REVIEW OF SCIENTIFIC INSTRUMENTS **88**, 011301 (2017)

Invited Article: Concepts and tools for the evaluation of measurement uncertainty

Antonio Possolo[a] and Hari K. Iyer[b]
Statistical Engineering Division, Information Technology Laboratory, National Institute of Standards and Technology, Gaithersburg, Maryland 20899-8980, USA

(Received 14 October 2016; accepted 20 December 2016; published online 31 January 2017)

Measurements involve comparisons of measured values with reference values traceable to measurement standards and are made to support decision-making. While the conventional definition of measurement focuses on quantitative properties (including ordinal properties), we adopt a broader view and entertain the possibility of regarding qualitative properties also as legitimate targets for measurement. A measurement result comprises the following: (i) a value that has been assigned to a property based on information derived from an experiment or computation, possibly also including information derived from other sources, and (ii) a characterization of the margin of doubt that remains about the true value of the property after taking that information into account. Measurement uncertainty is this margin of doubt, and it can be characterized by a probability distribution on the set of possible values of the property of interest. Mathematical or statistical models enable the quantification of measurement uncertainty and underlie the varied collection of methods available for uncertainty evaluation. Some of these methods have been in use for over a century (for example, as introduced by Gauss for the combination of mutually inconsistent observations or for the propagation of "errors"), while others are of fairly recent vintage (for example, Monte Carlo methods including those that involve Markov Chain Monte Carlo sampling). This contribution reviews the concepts, models, methods, and computations that are commonly used for the evaluation of measurement uncertainty, and illustrates their application in realistic examples drawn from multiple areas of science and technology, aiming to serve as a general, widely accessible reference. [http://dx.doi.org/10.1063/1.4974274]

I. MOTIVATION

On the last night of the month of May, 2009, Air France flight AF447, en route from Rio de Janeiro to Paris, crashed into the Atlantic Ocean, causing the deaths of all 228 people on board. Only on April 2nd, 2011, would the bulk of the wreckage be found resting on an abyssal plain, at a depth of 3900 m. The first in the sequence of events that caused the accident was an "inconsistency between the airspeed measurements, likely following the obstruction of the Pitot probes by ice crystals."[1]

The Pitot probes were part of a measurement system intended to measure the values of six quantities: calibrated air speed (CAS), Mach number, standard altitude, static air temperature, true air speed, and wind speed. The CAS and Mach number are the main sources of speed information available to the pilots.

The estimates of the values of these quantities are produced by on-board computers as functions of inputs generated by the Pitot probes and other sensors, including static pressure and total air temperature sensors, together with the estimates of ground speed and altitude provided by inertial reference systems.

None of the estimates of the quantities of primary interest were qualified with realistic, real-time, fit-for-purpose assessments of their associated uncertainties, and the pilots may not have realized that they should have adopted procedures appropriate for a situation where the airspeed values had become unreliable.

There were three independent measurement systems using three different Pitot probes on the aircraft. In principle, this redundancy might have enabled a *top-down* uncertainty evaluation of the measurements of air speed, based on the differences between the values measured by the three systems. However, the systems were similarly vulnerable to icing under the conditions encountered in this flight, even though all the probes had drains to allow the removal of accumulated water and an electrical heating system designed to prevent any accumulated water from icing up.

Subsequent studies[1] revealed that, in consequence of icing of the Pitot probes under flight conditions similar to those prevailing at the time of the accident, the relative error in Mach number may have been 62.5% of its true value, and the relative error in CAS may have been 64.3% of its true value, in both cases the measured speed having been too low.

II. INTRODUCTION

Section I emphasizes that measurements are made to inform decision-making. In the case of the ill-fated AF447 flight, the decisions were being made variously, and at different times, by the computers and by the pilots controlling the aircraft. (In the aircraft used in this flight, an Airbus A300-200,[2]

[a]Electronic mail: antonio.possolo@nist.gov. URL: http://www.nist.gov/itl/sed/possolo.cfm.
[b]Electronic mail: hariharan.iyer@nist.gov. URL: http://www.nist.gov/itl/sed/gsg/iyer.cfm.

several computer systems play major control roles.) Faced with inconsistent measured values of airspeed, the computers decided to disconnect the autopilot and change the flight control mode from *normal law* to an *alternate law* according to which the pilots' side-sticks control the ailerons and spoilers directly, but the control inputs that the pilots decided to apply were inappropriate.[1]

In a medical setting, the measurement of the troponin concentration in the blood of a patient with chest pain may support the decision to administer anticoagulant and antiplatelet drugs. In a commercial setting, the decision may be to reject a shipment of breakfast cereal boxes because they are underweight or to certify a lot of painted wooden toys as lead-free. In other settings, the decision may be to stop at a gas station because an automobile's engine is overheating, to continue the operation of a manufacturing line, to change the composition of an input being fed to a chemical reactor, to harvest a crop, to take cover if severe weather is anticipated, and so on.

Measurement uncertainty is the doubt about the true value of the property being measured that remains after making a measurement. Since decisions have consequences, ranging from a mere inconvenience to the loss of life, and including economic loss or profit, measurement uncertainty ought to be taken into account whenever a measurement result is used to inform decisions and actions. White[3] suggests that "measurement uncertainty is a measure of the quality of a measurement; it enables users of measurements to manage the risks and costs associated with decisions influenced by measurements."

In the case of AF447, the airspeed measurement systems were faulty and their redundancy was insufficient because all were vulnerable to the same failure mode. Since there were no trustworthy, fit-for-purpose means to evaluate the uncertainty associated with airspeed, the measured values were not only erroneous but also essentially meaningless. It is all well and good that the reliability of a measurement system involving a Pitot probe will have been characterized under "normal" operating conditions, and that the margin of doubt surrounding its output may be, under such "normal" conditions, inconsequential. It is quite another to characterize such reliability under the full range of realistic operating conditions.

The presentation of estimates of CAS or Mach number, without such a qualification, presumes that measurement uncertainty is negligible for practical purposes. Or, at a minimum, if the presentation is numeric, that all digits reported are trustworthy. A similar interpretive assumption is made if the presentation is graphical.

The characterization of reliability in measurement typically involves the following: (i) calibration, whereby a rule is developed that validly maps instrumental outputs (for example, readings of differential pressure) to values of the quantity intended to be measured (say, airspeed) that are traceable to a recognized reference value; (ii) a study leading to the identification of relevant sources of uncertainty, followed by evaluation of the contributions these sources make to the overall measurement uncertainty (Sec. VII).

The *International Vocabulary of Metrology* (VIM)[4] defines *traceability* as a "property of a measurement result whereby the result can be related to a reference through a documented unbroken chain of calibrations, each contributing to the measurement uncertainty" (VIM 2.41), and also notes that "traceability of a measurement result does not ensure that the measurement uncertainty is adequate for a given purpose or that there is an absence of mistakes."

Instrument calibration and evaluations of uncertainty depend on models that describe how measurement outputs relate to inputs, and how these inputs are produced in the context of the specific physical structure and operation of the instruments used for measurement. Uncertainty evaluation involves using these models to propagate the contributions that sources of uncertainty make to the measured values.

Finally, evaluations of measurement uncertainty must be expressed in a manner that is informative and actionable. In the case of AF447, the pilots apparently were surprised[1] (p. 172) by the fact that the autopilot had been disconnected, but seemed not to have realized right away that the all-important airspeed indications had become unreliable and that both Mach number and CAS might have been very different from what the flight deck displays indicated.

Section III discusses the concept of measurement and what distinguishes measurement from other, apparently similar activities. Section IV reviews a few basic concepts and tools from probability theory and statistics that are used in Secs. V–IX. Section V explains the meaning of measurement uncertainty, and Sections VII and VIII explain how it may be evaluated and expressed. Section VI describes and provides examples of different kinds of measurement models and gives some guidance about when they should be used. Concrete and realistic examples of measurement pervade all the sections and constitute an indispensable part of the narrative.

We provide a generally informal introduction to all of these topics, often simplifying issues and neglecting subtleties that the specialized literature may discuss at great length. For this reason, we provide copious references that may satisfy the reader interested in a treatment more detailed and comprehensive than we can provide here.

III. MEASUREMENT

The word "measurement" is commonly used either to indicate a series of actions undertaken to determine the value of a property or to designate the resulting value. The National Institute of Standards and Technology (NIST) defines a measurement result as comprising (a) a measured value of a property and (b) a qualification of this value with an evaluation of measurement uncertainty (Sec. V).[5–7]

In this article, we focus on the measurement of properties of interest to the physical sciences and to derivative technologies and their applications, including physics, chemistry, biology, engineering, agriculture, manufacturing, etc. We do not discuss measurements of psychological or sensory properties of living organisms, like intelligence, preference, taste, or smell, nor measurements of properties of human or animal populations or of their economic or social interactions.

The properties of interest may be quantitative (bore of an engine cylinder, mass of cocaine in a package, temperature of an oven, efficiency of a chemical reactor, etc.), ordinal

(hardness of a blade, intensity of an earthquake in the Modified Mercalli scale,[8] etc.), or qualitative (gender of a parrot, provenance of a glass fragment, identity of a chemical compound, etc.).

Measurement determines an estimate (believed to be "best" according to some criterion of optimality) of the true value of a property of a particular object, system, or process, by comparing it with the reference value of the same property instantiated in a standard.[9] For example, measuring the length of a box involves comparing its longest dimension against a measuring tape inscribed with markings every millimeter.

The validity and transferability (between people or intelligent machines) of the measurement result rests on the establishment of a chain of comparisons that ultimately relates the distances between consecutive markings on the measuring tape to a reference value recognized by all who may use the measurement result. In this case, the reference value would be the "the length of the path traveled by light in vacuum during a time interval of 1/299 792 458 of a second," which is 1 m by definition,[10] and may be realized using an iodine stabilized helium-neon laser[11] or optical frequency combs.[12,13]

The comparison involved in measurement may be direct, as in the example just given, or it may be indirect, as when the volume of the box is determined by multiplying its length, width, and height, rather than by comparing the volume of the box against the volume of water that the box displaces when fully immersed in it.

The oldest civilizations on record measured length, volume, mass, and time routinely. In Ancient Egypt cubit rods, and ropes with subdivisions marked by knots, were used to measure length. Values of area were derived from such measurements and were required frequently, owing to the changes to agricultural land brought about by recurring flooding of the Nile. Some of the measurements done in classical antiquity revealed extraordinary insight and ingenuity: for example, the measurement of the circumference of the earth by Eratosthenes of Cyrene (276–194 BC).[14]

Two remarkable accomplishments in measurement science, achieved early on during the Industrial Revolution, were Henry Cavendish's determination of the mass of the earth[15] and John Harrison's invention of a reliable, portable chronometer that, by enabling accurate determination of longitude, made long-distance maritime navigation possible without sight of reference points on land.[16] (Up until then, including in the renowned exploits of Portuguese navigators that began in the 15th century, most maritime navigation occurred in close proximity to shores, except for the occasional forays into the western Atlantic to profit from favorable winds, either in return trips from Africa, or in voyages to India.[17])

A. Definition

The VIM 2.1 defines *measurement* as a process of experimentally obtaining one or more values that can reasonably be attributed to a quantity. In addition, it notes that measurement (i) implies comparison of quantities or counting of entities, and (ii) presupposes a description of the quantity commensurate with the intended use of a measurement result, a measurement procedure, and a calibrated measuring system operating according to the specified measurement procedure, including the measurement conditions.

von Helmholtz[18] formulated the conventional understanding of measurement as applying to quantifiable properties and involving the determination of a number together with a unit of measurement to express a magnitude.[19] Accordingly, the VIM (2.3) restricts the concept of measurement to quantitative properties only.

The understanding of measurement that we favor is much broader than what the VIM contemplates, and generally it agrees with the definitions suggested or entertained by Nicholas and White,[20] Dybkaer,[21] White,[22] and Mari and Carbone[23] to address the evolving and expanding needs of measurement science. In this conformity, we propose the following definition, where the word "measurand" (cf. VIM 2.3) means the property intended to be measured.

> Measurement is an experimental or computational process that, by comparing the measurand with a standard, produces an estimate of the true value of a property of a material or virtual object or collection of objects, or of a process, event, or series of events, together with an evaluation of the uncertainty associated with that estimate, and intended for use in support of decision-making.

The Working Group on the International Vocabulary of Metrology (VIM) of the Joint Committees for Guides in Metrology (JCGM) and the Committee on Reference Materials (REMCO) of the International Organization for Standardization (ISO) have been considering whether the term "measurement" may also apply to the assignment of value to qualitative (nominal) properties, or whether a different term should be used instead.

Next we clarify the several key elements of this definition of measurement. In particular, we will discuss what distinguishes measurement from other experimental or computational processes in science and technology, and the meaning of several qualifiers used in the definition.

Counting is a form of measurement where the reference used for comparison is a set of consecutive positive integers starting with the number 1. Counting consists of establishing a one-to-one correspondence between the entities being counted and such set. This understanding is contained in Bertrand Russel's definition of measurement as "any method by which a unique and reciprocal correspondence is established between all or some of the magnitudes of a kind and all or some of the numbers, integral, rational, or real."[24] The result of counting is very much like the result of measuring any other quantity: comprising the number of objects being counted qualified with a specification of their nature ("7 hills of Rome").

Not all scientific experiments produce measurement results. For example, the demonstration that Hans Christian Ørsted carried out in April of 1820,[25] of the effect that the passage of an electrical current through a conductor allegedly exerted upon a nearby compass needle, suggested the existence of a property (magnetism) but neither did it explain why the property manifested itself nor did it quantify the value of the property by comparison with any standard.

Similarly, not all computations are done for purposes of measurement. For example, the computations used to control the autonomous (self-driving) automobile called "Stanley," which won the 2005 DARPA Grand Challenge,[26] solved a series of optimization problems, not a measurement problem. The goal was not to measure Stanley's properties (even if some of them, like the geographical coordinates of its location, were measured in the process) but to navigate successfully from one location to another. In other words, Stanley was not regarded as a system under measurement,[27] although its control relied on inputs from several on-board measurement systems.

Some computations do produce *bona fide* measurement results, and for this reason the foregoing definition lists computational processes alongside experimental processes as possible sources of measurements. For example, the AKS primality test[28] will determine with full certainty, and independently of any unproven assumption, whether any given integer is a prime number. To the extent that it evaluates a qualitative property (primality) and qualifies the result with a statement of uncertainty (full certainty), it performs a computational measurement of primality.

NIST's *Fire Dynamics Simulator* is a paradigmatic example of how measurements may be made in a virtual environment entirely via numerical computation: it involves a model of the physical world (which includes specification of how materials burn and of how air circulates in a room on fire) and rests firmly on validating empirical evidence.[29,30]

The experiments and computations that are performed for purposes of measurement necessarily involve comparison with a reference value instantiated in a standard because this standard either defines the unit of measurement or is a link in the traceability chain that ultimately relates the measurement result to its definition (VIM 2.42). For example, measuring distances between geodetic marks by conventional triangulation requires that the endpoints of a *baseline* also be included in the triangulation. The length of this baseline is determined by comparison with a standard, for example, a calibrated steel tape, whose length in turn has been determined by comparison with higher-order standards, in a chain of inter-comparisons and calibrations that ultimately relates the measurand (the distances between the geodetic marks) to a realization of the meter.

Similarly, a virtual measurement of the temperature at a particular location in the model of a room being consumed by fire, made using the Consolidated Model of Fire and Smoke Transport (CFAST) simulator developed by NIST,[31] rests on comparisons that were made between simulator predictions and data gathered in full-scale fire experiments.[32] In the course of these experiments, measurements were made using calibrated instruments (thermometers, anemometers, barometers), thus enabling comparison with the relevant measurement units in the International System of Units (SI).[10]

The property intended to be measured (*measurand*) may be qualitative: for example, the provenance of a glass fragment collected in a forensic investigation, the identity of the nucleobase at a particular location of a strand of DNA, the shape of a cement particle, or the structure of a molecule. Or it may be quantitative: for example, the mass concentration of 25-hydroxyvitamin D_3 in NIST SRM 972a, Level 1, whose certified value is 28.8 ng ml^{-1} with standard uncertainty 0.55 ng ml^{-1}. The measurand may also be an ordinal property (for example, the Rockwell C hardness of a material) or a function (for example, relating the response of a force transducer to an applied force).

The evaluation of measurement uncertainty (discussed in Sec. V) is an essential part of measurement because it delineates a margin of doubt that qualifies the reliability (or trustworthiness) of the assignment of a value (*estimate*) to the measurand and suggests the extent to which the measurement result conveys the same information for different users in different places and at different times.[23] For this reason, a measurement result must comprise both an estimate of the measurand and an evaluation of the associated uncertainty.

White[22] explains that the intention to influence an action or to make a decision "is an important reminder that measurements have a purpose that impacts on the definition of the measurand (fitness for purpose), and that a decision carries a risk of being incorrect due to uncertainty in the measurements, and a decision implies a comparison against pre-established performance criteria, a pre-existing measurement scale, and the need for metrological traceability." The authors of the aforementioned NIST CFAST simulator state that it is intended as an aid in the fire safety decision-making process.[32]

The following example illustrates the impact that measurement uncertainty may have on decisions that involve measurement results. According to 21 U.S.C. §841(b) (1) (A), a statutory range of ten years to life applies to offenses involving any person that knowingly or intentionally manufactures, distributes, or dispenses 10 g or more of a mixture or substance containing a detectable amount of lysergic acid diethylamide (LSD). The statute seems to be concerned only with the detectability of LSD in the material, not with its mass fraction. Therefore, the uncertainty surrounding the determination of LSD in a sample and the uncertainty surrounding the definition of the detection limit together quantify a risk of the greatest consequence for a defendant (being jailed for ten years or more).

Measurement should also be fit-for-purpose. In particular, it should be sufficiently informative in the sense that the associated uncertainty should be small enough to make the measured value relevant for its intended practical use. Fitness-for-purpose also determines the extent of the effort that needs to be put into the evaluation of measurement uncertainty: for example, the uncertainty requirements when measuring air temperature with a liquid-in-glass thermometer in a Stevenson screen for purposes of long-term climate monitoring are much less onerous than the requirements concerning the uncertainty associated with the measurement of length, when substantiating the detection of gravitational waves.[33]

Mari and Carbone[23] suggest that *objectivity* and *intersubjectivity* are defining traits of measurement. The former meaning "that measurement results actually provide information about the measurand and not of some other property." In other words, "objectivity" means that they are "on target" and are not corrupted by persistent, uncorrected error that makes them deviate systematically from the true value of the measurand.

Inter-subjectivity guarantees "that the meaning of a measurement result is unambiguous and can be easily reconstructed in principle by anyone, possibly on the basis of suitable conventions." In other words, inter-subjectivity guarantees that measurement results are not private constructs but instead should be transferable within a community that recognizes the same set of reference values (for example, the SI). The fact that different metrologists may and often do produce different measurement results for the same measurand does not deny inter-subjectivity: the measured values may disagree or the evaluations of uncertainty associated with them may differ in value or in meaning. Such plurality of results often stimulates dialog and consensus-building to improve the collective knowledge of the measurand (cf. Secs. VI B 5 and VII B 5).

Furthermore, measurement results carry an implied promise of reproducibility by others to within the stated uncertainty and have an implied intrinsic practical value: predictions made taking measurement results into account should be more accurate and generally more reliable than corresponding predictions made in the absence of the measurement results.

IV. PROBABILITY AND STATISTICS

Section V introduces a probabilistic characterization of measurement uncertainty, and Secs. IV A–IV D employ a variety of statistical models and methods. Accessible references on these topics are readily available in excellent books and online resources. Among the former, Freedman, Pisani, and Purves[34] and DeGroot and Schervish[35] are widely used in college courses, and Wasserman[36] provides a more advanced, but still generally accessible overview. Among the latter, NIST/SEMATECH[37] and Possolo and Toman[38] cater to the needs of practitioners of measurement science. In this section, we review a few basic concepts and tools of probability and statistics that are used in Secs. IV A–IV D.

Lindley[39] has argued cogently and passionately in favor of "the inevitability of probability" as description of uncertainty: "*the only satisfactory description of uncertainty is probability*. By this I mean that every uncertainty statement must be in the form of a probability, that several uncertainties must be combined using the rules of probability, and that the calculus of probabilities is adequate to handle all situations involving uncertainty. In particular, alternative descriptions of uncertainty are unnecessary."

However, the choice of probability as representation of uncertainty is not made universally. For example, Mauris, Lasserre, and Foulloy[40] use fuzzy logic to represent measurement uncertainty, and Zadeh[41] argues in favor of the same technology for an even wider set of applications.

A. Probability distributions

A probability distribution on the set S of possible values for the value of a property is a mathematical function P that assigns, to every subset A of S (or, at least, to the so-called *measurable* subsets of S), a probability $P(A)$ such that (i) $P(A) \geqslant 0$, (ii) $P(S) = 1$, and (iii) if $A_1, A_2, \ldots$ are subsets of S with no elements in common, then $P(A_1 \cup A_2 \cup \cdots) = P(A_1) + P(A_2) + \cdots$.[42] The more common version of this axiomatization stipulates that (iii) should apply both to finite and infinite sequences of subsets of S, but some authors favor restricting it to finite sequences only.[43,44]

A probability distribution may be visualized by analogy with a distribution of mass in a region of space. For example, imagine that a canvas has been coated with oil paint applied with a painter's palette knife, in such a manner that the thickness of the coating is uneven, being thicker in some places than in others. If the canvas represents S and the total mass of applied paint represents the unit of probability, then the probability of a subset of the canvas is like the mass of paint on the subset, and it is expressed as mass per unit of area.

Similarly to how one may speak of the density of the paint on the canvas, expressed as mass per unit of area, one may also introduce the probability density of the distribution P as a non-negative function p such that $P(A) = \int_A p(s)\mathrm{d}s$. The often portrayed "bell-shaped curve" is the probability density of a Gaussian (or, normal) distribution. The shaded area under the trapezoidal probability density depicted in Figure 8 comprises 95% of the total area under it; hence the corresponding probability distribution assigns 95% probability to the footprint of this area. Figure 13 depicts the probability density of a vectorial quantity with two elements.

Distributions like the trapezoidal distribution just mentioned, and the Weibull distribution considered in Sec. VI B 2, are continuous distributions (they spread the unit of probability throughout their range without leaving "lumps" at any individual points). The probability distributions of principal interest in the example discussed in Sec. VI B 6 are discrete because they are concentrated on a countable (indeed finite in this case) set of labels of possible sources for a glass fragment of interest in a forensic investigation, and they describe the confidence with which a fragment of unknown origin may be assigned to a particular source.

B. Random variables

A random variable is a mathematical model for a value that has a probability distribution as an attribute. For example, the standard atomic weight of carbon is the interval [12.0096, 12.0116][45] (Sec. VII A 5). This may be interpreted by saying that the atomic weight of carbon in a specimen of Indiana limestone is represented by a random variable with a uniform (or, rectangular) distribution on this interval. In other words, we use a probability distribution to express our state of knowledge about the average atomic weight of the isotopes of carbon actually present in the sample: in particular when all we know is that the specimen is from a "normal material" as defined in Sec. VI A 1.

Two random variables U and V are said to be (stochastically) independent if the probability that U and V jointly take values in sets A and B, respectively, equals the product of the probabilities of their taking values in these sets separately. This means that knowledge that U has taken a value in A does not alter the probability of V taking a value in B.

C. Statistical models

The examples discussed in Secs. VI B 1–VI B 5 all involve statistical models, that is, descriptions of relations between random variables, some observable, others not. The statistical models used for uncertainty evaluations in measurement science typically describe the relationship between the measurand, the property intended to be measured, and other properties, some observed directly in the course of the measurement experiment, others having been measured previously, others still having to be estimated from empirical data, yet all regarded as random variables whose probability distributions describe their associated uncertainties.

The following four types of statistical models arise often in practice, in particular in the examples discussed in Secs. IV D 1–IV D 4. However, in some instances, the models may be applicable only to suitably re-expressed data, for example, to the logarithms of the observations, rather than to the observations themselves.

(i) *Additive measurement error model.* Each observation made under conditions of repeatability (VIM 2.20) $x_i = g(y)+\varepsilon_i$, for $i = 1, \ldots, m$, is the sum of a known function g of the value y of a measurand, and a random variable ε_i that represents measurement error. The measurement errors corresponding to different observations may have the same or different standard deviations, may be correlated or uncorrelated, and they may have a Gaussian distribution or some other distribution. Example E2 in the *Simple Guide for Evaluating and Expressing the Uncertainty of NIST Measurement Results* (subsequently referred to as *Simple Guide*)[7] describes an instance of this model.

