Addressing Nonnegligible Model Uncertainty in Machine Learning-Assisted Measurements

Leopoldo Angrisani, *Fellow, IEEE*, Maria Cacciapuoti, *Graduate Student Member, IEEE*, Sabatina Criscuolo, *Member, IEEE*, Mauro D'Arco, *Senior Member, IEEE*, Egidio De Benedetto, *Senior Member, IEEE*, Luigi Duraccio, *Member, IEEE*, and Annarita Tedesco

***Abstract*—Machine learning (ML) techniques are increasingly used in measurement science to develop novel measurement models. Such *ML-based measurement models* are particularly useful when *conventional* ones are unavailable, enabling measurement operations that would otherwise be impractical. However, the resulting relationship between the measurand and the input quantities is not directly selected by the practitioner; rather, it is inherently determined by a data-driven training process. This gives rise to an additional uncertainty component (denoted in previous work as *measurement-model uncertainty*) which, when adherence to the Joint Committee for Guides in Metrology (JCGM) is required, shall be negligible with respect to the component associated with the measurement of the physical input quantities (denoted in previous work as *physical quantity uncertainty*). When this condition cannot be met, a direct application of the JCGM framework may understate the measurement uncertainty, thereby yielding coverage intervals that are not representative of the values that can be reasonably attributed to the measurand. Accordingly, to address the limitations of a direct JCGM application while remaining consistent with its framework, this work proposes a JCGM-compliant method that leverages joint probability distributions, sampled via nested Monte Carlo simulations, to obtain a fully fledged measurement result for ML-assisted measurements. Two illustrative case studies based on deterministic ML regressors demonstrate the applicability of the proposed method and show that the resulting coverage intervals explicitly account for the variability introduced by the ML model, thereby avoiding any understatement of the reported measurement uncertainty.**

***Index Terms*—Artificial intelligence, guide to the expression of uncertainty in measurement (GUM), instrumentation and measurement, Joint Committee for Guides in Metrology (JCGM), machine learning (ML), measurement uncertainty, metrology, Monte Carlo method (MCM), regression.**

Received 12 January 2026; revised 20 April 2026; accepted 22 April 2026. Date of publication 19 May 2026; date of current version 8 June 2026. This work was supported in part by the Italian Ministry of University and Research (MUR) through the Projects: "Made in Italy Circolare e Sostenibile," PNRR PE11 (CUP E63C22002130007); and in part by the 2025 IEEE I&M Graduate Fellowship Award, which was awarded to Maria Cacciapuoti. The Associate Editor coordinating the review process was Dr. Guolin He. *(Corresponding author: Egidio De Benedetto.)*

Leopoldo Angrisani, Maria Cacciapuoti, Mauro D'Arco, Egidio De Benedetto, and Luigi Duraccio are with the Department of Information Technology and Electrical Engineering (DIETI), University of Naples Federico II, 80125 Naples, Italy (e-mail: angrisan@unina.it; maria.cacciapuoti@unina.it; mauro.darco@unina.it; egidio.debenedetto@unina.it; luigi.duraccio@unina.it).

Sabatina Criscuolo is with the Institute of Intelligent Industrial Technologies and Systems for Advanced Manufacturing, National Research Council of Italy, 23900 Lecco, Italy (e-mail: sabatina.criscuolo@stiima.cnr.it).

Annarita Tedesco is with the Department of Public Health, University of Naples Federico II, 80131 Naples, Italy (e-mail: annarita.tedesco@unina.it).

Digital Object Identifier 10.1109/TIM.2026.3694743

I. Introduction

MODERN measurement science increasingly integrates machine learning (ML) as a fundamental element of measurement processes [1], [2]. In particular, the capability of ML to infer relationships between a set of input data (the *predictors*, in ML terminology) and one or more output quantities has enabled the development of *ML-based measurement models*. In these models, the predictors correspond to the physical input quantities involved in the measurement process, and the output quantities correspond to the measurands [3]. Such ML-based measurement models are particularly useful in situations where conventional measurement models fail, either because no closed-form relationship exists between the input quantities and the measurand, or because such a relationship exists but does not adequately satisfy the requirements of the measurement application.

However, the deployment of ML-based measurement models raises important considerations when compliance with established metrological standards is required, particularly those issued by the Joint Committee for Guides in Metrology (JCGM), which provides internationally adopted guidance for measurement science, including the evaluation and expression of measurement uncertainty. In particular, the international vocabulary of metrology (VIM), a key JCGM guidance document, states that the measurand should be defined such that the *definitional uncertainty* can be regarded as negligible with respect to the other components of measurement uncertainty [4]. The guide to the expression of uncertainty in measurement (GUM), another key set of JCGM guidance documents, adopts the same principle but uses a different terminology, referring to it as *intrinsic uncertainty* [5]. Conventionally, this requirement is addressed by selecting an adequate measurement model grounded in physical laws or, more broadly, in available knowledge, although the GUM itself explicitly acknowledges that this may not always be practicable (Clause D3.4) [5].

Therefore, when an ML-based measurement model is employed in place of conventional measurement models, a careful verification of the requirement described in the VIM becomes necessary. Indeed, ML-based measurement models are obtained through a *learning* (or *training*) process that is inherently influenced by multiple factors, including the quantity and quality of the available data and the computational capabilities. This variability can be modeled as an uncertainty component directly related to the *definitional* (or

intrinsic) one, referred to in previous work as *measurement-model uncertainty* [6], which must be treated as an additional component of the overall uncertainty budget, alongside that associated with the input quantities, referred to as *physical quantity uncertainty* component [6]. For this reason, in previous work [6], a novel methodology was proposed with the dual objective of: 1) evaluating the *measurement-model uncertainty* component and 2) implementing corrective actions to reduce it until it is negligible with respect to the *physical quantity uncertainty* one, thereby satisfying the assumption stated in the VIM and enabling the direct application of the GUM for uncertainty evaluation in measurements based on ML deterministic regressors. Nevertheless, a significant gap still needs to be addressed: it may occur that, despite implementing the proposed solutions, such as increasing the amount of data or improving its quality, the *measurement-model uncertainty* component remains significant and cannot be neglected within the overall uncertainty budget. Consequently, this component must be appropriately combined with the *physical quantity uncertainty* one, to obtain a measurement uncertainty that also accounts for the variability introduced by the adoption of an ML model. It is also advisable, as suggested by the GUM, specifically its *Supplement 1* (GUM-S1) [7], to represent the measurand in terms of a probability distribution, since it may, in general, differ significantly from the assumption of a normal or Student's t-distribution typically made [5].

Based on these considerations, this work proposes a method to combine measurement uncertainty components in ML-assisted measurements. In particular, the methodology leverages GUM-S1 and the Monte Carlo method (MCM) to provide a probability distribution of the measurand that reflects the state of knowledge about the measurand and correctly accounts for all relevant components of the measurement uncertainty, i.e., both the *physical quantity uncertainty* and the *measurement-model uncertainty*. Through two case studies with deterministic ML regressors, specifically, the estimation of: 1) the electrical power in a power plant and 2) the state-of-charge of batteries, this article demonstrates the feasibility of the method and highlights its advantages over conventional approaches. Such case studies are based on well-known and publicly available datasets, to facilitate reproducibility and isolate the methodological contribution from application-specific complexities.

The article is organized as follows: Section II provides a background on uncertainty evaluation in ML-based measurements, detailing the problem statement and the necessity for a JCGM-compliant approach. Section III describes the proposed method for combining the measurement uncertainty components and representing the measurand in terms of a probability distribution. Sections IV and V present the two case studies. Finally, conclusions are drawn, and future work is outlined.

