TC7 - Measurement Science
TC11 - Testing, Inspection and Certification
TC13 - Biology and Medicine
TC21 - Mathematical Tools

In association with

TC6 - Digitalization

- Mathmet

Joint International Conference 2026

Trusted Measurement, Trusted Analysis

Supporting Science-driven Decision-making

Incorporating

AMCTM 26

Book of Abstracts

27 - 29 October

FARO – ALGARVE - PORTUGAL

Conference Scope

The metrology community is tasked with providing a coherent framework in which measurement results can be made interoperable through metrological traceability to an agreed set of standard units. Most technical data used in decision-making is the result of workflows that depend on measurement data, models, algorithms and software, and for such outputs to be trustworthy and used effectively, it is necessary that all elements in the workflow are transparent, principled, explainable and interpretable and can be quality assured. The overall aim of the conference is to provide a multi-disciplinary forum to address how the metrology community can support trustworthy workflows across all scientific and technical disciplines, through

- Extending metrology traceability and uncertainty quantification concepts to encompass complete workflows, from data to decisions
- Using a multidisciplinary approach to address measurement and analysis challenges relating to new measurement modalities and applications
- Understanding better the measurement theory, modelling, statistical and computational requirements in metrology and develop new mathematical and computational tools for measurement and testing.

The conference was open to the submission of papers related to the general themes of

T1 - Measurement Theory and Foundations of Measurement Science: concepts, axioms, epistemological foundations, social and ethical aspects, rules, principles, methods, and procedures underlying the obtaining, processing, expression, communication, and interpretation of correct and reliable results of measurements of physical and non-physical quantities, **including psychosocial measurements**, as well as the assessment of nominal properties.

T2 - Mathematical and statistical modelling: data analysis, machine learning and uncertainty quantification, AI in measurement systems and data processing, knowledge representation and reasoning, trustworthiness of data, models, algorithms, and software

T3 - Testing, inspection and certification: developing measurement science and best practice applied in industry, environmental protection, healthcare, safety and quality of life.

T4 - Measurements in Life Sciences: measurement of whole-body, organ, and cellular function; measurements in the clinical laboratory; medical imaging; medical information systems; diagnosis and medical decision-making; knowledge-based systems; intelligent instrumentation in medicine and healthcare.

T5 - Digitalization: fundamental concepts of measurement science that relate to digitalisation and digital transformation in science, industry, and society: digital representation and use of metrological information, digital transformation in traceability and the quality infrastructure, metrology for digitalisation and digital technologies in industry and science.

Partners & Supporters

This event is proudly supported by the following institutions:

Organized by:

Portuguese Association of Accredited Laboratories

Portuguese Institute for Quality

Host University & Venue:

University of Algarve

Sponsored by:

Sociedade de Representações de Química, Lda.

Endorsed by:

Manifesto Quality Portugal

Portuguese Association for Quality

- Provisional Programme -
(2026-10-25)

Day 1 — Tuesday, 27 October 2026

9:00 REGISTRATION, COFFEE, POSTER HANGING - Adjacent to Auditorium **1.4** and **1.5**

10:00 CONFERENCE OPENING

	Welcome Plenary Auditorium **1.5**
10:00	***Welcome*** Conference chairs: Alistair Forbes, Alvaro Ribeiro, Elisabeth Monteiro, Francesca Pennecchi, Joao Sousa
	University of Algarve Representative
	Plenary Session Chairs: **Joao Sousa, Alistair Forbes**
10:30	Giovanni Battista Rossi, ***"On the nature of modern science including measurement"***

11:10–11:30 Coffee Break, Posters

11:30	Miles McCrory, Weronika Pryciuk, Liam Wright, ***"Quantitative Image Analysis: An Approach to Estimating Uncertainties Associated with Image Segmentation"***
12:10	Katsuhiro Shirono, ***"Can predictive uncertainty in machine learning be measurement uncertainty?"***

12:50–14:00 Lunch, Posters

	Plenary Session Auditorium 1.5 Chair**: Katsuhiro Shirono**
14:00	Ignacio Lira, ***"Simple GUM-Compliant Uncertainty Evaluation for Errors-in-Variables Regression"***
14:40	Liam Wright, ***"Bridging the Gap Between Chemometrics and Metrology"***

15:20–15:50 Coffee Break, Posters

	Session 1: Mathematics and Statistics Auditorium **1.5** Chair: **Ignacio Lira**	**Session 2: Digitalisation** Auditorium **1.4** Chair**: Michael Chrubasik**
15:50	Alistair Forbes, ***"Distributions associated with nonlinear least-squares problems"***	Marina Romanchikova, ***"Defining minimum metadata for nanometrology"***
16:15	Eva Fišerová, Daniel Gogola, Gejza Wimmer, Viktor Witkovský, ***"Ellipsoid Fitting for Geometric Calibration: Uncertainty Quantification and Sphericity Testing"***	Teodor Tóth, ***"Impact of Acquisition Parameters on the Metrological Accuracy of xCT Workflows"***
16:40	Igor Zakharov, ***"Theta-theta position 3D fixing: measurement uncertainty evaluation"***	Sai Hin Matthew Chan, Russell Miller, ***"Applying Data Quality Metrics to AIS Communications"***
17:05	Igor Zakharov, ***"Rho-rho position 2D fixing: measurement uncertainty evaluation"***	Hugo Gasca-Aragón, ***"THE DIGITAL METROLOGICAL EXPERT (DME)"***

17:30 Day 1 Technical sessions Close

18:00	**Reception, University of Algarve**
20:00	End of Day 1

Day 2 — Wednesday, 28 October 2026

	Plenary Session — Auditorium **1.5** — Chair: **Severine Demeyer**
09:30	**Welcome to Day 2**
09:40	Alistair Forbes, ***"The Central Limit Theorem and sums of t-variates"***
10:20	Michael Chrubasik, ***"Towards an End-to-End Semantic Model For Calibration Services"***

11:00–11:30 Coffee Break, Posters

	Session 3: Digitalisation — Auditorium **1.5** — Chair: **QingPing Yang**	**Session 4: Measurement Science** — Auditorium **1.4** — Chair: **Sanowar Khan**
11:30	Priyanka Bharti, Marina Romanchikova, Beth Tiplady, ***"Multi-Purpose Semantic Modelling in Radiofrequency Measurements"***	Francesco Crenna, Giovanni Battista Rossi, ***"Toward an operative definition of spectral resolution"***
11:55	Moulham Alsuleman, Michael Chrubasik, Joao Gregorio, Harvey Whelan, ***"Agent-Based Knowledge Graph Generation for Scientific Experiments"***	João Alves e Sousa, Elsa Batista, Rui Maldonado, ***"Uncertainty calculation for microchannel hydraulic resistance determination"***
12:20	Carlos Pires, Gonçalo Carvalhal, João Abrantes, João A. Sousa, Isabel Lopes, Amauri Camargo, ***"Development and implementation of digital calibration certificates at IPQ: from NMI to end-user"***	Soheil Ahmadi, Kenneth Grattan, Sanowar Khan, ***"Geometry-Resolution Effects in 3-D Finite-Element Simulations of Twisted-Conductor Loss"***

12:45-14:00 Lunch

	Plenary Session — Auditorium **1.5** — Chair: **João Sousa**
14:00	Francesca Pennecchi, ***"Extension of the JCGM 106 approach: assessing the number of truly conforming items in a lot"***
14:40	Elisabeth Costa Monteiro, ***"Metrological Traceability Infrastructure for Diabetes-Related Laboratory Measurements: A Pilot Study of Global Availability and Access"***

15:20–15:40 Tea & Coffee, Posters

15:40–16:20 Poster walk *

	Session 5: Measurement — Auditorium **1.5** — Chair: **Francesco Crenna**	**Session 6: Life Sciences** — Auditorium **1.4** — Chair: **Elisabeth Costa Monteiro**
16:20	Marija Cundeva-Blajer, ***"Expert Conformity Assessment Algorithm in Calibration by MonteCalc Uncertainty Toolkit"***	Padmini Krishnadas, Urs Hackstein, Alen Bosnjakovic, Philip Aston, ***"Light vs. Dark Skin Tone Classification Using Photoplethysmography Signals"***
16:45	Zeliha Hanım Karapınar, ***"Self-Validating Bayesian Recalibration Interval Optimization of Mass Standards"***	Thábata Rayane Gusmão, ***"Contextual Exploration of Hair Mineral Data: Integrating Group Comparisons and Reference Intervals"***
17:10		Bartlomiej Cieszynski, ***"Multi-Dimensional Data Quality Framework for Machine Learning in Arrhythmia Classification"***

18:00 Day 2 Technical sessions Close

19:00	**Drinks reception - Hotel Eva Senses**
19:45	**Conference dinner - Hotel Eva Senses**
22:00	End of Day 2

Day 3 — Thursday, 29 October 2026

	Plenary Session — Auditorium **1.5** Chair: **Hugo Gasca-Aragon**
09:30	Welcome to Day 3
09:40	Tameem Adel, ***"Metrological Validation in Medical AI: Using Counterfactuals to Isolate Sensor-Induced Shortcut Learning"***
10:20	Hannah Strassburg, ***"Data Quality and Uncertainty in Climate Modelling: A Case Study Using the Input4MIPS Emissions Dataset"***

11:00–11:30 Coffee Break, Posters

	Session 7: Digitalisation Auditorium **1.5** Chair: **Liam Wright**	**Session 8: Measurement** Auditorium **1.4** Chair: **Marija Cundeva-Blajer**
11:30	Joao Gregorio, ***"Image Data Quality: An Empirical Approach to Metadata Reliability in Multi-Site Imaging"***	QingPing Yang, ***"Principled Analysis of Error and Uncertainty for General Staged Measurement Chains"***
11:55	Johannes Kombe, ***"Virtual Knowledge Graph for Air Quality Measurement Datasets"***	Ahmad Barari, Wahab Janjuaa, ***"Loosely-Coupled Camera-Anchored Residual Learning for Three-Dimensional Handheld Inertial Navigation"***
12:20	Callum Whyte, ***"Validation of LiDAR sensors in virtual testing environments"***	

12:45–14:00 Lunch, poster take down

	Plenary Session — Auditorium **1.5** Chair: **Francesca Pennecchi**
14:00	Vivek Desai; Andrew Thompson, ***"Extending Confidence Calibration to Generalised Measures of Variation"***
14:40	Severine Demeyer, ***"Interpreting interlaboratory comparison results: current challenges and future perspectives"***
15:20	**Uncertainty in Measurement**, Kharkiv, December 2026 **IMEKO WC 2027**, Rimini, Italy
15:40	Close
	Coffee & Tea
16:00	End of Conference

*

Poster Walk — Poster Presentations *Presenters attending Wednesday 28 October at 15:40*
Edita Raudienė, ***"Trusted Measurement Data through Optimized Uncertainty Evaluation for Science-Driven Decision-Making"***
Kweonho Kang, Chang Hwa Lee, Jae Soo Ryu, ***"Standard Reference Data for Thermal Diffusivity of HANA-6"***
Asta Meškuotienė, "***Technical challenges and interpretations: the interaction of legal metrology and accreditation requirements in laboratory practice"***
Saowaluck Ukrisdawithid, ***"A Metrological Framework for the Comprehensive Annual Calibration of pH Meters"***
Livio D'Alvia, venla korkko, ***"Experimental Microwave Array Sensing for Brain Tumor Detection: Sensitivity and Repeatability Evaluation in a Head Phantom Model"***
Salah BOUHOUCHE, ***"Extended robust algorithm (A) ISO 13528 to dynamic measurement for interlaboratory comparison"***
Luís Alves, Dora Carinhas, Paulo Infante, José Rocha, ***"A Statistical Workflow for the Metrological Evaluation of a Thermoregulated Bath Used in Hydro-Oceanographic Sensor Calibration"***

List of abstracts accepted as ***oral presentations***

T1 - Measurement Theory and Foundations of Measurement Science

		page
#6	Giovanni Battista Rossi, ***"On the nature of modern science including measurement***"	*12*
#11	Francesco Crenna, Giovanni Battista Rossi, ***"Toward an operative definition of spectral resolution"***	*16*
#12	João Alves e Sousa, Elsa Batista, Rui Maldonado, *"Uncertainty calculation for microchannel hydraulic resistance determination"*	*20*
#15	QingPing Yang, ***"Principled Analysis of Error and Uncertainty for General Staged Measurement Chains"***	*22*
#27	Igor Zakharov, ***"Theta-theta position 3D fixing: measurement uncertainty evaluation"***	*24*
#28	Igor Zakharov, ***"Rho-rho position 2D fixing: measurement uncertainty evaluation"***	*26*

T2 - Testing, inspection and certification

		page
#37	Marija Cundeva-Blajer, ***"Expert Conformity Assessment Algorithm in Calibration by MonteCalc Uncertainty Toolkit"***	*29*
#46	Fernando Oliveira, ***"Windowed Cross-Correlation for Assessing Relative Timing Drift in DAQ Systems"***	*31*

T3 - Measurements in Life Sciences

		page
#17	Padmini Krishnadas, Urs Hackstein, Alen Bosnjakovic, Philip Aston, ***"Light vs. Dark Skin Tone Classification Using Photoplethysmography Signals"***	*34*
#29	Elisabeth Costa Monteiro, ***"Metrological Traceability Infrastructure for Diabetes-Related Laboratory Measurements: A Pilot Study of Global Availability and Access"***	*36*
#45	Thábata Rayane Gusmão, ***"Contextual Exploration of Hair Mineral Data:Integrating Group Comparisons and Reference Intervals"***	*38*

List of abstracts accepted as ***oral presentations***

T4 - Mathematical and statistical modelling

page

#7 Ignacio Lira, ***"Simple GUM-Compliant Uncertainty Evaluation for Errors-in-Variables Regression"*** *41*

#8 Katsuhiro Shirono, ***"Can predictive uncertainty in machine learning be measurement uncertainty?"*** *43*

#10 Severine Demeyer, "***Interpreting interlaboratory comparison results: current challenges and future perspectives"*** *45*

#16 Alistair Forbes, ***"Distributions associated with nonlinear least-squares problems"*** *47*

#19 Alistair Forbes, ***"The Central Limit Theorem and sums of t-variates"*** *49*

#22 Ahmad Barari, Wahab Janjuaa, ***"Loosely-Coupled Camera-Anchored Residual Learning for Three-Dimensional Handheld Inertial Navigation"*** *51*

#25 Franco Pavese, ***"Controversial new definition of "Measurement Uncertainty""*** *52*

#30 Hugo Gasca-Aragón, ***"THE DIGITAL METROLOGICAL EXPERT (DME)"*** *53*

#33 Zeliha Hanım Karapınar, ***"Self-Validating Bayesian Recalibration Interval Optimization of Mass Standards"*** *55*

#34 Tameem Adel, ***"Metrological Validation in Medical AI: Using Counterfactuals to Isolate Sensor-Induced Shortcut Learning"*** *57*

#36 Soheil Ahmadi, Kenneth Grattan, Sanowar Khan, ***"Geometry-Resolution Effects in 3-D Finite-Element Simulations of Twisted-Conductor Loss"*** *60*

#39 Vivek Desai; Andrew Thompson, ***"Extending Confidence Calibration to Generalised Measures of Variation"*** *62*

#40 Francesca Pennecchi, ***"Extension of the JCGM 106 approach: assessing the number of truly conforming items in a lot"*** *64*

#41 Miles McCrory, Weronika Pryciuk, Liam Wright, ***"Quantitative Image Analysis: An Approach to Estimating Uncertainties Associated with Image Segmentation"*** *67*

#42 Liam Wright, ***"Bridging the Gap Between Chemometrics and Metrology"*** *69*

#44 Eva Fišerová, Daniel Gogola, Gejza Wimmer, Viktor Witkovský, ***"Ellipsoid Fitting for Geometric Calibration: Uncertainty Quantification and Sphericity Testing"*** *71*

List of abstracts accepted as ***oral presentations***

T5 – Digitalization

page

#2	Joao Gregorio, ***"Image Data Quality: An Empirical Approach to Metadata Reliability in Multi-Site Imaging"***	*74*
#3	Hannah Strassburg, ***"Data Quality and Uncertainty in Climate Modelling: A Case Study Using the Input4MIPS Emissions Dataset"***	*76*
#9	Michael Chrubasik, ***"Towards an End-to-End Semantic Model For Calibration Services"***	*78*
#13	Marina Romanchikova, ***"Defining minimum metadata for nanometrology"***	*80*
#14	Bartlomiej Cieszynski, ***"Multi-Dimensional Data Quality Framework for Machine Learning in Arrhythmia Classification"***	*82*
#20	Priyanka Bharti, Marina Romanchikova, Beth Tiplady, ***"Multi-Purpose Semantic Modelling in Radiofrequency Measurements"***	*84*
#23	Carlos Pires, Gonçalo Carvalhal, João Abrantes, João A. Sousa, Isabel Lopes, Amauri Camargo, ***"Development and implementation of digital calibration certificates at IPQ: from NMI to end-user"***	*86*
#24	Johannes Kombe, ***"Virtual Knowledge Graph for Air Quality Measurement Datasets"***	*88*
#31	Callum Whyte, ***"Validation of LiDAR sensors in virtual testing environments"***	*90*
#35	Sai Hin Matthew Chan, Russell Miller, ***"Applying Data Quality Metrics to AIS Communications"***	*92*
#38	Teodor Tóth, ***"Impact of Acquisition Parameters on the Metrological Accuracy of xCT Workflows"***	*94*
#43	Moulham Alsuleman, Michael Chrubasik, Joao Gregorio, Harvey Whelan, ***"Agent-Based Knowledge Graph Generation for Scientific Experiments"***	*96*

List of abstracts accepted as ***poster presentations***

T1 - Measurement Theory and Foundations of Measurement Science

page

#4 Edita Raudienė, ***"Trusted Measurement Data through Optimized Uncertainty Evaluation for Science-Driven Decision-Making"*** *99*

#26 Kweonho Kang, Chang Hwa Lee, Jae Soo Ryu, ***"Standard Reference Data for Thermal Diffusivity of HANA-6"*** *101*

T2 - Testing, inspection and certification

#5 Asta Meškuotienė, ***"Technical challenges and interpretations: the interaction of legal metrology and accreditation requirements in laboratory practice"*** *103*

#47 Saowaluck Ukrisdawithid, ***"A Metrological Framework for the Comprehensive Annual Calibration of pH Meters"*** *106*

T3 - Measurements in Life Sciences

#21 Livio D'Alvia, venla korkko, ***"Experimental Microwave Array Sensing for Brain Tumor Detection: Sensitivity and Repeatability Evaluation in a Head Phantom Model"*** *108*

T4 - Mathematical and statistical modelling

#1 Salah BOUHOUCHE, ***"Extended robust algorithm (A) ISO 13528 to dynamic measurement for interlaboratory comparison"*** *110*

#32 Luís Alves, Dora Carinhas, Paulo Infante, José Rocha, ***"A Statistical Workflow for the Metrological Evaluation of a Thermoregulated Bath Used in Hydro-Oceanographic Sensor Calibration"*** *117*

Abstracts accepted as ***oral presentations***

T1 - Measurement Theory and Foundations of Measurement Science

		page
#6	Giovanni Battista Rossi, ***"On the nature of modern science including measurement***"	*12*
#11	Francesco Crenna, Giovanni Battista Rossi, ***"Toward an operative definition of spectral resolution"***	*16*
#12	João Alves e Sousa, Elsa Batista, Rui Maldonado, *"Uncertainty calculation for microchannel hydraulic resistance determination"*	*20*
#15	QingPing Yang, ***"Principled Analysis of Error and Uncertainty for General Staged Measurement Chains"***	*22*
#27	Igor Zakharov, ***"Theta-theta position 3D fixing: measurement uncertainty evaluation"***	*24*
#28	Igor Zakharov, ***"Rho-rho position 2D fixing: measurement uncertainty evaluation"***	*26*

2026 IMEKO Joint Conference
TC7 - TC11 – TC13 - TC21 in conjunction with TC6 and Mathmet
Faro, Portugal, 27-29 October, 2026

On the nature of modern science including measurement

Giovanni Battista Rossi

Università degli Studi di Genova – DIME, via all'Opera Pia 15a, 16145 Genova, Italy

g.b.rossi@unige,it

***Abstract* – It is discussed the scientificity of measurement. The question is not trivial, and the approach is not unanimous. Yet this aspect is, for the author, vital for the future of measurement, especially for education and research. Some critical aspects are highlighted and the need of overcoming them is stressed.**

I. INTRODUCTION: A CONCISE HISTORICAL PERSPECTIVE

Measurement has always been present and decisive in human activities since the dawn of mankind [1]. Establishing fair borders of cultivated territories, enabling fair commerce of goods, studying the asset of the sky, for make forecasting, are just a few examples. Interestingly enough, Holy Scriptures considered fair measurements as a irreducible requirement for an ethical behavior. Yet, at that time, distinguishing between the practical (today we say: technological) and the "scientific" aspect of such activities is impossible so intertwined they are. Yet definitely scientific text started to appear, such as "The Elements", by Euclid, or the "Physics", by Aristoteles. But the real change, at least in Western culture, occurred between the fifteenth to the seventeenth century, with what was properly called "the scientific revolution", which constituted the birth of what we can call the "modern science" (MS). Here the approach to the study of Nature changed radically, and measurement played a key role in that.

Yet, interestingly enough, measurement theories appeared only (much) later, starting in the nineteenth century and continuing also at the present day. There is a reason for such a delay, to be discussed at a later stage, Therefore, in the following, after a quick revision of the principles of MS, necessary for highlighting some critical aspects that continue to be relevant, it will be discussed in which way measurement, in the present day, poses itself in regard to MS, and some, non-negligible, in the author's vision, criticalities are present.

II. OVERVIEW OF THE MAIN CHARACTERS OF MODERN SCIENCE

A noteworthy characterization of modern science due to Antiseri, is here summarized [2]. According to him, MS can be considered as a kind of knowledge which is:

- public,
- progressive, and
- controllable.

Very briefly, public mean that this kind of knowledge can be accessed by anyone (through, obviously, a proper gradual preparation) and thus it is not reserved to any elite having exclusive characters (such as, typically, some kind of initiation).

But most important for the present discussion are the following two aspects.

Progressivity is extremely important in two key aspects.

Firstly, any scientific achievement cannot be considered as definitive, since it can always be overcome by successive findings.

Secondly, this provides an important link, with Kuhn's theory of scientific revolutions (which, in fact, does not concerns MS alone, but science in general) [3].

According to him, science does not progress in an "incremental" way, that is by acquiring additional information, that that can be included in an already accepted framework, by sometimes it implies real "scientific revolutions", which means dramatic change of perspective. Such general perspectives are called leading "paradigms", by the mentioned author. A typical, although somewhat extreme, example is the change between the geocentric to the heliocentric vision of the cosmos.

An important associated concept is that of "normal science". It represents the vision shared by the majority of the scientists, when such a revolution appears at the horizon: normal science tries to oppose in any way to the change implied by the new vision, unless it is forced to accept it, in consequence of what is ongoing.

Lastly, controllability appears, to some extent, the most important feature, and it involves what follows. When a class of phenomena is considered, MS describes/explains them by a model/theory. But such a proposal must be checked in the following way. One, or more, experiment(s) is/are designed and performed, whose results can be predicted by the theory. If the results obtained are not "compatible" with the predictions, the theory is rejected. Technically this, according to Popper, may be stated by saying that the theory has been "falsified" [4]. If instead, the results are conformal to what expected, it can be stated that the theory has been "enforced". But not definitely

"verified" since new experiments could, in principle, falsify it. It goes without saying that, when a theory satisfies various tests, it is usually assumed as highly reliable. This aspect is particularly important. Sometimes, the way of controlling scientific models/theories is called "the scientific method", and the possibility to apply it to a proposed theory is one of the main criteria for claiming the scientificity of such theory.

Now, after this, necessary short, review of the basic principles of modern science, we can go back to measurement, but, prior of then, there is a key concept that needs to be considered attentively, since it is crucial for the present discussion.

III. SCIENTIFIC MODELS

A. Scientific models and the problem of realism

In which way MS expresses its contents? Definitely, through models/theories. A model can be described as an abstract system, associated to a real system. In this perspective, a scientific theory can be seen just as a very general model, capable of describing many real situations. Scientific models immediately rise the problem of their relationship with (empirical) reality, which may be called the problem of "realism".

There are a few possibilities in respect of that; here only three positions are mentioned. But prior of that, some historical facts need recalling, in particular the dispute between Galileo Galilei and Cardinal Bellarmino [2].

In fact, Cardinal Bellarmino did not totally deny the value of MS and he admitted that MS might be good in making predictions, but it was unable to explain the true «nature» of things/phenomena, which is instead revealed by the Holy Scriptures. This paves the way to discussing the mentioned problem of realism, i.e., the problem of the relationship between scientific models/theories and reality.

This is an ample problem, yet only three main positions are here recalled.

• Radical or "essential" realism. This is Galileo's position: «The book of Nature is written in the language of mathematics and geometry», i.e., there is a strict correspondence between models/theories and reality. It may be noted that this is related to the controversial notion of "true value", in measurement.

• Instrumentalism. In its extreme version, this was the position expressed by Cardinal Bellarmino, in his dispute with Galileo. Basically, he held that scientific models are mathematical devices that can make provisions, with some practical utility, but they cannot reveal the intimate essence of reality, which is instead revealed by the Holy Scriptures. Models have practical utility but no knowledge value. (Interestingly enough, this position is to some extent also supported by modern pragmatism).

• Moderate realism. This was best expressed by Francesco Barone, who held that if scientific models allow us to make reliable provisions, this means that they somehow «byte into reality» [5]. This is also the position of the present authors.

A remarkable paper on the problem of realism, in connection with measurement, was issued by Morawski [6].

B. Controllability of probabilistic model

Models can be either deterministic or non-deterministic, the most remarkable example of the latter being probabilistic models. A model can be considered deterministic if, in any situation it can be applied – we will say in any instance of the model – all the statement concerning the quantity involved need being either "true" or "false". The attribution of a measurement value to a quantity is an example of such statements.

In the case of a probabilistic model instead, the category of truth is substituted with a degree of probability (for, at least, one relevant statement associated to the model. Probabilistic models are very important in MS, including measurement. Therefore, the problem of their controllability is very important and non-trivial, since a probabilistic model does not allow us to exactly predict the result of an experiment based on it. So, how can that model by falsified?

According to Costantini [7], this can be done through a significance test. In such a test, the set of possible results is parsed in two subsets, one having a high probability, the other having the complementary, very low, probability. The model is falsified if the result fall in the low probability set. It appears how the logic of this kind of testing is very close to the general approach of the scientific method, although there is a component of conventionality associated to what is considered a low probability, which must be agreed among the scientists involved. This kind of testing may be called a hypothetic-deductive inference, as the general approach to testing scientific models.

IV. CONSIDERATIONS ON THE SCIENTIFICITY OF MEASUREMENT

Let us now consider the main question we are interested in, the scientificity of measurement. The question is not trivial for, at least, two main issues.

As previously discussed, measurement plays a key role in assessing the scientificity of proposed theory,because it is essential for performing hypothetic-deductive inferences, since the adherence of previsions of the model with observed results of an experiment are typically made through measurements. So, measurement is essential for founding MS. But if measurement is also a science, does measurement found itself [8]?

The second issue concerns the relevant importance of measurement in practical activities, which emerged even from ancient times. But this aspect may be, at the same time, a strength and a weakness. A strength because the allocation of relevant resources to it – for example the international system of metrology, including the BIPM, the NMIs and the networks of calibration and testing

laboratories – appears as fully justified. But at the same time the very importance of measurement poses a need for standardisation. But standardisation may typically produce an inertia to change and evolution.

Therefore, these and other issued will be briefly discussed in the following.

A. On the foundation of measurement.

Sometimes it is said that measurement establishes a bridge between reality and models [9]. This statement can be accepted, if understood in a highly informal way. But if analysed in greater depth, it shows criticalities. In fact, if measurement is intended as a science, it should be modelled as well and it cannot include reality as such, since models and reality lye at different levels. Rather reality should be substituted by a model of the measurand, which, combined with a model of the measurement system, forms a model of the measurement process. The relationship with this model with reality is just a case of the problem of realism, already discussed.

B. The model of measurement.

If measurement is considered a science, in the author's opinion, it must refer to a general model/theory. The alternative is to consider it rather a collection of "good practices" taken from other disciplines. Obviously, in each historical period there can be different models associated to a discipline, but a scientist and also a practitioner operating in that area must always have at least one to refer to.

As mentioned, a model of the measurement process can, in full generality, be considered as the combination of a model of the measurand and a model of the measurement "system". The former is strongly dependent from disciplines; what can be said of it is that it must be consistent with the condition of measurability on one of the universally recognised measurement scales, i.e., ordinal, interval or ratio, with some attention to nominal scales also. On the other hand, the model of the measurement "system", should be as unified as possible. In fact, recently, a few such models have been proposed.

Although they are not totally superposable, they all agree un some key points, such as:

• The model basically consider an interaction between the measurand and the measuring system, which, on the basis of a cause-effect relation, produces observable outputs, on the basis of which a measurement value can be obtain, thanks to a previous calibration of the measurement system.

• Such a structure is fully general and thus holds true for both static and dynamic measurements and for either direct or indirect ones.

It will be suggested in the following that a consciousness of the model of the measurement process is, at present, far from being universally acknowledged.

C. Controlling measurement models

The controllability of measurement models seems to be, at present, quite well accessed, in the Physics and Engineering areas. In fact:

• A single model, underlying a specific measurement, can be controlled by a "calibration verification" of the measurement system used.

• The above implies using input standards, whose value can be ascertained by a more accurate system. Note, in this regard, the possible use of "check standards" that allow verification to take place in the actual measuring conditions.

• At the top of the traceability chain, primary scales exist, which cannot be checked in respect a standard having a higher accuracy. Yet, intercomparisons (key comparisons) allow to define a "virtual standard", which consists of a travelling object, to which a value can be associated, as some kind of mean of the values of the primary laboratories involved in the exercise. This also has the structure of a hypothetic- deductive inference.

Unfortunately, the system, at present, does not seem to be equally developed for measurement in social sciences. In that case, the agreement on the model of the measurand seems to be the main obstacle. In consideration of this, we can accept that different models, can coexist in the same historical period, but it is essential that any measurement be defined in respect of one of them, clearly stated.

D. Standardization, progress in measurement and normal science

As mentioned above, the great and amply recognised character of measurement as a technology, has advantages and disadvantages. The former ones have already been mentioned, let us then deals shortly with disadvantages also.

A technology, today, implies standardization, and this, in the case of measurement, has been implemented, at the highest level, through the Joint Committee for Guides in Metrology, directly depending upon the BIPM [10]. Although of clear practical value, such a powerful organism, may have also influenced research, thus playing the role, in Kuhn's vision, of representative of normal science. It is here suggested that there is a risk of this structure to be a possible obstacle to the process of measurement, intended as a science.

V. CONCLUSIONS: ON THE FUTURE OF MEASUREMENT

The importance of measurement is so evident, that it is reasonable to assume that in the future it will continue to play a key role in society.

Not equally straightforward is its future as an autonomous science. Some major steps should be taken, in the author's opinion, in this regard.

Firstly, the acknowledgement of one (or more) reference model(s): in defect of this, measurement will be regarded as a technology only, supported by "good practices", borrowed by other disciplines. This would have a strong negative effect, both in research and in education.

Secondly, for measurement to become really transdisciplinary, the issue of controllability (of the associated models) should be more carefully considered, especially in social sciences.

Lastly, the role of standardization bodies should in no way influence and condition research in measurement.

REFERENCES

[1] M.E Himbert, "A brief history of measurement", Eur. Phys. J. Special Topics 172, 25–35 (2009) Springer-Verlag 2009 DOI: 10.1140/epjst/e2009-01039-1

[2] G. Reale, D. Antiseri, "Storia della Filosofia, Vol. 4, Umanesimo, Rinascimento e Rivoluzione Scientifica", Bompiani, 2008.

[3] T. Kuhn, "The structure of scientific revolutions", Chicago University Press, 1962 (Italian translation by A. Curgo, 5th ed., Einaudi, 1999).

[4] K. Popper, "Logica della scoperta scientifica", 1934 (Italian edition by Einaudi, Torino, 1970)

[5] G. Boniolo, P. Vidali, "Introduzione alla filosofia della scienza, Bruno Mondadori", 2003.

[6] R. Morawski, "Realists' versus instrumentalists' understanding of measurement", XVII IMEKO World Congress, 2003.

[7] D. Costantini, "Verso una rappresentazione probabilistica del mondo", EMMEBI, 2011.

[8] L. Mari, "The problem of the foundation of measurement", Measurement 38 (2005) 259–266

[9] L. Mari, D. Petri, "Measurement Science: Constructing Bridges between Reality and Knowledge", December 2014, IEEE Instrumentation & Measurement Magazine, doi: 10.1109/MIM.2014.6968922.

[10] https://www.bipm.org/en/committees/jc/jcgm/.

2026 IMEKO Joint Conference
TC7 - TC11 – TC13 - TC21 in conjunction with TC6 and Mathmet
Faro, Portugal, 27-29 October, 2026

Toward an operative definition of spectral resolution

Francesco Crenna, Giovanni Battista Rossi

Università degli Studi di Genova – DIME, via all'Opera Pia 15a, 16145 Genova, Italy, francesco.crenna@unige.it, g.b.rossi@unige.it

***Abstract* – The spectral resolution information is important when setting up the acquisition parameters for a spectral measurement. Different definitions for spectral resolution are available usually based on theoretical concepts. Here we propose an operative approach focused on the ability of the algorithms to distinguish automatically two spectral contributions with very near frequencies. A standard boxcar window will be considered with and without zero padding. Further investigations will include special observation windows to limit the spectral leakage and the presence of noisy signals.**

I. INTRODUCTION

The aim of this study is to give operative indications about the resolution capabilities of a spectral measurement algorithm. Spectral resolution is usually defined considering the observation window used to acquire the dynamic signal [1]. The duration and the form of the observation window affects the spectral resolution. From the point of view of information, considering a periodic signal, the more periods are considered the more information is gathered to measure its periodicity, or frequency, thus improving the spectral resolution.

From a mathematical point of view when a periodic signal is acquired for a limited time, it can be considered as multiplied by a rectangular window of duration equal to the acquisition time:

$$S(t) = \mathrm{P(t)} \cdot \mathrm{W(t)} \tag{1}$$

where $\mathrm{P(t)}$ is the periodical (in the simplest case a pure sinusoidal) signal, ideally extending from $\pm\infty$; $\mathrm{W(t)}$ is the observation window, extending from $\pm\frac{T}{2}$, and $S(t)$ is the acquired or observed signal. Consequently, the resulting spectrum will present the convolution of the frequency components included in the periodical signal, with the spectrum $\mathrm{W(\omega)}$ of the observation window.

$$S(\omega) = \frac{1}{2\pi}\mathrm{P(\omega)} * \mathrm{W(\omega)} =$$

$$= \frac{1}{2\pi}\int_{-\infty}^{+\infty} \mathrm{P}(x)W(\omega - x)\, dx. \tag{2}$$

In the case of a periodic signal constituted by a sum of two or more sinusoids the spectrum is the sum:

$$S_{sin}(\omega) = \sum a_i \delta(\omega_i) * \mathrm{W(\omega)}. \tag{3}$$

In case of a standard boxcar window:

$$W(t) = 1\ for -\frac{T}{2} < t < \frac{T}{2} \text{ and } W(t) = 0 \text{ elsewhere} \tag{4}$$

the spectrum $\mathrm{W(\omega)}$ is

$$\mathrm{W(\omega)} = \frac{\sin(\omega)}{\omega} e^{j\omega\frac{T}{2}}, \tag{5}$$

while special observation windows have different [2,3] and specific shapes, designed for the intended purposes, such as reducing spectral leakage, or measuring the signal amplitude in the central part of its power. The spectral resolution is generally defined considering a shape parameter of the $\mathrm{W(\omega)}$ spectrum. One of the most common is the half power bandwidth, or -3dB bandwidth [3]. Two signals of equal amplitude cannot be distinguished if their frequency difference is less than such a bandwidth. Considering such definition as theoretical resolution we can roughly define it as:

$$\Delta f_{3dB} = \frac{\alpha_w}{T} \tag{6}$$

where the resolution computed as -3dB bandwidth of the observation window, is obtained considering the frequency limit corresponding to the window duration multiplied by a reduction factor typical of the shape of the window in use. In case of a rectangular window such value is

$$\alpha_w = 0.89. \tag{7}$$

Although windows performances and the associated spectral resolution have been amply investigated in the literature, from different standpoints, the approach here proposed is focused on an operative definition of resolution, considering, to this purpose, the capability of

the spectral analysis algorithm to detect as separate two peaks with nearby frequencies.

To avoid possible influences of the operator who can, with her or his perception, decide differently about the presence of two distinct peaks in the spectrum, an automatic algorithm instead is here considered, for both spectral analysis and peak detection.

In the first part of the paper the automatic peak detection procedure is described and its parameters are defined. In the second part it is applied to the detection of two peaks corresponding to two sinusoidal signals, observed through a rectangular window.

As further developments, the overall investigation could be extended to different windows, or considering the presence of noise, or other signal compositions.

II. PEAK-FINDING ALGORITHM

As a first step, the automatic peak finding algorithm has been studied, to understand its dependence upon the different available parameters.

Peak find automatic procedures are often based on the change of sign of the first derivative. This is the case for the MatLab® function *findpeaks* [4]. Similar approach is offered by the Python package PeakUtils [5]. Such procedures find out every peak in the signal including non-significative fluctuations due to noise or spectral leakage. A set of criteria helps in the discrimination of interesting peaks among the multitude of raw ones. In Matlab® the instruction *findpeaks* is associated with different peak selection options. In the following the minimum peak prominence and minimum peak height are considered.

The latter is an absolute parameter useful to discard small amplitude peaks, for example due to noise. In the following it has limited interest, but it will be important in further investigations including peaks of different amplitudes.

The former instead is a relative measure working on the height of the peak with respect to the surrounding peaks or signal behavior. The prominence is defined as the height respect to a reference level obtained considering the two sides on the left and right of the potential peak. Each side defines an interval considering or the signal limit (for example zero / maximum time or, in case of a spectrum, zero /maximum frequency) or the horizontal crossing from the potential peak amplitude to the eventual left or right crossing of the signal. In these two intervals the minimum signal amplitude is identified. These two values define the reference level over which the peak prominence is evaluated.

The peak prominence value is critical, since it determines the height of the detected peak over the reference level, and, in our case, the difference in height with respect to the neighbor signal peak. To understand the prominence effect on peak detection and setup the most suitable value for studying the spectral resolution of the measurement system, a preliminary test with different prominence values was conducted, as presented in the following figure.

Here the spectrum of two sinusoidal signals is considered: the lowest frequency signal is fixed while the higher frequency is moving in short steps until the *findpeaks* algorithm is able to detect the two peaks, using the selected prominence. The spectra for which two peaks are identified is presented in blue, together with the couple of spectra with frequency differences slightly lower and higher respect to the detected one.

It appears that the lowest prominence, 0.02, is not sufficient for an evident peak separation; 0.04 is at the limit, 0.05 seems a robust quantity to avoid misdetection and to have a situation easily interpretable by a human expert also. Over 0.05 the situation does not improve significantly; thus the value 0.05 is considered for all the following investigations.

