Review #1: 'Truth and Uncertainty at the Crossroads' by Antonio Possolo

Comment: Recommendation
The article's Abstract sums up the central claims accurately that the authors develop and substantiate in their narrative, including what I believe to be the correct conclusion that the "error" and "uncertainty" concepts are not fundamentally different, and may be regarded as alternative and complementary interpretations of the doubt about the true value of the measurand that remains after measurement.
However, my reading of the Guide to the Expression of Uncertainty in Measurement (GUM) [Joint Committee for Guides in Metrology (JCGM), 2008] suggests less polarizing views about this issue than the views that the authors of the article under review derive from the same Guide.
Their take on things brings to mind the acerbic discussion of essentially the same issues that took place in meetings of the ISO/TAG-4 Working Group 3, around 1986-87 [Collé, 1987a,b] [Schumacher, 1987].
The article should be published after it will have been shortened and more sharply focused to convey its message most effectively, and after improvements will have been made to deficient passages that are discussed under Specific Comments.

Reply: The authors thank the reviewer for this insightful and thorough review of their manuscript.

Action: See below, under specific comments.

Comment: The Technical Corrections offer an assortment of suggestions concerning English usage that the authors should consider.

Reply: The authors appreciate these corrections.

Action: See below for details.

Comment: General Comments
Acknowledgment should be made of an understanding of the relation between measurement uncertainty and measurement error that predates the GUM, and that the authors of the article under review likely will find agreeable: "The uncertainty of a reported value is meant to be a credible estimate of the likely limits to its actual error, i.e., the magnitude and sign of its deviation from the truth" [Eisenhart and Collé, 1980]. Churchill Eisenhart was my most illustrious predecessor at NIST, and Ronald Collé, a distinguished and esteemed

NIST colleague, served as convener of the working group (ISO-TAG-4/WG3) that laid the groundwork for the creation of the GUM [Collé and Karp, 1987].
A discussion whose tenor places the "error approach" on Mars and the "uncertainty approach" on Venus sounds more like the discussions that inflamed the metrological community thirty-five years ago, than a useful discussion that we can engage in today with the benefit of the experience accumulated in these many intervening years [Eisenhart and Collé, 1980] [Collé, 1987a] [Colclough, 1987] [Schumacher, 1987].
The viewpoint that the authors of the article under review wish to convey, can be conveyed quite simply also by means of an allegory: measurement errors are the "carriers" of measurement uncertainty, in a sense analogous to how photons are the "carriers" of light waves and, more generally, of the electromagnetic force.
Accepting such dualism between errors and uncertainty facilitates the scientific discourse without excluding individual or cultural preferences, and tones down the drama that has been unfolding in the literature and that, at times, this article also exacerbates unnecessarily.

Reply: Agreed.

Action: Relevant parts of the paper were reorganized and rewritten. Care has been taken not to give the impression of a split of the community in two blocs in hostile opposition. Relevant references have been included.

Comment: The 26 pages of text of the article under review arguably are overkill to convey this simple, conciliatory message: what they do prompt is a review almost as long as the article itself, thus making this review much too long by any standard.
In fact, the key message of the article will be delivered more effectively, and the article will have greater impact, if the article is shortened and its arguments are streamlined.
The article's length can be reduced at least by deleting those portions that distract more than they add insight: for example, the digressions in section 2 and in subsections 5.2 and 5.3.

Reply: We agree to shorten/rewrite particularly Section 2. We are not really convinced that Sections 5.2 and 5.3 are disgressions. The impossibility to evaluate errors/uncertainties in the nonlinear case without approximate knowledge of the true value and the inadequacy to conceive the measured value as the most probable value without consideration of the a priori probability seem essential to us. Instead we shortened the paper by reorganizing it in a way that repetition could better be avoided. Further, we have deleted the Section on Bayesianism versus non-Bayesianism, because the current version of GUM is not fully Bayesian, and this Section was thus connected only loosely to the

main topic of the paper.

Action: The paper has been substantially reorganized and shortened.

Comment: The authors may wish to extend their criticism to the International Vocabulary of Metrology (VIM) [Joint Committee for Guides in Metrology, 2012], whose Introduction states: "The change in the treatment of measurement uncertainty from an Error Approach (sometimes called Traditional Approach or True Value Approach) to an Uncertainty Approach necessitated reconsideration of some of the related concepts appearing in the second edition of the VIM."

Reply: Agreed.

Action: Added before the criticism is formulated: "The International Vocabulary of Metrology document (BIPM, 2012) points in the same direction"

Comment: The authors also seem to be unaware of the critical evaluation of the GUM that Gleser [1998] published shortly after the original, 1993 edition of the GUM was corrected and reprinted, in 1995 [BIPM et al., 1995]. References to suitable portions of this evaluation will add value to the article under review, and will also facilitate shortening it.

Reply: Yes, indeed. We were unaware of this review.

Action: Reference to Gleser (1998) has been made.

Comment: The article is very repetitive in the multiple instances where it rehashes the relations between the concepts of error, uncertainty, true value, and Bayesianism. Consolidating and refocusing the fragmentary discussion of these relations would make the article much easier to read and would enhance the cogency of its arguments. However, accomplishing this would involve a major rewrite.

Reply: Agreed.

Action: The paper has been reorganized, and parts of the article have been rewritten, in order to avoid repetition and to make the structure of the argument clearer.

Comment: In its Annex E (E.5.1) the GUM addresses the issue that is the main focus of the article under review, when it states that "The focus of this Guide is on the measurement result and its evaluated uncertainty rather than on the unknowable quantities "true" value and error (see Annex D). By taking the operational views that the result

of a measurement is simply the value attributed to the measurand and that the uncertainty of that result is a measure of the dispersion of the values that could reasonably be attributed to the measurand, this Guide in effect uncouples the often confusing connection between uncertainty and the unknowable quantities "true" value and error."
The authors of the article under review quite correctly point out that the uncertainty is neither a property of the measured value nor is it about the measured value. The uncertainty surrounds or clouds the true value, and qualifies the state of knowledge that the metrologist has of the true value. To the extent that the target of measurement is the true value, the measurement error is meaningful even if not observable (however, it can be estimated in some cases, as discussed below in relation with Line 180).

Reply: Fully agreed.

Comment: The suggestion, made in the aforementioned E.5.1, that "uncertainty," "error," and "true value" should be uncoupled from one another seems at odds with what is actually done in the practice of measurement science. For example, in relation with certified reference materials, "NIST asserts that a certified value provides an estimate of the true value of a defined measurand" [Beauchamp et al., 2020, 1.2.4].
Therefore, the implied understanding of the scientists developing these materials is that the uncertainties reported in the corresponding certificates are informative about the relation between the measured value and the true value, the difference between the former and the latter being the measurement error.

