Review #3:

Comment: Manuscript title: Truth and Uncertainty. A critical discussion of the error concept versus the uncertainty concept
In this manuscript the authors present an argument that the the "error concept" and the framework put forth by the GUM (deemed the "uncertainty concept") are the same. The major issue I understand that the authors take with the GUM is in the recommendation that uncertainties reported with estimates of measurement need not be specifically interpreted with respect to errors and "true values." The authors refute this claim by seemingly arguing that uncertainties or "estimated errors" cannot be interpreted without reference to true values and therefore the concepts must be the same. To be honest, it was extremely difficult to parse through the unnecessarily lengthy 23 pages of text to come to the understanding that this is (I believe) the authors main argument, and critically I do not believe this argument is effectively made. In general the manuscript is too long with repetitive sections that are often confusing and in some places contradictory.

Reply: We agree that the structure of the manuscript was not sufficiently clear and contained unnecessary repetition.

Action: We have restructured and partly rewritten the paper in order to make the arguments clearer, to avoid repetition, and to shorten the manuscript.

Comment: The message is often lost in unnecessary language arguments between the GUM and the authors' definition of 'error' and in generally narrow and misconceived discussions about frequentist vs Bayesian statistical methods.

Reply: We agree that language arguments and conceptual arguments should not be merged. For the frequentist vs. Bayesian discussion, see below.

Action: The manuscript has been restructured. Terminological issues and conceptual issues are now discussed separately.

Comment: More seriously, the language used in reference to statistical concepts is imprecise and in some areas completely incorrect. The authors should define their terms with equations where applicable and adhere to commonly accepted mathematical/statistical/probabilistic definitions. In several places the statistical interpretations of their "error estimates" in relation to "true" values are overly simplified and likely to be misinterpreted, in particular when models are misspecified.

Reply: Often the point is not how quantities are mathematically specified but how the involved terms are interpreted, i.e., to which entities in te real world the quantities refer. For example, the variance of a frequentist pdf and the variance of a pdf representing personal belief are calculated according to exactly the same formalism, but they connote an entirely different thing.

Action: During rewriting, we have taken care to use an unambiguous language.

Comment: Rather than focusing on an argument that the uncertainties typically reported under the "error concept" can be also be interpreted as under the "uncertainty concept", they seem to miss the point of GUM (as I interpret GUM, but I would also argue more broadly the understanding of these concepts in the field of statistics) that reported uncertainties need not come with inferential statements about how close estimates are to the true value (e.g. actual errors) to be useful for comparison to other estimates.

Reply: If the purpose of uncertainty reporting is not to provide an idea how close the estimate is to the true value, then we do not understand what the purpose of uncertainty reporting is. And even worse: Then we need a different concept that provides this. An estimate how close the estimate is to the true value is exactly what the data user needs.

Comment: Instead the focus seems to be mainly a language argument that "true value" is the same as "value of the measurand" and so the concepts must be the same – a not particularly useful argument in my opinion.

Reply: We disagree. The bulk of our argument is conceptual, not terminological.

Action: In the revised version, we separate language arguments and conceptual arguments. The terminological part is only a small fraction of the paper.

Comment: Without considerable revision and restructuring I do not believe the manuscript provides a useful contribution to AMT. In fact, I am concerned that publication in its present form would propagate dangerous misconceptions about statistical methods and uncertainty quantification to the community.

Reply: We agree that the original structure did not optimally support our argument.

Action: The paper has been restructured and partly rewritten.

Comment: In addition, the authors do not provide a concise, understandable overview of the "error" and "uncertainty concepts" and

the supposed differences, which narrows the manuscript's audience to those who are already well-versed in the GUM and the specific error analysis framework the authors consider.

Reply: It is exactly the main problem of GUM that they state that the error concept and the uncertainty concept are different but fail to clearly specify in what the difference consists. This is what we criticize. Since GUM is vague with respect to what the difference is, we posit working hypotheses what the difference might be.

Action: In the revised version we hope to have made our argument form – to posit working hypotheses and to refute them – clearer.