(ii) *Random effects model.* The value $x_j = y + \lambda_j + \varepsilon_j$ measured by laboratory j, or using measurement method j, is equal to the value y of the measurand, plus the value λ_j of a random variable representing a laboratory or method effect, plus the value ε_j of a random variable representing measurement error, for $j = 1, \ldots, n$ laboratories or methods. The $\{\lambda_j\}$ are usually assumed to have mean 0 and the same standard deviation τ. The probability distributions of the $\{\lambda_j\}$ and of the $\{\varepsilon_j\}$ are often assumed to be Gaussian. If the data were only the $\{x_j\}$, then it would be impossible to distinguish the laboratory effects $\{\lambda_j\}$ from the laboratory-specific measurement errors $\{\varepsilon_j\}$. Since the evaluations of uncertainty $\{u(x_j)\}$ also are part of the data, and we know that the absolute values of the $\{\varepsilon_j\}$ are generally comparable to the $\{u(x_j)\}$, we can conclude that any "excess variance" exhibited by the $\{x_j\}$ is attributable to the $\{\lambda_j\}$, whose dispersion (or scatter) is gauged by τ. We illustrate the use of this model in Sec. VI B 5.

(iii) *Regression model.* The measurand y is a function relating corresponding values of two quantities at least one of which is corrupted by measurement error. For example, y is a third-degree polynomial and the amount-of-substance fraction of a component of a gas mixture is given by $x = y(r) + \varepsilon$, where r denotes an instrumental indication and the random variable ε denotes measurement error. Many calibrations involve the determination of such a function y using methods of statistical regression analysis. We will illustrate a regression model in the context of instrument calibration in Sec. VI B 3. Example E17 in the *Simple Guide*[7] describes a different regression model that is used for a similar purpose.

(iv) *Time series model.* The observations are indexed by discrete, typically equispaced epochs in time, and the measurand y may be their common mean, as in $x_t = y + \varphi_\alpha(x_{t-1}, x_{t-2}, \ldots) + \psi_\beta(\varepsilon_t, \varepsilon_{t-1}, \varepsilon_{t-2}, \ldots)$, where α and β denote adjustable (vector) parameters, and φ and ψ are commonly assumed to be linear functions of their arguments. According to this model, the observation made at epoch t is influenced by the values of observations made at previous epochs, and also by measurement errors affecting them. The example discussed in Sec. VI B 1 involves a model of this kind.

D. Statistical methods

1. Model selection

In just about all applications where a statistical model is called for, multiple models may reasonably be entertained that typically yield somewhat different results. In the example discussed in Sec. VI B 2, Weibull, lognormal, and gamma models are reasonable candidates for the lifetime of an aircraft component. Therefore, two questions arise: (i) how to choose one among several alternative models and (ii) how to evaluate and propagate the uncertainty surrounding model selection.

The first question is often answered by comparing values that a numerical model-selection criterion takes for the different alternatives being contemplated.[46] One such criterion is the Bayesian Information Criterion (BIC), which is used in Secs. VI B 1 and VII B 2. Another is Akaike's Information Criterion or a variant thereof, which is used in Secs. VI B 1 and VI B 3. Ideally, these numerical criteria should be used in conjunction with the examination of graphical diagnostics that are revealing about model performance, as in Figure 4.

A model may also be selected based on a comparison of the predictive skill of the alternative models, when they are applied to data similar to, but different from the data used to calibrate them: this approach, called *cross-validation*,[47] is illustrated in Sec. VI B 6.

Since selection of a particular model does not imply certainty in the "truth" of the selected model (in fact, and in just about all cases, there will be no "true" model), the question remains of how to recognize and propagate model uncertainty.[48,49] As the example in Sec. VII B 2 illustrates, this may be done in the course of the uncertainty evaluation by including model resampling and averaging during application of the statistical bootstrap.

2. Estimation

In measurement models called observation equations (reviewed in Sec. VI B), which in fact are statistical models, the measurand either appears as an adjustable parameter of a statistical model (Sec. VI B 1) or is a function of several

such parameters (Sec. VI B 2). The values of these parameters are determined by application of a statistical estimation method. The example in Sec. VI B 2 introduces and applies maximum likelihood estimation, and the method is also used in the examples presented in Secs. VI B 3 and VII B 1.

If the data $\boldsymbol{v}$ (possibly a vector of observations of the same value of a property made under conditions of repeatability, VIM 2.20) are modeled as an outcome of a random variable whose probability density function is p_θ, then a maximum likelihood estimate of θ is a value of this parameter that maximizes $p_\theta(\boldsymbol{v})$.

The example in Sec. VII B 5 defines and illustrates a Bayesian estimation procedure. Both maximum likelihood estimation and Bayesian estimation procedures produce not only estimates of parameters but also evaluations of the uncertainty surrounding the estimates: the former recognizing only the components of uncertainty that are expressed in the likelihood function, the latter recognizing also the *a priori* uncertainty encapsulated in the prior distribution that is used.

3. Statistical intervals

Both the GUM and the VIM characterize *coverage intervals* and take pains to distinguish them from other types of intervals often used to express results of statistical analyses, which include confidence intervals, (Bayesian) credible intervals, prediction intervals, and tolerance intervals.[50,51]

The GUM offers the need "to provide 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" as motivation for the definition of a *statistical coverage interval* (GUM C.2.30) as "an interval for which it can be stated with a given level of confidence that it contains at least a specified proportion of the population."

The VIM (2.36) defines *coverage interval* as an "interval containing the set of true quantity values of a measurand with a stated probability, based on the information available." The corresponding definition in the GUM Supplement 1 (GUM-S1, 3.12)[52] is "interval containing the value of a quantity with a stated probability, based on the information available," and it is accompanied by a note suggesting that "a coverage interval is sometimes known as a credible interval or a Bayesian interval."

The statistical model underlying the uncertainty evaluations described in the GUM, which will be discussed in Section VI, involves n random variables $X_1, \ldots, X_n$ (input quantities), and a function of them, $Y = f(X_1, \ldots, X_n)$ (output quantity), that is used to estimate the value θ of a property of interest, which could be the mean (or, expected value) of the probability distribution of Y, or some other attribute of this distribution. The GUM focuses on the case where the $\{X_i\}$, Y, and θ all are scalars.

Confidence Interval. The endpoints of a confidence interval for θ are random variables A and B such that $\Pr\{A < \theta < B\}$ has a specified value, say 95%. The interval whose endpoints are the realized values of the random variables A and B is called a confidence interval. In practice, A and B are functions of the $\{X_i\}$ that do not depend on θ.

Credible Interval. A (Bayesian) credible interval is an interval of possible values of θ that this parameter (now regarded as a random variable) has a specified posterior probability of belonging to, given the data, which may be measured values of the $\{X_i\}$, or observations underlying these measured values.

Prediction Interval. The endpoints of a prediction interval for Y are two random variables A and B such that $\Pr\{A < f(Z_1, \ldots, Z_n) < B\}$ has a specified value, say 95%, and where the $\{Z_i\}$ are independent of the $\{X_i\}$ and have the same probability distribution as the $\{X_i\}$. This is often described informally by saying that a prediction interval for the estimate of θ is an interval that has a specified probability of including the value of the estimate corresponding to a set of *future* observations.

Tolerance Interval. A tolerance interval for the probability distribution of Y is an interval whose endpoints are random variables and that, with a specified probability (*confidence*), includes at least a specified fraction (*content*) of the unit of probability that the distribution of Y distributes throughout its range.

Therefore, the definition of coverage interval in the GUM (C.2.30) is closest to the definition of tolerance interval. However, the GUM (2.3.5, Note 1, and 6.2.2) calls "coverage probability" (or "level of confidence") the "content" in the foregoing definition of tolerance interval, and in examples (say, in 7.2.4 and in H.1.6) the GUM only specifies the coverage probability. And the coverage intervals that the GUM produces, for example, 100.021 47 g ± 0.000 79 g in 7.2.4, and 50.000 838 mm ± 0.000 093 mm in H.1.6, are approximate Student's t confidence (not tolerance) intervals for the mean of the probability distribution of the output quantity. The definition of coverage interval in the VIM is closest to the definition of a Bayesian credible interval.

a. Illustration. Five weighings of a standard #10 mailing envelope enclosing four sheets of paper, made using a postal scale under conditions of repeatability (VIM 2.20), produced 25.7 g, 24.1 g, 24.8 g, 25.5 g, and 23.5 g. Suppose that these are a sample from a Gaussian distribution whose mean and standard deviation both are unknown.

- A 95% confidence interval[50] for the mean is (23.6 g, 25.9 g). The endpoints are given by $\overline{w} \pm t_{4,97.5\%} s/\sqrt{5}$, where $\overline{w}$ denotes the average of the five weighings, s denotes their standard deviation, and $t_{4,97.5\%}$ denotes the 97.5th percentile of a Student's t distribution with four degrees of freedom (R function t.test may be used to compute such confidence intervals). This is the coverage interval specified in the GUM (G.3.2).
- Since the same interval is also a 95% Bayesian credible interval corresponding to the Jeffreys' prior distribution,[53] (23.6 g, 25.9 g) satisfies the substance and the spirit of the definition of coverage interval given in the VIM 2.36. (However, many other Bayesian credible intervals, corresponding to other prior distributions, would qualify as well.)

- A 95% prediction interval[50] for the average of five future observations that may be drawn from the same distribution as the observations in hand ranges from 23.1 g to 26.4 g, which are $\overline{w} \pm t_{4,97.5\%} s\sqrt{2/5}$.
- A tolerance interval[50] for the probability distribution of the average of five observations from the same distribution that those five weighings were drawn from, with 95% confidence and 95% content, ranges from 21.8 g to 27.6 g.

4. Monte Carlo and bootstrap

Both Monte Carlo simulation methods and the statistical bootstrap[54] are often used for uncertainty evaluation.

The version of the Monte Carlo method used for "propagation of distributions," in the sense of the GUM-S1, is discussed in Sec. VII A 3 and applied in Sec. VII A 6. The Markov Chain Monte Carlo (MCMC) method[55] is used in Sec. VII B 5 to perform a Bayesian uncertainty evaluation for the mass fraction of PCB 28 in a sediment.

The statistical bootstrap is a widely applicable, the most useful statistical procedure, with two main variants: one does not require that an assumption be made about the specific form of the probability distribution describing the dispersion or scatter of the data (hence is called "non-parametric bootstrap"), while the other does require that a particular class of probability distributions be specified whose elements are determined by values of a parameter whose unknown true value is estimated using data gathered during measurement (and for this reason it is called "parametric bootstrap"). Hesterberg[56] provides an instructive review of bootstrap methods.

Both variants can be explained succinctly by considering a sample $v_1, \ldots, v_m$ drawn from a probability distribution P_θ that depends on a parameter θ. Suppose that θ is estimated by $s = S(v_1, \ldots, v_m)$, for some fully specified function S (that does not involve θ). The form of P_θ (for example, that it is a two-parameter Weibull distribution) needs to be specified for the parametric bootstrap, but not for the non-parametric bootstrap.

For the non-parametric bootstrap, draw a sample of size m, uniformly at random and with replacement, from the set $\{v_1, \ldots, v_m\}$. This means that the m elements in this set have the same probability of being selected, and that each may be selected more than once. Let $\{v_1^*, \ldots, v_m^*\}$ denote the resulting sample and evaluate $s^* = S(v_1^*, \ldots, v_m^*)$. Repeat this process a large number K of times to obtain $s_1^*, \ldots, s_K^*$, and then treat these K values as a sample from the probability distribution of the random variable $S(V_1, \ldots, V_m)$. In particular, when θ is a scalar and S is real-valued, the standard deviation of the $\{s_k^*\}$ is an estimate of the standard deviation of $S(V_1, \ldots, V_m)$.

For example, the Analytical Methods Committee[57] of the Royal Society of Chemistry lists the following determinations of the mass fraction of copper (expressed in μg/g) in wholemeal flour obtained under the conditions of repeatability (VIM 2.20): 2.9, 3.1, 3.4, 3.4, 3.7, 3.7, 2.8, 2.5, 2.4, 2.4, 2.7, 2.2, 5.28, 3.37, 3.03, 3.03, 28.95, 3.77, 3.4, 2.2, 3.5, 3.6, 3.7, 3.7. Function huberM defined in R package robustbase[58] implements a robust alternative to the arithmetic average as indication of the typical value of the mass fraction, which it estimates as 3.21 μg/g. (The term "robust" means that the performance of the estimate, in particular, the associated uncertainty, by and large is unaffected by the set of replicated observations possibly not being a sample from a Gaussian distribution.)

Application of the non-parametric bootstrap with $K = 1000$ produced a set of bootstrap replicates with standard deviation 0.14 μg/g, which characterizes summarily the measurement uncertainty associated with the estimate.

For the parametric bootstrap, first compute an estimate $\widehat{\theta}$ of θ (for example, by application of the method of maximum likelihood) using the data $v_1, \ldots, v_m$. Then repeatedly draw samples of size m from $P_{\widehat{\theta}}$: that is, from the probability distribution of the data with the estimate of the parameter "plugged-in" and treated as if it were the true value of the parameter. Evaluate S for each of these samples, obtaining $s_1^*, \ldots, s_K^*$, and then proceed as above. The parametric bootstrap is employed for uncertainty evaluations in Secs. VII B 2, VII B 4, and VII B 5, for example.

V. MEASUREMENT UNCERTAINTY

Measurement uncertainty is the doubt about the true value of the measurand that remains after making a measurement. Bell[59] points out that to characterize the margin of this doubt, we need to answer two questions: "*How big is the margin?*" *and* "*How bad is the doubt?*" Answering these questions involves specifying (i) a range of values of the measurand that are consistent with the measured value (that is, a range of values that the measured value does not differ significantly from as judged by a suitable statistical test that takes measurement uncertainty into account); (ii) the strength of the conviction, or the degree of belief, that the true value of the measurand indeed lies within that range.

For example, the certificate for NIST SRM 972a[60] states explicitly that, with probability 95%, the mass concentration of 25-hydroxyvitamin D_3 in Level 1 of the material lies within the interval 28.8 ng ml^{-1} ± 1.1 ng ml^{-1}. The 1.1 ng ml^{-1} is the size of the margin of doubt: it is called *expanded uncertainty* for 95% coverage. The 95% probability expresses the degree of belief that the true value of the mass concentration lies in that interval, and its complement, 100% − 95% = 5% characterizes the doubt.

The general concept of measurement uncertainty may be explained in several different ways that mutually illuminate each other. For example, Tal[61] suggests a characterization focused on predictive acumen, "measurement uncertainty is a measure of the predictability of measurement outcomes under an idealized model of the measurement process," in accordance with the intuitive notion that a measurement result contains a promise about what others should expect when they evaluate the same property.

A distinction should be kept clearly in mind between what measurement uncertainty is and how it may be represented. In the example given above concerning the mass concentration of a vitamin D constituent, the severity of the doubt surrounding the measurand is represented probabilistically. The *Guide to the expression of uncertainty in measurement (GUM),*[62] Taylor and Kuyatt,[6] and the *Simple Guide*[7] all adopt a probabilistic representation of measurement uncertainty.

The definition of measurement uncertainty as doubt that remains after making a measurement is in syntony with the definition of uncertainty adopted by the many scientists, from a very wide range of disciplines, participating in the Intergovernmental Panel on Climate Change (IPCC), an international organization that assesses the scientific basis of the observations of climate change. Its Working Group I (WG-I) defines *uncertainty as* "a state of incomplete knowledge that can result from a lack of information or from disagreement about what is known or even knowable." And this WG-I also distinguishes the concept of uncertainty from its representation, explaining that uncertainty can be "represented by quantitative measures (e.g., a probability density function) or by qualitative statements (e.g., reflecting the judgment of a team of experts)"[63] (p. 1464).

A. Probabilistic representation of measurement uncertainty

Adopting a probabilistic representation of measurement uncertainty requires some understanding of the mathematical structure of probability distributions and involves assigning a specific meaning to probabilistic statements. The former is mostly a technical issue that we have reviewed briefly in Sec. IV A, while the latter is a philosophical issue with practical implications. For this reason, here we discuss, however briefly, some of the philosophical issues underlying the interpretation of probabilistic statements, justified by the following observation made by Savage:[64] "We statisticians, with our specific concern for uncertainty, are even more liable than other practical men to encounter philosophy, whether we like it or not."

A probabilistic statement like "with probability 95%, the mass concentration of 25-hydroxyvitamin D_3 in Level 1 of the material lies within the interval 28.8 ng ml^{-1} ± 1.1 ng ml^{-1}" may be interpreted in any one of several different ways, depending on the meaning of probability.

Two common interpretations are as follows: (a) "Based on all the evidence in hand, NIST scientists are prepared to bet 95 cents on the dollar that the true value of that mass concentration lies between 27.7 ng ml^{-1} and 29.9 ng ml^{-1}"; (b) "no fewer than 95% of similar intervals that NIST has produced historically are expected to have covered their respective targets, and no more than 5% missed them." A narrower version of (b) claims only that, if the process of certification of the reference material were repeated many times over, in essentially the same conditions, and a nominal 95% coverage interval were produced each time, then about 95% of these intervals would cover the true value of the measurand, and 5% would miss it.

Hájek[65] describes and compares no fewer than six different interpretations of probability. Two of them are commonly used to characterize statistical inferences: (i) subjective degree of belief (manifested in (a) above, often called *credences*); (ii) long-run frequency (implied in (b), frequently called *chances*).

The concept of probability (*epistemic probability*) as a means to provide a continuum of gradations between the values of true and false (for propositions) entertained by classical logic, conveying partial entailment, originated with Keynes[66] and was developed vigorously by Carnap[67] in philosophy, and by Jeffreys[68] and Cox[69] in statistics. Mellor[70] explains that "in civil courts, where plaintiffs are required to establish their case 'on the balance of probabilities,' these probabilities are also epistemic."

Interpretation (i) is subjective or personalistic,[44,71] with probability conveying the depth of someone's belief, or the strength of someone's confidence. Interpretation (ii) is often called *frequentist* because it equates probability with the frequency with which some repetitive event occurs in the long run over time, or across a possibly arbitrarily large population (either real or hypothetical).

Much of conventional statistical practice relies on (ii), and it is often portrayed as objective and empirical. However, the practice of frequentist statistics comprises subjective elements. Some are implicit in model selection. Others serve to enable the frequentist interpretation itself, for example, by invoking a hypothetical or imaginary population that the data are regarded as having been sampled from. In such cases it is not the probability that is subjective, but the very hypothetical world that defines the supposedly objective probability.

The meaning of the probabilities used to characterize measurement uncertainty is consequential in practice because measurement results are typically intended to be communicated and shared by different people, and to guide their actions and interactions as they advance science, develop and apply technologies, or engage in trade and commerce. If measurement uncertainty is to serve as a guide for action in society, then its meaning and expression need to be transferable from one person to another without requiring lengthy or complicated explanations.

The views are strikingly divided on how probabilities should be interpreted when they are used to characterize measurement uncertainty. For example, O'Hagan[72] argues persuasively that only a subjective interpretation of probability, reflecting a state of knowledge, seems capable of addressing all aspects of measurement uncertainty comprehensively. But spirited arguments have also been advanced favoring the frequentist interpretation as most adequate for applications in measurement science. For example, Willink and White[73] state that "the purpose of an uncertainty analysis is to characterise the behaviour of real world events, including the error processes in our measurements, it seems ill-advised to forgo the use of frequentist statistics where its methods are proven." White[3] provides a captivating, thought-provoking narrative favoring the frequentist viewpoint.

The key questions in this dilemma are the following: (a) Whose is the uncertainty captured in evaluations of measurement uncertainty? (b) How will this uncertainty be communicated and used by producers and consumers of measurement results?

Concerning the former — *whose uncertainty is it?* — the Merriam-Webster online dictionary (www.merriam-webster.com) defines "uncertainty" as "the quality or state of being uncertain," and define "uncertain" as "not known beyond doubt." Since measurement uncertainty is a kind of uncertainty, by rights its meaning ought to conform to the common meaning of the general concept, which applies to states of knowledge and to the relation between a subject and an object of this subject's interest. For example, when we admit to being

very uncertain about the identity of the fiftieth digit in the decimal representation of the number π, we are indicating that we really do not know what that digit is, not that there is anything intrinsically "uncertain" that prevents us from determining it with full certainty if only we put our minds and computing resources to it.

Uncertainty is the absence of certainty, and certainty is either a mental state of belief that is incontrovertible for the holder of the belief (like, "I am certain that my eldest son was born in the month of February"), or a logical necessity (like, "I am certain that 426 389 is a prime number"). In summary, uncertainty in general and measurement uncertainty in particular qualify knowledge or incomplete knowledge; hence it is yours or mine, or possibly a group of people's.

For example, the recommended values of the fundamental physical constants, which are released periodically by the Committee on Data for Science and Technology (CODATA), are qualified with uncertainties that reflect how well CODATA as a collective believe to know their values, based on the published information that they take into account in preparation for each release:[74] for the Newtonian constant of gravitation, the stated relative standard uncertainty is 0.012%.

Standard uncertainties, expanded uncertainties, and coverage intervals are common expressions of measurement uncertainty for scalar measurands. The standard measurement uncertainty associated with a scalar measurand is the standard deviation (or an analog thereof) of the probability distribution that expresses the state of knowledge about the measurand (cf. VIM (2.30)).

In the absence of specific assumptions about the probability distribution of a measurand y (say, that it is Gaussian or rectangular), the associated standard uncertainty $u(y)$ has no probabilistic meaning and conveys only an indication of the spread of the underlying distribution around its mean. In any case, $u(y)$ is a summary descriptor of the dispersion (or scatter) of the probability distribution of the measurand. An expanded uncertainty typically is a multiple $ku(y)$ of the standard uncertainty, where the multiplier k (called *coverage factor*) is chosen large enough so that $y \pm ku(y)$ is a coverage interval for the true value of y with a specified probability.

A coverage region (which, in the case of a scalar measurand, may be an interval) is a set of possible values of the measurand that, with specified probability, is believed (by a metrologist or by a community of metrologists) to include the true value of the measurand (cf. VIM 2.36) with a specified probability.

Concerning the question of how uncertainty will be communicated and used, we are willing to concede that a frequentist interpretation may be easier to explain and in closer accord with popular sentiment than any other because most people are familiar with games of chance, which is where frequentism appears to be most relevant. However, a close and critical inspection reveals fundamental deficiencies in the common frequentist concept of probability.[75,76]

Furthermore, a frequentist interpretation need not be more credible than any other because people generally neither keep score nor can present any other evidence about the frequency with which the probabilistic statements they have made have turned out to be true. A notable exception is the *skill* of weather forecasting systems, which is tracked meticulously.[77] But even here, what empirical observations verify is not the stated probabilities of rain, but merely the "*success*" rate of predictions of rain or no-rain. The daily probability of rain that forecasts typically provide cannot possibly be verified empirically because the forecast takes into account the particular, one-off circumstances that differ from day to day and will not repeat over the long run.

Provided that an agreement has been reached about measurement uncertainty as characterization of a state of knowledge (about the measurand), and that a choice has been made to use probabilities to quantify measurement uncertainty, a subjective interpretation of these probabilities may be more credible than a frequentist interpretation because it relies only on the ability of the knowing subject or subjects to express their uncertainty. Of course, the question remains of the practical usefulness (including transferability) of the underlying belief. This usefulness is contingent on the publicly demonstrable practical value of predictions made consistently with said belief.

Our position is that measurement uncertainty refers to a state of knowledge held by an individual person, by a group of persons comparably familiar with the same shared, relevant evidence about the object of interest, or by an artificial intelligence (for example, Google's *DeepMind* or IBM's *Watson*), hence quantifies a (subjective) credence. We also believe that this character does not in any way discredit its relevance to science, technology, or commerce. Quite the contrary, the explicit recognition that different people interested in the true value of the same measurand believe to know it to varying degrees stimulates the exchange of information and indicates where efforts may most advantageously be expended to fill gaps in knowledge. Inter-comparisons of measurement results, as discussed in Section VII B, give the participants the means to gauge the trustworthiness of their own uncertainty evaluations, regardless of how they will have been produced or interpreted.