Reply: We fully agree.

Action: None, because the Beauchamp-statement seems to address the uncertainty issue only indirectly. Although this quotation could possibly strengthen our point, its inclusion would imply considerably more text, which we want to avoid for reasons of brevity.

Comment: Specific Comments
The numbers in boldface refer to line numbers in the version of the preprint made available for discussion on June 29, 2021.
002 + 245 Here and elsewhere throughout the article, "GUM8" should be replaced by "GUM" because the GUM and its existing and planned supplements are being rearranged and renumbered, and "GUM8" is already reserved to refer to something other than the current GUM. For similar reasons, "GUM09" is likely to be misinterpreted, and should not be used: in fact, it is not needed at all because the authors use this acronym only in the very same line (245)

where they introduce it.

Reply: Thank you for bringing this to our attention.

Action: To be able to still distinguish between GUM in general and the 2008 version, and not to clash with the GUM numbering system, we have changed 'GUM08' to 'GUM-2008'. The abbreviation GUM09 is not used any longer.

Comment: Specific Comments
010 *the term 'error' was used, with some caveats, for designating a statistical estimate of the expected difference between the measured and the true value of a measurand*
The traditional and still customary meaning of "error" in statistical models is of a non-observable difference between the observed and the (generally also non-observable) true value of a quantity [Davison, 2008, Example 1.1].
For example, in the relationship $m = \mu + \epsilon$ between a measured value, m, and the true value, μ, of the mass of a massive entity, ϵ is the error.
The error is generally neither known nor observable, but in many situations it can be estimated, with ϵ being commonly used to denote the estimate (refer to the discussion of Line 180). In the discussion of Lines 518 + 738 below, it will become clear how useful the explicit consideration of error can be, by allowing one conceptually to separate contributions made by different sources of uncertainty.

Reply: We fully agree.

Action: In the course or rewriting the terminology part, we have made clearer that in the traditional terminology, the term 'error' implies an equivocation.

Comment: 015 *stipulated a new terminology, where the term 'measurement uncertainty' is used in situations where one would have said 'measurement error'*
The word "error" occurs 131 times throughout the GUM, and not always deprecatingly. For example, in its 2.2.4, the GUM acknowledges that "The definition of uncertainty of measurement [...] is not inconsistent with other concepts of uncertainty of measurement, such as a measure of the possible error in the estimated value of the measurand."

Reply: In the definitions part, GUM defines error as the difference between the measured value and the measurand. There is no instance in GUM where they acknowledge the equivocation that the term 'error' can also refer to a statistical estimate of this difference. On page 5, Note 3, GUM says "In this Guide, great care is taken to distinguish between the terms "error" and "uncertainty". They

are not synonyms, but represent completely different concepts; they should not be confused with one another or misused." This, we think, justifies our statement.

Action: We have not taken any action in reply to this specific comment, but our rewriting of the introductory sections in reply to the general comment should have made our point clearer.

Comment: 025 *the error statisticians and the uncertainty statisticians*
This classification of statisticians into these two classes is an invention of the authors that is more reflective of their imagination than of reality. In fact, the principal participants in the debates that took place in and around the aforementioned ISO/TAG-4/WG-3 were not statisticians.

Reply: Agreed.

Action: In the new version of the manuscript, the terms "error statistician" and "uncertainty statistician" are no longer used.

Comment: Furthermore, the issue of "systematic" versus "random" errors (which we will discuss below, in relation with Line 597) may have been even more divisive than the issue of "error" versus "uncertainty." Therefore, I urge the authors to devise a different way of characterizing the two camps they are alluding to here. A reference to Mayo and Spanos [2011] would be appropriate.

Reply: We use a definition of random vs. systematic errors which is based fully on observational grounds. This terminology has been agreed by the entire TUNER consortium. To avoid confusion, we do not want to change the terminology again.

Action: We mention that our random errors correspond to the volatile errors and that our systematic errors correspond to the persistent errors. However, we use the terms 'volatile' and 'persistent' as purely descriptive terms, not as new technical terms.

Comment: 046 *according to Bayesian statistics (Bayes, 1763) the measured value cannot always be interpreted as the most probable value of the measurand*
Since one does not need to invoke Bayesian statistics to reach the same conclusion [Possolo and Iyer, 2017, Page 011301-12], this remark is spurious.

Reply: Here we have to respectfully disagree. The argument of Possolo and Iyer, 2017, Page 011301-12 is based on the assumption that the error correlations between multiple measured values are not known or not considered when

a higher-level data product is produced. Under this assumption the conclusion that the resulting value is not the most probable one is correct. However, von Clarmann et al. (2020) offer a method to estimate the higher-level data product (here: trace gas mixing ratios) under consideration of the full measurement error covariance matrix. Even if the latter does include all uncertainties and covariances, and the Possolo an Iyer argument thus does not apply, still the base rate problem is there, and without consideration of the a priori probabilities the estimate will **not** render the most probable mixing ratios but only the most likely ones. Thus, we do not see what is spurious about the base rate argument.

Comment: 071 Recapitulation of the concept of indirect measurements I believe that this long foray into inverse problems adds nothing of value to the discussion, hence suggest that section 2 be deleted. The discussion in subsection 6.3, The causal arrow, can easily be reformulated, and in the process also shortened, to drive the same points across — refer to specific suggestions below, for Line 618.

Reply: We agree in part. This section contained unnecessary formalism and detail, and it was too long. However, to understand how a measured or estimated value is probabilistically related to the true value, we still think that it is important to highlight the inverse nature of a measurement process.

Action: This section has been merged with other sections and has been considerably shortened. All unnecessary formalism has been removed.

Comment: 119 *ancient researchers realized that measurement results always have errors*
It is all a matter of perspective, of course, but I am of the opinion that it is unfair to call Gauss or Legendre "ancient." In the context of European history, the word is typically reserved for the period ending with the fall of the Western Roman Empire (around 500 CE). In addition, and in particular concerning Gauss, with whose works I am more familiar than with Legendre's, I can only say that it is difficult for me to imagine a person of more luminous modernity, or with a better sense for what is relevant in scientific practice (of his time or contemporary), than Gauss.

Reply: The term 'ancient' was by no means meant in any dismissive way. We realize that it can be understood in this way and replace it.

Action: New wording: "Investigators realized already in the 19th century that measurement results always have errors."