III. PEAK IDENTIFICATION

To investigate the capability to identify two distinct peaks corresponding to two spectral components, two sinusoidal signals are simulated on a fixed time window of 10 s. One signal has a fixed, or base frequency, the second one has a slightly higher frequency. Frequencies are in the range between 1 Hz and 2 Hz and the sampling frequency

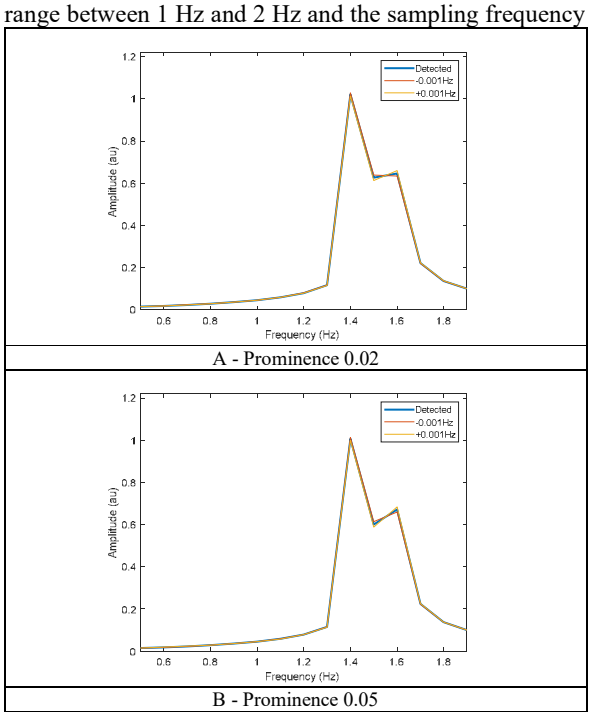


Fig. 1. Amplitude spectra for double sinusoidal signals identified by the 'findpeaks' function with two prominence values: Case A does not present evident peak distinction, while in Case B peaks are effectively discriminated.

is 500 Hz, far higher than the Nyquist one. The period

corresponding to the fixed, base, frequency can be set as a multiple of the observation window or not, while the higher will increase in short steps from the base frequency, as it will be described hereafter.

Once simulated the spectral analysis is conducted considering a standard boxcar window of the same width of the simulation window, of course a special Hanning window can be introduced, to take into account possible leakage effects.

The spectral analysis is carried out on the simulated signal as it is, i.e., 10 s signal sampled at 500 Hz equal to 5000 points, or on a zero padded signal with a number of points equal to four times the original number of points, i.e., 20000 in the actual configuration. The analyzer is normalized in amplitude, showing distinct sinusoids of the same unitary amplitude.

The peak identification operation is conducted on the spectra, using the *findpeaks* function with the standard parameter identified above. If the two frequencies are too near, only one peak is detected. Therefore, the second frequency has been slightly incremented by Δf, then the procedure is repeated until two peaks are identified. Since the second frequency is increased in each step, in this situation the spectral distance between the two peaks is the minimum to enable the automatic procedure to identify two distinct peaks. A very small frequency increment is selected $\Delta f = 0.01$ Hz, in a range for between zero (second frequency equal to the base one), up to maximum +0.6 Hz from the first, base, frequency.

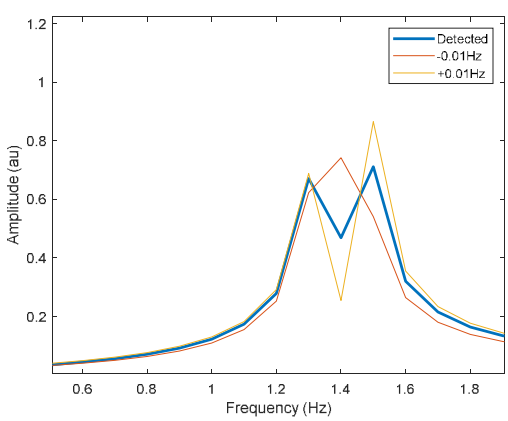


Fig. 2. Amplitude spectra and peak identification for a base frequency of 1.3 Hz. Color description n the text

IV. RESULTS

The rectangular observation window of 10 s, limits the spectral resolution, when operating without a zero padding, to

$$\Delta f_{3dB} = \frac{0.89}{10} = 0.089 \text{ Hz} \tag{8}$$

and the minimum distance between two peaks is expected to be two times the resolution, or, in this case, 0.18 Hz. Possibly a longer observation window will improve this situation and can be investigated in future, with the same parameters.

The following figure presents three colored spectra: the one with the minimum distance Δf to detect two peaks (blue), and the two nearby spectra for which only one peak is detected (red), or two peaks are detected (orange) but having a double Δf. In the presented case a 1.3 Hz base frequency, corresponding to 13 periods in the observation window, is considered.

When changing the base frequency increasing it by $\Delta f = 0.01$ Hz, a non-integer number of periods are included in the observation window, and the minimum Δf for peak separation varies. When an integer number of periods is considered, a minimum distance of 0.208 Hz is required to separate the two peaks and it reduces to 0.08 Hz when non integer periods are considered. Nevertheless the detected frequency is the same since the algorithm resolution is wider than the obtained Δf.

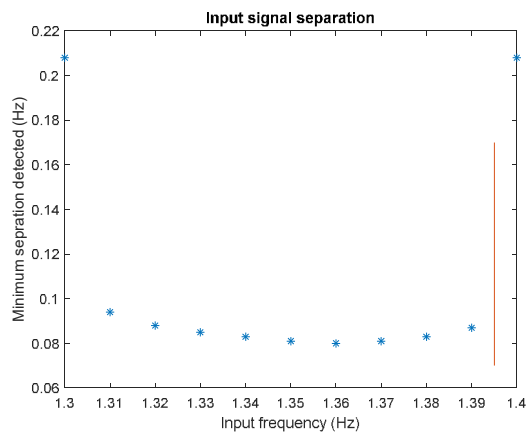


Fig. 3. Minimum Δf separation for the two sinusoids required fto detected two peaks in the spectra, as a function of the base frequency. 1.3 and 1.4 Hz corresponds to an integer number of periods in the observation window.

V. CONCLUSIONS

A common guidance when operating with observation window of duration T, is to consider a spectral resolution equal to Δf_{3dB} (6, 8). This limits the possibility to distinguish two nearby peaks at a distance less than such resolution.

Obtained results for a standard rectangular window, suggest that it is worth noting what follows.

- Two original peaks with spectral distance near the expected resolution of 0.089 Hz, can be distinguished, when an observation of a non-integer number of periods is considered. In this case, the algorithm will assign peaks to two frequencies with a spectral distance given by the algorithm resolution. In such cases the spectral analyzer

can distinguish peaks with distance around the reference resolution (Figure 3).

-In the more theoretical situation in which an integer number of periods is considered, the minimum distance to detect two peaks is far larger and more than two times the expected resolution (Figure 3 points at 1.3 and 1.4 Hz).

In the final paper the analysis will be extended considering the detected peaks' amplitude both from an absolute and relative points of view.

The proposed approach can be extended to special observation windows for which the leakage effect is certainly limited, as compared with the rectangular. Moreover, some noise can be added to the signals to understand how the noise limits the capability of the algorithm to detect two nearby peaks.

REFERENCES

[1] G.B. Rossi, F. Crenna, M. Khalil, Foundational issues in frequency spectrum measurement – Preliminary considerations, Measurement: Sensors, 38, 2025, 101502,https://doi.org/10.1016/j.measen.2024.101502.

[2] F.J. Harris, On the use of windows for harmonic analysis with the discrete Fourier transform, IEEE Proceedings, 66, 1, 51–83,1978.

[3] K.M.M. Prabhu, *Window functions and their applications in signal processing*. Taylor & Francis, 2014, ISBN 978-1-4665-1583-3

[4] Peak Analysis in Matlab, https://it.mathworks.com/help/signal/ug/peak-analysis.html, (verified 2026 07 20)

[5] L. H. Negri, Peakutils, https://peakutils.readthedocs.io/en/latest/#license (verified 2026 07 20)

2026 IMEKO Joint Conference
TC7 - TC11 – TC13 - TC24 in conjunction with TC6 and Mathmet
Faro, Portugal, 27-29 October, 2026

Uncertainty calculation for microchannel hydraulic resistance determination

Rui Maldonado[1], Elsa Batista[1], João Alves e Sousa[1]

[1] *IPQ, Caparica, Portugal, rui.maldonado@ipq.pt, ebatista@ipq.pt, jasousa@ipq.pt*

***Abstract* – Within the EPM MFMET II project, establishing traceable metrology for microfluidic systems requires accurate experimental tests. During these experiments, hydraulic resistance is determined as the slope of the linear relationship between the applied flow rate and the resulting pressure drop across a microchannel. However, experimental data inherently contains measurement uncertainties in both the independent variable - flow rate and the dependent variable - pressure drop. Traditional Ordinary Least Squares (OLS) regression is inadequate for this analysis, as it assumes the x-axis is error-free, ultimately leading to a biased slope and underestimated resistance uncertainty. To overcome this limitation, this work applies the Generalized Least Squares (GLS) method, strictly complying with the ISO/TS 28037 standard.**

I. INTRODUCTION

Hydrodynamic resistance serves as the fluidic counterpart to electrical resistance, acting as a key parameter to govern liquid behavior within microscale networks. In advanced microfluidic systems like Organ-on-Chip (OoC), precise management of this parameter is vital to replicate physiological shear stresses.

Within the scope of the European project MFMET II, this work aims to develop protocols and guidelines for the standardization and quality control of microfluidic devices. To achieve this, an experimental setup was built at IPQ to characterize flow resistance across different channel geometries in different types of chips materials [1].

II. THEORETICAL FRAMEWORK

In microfluidic regimes, the flow is characterized by extremely low Reynolds numbers, where viscous forces completely dominate inertial forces. Under these conditions, the flow of an incompressible, Newtonian fluid through a microchannel is laminar, steady and unidirectional according to the Poiseuille law.

The pressure drop along the channel is governed by the fluidic analogue of Ohm's Law, establishing a linear relationship between pressure drop (ΔP) and volumetric flow rate (Q) through the hydraulic resistance (R_{hyd}) [2]:

$$\Delta P = R_{hyd} \cdot Q \qquad (1)$$

III. EXPERIMENTAL SETUP

The experimental apparatus deployed at IPQ utilizes a high-precision, two-phase differential measurement protocol:

- Phase 1 (Total): The pressure drop of the entire fluidic network, including the target microfluidic chip, is recorded.
- Phase 2 (Baseline): The target chip is replaced by a channel with a very large cross-sectional area, presenting negligible resistance, while keeping all other system parameters identical.

The subtraction of Phase 2 from Phase 1 isolates the precise pressure drop across the studied microchannel, systematically removing parasitic resistances from the tubing and fittings.

The setup is composed of a flow generator, a pressure sensor and the microfluidic chip (see figure 1). During the measurement campaign two flow generators will be used:

- The pressure-driven Beetle developed by Microfluidic Innovation Center (MIC), which must be coupled with a flow sensor, properly calibrated.
- The Nexus 3000 high-precision perfusion syringe.

Two different types of chips were used; one made of glass and two made of TOPAS. Several microchannels with these chips were tested.

Figure 1 – Flow resistivity setup with a pressure-driven generator.

IV. RESULTS

For each tested microchannel, the experimental differential pressure drop, ΔP_{chip}, is plotted against the calibrated volumetric flow rate, Q. The experimental hydraulic resistance, $R_{\text{hyd, exp}}$ is extracted directly from the slope (m) of the resulting linear regression calculated with the Ordinary Least Squares (OLS) method.

$$\Delta P = m \cdot Q + b \quad (2)$$

To evaluate the reliability of the determined R_{hyd}, a rigorous uncertainty calculation method will be developed in strict compliance with the ISO/TS 28037 standard [3].

Unlike classical OLS regression, which assumes the independent variable (flow rate, Q) has no uncertainty, the ISO 28037 framework provides a standard for linear regressions with uncertainties in both coordinates. Which is relevant in this case, given that both the flow rate and the pressure drop have associated uncertainties. Thus, to estimate the measurement uncertainty associated with the equation (2) above, using the GUM uncertainty framework, it will be necessary to have access to the uncertainty associated with quantities “m” and “b” that define the linear regression. This approach is called forward evaluation where, given an estimate of X and its associated standard uncertainty, the corresponding value of Y is obtained together with its associated standard uncertainty. Another possibility would be the inverse evaluation, when we seek a value for X given an estimate of Y.

Additionally, to ensure the high reliability of these results, a conservative upper-bounding approach will be implemented to estimate the maximum realistic uncertainty limits.

Finally, a comparison between the two approaches being developed at IPQ and the open-source application developed by National Physical Laboratory (NPL) should be made.

References

[1] EURAMET, "Publishable Summary for 24NRM03 MFMET II: Establishing metrology standards in microfluidic devices II," May 2025. [Online]. Available: https://mfmet.eu/

[2] Ogheard, F., Batista, E., Büker, O., Daugbjerg, T. S., Romieu, K., & Silverio, V. (2026). *MFMET II A1.2.1 — Review of existing standardised methods for characterising pressure drop and flow resistance*. Zenodo. https://doi.org/10.5281/zenodo.18155039June 2012, pp.1613-1624.

[3] ISO/TS 28037:2011 — *Determination of uncertainty of measurement — Use of straight-line calibration relationships*. International Organization for Standardization, Geneva, Switzerland.

2026 IMEKO Joint Conference
TC7 - TC11 – TC13 - TC21 in conjunction with TC6 and Mathmet
Faro, Portugal, 27-29 October, 2026

Principled Analysis of Error and Uncertainty for General Staged Measurement Chains

Qingping Yang

Brunel University of London, Uxbridge UB8 3PH, United Kingdom; qingping.yang@brunel.ac.uk

I. INTRODUCTION

Trustworthy measurement and trustworthy analysis require both reliable acquisition of measurement information and the transparent, principled use of models and computational procedures to obtain reliable and high-quality results. Error decomposition and uncertainty evaluation are central to metrology and measurement theory and are two of the most important forms of analysis in measurement and instrumentation science. Measurement processes can generally be modelled within a broader framework of measurement science [1] and as staged measurement chains [2]. Building on the basic measurement process, the author has presented the fundamental principles of measurement processes as three laws, general measurement chains based on a general communication model and an example error analysis of measurement chains [2]. This work further develops the results with two analysis theorems for general measurement chains. Theorem 1 decomposes the measurement error by source and propagation path. Theorem 2 evaluates uncertainty from the implemented measurement chain model. A case study of CMM (coordinate measuring machine) is provided to demonstrate the use of both theorems for trustworthy error analysis and uncertainty evaluation.

II. BASIC PROCESS, LAWS AND STAGED CHAIN

The basic measurement process maps the true attribute of an entity, through a binding indication, to an observer's representation:

$$\mathrm{T.A} \xrightarrow{s} \mathrm{O.B} \xrightarrow{p} \mathrm{O.C}$$

where $T.A$ is the attribute or measurand of the entity T; $O.B$ is the sensor output as the binding attribute between the entity and the observer O; O.C is the observer or agent's concept or representation. s is the sensing function, p is the processing function.

Three laws have been formulated for measurement processes [2]:

Law 1 (Oneness). *Given the measurable attributes of an entity in its environment, an agent's representation of these attributes converges asymptotically towards their true states.*

$$\lim_{t\to\infty} O.C(t) = T.A$$

Law 2 (Dual Processes). *Every agent–entity interaction necessarily involves a binding attribute and two subprocesses which are asymptotically inverse to each other.*

$$\mathrm{T.A} \xrightarrow{s(t)} \mathrm{O.B} \xrightarrow{p(t)} \mathrm{O.C}, \qquad \lim_{t\to\infty} p(t) \circ s(t) = I.$$

The inverse relation is understood to apply where the original state can be uniquely recovered in principle.

Law 3 (Measurement–Control Duality). *Under appropriate regularity, measurement (estimation) and control are dual processes, differing primarily in the direction of information flow.*

$$\mathcal{M}: T.A \to O.C \overset{dual}{\longleftrightarrow} \mathcal{C}: O.C \to T.A$$

In [2], general communication was further represented as a control/encoding process, a channel and a measurement/decoding process:

$$\mathrm{O_1.C_1} \xrightarrow{\pi} \mathrm{O_1.B_1} \xrightarrow{d} \mathrm{T_1.A_1} \xrightarrow{H} \mathrm{T_2.A_2} \xrightarrow{s} \mathrm{O_2.B_2} \xrightarrow{p} \mathrm{O_2.C_2}$$

A general measurement chain can be represented as this type of communication structure. The observation, interpretation and complete chains are respectively:

$$x^* = x_0^* \xrightarrow{T_1} x_1^* \xrightarrow{T_2} \cdots \xrightarrow{T_m} x_m^* = y^* \tag{1}$$

$$y = \hat{x}_m \xrightarrow{P_m} \hat{x}_{m-1} \xrightarrow{P_{m-1}} \cdots \xrightarrow{P_1} \hat{x}_0 = x \tag{2}$$

$$x^* \xrightarrow{T_1,\ldots,T_m} y^* \xrightarrow{+\Delta y} y \xrightarrow{P_m,\ldots,P_1} x \tag{3}$$

All x_k^* and $\hat{x}_k$ are vectors, whose dimensions may vary. The T_k are generally nonlinear local observation functions, and the P_k are generally nonlinear, locally differentiable implemented interpretations. The notation T_k^{-1} denotes a local inverse of T_k, valid in the neighbourhood considered. The error-free indication is y^* and the realised observation model is $y = y^* + \Delta y$.

III. THEOREM 1: ERROR ANALYSIS

Theorem 1. For chains (1)–(3), suppose each T_k is locally invertible in the measurement neighbourhood considered, and each implemented P_k is locally differentiable. Define the local inverse mismatch and Jacobian by $\Delta P_k = P_k - T_k^{-1}$, $J_k = \left.\frac{\partial P_k(u)}{\partial \mathrm{u}}\right|_{u=x_k^*}$ and let $\Phi_{1:k-1} = J_1 \cdots J_{k-1}$, $\Phi_{1:0} = I$, $e := x - x^*$, then, to first order, the measurement error is:

$$e \approx \sum_{k=1}^{m} \Phi_{1:k-1}\, \Delta P_k(x_k^*) + \Phi_{1:m}\Delta y \tag{4}$$

Using the standard local affine approximation $\Delta P_k(u) \approx \Delta A_k u + \Delta t_k$, where ΔA_k and Δt_k are the local linear and offset components, equation (4) becomes:

$$e \approx \sum_{k=1}^{m} \Phi_{1:k-1}\, (\Delta A_k x_k^* + \Delta t_k) + \Phi_{1:m}\Delta y \tag{5}$$

Although x_k^* denotes the local true state, practical calculations evaluate the local Jacobians and affine terms at the corresponding nominal or estimated states.

Proof Sketch. Let $e_k = \hat{x}_k - x_k^*$ and $e_m = \Delta y$. A first-order expansion of P_k gives $e_{k-1} \approx \Delta P_k(x_k^*) + J_k e_k$. Repeated backward substitution proves (4) and substitution of the local affine approximation gives (5).

Theorem 1 provides a stage-wise error decomposition: each contribution is associated with a local source, state space and propagation path. Local inverse mismatches quantify departures from Law 2, while their propagated sum quantifies departure from the target given by Law 1.

IV. THEOREM 2: UNCERTAINTY EVALUATION

Theorem 2. Let the implemented staged measurement equation be $X_m = Y$, $X_{k-1} = P_k(X_k, Q_k)$ and $X = X_0 = F(Y, Q_1, \dots, Q_m)$, where Q_k collects the additional input quantities contributing to uncertainty at stage k, including calibration parameters, corrections and environmental influence quantities. At the estimated values, define the sensitivity matrices: $J_k = \frac{\partial P_k}{\partial X_k}$ and $B_k = \frac{\partial P_k}{\partial Q_k}$. Let $H_Y = \Phi_{1:m}$, $H_k = \Phi_{1:k-1} B_k$, $C = [H_Y, H_1, \dots, H_m]$, $U = [Y^T, Q_1^T, \dots, Q_m^T]^T$. The centred linearised model is

$$\delta X \approx C\, \delta U. \tag{6}$$

If U has covariance matrix Σ_U, the covariance associated with the measurement result is

$$\Sigma_X = C \Sigma_U C^T \tag{7}$$

If $Y, Q_1, \dots, Q_m$ are mutually uncorrelated,

$$\Sigma_X = \Phi_{1:m} \Sigma_Y \Phi_{1:m}^T + \sum_{k=1}^{m} \Phi_{1:k-1} B_k \Sigma_{Q_k} B_k^T \Phi_{1:k-1}^T \tag{8}$$

Proof Sketch. First-order linearisation of the complete measurement chain equation gives (6). Covariance propagation gives (7), and block-diagonal Σ_U gives (8).

Theorem 2 evaluates uncertainty directly from the implemented staged measurement equation, without requiring knowledge of the measurement error. It is consistent with the GUM and retains correlations among shared input quantities [3].

V. CMM CASE STUDY

A case study of a moving-bridge MB1 CMM is performed to demonstrate the above approach for error analysis and uncertainty evaluation. To avoid confusion with the CMM X- and Y-axes, let $\mathbf{q}^*, \mathbf{q}$ be the true and reported homogeneous workpiece positions and $\mathbf{r}^*, \mathbf{r}$ the ideal and realised machine indications (Fig. 1). The CMM position measuring chain is shown in (9).

$$\mathbf{q}_0^* \xrightarrow{\mathbf{T}_y} \mathbf{q}_1^* \xrightarrow{\mathbf{T}_x} \mathbf{q}_2^* \xrightarrow{\mathbf{T}_z} \mathbf{r}^* \xrightarrow{+\Delta\mathbf{r}} \mathbf{r} \xrightarrow{\mathbf{P}_z} \mathbf{q}_2 \xrightarrow{\mathbf{P}_x} \mathbf{q}_1 \xrightarrow{\mathbf{P}_y} \mathbf{q} \tag{9}$$

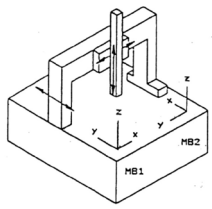


Fig. 1. MB1 moving-bridge CMM.

For $k \in \{y, x, z\}$, $\mathbf{T}_k$ is the complete 4×4 homogeneous axis transform; $\mathbf{R}_k$ and $\mathbf{t}_k$ are its 3×3 rotation and 3×1 translation. The nominal and real transforms are:

$$\mathbf{T}_k^{(0)} = \begin{bmatrix} \mathbf{R}_k^{(0)} & \mathbf{t}_k^{(0)} \\ \mathbf{0}^T & 1 \end{bmatrix}, \mathbf{T}_k = \begin{bmatrix} \mathbf{R}_k^{(0)} + \Delta\mathbf{R}_k & \mathbf{t}_k^{(0)} + \Delta\mathbf{t}_k \\ \mathbf{0}^T & 1 \end{bmatrix}. \tag{10}$$

The implemented interpretation is normally the inverse nominal or compensated transform, rather than the inverse real motion:

$$\mathbf{P}_k = \left(\mathbf{T}_k^{(0)}\right)^{-1}, \qquad \Delta\mathbf{P}_k = \mathbf{P}_k - \mathbf{T}_k^{-1}. \tag{11}$$

Thus the nominal model defines the implemented $\mathbf{P}_k$, while its departure from the inverse real transform quantifies the deviation from the inverse relation in Law 2. All $\mathbf{q}_k$ and $\mathbf{r}$ are 4×1 homogeneous vectors. With $\mathbf{q}_k^* = [(\mathbf{c}_k^*)^T, 1]^T$ and $\Delta\mathbf{q} = \mathbf{q} - \mathbf{q}_0^* = [\Delta\mathbf{c}^\mathbf{T}, \mathbf{0}]^\mathbf{T}$, the CMM real coordinates (including probe offset) at each axis are $\mathbf{c}_z^* = [X_p, Y_p, Z_p]^T$, $\mathbf{c}_x^* = [X_p, Y_p, Z + Z_p]^T$, $\mathbf{c}_y^* = [X + X_p, Y_p, Z + Z_p]^T$, we have the measurement position error:

$$\Delta\mathbf{c} \approx \Delta\mathbf{R}_y \mathbf{c}_y^* + \Delta\mathbf{t}_y + \Delta\mathbf{R}_x \mathbf{c}_x^* + \Delta\mathbf{t}_x + \Delta\mathbf{R}_z \mathbf{c}_z^* + \Delta\mathbf{t}_z. \tag{12}$$

Equation (12) combines rotational and translational errors; its expansion reproduces $(\Delta X, \Delta Y, \Delta Z)$ in [4, Eq. (4)].

Let $\boldsymbol{\alpha}$ collect the geometric-error quantities and $\mathbf{G}_{\mathrm{CMM}} = \frac{\partial \mathbf{c}}{\partial \boldsymbol{\alpha}}$. Theorem 2 gives

$$\delta\mathbf{c} \approx \mathbf{G}_{\mathrm{CMM}} \delta\boldsymbol{\alpha}, \qquad \boldsymbol{\Sigma}_\mathbf{c} = \mathbf{G}_{\mathrm{CMM}} \boldsymbol{\Sigma}_{\boldsymbol{\alpha}} \mathbf{G}_{\mathrm{CMM}}^T. \tag{13}$$

This retains correlations and, under the independence assumptions of [4], reduces diagonally to R_X^2, R_Y^2, R_Z^2 in [4, Eq. (5)]. The full paper will present the explicit expansions and demonstrate their term-by-term correspondence with the error and uncertainty results obtained using the traditional approach in [4].

VI. CONCLUSION

Theorems 1 and 2, derived from the fundamental principles, provide stage-wise error analysis and GUM-consistent uncertainty evaluation for general measurement chains. The CMM case study reproduces the established volumetric-error expressions and the corresponding random-error variances. Compared with the existing approaches, the proposed framework provides a simpler and more systematic account of error and uncertainty sources and their propagation paths, together with a clearer epistemological perspective on measurement processes.

REFERENCES

[1] Q. Yang, C. Butler, "On the framework of measurement science". Proc. XIV IMEKO World Congress, Tampere, Finland, 5, 1997, pp.101-106.
[2] Q. Yang, "Towards a unification of measurement, AI, quality and sustainability: Foundations and principles", MATEC Web Conf., 413, 2025, 00002.
[3] JCGM 100:2008, GUM, 2008.
[4] H. J. Pahk et al., "Development of virtual coordinate measuring machines incorporating probe errors", Proc. IMechE Part B, 212(7), 1998, pp. 533-548.

2026 IMEKO Joint Conference
TC7 - TC11 – TC13 - TC21 in conjunction with TC6 and Mathmet
Faro, Portugal, 27-29 October, 2026

Theta-theta position 3D fixing: measurement uncertainty evaluation

Igor Zakharov[1], Olesia Botsiura[2]

[1]*Kharkiv national university of radio electronics, Nauka ave. 14, Kharkiv, Ukraine, newzip@ukr.net*
[2]*Kharkiv national university of radio electronics, Nauka ave. 14, Kharkiv, Ukraine, olesia.botsiura@nure.ua*

***Abstract* – A models for measuring object coordinates in 3D space using the theta-theta fixation method is obtained. Expressions are presented for estimating the numerical values of object coordinates in a local rectangular coordinate system using the coordinates of the locations of two reference stations, as well as the azimuths and elevation angles of the object measured by them. Based on the law of uncertainty propagation, formulas for evaluation the standard uncertainties of the measurands are obtained. Mathematical expressions for the sensitivity coefficients and the uncertainty contributions of the input quantities to the measurands are given. Formulas for the expanded uncertainty evaluation of the measurands are obtained using the kurtosis method. Uncertainty budgets for measuring the abscissa, ordinate, and applicate of the object are constructed.**

I. INTRODUCTION

The theta-theta location (goniometric) method requires at least two reference stations equipped with only one type of measuring instruments – a goniometer. Its advantage is the absence of radiation from the reference stations (RS), which is necessary for the rangefinder (rho-rho) and rangefinder-goniometric (rho-theta) methods, as well as the absence of ambiguity in determining the object's coordinates, inherent in the rangefinder method. The estimates of the object's Cartesian coordinates (x, y, z) obtained by this method will depend on the coordinates of the RSs locations (x_1, y_1, z_1), (x_2, y_2, z_2), as well as on the target azimuths α_1 and α_2, measured by both RSs, and the elevation angle β, measured by one of the RSs. Corrections related to the inaccuracy of the RSs northward direction α_N and the inaccuracy of their platform horizontalizing β_H should be introduced into the angle measurement results. Considering these quantities as inputs to a mathematical measurement model, expressions for the standard uncertainties of the abscissa, ordinate, and applicate of an object can be derived using the law of uncertainty propagation [1]. When simultaneously and synchronously measuring azimuths and elevation angle, the correlation between the repeated results of these measurements should be taken into account. Estimates of expanded uncertainties are best obtained using the kurtosis method [2], since in this case they will be compatible with estimates obtained by the Monte Carlo method [3].

II. BASIC RELATIONSHIPS

A. Obtaining mathematical models for measuring the coordinates of an object

The Cartesian coordinates of an object (x, y, z) are determined through the parameters of a spherical coordinate system, tied to the first RS by a system of equations:

$$\left.\begin{aligned} x &= x_1 + \rho_1 \cos\beta_1 \cos\alpha_1 \\ y &= y_1 + \rho_1 \sin\beta_1 \\ z &= z_1 + \rho_1 \cos\beta_1 \sin\alpha_1 \end{aligned}\right\}, \quad (1)$$

where ρ_1 are the distances to the object from the location of the first RS.

A similar system of equations can be written for the second RS.

By solving these systems, we obtain expressions for the object's Cartesian coordinates, taking into account the above corrections. For example, the expression for the object's abscissa will look like this:

$$x = x_1 + \mathrm{ctg}(\alpha_1 + \alpha_{N1})\left[\frac{(x_2 - x_1) + (z_1 - z_2)\mathrm{ctg}(\alpha_2 + \alpha_{N_2})}{\mathrm{ctg}(\alpha_1 + \alpha_{N_1}) - \mathrm{ctg}(\alpha_2 + \alpha_{N_2})}\right]. \quad (2)$$

B. Evaluation the numerical values of an object's coordinates

The numerical values of the measurands $\hat{x}, \hat{y}, \hat{z}$ can be determined from the expressions of the mathematical models for measuring the coordinates of an object by substituting into them the estimates of the input quantities $\hat{x}_1$, $\hat{y}_1$, $\hat{z}_1$, $\hat{x}_2$, $\hat{y}_2$, $\hat{z}_2$, $\overline{\alpha}_1$, $\overline{\alpha}_2$, $\overline{\beta}_1$ and taking the mathematical expectations of the corrections $\hat{\alpha}_{N_1}$, $\hat{\alpha}_{N_2}$, $\hat{\beta}_{H_1}$ equal to zero.

For the abscissa of the object, we will have:

$$\hat{x} = \hat{x}_1 + \mathrm{ctg}(\overline{\alpha}_1)\left[\frac{(\hat{x}_2 - \hat{x}_1) + (\hat{z}_1 - \hat{z}_2)\mathrm{ctg}(\overline{\alpha}_2)}{\mathrm{ctg}(\overline{\alpha}_1) - \mathrm{ctg}(\overline{\alpha}_2 + \overline{\alpha}_{N_2})}\right]. \quad (3)$$

C. Evaluation of standard uncertainties of object coordinates

Using the mathematical models for measuring the object's coordinates (x, y, z) obtained in Subsection *A*, in accordance with the rule for summing variances and covariances [1], we can write expressions for the combined standard variances of their measurement, taking into account the correlations between the results of simultaneous measurements of angles α_1, α_2, and β_1.

For the coordinate x this expression will have the form:

$$\begin{aligned} u^2(\hat{x}) = \{&[c_{x_1}(\hat{x})u_B(\hat{x}_1)]^2 + [c_{x_2}(\hat{x})u_B(\hat{x}_2)]^2 + [c_{z_1}(\hat{x})u(\hat{z}_1)]^2 + \\ &+[c_{z_2}(\hat{x})u_B(\hat{z}_2)]^2 + c^2_{\alpha_1}(\hat{x})[u^2_A(\overline{\alpha}_1) + u^2_B(\hat{\alpha}_1) + u^2_B(\hat{\alpha}_{N_1})] + \\ &+c^2_{\alpha_2}(\hat{x})[u^2_A(\overline{\alpha}_2) + u^2_B(\hat{\alpha}_2) + u^2_B(\hat{\alpha}_{N_2})] + \\ &+2r_{\alpha_1,\alpha_2}c_{\alpha_1}(\hat{x})c_{\alpha_2}(\hat{x})u_A(\overline{\alpha}_1)u_A(\overline{\alpha}_2)\,, \end{aligned} \quad (4)$$

where the sensitivity coefficients $c_{x_1}(\hat{x})$, $c_{x_2}(\hat{x})$, $c_{z_1}(\hat{x})$, $c_{z_2}(\hat{x})$, $c_{\alpha_1}(\hat{x})$, $c_{\alpha_2}(\hat{x})$ are defined as the values of the partial derivatives of expression (2) with respect to the corresponding input quantities; $u_B(\hat{x}_1)$, $u_B(\hat{x}_2)$, $u_B(\hat{z}_1)$, $u_B(\hat{z}_2)$ are the standard uncertainties of type *B* of the coordinates of the corresponding RSs; $u_A(\overline{\alpha}_1)$, $u_A(\overline{\alpha}_2)$ are the standard uncertainties of type *A* of the measurement of the corresponding angles; $u_B(\hat{\alpha}_1)$, $u_B(\hat{\alpha}_2)$ are the standard instrumental uncertainties of the corresponding goniometers; $u_B(\hat{\alpha}_{N_1})$, $u_B(\hat{\alpha}_{N_2})$ are the standard uncertainties of type *B* of the corresponding corrections for the stations' referencing to the north; r_{α_1,α_2} are the value of the observed correlation between the results of n repeated simultaneous measurements of angles α_1 and α_2, respectively.

D. Evaluation of the expanded uncertainty of object coordinates

We will determine the expanded uncertainties of coordinate measurements (x, y, z) using the kurtosis method [2].

For the abscissa of an object, this expression has the form:

$$U(\hat{x}) = k_{0,9545}u(\hat{x})\,; \quad (5)$$

where the coverage factor for the confidence level of 0.9545 is found using the formula [4]:

$$k_{0,9545} = 0{,}12[\eta(\hat{x})]^3 + 0{,}12\eta(\hat{x}) + 2{,}0\,, \quad (6)$$

and the kurtosis of the distribution $\eta(\hat{x})$ will be equal to:

$$\begin{aligned} \eta(\hat{x}) = \frac{1}{u^4(\hat{x})}\{&\eta_x[c^4_{x_1}(\hat{x})u^4_B(\hat{x}_1) + c^4_{x_2}(\hat{x})u^4_B(\hat{x}_2)] + \eta_z[c^4_{z_1}(\hat{x})u^4_B(\hat{z}_1) + \\ &+c^4_{z_2}(\hat{x})u^4_B(\hat{z}_2)] + \eta_\alpha[c^4_{\alpha_1}(\hat{x})u^4_B(\hat{\alpha}_1) + c^4_{\alpha_2}(\hat{x})u^4_B(\hat{\alpha}_2)] + \eta_{\alpha_N}[c^4_{\alpha_1}(\hat{x})u^4_B(\hat{\alpha}_{N_1}) + \\ &+c^4_{\alpha_2}(\hat{x})u^4_B(\hat{\alpha}_{N_2})] + \eta_t[c^2_{\alpha_1}(\hat{x})u^4_A(\overline{\alpha}_1) + c^2_{\alpha_2}(\hat{x})u^4_A(\overline{\alpha}_2) + \\ &+2r_{\alpha_1,\alpha_2}c_{\alpha_1}(\hat{x})c_{\alpha_2}(\hat{x})u_A(\overline{\alpha}_1)u_A(\overline{\alpha}_2)]^2\}\,. \end{aligned} \quad (7)$$

In formula (7), η_x, η_z are the kurtosis of the distributions of the estimated abscissas and applicates of the RS locations, respectively; η_α are the kurtosis of the distributions of the estimated angles $\hat{\alpha}_1$, $\hat{\alpha}_2$, respectively; η_t is the kurtosis of the t-distribution for a probability of 0.9545 and the degrees of freedom $\nu = n - 1$.

Uncertainty budgets for the measurement of the abscissa, ordinate, and applicates of an object are constructed. An example of the uncertainty evaluation of three-dimensional coordinates of an object for the actual characteristics of RS is considered.

REFERENCES

[1] ISO/IEC Guide 98-3:2008. Uncertainty of measurement — Part 3: Guide to the expression of uncertainty in measurement (GUM:1995).

[2] I.Zakharov, O.Botsyura, "Calculation of Expanded Uncertainty in Measurements Using the Kurtosis Method when Implementing a Bayesian Approach," / Measurement Techniques, 2019, vol.1, issue 4, pp. 327-331

[3] ISO/IEC Guide 98-3:2008/Suppl 1:2008. Uncertainty of measurement — Part 3: Guide to the expression of uncertainty in measurement (GUM:1995) — Supplement 1: Propagation of distributions using a Monte Carlo method.

[4] COOMET Recommendation R/GM/35:2022 "Expression of the expanded measurement uncertainty (method of kurtosis)".

2026 IMEKO Joint Conference
TC7 - TC11 – TC13 - TC21 in conjunction with TC6 and Mathmet
Faro, Portugal, 27-29 October, 2026

Rho-rho position 2D fixing: measurement uncertainty evaluation

Igor Zakharov[1], Olesia Botsiura[2]

[1]*Kharkiv national university of radio electronics, Nauka, ave. 14, Kharkiv, Ukraine, newzip@ukr.net*
[2]*Kharkiv national university of radio electronics, Nauka, ave. 14, Kharkiv, Ukraine, olesia.botsiura@nure.ua*

***Abstract* – The mathematical models for measuring object coordinates in 2D space using the rho-rho fixation method is obtained. Expressions are presented for estimating the numerical values of the object coordinates in a local Cartesian coordinate system using the coordinates of the locations of two reference stations, as well as the distances to the object measured by them. Based on the law of uncertainty propagation, formulas for standard uncertainties evaluation of the measurands are obtained. Expressions for the sensitivity coefficients and the uncertainty contributions of the of the input quantities to the measurands are given. Formulas for the expanded uncertainty evaluation of the measurands are obtained using the kurtosis method. Uncertainty budgets for measuring the abscissa and ordinate of the object are constructed.**

I. INTRODUCTION

The rho-rho method of fixing an object on a plane (rangefinder method) requires the presence of at least two reference stations (RSs) equipped with only one type of measuring instrument – a rangefinder. Its advantage is the simplicity of implementation. In this case, no corrections should be introduced into the results of distance measurements associated with the inaccuracy of referencing the RSs to the direction to the north, as in the goniometric and goniometric-ranging methods. The estimates of the Cartesian coordinates of the object (x, y), obtained by this method, will depend on the coordinates of the location of the station points (x_1, y_1), (x_2, y_2), as well as on the distances from them to the object ρ_1 and ρ_2, measured by both RSs. Considering these quantities as input in the mathematical measurement model, it is possible to obtain expressions for the standard uncertainties of the abscissa and ordinate of the object based on the law of uncertainty propagation [1]. During simultaneous synchronous repeated measurements of distances ρ_1 and ρ_2, the correlation between the results of their measurements $\overline{\rho}_1$ and $\overline{\rho}_2$ should be taken into account. We will obtain estimates of expanded uncertainties using the kurtosis method [2], since in this case they will be compatible with estimates obtained using the Monte Carlo method [3].

II. BASIC RELATIONSHIPS

A. Obtaining mathematical models for measuring the coordinates of an object

The rangefinder method allows you to determine the Cartesian coordinates of an object $\{x, y\}$ by measuring distances ρ_1, ρ_2 to two base points with known coordinates $\{x_1, y_1\}$, $\{x_2, y_2\}$:

$$\begin{cases} \rho_1^2 = (x - x_1)^2 + (y - y_1)^2; \\ \rho_2^2 = (x - x_2)^2 + (y - y_2)^2. \end{cases} \quad (1)$$

We obtain the solution to system (1) by eliminating one of the unknowns:

$$x = x_1 + \frac{(\rho_2^2 - \rho_1^2) + (x_1^2 - x_2^2) + (y_1^2 - y_2^2)}{2\sqrt{(x_1^2 - x_2^2) + (y_1^2 - y_2^2)}}(x_2 - x_1) - (y_2 - y_1) \times$$
$$\times \frac{\sqrt{2[(x_1^2 - x_2^2) + (y_1^2 - y_2^2)]\rho_1^2 - [(\rho_2^2 - \rho_1^2) + (x_1^2 - x_2^2) + (y_1^2 - y_2^2)]^2}}{\sqrt{2}[(x_1^2 - x_2^2) + (y_1^2 - y_2^2)]}; \quad (2)$$

$$y = y_1 + \frac{(\rho_2^2 - \rho_1^2) + (x_1^2 - x_2^2) + (y_1^2 - y_2^2)}{2\sqrt{(x_1^2 - x_2^2) + (y_1^2 - y_2^2)}}(y_2 - y_1) \pm (x_2 - x_1) \times$$
$$\times \frac{\sqrt{2[(x_1^2 - x_2^2) + (y_1^2 - y_2^2)]\rho_1^2 - [(\rho_2^2 - \rho_1^2) + (x_1^2 - x_2^2) + (y_1^2 - y_2^2)]^2}}{\sqrt{2}[(x_1^2 - x_2^2) + (y_1^2 - y_2^2)]}. \quad (3)$$

B. Evaluation the numerical values of an object's coordinates

The numerical values of the measurands $\hat{x}, \hat{y}$ can be determined from the expressions of mathematical models for measuring the coordinates of an object by substituting into them the estimates of the input quantities $\overline{\rho}_1$, $\overline{\rho}_2$, $\hat{x}_1$, $\hat{y}_1$, $\hat{x}_2$, $\hat{y}_2$.