Neither does subjectivity preclude consensus, which may be reached by application of statistical methods, although there is yet no agreement on which method may be best to reach consensus. For example, opinion pooling techniques have been in use for this purpose for a very long time[78,79] and continue to be deemed useful.[80] The example presented in Sec. VI B 5 describes a statistical analysis of measurement results for the mass fraction of a polychlorinated biphenyl (PCB) in sediment that produces a consensus value and qualifies it with an uncertainty evaluation that reflects both observed differences between the values measured by the participating laboratories and the stated laboratory-specific uncertainties.

VI. MEASUREMENT MODELS

Our approach to the evaluation of measurement uncertainty is model-based, and so are the approaches described in the GUM, by Taylor and Kuyatt in 1994 for NIST,[6,7] and by EURACHEM[81] (a network of European organizations aiming to establish a system for the international traceability of chemical measurements and the promotion of good quality practices, www.eurachem.org), among many others. Therefore, we begin

by reviewing the roles that models play in measurement, generally inspired by Tal.[19] Sections VI A and VI B introduce the two classes of models used most often for uncertainty evaluations.

Measurement is typically done in the context of an experiment designed to make the property of interest accessible to a sensor in a measuring instrument. For example, to measure the amount-of-substance fraction of sulfur dioxide in a gas mixture, the mixture is introduced into a flow-through process analyzer with a pulsed UV fluorescence detector.

The indications produced by the instrument usually require further processing to obtain a measurement result. The steps that must be taken to transform instrumental indications into a measurement result depend on an understanding that relates the instrumental response to the measurand, or, more generally, on an understanding of the relationship between the inputs, possibly of very different natures, and the measurand. This understanding generally comprises physical theories and statistical modeling assumptions.

The indications produced by the flow-through process analyzer are corrected for an additive "background" contribution that is the result of electrical noise and scattered light. Similarly, a contribution from *dark current* is subtracted from the indications produced by the charge-coupled device (CCD) in an imaging photometer. In both cases, even this first and fairly trivial step already presumes an understanding of physical processes operating in the measuring instrument that influence the indications it produces.

The measurement of airspeed using a Pitot tube, which is discussed in Sec. VI A, rests on the theory of fluid dynamics that describes how differences in pressure between suitably positioned ports are informative about airspeed and that also describes the role that temperature plays in the process via the ideal gas law.

The characterization of the thermal stability of a thermal bath, considered in Sec. VI B, depends on an understanding of the pattern of correlations between observations of temperature made at regular intervals and involves specific assumptions that support the selection and use of a particular statistical model for a time series.

Calibration (VIM 2.39) is a procedure that establishes a relationship between values of a property realized in measurement standards, and indications provided by measuring devices, or property values of artifacts or material specimens, taking into account the measurement uncertainties of the participating standards, devices, artifacts, or specimens. For a measuring device, the relationship is usually described by means of a calibration function that maps values of the property realized in the standards, to indications produced by the device being calibrated. White and Saunders[82] provide an account of calibration based on interpolation equations.

However, to use a calibrated device in practice, the (mathematical) inverse of the calibration function is required, which takes an indication produced by the device as input, and produces an estimate of the property of interest as output. Bartel, Stoudt, and Possolo[83] illustrate this process in the context of force measurement.

For example, to measure temperature using a platinum resistance thermometer (PRT), a relationship must be established that maps values of electrical resistance into values of temperature. One such relationship is of the form $R(t) = R(t_0)(1 + \alpha(t - t_0) + \beta(t - t_0)^2)$, where $R(t)$ denotes the resistance at temperature t, $R(t_0)$ denotes the resistance at a reference temperature t_0 (usually 0 °C), and α and β are instrument-specific, adjustable parameters.

Calibrating the PRT amounts to estimating the values of these parameters based on a collection of pairs of values $\{(t_i, R(t_i))\}$, where the $\{t_i\}$ denote known temperatures, for example, the triple point of water (0.01 °C), the melting point of gallium (29.7646 °C), and the freezing point of indium (156.5985 °C), and the $\{R(t_i)\}$ denote the corresponding resistances.

Calibration also involves evaluating the uncertainty associated with the estimates of α and β: uncertainty that they inherit from the uncertainties associated with the realization of the $\{t_i\}$, the measurement of the $\{R(t_i)\}$, and the selection of functional form for the calibration function.

A. Measurement equations

A *measurement equation* expresses the measurand as a function of a finite set of input variables for which estimates and uncertainty evaluations are available. This is the only measurement model considered in the GUM, formulated as $y = f(x_1, \ldots, x_n)$, where $x_1, \ldots, x_n$ denote values of the inputs, y denotes the corresponding value of the measurand (or, output), and f is assumed to be a fully specified function.

The GUM assumes that the inputs and the output all are quantitative, and the approximation technique it employs to evaluate the uncertainty associated with the output hinges on the possibility of computing partial derivatives of f with respect to its arguments. When the measurement equation involves some qualitative inputs, or when the output is qualitative, then uncertainty evaluation and propagation cannot be done using the technology in the GUM, but may be practicable instead using a customized version of the Monte Carlo method, as in Example E6 (DNA Sequencing) of the *Simple Guide*.[7]

The uncertainty evaluation hinges on modeling the inputs and the output as random variables, with the understanding that any quantity value surrounded by uncertainty may be so modeled regardless of its intrinsic nature (for example, whether it is "fixed" or "variable," in any senses of these words). The GUM explains (in Note 1 of 4.1.1) that "the same symbol is used for the physical quantity [. . .] and for the random variable [. . .] that represents the possible outcome of an observation of that quantity."

A more careful account distinguishes random variables ($X_1, \ldots, X_n$ and Y) from their values ($x_1, \ldots, x_n$ and y) and also distinguishes possible values of these random variables from the true values of the quantities that they represent: $\xi_1, \ldots, \xi_n$ for the inputs and η of the output. (The GUM assumes that these true values are essentially unique, and so do we.)

The measurement equation usually describes a physical law or relation known to hold for the true values of the inputs and the true value of the output. For example, suppose that a

pendulum is used to measure the local acceleration of gravity g using the measurement equation $g = 4\pi^2\ell/T^2$, where ℓ denotes the length of the pendulum and T denotes its period of oscillation at that location. In practice, however, and owing to measurement errors, ℓ and T will differ from their true values λ and τ, hence so will g differ from its true value $\gamma = 4\pi^2\lambda/\tau^2$. In this case, furthermore, the very relationship between true values of inputs and output already involves an approximation, based on the assumption that the angular amplitude of the oscillation does not exceed a small angle, say, 0.1 rad, so that all but the leading term of what otherwise would be an infinite series may be neglected.[84]

In general, then, we have on the one hand a true relationship between true values, $\eta = \varphi(\xi_1, \ldots, \xi_n)$, and on the other its counterpart between measured values, $y = f(x_1, \ldots, x_n)$, where f may be (i) identical to φ, (ii) an approximation to φ, or (iii) a function of the observations underlying the $\{x_i\}$, rather than of the $\{x_i\}$ themselves that produces an estimate of the measurand satisfying the selected optimality criterion (that is, the criterion according to which the estimate is "best"). The following example illustrates this third possibility.

ILLUSTRATION: Consider measuring the area $\eta = \alpha\gamma$ of a rectangle of length α and width γ, based on n pairs of determinations of its length and width $(a_1, c_1), \ldots, (a_n, c_n)$ made under conditions of repeatability. The $\{a_i\}$ and the $\{c_i\}$ may be correlated. Suppose that these determinations are like a sample from a bivariate Gaussian (or, normal) distribution with mean vector (α, γ) and a positive-definite,[85] not necessarily diagonal, covariance matrix. Suppose also that the goal is to estimate the product $\alpha\gamma$ using an estimator that is "best" in the sense that it is unbiased and has minimum variance among all unbiased estimators. Since $\eta = \alpha\gamma$, and $\overline{a}$ and $\overline{c}$ are the "best" estimates of α and γ, it seems reasonable to regard $\overline{a}$ and $\overline{c}$ as input quantities, and to define $y = \overline{a}\,\overline{c}$ as estimate of η. However, $\overline{a}\,\overline{c}$ is a biased estimate of the area. The "best" estimate is $\overline{a}\,\overline{c} - \sum_{i=1}^{n}(a_i - \overline{a})(c_i - \overline{c})/(n(n-1))$, and this cannot be expressed as a function of $\overline{a}$ and $\overline{c}$ alone.

This illustration also shows that there is no guarantee that $y = f(x_1, \ldots, x_n)$ will be the best estimate of η, even if the $\{x_i\}$ are the best estimates of the $\{\xi_i\}$, however "best" may be defined. For example, when the goal is to achieve minimum mean squared error $\mathbb{E}(Y - \varphi(\xi_1, \ldots, \xi_n))^2$, where "$\mathbb{E}$" denotes expected value, or mathematical expectation (of a random variable), and φ is a non-linear function, as it is in the foregoing examples of the pendulum and of the area of a rectangle, generally $\mathbb{E}(\varphi(X_1, \ldots, X_n)) \neq \varphi(\xi_1, \ldots, \xi_n)$. (A real-valued function φ of a real variable is said to be linear if $\varphi(au + bv) = a\varphi(u) + b\varphi(v)$ for real numbers a, b, u, and v such that u, v, and $au + bv$ are in the domain of the function.)

This implies that the estimate of the true value of the output quantity obtained by merely substituting what should be true values of the inputs with measured values need not be the best estimate of the true value of the output quantity. For example, suppose that we wish to estimate $\exp(\xi)$ using $\exp(X)$, where X has a Gaussian distribution with mean ξ and standard deviation σ. In these circumstances, $\exp(X)$ has a lognormal distribution with mean $\exp(\xi + \sigma^2/2) > \exp(\xi)$. That is, $\exp(X)$ is biased high for $\exp(\xi)$.

The GUM 4.1.4 considers a situation where there are m independent, identically distributed replicates of the input quantities, $(x_{1,1}, \ldots, x_{1,n})$, $\ldots$, $(x_{m,1}, \ldots, x_{m,n})$. The goal is to estimate $f(\xi_1, \ldots, \xi_n)$ using either $y_A = \sum_{i=1}^{m} f(x_{i,1}, \ldots, x_{i,n})/m$ or $y_B = f(\overline{x}_1, \ldots, \overline{x}_n)$, where $\overline{x}_j = \sum_{i=1}^{m} x_{i,j}/m$ for $j = 1, \ldots, n$. The GUM suggests that y_A "may be preferable" to y_B. However, Wang and Iyer,[86] developing a suggestion made by Duane C. Boes, show that when $n = 2$ and $f(x_1, x_2) = x_1/x_2$, the preferable estimator (in the sense of having the smaller mean squared error) may be y_A or y_B, depending on whether the conditional variance of X_1 given $X_2 = x$ is proportional to x^2 or to x, respectively.

For the purpose of evaluating the uncertainty associated with the output quantity, the following assumptions are usually made: (i) the $\{x_i\}$ are regarded as observed values of random variables $\{X_i\}$; (ii) X_i has mean ξ_i and standard deviation $u(\xi_i)$ (called "standard uncertainty" and discussed in Sec. VII A) for $i = 1, \ldots, n$; (iii) the $\{X_i\}$ have some fully specified (joint) probability distribution, which describes both their individual (or, marginal) uncertainty, as well their stochastic inter-relations (for example, that when one of them takes a value above its mean, another tends to follow suit).

1. Example: Molecular weight of CO_2

The relative molecular mass (or, molecular weight) of carbon dioxide is $M_r(CO_2) = A_r(C) + 2A_r(O)$ (neglecting the minuscule mass deficiency attributable to the molecule's binding energy), where $A_r(C)$ and $A_r(O)$ denote the relative standard atomic masses (or, atomic weights) of carbon and oxygen. The output quantity, $M_r(CO_2)$, is a linear function of the input quantities, $A_r(C)$ and $A_r(O)$, in the sense that they are combined after multiplying each one by a scalar (1 and 2, respectively), and adding the results. We are interested in evaluating the uncertainty associated with $M_r(CO_2)$ in a "normal" material.

The IUPAC Commission on Isotopic Abundances and Atomic Weights (CIAAW) defines a "normal" material as any terrestrial material that "is a reasonably possible source for this element or its compounds in commerce, for industry or science; the material is not itself studied for some extraordinary anomaly and its isotopic composition has not been modified significantly in a geologically brief period."[87]

For CO_2, the "normal" material could be atmospheric air, volcanic gas, landfill emissions, or commercial tank gas, among many others. Because these different sources have different isotopic compositions,[88] the molecular weight of CO_2 in a particular sample can take any value within an interval of non-negligible width. The measurement uncertainty associated with $M_r(CO_2)$ is intended to capture the diversity of isotopic compositions of carbon and oxygen (this being the major source of uncertainty by far), as well as the uncertainty surrounding the relative atomic weights of the different isotopes of these elements.

2. Example: Pitot tube

A typical Pitot tube used to measure airspeed has an orifice facing directly into the air flow to measure total pressure, and at least one orifice whose surface normal is orthogonal to the flow to measure static pressure (Figure 1). Airspeed v is determined

FIG. 1. Pitot tube mounted on a helicopter (Zátonyi Sándor, en.wikipedia.org/wiki/Pitot_tube) showing one large, forward-facing, circular orifice to measure total pressure, and several small circular orifices behind a trim ring, to measure static pressure.

by the difference Δ between the total and static pressures, and by the mass density ρ of air, according to the measurement equation $v = \sqrt{2\Delta/\rho}$. In this example the output quantity, v, is a non-linear function of the input quantities, Δ and ρ, involving the square root of a ratio.

Since ρ is usually estimated by application of the ideal gas law, the measurement equation becomes $v = \sqrt{2\Delta R_s T/p}$, where p and T denote air pressure and temperature, and $R_s = 287.058\,\mathrm{J\,kg^{-1}K^{-1}}$ is the specific gas constant for dry air.

B. Observation equations

An *observation equation is a statistical model* that expresses the measurand as a known function of the parameters of the probability distribution of the inputs.

For example, in Sec. VI B 2 we model the lifetime of a component of an aircraft as an outcome of a Weibull random variable, which is characterized by two parameters: shape α and scale β. The measurand is the expected value of the lifetime, $\tau = \beta\Gamma(1/\alpha)$, where Γ denotes the gamma function of mathematical analysis.[89]

When the relationship between inputs and output is expressed via an observation equation, a statistical method, possibly involving a Monte Carlo procedure, may suffice to estimate the value of the measurand and to evaluate the associated uncertainty. The example concerning the performance of body armor, discussed in Secs. VI B 4 and VII B 4, illustrates this case.

The most demanding task when performing uncertainty evaluations in the context of observation equations is the development of a statistical model for the experimental data. Typically this requires close collaboration between a scientist and a statistician. Once the model has been defined and deemed to be adequate for the data, a choice needs to be made about how to estimate the adjustable parameters that determine the value of the measurand. The estimation of parameters in the model may be called *model calibration*.

The following examples introduce situations where observation equations should be used and describe the corresponding model-building exercises. Later on, in Sec. VII B, we will explain how these data and statistical models may be used to evaluate the uncertainty of the measurands involved.

1. Example: Thermal bath

The readings of temperature listed in Table I and depicted in Figure 2 were taken every minute with a thermocouple immersed in a thermal bath during a period of 100 min, to characterize the state of thermal equilibrium of the bath and estimate its mean temperature at the thermocouple location.

Owing to exchanges of thermal energy with the environment external to the bath, typically the temperature at any location will not remain constant, possibly due to convection currents in the bath. If the temperature does not drift but merely oscillates, however irregularly, around some central value, then the bath is deemed to be in equilibrium, and characterizing the state of thermal equilibrium amounts to describing the pattern and amplitude of such oscillations precisely.

The question may naturally be asked of why this situation is not amenable to modeling using a measurement equation where the output τ is the long-term mean temperature of the bath, the inputs are the $m = 100$ readings of temperature $t_1, \ldots, t_m$, and the measurement function f in $\tau = f(t_1, \ldots, t_m)$ is the arithmetic average.

TABLE I. Time series of temperature readings (expressed as deviations from 50 °C, all positive, therefore the first reading of temperature was 50.1024 °C) produced every minute by a thermocouple immersed in a thermal bath. The temporal order is from top to bottom between rows and from left to right within each row. Data kindly shared by Victor Eduardo Herrera Diaz (Centro de Metrología del Ejército Ecuatoriano, CMME, Quito, Ecuador) during an international workshop held at the Laboratorio Tecnológico del Uruguay (LATU, Montevideo, Uruguay) in March, 2013.

0.1024	0.1054	0.1026	0.1042	0.1026	0.1039	0.1065	0.1052	0.1067	0.1072
0.1054	0.1049	0.1082	0.1039	0.1052	0.1085	0.1088	0.1075	0.1085	0.1098
0.1070	0.1060	0.1067	0.1065	0.1072	0.1062	0.1085	0.1062	0.1034	0.1049
0.1044	0.1057	0.1060	0.1082	0.1052	0.1060	0.1057	0.1072	0.1072	0.1077
0.1103	0.1090	0.1077	0.1082	0.1067	0.1098	0.1057	0.1060	0.1019	0.1021
0.0993	0.1014	0.0965	0.1014	0.0996	0.0993	0.1003	0.1006	0.1026	0.1014
0.1039	0.1044	0.1024	0.1037	0.1060	0.1024	0.1039	0.1070	0.1054	0.1065
0.1072	0.1065	0.1085	0.1080	0.1093	0.1090	0.1128	0.1080	0.1108	0.1085
0.1080	0.1100	0.1065	0.1062	0.1057	0.1052	0.1057	0.1034	0.1037	0.1009
0.1009	0.1044	0.1021	0.1021	0.1029	0.1037	0.1049	0.1082	0.1044	0.1067

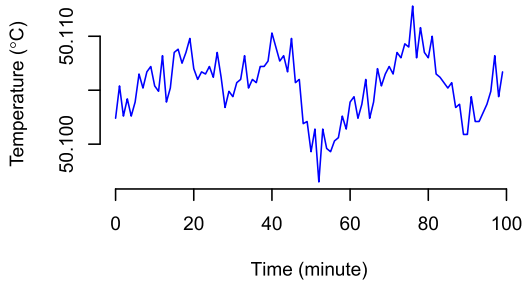

FIG. 2. Time series of temperature readings listed in Table I.

The reasons are that there is no scientific motivation for f to be of this form or of any other, and the fact that the quantity of interest being the mean temperature does not imply that f should be the arithmetic average of its arguments. In the example discussed in Sec. VI B 2, the measurand is the mean lifetime of an aircraft component, and yet it is not estimated by the arithmetic average of the observed lifetimes (and it would still not be the arithmetic average even if there were no censored observations).

The simplest measurement model, explaining how the different values of temperature observed under conditions of repeatability may be consistent with a single true mean temperature, is the additive measurement error model introduced in Sec. IV C: $t_i = \tau + \delta_i$, for $i = 1, \ldots, m$, where each δ_i is assumed to be the value of a random variable with mean zero.

Figure 3 shows estimates of the autocorrelation function (ACF) and of the partial autocorrelation function (PACF) of the temperature readings. The value of the ACF at lag h is the correlation between t_i and t_{i+h}; the corresponding value of the PACF at the same lag is the autocorrelation between t_i and t_{i+h} after adjusting t_i and t_{i+h} for their dependence on linear functions

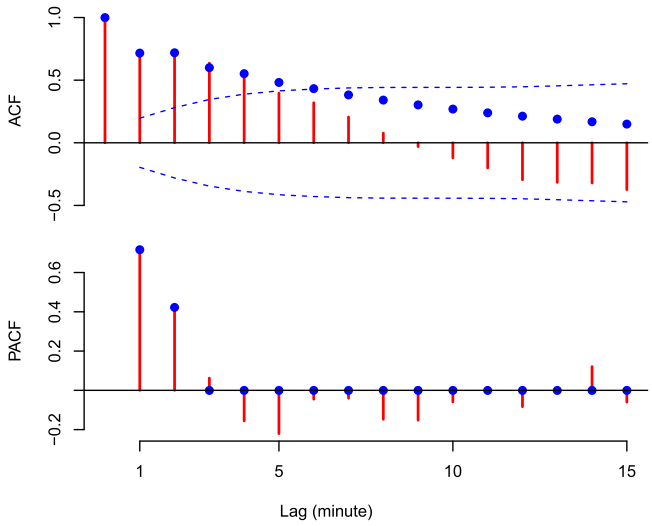

FIG. 3. Autocorrelation (ACF) and partial autocorrelation (PACF) functions for the time series of temperature readings listed in Table I. The (blue) dashed lines are approximate, 95% coverage intervals for the true ACF, by R function acf with the option `ci.type = "ma"` because, according to the function's documentation, in this way they "may be less potentially misleading" than if they were based on uncorrelated observations.[90] The blue dots are the values of the ACF and of the PACF corresponding to the autoregressive model of order 2 fitted to the data.

of $t_{i+1}, \ldots, t_{i+h-1}$. Both plots suggest that if a simple, additive measurement error model were to be used, the errors $\{\delta_i\}$ could not possibly be independent random variables.

These plots, together with conventional guidance for model selection,[91] suggest a different candidate model for these data: an auto-regression of order 2, often denoted AR(2). This is also the model that minimizes BIC (Sec. IV D) over a wide class of commonly used models for time series. Other model selection criteria, for example, AIC and AICC,[46] would suggest other models. Example E20 of the *Simple Guide*[7] considers one of these other models, but produces the same estimate of τ and the same evaluation of the associated uncertainty that the AR(2) model produces.

The AR(2) model, which is the observation equation selected for these data, has the following form: $t_i = \tau + \varphi_1(t_{i-1} - \tau) + \varphi_2(t_{i-2} - \tau) + \varepsilon_i$, where the $\{\varepsilon_i\}$ are assumed to be uncorrelated Gaussian random variables with mean 0 and standard deviation σ. According to this model, the correlations between the observations arise owing to a memory effect: the temperature at time t is determined by the values of temperature measured during the previous 2 min, plus a volatile (unpredictable, or "random") measurement error.

2. Example: F-100 component lifetime

A study of the lifetime of a component of the F-100 *Super Sabre* fighter aircraft,[92] which was built by North American Rockwell Corporation, involved testing 13 specimens of this component until the 10th failure was observed: their lifetimes were 0.22 h, 0.50 h, 0.88 h, 1.00 h, 1.32 h, 1.33 h, 1.54 h, 1.76 h, 2.50 h, and 3.00 h. Of the other three it is known only that their lifetimes exceeded 3.00 h: that is, these three are (right) *censored*, and the censoring is said to be of Type II (experiment terminates once a predetermined number of failures will have occurred).[93,94]

There are three challenges here: first, how to model the sampling variability of the values of lifetime observed in the experiment; second, how to define the measurand; third, how to extract the information that the censored data provide about this measurand and merge it with the information that the uncensored data provide.

In the original study,[92] the probabilistic model treated the thirteen values of lifetime $t_1, \ldots, t_{13}$ as a sample from a Weibull probability distribution[35] with unknown scale $\alpha > 0$ and shape $\beta > 0$.

Consistently with this definition of observation equation, the measurand will be a function of the two parameters. For example, the expected value of the Weibull distribution fitted to the data is $\tau = \beta\Gamma(1/\alpha)$, also called *mean time before failure* (MTBF), where Γ denotes the gamma function.

Other measurands may also be of interest: for example, the *reliable life*, which is the duration of the longest period of operation when the probability of failure remains below some given threshold. For these data, and as we shall explain next, the maximum likelihood estimate of the MTBF is 2.07 h. The 95% reliable life is 0.28 h, meaning that, with 95% probability, a component will not fail during the first 0.28 h of operation.

For the remainder of this example, including its continuation in Sec. VII B 2, we will focus on the MTBF τ as

the measurand. Thus, the task at hand consists of estimating α and β, computing the corresponding τ, evaluating the measurement uncertainties associated with α and β, and finally propagating these to evaluate the uncertainty associated with τ.

The MTBF is an attribute of the batch of components manufactured under the same conditions. The associated uncertainty, $u(\tau)$, may be reduced by increasing the size of the sample used to estimate τ. The user of a particular component may be more interested in the expected lifetime of this component than in the batch's MTBF, and in the uncertainty associated with this particular lifetime. Even though the expected lifetime of a particular component is the same as the MTBF for the batch, the associated uncertainty typically will be much larger. For these data, and as we shall see, $u(\tau) = 0.50$ h, but the standard uncertainty associated with the lifetime of a particular component is the standard deviation of the fitted Weibull distribution, 1.48 h in this case.

