peeling are the Mill-Drayson method (1978) and the ‘interleave method’ (Thompson and Gordley, 2009). The Mill-Drayson method starts with the lowermost tangent altitudes and scales the entire profile of the atmospheric state variables above to minimize the residual between measurement and modeled signal. Next, the second tangent altitude from bottom is used to scale the related upper segment of the profile above the related tangent altitude. Several iterations over the limb scan are made. The goal is to avoid the typical onion-peeling error propagation which tends to trigger oscillations in the profiles. This method became somewhat obsolete with the advent of numerical regularization. Without knowledge of the original method by Mill and Drayson this method has been applied to the SOFIE instrument by Marshall et al. (2011).

The interleave method decomposes the limb scan in multiple disjoint subsets of measurements, e.g., such that one set contains the tangent altitudes with even numbers and the other those with the odd numbers. For each subset of measurements an independent onion peeling retrieval is performed. Finally both the resulting profiles are merged to give one profile. The goal of this method is to get rid of the onion-peeling oscillations, which is achieved by having thicker layers and thus better sensitivity – at the cost of degraded vertical resolution – in each retrieval step. The interleave method has been used, e.g., for HALOE and SABER.

As will be seen below, rigorous error propagation for onion peeling retrievals and its variants is tedious and thus rarely performed. Instead, Monte-Carlo-type sensitivity studies can be performed on the basis of simulated measurements superimposed with artificial noise, which are analyzed using the onion-peeling scheme. The error estimate is then provided by the variance of the ensemble results around the reference value at each altitude.

5.4.6 Chahine’s Relaxation Method

The Chahine relaxation method (Chahine, 1968, 1970) was originally suggested to retrieve vertical profiles of the temperature from measurements of the emerging specific intensity at several frequencies in the infrared spectral range. Later this method has been adapted by employing the geometrical decomposition to the retrieval of vertical distributions of atmospheric trace gases from the measurements of the scattered solar light in limb viewing geometry (e.g., Sioris et al., 2003, 2004).
Essentially, the measurement and state vectors have to be constructed in a way that for each of its components the following linear relationship can be considered as an acceptable approximation:

\[
\hat{x}_j = y_j F(x_a)_j,
\]

(11)

where \([\ldots]_j\) denotes the \(j\)-th component of the corresponding vector. To obtain the solution, Eq. 11 needs to be solved for each component \([x_j]\) of the state vector independently. In the original approach the number of measurements and the number of the retrieved values need to be the same. However, Sioris et al. (2004) suggested an extension of the method which solved a slightly underestimated problem with larger number of state vector components and a combination of the components of the measurement vector in the right hand side.

In the original approach of Chahine, the measurement vector \(y\) comprised measured radiances, \(F(x_a)\) the related modeled radiances, the state vector \(x\) comprised Planck functions at certain pressure levels, and spectral decomposition was applied, i.e., Eq. 11 was solved for each frequency independently. In the approach of Sioris, the measurement vector contained trace gas slant columns at each line-of-sight, the state vector contained trace gas number densities at altitude levels and Eq. 11 needed to be solved for each line-of-sight independently, i.e., the geometrical decomposition was employed.

The Chahine relaxation method is a nested iteration of the type

\[
[x_{i+1}]_j = [x_i]_j \frac{[y]_j}{[F(x_i)]_j},
\]

(12)

The inner loop runs over the altitude indices \(j\) and is usually started at the top of the atmosphere and proceeds downwards, similarly to the onion peeling method. However, in this inner loop, the information retrieved at higher levels is not directly used when Eq. 11 is solved for lower layers. Instead, the same current guess profile \(x_i\) is used to evaluate \(\hat{x}_{i+1}\) for all altitudes \(j\). Only after finishing the inner iteration over the altitudes \(j\), the state vector \(x_i\) is updated with \(x_{i+1}\). The outer iteration over \(i\) is repeated until convergence is reached.

Similarly to the onion peeling, rigorous error propagation for the Chahine relaxation method is challenging and the same approach as suggested for the onion peeling method can be used instead.

### 5.5 Nonlinearity Issues

The radiative transfer equation is nonlinear. This problem can be remedied by putting the retrieval equation used in an iterative context, e.g.,

\[
\hat{x}_{ML:i+1} = x_i + \left(K_i^T S_{y,\text{total}}^{-1} K_i\right)^{-1} \times \\
\left(K_i^T S_{y,\text{total}}^{-1} (y - F(x_i;b))\right)
\]

or

\[
\hat{x}_{reg;i+1} = x_i + \left(K_i^T S_{y,\text{total}}^{-1} K_i + R\right)^{-1} \times \\
\left(K_i^T S_{y,\text{total}}^{-1} (y - F(x_i;b)) - R(x_i - x_a)\right)
\]

(13)

(14)
for maximum likelihood or regularized problems, respectively, where $i$ is the iteration index. The last term in this equation assures that the prior information will not be ‘forgotten’ during the iteration (see, Rodgers 2000, p. 88).

To avoid seeking an $\hat{x}$ that is beyond the range of validity of the linear approximation $y(x_{i+1}) \approx F(x_i) + K(x_{i+1} - x_i)$, Levenberg (1948); Marquardt (1963) suggested a method that limits the stepwidth $x_{i+1} - x_i$ and turns it towards the direction of the steepest descent of the object function:

$$\hat{x}_{\text{ML};i+1} = x_i + \left(K_i^T S_{y,\text{total}}^{-1} K_i + \lambda I\right)^{-1} \times$$

$$\left(K_i^T S_{y,\text{total}}^{-1} (y - F(x_i; b))\right)$$

or

$$\hat{x}_{\text{reg};i+1} = x_i + \left(K_i^T S_{y,\text{total}}^{-1} K_i + R + \lambda I\right)^{-1} \times$$

$$\left(K_i^T S_{y,\text{total}}^{-1} (y - F(x_i; b)) - R(x_i - x_a)\right),$$

where $\lambda$ is a scalar that is adjusted during the iteration according to the local non-linearity of $F$ and $I$ is unity. Marks and Rodgers (1993) and Rodgers (2000) suggest the following variant:

$$\hat{x}_{\text{reg};i+1} = x_i + \left(K_i^T S_{y,\text{total}}^{-1} K_i + \lambda R\right)^{-1} \times$$

$$\left(K_i^T S_{y,\text{total}}^{-1} (y - F(x_i; b)) - R(x_i - x_a)\right).$$

Butz et al. (2012) have found that in some cases a reduced step-size Gauss-Newton algorithm works much better than the Levenberg-Marquardt algorithm (Eq. 15).

Many inverse radiative transfer problems are only “moderately non-linear” (in the sense of Rodgers, 2000) in that the retrieval equations are solved iteratively, to cope with non-linearity, but linear error estimation around the best estimate is considered adequate. If error bars are so large that they exceed the range around the best estimate where the true function $y = F(x)$ is sufficiently well approximated by the tangent $y \approx F(x_0) + K\Delta x$, then Monte Carlo or ensemble type sensitivity studies are the only remaining options. A further benefit of Monte Carlo methods, and in particular Markov Chain Monte Carlo methods, is that the posterior distributions, which can significantly deviate from the Gaussian ones, can be explored and characterized in detail, as demonstrated by Tamminen and Kyrölä (2001), Tamminen (2004), Brynjarsdottir et al. (2018), Tamminen and Kyrölä (2001), and Tamminen (2004), and Brynjarsdottir et al. (2018). Also neural network based concepts have been developed and investigated in this context (see, e.g. Pfreundschuh et al., 2018). Monte Carlo error estimates exceed the computational resources needed for the retrieval by far. Thus, they are often not apt for routine applications but their range of application remains limited to representative test cases.
6 Sources of Errors

There are multiple categories of errors and uncertainties in atmospheric state variables retrieved from satellite measurements. These are:

1. errors caused by less than perfect measurements, which include measurement noise and calibration errors,

2. errors caused by inaccuracies of the radiative transfer model used in the data analysis, which include numerical approximations, missing physical processes, or uncertainties in the values used as constants by the model, particularly spectroscopic parameters,

3. errors caused by decomposing the inverse problem, giving rise to parameter errors,

4. errors caused by the constraint applied to the retrieval, which does not allow the retrieval to produce the solution that is best compatible with the measurements.

Another factor that can cause discrepancies between two sets of measurements is that the measurements might not refer to exactly the same air mass or the same time. This, along with natural variability, often explains the differences encountered (see, e.g., Sofieva et al., 2008; Verhoelst et al., 2015; Laeng et al., 2019). In the following sections, these categories of errors and uncertainties are discussed in more detail.

6.1 Measurement Errors

In remote sensing a number of processing steps are necessary to obtain a calibrated signal in physical units from the raw data. The latter are usually referred to as the Level-0 data. Their units depend on the instrument type and the related quantities can be detector voltages, photon counts or similar. Level-1 processing transforms the Level-0 data into calibrated measurement data, which no longer depend on the particular measurement device used, such as radiance units or transmission. These are conventionally referred to as Level-1 data. If multiple processing steps are required, distinctions can be made between Level-1a, Level-1b, etc. data, but this distinction is of no relevance here. These Level-1 data come with auxiliary data describing the geolocation and time of the measurement, the measurement geometry, and so forth. The Level-1 data are the input to the retrieval of the atmospheric state. Estimates of the atmospheric state variables are referred to as the Level-2 data product. We use a convention that all uncertainties in the Level-1 data – including metadata – fall into the category “measurement uncertainties”. The main sources of measurement uncertainties include but are not limited to measurement noise, including discretization noise; zero calibration error (i.e., that the measurement signal is non-zero even though the true radiance signal is zero, which can be understood as an additive calibration error); gain calibration (this is a multiplicative calibration error); higher order errors (e.g., nonlinear detector response); uncertainties in auxiliary data, such as measurement geometry in terms of tangent altitude, the exact time of the measurement, etc.; and straylight. Further, all these errors can be subject to a drift, i.e., there can be some time-dependence.
6.1.1 Measurement Noise

Measurement noise is usually conceived as a statistical uncertainty which is described by the error variance of each single spectral data point. Usually, the uncertainties are considered as uncorrelated between the single components of the measurement vector, which implies a diagonal noise covariance matrix. In some cases, however, the measurement noise covariance matrix $S_y$ has off-diagonal elements, e.g., in Fourier transform spectrometry if apodization (see, e.g., Norton and Beer 1976) and/or zero-filling is applied.

According to generalized Gaussian error analysis, the mapping of measurement noise $\epsilon$ onto the result $\hat{x}_{\text{reg}}$ is

$$S_{x,\text{noise}} = GS_{y,\text{noise}}G^T,$$

(18)

with $G$ as defined in Eq. 7. This method is used by the MIPAS-IMK, TES, GOMOS, OMPS-LP (NASA, IUP Bremen, and Saskatchewan), OSIRIS, SBUV or SCIAMACHY-Greifswald (Lednyts’kyy et al., 2015) and SCIAMACHY-IUP data processors. Equation (18) is applicable also to maximum likelihood retrievals just by setting the $R$ term in the gain function $G_{\text{tep}}$ to zero. After excessive and cheerful cancellation this finally gives

$$S_{x,\text{noise,ML}} = \left[K^TS_{y,\text{noise}}^{-1}K\right]^{-1}.$$

(19)

Error correlations between the elements of $x$ are implicitly considered. It is important to note that such correlations will typically be present even if the measurement errors are uncorrelated and if no regularization is applied. For some retrievals, the so-called retrieval covariance matrix $S_x$ is evaluated using Eq. (10). These error estimates, however, represent not only the mapping of the measurement noise onto the retrieved quantity but also the error introduced by the application of the constraint, i.e., the ‘smoothing error’ in the terminology by Rodgers (2000). Related problems are discussed in Section 6.4. The retrieval error evaluated by this method will represent a meaningful quantity only if the a priori covariance matrix $S_a$ represents the actual variability of the atmospheric state rather than any ad hoc assumptions.

For some instruments the error estimate is based on the analysis of the residuals between the measurements and the best fitting modeled spectrum. Gauss (1821) has proven that the “residual sum of squares divided by the number of degrees of freedom is an unbiased estimator of $\sigma^2$” (translation into modern terminology by Aldrich 1998). This Gaussian $\sigma$ contains not only measurement noise but also other error components. Application of Eq. (18) to a residual-based noise characterization may be deemed more realistic than the application of this equation to pure measurement noise. However, not all uncertainties will show up in the residual. For example, spectroscopic band intensity errors of the target species will be fully compensated by erroneous retrieved concentrations and will thus create no additional spectral residual. Thus the residual-based error analysis will not provide the total uncertainty of the retrieved state variable, nor does it allow for decomposition of the error budget into its components. It is suitable to estimate the retrieval noise error $S_{x,\text{noise}}$ only if it can be assumed that the residual is dominated by the measurement noise. Residual based uncertainty estimation is used for, e.g., SCIAMACHY (U. Bremen) or ACE-FTS.