C. Evaluation of standard uncertainties of object coordinates

Using the models for measuring object coordinates (x, y) obtained in Subsection A in accordance with the rule for summing variances and covariances [1], we can write expressions for the combined standard variances of their measurements, taking into account the correlations between the results of simultaneous distance measurements $\overline{\rho}_1, \overline{\rho}_2$.

For a coordinate *x*, this expression will have the form:

$$u^2(\hat{x}) = c_{x_1}^2(\hat{x})u_B^2(\hat{x}_1) + c_{x_2}^2(\hat{x})u_B^2(\hat{x}_2) + c_{y_1}^2(\hat{x})u_B^2(\hat{y}_1) + \\ + c_{y_2}^2(\hat{x})u_B^2(\hat{y}_2) + c_{\rho_1}^2(\hat{x})u_B^2(\hat{\rho}_1) + c_{\rho_2}^2(\hat{x})u(\hat{\rho}_2) + c_{\rho_1}^2(\hat{x})u_A^2(\overline{\rho}_1) + \\ + c_{\rho_2}^2(\hat{x})u_A^2(\overline{\rho}_2) + 2r_{\rho_1,\rho_2}c_{\rho_1}(\hat{x})c_{\rho_2}(\hat{x})u_A(\overline{\rho}_1)u_A(\overline{\rho}_2), \quad (4)$$

where the sensitivity coefficients $c_{x_1}(\hat{x})$, $c_{x_2}(\hat{x})$, $c_{y_1}(\hat{x})$, $c_{y_2}(\hat{x})$, $c_{\rho_1}(\hat{x})$, $c_{\rho_2}(\hat{x})$ are defined as the values of the partial derivatives of the system of equations (2) with respect to the corresponding input quantities; $u_B(\hat{x}_1)$, $u_B(\hat{x}_2)$, $u_B(\hat{y}_1)$, $u_B(\hat{y}_2)$ are the standard uncertainties of type *B* in the estimates of the coordinates of the corresponding RSs; $u_A(\overline{\rho}_1)$, $u_A(\overline{\rho}_2)$ are the standard uncertainties of type *A* in the measurement of the corresponding distances; $u_B(\hat{\rho}_1)$, $u_B(\hat{\rho}_2)$ are the standard instrumental uncertainties of the corresponding rangefinders; r_{ρ_1,ρ_2} is the value of the coefficient of observed correlation between the results of n repeated simultaneous measurements of distances ρ_1 and ρ_2, respectively.

C. Evaluation of the expanded uncertainty of object coordinates

We will calculate the expanded uncertainties of coordinate measurements (x, y) using the kurtosis method [2].

For the abscissa of an object, this expression has the form

$$U(\hat{x}) = k_{0,9545}u(\hat{x}) \quad (5)$$

where the coverage factor for the confidence level of 0.9545 is found using the formula [4]:

$$k_{0,9545} = 0{,}12[\eta(\hat{x})]^3 + 0{,}12\eta(\hat{x}) + 2{,}0, \quad (6)$$

and the kurtosis of the distribution *x* will be equal to:

$$\eta(\hat{x}) = \frac{1}{u^4(\hat{x})}\{\eta_x[c_{x_1}^4(\hat{x})u_B^4(\hat{x}_1) + c_{x_2}^4(\hat{x})u_B^4(\hat{x}_2)] + \eta_z[c_{z_1}^4(\hat{x})u_B^4(\hat{y}_1) + \\ + c_{y_2}^4(\hat{x})u_B^4(\hat{y}_2)] + \eta_\rho[c_{\rho_1}^4(\hat{x})u_B^4(\hat{\rho}_1) + c_{\rho_2}^4(\hat{x})u_B^4(\hat{\rho}_2)] + \\ + \eta_t[c_{\rho_1}^2(\hat{x})u_A^4(\overline{\rho}_1) + c_{\rho_2}^2(\hat{x})u_A^4(\overline{\rho}_2) + \\ + 2r_{\alpha_1,\alpha_2}c_{\alpha_1}(\hat{x})c_{\alpha_2}(\hat{x})u_A(\overline{\alpha}_1)u_A(\overline{\alpha}_2)]^2\}. \quad (7)$$

In formula (7) η_x, η_y are the kurtosis of the distributions of the estimated abscissa and ordinate positions of station locations, respectively; η_ρ is the kurtosis of the distributions of the instrumental errors of the rangefinders; η_t is the kurtosis of the t-distribution for a probability of 0.9545 and the of degrees of freedom $\nu = n - 1$ [5].

Uncertainty budgets for the measurement of the abscissa, and ordinate of the object are constructed.

An example of estimating the uncertainties of two-dimensional coordinates of an object for the actual characteristics of reference stations is considered.

REFERENCES

[1] ISO/IEC Guide 98-3:2008. Uncertainty of measurement — Part 3: Guide to the expression of uncertainty in measurement (GUM:1995).

[2] I.Zakharov, O.Botsyura, "Calculation of Expanded Uncertainty in Measurements Using the Kurtosis Method when Implementing a Bayesian Approach", Measurement Techniques, 2019, vol.1, issue 4, pp. 327-331

[3] ISO/IEC Guide 98-3:2008/Suppl 1:2008. Uncertainty of measurement — Part 3: Guide to the expression of uncertainty in measurement (GUM:1995) — Supplement 1: Propagation of distributions using a Monte Carlo method.

[4] Expression of the expanded measurement uncertainty (method of kurtosis) // COOMET Recommendation R/GM/35:2022.

[5] I. Zakharov, P. Neyezhmakov, O. Botsiura, "Expanded uncertainty evaluation taking into account the correlation between estimates of input quantities", Ukrainian Metrological Journal, 2021, No 1, pp. 4-8.

Abstracts accepted as ***oral presentations***

T2 - Testing, inspection and certification

page

#37 Marija Cundeva-Blajer, ***"Expert Conformity Assessment Algorithm in Calibration by MonteCalc Uncertainty Toolkit"*** *29*

#46 Fernando Oliveira, ***"Windowed Cross-Correlation for Assessing Relative Timing Drift in DAQ Systems"*** *31*

2026 IMEKO Joint Conference
TC7 - TC11 – TC13 - TC21 in conjunction with TC6 and Mathmet
Faro, Portugal, 27-29 October, 2026

Expert Conformity Assessment Algorithm in Calibration by MonteCalc Uncertainty Toolkit

Marija Cundeva – Blajer, Monika Nakova, Kiril Demerdziev, Gjorgji Dimitrovski, Viktor Sapundziovski, Ilija Stojkovski

Ss. Cyril and Methodius University in Skopje, Faculty of Electrical Engineering and Information Technologies, str. "Rugjer Boshkovikj" No.18, Skopje, Republic of North Macedonia,
mcundeva@feit.ukim.edu.mk, monika_nakova@yahoo.com, kdemerdziev@feit.ukim.edu.mk, dzole10.mk@gmail.com, viktor.sapundziovski@gmail.com, ilijastojkovski996@gmail.com

I. INTRODUCTION

Critical infrastructure industries like power supply systems, telecommunications, or high-technology sectors like electronics, are strongly dependent on resiliant metrology capacities of electrical quantities [1], imposing confident calibration of instruments for electrical current, voltage, resistance, capacitance, inductance or frequency. A contantly present challenge is building robust models for uncertainty budget estimation, by developing advanced calculation methods, [2]. The linearization of the measuring model allows using the GUM [3], adequate for numerous cases; however, in certain complex cases it is more reliable to conduct Monte Carlo simulations as in the GUM Supplement 1, [4]. The Monte Carlo (MC) method, deploys random sampling techniques for propagation of uncertainties through complex measurement models, [5]. It shows visible advantages for electrical calibrations, due to the nonlinear, stochastic and often correlated parameters. Some initially conducted uncertainty estimations by the MC approach in the Laboratory for Electrical Measurements (LEM) at the Ss. Cyril and Methodius University in Skopje (UKIM), are presented in [6]. The standard ISO/IEC 17025 imposes the necessity of conformity statements and the guideline ILAC-G8:09/19 provides general instructions how to select appropriate decision rules and on compiling the required elements of an assessment rule if no standard rules apply, which is in compliance to the guide JCGM 106:2012. There is a risk of delivering a wrong decision because of greater measurement uncertainty, but using of guard bands decreases the likelihood of making an inappropriate decision. As in ILAC-G8, the guard band *w* is defined as:

$$w = |TL - AL| \tag{1}$$

where *TL* is the tolerance/specification limit, and *AL* is the acceptance limit.

The accredited calibration laboratory for electrical instruments LEM has established decision-making rules for conformity assessment of the calibrated instruments, [2], [6]. However, a necessity for relating the assessment with different uncertainty evaluation approaches (GUM and MC) is identified. Software developments in LEM for additional computer-intensive calculations and analysis of various calibration indicators by using different mathematical approaches, are realized and empirically validated, [6]. Here, an effort for development of an expert conformity assessment algorithm in calibration of electrical instruments in LEM will be presented.

II. UNCERTAINTY ESTIMATION IN LEM - MONTECALC UNCERTAINTY TOOLKIT

The starting calculation of the uncertainty budget is conducted according to GUM, [3]. An original LEM data fusion concept presented in [2] is applied to gain the uncertainty budget. However, if there are significant deviations from the normal distribution, like in certain calibrations of instruments for extreme (very high or very low) electrical quantities, GUM does not provide an adequate representation, [1], [6]. When one of the uncertainty contributions is significantly larger than the others, or when the magnitude order of the output quantity is the same or close order to the uncertainty, the only reliable option is the MC approach. An original universal software MonteCalc Uncertainty Toolkit, is developed by the LEM team in LabVIEW™, [7]. The two uncertainty evaluation methods, GUM and MC are embedded in the Toolkit. The software is executed on real experimental data derived from calibration of different electrical instruments. The outcomes gained from the MC methodology are compared against the results from the GUM approach and the deviations are analyzed. An important feature of the Toolkit are the expert built-in algorithms for decision making of conformity assessment against different acceptance criteria compliant with the ILAC G8:09/19. The Toolkit has no restrictions on the instrument type, the quantity measured, the range, the probability distribution etc., as it is of universal application on any case of uncertainty estimation. The results presented here are only selected specific points from the overall calibration protocols. For the algorithm validation, a RLC meter (UUT) calibration is selected in the range of electrical inductance, at the point of 20 mH. In Table 1 are

the uncertainty results from the Toolkit, by the GUM approach and by the MC simulation, respectively. The derived results with the accompanying uncertainty are subjected to conformity assessment by the Toolkit, against different decision making rules. The compliance graphs from the Toolkit, by GUM in Fig. 1 and by MC in Fig. 2 depending on the limit value, set with an appropriate control for the decision making rule, i.e. guard band *w=3U* (*U*-expanded uncertainty), at the calibration point of 20 mH the UUT, are depicted.

Table 1. Extended measurement uncertainty calculated according to the GUM and MC at point of 20 mH.

GUM	Monte Carlo
0.069 mH	0.39 mH

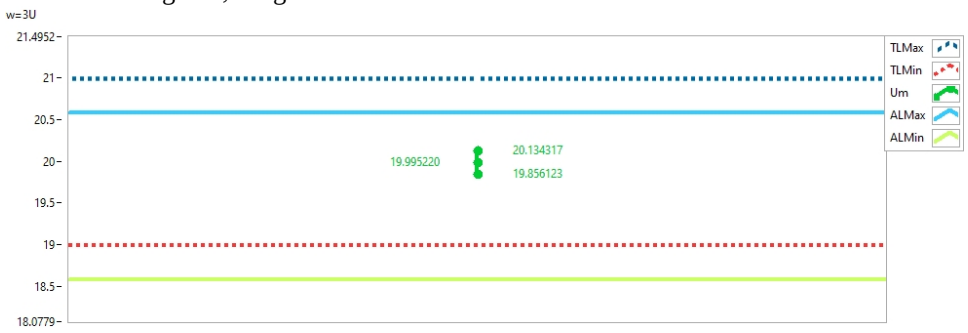


Fig. 1. Measured result at a point of 20 mH for w=3U (6-Sigma rule) with uncertainty calculated by GUM

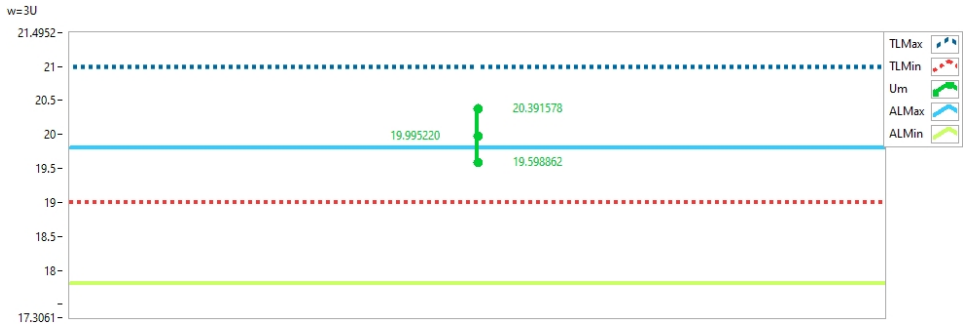


Fig. 2. Measured result at a point of 20 mH for w=3U (6-Sigma rule) with uncertainty calculated by Monte Carlo

III. CONCLUSION

The LEM originally developed software, MonteCalc Uncertainty Toolkit, implementing both GUM and Monte Carlo is applied in calibration of electrical instruments. The uncertainty results derived, are entries in the decision process, a built-in algorithm module within the MonteCalc Uncertainty Toolkit. The study provided that the two evaluation methods, generated, evidently different uncertainty values, leading to opposite decisions (pass by GUM and fail by MC). So, the MC approach is deemed more suitable as it accounts for actual distributions, resulting in a broader and more conservative uncertainty range providing more reliable conformity assessments.

REFERENCES

[1]M. Cundeva-Blajer, M. Nakova, V. Sapundziovski, K. Demerdziev, Improvement of metrology infrastructure in the area of extreme impedance calibrations, Acta IMEKO, vol. 13 (2024) no. 3, pp. 1-6. DOI: 10.21014/actaimeko.v13i3.1765

[2]M. Cundeva-Blajer "Data Fusion for Uncertainty Evaluation in Extreme Electrical Metrology", Proc. of Int. Cong. on Measurement, Quality and Data Science MQDS 2023, 5-7.06.2023, Bordeaux, France

[3]JCGM 100:2008 with member organizations (BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP and OIML), "Evaluation of measurement data – Guide to the expression of uncertainty in measurement – GUM".

[4]JCGM 101:2008 "Supplement 1 to the "Guide to the expression of uncertainty in measurement" - Propagation of distributions using a Monte Carlo method"

[5]M. G. Cox, A. B. Forbes, P. M. Harris, I. M. Smith "Measurement uncertainty evaluation associated with calibration functions" XIX IMEKO World Congress Fundamental and Applied Metrology, 2009

[6]M. Cundeva-Blajer, M. Nakova, K. Demerdziev "Conformity Assessment of Impedance Parameters Meters by MonteCalc Uncertainty Toolkit", Proc. of the IMEKO TC8&TC11&TC24 Joint Int. Conf., 14-17.09.2025, Torino, Italy

[7] "LabVIEW Programming Reference Manual", National Instruments Corporation, 2024

2026 IMEKO Joint Conference
TC7 - TC11 – TC13 - TC21 in conjunction with TC6 and Mathmet
Faro, Portugal, 27-29 October, 2026

Windowed Cross-Correlation for Assessing Relative Timing Drift in DAQ Systems

Fernando Oliveira[1], João dos Reis[1], Pedro Ferreira[1], Álvaro Ribeiro[1]

[1] *Laboratório Nacional de Engenharia Civil, Av. Brasil, 101 Lisboa, fvoliveira@lnec.pt, jpreis@lnec.pt, pmferreira@lnec.pt, asribeiro@lnec.pt*

***Abstract* – Data quality is essential in engineering and scientific applications, particularly when measurements originate from multiple acquisition systems. Ensuring that different quantities are recorded at the same time instant is critical for reliable analysis and decision-making. This work presents a method to evaluate the synchronization of signals obtained from distinct channels, sensors, or systems, by computing successive windowed normalized cross-correlation function throughout the measurement period. The method is first demonstrated through numerical simulations of a Kelvin–Voigt model, allowing the assessment of acquisition-start offsets and variable sampling periodicity. It is then applied to measurement data from two independent data-acquisition systems for testing a passive energy-dissipation device. Results show that the method effectively detects asynchronous signals and characterizes how it varies over time, particularly in systems governed by independent clocks. These findings highlight the potential of the proposed approach for synchronization assessment and motivate further developments in data-quality evaluation.**

I. INTRODUCTION

When measurements are acquired using different data-acquisition systems, two main factors can compromise synchronization. First, the systems typically do not share a common absolute time reference, meaning that their internal time bases are offset. Second, the characteristics of each internal clock may differ, leading to unequal or unstable sampling intervals due to jitter. Together, these effects result in asynchronous measurements.

The impact of clock behavior on sampling accuracy has been extensively studied. Azeredo-Leme [1] has made a comprehensive tutorial on how clock jitter affects data conversion, detailing its mechanisms, mathematical implications, and practical consequences. On the other hand, several synchronization methods have been proposed, evaluated and compared ([2], [3], [3]). Moreover, several synchronization strategies have been developed for modern data-acquisition (DAQ) systems, often relying on external timing sources. Manufacturers such as HBM and Gantner provide multiple synchronization options for their devices, including IEEE1394b FireWire, Precision Time Protocol (PTPv2), Network Time Protocol (NTP), IRIG-B, AFNOR, and EtherCAT, among others.

The National Laboratory for Civil Engineering (LNEC) in Portugal has been actively implementing systems based on those solutions in large civil engineering projects for structural health monitoring (SHM), particularly in bridges and dams, as well as applications in laboratory environment. A central concern in these contexts is the simultaneous accuracy of different DAQ systems. A preliminary evaluation of the systems synchrony was performed under laboratory conditions, and the results revealed deviations that may affect real-world measurements and post-processing, namely the identification of natural frequencies and mode shapes. Similar effects were observed in laboratory tests involving the characterization of hysteresis loops in passive energy-dissipation devices, where synchronization errors can distort the force–displacement relationship and compromise the evaluation of mechanical properties.

II. SYNCHRONIZATION EVALUATION METHOD

The proposed methodology relies on evaluating the normalized cross-correlation function between the signals. The deviation of the cross-correlation peak from zero indicates the time delay between signals. By dividing the signals into multiple windows (at least 10 times the period of the time signal under evaluation), the time delay can be tracked throughout the measurement period. The initial lag value reflects the starting offset between the signals, while the evolution over time reveals how the lag changes during acquisition. The initial lag may arise from differences in the absolute time reference of the DAQ systems, whereas the temporal evolution of the lag may result from discrepancies in sampling intervals caused by approximate sampling periods or DAQ clock jitter. It should also be noted that any physical phase difference between the signals will manifest as a lag in the cross-correlation analysis. Therefore, absolute time offsets can only be identified when a reference value is known beforehand.

To illustrate the concept, a simple numerical example is introduced, allowing a direct parallel with the practical case presented later. Two systems are considered to characterize the mechanical behavior of a passive energy-dissipation device. The first system measures displacement at a sampling rate of 2399,52 Hz, imposed by a sinusoidal excitation with 5 Hz frequency and 7,5 mm amplitude. The second system measures force, defined by the Kelvin–Voigt model with a stiffness of 10 kN/mm and a damping coefficient of 0,1 kN·s/mm, at a sampling rate of 2400 Hz. An absolute time offset of −10 ms between the two systems is also assumed. The resulting hysteresis loop is shown in Fig. 1 (left). It can be observed that the loop gradually changes from the red ellipse to the blue ellipse over the duration of the test, as a consequence of the simulated asynchronous sampling. In fact, if only one measurement system were used, the response would remain represented solely by the red ellipse. The right-hand plot of the same figure shows the lag estimated in 2-second windows (corresponding to ten cycles of the 5 Hz excitation), together with the maximum correlation values. A linear decreasing trend is observed, with a slope matching the imposed variation in sampling period, i.e., (1/2399,52-1/2400)*2400=0,0002 s/s. Furthermore, the initial lag of −0,42 ms corresponds to the difference between the intrinsic phase lag between force and displacement (9,58 ms, estimated numerically) and the imposed absolute time offset of −10 ms.

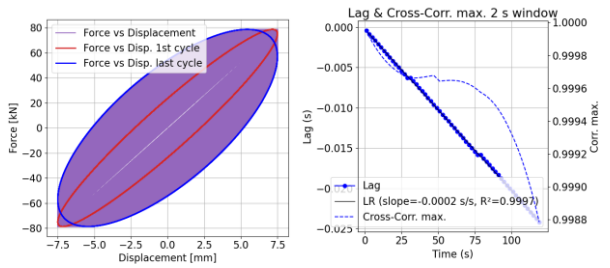


Fig. 1. Hysteresis loop (left) and the lag evolution over time, from the Kelvin–Voigt model.

III. RESULTS OF MEASUREMENTS AT THE LAB.

At LNEC, a passive energy-dissipation device was tested to evaluate its mechanical behavior. Two independent measurement systems were used: displacement was recorded using a NI PXI system (card PXIe-4497), while force was measured with an HBK QuantumX 840B. Both systems were configured for a nominal sampling rate of 2400 Hz; however, the recorded data show effective sampling frequencies of 2399.98 Hz for the displacement system and 2400 Hz for the force system. After the tests, it was observed that the hysteresis loop changed progressively over time – see Fig. 2 (left). The right-hand plot shows the evolution of the lag estimated in 2-second windows, where both the red and blue curves exhibit a monotonic decrease. The magnitude of this lag variation exceeds what would be expected solely from the sampling-period mismatch, indicating the presence of additional factors affecting the measurements, such as clock skew, jitter, or other, that requires further investigation. After resampling the data to enforce a common sampling grid, the lag curve becomes nearly horizontal (green), more closely representing the actual behavior of the dissipation device. The corresponding hysteresis loop also becomes stable, confirming that the observed drift was primarily due to asynchronous sampling rather than physical changes in the device response.

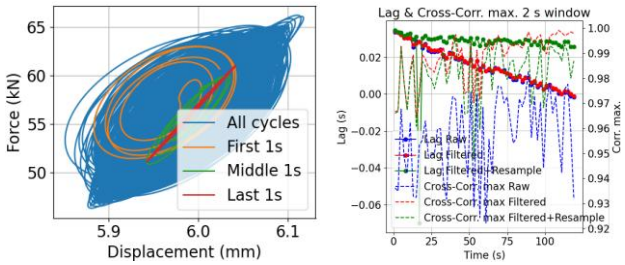


Fig. 2. Hysteresis loop (left) and the lag evolution over time, for the passive device testing.

IV. CONCLUSIONS AND OUTLOOK

The simulation showed that the estimated lag slope accurately reproduced the imposed sampling-rate mismatch, whereas the intercept reflected the combined effect of the physical phase lag and the acquisition-start offset. In the laboratory dataset, the windowed analysis revealed a progressive relative lag that was substantially reduced after resampling. The method therefore provides a diagnostic of relative timing consistency between sufficiently correlated signals. However, an independent timing reference is required to quantify absolute synchronization error and to distinguish clock-related effects from genuine physical phase variations.

REFERENCES

[1] C. Azeredo-Leme, "Clock Jitter Effects on Sampling: A Tutorial,", doi: 10.1109/MCAS.2011.942067.

[2] T. Kreuz, F. Mormann, R. Andrzejak, A. Kraskov, K. Lehnertz, P. Grassberger, "Measuring synchronization in coupled model systems: A comparison of different approaches," doi: 10.1016/j.physd.2006.09.039.

[3] B. Nalepa and A. Gwiazda, "Synchronization of Measurement Data with the Reference System", DOI 10.1088/1757-899X/1182/1/012052

[4] N. A. Paul, R. S. Rai, S. J. Vijay, "A Comparative Study on Mathematical Approaches to Determine the Time Lag and Synchrony Between Two Time-Series Data in Different Engineering Applications", https://doi.org/10.1007/978-981-16-4083-4_22

Abstracts accepted as ***oral presentations***

T3 - Measurements in Life Sciences

page

#17 Padmini Krishnadas, Urs Hackstein, Alen Bosnjakovic, Philip Aston, ***"Light vs. Dark Skin Tone Classification Using Photoplethysmography Signals"*** *34*

#29 Elisabeth Costa Monteiro, ***"Metrological Traceability Infrastructure for Diabetes-Related Laboratory Measurements: A Pilot Study of Global Availability and Access"*** *36*

#45 Thábata Rayane Gusmão, ***"Contextual Exploration of Hair Mineral Data:Integrating Group Comparisons and Reference Intervals"*** *38*

2026 IMEKO Joint Conference
TC7 - TC11 - TC13 - TC21 in association with TC6 and Mathmet
Faro, Portugal, 27–29 October, 2026

Light vs. Dark Skin Tone Classification Using Photoplethysmography Signals

Padmini Krishnadas[1], Urs Hackstein[2], Alen Bosnjakovic[3], Philip J. Aston[1,4]

[1]*Data Science and AI Department, National Physical Laboratory, Hampton Road, Teddington, TW11 0LW, UK, Padmini.Krishnadas@npl.co.uk, Philip.Aston@npl.co.uk*
[2]*Department of Life Science Engineering, Technische Hochschule Mittelhessen, University of Applied Sciences, Wiesenstrasse 14, 35390 Giessen, Germany, Urs.Hackstein@lse.thm.de*
[3]*Institute of Metrology of Bosnia and Herzegovina, Branilaca Sarajeva 25, 71000 Sarajevo, Bosnia and Herzegovina, alen.bosnjakovic@met.gov.ba*
[4]*School of Mathematics and Physics, University of Surrey, Guildford GU2 7XH, UK*

I. INTRODUCTION

It is well known that optically generated photoplethysmography (PPG) signals, which are generated by smartwatches and other wearable devices as well as clinical pulse oximeters, are affected by skin tone. To illustrate this, we previously tried to classify PPG signals by skin tone using the 6 class Fitzpatrick skin tone scale [1]. In this work, we used a fuzzy accuracy to account for the subjective nature of the skin tone classification, and obtained 96 % fuzzy accuracy using tree-based models on raw signals [2]. In this work, we try a simpler classification using the two classes of light and dark skin tones. Three machine learning approaches were used for this binary classification problem namely machine learning on features extracted from the signals, tree-based approaches on raw PPG signals and deep learning on image representations of the signals generated by the Symmetric Projection Attractor Reconstruction (SPAR) method.

II. FITZPATRICK SKIN TYPE

The Fitzpatrick skin type scale consists of six classes, ranging from very fair (class 1) to dark brown (class 6). It was originally developed to classify people with white skin in order to select the correct UV dose for treating skin disease, not for characterising skin tone [1]. However, it is now widely used for skin tone characterisation and there are a few datasets available that record the Fitzpatrick skin tone for subjects, along with other metadata.

A severe limitation of the Fitzpatrick skin tone assessment is its subjective nature. Consequently, there is often uncertainty around class assignments due to a lack of well-defined boundaries between classes. This is a potential problem when training and evaluating machine learning models against Fitzpatrick labels as the designated class could easily be one class different from the predicted class.

For our classification of light vs. dark skin tones, we define light skin tone as Fitzpatrick classes 1, 2, 3 and dark skin tone as Fitzpatrick classes 4, 5, 6. However, since some class 3 subjects will inevitably be misclassified as class 4, and some class 4 subjects will be misclassified as class 3, which results in an incorrect light/dark label in both cases, we also consider a classification of skin tone classes 1, 2 vs. classes 5, 6 so that there is a greater accuracy in the light and dark skin tone labels.

III. DATA

For this study, we used the AuroraBP dataset [3] which includes PPG waveforms and Fitzpatrick skin tone labels for 823 subjects (51.6 % female). Two distinct protocols were used, namely auscultatory and oscillometric, and we combined these into a single dataset giving a combined total of 19,017 records. The dataset is highly imbalanced with respect to skin tone, with 90.3 % of subjects and 88.2 % of records in the lighter skin tone classes 1, 2 or 3.

IV. DATASETS

Good quality PPG signals are often identified by requiring optical quality (a quality measure provided with the dataset) to be ≥ 0.65. This gave a total of 14,667 records, but only 35 of these were in skin tone class 6 (darkest skin). Thus, we used instead an autocorrelation quality measure [4] and required this to be ≥ 0.6. This resulted in a total of 16,898 records with 128 records in class 6 which, although still small, is much higher than when using the optical quality. Using these records, we created two datasets.

- **Dataset 1** We selected records to create two balanced classes of light (classes 1, 2, 3) and dark (classes 4, 5, 6) skin tone with 1.718 records in each class.

- **Dataset 2** We selected records from classes 1, 2 for light skin and classes 5, 6 for dark skin tone to reduce uncertainty from boundary labels between light and dark skin tones (classes 3 and 4). A balanced dataset was created with 628 records in each class.

Table 1. Classification accuracies for Dataset 1 (123 vs. 456) and Dataset 2 (12 vs. 56).

Method	Dataset 1 Accuracy (%)	Dataset 2 Accuracy (%)
PPG features	67.2	72.5
Raw signals (GBDT)	84.1	96.1
Raw signals (RF)	82.3	95.7
SPAR images	71.3	81.0

Table 2. UCE results for Dataset 2 (12 vs. 56).

Method	Machine/Deep Learning	UCE
Raw signals	GBDT	0.6689
Raw signals	RF	0.0039
SPAR images	ResNet50	0.4749

V. CLASSIFICATION OF SKIN TONE

We considered three machine learning approaches for this task, which employed different inputs, namely interpretable PPG features (extracted using the PulseAnalyse code [5]), raw PPG signals, and Symmetric Projection Attractor Reconstruction (SPAR) images, with one SPAR attractor image being generated from each PPG signal [6]. In each case, multiple machine learning methods were investigated but we only present results for the best performing methods in each case. The interpretable PPG features were classified using a neural network with two fully connected layers and a ReLU activation function. The classification of the raw signals was performed using a gradient-boosted decision tree (GBDT) with a binary focal loss function and a random forest (RF) approach using Gini impurity. Deep learning was used to classify the SPAR images with the best results being obtained by VGG16 (Dataset 1) and ResNet50 (Dataset 2). A stratified 5-fold cross-validation by subject was employed in all cases to ensure no data leakage between training, testing and validation sets.

VI. RESULTS

The classification results are shown in Table 1. There is a similar pattern to the accuracies for both datasets, with the highest accuracy being obtained by the raw signals, followed by the SPAR images and with poorest performance obtained using PPG features. While the PPG features are easily understood, they clearly do not capture all the information about the skin tone variation in the signals. As expected, the accuracies for Dataset 2 were higher than for Dataset 1.

VII. UNCERTAINTY QUANTIFICATION

In addition to classification accuracy, we also considered uncertainty quantification for the raw signals and SPAR results on Dataset 2. In particular, we considered the Uncertainty Calibration Error (UCE) [7]. We see from Table 2 that the GBDT applied to raw signals had the best accuracy but the worst UCE value. The SPAR images had a lower accuracy but an improved UCE, while the RF using raw signals had only slightly reduced accuracy compared to the GBDT methods, but a very small UCE value.

VIII. CONCLUSIONS

Removing skin tone classes 3 and 4 for Dataset 2 ensured both greater label accuracy and a greater difference between light and dark skin tone classes which resulted in a 12.0 % increase in accuracy for raw signals (GBDT) with a very high 96.1 % accuracy for Dataset 2. It is important to consider uncertainty quantification metrics in addition, as these results showed that the GBDT method had the best accuracy but the worst UCE value while the RF method had a slightly reduced accuracy but vastly improved UCE.

These results show there are clearly discernible differences between the PPG signals for light and dark skin. Thus, in future, diagnostic assessments or research studies using PPG signals should account for variability arising from the skin tone of subjects to ensure robust measurements and subsequent clinical interpretation.

REFERENCES

[1] T. Fitzpatrick, "The validity and practicality of sun-reactive skin types I through VI", Arch. Dermatol., vol 124, 1988, pp.869-871.

[2] P. Krishnadas, U. Hackstein, A. Bosnjakovic, P.J. Aston, "Fuzzy accuracy compensates for label subjectivity in classification of skin tone using wearable photoplethysmography signals", in Proc. IEEE Int. Conf. Fuzz, Sys., 2026.

[3] B. Mieloszyk, H. Twede, J. Lester, et al., "A comparison of wearable tonometry, photoplethysmography, and electrocardiography for cuffless measurement of blood pressure in an ambulatory setting", IEEE J. Biomed. Hlth. Inform., vol.26, 2022, pp.2864â2875.

[4] "QUMPHY D3 report", 2026. https://www.qumphy.ptb.de/publications Accessed 26th June 2026.

[5] P.H. Charlton, "Pulse-Analyse", 2023. https://github.com/peterhcharlton/pulse-analyse Accessed 20th January 2026.

[6] P.J. Aston, M. Christie, Y. Huang, M. Nandi, "Beyond HRV: Attractor reconstruction using the entire cardiovascular waveform data for novel feature extraction", Phys. Meas., vol. 39, 2018, 024001.

[7] M.-H. Laves, S. Ihler, K.-P. Kortmann, T. Ortmaier, "Uncertainty calibration error: A new metric for multi-class classification", 2021.

2026 IMEKO Joint Conference
TC7 - TC11 – TC13 - TC21 in conjunction with TC6 and Mathmet
Faro, Portugal, 27-29 October, 2026

Metrological Traceability Infrastructure for Diabetes-Related Laboratory Measurements: A Pilot Study of Global Availability and Access

Elisabeth Costa Monteiro[1], Janclei dos Santos Martins[1]

[1]*Postgraduate Programme in Metrology, Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Rio de Janeiro, Brazil, beth@puc-rio.br*

***Abstract* – Metrological traceability is essential to ensure the comparability and reliability of laboratory measurement results used for diagnosis and clinical decision-making. However, the availability and geographical distribution of higher-order metrological infrastructure vary among measurands and may affect countries' access differently. This study develops an analytical approach for mapping such infrastructure and relating its distribution to clinical priorities and geoeconomic vulnerabilities. As a preliminary application, four diabetes-related measurands were selected: glucose, because of its central role in diabetes diagnosis, and HbA1c, C-peptide and insulin, which represent progressively more complex biological measurands and different challenges for establishing higher-order traceability. Data from the KCDB, JCTLM and WHO databases were examined to identify internationally available measurement capabilities, comparisons, reference measurement procedures, services and materials. The findings reveal marked differences among the four measurands and a geographical mismatch between the availability of higher-order metrological resources and regions with major unmet needs for reliable diabetes diagnosis.**

I. INTRODUCTION

Reliable laboratory results are essential for accurate diagnosis, monitoring, and clinical decision-making, as measurement errors can lead to misclassification and inappropriate therapies [1,2]. Metrology ensures confidence in these results through metrological traceability. In laboratory medicine, tracing results to primary standards is challenging due to the complex chemical or biological nature of measurands, requiring higher-order reference measurement procedures, materials, and services [1-3]. Since the establishment of the Joint Committee for Traceability in Laboratory Medicine (JCTLM) in 2002, international infrastructure has expanded to support this integration [1-4]. This study develops a framework to evaluate whether the geographical distribution of this infrastructure aligns with clinical priorities and if geoeconomic conditions limit its national integration. As a preliminary application, we focus on diabetes-related measurands: glucose (the primary diagnostic marker), alongside glycated haemoglobin (HbA1c), C-peptide, and insulin, which represent diverse biological complexities in establishing higher-order traceability.

II. MATERIALS AND METHODS

International databases were systematically reviewed for the four diabetes-related measurands to map higher-order metrological infrastructure. The BIPM Key Comparison Database (KCDB) was analysed for published Calibration and Measurement Capabilities (CMCs) and CIPM MRA international comparisons to formally verify recognised capabilities and institutional equivalence. The JCTLM database was consulted to identify higher-order Reference Measurement Procedures (RMPs), Reference Measurement Services (RMS), and Certified Reference Materials (CRMs). The World Health Organisation (WHO) catalogue was examined for International Biological Reference Standards to account for complex biomolecules where traceability relies on International Units (IU) rather than direct SI units. Identified resources were categorised by database, provider, and country. Their geographical distribution was cross-referenced with World Bank income classifications and epidemiological indicators (diabetes prevalence, undiagnosed cases, and treatment coverage) to detect disparities between metrological resource availability and regional clinical needs.

III. DISCUSSED RESULTS AND CONCLUSION

Marked differences were identified in the maturity, diversity and geographical availability of internationally recognised higher-order metrological resources for the four diabetes-related measurands. Glucose exhibited the most comprehensive metrological framework. It was the only measurand with declared KCDB CMCs and was represented in KCDB comparisons and across all three JCTLM categories considered: RMPs, RMS, and CRMs. Twelve countries participated in the three identified KCDB glucose comparisons, distributed across the

Americas, Europe, and Asia. South Korea was the only country to participate in all three comparisons conducted to date. This comparatively mature infrastructure supports the feasibility of establishing an SI-traceable glucose measurement system using approaches closely related to classical chemical metrology.

For HbA1c, the infrastructure was more restricted. Internationally recognised JCTLM Reference Measurement Services were identified in only four countries—China, France, Germany and Japan—while JCTLM reference materials were associated with France and Belgium. France was the only country identified as providing JCTLM reference materials for both glucose and HbA1c. Despite these limitations, the BIPM international comparison for HbA1c demonstrates that efforts are underway to expand metrological competence beyond the limited number of countries that provide higher-order reference services and materials.

The infrastructure identified for C-peptide was considerably more limited. JCTLM Certified Reference Materials were associated with a single provider country, Japan, while WHO biological reference standards are supplied through the United Kingdom. No KCDB comparison dedicated specifically to C-peptide was identified. The limited diversity and geographic availability of higher-order resources underscore the additional challenges of establishing traceability systems for peptide biomarkers.

Insulin presented a different and particularly challenging metrological configuration. A KCDB international comparison conducted in 2017–2018 involved nine countries, demonstrating internationally distributed measurement competence. Nevertheless, no CMCs, JCTLM Reference Measurement Services or JCTLM Certified Reference Materials were identified for insulin. International standardisation therefore relies substantially on WHO International Biological Reference Preparations expressed in International Units. This distinction illustrates that participation in international measurement comparisons does not necessarily imply the availability of higher-order resources that can directly support routine traceability chains. However, it may suggest that efforts are underway to address challenges in establishing a primary calibrator material.

Taken together, the four measurands reveal a progressive restriction in the availability and diversity of higher-order traceability resources as the biological and metrological complexity of the measurand increases. The transition extends from the comparatively comprehensive SI-based infrastructure available for glucose to increasingly restricted reference-material and reference-service systems for HbA1c and C-peptide, and finally to the predominance of biological standardisation through WHO International Units for insulin.