II. Background and Problem Statement

According to the JCGM, measurements are grounded on the availability of a measurement model, which defines the relationship between the measurand Y and a set of input quantities $(X_1, X_2, \ldots, X_N)$ [5]. As anticipated, such a relationship is typically selected by the operator, based on available knowledge, to fulfill the assumption, described in the VIM that the uncertainty component associated with the definition of the measurand (i.e., the measurement model) is negligible with respect to that arising from the measurements of the input quantities [4]. In general, guidance on how to develop measurement models is provided in the GUM-6 document [8]. According to this framework, the measurement uncertainty is obtained by propagating, through the measurement model, the uncertainties related to the measurements of the input quantities. To do this, the GUM introduces the law of propagation of uncertainty (LPU) [5] and the *propagation of distributions*, this latter described in GUM-S1 [7] and applicable through MCM when the measurement model is strongly nonlinear or when the input quantities are represented by strongly asymmetric probability distributions, making the application of the LPU inadvisable.

Therefore, in conventional measurement contexts, the dominant component of the uncertainty budget is that associated with the measurement of the input physical quantities and propagated through the measurement model (the *physical quantity uncertainty* component described in [6]), and the uncertainty component related to the measurement model is generally neglected [5]. It can thus be summarized that, within the current JCGM framework, the measurand Y is treated as a random variable, being a well-defined function of the random variables $(X_1, X_2, \ldots, X_N)$.

However, in the case of ML-assisted measurements, the measurement model is inferred from data through a learning process. This process typically assumes the existence of an *unknown* functional relationship f^* linking the input quantities to the measurand. Accordingly, let: 1) $\mathcal{X} \subseteq \mathbb{R}^N$ denote the input space of admissible realizations of the input quantities $(X_1, X_2, \ldots, X_N)$; 2) $\mathcal{Y} \subseteq \mathbb{R}$ denote the output space of the measurand Y; and 3) $\mathcal{H}$ denote a prescribed hypothesis space of candidate functional relationships $h : \mathcal{X} \to \mathcal{Y}$. The learning objective is then to identify an optimal function $\hat{h} \in \mathcal{H}$ that best approximates f^* depending on the adopted learning criterion [9]. In compliance with GUM-6, the resulting relationship $\hat{h}$ is only a fit-for-purpose approximation of the *unknown* relationship f^* [8]. Hence, the measurand can be expressed as

$$Y = \hat{h}(X_1, X_2, \ldots, X_N) = f^*(X_1, X_2, \ldots, X_N) + E \quad (1)$$

where the term E represents an error contribution introduced by $\hat{h}$[1], and may consist of both a systematic component (E_{sist}, compensable) and a random component (E_{rand}). According to GUM-6, these two components can be related to well-understood and poorly understood effects, respectively [8].

On this basis, it becomes reasonable to consider the measurand Y as a random variable not only because the input

[1]Within this perspective on measurements, which can be framed as *model-based realism*, a *true* relationship is assumed to exist between the measurand and the input quantities, and the resulting ML-based measurement model is regarded as its estimate. Accordingly, the measurand definition is not affected by modeling and learning choices; rather, their impact is captured by the discrepancy between the *true* and the estimated relationships, i.e., by the error term E. Further details on measurement perspectives are provided in [10].

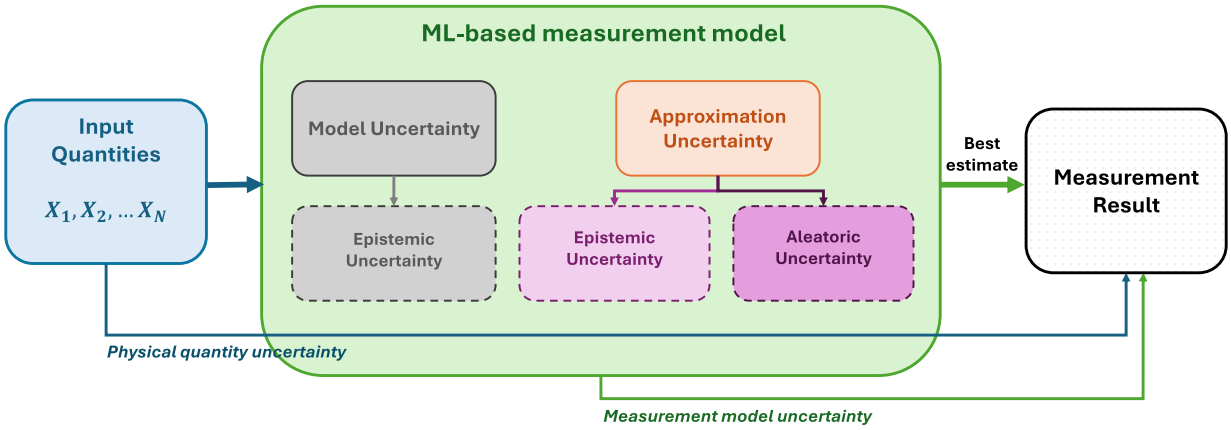

Fig. 1. Overview of the uncertainty components in ML-assisted measurements [6]. The *physical quantity uncertainty* component represents the uncertainty propagated from the measurement of the input quantities. An additional component is given by the *measurement-model uncertainty*, which accounts for both model-related and approximation-related effects. This component can be further decomposed into epistemic and aleatoric contributions. When the *measurement-model uncertainty* is not negligible with respect to the *physical quantity uncertainty*, the two components must be combined to express the overall measurement uncertainty and associate it with the best estimate of the measurand, thereby obtaining the measurement result.

quantities $(X_1, X_2, \ldots, X_N)$ are random variables, but also because the measurement model $\hat{h}$ itself introduces a random component E_{rand}. Consequently, in previous work [6], the concept of *measurement-model uncertainty* was introduced. This uncertainty component, in ML literature, is decomposed into *model uncertainty* (related to the selection of hypothesis space H, which is often neglected under the assumption of a correct specification) and *approximation uncertainty* [9], [11]. In turn, the approximation uncertainty consists of both *epistemic* and *aleatoric* contributions [9]. The first contribution arises from the lack of sufficient training data, an inadequately refined model (e.g., due to suboptimal hyperparameter selection), or choices made during the training process. On the other hand, aleatoric contribution is related to the inherent stochasticity of the observations and training process.

For clarity, a schematic representation of the uncertainty components in ML-assisted measurements is shown in Fig. 1.

To verify the assumption of negligible definitional uncertainty for the measurand, described in the VIM, the *measurement model uncertainty* component should be evaluated and compared with the *physical quantity uncertainty* one. For this reason, as anticipated in Section I, in previous work [6] a methodology was proposed to evaluate the *measurement-model uncertainty*, with a specific focus on ML deterministic regression models.[2] A key element of the methodology proposed in [6] was the *model calibration* procedure. Through *model calibration*, the predictions provided by the ML-based measurement model were compared with a set of reference values of the measurand. This enabled the determination of a fully fledged calibration diagram for the ML-based measurement model, allowing the compensation for systematic effects (E_{sist}) and the attribution of an uncertainty value (the *measurement-model uncertainty* component) to each prediction of the ML-based measurement model, reflecting the aleatoric contribution of the measurement model itself. By analogy, model calibration allows for assigning an uncertainty value to each *reading* obtained from the ML-based measurement model, in the same way as instrument calibration conventionally does for the readings of a measuring instrument [6]. Such model calibration procedure is also closely related to the verification of measurement-model adequacy discussed in GUM-6 [8], namely assessing that the measurement model is corrected for all known effects and that the associated uncertainty reflects all factors that could reasonably affect the estimate. Moreover, the methodology described in [6] also incorporated *corrective actions* aimed at reducing the evaluated *measurement-model uncertainty*. Those actions included, for example, improving the quality of the training data to mitigate the *aleatoric* contribution of uncertainty; enlarging the training dataset or refining the model structure to reduce the *epistemic* contribution, and optimizing the training and validation processes.