The question may and should be asked whether, in the absence of cogent substantive reason to adopt the Weibull model, probability distributions other than the Weibull could reasonably be entertained for these data. Two commonly considered alternatives are the lognormal and the gamma distributions. In this case, the lognormal achieves a considerably poorer fit to the data than the Weibull, but the gamma provides a tenable model for these data. In Sec. VII B 2 we will discuss how this model uncertainty may be recognized and propagated.

The Weibull model may be fitted to the data by the method of maximum likelihood.[35] Let $f_{\alpha,\beta}$ denote the corresponding probability density function, and $F_{\alpha,\beta}$ the cumulative distribution function[95] (Chap. 21). The maximum likelihood estimates (MLES) of α and β are the values that maximize the corresponding likelihood function, whose logarithm in this case is $\ell(\alpha, \beta) = \log((m_{\mathcal{U}} + m_{\mathcal{C}})!/m_{\mathcal{C}}!) + \sum_{i\in\mathcal{U}} \log f_{\alpha,\beta}(t_i) + \sum_{i\in\mathcal{C}} \log(1 - F_{\alpha,\beta}(t_i))$, where $\mathcal{U}$ corresponds to the $m_{\mathcal{U}} = 10$ uncensored observations, and $\mathcal{C}$ corresponds to the $m_{\mathcal{C}} = 3$ censored observations. Lawless[93] (Sec. 1.4.1a) explains how this expression follows from the design of this particular experiment (Type II censoring). The MLES of the Weibull parameters are $\widehat{\alpha} = 1.42$ and $\widehat{\beta} = 2.27$ h, hence the estimate of the expected lifetime is given by the measurement equation $\widehat{\tau} = \widehat{\beta}\Gamma(1/\widehat{\alpha}) = 2.07$ h.

3. Example: Thermistor calibration

Whetstone *et al.*[96] employed thermistor probes to measure the temperature of water flowing through a weighing apparatus used to measure coefficients of discharge of orifice plates and the temperature of the atmosphere surrounding the apparatus. The thermistors were calibrated by comparison with a platinum resistance thermometer (PRT) that had previously been calibrated by the Pressure and Temperature Division of what was then the National Bureau of Standards.

Table II lists the data used for the calibration of thermistor 775 008 from Table 17 of Whetstone *et al.*,[96] comprising readings taken simultaneously with the thermistor and the PRT immersed in a thermostatically controlled thermal bath filled with mineral oil.

TABLE II. Values of temperature of a thermostatically controlled bath measured simultaneously by a calibrated PRT and by a thermistor.

	Temperature (°C)					
PRT	20.91	25.42	30.50	34.96	40.23	34.93
Thermistor	20.85	25.52	30.70	35.22	40.47	35.18
PRT	30.05	25.03	20.87	16.41	16.40	39.34
Thermistor	30.25	25.10	20.81	16.23	16.22	39.56

Calibrating the thermistor amounts to building a *calibration function* φ that maps values of temperature indicated by the PRT to values of temperature indicated by the thermistor. The function characterizes how the thermistor responds when immersed in a medium at the temperature indicated by the PRT that acts as a reference.

In practice, however, the thermistor produces a reading of the temperature of the medium that it is immersed in and with which it is assumed to be in thermal equilibrium. This is done by reading the temperature that the thermistor produces, and then applying to this reading the inverse $\psi = \varphi^{-1}$ of the calibration function. In the context of gas analysis,[97] this function is called the *analysis function*. In the context of calibration of force transducers, Bartel, Stoudt, and Possolo[83] call the corresponding function *measurement function*.

The question may naturally be asked why not build ψ directly, given that it is the function needed to use the thermistor in practice, instead of determining φ first, and then inverting it to obtain ψ. There are two reasons why. The first is a common legal requirement: the calibration must characterize the response of the device being calibrated to known inputs, hence the need for φ. The second is statistical and practical: the temperature t indicated by the thermistor is modeled as $t = \varphi(\tau)+\delta$, where τ denotes the value of temperature measured by the PRT, and the uncertainty associated with τ is much smaller (almost 10 times smaller) than the uncertainty associated with t that is attributed to measurement error δ. In these circumstances, estimating φ is a standard regression problem.

In many calibrations, the uncertainties associated with the reference values are not negligible by comparison with the uncertainties associated with the instrumental indications. In such cases, estimating φ or ψ should be done via errors-in-variables regression.[98,99] This statistical procedure is used for value assignment to gas mixture reference materials[97,100] and in the calibration of force transducers.[83]

The thermistor calibration data are $m = 12$ pairs of temperature values $(\tau_1, t_1), \ldots, (\tau_m, t_m)$ measured simultaneously by the PRT and by the thermistor. Since there is no substantive reason to choose any particular functional form for the measurand φ, we will select it from among low-order polynomials, as described next.

To the generic regression model introduced in Sec. IV C, $t_i = \varphi(\tau_i) + \delta_i$, we will now add the assumption that the measurement errors $\{\delta_i\}$ are like a sample from a Gaussian distribution with mean 0 and unknown standard deviation σ. The adequacy of this assumption should be evaluated by examining a QQ-plot[101] of the regression residuals, and a plot of the residuals against the fitted values, to check that there are

no indications of violation of the assumption that the same σ indeed seems to apply across the whole range of the variables in play. Under this assumption, the maximum likelihood estimates of the coefficients of the polynomial φ are the ordinary least squares estimates.

We will entertain polynomials of degree $p = 1, 2, \ldots, 6$ as candidates for the calibration function. To select the "best" degree, we rely on three criteria: (i) the second-order version of Akaike's Information Criterion (AICC)[46] as implemented in R function AICc defined in package AICcmodavg;[102] (ii) graphical examination of the residuals from the candidate regression models; and (iii) comparison of the alternative models using the analysis of variance technique implemented in R function anova.[103] Generally, the smaller the AICc, the better the model.

All three model selection criteria single out a cubic polynomial as achieving the best compromise between goodness-of-fit (that is, small residuals) and model parsimony (that is, small number of adjustable parameters). Figure 4 shows the results of (i) and (ii). Hence, $\varphi(\tau) = \beta_0 + \beta_1\tau + \beta_2\tau^2 + \beta_3\tau^3$. The least squares estimates of the coefficients are $\widehat{\beta}_0 = -0.2785$ °C, $\widehat{\beta}_1 = 0.9722$ °C^{-1}, $\widehat{\beta}_2 = 0.002\,773$ °C^{-2}, and $\widehat{\beta}_3 = -4.404 \times 10^{-5}$ °C^{-3}. All except the intercept $\widehat{\beta}_0$ are statistically significantly different from 0. (These coefficients differ from the coefficients in Table 18 of Whetstone *et al.*[96] because the latter pertain to a polynomial of the 3rd degree fitted to the $\{\tau_i\}$ as a function of the $\{t_i\}$).

Conventional graphical diagnostics — a plot of estimated residuals $\{\widehat{\delta}_i\}$ against fitted values $\{\widehat{t}_i\}$, and a QQ-plot of the residuals — reveal no obvious inadequacy of the model for these data. Since the temperatures indicated by the thermistor during calibration ranged from 16.40 °C to 39.43 °C, the calibration function should not be trusted for values outside this range.

A polynomial function is not necessarily invertible. Figure 5 depicts the fitted calibration function, and the corresponding analysis function, and confirms that this polynomial approximation to φ increases monotonically over the range of τ, hence is invertible. The inverse of a polynomial is not a polynomial, however, and generally there will be no closed-form expression for $\psi = \varphi^{-1}$. For use in practice, values of ψ are determined at a set of t values chosen based on the intended use of the instrument, say at every half degree Celsius, and then an interpolant may be developed, a spline, for example, to allow computing $\psi(t)$ for any value of t within the calibration range. (A spline is a smooth, piecewise polynomial function that passes through all of the given points without any unwieldy oscillations between them.[104,105])

To find the calibrated value τ of temperature that corresponds to a reading t made by the thermistor involves solving the following equation for τ: $\widehat{\beta}_0 + \widehat{\beta}_1\tau + \widehat{\beta}_2\tau^2 + \widehat{\beta}_3\tau^3 = t$. Of the three solutions (generally complex numbers) that this equation will have, we select a real solution approximately within the calibration range (16 °C to 40 °C in this case), provided such solution exists.

For example, if $t = 27.68$ °C, computing $\psi(t)$ involves solving the cubic equation $-0.2785 + 0.9722\tau + 0.002\,773\tau^2 - 0.000\,044\,04\tau^3 = 27.68$. Of the three roots of this equation, 27.54 °C, −135.14 °C, and 170.57 °C, only the first is within the calibration range. Even though the roots of polynomials of degree up to 4 may be determined algebraically (that is, using radicals and arithmetic operations), in practice the roots of polynomials of degree greater than 2 are usually determined using a numerical solver, for example, polyroot in R.

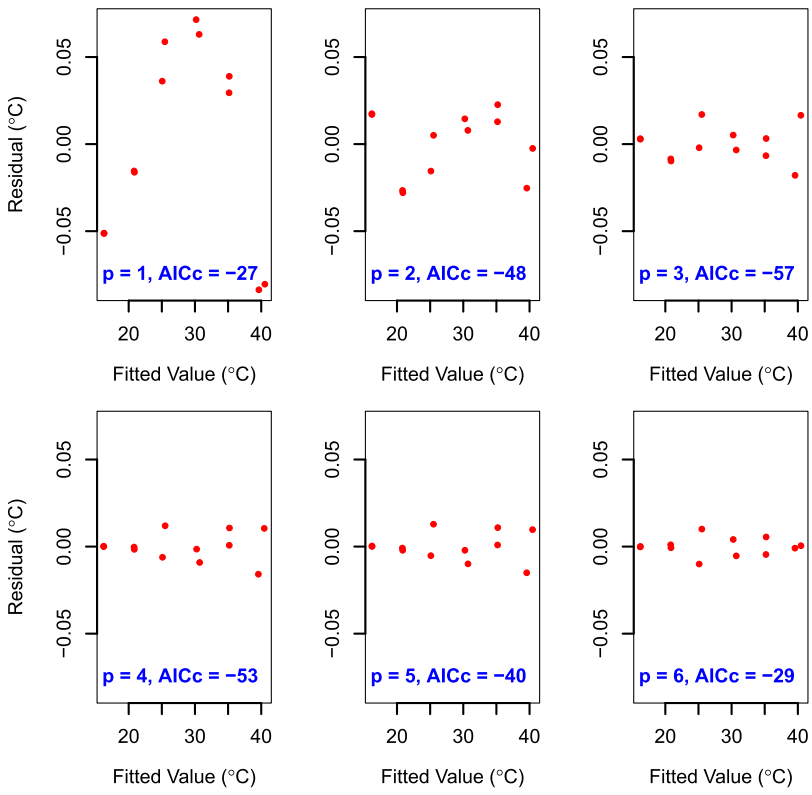

FIG. 4. Residuals and values of AICC for polynomials of degree $p = 1, 2, \ldots, 6$ that are candidate models for the thermistor calibration function φ. The AICC achieves a minimum when $p = 3$, and the scatter (in the vertical direction) of the residuals is not reduced appreciably for larger values of p.

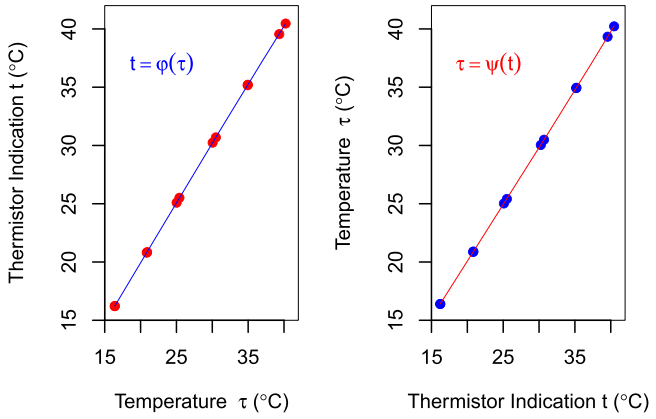

FIG. 5. Left panel: Selected and fitted cubic polynomial that approximate the thermistor calibration function φ. Right panel: Analysis function that produces the value $\tau = \psi(t)$ of temperature corresponding to an indication t produced by the thermistor. The two functions appear to be identical only because t and τ have the same units of measurement and corresponding values are numerically close to one another.

4. Example: Ballistic limit of body armor

The *ballistic limit* v_{50} of a particular type of bullet-proof vest, and for a particular type of bullet, is the bullet velocity at which the probability of perforation is 50%. Similar to the mean time before failure of a component of the F-100 *Super Sabre*, discussed in Sec. VI B 2, the ballistic limit is an attribute of a lot of vests manufactured under identical conditions. To measure v_{50}, several physically identical bullets are fired with different velocities at identical vests under standardized conditions, for example, as specified by OLES.[106] For each bullet, the result is recorded as a binary (nominal) outcome indicating whether the vest stopped the bullet or not. The inputs for the measurement model are bullet velocity and this binary outcome.

The measurement results, depicted in Figure 6, for a particular type of Zylon vest, from a test conducted at NIST[107]

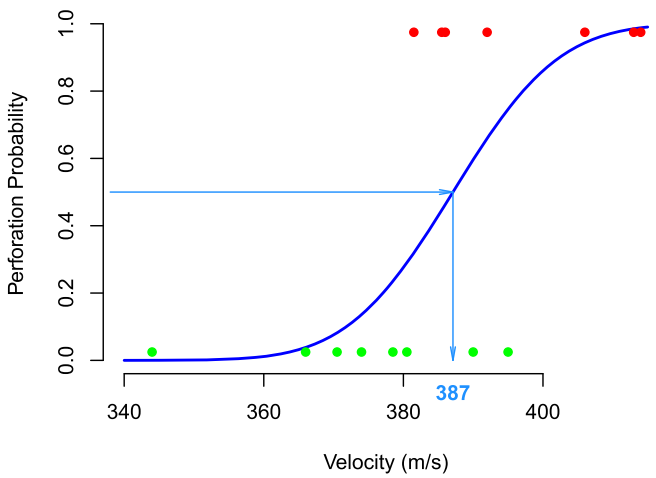

FIG. 6. Probit regression model fitted to the results of a test to measure the ballistic limit of a particular type of bullet-proof vest. The red dots indicate the bullet velocities that achieved perforation, and the green dots those that the vest stopped. The blue arrows indicate how to find the velocity that corresponds to a 50% penetration probability.

that involved $m = 15$ bullets, are (344, 0), (380.5, 0), (406, 1), (386, 1), (366, 0), (381.5, 1), (374, 0), (385.5, 1), (370.5, 0), (395, 0), (414, 1), (390, 0), (413, 1), (392, 1), (378.5, 0). The first value in each pair is the bullet velocity (expressed in m/s), and the second indicates whether the bullet did (1) or did not (0) penetrate the vest.

The measurement model for v_{50} involves an observation equation and a measurement equation. The observation equation is a Bernoulli model for bullet perforation, which states that the results from different shots are like outcomes of independent tosses of different coins, where the coin corresponds to a bullet of velocity v having probability $\pi(v)$ of "heads," denoting perforation. The probability of penetration in turn is assumed to be of the probit form: $\pi(v) = \Phi(\alpha + \beta v)$, where Φ denotes the cumulative probability distribution function of a Gaussian distribution with mean 0 and standard deviation 1, and α and β are parameters to be estimated. The measurement equation is $v_{50} = -\alpha/\beta$, which follows from $\Phi^{-1}(0.5) = 0$ (cf. with the measurement equation mentioned near the end of Sec. VI B 2).

Fitting the model to the data by the method of maximum likelihood as implemented in R function glm produces $\widehat{\alpha} = -32$ and $\widehat{\beta} = 0.084$ s/m, hence $\widehat{v}_{50} = -\widehat{\alpha}/\widehat{\beta} = 387$ m/s. Figure 6 depicts the data and the fitted probit regression function.

In this case, too, alternative models may reasonably be entertained, and the sensitivity of the results to the choice of model should be assessed. Differences between the results that correspond to different models are a quantifiable contribution to measurement uncertainty attributable to model uncertainty. For example, if, in the model for $\pi(v)$, a logit had been used instead of a probit, the estimate of v_{50} would still have been the same 387 m/s. The AICC criterion,[46] also used in Sec. VI B 3, gives an edge to the probit over the logit in this case.

5. Example: PCB in sediment

A key *comparison* is an interlaboratory study defined and organized by a Consultative Committee of the International Committee for Weights and Measures (CIPM), involving national metrology institutes of different countries, to test the principal techniques and methods in a particular field of measurement science.[108] Key comparison CCQM–k25 was organized by the Consultative Committee for the Amount of Substance (Metrology in Chemistry and Biology, CCQM), to compare the results of the determination of the mass fractions of five different polychlorinated biphenyl (PCB) congeners in sediment.[109]

Only the measurement results produced by $n = 6$ participating laboratories were deemed suitable (owing to substantive reasons detailed in Ref. 109) for inclusion in the calculation of a consensus value for the mass fraction of PCB 28 (2,4,4′-trichlorobiphenyl) (measurand). Table III lists the selected results, and Figure 7 displays them, the consensus value, and the associated standard uncertainty.

The data are the measured values $w_1, \ldots, w_m$ of mass fraction, the associated standard uncertainties $u(w_1), \ldots, u(w_m)$, and the numbers of degrees of freedom $\nu_1, \ldots, \nu_m$ that the standard uncertainties are based on.

TABLE III. Measured values w_j of the mass fraction (ng/g) of PCB 28 obtained in CCQM–K25, standard uncertainties $u(w_j)$, and numbers of degrees of freedom ν_j that these standard uncertainties are based on, in CCQM–K25.

Lab	w_j (ng/g)	$u(w_j)$ (ng/g)	ν_j
IRMM	34.30	1.03	60
KRISS	32.90	0.69	4
NARL	34.53	0.83	18
NIST	32.42	0.29	2
NMIJ	31.90	0.40	13
NRC	35.80	0.38	60

The notion of degrees of freedom, as it is used in this context, often is a source of confusion. When $u(w_j)$ is the result of a Type A evaluation of uncertainty (reviewed in Sec. VII), the corresponding ν_j is a function of the number of observations used to compute $u(w_j)$. For example, when $u(w_j)$ is the standard deviation of the sampling distribution of the average of a set of m_j observations obtained under repeatability conditions (VIM 2.20), then $\nu_j = m_j - 1$.

But when, as is most often the case, $u(w_j)$ combines the results of Type A and Type B evaluations of several different components of uncertainty, any assignment of value to ν_j is likely to prove controversial. The GUM G.4.2 provides useful guidance for how to assign an effective number of degrees of freedom to a standard uncertainty resulting from a Type B evaluation.

In Sec. VII B 5, the numbers of degrees of freedom listed in Table III are used in a Monte Carlo method for uncertainty evaluation that recognizes the limited amount of information that the laboratory-specific uncertainty evaluations are based on.

The standard deviation of the measured values $\{w_j\}$ listed in Table III equals 1.48 ng/g, while the laboratory-specific standard uncertainties are in the range 0.29 ng/g to 1.03 ng/g, and their median is 0.545 ng/g. Therefore, the measured values are almost 3 times more dispersed than the typical, within-laboratory standard uncertainty. This "excess" variance is often interpreted as suggesting that the laboratories have failed to identify and evaluate one or more important sources of uncertainty, whose combined contribution Thompson and Ellison[111] have called *dark uncertainty*.

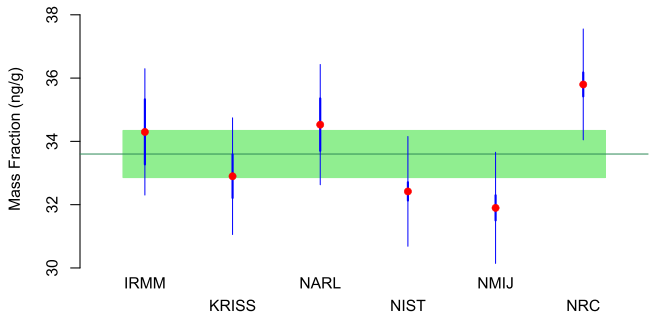

FIG. 7. Measurement results for the mass fraction (ng/g) of PCB 28 obtained in CCQM–K25. Each large (red) dot represents the value w_j measured by a participating laboratory; the thick, vertical (blue) line segment depicts $w_j \pm u(w_j)$; and the thin, vertical line segment depicts the corresponding uncertainty including the contribution from dark uncertainty, estimated as $\hat{\tau} = 1.71$ ng/g, extending to $w_j \pm (\hat{\tau}^2 + u(w_j)^2)^{\frac{1}{2}}$. The thin, horizontal (dark green) line marks the consensus value $\hat{\mu} = 33.6$ ng/g, and the shaded (light green) band around it represents $\hat{\mu} \pm u(\mu)$, where $u(\mu)$ is the standard uncertainty computed by R function rma defined in package metafor.[110]

A laboratory random effects model is capable of capturing and expressing this contribution quantitatively.[112–114] The model has many variants, and a long history of use and proven utility, often being referred to as a variance components model for the analysis of variance.[115] The version we shall use (which is the observation equation for this example) represents the value of mass fraction measured by each laboratory as $w_j = \mu + \lambda_j + \varepsilon_j$ for $j = 1, \ldots, n$, where $n = 6$ is the number of laboratories, μ denotes the true value of the measurand, $\lambda_1, \ldots, \lambda_n$ represent the laboratory effects, and $\varepsilon_1, \ldots, \varepsilon_n$ represent measurement errors.

In many applications (and for the uncertainty evaluation described in Sec. VII B 5), the $\{\lambda_j\}$ are assumed to be a sample from a Gaussian distribution with mean 0 and standard deviation τ, and the $\{\varepsilon_j\}$ are assumed to be outcomes of Gaussian random variables with mean 0 and standard deviations $\{\sigma_j\}$. The $\{u(w_j)\}$ are regarded as estimates, or imperfect evaluations of the $\{\sigma_j\}$. The random effects model may also be used under different assumptions.[116,117]

In all cases, diagnostics should be examined that provide clues about the violation of any assumptions that are made. Useful graphical diagnostics include QQ-plots of the predicted values of the random effects $\{\widehat{\lambda}_j\}$, and of the estimated residuals $\{\widehat{\varepsilon}_j\}$. We also recommend that the sensitivity of the results be evaluated by comparing results produced by different methods for fitting the random effects model to the data.

If the data were only the $\{w_j\}$, it would not be possible to distinguish the laboratory effects $\{\lambda_j\}$ from the measurement errors $\{\varepsilon_j\}$. As it is, we know that the absolute values of the $\{\varepsilon_j\}$ are generally comparable to the $\{u(w_j)\}$ and conclude that any "excess variance" the $\{w_j\}$ may exhibit is attributable to the $\{\lambda_j\}$, whose scatter is gauged by the standard deviation τ of their common probability distribution.

DerSimonian and Laird[118] suggested the procedure most widely used in meta-analysis to fit this type of model. The estimate of μ (*consensus value*), $\widehat{\mu}_{\mathrm{DL}} = 33.6$ ng/g, is a weighted average of the values measured by the participating laboratories, with weights inversely proportional to $\{\tau^2 + \sigma_j^2\}$. Since both τ and the $\{\sigma_j\}$ are unknown, they are substituted by estimates, $\widehat{\tau}_{\mathrm{DL}}$ and $\widehat{\sigma}_j = u(w_j)$, for $j = 1, \ldots, n$.

The estimate of τ is obtained by equating observed and expected values of a particular function of the measurement results. The resulting $\widehat{\tau}_{\mathrm{DL}}$ is then used in the definition of the weights as if it were not surrounded by uncertainty, which is an obvious weakness of the procedure (addressed in Sec. VII B 5). For the PCB 28 data, $\widehat{\tau}_{\mathrm{DL}} = 1.71$ ng/g. The DerSimonian-Laird procedure is implemented in R function rma, defined in package metafor:[110] its main results are depicted in Figure 7.

The fact that the estimate of τ is about three times larger than the median of the $\{u(w_j)\}$ corroborates the reality and importance of dark uncertainty in this case. The random effects model provides the technical machinery necessary to recognize and propagate this contribution, as we shall explain in Sec. VII B 5.

6. Example: Forensic glass

Evett and Spiehler[119] describe and demonstrate the use of a reference collection of glass fragments in forensic investigations, which was assembled by the Birmingham laboratory of the Home Office Forensic Science Service (FSS) of the United Kingdom. (The FSS, which used to have seven main laboratories across England and Wales, was closed at the end of March, 2012.)