A second structural feature concerns geographical distribution. The internationally recognised resources identified were provided predominantly by countries in Europe, Asia and, for some categories, the Americas. No provider of the higher-order reference materials or services examined was identified in Africa, and none was identified among low-income or lower-middle-income economies. This distribution contrasts with the global epidemiology of diabetes. Treatment coverage is particularly low in parts of sub-Saharan Africa and South Asia, while a very high proportion of untreated diabetes remains undiagnosed in these regions [5]. Countries combining substantial diagnostic needs with limited economic resources may therefore depend on higher-order metrological infrastructure located abroad, potentially increasing technical, logistical and economic barriers to implementing complete traceability chains in routine laboratory practice.

This pilot study consequently identifies two complementary challenges: differences in the maturity of higher-order metrological infrastructure by measurand complexity, and a geographical mismatch between the locations of such infrastructure and regions with substantial unmet clinical and economic needs. Although limited to four diabetes-related measurands, the proposed approach provides a basis for investigating whether similar patterns occur across laboratory medicine and for informing strategies to expand equitable access to metrological traceability worldwide.

REFERENCES

[1] E.Costa-Monteiro, and R.Summers, “Metrological requirements for biomedical device assessment and their ethical implications”, Measurement: Sensors, Vol. 24, December 2022, 100574.

[2] E.Costa-Monteiro, “150 Years of the Evolution of Metrology in Health: An Overview from the Metre Convention (1875) to Today (2025)”, Plenary Lecture at the Test and Measurement International Conference 2025. Pretoria, South Africa: National Laboratory Association (NLA).

[3] M.S.Lantman, C.Cobbaert , M.Panteghini, M.Berkel, R.L.Smeets, J.J.Hellemond and M.H.M.Thelen, “Why Metrological Traceability Matters in Medical Laboratory Diagnostics”, JALM, 2026, pp. 632–646

[4] G.R.D.Jones, C.Jackson, “The Joint Committee for Traceability in Laboratory Medicine (JCTLM)–its history and operation”, Clinica Chimica Acta 453, 2016, pp. 86-94

[5] F.Teufel, K.Orgutsova, I.Genitsaridi, R.Carrillo-Larco, J.Varghese, M.Marcus, J.W.Sacre, S.F.Sajjadi, F.Chiwanga, J.Manne-Goehler, J.Seiglie, D.Flood, J.Harding, P.Salpea, A.Salim, E.Boyko, D.Magliano, M.K.Al, “Global, Regional, and National Estimates of Undiagnosed Diabetes in Adults: Findings From the 2025 IDF Diabetes Atlas ”, Diabetes Care, 49, 2026, pp. 490–496.

2026 IMEKO Joint Conference
TC7 - TC11 – TC13 - TC21 in conjunction with TC6 and Mathmet
Faro, Portugal, 27-29 October, 2026

Contextual Exploration of Hair Mineral Data:Integrating Group Comparisons and Reference Intervals

Thábata Rayane Gusmão[1,2], Tatiana Dillenburg Saint'Pierre[3], Elisabeth Costa Monteiro[4]

[1] *Department of Informatics, Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Brazil*
[2] *Tecgraf Institute, Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Brazil*
[3] *Department of Chemistry, Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Brazil*
[4] *Postgraduate Programme in Metrology, PUC-Rio, Brazil*

***Abstract* – Hair mineral analysis can reveal multi-element patterns, but interpretation requires both statistical comparison across participant characteristics and contextualisation relative to reference intervals. We developed a graphical application integrating these two perspectives and applied it retrospectively to three ICP-MS hair datasets collected in 2013, 2018, and 2023, comprising 116, 176, and 169 women, respectively. Elemental concentrations were compared across available contextual variables using one-way ANOVA and Tukey-Kramer comparisons, and visualised both in their original units and relative to established hair reference intervals. Across datasets, cosmetic treatment showed the most consistent associations. Significant group differences recurred for Ca, Mg, Sr, Na, and K, with treated-hair groups generally showing higher concentrations; for Ca, Mg, and Sr, some treated groups approached or exceeded upper reference limits. Other associations varied among datasets. The application supports exploratory identification of recurrent and context-dependent patterns without assigning causal or clinical significance to the observed associations.**

I. INTRODUCTION

Hair is a non-invasive matrix for multi-element analysis and has been used in biomonitoring [1] and health-related studies [2]–[4]. Literature indicates that hair composition varies with demographic characteristics and cosmetic treatments [5], including hair colouring and bleaching [6]. Furthermore, measurements from hair fibre matrices may not correspond to blood analysis results, as blood reflects a specific point in time, while hair fibres integrate exposures over a longer period [7]. Beyond quantifying elements and assessing differences in their concentrations, multi-element analysis requires interpreting observed differences in light of conditions and factors associated with variation in the measured concentrations.

The complexity of interpreting hair mineral data is particularly evident in physiological and clinical contexts. Elemental profiles have been investigated in relation to obesity and hypertension [2], osteoporosis [3], and low bone mass, with hair measurements incorporated into predictive models alongside broader health-assessment variables [4]. Additional studies have compared hair measurements with serum, dietary, and functional indicators, such as in assessing iodine status [8]. Although these studies address distinct research questions and populations, collectively they illustrate the importance of considering relevant contextual variables when interpreting hair mineral results.

A graphical application was developed to integrate statistical comparisons across groups defined by participant characteristics with visualisation of elemental concentrations relative to established reference intervals. It was used to examine three retrospective hair-mineral datasets obtained by ICP-MS in 2013, 2018, and 2023, with the aim of identifying recurrent and dataset-dependent patterns for further investigation.

II. MATERIALS AND METHODS

Group comparisons show whether elemental concentrations differ among participants but not where they fall relative to established reference intervals. We therefore incorporated intervals established for minor and trace elements in human hair from an urban population of Rio de Janeiro (N = 1775) [9]. That study employed ICP-MS and statistical procedures based on IFCC recommendations and evaluated the analytical method using the GBW-09101 human-hair certified reference material, with results agreeing with certified values within the associated uncertainties.

Concentrations are displayed in their original units and relative to the corresponding reference interval. For intervals with lower (L) and upper (U) limits, concentration x is expressed as $r = (x - L)/(U - L)$, so the interval spans 0 to 1; when only an upper limit U is available, $r = x/U$. Participant information was used to define comparison groups. Non-numeric entries were treated as missing, negative analytical values were set to

zero, and outliers identified for each element × group combination using the 1.5-IQR rule were excluded from subsequent analyses. The remaining observations were displayed in boxplots and analysed by one-way ANOVA, with Tukey-Kramer pairwise comparisons when appropriate.

The datasets comprised 116 women in 2013 (4–88 years), 176 in 2018 (15–21 years), and 169 in 2023 (15–59 years), and were analysed separately because participant characteristics and available contextual information differed among years. Cosmetic treatment was documented in all datasets, while smoking, oral-contraceptive use, and alcohol consumption were available in selected datasets. The reference catalogue provided source-specific intervals for 29 elements.

III. RESULTS

Cosmetic treatment showed the most consistent association with elemental concentrations across all three datasets. Significant differences among treatment groups were observed in each year for Ca, Mg, Sr, Na, and K. For Mg and Sr, groups with treated hair, particularly those with combined straightening and colouring or colouring alone, consistently showed higher concentrations, with some treated groups exceeding the upper reference limits. In contrast, untreated hair generally remained within the corresponding intervals. Ca showed a similar pattern, with higher concentrations and values above the reference interval in specific treated groups. For Na and K, combined straightening and colouring consistently corresponded to higher-concentration groups across all datasets.

Other elements showed less reproducible patterns across years. Treatment-related differences for As and Se were observed in 2013 and 2018 but not in 2023, while Zn differed among treatment groups in 2013 and 2023 but not in 2018. Associations with smoking and oral-contraceptive use were identified for selected elements, although these findings were less consistently reproduced. Statistical group comparisons and positioning relative to reference intervals therefore provided complementary perspectives, helping to distinguish recurrent patterns from dataset-specific findings.

IV. DISCUSSION AND CONCLUSION

The results highlight two complementary questions: whether concentrations differ among groups defined by contextual variables and where their distributions fall relative to reference intervals. Recurrent treatment-associated patterns across datasets with different age distributions support further investigation, whereas less reproducible findings illustrate the importance of dataset-specific context. The graphical application integrates these perspectives, facilitating exploratory identification and prioritisation of recurrent patterns. Given the group-specific outlier exclusion and differences among datasets, the analyses remain exploratory; their retrospective and observational nature precludes causal or diagnostic interpretation of the identified associations.

REFERENCES

[1] A. Lallmahomed, F. Mercier, N. Costet, C. Fillol, N. Bonvallot, and B. Le Bot, “Characterization of organic contaminants in hair for biomonitoring purposes,” Environ. Int., vol. 183, Art. no. 108419, 2024, doi: 10.1016/j.envint.2024.108419.

[2] A. V. Skalny et al., “Interactive Effects of Obesity and Hypertension on Patterns of Hair Essential Trace Element and Mineral Content in Adult Women,” Biol. Trace Elem. Res., vol. 201, pp. 4677–4687, 2023, doi: 10.1007/s12011-023-03561-y.

[3] A. V. Skalny et al., “Hair and Serum Trace Element and Mineral Levels Profiles in Women with Premenopausal and Postmenopausal Osteoporosis,” Biol. Trace Elem. Res., vol. 202, pp. 3886–3899, 2024, doi: 10.1007/s12011-023-03970-z.

[4] S. J. Kang et al., “Preventive machine learning models incorporating health checkup data and hair mineral analysis for low bone mass identification,” Sci. Rep., vol. 14, Art. no. 18792, 2024, doi: 10.1038/s41598-024-69090-3.

[5] G. Peregrino, C. G. Massone, A. H. Nudi, and T. D. Saint’Pierre, “Hair Mineralogram Analysis for Health Assessment: Statistical Bias from Gender and Aesthetic Treatments,” Braz. J. Anal. Chem., vol. 8, no. 33, pp. 71–88, 2021, doi: 10.30744/brjac.2179-3425.AR-20-2021.

[6] J.-Y. Shim et al., “The effects of dyes and bleaches on the sorption and desorption of trace elements in human hair,” Environ. Geochem. Health, vol. 45, pp. 2349–2358, 2023, doi: 10.1007/s10653-022-01358-x.

[7] A. Shahverdian et al., “Comparison of Mineral Levels in Blood and Hair Samples of Healthy Adults: Evaluating the Clinical Utility of Hair Mineral Analysis,” Biol. Trace Elem. Res., vol. 204, pp. 1915–1927, 2026, doi: 10.1007/s12011-025-04793-w.

[8] Y. Fuse, Y. Kunii, N. Tsukada, Y. Ito, and Y. Shishiba, “Scalp hair and serum iodine as potential biomarkers for iodine intake and their association with the thyroid function,” J. Endocr. Soc., vol. 10, no. 5, Art. no. bvag084, 2026, doi: 10.1210/jendso/bvag084.

[9] M. T. W. D. Carneiro, C. L. Porto da Silveira, N. Miekeley, and L. M. de Carvalho Fortes, “Intervalos de referência para elementos menores e traço em cabelo humano para a população da cidade do Rio de Janeiro-Brasil,” Quim. Nova, vol. 25, pp. 37–45, 2002.

Abstracts accepted as ***oral presentations***

T4 - Mathematical and statistical modelling

page

#7 Ignacio Lira, ***"Simple GUM-Compliant Uncertainty Evaluation for Errors-in-Variables Regression"*** *41*

#8 Katsuhiro Shirono, ***"Can predictive uncertainty in machine learning be measurement uncertainty?"*** *43*

#10 Severine Demeyer, "***Interpreting interlaboratory comparison results: current challenges and future perspectives"*** *45*

#16 Alistair Forbes, ***"Distributions associated with nonlinear least-squares problems"*** *47*

#19 Alistair Forbes, ***"The Central Limit Theorem and sums of t-variates"*** *49*

#22 Ahmad Barari, Wahab Janjuaa, ***"Loosely-Coupled Camera-Anchored Residual Learning for Three-Dimensional Handheld Inertial Navigation"*** *51*

#25 Franco Pavese, ***"Controversial new definition of "Measurement Uncertainty""*** *52*

#30 Hugo Gasca-Aragón, ***"THE DIGITAL METROLOGICAL EXPERT (DME)"*** *53*

#33 Zeliha Hanım Karapınar, ***"Self-Validating Bayesian Recalibration Interval Optimization of Mass Standards"*** *55*

#34 Tameem Adel, ***"Metrological Validation in Medical AI: Using Counterfactuals to Isolate Sensor-Induced Shortcut Learning"*** *57*

#36 Soheil Ahmadi, Kenneth Grattan, Sanowar Khan, ***"Geometry-Resolution Effects in 3-D Finite-Element Simulations of Twisted-Conductor Loss"*** *60*

#39 Vivek Desai; Andrew Thompson, ***"Extending Confidence Calibration to Generalised Measures of Variation"*** *62*

#40 Francesca Pennecchi, ***"Extension of the JCGM 106 approach: assessing the number of truly conforming items in a lot"*** *64*

#41 Miles McCrory, Weronika Pryciuk, Liam Wright, ***"Quantitative Image Analysis: An Approach to Estimating Uncertainties Associated with Image Segmentation"*** *67*

#42 Liam Wright, ***"Bridging the Gap Between Chemometrics and Metrology"*** *69*

#44 Eva Fišerová, Daniel Gogola, Gejza Wimmer, Viktor Witkovský, ***"Ellipsoid Fitting for Geometric Calibration: Uncertainty Quantification and Sphericity Testing"*** *71*

Simple GUM-Compliant Uncertainty Evaluation for Errors-in-Variables Regression

Ignacio Lira[1]

[1] *Pontificia Universidad Católica de Chile, Santiago, Chile*

Abstract – **The classical errors-in-variables (EIV) regression problem has long been central to metrology, though parameter uncertainty remains difficult to quantify when models are non linear or depart from Gaussian assumptions. This paper presents a transparent, three-stage Monte Carlo framework, compliant with GUM Supplement 1, that address these challenges. By treating input coordinates as probability density functions that encode epistemic uncertainty, the method directly generates joint parameter distributions without the computational burden of full Bayesian MCMC samplers. The framework is validated against the Pearson–York straight-line benchmark under non-Gaussian inputs and applied to the non-linear calibration of an industrial R-type thermocouple. Results show that the method outperforms geometric fitting and standard least-squares linearizations, offering a practical mathematical engine for uncertainty propagation in digital calibration certificates.**

I. INTRODUCTION

Fitting a curve to empirical data where both the stimulus and response are subject to measurement uncertainties—the classical errors-in-variables (EIV) problem—is a ubiquitous task in metrology. Conventional weighted least-squares (WLS) and Orthogonal Distances (OD) methods provide point estimates for calibration coefficients but often fail to provide rigorous uncertainty evaluations, particularly when dealing with heteroscedasticity or non-linear models. While full Bayesian inference solves this robustly [1], the required Markov Chain Monte Carlo (MCMC) simulations can be computationally prohibitive and opaque for industrial practitioners.

Recent attempts to solve the EIV problem analytically highlight significant mathematical bottlenecks. For instance, applying standard least-squares alongside GUM-S1 often fails for non-linear models or unknown error variances, yielding structural divergences from exact Bayesian posteriors [2]. Furthermore, rigorous analytical treatments of EIV models—such as thermocouple calibration—require massive block matrices with Type II constraints and iterative Taylor linearizations [3]. Even the most recent analytical frameworks rely on implicit differentiation and local first-order approximations, yielding only standard uncertainties that are blind to the topological shifts in coverage intervals caused by non-Gaussian inputs [4].

This work introduces a transparent, three-stage forward Monte Carlo framework fully compliant with Supplement 1 to the Guide to the Expression of Uncertainty in Measurement (GUM S1). The method generates complete joint probability density functions (PDFs) for the calibration coefficients, bypassing the need for complex priors and unwieldy linearizations while seamlessly handling non-Gaussian input uncertainties.

II. THE THREE-STAGE METHODOLOGY

The framework treats coordinate data not as fixed mathematical points, but as epistemic state-of-knowledge PDFs.

Stage 1: Each coordinate (x_i, y_i) is assigned a specific PDF based on calibration histories, instrument resolution, or type B evaluations. This allows for the natural inclusion of asymmetric, uniform, or heavy-tailed distributions.

Stage 2: A direct Monte Carlo loop samples from these input PDFs, solving a deterministic non-linear optimization for each iteration. The aggregated results form the joint empirical PDF of the model's parameters.

Stage 3: The resulting joint distribution is stored (e.g., within a Digital Calibration Certificate), allowing end-users to invert the model dynamically and calculate future stimulus values with full uncertainty traceability.

III. CASE STUDY 1: THE PEARSON-YORK BENCHMARK WITH NON-GAUSSIAN INPUTS

To validate the framework's mathematical consistency, it was first applied to the historical benchmark of Pearson's data with York's weights. This specific problem has been treated countless times as a benchmark for straight-line EIV regression, as extensively

reviewed by Cantrell [5], and has been tackled through probabilistic and least-squares inference by the author [6], as well as more recently using full Bayesian methods for Type A evaluation [7].

While extensively analyzed under Gaussian assumptions in the aforementioned literature, industrial measurements often suffer from uncorrected systematic effects or resolution limits better described by non-Gaussian PDFs. By assigning uniform and Student-t distributions to the classical Pearson-York variances, the proposed Monte Carlo framework reveals significant structural shifts and expanding coverage intervals in the resulting parameter distributions, exposing the risks of blindly applying standard WLS linearizations to non-ideal data.

IV. CASE STUDY 2: INDUSTRIAL R-TYPE THERMOCOUPLE CALIBRATION

Moving from abstract benchmarks to applied metrology, the framework was deployed to calibrate an R-type thermocouple. Evaluating calibration uncertainty for interpolating polynomials between fixed points is critical, as excursions between these points contribute significantly to interpolated uncertainty [8]. While recent literature has explored this non-linear calibration using Orthogonal Distances [9], such purely geometric approaches fundamentally neglect the rigorous propagation of measurement uncertainty. Furthermore, even when standard frameworks are applied to EIV problems, approximations inherent in documents like ISO/TS 28037 can severely understate the actual uncertainty, making Monte Carlo methods highly preferable [10]. Monte Carlo methods are superior for this exact application, especially when distributions are non-normal and correlation is present, confirming that analytical LPU methods have limitations.

In this study, realistic industrial uncertainties were assigned to the standard calibration data (e.g., $u(T) = 0.5°\mathrm{C}$ for the thermal bath and $u(E) = 1\mu\mathrm{V}$ for the digital multimeter). The three-stage framework successfully extracted the joint PDFs for the polynomial coefficients. The results demonstrate how propagating actual measurement uncertainties drastically alters the calibration landscape compared to purely geometric fitting, ensuring that the sensor's digital twin accurately reflects the physical reality of the laboratory.

V. CONCLUSIONS

The proposed framework offers a practical bridge between rigorous probability theory and everyday metrological practice. By successfully managing both non-Gaussian theoretical benchmarks and real-world sensor calibrations, it provides a vital computational tool for the emerging infrastructure of digital metrology.

VI. REFERENCES

[1] I. Lira and D. Grientschnig,"Error-in-variables models in calibration," *Metrologia*, vol. 54, pp. S133–S145, 2017.

[2] C. Elster and B. Toman, "Bayesian uncertainty analysis for a regression model versus application of GUM Supplement 1 to the least-squares estimate," *Metrologia*, vol. 48, pp. 233–240, 2011.

[3] G. Wimmer, S. Ďuriš, R. Palenčár, and V. Witkovský, "EIV calibration model of thermocouples," in *Advanced Mathematical and Computational Tools in Metrology and Testing XI*, World Scientific, pp. 401–408, 2018.

[4] Y. Zhang *et al.*, "A GUM-based method for uncertainty evaluation of the slope in linear measurement models," *Meas. Sci. Technol.*, vol. 37, 275001, 2026.

[5] C. A. Cantrell, "Technical Note: Review of methods for linear least-squares fitting of data and application to atmospheric chemistry problems," *Atmos. Chem. Phys.*, vol. 8, pp. 5477–5487, 2008.

[6] I. Lira, C. Elster, and W. Wöger, "Probabilistic and least-squares inference of the parameters of a straight-line model," *Metrologia*, vol. 44, pp. 379–384, 2007.

[7] A. M. H. van der Veen, "Bayesian methods for type A evaluation of standard uncertainty," *Metrologia*, vol. 55, pp. 670–684, 2018.

[8] J. V. Pearce, P. M. Harris, and J. C. Greenwood, "Evaluating Uncertainties in Interpolations Between Calibration Data for Thermocouples," *Int. J. Thermophys.*, vol. 31, pp. 1517–1526, 2010.

[9] L. A. Cantera-Cantera, C. Vargas-Jarillo, S. I. Palomino-Reséndiz, Y. Lozano-Hernández, and C. M. Montelongo-Vázquez, "A Polynomial Fitting Problem: The Orthogonal Distances Method," *Mathematics*, vol. 10, no. 4596, 2022.

[10] K. Klauenberg, S. Martens, A. Bošnjaković, M. G. Cox, A. M. H. van der Veen, and C. Elster, "The GUM perspective on straight-line errors-in-variables regression," *Measurement*, vol. 187, 110340, 2022.

2026 IMEKO Joint Conference
TC7 - TC11 – TC13 - TC21 in conjunction with TC6 and Mathmet
Faro, Portugal, 27-29 October, 2026

Abstract/Can predictive uncertainty in machine learning be measurement uncertainty?

Katsuhiro Shirono[1], Junnosuke Takai[1], Yasuhiro Mizutani[2]

[1] *National Institute of Advanced Industrial Science and Technology (AIST)/National Metrology Institute of Japan (NMIJ), k.shirono@aist.go.jp, takai.jun@aist.go.jp*
[2] *Department of Mechanical Engineering, Graduate School of Engineering, Osaka University, mizutani@mech.eng.osaka-u.ac.jp*

***Abstract* – Machine learning (ML), particularly neural networks, has been increasingly applied in measurement domain. In such applications, uncertainty quantification is essential for evaluating reliability. Bayesian neural networks (BNNs) are often used for this purpose, as they provide probabilistic outputs. However, it remains unclear whether the uncertainty obtained from BNNs can be interpreted as measurement uncertainty. In this study, we examine this issue from five perspectives: out-of-distribution behavior, lack of uncertainty propagation in training data, mismatch in inferential frameworks, computational limitations of approximate Bayesian methods, and difficulty in estimating heteroscedastic variance. These findings suggest that ML uncertainty and measurement uncertainty are conceptually distinct.**

I. INTRODUCTION

Recently, machine learning (ML) has been widely applied in the field of measurement, and highly accurate predictions based on neural networks have become possible. In such applications, it is important not only to obtain accurate predictions but also to quantify their uncertainty in order to evaluate their reliability. Bayesian neural networks (BNNs) are a representative method in this context, as they provide uncertainty in the form of probability distributions over outputs, and are therefore widely used for quantifying uncertainty, e.g. [1].

However, it remains unclear to what extent the uncertainty obtained from BNNs can be interpreted as measurement uncertainty. In this study, we aim to systematically examine these issues and to clarify under what conditions the uncertainty derived from BNNs may be interpreted as measurement uncertainty.

II. ANALYSIS OF PREDICTIE UNCERTAINTY IN MACHINE LEARNING

This section examines the extent to which uncertainty derived from ML, particularly BNNs, can be interpreted as measurement uncertainty.

A. Out-of-distribution behavior

ML models are typically trained under the assumption that training and deployment data follow the same distribution. When inputs fall outside this domain, the model may still produce apparently reasonable outputs, but these cannot be interpreted as reliable inferences. Although some BNNs may reflect OOD conditions through increased epistemic uncertainty, defined here as uncertainty ascribed to the posterior, this behavior is not consistently observed. As a result, quantified uncertainty may fail to represent the true lack of knowledge.

However, from a measurement perspective, this is not primarily an uncertainty issue but a question of applicability. Measurement results are valid only within the domain where the instrument has been validated. Therefore, OOD should be treated as an issue of method validation, rather than as an issue to be compensated for by uncertainty quantification.

B. Propagation of uncertainty in training data

In measurement, observations are inherently uncertain, and this uncertainty propagates through the model. In contrast, most ML approaches treat training data as deterministic. The uncertainty associated with observations, including noise and correlation structure, is generally not explicitly modeled.

Although this issue has been discussed [2], there is no widely accepted framework to propagate input uncertainty consistently through ML models. As a result, quantified uncertainty may not fully reflect the uncertainty of the underlying data-generating process.

This leads to a fundamental limitation: ML predictions are not derived from a probabilistically consistent treatment of measurement uncertainty, and their interpretation as such remains unclear.

C. Mismatch in inferential frameworks

The most fundamental issue lies in the difference

between predictive and inferential problem settings. BNNs typically provide posterior predictive distributions, which describe variability in possible observations. In contrast, measurement aims to infer the value of a measurand represented in a physical relationship model. In the common framework of Bayesian statistics, including the JCGM 101 [3] approach, it is appropriate to investigate a posterior distribution over the measurand itself to quantify measurement uncertainty.

We examined the conditions under which predictive distributions and measurand posteriors may coincide. We found that a prior distribution for the measurand is key, and that the conditions for coincidence are rarely satisfied in metrological measurement practice, which differs from typical ML applications. This implies that predictive uncertainty obtained from BNNs does not directly correspond to measurement uncertainty. This is not because the BNN approach is incorrect, but because it addresses a different problem: prediction rather than inference.

D. Computational limitations of variational inference approaches

Most BNNs rely on approximate inference methods such as variational inference, one representative method of which is Monte Carlo dropout (MCDO). There have been reports on the computational limitations of MCDO [4]. We also investigated this property by applying BNNs with MCDO to a simple prediction of $f(x) = -100x + x^3$ for $-10 < x < +10$. Here x is the only input and $f(x)$ is the target output. The standard deviation of the observation error was set to 5, and a few outliers were included. The ML model consisted of two hidden layers with 20 nodes each and a fully connected output layer. The dropout rate was set to 0.1. For prediction at $x = 2.5$, Fig. 1 indicates that increasing the size of the training data does not necessarily reduce the quantified standard deviation, which corresponds only to the epistemic uncertainty determined by MCDO.

Moreover, standard BNN frameworks do not capture structural uncertainties such as model rigidity and regularization effects. These factors can introduce systematic errors that are indistinguishable from aleatoric uncertainty, i.e., the uncertainty associated with observation error.

E. Computational limitations in heteroscedastic variance estimation

Heteroscedasticity refers to the situation in which the variance of observation errors depends on the input conditions. In measurement, such input-dependent variability should be explicitly quantified as part of the uncertainty. However, in machine learning, estimating this variance accurately is not straightforward [5]. As a result, the estimation of input-dependent variance can be unstable or biased. This problem is not specific to BNNs, but is common to ML approaches in general.

III. CONCLUSION

In this study, we examined the extent to which uncertainty derived from ML models, particularly BNNs, can be interpreted as measurement uncertainty. By analyzing five issues, we clarified the fundamental limitations of such an interpretation. It may not be fair to regard computational limitations as problems specific to BNNs, because no common alternative methodology for measurement uncertainty quantification has been established. However, these results indicate that ML uncertainty obtained through BNNs can be useful within a predictive framework, but should not be directly interpreted as measurement uncertainty.

REFERENCES

[1] J.Takai, Y.Tanaka, M.Yoshioka, and K.Shirono, "Approximative Bayesian approach for uncertainty evaluation in machine learning-based hardness measurement," Proc. of M4Dconf2025.

[2] A.Thompson, "Analytical results for uncertainty propagation through trained machine learning regression models," Measurement, vol.234, 2024, 114841.

[3] JCGM 101, BIPM, 2008.

[4] F.Verdoja and V.Kyrki, "Notes on the Behavior of Monte Carlo Dropout," Proc. of ICML 2021 Workshop on Uncertainty and Robustness in Deep Learning, 2021.

[5] M.Seitzer, A.Tavakoli, D.Antic, and G.Martius, "On the Pitfalls of Heteroscedastic Uncertainty Estimation with Probabilistic Neural Networks," Proc. of ICLR, 2022.

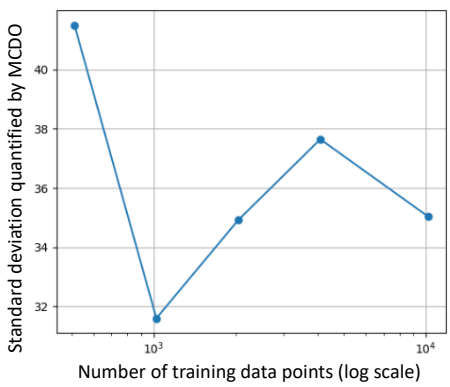


Fig. 1. Quantified standard deviation with Monte Carlo Dropout approach (epistemic uncertainty)

2026 IMEKO Joint Conference
TC7 - TC11 – TC13 - TC21 in conjunction with TC6 and Mathmet
Faro, Portugal, 27-29 October, 2026

Interpreting interlaboratory comparison results: current challenges and future perspectives

Séverine Demeyer[1], Maurice G Cox[2], Stephen LR Ellison[3], Steffen Uhlig[4], Adriaan M H van der Veen[5], Gerd Wübbeler[6]

[1] LNE, 29 avenue Roger Hennequin, 78197 Trappes Cedex, France, severine.demeyer@lne.fr, [2] National Physical Laboratory, Hampton Road, Teddington TW11 0LW, UK, maurice.cox@npl.co.uk, [3] Government Chemist Programme, LGC, Queens Road, Teddington, Middlesex TW11OLY, S.Ellison@LGCGroup.com, [4] QuoData GmbH – Quality & Statistics, Dresden, Germany, steffen.uhlig@quodata.de, [5] VSL, Thijsseweg 11, 2629 JA Delft, The Netherlands, avdveen@vsl.nl, [6] Physikalisch-Technische Bundesanstalt, Braunschweig and Berlin, Germany, gerd.wuebbeler@ptb.de

***Abstract* – Laboratory competence in measurement, including the evaluation of measurement uncertainty, is often assessed through interlaboratory comparisons, where participants report measured values alongside their associated uncertainties. Currently, the evaluation of such comparison data lacks a harmonized framework, in particular with respect to forming a consensus value, leading to disparate interpretations of results and unreliable feedback to participants. To address these gaps, standardized statistical approaches are required to ensure robust evaluation and to increase confidence in comparison outcomes. This work specifically targets proficiency testing and key comparisons with CIPM.**

I. INTRODUCTION

Interlaboratory comparisons are widely used to demonstrate laboratory competence [1], support underpinning calibration and measurement capabilities, and provide confidence in measurement results across industrial and regulatory sectors. The assessment of interlaboratory results commonly relies on statistical indicators such as degrees of equivalence and E_n scores, which compare measurement deviations with their associated uncertainties. However, standard compatibility indicators may be difficult to interpret in practice especially when uncertainties are heterogeneous, data sets are limited or show overdispersion, or correlations exist between participants' data. In such scenarios, interpretation difficulties can lead to unnecessary investigations, unwarranted revision of reported uncertainties, or superfluous comparison activities.

After reviewing key limitations of current practices with respect to evaluating comparison data with associated uncertainties and metrics for performance evaluation, the paper will explore emerging perspectives for moving from rigid compatibility indicators toward confidence-based assessments of laboratory performance and equivalence. The evolving methodology could improve the interpretability, robustness and practical value of future exercises, contributing to greater confidence in the conclusions drawn from the comparison results.

II. LIMITATIONS OF CONVENTIONAL COMPATIBILITY METRICS

Procedures to handle heterogeneity in comparison data sets are available in the key comparison literature where they are mainly applied to comparisons of NMIs (national metrology institutes) and DIs (designated institutes) [2], [3], [4]. For key comparisons organised by the CIPM (International Committee for Weights and Measures), however, no single procedure or guideline has been adopted consistently. Instead, consultative committees have adopted differing and often incompatible practices. For example, the Consultative Committee for Amount of Substance: Metrology in Chemistry and Biology (CCQM) covering chemical measurement published in 2013 a guideline [5] describing statistical approaches that incorporate participant uncertainties and several methods for handling outliers. The Consultative Committee for Photometry and Radiometry (CCPR) [6] issued a guidance note that specifies only one procedure and provides basic advice on consistency checking. These two procedures differ appreciably, and other committees also apply approaches that vary in both content and underlying methodology. It is unclear whether these differences are rooted in technical differences in the data or in traditions in the different metrological areas.

For general proficiency testing and external quality assessment, the principal standard is ISO 13528 [7]. The standard states that, when laboratories provide results that include uncertainties, estimating a consensus value becomes a complex problem for which many different approaches have been proposed. Neither ISO/IEC 17043

[8] nor ISO 13528 provides PT organizers with specific procedures for addressing this problem. ISO 13528 offers guidance on performance metrics that incorporate measurement uncertainty, such as E_n and zeta-scores, but gives little direction on how to calculate consensus values from measurement results including uncertainties.

There is accordingly an established literature offering a range of candidate approaches, but no consistent recommendation on which should be preferred. Moreover, the number and diversity of available methods continue to grow, particularly with the increasing use of Bayesian techniques for reference-value assignment [9] and for the evaluation of measurement uncertainties [10].

III. TOWARDS CONFIDENCE BASED ASSESSMENTS OF LABORATORY PERFORMANCE AND EQUIVALENCE

Several methodological developments could contribute to the probabilistic assessment of confidence in subsequent decisions. First, richer, uncertainty-aware statistical models for consensus values of interlaboratory comparisons are becoming increasingly feasible. Hierarchical and mixed-effects models provide natural frameworks for representing multiple sources of variability arising from laboratories, instruments, operators, measurement campaigns, or reference standards. Such models allow uncertainty to be propagated consistently while explicitly accounting for dependence structures that are often ignored in conventional analyses.

Second, probabilistic approaches offer opportunities to express conclusions in forms that are more directly interpretable by technical experts and decision makers. Rather than reducing comparison outcomes to binary statements of compatibility, future analyses may quantify probabilities of practical equivalence, confidence that laboratory biases remain within application-specific tolerances, or predictive probabilities describing expected performance in future comparison exercises. Such measures may be more closely aligned with the questions that laboratories, accreditation bodies, regulators and industrial users ultimately seek to answer.

Third, increasing attention is likely to be devoted to the integration of multiple sources of prior information. Interlaboratory comparisons rarely constitute isolated experiments; rather, they are embedded within broader measurement systems that include previous comparison results, routine quality control data, calibration histories, digital calibration certificates, reference material characterization, and process knowledge. Future assessment frameworks may therefore exploit this diversity of information to produce more robust and informative evaluations of laboratory performance than analyses based solely on a single comparison exercise.

Looking further ahead, confidence-based assessment naturally aligns with the ongoing digital transformation of metrology. Digital calibration certificates, machine-readable uncertainty information, digital twins, and interconnected measurement infrastructures will increasingly generate rich streams of measurement evidence. Future interlaboratory comparisons may evolve from isolated benchmarking exercises towards continuously updated assessments in which confidence is progressively refined as new evidence becomes available.

IV. CONCLUSION

This paper proposes a paradigm shift by moving from compatibility-based indicators towards confidence-based assessment of laboratory performance and equivalence in interlaboratory comparison exercises. Rather than asking whether laboratories are statistically compatible according to a predefined criterion, future methodologies may aim to quantify the confidence that available evidence supports a given scientific or operational conclusion. Such conclusions may concern laboratory equivalence, conformity with specified performance requirements, long-term measurement stability, or the reliability of calibration services for particular applications.

REFERENCES

[1] ISO/IEC 17025:2017, General requirements for the competence of testing and calibration laboratories.

[2] AMH van der Veen. Dark uncertainty in key comparisons in the gas analysis area. Metrologia, 60, 2023.

[3] MG Cox. The evaluation of key comparison data: Determining the largest consistent subset. Metrologia, 44, 2007.

[4] Wübbeler G, Bodnar O, Mickan B, and Elster C. Explanatory power of degrees of equivalence in the presence of a random instability of the common measurand. Metrologia, 52, 2015.

[5] Consultative Committee for Amount of Substance (Metrology in Chemistry and Biology) (CCQM). CCQM 13-22 guidance note: Estimation of a consensus KCRV and associated degrees of freedom, 2013.

[6] Consultative Committee for Photometry and Radiometry (CCPR). CCPR G4: Guidelines for preparing CCPR key comparisons.

[7] ISO 13528:2022, Statistical methods for use in proficiency testing by inter-laboratory comparison. 3rd ed.

[8] ISO/IEC 17043:2023. Conformity assessment – General requirements for the competence of proficiency testing providers. 2nd ed.

[9] K Shirono, M Shiro, H Tanaka, and K Ehara. Proficiency tests with uncertainty information: Detection of an unknown random effect. Measurement, 83, 2016.

[10] S Demeyer, N Fischer, and C Elster. Guidance on Bayesian uncertainty evaluation for a class of GUM measurement models. Metrologia, 58, 2020.

2026 IMEKO Joint Conference
TC7 - TC11 - TC13 - TC21 in association with TC6 and Mathmet
Faro, Portugal, 27–29 October, 2026

Distributions associated with nonlinear least-squares problems

Alistair Forbes[1]

[1]*National Physical Laboratory, Teddington, UK, alistair.forbes@gmail.com*

I. INTRODUCTION

A number of papers, e.g., [4, 5, 6], have considered the relationship between the Monte Carlo method (MCM), as described in Supplement 1, GUMS1 [2], and Supplement 2, GUMS2 [3] to the GUM [1], and Bayesian inference in the context of uncertainty evaluation. Both GUMS1 and GUMS2 can be viewed as applying to square systems in which there is exactly the same number of observations as there are parameters to be estimated. The analysis shows that the Monte Carlo method produces a sample from a distribution that coincides with a Bayesian posterior distribution [7] for a particular choice of prior. This paper continues the analysis for the case of over-determined systems in which there are more observations than parameters. We show that, in general, the sampling distributions associated with nonlinear least-squares estimation cannot be regarded as Bayesian posterior distributions.

II. NONLINEAR LEAST-SQUARES PROBLEMS

We restrict ourselves mainly to models of the form

$$\boldsymbol{y}|\boldsymbol{\alpha} \sim \mathcal{N}(\boldsymbol{\phi}(\boldsymbol{\alpha}), \sigma^2 I) \tag{1}$$

with σ^2 known. The model involves parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)^\top$ and data vector, $\boldsymbol{y} = (y_1, \ldots, y_m)^\top$, $m \geq n$. The mapping $\boldsymbol{\alpha} \mapsto \boldsymbol{\phi}(\boldsymbol{\alpha})$ defines an n-dimensional surface in $\mathcal{R}^m$, the model surface or response surface. The data vector $\boldsymbol{y}$ represents the result of measurement of the response with the system in an unknown state $\boldsymbol{\alpha}$ and the data vector is used to make inferences about or estimate the state. The model (1) defines the likelihood

$$p(\boldsymbol{y}|\boldsymbol{\alpha}) \propto \exp\left\{-\frac{1}{2\sigma^2} d^2(\boldsymbol{y}, \boldsymbol{\alpha})\right\} \tag{2}$$

involving the Euclidean distance $d(\boldsymbol{y}, \boldsymbol{a})$ from the data vector $\boldsymbol{y}$ to the model surface:

$$d^2(\boldsymbol{y}, \boldsymbol{\alpha}) = (\boldsymbol{y} - \boldsymbol{\phi}(\boldsymbol{\alpha}))^\top (\boldsymbol{y} - \boldsymbol{\phi}(\boldsymbol{\alpha})). \tag{3}$$

In Bayesian inference, the posterior distribution $p(\boldsymbol{\alpha}|\boldsymbol{y})$ is defined in terms of the likelihood and a prior $p(\boldsymbol{\alpha})$ for $\boldsymbol{\alpha}$:

$$p(\boldsymbol{\alpha}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\alpha}) p(\boldsymbol{\alpha}),$$

with $p(\boldsymbol{y}|\boldsymbol{\alpha})$ given in (2). If the non-informative prior $p(\boldsymbol{\alpha}) \propto 1$ is chosen, then this posterior distribution is defined by the distance function $d(\boldsymbol{y}, \boldsymbol{\alpha})$. The shape of the posterior distribution thus depends on the shape of the model surface $\boldsymbol{\alpha} \mapsto \boldsymbol{\phi}(\boldsymbol{\alpha})$ near $\boldsymbol{y}$ and the relationship of the data vector $\boldsymbol{y}$ to the model surface.