Comment: I do agree with the authors that the quantitative methods laid out under the GUM framework are not inconsistent with the traditional error analysis framework and, if properly understood, the interpretations of such quantities under both frameworks are generally in agreement. It is my belief that a useful manuscript would argue these points very concisely, showing that the recommendation in GUM are not inconsistent with traditional methods in the atmospheric remote sensing community, and would focus more attention on addressing how the GUM principles apply to atmospheric retrievals and where GUM may fall short.

Reply: Our conclusion is that we do not see relevant differences between the error concept and the uncertainty concept. Thus, the agreement of the interpretations follows a fortiori. Problem areas where GUM falls short were already discussed in Section 6 of the original manuscript.

Action: The Section on the applicability of GUM-2008 to remote sensing of the atmosphere (Section 4 in the revised version) has been partly rewritten, and has become shorter and more focused in the course of restructuring the paper.

Comment: General comment about "true values" and uncertainties The authors need to clearly state what they mean by "true value" in their arguments. Specifically, when discussing true values are they referring to the truth in terms of reality (if such a quantity exists)...

Reply: Of course we cannot prove the existence of an external truth. However, the purpose of the entire endeavor of taking measurements is simply to create a link between our mind and the external reality. If we deny the existence of an external reality, then we need no measurements. Since the journal AMT has the term "measurement" in its title, we think that the existence of an external reality is uncontroversial among the AMT readership. In the spirit of reviewer #2, who sees the risk that the paper drifts too far towards philosophy, we prefer not to dwell on this issue in the paper[1].

Comment: ... or the true value in terms of the specified statistical model and resulting theory? The latter are the only “true values” that have any statistical guarantees in the interpretation of uncertainty estimates and are only equivalent to the true value in reality if the statistical model perfectly describes the true data generating process (i.e. is the “correct” model), which we know is unlikely to be the case particularly in atmospheric remote sensing retrievals.

Reply: Among all concepts of truth (logical truth, analytical truth, factual truth, see R. Carnap, New York, 1966 ”Philosophical foundations of physics”, for a deeper discussion on this), only the latter, the factual truth, is relevant. Among the theories of factual truth, the correspondence theory is the most intuitive one, and is probably accepted by most empirical scientists. Correspondence theory follows largely the pattern ”The statement ‘snow is white’ is true if and only if snow is white”. However, we see no conflict of our paper with the coherence theory of truth. In order to avoid disgressions and to drift away too far towards philosophical aspects not relevant to the paper, we prefer not to include in the paper the discussion of different theories of truth.
We think that in the context of measurements it is sufficiently clear that the true value refers to reality. We are not aware that in the context of measurements the term “true value” is used for anything else. Values resulting from a specified statistical model and resulting theory are not usually called true values. The fact that models usually do not fully represent reality is exactly the reason why in the TUNER project the need of the consideration of model errors is highlighted. Since in our paper the true value bears the attribute “unknowable”, it should be clear what the connotation of “true value” is in our paper. The attribute “unknowable” would simply make no sense for any other connotation of the term “truth”.

Action: To remove all residual ambiguity, we have changed in the introduction “the true value of the measurand” to “the true value the measurand has in reality”. For the sake of shortness, we avoid any discussion on what ‘reality’ is.

Comment: As an example, consider maximum likelihood estimation for atmospheric remote sensing retrievals. Measured radiances y are assumed to be generated from a true state of the atmosphere x through a “true” radiative transfer function f , and the true data generating process may be idealized as

$$y = f(x, b) + \epsilon$$

[1] Strictly speaking, the solipsist denial of external truth would make the reviewing of papers absurd. If a paper to be reviewed would not be part of the external truth but a product of the reviewer’s thinking, then it remains unintelligible why the paper needs to be reviewed.

assuming correctly specified Gaussian random errors, ϵ. In practice, as the authors point out, f is not fully known and is replaced by a function F that represents the radiative transfer function to the best of the scientists knowledge. In this case, the observation equation used for inference is

$$y = F(x, b) + \epsilon$$

and the statistical model is "misspecified" in relation to the true model. Under both models, the radiances are assumed to be generated from the specified distribution with unknown true state x_0 , but crucially these "true values" are not the same under both models!