Non-optimal decomposition of the inverse problem, such as single profile retrieval, single species retrieval, etc. (Section 5.4.2), however, causes the following problem: $S_{x,\text{noise}}$ contains only the noise-induced uncertainties associated with

19
the current step of the inversion process. Propagated noise from preceding retrieval steps is formally dealt with as parameter error (see Section 6.3).

The mapping of measurement noise into the retrieval domain depends on the retrieval approach chosen. Naturally, noise has a larger effect when regularization is kept small in order to get the best possible spatial resolution, because noise and resolution are competing quantities. However, there are also other choices in the retrieval scheme which have bearing on the measurement noise as evaluated above. In the ideal case, when the retrieval vector represents the entire atmospheric state with all its relevant variables, $S_{x,\text{noise}}$ covers all uncertainties associated with everything other than the target variable. For example, if one is interested in the error of ozone abundances, any uncertainty in the ozone mixing ratio caused by water vapour uncertainties is implicitly included in $S_{x,\text{noise}}$, as suggested by Marks and Rodgers (1993), Tarantola and Valette (1982), Eriksson (2000), or von Clarmann et al. (2001); for a different perspective on this issue, see Section 4.1.2 in Rodgers (2000). The situation is different in a decomposed retrieval (Section 5.4). In the case of species-wise decomposition, the uncertainty entailed by the uncertainty of an interfering species is evaluated as parameter error. The same holds for onion peeling error propagation (Section 5.4.5). Here retrieval noise, i.e., the mapping of the measurement noise on the retrieval, accounts only for the noise of the analysis of a single tangent altitude, while the noise propagated downwards from higher altitudes is formally considered to be a parameter error. As a consequence, retrieval noise estimates from two datasets are not necessarily intercomparable. A sensible comparison is only possible between the total random errors, because the partitioning between noise and parameter errors depends on the retrieval system chosen and in particular how the inverse problem is decomposed into sub-problems.

In the context of error propagation in the the Levenberg-Marquardt algorithm (Section 5.5), it is important to distinguish two different applications.

(a) If the Levenberg-Marquardt algorithm is used only to dampen each iteration step and the iteration is only truncated after full convergence has been reached, then the $\lambda I$ term has no sizeable impact on the solution, even if $\lambda \neq 0$ at the final iteration. Thus, $\lambda I$ must not be included in the gain matrix $G$ used for error estimation.

(b) Sometimes the Levenberg-Marquardt iteration is intentionally stopped before full convergence is reached. The rationale is to use the regularizing characteristics of the $\lambda I$ term which would be lost after too many iterations. The discussion of this approach of regularization is beyond the scope of this paper, and it must suffice to mention that in this case the retrieval error has to be evaluated as suggested by Ceccherini and Ridolfi (2010).

### 6.1.2 Calibration Uncertainties

Besides measurement noise, calibration uncertainties also contribute to the measurement error (see, e.g., Kleinert et al., 2018). Often the transformation from the raw data $y_{\text{raw}}$ (such as detector voltage) to the data in physical units $y$ (such as spectral radiance) uses a linear scheme such as

$$y = y_{\text{raw}}/b - a,$$

(20)
where \(a\) is a zero level offset correction and \(b\) is a gain calibration coefficient (e.g., Revercomb et al., 1988, their Eq. 2). In the case of spectral measurements, both \(a\) and \(b\) are usually a function of frequency. Even after careful radiometric calibration, there will always be a residual zero level and gain calibration uncertainty.

Among the satellite missions considered here, the following schemes to assess the zero level calibration error are in use, or at least possible:

- Propagation of the assumed zero level calibration error in the retrieved target quantity \(\sigma_{x,\text{zero}}\) from the zero level calibration uncertainty in the measurement domain \(\sigma_{y,\text{zero}}\), using linear mapping of the type

\[
\sigma_{x,\text{zero}} = G\sigma_{y,\text{zero}}. \tag{21}
\]

- A zero level correction is jointly fitted along with the target variables. In this case, this error component does not need to be assessed separately but is automatically included in the noise-induced error, at least if no constraint is applied to the zero offset correction. Since this additional fit variable tends to destabilize the retrieval, noise-induced errors will become larger. This approach has been chosen for MIPAS-IMK, Odin/SMR, and for some of the MLS data products.

- The zero level uncertainty is added as a fully correlated component to the measurement error covariance matrix \(S_{x,\text{noise}}\) and thus needs no extra treatment. It is then accounted for by the error evaluated using Eq. 18. We are not aware of any processor using this method.

- The zero level uncertainty is deemed negligibly small and thus not evaluated. This approach has been chosen by SAGE I, SAGE II, SAGE III, SCIAMACHY, ACE-FTS, and OMPS LP.

Similar arguments hold for the gain calibration uncertainty, and in theory the same methods can be applied. In emission spectroscopy, however, gain calibration uncertainty is much harder to distinguish from concentration changes of the target species or temperature changes than offset calibration. For MIPAS-IMK the linear mapping method is used. On the contrary, for many limb-scatter retrievals a normalization with respect to a higher tangent height is done. As a result, the gain correction, \(b\), mostly cancels out (von Savigny et al., 2003, e.g.).

Occasionally, application of Eq. (20) is inadequate, e.g., if the detector response function is nonlinear (see, e.g., Kleinert et al., 2018). We are not aware of any data product where uncertainties of the coefficients of the nonlinear detector response function are routinely considered in the error budget of the Level-2 products. Arguably all calibration constants can be time-dependent and thus cause a drift. This issue is discussed in Section 6.7.

Another issue is frequency calibration. A spectral shift translates into a radiometric error that is highly correlated across the spectral line. The impact of such an error on the retrieval result is highly dependent on the retrieval setup and the selection of microwindows. A spectral shift correction can be jointly fitted with the target variables as it can be done in the framework of the the zero level correction. Residual frequency calibration errors after correction are still an issue of the Level-2 error budget. Since the radiometric error induced by a spectral calibration error is antisymmetric to the line center, its effect on the retrieval results will be different when the microwindow contains only part of the line.
For Odin/SMR and MIPAS a frequency offset is fitted as a scalar value characterizing a complete limb scan. Where necessary, for SCIAMACHY and OMPS (IUP Bremen), in addition to the Level-1 correction from NASA or ESA, respectively, a spectral shift/squeeze correction is determined during the pre-processing step by performing spectral fits for each line of sight and spectral window individually. IUP-DOAS and BIRA retrievals also use a shift/squeeze correction.

For TES, the frequency calibration is performed as part of the Level-1B processing and is not included in the error covariances supplied with the Level 2 product. OMPS LP depends on the well-characterized Fraunhofer structure in the solar spectrum to establish and maintain its spectral registration (Jaross et al., 2014), and this work is done as a part of the Level-1 processing. For SAGE II, the filter used for the water vapor retrieval changed after launch but appeared to stabilize and a static correction for the filter spectral location and bandpass is applied in the retrieval (Thomason et al., 2004). For SAGE III/ISS, spectral calibration is performed for each observation by analyzing the apparent unobstructed solar spectrum.

### 6.1.3 Instrument Characterization Errors

Under instrument characterization errors we subsume instrument line shape errors (uncertainties in the spectral response function of the instrument), uncertainties in the field of view characterization, and so forth. Which of the error sources in this category are relevant depends on the particular instrument under assessment.

The preflight characterization of the spectral response function of the instrument typically relies on a monochromatic signal. Once in space, narrow spectral lines can be used to determine possible drifts in the instrument spectral line shape.

Depending on the field-of-view width and a shape of the response function, the field-of-view characterization can be of crucial importance for limb-scatter sensors, because the limb-scatter radiance can vary by more than 5 orders of magnitude between tangent altitudes of 0 km and 100 km. In this case, small errors in the field-of-view characterization may lead to large errors in the measured limb radiances at higher tangent altitudes. Also limb scanning emission and solar occultation measurements show a sizeable sensitivity to field-of-view uncertainties.

Instrument characterization errors are, if at all, typically evaluated using linear mapping.

### 6.1.4 Auxiliary data errors

We understand auxiliary data errors to refer to quantities that come along with the measurement data but are not part of the \( \mathbf{y} \) vector. Typical examples are time registration errors, uncertainties in the measurement geometry such as tangent altitude pointing, and so forth. Due to the variable nature of the errors under this category, it is impossible to suggest a common scheme. Some of these errors can be assessed by sensitivity studies or linear mapping, using the same formalism as discussed under parameter errors. Alternatively, the uncertain auxiliary data can be jointly retrieved with the target variables. In the following, the most prominent auxiliary data uncertainties are listed, and their treatment by the instrument groups is documented.

In limb sounding, pointing errors propagate to the result for various reasons. Depending on the design of the retrieval scheme, different mechanisms may play a role. For example, the amount of air seen along the line of sight and the atmospheric state variables depend crucially on the tangent altitude. In the case of vertical gradients of atmospheric state variables, the assignment of a value which is \textit{per se} correct to an erroneous altitude causes an error. Occultation measurements using the sun
as background radiation source can depend on which part of the solar disk is seen by the instrument. The residual pointing error to be considered in the error estimation depends on the pointing correction schemes applied. For MIPAS-IMK limb emission measurements, the first step of the retrieval chain is the simultaneous retrieval of temperature and tangent altitudes (von Clarmann et al., 2003). Results are used as known parameters in subsequent retrieval steps where trace gas abundances are retrieved. Residual errors of temperature and tangent altitudes are treated as parameter error in subsequent steps. For OMPS-LP measurements, the pointing correction is derived from radiance measurements using the absolute radiance residual method (ARRM) and the Rayleigh scattering attitude sensor (RSAS) methods (Scott et al., 1996; Moy et al., 2017). Also OSIRIS uses the RSAS method (Bourassa et al., 2018).

For SCIAMACHY a correction to the pointing information is derived by analyzing measurements in the occultation geometry (Bramstedt et al., 2017) and is implemented to Level 0 to 1 data processing. The effect of residual pointing errors is assessed via Monte-Carlo-type studies for representative profiles (Rahpoe et al., 2013). Earlier SCIAMACHY analysis relied on a pointing retrieval using limb radiances below 300 nm. This so-called “knee-method” uses the known altitude of the maximum of the limb radiances originating from Rayleigh-scattering (Kaiser et al., 2004; von Savigny et al., 2005). For ODIN/SMR a scalar pointing correction is fitted for the entire limb scan.

6.2 Model Errors

The radiative transfer model used in the retrieval solves the radiative transfer equation and makes the signal comparable to what the instrument would see by integration over the finite field of view and by convolution with the spectral instrument response function. A lot can go wrong here as our knowledge on radiative transfer can be erroneous or inaccurate. Some known physics, such as non-local thermodynamic equilibrium, line coupling, or more sophisticated than usual line-shape functions, may be disregarded for reasons of computational efficiency. Time constraints can also lead to numerical integration being performed with limited precision or weak spectral transitions being ignored. The goal in formulating the radiative transfer model is to keep model errors from known sources much smaller than the measurement error while maintaining computational efficiency. Naturally, any unknown sources of model error are hardest to quantify. In the following, the most relevant types of known model errors are discussed.

6.2.1 Incomplete Models

Some relevant physical processes included in \( f \) may be left unaccounted for by the radiative transfer model \( F \) in use (Section 4). Typical examples are non-local thermodynamic equilibrium (Non-LTE) emission, line coupling, or line shape issues. Non-LTE emissions occur when air density is so low that the excited molecule after absorption of a photon or in its nascent state will re-emit radiation before quenching redistributes the energy towards a Boltzmann distribution (e.g., López-Puertas and Taylor, 2001). Line mixing is a high pressure phenomenon where collisions transfer angular momentum, entailing energy transfer between energetically adjacent transitions (e.g., Armstrong, 1982; Bulanin et al., 1984; Strow and Gentry, 1986; Hartmann and Boulet, 1991; Hartmann et al., 2009; Thompson et al., 2012; Alvarado et al., 2013). The usual line-shape models, such as the Voigt lineshape (Voigt, 1912), may not adequately represent the true line shape (e.g., Galatry, 1961; Berman,
Critical issues in ultraviolet or visible remote sensing are scattering and polarization. Different levels of sophistication of models refer to the treatment of sphericity of the atmosphere and orders of scattering accounted for.

If a complete model is available but not used for the operational retrieval for reasons of computational efficiency, the effect of the missing processes can be assessed via sensitivity analyses based on the complete model and considered in the error budget. If the error is of a systematic nature, the related bias can even be corrected for, and only the residual scatter begs consideration in the error analysis.