The least-squares estimate $\boldsymbol{a}$ is given by the $\boldsymbol{\alpha}$ that minimises $d^2(\boldsymbol{y}, \boldsymbol{\alpha})$ with respect to $\boldsymbol{\alpha}$ and corresponds to the maximum likelihood estimate of $\boldsymbol{\alpha}$. For $\boldsymbol{y}$ close to the model surface, the least-squares estimate is a function of the data vector $\boldsymbol{y}$ and we denote this function by A: $\boldsymbol{a} = A(\boldsymbol{y})$. The conditions for $\boldsymbol{a}$ to be the least-squares solution are that

$$J(\boldsymbol{a})^\top (\boldsymbol{y} - \boldsymbol{\phi}(\boldsymbol{a})) = \boldsymbol{0}, \quad J_{ij}(\boldsymbol{a}) = -\frac{\partial \phi_i}{\partial \alpha_j}(\boldsymbol{a}), \tag{4}$$

involving the Jacobian matrix $J(\boldsymbol{\alpha})$ whose columns are a basis for the tangent space to the surface at $\boldsymbol{\alpha}$.

III. SAMPLING DISTRIBUTIONS ASSOCIATED WITH LEAST-SQUARES ESTIMATES

The least-squares estimate $\boldsymbol{a}$ defines the point $\boldsymbol{\phi}(\boldsymbol{a})$ on the surface $\boldsymbol{\alpha} \mapsto \boldsymbol{\phi}(\boldsymbol{\alpha})$ closest to $\boldsymbol{y}$ and the optimality conditions state that the vector $\boldsymbol{y} - \boldsymbol{\phi}(\boldsymbol{a})$ is orthogonal to the surface at $\boldsymbol{\alpha} = \boldsymbol{a}$.

Let the $m \times (m-n)$ matrix $N(\boldsymbol{\alpha})$ define a basis for the vectors orthogonal to the surface at $\boldsymbol{\phi}(\boldsymbol{\alpha})$. The optimality condition states that

$$\boldsymbol{y} - \boldsymbol{\phi}(\boldsymbol{a}) = N(\boldsymbol{a})\boldsymbol{q} \tag{5}$$

for some unique $\boldsymbol{q}$ in $\mathcal{R}^{m-n}$. We can also regard $\boldsymbol{q}$ as a function of $\boldsymbol{y}$: $\boldsymbol{q} = B(\boldsymbol{y})$. For any $\boldsymbol{y} \in \mathcal{R}^m$ close to the model surface, the mapping $G : \mathcal{R}^m \longrightarrow \mathcal{R}^m$

$$\boldsymbol{y} \mapsto T(\boldsymbol{y}) = \begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{q} \end{bmatrix} = \begin{bmatrix} A(\boldsymbol{y}) \\ B(\boldsymbol{y}) \end{bmatrix}, \tag{6}$$

taking $\boldsymbol{y}$ to the corresponding least-squares estimate $\boldsymbol{a}$ and $\boldsymbol{q}$ defined by (5) represents a change of variables, with inverse

$$\begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{q} \end{bmatrix} \mapsto S(\boldsymbol{a}, \boldsymbol{q}) = \boldsymbol{\phi}(\boldsymbol{a}) + N(\boldsymbol{a})\boldsymbol{q}. \tag{7}$$

Any distribution for $p_Y(\boldsymbol{y})$ for $\boldsymbol{y}$ defines the joint distribution $p(\boldsymbol{a}, \boldsymbol{q})$ through the change of variable formula [8]

$$p(\boldsymbol{a}, \boldsymbol{q}) = |J_S| p_Y(\boldsymbol{\phi}(\boldsymbol{a}) + N(\boldsymbol{a})\boldsymbol{q}),$$

where J_S is the $m \times m$ matrix of partial derivatives of S evaluated at $(\boldsymbol{a}, \boldsymbol{q})$ and $|J_S|$ its determinant. In particular, if $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{y}_0, \sigma^2 I)$, then the propagated distribution is such that

$$p(\boldsymbol{a}, \boldsymbol{q}|\boldsymbol{y}_0) \propto |J_S| \exp\left\{-\frac{d^2(\boldsymbol{a}, \boldsymbol{q}, \boldsymbol{y}_0)}{2\sigma^2}\right\}, \quad (8)$$

where

$$d^2(\boldsymbol{a}, \boldsymbol{q}, \boldsymbol{y}_0) = (\boldsymbol{y}_0 - S(\boldsymbol{a}, \boldsymbol{q}))^\top (\boldsymbol{y}_0 - S(\boldsymbol{a}, \boldsymbol{q})). \quad (9)$$

The term $d^2(\boldsymbol{a}, \boldsymbol{q}, \boldsymbol{y}_0)$ is the square of the Euclidean distance between $S(\boldsymbol{a}, \boldsymbol{q})$ and $\boldsymbol{y}_0$ in $\mathcal{R}^m$. The distribution $p(\boldsymbol{a}|\boldsymbol{y}_0)$ for $\boldsymbol{a}$ alone is given by marginalisation:

$$p(\boldsymbol{a}|\boldsymbol{y}_0) = \int_{\boldsymbol{q} \in \mathcal{R}^{m-n}} p(\boldsymbol{a}, \boldsymbol{q}|\boldsymbol{y}_0) \mathrm{d}\boldsymbol{q}. \quad (10)$$

The distribution $p(\boldsymbol{a}|\boldsymbol{y}_0)$ describes the sampling distribution for $\boldsymbol{a}_q = A(\boldsymbol{y}_q)$, where $\boldsymbol{y}_q \in \mathcal{N}(\boldsymbol{y}_0, \sigma^2 I)$. The distribution $p(\boldsymbol{a}|\boldsymbol{y}_0)$ is defined for any $\boldsymbol{y}_0 \in \mathcal{R}^m$ but in practice we are only interested in $\boldsymbol{y}_0$ close to the model surface $\boldsymbol{\phi}(\boldsymbol{\alpha})$. Performing the integration in (10) is only practical if m is close to n. For other cases, we can apply the law of propagation (LPU, [1]) to a linearisation of the function T in (6) to estimate the mean and variance of the distribution $p(\boldsymbol{a}|\boldsymbol{y}_0)$ or use the Monte Carlo method (MCM).

A. *Two implementations of LPU and MCM*

Suppose data $\boldsymbol{y}$ has been observed and least-squares estimate $\boldsymbol{a}$ calculated. There are two common approaches to implementing both the LPU and MCM methods to say something about the sampling distribution associated with the estimate. The first approach, which we label A, assigns $\boldsymbol{y}_0 = \boldsymbol{\phi}(\boldsymbol{\alpha} = \boldsymbol{a})$, the least-squares estimate. Approach A, sampling $\boldsymbol{y}_q \in \mathcal{N}(\boldsymbol{\phi}(\boldsymbol{a}), \sigma^2 I)$, simulates repeating an experiment involving physical system defined by state $\boldsymbol{\alpha} = \boldsymbol{a}$ many times, gathering data vectors $\boldsymbol{y}_q$ and calculating $\boldsymbol{a}_q = A(\boldsymbol{y}_q)$. The second approach, which we label B, is to set $\boldsymbol{y}_0 = \boldsymbol{y}$, the observed data vector. The interpretation of how approach B relates to inferences about $\boldsymbol{\alpha}$ is not so straightforward as that of approach A.

In this paper, we show that for linear systems and square systems ($m = n$), approaches A and B are the same and can be interpreted in terms of Bayesian posterior distributions but for general nonlinear, over-determined systems, there is no choice of prior distributions that enable the sampling distributions associated with either approaches A and B to be interpreted as Bayesian posterior distributions. We also discuss the interpretation of approaches A and B and the relationship between the GUMS1 [2] and Bayesian approaches for over-determined ($m > n$) and square systems ($m = n$).

This work has been supported by the UK's National Measurement System programme for Data Science.

REFERENCES

[1] BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP, and OIML. Evaluation of measurement data — Guide to the expression of uncertainty in measurement. Joint Committee for Guides in Metrology, JCGM 100:2008, 2008.

[2] BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP, and OIML. Evaluation of measurement data — Supplement 1 to the "Guide to the expression of uncertainty in measurement" — Propagation of distributions using a Monte Carlo method. Joint Committee for Guides in Metrology, JCGM 101:2008, 2008.

[3] BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP, and OIML. Evaluation of measurement data — Supplement 2 to the "Guide to the expression of uncertainty in measurement" — extension to any number of output quantities. Joint Committee for Guides in Metrology, JCGM 102:2011, 2011.

[4] C. Elster. Bayesian uncertainty analysis compared with the application of the GUM and its supplements. *Metrologia*, 51:S159–S166, 2014.

[5] C. Elster, W. Wöger, and M. G. Cox. Draft GUM supplement 1 and Bayesian analysis. *Metrologia*, 44:L31–L32, 2007.

[6] A. B. Forbes and J. A. Sousa. The GUM, Bayesian inference and the observation and measurement equations. *Measurement*, 44:1422–1435, 2011.

[7] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC, Boca Raton, Fl., third edition, 2014.

[8] G. R. Grimmett and D. R. Stirzaker. *Probability and Random Processes*. Oxford University Press, third edition, 2001.

2026 IMEKO Joint Conference
TC7 - TC11 - TC13 - TC21 in association with TC6 and Mathmet
Faro, Portugal, 27–29 October, 2026

The central limit theorem and sums of t-variates

Alistair Forbes[1]

[1]*National Physical Laboratory, Teddington, UK, alistair.forbes@gmail.com*

I. INTRODUCTION

The Central Limit Theorem (CLT) [1, 2, 4] is often used, and in particular by the GUM [3, Annex G.2], to justify the association of a Gaussian distribution with a measurand on the basis that if the measurand depends on even a modest number of input/influence factors the output distribution will be approximated well be a Gaussian. The CLT guarantees only 'convergence in distribution', that is the cumulative distribution functions (CDFs) converge, but does not say much about the rate of convergence and, in fact, the convergence can be surprisingly slow for platykurtic distributions with fatter tails. The CLT does not imply that the corresponding probability density functions (PDFs) converge nor does it say anything about the asymptotic behaviour associated with the convergence. In this paper, we examine issues associated the CLT and the application of the Welch-Satterthwaite formula [6],[3, Annex G.4] to sums of t-variates.

II. SUMS OF t-VARIATES

Consider the weighted sum b_k of 2^{k-1} variates a_j, where

$$a_j \sim t_\nu(0, (\nu-2)/\nu), \quad \nu > 2, \quad k = 1, \ldots, K,$$

with $b_1 = a_1$, $b_2 = (a_1 + a_2)/\sqrt{2}$,

$$b_3 = \frac{a_1 + \cdots + a_4}{\sqrt{4}}, \quad b_4 = \frac{a_1 + \cdots + a_8}{\sqrt{8}},$$

etc. The PDFs p_k associated with the b_k are given by a sequence of convolutions which we can perform numerically using the fast Fourier transform. Since $t_\nu(0, (\nu-2)/\nu)$ has mean zero and variance one, each distribution p_k also has mean zero and variance one, a consequence of the law of propagation of uncertainty (LPU) [3, Section 3].

Figure 1 graphs the 2^{k-1} convolutions $p_k(\xi)$ of $t_\nu(0, (\nu-2)/\nu)$ for sums of 1, 4, 16 and 64 t-variates for $\nu = 3$. As k increases, the PDFs visually become closer to that of $\mathcal{N}(0, 1)$, also graphed. However, even for the convolution of 64 t-distributions, the PDF associated with the sum is visually different from the limit PDF of $\mathcal{N}(0, 1)$. We are generally used to thinking that the distribution associated with a sum of variates converges quite quickly to the target Gaussian distribution, for example, the sum of variates each associated with a uniform distribution. However, the moments associated with a uniform distribution are all finite and, in particular, a uniform distribution has small kurtosis relative to a Gaussian (leptokurtic – having slender tails). By contrast t-distributions are platykurtic, having fatter tails, relative to a Gaussian and the kurtosis associated a t-distribution is finite only if $\nu > 4$; for $\nu = 3$, as in this example, it is infinite. For such platykurtic distributions, the convergence, in the central limit sense, is much slower [5].

Figure 2 graphs the $\log_{10} p(\xi)$ of the PDFs graphed in figure 1. The asymptotic behaviour of p_k is seen to be similar for large $|\xi|$ and in fact the asymptotic behaviour of $p_k(\xi)$ is such that $p_k(\xi) \propto 1/\xi^{\nu+1}$ and is independent of k. For large ξ, $p_k(\xi) \approx p_{k-1}(\xi)/\sqrt{2}$ so the differences in the values in the tails of these distributions are modest.

III. THE WELCH-SATTERTHWAITE FORMULA

The GUM uses the Welch-Satterthwaite (WS) formula [6] to associate a t-distribution with a sum of variates. We can apply the WS approach to determine the t-distributions $p_{WS,k}$ associated with each b_k. The kth such distribution is $t_{d_k}(0, (\nu-2)/\nu)$ where the effective degrees of freedom d_k is given by $d_k = 2^{k-1}\nu$. As the number of degrees of freedom associated with a t-distribution, the PDF associated with the t-distribution converges to that of a Gaussian distribution, a consequence of the fact that $\lim_{n\to\infty} (1 + x/n)^n = e^x$. So as k increases, $p_{WS,k}$ becomes closer to a Gaussian distribution. One might think that this convergence is a consequence of the CLT. However, the limit of the distributions $p_{WS,k}$ as $k \to \infty$ is the Gaussian distribution $\mathcal{N}(0, (\nu-2)/\nu)$ whereas the CLT implies that the limit CDF is that associated with $\mathcal{N}(0, 1)$. The convergence of the distributions $p_{WS,k}$ is not related to the CLT but instead is the result of the fact that the convolution of two gamma distributions $\mathcal{G}(\alpha_1, \beta)$ and $\mathcal{G}(\alpha_2, \beta)$ is the gamma distribution $\mathcal{G}(\alpha_1 + \alpha_2, \beta)$.

Figure 3 graphs the t-distributions $p_{\mathrm{WS},k}(\xi)$ for associated with sums of 1, 4, 16 and 64 t-variates for $\nu = 3$. As k increases, the distributions become closer to $\mathcal{N}(0, (\nu-2)/\nu)$ also graphed. For sums of 16 and particularly 64 variates, the corresponding distribution $k_{\mathrm{WS},k}$ is hardly distinguishable from $\mathcal{N}(0, (\nu-1)/\nu)$, having degrees of freedom 48 and 192, respectively. The figure also graphs the PDF associated with $\mathcal{N}(0, 1)$ and is significantly different $\mathcal{N}(0, (\nu-1)/\nu)$.

IV. INTERPRETATION OF UNCERTAINTY STATEMENTS DETERMINED USING WS

The fact that the distributions $p_{WS,k}$ converge to a completely different distribution to that determined by the LPU

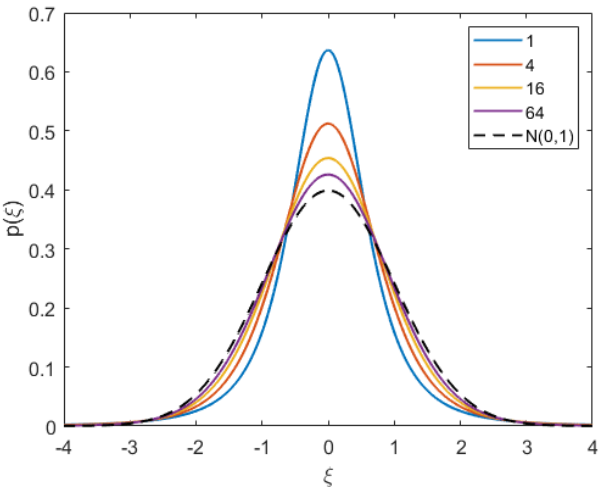


Fig. 1. Graphs of the 2^{k-1} convolutions $p_k(\xi)$ of $t_\nu(0,(\nu-2)/\nu)$ for $k = 1$, 4, 16 and 64, for $\nu = 3$. As k increases, the distributions visually become closer to $\mathcal{N}(0,1)$ also graphed.

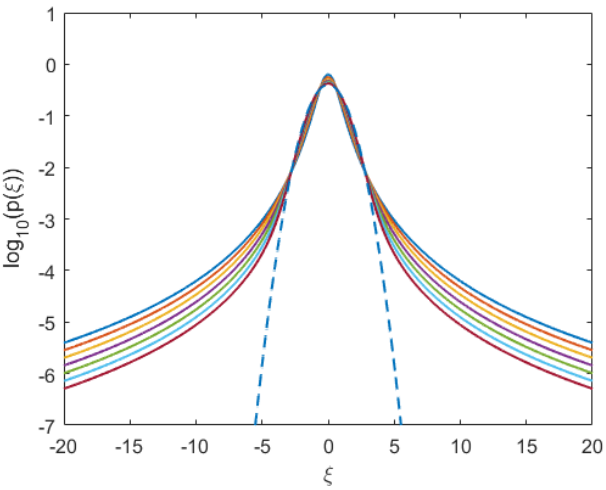


Fig. 2. Graphs of $\log_{10} p_k(\xi)$, solid, along with $\log_{10} p_{\mathcal{N}}(0,1)$ a negative quadratic, dashed. As k increases the tails of the distributions become somewhat smaller in magnitude.

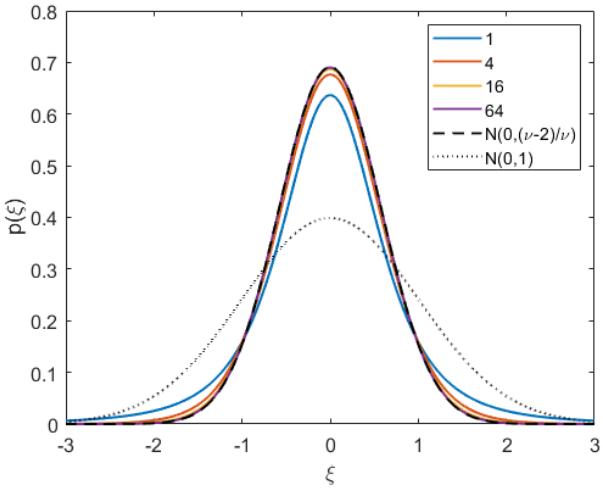


Fig. 3. Graphs of the t-distributions $p_{\mathrm{WS},k}(\xi)$ for $k = 1$, 4, 16 and 64, for $\nu = 3$. As k increases, the distributions become closer to $\mathcal{N}(0,(\nu-2)/\nu)$, also graphed. The figure also graphs $\mathcal{N}(0,1)$ for comparison.

and the CLT points to the fact that uncertainty statements that depend on the application of the WS formula cannot be interpreted in terms of the propagation of means and variances or the propagation of distributions. This paper will explore further issues associated with the application of the Welch-Satterthwaite formula, as recommended by the GUM, in uncertainty evaluation.

This work has been supported by the UK's National Measurement System programme for Data Science.

REFERENCES

[1] S. Artstein, K. M. Ball, F. Barthe, and A. Naor. On the rate of convergence in the entropic central limit theorem. *Probability Theory and Related Fields*, 129:381â–390, 2004.

[2] P. Billingsley. *Probability and Measure*. A Wiley-Interscience publication. Wiley, New York, 3rd edition, 1995.

[3] BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP, and OIML. Evaluation of measurement data — Guide to the expression of uncertainty in measurement. Joint Committee for Guides in Metrology, JCGM 100:2008, 2008.

[4] G. R. Grimmett and D. R. Stirzaker. *Probability and Random Processes*. Oxford University Press, third edition, 2001.

[5] R. M. Vogel, S. M. Papalexiou, J. Lamontagne, and F. Dolan. When heavy tails disrupt statistical inference. *The American Statistician*, 79:1–42, 09 2024.

[6] B. L. Welch. The generalization of "student's" problem when several different population variances are involved. *Biometrika*, 34:28–35, 1947.

2026 IMEKO Joint Conference
TC7 - TC11 – TC13 - TC21 in conjunction with TC6 and Mathmet
Faro, Portugal, 27-29 October, 2026

Loosely-Coupled Camera-Anchored Residual Learning for Three-Dimensional Handheld Inertial Navigation

Ahmad Barari[1], Wahab Janjuaa[2]

[1] *ahmad.barari@ontariotechu.ca*
[2] *wahab.janjuaa@ontariotechu.ca*

Abstract

The work addresses the current and technically interesting problem of inertial navigation and proposes an ambitious combination of ZUPT-enabled inertial navigation, camera-based optical referencing, and residual error correction using machine learning. Using a high-precision laser tracker to generate reference data and performing statistical comparisons with multiple baseline methods are potential strengths of the work.

However, abstract in its current form is overloaded and difficult to understand. It presents too many technical components and statements without sufficiently explaining their relationships and individual contributions to the proposed method. The authors should more clearly highlight the main methodological novelty of the work and separate it from a combination of already known approaches.

The Abstract would also be greatly enhanced by providing specific quantitative results, such as the trajectory error for the proposed method and the baseline comparison methods, the magnitude of the improvement achieved, the length of the trajectories examined, and the size of the experimental data set. The broad claim that classical visual-inertial filters are unable to account for relevant intersensory relationships needs either to be substantiated in more detail or formulated less categorically.

Additionally, the practical limitations of the optically referenced approach should be discussed, particularly the need for observable ArUco markers, as well as the ability of the trained model to generalize to previously unencountered trajectories and operating conditions.

Overall, the work appears to be potentially original and relevant, but the abstract requires significant improvement in terms of clarity of presentation and the provision of more specific evidence to support the claimed technical advantages.

2026 IMEKO Joint Conference
TC7 - TC11 – TC13 - TC21 in conjunction with TC6 and Mathmet
Faro, Portugal, 27-29 October, 2026

Controversial new definition of "Measurement Uncertainty"

Franco Pavese

Former Research Director in Metrology at CNR, Torino, Italy

The extension of the concept of uncertainty to the nominal quantities (i.e., of the qualitative type) has been one of the most important modern developments of modern metrology (and of AI applications). In nominal quantities numerical values are not measured, but a category or class is assigned. Therefore, the uncertainty is not anymore representing a numerical interval of results, but instead establish the process required to assign a "category" or "class". The uncertainty becomes the probability of wrong classification, thus requiring specific methodologies: e.g., the insufficient resolution of the measuring tool may increase the doubt about the category to be assigned. This is a correct use of the concept of "doubt". Instead, when uncertainty concerns numerical quantities, i.e. obtaining from the measuring instruments an interval of numerical values, it cannot be defined for the general definition of uncertainty in measurement ("Doubts about the position of the true value"), as recently adopted, because this expression is metrologically incorrect. The Author does not agree about the use of words such as "doubt" or "position" looking more "colloquial" than scientific. Secondly, the Author hesitates accepting the concept of "true value": science should not embrace any specific philosophical position for one of its fundamental concepts. The presentation will expand and justify the above, discussing author's uncertainty in accepting such a recent general change of the concept of measurement uncertainty outside the limited requirements concerning "nominal quantities".

THE DIGITAL METROLOGICAL EXPERT (DME)

H. Gasca-Aragon[a], D. Hutzschenreuter[a], W. El Jaoua[a], D. Urban[a] , M. Gafert[a], F. Haertig[a]*

[a] Physikalisch-Technische Bundesanstalt, Bundesallee 100, 38116 Braunschweig, Germany,

* Corresponding author

Abstract − Key comparisons guarantee that different National Metrology Institutes and Designated Institutes can produce equivalent measurement results for a specific quantity to, e.g., confirm Calibration and Measurement Capabilities. However, analysis of comparisons can be very time-consuming, error-prone and is not easily manageable by smaller metrology organizations. The Digital Metrological Expert (DME) is an open-access software tool for automated comparison evaluation with the aim to greatly facilitate this traditionally manual process. Participants of the comparison produced Digital Calibration Certificates (DCC) as report files for evaluation with the DME. The report produced by the DME contains measurement data, processed data and calculated statistical values, such as the key comparison reference value and degrees of equivalence, as well as supplemental plots. This paper covers the DME workflow based on the comparisons. Supported compared quantities up to date include the areas of mass, temperature, and length.

***Keywords*:** Digital Metrological Expert, DME, key comparison, DCC, D-SI, smart standards.

1. INTRODUCTION

Key comparisons are a core concept of the Mutal Recognition Arrangement [1] as they guarantee that different National Metrology Institutes (NMIs) and Designated Institutes (DIs) can produce equivalent measurement results for a specific quantity and thereby fostering traceability and comparability of measurements worldwide. Successful comparisons may confirm Calibration and Measurement Capabilities (CMCs) that allow NMIs and DIs to offer recognised services to support legal-metrology and strengthen end-users confidence in measurement results.

Traditional electronic "paper-based" comparison workflows and manual statistical analysis are time consuming and error-prone. While guidelines for the traditional evaluation of comparisons exist, they might not be easily applicable to real world examples, since there is no "one size fits all" approach. Furthermore, the implementation of traditional and advanced statistical models, robust handling of inconsistent data and decision making might not be straightforward and limit the participation of smaller or emerging organizations in metrology, especially in the role of the pilot laboratory.

Software like the Digital Metrological Expert (DME) developed by PTB strives to automate the evaluation of key comparisons starting with the report files of individual participants and ending in a comprehensive evaluation report. This leads to a robust and more efficient data processing as well as reproducible and transparent data analysis, enabling more NMIs and DIs to pilot comparisons. Due to this streamlined approach the duration of evaluation process is greatly shortened, leading to faster turnaround times and reduced workload for the pilot laboratory while the consistency across comparisons is improved.

The DME is an open-access software tool for comparison analysis [2, 3], that leverages existing and evolving digital standards, such as the SI Digital Framework [4] and FAIR (findable, accessible, interoperable, reusable) data principles [5]. It takes machine-readable DCCs [6] as input data and automatically computes key comparison reference values (KCRVs) and other statistical parameters. Generated results are provided as a machine-actionable DCC report also documenting the evaluation process and data flows.

A first proof-of-concept use-case was established for a virtual comparison of mass measurements by Rodiek et al. [7]. Since then, other real-world comparison had taken place: radiation temperature measurements by Müller et al. [8,9,10], Kibble balances comparison by Konrad et al [11]. Current efforts include ongoing comparison on quantities such as density, ultrasonic power and content of amount of substance.

Section 2 gives an overview of how the comparison is evaluated using DCCs in conjunction with the DME. Section 3 will cover the summary results of virtual comparisons. Section 4 will finalize with conclusions.

2. METHODS AND PROCEDURES

2.1. Utilizing Digital Calibration Certificates

Given recent developments of the DCC and the harmonized representation of measurement data following the good practice recommendation for DCCs by the German Calibration Service DKD [13]. With an increasing number of participants, that are already involved in Digitization efforts in metrology, only a limited number of measurands and a straightforward digitization of the report form. Comparisons are taking advantage of reporting the results in DCC file format instead of an analogue report. They provide ideal test cases for the evaluation of the comparison using DCCs in conjunction with the DME [7, 8,9,10,11].

DCCs are based on a highly flexible, machine-readable, XML schema [6] that allows to represent, among others, administrative data, measurement results, influence conditions and information on the calibration items. Measurement data are expressed using the D-SI metadata model [4,12] that requires at least a numerical value and a unit, optionally accompanied by the expanded uncertainty value and coverage probability.

Each measurand of the comparison is digitally represented according to the DCC schema [7, 8, 9, 10,11]. DME take advantage of recent developments in semantic vocabulary and its use by the DCC, the so called *refTypes* (most of them comply to the terms in the DKD metrology

Thesaurus [13]), as well as *ids* and *refIds*. This enables downstream software tools like the DME to easily import and process the supplied data. A comprehensive list of the used refTypes can be found in [8,13].

Participants added their data either by using a text editor or by means of the GEMIMEG tool [14]. For the administrative data the participants filled in the starting and end date of the measurements, creation date of the DCC, name and address of their institute and contact information for the person responsible for the measurements.

By reporting data this way, the DME can use its usual evaluation workflow. In some cases, additional contribution was manufactured to have a larger deviation to stimulate the outlier filter mechanism. This contribution does not belong to a specific institute.

2.2. Comparison evaluation using the Digital Metrological Expert tool

The DME extracts measurement data from DCCs, computes statistical parameters of the comparison and finally produces a DCC-like, machine-readable report. It is developed in JAVA using the Spring Boot framework for the backend and Angular realizing a browser-based web user interface [2, 3]. Internally it leverages already existing digital standards as a foundational layer (like the SI Digital Framework, expert reports and CODATA constants).

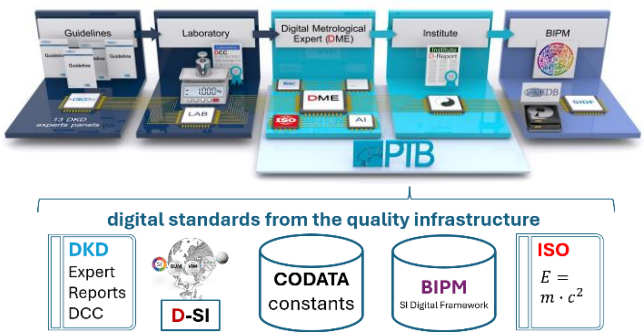


Figure 1. Schema of the DME digital workflow.

The schema of the digital workflow of the DME is shown in Figure 1. DCC reports containing the measurement data are provided by the participants in a machine accessible way via an online data storage provided as a demonstration (e.g. PTBs DCC database; in the future different storage solution will be possible). The DME accesses the reports through Application Programming Interfaces (APIs) using a dedicated Reader-Service and loads the data into its internal data model.

The comparison is evaluated using domain specific comparison analysis services. These services are implementations of key comparison evaluations for specific metrological areas and quantities. Finally, an Output Writer-Service creates the machine-readable report, containing information on all analysis steps, decisions and plots of results (see section 3.3).

A detailed schema of a comparison evaluation workflow is shown in Figure 2. Steps 1 and 2 are handled by the Input-Reader and data is identified in the DCCs by specific refTypes, ids and refIds of the DCC elements. The domain specific evaluation consists of step 3 to 6 and are run for each of the target points. First preprocessing the data if required, the reported data can be transformed by subtraction or multiplication, that are supplied to the DME (step 3). The key comparison reference value (KCRV) is determined, as documented in the technical protocol, in step 4.

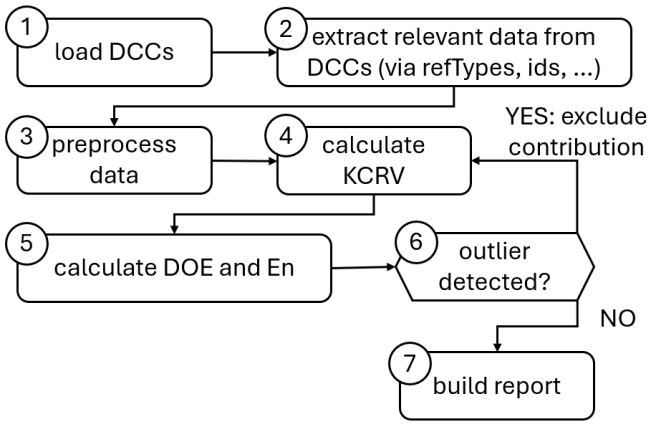


Figure 2. Evaluation workflow.

Currently available methods include deterministic weighted mean and median, and one-way random effects model implemented with Hierarchical Bayesian methods via interoperability with the NIST Decision Tree [15] if applicable or other implementations [16]. Computation continues with the unilateral and bilateral degrees of equivalence (DoE) and unilateral and bilateral E_n scores (E_n). In step 6 absolute value of the E_n are calculated. An outlier filter is applied in step 6: First the contribution with the largest unilateral E_n is identified. If the E_n is larger than one the contribution is excluded from the KCRV calculation and steps 4 to 6 are repeated. This outlier filter implements one of the hypothesis testing modules available in the DME. Note that the DoE and E_n are calculated as correlated or uncorrelated values, depending on weather a contribution is included or excluded from the KCRV calculation [17]. Once no more outliers are identified the output report will be built by the Output Writer-Service (step 7).

3. RESULTS AND DISCUSSION

The results demonstrate that the DME was successful at evaluating the comparison. Additional effort is required for generalized application, such as monitoring stability of items under comparison, linking multiple loops within a comparison, and linking to hierarchical parent comparisons.

4. CONCLUSIONS

The DME was extended to fit the needs of the evaluation of the referenced comparisons [7,8,9,10,11], demonstrated by the successful comparison respectively reported. As shown, the adoption of digital tools for the evaluation of key comparisons can greatly facilitate the analysis and reporting process with the DME. Automatic report/plot generation promises to save a significant amount of time and tedious manual labour, providing a transparent, reproducible and traceable way to document evaluation chains. A first demonstration of handling private data and secure treatment was showed during the Kibble balance doubly blind comparison in Nov 2025 at the BIPM meeting [11].

REFERENCE

2026 IMEKO Joint Conference
TC7 - TC11 – TC13 - TC21 in conjunction with TC6 and Mathmet
Faro, Portugal, 27-29 October, 2026

Self-Validating Bayesian Recalibration Interval Optimization of Mass Standards

Zeliha H. Karapınar[1], Beste Korutlu[2], Emrecan Keskin[3]

[1,2,3] *TUBITAK National Metrology Institute (UME), Kocaeli, Türkiye,*
zeliha.karapinar@aday.tubitak.gov.tr, beste.korutlu@tubitak.gov.tr, keskin26@itu.edu.tr

I. INTRODUCTION

Reference mass standards are recalibrated at predefined intervals. The interval is initially one year for newly acquired standards and may be extended after long-term stability is demonstrated. This approach is inefficient in both directions: stable standards are recalibrated more often than necessary, while degrading ones may exceed acceptance limits between recalibrations. ISO/IEC 17025 requires laboratories to ensure measurement standard validity, and ILAC-G24 provides guidance for establishing and reviewing calibration intervals. Rigorous, traceable and auditable decision-support can address this problem.

We report the decision core of a physics-informed machine learning (PIML) framework [1], where deterministic metrological knowledge (mass calibration measurement model) is separated from data-driven temporal drift and residual modelling. This separation keeps the metrological model explicit and auditable. The drift model and recalibration decision logic are validated using 10 years of archived calibration data.

The dataset comprises two OIML R111-1 accuracy classes, E_1 and E_2, each containing nineteen mass standards from 1 mg to 1 kg with ten annual calibration results (2016–2025), with synthetic cases introduced to represent specific drift behaviours under controlled conditions. Each record includes conventional mass deviation, repeatability standard deviation, expanded uncertainty (U), and maximum permissible error (δm). Selected time series were synthetically perturbed to evaluate different decision scenarios.

II. METHOD

A. Bayesian drift model

For each standard the conventional mass deviation Δ_{mc} at year t is modelled as a Gaussian random error with the standard uncertainty, where U is the expanded uncertainty ($k = 2$) reported in the calibration certificate $\mu(t) = a + bt$ or $\mu(t) = a + b + ct^2$s: The analytic posterior distribution enables model selection using the Bayes factor [2]. However, diffuse priors may suffer from the Lindley–Bartlett paradox and over-penalize the quadratic model. Therefore, a Zellner g-prior [3] is adopted with $P_0 = X^{\mathrm{T}}WX/g, g = n^2$, where X and W are the design and uncertainty-weight matrices, respectively. This reduces sensitivity to the paradox while causing only approximately 1% shrinkage of the posterior mean for $n = 10$. The quadratic model is selected only when its maximum deviation from the linear model reaches approximately $3u$ and $BF > 10$, indicating strong evidence according to Jeffreys [2,4].

B. Trustworthiness gates

Gate 1, balance health. Gate 1 is evaluated first, because subsequent decisions rely on valid weighing results. The standard deviation of the weighing difference between the test and the reference weights must satisfy

$$0.3d \leq \sigma \leq 3d \tag{1}$$

where d is the comparator scale interval. The lower bound follows from the quantisation limit ($d/\sqrt{12}$, approximated as $0.3d$), while the upper bound reflects accepted balance suitability. Because gradual degradation may remain within this range, repeatability exceeding twice the historical median for the same comparator indicates deteriorating balance performance. Grouping such events by calibration year distinguishes systemic from isolated faults.

Gate 2, model adequacy. The fitted model is accepted if the reduced chi-square satisfies $\chi_r^2 \leq 2$. For $\nu = 8$ degree of freedom (ten calibration points minus two fitted parameters), this threshold corresponds to a false-rejection probability of about 4%, consistent with Gate 3. Because residuals are normalized by the reported standard uncertainty u, adequacy is assessed relative to measurement uncertainty rather than absolute residuals.

Gate 3, self-validation. Gate 3 evaluates predictive performance using the normalized prediction error

$$z = \frac{y_{obs} - x_*^T \mu}{\sqrt{x_*^T \sum x_* + u_{obs}^2}} \tag{2}$$

where y_{obs} is the observed conventional mass deviation, x_* the design vector at the prediction time, μ and Σ the posterior mean and covariance, and u_{obs} the standard uncertainty of the observed calibration result. For a well-calibrated model, $z \sim N(0,1)$; therefore, about 4.6% of predictions exceed $z > 2$ by chance. Interval extension is rejected only when $z > 3$ or when the observed number of exceedances is significantly greater than expected according to a one-sided binomial test ($\alpha = 0.05$). Gates 2

and 3 are complementary: a good fit does not guarantee good predictive performance.

C. Risk index and decision rule

The predicted mass deviation and its total uncertainty are normalized by the maximum permissible error (MPE)

$$R(t) = \frac{|\Delta_{md}(t)| + 2u_{total}(t)}{U_{max}} \quad (3)$$

Where $\Delta_{md}(t)$ the predicted mass drift. Recalibration decisions are made after satisfying all three validation gates and are based on $R(t)$. Standards with $R(t) \geq 1$ are recalibrated immediately, while those with $0.7 \leq R(t) < 1$ remain under monitoring. Otherwise, the calibration interval is extended by one ILAC-G24 step, up to twice the current interval or three years, provided that $R(t) < 0.7$ during extension. The predicted drift correction is applied and its uncertainty is included in the uncertainty budget. Thus, recalibration is driven by uncertainty growth rather than predictable deterministic drift.

D. Validation Protocol

Random splitting and leave-one-out cross-validation are unsuitable for calibration time series because they introduce future information or evaluate interpolation rather than the extrapolation required for interval decisions. Therefore, rolling-origin validation was adopted: models trained from 2016 to each origin Y predicted $Y+1$ and $Y+2$. Validation started after six calibration results, yielding 133 evaluations per class. The two-year horizon corresponds to the maximum interval extension, and z-scores were used in Gate 3.

III. RESULTS

Rolling-origin validation resulted in ($k = 2$) coverage values of 95% and 89% for E_2 at one- and two-year horizons, respectively, and 95% and 93% for E_1. After applying all three validation gates, coverage increased to 100% for E_2 and 98% for E_1. Except for one E_1 case, all predictions outside the coverage interval corresponded to series reject by at least one validation gate, demonstrating the effectiveness of the proposed filtering. Table 1 summarizes the validation gate outcomes, and Fig. 1 presents representative examples. Gate 1 detected all injected balance faults, including repeatability deviations and quantisation-limit violations. The 50 g series was selected as quadratic and qualified for interval extension because its predicted drift approached zero within two-years, keeping the risk index below the warning threshold. In contrast, the 10 g quadratic series was rejected by Gate 3 because increasing curvature caused three rolling forecasts to exceed the ($|z| = 2$) limit. Thus interval extension depends on demonstrated predictive performance rather than the selected model class. The oscillating series further illustrates the uncertainty-relative nature of the framework. In E_2, the observed variation remained below the uncertainty level (approximately 40 µg versus 135 µg uncertainty) resulting in a reduced $\chi_r^2 \leq 2 = 0.91$, and acceptance by all gates. In E_1, the same variation exceeded the lower uncertainty level, resulting in rejection with a reduced $\chi_r^2 \leq 2 = 43{,}7$. Thus, identical physical behaviour can lead to different decisions because acceptance is evaluated relative to measurement uncertainty rather than absolute residual magnitude.