Reply: Not quite. F and f will not produce the same y, even when applied to the same x. There is only one true x. Results inferred from $y = F(x, b) + \epsilon$ are estimates of x but not the true x. The true x is not model dependent.

Comment: The MLE, $\hat{x}$, has the interpretation of the value of x such that the model (correct or misspecified) generates radiances most similar to what is observed (i.e. given $\hat{x}$ the observed radiances are most probable). Assuming regularity conditions hold and reasonably large sample size, the sampling distribution of $\hat{x}$ is approximately Gaussian with mean equal to x_0 (under the model) and the standard deviation represents an estimate of the expected deviation of the estimate from that true value, x_0 . Under the correct model, x_0 is interpreted as the true state of the atmosphere, but under the misspecified model x_0 is the value of x that minimizes the difference between the true data generating model and the misspecified model. The degree to which this true value matches the true target depends on the degree of misspecification which is not known.

Reply: Up to this point we agree.

Comment: Therefore, statements about unknown true values (reality) based on misspecified models ("all models are wrong") are inferential and conditional on all of the assumptions and uncertainties in the measurement system. I do not read GUM as dispensing with the concept of the true value, I understand GUM to recommend that when reporting uncertainties associated with estimates of a value of a measurand (GUM agrees "value of a measurand" can be synonymous with "true" value of the measurand) it is not necessary to make inferential statements about actual errors specifically when reported uncertainties are meant to be used to assess reliability/consistency with other measurement systems. That is, if I have two different measurement frameworks providing interval (uncertainties) of plausible values of a measurand, these intervals can be used to compare consistency with each other without needing to know the "true value."

In this case, it is only necessary to describe uncertainty estimates as summarizing a range of estimates that would also be plausible for the measurand under the measurement system, which is still consistent with quantities reported under the "error concept." Consider an uncertainty (or 'error') estimate, σ_x , related to parameter/measurand x, e.g. the standard deviation of a sampling distribution of $\hat{x}$ (frequentist) or the posterior standard deviation of the posterior $p(x|y)$ (Bayesian).

Reply: The conventional error estimate may also include systematic effects. Thus, σ_x is not necessarily the standard deviation of a sampling distribution of $\hat{x}$. It has long been recognized that the sample standard deviation covers only the random part of the total error.

Comment: Under a frequentist approach, σ_x describes how much the estimate is expected to vary around its statistical expectation $E(\hat{x})$ and represents the spread of values that would also be plausible values of the estimator if the experiment were to be repeated, given the same assumptions in the measurement system. Under a Bayesian paradigm, σ_x describes variability around the posterior mean and provides information on the spread of plausible values (estimates) of the measurand that are also consistent with the scientists' knowledge given the observations, assumptions and prior knowledge. Of course, you can argue that σ_x also describes the spread of the "error distribution" $\hat{x} - x_0$ but this doesn't describe the expected magnitude of actual errors, the mean of the "error distribution" $E(\hat{x}) - x_0$, unless the estimator is unbiased (see related comment about MSE vs variance below). Given this, I do not understand what the authors' issue with this GUM recommendation is, unless they are simply arguing that "value of a measurand" also means "true value of a measurand" (that the GUM agrees with) in which case I see this as quibbling about words and not addressing the larger concept of whether it is necessary to make inferential statements of the form e.g. "95% confident that the true value is within some interval" when reporting uncertainties.

Reply: We broadly agree, and the whole dispute seems to be based on the misunderstanding that we take the sample standard deviation as the estimate of the total error. Instead, our error estimate is the standard deviation of the density function representing the estimated total error.
Our main point is, that GUM is not clear about why it is a problem that the true value of the measurand is unknown and unknowable. Without this piece of information, however, it is unintelligible what the difference between an estimated error and uncertainty is. Many interpretations are possible, yours is one of them. In the literature we quote, we find some more. We formulate working hypotheses how the unknownness of the true value might affect er-

ror/uncertainty estimation and discuss them. We still do not see what should be wrong with this approach.
Interestingly, GUM explicitly supports level-of-confidence approaches (e.g. p. viii). Thus the dismissal of these cannot be quoted as a key difference between the error concept and the uncertainty concept.