In stellar occultation, the forward model for retrievals of traces gases from UV-VIS measurements does not include the deterministic description of stellar spectra perturbations due to scintillations. This omission is not only due to complicated description of wave propagation in random media, but also by a stochastic nature of small-scale air density irregularities generated by small-vertical-scale gravity waves and turbulence. These perturbations can be, however, characterized and added as an additional, correlated in wavelength component to the measurement noise, as shown in Sofieva et al. (2009).

If no complete model is available, then it can only be hoped that the related error is sufficiently small compared to the other error sources that it has no bearing on the total error budget.

6.2.2 Numerical Issues

The numerical solution of the radiative transfer equation requires a lot of integration, e.g., to integrate the spectral radiances over the field of view based on a finite number of so-called pencil beams; the spectral grid on which the radiative transfer is calculated has a finite width; radiative transfer through the atmosphere is in most models based on a finite number of layers or levels, just to name a few. Any improvement of computational accuracy goes along with increased computational effort. For most satellite data processors, the setting is chosen in a way that these issues produce a retrieval error which is so small compared to the leading error sources that it can be ignored in the error budget.

6.2.3 Model Constants

The main constants of relevance here include spectroscopic data, quenching rates, refractive indices, etc. The values of other constants (radius of Earth, gas constant, molecular weights, etc.) are known at an accuracy which renders analysis of related retrieval errors unnecessary. Estimation of the impact of spectroscopic errors poses some serious problems.

A major problem in the propagation of spectroscopic data errors is that, in some cases, no uncertainties of cross-sections are available. Also, when they are available, information on error correlations is not provided. If a retrieval uses, say, a large number of ozone lines, it would be of utmost importance to know whether errors in the intensity of these lines are correlated (e.g., because the uncertainties are attributed to uncertainties in the gas amount in the cell used in the lab where the spectroscopic parameters were measured) or uncorrelated (because errors are dominated by noise in the lab measurement or because the
spectroscopic information stems from different lab measurements). In the uncorrelated case the errors would randomize while in the correlated case they would fully survive the error propagation for a retrieval using multiple spectral lines.

To exemplify another issue, consider a gas-wise sequential retrieval where H$_2$O is retrieved first, and this H$_2$O profile is then used as a known parameter in a retrieval of ozone in another spectral region. It is possible for the spectroscopic errors of H$_2$O to cancel out in the ozone retrieval if these errors are consistent over the entire spectrum. For example, if H$_2$O line intensities are too high, too little H$_2$O will be retrieved. Subsequently, during the ozone retrieval, the combination of the too little H$_2$O with the too large line intensities produce the correct impact of H$_2$O on the modeled spectra. This results in the H$_2$O line intensity errors not propagating into the retrieved ozone concentrations.

The usual way to estimate the propagation of spectroscopic data errors is to conduct sensitivity studies with perturbed spectroscopic data. Since, as stated above, the correlations between spectroscopic data errors are unknown and not reported in commonly used spectroscopic databases, these sensitivity studies render only a crude estimate of the related retrieval error.

The OCO-2 team is currently working on an effort to characterize retrieval uncertainties associated with the sets of laboratory measurements used in the multi-spectrum fitting analysis that was used to derive the spectroscopic parameters used in the OCO-2 forward model (unpublished work).

In the case of retrievals of trace gas abundances, one might argue that uncertainties of the line intensity can be mapped directly onto the target concentration retrieval. Because both the line intensity and abundance appear reciprocally in the exponent of Beer’s law, the non-linearity of the radiative transfer equation has no bearing on the line intensity error propagation. It has, however, been shown that it is not sufficient to restrict related error analysis to the line intensities. For example, pressure broadening has a sizeable effect (e.g. Urban et al., 2005; Glatthor et al., 2018) in the infrared and microwave regions. Connor et al. (2016) had found, in their linear error analysis for OCO-2 retrievals, that a constant perturbation even to CO$_2$ line intensities did not in fact map to a constant impact on the XCO$_2$ retrievals within the NASA OCO-2 algorithm. Within that algorithm, the impact of line intensity perturbation on the retrieved XCO$_2$ varies spatially and appears to depend on the surface brightness. Inclusion of surface albedo terms in the state vector for the OCO-2 algorithm gives rise to this information cross-talk.

The propagation of uncertainties of model constants follows the same formalisms as proposed for uncertainties in atmospheric parameters (Section 6.3).

In the NASA ACOS/OCO-2 and OCO-3 CO$_2$ retrieval algorithm spectral residuals caused by imperfect spectroscopy, solar model and instrument characterization are dealt with by fitting scaling factors to fixed spectral residual patterns. These patterns are the Empirical Orthogonal Functions (EOFs) that result from a singular value decomposition of spectral residuals from training retrievals (O’Dell et al., 2018). A similar approach was adopted independently by Lange and Landgraf (2018) for retrievals of methane from GOSAT thermal infrared spectra.
6.3 Parameter errors

We define parameter errors as those errors caused by the decomposition of the full retrieval problem such that later retrieval stages assume that a part of the atmospheric state is already known and thus not included in the retrieval vector \( \hat{x} \) (see, Section 5.4). The assumed values can derive from either climatologies or results of a preceding retrieval step or from climatologies or any other source of prior information. The ideal sequence of operations has the first atmospheric state variables retrieved being those whose signal is only weakly dependent on or interfered by other state variables. Once known, these values can be used for subsequent retrieval steps as “fixed” parameters. Error propagation has to be considered.

The impact \( \Delta x \) of errors in parameters can be estimated via sensitivity studies, where a measurement is simulated with parameter \( b \) that is perturbed by a certain amount \( \Delta b \), e.g., one standard deviation of its uncertainty.

\[
\Delta x = G (F(\hat{x}; b + \Delta b) - F(\hat{x}; b))
\]  

(22)

This scheme is used, e.g., for MIPAS IMK, SCIAMACHY-Greifswald, SCIAMACHY IUP (see Bremen, Rahpoe et al. 2013 or E. Malinina et al. 2018 and SMILES-NICT (See, e.g., Sato et al., 2012, 2014)). If the parameter \( b \) is a vector, whose elements’ error correlations are known and relevant, generalized Gaussian error propagation can be applied

\[
S_{x,b} = G K_b S_b K^T_b G^T,
\]  

(23)

where \( K_b \) is the Jacobian matrix representing the sensitivities \( \frac{\partial y_m}{\partial b_j} \) of the measurements with respect to a changing parameter \( b_j \).

Depending on the source of the information on the parameter vector — climatology, preceding retrieval step, independent measurements, or whatsoever, the parameter errors can be correlated or uncorrelated in space and time.

Occasionally errors are of a mixed nature, e.g., if a quantity is jointly retrieved along with the target quantity but strongly constrained. In this case, the parameter actually is part of the retrieval vector \( \hat{x} \) but its value still depends largely on the a priori information. Uncertainties that derive for this situation are discussed in Section 6.4.

6.3.1 Error propagation in onion peeling

In onion peeling (Section 5.4.5) the ray path with the highest tangent altitude is analyzed first. In the second step, the results of the first step are used as known parameters. Thus the retrieval error of the first step has to be considered as a source of parameter error in the second step, and so forth. Explicit error propagation through an onion peeling retrieval has been studied, e.g., by Noël et al. (2016).

Instead, Alternatively, the onion peeling retrieval error can be estimated using a Monte Carlo method. For the solution profile \( x \) a limb sequence of measurements is calculated. Artificial noise with the same characteristics of as the real measurement
noise is superimposed upon the measurements. A sample of limb sequences is generated, based on the same forward radiative transfer calculations but different in the actual realization of the random noise. For each of these simulated limb sequences a retrieval is performed and, from the scatter of these results, the retrieval error covariance matrix $S_{x,\text{noise}}$ is calculated.

### 6.4 A Priori Information

In order to avoid wording that is too abstract, we assume that the retrieval vector represents vertical profiles of atmospheric state variables. However, with some adjustments the mathematical concept is applicable to 2-D or 3-D fields of atmospheric state variables as well. The framework is also applicable to column retrievals. In this case, the retrieval vector has only one element.

By performing regularized retrievals invoking Eq. (4) or variants of it, the retrieved atmospheric state will deviate from the one which is most consistent with the pure measurement information. As a consequence, this can introduce additional bias and distortion, and the resolution can be degraded with respect to the true state of the atmosphere beyond the degradation caused by the finite retrieval grid. The resulting profile is a mixture of the measurement information and the a priori information used. For the interpretation of constrained retrievals it is of utmost importance to have tools available to diagnose the content of a priori information in the retrievals. As in Rodgers (1976, 1990, 2000), one can calculate the derivative of the retrieved state with respect to the true state, and call the resulting matrix the “averaging kernel matrix”

$$A = \frac{\partial \hat{x}_i}{\partial x_j} = GK = (K^T S_{y,\text{total}}^{-1} K + R)^{-1} K^T S_{y,\text{total}}^{-1} K. \tag{24}$$

The rows of the averaging kernel represent the weighting functions, which determine to what degree the result at altitude level $i$ depends on the true atmospheric state at altitude level $j$. Its columns represent the response of the retrieval to a delta perturbation at a single altitude level. If a joint retrieval of profiles of multiple different quantities is made, the above refers to the diagonal blocks of the averaging kernel matrix which refer to the quantity under consideration. The presence of non-negligible off-diagonal blocks indicates a significant interference between the species introduced by the regularization scheme.

The averaging kernel of a fully converged Levenberg-Marquardt retrieval equals that of the respective retrieval without the $\lambda I$ term because after convergence this term has no impact on the solution. If the iteration is ended prematurely in order to use the Levenberg-Marquardt method to regularize ill-posed inverse problems, then the averaging kernel has to be calculated as suggested by Ceccherini and Ridolfi (2010).

When $S_{y,\text{total}}$ is approximated by $S_{y,\text{noise}}$ in the retrieval, the same approximation must be used for the evaluation of the averaging kernel. That is to say, in this case the averaging kernel should be calculated involving $S_{y,\text{noise}}$ instead of $S_{y,\text{total}}$.

Conversely, the derivative of the retrieved state with respect to the a priori information is $I - A$. With this, the retrieval can be rewritten as

$$\hat{x} = Ax + (I - A)x_a. \tag{25}$$
For the retrieval of column amounts, the sensitivity of the column to the true state values at different altitudes can be represented by the column averaging kernel $A_{\text{col}}$, which is, as opposed to the averaging kernel described above, not a square matrix but a row vector (Wunch et al., 2010, see, e.g.)

$$A_{\text{col}} = h^T A,$$  \hspace{1cm} \text{(26)}

where $h^T$ is the column operator whose multiplication with the vertical profile yields the vertical column density and where $A$ is the regular profile averaging kernel as described above.

Usually regularization will entail that the retrieved state $\hat{x}$ is a smoothed and possibly biased representation of the true state $x$. Rodgers (2000, p. 48) offers two possible interpretations of the retrieved state $\hat{x}$. It can either be conceived as a smoothed estimate of the true state, or it can be construed as an estimate of the smoothed true state. The choice of the interpretation has major impacts on the error budget, which are discussed below. All this is not to say that the effect of the prior information is restricted to smoothing. The resulting profile shape can be distorted in a sense that the extrema of a profile can be shifted upward or downward (see, e.g., the HOCl profiles in Jackman et al., 2008, their Fig. 12) or bias the result (see, e.g., Bhartia et al., 2013). The averaging kernel matrix contains information on the dependence of the result on the true state and the a priori assumption, the vertical resolution, and the information displacement.

### 6.4.1 The Retrieved State as a Smoothed Estimate of the Truth

As stated above, a retrieval can be understood as a smoothed estimate of the truth or an estimate of the smoothed truth. In the first case, any deviation between the estimate and the truth which is caused by the regularization of the retrieval has to be included in the error budget. Rodgers (2000) calls this error component ‘smoothing error’ and has suggested the following formalism to estimate it (Rodgers, 1990):

$$S_{x, \text{smoothing}} = (I - A) S_a (I - A)^T.$$  \hspace{1cm} \text{(27)}

While in principle this formulation complies with generalized Gaussian error propagation, this concept has been criticized (von Clarmann, 2014). The main criticism refers to the fact that this estimate does not refer to the difference between the retrieved and the true state but only to the difference between the estimate and the true state as represented on the grid on which $S_a$ has been evaluated. This leads to the undesirable effect that a smoothing error evaluated on a coarse grid will be smaller than a smoothing error evaluated on a fine grid. Further, a smoothing error evaluated on a coarse grid and then propagated onto a fine grid, will be smaller than the smoothing error evaluated directly on the fine grid, although the interpolation between the grids is a linear operation, which is another undesirable outcome.