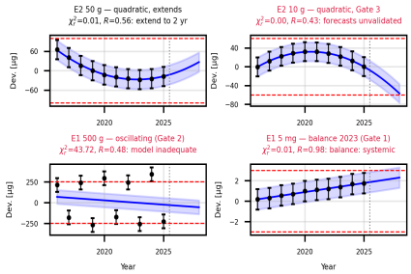


Fig.1. Representative examples of accepted and rejected recalibration decisions.

Table 1. Closed-loop decisions in both classes; every mechanism exercised at least once.

Series	Class	Model	χ_r^2	Governing mechanism	Decision
Stable / mild / long-term drifts (13)	E2	linear	≤ 0.15	all gates pass, $R \leq 0.60$	extend to 2 yr
50 g recovering, returns to zero	E2	**quadratic**	0.01	model selection + low risk	**extend to 2 yr**
10 g recovering, growing curvature	E2	**quadratic**	0.00	Gate 3, 3/7 $\lvert z\rvert > 2$	revert to default
20 g regime change	E2	linear	1.30	Gate 3, 3/7 $\lvert z\rvert > 2$, $p = 0.003$	revert to default
1 g / 2 g injected outliers	E2	lin. / quad.	18.0 / 3.7	Gate 2, χ_r^2	revert / recalibrate
100 g STD above ceiling	E2	linear	0.00	Gate 1, corridor ceiling	revert; check balance
500 g oscillating (sub-u), 5 g	E2	linear	≤ 0.91	all gates pass, $R \approx 0.7$	keep current
500 g oscillating, same series	E1	linear	43.7	Gate 2, χ_r^2	revert to default
Balance faults: mg range 2023, 10 g 2021, 200 g	E1	—	—	Gate 1: systemic / isolated / floor	revert; investigate
200 mg out-of-tolerance, 100 g step	E1	linear	≤ 1.3	risk, $R \approx 1.79$ / 1.40	recalibrate now
Near tolerance (4), capped (2), 1 kg	E1	linear	≤ 1.0	$0.7 \leq R < 1$, $R_{+2} \geq 0.7$, pass	monitor, keep, extend

IV. CONCLUSION

A Bayesian, uncertainty-aware decision module for recalibration interval optimization within a PIML framework was validated using ten years of calibration data from two OIML accuracy classes. Under the proposed rule, 13 of 19 E_2 standards qualified for a two-year interval, while the more stringent E_1 tolerances left remaining extensions withheld. The results reveal the current limitation: validation gates assess model adequacy only relative to measurement uncertainty and cannot exploit systematic residual behaviour below this level. This information will be addressed by the residual-learning component of the complete PIML framework. Future work will develop per-standard digital twins to improve interval optimization while preserving metrological traceability and uncertainty evaluation.

REFERENCES

[1] G. E. Karniadakis et al., "Physics-informed machine learning", Nat. Rev. Phys., vol. 3, pp. 422--440, 2021.

[2] R. E. Kass, A. E. Raftery, "Bayes factors", J. Am. Stat. Assoc., vol. 90, pp. 773--795, 1995.

[3] A. Zellner, "On assessing prior distributions and Bayesian regression analysis with g-prior distributions", Bayesian Inference and Decision .Techniques, 1986.

[4] H. Jeffreys, "Theory of Probability", 3rd ed., Oxford University Press, Oxford, 196.

2026 IMEKO Joint Conference
TC7 - TC11 - TC13 - TC21 in association with TC6 and Mathmet
Faro, Portugal, 27–29 October, 2026

Metrological Validation in Medical AI: Using Counterfactuals to Isolate Sensor-Induced Shortcut Learning

Tameem Adel[1]

[1]*National Physical Laboratory (NPL), United Kingdom, tameem.adel@npl.co.uk*

***Abstract* – Machine learning (ML) models in healthcare frequently suffer from shortcut learning, where algorithms overfit to spurious correlations rather than true physiological signals. In medical imaging, the most pervasive shortcuts are fundamentally metrological: Models learn to detect the specific instrument, sensor noise profile, or calibration artifact rather than the underlying pathology. This case study introduces a counterfactual explanation framework to audit ML models for measurement-induced biases. By generating synthetic counterfactuals that isolate the metrological footprint of the acquisition device, we can quantify the exact degree to which an algorithm relies on measurement artifacts, bridging metrological traceability with Responsible artificial intelligence (AI).**

I. THE METROLOGICAL FAILURE MODE: INSTRUMENT AS A SHORTCUT

Shortcut learning [1] is a failure mode where an ML model exploits unintended, superficial correlations in a training dataset to solve a task, rather than learning the actual underlying semantics or causal relationships. For example, models diagnosing pneumonia from chest X-rays often learn to distinguish between portable ICU scanners and fixed outpatient scanners, using the portable sensor's noise profile as a proxy for severe illness [2].

Fundamentally, this phenomenon represents a profound failure of measurement invariance. Measurement invariance is the statistical property ensuring that an ML model, used for diagnosis, extracts clinical features from medical data in the exact same way, regardless of the hardware, manufacturer, or acquisition protocol used to capture that data. When measurement invariance is lacking, the ML model exploits vendor-specific variations in the training data as predictive shortcuts. Rather than learning the complex, invariant biological representations of a disease, the ML model learns to identify the machine, the clinic, or the scanning protocol, mistakenly treating these hardware fingerprints as diagnostic criteria.

When an ML model used for medical diagnosis relies on scanner-specific artifacts, it is effectively evaluating the measuring instrument rather than the biological subject. In metrology, a measurement system is expected to yield invariant results regardless of the specific apparatus used, provided the devices are properly calibrated. In medical AI, however, the high-dimensional capacity of deep neural networks (DNNs) allows the model to weaponise these subtle metrological discrepancies. In such a case, the ML model conflates the physics of the hardware detector with the pathophysiology of the patient, leading to artificially inflated performance metrics during internal validation.

Consequently, there is an urgent need for rigorous auditing frameworks that treat algorithmic validation as a metrological exercise. We must evaluate model bias by characterising its mistakes specifically under metrological distribution shifts. Standard cross-validation fails to expose this vulnerability because the spurious correlation between the sensor artifact and the clinical label is preserved across random data splits. To ensure algorithm robustness, auditing methodologies must force the model to confront the isolated variable of the instrument itself, deliberately decoupling the sensor's physical signature from the diagnostic target.

II. METHODOLOGY: GENERATING METROLOGICAL COUNTERFACTUALS

To quantify the degree of reliance of an ML model on sensor-induced shortcuts, we ask a counterfactual question: What would the algorithm predict if this exact same patient had been scanned by a different machine?

Because it is clinically impossible to subject a patient to simultaneous, identical scans across multiple disparate machines at the exact same moment in time, we must generate these counterfactuals synthetically. This methodology relies on decoupling the image into two latent spaces: the physiological signal (the patient's anatomy and pathology) and the metrological footprint (e.g. the sensor's noise profile, contrast signature, and calibration artifacts).

A. *Counterfactual Generation Framework*

We utilise an unpaired image-to-image translation architecture, specifically a Cycle-Consistent Generative Adversarial Network (CycleGAN) [3]. Let X_A denote the domain of images acquired by Sensor A (e.g. a portable ICU

X-ray machine), and X_B represent images from Sensor B (e.g. a high-fidelity fixed clinical scanner).

The framework learns a mapping function $G : X_A \rightarrow X_B$ that translates the sensor characteristics while preserving the underlying anatomical structure. For a given original image $x \in X_A$ with a known clinical label y, the generator creates a synthetic counterfactual x'. This counterfactual x' contains the exact same biological structures as x, but possesses the metrological footprint of Sensor B.

The Auditing Metric: Metrological Reliance Score Let $f(x)$ be the probabilistic prediction of our diagnostic ML model, for a given input image x. By passing both the original image and its counterfactual through the model, we calculate the Metrological Reliance (MR):

$$\text{MR} = |f(x) - f(x')| \tag{1}$$

A highly robust, biologically grounded model will yield an MR close to 0, indicating measurement invariance. In contrast, a model suffering from shortcut learning will exhibit a high MR, demonstrating that shifting the sensor's footprint drastically alters the diagnostic prediction, even when the underlying biology remains identical.

III. EXPERIMENTAL SETUP

To empirically demonstrate this framework, we construct a controlled experiment targeting the classic "ICU vs. Outpatient" pneumonia shortcut.

A. *Baseline Model Training*

We train a standard convolutional neural network (e.g. ResNet-50) to classify images as Normal or Pneumonia. The training data inherently contains a metrological bias: severe pneumonia cases are predominantly sourced from portable, low-resolution X-ray machines (representing bedridden ICU patients), while normal cases are predominantly sourced from high-resolution fixed scanners.

B. *Counterfactual Generation*

Using a subset of the data, we train a CycleGAN [3] to learn the translation between the Portable Sensor Domain and the Fixed Sensor Domain. We then generate a test set of metrological counterfactuals. Every portable X-ray of a pneumonia patient is translated to look as though it was taken on a fixed scanner, and every normal fixed X-ray is translated to look as though it was taken on a portable scanner.

C. *Auditing and Evaluation*

The baseline classifier is evaluated on both the original test set and the counterfactual test set. We analyse the drop in the Area Under the Receiver Operating Characteristic Curve (AUC-ROC).

- **H**ypothesis: If the model properly learnt true physiological markers of pneumonia (e.g. opacity, consolidation), its performance on the counterfactual set will remain stable.
- **E**xpected Failure Mode: If the model learnt to use the portable sensor's noise profile as a proxy for disease, it will misclassify the counterfactuals. It will predict "Normal" for sick patients whose images have been artificially cleaned to look like fixed-scanner images, and "Pneumonia" for healthy patients whose images have been degraded to mimic portable scanners.

IV. DISCUSSION: BRIDGING TRUSTWORTHY AI AND MEASUREMENT SCIENCE

The introduction of metrological counterfactuals offers a necessary paradigm shift in how we validate medical AI. Current regulatory frameworks and validation standards heavily emphasise the diversity of patient demographics (age, sex, ethnicity) to ensure algorithmic fairness. However, they frequently overlook the metrological diversity of the hardware generating the data.

Redefining Traceability in AI In traditional metrology, traceability ensures that a measurement result can be related to a reference through a documented unbroken chain of calibrations. In medical AI, we must establish algorithmic traceability: Proving that a model's output traces back to biological reality, not to the idiosyncrasies of the data acquisition pipeline. The Metrological Reliance (MR) score provides a quantifiable metric for this traceability, serving as a boundary for the model's uncertainty under real-world clinical shifts.

Implications for Deployment Hospitals currently struggle when ML models that performed exceptionally well during retrospective validation fail abruptly when deployed in a new hospital ward or when scanner hardware is upgraded. By incorporating counterfactual sensor auditing into the ML lifecycle, developers can identify and penalise shortcut learning before deployment. This bridges the gap between the computational mechanics of deep learning and the rigorous standards of measurement science, ultimately fostering trustworthy AI systems that are resilient, scientifically valid, and safe for patient care.

References

[1] R.Gierhos, J.Jacobsen, C.Michaelis, R.Zemel, W.Brendel, M.Bethge, F.Wichmann, "Shortcut learning in deep neural networks", Nature Machine Intelligence, vol.2, No.11, 2020, pp.665-673.

[2] J.Zech, M.Badgeley, M.Liu, A.Costa, J.Titano, E.Oermann, "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study", PLoS medicine, vol.15, No.11, 2018.

[3] M.Zhu, S.Gong, Z.Qian, L.Zhang, "A brief review on cycle generative adversarial networks", IIAE international conference on intelligent systems and image processing (ICISIP), 2019, pp.235-242.

2026 IMEKO Joint Conference
TC7 - TC11 - TC13 - TC21 in association with TC6 and Mathmet
Faro, Portugal, 27–29 October, 2026

Geometry-Resolution Effects in 3-D Finite-Element Simulations of Twisted-Conductor Loss

Soheil Ahmadi [1], S. H. Khan[1], K. T. V. Grattan[1]

[1]*School of Science & Technology, City St George's, University of London, London EC1V 0HB, UK*

soheil.ahmadi@citystgeorges.ac.uk; s.h.khan@citystgeorges.ac.uk; k.t.v.grattan@citystgeorges.ac.uk

***Abstract* – Automated optimisation can converge on data from a geometrically biased simulation. This paper separates solid-model resolution, finite-element resolution and post-processing in a three-dimensional twisted-conductor loss model. One turn is assembled from N straight rotational-extrusion increments before meshing. A derived volume factor, $1-(2/3)\sin^2(\pi/N)$, predicts the 2.537349% $N = 16$ twisted-to-parallel deficit within the stored volume precision. With a fixed 4 mm element-size control, increasing N from 12 to 16 changed 50 Hz loss by 1.056–1.058% for the tested 2, 3 and 4 m models, while elapsed time rose by 9.8–16.5% for the timed cases. A separate mesh-size sweep spanned 0.245%. Transverse-cut calculations are consistent only when strand-normal and transverse area–length bases are paired. The effective length cancels in the paired $R_{\mathrm{ac}}/R_{\mathrm{dc}}$ ratio, but inferring it from DC loss makes the unity DC check self-constrained. Independent geometry, field and post-processing checks are therefore required before simulation outputs are used in automated optimisation.**

I. MOTIVATION AND METHOD

Optimisation of expensive black-box simulations assumes that each evaluation is a faithful objective value [1]. Verification must also test the domain and reported measurand, not only solver convergence [2, 3]. This is especially relevant to helical conductors, for which several analytical, explicit 3-D and coordinate-transformed formulations exist [4, 5, 6]. Curved-domain approximation can influence Maxwell solutions independently of the field mesh [7, 8].

The model has nominal copper area $A_0 = 2711$ mm^2 and five insulated 85-strand segments (425 regular-hexagonal copper strands in total); the dimensional drawing gives a nominal strand side length $s = 1.567$ mm. It contains one turn, so axial length L_z equals lay length λ. A 72° one-fifth cross-section is imported in cylindrical coordinates. For

$$\Delta\theta = 360^\circ/N, \qquad \Delta z = L_z/N, \tag{1}$$

The ANSYS Mechanical APDL commands `VEXT` and `VGEN` create the first volume and N total increment positions; the volumes are united, partitioned and copied through five 72° rotations [9, 10]. ANSYS specifies that `VEXT` extrudes straight between the end faces. Only then is the domain discretised with SOLID237 finite elements (FE) (Fig. 1). Copper and the numerical air region use relative permeability $\mu_r = 1$, with $\rho_{\mathrm{Cu}} = 1.7241 \times 10^{-8}$ and $\rho_{\mathrm{air}} = 1.0\times10^{10}\ \Omega\,\mathrm{m}$. All runs use 10 A root-mean-square (RMS) total; the matched N study imposes 2 A RMS on each segment. Consequently, N counts geometry divisions per turn, h is the nominal element-size control and n_e is the resulting total conductor-plus-air FE count.

The central void means that every conductor strand lies at $r > 0$. For the intended smooth helix, a transverse terminal plane cuts each strand obliquely, with $\tan\beta = 2\pi r/\lambda$, $L_c = L_z \sec\beta$ and $A_z = A_\perp \sec\beta$. Obliquity is not intrinsically an error because $A_z L_z = A_\perp L_c$; the transverse and strand-normal area–length conventions must instead be declared and paired consistently.

For summed copper-element volume V_h, geometry closure is measured by

$$A_{\mathrm{eq}} = V_h/L_z, \qquad \delta_A = 100(A_{\mathrm{eq}}/A_0 - 1). \tag{2}$$

For total loss P_f, RMS current I and a declared length L,

$$r_f^{(L)} = \frac{P_f}{I^2 L}, \qquad \frac{r_{50}^{(L)}}{r_0^{(L)}} = \frac{P_{50}}{P_0}. \tag{3}$$

The common length cancels from the paired ratio but not from W/m or Ω/m.

II. RESULTS AND IMPLICATIONS

At fixed $h = 4$ mm, increasing N from 12 to 16 reduced P_{50} by 1.0562%, 1.0579% and 1.0579% for $L_z =$ 2, 3 and 4 m, respectively. Because each solid was remeshed, this is the response to N at fixed mesh control, not an isolated geometry-error component. For the 2 m case, P_0 decreased from 1.338519 to 1.311707 mW, while $R_{\mathrm{ac}}/R_{\mathrm{dc}} = P_{50}/P_0$ increased from 1.698522 to 1.714935 (+0.9663% relative). For the timed 3 and 4 m jobs (48 distributed-memory processes), Table 1 shows that the change raised elapsed time by 9.8% and 16.5%; ANSYS PREP7 preprocessing rose by 25.6% and 29.6%, although n_e rose only 0.15% and 2.15%. The additional cost was therefore not captured by FE count alone. These single-job timings are indicative, not a hardware-scaling benchmark; percentages use the unrounded solver times.

Across eight $N = 16$ lengths, $A_{\mathrm{eq}} = 2642.4130$ mm^2 to within 10^{-5} mm^2, giving $\delta_A = -2.5300\%$. Let $V_\parallel$

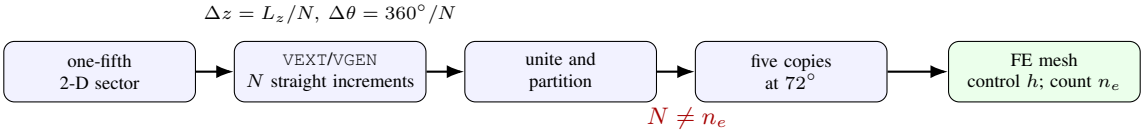


Fig. 1. Piecewise twist construction followed by independent FE meshing. $N = 12$ and 16 mean 30° and 22.5° geometry increments, not 12 or 16 elements.

Table 1. Response to geometry resolution and recorded cost at $h = 4$ mm.

L_z (m)	N	n_e (million)	P_{50} (mW)	PREP7 (min)	Elapsed (min)
3	12	6.744	3.404482	26.4	59.2
3	16	6.754	3.368465	33.2	65.0
4	12	8.966	4.536718	37.2	91.5
4	16	9.159	4.488724	48.2	106.6

denote the summed element volume of the like-for-like parallel control, whose equivalent area is 2711.2058 mm^2 ($+0.0076\%$). For straight interpolation between faces rotated by $2\pi/N$, integration through one increment gives $V_N/V_\parallel = 1 - (2/3)\sin^2(\pi/N)$. The measured and predicted $N = 16$ deficits relative to the control both round to -2.5373489%, agreeing within the stored volume precision; this numerical identity is not an uncertainty estimate. A separate $N = 16$, 4 m sweep from $h = 2$ to 10 mm changed P_{50} by 0.245%; the two perturbations are not added as uncertainty components.

For every DC case, the processing length satisfies

$$L_{\text{eff}} = \frac{P_0}{I^2(\rho/A_0)}. \tag{4}$$

The resulting unity DC check is imposed algebraically, while Eq. (3) becomes P_{50}/P_0. Across the 14 sampled $N = 16$ pitches, the paired ratio increases monotonically from 1.323948 at 0.1 m to 1.717841 at 5 m, whereas axial AC resistance decreases monotonically from 22.6308 to 11.2190 $\mu\Omega$/m [5]. The two declared objectives therefore select opposite sampled endpoints.

A pre-optimisation gate should require independent volume, transverse/strand-normal area and topology checks; refinement in both N and h; dimensional tests of post-processing; explicit axial/path length conventions; and stable candidate ranking over accepted numerical and input variations. Future work should use a crossed $N \times h$ design and compare the piecewise solid with an exact CAD or helicoidal-coordinate reference.

III. CONCLUSION

Geometry construction, FE discretisation and post-processing are separate parts of the simulation measurement model. Here, transverse-cut calculations are consistent only when area and length bases are paired, and the straight-generator factor explains the measured $N = 16$ volume deficit within the stored numerical precision. Increasing N from 12 to 16 produced a consistent 1.056–1.058% loss response for the tested 2, 3 and 4 m cases, and the 2 m paired ratio increased by 0.966%. The normalisation made the DC check non-independent. All three stages must therefore pass explicit verification criteria before their outputs are used for optimisation.

IV. REFERENCES

[1] Donald R. Jones, Matthias Schonlau, and William J. Welch. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13(4):455–492, 1998.

[2] Christopher J. Roy and William L. Oberkampf. A comprehensive framework for verification, validation, and uncertainty quantification in scientific computing. *Computer Methods in Applied Mechanics and Engineering*, 200(25–28):2131–2144, 2011.

[3] Joint Committee for Guides in Metrology. Guide to the expression of uncertainty in measurement—part 6: Developing and using measurement models. Technical Report JCGM GUM-6:2020, Joint Committee for Guides in Metrology, 2020.

[4] Emilio Plumed, Ignacio Lope, Claudio Carretero, and Jesús Acero. A recursive methodology for modelling multi-stranded wires with multilevel helix structure. *Applied Mathematical Modelling*, 83:76–89, 2020.

[5] Soheil Ahmadi, Sanowar H. Khan, and Kenneth T. V. Grattan. Impact of twisting on skin and proximity losses in segmented underground cables: A 3D finite-element study. *Applied Sciences*, 15(5):2814, 2025.

[6] Albert Piwonski, Julien Dular, Rodrigo Silva Rezende, and Rolf Schuhmann. 2-D eddy current boundary value problems for power cables with helicoidal symmetry. *IEEE Transactions on Magnetics*, 59(5):1–4, 2023. Art. no. 6300204.

[7] Rubén Aylwin and Carlos Jerez-Hanckes. Finite-element domain approximation for maxwell variational problems on curved domains. *SIAM Journal on Numerical Analysis*, 61(3):1139–1171, 2023.

[8] Shingo Hiruma. Extended finite element method enhanced by exact boundary representation for analysis of eddy currents. *IEEE Transactions on Magnetics*, 60(3):1–4, 2024. Art. no. 7400704.

[9] Ansys, Inc. *VEXT: Generates Additional Volumes by Extruding Areas*. Ansys Mechanical APDL Command Reference, Release 2025 R2, 2025. Accessed 16 August 2026.

[10] Ansys, Inc. *VGEN: Generates Additional Volumes from a Pattern of Volumes*. Ansys Mechanical APDL Command Reference, Release 2025 R2, 2025. Accessed 16 August 2026.

2026 IMEKO Joint Conference
TC7 - TC11 - TC13 - TC21 in association with TC6 and Mathmet
Faro, Portugal, 27–29 October, 2026

Extending Confidence Calibration to Generalised Measures of Variation

Vivek Desai[1] and Andrew Thompson[1]

[1]*Data Science and AI Department, National Physical Laboratory, Hampton Road, Teddington, TW11 0LW, United Kingdom*

I. INTRODUCTION

Reliable uncertainty quantification is increasingly recognised as a key requirement for trustworthy machine learning (ML) systems. For classification tasks, uncertainty is typically represented through class probability distributions. A classifier is said to be *calibrated* if its probabilities match the empirical frequency of outcomes. While many classifiers produce probability-like outputs, these values are often poorly calibrated and therefore cannot always be interpreted as meaningful statements of uncertainty. Indeed, it has been shown that deep learning models exhibit overconfidence [1], and that their predictions often require recalibrating with post-hoc methods.

Most existing calibration metrics focus on confidence calibration, where confidence is defined as the maximum predicted probability. The most widely used metric is the Expected Calibration Error (ECE) [1], which measures the agreement between predicted confidence and observed accuracy. Although ECE has become a standard tool for uncertainty evaluation, it relies exclusively on the largest probability and ignores the remainder of the predictive probability distribution.

In this work, we introduce the Variation Calibration Error (VCE), a general framework that extends confidence calibration to arbitrary measures of variation defined on probability distributions. The framework enables calibration assessment using uncertainty measures that utilise the entire probability distribution, including entropy, while retaining confidence calibration as a special case.

II. BACKGROUND

Confidence calibration evaluates whether the confidence values produced by a classifier correspond to the observed frequency with which predictions are correct. Metrics such as ECE achieve this by grouping predictions according to confidence and comparing average confidence with empirical accuracy within each group.

For multiclass classification, however, confidence captures only a limited aspect of predictive uncertainty. Alternative measures of variation have been proposed to quantify the spread or concentration of a probability distribution, including Shannon entropy, Wilcox's Variation Ratio (WVR), and the Index of Qualitative Variation (IQV). For further details on the WVR or IQV measures, we refer the reader to [2]. Unlike confidence, these quantities take into account the entire probability mass distribution rather than only the most probable class.

Previous work proposed the Uncertainty Calibration Error (UCE) [3], which evaluates calibration by comparing predicted entropy values with observed misclassification rates. However, the relationship between entropy and classification error is not necessarily direct, even for a perfectly calibrated classifier. This raises the question of whether entropy calibration should instead be defined through a framework that more closely parallels the probabilistic interpretation underlying confidence calibration.

III. METHODS

Here we present the formulation of the VCE metric and illustrate how it can be constructed as a multiclass extension of the ECE metric. The broad approach we adopt to define the VCE could also be applied to other calibration metrics with different binning schemes.

Given N samples indexed by $i = 1, \ldots, N$, we write $\hat{p}_i$ for the largest probability value for sample i. To extend the ECE to generalised measures of variation, we choose a measure of variation $\mathcal{V}$. Given number of bins M, we define the set B_m to be $B_m := \left\{ i : \mathcal{V}(p_i) \in \left(\frac{m-1}{M}, \frac{m}{M}\right] \right\}$. Given C classes, we write q_i for the length-C vector consisting of the reverse-ordered probabilities for sample i. Define, for any non-empty set $B \subseteq \{1, \ldots, N\}$,

$$\mathcal{P}(B) := \mathcal{V}\left(\frac{1}{|B|} \sum_{i \in B} q_i \right). \tag{1}$$

For sample i, we write r_i for the length-C vector whose c^{th} entry is $(r_i)_c := \mathbf{1}(y_i = k_{ic})$, where k_{ic} is the class corresponding to the cth entry of q_i, and then define, for any non-empty set $B \subseteq \{1, \ldots, N\}$,

$$\mathcal{O}(B) := \mathcal{V}\left(\frac{1}{|B|} \sum_{i \in B} r_i \right). \tag{2}$$

Here $\mathcal{P}$ and $\mathcal{O}$ stand for *predicted variation* and *observed variation* respectively. Analogous to the ECE, the VCE is then given by

$$\text{VCE} := \sum_{m=1}^{M} \frac{|B_m|}{N} |\mathcal{O}(B_m) - \mathcal{P}(B_m)|. \tag{3}$$

It is straightforward to check that the VCE reduces to the ECE when the measure of variation is chosen to be confidence.

To test the metric behaviour, we generate predictions for a C-class classification task which are perfectly calibrated by design. For each sample $i \in \{1, \ldots, N\}$, we sample a predicted probability vector

$$p_i = [(p_i)_1 \cdots (p_i)_C]^\mathsf{T} \tag{4}$$

from the same Dirichlet distribution with parameters $\alpha = (\alpha_1, \ldots, \alpha_C)$. The α parameters determine the relative frequency of the classes. For each sample $i \in \{1, \ldots, N\}$, we then use these predicted probabilities to draw a "true" class; in other words, $y_i = c$ with probability $(p_i)_c$.

For each experiment, characterised by a choice of α, we calculate ECE [1], UCE [3] and VCE with normalised Shannon entropy as the variation measure, for varying choices of N. We fix $M = 10$ for all experiments.

IV. RESULTS AND DISCUSSION

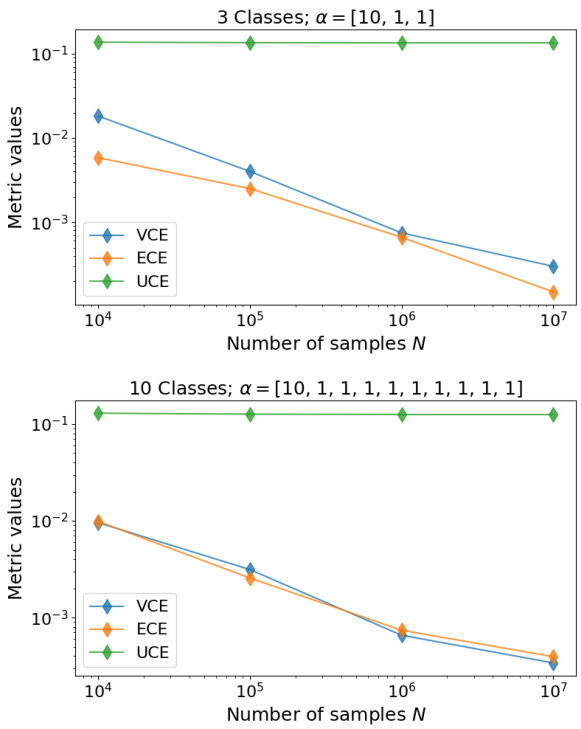


Fig. 1. Metric results for an equal-width binning strategy for the VCE, ECE, and UCE metrics. Results are shown for a 3*-class and* 10*-class classification problem for a heavily-skewed Dirichlet distribution, across* 4 *different sample numbers* $N \in \{10^4, 10^5, 10^6, 10^7\}$.

Figure 1 shows the results of evaluating the distinct calibration metrics for different sample sizes for two different class numbers. We examine the behaviour of the metrics with equal-width binning schemes for C-class ($C \in \{3, 10\}$) classification tasks, across one set of α parameters that parameterise a heavily-skewed Dirichlet distribution, i.e., mirroring a set of predictions for significant class imbalance.

Figure 1 demonstrates that the VCE metric obeys the expected asymptotic relationship $\text{VCE}^n \to 0$ as $n \to \infty$. The pathology of the UCE can be clearly seen, as a noise floor persists despite the use of perfectly calibrated data, even under increasing sample number. The ECE and VCE behave similarly, decreasing monotonically with increasing sample number N. Further experiments for different binning schemes, and various parametrisations of the Dirichlet distribution, have been conducted, however are omitted from this abstract for brevity. Importantly, we find that the VCE metric confirms perfect calibration, i.e, metric values close to 0, under all configurations of class number, sample number, and Dirichlet parameters α.

V. CONCLUSION

We propose the VCE to assess the calibration of generalised variation metrics for ML classification predictions. In particular this allows for an extension of the ECE to variation metrics which take into account the full probability distribution. We provide numerical evidence that the metric with entropy behaves as desired for predictions which are perfectly calibrated by design, in contrast to the UCE, which exhibits an undesirable noise floor even under perfect calibration.

REFERENCES

[1] Guo, C., Pleiss, G., Sun, Y. and Weinberger, K.Q., 2017, July. On calibration of modern neural networks. In International conference on machine learning (pp. 1321-1330). PMLR.

[2] Bilson, S., Cox, M., Pustogvar, A. and Thompson, A., 2025. A metrological framework for uncertainty evaluation in machine learning classification models. Metrologia, 62(6), p.064001.

[3] Laves, M.H., Ihler, S., Kortmann, K.P. and Ortmaier, T., 2019. Well-calibrated model uncertainty with temperature scaling for dropout variational inference. arXiv preprint arXiv:1909.13550.

2026 IMEKO Joint Conference
TC7 - TC11 – TC13 - TC21 in conjunction with TC6 and Mathmet
Faro, Portugal, 27-29 October, 2026

Extension of the JCGM 106 approach: assessing the number of truly conforming items in a lot

Francesca Pennecchi[1], Anna Ballario[1,2]

[1] *Istituto Nazionale di Ricerca Metrologica (INRiM), Str. delle Cacce 91, 10135 Torino, Italy*
f.pennecchi@inrim.it
[2] *Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy,*
anna.ballario@polito.it

Abstract – This paper extends conformity assessment framework based on JCGM 106:2012 from single items to an entire lot of N items. It introduces Model C to estimate the number of good true values (truly conforming items) when only a sample of n items is measured. Model C combines Model A, which uses specific risks requiring full inspection ($n = N$), and Model B, which uses global risks to assess the lot without any testing ($n = 0$). Defined as the convolution of the probability distributions of Models A and B, Model C serves as an intermediate approach. Monte Carlo simulations demonstrate that Model C improves estimation accuracy as sample size increases, offering a practical foundation for acceptance sampling plans when full inspection is unfeasible.

I. INTRODUCTION

Recent publications proposed statistical frameworks that extend conformity assessment (CA) from individual items - as addressed by JCGM 106:2012 [1] - to a whole batch/lot of items.

A first study [2] introduced two distinct models based on discrete probability distributions. Hereby, we are interested in the first one. Applying a Poisson binomial distribution, this model infers the actual number of truly conforming items within a lot of items, based on the observed measured values (affected by measurement uncertainty) and corresponding specific risks (calculated as prescribed by [1]). Those risks are the parameters of the Poisson binomial distribution.

A more recent publication [3] updates CA risk evaluation by embedding it into a generalized, fully Bayesian model originally conceived for measurements by counting. The number of truly conforming items is treated as the underlying measurand, while the accepted items represent the observed count (i.e., the number of items measured as conforming). By translating global consumer's and producer's risks (calculated as prescribed by [1]) into probabilities of false positives and false negatives, the model defines a discrete likelihood function that accounts for misclassification errors. This allows the integration of prior information through flexible uniform or binomial distributions. Ultimately, this Bayesian approach delivers a posterior probability mass function for the actual number of truly conforming items within the lot.

Both models (Model A based on the Poisson binomial distribution, and Model B based on the Bayesian approach for counting items) aim at making inference on the number of actually conforming items in a lot of N items, that is, on the number of good true values (GTV) in the lot. The difference between the two is that the former, relying on specific risks, requires that all the N items are measured and tested against tolerance limits, whereas the latter model, relying on global risks, provides inference without/before performing any actual measurement. The present work proposed a further model—Model C, which is the combination of Models A and B addressing the (intermediate) case when a sample of n items drawn from the lot is available and measured. This model conveys more information on the number of GTV in the lot with respect to model B, providing increasingly accurate estimates when sample size n increases.

II. A NEW MODEL FOR THE NUMBER OF TRULY CONFORMING ITEMS IN A LOT

Guide [1] provides a Bayesian framework for the conformity assessment of the true value η of a property of interest of a single item by inclusion of prior knowledge on η, modelled by a probability density function $g_0(\eta)$, and the knowledge acquired in the measurement, leading to a measured value η_m which is typically different from η. Global consumer risk R_c (false positive error) is defined as the joint probability of η being outside a tolerance interval TI and the corresponding measured value η_m being inside an acceptance interval AI [1, eq. (17)]. A counterpart definition holds for the global producer risk R_p (false negative error) [1, eq. (18)]. When a measurement result η_m is obtained, a posterior distribution can be formed based on the available prior information $g_0(\eta)$ and the likelihood function $h(\eta_m|\eta)$ (typically conveying the measurement uncertainty u_m). Specific consumer risk R_c^* and producer risk R_p^* are the integral of the posterior distribution outside and within TI, respectively [1, Clause 9.3.2].

A. Model A

In the most informative case of a full inspection of the lot, i.e., when all the N items are measured and corresponding specific risks can be calculated, Model A provides the probability of the number of GTV (between 0 and N), given by the discrete random variable $Y_A \sim$ Poisson binomial $(1 - R_{c1}^*, \ldots, R_{ck}^*, R_{p(K+1)}^*, \ldots, R_{pN}^*)$, where K is the number of items in the sample showing a good measured value (GMV), i.e., laying within the AI (AI could be equal or different from TI).

B. Model B

In the opposite situation, no sample is performed (and no actual measurement is taken) and the quality of the whole lot can be assessed only based on global risks. Model B provides the probability of the number of GTV (between 0 and N), given by a (posterior) discrete random variable Y_B whose parameters are the probabilities of incorrect CA decisions, which depend on the global risks and the probability $P(\eta \in \mathrm{TI}) = \int_{\mathrm{TI}} g_0(\eta)\, d\eta$.

C. Model C – the new model

Considering the intermediate case where only a sample of n items undergo CA and the remaining $N - n$ are left unseen, the random variable counting the number of GTV items within the whole lot is given by $Y_C = Y_A + Y_B$, where Y_A is applied on the sample of size n and Y_B on the remaining $N - n$ items. The distribution of Y_C is hence the convolution of the probability distributions of Y_A and Y_B.

III. SIMULATION STUDY

To study the efficacy of the proposed Model C, several Monte Carlo simulations were performed assuming a standard normal distribution N(0,1) as the prior for the quantity of interest η; a normal distribution as the likelihood function of the observed values (with standard deviation equal to measurement uncertainty u_m); AI equal to TI, specifically $(-\infty, 2)$; lot size equal to $N = 1000$. At each simulation, N random values η were drawn from the prior distribution, simulating the true values of the N items in the lot – the ground truth, and just as many η_m values were drawn from the normal likelihood $N(\eta, u_m)$, simulating the corresponding measured values. To consider the possible discrepancy between the assumed prior distribution and the one that might be really underlying the population, additional simulations were performed by drawing the 1000 η values from different prior distributions, while maintaining the assumed N(0,1) prior when computing the posterior distribution (and hence specific and global risks).

For each configuration of the parameters involved, 10000 MC simulations were performed. The percentage of simulated GTVs being encompassed in the 95 % coverage interval of random variables Y_A, Y_B and Y_c was evaluated as a performance metric of the corresponding models.

IV. RESULTS AND DISCUSSION

Table 1 reports the performance metrics of the three models, when prior standard deviation is equal to one and $u_m = 0.1$. It can be appreciated that for $n = 0$ (Model B), 100 (Model C), 500 (Model C), and N (Model A), there is an increasing performance, showing that the best solution would be to test all the items, naturally. Should a full inspection not be possible, an appropriate choice of the sample size n could anyway lead to acceptable results. Note that for prior distribution departing from the assumed standard normal, performance metrics decrease, except for Model A that relies on all the measured values that are consistent with the corresponding true values. When prior mean is 0.5 (substantially different from the assumed 0), Model B and C fail, irrespectively of the sample size. Nonetheless, the mean values of the 10000 expectations of the random variables corresponding to the three models approach the mean values of the simulated GTVs (i.e., 933 - the ground truth): 978 (Model B), 974 (Model C, $n = 100$), 956 (Model C, $n = 500$), and 934 (Model A). That is, on average, the expected number of estimated GTVs tends to the true mean value, even if, individually, the simulated true values are almost never included in the 95 % coverage interval of the random variables corresponding to Model B and C.

Table 1. Performance metric of the three Models, when prior standard deviation is equal to one and $u_m = 0.1$.

Prior mean	Model B	Model C ($n = 100$)	Model C ($n = 500$)	Model A
0	66 %	72 %	82 %	99 %
0.1	40 %	43 %	66 %	98 %
0.5	0 %	0 %	0 %	98 %

V. CONCLUSION

The introduction of Model C as the convolution of the distributions underlying Model A and B, and as a compromise between the two, can offer a base for determination of acceptance sampling plans. All the models, making inference on the number of GTVs in a lot, generalize the approach in [1], extending the CA to a whole set of items rather than a single item at a time.

Future work will deal with the impact on the proposed model by further parameters, like prior standard deviation, measurement uncertainty, and the ratio n/N.

REFERENCES

[1] JCGM 106:2012 "Evaluation of measurement data – The role of measurement uncertainty in conformity assessment".

[2] F Pennecchi and I Kuselman, Probabilities of true and false decisions in conformity assessment of a finite sample of items, Meas. Sci. Technol. 35 (2024) 056003.

https://iopscience.iop.org/article/10.1088/1361-6501/ad2439
[3] Pennecchi, F., Bich, W. A Bayesian model for measurements by counting. Qual Quant (2026). https://doi.org/10.1007/s11135-026-02773-3

2026 IMEKO Joint Conference
TC7 - TC11 - TC13 - TC21 in association with TC6 and Mathmet
Faro, Portugal, 27–29 October, 2026

Quantitative Image Analysis: An Approach to Estimating Uncertainties Associated with Image Segmentation

Weronika Pryciuk[1], Miles McCrory[2], Liam Wright[3]

[1]*Climate and Earth Observation, National Physical Laboratory, weronika.pryciuk@npl.co.uk*
[2]*Data Science, National Physical Laboratory, miles.mccrory@npl.co.uk*
[3]*Data Science, National Physical Laboratory, liam.wright@npl.co.uk*

***Abstract* – Image segmentation is a key image processing step in a typical data analysis pipeline, and has been widely adopted by many sectors of research. However, image segmentation is susceptible to uncertainties which need to be addressed in order to build metroligically robust image analysis routines. This work focuses on image segmentation using clustering, with input uncertainties originating from the input images and inherent uncertainties associated with the analysis pipeline. A framework is presented for propagating these uncertainties to derive probabilistic segmentation maps with associated uncertainty estimates.**

I. INTRODUCTION

Estimating uncertainties associated with a measurand is crucial in building trust in the results inferred, or decisions made, from the measured data: well established metrological practices are carried out in a wide range of sectors in academia and industry. However, often overlooked is image data as a measurand, and uncertainties can be left unexplored or undetermined. Imaging science is widely used throughout academic and industrial research (for instance using magnetic resonance imaging as a diagnostic tool for patients [1] or investigating the effects of climate change using satellite images [2]).