The collection comprised 214 glass fragments of known provenance and included measured values of the refractive index and of the mass fractions of oxides of the major elements (Na, Mg, Al, Si, K, Ca, Ba, Fe). This is the *Glass Identification Data Set* in the Machine Learning Repository of the University of California at Irvine,[120] also available in object glass of the R package mda,[121] which is used in Example E29 of the *Simple Guide*.

The provenance of the glass fragments in this collection is as follows (with the number of fragments in each class between parentheses): float processed building windows (70), non-float processed building windows (76), float processed vehicle windows (17), containers (13), tableware (9), and headlamps (29). Contemporary windows are made of glass that, in its molten state, is poured onto a bed of molten metal (hence floated): this process, developed in the 1950s, produces glass sheets of very uniform thickness and superior flatness.[122]

Since it is possible to determine the refractive index and chemical composition of even a very small glass fragment as may be found in the clothing of a suspect, a forensic investigation may compare these attributes against a reference collection like the one mentioned above, to gain insight into the provenance of the fragment. In this example the measurand is the class (among the six listed above) that the fragment may belong to.

The classifier that we shall use to assign glass fragments to the six classes models the $n = 9$ quantitative inputs (refractive index and mass fractions of major oxides) as an outcome of an n-dimensional random vector whose probability distribution is a mixture of multivariate Gaussian distributions described below. Many other classifiers could be used for the same purpose:[121] the results likely depend markedly on the modeling choice that is made.

The reference collection comprises fragments belonging to six classes. Given that a fragment belongs to class j, the probability distribution of the vector of its inputs is a mixture of k multivariate Gaussian distributions with different mean vectors but the same covariance matrix. That is, each class is represented by a union of k sub-classes all with the same hyper-ellipsoidal shape, centered at different points in the n-dimensional space of inputs: this model can accurately track and distinguish classes whose representative sets in n-dimensional space have complex shapes.

This model for mixture discriminant analysis[123,124] is implemented in R function mda, which is defined in the package of the same name.[121] Given values of the refractive index and of the mass fractions of the major oxides for a glass fragment of unknown provenance, the calibrated (or, trained) model computes the probability that the fragment belongs to each of the six classes represented in the reference collection, and the fragment is then assigned to the class to which it is most likely to belong.

For example, for a fragment with refractive index 1.516 13, and mass fractions of the major oxides 13.92 cg/g (Na), 3.52 cg/g (Mg), 1.25 cg/g (Al), 72.88 cg/g (Si), 0.37 cg/g (K), 7.94 cg/g (Ca), 0 cg/g (Ba), 0.14 cg/g (Fe), the model fitted with $k = 4$ produces the following probabilities of class membership: float processed building windows, 0.29; non-float processed building windows, 0.64; vehicle windows, 0.075; the other classes having negligible probabilities. Therefore, the fragment is deemed to be from a building window, more likely from an old building than from a modern building.

As implemented, the model is determined by the number of sub-classes k per class. To choose the best value for k, we employed 10-fold cross-validation (refer to Wasserman[36] Sec. 22.8 for background information about this statistical technique). First, we partitioned the data into 10 subsets $\mathcal{S}_1, \ldots, \mathcal{S}_{10}$, each comprising about 10% of the data. Then we repeated the following steps for $r = 1, \ldots, 10$ and for each number $k = 1, \ldots, 5$ of sub-classes per class.

1. We built a mixture discriminant model D_r with k sub-classes for each of the six classes of glass, using all the data except those in $\mathcal{S}_r$;
2. We used D_r to predict the classes of the glass samples in $\mathcal{S}_r$, and computed the corresponding error rate, defined as the number of incorrect predictions divided by the number of glass samples in $\mathcal{S}_r$.

This process produced 10 error rates (one for each subset left aside) for each value of k. The averages of these decuplets of rates were 36%, 35%, 33%, 29%, and 33% for the models with $k = 1, 2, 3, 4, 5$ sub-classes, respectively. Therefore, the model with $k = 4$ sub-classes per class is selected as the most accurate, with 29% error rate.

The performance of the selected classifier model may be evaluated in greater detail using *leave-one-out cross-validation*,[47] as follows: for each glass fragment in turn, build a classifier using the data for the other fragments only, and use it to predict the type of the glass fragment left out. The overall error rate, which is estimated by the proportion of glass fragments that were misclassified, turned out to be 31%. This result is comparable to the result of the 10-fold cross-validation used during model selection. Leave-one-out cross-validation could have been used during model selection instead of 10-fold cross-validation, but the corresponding computations would have been considerably more time-consuming. A naive evaluation of the error rate, based on how accurately the model fitted to the data for all 214 glass fragments predicts their class memberships, produces an overoptimistic 21%.

VII. EVALUATING MEASUREMENT UNCERTAINTY

Several different techniques may be used to evaluate the measurement uncertainty of inputs and outputs of measurement models. The choice of technique depends on (i) the measurement model (in particular whether it is a measurement equation or an observation equation), (ii) the intended use of the measurement result (measured value and

associated uncertainty together), and (iii) the resources available to carry out the uncertainty propagation exercise. Most of the preparatory work goes into defining the measurement model, and producing estimates and uncertainty evaluations for the model inputs.

The evaluations of uncertainty associated with inputs of measurement models may produce only standard uncertainties, coverage regions, or fully specified probability distributions that provide a complete description of the associated uncertainty. (Monte Carlo uncertainty evaluations require that the uncertainty associated with the inputs be expressed in the form of fully specified probability distributions.)

These evaluations are often classified into *Type A* or *Type B* depending on how they are performed:

- *Type A* evaluations involve the application of statistical methods to experimental data, consistently with a measurement model;
- *Type B* evaluations involve the elicitation of expert knowledge (from a single expert or from a group of experts, also from authoritative sources including calibration certificates, certified reference materials, and technical publications) and its distillation into probability distributions, or into summaries thereof like standard uncertainties or coverage regions, that describe states of knowledge about the true values of the inputs.

Taylor and Kuyatt[6] discuss this classification in Sec. 3, the GUM in Sec. 4.2, and Possolo[7] in Sec. 5. Unfortunately, the meaning of these types is much too often misunderstood or misapplied. Therefore, it seems preferable to state the original source of the uncertainty evaluation explicitly in each case: for example, experimental data (even if more than one step removed from the immediate source of the uncertainty evaluation), meta-analysis,[125] literature survey, expert opinion, or mere guess.

A commonly used classification of measurement errors (hence of uncertainty components) into *random* or *systematic* is problematic because it involves a judgment about the essence of the errors and presupposes that there is a widely accepted, shared understanding of the meaning of these terms, when in fact there is not. It seems preferable to classify uncertainty components according to the behavior of their effects, as being either *persistent* or *volatile*. For example, when calibrating a force transducer, the orientation of the transducer relative to the loading platens of the deadweight machine is a persistent effect because a change in such orientation may shift the transducer's response up or down at all set-points of the applied force, by unknown and possibly variable amounts, but all in the same direction, as illustrated in Figure 5 of Bartel.[126]

An alternative classification of uncertainty evaluation modalities recognizes and distinguishes *bottom-up* and *top-down* evaluations.

- Bottom-Up evaluations involve (i) the complete enumeration of all relevant sources of uncertainty, (ii) a description of how they contribute to the uncertainty of the measurement result, which may be depicted in a cause-and-effect diagram[81] (Appendix D), and (iii) the quantification of the contributions they make to the uncertainty of the result. These elements are often summarized in an uncertainty budget.
- Top-Down evaluations are based on inter-comparisons of measurement results for the same measurand obtained by different laboratories, or by different scientists at the same laboratory, in either case all working independently, employing the same or different measurement methods. Interlaboratory studies and comparisons with a standard provide evaluations of measurement uncertainty that do not rely on the identification and characterization of the underlying sources of uncertainty.

Even though Gaussian and rectangular probability distributions are very commonly used to describe the uncertainty associated with inputs to measurement models, any other probability distribution is a potential model for the uncertainty associated with an input or output. The example discussed in Sec. VII B 2 uses distributions other than Gaussian or rectangular. Possolo and Elster[127] and Possolo[7] provide guidance for how to assign probability distributions to input quantities.

Automatic methods for assignment of distributions to inputs (for example, "rules" based on maximum entropy considerations) should be avoided. In all cases, the choice should be the result of deliberate model selection exercises, informed by specific knowledge about the inputs, and taking into account the pattern of dispersion apparent in relevant experimental data. The advice that the GUM (3.4.8) offers is particularly relevant here: that any framework for assessing uncertainty "*cannot substitute for critical thinking, intellectual honesty and professional skill. The evaluation of uncertainty is neither a routine task nor a purely mathematical one; it depends on detailed knowledge of the nature of the measurand and of the measurement*."

A. Uncertainty evaluation for measurement equations

For a measurand that appears as output in a measurement equation model, $y = f(x_1, \ldots, x_n)$, and when the (joint) probability distribution of the random variables modeling the inputs does not depend on y, evaluating the uncertainty associated with y reduces to computing the probability distribution of a function of n random variables, or some suitable summary of the corresponding dispersion of values.

This evaluation may be undertaken (i) exactly and in closed form (but only in rare, particularly simple cases) employing the change-of-variable formula,[38,128] (ii) using a locally linear (or locally quadratic) approximation to f, which is the approach described in the GUM, or (iii) by application of a Monte Carlo method, described by Morgan and Henrion[129] and used in the GUM-S1 for scalar measurands, and in the GUM Supplement 2 (GUM-S2)[130] for multivariate measurands.

1. Change-of-variable formula

When there is a single input quantity x that is modeled by a random variable with probability density p_X, and the measurement equation is $y = f(x)$, where f is either increasing or decreasing (but not both) on the range of the input quantity, and f's inverse g (such that $g(f(x)) = x$ for all values of x) has a continuous first derivative $\dot{g}$, then the probability density of y

is p_Y such that $p_Y(y) = p_X[g(y)]|\dot{g}(y)|$, for all values of y, where $|\dot{g}(y)|$ denotes the absolute value of the derivative of g at y.[128]

ILLUSTRATION: Ballistic chronographs are used to measure bullet velocity in studies of the performance of ballistic armor ("bullet-proof" vests)[131] (cf. Secs. VI B 4 and VII B 4). The relative standard measurement uncertainty of typical commercial ballistic chronographs is 0.5%. Suppose that the uncertainty associated with the muzzle velocity of a 9 mm bullet is described by a Gaussian probability distribution with mean $\mu = 410$ m/s and standard deviation $\sigma = 2$ m/s, and that the mass of the bullet is $m = 7.45$ g with negligible uncertainty. The change-of-variable formula given above may be used to derive the probability density of the kinetic energy $E = \frac{1}{2}mv^2$ of the bullet because the function that maps the velocity to the energy is increasing for all positive values of the velocity. Since $v = \sqrt{2E/m}$ and the derivative of v with respect to E is $1/\sqrt{2mE}$, the probability density of the kinetic energy takes the value $\exp(-(\sqrt{E} - \mu\sqrt{m/2})^2/(m\sigma^2))/(2\sigma\sqrt{m\pi E})$ at E. As it turns out, for the particular values of μ, σ, and m given above, this density is approximated very closely by the density of a Gaussian distribution with mean 626 J and standard deviation 6 J.

The multivariate change-of-variable formula can be used to derive the probability density analytically of an output quantity that is a function of $n > 1$ input quantities with non-negligible uncertainties:[38] for example, in the illustration described above, if not only E but also m had positive standard uncertainties. However, since this derivation involves an $(n-1)$ -dimensional integration, it often is impracticable. However, a Monte Carlo method provides an easy workaround because it draws an arbitrarily large sample from the probability distribution of the output quantity without computing this distribution explicitly.

2. Gauss's formula

The following formula, appearing as Equation (13) in the GUM, may be used when the function f is differentiable and both the inputs and the output are quantities. It provides an approximation to the standard deviation (*standard uncertainty*) $u(y)$ of the probability distribution of the output quantity as a function of the standard uncertainties $\{u(x_i)\}$ and (Pearson's product-moment) correlation coefficients $\{r_{ij}\}$ of the input quantities,[132]

$$u^2(y) \approx \sum_{i=1}^{n} c_i^2 u^2(x_i) + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} c_i c_j u(x_i) u(x_j) r_{ij}. \quad (1)$$

The *sensitivity coefficient* $c_i = \partial f(x_1, \ldots, x_n)/\partial x_i$ denotes the value at $(x_1, \ldots, x_n)$ of the first-order partial derivative of f with respect to the ith input quantity x_i, for $i = 1, \ldots, n$. The name *sensitivity coefficient* is apt because c_i expresses how the value of the output quantity varies with changes in the value of x_i: its measurement unit is the ratio of the measurement unit of y to the measurement unit of x_i. More importantly, the absolute value of c_i quantifies how f either amplifies (if $|c_i| > 1$) or mutes (if $|c_i| < 1$) the effect of the uncertainty associated with x_i, upon the uncertainty associated with the output quantity.

When the random variables modeling the inputs all are uncorrelated ($r_{ij} = 0$ for all $1 \le i < j \le n$), the formula reduces to $u^2(y) \approx \sum_{i=1}^{n} c_i^2 u^2(x_i)$, the form in which it was introduced by Gauss.[133]

The approximation defined in Equation (1) is just part of the more rigorously formulated, and farther reaching *Delta Method of* probability theory, which Casella and Berger[128] formulate as their Theorem 5.5.28. The quality of the approximation depends on how close to linear the function f may be in a neighborhood of $(x_1, \ldots, x_n)$ sufficiently large to include most of the "mass" of the joint probability distribution of the input quantities.[134]

Kline and McClintock[135] use Gauss's formula and much of the technical apparatus that only much later would be gathered into the GUM. Mandel[136] calls "derived quantity" what the GUM calls "output quantity," and illustrates the use of this same "law of propagation of errors," suggesting that it may be reliably used when the relative measurement uncertainties associated with the input quantities, $\{u(x_i)/x_i\}$, are no larger than 10%.

When the output quantity is a ratio of products of uncorrelated input quantities, or, more generally, $y = \kappa x_1^{\alpha_1} \ldots x_n^{\alpha_n}$, where κ and the $\{\alpha_i\}$ are (positive or negative) constants, then Equation (1) takes a particularly simple, easy-to-remember form,

$$(u(y)/y)^2 \approx (\alpha_1 u(x_1)/x_1)^2 + \cdots + (\alpha_n u(x_n)/x_n)^2. \quad (2)$$

The probabilistic interpretation of $u(y)$ depends on the probability distribution of y and how $u(y)$ will have been evaluated. In many cases, but by no means in all cases, $y \pm 2u(y)$ is an approximate 95% coverage interval for the true value of y, the more accurate the approximation the closer to Gaussian the probability distribution of y is. However, the approximation may be surprisingly good even when the distribution of y is markedly skewed (or, asymmetrical) with the right tail (of the corresponding probability density) heavier than the left tail.[34]

The *NIST Uncertainty Machine*, available at uncertainty.nist.gov as a Web-based application accessible worldwide via any of the more commonly used Web browsers, provides an easy-to-use implementation of Gauss's formula[134] and the Monte Carlo method described in Sec. VII A 3.

3. Monte Carlo method

The Monte Carlo method for uncertainty evaluation in the context of measurement equations, which the GUM-1 calls *propagation of distributions*, offers several important advantages over the approach described in Sec. VII A 2: (i) it can produce as many correct significant digits in its results as may be required; (ii) it does not involve the computation of derivatives, either analytically or numerically; (iii) it is applicable even when f is markedly non-linear; and (iv) it provides the raw materials necessary to characterize the whole probability distribution of the output quantity.

Even when the input quantities have (scaled and shifted) Student's t distributions, there is no need to use any function of their numbers of degrees of freedom to produce an approximation to the probability distribution of the output quantity: the Monte Carlo sample alone suffices.[169]

Application of the Monte Carlo method does, however, require more information than the application of Equation (1):

in particular, it requires that the (joint) probability distribution of the input quantities be specified fully.

The general version of the Monte Carlo method dates back to the middle of the twentieth century.[137] Morgan and Henrion[129] described and employed it for the evaluation of measurement uncertainty well before it found its way into the GUM-S1. Here it serves to replace the change-of-variable formula to characterize the probability distribution of the output quantity when it is impracticable to do so analytically and involves the following steps.

MC1: Model the uncertainty surrounding the input quantities by means of a joint probability distribution. (This step is simplified considerably when the random variables modeling the input quantities may be treated as stochastically independent, because in such a case all one has to do is to specify the probability distribution of each one of them).

MC2: Choose a positive integer K sufficiently large to ensure that the results of the uncertainty evaluation will have the required number of significant digits and draw a sample of size K from the joint distribution of the input quantities (if they happen to be independent, then this amounts to drawing a sample of size K from the distribution of each of the input quantities separately), to obtain $(x_{1,1}, \ldots, x_{n,1}), \ldots, (x_{1,K}, \ldots, x_{n,K})$.

MC3: Compute $y_1 = f(x_{1,1}, \ldots, x_{n,1}), \ldots, y_K = f(x_{1,K}, \ldots, x_{n,K})$, which are a sample from the probability distribution of the output quantity.

All that the Monte Carlo method produces is an arbitrarily large sample $\{y_k\}$ drawn from the probability distribution of the output quantity. Deciding how this sample should be reduced to obtain an estimate of the true value of the output quantity, and a particular uncertainty evaluation to qualify it, requires the application of supplementary criteria of optimality and fitness for purpose.

In particular, and for scalar measurands, there is no guarantee that the average of the Monte Carlo sample will be "best" in any useful sense, or that coverage intervals whose endpoints are the mth smallest and the mth largest values in this sample (for some value of m smaller than one half of the sample size) will include the mean of the distribution of the output quantity, or any other particular indication of its typical value (median, etc.) with the specified coverage probability. This limitation will be discussed further below, in Sec. VII A 4.

Summaries often found useful include estimates of the probability density of the output quantity, estimates of its standard deviation (for scalar measurands), and coverage regions.

(a) Probability density: The most inclusive summarization is an estimate of the corresponding probability density function: either a simple histogram, or a more sophisticated kernel density estimate,[138] as in Figure 9, and their generalizations when the output quantity is multivariate, as in Figure 13.

(b) Standard measurement uncertainty: When the output quantity is a scalar, the standard measurement uncertainty associated with it may be evaluated as the sample standard deviation of $\{y_1, \ldots, y_K\}$, or as an estimate of the standard deviation that is resistant to outliers, for example as implemented in R functions `mad` or `hubers` (the latter is defined in package MASS[139]).

(c) Coverage region: If the output quantity is a scalar, and $y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(K)}$ denote the sorted values of the Monte Carlo sample, and $0 < \alpha < \frac{1}{2}$, then the interval $(y_{(K\alpha/2)}, y_{(K(1-\alpha/2))})$ is one of the many different intervals that, with probability $1 - \alpha$, is believed to include the true value of the output quantity. (Since $K\alpha/2$ and $K(1 - \alpha/2)$ need not be integers, the end-points of this coverage interval may be calculated by interpolation of adjacent $y_{(k)}$ s.) When the output quantity is multivariate, then the (hyper-)ellipsoid with the smallest volume that includes $100(1 - \alpha)\,\%$ of the $\{y_k\}$ is one possible coverage region that, with probability $1 - \alpha$, is believed to include the true value of the output quantity.

The adaptive procedure explained in the GUM-S1, the technique that Wübbeler *et al.*[140] describe, or sub-sampling methods,[141] all may be used to ascertain whether K is large enough to support the number of significant digits required for the characterization of the evaluation of measurement uncertainty, or needs to be increased. The *NIST Uncertainty Machine* employs a sub-sampling procedure to determine the number of significant digits to report for the uncertainty evaluations that it produces by application of the Monte Carlo method.

4. Monte Carlo method—Cautionary remarks

The validity of coverage intervals derived from Monte Carlo samples has been questioned because, in some situations, probabilistically symmetric coverage intervals as defined in (c) above cover the true value of the measurand with frequency much smaller than their nominal coverage probability, in a long sequence of simulated repetitions of the measurement process and corresponding uncertainty evaluation.[142,143]

In some cases this is a consequence of a choice of coverage interval that is not fit for the intended purpose, as Possolo, Toman, and Estler[144] explain. However, since other issues or limitations may arise,[3] next we describe an illustrative, cautionary tale that should be kept in mind when choosing a particular expression of uncertainty to summarize the scatter of the sample drawn from the probability distribution of the output quantity.

Suppose that the measurand is the mean value $\mu = \alpha\kappa/(\kappa - 1)$ of a Pareto probability distribution with scale $\alpha > 0$ and shape $\kappa > 1$, and that μ is to be estimated based on 5 replicated observations from this distribution, by the method of maximum likelihood. That is, $\widehat{\mu} = \widehat{\alpha}\widehat{\kappa}/(\widehat{\kappa} - 1)$, with $\widehat{\alpha} = x_{(1)}$ and $\widehat{\kappa} = 5/\sum_{i=1}^{5} \log(x_i/x_{(1)})$, where $x_{(1)} = \min\{x_1, \ldots, x_5\}$.

In this case, the Monte Carlo method is the parametric statistical bootstrap, and involves simulating K samples of size 5 drawn for a Pareto distribution with the scale $\widehat{\alpha}$ and shape $\widehat{\kappa}$ specified above, and for each sample estimating the scale and shape by the method maximum likelihood, and then computing the corresponding estimate of μ. This process produces $\mu_1, \ldots, \mu_K$ as a sample of size K drawn from the probability distribution of the estimate of the measurand.

If $\alpha = 1$ and $\kappa = 2.25$, then the true value of the measurand is $\mu = 9/5$. With $K = 10\,000$, the probabilistically symmetric

95% coverage interval for μ ranges from the 250th smallest to the 250th largest of the $\{\mu_k\}$. However, this interval includes the true value of μ with only 75% probability. One might conjecture that this interval is just a bad choice of coverage interval for μ, and opt instead for the 95% coverage interval ranging from the 500th smallest to the largest of the $\{\mu_k\}$: the effective coverage of this interval is 95%. However, this felicitous choice requires the gift of prophecy about where the true value of the output quantity is located, and few and far between are the metrologists endowed with such a gift.

The main factors causing the "failure" of the probabilistically symmetric 95% coverage interval in this case are as follows: (i) the marked asymmetry of the distribution of the MLE of μ, whose right tail is much heavier than the left tail; (ii) the large coefficient of variation (that is, relative standard uncertainty), about 19%, associated with $\widehat{\mu}$; and (iii) the bias affecting $\widehat{\mu}$ (persistent error defined as the difference between the expected value of $\widehat{\mu}$ and μ, whose relative size is $-12\,\%$, hence indicating that $\widehat{\mu}$ tends to be too small). Efron and Tibshirani[54] and Davison and Hinkley[145] describe other techniques to construct bootstrap intervals that may be able to overcome shortcomings including these just discussed.

The question may then reasonably be asked of what it is that a coverage interval derived from a Monte Carlo sample drawn from the probability distribution of the output quantity achieves in general, and not just in "well-behaved" cases such as those that we shall present in Secs. VII A 5 and VII A 6. The answer to this question rests on a universal attribute of samples drawn from probability distributions for scalar quantities, which we describe next.

Consider two independent random variables V_1 and V_2 with the same probability distribution P. Each realization of the pair is a sample drawn from P. Now, consider a random variable W independent of those two, and with the same distribution P. Owing to independence and to the commonality of the distribution, all six orderings of the values of V_1, V_2, and W are equally likely. Therefore, W is equally likely to be to the left of $\min\{V_1, V_2\}$, between V_1 and V_2, or to the right of $\max\{V_1, V_2\}$: the probability of each of these three events is 1/3. In other words, each of the three intervals (of which two have infinite length) defined by V_1 and V_2 is a *prediction interval* for W with probability 1/3.

The same reasoning, applied to the Monte Carlo sample $\mu_1, \ldots, \mu_K$, leads to the conclusion that, once these K values are ordered from smallest to largest, every little interval between two consecutive ordered values has probability $1/(K+1)$ of including another value that may subsequently be drawn from the same distribution. The union of any $0.95(K+1)$ of these little intervals is a prediction region with 95% probability for a new drawing from the distribution of the output quantity.[146,147] And if these many little intervals are chosen so that no gaps are left between them, then their union will be a prediction interval with 95% coverage probability. This "coverage," however, is not of the mean, median, or any other particular indication of location of the measurand, but of the unit of probability of the distribution itself; or, in other words, of the true value of the measurand in the absence of any specific information about it other than what is encapsulated in the probability distribution sampled by the Monte Carlo method.