Comment: 149 *In the case of 'error', its statistical estimate is mostly understood to be a quadratic estimate and thus does not carry any information about the sign of the error.*

The authors may like to replace this awkward sentence with something along the following lines: "In most cases, errors are not estimated individually. Instead, their typical size is summarized by the square root of their mean squared value, or by the median of their absolute value. Such summaries do not preserve information about the signs of any individual errors."

Reply: We agree to reword this statement:

Action: New wording: In the case of 'error', its statistical estimate is mostly understood to be the square root of the variance of the probability density function of the error and thus does not carry any information about the sign of the error.

Comment: 154 *the term 'error' has commonly been used to signify a statistical estimate of the size of the difference between the measured and the true value of the measurand*
This is repetitive of the material around Line 10 that was discussed above. In both instances, the authors are unnecessarily turning something simple into something complicated. One thing is the error ϵ in the example discussed above, $m = \mu + \epsilon$. Another thing is how this error may be characterized or quantified.

Reply: Agreed; however, we prefer to leave this part and avoid the repetition elsewhere in the paper.

Action: The text has been restructured and shortened in order to make the arguments clearer and to avoid repetition. The new structure is intended to make our arguments w.r.t. terminological versus structural arguments clearer.

Comment: For example, the possible errors may be characterized by the probability distribution of ϵ, like when one says: the signal was corrupted by white noise with mean 0 and standard deviation σ.
The sizes of possible errors may be summarized by the mean squared error (MSE) of the estimator of the measurand, which captures the difference between expected value of the estimator and the true value of the measurand, as well as dispersion around that expected value. Other summaries include the standard deviation of the error distribution, or the now outmoded probable error.
The error may also be characterized indirectly, by an expression of the uncertainty surrounding the quantity of interest. For example, Yoshino et al. [1988] reported the measurement result for the absorption cross-section of ozone at 253.65 nm as $1145^{+7.1}_{-144}10^{-20}$ cm^2/molecule, which says that the measurement error has an asymmetric distribution.

Reply: Agreed. Most such cases in our context are caused by the non-linearity of Beer's law. Symmetrically distributed errors in the transmission measurements cause an asymmetric distribution of the inferred cross section errors. This highlights how important the non-linearity issue actually can be.

Action: Added: "Nonlinear error propagation may in some cases make asymmetric error estimates adequate."

Comment: 180 *Since the true value is not known, the actual difference between the measured or estimated value and the true value of the measurand cannot be calculated.*
The authors quite correctly point out that this argument lacks cogency. In fact, more can be said further to dismiss this claim as being no more than a myth. Consider the simplest of cases of statistical estimation, where one has replicated determinations of the same quantity, $r_1, \ldots, r_m$, which are then combined to obtain an estimate $t = T(r_1, \ldots, r_m)$ of a quantity τ. The estimate t could be as simple as the average or the median of the replicates, or it could be their coefficient of variation (standard deviation divided by the average). It is then possible, using the statistical jackknife [Mosteller and Tukey, 1977, Chapter 8] or the statistical bootstrap [Efron and Tibshirani, 1993], to estimate not only the standard deviation of t (based on this single set of replicates r_i), but also both the sign and the magnitude of the error $t - \tau$.

Reply: We are happy that we agree on this important point. With respect to the example presented in the review, however, GUM defenders would probably object that these methods provide a handle on the random part of the error but not on what they call "systematic" effects.

Action: None, for the sake of brevity.

Comment: 200 + 364 *our reading is that an error distribution is understood as a distribution whose spread is the estimated statistical error and whose expectation value is the true value, while an uncertainty distribution is understood as a distribution whose spread is the estimated uncertainty and whose expectation value is the measured or estimated value / The error distribution must not be conceived as a probability density distribution of a value to be the true value*
In the simple model for a measured mass, $m = \mu + \epsilon$, the "error distribution" generally refers to the probability distribution of ϵ, hence the expected value of the "error distribution" will not be μ, which denotes the true value of the measurand. Instead, this expected value will be the *bias*, which is the persistent offset of m from μ. (Refer to the discussion of Line 597, where I explain why I prefer "persistent" to "systematic," and "volatile" to "random.")
Neither "error distribution" nor "uncertainty distribution" are men-

tioned in the GUM. While the GUM offers considerable guidance about the assignment of distributions to input quantities, x_j, in its first 69 pages (out of a total of 120) all that it provides about the probability distribution of the output quantity, y, is an approximation to its standard deviation, in Equations (10) and (13).
Furthermore, the GUM seems to be more concerned with evaluating $u(y)$ than with estimating the measurand optimally, because the "substitution" estimate of the measurand, which is obtained by substituting the x_j by their best estimates in $y = f(x_1, \ldots, x_n)$, generally will not yield the best estimate of the measurand in the sense of minimizing mean squared error, mean absolute error, or other similar criteria [Possolo and Iyer, 2017, Page 011301-12].
In the course of those initial 69 pages, the GUM touches upon the topic of the distribution of y tangentially – for example when it discusses expanded uncertainty, coverage factors, and coverage interval –, but only in its Annex G (beginning on Page 70) does the GUM venture into a discussion of how to characterize the probability distribution of y. Annex G invokes the Central Limit Theorem based on a first-order Taylor approximation of the measurement function f in $y = f(x_1, \ldots, x_n)$, to claim that y's distribution may be taken as being approximately Gaussian. This argument can, on occasion, be spectacularly inaccurate [Possolo, 2015, Example E11].
Since $u(y)$ typically is based on finitely many degrees of freedom, the GUM argues (using a slightly different notation) that $(y - \nu)/u(y)$, where ν denotes y's true value, should have a Student's t distribution approximately, wherefrom coverage intervals then issue readily, thus achieving the goal, stated in its clause 0.5, of providing "an interval about the measurement result that may be expected to encompass a large fraction of the distribution of values that could reasonably be attributed to the quantity subject to measurement."
The meaning of this distribution that the GUM, by hook or by crook assigns to y, and, even more importantly, the meaning of the distributions derived for y by application of the Monte Carlo method, and of the coverage intervals based on them, should be the more appropriate and productive targets for critical review, similarly to what Stoudt et al. [2021] have done.