Image segmentation is widespread in the analysis pipeline , however, to robustly assert which segment a particular pixel is in, metrology practices must be applied. Segmentation methods are vast, ranging from simple segmentation methods based on pixel intensity values and thresholding, to much more advanced methods such as Segment Anything Model [3]; at the forefront of artificial intelligence and machine learning.

In this work, clustering segmentation was opted as the most viable and applicable method to a variety of image datasets. The method proposed here introduces metrological practices into a clustering segmentation routine to estimate uncertainty associated with the well-established image processing step. Propagation of uncertainty from pixel-level measurements to region-level decisions is thereby achieved, supporting trusted downstream analysis.

II. METHODOLOGY

The proposed approach models both random and systematic sources of uncertainty associated with input image data and propagates them through successive processing stages. Image-derived features, including local intensity statistics, texture metrics and gradient-based descriptors, are computed under uncertainty and used as inputs to a clustering-based segmentation step. The proposed approach also accounts for uncertainty in the segmentation algorithm itself. Due to the non-linear and non-smooth nature of segmentation, a Monte Carlo approach consistent with established metrological guidance is adopted for uncertainty propagation [4].

A. *Feature extraction uncertainty*

To characterise uncertainty in the input image, a multidimensional feature space is constructed using a set of image-derived features capturing local intensity and structural properties. The feature set includes local mean and variance, entropy and the mean and variance of the gradient magnitude field derived from Sobel-filtered image gradients [5]. Since these features are derived from uncertain input data, uncertainty associated with both the image measurements and the feature extraction operations must be considered and propagated through subsequent analysis stages using Monte Carlo sampling. These features were chosen for their widespread use in image analysis and their ability to capture complementary aspects of image content. Importantly, these features are computationally tractable for Monte Carlo evaluation.

B. *Clustering algorithm uncertainty*

K-means clustering is employed as a well-established unsupervised clustering method which defines a mapping from the feature space to discrete cluster labels [6]. The random initialisation of the K-means clustering algorithm often results in different cluster labels which can be considered as one source of uncertainty in the model. Random

Monte Carlo sampling is used to quantify the model uncertainty by obtaining a set of cluster labels that capture the variability of the segmentation. Whilst K-means can be computationally expensive, capturing the uncertainty in the segmentation from the clustering algorithm is necessary for confident and robust decision making in downstream tasks.

C. *Probabilistic segmentation map*

The guide to the expression of uncertainty in measurement supplement document 'Propagation of distributions using a Monte Carlo method' [4] provides guidance on defining the measurement model through which uncertainty will be propagated, as well as defining the probability density functions for the inputs. Here, input uncertainties are defined as random and systematic sources following a Gaussian distribution. Image uncertainty is propagated using random Monte Carlo sampling through texture metrics using punpy [7]. Feature perturbations are sampled from Gaussian distributions using the propagated standard uncertainties. Using a Monte Carlo scheme, realisations of the derived feature space are clustered using K-means with seeded random initialisation to obtain a set of cluster labels. A consensus matrix is constructed from the set of cluster labels [8] which captures the tendency of pixels appearing in the same cluster over n Monte Carlo trials determined by convergence in the normalised mutual information [9] between cluster labels. For a given clustering solution, a mean and a standard deviation is calculated from the consensus matrix for each pixel. These statistics are used to form a probabilistic segmentation map with associated uncertainty. The probabilistic segmentation quantifies the average tendency of a clustering solution and the stability of each pixel being in a cluster.

III. RESULTS

Preliminary evaluation was performed using a synthetic four-class texture image comprising sinusoidal, checkerboard, striped, and granular regions. Random measurement uncertainty was introduced into the image observations and propagated through feature extraction and clustering using Monte Carlo simulation. Here, the standard deviation computed from the consensus matrix is shown for the segmentation. Figure 1 shows (a) the original input image, (b)(i) the uncertainty from the input image, (b)(ii) the uncertainty in the clustering and (b)(iii) the uncertainty from both the input image and the clustering. The greatest instability in the segmentation as a result of uncertainty inputs is captured through a larger standard deviation which is observed along the boundaries of the four textures.

IV. DISCUSSION

In this paper, a new and novel framework for propagating uncertainty through a segmentation workflow is proposed. The probabilistic segmentation maps obtained from this framework capture the stability of a segmentation to support downstream analysis and also allow for traceability through an uncertainty pipeline. The method will be applied to satellite and hyper-spectral imaging data to validate the suitability of the approach to real-world tasks. Further investigation into methods for propagating uncertainties that remain computationally tractable will be tested. Whilst the focus here has been on image-based texture metrics and K-means clustering, this framework can be generalised to alternative feature spaces and segmentation algorithms, providing a broad foundation for robust and traceable image analysis routines.

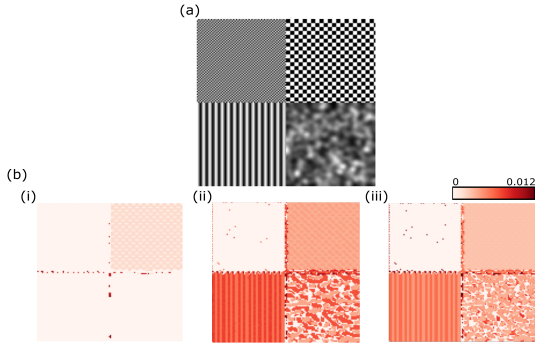


Fig. 1. Synthetic four-class texture image with associated probabilistic segmentation uncertainty maps.

REFERENCES

[1] S. Hussain, et al. Modern diagnosistic imaging technique applications and risk factors in the medical field: a review. BioMed research international. 2022.

[2] J. Yang, et al. The role of satellite remote sensing in climate change studies. Nature climate change. 2013.

[3] A. Kirillov, et al. "Segment anything." Proceedings of the IEEE/CVF. 2023.

[4] JCGM. Evaluation of measurement data - Guide to the expression of uncertainty in measurement. 2008.

[5] O. R. Vincent, et al. A descriptive algorithm for sobel image edge detection. In Proceedings of informing science and IT education conference. 2009.

[6] M. Ahmed, et al. The k-means Algorithm: A Comprehensive Survey and Performance Evaluation, Electronics. 2020.

[7] P. D. Vis, et al. Punpy: Propagating Uncertainties with PYthon. Python.

[8] S. Monti, et al. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. Springer. 2003.

[9] M. McCrory and S. A. Thomas, "Cluster Metric Sensitivity to Irrelevant Features''. Computational Problems in Science and Engineering II. 2024.

2026 IMEKO Joint Conference
TC7 - TC11 - TC13 - TC21 in association with TC6 and Mathmet
Faro, Portugal, 27–29 October, 2026

Bridging the Gap Between Chemometrics and Metrology

Liam Wright[1 *], Miles McCrory[1], Xavier Loizeau[1], Alison Nordon[2], Peter Harris[1]

[1]*Data Science, National Physical Laboratory, liam.wright@npl.co.uk*
[2]*Centre for Process Analytics and Control Technology, University of Strathclyde*
[*]*corresponding author: liam.wright@npl.co.uk*

***Abstract* – This work presents the methods and results of building metrological practices into analytical chemistry data analysis pipelines. Using a Monte Carlo approach, uncertainties associated with a well-established calibration model generation method's estimations are quantified and examined for various levels of available input information. Advancing the uncertainty quantification routine to be suitable for real-time analysis was carried out using multi-fidelity Monte Carlo - for in-line spectroscopy applications.**

I. INTRODUCTION

Chemometrics is an area of analytical chemistry which covers a wide range of subject areas: from pharmaceuticals and the food industry to chemical manufacturing and waste management - spanning a large range of methods, techniques, and practices in order to acquire and analyse data with regard to the chemistry of the system in question. Within it, there is a large range of mathematical techniques used to process collected data into results useful to the end user. Building trust in the calculated results requires a robust and thorough uncertainty analysis pipeline to be devised. Metrology within chemometrics, however, is generally poorly addressed, and therefore ground must be broken in order to meet the need.

This work focuses on building metrological frameworks in a key chemometrics analytical technique: calibration model generation. The methods and results presented here are two-fold: firstly, propagating uncertainty through a calibration model for mid-infrared spectroscopy data analysis, and secondly employing multi-fidelity Monte Carlo as a technique to quicken the estimation of uncertainties for real-time analysis purposes.

II. METHODOLOGY

A. Data acquisition

Experimental data was acquired using mid-infrared (MIR) spectroscopy for a ternary mixture of ethanol, acetone, and ethyl-acetate. MIR spectroscopy is a powerful tool, can be cheap to run in comparison to other methods, and can be used as an analytical probe into the make-up of a chemical - known or unknown [2]. A ternary mixture of ethanol, acetone, and ethtyl-acetate were investigated, and the MIR wavenumber range spanned [800, 2000] nm^{-1} - a range in which the three chemicals have identifiable peaks.

Uncertainties associated with the spectral measurands were estimated using a top-down approach, as a bottom-up, analytical solution was unfeasible [1]. Part of this work explores the amount of information given to the calibration model, and what effect it has on the resulting output and associated uncertainties.

B. Calibration model generation

To estimate an unknown chemical mixture, one can build a calibration model using a set of known mixtures and corresponding spectroscopic response (known from hereon in as the calibration set). Here, calibration model generation was carried out using partial least squares (PLS) regression - a popular method in chemometrics [4]. PLS is a multivariate linear regression method used to determine the model relating a predictor matrix, $\mathbf{X}$, and response $\mathbf{Y}$ via $\mathbf{Y} = \beta\mathbf{X} + \beta_{\mathbf{0}}$ where β and $\beta_{\mathbf{0}}$ are the regression coefficient matrix and error term respectively. Here, $\mathbf{X}$ is a matrix of the measured MIR spectra for each composition mixture matrix $\mathbf{Y}$.

Ethanol (%)	Acetone (%)	Ethyl-acetate (%)
100	0	0
50	50	0
17	33	50
⋮	⋮	⋮

Table 1. Example of the calibration mixture matrix, consisting of compositions for the ternary mixture. Each entry (i.e. row) has measured spectra associated.

Here, β is calculated using the calibration set of spectra and associated composition mixture matrix (an example of this matrix is given in Table 1) and the unknown (or 'test') composition mixture matrix can be calculated using the measured spectra and calculated β. SIMPLS algorithm [3] was employed to solve the regression problem above, and provide a calibration model for the chemometrics problem stated. The SIMPLS algorithm was treated as the measurement model in this work, through which uncertainties were

propagated.

C. Classic and multi-fidelity Monte Carlo

A Monte Carlo (MC) approach to uncertainty propagation [1] was used as a numerical method was the only viable solution - analytical uncertainty propagation was not possible. The routine in the first instance sampled input spectral data from the real data through bootstrapping or from i.i.d. normal distributions, and estimates of the unknown mixture concentrations were calculated with associated uncertainties.

The classic MC method was implemented and tested by varying levels of information with respect to the input data and different sampling methods. Typical measurements taken in chemometrics consist of three repeats of each spectra, which from a metrology perspective requires assumptions to be made about the underlying distribution function in order to propagate uncertainty. Here, larger repeated datasets were available which allowed for other sampling techniques to be tested.

Extending on the MC approach, a multi-fidelity MC (MFMC) framework was devised as a method to propagate uncertainty for real-time applications [5]. MFMC generally works by splitting the calculation into high- and low-fidelity models or data, estimating the desired measurand, and calculating a statistical comparison between the fidelity models. High- and low-fidelity in this instance was defined as the resolution of the spectra used in building the calibration model (i.e. the number of wavenumber values in the spectra); low-fidelity being lower resolution, and high-fidelity being higher resolution.

III. RESULTS

This section presents some exemplary results highlighting the some of the methods and their capabilities - only the classic MC uncertainty propagation is considered. The ternary mixture of ethanol, acetone, and ethyl-acetate were used as inputs, and calibration models were created using PLS regression and the SIMPLS algorithm.

The estimations could be compared to a ground truth (through gravimetric measurements) which allowed for assessment of the methods to be probed, and for adaptive MC to be employed. The results presented in Figure 1(a) show a subset of the output from the classic MC approach, with the back-calculated calibration mixture composition highlighted in green, and the unknown mixture given in red. In each sub image of Figure 1(a), the red line indicates the known ground truth value.

For each output distribution of the back-calculated and unknown test mixtures, Pearson's correlation coefficients were calculate, presented in Figure 1(b) as colour maps.

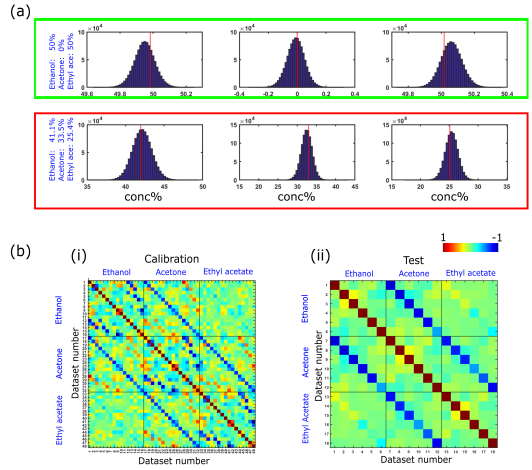


Fig. 1. Classic Monte Carlo approach to uncertainty propagation through the calibration model. (a) shows a subset out estimations for the back-calculated calibration mixtures (green) and unknown test mixtures (red). (b) shows the calculated Pearson's correlation coefficients for each output distribution of the approach.

IV. CONCLUSION & DISCUSSION

The foundations of improving metrological practices within one aspect of chemometrics are presented. MC and MFMC techniques showcase the viability of uncertainty analysis within an application area like spectroscopic chemistry, with the additional benefit of improving calculation speed with the latter. Examples of the proposed methods are given for the ternary mixtures of ethanol, acetone, and ethyl-acetate - building calibration models using the well established PLS regression method.

REFERENCES

[1] BIPM et al. Evaluation of measurement data, Supplement 1 to the "Guide to the expression of uncertainty in measurement: propagation of distributions using a Monte Carlo method". Joint Committee for Guides in Metrology, JCGM 101:2008

[2] H. Pollock and S. Kazarian, "Microspectroscopy in the Mid-Infrared", In Encyclopedia of Analytical Chemistry, pp. 1-26, 2014.

[3] S. De Jong, "SIMPLS: An alternative approach to partial least squares regression", Chemometrics and Intelligent Laboratory Systems, vol. 18, no. 3, pp. 251-263, 1993.

[4] S. Wold, et al., "PLS-regression: a basic tool of chemometrics", Chemometrics and Intelligent Laboratory Systems, vol. 58, no. 2, 2001, pp. 109-130,

[5] B. Peherstorfer, et al., "Multifidelity Monte Carlo estimation for large-scale uncertainty propagation", AIAA Non-Deterministic Approaches Conference (p. 1660), 2018.

2026 IMEKO Joint Conference
TC7 - TC11 - TC13 - TC21 in association with TC6 and Mathmet
Faro, Portugal, 27–29 October, 2026

Ellipsoid Fitting for Geometric Calibration: Uncertainty Quantification and Sphericity Testing

Viktor Witkovský[1], Eva Fišerová[2], Gejza Wimmer[3], Daniel Gogola[1]

[1]*Institute of Measurement Science, Slovak Academy of Sciences, Dúbravská cesta 9, 841 04 Bratislava, Slovakia; witkovsky@savba.sk*
[2]*Department of Mathematical Analysis and Applications of Mathematics, Faculty of Science, Palacký University Olomouc, 17. listopadu 12, 771 46 Olomouc, Czech Republic; eva.fiserova@upol.cz*
[3]*Mathematical Institute, Slovak Academy of Sciences, Štefánikova 49, 814 73 Bratislava, Slovakia; wimmer@mat.savba.sk*

***Abstract* – Reliable geometric calibration requires statistical methods that distinguish actual shape deformation from variability caused by measurement uncertainty. We consider fitting a three-dimensional ellipsoid to measured surface points within a nonlinear errors-in-variables framework, in which all spatial coordinates are treated as uncertain measured quantities. The approach uses iterated linearization and the OEFPIL methodology, allowing covariance information to enter the fit and providing uncertainty estimates for the fitted parameters. Algebraic ellipsoid parameters can be further transformed into the center, principal semi-axes, and orientation, with uncertainty propagation to these geometric quantities. The resulting covariance information also provides a basis for statistical testing of sphericity. The methodology is motivated by calibration of magnetic resonance imaging systems using spherical phantoms, where deviations of the reconstructed surface from a sphere reveal geometric distortions associated with the gradient system.**

I. INTRODUCTION

Geometric fitting is a basic operation in dimensional metrology, imaging and instrument calibration. A common simplification is to regard some measured coordinates as error-free predictors. For surface reconstruction this distinction is artificial: all coordinates are measurement results and may be affected by uncertainty and correlation. Consequently, ordinary regression or a purely algebraic fit need not represent the measurement process adequately.

Ellipsoid fitting provides a useful example. A general non-degenerate ellipsoid can be written as

$$x^2+By^2+Cz^2+Dxy+Exz+Fyz+Gx+Hy+Kz+L=0, \tag{1}$$

where the normalization of the coefficient of x^2 removes the arbitrary scale. The coefficients determine the centre, lengths and directions of the principal semi-axes. Their estimation is therefore directly linked to quantitative assessment of geometric distortion.

The purpose of this contribution is to formulate ellipsoid fitting as an uncertainty-aware errors-in-variables (EIV) problem and to connect the fitted geometry with a metrological decision on sphericity. The work builds on earlier works on ellipse fitting by linearized nonlinear constraints [3], on ellipsoid fitting to data [1], and on the OEFPIL (*Optimal Estimation of Function Parameters by Iterated Linearization*) framework for nonlinear EIV regression with correlated variables [2]. This contribution is model and application-oriented; a full methodological details and numerical validation study is still in progress.

II. EIV ELLIPSOID FITTING WITH OEFPIL

Let $\mathbf{x}_i = (x_i, y_i, z_i)'$ denote an observed surface point and $\mu_i = (\mu_i, \nu_i, \xi_i)'$ its unknown error-free counterpart. The measurement model is

$$\mathbf{x}_i = \mu_i + \varepsilon_i, \qquad f(\mu_i;\theta) = 0, \tag{2}$$

where $f = 0$ is the ellipsoid constraint corresponding to (1) and $\theta = (B, C, D, E, F, G, H, K, L)'$. In stacked form, the errors may be assigned a covariance matrix Σ describing uncertainties and correlations of the measured coordinates.

The nonlinear constraints are linearized by a first-order Taylor expansion at current estimates of the error-free coordinates and ellipsoid parameters. The resulting locally linear model with type-II constraints is solved repeatedly until convergence. This is the central computational idea used by OEFPIL, an iterated-linearization approach for nonlinear EIV models. The general algorithm accommodates correlated EIV data and yields parameter estimates together with their covariance matrix [2]. A MATLAB implementation is openly available [4].

For the ellipsoid problem, the fitted algebraic coefficients are subsequently transformed to geometric quantities. The covariance matrix of the algebraic estimates can be propagated through these transformations to quantify uncertainty of the centre, axes and other derived characteristics.

III. UNCERTAINTY AND SPHERICITY TESTING

For a nominal spherical object, the fitted ellipsoid should be compatible with equal principal semi-axes. In the normalized algebraic representation (1), a sphere satisfies

$$B = C = 1, \qquad D = E = F = 0. \tag{3}$$

Thus sphericity can be formulated as a joint statistical hypothesis concerning five fitted coefficients. The covariance matrix supplied by the EIV fit provides the information needed for a Wald-type test of these restrictions. Under the homoscedastic, uncorrelated model, the working theoretical development leads to an F-type calibration; extension and validation for more general covariance structures form part of the ongoing work.

This distinction is metrologically important. Differences between estimated semi-axis lengths describe the magnitude of deformation, whereas a hypothesis test addresses whether the apparent deformation is distinguishable from measurement uncertainty. The analysis therefore connects parameter estimation, uncertainty quantification and a decision criterion within a single measurement model.

IV. MRI PHANTOM CALIBRATION

The motivating application arises in magnetic resonance imaging (MRI). Spatial information is encoded by gradient magnetic fields along the three coordinate axes. Imperfect gradient linearity, unequal gradient strengths or departures from the intended geometric configuration can lead to systematic deformation of reconstructed objects. A spherical phantom with known geometry is therefore a natural reference object for assessing spatial fidelity.

Thin contiguous spin-echo slices of the phantom are used to reconstruct a three-dimensional point cloud describing its surface. Measurements available at the Institute of Measurement Science were obtained on three different MRI systems: a permanent-magnet ESAOTE Opera system, a resistive whole-body research tomograph, and a 4.7 T Varian research system. The systems use different magnet and RF-coil configurations, which also produce different spatial distributions of detected surface points. Volume RF coils provide comparatively homogeneous surface coverage, while a surface coil may lead to pronounced non-uniform point density.

Figure 1 illustrates the reconstructed point cloud for the smaller spherical phantom measured on the Varian 4.7 T system. In the intended calibration workflow, OEFPIL-based fitting separates the latent ellipsoid geometry from coordinate measurement errors. Estimated semi-axis ratios characterize axis-dependent scaling; the estimated principal directions can reveal cross-axis effects; and the sphericity test provides an uncertainty-aware criterion for deciding whether a significant geometric deformation is present. These quantities can then support subsequent determination of corrections for the MRI gradient system.

V. CONCLUSIONS

Ellipsoid fitting is a compact example of the complete data-to-decision workflow emphasized by trusted measurement and trusted analysis. The measurement coordinates, nonlinear surface model, numerical fitting algorithm, covariance information, uncertainty propagation and final sphericity decision are treated as connected elements rather than as separate processing steps.

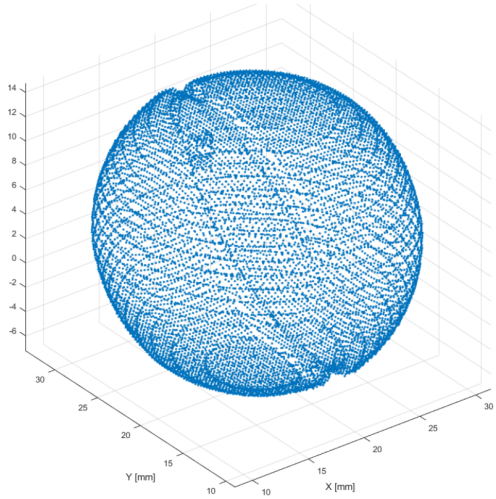


Fig. 1. Measured surface points of a spherical phantom reconstructed from MRI data acquired on the Varian 4.7 T system with a volume RF coil. The nearly uniform point coverage makes the data suitable for uncertainty-aware assessment of geometric deformation.

The proposed approach combines current theoretical work on ellipsoid parametrization and inference, the OEFPIL methodology for nonlinear EIV models, and experimentally available MRI phantom data. The immediate objective is not to claim a fully validated MRI calibration procedure, but to establish a statistically coherent framework in which geometric deformation can be estimated and judged relative to its uncertainty. The next stage will compare alternative covariance models, investigate sensitivity to non-uniform surface sampling and validate the sphericity test by simulation and repeated phantom measurements.

REFERENCES

[1] Turner, D. A. Anderson, I. J. Mason, J. C. and Cox, M. G, “An algorithm for fitting an ellipsoid to data”, *National Physical Laboratory, UK*, 1999.

[2] V. Witkovský, G. Wimmer, A. Charvátová Campbell, P. Klapetek and R. Šlesinger, “Estimation of function parameters through iterated linearization for nonlinear errors-in-variable regression with correlated variables”, *Measurement: Sensors*, vol. 38, 101728, 2025. doi:10.1016/j.measen.2024.101728.

[3] R. König, G. Wimmer and V. Witkovský, “Ellipse fitting by linearized nonlinear constraints to demodulate quadrature homodyne interferometer signals and to determine the statistical uncertainty of the interferometric phase”, *Measurement Science and Technology*, vol. 25, 115001, 2014.

[4] V. Witkovský, “OEFPIL-MATLAB”, open-source MATLAB software, github.com/witkovsky/OEFPIL-MATLAB, accessed August 2026.

Abstracts accepted as ***oral presentations***

T5 – Digitalization

page

#2 Joao Gregorio, ***"Image Data Quality: An Empirical Approach to Metadata Reliability in Multi-Site Imaging"*** *74*

#3 Hannah Strassburg, ***"Data Quality and Uncertainty in Climate Modelling: A Case Study Using the Input4MIPS Emissions Dataset"*** *76*

#9 Michael Chrubasik, ***"Towards an End-to-End Semantic Model For Calibration Services"*** *78*

#13 Marina Romanchikova, ***"Defining minimum metadata for nanometrology"*** *80*

#14 Bartlomiej Cieszynski, ***"Multi-Dimensional Data Quality Framework for Machine Learning in Arrhythmia Classification"*** *82*

#20 Priyanka Bharti, Marina Romanchikova, Beth Tiplady, ***"Multi-Purpose Semantic Modelling in Radiofrequency Measurements"*** *84*

#23 Carlos Pires, Gonçalo Carvalhal, João Abrantes, João A. Sousa, Isabel Lopes, Amauri Camargo, ***"Development and implementation of digital calibration certificates at IPQ: from NMI to end-user"*** *86*

#24 Johannes Kombe, ***"Virtual Knowledge Graph for Air Quality Measurement Datasets"*** *88*

#31 Callum Whyte, ***"Validation of LiDAR sensors in virtual testing environments"*** *90*

#35 Sai Hin Matthew Chan, Russell Miller, ***"Applying Data Quality Metrics to AIS Communications"*** *92*

#38 Teodor Tóth, ***"Impact of Acquisition Parameters on the Metrological Accuracy of xCT Workflows"*** *94*

#43 Moulham Alsuleman, Michael Chrubasik, Joao Gregorio, Harvey Whelan, ***"Agent-Based Knowledge Graph Generation for Scientific Experiments"*** *96*

2026 IMEKO Joint Conference
TC7 - TC11 - TC13 - TC21 in association with TC6 and Mathmet
Faro, Portugal, 27–29 October, 2026

Image Data Quality: An Empirical Approach to Metadata Reliability in Multi-Site Imaging

João Gregório*[1], Elizabeth A. Cooke[2], Liam Wright[3]

[1]*Informatics, Data Science & AI Department, National Physical Laboratory, Glasgow, UK*
[2]*Data Analytics and Modelling, Data Science & AI Department, National Physical Laboratory, Cambridge, UK*
[3]*Data Analytics and Modelling, Data Science & AI Department, National Physical Laboratory, Teddington, UK*
**joao.gregorio@npl.co.uk*

I. BACKGROUND

Multi-site imaging research projects depend on heterogeneous acquisition systems, institutional practices, and data management conventions. While such studies enable scale and comparative analysis, they introduce persistent challenges in metadata consistency, interoperability, and data usability [1, 2].

Imaging standards such as Digital Imaging and Communications in Medicine (DICOM) provide extensive metadata schemas intended to support interoperability [3]. However, these standards define what can be recorded rather than what is consistently available in practice. Differences in equipment manufacturers, software configurations, and site-specific workflows result in metadata that is unevenly populated, inconsistently formatted, or unsuitable for automated use. Consequently, workflows that assume the presence of standard-defined attributes frequently encounter gaps that only become apparent during downstream analysis.

In a multi-site quantitative magnetic resonance imaging (MRI) context, this manifests as difficulty identifying data subsets matching specific criteria, ambiguity in interpreting acquisition context, and a lack of systematic mechanisms for tracking provenance. These issues arise during data integration and use rather than at acquisition.

To address these challenges, this work develops a DICOM metadata processing pipeline as a case study to derive practical guidance for image data quality [4, 5]. The objective is to identify empirically grounded minimum metadata requirements, augmentation strategies, and reproducibility practices applicable across imaging domains.

II. METHODS

Rather than assuming theoretical DICOM standard compliance, an empirical approach, shown in Figure 1, was employed to identify reliably available metadata acquired across four institutions and three MRI scanner manufacturers. All available metadata fields were extracted and compared across sites to assess consistency.

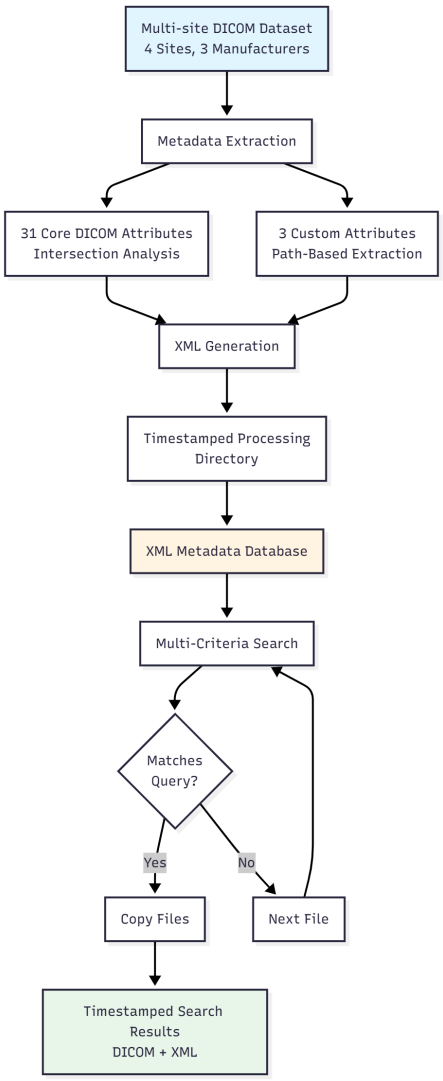


Fig. 1. DICOM metadata processing and search workflow. The pipeline combines intersection analysis of DICOM headers with path-based extraction of contextual attributes, storing results in timestamped XML records and enabling multi-criteria search.

A reliable baseline was established through intersection analysis, identifying only those attributes present in 100 % of files across all sites. This conservative approach ensures that the resulting metadata schema contains only consistently available and usable attributes.

To address missing contextual information, additional metadata were derived from structured file paths. This included extraction of acquisition site and measurement type using pattern matching and keyword detection. This augmentation recognises that relevant contextual information is often encoded outside formal metadata fields.

All extracted metadata were recorded in structured XML format, with outputs organised in timestamped directories to ensure traceability and reproducibility. A query mechanism was implemented to support filtering based on multiple attribute criteria, enabling consistent data discovery across the dataset.

III. RESULTS

The analysis revealed that only a small subset of metadata attributes was consistently available across all datasets. Of the many fields defined by DICOM, only 31 attributes were present in all files across all sites and manufacturers.

Attributes commonly assumed to be available, including acquisition parameters (e.g., echo time, repetition time) and spatial descriptors (e.g., pixel spacing, slice thickness), were frequently missing or inconsistently populated. This highlights a substantial gap between theoretical standard definitions and practical data availability.

Augmenting metadata with file path-derived information enabled recovery of key contextual attributes, including acquisition site and measurement type. This additional information supported consistent identification, filtering, and grouping of data across sites, improving usability for downstream analysis.

IV. DISCUSSION

The findings highlight a fundamental limitation of relying on formal standards alone: metadata availability is determined by implementation rather than specification. As a result, assumptions about metadata completeness can introduce hidden uncertainty into data analysis workflows. Three key implications follow.

First, minimum metadata requirements should be established through empirical validation rather than theoretical expectations. Intersection-based approaches provide a robust mechanism for identifying consistently available attributes across heterogeneous datasets.

Second, contextual information outside formal metadata fields should be systematically incorporated. Structured file paths, naming conventions, and other external sources can provide critical information not reliably captured within standard metadata fields.

Third, simple traceability practices, such as timestamped processing and structured outputs, can significantly improve reproducibility without requiring complex data management infrastructure.

While demonstrated on DICOM data, these observations are not domain-specific. Similar challenges arise across scientific and engineering domains where heterogeneous systems generate data under varying conventions. Addressing metadata reliability is therefore essential for improving the robustness and trustworthiness of data-driven methods and supporting FAIR data principles [6].

REFERENCES

[1] A. Kosvyra, D. T. Filos, D. T. Fotopoulos, O. Tsave, and I. Chouvarda, "Toward ensuring data quality in multi-site cancer imaging repositories," *Information*, vol. 15, no. 9, p. 533, 2024.

[2] S. Langer and B. Bartholmai, "Imaging informatics: challenges in multi-site imaging trials," *Journal of Digital Imaging*, vol. 24, no. 1, pp. 151–159, 2011.

[3] National Electrical Manufacturers Association (NEMA), "DICOM PS3.1 Digital Imaging and Communications in Medicine Standard," tech. rep., NEMA, 2025.

[4] R. Miller, H. Whelan, M. Chrubasik, D. Whittaker, P. Duncan, and J. Gregório, "A framework for current and new data quality dimensions: An overview," *Data*, vol. 9, no. 12, p. 151, 2024.

[5] J. Gregório, P. Duncan, M. Alsuleman, M. Chrubasik, B. Cieszynski, V. Desai, J.-L. Hippolyte, P. Krishnadas, K. Lines, R. Miller, N. Peric, and H. Whelan, "Data quality good practice guide," Tech. Rep. MS 63, National Physical Laboratory, 2026.

[6] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, *et al.*, "The FAIR guiding principles for scientific data management and stewardship," *Scientific Data*, vol. 3, no. 1, pp. 1–9, 2016.

2026 IMEKO Joint Conference
TC7 - TC11 - TC13 - TC21 in association with TC6 and Mathmet
Faro, Portugal, 27–29 October, 2026

Data Quality and Uncertainty in Climate Modelling: A Case Study Using the Input4MIPS Emissions Dataset

Hannah Strassburg*[1] and João Gregório[2]

[1]*Informatics, Data Science & AI Department, National Physical Laboratory, Teddington, UK*
[2]*Informatics, Data Science & AI Department, National Physical Laboratory, Glasgow, UK*
**hannah.strassburg@npl.co.uk*

I. BACKGROUND

Climate science requires large amounts of temporal and spatial data from different sources to run its modelling workflows. The use of poor-quality data in these workflows can lead to the introduction of bias to modelling outputs or create incorrect uncertainty estimates, causing results to be non-reproducible [1]. These issues affect the decision-making processes that rely on this data, as well as the modelling workflows through which poor-quality data propagates, ultimately undermining trust in the results. For climate modelling, the most common domains where data quality issues arise are sensor failure, leading to issues such as missing data points; inaccuracies in measurement readings; temporal inconsistencies; and a lack of data standardisation between different locations and times.

Addressing data quality issues identified in climate datasets before they are employed in modelling workflows can serve as a mitigation strategy for preventing the propagation of bias and incorrect uncertainty estimates through the end of the decision-making processes. Such mitigation is hindered by the lack of robust data quality frameworks for systematically identifying and representing data quality issues in model inputs. In response, this study proposes a structured quality framework for systematically identifying and representing data quality issues in climate modelling inputs, and it demonstrates its utility using a representative climate emissions dataset [2, 3, 4].

II. METHODS

The proposed framework was developed and evaluated using the Input4MIPS dataset, comprising of both global proxy-derived (e.g. from population density maps or prior emission inventories) and measured emissions data of carbon monoxide, methane, sulphur dioxide, and black carbon spanning 1750–2014 [4]. Like most climate datasets, the data are provided in NetCDF format, containing both metadata and atmospheric values [5]. The machine-readability of the data and metadata allows for a framework that encompasses automated data ingestion, statistical validation, anomaly detection and scenario clarity to perform structured checks on the quality of the dataset, allowing for aspects such as completeness, metadata compliance, consistency and traceability to be characterised [1].

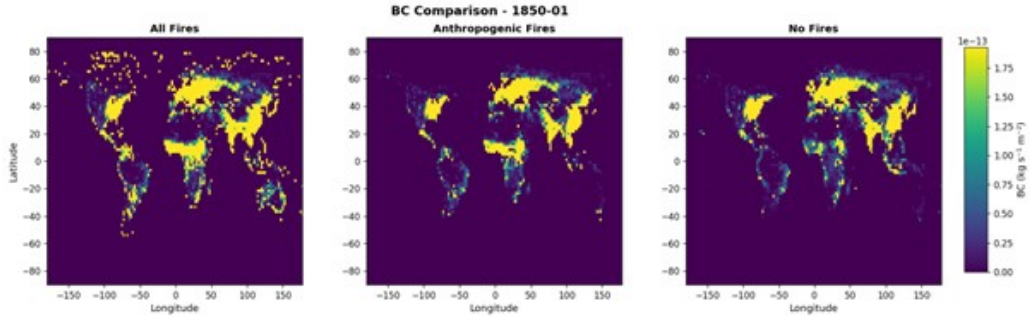
III. RESULTS

The characterisation of the dataset revealed a clear threshold through which a bimodal distribution could be defined. The threshold was defined to be 10^{-7} $\mathrm{kg\,m^{-1}\,s^{-2}}$, with the values below this representing emission values within plausible ranges, while those above being considered anomalous artefacts with concentrations of up to 10^{+18} $\mathrm{kg\,m^{-1}\,s^{-2}}$ [3]. These anomalous artefacts were clustered around coastlines, aligning with current literature [3]. Leveraging this distribution to identify modelling artefacts avoids reliance on externally defined physical bounds and offers a generalisable approach for comparable datasets. This is supported by Figure 1, where high-magnitude values persist across time and are confined to the all-fires scenario. Based on this classification, per-file statistics were derived to characterise the plausible data range, with the minimum defined as the smallest non-zero value and the maximum as the largest value within the valid domain, allowing for a proper understanding of the real fluctuations of the recorded chemical species.

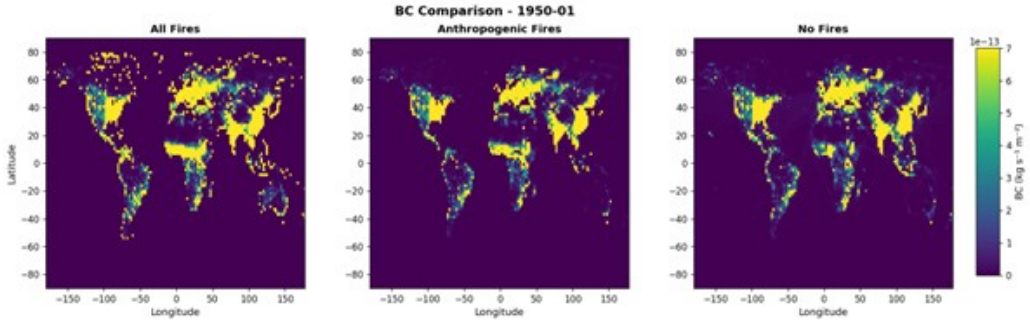
The dataset metadata was found to be structured, incorporating descriptors that support traceability and interoperability, thereby enabling reproducible analyses. This type of metadata structure facilitates both automated data handling in climate modelling workflows and manual quality assurance, ensuring consistency with the Climate Model Intercomparison Project standards and enabling the efficient identification of datasets for specific applications.