5. Example: Molecular weight of CO_2

At its 2009 meeting in Vienna, the CIAAW resolved to express the standard atomic weight of ten elements — including carbon and oxygen — by means of intervals that characterize the span of their atomic-weight values in "normal" materials. The intervals are [12.0096, 12.0116] for carbon and [15.999 03, 15.999 77] for oxygen.[45] (Note that atomic and molecular weights, being ratios of masses, are dimensionless quantities.)

If $A_r^*(\mathrm{C})$ and $A_r^*(\mathrm{O})$ denote independent random variables with uniform (or, rectangular) distributions over these intervals, then their mean values are 12.0106 and 15.9994 (which are the midpoints of the intervals), and their standard deviations are $u(A_r(\mathrm{C})) = 0.0006$ and $u(A_r(\mathrm{O})) = 0.0002$ (the standard deviation of a uniform distribution equals the length of the interval where the distribution is concentrated, divided by $\sqrt{12}$).

Therefore, $M_r(\mathrm{CO_2}) = 12.0106 + 2(15.9994) = 44.0094$ is an estimate of $M_r(\mathrm{CO_2})$. Neglecting the diminutive correlation between $A_r^*(\mathrm{C})$ and $A_r^*(\mathrm{O})$ that is induced by the implied normalization relative to the atomic mass of $^{12}\mathrm{C}$, $M_r(\mathrm{CO_2})$ is a function of two uncorrelated random variables. Since this function is linear, a result from probability theory implies that the variance of $M_r(\mathrm{CO_2})$ is equal to the variance of $A_r(\mathrm{C})$ plus 4 times the variance of $A_r(\mathrm{O})$: $u^2(M_r(\mathrm{CO_2})) = u^2(A_r(\mathrm{C})) + 4u^2(A_r(\mathrm{O})) = (0.0006)^2 + 4(0.0002)^2 = (0.000\,721)^2$. Therefore, $u(M_r(\mathrm{CO_2})) = 0.0007$.

In this case, it is also possible to derive analytically not only the standard uncertainty $u(M_r(\mathrm{CO_2}))$, but the whole probability distribution that characterizes the uncertainty associated with the molecular weight of CO_2. In fact, $M_r^*(\mathrm{CO_2}) = A_r^*(\mathrm{C}) + 2A_r^*(\mathrm{O})$ is a random variable with a symmetrical trapezoidal distribution[148] with the mean and standard deviation given above, and whose probability density is depicted in Figure 8. Using this fact, exact coverage intervals can be computed: for example, [44.0080, 44.0108] is the shortest (among infinitely many alternatives) 95% coverage interval for the molecular weight of carbon dioxide.

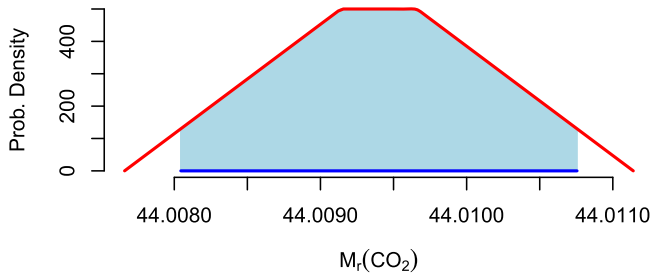

FIG. 8. Trapezoidal probability density that characterizes the uncertainty associated with the molecular weight of carbon dioxide, assuming that the atomic weights of carbon and oxygen are independent random variables distributed uniformly over the corresponding standard atomic weight intervals. The shaded region comprises 95% of the area under the trapezoid, and its footprint on the horizontal axis is the shortest, exact 95% coverage interval.

6. Example: Pitot tube

The pioneering contribution made by Kline and McClintock[135] already comprises most of the technical elements that, many years later, would come to be codified in the GUM. These authors consider a case where a measurement result for pressure is specified as 50.2 ± 0.5 psi(absolute) (20 to 1), indicating "that the best value for the pressure is believed to be 50.2 psia and the odds are 20 to 1 that the true value lies within ±0.5 psia(absolute) of this best estimate."

Kline and McClintock[135] illustrate the approximate method to evaluate the uncertainty associated with the airspeed v using Gauss's formula, where $\Delta = 1.993$ kPa was measured with a U-tube manometer, $p = 101.4$ kPa was measured with a Bourdon gage, and $T = 292.8$ K was measured with a mercury-in-glass thermometer. The corresponding estimate of airspeed was $v = 57.48$ m/s.

The expanded uncertainties for 95% coverage (V) were $U_{95\%}(\Delta) = 0.025$ kPa, $U_{95\%}(p) = 2.1$ kPa, and $U_{95\%}(T) = 0.22$ K. We take the corresponding standard uncertainties as one half of these expanded uncertainties. (Since the original treatment disregards the uncertainty component affecting R_s that is attributable to lack of knowledge about the actual humidity of air, below we assume that $u(R_s) = 0$.)

Since the measurement equation $v = \sqrt{2\Delta R_s T/p}$ is a ratio of products of powers of the input quantities, Gauss's formula reduces to Equation (2), yielding $(u(v)/v)^2 \approx ((\frac{1}{2})u(\Delta)/\Delta)^2 + ((\frac{1}{2})u(R_s)/R_s)^2 + ((\frac{1}{2})u(T)/T)^2 + ((-\frac{1}{2})u(p)/p)^2 = ((\frac{1}{2})0.0125/1.993)^2 + ((\frac{1}{2})0/287.058)^2 + ((\frac{1}{2})0.11/292.8)^2 + ((-\frac{1}{2})1.05/101.4)^2 = (0.006\,06)^2$, hence $u(v) = 0.006\,06(57.48$ m/s$) = 0.35$ m/s.

The *NIST Uncertainty Machine* produces $v = 57.48$ m/s and $u(v) = 0.35$ m/s according to both Gauss's formula and the Monte Carlo method (for which the input variables were modeled as Gaussian random variables), hence relative standard uncertainty $u(v)/v = 0.6\%$. An approximate 95% coverage interval defined as $v \pm 2u(v)$ ranges from 56.78 m/s to 58.18 m/s. Its counterpart based on the results of the Monte Carlo method, with endpoints given by the 2.5th and 97.5th percentiles of a sample of size 1×10^6 drawn from the probability distribution of v, ranges from 56.81 m/s to 58.17 m/s.

B. Uncertainty evaluation for observation equations

The evaluation of uncertainty for measurands defined by observation equations is an exercise in statistical inference because observation equations are statistical models where the measurand appears either as a parameter of a probability distribution or as a known function of parameters of a probability distribution. These parameters need to be estimated from experimental data, possibly incorporating other relevant information, and the uncertainty evaluation typically is a by-product of the statistical exercise of fitting the model to the data.

When any of several alternative statistical models are similarly adequate for the experimental data, and the results (estimate of the measurand and uncertainty evaluation) are sensitive to model choice, then model uncertainty should be taken into account and propagated to the final results. The example concerning the lifetime of an aircraft component, discussed in Sec. VII B 2, illustrates how model uncertainty may be evaluated quantitatively and propagated.

1. Example: Thermal bath

The maximum likelihood estimates of the parameters of the AR(2) model selected for the data listed in Table I, $t_i = \tau + \varphi_1(t_{i-1} - \tau) + \varphi_2(t_{i-2} - \tau) + \varepsilon_i$, are $\widehat{\varphi}_1 = 0.41$, $\widehat{\varphi}_2 = 0.42$, and $\sigma = 0.002$ °C. These were computed using the R function Arima defined in package forecast.[149,150] The function also provides evaluations of the uncertainties associated with the estimates of the auto-regressive coefficients, $u(\varphi_1) = u(\varphi_2) = 0.09$, using approximate methods of mathematical statistics.

Most importantly, the function estimates the mean temperature of the bath as 50.105 °C with standard uncertainty 0.001 °C, which is three times larger than the naive (and incorrect) uncertainty evaluation obtained by neglecting the autocorrelations, which is $s/\sqrt{100} = 0.0003$ °C, where s denotes the standard deviation of the 100 readings of temperature listed in Table I.

The Gaussian stochastic process that we fitted to the data happens to be *stationary*, which is the reason we are entitled to conclude that the bath is in thermal equilibrium. Loosely speaking, stationarity means that the level, magnitude, and pattern of the oscillations of temperature observed in a time window that is, for example, 20 min long, do not depend on when we start observing. (Note that not all AR(2) processes are stationary.) Furthermore, the model suggests that the heat transfer mechanisms that drive the oscillations in temperature operate at a temporal scale of approximately 3 min, and that the memory of such recent past is clouded by transient perturbations ("innovations") $\{\varepsilon_i\}$ whose typical magnitude is (plus or minus) 0.002 °C.

2. Example: F-100 component lifetime

Since the estimate of the expected lifetime is a function of the maximum likelihood estimates of the parameters of the Weibull distribution, $\widehat{\tau} = \widehat{\beta}\Gamma(1 + 1/\widehat{\alpha}) = 2.07$ h, one may use Gauss's formula (Equation (1)) to compute an approximation to $u(\tau)$ as a function of $u(\alpha)$, $u(\beta)$, and the correlation $r(\alpha, \beta)$ between them.

According to a result of mathematical statistics that applies to maximum likelihood estimates with great generality,[36] approximations to these standard uncertainties and correlation may be extracted from the determinant of the matrix of second-order partial derivatives (Hessian) of the logarithm of the likelihood function evaluated at $(\widehat{\alpha}, \widehat{\beta})$. The quality of these approximations depends on the size of the sample (13 in this case) and on the actual sample values from which the maximum likelihood estimates are derived. In this case, the approximate standard uncertainties and correlation associated with the estimates of the shape and scale parameters are $u(\alpha) = 0.385$, $u(\beta) = 0.507$ h, and $r(\alpha, \beta) = 0.0101$. Then, either Equation (1) or the *NIST Uncertainty Machine* yields $u(\tau) = 0.50$ h.

The Monte Carlo method provides an alternative, particularly convenient way to characterize the uncertainty that surrounds the estimate of τ. The method is an instance of the

so-called parametric statistical bootstrap,[54] and involves first repeating the following steps for $k = 1, \ldots, K$, where we chose $K = 50\,000$.

1. Draw a sample of size 13 from a Weibull probability distribution with shape $\widehat{\alpha}$ and scale $\widehat{\beta}$ as estimated by maximum likelihood from the experimental data.
2. Take the smallest 10 sample values as uncensored data, and the 3 largest as right-censored at the value of the largest of the smallest 10.
3. Use this censored sample to compute maximum likelihood estimates α_k and β_k and the corresponding value $\tau_k = \beta_k \Gamma(1 + 1/\alpha_k)$.

The resulting sample of expected lifetimes $\tau_1, \ldots, \tau_K$ may be summarized in any one of the several different ways. Their standard deviation is an evaluation of $u(\tau) = 0.51$ h, which is quite close to the approximation derived from the large-sample properties of the maximum likelihood estimates, followed by application of Gauss's formula. But this Monte Carlo sample provides so much more than just an evaluation of $u(\tau)$: Figure 9 shows an estimate of the corresponding probability density, fully characterizing the uncertainty associated with τ, and a 95% coverage interval ranging from 1.6 h to 3.6 h (these are the 2.5th and 97.5th percentiles of the Monte Carlo sample).

It just so turns out that the average of $\tau_1, \ldots, \tau_K$ is 2.56 h, while the MLE for the lifetime is 2.07 h. Even though the two estimates are not (statistically) significantly different, one may wonder why there should be a difference, and in particular a difference that suggests that the MLE may be too small. This difference may be attributable to the conjunction of two facts: (i) in samples of size 13 drawn from a Weibull distribution with shape and scale equal to the mle's given in Sec. VI B 2, the 10th smallest observation will be greater than 3 h with probability 0.34; (ii) when the 10th smallest observation (hence the censoring value) is greater than 3 h, the corresponding MLE of τ is almost certain to be larger than the $\widehat{\tau}$ we found. The *double bootstrap*[145] (Algorithm 3.3) — an advanced topic that we will not pursue — may be used for bias adjustment.

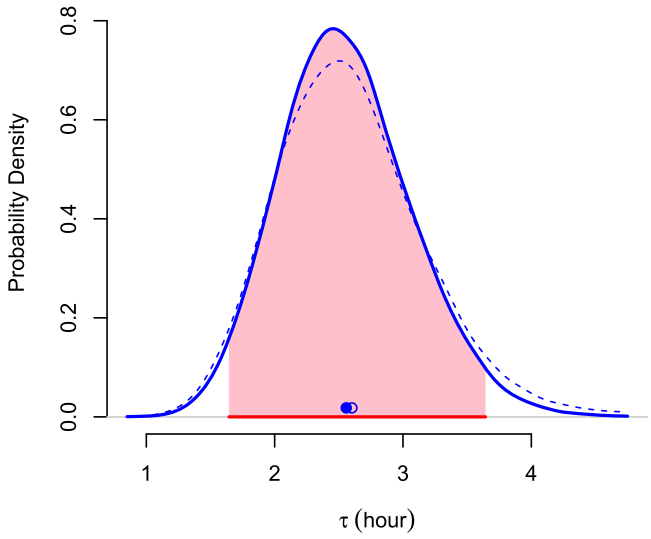

FIG. 9. Kernel estimate[138] of the probability density of the parametric bootstrap distribution of the expected lifetime τ. The shaded (pink) region comprises 95% of the area under the curve: its projection onto the horizontal axis is a 95% coverage interval for the expected lifetime. The (blue) dot indicates the average of the bootstrap sample. The dashed line depicts an estimate of the probability density of the parametric bootstrap distribution of the expected lifetime taking model uncertainty into account, and the (blue) open circle indicates the corresponding average.

Up until now we have taken the Weibull model for granted when in fact other models would be reasonable candidates for the experimental data: for example, the gamma and lognormal distributions. The corresponding maximum likelihood estimates of τ are 2.11 h and 2.68 h (both computed taking censoring into account) versus 2.07 h for the Weibull.

The relative adequacy of the gamma, lognormal, and Weibull models may be gauged using the Bayesian Information Criterion (BIC), which is a guide for model selection.[46] The values of BIC are −1.77 for the gamma, 1.15 for the lognormal, and −1.65 for the Weibull, which suggest that the gamma model indeed may be best for these data (the smaller the value of BIC the better the model).

The values of BIC may be transformed into approximate (Bayesian posterior) probabilities for the alternative models given the data. Assuming that the three models are equally likely *a priori*, Equation (7.41) from Hastie, Tibshirani, and Friedman[124] produces the following probabilities: 46% for gamma, 11% for lognormal, and 43% for Weibull. This discrete probability distribution describes model uncertainty.

The Monte Carlo method may be used to propagate both model uncertainty and the uncertainty expressed in the scatter of the experimental data. The process is similar to what we described above for the estimate of τ based on the Weibull model. For each iteration k, we begin by selecting a model (among the gamma, lognormal, and Weibull) at random according to the foregoing probabilities and then take the same steps described above for the Weibull, but using the selected distribution instead.

The resulting bootstrap sample $\{\tau_k\}$ will then comprise maximum likelihood estimates from the three models in proportions roughly equal to those probabilities. Figure 9 shows an estimate of the corresponding probability density and compares it with the distribution obtained when sampling only from the calibrated Weibull distribution. The standard uncertainty associated with τ becomes 0.61 h, and the 95% coverage interval, whose endpoints are the 2.5th and 97.5th percentiles of the Monte Carlo sample, ranges from 1.6 h to 3.9 h. Thus, incorporation of model uncertainty does induce larger uncertainty than when a single, particular model is assumed.

3. Example: Thermistor calibration

Whetstone *et al.*[96] (p. 62) report that the standard uncertainty associated with the values of temperature measured by the PRT is $u_{\text{prt}}(t) = 0.0015\,^{\circ}\text{C}$, and point out that extension cables used to connect the thermistor probe to the location where the indications were read also are a source of uncertainty with standard uncertainty $u_{\text{cable}}(t) = 0.01\,^{\circ}\text{C}$. These, and the contributions from the estimated residuals $\{\widehat{\delta}_i\}$ from the fitted model $t_i = \varphi(\tau_i) + \delta_i$ for $i = 1, \ldots, m$, will be propagated using the Monte Carlo method by taking the following steps.

1. Compute $\gamma = \left(\widehat{\sigma}^2 + u^2_{\text{cable}}(t)\right)^{\frac{1}{2}} = 0.016\,^{\circ}\text{C}$, where $\widehat{\sigma} = 0.012\,^{\circ}\text{C}$ is the estimated standard deviation of the residuals $\{\widehat{\delta}_i\}$.

2. Let $\theta_1, \ldots, \theta_m$ denote a set of $m = 100$ indication values for the thermistor equispaced from 16.22 °C to 40.47 °C. (The inverse of the calibration function will be evaluated for purposes of graphical display as in Figure 10.)
3. Choose a positive integer K sufficiently large to ensure that the results of the uncertainty evaluation will have the required number of significant digits (in this example $K = 10\,000$), and then for $k = 1, \ldots, K$, we have the following:
   (a) Draw $\tau_{1,k}, \ldots, \tau_{n,k}$ independently from $n = 12$ Gaussian distributions with means $\tau_1, \ldots, \tau_n$ (the values of temperature reported by the PRT) and standard deviations all equal to $u_{\text{prt}}(\tau)$.
   (b) Draw $t_{1,k}, \ldots, t_{n,k}$ independently from $n = 12$ Gaussian distributions with means $\widehat{t}_1, \ldots, \widehat{t}_n$ (the thermistor indications predicted by the calibration function φ at the values of temperature reported by the PRT) and standard deviations all equal to γ.
   (c) Estimate the coefficients of a third order polynomial φ_k by least squares that expresses the $\{t_{j,k} : j = 1, \ldots, n\}$ as a function of the $\{\tau_{j,k} : j = 1, \ldots, n\}$.
   (d) For each $i = 1, \ldots, m$, compute $\tau_{i,k} = \widehat{\psi}_k(\theta_i)$, where ψ_k denotes the inverse of φ_k. This step involves solving a cubic equation for each i, and determining the suitable root to assign to $\tau_{i,k}$.
4. Determine coverage intervals, depicted in Figure 10, for all values of $i = 1, \ldots, m$ simultaneously, applying the method described by Davison and Hinkley[145] (Sec. 4.2.4) and implemented in R function envelope,[151] using the data in the $m \times K$ array with the $\{\tau_{i,k}\}$.

4. Example: Ballistic limit of body armor

The maximum likelihood procedure used to estimate the intercept α and slope β of the probit regression model considered in Sec. VI B 4 also provides evaluations of the standard uncertainties and covariance for $\widehat{\alpha}$ and $\widehat{\beta}$ based on the large-sample theory of maximum likelihood estimation:[36] $u(\alpha) = 16$ and $u(\beta) = 0.042$ s/m, and $\text{cov}(\widehat{\alpha}, \widehat{\beta}) = -0.6697$. Gauss's formula, Equation (1), taking the correlation (−0.9997) between the estimates of α and β into account, then yields $u(v_{50}) = 5$ m/s.

A logit regression model fits the data just about as well as the probit regression model and essentially reproduces the probit's estimate of v_{50} and the evaluation of associated uncertainty. Therefore, unless one should like to entertain still other alternative models, there will be no reason to perform model averaging for the uncertainty evaluation, as was necessary in the case of the lifetime of a component of the F-100 Super Sabre (Sec. VII B 2).

Application of the parametric statistical bootstrap[54] involves first repeating the following steps for $k = 1, \ldots, K$, where we chose $K = 50\,000$.

1. Simulate values of binary random variables $B_{1,k}, \ldots, B_{m,k}$ to synthesize data $(v_1, B_{1,k}), \ldots, (v_m, B_{m,k})$, where $v_1 = 344$ m/s, …, $v_m = 378.5$ m/s are kept fixed at the actual bullet velocities achieved in the experiment, and $B_{i,k}$ has a Bernoulli probability distribution with probability of "success" (that is, vest perforation) $\widehat{\pi}(v_i) = \Phi(\widehat{\alpha} + \widehat{\beta} v_i)$, for $i = 1, \ldots, m$.
2. Fit a probit regression to these data, and use the resulting estimates of α and β to obtain the corresponding estimate $v_{50,k}$ of the ballistic limit.

The question may be asked whether, in step (a), the bullet velocities should be kept fixed at the observed values or varied either by resampling or by sampling from some appropriate probability distribution. Since these velocities are determined by the powder loads dispensed into the bullet casings, and the individual loads are weighed by the testing laboratory,[107] they are believed to be sufficiently reproducible to be treated as controlled and fixed for the present purpose.

Figure 11 shows a smooth histogram of the bootstrap sample $\{v_{50,1}, \ldots, v_{50,K}\}$. The corresponding standard deviation $u(v_{50}) = 5$ m/s is identical to the large-sample approximation indicated above. A corresponding 95% coverage interval, whose endpoints are the 2.5th and 97.5th percentiles of this sample, ranges from 377 m/s to 397 m/s.

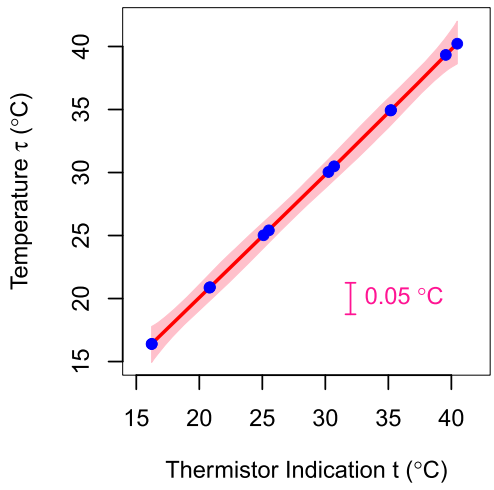

FIG. 10. Analysis function and simultaneous coverage (pink) band for τ, with 95% coverage probability. Since this band in fact is very narrow, in this display it appears magnified 50-fold in the vertical direction, with the inset scale bar representing a difference of 0.05 °C with the same magnification as the uncertainty band. The blue dots represent the measured values.

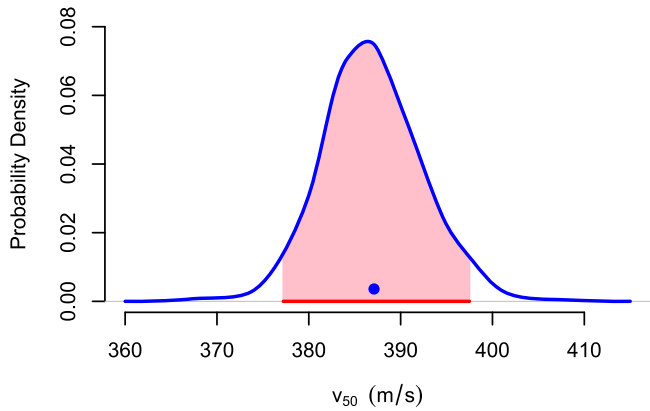

FIG. 11. Kernel estimate[138] of the probability density of the parametric bootstrap distribution of the ballistic limit v_{50}. The shaded (pink) region comprises 95% of the area under the curve: its projection onto the horizontal axis is a 95% coverage interval for the ballistic limit. The (blue) dot indicates $\widehat{v}_{50}$.

5. Example: PCB in sediment

In Sec. VI B 5 we noted that the values measured by the participating laboratories are almost 3 times more dispersed than the typical, within-laboratory standard uncertainty. The parameter $\widehat{\tau}_{\mathrm{DL}} = 1.71$ ng/g quantifies this "extra" uncertainty, which is a form of dark uncertainty because the participants did not recognize it in their individual uncertainty evaluations. The laboratory-specific standard uncertainties $\{u(w_j)\}$ typically result from a *bottom-up* uncertainty evaluation (Sec. VII).

The laboratory random effects model introduced in Sec. VI B 5 provides not only the means to recognize the presence of dark uncertainty, but it also produces uncertainty evaluations that express all relevant sources of uncertainty. R function rma, defined in package metafor,[110] computes the DerSimonian-Laird estimate of the measurand, $\widehat{\mu} = 33.6$ ng/g, an evaluation of the associated standard uncertainty $u(\mu) = 0.75$ ng/g, and an approximate 95% coverage interval for μ ranging from 32.1 ng/g to 35.1 ng/g.

Guolo and Varin[152] and Hoaglin,[153] among others, have pointed out that both the estimate of the consensus value and the associated uncertainty produced by the DerSimonian-Laird procedure may be unreliable especially when the number of studies being pooled is small (less than 10). Two important shortcomings of the uncertainty evaluation produced by the DerSimonian-Laird procedure are as follows: (i) it does not take the numbers of degrees of freedom associated with the $\{u(w_j)\}$ into account and (ii) it does not recognize that the between-laboratory standard deviation τ often is estimated based on inter-comparing a rather small number of measured values (6 in this case).