Another criticism that may be applied to the concept of the smoothing error is that it forces one to meander around between a subjective (personalist) and an objective concept of probability. In the retrieval (Eq. 6), $S_a$ represents parameters of a personalist’s probability distribution. That is to say, the underlying concept of probability is a subjective one, describing the agent’s (lack of) knowledge or information about a value which is determined in the true world. Conversely, $S_a$ in Eq. (27) is required to represent parameters of an objective probability distribution, equivalent to a frequency distribution.
Since interpolation of profiles to other grids is a standard operation, it is not advisable to include the smoothing error in the error budget without a caveat. Instead, the averaging kernels should be communicated to the user, allowing them to evaluate the smoothing error on the final working grid.

In this context it should be mentioned that error estimates according to Eq. (10) include a smoothing error component and should not be used to calculate the error budget because the data user might not be aware of related problems and might, when interpolating profiles on a finer grid, propagate these error estimates to the finer grid.

Further, Rodgers (2000) points out that Eq. (27) will only yield a meaningful smoothing error if $S_a$ is not just a constraint matrix chosen ad hoc to regularize the inversion but a real statistical description of the variability of the actual states around the mean state used as $x_a$. This criterion should even be more rigorous: The maxim of the most specific reference class has to be applied (Hempel, 1965). For example, to calculate the smoothing error of a midlatitude ozone profile retrieval we cannot use a global ozone climatology. To calculate the smoothing error of a polar winter ozone retrieval, we must not use an ozone climatology built from a whole year of polar ozone data. The most specific reference class will be a homogeneous reference class whose internal variability is, as far as known, purely random.

Not all applications of a retrieval scheme of the type Eq. (6) use a climatological mean profile as a priori. For example, for the upcoming TEMPO mission, actual ozone measurements have been tested to be used as a priori (see, e.g., Johnson et al., 2018). In such applications, the $S_a$ matrix contains the estimated uncertainties of the individual ozone measurements used instead of the climatological variability. The standard approach of maximum a posteriori retrievals with climatological prior is based on the assumption that a climatology based on data collected in the past will also be true for the actual case. Hume (1748) was the first to show that this assumption cannot conclusively be inferred from anything. The use of actual measurement data from independent sources as prior information dispenses with this assumption and reduces the maximum a posteriori retrieval to a sort of optimal average of two independent measurements. The smoothing error evaluated for this kind of retrieval scheme represents the propagated uncertainties of the measurement(s) used as prior information.

A particular problem is the evaluation of the smoothing error difference (occasionally, perhaps more adequately, called ‘smoothing difference error’) of a pair of measurements. For this purpose, Rodgers and Connor (2003) suggest that the retrieved profiles should first be transformed to the same a priori profile $x_c$, using the general transformation scheme Eq. (10.48) of Rodgers (2000)

$$
\hat{x}_{\text{new}} = \left( \hat{S}_x^{-1} - \hat{S}_{a,\text{old}}^{-1} + \hat{S}_{a,\text{new}}^{-1} \right)^{-1}
\left[ \hat{S}_x^{-1}\hat{x}_{\text{old}} - \hat{S}_{a,\text{old}}^{-1}x_{a,\text{old}} + \hat{S}_{a,\text{new}}^{-1}x_{a,\text{new}} \right],
$$

where profile and covariance matrix $x_{a,\text{old}}$ and $S_{a,\text{old}}$ represent the initially used prior information to be removed, and where $x_{a,\text{new}}$ and $S_{a,\text{new}}$ represent the new prior information be included instead. In the given application, the old prior information is that used for the retrieval, and the new one, $x_{a,\text{new}}$, is the prior information $x_c$, valid for the comparison profile. This transformation is possible within linear theory and adequate if and only if one result is in the linear domain of the other. If the profiles are provided on different grids, a transformation of the profiles, covariance matrices, and averaging kernels to a common grid must precede the above transformation (see, e.g., Stiller et al. 2012 or Eckert et al. 2014 for sample applications,
or Keppens et al. 2019 for a summary of methods.) Then the smoothing error difference is evaluated, where $A_1$ and $A_2$ are the averaging kernels of the retrieved profiles $x_1$ and $x_2$, all after application of the transformations outlined before.

$$S_{x_1 - x_2; \text{smoothing}} = (A_1 - A_2)S_c(A_1 - A_2)^T$$ (29)

Here $S_c$ is the a priori covariance matrix describing the variability of the atmospheric state around a priori profile $x_c$ valid for the comparison profile. Rodgers and Connor (2003), however, do not specify what type of $x_c$ profile is adequate. The common a priori profile $x_c$ can by no means be freely chosen. Here the maxim of the most specific reference class Hempel (1965) becomes important again. When large samples of collocated measurements are compared, the appropriate reference class for instrument 1 is not necessarily the appropriate reference class for instrument 2, and vice versa, because both instruments might typically sample different parts of the atmosphere. The adequate sample with which to build the statistics needed to evaluate the smoothing error of the difference is that which is representative for the actual collocations of both measurements.

The criticism of the smoothing error as formulated above (After Eq. 27) does not apply to the smoothing error difference.

6.4.2 The Retrieved State as an Estimate of the Smoothed Truth

With the interpretation of the retrieved state as an estimate of the smoothed truth, we accept that measurements can only provide a finite-resolution representation of the truth and do not consider this as an error component of the measurement (not to mention the philosophical problems associated with what an infinitely resolved atmospheric state shall be; see von Clarmann 2014 for a critical discussion). The only important thing to consider is to avoid comparison of apples and oranges: Differences of atmospheric state variables are only meaningful if the data contain the same amount of the same a priori information and have the same vertical resolution. This is not typically given when two measurements are compared, and a part of the observed differences is thus due to related artefacts.

If the contrast in resolution is large enough to consider the better resolved measurement as both practically ideal compared to the other one and practically free of a priori information, then it is common practice to apply the averaging kernel of the coarser resolved measurement to the better resolved measurement (see Section 6.4.3 for concepts of altitude resolution).

$$\hat{x}_{1,\text{smoothed}} = A_2\hat{x}_{1,\text{original}} + (I - A_2)x_{a,2}$$ (30)

Here index 1 refers to the better resolved measurement and index 2 to the coarser resolved one. The other indices are self-explanatory. To our best knowledge, this approach was first suggested by Connor et al. (1994). This approach is also commonly applied when measurements are compared to model data. In this case the averaging kernel of the measurement is applied to the modeled atmospheric state. In data assimilation the averaging kernel has to be included in the observation operator. Within linear theory and the assumption in force that the better resolved data set contains no sizeable amount of a priori information, the Connor et al. approach indeed solves the problem that the original datasets are not directly comparable due to different vertical resolutions. The problem of interpolability of averaging kernels is discussed in Arosio et al. (2018).

Problems occur when linear theory is no longer adequate to describe the problem. For example, B. Funke has, during the preparation of Funke et al. (2017), encountered the following difficulty: When the MIPAS averaging kernels were applied to
modeled Nitric Oxide (NO) distributions, the discrepancies between the modeled and the measured NO distributions were found to be larger than in the comparison without application of MIPAS averaging kernels. The following reason has been identified. The measured and the modeled NO distributions were so different that the application of the MIPAS averaging kernel to the modeled NO distributions was no longer justified. That is to say, the Connor et al. method is only valid if the data sets to be compared are similar enough to justify the assumption of linear theory. What would have been needed were MIPAS averaging kernels calculated for the modeled NO distribution. The latter approach, i.e., to calculate dedicated averaging kernels for model atmospheres, has been chosen, e.g., by Schneider et al. (2017) and references therein.

In this context another caveat is in order: Averaging kernels usually depend on the units in which the atmospheric state is expressed. For example, averaging kernels evaluated for volume mixing ratios must not be applied to number density profiles. Some authors prefer to use so-called ‘fractional averaging kernels’ instead, which refer to the relative instead of the absolute change of the state variable and are thus unit-independent (Keppens et al., 2015). However, again the caveat applies that these can be calculated only within linear theory, with the assumption in force that the retrieved profile is sufficiently close to the true profile.

6.4.3 Altitude Resolution

Often the full information contained in the averaging kernel is summarized in simpler terms. The most important simple diagnostics that partially describe the content of the averaging kernel are vertical resolution, information displacement, and measurement response. We first discuss the concept of vertical resolution.

Vertical resolution of the retrieval, not to be confused with the vertical resolution of the instrument itself, describes the ability to distinguish separate features in a vertical profile. It is limited by both the vertical retrieval grid on which the results are presented and by correlations of the data at adjacent grid points caused by the regularization term of the retrieval. Contrary to common belief, a wide field of view or an observation geometry other than limb or with coarse vertical sampling do not per se degrade the vertical resolution of the measurements. The altitude resolution of the retrieval is determined only by the vertical grid and the regularization. It goes without saying, however, that a wide field of view or any sub-optimal observation geometry often forces the retrieval scientist to use a stronger regularization to get useful results, which, in turn, will degrade the altitude resolution. Thus, the field of view geometry or sampling have an indirect influence on the vertical resolution of the retrieval, which is fully accounted for by the averaging kernel matrix and does not need extra treatment. Vertical oversampling in limb sounding, i.e. the use of a tangent altitude spacing finer than the width of the instantaneous field of view of the instrument still allows a useful vertical resolution finer than the field of view (Roscoe and Hill, 2002). A measurement mode of this type has been employed, e.g., for MIPAS for the measurements recorded after 2004 (Fischer et al., 2008), and is standard for sub-millimeter and microwave measurements such as MLS (Barath et al., 1993; Waters et al., 2006) or ODIN/SMR (Urban et al., 2005).

Rodgers (2000) reports four measures of the vertical resolution, all based on the averaging kernel matrix. Two of these measures are commonly used. The first is the full width at half maximum of the respective row of the averaging kernel matrix, or, in the case of a retrieval of multiple quantities, the part of the row associated with diagonal block associated with the
quantity of interest. The second is the reciprocal data density, which is the local grid width divided by the respective diagonal value of the averaging kernel matrix \( (\text{Purser and Huang, 1993}) \). In less than well-behaved retrievals or at the extreme ends of the profiles, where the maximum of the averaging kernel does not coincide with its nominal altitude, the latter provides better intelligible results.

The Backus and Gilbert (1970) spread (shown here in a generalized variant introduced by Rodgers 2000)

\[
s(z) = 12 \int (z - z')^2 A^2(z, z')dz' / (\int A(z, z')dz')^2,
\]

where \( z \) is altitude and \( A \) the respective element of the averaging kernel matrix \( A \), was found by Keppens et al. (2015) to be most informative under certain circumstances. Obviously in the case of the retrieval of different state variables the summation here and all similar applications should only be performed inside the diagonal sub-blocks, corresponding to each retrieval quantity. Among other reasons discussed towards the end of the section, going outside the sub-blocks could mean that even different units would be mixed.

A drawback of the Backus-Gilbert spread is that it depends largely on the grid on which the retrieval is performed. The averaging kernel of a retrieval performed on a finer vertical grid will have more pronounced side-lobes which are simply not resolved by an averaging kernel evaluated on a coarser grid. The Backus-Gilbert spread is very sensitive to such side-lobes and will thus inadequately ‘punish’ the fine-grid retrieval by giving large weight to these side-lobes and thus assigning a large ‘spread’ to them. It thus does not seem suitable for a largely grid-independent measure of the vertical resolution.

Obviously, the altitude resolution can be altitude dependent. Usually, the averaging kernel matrix is evaluated on the grid on which the retrieval is performed, because the Jacobians needed are often a by-product of the retrieval. The disadvantage of this approach, however, is that the averaging kernel does not represent any subgrid smoothing effects. Averaging kernels evaluated on a finer grid, which, by the way, are no longer square, can in principle be provided if the related Jacobians are made available, but this is hardly ever done. The ideal averaging kernel is the identity matrix. This averaging kernel matrix corresponds to a maximum likelihood retrieval, where the weight of prior information is zero. Here the altitude resolution is equal to the gridwidth of the retrieval. In agreement with our intuition, the altitude resolution cannot be better than the width of the grid on which the retrieval is performed.

It is a common misconception that the averaging kernel characterizes the vertical resolution of the estimated profiles \( \hat{x} \). The retrieval as represented by Eqs (3), (4), (6), or (25) is a correction of an initially guessed or a priori assumed profile. The altitude resolution obtained from the averaging kernel characterizes only the correction term but not the a priori component. If the a priori profile is highly structured and thus resolves fine scales, these structures are propagated onto the result \( \hat{x} \).

As is often the case, precision and resolution share a trade space in remote sounding retrievals. We see from Eq. (18) (with Eq. 7 inserted) that weaker \textbf{regularisation} \textbf{regularization} will increase the impact of measurement noise. Conversely, weaker \textbf{regularisation} \textbf{regularization} will, according to Eq. (24), push the averaging kernel towards the identity matrix, which is associated with the optimally obtainable resolution of a profile at a given discretization.