Reply: The example discussed by [Possolo and Iyer, 2017, Page 011301-12] inquires into a case where a higher level data product (here: the area of a rectancle) is calculated from the direct measurements of the lengths of the vertices. Possolo and Iyer demonstrate that the most probable area cannot be determined without knowledge of the error correlation of the length measurements of both vertices. Von Clarmann et al. (2020) concede that retrieved values of atmospheric state variables are not optimal as long as the measurement error covariance matrix includes only measurement noise but not the other error sources, mapped into the measurement space. The problem mentioned by

Possolo and Iyer falls in this category of problems. In the paper under review, we indeed forgot to consider this problem. But independent from this, our argument invoking the base rate still holds.
Gaussian error propagation (extended to consider covariances) holds, regardless of the error distributions of the ingoing quantities, as long as the function through which the errors are propagated is sufficiently linear. In particular, it is not required that the errors of the ingoing quantities follow a Gaussian distribution. Monte Carlo methods are needed either if the function is too nonlinear for Gaussian error propagation (we tackle this issue in old Section 5.3, new Section 3.3), if more information than only the expectation value and the variance shall be inferred (beyond the scope of our paper) or if the measurement error covariance matrix used for the inversion does not include all error components. In the latter case MC methods can be used to correct the result, but that is somewhat beyond the issue of error estimation. However, MC methods do not solve the problem of the base rate fallacy.

Action:. Added "[...yield the probability distribution of any value to be the true value.] This holds even if the error distribution is extended to include also systematic effects, and if all error correlations are adequately taken into account in the case of multi-dimensional measurements."

Comment: 213 *Frequentist statistics, we understand, is a concept where the term 'probability' is defined via the limit of frequencies for a sample size approaching infinity. This definition is untenable because it involves a circularity*
The authors oversimplify and are unacceptably dismissive. If the Frequentist interpretation of probability were this "obviously" defective, then none of John Venn, Richard von Mises, Andrey Nikolaevich Kolmogorov, Jerzy Neyman or Jack Kiefer – all intellectual giants in their own right – would have embraced it.
I suggest that the authors avoid embarrassment by considering the excellent overview of the interpretations of probability compiled by Hájek [2007].

Reply: If this argument was conclusive, then we should either believe in the ether theory or we should deny Lorentz and Poincaré the status of an intellectual giant. Scientific knowledge is approximately accumulative, and thus it is not astonishing that later generations know more than earlier generations. Stegmüller's argument runs as follows: Hypothetical frequentism (i.e. frequentism involving an infinite reference class; this seems necessary to be able to extend the probability concept to probability density functions of continuous random variables) rely on the large number theorem. Both the weak and the strong version of the large number theorem involve probabilities. Thus the definition is circular or involves an infinite regress. We do not see any flaw in his argument, and this argument is refuted neither by Hájek [2007] nor by any other literature of our knowledge.
It is important to note that the distinction between frequentism and other con-

ceptions of probability such as subjective probability is about how the term 'probability' is defined, and not about the methods used. Methods still can work well, even if the definition of the underlying key term is flawed.

Action: As a compromise, we have replaced "is untenable" with "challenged".

Comment: 265 *The values the rational agent believes to be true are sufficient in this case, because the error distribution does not tell us anything about the truth anyway but only about the agent's believe of what truth is.*
The simplest measurement error model mentioned above, $m = \mu + \epsilon$, is meaningful under essentially all paradigms of statistical inference. In neither the classical (Frequentist) nor in the Bayesian approaches does the probability distribution of ϵ convey any information about μ, other than in special cases: for example, when the variance of ϵ depends on μ.
Both approaches involve assigning a probability distribution to ϵ, which then determines the likelihood function. The Bayesian approach involves also assignments of probability distributions to μ and to any parameters in the distribution of ϵ whose values are unknown. The marginal distribution of m typically will differ in the classical and Bayesian approaches even when the same choice is made for the distribution of ϵ.

Reply: Agreed for the mathematical part but we are afraid that our original text did not make clear the point we intended to make. Thus we reword our statement for clarity.

Action: Reworded: "In GUM the error concept is discarded because the capability of conducting an error estimate allegedly depends on the knowledge of the true value. However, once having invoked the concept of subjective probability, no objective knowledge of the unknowable true value is needed any longer. The subjectivist can work with the value they believes to be true. This solves the alleged problem of the error concept, namely, that the true value is unknown."

Comment: 317 *Monte Carlo uncertainty estimation, however, is in its heart a frequentist method, because it estimates the uncertainty from the frequency distribution of the Monte Carlo samples.*
The authors are quite wrong on this one.
Of course, the extent of how wrong depends on what they mean by "Monte Carlo uncertainty estimation." I assume that they mean it in the sense and context in which it was introduced into uncertainty analysis by Morgan and Henrion [1992], subsequently having been incorporated into the GUM Supplement 1 [Joint Committee for Guides in Metrology, 2008]. In such sense and context, the Monte Carlo method is purely mathematical, and non-denominational (neither Frequentist nor Bayesian), and solves the following problem:

given a random vector X whose probability distribution has been fully specified, and a real-valued, measurable function f defined on the range of X, determine the probability distribution of $Y = f(X)$. The Monte Carlo method solves this problem using numerical methods and sampling driven by pseudo-random numbers. It solves it in the sense that it can produce the value of $Pr(Y \in B)$ to within any specified accuracy, for any measurable subset B in the range of Y . The fact that its accuracy is guaranteed by the Law of Large Numbers does not make it Frequentist because the Law of Large Numbers is neither Frequentist nor Bayesian. The Law of Large Numbers is a mathematical result about sums of random variables based on Kolmogorov's axioms for probability measures [Kolmogorov, 1933]. If the authors' views on the Monte Carlo method were correct, then Markov Chain Monte Carlo sampling, which is the workhorse of contemporary Bayesian inference, would be "in its heart a frequentist method" too!

Reply: We still think that, by employing samples, MC methods do use the frequentist toolbox. A Monte Carlo sample is a sample where values considered as more probable occur more frequently. But since we think that the use of the frequentist toolbox does not make a subjectivist a frequentist, and since this statement is not needed to support our stance, we decided to withdraw our argument on MC methods.

Action: Argument deleted.