The iterative application of these actions allowed, in many cases, to achieve a *measurement-model uncertainty* component negligible with respect to the *physical quantity uncertainty* one, thus validating the ML-based measurement model under the VIM assumption. However, the results obtained also demonstrated that this condition cannot always be fulfilled. Even after applying the corrective actions, the *measurement-model uncertainty* component can remain comparable to (or even greater than) the *physical quantity uncertainty* one, particularly when the available data are limited, noisy, or not fully representative of the measurement conditions, or when the ML model exhibits significant nonlinearity. In such circumstances, neglecting the contribution of the measurement model would lead to an incomplete or biased uncertainty evaluation,

[2]In deterministic ML models, once deployed, a fixed set of input quantities yields a single point estimate; thus, repeated evaluations at the same input return the same output. This assesses model uncertainty more challenging than in stochastic ML models, which can generate multiple outputs for the same fixed input (e.g., via Monte Carlo dropout or deep ensembles) and thereby directly characterize output variability [3], [12]. The focus on deterministic models is due to their widespread use, lower computational cost, and easier reproducibility; moreover, stochastic approaches often rely on sampling-based approximations that may introduce additional variance and require longer inference time to obtain stable estimates [9].

thereby undermining the trustworthiness of the measurement result, since the reported uncertainty would no longer adequately represent all relevant sources of variability and influence.

III. Method

The proposed method enables the uncertainty evaluation in ML-assisted measurements under the general assumption that the *measurement-model uncertainty* component is nonnegligible with respect to the *physical quantity uncertainty* one. It is designed to be model- and domain-independent, while remaining compliant with the JCGM framework and is based on the principles underlying the GUM-S1 approach to obtain a probabilistic representation of the measurand.

According to the conventional GUM-S1 approach, each input quantity $X_i \in (X_1, X_2, \ldots, X_N)$ is characterized by an assigned probability distribution [7]. These distributions are then sampled by generating a finite number M of *Monte Carlo samples*, i.e., realizations of the input quantities obtained by random sampling from the assigned distributions [7]. The resulting set of sampled inputs is finally propagated through the measurement model (in this case, the ML-based measurement model $\hat{h}$). This procedure yields a numerical approximation of the probability distribution of the measurand Y, obtained as an empirical representation based on the simulated results, namely the Monte Carlo samples of the measurand. The resulting approximate probability distribution enables the determination of the best estimate of the measurand (given by the arithmetic mean of the Monte Carlo samples) and the associated standard uncertainty (given by the sample standard deviation of the Monte Carlo samples). It also enables the determination of a coverage interval, representing the range of values that can reasonably be attributed to the measurand.

However, as discussed above, this holds only when the *measurement-model uncertainty* component is negligible with respect to the *physical quantity uncertainty* one. Since the ML-based measurement model $\hat{h}$ is only an approximation of the *true* relationship f^*, this assumption may not always be satisfied. In particular, recalling the expression of the measurand introduced in (1), the random component of the error E_{rand} may indeed be relevant, thus making the *measurement-model uncertainty* component nonnegligible. In such cases, each of the Monte Carlo M samples of the measurand is subject to uncertainty and should therefore be regarded as a random variable.

If a probability distribution is available for the random error component E_{rand} (e.g., inferred from calibration results for deterministic ML regressors, or obtained via Monte Carlo dropout or deep ensembles for stochastic ML regressors, as introduced in Section II), each of the M Monte Carlo samples of the measurand can itself be represented by a probability distribution (reflecting the *measurement-model uncertainty* component). This distribution can be sampled by generating a nested set of Q Monte Carlo extractions, following the same guidelines adopted for selecting M as described in the GUM-S1 document [7].

In summary, the proposed method allows for the construction of a 2-D matrix of Monte Carlo samples of size $M \times Q$. Here, M is the number of samples used to represent the probability distributions of the input quantities, while Q denotes the number of samples representing the probability distribution associated with each Monte Carlo sample produced by the ML-based measurement model affected by uncertainty. The resulting matrix of $M \times Q$ Monte Carlo samples corresponds to the empirical probability distribution $\hat{g}_Y(y)$ of the measurand Y. This distribution is obtained via Monte Carlo sampling from the joint probability distribution $\pi(Y, X)$, which accounts for both the variability introduced by the measurement model and that arising from the measurement of the input quantities. By arranging all the samples in a single row [13], the best estimate of Y is evaluated as the arithmetic mean, while the overall measurement uncertainty is obtained from the sample standard deviation, in compliance with the traditional GUM-S1 framework [7] but considering both the *physical quantity uncertainty* and the *measurement-model uncertainty* components.

Notably, an alternative and well-established method for combining measurement uncertainty components is the variance decomposition approach,[3] expressed as

$$u(Y) = \sqrt{u^2_{phys}(Y) + u^2_{mod}(Y)} \tag{2}$$

where $u^2_{phys}(Y)$ represents the *physical quantity uncertainty*, and $u^2_{mod}(Y)$ represents the *measurement-model uncertainty*. In fact, these two components of the measurement uncertainty can also be evaluated separately, as shown in previous work [6]. In summary, $u^2_{phys}(Y)$ can be obtained by applying the traditional GUM-S1 approach, thus neglecting the contribution due to model uncertainty. By contrast, $u^2_{mod}(Y)$ is derived from the best estimate of the Monte Carlo samples of the measurand, obtained according to the traditional GUM-S1 approach, and from the calibration diagram of the ML-based measurement model obtained during the calibration phase, thus neglecting the variability of the input physical quantities.

Overall, the variance decomposition approach [(2)] is commonly used as a formal benchmark for the combination of different variability contributions. Accordingly, it may also be employed in scenarios such as ML-assisted measurements, where multiple components of uncertainty coexist. However, the main advantage of the proposed method over variance decomposition lies in its ability to provide a numerical approximation of the entire probability distribution of the measurand, in accordance with GUM-S1, at the cost of potentially increased computational effort. This enables the determination of coverage intervals without relying on the assumption that the measurand follows either a Student's t-distribution, as implied by the Welch-Satterthwaite equation [15], or a normal distribution, as suggested by the central limit theorem [5], [16], assumptions that may not always hold.

As a representative example, this work considers 95% coverage intervals. According to GUM-S1 [7], this interval is not necessarily symmetric about the best estimate and corresponds

[3]The variance decomposition approach involves determining the total combined variance by summing the individual variances of the uncorrelated contributing factors, in this case, the ML-based measurement model and the input quantities [14].