Action: Terminological and conceptual issues have been separated in the revised version.

Comment: Additional comments
1. The authors spend several pages (sections) arguing that in addition to the universally accepted statistical definition of error as the difference between measured/estimated and the "truth", a second definition of the word error be accepted (deemed 'error') to refer to statistical estimate of the expected differences between the observed/estimated and true value. This secondary 'error' definition proves confusing in multiple places as it is unclear to which error the authors are referring to, be it actual error or 'error', thus inadvertently making an argument for GUM's choice of separation in language of uncertainty estimates and actual errors.

Reply: We have provided considerable evidence of the use of the term "error" as a statistical quantity in the literature. These examples prove that the "statistical definition" is not so universally accepted as the only meaning of the term "error". We do agree that the implied equivocation may in some cases cause confusion.

Action: In the revised version we take care not to use the term 'error' without the attributes 'estimated' (for the unsigned statistical estimate) or 'actual' (for the realization of the respective random variable), where relevant. This removes all ambiguity.

Comment: In general, the arguments about language definitions of "uncertainty" and "error" could be summarized much more concisely in about a paragraph, acknowledging that the GUM definition of 'uncertainty (of measurement)' encompasses the same quantities that have have often been shorthandedly referred to with reference to the word error as "error estimates", "error bars", etc. Therefore, I think large portions of sections 3 and 4 are repetitive and could be removed.

Reply: We agree that the original version of the paper was repetitive.

Action: Terminological issues have been separated from the conceptional stuff and the related text has been shortened and partly rewritten.

Comment: 2. Page 1, lines 10-11: I find the definition of 'error' as des-

ignating a statistical estimate of the expected difference between the measured and the true value of a measurand to be not in agreement with standard deviations they later reveal are often use as "error estimates" in remote sensing retrievals (e.g. section 5.3). The authors definition is consistent with statistical summaries of error like root mean squared error (RMSE) which estimates the square root of the expected squared difference of actual errors, or median absolute difference the mean of the absolute value of actual errors. The variance of an estimator is only theoretically equal to the MSE if the estimator is unbiased, and even in that case the variability is around the true model parameter of a potentially misspecified model, not necessarily reality. Any inferential statements about the true value in reality and distributions of actual errors are conditioned on all assumptions and uncertainties in the measurement system being reasonably correct. The authors need to clarify their language in regards to what they mean by 'error', "true values", and how these definitions apply to the uncertainty estimates they reference later. Otherwise I am concerned that there is a serious underlying misunderstanding of how to interpret uncertainties they report.

Reply: When we write 'standard deviation' we do not mean a sample standard deviation. In accordance with GUM we conceive the standard deviation as a characteristic of a distribution representing subjective probabilities without any link to multiple measurements. These standard deviations can (and should, if supposed to characterize the total error) – also in agreement with GUM – also include the effects of the systematic effects and model deficiencies, as specified, e.g., in von Clarmann et al., Atmos. Meas. Tech., 13, 4393-4436, https://doi.org/10.5194/amt-13-4393-2020, 2020.
In the criticized Section 5.3, however, the covariance matrices explicitly represent only isolated components of the total error budget. Thus the criticism seems not applicable here.

Action: To avoid such kind of misunderstanding, we have added a footnote, "When we use variances and standard deviations, we do not mean sample variances and sample standard variations but simply the second central moment of a distribution or its square root. In accordance with GUM-2008, this distribution can represent a probability in the sense of personal belief, and thus can include also systematic effects.".

Comment: 3. Page 1, lines 10-11: This is also the first place in the paper where the failure to use consistent mathematical notation is problematic. Consider the simple statistical model

$$X = \mu + \epsilon,$$

where ϵ is a random variable representing actual measurement error. What the authors contend is 'error' could be written, $E(X - \epsilon)$ where

$E()$ denotes the statistical expectation and μ is the "true value" of the measurand. This quantity is equivalent to $E(e)$ and would represent measurement bias.
Or do they mean this to represent var(ϵ), that is $E(\epsilon - E(\epsilon))^2$? Or instead do they intend to refer to the same manner of quantities but with respect to an estimator of μ given a set of observations of $X, x_1, \ldots, x_n$, say $\hat{x}$? The latter would be consistent with what the authors presents in Section 2, but it would help immensely if the authors provided some manner of illustrative model, and used it to clarify their ensuing arguments.