In the context of altitude resolution, a cautionary note is in order. The altitude resolution is neither identical to the grid width nor with the information smearing. In a regularized retrieval the vertical resolution is coarser than the retrieval grid. Only
in an unconstrained maximum likelihood retrieval is the vertical resolution equal with the gridwidth. Conversely, the vertical resolution of measurements that are sensitive to a very small air parcel is only limited by the vertical grid, and the sampling theorem (Shannon, 1948) applies. That is to say, in situ measurements from a cruising aircraft have no altitude resolution in the sense as defined here although the measurements may be practically point measurements due to the small vertical extent of the air parcel probed. In remote sensing the radiative transfer equation, which is integrated over all altitudes relevant to the retrieved profile, acts as an anti-aliasing filter, and the sampling theorem is of no concern.

Another concept closely related to the concept of altitude resolution is that of the degrees of freedom of the retrieval. This number is calculated as the trace of the averaging kernel matrix (Rodgers, 2000)\(^7\).

### 6.4.4 Information Displacement

Ideally the maximum, the mean, and the median of the averaging kernel coincide with the nominal altitude but “it ain’t necessarily so” (George and Ira Gershwin, 1935). Any displacement reflects the fact that the interpretation of the retrieved profile without consideration of the averaging kernel is, mildly speaking, misleading. This problem can often be remedied by comparing the retrieved profile not with any reference profile, but with a reference profile to which the averaging kernel matrix of the remotely sensed profile has been applied according to Eq. (30). Again, the caveat that this method is valid only within linear theory applies. A measure of the information displacement is the centroid offset of the averaging kernel (see, e.g., Keppens et al. 2015 and references therein).

An example of the importance of this issue is found in Jackman et al. (2008) who compared modeled HOCl distributions to those measured by MIPAS (von Clarmann et al., 2006). Maximum concentrations are displaced by more than 5 km before consideration of the averaging kernels.

### 6.4.5 Measurement Response Regularization bias and measurement response

As mentioned above, regularization may not only smooth the retrieval but may bias or distort it. The related component of the bias, \(\langle \hat{x} - x \rangle_{\text{regul}}\),

\[
\langle \hat{x} - x \rangle_{\text{regul}} = \langle (I - A) (x_0 - x) \rangle
\]

We have to distinguish two cases: Firstly, smoothing can cause biases, because, e.g., a sharp maximum of the true profile will always be reproduced too low, and the wings of the maximum will be reproduced too high. This type of bias, however, is only relevant if the retrieval is conceived as a smoothed representation of the truth as discussed in Section 6.4.1. If the retrieval is conceived as an estimate of the smoothed truth as discussed in Section 6.4.2, this type of bias is of no concern.

Secondly, regularization can cause a bias by pushing the result systematically towards higher or lower values. Any such effect besides mere smoothing is characterized by the measurement response function \(q\) (occasionally also called ‘vertical sensitivity’), a concept which goes back to an idea of Eriksson (2000) and Baron et al. (2002). It is defined as the sum over the

\(^7\)Rodgers, 2000 uses the term ‘degrees of freedom of the signal’ but this term is defined for Bayesian maximum a posteriori retrievals only. To avoid clashing with Rodgers’ terminology, we use the term ‘degrees of freedom of the retrieval’, which is applicable in a wider context.
row of the averaging kernel matrix.

\[ q_i = \sum_{j=1}^{n} a_{i,j} \]  

In the case of a multi-profile-multi-species profile retrieval, the sum is calculated over the subblock or the averaging kernel matrix referring to the profile under assessment. If the regularization of a retrieval provides a smoothed version of the truth, without systematically pushing results towards greater or smaller values, the sum of the elements over each row of the averaging kernel should be unity. Any deviation of the row-sums from unity thus hints at an influence of the constraint that is beyond pure smoothing. The measurement response function is retrieval-unit-dependent.

Even if the averaging kernel matrix is far from unity, a measurement response function close to unity indicates that the retrieval is, putting measurement errors aside for a moment, a smoothed but unbiased representation of the true profile. Conversely, small values of the measurement response function deviating by an appreciable amount from unity indicate a large influence of the prior information not only on the profile shape but also on the integrated values. Interpretation of the measurement response, however, requires some caution. Any non-zero \( (x_{a} - x) \) will cause a bias in this case.

The row sum of the averaging kernel, which makes up the measurement response, consists of summands which refer to a perturbation by the same amount in each layer, where, again, the ‘sameness’ is unit-dependent. Such a perturbation can be fully realistic in one layer and fully unrealistic in other layers, depending on the retrieval-units. The evaluation of the measurement response is particularly problematic in cases where the profile values cover a wide dynamic range. This is the case, e.g., for the H\(_2\)O mixing ratios or when the retrieval units are number density. For example, a certain perturbation in terms of number density of a certain trace gas at lower altitudes can correspond to merely moderate changes in the mixing ratio, while the same perturbation in terms of number density at high altitudes where air density is low will correspond to fully unrealistic mixing ratios. Thus, averaging kernels evaluated in units of number density can fake a large dependence of the result on values at higher altitudes which does not exist in the real world. These large contributions from higher layers can lead to unrealistic large values of the measurement response at lower altitudes.

### 6.4.6 Regularization Crosstalk

The discussion of the averaging kernel matrix and smoothing error was focused on the retrieval of single quantities so far, e.g., vertical profiles of a single state variable. Often, however, multiple different state variables are jointly retrieved in one leap. In this case the regularization constraining one state variable can affect the result of the other and vice versa. More specifically, the smoothing error of one variable can propagate onto the result of the other variable and thus give rise to regularization crosstalk. If the full (multi-variable) averaging kernel matrices are stored, the resulting parameter errors can be evaluated using Eq. 27. The case under discussion lies between the extremes of treating the other variable as a known parameter during the retrieval and the unconstrained joint-fit of both quantities.

### 6.4.7 Related Issues

In the context of averaging kernels and vertical resolution a few further remarks are in order.
Time series of state values at a given altitude are particularly problematic when the averaging kernel is time-dependent in itself. Here it may help to remove the prior information from the data along with resampling in order to achieve $A = I$ as suggested by von Clarmann et al. (2015).

While averaging kernels of maximum likelihood retrievals are unity on the native grid on which the retrieval has been performed, any interpolation to finer grids will entail non-unity averaging kernels.

Averaging optimal estimates will not usually create optimal averages. This is particularly true when the prior information is the same for each retrieval, e.g., a climatological data set. This is because the weight of the prior information will be too large in the average (see, e.g., Ceccherini et al., 2014).

Even if optimal estimation is conceived in an objectivist sense, where the prior information can be conceived as the frequency distribution of true states, any deviation of the assumed frequency distribution from the true one is an additional error source which is not typically considered in estimated error budgets.

### 6.5 Unknown Error Components

Error estimation will never be perfect, not only because the input variables of error estimation are uncertain in themselves, but also because there always are error sources that those responsible for the error estimation may not be aware of. Povey and Grainger (2015) propose “to present multiple self-consistent realisations of a data set as a means of depicting unquantified uncertainties.” It is obvious that such ensemble techniques are well suited to investigate the non-linear interaction of multiple known error sources, to obtain sensitivity information if the data processors contributing to the ensemble consider different types of known uncertainties, or to identify the spread of results which may result from different numerical implementations. These authors, however, fall short of telling us how such ensemble techniques should provide information on the effect of unknown error sources. The problem is that none of the data processors contributing to the ensemble has the unknown mechanism implemented, and the unknown uncertainty will cause an unknown bias of the ensemble mean rather than scatter of the ensemble.

The only way known to us to gain confidence that all relevant error sources have been considered is to compare multiple independent measurements based on different measurement systems where we can fairly safely exclude that they all are affected by the same type of systematic effect. If the discrepancies between the results of different instruments can be explained by the combined error budgets, we have reason to believe that the error budgets of the instruments under comparison are fairly complete (e.g., Rodgers and Connor, 2003; von Clarmann, 2006). For at least three independent measurements the random components of the error can be pinpointed quite safely (Laeng and von Clarmann, 2019, and references therein) (Loew et al., 2017, and references therein), and for a large number of independent instruments one can assume that even the bias of the mean will approach zero.
6.6 Natural Variability

It goes without saying that natural variability in a sense that the atmospheric state at place \( s_1 \) and time \( t_1 \) differs from the one at \( s_2 \) and \( t_2 \) is not a genuine retrieval error. However, when in a validation context two independent measurements of the same state variable are compared and the measurements do not refer to exactly the same airmass, the spatial or temporal mismatch of the measurements along with natural variability will contribute to the error of the difference. Often, natural variability is invoked as a universal excuse if validation studies hint at unexplained discrepancies. To allow a more quantitative assessment of the role of natural variability in validation, tools to assess the impact of less-than-perfect collocations is provided by, e.g., Sofieva et al. (2008), Verhoelst et al. (2015) or Laeng et al. (2019). The latter tool estimates the difference between two measurements that is explained by natural variability and is based on a parametrization of high-resolution model data. It saves the validation scientist from the need of dedicated model studies for each comparison. Also dense and precise high-resolution measurements can be used as so-called ‘fiducial reference measurements’. The latter approach allows simultaneous evaluation of natural variability and validation of error estimates, as discussed in Staten and Reichler (2009) and Sofieva et al. (2014).

6.7 Drifts

Instrument drift we understand is a false trend in the derived state variables which is caused by an unstable instrument. At first order, a drift can be avoided if regular and frequent calibration are performed or if the self-calibrating measurement procedures are employed. However, higher order effects, e.g., related to the non-linearity of the calibration curve, can lead to noticeable drifts. Eckert et al. (2014), e.g., found drifts in MIPAS ozone even though regular calibration was performed. This was for the reason that, due to detector aging, the non-linearity of the detector sensitivity changed with time. Also the notorious aging problem of space-based UV-measurements, degradation in the sense of a reduction in throughput due to intense solar radiation can cause drifts. Most often the cause are the coatings of optical elements.

Whenever ex ante drift estimates are available, they should of course be communicated to the data user. Since, however, drifts usually can be determined only reliably towards the end of a mission, it does not make sense to require drift estimates in data characterization papers, which are typically written in the early phase of a mission. A deeper discussion of drifts is found, e.g., in Hubert et al. (2016).

Spaceborne UV measurements are typically affected by particularly severe instrumental degradation, i.e., loss of throughput. This is usually caused by optical coatings degrading when exposed to UV radiation. If a tangent altitude normalization approach or another self-calibration approach is used, this degradation is not necessarily a big problem, but the signal to noise ratio will decrease over time.

The SBUV/2 instruments use an on-board calibration system to track relative spectral and temporal changes in diffuser reflectivity using a mercury lamp (e.g., DeLand et al., 2012). Since the solar diffuser is the only additional optical element between radiance and irradiance measurements, this system enables an accurate throughput change correction to be derived from SBUV/2 solar measurements. This correction is applied in the Level 1 processing and is not included in the error covariances.
Intrinsically self-calibrating measurement geometries such as solar and stellar occultation or regular calibration measurements using internal sources at first order remove this error. This does, however, not apply to drifts of the shape of the nonlinear detector response function as discussed above. To date, these drifts are not evaluated as part of the routine error analysis of the Level-2 product but they are assessed by careful comparison with other instruments. While it is not easily possible to get absolute drift estimates from this, at least the relative drifts between instruments can be estimated (e.g., Eckert et al., 2014; Laeng et al., 2017; Hubert et al., 2012; Rahpoe et al., 2015; DeLand et al., 2012). It is important to note that relative drifts between instruments may have causes beyond time-dependent calibration changes (e.g., a drift in tangent height registration as shown in Livesey et al. 2018, Bourassa et al. 2018, or Kramarova et al. 2018.).

6.8 Combination of Error Components

Within linear theory, errors of different sources combine additively and follow Gaussian error propagation. We have \( k \) covariance matrices of the dimension \( n \times n \) representing the errors of \( k \) different sources \( S_{x,1}, S_{x,2}, \ldots, S_{x,k} \) and get

\[
S_{x,\text{total}} = \begin{pmatrix}
I_n & I_n & \cdots & I_n
\end{pmatrix} \times
\begin{pmatrix}
S_1 & C_{1,2} & \cdots & C_{1,k} \\
C_{2,1} & S_2 & \cdots & C_{2,k} \\
\vdots & \vdots & \ddots & \vdots \\
C_{k,1} & C_{k,2} & \cdots & S_k
\end{pmatrix}
\begin{pmatrix}
I_n \\
I_n \\
\vdots \\
I_n
\end{pmatrix},
\]

where \( I_n \) is an \( n \times n \) identity matrix and where the \( C \) matrices represent covariances among the error sources. For independent error sources, these covariances are zero and combination of errors comes down to summing up the error covariance matrices. Beyond linear theory, the interaction of various error sources is best studied by means of ensemble sensitivity studies (see, e.g., Kulawik et al., 2019).