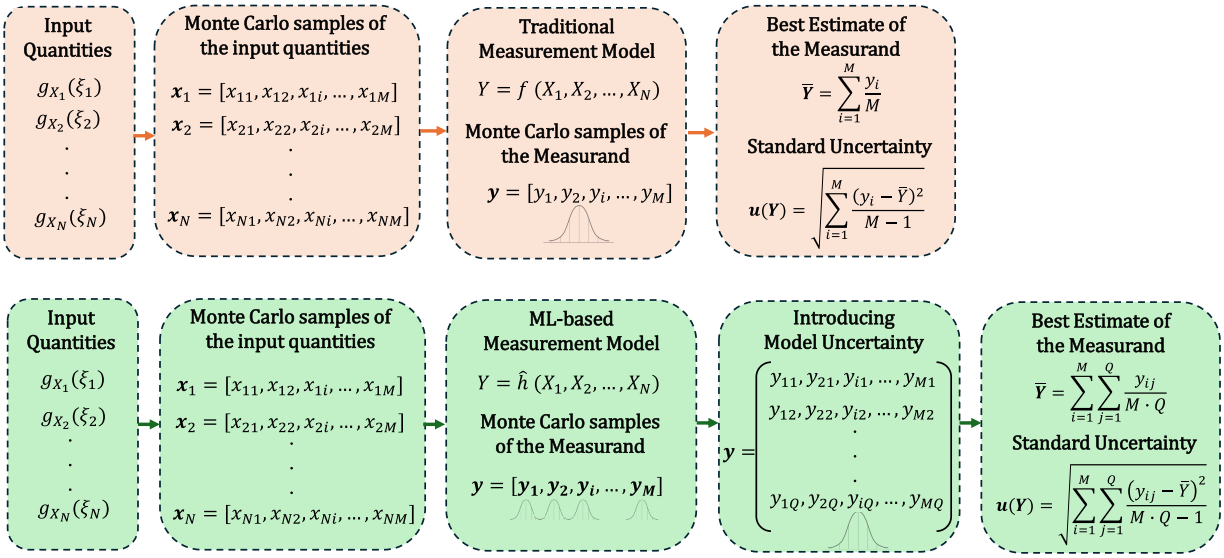

Fig. 2. Illustration of the proposed method (in green) compared with the traditional GUM-S1 approach (in orange). Notably, in the proposed method, each of the M Monte Carlo samples is represented by a probability distribution arising from the variability in the measurement model.

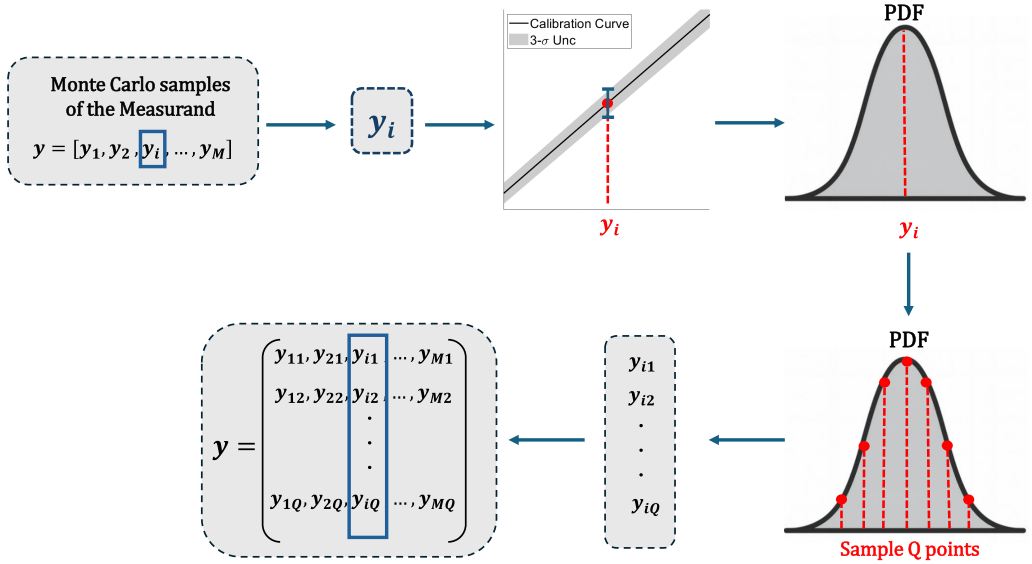

Fig. 3. Procedure for introducing model uncertainty in each Monte Carlo sample of the measurand. Given a Monte Carlo sample y_i, the corresponding value is mapped through the model calibration curve, from which the probability density function (pdf) of the sample is obtained. The resulting pdf is then sampled at Q points, generating a vector $(y_{i1}, y_{i2}, \ldots, y_{iQ})$ that forms the ith column of the $M \times Q$ matrix.

to the narrowest interval containing 95% of the probability mass. More generally, the same approach can be used to determine coverage intervals at any required probability level, directly accounting for the actual shape of the distribution.

For the sake of clarity, Fig. 2 illustrates the proposed method in comparison with the traditional GUM-S1 approach for the case of deterministic ML regressors. Moreover, for a more detailed overview of the methodology, Fig. 3 shows the sequential steps involved in constructing the $M \times Q$ matrix necessary to combine model and input-related uncertainty components. In addition, to provide a clear operational guide for the implementation of the proposed method, the sequential steps for constructing the $M \times Q$ matrix and evaluating the final measurement result are summarized in the form of a pseudocode (see Algorithm 1).

IV. Case Study I

The first case study focuses on the measurement of the full-load electrical power output of a combined cycle power plant (CCPP) operating under base-load conditions [17], already introduced in [6]. The measurement of this quantity is essential for optimizing plant efficiency, maximizing the economic yield from available energy resources, and ensuring the reliability and sustainability of the overall system. In this context, a physically-based measurement model would require the solution of a complex set of nonlinear thermodynamic equations.

Algorithm 1 Pseudocode of the Proposed Method

Require: (i) PDFs $[g_{X_1}(\xi_1), \ldots, g_{X_N}(\xi_N)]$ of the physical input quantities $(X_1, \ldots, X_N)$; (ii) trained ML model $\hat{h}$; (iii) model calibration diagram.

Ensure: (i) Best estimate $\bar{Y}$, (ii) standard uncertainty $u(Y)$, and (iii) empirical pdf $\hat{g}_Y(y)$ of the measurand Y.

Phase 1 — Propagation of Input Quantity PDFs (M Monte Carlo samples)
for $i = 1$ **to** M **do**
  Generate the ith input vector $[x_{1i}, x_{2i}, \ldots, x_{Ni}]$ by independently sampling each input quantity $(X_1, \ldots, X_N)$ from the assigned PDFs $[g_{X_1}(\xi_1), \ldots, g_{X_N}(\xi_N)]$.
  Compute the ith Monte Carlo sample of the measurand: $y_i = \hat{h}(x_{1i}, \ldots, x_{Ni})$
end for
Phase 2 — Introduction of Measurement Model Uncertainty (Q Monte Carlo samples)
for $i = 1$ **to** M **do**
  Map y_i through the model calibration diagram to identify its pdf $g_{Y_i}(y)$
  for $j = 1$ **to** Q **do**
    Extract the jth sample y_{ij} from $g_{Y_i}(y)$
  end for
  Form the ith column of the $M \times Q$ matrix: $[y_{i1}, y_{i2}, \ldots, y_{iQ}]^T$
end for
The collection $\{y_{ij}\}$, with $i = 1, \ldots, M$ and $j = 1, \ldots, Q$, defines the empirical distribution $\hat{g}_Y(y)$ of the measurand Y
Phase 3 — Measurement Uncertainty Evaluation
Compute the best estimate

$$\bar{Y} = \sum_{i=1}^{M} \sum_{j=1}^{Q} \frac{y_{ij}}{MQ}$$

Compute the standard uncertainty

$$u(Y) = \sqrt{\sum_{i=1}^{M} \sum_{j=1}^{Q} \frac{(y_{ij} - \bar{Y})^2}{MQ - 1}}$$

Determine the shortest coverage interval at the selected coverage probability from the empirical distribution $\hat{g}_Y(y)$

The associated computational demand and limited analytical tractability make the formulation of an explicit analytical model impractical for real-world applications. This challenge motivates the use of data-driven models and underscores the need to extend the GUM metrological framework to encompass such approaches. In particular, the input quantities for the indirect measurement of this physical quantity are the ambient temperature (AT), atmospheric pressure (AP), relative humidity (RH), and vacuum pressure (V). In Fig. 4, a schematic representation of the power plant is provided.