Reply: We refer to neither of the three mentioned quantities. The options suggested in the review rely on sample standard deviations from multiple measurements. We understand that the estimated error is the square root of the second moment of a distribution that characterizes the personal belief of an agent. This is in accordance with GUM.

Action: Part of the problem should have been solved by the footnote mentioned above. We further add: "The estimate of the total error includes both measurement noise and all known components of further errors, random or systematic, caused by uncertainties in the measurement and data analysis system"

Comment: 4. Page 1, lines 14-16: I do not believe GUM presents a "contrasting" definition of the term error. GUM presents the universally accepted statistical definition of error, and defines "uncertainty" to quantify the spread of plausible values given uncertainties in the system.

Reply: Agreed that 'contrasting' is not the adequate term. We disagree that the GUM definition is universally accepted as the only meaning of the term 'error'. The literature we quote furnishes evidence of the contrary.

Action: 'Contrasting' replaced by 'narrower'.

Comment: What do the authors mean here by "measurement error" that the term "measurement uncertainty" is replacing? $Var(\epsilon)$?

Reply: We do agree that the term 'measurement error' is misleading because TUNER is interested in the errors/uncertainties whatsoever of the retrieved value $\hat{x}$, while the term 'measurement error' can be understood as errors in the measured signal y. But the mathematical definition does not help us here, because the quantity $Var(\epsilon)$ can mean two very different things. Conceived as sample variance it would include only the random, or volatile, part of the error, while within the concept of a personal-belief-probability it would include also the systematic (persistent) parts.

Action: In the introduction, "measurement errors in satellite date" is replaced by "errors in estimates of atmospheric state variables retrieved from satellite measurements". Due to the restructuring of the manuscript, the statement at issue has been moved to new Section 2.1, after a statement that the estimate of the total error includes both measurement noise and all known components of further errors, random or systematic, caused by uncertainties in the measurement and data analysis system.

Comment: Then, on page 3, lines 75-76, the authors refer to ϵ as the actual "measurement error" in the y-domain. Is this the same reference to measurement error as in line 16 or there is 'measurement error' meant to refer to the variance or standard deviation of the actual measurement errors (ϵ), If the latter, this inconsistency makes more of an argument for GUM's separation of language definitions of 'error' and 'uncertainty' than for the authors' definition.

Reply: In the original version of the paper, ϵ was consistently used for the actual error in the y-domain. A problem in the old version was not so much that there was an ambiguity between error as a statistical description of a random variable and error as an actual realization of a random variable. This was quite clear from the context. Instead, the problem was that we used the term 'measurement error' both for the error in the measured signal y and the error in the inferred state variable x. GUM does not help here, because GUM is about direct measurements where the x-domain and the y-domain need not to be distinguished.

Action: On suggestion of Reviewer #1 the formal part (where the epsilons appeared) has been shortened, and the statements at issue do no longer appear. Further, we use, wherever relevant, the terms error only with attributes 'estimated' or 'actual'. And we have taken care to reserve the term measurement errors to general contexts, for direct measurements, or in the case of indirect measurements in the y-domain but not for errors in estimates resulting from analyses invoking inversion.

Comment: 5. Page 1, lines 23-24: The authors state, "The claim is made that the uncertainty concept can be construed without reference to the unknown and unknowable true value while the error concept can not." What specifically is the error concept to which they are referring?

Reply: This is an excellent question that GUM fails to answer. Our paper is an endeavour to find out what they mean.