Some data providers publish total error estimates. This practice is also endorsed by Joint Committee for Guides in Metrology (JCGM) (2008a). There are, however, compelling arguments in favour of publishing the individual error components. Specifically, depending on the application of the data, the same type of error can act as random or systematic error. For example, in trend estimation constant biases of the target gas will fully cancel out. Conversely, if, e.g., the total chlorine budget is calculated, the systematic (i.e., time-independent) error components of the parent chlorine species can be fully uncorrelated among the species and thus have to be treated like fully uncorrelated random errors when the error of the total chlorine budget is estimated. In other words, the extent to which error components are ‘systematic’ is domain-dependent. An error which is systematic in time can be random in the altitude, species, or some other domain. Thus, the data user may be better helped by being given access to the individual error components and some advice on systematicity in the various domains.
7  **Recommendations**

The goal of the TUNER effort has always been to try to bring the atmospheric remote sensing community together to enable better science. While a great deal of work has been performed over many decades, certain questions about the intercomparability of different data sets continue to linger and can only be answered if the data provided satisfy the conditions of adequacy described in this paper. While TUNER is not the first attempt at achieving this lofty goal (and may not be the last), we believe that the TUNER group is well-suited to this task. With the aim of establishing a consensus on error reporting, the TUNER group is comprised of remote sensing retrieval experts representing instruments with well over a century of combined operational time and experience. Comprising both data providers and data users, the TUNER consortium aims to “practice what it preaches they preach” in the hopes that data from past, present, and future instruments may finally be used in a consistent and intercomparable fashion.

Based on the framework and consensus terminology outlined above, and in response to the conditions of adequacy formulated in Section 2, recommendations have been developed on how uncertainties shall be assessed and data characterization shall be reported. These recommendations may seem less specific than the reader might expect, but one-size-fits-all recommendations were found to be inadequate for the variety of instruments under consideration. In the following, we state the general principles that we consider to be useful. Further, we formulate recommendations with respect to the evaluation and reporting of random errors, systematic errors, and further diagnostic data. The respective conditions of adequacy which led to a particular recommendation are listed in brackets (see Section 2). When appropriate, the recommendation is followed by an example or a short discussion in order to elucidate the rationale behind the recommendation.

**R 1.** The language and notation used to describe the error budget must be clearly defined. This can be accomplished either by explicit definitions of all terms and symbols used or by reference to any available document that lays down a self-consistent terminology. We hope that this paper serves that purpose and that the terminology and notation introduced here will be found useful [CoA 1, CoA 3 CoA 4].

**R 2.** The choice of which error estimation scheme is adequate depends on the instrument and the specific retrieval scheme. Thus, no ‘one-size-fits-all’ error estimation scheme is recommended here. The responsibility for judging which treatment of uncertainties is adequate lies with the retrieval scientist, because only they can judge which error sources and error propagation mechanisms are relevant for a particular instrument or data product. Every effort should be made to make the error budget as complete as possible in the sense that all sizeable sources of uncertainty are included, either via linear mapping, sensitivity studies, or whatever is appropriate for the particular case under assessment [CoA 1, CoA 5, CoA 3]. An overview of the most commonly used retrieval schemes is given in Sections 4 and 5. Error sources are discussed in Section 6.

**R 3.** The ideal approach is to report the substantive contributions from each relevant error component separately [CoA 5, CoA 4].

The meaning of the reported uncertainties shall be clarified. Do they refer to $\pm 1\sigma$, $\pm 2\sigma$, etc. or to a specified confidence limit, such as 95% or 99%? Note that generalized Gaussian error propagation will usually produce error estimates in terms...
of variances, while Monte-Carlo type sensitivity studies enable the confidence limits to be directly estimated. If the one is transformed to the other, the assumed underlying distribution shall be reported CoA 1; CoA 4.

If representative error budgets are The reason for this recommendation is that an estimated error component due to one particular error source can be of random characteristic in one application and of systematic characteristic in another application. For example, errors due to uncertain strengths of spectral lines are random if, say, the chlorine budget is calculated from multiple chlorine-containing constituents, each having its own uncertainty due to spectroscopic data. Conversely, in the analysis of a time series of one species the estimated errors due to erroneous line intensities act as a systematic error. The data user is able to consider the relevant error components only if the error contributions are reported separately. If, in addition, the total error is reported, it is of utmost importance to tell the data user if the nature of each error component is chiefly additive (i.e., independent of the actual state value reported) or is chiefly relative (i.e., a scaling factor should include the systematic and the random components.

R 4. For each error source, it is often necessary to know if the resulting error components are independent between two subsets of data within a certain domain (time, space, species, etc.). For the first type, the estimated errors shall be reported in the same units as the state variable (e.g., Kelvin, ppmv, molec/cm³); for the second type, estimated errors shall be reported as percentage errors. With this information, the data user can adjust the error estimates to the particular scientific study example, the error component due to tangent altitude uncertainties can be correlated between different species retrieved from the same measurement. The error component due to spectroscopic data may be correlated in the altitude domain but uncorrelated between different species, etc. We recommend that data providers describe the correlation within each relevant domain either qualitatively or quantitatively, wherever possible, [CoA 3, CoA 65; CoA 3]. The need of this is illustrated by the example already described under Recommendation R 3. Another example are quasi-systematic errors which are random in the long run only but can be highly correlated on shorter time scales.

R 5. When instrument groups make the error components available, they should also indicate how they contribute which of them contribute primarily to the random and/or systematic error and which contribute primarily to the systematic error. Classification and combination of errors is most helpful to the data user if it is made by their systematic vs. random nature rather than by origin [CoA 5; CoA 3]. This is important, e.g., in the context of validation. If estimated errors are reported as aggregated parameter errors, and some of them are of systematic nature while the others are of random nature, the data user will not be able to judge which fraction of the bias or the standard deviation of the differences between two measurement systems is explained by the systematic or random error, respectively. On the face of it, this recommendation looks redundant with Recommendation R 4 applied to the time domain but it is not. Components of the error budget may be strongly autocorrelated in the time domain but still lead to zero bias and thus contribute to the random error only.

For each error source, it is often necessary to know if the resulting error components are independent between two subsets within a certain domain (time, space, species, etc.) We recommend that data providers describe the correlation within each relevant domain either qualitatively or quantitatively, wherever possible. CoA 5; CoA 3.
If certain estimated errors or other characterization data are known or suspected to depend systematically on time, latitude, or other parameters, this dependence should be reported.

R 6. The meaning of the reported uncertainties shall be clarified. Do they refer to $\pm 1\sigma$, $\pm 2\sigma$, etc. or to a specified confidence limit, such as 95% or 99%? Note that generalized Gaussian error propagation will usually produce error estimates in terms of variances, while Monte-Carlo-type sensitivity studies enable the confidence limits to be directly estimated. If the one is transformed to the other, the assumed underlying distribution shall be reported [CoA 31; CoA 4].

R 7. For all error components, the assumed ingoing uncertainties shall be reported in the relevant documentation, otherwise error propagation would not be traceable. It should also be reported which correlation characteristics were assumed (e.g., scalar perturbation of a profile, individual perturbation of its elements, or consideration of its full covariance matrix) [CoA 4]. With this information a data user can re-scale error estimates if there is some doubt about the assumption on ingoing uncertainties (e.g., the $\Delta b$ in Eq. 22 or $S_b$ in Eq. 23) or if in a comparison study the error estimates of one instrument are more optimistic or pessimistic than those of the other.

R 8. If the retrieval uses prior information in the sense of Eq. 4 or Eq. 6, the a priori profiles must be reported to allow the data user to apply Eq. (30) or to perform similar operations—variants of it [CoA 5, CoA 4]. Also for column retrievals where an a priori profile is scaled to obtain the best fit and then integrated over altitude to render the column, the a priori profile should be reported, because the column can depend on the assumed profile shape.

If the retrieval is performed in any other space than state variable over altitude, pressure, or likewise (e.g., eigenvectors or similar), then the final result should be presented as vertical profiles and also all diagnostic data (error estimates, averaging kernels) should be transformed to an altitude-dependent representation CoA 5, CoA 2, CoA 3.

R 9. In addition to the error budget, averaging kernels (Eq. 24) should be reported. If a certain retrieval scheme does not give direct access to averaging kernels (e.g., onion peeling) then averaging kernels shall be determined by sensitivity studies based on delta perturbations of the profile. For retrieval approaches using truncated singular value decomposition or related approaches, the final altitude resolution shall be expressed as averaging kernels. For global fit maximum likelihood retrievals (no regularization) the averaging kernels are by definition unity, but only in the native retrieval grid. In such cases, regridding of data will give rise to non-unity averaging kernels. At the very least, the original grid and the interpolation scheme shall be reported. Ideally the data provider calculates the averaging kernels on the final grid on which the data are provided to the user. For retrieval of vertical columns the column averaging kernels (Eq. 26) are the respective diagnostic [CoA 1, CoA 5, CoA 2, CoA 3].

R 10. The space to which the averaging kernel applies (e.g., linear/logarithmic, mixing ratio/density, absolute/relative, etc.) shall be reported. This is particularly important when data are reported in a form that differs from that of the retrieval state vector [CoA 1, CoA 5, CoA 3, CoA 4]. E.g., the averaging kernels resulting from a retrieval of the logarithms of mixing ratios must not be applied to the mixing ratios themselves. It is thus of utmost importance to communicate to the data user to which quantities the averaging kernels refer.
If, for application to mean profiles, mean averaging kernels are provided in conjunction with mean profiles instead of individual ones, then the correlation profiles between the averaging kernels and the retrieved profiles shall be provided (von Clarmann and Glatthor, 2019) CoA 6.

R 11. Publishing combined error estimates that include the smoothing error is insufficient. Error propagation of the smoothing error in the context of interpolation to finer grids will usually fail (von Clarmann, 2014) to produce the full smoothing error on the fine grid (von Clarmann, 2014). Instead, if smoothing error is considered, it should be reported separately, allowing the data user to propagate all errors except the smoothing error through the interpolation and to evaluate the smoothing error directly on the fine grid, if desired. A caveat on the peculiarities related to the interpolation of the smoothing error is adequate if the smoothing error is reported. If the data are understood to be a representation of the smoothed state of the atmosphere (Section 6.4.2), the smoothing error is not needed and averaging kernels along with the prior information are sufficient [CoA 2]. For example, in data assimilation, where the averaging kernel is part of the observation operator, inclusion of the smoothing error into the observation error covariance matrix would be redundant and lead to incorrect double counting of the smoothing effect.

R 12. If an altitude-resolved retrieval is performed in any other space than state value over altitude, pressure, or likewise (e.g., if eigenvectors or similar are used, see Section 5.1), 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 the respective representation [CoA 5, CoA 2, CoA 3]. While these alternative representations certainly have their advantages, the data producer is in a better position than the data user to provide the diagnostic data for a profile representation.

R 13. Communication of a complete error budget for each profile, broken down to all components with all correlation information, along with averaging kernels and a priori information used, is not always technically feasible and often creates unnecessary data traffic. Retrieval scientists should judge whether evaluation of error budgets and averaging kernels for a limited number of representative cases is adequate. If averaging kernels are only provided for a few representative cases, one might still consider to show at least the vertical resolution profiles for each profile [CoA 6].

If the following recommendations R 14–17 are applicable to the case when only representative diagnostic data are available.

In this context we would like to mention that there exist methods to convey the information content of a measurement at drastically reduced data volume (Migliorini et al., 2008). Such methods are particularly convenient in the context of data assimilation.

R 14. If representative error estimates are reported instead of error estimates for each single profile or data point, it is of utmost importance to tell the data user if the nature of each error component is chiefly additive (i.e., independent of the actual state value reported) or is chiefly relative (i.e., a scaling factor). For the first type, the estimated errors shall be reported in the same units as the state variable (e.g., Kelvin, ppmv, molec./cm³); for the second type, estimated errors shall be reported as percentage errors. With this information, the data user can adjust the error estimates to the particular scientific study [CoA 3, CoA 6]. For
example, measurement noise often leads to an additive error component, i.e., the estimated error is approximately of the same size, regardless how large the mixing ratio of the target gas is. Conversely, errors representing spectroscopic uncertainties are often multiplicative. That is to say, larger profile values have larger errors.

**R 15.** If certain estimated errors or other characterization data are known or suspected to depend systematically on time, latitude, or other parameters, this dependence should be reported, particularly if only representative errors are reported [CoA 3]. For example, in infrared emission spectroscopy the precision of concentration retrievals is usually worse for a colder atmosphere. With this information a data user who is using a retrieval of a particular cold day which is not well represented by the sample error estimates is warned that the actual precision may be worse than the reported one.

**R 16.** If, for application to mean profiles, mean averaging kernels are provided in conjunction with mean profiles instead of individual ones, then the correlation profiles between the averaging kernels and the retrieved profiles shall be provided (von Clarmann and Glatthor, 2019) [CoA 6].