A. Description of the Dataset

The dataset employed in this first case study is publicly available in [18]. Regarding the considered input quantities,

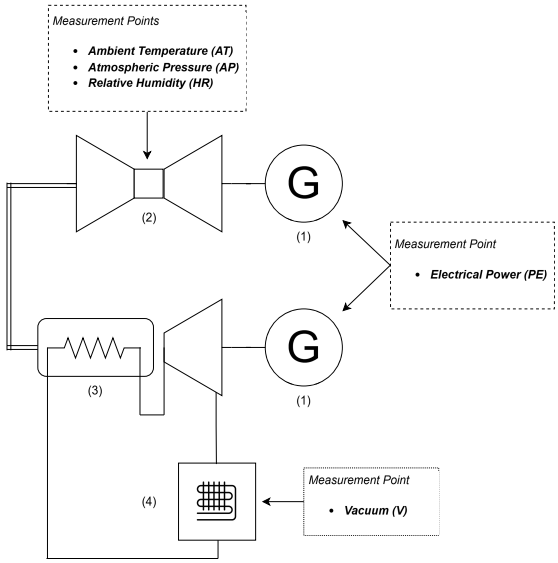

Fig. 4. Block diagram of the combined cycle power plant considered in Case Study I (Legend: 1-electric generators, 2-gas turbine, 3-heat recovery steam generator, and 4-condenser).

TABLE I
SEARCH SPACE ADOPTED FOR HYPERPARAMETERS FINE-TUNING IN CASE STUDY I

Tuned Hyperparameters	Search Space
number of hidden layers	{1, 2, 3, 4}
number of neurons in each layer	{10, 20, 50, 100}
learning rate	$\{10^{-2}, 10^{-3}, 10^{-4}\}$
batch size	{32, 64, 100}

AT ranges from 1.81 to 37.11 °C, AP from 992.89 to 1033.30 mbar, RH from 25.56% to 100.00%, and V from 25.36 to 81.56 cmHg. The measurand ranges from 420.26 to 495.76 MW. In accordance with the approach outlined in [6] and previously discussed therein, the following data preparation steps were conducted. First, to simulate the measurement uncertainty inherent to data acquisition, zero-mean Gaussian noise with a relative standard deviation of 5% was added to each input quantity. This noise model was selected due to the absence of detailed information regarding the specific instrumentation used for data collection. Then, approximately 70% of the available data were used for the identification and training of the ML-based measurement model, while the remaining samples were employed for model calibration and subsequent on-field employment. Finally, the measurand values were rounded to ensure a sufficient number of samples, specifically, 100 instances in the training dataset and 30 instances in the calibration dataset for each measurand value.

B. Training and Calibration of the ML-Based Measurement Model

The ML model adopted was a feedforward artificial neural network (ANN) employing the rectified linear unit (ReLU) as the activation function in its hidden layers. The choice of an ANN architecture was driven by its proven capability to effectively capture the nonlinearities between input quantities and the electrical power output that lack a closed-form

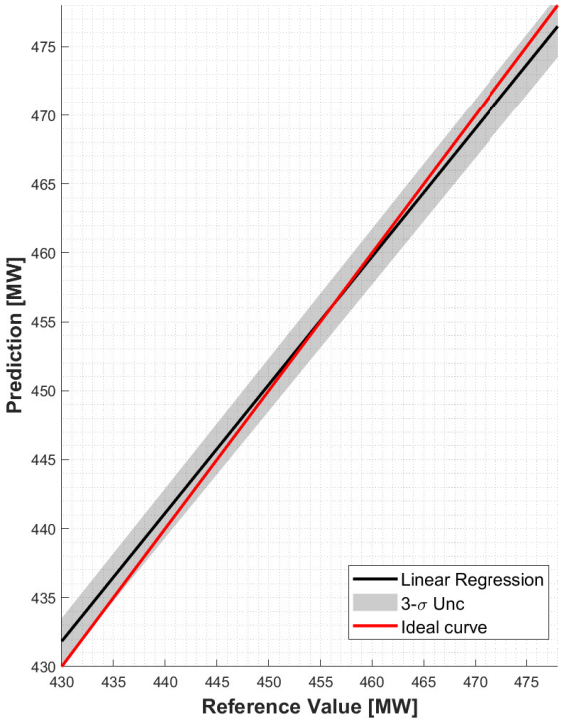

Fig. 5. Calibration diagram of the ML-based measurement model for Case Study I. The diagram includes: the ideal 1:1 curve (red line), the linear regression fit (black line), and the 3-σ uncertainty band (gray area).

analytical description. The training data were divided into training and validation subsets. A min–max normalization was applied, with the scaling parameters (minimum and maximum values) computed from the training subset. To identify the optimal hyperparameter configuration, a grid search approach combined with a hold-out validation strategy was employed, considering the search space reported in Table I. The optimal ANN configuration consisted of two hidden layers with 10 neurons each, trained using a learning rate of 10^{-2} and a batch size of 64, achieving a root mean squared error (RMSE) of 4.7 MW. The resulting ML-based measurement model was subsequently calibrated to obtain the corresponding calibration diagram, to assess whether the *measurement-model uncertainty* could be considered negligible compared to the *physical quantity uncertainty*. In particular, in [6] the model calibration procedure was carried out as follows. First, a calibration dataset was developed containing 30 distinct realizations of the input quantities for each reference (target) value of the measurand. Thus, for each target value, the ML model produced 30 predictions, from which: 1) the measurement error was computed as the difference between the reference value and the mean prediction and 2) the associated standard uncertainty was obtained via a Type-A evaluation from the dispersion of the predictions. The resulting calibration diagram (Fig. 5) can be used to correct systematic effects and to assign a pointwise *measurement-model uncertainty* to subsequent model outputs in actual measurement operations. In addition, the normality

TABLE II

MEASUREMENTS OF THE INPUT QUANTITIES FOR THE EMPLOYMENT OF THE ML-BASED MEASUREMENT MODEL FOR THE CASE STUDY I

Input Quantity	Best Estimate	Standard Uncertainty	PDF
AT	18.78 °C	0.58 °C	Uniform
AP	1019.2 mbar	0.6 mbar	Uniform
RH	80.20 %	2.9 %	Uniform
V	63.09 cm Hg	0.04 cm Hg	Uniform

of the prediction dispersion (i.e., the calibration residuals) was verified using a χ^2 goodness-of-fit test [19], thereby supporting the assignment of a pdf to each ML-model output.

C. Combination of the Uncertainty Components

For the experimental scenario, the considered measured values of the input quantities are reported in Table II, together with their corresponding standard uncertainties and assumed probability density functions (PDFs), based on typical off-the-shelf measuring instruments. In particular, for illustrative purposes, it was assumed that the AT and the RH were measured using a *DHT22* sensor, the vacuum level (V) using a *DVR2pro* vacuum gauge, and the AP using an *LPS22DF* sensor.

As a first step, Monte Carlo simulations were carried out according to the traditional GUM-S1 framework, using 10^6 pseudorandom samples generated in the MATLAB environment. By propagating the input PDFs through the ML-based measurement model, the empirical distribution of the measurand was obtained. From this distribution, the best estimate was determined to be 451.0 MW, along with the *physical quantity uncertainty* of 1.5 MW, ensuring that the uncertainty was reported with at maximum two significant figures. According to GUM-S1, the measurement result can therefore be expressed as (451.0 $\pm$ 1.5) MW. However, as previously discussed, the GUM-S1 framework relies on the assumption of negligible *measurement-model uncertainty*. From the calibration diagram, a representation of which is shown in Fig. 5, the *measurement-model uncertainty* was evaluated to be approximately 0.6 MW.

