Comments and suggestions for the Authors

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. 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. 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. 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. 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.

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. 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. 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.

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) 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. As an example, consider maximum likelihood estimation for atmospheric remote sensing retrievals. Measured radiances $\boldsymbol{y}$ are assumed to be generated from a true state of the atmosphere $\boldsymbol{x}$ through a "true" radiative transfer function $\boldsymbol{f}$, and the true data generating process may be idealized as

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

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

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

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 $\boldsymbol{x}_0$, but crucially these "true values" are not the same under both models! The MLE, $\hat{\boldsymbol{x}}$, has the interpretation of the value of $\boldsymbol{x}$ such that the model (correct or misspecified) generates radiances most similar to what is observed (i.e. given $\hat{\boldsymbol{x}}$ the observed radiances are most probable). Assuming regularity conditions hold and reasonably large sample size, the sampling distribution of $\hat{\boldsymbol{x}}$ is approximately Gaussian with mean equal to $\boldsymbol{x}_0$ (under the model) and the standard deviation represents an estimate of the expected deviation of the estimate from that true value, $\boldsymbol{x}_0$. Under the correct model, $\boldsymbol{x}_0$ is interpreted as the true state of the atmosphere, but under the misspecified model $\boldsymbol{x}_0$ is the value of $\boldsymbol{x}$ 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.

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 $\boldsymbol{x}$, e.g. the standard deviation of a sampling distribution of $\hat{\boldsymbol{x}}$ (frequentist) or the posterior standard deviation of the posterior $p(x|y)$ (Bayesian). Under a frequentist approach, σ_x describes how much the *estimate* is expected to vary around its statistical expectation $E(\hat{\boldsymbol{x}})$ 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" $\hat{\boldsymbol{x}} - \boldsymbol{x}_0$ but this doesn't describe the expected magnitude of actual errors, the mean of the "error distribution" $E(\hat{\boldsymbol{x}}) - \boldsymbol{x}_0$, 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.

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. 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.

2. Page 1, lines 10-11: I find the definition of ’error’ as *designating 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.

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, $E(X - \mu)$ where $E(\cdot)$ 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 $var(\epsilon)$, 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 $\hat{\boldsymbol{x}}$? 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.

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. What do the authors mean here by “measurement error” that the term “measurement uncertainty” is replacing? $Var(\epsilon)$? 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.

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? 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 $\hat{\boldsymbol{x}}$ and $\boldsymbol{mu}$. 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?

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. $\hat{\boldsymbol{x}} - \mu_0$?

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.

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.

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.

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.

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.

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).

13. Line 265: Again, what is meant by "error distribution"?

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.

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.

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.