Comment: I read GUM as saying "errors" (as defined as actual measurement errors) cannot be construed without knowledge of the true value (in the example above, μ), meaning that value of a realization

of the random variable ϵ in the example above can't be known because μ is unknown, and analogously the resulting difference between and estimate $\hat{x}$ and μ. However, the parameters of the distribution of ϵ (describing its mean and variance for example) can be discussed, reasoned about, and even estimated from data with some additional assumptions. How does the claim in the following sentence (lines 25-26) follow from this?

Reply: It is trivially true that the actual error cannot be inferred because the true value is not known. But this is not what conventional error estimation aimed at. Conventional error estimation has always aimed at providing statistical error estimates. This is why we distinguish in the revised version of the paper between terminological issues and conceptual issues. The claim in (old) lines 25-26 is quoted directly from GUM.

Action: Terminological and conceptual issues are now discussed separately.

Comment: 6. Page 2, lines 25-26: The authors state that the dispute comes down to "the question if and how the error (or uncertainty) distribution is related to the true value of the measurand." Again, here is where imprecise terminology is confusing. By "error distribution" I would assume they mean the distribution of the random measurement errors ϵ, but what do they mean by "uncertainty distribution" (or do they mean that the word "uncertainty" is now a synonym for "error distribution")? So the dispute is about the relationship between ϵ and μ_0? How? Or by error distribution do the mean some distribution of the estimator - the truth, e.g. $\hat{x} - \mu_0$?

Reply: Again, the misunderstanding arises because the reviewer conceives 'distribution' in a frequentist sense, as obtained from a sample of multiple measurements. We conceive 'distribution' in a wider sense (as GUM does!) as a probability distribution where the probability represents the degree of belief of an agent.

Action: The footnote mentioned above should solve this issue.

Comment: 7. Lines 24-27: The distinction between "error" vs "uncertainty" statisticians is artificial, I am not aware of any such distinction nor do I believe any such dispute or "rift" along these lines exists in the statistical community (I am a practicing statistician). Please cite a reference for the existence of this rift, if you have one.

Reply: As agreed with reviewer #1, we no longer use these terms.

Action: The terms 'error statistician' and 'uncertainty statistician' do no longer occur in the revised manuscript. We no longer refer to this rift and have toned down our statements made in this context.

Comment: 8. Line 128: It has yet to be clearly stated what the authors define to be the debated difference between “error estimation and uncertainty assessment.” A concise definition of both and the argued against definitions at the beginning of the paper would greatly improve the presentation.

Reply: We disagree. The missing definition what error estimation is (in contrast to uncertainty estimation) is exactly what we criticize in GUM. Our paper tries to find out what this difference might be. We find no relevant difference in the concepts (only in the terminology).

Action: In order to avoid misunderstandings we now state that the total error includes both measurement noise and all known components of further errors, random or systematic, caused by uncertainties in the measurement and data analysis system.

Comment: 9. Line 138-148: I can only assume the second meaning of the term ‘error’ the authors are referring to are shorthand statements that have been historically made in the literature such as “the estimated error of quantity of interest is X.” These statements typically use terms like “estimated error” to represent a quantity like a standard deviation of a sampling distribution or a posterior standard deviation, and it is assumed that the reader/community understands this implicit definition (and that is does not provide information about the actual truth without inference). Again, I do not find the argument over whether this specific quantity should be referred to as “uncertainty” or “estimated error” to be particularly compelling, but rather a discussion of the interpretation of these quantities seems to be needed.

Reply: We think that our separation of the manuscript in a (shorter) section on terminological issues and a (longer) one on conceptual aspects solves this issue. Further, we have made clear that the estimated error, as we use this term, does not only include the random part of the error that shows up in the sample standard deviation.

Action: Manuscript reorganized as described above.

Comment: 10. Line 190: I do not see how this is not quibbling about words. What does referring to “uncertainty” under the GUM definition as ‘error’ under the error concept provide that the GUM definition does not? Other than what seems to be a generally misused definition regarding the “truth” in the authors’ definition of ‘error’. I think the authors would agree that the two meanings of error set

forth in this manuscript refer to different concepts.

Reply: We do not see where a misused definition of “truth” comes into play. The two meanings of error refer to the same concept. The error is a random variable, and the term ‘error’ is, depending on the context, used for a specific actual realization as well as for a statistical characterization of its distribution. There is no concept that tries to estimate the error by subtracting the measurement from the true value.