Hence, it is questionable whether this uncertainty component can actually be considered negligible compared with the *physical quantity uncertainty*. By applying the decomposition of variance through (2), a combined uncertainty of 1.7 MW is obtained, again reported with at maximum two significant figures. Therefore, it can be concluded that the GUM-S1 framework provided a measurement uncertainty value lower than what would be more appropriate to consider, and the measurement result should be reported as (451.0 $\pm$ 1.7) MW. Nevertheless, as discussed, the variance decomposition used in (2) does not provide information about the output pdf of the measurand, which is essential for determining reliable coverage intervals. For this reason, the proposed method was applied, considering a combination of parameters M and Q for the Monte Carlo simulations.

The results are reported in Table III. As visible, the best estimate is consistent with the value obtained by applying the traditional GUM-S1 framework. Concerning the measurement uncertainty, even with a relatively low number of Monte

TABLE III

OBTAINED BEST ESTIMATE AND MEASUREMENT UNCERTAINTY AS A FUNCTION OF THE PARAMETERS M AND Q, AND THE COMPUTATIONAL TIME NEEDED FOR THE CASE STUDY I

M	Q	Time [s]	Best Estimate [MW]	Measurement Uncertainty [MW]
10^3	10^3	< 0.5	$451.0008 \approx 451.0$	$\mathbf{1.6}954 \approx 1.7$
10^4	10^3	< 1.0	$450.9586 \approx 451.0$	$\mathbf{1.6}833 \approx 1.7$
10^4	10^4	< 1.0	$450.9536 \approx 451.0$	$\mathbf{1.6}739 \approx 1.7$
10^5	10^3	< 5.0	$450.9603 \approx 451.0$	$\mathbf{1.6}887 \approx 1.7$
10^6	10^3	< 5.0	$450.9561 \approx 451.0$	$\mathbf{1.6}879 \approx 1.7$
$5 \cdot 10^4$	$5 \cdot 10^4$	< 50.0	$450.9573 \approx 451.0$	$\mathbf{1.6}779 \approx 1.7$

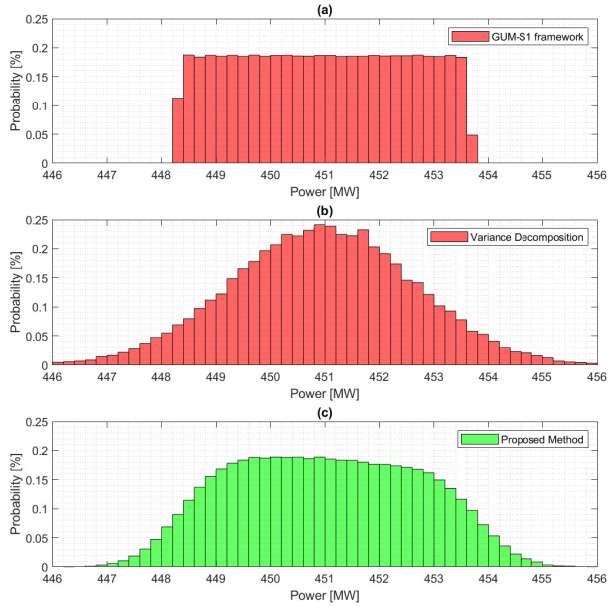

Fig. 6. Case Study I: resulting probability density function of the measurand according to (a) GUM-S1 framework, (b) variance decomposition approach, and (c) proposed method.

Carlo samples, the results obtained with the proposed method (1.7 MW) are consistent with the analytical benchmark derived from the variance decomposition approach [(2)]. Furthermore, the stability of the result across different values of M and Q demonstrates the convergence of the algorithm and its robustness even at lower computational scales. Therefore, even with computation times below 1 s, the proposed numerical method allows for a proper evaluation of the uncertainty, with the additional advantage of providing the probability distribution of the measurand. The computations were performed on an off-the-shelf PC equipped with an Intel i7-10850H CPU and 16 GB of RAM.

Fig. 6 shows a comparison between the PDFs obtained using the traditional GUM-S1 framework Fig. 6(a), the variance decomposition approach Fig. 6(b), and the proposed method Fig. 6(c). As shown, the proposed method accounts for the model-uncertainty component, resulting in a wider pdf than that derived from the traditional GUM-S1 approach. Furthermore, the resulting pdf differs from the normal distribution that would be obtained by applying (2) and relying on the central limit theorem. The practical impact is further highlighted by comparing the 95% shortest coverage intervals obtained with

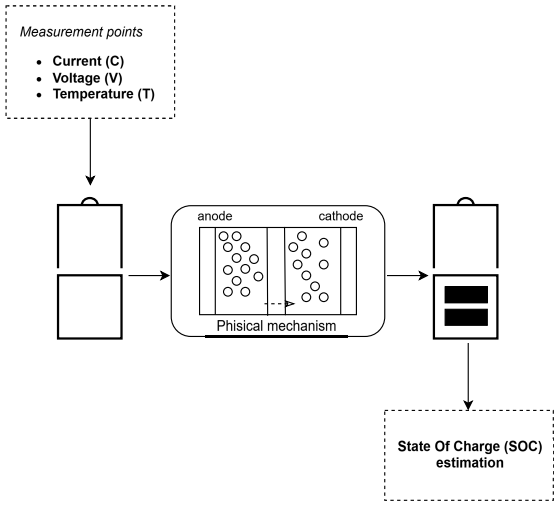

Fig. 7. Block diagram of the SoC model for the lithium battery considered in Case Study II.

TABLE IV

SEARCH SPACE ADOPTED FOR HYPERPARAMETERS FINE-TUNING IN CASE STUDY II

Tuned Hyperparameters	Search Space
number of neurons	{32, 64, 128}
dropout probability	{0.1, 0.3, 0.5}
learning rate	$\{10^{-2}, 5\times\ 10^{-3}, 10^{-3}\}$
batch size	{16, 32, 64}

the three approaches. In particular, the traditional GUM-S1 approach (neglecting model uncertainty) results in a narrower interval of [448.3 and 453.4] MW, while the analytical variance decomposition, which relies on the assumption of a normal distribution, provides an interval of [447.8 and 454.4] MW. On the contrary, the proposed method yields a more reliable range of [448.0 and 454.1] MW by accounting for the actual shape of the distribution. In power plant management, relying on the underestimated interval of the traditional GUM-S1 would lead to an overconfident estimation, potentially causing grid dispatching errors or financial risks due to unfulfilled commitments.

V. CASE STUDY II

The second case study concerns the indirect measurement of the state of charge (SoC) of a lithium-ion battery cell. The SoC quantifies the remaining capacity available in a battery relative to its nominal capacity and is typically expressed as a percentage ranging from 0% (fully discharged) to 100% (fully charged). Accurate knowledge of SoC is crucial in many applications, such as electric vehicles and stationary storage systems, as it directly impacts energy management strategies, safety margins, and the long-term durability of the storage system. From a metrological standpoint, SoC is a latent quantity that cannot be directly measured; instead, it must be inferred from other measurable electrical and thermal quantities, such as terminal voltage, current, and cell temperature, typically via equivalent-circuit models or electrochemical models (see Fig. 7). These models, however, require detailed parameter

identification and can become complex when realistic operating conditions are considered (e.g., varying temperature, aging, and dynamic load profiles). This motivates the adoption of data-driven approaches, where an ML model is trained to map input physical quantities to SoC without explicitly formulating the underlying electrochemical dynamics.

A. Description of the Dataset

The dataset used in this second case study is publicly available [20] and consists of experimental measurements obtained from LG 18650HG2 lithium-ion cells. The data were collected under a variety of discharge conditions and ambient temperatures, simulating realistic battery usage scenarios, including constant current discharge, dynamic load profiles, and standardized drive cycles. The dataset provides time-series recordings of voltage (V), current (I), and temperature (T), enabling the development and assessment of models aimed at learning the relationship between these inputs and the SoC of the battery.