**R 17.** If, in order to reduce the data volume of profile data characterization, only standard deviations are reported for the individual profiles instead of the full covariance matrices, then a representative random error correlation pattern in the altitude domain (correlation matrix) shall be made available. With this, the user can approximate individual covariance matrices [CoA 5, CoA 3, CoA 6].

**R 18.** The final criterion of adequacy of error reporting is whether discrepancies between measurements of the same atmospheric state variable by independent measurement systems can be explained by the error estimates. This practical and empirical criterion of completeness of the error budget does not require knowledge of the unknowable true value of the measurand [CoA 1, CoA 5,]. In this context we distinguish between random and systematic errors.

1. We consider random error estimation schemes as adequate if a combination of the deduced error and the less-than-perfect spatial or temporal coincidences between two data sets and natural variability together explain the observed standard deviation of the differences between two data sets. If predicted random errors fail to explain observed differences, they should be reassessed. Methods to find out which of the compared data sets has an inadequate random error estimate have been described in, e.g., Fioletov et al. (2006); Sofieva et al. (2014); Laeng and von Clarmann (2019) [CoA 5].

2. We consider estimates of the systematic errors to be adequate if they, along with sampling biases and after accounting for different vertical/horizontal/temporal resolutions and content of a priori information, explain the observed biases between independent instruments [CoA 5].

We consider it undesirable and a source of confusion to still report over-optimistic or pessimistic ex-ante error budgets without a related caveat if validation studies show that there is strong indication that the actual errors are significantly larger or smaller.

On the face of it, the list of recommendations appears quite weak, leaving a lot of freedom to the data provider. This is, however, not the case. Recommendation R 2, that the error budget should be as complete as possible, along with
Recommendation R 18, which gives a criterion for the completeness of the error budget quickly make the apparent freedom disappear.

Admittedly, these recommendations will not guarantee perfect compliance with the conditions of adequacy, but due to the competing needs of rigor versus practicability the problem seems overconstrained. In other words, you ‘can’t always get what you want’ (Jagger and Richards, 1969). However, we are still confident that they help to unify uncertainty reporting in the community of remote sensing of atmospheric composition and temperature. 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.

8 Discussion and Outlook

In this paper we have discussed conventional (as opposed to machine learning and artificial-intelligence based approaches) error estimation methods for Bayesian and non-Bayesian retrieval methods. The choice of the retrieval method is a dilemma. If likelihood-based methods are chosen, the retrieval lacks a probabilistic interpretation and ad hoc constraints will imply a bias, at least if the retrieval is conceived as a smoothed estimate of the true state. This horn of the dilemma is avoided by Bayesian methods, which use probabilistic constraints. Adherents of likelihood-based methods, however, will point out the second horn of the dilemma, which is, that it is never warranted that the a priori statistics chosen indeed represents the true background state. Further, they will raise the concern that Bayesian methods, even if based on the true background statistics, may render bias-free estimates in the long run, but may be off the true atmospheric state in a single case. The decision for the acceptance of the one or the other horn of the dilemma is a philosophical one and in most cases it cannot be based on scientific grounds. The only recommendation we can offer in this respect is a plea for mutual tolerance. Regardless which approach is chosen, the data characterization has to be consistent with the retrieval method chosen. This paper tries to provide the scientific basis for this.

This paper is mainly addressed to providers of Level-2 data, i.e., data on atmospheric state variables. Some data users, however, prefer to work directly with Level-1 data, i.e., with measured radiances or transmissions. For example, the direct data assimilation of measured signals is sometimes preferred over the assimilation of retrieved state variables (e.g., Andersson et al., 1994). The radiative transfer forward model is in this case included in the assimilation scheme. The advantage of this method is that it avoids all problems related to a priori knowledge and regularization. We have touched this approach only upon passing in this paper and do not want do delve deeper into this. The only caveat we wish to add is that the observation error covariance matrix should not include measurement noise only but also contributions by uncertain parameters not assimilated (Section 5.2). This will typically lead to non-sparse observation error covariance matrices which may be the source of some further headache.

In some fields of remote sensing of the atmosphere, retrieval methods based on artificial intelligence, neural networks and machine learning are explored (Lary et al., 2016). A precondition for unification of error reporting of classical and artificial-intelligence based retrieval schemes seems to be semantic connectibility. The glossary by
Stanford University (2020) is considered as an important first step. With respect to the data characterization of retrieval products generated with such algorithms, two cases have to be distinguished. The first case is that a neural network is used as a surrogate radiative transfer forward model, while the retrieval still follows the concepts presented in this paper. In this case, the error estimation and data characterization strategies discussed in Section 6 are still applicable, and the approximative nature of the neural-network based radiative transfer calculation can simply be conceived as a further source of forward modelling error. The second case is that machine learning algorithms are directly used for the retrieval. In this case, complete data characterization appears to be more challenging to us. Sensitivity studies or supervised learning of uncertainty prediction may be two possible pathways towards data characterization of artificial-intelligence based retrievals. In either case it seems important to us that the data user is provided with the same full data characterization as required for the conventional retrieval schemes.

But even with the conventional retrieval and error estimation schemes there is a lot of homework to do. We hope that this review paper has identified the most relevant problems in this field and provides a conceptual framework to adequately characterize remotely sensed atmospheric temperature and composition data.
<table>
<thead>
<tr>
<th>Instrument</th>
<th>Processor</th>
<th>Geometric Regularization</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACE-FTS</td>
<td>Version 3.6/3.6 global fit</td>
<td>0</td>
<td>Boone et al. (2005)</td>
</tr>
<tr>
<td>ATMOS</td>
<td>Version 1-2 onion peeling</td>
<td>0</td>
<td>Norton and Rinsland (1991)</td>
</tr>
<tr>
<td></td>
<td>Version 3 global fit</td>
<td>0</td>
<td>Irion et al. (2002)</td>
</tr>
<tr>
<td>HALOE</td>
<td>NASA onion peeling</td>
<td>0</td>
<td>Thompson and Gordley (2009)</td>
</tr>
<tr>
<td>HIRDLS</td>
<td>HIRDLS global fit</td>
<td>0</td>
<td>Gille et al. (2008)</td>
</tr>
<tr>
<td>ILAS</td>
<td>Version 1 onion-peeling</td>
<td>0</td>
<td>Yokota et al. (2002)</td>
</tr>
<tr>
<td>ILAS-2</td>
<td>Version 1.4 onion-peeling</td>
<td>0</td>
<td>Nakajima (2006)</td>
</tr>
<tr>
<td>MIPAS</td>
<td>UBologna geofit</td>
<td>0</td>
<td>Dinelli et al. (2010)</td>
</tr>
<tr>
<td></td>
<td>ESA global fit</td>
<td>0</td>
<td>Raspolli et al. (2013)</td>
</tr>
<tr>
<td></td>
<td>IMK/IAA global fit³</td>
<td>0</td>
<td>von Clarmann et al. (2003, 2009)</td>
</tr>
<tr>
<td></td>
<td>UOxford global fit⁴</td>
<td>0</td>
<td>Dudhia (2019)</td>
</tr>
<tr>
<td>MLS-AURA</td>
<td>NASA geofit⁵</td>
<td>0</td>
<td>Livesey et al. (2003)</td>
</tr>
<tr>
<td>MLS-UARS</td>
<td>NASA global fit</td>
<td>0</td>
<td>Livesey et al. (2003)</td>
</tr>
<tr>
<td>Odin-SMR</td>
<td>Chalmers global fit</td>
<td>0</td>
<td>Urban et al. (2005)</td>
</tr>
<tr>
<td></td>
<td>NASA global fit</td>
<td>0</td>
<td>Degenstein et al. (2009)</td>
</tr>
<tr>
<td>OSIRIS</td>
<td>Version 5 global fit</td>
<td>0</td>
<td>Degenstein et al. (2009)</td>
</tr>
</tbody>
</table>
Table 1: Satellite Data Processors, Limb Geometry (Emission, Occultation and Scattering) continued:

