Atmos. Meas. Tech. Discuss.,
doi:10.5194/amt-2019-350-RC1, 2019

Atmospheric Measurement Techniques Discussions

Interactive comment on "Estimating and Reporting Uncertainties in Remotely Sensed Atmospheric Composition and Temperature" *by* Thomas von Clarmann et al.

Anonymous Referee #1

Received and published: 3 December 2019

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.

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.

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.

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

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. 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.

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. 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?

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).

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

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).

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. 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. 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.

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.

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.

Introduction

l6: "reduction"? Should this be "deduction"?

l16: "The project ... is a consortium of". Please modify

l24: "atmospheric composition and temperature profiles". What about other profiles, e.g. water vapour? Is the paper limited to profiles, or are single property (column) retrievals also included?

l37: "are do not need to be", please correct.

Section 2:

l82, CoA 1: "and/or error estimation schemes". Would it not be better to say "and/or retrieval schemes"?

CoA 2: "independent of the vertical grid". But I assume at this point that error covari-

ances are specified on a specific grid used in the retrieval ?!

CoA 5: "and different amounts of prior information". Do you mean "and different sources of prior information"?

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.

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.

sec 3.1, l34: "actually claimed that there are conceptual differences between error analysis and un- certainty 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.

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).

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

l70: "macrorcopic"

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?

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

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".

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

eq. 5: What is L1? What are its properties?

Sec.5:

l22: mention the loss of information

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.

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).

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

Sec 6.1.1, l37: "cheerful" ...

Sec 6.1.3, l33: "measurments"

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).

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.

l24: "retrived"

p16, l40: "the derivative". I do not understand how to take such a derivative.

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.

6.3: Why is this section called "parameter errors" instead of something like "Inverse model decomposition errors"

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.

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).

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

p19, l10: "reasons reasons"

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

Section 7

Point 2: A bit weak, it leaves a lot of room for different approaches.

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

Point 4: Again leaves much freedom. What about proposing a 1 sigma as default?

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? What does "indicate" mean in practice? Please be more specific.

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. What is the difference between an error source and an error component?

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?

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?

Point 10: What are "similar operations"? Please be more explicit.

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.

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.

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?

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.

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?

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.

Interactive comment on Atmos. Meas. Tech. Discuss., doi:10.5194/amt-2019-350, 2019.