The voltage V ranges from 2.79 to 4.24 V, the current C from −18.09 to 6.00 mA, and the temperature T from −0.84 °C to 26.82 °C. Each instance in the dataset consists of time-series features, each aligned with the corresponding SoC target. These features are organized as fixed-length sequences, with a length of 300 time samples.

As a data preparation step for illustrating the proposed methodology, the values of SoC were rounded to the nearest integer (e.g., 1.0%, 2.0%, ..., 100.0%) to ensure a sufficient number of samples for each measurand value, necessary for the *model calibration* stage [6]. In particular, this procedure guaranteed a number of 30 instances per measurand value in the calibration dataset, for a total of 3000 instances. The remaining instances were allocated to the training dataset (approximately 4780 instances).

B. Training and Calibration of the ML-Based Measurement Model

The ML model adopted was a long short-term memory (LSTM) neural network, a type of recurrent neural network. This architecture was selected for its specialized ability to model temporal dependencies in the time-series measurement data provided by the battery sensors. The training dataset was partitioned by allocating 20% of the samples to the validation subset. Subsequently, a min–max normalization was applied, with the scaling parameters (minimum and maximum values) computed from the training subset. To determine the optimal hyperparameter configuration, a grid search approach combined with a hold-out validation strategy was employed. The set of tuned hyperparameters and the corresponding search space are reported in Table IV. The resulting optimal LSTM architecture consisted of a single hidden layer with 64 neurons, trained using a learning rate of 10^{-2}, a batch size of 16, and a dropout probability of 0.1. This configuration achieved an RMSE of 2.69%.

Therefore, the derived ML-based measurement model was subsequently calibrated to derive the corresponding calibration diagram, aiming to evaluate whether the *measurement-model uncertainty* could be regarded as negligible in comparison with the *physical quantity uncertainty*. The calibration results are presented in Fig. 8. Also in this case study, with respect to the evaluated *measurement-model uncertainty*, the hypothesis of a normal distribution was assessed by using the χ^2 test [19].

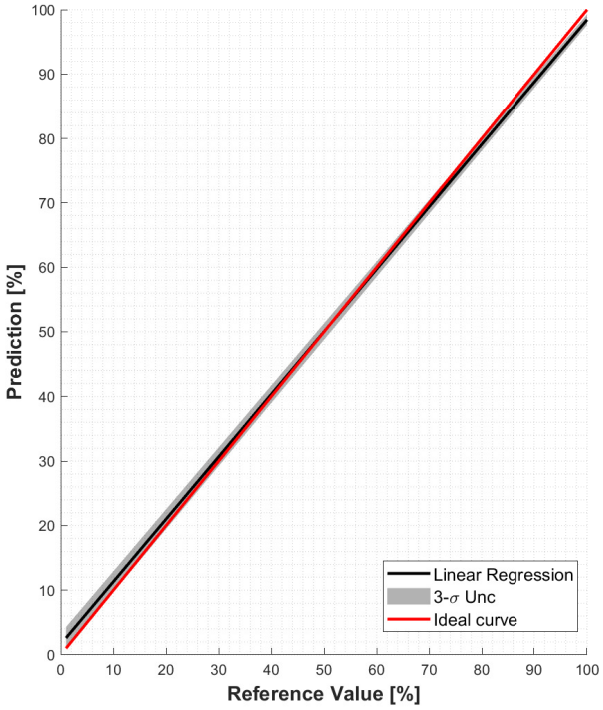

Fig. 8. Calibration diagram of the ML-based measurement model for Case Study II. The diagram includes: the ideal 1:1 curve (red line), the linear regression fit (black line), and the 3-σ uncertainty band (gray area).

C. Combination of the Uncertainty Components

For the experimental scenario, the considered measured values of the input quantities (composed of 300 time samples for V, C, and T) are reported in Table V. For the sake of brevity, only the range from the minimum to the maximum value is provided. Each value of the time sample is associated with a standard uncertainty (expressed as a percentage) and a pdf based on typical off-the-shelf measuring instruments. In particualr, for illustrative purposes, it was assumed that the voltage V, and the current I were measured with a *Kaiweets* multimeter, while the temperature T was measured with a *DHT11* sensor.

Similar to Case Study I, as a first step, Monte Carlo simulations were carried out in the MATLAB environment according to the traditional GUM-S1 framework, considering several samples equal to 10^6. By propagating the input PDFs through the ML-based measurement model, the empirical distribution of the measurand was obtained. From this distribution, the best estimate was determined to be 85.8%, along with the *physical quantity uncertainty* of 1.4%, ensuring that the uncertainty was reported with at maximum two significant figures. According to GUM-S1, the measurement result can therefore be expressed as (85.8 ± 1.4)%. However,

TABLE V

MEASUREMENTS OF THE INPUT QUANTITIES FOR THE EMPLOYMENT OF THE ML-BASED MEASUREMENT MODEL FOR THE CASE STUDY II

Input Quantity	Min-Max Values	Standard Uncertainty [%]	PDF
V	(3.9318 - 4.1158) V	1.0	Uniform
C	$(-2.092$ - $1.8671)$ mA	1.5	Uniform
T	$(-0.3155$ - $-0.2103)$ °C	5.0	Uniform

TABLE VI

OBTAINED BEST ESTIMATE AND MEASUREMENT UNCERTAINTY AS A FUNCTION OF THE PARAMETERS M AND Q, AND THE COMPUTATIONAL TIME NEEDED FOR CASE STUDY II

M	Q	Time [min]	Best Estimate [%]	Measurement Uncertainty [%]
10^3	10^3	$(5-10)$	$85.86 \approx 85.9$	$\mathbf{1.46} \approx 1.5$
10^4	10^3	$(5-10)$	$85.82 \approx 85.8$	$\mathbf{1.44} \approx 1.4$
10^4	10^4	$(5-10)$	$85.81 \approx 85.8$	$\mathbf{1.44} \approx 1.4$
10^5	10^3	$(5-10)$	$85.82 \approx 85.8$	$\mathbf{1.45} \approx 1.5$
10^6	10^3	$(5-10)$	$85.82 \approx 85.8$	$\mathbf{1.45} \approx 1.5$
$5 \cdot 10^4$	$5 \cdot 10^4$	$(5-10)$	$85.82 \approx 85.8$	$\mathbf{1.44} \approx 1.4$

as previously discussed, the GUM-S1 framework relies on the assumption of negligible *measurement-model uncertainty*. From the calibration diagram, whose representation is reported in Fig. 8, it was possible to observe that the *measurement-model uncertainty*, evaluated at the model prediction, was equal to about 0.3%. Hence, it is again questionable whether this uncertainty component can actually be considered negligible compared with the *physical quantity uncertainty* of 1.4%. By applying the decomposition of variance through (2), a combined uncertainty of 1.5% is obtained, again reported with at maximum two significant figures. Therefore, also in Case Study II, the GUM-S1 framework provided a measurement uncertainty value lower than what would be more appropriate to consider, as the measurement result should be reported as $(85.8 \pm 1.5)\%$. Nevertheless, as already discussed, information about the output pdf of the measurand is missing. For this reason, the proposed method was applied, considering a combination of parameters M and Q for the Monte Carlo simulations.

The results are summarized in Table VI. Again, the best estimate is consistent with the value obtained by applying the traditional GUM-S1 framework. Concerning the measurement uncertainty, the resulting value (1.5%) is consistent with the analytical benchmark [(2)]. As observed in the previous case study, this consistency across different M and Q configurations serves as a formal verification of the method's reliability and numerical robustness. Moreover, the proposed method offers the additional advantage of providing the full probability distribution of the measurand, with total computation times ranging between 5 and 10 min. The simulations were performed on a workstation equipped with an Intel i7-10700 CPU and 32 GB of RAM.