The *NIST Consensus Builder*,[154] soon to be available at `consensus.nist.gov`, is a Web-based application for the statistical analysis of measurement results from inter-laboratory studies. It employs the parametric statistical bootstrap[54] for uncertainty evaluation. This is consistent with the GUM-S1 and offers the ability to take into account the finite numbers of degrees of freedom $\{\nu_j\}$ associated with the $\{u(w_j)\}$, which addresses shortcoming (ii) described above. In particular, it produces $u_{\mathrm{DL}}(\mu) = 0.76$ ng/g, and a 95% coverage interval for μ ranging from 32.0 ng/g to 35.2 ng/g.

The raw materials for the uncertainty evaluation via the parametric statistical bootstrap are obtained by repeating the following steps a large number K of times, for $k = 1, \ldots, K$.

1. Draw τ_k from the approximate sampling probability distribution of $\widehat{\tau}_{\mathrm{DL}}$ that reflects the fairly small number of measured values involved in estimating τ, as suggested by Biggerstaff and Tweedie[155] and Biggerstaff and Jackson.[156]
2. Draw $x_{j,k}$ from a Gaussian distribution with mean $\widehat{\mu}$ and variance $\tau_k^2 + u^2(w_j)$, for $j = 1, \ldots, n$.
3. If ν_j is either infinity or unspecified, then $u_{j,k} = u(w_j)$, otherwise $u_{j,k} = u(w_j)\sqrt{\nu_j/\chi^2_{\nu_j}}$, where $\chi^2_{\nu_j}$ denotes a value drawn from a chi-squared distribution with ν_j degrees of freedom, for $j = 1, \ldots, n$.
4. Compute the DerSimonian-Laird consensus value μ_k corresponding to the triplets $(x_{1,k}, u_{1,k}, \nu_1), \ldots, (x_{n,k}, u_{n,k}, \nu_n)$.

The standard uncertainty associated with the DerSimonian-Laird consensus value is the standard deviation of the $\{\mu_k\}$, and a 95% coverage interval for μ ranges from the 2.5th to the 97.5th percentile of the $\{\mu_k\}$.

Alternatively, a Bayesian treatment can be adopted that also remedies the defects of the conventional DerSimonian-Laird uncertainty evaluation. The distinctive traits of a Bayesian treatment are these:[55] (i) all quantities whose values are unknown are modeled as non-observable random variables, and data are modeled as observed values of random variables; (ii) estimates and uncertainty evaluations for unknown quantity values are derived from the conditional probability distribution of the unknowns given the data (the so-called *posterior distribution*).

Enacting (i) involves specifying probability distributions for all the quantities in play (unknowns as well as data) and (ii) involves the application of Bayes's rule, typically via Markov Chain Monte Carlo (MCMC) sampling that produces an arbitrarily large sample from the posterior distribution.[55] Carrying this out successfully requires familiarity with probability models and with their selection for the intended purpose, and also with suitable, specialized software for statistical computing. We have relied on the implementation of MCMC in JAGS,[157] via R package R2jags.[158]

The distributions selected for the Bayesian analysis likely will be applicable to many consensus-building exercises, and for this reason have been implemented in the NIST Consensus Builder. The required assumptions are as follows.

- μ, τ, and the $\{\sigma_j\}$ are mutually independent *a priori*.
- μ has a prior Gaussian distribution with mean 0 and a very large standard deviation.
- τ and the $\{\sigma_j\}$ have prior half-Cauchy distributions as suggested by Gelman,[159] the former with median equal to the median of the absolute values of the differences between the measured values and their median, the latter with median equal to the median of the $\{u(w_j)\}$.
- Given τ, the $\{\lambda_j\}$ are Gaussian with mean 0 and standard deviation τ.
- Given μ, τ, and the $\{\lambda_j\}$, the measured values $\{w_j\}$ are modeled as outcomes of Gaussian random variables with means $\{\mu + \lambda_j\}$ and standard deviations $\{\sigma_j\}$.
- When the standard uncertainties associated with the measured values are based on finitely many numbers of degrees of freedom $\{\nu_j\}$, and given the $\{\sigma_j\}$, the $\{\nu_j u(w_j)^2/\sigma_j^2\}$ are modeled as outcomes of chi-squared random variables with $\{\nu_j\}$ degrees of freedom. When they are based on infinitely many numbers of degrees of freedom (that is, are regarded as known), $\sigma_j = u(w_j)$. Here, and similarly elsewhere, the numbers of degrees of freedom indicate the size of the evidentiary basis supporting the $\{u(w_j)\}$, which will be sample sizes (minus 1) in the case of Type A evaluations.

The estimate of μ is the mean, 33.6 ng/g, of the sample drawn from the corresponding posterior distribution by MCMC, and the associated standard uncertainty $u_{\mathrm{B}}(\mu) = 0.80$ ng/g is the standard deviation of the same sample. A 95% probability interval ranges from 32.0 ng/g to 35.2 ng/g, which happens to coincide with the corresponding interval obtained by

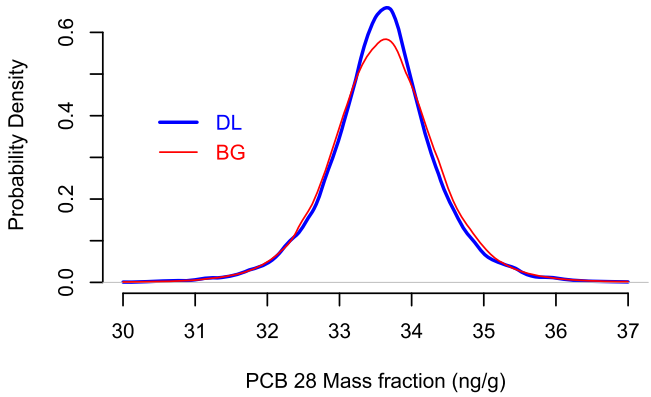

FIG. 12. Kernel estimates[138] of the probability density of the mass fraction of PCB 28 in a sediment, based on the samples produced by application of the parametric bootstrap for the DerSimonian-Laird procedure (DL), and by MCMC for the Bayesian procedure (BG).

application of the parametric bootstrap, to within the significant digits reported. However, Figure 12 shows that the probability densities for the measurand, corresponding to the parametric bootstrap for the DerSimonian-Laird procedure and to the Bayesian procedure, are not quite identical, the latter having slightly heavier tails than the former. These differences show that different ways of fitting the same (Gaussian random effects) model to the measurement results can impact the final results perceptibly.

6. Example: Forensic glass

In Sec. VI B 6 we concluded that, for a glass fragment with refractive index 1.51613 and mass fractions of the major oxides 13.92% (Na), 3.52% (Mg), 1.25% (Al), 72.88% (Si), 0.37% (K), 7.94% (Ca), 0% (Ba), 0.14% (Fe), the mixture discriminant model produces the following class membership probabilities: float processed building windows, 0.29; non-float processed building windows, 0.64; vehicle windows, 0.075; the other classes having negligible probabilities. On the other hand, application of leave-one-out cross-validation indicated that the classifier has approximately 30% error rate.

That probability distribution and this error rate provide complementary characterizations of measurement uncertainty associated with the measured value, which in this case was "non-float processed building windows." The former suggests that there is a substantial probability of the fragment being from a modern building window, and a non-negligible probability of its being from a vehicle window. The latter reminds us that the probability of an incorrect assignment is about 30%.

VIII. EXPRESSING MEASUREMENT UNCERTAINTY

Measurement uncertainty should be expressed in a manner that is fit for the purpose that the measurement result is intended to serve. In most cases, specifying a set of values of the measurand believed to include its true value with a specified probability (say, a 95% coverage region) suffices as expression of measurement uncertainty. The certificates accompanying NIST Standard Reference Materials (www.nist.gov/srm/) usually report 95% coverage intervals for scalar measurands, or analogous constructs for other measurands.

When the result of an evaluation of measurement uncertainty is intended for use in subsequent uncertainty propagation exercises involving Monte Carlo methods, the ideal expression of measurement uncertainty is a fully specified probability distribution characterizing the state of knowledge about the measurand, or a sample drawn from such probability distribution, sufficiently large to support the required number of significant digits for the results.

The techniques described in the GUM and by Taylor and Kuyatt[6] produce approximate coverage intervals for scalar measurands. The latter, which applies specifically to NIST measurement results, indicates that, by convention, the expanded uncertainty should be twice the standard uncertainty. This is motivated by the fact that, in many cases, a coverage interval of the form $y \pm 2u(y)$ achieves approximately 95% coverage probability even when the probability distribution of the measurand either has tails considerably heavier than the Gaussian distribution, or is markedly skewed (that is, has one tail longer or heavier than the other).[160] Taylor and Kuyatt[6] (Appendix B) also discuss when and how coverage intervals of the form $y \pm ku(y)$, with coverage factors k other than 2, may or should be used.

Coverage intervals or regions need not be symmetrical relative to the estimate of the measurand. Often the shortest or otherwise smallest such interval or region will not be symmetrical, especially when the measurand is constrained to be non-negative or to lie in a bounded region. The NIST Uncertainty Machine implements the Monte Carlo method of the GUM-S1 and will produce coverage intervals that are symmetrical relative to y upon request. Asymmetric intervals are commonly used in nuclear physics. For example, Hosmer *et al.*[161] report the result of measuring the half-life of ^{80}Cu as 170^{+110}_{-50} ms.

When it is desired to propagate the uncertainty expressed in an asymmetric coverage interval while preserving the asymmetry, a Monte Carlo method can be used, for example, by drawing samples from a probability distribution whose median (or mean, depending on the situation and application) is equal to y and that otherwise assigns the stated coverage probability to the coverage interval given.

Audi *et al.*[162] and Barlow[163] describe several symmetrization techniques that may be used to propagate the uncertainty expressed in asymmetric 95% coverage intervals glossing over the asymmetry. A simple rule that is often used for the same purpose defines an approximate, effective standard uncertainty as one-fourth of the width of the asymmetric interval.

For functional measurands (for example, the analysis function for a thermistor discussed in Sec. VII B 3), an informative expression of measurement uncertainty takes the form of a coverage band that, with specified probability, includes the graph of the whole function of interest over the relevant range of its argument. Figure 10 depicts one such band.

For multivariate measurands, the region of smallest volume that, with specified probability, is believed to include the true value of the measurand, provides a useful and concise expression of measurement uncertainty. For example, in the context of the example about the F-100 component lifetime discussed in Secs. VI B 2 and VII B 2, a bivariate measurand of possible interest is the vector (α, β), whose components are

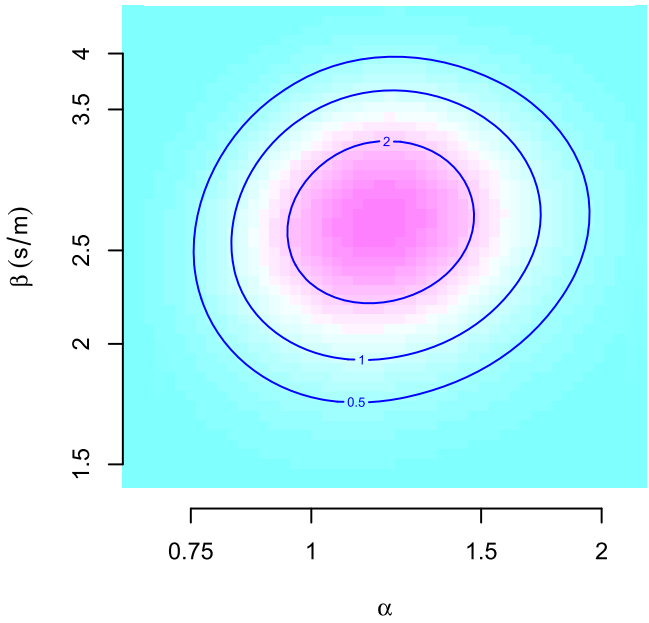

FIG. 13. Bivariate ash estimate (product polynomial kernel)[165] of the joint probability density of the shape and scale parameters of the Weibull distribution for the lifetime of an F-100 component, based on a parametric bootstrap sample of size 50 000.

the shape and scale of the Weibull distribution that describes the variability of the data.

Figure 13 shows an estimate of the bivariate probability density of (α, β) based on a parametric bootstrap sample of size 50 000, computed using R function ash2 defined in package ash.[164] The region inside the curve labeled "1" is an approximate 67% coverage region for (α, β).

IX. RECAPITULATION

The sole measurement model considered in the GUM, which is often called the *measurement equation*, $y = f(x_1, \ldots, x_n)$, cannot address situations where the observations made in a measurement experiment are mutually inconsistent (mass fractions of PCB 28 in sediment, determined by several laboratories), or where the measurand is a function of parameters of the probability distribution describing the variability of the experimental data (censored observations of the lifetime of a component of the F-100 Super Sabre fighter jet). Situations like these, and many others, demand that the measurement model be a statistical model (*observation equation*).

Since several, alternative statistical models often are comparably reasonable candidates for the data gathered in a measurement experiment, and these alternatives may produce materially different results, the corresponding *model uncertainty* must be evaluated and propagated just like the contribution from any other source of uncertainty. In Sec. VII B 2 we have shown how the Monte Carlo method can be used to carry out this task.

In contrast to how measurement uncertainty is defined in the VIM (2.26), we have advanced the simple and obvious notion that measurement uncertainty is the doubt about the true value of the measurand that remains after making a measurement. This doubt may be quantified and represented in any one of the several different ways, depending on the intended use. The most thorough representation of measurement uncertainty is afforded by a probability distribution on the set of possible values of the measurand, and it is applicable to measurands of all kinds, qualitative or quantitative, the latter possibly being scalar, vectorial, or functional. In many applications, a suitable summary of the dispersion of values of this distribution suffices. The standard deviation is one such summary for scalar measurands, and it is one example of the "non-negative parameter characterizing the dispersion of the quantity values being attributed to a measurand" that the VIM (2.26) mentions when defining measurement uncertainty.

We discussed eight examples of measurement in detail and used them to illustrate different measurement models and different techniques that are available for uncertainty evaluation: (1) molecular weight of CO_2; (2) airspeed using a Pitot tube; (3) average temperature of a thermal bath; (4) lifetime of a component of the F-100 aircraft; (5) calibration of a thermistor; (6) ballistic limit of body armor; (8) consensus value of the mass fraction of a PCB in sediment; and (9) provenance of a glass fragment. We have also used (in Sec. IV D 4) an instance of measurement of the mass fraction of copper in wholemeal flour to illustrate an application of the non-parametric statistical bootstrap.

Model-based approaches provide a unified solution to both problems of estimating the measurand and evaluating the associated measurement uncertainty. Tal[19] suggests that "a central motivation for the development of model-based accounts is the attempt to clarify the epistemological principles underlying aspects of measurement practice," giving as example the fact that "metrologists employ a variety of methods for the calibration of measuring instruments, the standardization and tracing of units and the evaluation of uncertainties."

The measurement models in Examples (1) and (2) were conventional measurement equations, linear for the first, nonlinear for the second. In (1), the probability distribution of the measurand could be characterized analytically using tools from the theory of probability. Both Gauss's formula and the Monte Carlo method of the GUM-S1 were applied to (2), yielding very much the same results. The NIST Uncertainty Machine facilitates the application of both of these techniques to evaluate the uncertainty of scalar measurands.

The measurement models in the other six examples all required observation equations. A Gaussian auto-regression of order 2 in (3), a Weibull distribution in (4), a cubic polynomial regression in (5), a generalized linear model[166] (probit regression) in (6), and a random effects model in (7). Example (6) exhibited several unusual features, including nominal observations (perforation of a bullet-proof vest, or not) and involved a hybrid measurement model, comprising both observation and measurement equations. In Example (9), the measurand itself was a nominal property.

The techniques presented for uncertainty evaluation included: (a) Gauss's formula, which provides an approximation for the standard uncertainty of a scalar measurand and is a special case of the Delta Method of probability theory; (b) the change-of-variable formula; (c) the parametric statistical bootstrap, which is the Monte Carlo method described

in the GUM-S1 for the "propagation of distributions"; (d) maximum likelihood for observation equations; and (e) Bayesian procedures, including MCMC sampling.

These techniques were employed as follows: (a) in Examples (1), (2), (4), and (6); (b) underlies the analytical derivation of the probability distribution of the measurand in (2); (c) in (2), (4), (5), (6), and (7); (d) in (3), (4), (5), (6), (7), and (9); and (e) in (7).

Our starting point was a broader concept of measurement than is entertained in the GUM and in the VIM, intended to accommodate the vast growth in depth and breadth of measurement science since the GUM was first published. In particular, we recognize purely computational processes as capable of producing *bona fide* measurement results. We regard the values of all properties (qualitative as well as quantitative) as legitimate measurands, even though some authors have argued in favor of a different name for the assignment of value to qualitative properties (for example, *examination*[167]). One of our motivations to adopt this wide scope is the expansion of the domain of measurement science as practiced at NIST, which now includes measurement services catering to proteomics and genomics,[168] and forensic science.

ACKNOWLEDGMENTS

The authors are much indebted to their NIST colleague Amanda Koepke for sharing her R implementation of the procedure suggested by Biggerstaff and Tweedie[155] and Biggerstaff and Jackson[156] for the uncertainty evaluation described in Sec. VII B 5. We have also heeded David Duewer's advice, also from NIST, to the effect that it is preferable to state the original source of the uncertainty evaluation explicitly in each case, rather than merely classifying the evaluation as of Type A or Type B. Chuck Ehrlich, Amanda Koepke, Jolene Splett, and Jack Wang, all from NIST, provided a very large number of very detailed and most helpful corrections and suggestions for improvement that we are immensely grateful for. Two anonymous referees kindly complimented and welcomed our contribution. One of them was extraordinarily generous in providing copious comments and many excellent suggestions for improvement. Finally, we wish to express our appreciation for the encouragement that we received from James Matey (Consulting Editor for Invited Reviews and Invited Articles of the *Review of Scientific Instruments*, and NIST colleague) from inception until the final stage of production of this article.

[1]BEA, Final report on the accident on 1st June 2009 to the Airbus A330-203 registered F-GZCP operated by Air France flight AF 447 Rio de Janeiro – Paris, Technical Report No. F-GZCP (Bureau d'Enquêtes et d'Analyses pour la sécurité de l'aviation civile, Le Bourget, France, 2012).

[2]Any mention of commercial products is for information only and does not imply recommendation or endorsement by NIST.

[3]D. R. White, "In pursuit of a fit-for-purpose uncertainty guide," Metrologia **53**, S107–S124 (2016).

[4]ISO/IEC, *International Vocabulary of Metrology — Basic and General Concepts and Associated Terms (VIM)*, 1st ed. (International Organization for Standardization (ISO)/International Electrotechnical Commission (IEC), Geneva, Switzerland, 2007), ISO/IEC Guide 99.

[5]NIST, *NIST Quality Manual for Measurement Services — NIST QM-I* (National Institute of Standards and Technology, U.S. Department of Commerce, Gaithersburg, Maryland, 2015), Version 9.

[6]B. N. Taylor and C. E. Kuyatt, *Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results*, NIST Technical Note 1297 (National Institute of Standards and Technology, Gaithersburg, MD, 1994).

[7]A. Possolo, *Simple Guide for Evaluating and Expressing the Uncertainty of NIST Measurement Results*, NIST Technical Note 1900 (National Institute of Standards and Technology, Gaithersburg, MD, 2015).

[8]H. O. Wood and F. Neumann, "Modified Mercalli intensity scale of 1931," Bull. Seismol. Soc. Am. **21**, 277–283 (1931).

[9]C. Ehrlich, "Terminological aspects of the guide to the expression of uncertainty in measurement (GUM)," Metrologia **51**, S145–S154 (2014).

[10]BIPM, *The International System of Units (SI)*, 8th ed. (International Bureau of Weights and Measures (BIPM), Sèvres, France, 2006).

[11]D. A. Swyt, "Length and dimensional measurements at NIST," J. Res. Natl. Bur. Stand. **106**, 1–23 (2001).

[12]T. J. Quinn, "Practical realization of the definition of the metre, including recommended radiations of other optical frequency standards (2001)," Metrologia **40**, 103–133 (2003).

[13]J. Ye, H. Schnatz, and L. W. Hollberg, "Optical frequency combs: From frequency metrology to optical phase control," IEEE J. Sel. Top. Quantum Electron. **9**, 1041–1058 (2003).

[14]I. Fischer, "Another look at Eratosthenes' and Posidonius' determinations of the earth's circumference," Q. J. R. Astron. Soc. **16**, 152–167 (1975).

[15]H. Cavendish, "Experiments to determine the density of the earth. by Henry Cavendish, Esq. F. R. S. and A. S.," Philos. Trans. R. Soc. London **88**, 469–526 (1798).

[16]D. Sobel, *Longitude: The True Story of a Lone Genius Who Solved the Greatest Scientific Problem of His Time* (Bloomsbury, New York, 1995).

[17]A. R. Disney, *A History of Portugal and the Portuguese Empire*, Vol. 2: The Portuguese Empire (Cambridge University Press, Cambridge, UK, 2009).

[18]H. von Helmholtz, Counting and Measuring (D. Van Nostrand, New Jersey, 1887), C. L. Bryan (translator, 1930).

[19]E. Tal, "Measurement in science," in *The Stanford Encyclopedia of Philosophy*, edited by E. N. Zalta, summer ed. (The Metaphysics Research Lab, Center for the Study of Language and Information (CSLI), Stanford University, 2015).

[20]J. V. Nicholas and D. R. White, *Traceable Temperatures*, 2nd ed. (John Wiley & Sons, Chichester, England, 2001).

[21]R. Dybkaer, "Definitions of measurement," Accredit. Qual. Assur. **16**, 479–482 (2011).

[22]R. White, "The meaning of measurement in metrology," Accredit. Qual. Assur. **16**, 31–41 (2011).

[23]L. Mari and P. Carbone, "Measurement fundamentals: A pragmatic view," IEEE Trans. Instrum. Meas. **61**, 2107–2115 (2012).

[24]B. Russell, *The Principles of Mathematics* (W. W. Norton, New York, NY, 1903).

[25]K. Jelved, A. D. Jackson, O. Knudsen, and A. D. Wilson, "Experiments on the effect of the electric conflict on the magnetic needle (July 21, 1820)," in *Selected Scientific Works of Hans Christian Orsted* (Princeton University Press, 1998), pp. 413–416.

[26]S. Thrun, M. Montemerlo, H. Dahlkamp, D. Stavens, A. Aron, J. Diebel, P. Fong, J. Gale, M. Halpenny, G. Hoffmann, K. Lau, C. Oakley, M. Palatucci, V. Pratt, P. Stang, S. Strohband, C. Dupont, L.-E. Jendrossek, C. Koelen, C. Markey, C. Rummel, J. Niekerk, E. Jensen, P. Alessandrini, G. Bradski, B. Davies, S. Ettinger, A. Kaehler, A. Nefian, and P. Mahoney, "Stanley: The robot that won the DARPA grand challenge," in *The 2005 DARPA Grand Challenge: The Great Robot Race*, edited by M. Buehler, K. Iagnemma, and S. Singh (Springer Berlin Heidelberg, Berlin, Heidelberg, 2007), pp. 1–43.

[27]A. Frigerio, A. Giordani, and L. Mari, "Outline of a general model of measurement," Synthese **175**, 123–149 (2010).

[28]M. Agrawal, N. Kayal, and N. Saxena, "PRIMES is in P," Ann. Math. **160**, 781–793 (2004).

[29]K. McGrattan, S. Hostikka, R. McDermott, C. Weinschenk, K. Overholt, and J. Floyd, *Fire Dynamics Simulator—Technical Reference Guide*, NIST Special Publication 1018-1, Vol. 1: Mathematical Model, 6th ed. (National Institute of Standards and Technology, Gaithersburg, MD, 2015).

[30]K. McGrattan, S. Hostikka, R. McDermott, C. Weinschenk, K. Overholt, and J. Floyd, *Fire Dynamics Simulator—Technical Reference Guide*, NIST Special Publication 1018-3, Vol. 3: Validation, 6th ed. (National Institute of Standards and Technology, Gaithersburg, MD, 2015).