Action: Terminological issues are now confined to one section, the remainder of the paper is on conceptual issues.

Comment: 11. Line 199: I cannot find reference to “error distribution” or “uncertainty distribution” in GUM. The definitions provided here are consistent with a sampling distribution of an estimator (“error distribution”) and a posterior distribution (“uncertainty distribution”). Is this what is intended? If so, please adhere to well-defined statistical definitions. If not, please clarify.

Reply: In general, we mean with ‘error distribution’ the probability distribution in terms of degree of belief, as endorsed by GUM on p. 57. However, in this particular context the term ‘distribution’ is indeed unnecessary.

Action: “...how the error (or uncertainty) distribution is related...” has been replaced by “how the measured or estimated value along with the estimated error (or uncertainty) are related ...”

Comment: 12. Lines 233-239: I see no reason why uncertainties reported as in GUM, along with assumptions of the statistical model, cannot be used for hypothesis testing. In (frequentist) hypothesis testing an assumption is made about the true state, in which case the truth is assumed known and inference is made based on how reasonable this assumption is given the variability (uncertainty) of plausible estimates under the measurement system. If the assumed value of the truth is outside what the scientist believes to be plausible based on their understanding of the measurement system and uncertainties, then a decision is made that the hypothesized value is unlikely to be the true value. A Bayesian hypothesis test would argue whether or not an assumed value or range of values for the parameter are consistent or not with posterior knowledge (uncertainties).

Reply: If the uncertainty does not state a statistical relation between the measured state and the true state, then it cannot be judged how consistent the measurement is with the assumption. The wording “what the scientist believes to be plausible” brings in the concept of the true value through the back door. And if the uncertainty provides a (statistical, estimated, whatsoever) relation between the measured value and the true value, what is then the difference between the uncertainty concept and the error concept?

Comment: 13. Line 265: Again, what is meant by “error distribution”?

Reply: The criticized sentence does no longer appear in the revised version.

Comment: 14. Lines 317-318: 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. This statement is fundamentally false. Monte Carlo methods are simply methods to solve numerical problems through sampling and are used in both frequentist and Bayesian statistics.

Reply: Monte Carlo methods realize probability distributions as frequency distribution and finally interpret resulting frequencies as probabilities. That is all we intended to say. But since this argument is not necessary for our case, we have decided to remove it.

Action: This argument has been deleted.

Comment: 15. Section 5.1: The GUM definition of “uncertainty” does not dispense with reference to the measurand only to its true value. To this end, GUM is consistent with the authors statement we conceive the definition of a quantity and the assignment of the value to a quantity as quite different things. In general this section reads more as a language ”gotcha” argument against the GUM’s use of the term operational definition rather than in a useful argument about the definition of uncertainty, and as such I’d suggest omitting.

Reply: In (old) Section 5.1 (new Section 3) we try to understand why the fact that the true value of the measurand might cause problems for error estimation. We find that it is a legitimate question, how and why the unknownness of the true value might cause problems. The answer is not as trivial as one might think, because, although according to GUM the only meaning of the term “error” is the actual difference between the measured value and the true value of the measurand, the conventional concept of error estimation has never been to try to calculate this actual difference.

Action: During the restructuring/rewriting we have tried to make our argument clearer.

Comment: 16. Section 5.2: This section should be omitted. It presents incomplete and oversimplified interpretations of Bayesian and frequentist methods that are distracting to the manuscript.

Reply: This section is not primarily about Bayesian vs. frequentist methods. The Bayes theorem is accepted both by frequentists and Bayesians and can be inferred directly from the Kolmogorov axioms, without invoking any particular interpretation of probability. We find our base-rate-fallacy argument essential. It is most probably the ONLY cogent argument why error bars around the estimated value must not be considered as descriptors of a pdf that tells one the probability of a given value to be the true value.

Action: The section has been completely reorganized. We have instead deleted the Section “Bayesian versus non-Bayesian”, because the current GUM is not clearly Bayesian; only some of its interpretations are.