Finally, in Fig. 9, a comparison between the PDFs obtained using the traditional GUM-S1 framework Fig. 9(a), the variance decomposition approach Fig. 9(b), and the proposed method Fig. 9(c) is shown. In the same way as in Case Study I, the proposed method accounts for the model-uncertainty

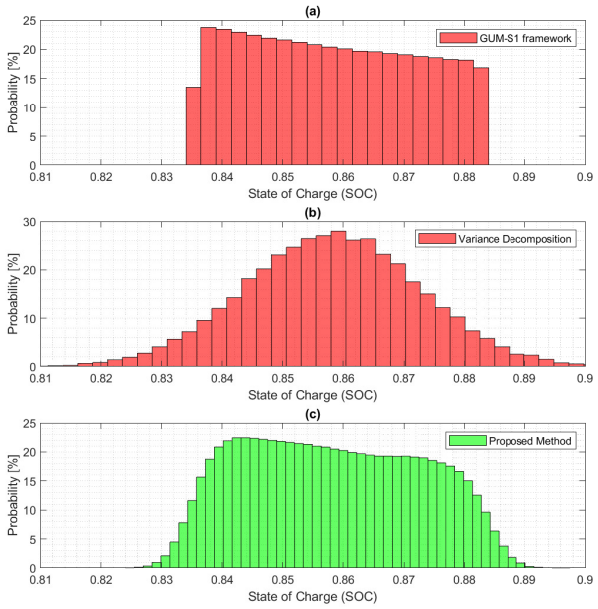

Fig. 9. Case Study II: resulting probability density function of the measurand according to (a) GUM-S1 framework, (b) variance decomposition approach, and (c) proposed method.

component, resulting in a wider pdf than that derived from the traditional GUM-S1 approach. Furthermore, the resulting pdf differs from the normal distribution that would be obtained by applying (2) and relying on the central limit theorem. Comparing the 95% shortest coverage intervals, the traditional GUM-S1 approach (neglecting model uncertainty) provides a range of [83.5 and 88.1]%, whereas the variance decomposition approach, assuming a Gaussian pdf, results in an interval of [83.0 and 88.6]%. On the other hand, the proposed nested MCM identifies a range of [83.5 and 88.4]%, reflecting a more reliable state of knowledge about the measurand. For a battery management system, using the more realistic interval provided by the proposed method is vital to ensure safety and longevity, as an overconfident estimate could lead to improper cell balancing or unexpected deep discharge, accelerating premature cell degradation.

VI. CONCLUSION AND FUTURE WORK

This article addressed the combination of uncertainty components in ML-assisted measurements using an approach consistent with both the GUM-S1 Monte Carlo framework and variance-decomposition principles. In particular, once the uncertainty contribution associated with the ML-based measurement model and that arising from the measurement of the input physical quantities are evaluated, they are combined via nested Monte Carlo simulations to obtain the empirical distribution of the measurand, accounting for all relevant contributions to measurement uncertainty. The proposed approach was validated through two case studies in which ML enables measurements that would otherwise be difficult to perform.

A key strength of the proposed methodology is its model- and domain-independent formulation, which makes it applicable to a broad class of ML regressors across different

measurement scenarios. This generality is supported by the two case studies, which differ in both problem structure and modeling complexity, and by the use of distinct architectures (ANN and LSTM).

The results indicate that the proposed approach provides a more comprehensive uncertainty evaluation than applying either the conventional GUM-S1 procedure or variance decomposition alone, without a significant increase in computational effort or dependence on specific backbone models. Future work will investigate extensions to stochastic regression models within both frequentist and Bayesian frameworks.

REFERENCES

[1] S. Shirmohammadi and H. Al Osman, "Machine learning in measurement part 1: Error contribution and terminology confusion," *IEEE Instrum. Meas. Mag.*, vol. 24, no. 2, pp. 84–92, Apr. 2021.

[2] H. Al Osman and S. Shirmohammadi, "Machine learning in measurement part 2: Uncertainty quantification," *IEEE Instrum. Meas. Mag.*, vol. 24, no. 3, pp. 23–27, May 2021.

[3] T. Adel, S. Bilson, M. Levene, and A. Thompson, "Trustworthy artificial intelligence in the context of metrology," in *Producing Artificial Intelligent Systems: The Roles of Benchmarking, Standardisation and Certification*. Cham, Switzerland: Springer, 2024, pp. 53–75.

[4] International Vocabulary of Metrology—Basic and General Concepts and Associated Terms (VIM), Standard JCGM 200:2012, 2012.

[5] Evaluation of Measurement Data—Guide to the Expression of Uncertainty in Measurement, Standard JCGM 100:2008, 2008.

[6] L. Angrisani et al., "Machine learning-based measurement models: A novel methodology for GUM-based uncertainty evaluation," *IEEE Trans. Instrum. Meas.*, vol. 74, pp. 1–11, 2025.

[7] Evaluation of Measurement Data—Supplement 1 To the 'Guide to the Expression of Uncertainty in Measurement' Propagation of Distributions Using a Monte Carlo Method, Standard JCGM 101:2008, 2008.

[8] Guide To the Expression of Uncertainty in Measurement—Part 6: Developing and Using Measurement Models, Standard JCGM GUM-6:2020, 2020.

[9] E. Hüllermeier and W. Waegeman, "Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods," *Mach. Learn.*, vol. 110, no. 3, pp. 457–506, Mar. 2021.

[10] L. Mari, M. Wilson, and A. Maul, *Measurement Across the Sciences: Developing a Shared Concept System for Measurement*. Cham, Switzerland: Springer, 2023.

[11] E. Buchicchio, P. Carbone, A. De Angelis, F. Santoni, and A. Moschitta, "Uncertainty quantification in AI-based measurement systems," *IEEE Instrum. Meas. Mag.*, vol. 28, no. 3, pp. 52–59, May 2025.

[12] J. Gawlikowski et al., "A survey of uncertainty in deep neural networks," *Artif. Intell. Rev.*, vol. 56, no. 1, pp. 1513–1589, 2023.

[13] M. G. Larson, "Analysis of variance," *Circulation*, vol. 117, no. 1, pp. 115–121, 2008.

[14] P. Willems, "Model uncertainty analysis by variance decomposition," *Phys. Chem. Earth, A/B/C*, vols. 42–44, pp. 21–30, Jan. 2012.

[15] B. Hall and R. Willink, "'Does' Welch-Satterthwaite' make a good uncertainty estimate?," *Metrologia*, vol. 38, no. 1, p. 9, 2001.

[16] R. M. Dudley, "Central limit theorems for empirical measures," *Ann. Probab.*, vol. 6, no. 6, pp. 899–929, 1978.

[17] P. Tüfekci, "Prediction of full load electrical power output of a base load operated combined cycle power plant using machine learning methods," *Int. J. Electr. Power Energy Syst.*, vol. 60, pp. 126–140, Sep. 2014.

[18] *Combined Cycle Power Plant [Dataset]*. Accessed: Jan. 11, 2026. [Online]. Available: https://archive.ics.uci.edu/dataset/294/combined+cycle+power+plant

[19] B. Y. Lemeshko, "Chi-square-type tests for verification of normality," *Meas. Techn.*, vol. 58, no. 6, pp. 581–591, Sep. 2015.

[20] *Lithium Ion Battery State of Charge Estimation*. Accessed: Jan. 11, 2026. [Online]. Available: https://www.kaggle.com/code/dhakalamrit/lithium-ion-battery-state-of-charge-estimation/notebook

Open Access funding provided by 'University of Naples Federico II' within the CRUI CARE Agreement