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Processor</th>
<th>Geometric Decomposition</th>
<th>Regularization</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>SABER</td>
<td>NASA</td>
<td>onion peeling</td>
<td>0</td>
<td>Bourassa et al. (2012)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>interleaved</td>
<td></td>
<td>Russell III et al. (1999)</td>
</tr>
<tr>
<td>SAGE I</td>
<td>NASA</td>
<td>Chahine&lt;sup&gt;6&lt;/sup&gt;</td>
<td>0</td>
<td>Rong et al. (2009)</td>
</tr>
<tr>
<td>SAGE II</td>
<td>NASA</td>
<td>global fit&lt;sup&gt;7&lt;/sup&gt;</td>
<td>0</td>
<td>Chu and McCormick (1979)</td>
</tr>
<tr>
<td>SAGE III</td>
<td>NASA</td>
<td>global fit</td>
<td>0</td>
<td>Damadeo et al. (2013)</td>
</tr>
<tr>
<td>SAMS</td>
<td>NASA</td>
<td>sequential estimation</td>
<td>S&lt;sub&gt;a&lt;/sub&gt;&lt;sup&gt;-1&lt;/sup&gt;</td>
<td>Wofsy et al. (2002)</td>
</tr>
<tr>
<td>SCIAMACHY</td>
<td>IMK</td>
<td>global fit</td>
<td>S&lt;sub&gt;a&lt;/sub&gt;&lt;sup&gt;-1&lt;/sup&gt; + L&lt;sub&gt;T&lt;/sub&gt;&lt;sup&gt;γ&lt;/sup&gt;&lt;sub&gt;1&lt;/sub&gt;&lt;sub&gt;L&lt;/sub&gt;&lt;sub&gt;1&lt;/sub&gt;</td>
<td>Bender et al. (2013)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Bender et al. (2017)</td>
</tr>
<tr>
<td></td>
<td>IUP (limb scatter) Bremen (limb scatter)</td>
<td>global fit</td>
<td>L&lt;sub&gt;T&lt;/sub&gt;&lt;sup&gt;γ&lt;/sup&gt;&lt;sub&gt;1&lt;/sub&gt;&lt;sub&gt;L&lt;/sub&gt;&lt;sub&gt;0&lt;/sub&gt;&lt;sub&gt;γ&lt;/sub&gt;&lt;sub&gt;1&lt;/sub&gt;&lt;sub&gt;L&lt;/sub&gt;&lt;sub&gt;0&lt;/sub&gt; + L&lt;sub&gt;T&lt;/sub&gt;&lt;sup&gt;γ&lt;/sup&gt;&lt;sub&gt;2&lt;/sub&gt;&lt;sub&gt;L&lt;/sub&gt;&lt;sub&gt;1&lt;/sub&gt;</td>
<td>Rozanov et al. (2011b)</td>
</tr>
<tr>
<td></td>
<td>IUP (occultation) Bremen (occultation)</td>
<td>onion peeling</td>
<td>0</td>
<td>Noël et al. (2016)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Noël et al. (2018)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>global fit</td>
<td>L&lt;sub&gt;T&lt;/sub&gt;&lt;sup&gt;γ&lt;/sup&gt;&lt;sub&gt;1&lt;/sub&gt;&lt;sub&gt;L&lt;/sub&gt;&lt;sub&gt;0&lt;/sub&gt;&lt;sub&gt;γ&lt;/sub&gt;&lt;sub&gt;1&lt;/sub&gt;&lt;sub&gt;L&lt;/sub&gt;&lt;sub&gt;0&lt;/sub&gt; + L&lt;sub&gt;T&lt;/sub&gt;&lt;sup&gt;γ&lt;/sup&gt;&lt;sub&gt;2&lt;/sub&gt;&lt;sub&gt;L&lt;/sub&gt;&lt;sub&gt;1&lt;/sub&gt;</td>
<td>Azam et al. (2012)</td>
</tr>
<tr>
<td>SMILES</td>
<td>JAXA</td>
<td>onion peeling, global fit</td>
<td>S&lt;sub&gt;a&lt;/sub&gt;&lt;sup&gt;-1&lt;/sup&gt;</td>
<td>Takahashi et al. (2010)</td>
</tr>
<tr>
<td>NICT</td>
<td>onion peeling, global fit</td>
<td>S&lt;sub&gt;a&lt;/sub&gt;&lt;sup&gt;-1&lt;/sup&gt;</td>
<td>Baron et al. (2011)</td>
<td></td>
</tr>
<tr>
<td>SOFIE</td>
<td>GATS</td>
<td>onion peeling</td>
<td>0</td>
<td>Gordley et al. (2009b)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>interleaved</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(1) For processors with multiple data products, the actual regularization may vary depending on the retrieved atmospheric parameter.
(2) sequential estimation using a Kalman filter
(3) under consideration of horizontal gradients
(4) sequential estimation in the spectral domain
(5) subsets of orbits are used.
(6) SAGE team’s best guess as original documentation was lost.
(7) onion peeling for H<sub>2</sub>O.
<table>
<thead>
<tr>
<th>Instrument</th>
<th>Processor</th>
<th>Column or Profile Retrieval</th>
<th>Regularization $^1$</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIRS</td>
<td>NASA v6</td>
<td>profiles</td>
<td>PCA</td>
<td>Susskind et al. (2014)</td>
</tr>
<tr>
<td>NUCAPS</td>
<td>profiles</td>
<td>PCA</td>
<td></td>
<td>Susskind et al. (2003)</td>
</tr>
<tr>
<td>CLIMCAPS</td>
<td>profiles</td>
<td>$S_a^{-1}$</td>
<td>Smith and Barnet (2019)</td>
<td></td>
</tr>
<tr>
<td>MUSES</td>
<td>profiles</td>
<td>$L_1^T \gamma_1 L_1$ or $L_0^T \gamma_2 L_0$</td>
<td>Fu et al. (2018)</td>
<td></td>
</tr>
<tr>
<td>CrIS</td>
<td>NUCAPS</td>
<td>profiles</td>
<td>PCA</td>
<td>Susskind et al. (2003)</td>
</tr>
<tr>
<td>CLIMCAPS</td>
<td>profiles</td>
<td>$S_a^{-1}$</td>
<td>Smith and Barnet (2019)</td>
<td></td>
</tr>
<tr>
<td>CFPR</td>
<td>profiles</td>
<td>$S_a^{-1}$</td>
<td>Shephard and Cady-Pereira (2015)</td>
<td></td>
</tr>
<tr>
<td>MUSES</td>
<td>profiles</td>
<td>$L_1^T \gamma_1 L_1$ or $L_0^T \gamma_2 L_0$</td>
<td>Fu et al. (2016)</td>
<td></td>
</tr>
<tr>
<td>GOME</td>
<td>AMC-DOAS</td>
<td>columns</td>
<td>0</td>
<td>Noël et al. (1999)</td>
</tr>
<tr>
<td>GODFIT(BIRA)</td>
<td>columns</td>
<td>0</td>
<td>Lerot et al. (2010)</td>
<td></td>
</tr>
<tr>
<td>DOAS</td>
<td>columns</td>
<td>0</td>
<td>Richter et al. (1998)</td>
<td></td>
</tr>
<tr>
<td>OPERA (KNMI)</td>
<td>profiles</td>
<td>$S_a^{-1}$</td>
<td>van Peet et al. (2014)</td>
<td></td>
</tr>
<tr>
<td>RAL</td>
<td>profiles</td>
<td>$S_a^{-1}$</td>
<td>Siddans (2003)</td>
<td></td>
</tr>
<tr>
<td>WFDOAS</td>
<td>columns</td>
<td>0</td>
<td>Weber et al. (2018)</td>
<td></td>
</tr>
<tr>
<td>GOME-2</td>
<td>AMC-DOAS</td>
<td>columns</td>
<td>0</td>
<td>Noël et al. (2008)</td>
</tr>
<tr>
<td>GODFIT(BIRA)</td>
<td>columns</td>
<td>0</td>
<td>Lerot et al. (2010)</td>
<td></td>
</tr>
<tr>
<td>DOAS</td>
<td>columns</td>
<td>0</td>
<td>Vrekoussis et al. (2009)</td>
<td></td>
</tr>
<tr>
<td>OPERA (KNMI)</td>
<td>profiles</td>
<td>$S_a^{-1}$</td>
<td>van Peet et al. (2014)</td>
<td></td>
</tr>
<tr>
<td>RAL</td>
<td>profiles</td>
<td>$S_a^{-1}$</td>
<td>Siddans (2003)</td>
<td></td>
</tr>
<tr>
<td>GOSAT</td>
<td>ACOS</td>
<td>XCO$_2$</td>
<td>$S_a^{-1}$</td>
<td>O’Dell et al. (2012)</td>
</tr>
<tr>
<td>BESD</td>
<td>XCO$_2$</td>
<td>$S_a^{-1}$</td>
<td>Heymann et al. (2015)</td>
<td></td>
</tr>
<tr>
<td>NIES</td>
<td>columns/profiles</td>
<td>$S_a^{-1}$</td>
<td>Yoshida et al. (2013)</td>
<td></td>
</tr>
<tr>
<td>RemoTeC</td>
<td>profiles</td>
<td>$L_1^T \gamma_1 L_1$ or $L_0^T \gamma_2 L_0$</td>
<td>Butz et al. (2011)</td>
<td></td>
</tr>
<tr>
<td>UOL</td>
<td>XCO$_2$</td>
<td>$S_a^{-1}$</td>
<td>Cogan et al. (2012)</td>
<td></td>
</tr>
<tr>
<td>IASI</td>
<td>ASIMUT-ALVL (BIRA)</td>
<td>profiles</td>
<td>$S_a^{-1}$</td>
<td>De Wachter et al. (2017)</td>
</tr>
<tr>
<td>Instrument</td>
<td>Processor</td>
<td>Column or Profile Retrieval</td>
<td>Regularization $^1$</td>
<td>Reference</td>
</tr>
<tr>
<td>------------</td>
<td>-----------</td>
<td>-----------------------------</td>
<td>----------------------</td>
<td>-----------</td>
</tr>
<tr>
<td>EUMETSAT</td>
<td>profiles</td>
<td>$S_a^{-1}$</td>
<td>August et al. (2012)</td>
<td></td>
</tr>
<tr>
<td>FORLI(ULB/LATMOS)</td>
<td>profiles</td>
<td>$S_a^{-1}$</td>
<td>Hurtmans et al. (2012)</td>
<td></td>
</tr>
<tr>
<td>MUSICA (IMK)</td>
<td>profiles</td>
<td>$S_a^{-1}$</td>
<td>Borger et al. (2018)</td>
<td></td>
</tr>
<tr>
<td>RAL</td>
<td>profiles</td>
<td>$S_a^{-1}$</td>
<td>Siddans et al. (2017)</td>
<td></td>
</tr>
<tr>
<td>δ-IASI</td>
<td>profiles</td>
<td>$S_a^{-1}$</td>
<td>Liuzzi et al. (2016)</td>
<td></td>
</tr>
<tr>
<td>MOPITT V8</td>
<td>profiles</td>
<td>$S_a^{-1}$</td>
<td>Deeter et al. (2019)</td>
<td></td>
</tr>
<tr>
<td>OCO-2</td>
<td>FOCAL</td>
<td>CO$_2$ profiles</td>
<td>Reuter et al. (2017b, a)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>NASA/V9</td>
<td>XCO$_2$</td>
<td>O’Dell et al. (2018)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>UOL</td>
<td>XCO$_2$</td>
<td>Boesch et al. (2011)</td>
<td></td>
</tr>
<tr>
<td>RemTeC</td>
<td>profiles</td>
<td>$L_{1}^T \gamma L_{1}$</td>
<td>Wu et al. (2018)</td>
<td></td>
</tr>
<tr>
<td>OMI DOAS</td>
<td>columns</td>
<td>0</td>
<td>Theys et al. (2015)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Boersma et al. (2018)</td>
<td></td>
</tr>
<tr>
<td>GODFIT (BIRA)</td>
<td>columns</td>
<td>0</td>
<td>Lerot et al. (2010)</td>
<td></td>
</tr>
<tr>
<td>Harvard-SAO V3</td>
<td>columns</td>
<td>0</td>
<td>González Abad et al. (2015)</td>
<td></td>
</tr>
<tr>
<td>MUSES</td>
<td>profiles</td>
<td>$L_{0}^T \gamma_2 L_{0}$ or $S_a^{-1}$</td>
<td>Fu et al. (2018)</td>
<td></td>
</tr>
<tr>
<td>NASA-GSFC</td>
<td>columns</td>
<td>0</td>
<td>Bhartia and Wellemeyer (2002)</td>
<td></td>
</tr>
<tr>
<td>OPERA (KNMI)</td>
<td>profiles</td>
<td>$S_a^{-1}$</td>
<td>van Peet et al. (2014)</td>
<td></td>
</tr>
<tr>
<td>RAL</td>
<td>profiles</td>
<td>$S_a^{-1}$</td>
<td>Siddans (2003)</td>
<td></td>
</tr>
<tr>
<td>OMPS-NM</td>
<td>NASA-GSFC</td>
<td>columns</td>
<td>Bhartia and Wellemeyer (2002)</td>
<td></td>
</tr>
<tr>
<td>OMPS-NP</td>
<td>NASA-GSFC</td>
<td>profiles $S_a^{-1}$</td>
<td>Bhartia et al. (2013)</td>
<td></td>
</tr>
<tr>
<td>SBUV</td>
<td>NASA</td>
<td>profiles $S_a^{-1}$</td>
<td>Bhartia et al. (2013)</td>
<td></td>
</tr>
<tr>
<td>SCIAMACHY</td>
<td>AMC-DOAS</td>
<td>columns</td>
<td>Noël et al. (2004)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BESD</td>
<td>profiles $S_a^{-1}$</td>
<td>Reuter et al. (2010)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Reuter et al. (2011)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>DOAS</td>
<td>columns</td>
<td>Afe et al. (2004)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Boersma et al. (2018)</td>
<td></td>
</tr>
<tr>
<td>OPERA (KNMI)</td>
<td>profiles</td>
<td>$S_a^{-1}$</td>
<td>van Peet et al. (2014)</td>
<td></td>
</tr>
<tr>
<td>RAL</td>
<td>O$_3$ profiles</td>
<td>$S_a^{-1}$</td>
<td>Siddans (2003)</td>
<td></td>
</tr>
<tr>
<td>WMF-DOAS</td>
<td>columns</td>
<td>0</td>
<td>Schneising et al. (2012)</td>
<td></td>
</tr>
</tbody>
</table>
Table 2: Satellite Data Processors: Nadir sounders (continued)

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Processor</th>
<th>Column or Profile Retrieval</th>
<th>Regularization&lt;sup&gt;1&lt;/sup&gt;</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>TES</td>
<td>v7</td>
<td>profiles</td>
<td>$L_T^T \gamma_1 L_1 /$</td>
<td>Bowman et al. (2006)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$L_2^T \gamma_2 L_2$ or</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$S_a^{-1}$</td>
<td></td>
</tr>
<tr>
<td>TROPOMI</td>
<td>DOAS</td>
<td>columns</td>
<td>0</td>
<td>Theys et al. (2017)</td>
</tr>
<tr>
<td>GODFIT (BIRA)</td>
<td>columns</td>
<td></td>
<td>0</td>
<td>Lerot et al. (2010)</td>
</tr>
<tr>
<td>RemoTeC</td>
<td>profiles</td>
<td></td>
<td>$L_T^T \gamma_1 L_1 /$</td>
<td>Hu et al. (2018)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$L_0^T \gamma_2 L_0$</td>
<td></td>
</tr>
<tr>
<td>SICOR</td>
<td>columns</td>
<td></td>
<td>0</td>
<td>Borsdorff et al. (2018)</td>
</tr>
<tr>
<td>WMF-DOAS</td>
<td>columns</td>
<td></td>
<td>0</td>
<td>Schneising et al. (2019)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Weber et al. (2018)</td>
</tr>
</tbody>
</table>

<sup>1</sup> For processors with multiple data products, the actual regularization may vary depending on the retrieved atmospheric parameter, and whether it is a column or profile.
**Code and data availability.** N/A

**Author contributions.** TvC, DAD and NJL organized the project. TvC, NJL and RD wrote major parts of the text. All authors contributed to the discussion of the paper and particularly the recommendations.

**Competing interests.** TvC, CvS and GPS are associate editors of AMT but have not been involved in the evaluation of this paper.

5 **Acknowledgements.** The World Meteorological Organization (WMO) has provided travel support through the Stratosphere-troposphere Processes And their Role in Climate (SPARC) project who have selected the TUNER project as a SPARC activity. The International Space Science Institute (ISSI) has funded two International Team meetings in Berne at their venue. Part of this work was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under contract with NASA. The authors would like to thank Andreas Richter for providing information on UV-VIS nadir sounding products.
References

Adrain, R.: Research Concerning the Probabilities of the Errors which Happen in Making Observations, &c., The Analyst, or Mathematical Museum, 1, 93–109, the original pagination of the paper could not be verified. We refer to a version available via https://www.cs.xu.edu/math/Sources/Adrain/1808-Analyst.pdf, retrieved on 8 Feb 2018, 1808.


George and Ira Gershwin: It ain’t necessarily so, in: Porgy and Bess, Colonial Theater, Boston, 1935.


Joint Committee for Guides in Metrology (JCGM): International vocabulary of metrology - Basic and general concepts and associated terms (VIM), Pavillon de Breteuil, F-92312 Sèvres Cedex, 3rd edn., 2008b.