Comment: 318 *it is astonishing why GUM08, if representing a Bayesian concept, does not in the first place require to apply the Bayes theorem*
The authors should reference Gleser [1998] who points out the mixed-bag of viewpoints coexisting in the GUM. Clearly the authors are well entitled to feel astonishment at the GUM not using Bayes rule at all, especially considering the whirlwind of claims about the GUM and its Supplements being Bayesian.
However, in fairness to the GUM, such whirlwind has been more of an afterthought than a consequence of the GUM itself. First, the word "Bayes" is nowhere to be found in the GUM, and the word "Bayesian" occurs exactly once: n the title of reference [14], on Page 115.
Only in Annex E (E.3.5) does the GUM venture into this controversial territory when it says "In contrast to this frequency-based point of view of probability, an equally valid viewpoint is that probability is a measure of the degree of belief that an event will occur." And then it adds: "Recommendation INC-1 (1980) upon which this Guide rests implicitly adopts such a viewpoint of probability."
The expression "degree of belief" occurs exactly once in the main body of the GUM (3.3.5), where it says: "Thus a Type A standard

uncertainty is obtained from a probability density function (C.2.5) derived from an observed frequency distribution (C.2.18), while a Type B standard uncertainty is obtained from an assumed probability density function based on the degree of belief that an event will occur [often called subjective probabil- ity (C.2.1)]. Both approaches employ recognized interpretations of probability.
The same expression occurs in Annex C, and again in Annex E, where E.3.6 comes the closest to advocacy by enumerating "three distinct advantages to adopting an interpretation of probability based on degree of belief." Therefore, and on the whole, the GUM is far more discreetly or ambiguously Bayesian than it has more recently been heralded to be (surprisingly, mostly by "born again," self-declared Bayesians).
The GUM's alleged Bayesianism in fact reduces to (i) entertaining (subjective) probability distributions for input quantities that are elicited from experts, and (ii) regarding the probability distribution of the measurand as quantification of degrees of belief about the true value of the measurand, even though it is not a Bayesian posterior distribution [Gleser, 1998, 2.2].

Reply: We do not claim that GUM is Bayesian. However, we recognize that there exist readings of GUM that see it as Bayesian. We do not say that we **are** astonished that the Bayes theorem plays no role in GUM but we say we should be astonished that the Bayes theorem plays no role in GUM if GUM really was Bayesian. Our argument follows the form of a reduction to the absurd: As a working hypothesis we assume that GUM is Bayesian, we then find that this assumptions lead to inconsistencies, and finally we conclude that the purported Bayesian turn cannot explain the difference between the error concept and the uncertainty concept.

Action: We have reworded the related text to make the argument clearer, and to make it more obvious that our argument form is a *reductio ad absurdum*.

Comment: 343 *This suggests that the uncertainty is an attribute of the true value while the error is associated with a measurement or an estimate. Because of the measurement error there is an uncertainty as to what the true value is. The uncertainty thus describes the degree of ignorance about the true value while the estimated error describes to which degree the measurement is thought to deviate from the true value*
The authors are quite right. Please consider the following rewrite, which, although allegorical, I believe further enhances the expression of the authors' sentiment – also compare with Possolo [2015, Note 3.2, Page 16]: This suggests that measurement uncertainty surrounds the true value of the measurand like a fog that obfuscates it, while measurement error is both the source of that fog and part and parcel of the measured value. Measurement uncertainty thus describes the

doubt about the true value of the measurand, while measurement error quantifies the extent to which the measured value deviates from the true value.

Reply: Agreed.

Action: Footnote added.

Comment: 379 *The weight of Thomas Bayes or the body height of David Hume at a certain time are well-defined quantities although we have no chance to measure them today*
I suggest that, for the sake of propriety and good taste, the authors abstain from referring to properties of the bodies of Thomas Bayes and David Hume, refined and excellent gentlemen both, long deceased, and use instead properties of other notable material entities that are no longer amenable to measurement, like the Colossus of Rhodes or the Lighthouse of Alexandria.

Reply: Agreed.

Action: Changed as suggested; footnote added with reference to the originator of this idea for this illustrative example.

Comment: 411 *5.2 Likelihood, probability, and the base rate fallacy*
I believe that this subsection is a digression from the main topic that would best be deleted. A shorter, better focused article will have greater impact than one with multiple digressions that are largely off-topic.

Reply: Since we conceive measurements (and their analysis) as estimation of the true value involving an inverse process, we think that this argument is quite essential. We do, however, agree that this section was too long, and that the original structure of the paper did not make our argument sufficiently clear.

Action: This concern has been considered when the manuscript was reorganized.

Comment: 481 *5.3 Nonlinearity issues*
The same suggestion as for subsection 5.2, for the same reasons.

Reply: The denial of the approximate knowledge of the true value poses serious problems particular in the case of non-linearity. In linear cases the statistical error can be estimated without knowledge of the true value because the Jacobian does not depend on the measurand. This is not the case if there is non-linearity. We do not think that this is a disgression or off-topic, but a serious challenge to an uncertainty concept that pretends to be able to avoid the concept of the

true value.

Comment: 518 + 738 *5.4 Incompleteness of the error budget*
This is an important issue that the authors should address in greater generality than in the context of inverse problems. The following example captures the key issues clearly and simply. The authors allude to the same ideas in Line 738.
The values measured in inter-laboratory studies are often modeled as $m_j = \mu + \lambda_j + \epsilon_j$ for $j = 1, \ldots, n$, where μ denotes the true value of the quantity of interest, and the λ_j and the ϵ_j are errors of different kinds: the former express laboratory effects [Toman and Possolo, 2009a,b, 2010], which in many cases will be persistent effects attributable to differences between measurement methods or between forms of calibration; the latter are laboratory-specific measurement errors quantified in the uncertainties reported by the participants.
The reality of the λ_j (that is, that they cannot all be zero) becomes apparent only when the measurement results are put on the table and inter-compared.
If the measured values are significantly more dispersed than the associated, reported uncertainties intimate that they should be, then this is an indication that there is some dark uncertainty [Thompson and Ellison, 2011] afoot that was not captured in the individual uncertainty budgets.
This dark uncertainty is "carried" (in the sense in which this term was used in the General Comments) by the λ_j. Refer to Koepke et al. [2017] and to Possolo et al. [2021] for more extended discussions of this concept.

Reply: The rationale behind limiting our discussion to inverse problems is that this manuscript has been written for an AMT special issue that deals exclusively with error reporting in remote sensing contexts. There are certainly many people who are better qualified than we are to discuss these issues in the context of general metrology. We see it as our task to scrutinize the applicability and usefulness of the BIPM guidelines to remote sensing of the atmosphere.
By the way: Those passages in our manuscript where the incompleteness of the error budget is discussed do not refer explicitly to inverse problems.

Comment: 548 *We have mentioned above that the uncertainty concept depends on the acceptance of the subjective probability in the sense of degree of rational belief. Without that, an error budget including systematic effects would make no sense because systematic effects cannot easily be conceived as probabilistic in a frequentist sense; that is to say, the resulting error cannot be conceived as a random variable in a frequentist sense.*
These statements are inaccurate.
First, the uncertainty concept may be contingent on a Bayesian perspective, but this perspective need not be subjective: it can be

a so-called "objective Bayesian" perspective, which Jeffreys [1946], Bernardo [1979], and Berger [2006], among others, have favored.