[31]R. D. Peacock, K. B. McGrattan, G. P. Forney, and P. A. Reneke, *CFAST—Consolidated Fire and Smoke Transport (Version 7)—Volume 1: Technical*

Reference Guide, NIST Technical Note 1889v1 (National Institute of Standards and Technology, Gaithersburg, MD, 2015).
[32]R. D. Peacock, K. B. McGrattan, G. P. Forney, and P. A. Reneke, *CFAST—Consolidated Fire and Smoke Transport (Version 7)—Volume 3: Verification and Validation Guide*, NIST Technical Note 1889v3 (National Institute of Standards and Technology, Gaithersburg, MD, 2015).
[33]B. P. Abbott *et al.*, "Observation of gravitational waves from a binary black hole merger," Phys. Rev. Lett. **116**, 061102 (2016).
[34]D. Freedman, R. Pisani, and R. Purves, *Statistics*, 4th ed. (W. W. Norton & Company, New York, NY, 2007).
[35]M. H. DeGroot and M. J. Schervish, *Probability and Statistics*, 4th ed. (Addison-Wesley, 2011).
[36]L. Wasserman, *All of Statistics: A Concise Course in Statistical Inference* (Springer Science+Business Media, New York, NY, 2004).
[37]NIST/SEMATECH, *NIST/SEMATECH e-Handbook of Statistical Methods* (National Institute of Standards and Technology, U.S. Department of Commerce, Gaithersburg, Maryland, 2006).
[38]A. Possolo and B. Toman, *Tutorial for Metrologists on the Probabilistic and Statistical Apparatus Underlying the GUM and Related Documents* (National Institute of Standards and Technology, Gaithersburg, MD, 2011).
[39]D. V. Lindley, "The probability approach to the treatment of uncertainty in artificial intelligence and expert systems," Stat. Sci. **2**, 17–24 (1987).
[40]G. Mauris, V. Lasserre, and L. Foulloy, "A fuzzy approach for the expression of uncertainty in measurement," Measurement **29**, 165–177 (2001).
[41]L. A. Zadeh, "Is there a need for fuzzy logic?," Inf. Sci. **178**, 2751–2779 (2008).
[42]A. N. Kolmogorov, *Foundations of the Theory of Probability*, 2nd ed., edited by N. Morrison (Chelsea Publishing Co., New York, NY, 1933).
[43]J. B. Kadane, M. J. Schervish, and T. Seidenfeld, "Statistical implications of finitely additive probability," in *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno De Finetti*, Studies in Bayesian Econometrics and Statistics, Vol. 6, edited by P. K. Goel and A. Zellner (Elsevier Science, 1986), Chap. 2.5.
[44]B. de Finetti, *Theory of Probability: A Critical Introductory Treatment*(John Wiley & Sons, Chichester, 1990), two volumes, translated from the Italian and with a preface by Antonio Machì and Adrian Smith, with a foreword by D. V. Lindley, Reprint of the 1975 translation.
[45]M. E. Wieser, N. Holden, T. B. Coplen, J. K. Böhlke, M. Berglund, W. A. Brand, P. D. Bièvre, M. Gröning, R. D. Loss, J. Meija, T. Hirata, T. Prohaska, R. Schoenberg, G. O'Connor, T. Walczyk, S. Yoneda, and X.-K. Zhu, "Atomic weights of the elements 2011 (IUPAC Technical Report)," Pure Appl. Chem. **85**, 1047–1078 (2013).
[46]K. Burnham and D. Anderson, *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd ed. (Springer-Verlag, New York, NY, 2002).
[47]F. Mosteller and J. W. Tukey, *Data Analysis and Regression* (Addison-Wesley Publishing Company, Reading, Massachusetts, 1977).
[48]C. Chatfield, "Model uncertainty, data mining and statistical inference," J. R. Stat. Soc., Ser. A **158**, 419–466 (1995).
[49]M. Clyde and E. I. George, "Model uncertainty," Stat. Sci. **19**, 81–94 (2004).
[50]G. J. Hahn and W. Q. Meeker, *Statistical Intervals: A Guide for Practitioners* (John Wiley & Sons, 1991).
[51]M. J. Schervish, *Theory of Statistics*, Springer Series in Statistics (Springer Verlag, New York, NY, 1995).
[52]Joint Committee for Guides in Metrology, *Evaluation of measurement data—Supplement 1 to the "Guide to the expression of uncertainty in measurement" — Propagation of distributions using a Monte Carlo method* (International Bureau of Weights and Measures (BIPM), Sèvres, France, 2008), BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP and OIML, JCGM 101:2008.
[53]G. E. P. Box and G. C. Tiao, *Bayesian Inference in Statistical Analysis* (Addison-Wesley, Reading, Massachusetts, 1973).
[54]B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap* (Chapman & Hall, London, UK, 1993).
[55]A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis*, 3rd ed. (Chapman & Hall/CRC, Boca Raton, FL, 2013).
[56]T. C. Hesterberg, "What teachers should know about the bootstrap: Resampling in the undergraduate statistics curriculum," Am. Stat. **69**, 371–386 (2015).
[57]Analytical Methods Committee, "Robust statistics—How not to reject outliers. Part I. Basic concepts," Analyst **114**, 1693–1697 (1989).
[58]P. Rousseeuw, C. Croux, V. Todorov, A. Ruckstuhl, M. Salibian-Barrera, T. Verbeke, M. Koller, and M. Maechler, robustbase: Basic Robust Statistics, R package version 0.9–7, 2012.
[59]S. Bell, *A Beginner's Guide to Uncertainty of Measurement, Measurement Good Practice Guide No. 11 (Issue 2)* (National Physical Laboratory, Teddington, Middlesex, United Kingdom, 1999), amendments March 2001.
[60]C. A. Gonzales and R. L. Watters, *Standard Reference Material 972a, Vitamin D Metabolites in Frozen Human Serum* (Office of Reference Materials, National Institute of Standards and Technology, Department of Commerce, Gaithersburg, Maryland, 2013).
[61]E. Tal, "The epistemology of measurement: A model-based account," Ph.D. thesis, Department of Philosophy, University of Toronto, Toronto, Canada, 2012.
[62]Joint Committee for Guides in Metrology, *Evaluation of Measurement Data—Guide to the Expression of Uncertainty in Measurement* (International Bureau of Weights and Measures (BIPM), Sèvres, France, 2008), BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP and OIML, JCGM 100:2008, GUM 1995 with minor corrections.
[63]IPCC, *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, edited by T. F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, and P. M. Midgley (Cambridge University Press, Cambridge, UK, 2013).
[64]L. J. Savage, "Difficulties in the theory of personal probability," Philos. Sci. **34**, 305–310 (1967).
[65]A. Hájek, "Interpretations of probability," in *The Stanford Encyclopedia of Philosophy*, edited by E. N. Zalta, Winter 2012 ed. (The Metaphysics Research Lab, Center for the Study of Language and Information, Stanford University, Stanford, California, 2012).
[66]J. M. Keynes, *A Treatise on Probability* (MacMillan and Co., London, 1929).
[67]R. Carnap, *Logical Foundations of Probability*, 2nd ed. (University of Chicago Press, Chicago, Illinois, 1962).
[68]H. Jeffreys, *Theory of Probability* (Oxford University Press, London, 1939).
[69]R. T. Cox, *The Algebra of Probable Inference* (The Johns Hopkins Press, Baltimore, Maryland, 1961).
[70]D. H. Mellor, *Probability: A Philosophical Introduction* (Routledge, New York, 2005).
[71]L. J. Savage, *The Foundations of Statistics* (Dover Publications, New York, New York, 1972).
[72]A. O'Hagan, "Eliciting and using expert knowledge in metrology," Metrologia **51**, S237–S244 (2014).
[73]R. Willink and R. White, "Disentangling classical and Bayesian approaches to uncertainty analysis," Technical Report No. CCT/12–07 (Bureau International des Poids et Mesures (BIPM), Sèvres, France, 2012), consultative Committee for Thermometry (CCT)—Working Documents, 26th Meeting.
[74]P. J. Mohr, B. N. Taylor, and D. B. Newell, "CODATA recommended values of the fundamental physical constants: 2010," Rev. Mod. Phys. **84**, 1527–1605 (2012).
[75]A. Hájek, "'Mises redux'—Redux: Fifteen arguments against finite frequentism," Erkenntnis **45**, 209–227 (1996).
[76]A. Hájek, "Fifteen arguments against hypothetical frequentism," Erkenntnis **70**, 211–235 (2009).
[77]S. Li and A. W. Robertson, "Evaluation of submonthly precipitation forecast skill from global ensemble prediction systems," Mon. Weather Rev. **143**, 2871–2889 (2015).
[78]M. Stone, "The opinion pool," Ann. Math. Stat. **32**, 1339–1342 (1961).
[79]M. H. DeGroot, "Reaching a consensus," J. Am. Stat. Assoc. **69**, 118–121 (1974).
[80]F. Dietrich and C. List, "Probabilistic opinion pooling," in *The Oxford Handbook of Probability and Philosophy*, edited by A. Hajek and C. Hitchcock (Oxford University Press, Oxford, UK, 2016), Chap. 25.
[81]S. L. R. Ellison and A. Williams, *Quantifying Uncertainty in Analytical Measurement*, EURACHEM/CITAC Guide CG-4, QUAM:2012.P1, 3rd ed. (Eurachem, 2012).
[82]D. R. White and P. Saunders, "The propagation of uncertainty with calibration equations," Meas. Sci. Technol. **18**, 2157–2169 (2007).
[83]T. Bartel, S. Stoudt, and A. Possolo, "Force calibration using errors-in-variables regression and monte carlo uncertainty evaluation," Metrologia **53**, 965–980 (2016).
[84]R. A. Nelson and M. G. Olsson, "The pendulum—Rich physics from a simple system," Am. J. Phys. **54**, 112–121 (1986).

[85]R. J. Muirhead, *Aspects of Multivariate Statistical Theory* (John Wiley & Sons, Hoboken, NJ, 2005).
[86]C. M. Wang and H. K. Iyer, "On non-linear estimation of a measurand," Metrologia **49**, 20–26 (2012).
[87]H. S. Peiser, N. E. Holden, P. D. Bièvre, I. L. Barnes, R. Hagemann, J. R. de Laeter, T. J. Murphy, E. Roth, M. Shima, and H. G. Thode, "Element by element review of their atomic weights," Pure Appl. Chem. **56**, 695–768 (1984).
[88]T. B. Coplen, J. K. Böhlke, P. D. Bièvre, T. Ding, N. E. Holden, J. A. Hopple, H. R. Krouse, A. Lamberty, H. S. Peiser, K. Révész, S. E. Rieder, K. J. R. Rosman, E. Roth, P. D. P. Taylor, J. R. D. Vocke, and Y. K. Xiao, "Isotope-abundance variations of selected elements," Pure Appl. Chem. **74**, 1987–2017 (2002).
[89]R. A. Askey and R. Roy, "Gamma function," in *NIST Handbook of Mathematical Functions*, edited by F. W. J. Olver, D. W. Lozier, R. F. Boisvert, and C. W. Clark (Cambridge University Press, Cambridge, UK, 2010).
[90]R Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Austria, Vienna, 2015).
[91]G. E. P. Box, G. M. Jenkins, and G. C. Reinsel, *Time Series Analysis: Forecasting and Control*, 4th ed. (John Wiley & Sons, Hoboken, NJ, 2008).
[92]N. R. Mann and K. W. Fertig, "Tables for obtaining weibull confidence bounds and tolerance bounds based on best linear invariant estimates of parameters of the extreme-value distribution," Technometrics **15**, 87–101 (1973).
[93]J. F. Lawless, *Statistical Models and Methods for Lifetime Data* (John Wiley & Sons, New York, NY, 1982).
[94]W. Meeker and L. A. Escobar, *Statistical Methods for Reliability Data* (John Wiley & Sons, New York, 1998).
[95]N. L. Johnson, S. Kotz, and N. Balakrishnan, *Continuous Univariate Distributions, Volume 1*, 2nd ed. (John Wiley & Sons, New York, NY, 1994).
[96]J. R. Whetstone, W. G. Cleveland, G. P. Baumgarten, S. Woo, and M. C. Croarkin, "Measurements of coefficients of discharge for concentric flange-tapped square-edged orifice meters in water over the reynolds number range 600 to 2 700 000," NIST Technical Note 1264 (National Institute of Standards and Technology, U.S. Department of Commerce, Gaithersburg, MD, 1989).
[97]ISO, *Gas Analysis—Comparison Methods for Determining and Checking the Composition of Calibration Gas Mixtures* (International Organization for Standardization (ISO), Geneva, Switzerland, 2001), International Standard ISO 6143:2001(E).
[98]W. A. Fuller, *Measurement Error Models* (John Wiley & Sons, New York, NY, 1987).
[99]R. J. Carroll, D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu, *Measurement Error in Nonlinear Models—A Modern Perspective*, 2nd ed. (Chapman and Hall/CRC, Boca Raton, Florida, 2006).
[100]F. R. Guenther and A. Possolo, "Calibration and uncertainty assessment for certified reference gas mixtures," Anal. Bioanal. Chem. **399**, 489–500 (2011).
[101]M. B. Wilk and R. Gnanadesikan, "Probability plotting methods for the analysis of data," Biometrika **55**, 1–17 (1968).
[102]M. J. Mazerolle, AICcmodavg: Model selection and multimodel inference based on (Q)AIC(c), R package version 2.0–3, 2015.
[103]J. M. Chambers, "Linear models," in *Statistical Models in S*, edited by J. M. Chambers and T. J. Hastie (Chapman and Hall/CRC, Boca Raton, FL, 1991), Chap. 4.
[104]Y. N. Subbotin, "Spline interpolation," in *Encyclopedia of Mathematics* (Springer & European Mathematical Society, 2002), last modified on 7 February 2011.
[105]C. de Boor, *A Practical Guide to Splines, Applied Mathematical Sciences No. 27* (Springer-Verlag, New York, NY, 2001).
[106]OLES, *Ballistic Resistance of Body Armor, NIJ Standard-0101.06* (National Institute of Justice, Office of Law Enforcement Standards, National Institute of Standards and Technology, Washington, DC, 2008).
[107]D. Mauchant, K. D. Rice, M. A. Riley, D. Leber, D. Samarov, and A. L. Forster, "Analysis of three different regression models to estimate the ballistic performance of new and environmentally conditioned body armor," Technical Report No. NISTIR 7760 (National Institute of Standards and Technology, Gaithersburg, MD, 2011).
[108]Comité International des Poids et Mesures (CIPM), *Mutual Recognition of National Measurement Standards and of Calibration and Measurement Certificates Issued by National Metrology Institutes* (Bureau International des Poids et Mesures (BIPM), Pavillon de Breteuil, Sèvres, France, 1999), Technical Supplement Revised in October 2003.
[109]M. Schantz and S. Wise, "CCQM–K25: Determination of PCB congeners in sediment," Metrologia **41**, 08001 (2004).
[110]W. Viechtbauer, "Conducting meta-analyses in R with the metafor package," J. Stat. Software **36**, 1–48 (2010).
[111]M. Thompson and S. L. R. Ellison, "Dark uncertainty," Accredit. Qual. Assur. **16**, 483–487 (2011).
[112]B. Toman and A. Possolo, "Laboratory effects models for interlaboratory comparisons," Accredit. Qual. Assur. **14**, 553–563 (2009).
[113]B. Toman and A. Possolo, "Erratum to: Laboratory effects models for interlaboratory comparisons," Accredit. Qual. Assur. **15**, 653–654 (2010).
[114]M. Borenstein, L. V. Hedges, J. P. Higgins, and H. R. Rothstein, "A basic introduction to fixed-effect and random-effects models for meta-analysis," Res. Synth. Methods **1**, 97–111 (2010).
[115]S. R. Searle, G. Casella, and C. E. McCulloch, *Variance Components* (John Wiley & Sons, Hoboken, NJ, 2006).
[116]J. C. Pinheiro, C. Liu, and Y. N. Wu, "Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate *t* distribution," J. Comput. Graphical Stat. **10**, 249–276 (2001).
[117]A. L. Rukhin and A. Possolo, "Laplace random effects models for interlaboratory studies," Comput. Stat. Data Anal. **55**, 1815–1827 (2011).
[118]R. DerSimonian and N. Laird, "Meta-analysis in clinical trials," Controlled Clin. Trials **7**, 177–188 (1986).
[119]I. W. Evett and E. J. Spiehler, "Rule induction in forensic science," in *KBS in Government* (Online Publications, Pinner, UK, 1987), pp. 107–118.
[120]K. Bache and M. Lichman, UCI Machine Learning Repository, 2013.
[121]T. Hastie, R. Tibshirani, F. Leisch, K. Hornik, and B. D. Ripley, mda: Mixture and flexible discriminant analysis, R package version 0.4–4, 2013.
[122]L. A. B. Pilkington, "Review lecture: The float glass process," Proc. R. Soc. London, Ser. A **314**, 1–25 (1969).
[123]T. Hastie and R. Tibshirani, "Discriminant analysis by Gaussian mixtures," J. R. Stat. Soc., Ser. B **58**, 155–176 (1996).
[124]T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. (Springer-Verlag, New York, 2009).
[125]H. Cooper, L. V. Hedges, and J. C. Valentine, *The Handbook of Research Synthesis and Meta-Analysis*, 2nd ed. (Russell Sage Foundation Publications, New York, NY, 2009).
[126]T. Bartel, "Uncertainty in NIST force measurements," J. Res. Natl. Inst. Stand. Technol. **110**, 589–603 (2005).
[127]A. Possolo and C. Elster, "Evaluating the uncertainty of input quantities in measurement models," Metrologia **51**, 339–353 (2014).
[128]G. Casella and R. L. Berger, *Statistical Inference*, 2nd ed. (Duxbury, Pacific Grove, California, 2002).
[129]M. G. Morgan and M. Henrion, *Uncertainty—A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*, First Paperback ed. (Cambridge University Press, New York, NY, 1992), 10th printing, 2007.
[130]Joint Committee for Guides in Metrology, *Evaluation of measurement data — Supplement 2 to the "Guide to the expression of uncertainty in measurement"—Extension to any number of output quantities* (International Bureau of Weights and Measures (BIPM), Sèvres, France, 2011), BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP and OIML, JCGM 102: 2011.
[131]N. G. Paulter, Jr. and D. R. Larson, "Reference ballistic chronograph," Opt. Eng. **48**, 043602 (2009).
[132]J. L. Rodgers and W. A. Nicewander, "Thirteen ways to look at the correlation coefficient," Am. Stat. **42**, 59–66 (1988).
[133]C. Gauss, "Theoria combinationis observationum erroribus minimis obnoxiae," in *Werke, Band IV, Wahrscheinlichkeitsrechnung und Geometrie* (Königlichen Gesellschaft der Wissenschaften, Göttingen, 1823), gdz.sub.uni-goettingen.de/.
[134]T. Lafarge and A. Possolo, "The NIST Uncertainty Machine," NCSLI Meas. J. Meas. Sci. **10**, 20–27 (2015).
[135]S. J. Kline and F. A. McClintock, "Describing uncertainties in single-sample experiments," Mech. Eng. **75**, 3–8 (1953).
[136]J. Mandel, *The Statistical Analysis of Experimental Data* (Interscience Publishers (John Wiley & Sons), New York, NY, 1964).
[137]C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*, 2nd ed. (Springer, New York, NY, 2004).
[138]B. W. Silverman, *Density Estimation* (Chapman and Hall, London, 1986).
[139]W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S*, 4th ed. (Springer, New York, 2002), ISBN: 0-387-95457-0.
[140]G. Wübbeler, P. M. Harris, M. G. Cox, and C. Elster, "A two-stage procedure for determining the number of trials in the application of a monte carlo method for uncertainty evaluation," Metrologia **47**, 317–324 (2010).

[141]D. N. Politis, J. P. Romano, and M. Wolf, *Subsampling* (Springer-Verlag, New York, 1999).
[142]B. D. Hall, "Evaluating methods of calculating measurement uncertainty," Metrologia **45**, L5–L8 (2008).
[143]R. Willink, "Probability, belief and success rate: comments on on the meaning of coverage probabilities," Metrologia **47**, 343–346 (2010).
[144]A. Possolo, B. Toman, and T. Estler, "Contribution to a conversation about the supplement 1 to the GUM," Metrologia **46**, L1–L7 (2009).
[145]A. C. Davison and D. Hinkley, *Bootstrap Methods and their Applications* (Cambridge University Press, New York, NY, 1997).
[146]S. S. Wilks, "Determination of sample sizes for setting tolerance limits," Ann. Math. Stat. **12**, 91–96 (1941).
[147]S. Preston, "Teaching prediction intervals," J. Stat. Educ. **8**(3) (2000).
[148]F. Killmann and E. von Collani, "A note on the convolution of the uniform and related distributions and their use in quality control," Econ. Qual. Control **16**, 17–41 (2001).
[149]R. Hyndman and Y. Khandakar, "Automatic time series forecasting: The forecast package for R," J. Stat. Software **27**, 1–22 (2008).
[150]R. J. Hyndman, forecast: Forecasting functions for time series and linear models, R package version 6.2, 2015.
[151]A. Canty and B. Ripley, boot: Bootstrap R (S-Plus) Functions, R package version 1.3-15, 2015.
[152]A. Guolo and C. Varin, "Random-effects meta-analysis: The number of studies matters," Stat. Methods Med. Res. (published online 2015).
[153]D. C. Hoaglin, "Misunderstandings about Q and 'Cochran'sQtest' in meta-analysis," Stat. Med. **35**, 485–495 (2016).
[154]A. Koepke, T. Lafarge, B. Toman, and A. Possolo, *NIST Consensus Builder—User's Manual* (National Institute of Standards and Technology, Gaithersburg, MD, 2016).
[155]B. J. Biggerstaff and R. L. Tweedie, "Incorporating variability in estimates of heterogeneity in the random effects model in meta-analysis," Stat. Med. **16**, 753–768 (1997).
[156]B. J. Biggerstaff and D. Jackson, "The exact distribution of Cochran's heterogeneity statistic in one-way random effects meta-analysis," Stat. Med. **27**, 6093–6110 (2008).
[157]M. Plummer, JAGS Version 4.0.0 user manual, 2015.
[158]Y.-S. Su and M. Yajima, R2jags: Using R to Run "JAGS," R package version 0.5–7, 2015.
[159]A. Gelman, "Prior distributions for variance parameters in hierarchical models," Bayesian Anal. **1**, 515–533 (2006).
[160]D. A. Freedman, *Statistical Models: Theory and Practice* (Cambridge University Press, New York, NY, 2009).
[161]P. Hosmer, H. Schatz, A. Aprahamian, O. Arndt, R. R. C. Clement, A. Estrade, K. Farouqi, K.-L. Kratz, S. N. Liddick, A. F. Lisetskiy, P. F. Mantica, P. Möller, W. F. Mueller, F. Montes, A. C. Morton, M. Ouellette, E. Pellegrini, J. Pereira, B. Pfeiffer, P. Reeder, P. Santi, M. Steiner, A. Stolz, B. E. Tomlin, W. B. Walters, and A. Wöhr, "Half-lives and branchings for β-delayed neutron emission for neutron-rich Co–Cu isotopes in the r-process," Phys. Rev. C **82**, 025806 (2010).
[162]G. Audi, F. Kondev, M. Wang, B. Pfeiffer, X. Sun, J. Blachot, and M. MacCormick, "The Nubase2012 evaluation of nuclear properties," Chin. Phys. C **36**, 1157–1286 (2012).
[163]R. Barlow, "Asymmetric errors," in *PHYSTAT2003: Statistical Problems in Particle Physics, Astrophysics and Cosmology* (SLAC National Accelerator Laboratory, Menlo Park, CA, 2003), pp. 250–255.
[164]D. W. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization*, 2nd ed. (John Wiley & Sons, Hoboken, NJ, 2015).
[165]D. W. Scott, A. Gebhardt, and S. Kaluzny, ash: David Scott's ASH Routines (2015), R package version 1.0–15.
[166]P. McCullagh and J. A. Nelder, *Generalized Linear Models*, 2nd ed. (Chapman and Hall/CRC, London, UK, 1989).
[167]G. Nordin, R. Dybkaer, U. Forsum, X. Fuentes-Arderiu, G. Schadow, and F. Pontet, "An outline for a vocabulary of nominal properties and examinations—Basic and general concepts and associated terms," Clin. Chem. Lab. Med. **48**, 1553–1566 (2010).
[168]A. L. Plant and R. L. Watters, *Standard Reference Material 2374, DNA Sequence Library for External RNA Controls* (Office of Reference Materials, National Institute of Standards and Technology, Department of Commerce, Gaithersburg, Maryland, 2013).
[169]J. Wang, NIST, personal communication (August 2016).