Reply: The uncertainty concept explicitly invokes subjective probability (Sect 3.3.5). We think that the uncertainty concept is contingent on the concept of subjective probability (i.e., a degree of belief concept of probability) but not necessarily on Bayesianism. While objective probability (both frequentist probability and Popper's propensity) is a characteristic of the event (the object), subjective probability is a characteristic of the knowledge of the agent (the subject) dealing with this event. Any concept of probability that conceives errors, uncertainties etc as a degree of ignorance and characterize them with a degree of belief are thus based on subjective probability, because the ignorance is a characteristic of the agent (the subject), not of the value (object).
Thus we think that even objective Bayesians employ the concept of subjective probability. That is to say, also objective Bayesianism depends on a concept of probability which describes the information (and its uncertainty) an agent (the 'subject') has. Thus one could argue that the term "objective Bayesianism" is a misnomer, or at least is misleading.
Our use of the terms 'subjective probability' and our understanding of 'objective Bayesianism' seem to be fully consistent with D. R. White (2016 Metrologia 53 S107) who states: "There are two main branches of Bayesian statistics, objective and subjective, both of which are founded on three main principles: the use of subjective probability, the use of Bayes' theorem to invert conditional probabilities, and the likelihood principle" (his Section 3.2). Thus, the contradiction between subjective probability and objective Bayesianism is only apparent but not real. Anyway, since GUM invokes subjective probability but not Bayesianism, the discussion of objective Bayesianism is not relevant to the paper.

Action: (Only indirectly linked to this comment) We have deleted the section on Bayesianism versus non-Bayesianism.

Comment: Second, the main difficulty facing a Frequentist approach to the characterization of measurement uncertainty concerns what the GUM calls Type B evaluations of uncertainty components, not the recognition of the contributions that persistent ("systematic") effects make to said uncertainty.

Reply: We must distinguish between methods that use frequency distributions as estimators of probability distributions and the way how the concept of 'probability' is defined. Type B evaluations work well even for frequency distributions. Take a finite sample of parameters b_i, calculate the variance, propagate the variance through the system using linear theory; and in parallel, propagate each parameter b_i through the system and calculate the variance in the result space. If the system is sufficiently linear, both variances will be approximately the same (and exactly the same in the linear case). We see no difficulty to apply Type B evaluations to frequency distributions.

The problem solved by GUM by abolishing frequentism and turning towards a subjective concept of probability is that systematic effects cannot be characterized by a probability distribution at all. In a frequentist's world it would simply make no sense to assign an error distribution to a systematic effect, because the systematic effect is only one number. To assign an error distribution to a systematic effect requires to conceive the error distribution as a characterization of the knowledge or belief of the rational agent instead of a frequency of events. And this is exactly what the subjective concept of probability does. For a frequentist it is absurd to assign a distribution to a systematic error.

Comment: In fact, the contributions from some persistent effects can be evaluated by Type A methods (refer to the comments above for line 180), and the contributions from some volatile ("random") effects can be evaluated by Type B methods (for example, the imprecision of a balance that a laboratory technician has great familiarity with).

Reply: We agree that there is no clear correspondence between Type A vs. Type B evaluation of uncertainty on the one hand and random vs. systematic error components on the other hand. TUNER recommends to evaluate all error components via Type B analysis. This seems necessary to disentangle the different error components (i.e. error from the different sources). For example, the error due to measurement noise is estimated by propagating known measurement noise through the retrieval. The same holds for, e.g., volatile parameter uncertainties. Type A evaluation is recommended to test the validity of the Type B estimates.

Comment: 597 *Von Clarmann et al. (2020) explicitly demand that error estimates be classified as random or systematic [. . .] In summary, the denial of the importance of distinguishing between random errors and systematic errors does not provide proper guidance, and altogether is a strong misjudgment.*
The word "demand" appears to be too strong a descriptor of what von Clarmann et al. [2020] actually did, which was to "formulate *recommendations* with respect to the evaluation and reporting of random errors, systematic errors, and further diagnostic data," where the emphasis on "recommendations" is mine.
We need to discuss two separate issues regarding this point: the first concerns the choice of terms ("systematic" and "random"); the second concerns whether and when to bundle them all into a single expression of uncertainty.
Concerning the first issue – the choice of terms:
My dislike of terms like "systematic" and "random" is that they are metaphysical:
they speak to the nature of the errors, which is often elusive and may be shifting.
For example, von Clarmann et al. [2020, R3, Page 4420] recognize that "depending on the application of the data, the same type of er-

ror can act as random or systematic error," and many other authors have acknowledged the same.
"Random," in particular, is a thorny concept, whose definition seems to be far from settled [Landsman, 2020] [Eagle, 2016] [Bennett, 2011] [Gács, 2005].
For these reasons, I recommend descriptive qualifiers instead, for example persistent (instead of "systematic") and volatile (instead of "random"). They are less committal and afford greater flexibility, in particular to address cases where a volatile error becomes persistent, or vice versa. Writing almost thirty-five years ago, Collé [1987a], summarized the two approaches to measurement uncertainty that were then dominant as follows: The "classical" approach is based on a central distinction between so-called random and systematic uncertainties. The uncertainties are presumably classified by the underlying physical error type [...] and the approach demands that the different uncertainty types be combined by different methods. Causing even further confusion, the uncertainties in these classical treatments are said to depend on one's "perspective" and hey possess chameleon-like properties, and may change from one type to another. In contrast, the "romantic" approach dispenses with the underlying error distinction, and classifies the uncertainties only on the basis of how the uncertainty estimates were made. All uncertainty components in this approach can be combined by the same general propagation formulae. The romantic approach underlies the BIPM/CIPM Recommendation.

Reply: We agree to replace "demand" with "recommend". We are reluctant to change terminology in favor of 'volatile' and 'persistent' errors because this would lead to inconsistent terminology within the AMT special issue this paper is written for. However the connotation of our terms 'random errors' and 'systematic' errors, as defined in von Clarmann et al. (2020) matches that of the 'volatile' and 'persistent' errors of Possolo. There, systematic errors are defined as bias-generating errors while random errors are defined as variance-generating errors. They can be distinguished fully on observational grounds.

Action: "demand" replaced with "recommend". Further we have added the following footnote: "In this context it is important to note that, in contrast to some older conceptions, von Clarmann et al. (2020) define 'systematic errors' as bias-generating errors and 'random errors' as variance-generating errors. To avoid confusion with the older conceptions, one can use instead the descriptive terms 'persistent' and 'volatile' errors as suggested by Possolo (2021). This is not done here to maintain consistency with von Clarmann et al. (2020).

Comment: Concerning the second issue – the bundling of contributions from all sources of uncertainty:
While agreeing with the romantic approach in principle, I believe that

it is advisable to consider how uncertainty evaluations will be used, before deciding whether to combine contributions from all sources of uncertainty into a single evaluation, or not. This is a more nuanced, less extreme approach than either of the two approaches aforementioned.
Consider an inter-laboratory study where several laboratories measure the same quantity independently of one another, or a meta-analysis of results from preexisting studies that were carried out and published independently of one another.
Suppose that the purpose is to blend the corresponding estimates into a consensus value: for example, as was done for the ozone absorption cross-section at 253.65 nm [Hodges et al., 2019].
Typically, the consensus value will be some form of weighted average. There- fore, the errors behind the uncertainties reported by the participants will "average out" in the process to some extent. This may be fine, or it may be inappropriate. Such "averaging out" will be fine if laboratory-specific persistent errors lead to estimates that are high for some laboratories and low for other laboratories, with the true value lying somewhere in the middle. But such "averaging out" will be inappropriate if a common bias, unbeknownst to all, affects all results similarly. Reporting separately the evaluation of the contributions made by persistent effects, and by volatile effects, as is commonly done in astrophysics and in particle physics, will then be an appropriate, prudent way to report uncertainty intended for use by a downstream user.
The need for such discretion, and the role that considerations of fitness-for-purpose of uncertainty evaluations should play in deciding what to do and when, is mentioned already in the pre-GUM literature [Ku, 1980].

Reply: We fully agree that the choice if errors of different type shall be bundled or not depends on the application. However, we still uphold our criticism that the denial of the importance of distinguishing between random errors and systematic errors does not provide proper guidance, and altogether is a strong misjudgment. The data provider does, in general, not know what the data user will do with the data, and since the data user may use the data for a purpose where it is essential to distinguish between systematic and random errors, the information which error component contributes to which category must be provided.

Action: Added: "[...is a strong misjudgment]. The data users must be provided with all information required to taylor the relevant error budget to the given application of the data."

Comment: 618 *6.3 The causal arrow – [. . .] We think that it is essential to appreciate the inverse nature of the problem, and this is much easier if the mea-*

surement equation describes the forward problem and thus does not suggest an unambiguous determination of the measurand from the measured quantity.
The measurement model in the GUM is only one of many kinds of measurement models to which the principles for uncertainty evaluation that are enunciated in the GUM apply. The GUM-6 [Joint Committee for Guides in Metrology, 2020], published recently, describes several other kinds of measurement models, including statistical measurement models.
Rodgers [2000, 2.3.2] explains how Bayesian statistical models can be used in general to solve inverse problems, and Ganesan et al. [2014] describe an application of hierarchical Bayesian methods to atmospheric trace gas inversions.
The Bayesian approach can be fruitful in such settings because the prior distribution acts as a regularization prescription.
Possolo [2015] gives examples of measurements involving models that are quite different from the conventional measurement model in the GUM. In particular, Examples E7 (Thermistor Calibration), E17 (Gas Analysis), and E32 (Load Cell Calibration) concern calibrations that are structurally similar to the thermometer example that the authors mention in Line 109.
Using $x_1, \ldots, x_n$ and y with the same roles that the GUM gives them, a statistical forward model can be formulated simply by saying $x_1, \ldots, x_n \sim L_y$, which is shorthand for "the joint probability distribution of (the random variables whose realized values are) the observable inputs $x_1, \ldots, x_n$ has y as a parameter and likelihood function L_y.
A Bayesian formulation will then add $y \sim P$, where P is the prior distribution of y, and application of Bayes's rule produces a solution for the inverse problem in the form of the posterior distribution, Q, of $y \sim Qx_1, \ldots, x_n$. Compare this formulation with the treatment of calibration via conventional regularization in Hagwood [1992].

Reply: Our point is that for a given measurand and a given realization of the measurement error, the measured signal is, putting quantum effects aside, which can be embraced by the measurement error, unambiguously determined. Conversely, for a given measured signal, the most likely or most probable value of the measurand is **not** unambiguously determined. Since traditionally a mathematical function (which is the theoretical core of the measurement model) maps an independent variable x unambiguously to a dependent variable y, while the opposite direction may be ambiguous, we find the GUM (2008) notation counterintuitive.

Action: We have inserted: "Many conceptions exist of measurement models, which relate the measured value to the true value, and depending on the context, one can be more adequate than another (Possolo, 2015). GUM recommends a model that conceives the estimate of the true value of the measurand as a function of the measured value. Since in remote sensing of the atmosphere

multiple atmospheric states can cause the same set of measurements, and the measurement function thus would be ambiguous, we prefer a different concept, as outlined in the following. Further, we have deleted the section on Bayesianism.

Comment: 662 *paradoxes shatter the bedrocks of Bayesian philosophy, namely the likelihood principle that says that all relevant evidence about an unknown quantity obtained from an experiment is contained in the likelihood. Others accept the theoretical validity of the Bayes theorem but challenge its applicability in real life because of the unknown and unknowable prior probabilities.*
The paradoxes alluded to often relate more to the adoption of so-called "non- informative" prior distributions than to the acceptance of the likelihood principle, as Cox [2006] points out, in a contribution referenced by White [2016].
All theories of inference have given rise to paradoxes, and nevertheless most often they produce valid and practically useful inferences. Regarding the likelihood principle in particular, at least one well-known "paradox" has been dismissed as a false alarm [Goldstein and Howard, 1991].
In any case, White [2016] does not come even close to suggesting that such paradoxes "shatter the bedrocks of Bayesian philosophy," in particular as applied in measurement science. I know for a fact that Rod White does not object to the use of Bayesian methods when these are warranted and there is genuine prior information that should be taken into account.
The objection, which is also raised by Bayesians [O'Hagan, 2006], is to the systematic reliance on "non-informative" prior distributions just for the sake of going through the motions of the Bayesian machinery or to pay lip service to scientific objectivity.
The Bayesian approach to problems of statistical inference is a choice among many that can be made, similarly to how some people choose to drink lemonade and others bourbon. Different approaches to statistical inference (be they frequentist, fiducial, or Bayesian) all can claim notable successes in solving problems of practical importance.
Bayesian methods, in particular, can boast a long and varied roster of accomplishments that prove beyond reasonable doubt that they are applicable in real life, and that they can be used to solve important practical problems, and that often they do so better than non-Bayesian alternatives [O'Hagan, 2008].
A particularly striking, recent accomplishment of Bayesian methods concerns the use of measurements of $\Delta 14$ CO_2 , in conjunction with atmospheric transport models, to demonstrate that several bottom-up approaches to the estimation of national inventories likely underestimate U.S. fossil fuel CO_2 emissions [Basu et al., 2020].
This study, which is based on methodological advances published in this very journal [Basu et al., 2016], includes rigorous, model-based

uncertainty evaluations, and also serves to show that the GUM and its supplements have much catching-up to do if they will ever come to play a role in addressing momentous issues like the measurement of greenhouse gas emissions. The suggestion that Bayesian methods are questionable because prior distributions are "unknown and unknowable" reveals a misconception about prior distributions: they are meant to encapsulate the knowledge that someone has about the quantity of interest, prior to performing an experiment that generates fresh information about it. Therefore, proper, informative, subjective prior distributions are known to who formulates them, by construction.
Of course, the Bayesian can be much mistaken and construct a prior distribution that reflects an erroneous conception of reality, in which case the "knowledge" that the prior encapsulates is false knowledge and its use will lead the inference astray. However, Bayesian methods cannot be blamed for delusions any more than Newton's laws can be blamed for accidental falls.

Reply: We do not deny the benefits provided by the Bayesian toolbox but we reported related critical issues. The aim was to provide an argument why we consider it as inadequate to commit the community to a full-blown Bayesian philosophy. The context of this argument are the question (a) if the "Bayesian turn" can explain the alleged difference between the error concept and the uncertainty concept, and (b), if so, if it is adequate to anchor Bayesianism in a normative document. Our reply to both these questions is negative. However, since the current version of GUM is not fully Bayesian, we now consider the section on Bayesianism as irrelevant for the title topic.

Action: The section on Bayesianism has been deleted.

Comment: 674 *the Bayesian philosophy relies on a couple of unwarranted assumptions, e.g., the likelihood principle and the indifference principle.*
The authors convey a wrong impression on both counts. Adherence to the likelihood principle is a choice that, in most applications, turns out to be a better choice than most alternatives. Still, it is only a choice, among many that can be made. Making such choice is necessary but not sufficient to be Bayesian. Many statisticians, physicists, chemists, and biologists adhering to the likelihood principle are not Bayesian.
Neither is adopting an indifference principle (or, more generally, using an allegedly non-informative prior distribution) necessary to qualify as being Bayesian.
In fact, quite the contrary is true: reliance on proper, informative, and suitably elicited subjective prior distributions, are the hallmarks of genuine Bayesian practice. But this, too, is only a choice [Robert, 2007].

Reply: We did not want in this paper to argue about which to positions make a user of the Bayesian toolbox a "real Bayesian" and which of the unwarranted assumptions belong to the bedrocks of Bayesian philosophy. We also did not want to judge if Bayesianism is good or bad. Instead, our point was that a philosophy that is challenged by a non-negligible part of the community shall not be made generally binding. However, since GUM itself (contrary to some of its readings) does not promote Bayesianism, we now consider this issue as irrelevant for the paper.

Action: As said above, the section on Bayesianism has been deleted.

Comment: Technical Corrections
025 Replace comes down to the question if and how with "comes down to the question of whether, and if so how"

Reply: Agreed.

Action: Changed as suggested.

Comment: 056 Replace Second we assess to which degree with "Second, we assess the degree to which"

Reply: Agreed.

Action: Corrected.

Comment: 068 Replace *we conclude to which degree* **with "we conclude the degree to which"**

Reply: Agreed.

Action: Corrected as suggested. In the same spirit we have changed "to which degree the measurement is thought..." to "the degree to which the measurement is thought..."

Comment: 128 Replace A rich methodical toolbox with "A rich methodological toolbox"

Reply: Agreed

Action: Corrected.

Comment: 180 Replace *This argument is often used to dispraise* **with "This argument is often used to disparage"**

Reply: Agreed but no longer relevant.

Action: None. because in the shortened manuscript this statement does no longer appear.

Comment: 267 Replace *agent's believe* **with "agent's belief"**

Reply: Thanks for spotting!

Action: Corrected.

Comment: 364 Replace *Quantities of which the value cannot determined* **with "Quantities whose values cannot be determined." This suggestion deliberately ignores the antiquated invective against using the possessive whose for inanimate objects, consistently with the recommendation in O'Conner [2019, Page 243].**

Reply: Agreed; the Copernicus editorial team usually does a great job with respect to language issues, and we trust that they will remove any language-related inconsistencies and outdatedness.

Action: Changed as suggested.

Comment: 373 Replace *Others have been formulated by us, serving, as arguments of the Devil's advocate, as working hypotheses in order to moot the error and uncertainty concepts in the context of indirect measurements* **with "We have formulated others as Devil's advocates, which are intended to serve as working hypotheses to MOOT the error and uncertainty concepts in the context of indirect measurements," except that "moot" needs to be replaced by a word that is suitable for this passage: maybe "merge" or "reconcile", depending on what the authors wish to express.**

Reply: Agreed.

Action: The new sentence now reads: "We have formulated others as Devil's advocates, which are intended to serve as working hypotheses to critically discuss the error and uncertainty concepts in the context of indirect measurements."

Comment: 413 Replace *measurements are not in the focus* **with either "measurements are not in focus" or "measurements are not the focus," depending on what the authors wish to say exactly.**

Reply: Agreed, but obsolete.

Action: None, because this sentence does no longer appear in the revised version of the manuscript.

Comment: 420 Replace *the probability that a person suffering fever to have Covid-19 is 50%* **with "the probability is 50% that a person with fever has COVID-19"**

Reply: Agreed.

Action: Corrected as suggested.

Comment: 429 Replace *distribution which is missing* **with "distribution that is missing"**

Reply: Agreed but obsolete.

Action: None, because this sentence does no longer appear in the revised version of the manuscript.

Comment: 470 Replace *the aggregation of random uncertainties* **with "the aggregation of random uncertainties"**

Reply: Agreed, thanks for spotting.

Action: corrected as suggested.

Comment: 612 Replace *strong misjudgement* **with "strong misjudgment" (unless the British spelling be preferred)**

Reply: Agreed.

Action: Corrected: "misjudgment".

Comment: 743 The sentence that includes *traditional error analysis can connotate a statistical quantity* **is unclear, and should be rewritten, taking into account the fact that the verb connotate is obsolete and has been replaced by connote. However, a term more generally familiar would be preferable, like suggest, possibly.**

Reply: Agreed.

Action: "[...] that the traditional error analysis can **deal with** a statistical quantity, and that [...]"