IV. DISCUSSION

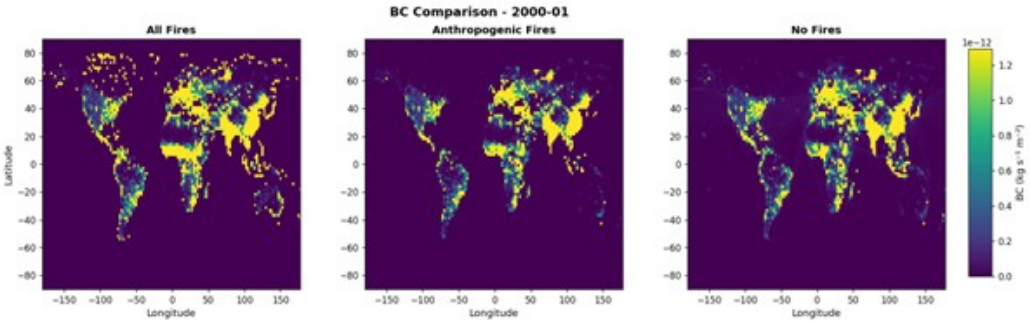
This study presents a structured data quality framework for climate modelling inputs, demonstrated on the Input4MIPS dataset. The results show that dataset characterisation can support the definition of meaningful, context-specific metrics for assessing data quality; enable the identification and handling of anomalous data; and support more reliable modelling outcomes. Future work will explore the development of data quality pipelines and the use



(a) 1850



(b) 1950



(c) 2000

Fig. 1. Temporal snapshots of Black Carbon emission flux across fire scenarios (columns within each row: all-fires, anthropogenic fires, no-fires) for selected years. Persistent high-magnitude values along coastlines appear only in the all-fires scenario and remain temporally invariant, indicating systematic non-physical artefacts in the dataset.

of semantic technologies to formally represent and operationalise data quality within climate modelling workflows.

REFERENCES

[1] R. Miller, H. Whelan, M. Chrubasik, D. Whittaker, P. Duncan, and J. Gregório, "A framework for current and new data quality dimensions: An overview," *Data*, vol. 9, no. 12, p. 151, 2024.

[2] J. Gregório, P. Duncan, M. Alsuleman, M. Chrubasik, B. Cieszynski, V. Desai, J.-L. Hippolyte, P. Krishnadas, K. Lines, R. Miller, N. Peric, and H. Whelan, "Data quality good practice guide," Tech. Rep. MS 63, National Physical Laboratory, 2026.

[3] R. M. Hoesly, S. J. Smith, L. Feng, Z. Klimont, G. Janssens-Maenhout, T. Pitkanen, J. J. Seibert, L. Vu, R. J. Andres, R. M. Bolt, T. C. Bond, L. Dawidowski, N. Kholod, J. I. Kurokawa, M. Li, L. Liu, Z. Lu, M. C. P. Moura, and P. R. O'Rourke, "Historical (1750–2014) anthropogenic emissions of reactive gases and aerosols from the community emissions data system (ceds)," *Geoscientific Model Development*, vol. 11, pp. 369–408, 2018.

[4] Lawrence Livermore National Laboratory (PCMDI), "Input4mips: Input datasets for model intercomparison projects," n.d. [Online].

[5] B. Eaton *et al.*, "Netcdf climate and forecast (cf) metadata conventions (version 1.13)," 2025. [Online].

2026 IMEKO Joint Conference
TC7 - TC11 – TC13 - TC21 in conjunction with TC6 and Mathmet
Faro, Portugal, 27-29 October 2026

Towards an End-to-End Semantic Model For Calibration Services

Michael Chrubasik[1], Hannah Strassburg[2], Jean-Laurent Hippolyte[3]

[1] *1 National Physical Laboratory, Glasgow, G1 1RD, michael.chrubasik@npl.co.uk*
[2] *National Physical Laboratory, Teddington, TW11 0LW, hannah.strassburg@npl.co.uk*
[3] *National Physical Laboratory, Cambridge, CB3 0HE, jean-laurent.hippolyte@npl.co.uk*

I. INTRODUCTION

Calibration laboratories hold large amounts of valuable measurement information and process context accumulated over years of operation, and making this data findable, accessible, interoperable and reusable (FAIR) remains a challenge [1]. Because laboratories governed by ISO/IEC 17025:2017 [2] operate under highly prescribed requirements, both the data and the processes that produce it are strong candidates for digitalisation. The standard sets out managerial and technical requirements for a calibration service, including Clause 6 on resources and Clause 7 on the process from request handling to the reporting of results.

In applied contexts, calibration records often end up split across spreadsheets, shared drives, and isolated Laboratory Information Management System setups. This fragmentation divorces actual measurement results and quality evidence from the operational context needed to audit and reuse the data [3]. Calibration data therefore remain fragmented across proprietary systems. A formal semantic representation of the end-to-end process offers a route through these challenges: machine-readable records enable automated integration, traceability verification and reusable administrative and measurement traces [4], [5].

Existing semantic models cover parts of the domain but remain scoped to single layers. The Digital Calibration Certificate (DCC) and its namesake ontology, DCCS, provide machine-readable certificate representations with result containers and participation patterns [4], [5]. Unit- and measurement-focused ontologies such as Ontology for Measurement Terminology (OMT), Digital-SI ontology (SIS), the SI Digital Framework and Quantities, Units, Dimensions and Types (QUDT) support metrology terminology, quantity values, units and uncertainty but not process flow [6], [7], [8]. Scientific metadata and provenance work, including Metadata4Ing and White's traceability scenarios, relate activities, methods and evidence through patterns built on provenance and sensor network ontologies (PROV-O and SOSA/SSN respectively [9], [10]). None formalises the end-to-end operationalisation of a calibration service, from customer request through laboratory execution to certificate provision. This work presents the Calibration Service Ontology (CSO) to address that gap. Contributions include ontology development following the Linked Open Terms (LOT) methodology, from scenario-driven requirements formalised in an Ontology Requirements Specification Document (ORSD) through to competency questions and basic validation; and the initial foundations of CSO via a hub-and-spoke approach. Validation at this stage is limited to synthetic competency-question coverage; full instantiation against representative calibration-service data is planned as future work.

II. METHOD

Development follows the LOT methodology, which structures ontology engineering into four activities [11]: requirement specification, implementation, publication and maintenance. This paper covers the first two; publication and maintenance are future work.

Requirements were drawn from three sources: an end-to-end operational data model from an NPL dimensional metrology service; a process breakdown from an NPL thermometry service covering method selection and step ordering; and White's published metrological provenance scenarios for traceability chains, measurement assurance and inter-laboratory certificate linking [10]. Scenario-specific terminology was abstracted into generic roles, classes and relations before deriving competency questions. An internal, access-controlled thesaurus was built to support analysis of ISO/IEC 17025 Clauses 3–8 alongside the VIM, GUM, ISO 17000 and ISO 9000 [12]. Candidate terms, source references and relationships were identified and validated for technical correctness and provenance. The implementation applies a hub-and-spoke architecture with external ontologies reused through direct import and subclassing rather than reinvention. Axioms are applied scarcely, with stricter validation left to post-processing. Coverage was assessed by mapping competency questions to executable query answer paths; questions without a complete path were recorded as gaps.

III. RESULTS

CSO operationalises ISO/IEC 17025 through a FAIR semantic model for calibration services and metrological traceability. It represents entities (methods, records, results), activities (processes, measurements) and agents (people, organisations) in their roles across the calibration lifecycle. Intended users are ontology engineers, metrology and quality experts, and laboratory data modelers; the intended use is to describe workflows across the calibration service lifecycle, providing a map of the resulting data flows. CSO builds on and links existing metrology-focused outputs into a unified calibration service model.

An initial terminology review resulted in the definition of an internal thesaurus, building on which the ontology is expressed through four modules: one hub and three spokes, with spokes importing only the hub and all external links held in the hub. The hub reuses DCCS and Metadata4Ing, which itself integrates PROV-O, SIS, SOSA/SSN and QUDT. The service-lifecycle spoke covers request, review, quotation, agreement, and item handling; the laboratory spoke covers method governance, measurement execution, evidence, result preparation and certificate authorisation; the traceability spoke covers VIM hierarchies, traceability chains, uncertainty budgets and inter-laboratory provenance. CSO classes carry source references to standards via mapping properties.

As part of the ontology requirement specification, competency questions, representing context and requirement-building questions, were defined, with 12 currently answerable through complete answer paths across seven areas of concern (method decision, reference materials, traceability chain, measurement uncertainty, certificate generation, customer information and measurement execution).

IV. DISCUSSION

CSO acts as a semantic-glue layer that aligns with DCCS, Metadata4Ing, SIS, SOSA/SSN and PROV-O in its hub and links standards-based concepts directly to those hierarchies. Centralising external ontology reuse within the hub simplifies maintenance and standards updates. The current model leans heavily on subclass and sub property axioms; later iterations should reduce unnecessary local terms and leverage direct reuse where it reduces complexity without weakening alignment. This is an early iteration: the thesaurus is under expert review, several competency questions still lack full representation, and the ontology has not been instantiated against a live service.

V. CONCLUSION AND FUTURE WORK

To the authors' knowledge, no existing ontology provides end-to-end semantic coverage of ISO/IEC 17025 calibration-service operations. CSO addresses this as an extended semantic-glue layer, supported by an internal thesaurus and source references to the VIM, GUM, ISO 17000, ISO 9000 and ISO/IEC 17025. Using LOT, CSO derives competency questions from three representative calibration scenarios. The architecture subclasses directly into DCCS, Metadata4Ing, SIS, SOSA and PROV-O rather than reinventing patterns. Future work aims to instantiate the ontology against representative calibration data with an application layer, add automated validation combining executable query tests with post-processing validation, and attempt a cross-NMI traceability demonstration using certificate-linking patterns. We present this work as a statement of intent and invite the metrology community to collaborate.

VI. ACKNOWLEDGEMENTS

This work was funded by the UK Department for Science, Innovation and Technology (DSIT) through the National Measurement Service (NMS) programme. The authors thank Joao Gregorio and Paul Duncan for their review and discussion input.

REFERENCES

[1] M. D. Wilkinson et al., "The FAIR Guiding Principles for scientific data management and stewardship", Sci. Data, vol. 3, p. 160018, 2016.

[2] ISO/IEC 17025:2017 — General Requirements for the Competence of Testing and Calibration Laboratories. ISO, Geneva, 2017.

[3] N. H.-Y. Yuen et al., "Laboratory Information Management System (LIMS) 4.0", Int. J. Eng. Bus. Manag., vol. 17, 2025.

[4] S. Hackel et al., "The Digital Calibration Certificate (DCC) for an End-to-End Digital Quality Infrastructure for Industry 4.0", Sci, vol. 5, no. 1, p. 11, 2023.

[5] DCCS Ontology – An OWL Ontology for Digital Calibration Certificates. PTB, 2024. https://ptb.de/dccs/v0.0.1/

[6] S. Wang et al., "Design and Implementation of an Ontology for Measurement Terminology in Digital Calibration Certificates", Sensors, vol. 24, no. 12, p. 3989, 2024.

[7] QUDT.org, Quantities, Units, Dimensions and Types (QUDT) Ontology, 2023. https://www.qudt.org/

[8] SIS Ontology – SI-Traceable Quantity Values, Units, and Uncertainty Declarations. PTB, 2024. https://ptb.de/sis/

[9] G. Lanza et al., Metadata4Ing: An ontology for describing the generation of research data within a scientific activity. Zenodo, 2025.

[10] R. M. White, "Provenance in the Context of Metrological Traceability", Metrology, vol. 5, no. 3, p. 52, 2025.

[11] M. Poveda-Villalón et al., "LOT: An industrial oriented ontology engineering framework", Eng. Appl. Artif. Intell., vol. 111, p. 104745, 2022.

[12] JCGM 200:2012 — International Vocabulary of Metrology (VIM). JCGM, Sèvres, 2012.

2026 IMEKO Joint Conference
TC7 - TC11 – TC13 - TC24 in conjunction with TC6 and Mathmet
Faro, Portugal, 27-29 October, 2026

Defining minimum metadata for nanometrology

Marina Romanchikova[1], Andrew Yacoot[1], Hannah Strassburg[1], Edward Heaps[1], Petr Klapetek[2], David Nečas[3], Bruno Torre[4], Virpi Korpelainen[5]

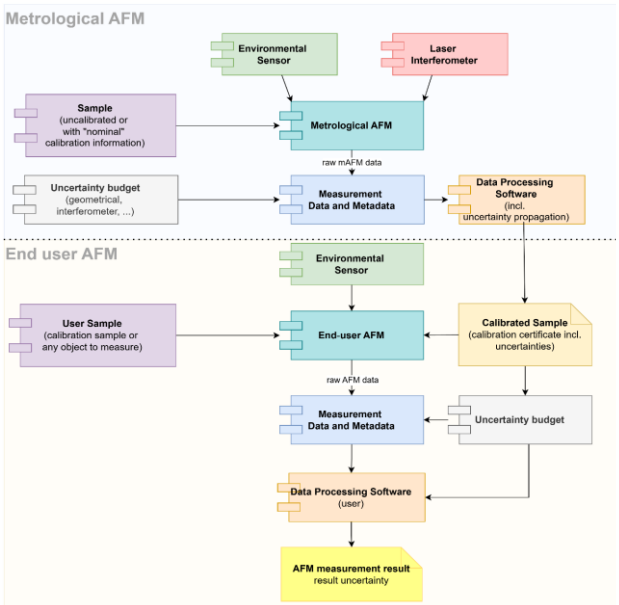
[1] *National Physical Laboratory, United Kingdom*
[2] *Czech Metrology Institute, Czech Republic*
[3] *Brno University of Technology, Czech Republic*
[4] *Istituto Nazionale di Ricerca Metrologica, Italy*
[5] *VTT Technical Research Centre of Finland Ltd, Finland*

I. INTRODUCTION

Although Atomic Force Microscopy (AFM) has been used for dimensional metrology since the 1990s, there are almost no formal descriptions (metadata schemas) of the AFM measurement process, with exception of the ISO 11952:2019 standard that contains guidelines on reporting of calibration results [1].

Metadata are vital for providing confidence in data and assessing the uncertainty of measurement results, as they describe what was measured and how (environmental conditions, instrument settings), in which form data are stored and how they were processed. A useful AFM process description needs to capture (just) enough information to support data processing and result interpretation, avoiding metadata collection for its own sake.

Fig. 1 Data processing pipeline in traceable AFM.



The information surrounding AFM measurements is typically scattered across proprietary instrument output and log files, laboratory notes, and calibration records. Lack of interoperable metadata increases documentation burden, leads to incorrect interpretation and processing, affects the validity of calibrations and measurements, limits data reuse, and contributes to the ongoing reproducibility crisis in science [2].

In the absence of agreed metadata specifications, NMIs, R&D laboratories, instrument manufacturers and end users to develop documentation practices that suit their needs. These practices may include the use of mixed digital and analogue metadata records in paper lab notebooks, photographs of the setup, spreadsheets and unformatted text. Furthermore, different metadata may be needed depending on the type of AFM and its use. An example of AFM data flow between an NMI and the end user is shown in Fig 1.

II. DINAMO APPROACH

Creating a minimal yet useful AFM measurement process description is the main task of the Digitalisation Route for Dimensional Nanometrology (DINAMO) project 24DIT02 funded by the European Partnership for Metrology. Work Package 1, “Digital Information Flow”, aims to develop the AFM Metadata Toolkit (AFM-MDT) - a set of metadata descriptors, rules and implementation examples to help users annotate AFM measurement data through the measurement pipeline.

The intended outputs of DINAMO are

1. Good practice guide for digital data management in nanometrology;
2. Example dataset showing how to implement the good practice using open-source file containers.

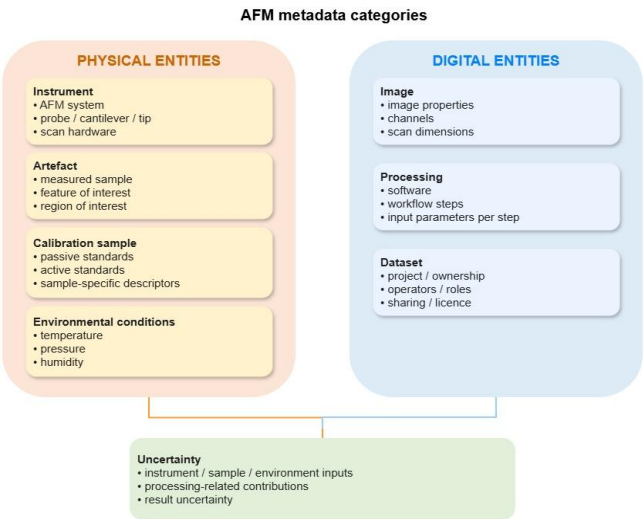
Metadata collection and revision were carried out by metrologists, software developers, manufacturers and other consortium partners to different requirements and perspectives. The working principle was to retain metadata only where it was needed to process AFM image data, interpret a measurement, assess calibration validity, or support reproducibility of result.

III. PRELIMINARY RESULTS

A. Metadata categories

The AFM-MDT groups metadata into physical, digital and cross-cutting categories illustrated in Figure 2.

Fig 2: Main AFM-MDT metadata categories.



Physical categories include the instrument, artefact, calibration standards and environmental conditions. Digital categories include image metadata, processing metadata and dataset description. Uncertainty is treated as cross-cutting because it may depend on outputs from both physical and digital information sources.

B. AFM-MDT coverage of real-life datasets

User's AFM metadata needs differ depending on the instrument type and use. For example, information about an interferometer may be crucial for a metrological AFM user, but of little value to an end user who uses a calibration sample, characterised by a metrological AFM, to calibrate their own instrument. To assess AFM-MDT performance, we compared it with the metadata sets obtained from a metrological AFM (metAFM) and a commercial AFM (comAFM).

The aim of the comparison was to show where AFM-MDT is "too much" (contains elements irrelevant to specific end user case) or "not enough" (does not provide all descriptors for the user scenario). The preliminary findings of the evaluation are shown in Table 1.

Table 1. Percentage of AFM-MDT metadata entries covered in the metadata of two example datasets.

AFM type	Instrument	Image	Dataset
Metrological	28%	40%	20%
Commercial	39%	80%	30%

Metadata coverage was incomplete in both datasets. The comAFM and metAFM metadata contained 14 and 10 of the 36 AFM-MDT fields, respectively. *Scan and stage motion* information was the only instrument category that was fully captured (all eight corresponding fields in both datasets). *Calibration* (two of 16 fields in both datasets) and *probe* metadata (one and zero of the eight AFM-MDT fields) were almost absent, indicating the need for additional manual logging by AFM users. *Image* metadata coverage was higher (four and two[1] out of five AFM-MDT fields).

IV. SUMMARY

This paper presents the interim results on developing minimum metadata definitions for traceable AFM measurements. The next steps will include:

- Continue testing and refining AFM-MDT metadata on real-life AFM measurements from metrological and commercial AFM instruments.
- Develop machine-readable AFM metadata representations;
- Implement and test AFM data and metadata container. A dataset will undergo the steps shown in Fig. 1 at different DINAMO partner sites to ensure interoperability.

The intended output of this work is a "simple but useful" open-source toolkit to help AFM manufacturers and users define and capture the metadata needed to process and interpret dimensional AFM measurements. The intended outcome is reduction in "data digging", increased efficiency and reliability in data processing, and improved reliability and reproducibility of dimensional AFM measurements as well as interoperability across AFM systems.

V. REFERENCES

[1] ISO 11952:2019 Determination of geometric quantities using SPM: Calibration of measuring systems.

[2] M Baker. "1,500 scientists lift the lid on reproducibility". Nature 533 (2016).

[1]: In metAFM metadata, scan time per pixel, pixel size, and number of pixels were inferable from other available metadata entries, fully covering AFM-MDT image metadata.

2026 IMEKO Joint Conference
TC7 - TC11 - TC13 - TC21 in association with TC6 and Mathmet
Faro, Portugal, 27–29 October, 2026

Multi-Dimensional Data Quality Framework for Machine Learning in Arrhythmia Classification

Bartłomiej Cieszyński*[1], Padmini Krishnadas[1], Vivek Desai[1], Nina Perić[1], Moulham Alsuleman[2], Michael Chrubasik[2], Paul Duncan[2], João Gregório[2]

[1]*Informatics, Data Science & AI Department, National Physical Laboratory, Teddington, UK*
[2]*Informatics, Data Science & AI Department, National Physical Laboratory, Glasgow, UK*
**bartlomiej.cieszynski@npl.co.uk*

I. BACKGROUND

The ISO 25012 standard defines data quality as the degree to which data is fit for its intended purpose and meets organisational requirements [1, 2]. In clinical settings, these requirements are shaped by patient safety, regulatory compliance, and the need to support timely and accurate decision-making. High-quality data is fundamental to reliable diagnostics, especially for systems that incorporate Machine Learning (ML), as failure to meet predefined quality criteria can compromise validity of results and safety of patients.

Electrocardiogram (ECG) data is widely regarded as a diagnostic gold standard for cardiac abnormality and is of increasing interest to ML researchers. The ECG data acquisition process is highly susceptible to degradation in data quality, making it a prime use case to study the impact of data quality on model performance. Signal reliability depends on recording duration, electrode placement, continuity across leads, and signal clarity. Degradation in these aspects may compromise diagnostics, leading to missed arrhythmias or misinterpretation of cardiac events [3].

Prior studies have proposed a variety of Signal Quality Indices (SQIs) to assess signal quality and investigated the impact of noise and artefacts on ML performance [4]. However, these approaches predominantly interpret signal quality as a measure of noise, thereby overlooking broader data quality dimensions defined in ISO 25012, such as accuracy, completeness, and precision [1, 5].

II. METHODS

This work aims to demonstrate how a correlation between a Data Quality Metric (DQM), model performance, and calibration may be obtained, by adopting a reference free approach to assess data quality. The DQM is defined through a weighted combination of multiple individual SQIs, a set of assessment metrics used to determine the suitability of a signal for its intended use case. The SQIs identified for this work are Signal-to-Noise Ratio (SNR), relative power of the QRS complex (QRS_RP), Sample Entropy (SampEn), Expected Precision (ExpPre), Coefficient of Variation (CV), Time Completeness (TimeComp), and Feature Completeness (FeatComp). These SQIs then map into the data quality dimensions of accuracy, completeness, and precision. The methodological workflow is shown in Figure 1.

The dataset used in this study is derived from PhysioNet's MIT-BIH Arrhythmia Database [6]. It was obtained from Kaggle [7], and has undergone pre-processing in the form of normalisation, segmentation into smaller signals of duration 1.2 times the nominal heartbeat period, and zero-padding to align signal array size.

The impact of data quality on model performance was evaluated using four commonly used modelling approaches for classification tasks: K-Nearest Neighbour (KNN), Random Forest (RF), Artificial Neural Network (ANN), and Convolutional Neural Network (CNN). These four model families were chosen to provide complementary perspectives on the effect of data quality: KNN as instance-based, RF as an ensemble of decision trees, ANN as a fully connected neural network, and CNN as a learned feature extractor for temporal morphology. Default hyperparameters were intentionally used to isolate the effect of data quality on model behaviour.

Accuracy is interpreted as the noisiness of the dataset and is quantified using SNR, SampEn, and QRS_RP. Completeness involved quantifying signal padding (zeros inside the signal) and dimension padding (zeros before or after the signal) through FeatComp and TimeComp respectively. Precision was interpreted through the number of significant figures present in each measurement and evaluated through consistency of decimal places across the dataset and comparison of the dataset's mean decimal precision to an expected value based on ECG monitor specifications and analogue-to-digital converter parameters. To weight the SQIs, Principal Component Analysis (PCA) was used.

III. RESULTS

The classification and calibration reports were implemented using SCIKIT-LEARN. Each combination of model and dataset version was used to generate respective classification and calibration reports. Due to the real source of the data, patients with certain conditions are less frequent

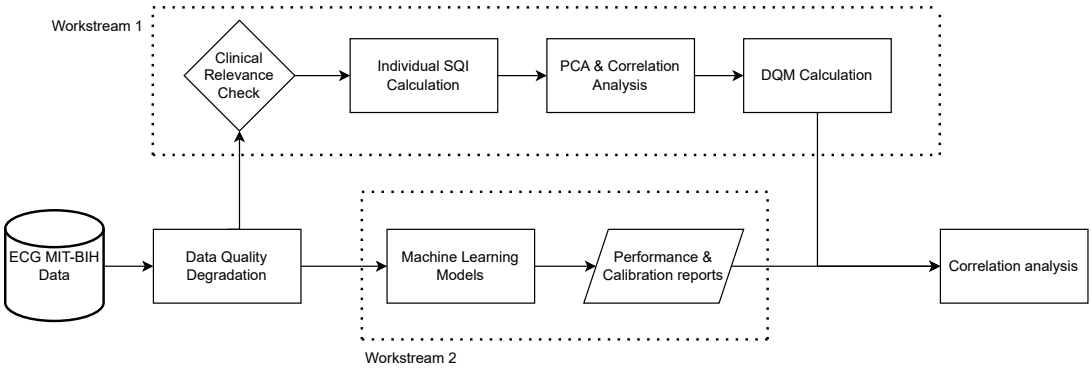


Fig. 1. Workflow for methodological implementation.

than normal and healthy patients, leading to significant class imbalance. To avoid this bias, macro-averaged values for Expected Calibration Error (ECE) and recall were used for comparison against the DQM.

SQIs were computed for each dataset. Individual SQIs were concatenated into a single matrix, which allowed for the use of PCA to establish weights. The PCA consistently determined that varying accuracy and signal duration, represented by TimeComp, had the greatest impact on data quality. This supports the results from the composite DQM where noisy datasets report the lowest DQM scores whilst the rounded datasets remained similar to the original.

When evaluating macro recall against the DQM score, a positive correlation was observed where recall improves with increases in composite SQI. When examining model calibration using macro ECE and the DQM score, a strong negative correlation was observed. Macro ECE values decrease as DQM scores increase. KNN displays the least reduction in macro ECE, whilst CNN and ANN show the most substantial decreases as data quality improves.

IV. DISCUSSION

The results of this study highlighted a relationship between the data quality of ECG signals and model performance. Calibration is closely linked to the inherent data quality of a signal, with degradation in data accuracy resulting in the most significant changes to model performance. This may be attributable to the degradation in data accuracy resulting in greatest variation, hence, the largest weight and contribution towards the DQM.

Higher DQM scores are associated with better class discrimination, suggesting models benefit from enhanced signal quality. These findings support the potential of a DQM to serve as an informative proxy to model robustness and calibration quality and leverage the possibility of determining a quality threshold for ML diagnostic tasks.

This work provides a foundational step toward a systematic understanding of the relationship between data quality and machine learning model behaviour. While several SQIs were tailored to ECG signals, the overall methodological framework could display broader applicability to signal quality assessment across domains.

REFERENCES

[1] "ISO/IEC 25012:2008 software engineering—software product quality requirements and evaluation (SQuaRE)—data quality model," technical report, International Organization for Standardization, Geneva, Switzerland, 2008.

[2] R. Miller, H. Whelan, M. Chrubasik, D. Whittaker, P. Duncan, and J. Gregório, "A framework for current and new data quality dimensions: an overview," *Data*, vol. 9, no. 12, p. 151, 2024.

[3] U. Satija, B. Ramkumar, and M. S. Manikandan, "A review of signal processing techniques for electrocardiogram signal quality assessment," *IEEE Reviews in Biomedical Engineering*, vol. 11, pp. 36–52, 2018.

[4] G. D. Clifford, J. Behar, Q. Li, and I. Rezek, "Signal quality indices and data fusion for determining clinical acceptability of electrocardiograms," *Physiological measurement*, vol. 33, no. 9, pp. 1419–1433, 2012.

[5] J. Gregorio, P. Duncan, M. Alsuleman, M. Chrubasik, B. Cieszynski, V. Desai, J.-L. Hippolyte, P. Krishnadas, K. Lines, R. Miller, N. Perić, and H. Whelan, "Data quality good practice guide," NPL Report MS 63, National Physical Laboratory, 2026.

[6] G. B. Moody and R. G. Mark, "The impact of the mit-bih arrhythmia database," *IEEE Engineering in Medicine and Biology Magazine*, vol. 20, no. 3, pp. 45–50, 2001.

[7] S. Fazeli, "ECG heartbeat categorization dataset." `https://www.kaggle.com/datasets/shayanfazeli/heartbeat`, 2018. Segmented and preprocessed ECG signals for heartbeat classification. Accessed: 2026-03-23.

2026 IMEKO Joint Conference
TC7 - TC11 – TC13 - TC24 in conjunction with TC6 and Mathmet
Faro, Portugal, 27-29 October, 2026

Multi-Purpose Semantic Modelling in Radiofrequency Measurements

Beth Tiplady[1], Priyanka Bharti[1], Marina Romanchikova[1]

[1] *National Physical Laboratory, Hampton Road, Teddington, Middlesex, TW11 0LW,*
beth.tiplady@npl.co.uk, priyanka.bharti@npl.co.uk, marina.romanchikova@npl.co.uk

I. INTRODUCTION

The National Physical Laboratory provides a service for the test and evaluation of microwave and radiofrequency sources (RFService). A proof-of-concept data management framework (DMF) was designed and implemented for a single RFService instrument to increase data accessibility to the end users [1]. The system implemented a single measurement workflow, using a Model-View-Controller pattern implemented in Python and using in-house designed user interface based on Qt for Python library. The persistent data storage was realised by combing a Kafka Message Broker with multi-model database that supported time series and nested metadata storage. One shortcoming of this prototype system was the lack of a unified data model between the user interface, the data handling software, and the database.

The implementation showed a reduction in access times by up to 24 hours, increased interoperability by bringing data from three previously disjoint sources into a single reference database and avoided one complete data loss over a 24-hour time interval.

Moving beyond the prototype, the new DMF needed to reflect the complexity of the full RFService measurement setup. A typical RF measurement involves multiple instruments, each with their own output formats, settings, and required metadata annotations.

An example RF measurement workflow involving three instruments (frequency counter, phase noise analyser and spectrum analyser) is illustrated in Figure 1. Each of the instrument can be used several times within a single measurement, with similar or different setups. An individual data collection episode involving a single instrument is described as *Run*. The data collected from all Runs undergoes many instrument-specific processing steps to yield the results to the customer. Note that the flow varies by instrument – not all instruments require further processing of data, as they perform on-instrument processing and return the data as already processed results.

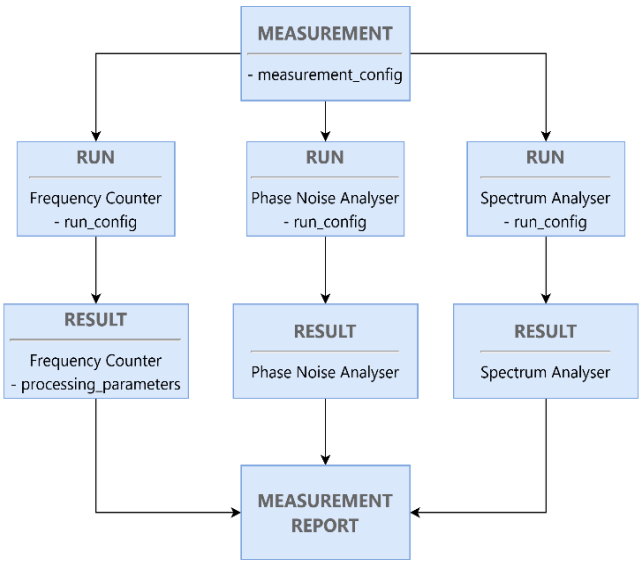


Fig 1: RF data pipeline for a measurement involving three instruments

At the moment, the traceability of the results to the raw data is largely manual and lacks a unified semantic model (or data schema) to represent these complex relationships and transformation steps. The new DMF design presented here aims to combine different elements of the RF measurement workflow – the control of RF instruments, the gathering of output data and metadata, and the documentation of processing steps – into a single repository to guarantee full traceability and reproducibility of results.

II. SEMANTIC MODELLING

Although the technical realisation of the data processing pipeline posed software and data engineering challenges, they are of low interest to scientific community. The technical challenge of this work was representing the multi-instrument multi-run multi-step measurement workflow using JavaScript Object Notation (JSON) schema - a tool not originally designed to describe complex scientific experiments.

The JSON schema approach was selected as it allowed to fulfil following goals:

1. **Validation**: It offers out-of-the box metadata model and validation toolkits.
2. **Adaptability**: JSON files can be understood and edited by the service users with minimal semantic technology experience.
3. **Metadata entry**: It will be used to automatically generate a graphical user interface (UI) for metadata entry with integrated data validation.
4. **Provenance**: Validated JSON (meta)data entered via the UI can be uploaded to the unified database where they are unambiguously associated with raw instrument outputs.
5. **Advanced semantic capabilities**: Schema files can be extended to JSON Linked Data (JSON-LD) format to explicitly assign metadata entries to ontology terms from commonly used ontologies such as QUDT, Semantic Sensor Networks etc.
6. **Re-use**: the same schema is used to automatically create the UI elements, the semantic model for the data handling software (controller), to store the metadata in the database, and to validate the data entry.
7. **Extensibility**: the schema can be easily expanded and modified to include new instruments and data sources.

The semantic model was refined in multiple rounds of revision with subject matter experts and service users. The updated model (Fig. 2) distinguishes between *static* properties which can be re-used between measurements (e.g. instrument configuration and reference signal definitions), and *dynamic* properties which vary between measurements (e.g. instrument roles) or single data acquisitions (*runs)*.

Instruments can assume various *roles* between measurements: the same instrument may act as Device Under Test (DUT) in one measurement and as a reference in another. To represent multi-instrument connectivity (one instrument can be connected to several others), we introduced the concept of *input sources*. An input source describes which channel of which instrument is connected to the target channel. This is a stark difference to the early prototype [1] that modelled channel inputs as static text labels.

The example in Fig. 2 shows a measurement that includes Spectrum Analyser *sa_02* as the acquisition instrument, configured as defined in "static properties" panel. The *static* metadata elements (f. e. channel IDs) change very rarely or never and can be re-used, rather than re-entered by the user in every measurement.

In the run 001, sa_02 is connected to UTC(NPL) reference signal on channel 1, to channel A of DUT 1 on channel 2, while channel 3 is not used. The input source properties are described in the static panel. The instruments and channel assignments may remain the same or change between the runs.

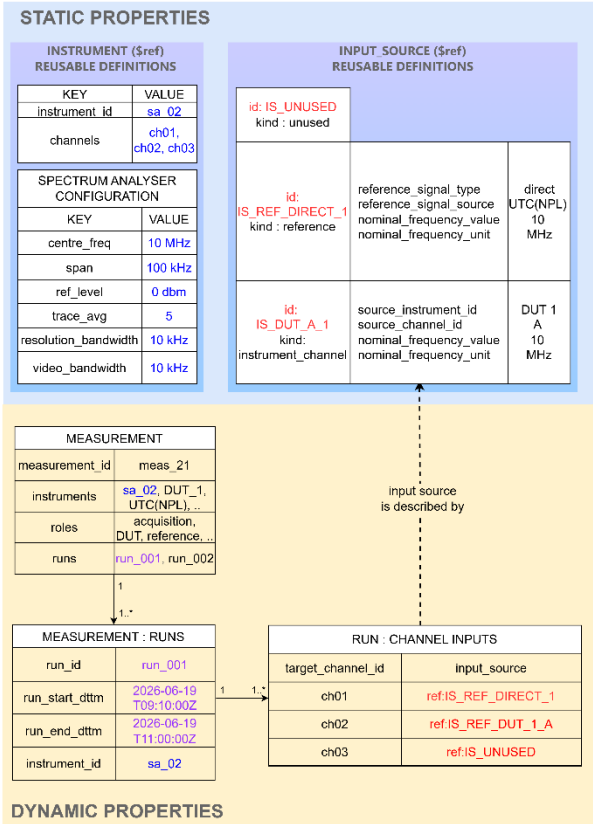


Fig 2: Static and dynamic instrument properties and connections in an example data acquisition.

III. CONCLUSIONS AND FUTURE WORK

This early stage work shows potential to use the JSON schema as the core of the RFService DMF. The model presented above features unified description of the instruments, instrument control settings, and data streams. RFService data post-processing and reporting models will be added later.

The prototype semantic model of the instruments will be integrated and trialled in backend (database-facing) and frontend (user-facing) interactions. If successful, further instruments will be added, as RFService capabilities develop. The semantic aspects of the work will include expanding the early schemas to JSON-LD and implementing a knowledge graph that will interact with the relational database where metadata are stored.

REFERENCES

[1] M Romanchikova et al, "Future-proofing digitalisation of a frequency measurement service", Measurement: Sensors, 2025, vol. 38, supplement

2026 IMEKO Joint Conference
TC7 - TC11 – TC13 - TC21 in conjunction with TC6 and Mathmet
Faro, Portugal, 27-29 October, 2026

Development and implementation of digital calibration certificates at IPQ: from NMI to end-user

Carlos Pires[1], Gonçalo Carvalhal[1], João Abrantes[1], Isabel Lopes[2], Amauri Camargo[2], João A. Sousa[1]

[1,] Instituto Português da Qualidade, Rua António Gião, 2, 2829-513 Caparica, Portugal, carlosp@ipq.pt
[2] Instituto de Soldadura e Qualidade, Av. Prof. Dr. Cavaco Silva, 33, 2740-120 Porto Salvo, Portugal, ialopes@isq.pt

***Abstract* – The digital transformation of metrology is changing how calibration information is produced, exchanged and used by laboratories, customers and automated measurement systems. Digital Calibration Certificates (DCC) enable calibration results, measurement uncertainty, metrological traceability and contextual metadata to be represented in a structured, machine-readable format. This paper presents the development and implementation of DCC-based solutions at Instituto Português da Qualidade (IPQ) in two metrological domains: time and frequency, and mass. It reviews recent european developments coordinated by EURAMET, describes case studies covering DCC generation, and reports on a real-world case in which a DCC was issued and transferred to an external customer, the Metrology Laboratory at Instituto de Soldadura e Qualidade (ISQ), an independent calibration and testing laboratory. The work assesses data consistency, interoperability, and usability, and is aligned with the best practices and guides from recent EURAMET projects.**

I. INTRODUCTION

Digital Calibration Certificates (DCC) are a key element of the digital transformation in metrology, allowing calibration results, measurement uncertainty, metrological traceability and contextual metadata to be exchanged in a structured and machine-readable form. For Instituto Português da Qualidade (IPQ), the national metrology institute of Portugal, the adoption of DCC supports the modernization of calibration services and contributes to more interoperable, data-driven metrological processes. As a customer and recipient of the DCC, the Metrology Laboratory at ISQ gains significant benefits in terms of optimisation of internal processes, reduction of human error risk and contribution to the end-to-end digitalisation of the calibration certificate management process.

This work presents the development and implementation of DCC-based solutions at IPQ in two metrological domains: time and frequency, and mass. These domains involve different measurement models, data structures and reporting requirements, providing a relevant basis for testing the flexibility of DCC within a national metrology institute. Implementation has focused on ensuring consistency with existing calibration procedures, preserving metrological traceability and complying with the requirements of ISO/IEC 17025:2017 [1]. The work is also aligned with the objectives of the most recent EURAMET projects, that addresses uncertainty propagation, data quality metrics, SI traceability, risk analysis and automated software frameworks for distributed sensor networks; DCC provide a complementary digital layer that can be automatically interpreted and integrated into sensor network assessment methods, e.g., the FunSNM project [6].

II. UPDATE ON DEVELOPMENTS IN EUROPE

The harmonization of DCC across Europe is coordinated by EURAMET Technical Committee for Interdisciplinary Metrology (TC-IM), through Project #1448 “Development of digital calibration certificates” [2]. Building on early work in the EMPIR project SmartCom (17IND02) [3], this initiative defines minimum content and interface requirements, a “Calibration Certificate Cover” describing mandatory information, and interfaces between national implementations. Since 2024 the work has been reorganized into a combined Structure, Harmonization and Software Tools work package, addressing interoperability of data formats and use cases across user types.

The most widely adopted implementation is the PTB-led DCC XML schema, now in version 3.3.0 [4], which meets ISO/IEC 17025 requirements and has been piloted across automotive, pharmaceutical and process industries in several European countries. DCC combines a machine-readable XML core with a human-readable rendering, secured through cryptographic digital signatures to preserve integrity and authenticity. Capacity-building has been reinforced by the completed EMPIR project

DCC2GO (21SCP01) [5], which delivered a training compendium and a starter kit for national metrology institutes, supporting laboratories such as IPQ in moving from paper-based to digital certificates.

III. CASE STUDIES

In the Laboratory of Time and Frequency at IPQ, DCC development targeted the calibration of frequency standards and time-interval measurements, Figure 1, requiring representation of stability metrics, environmental conditions and traceability chains to UTC(IPQ). In the Laboratory of Mass at IPQ, DCC development targeted mass calibration, structuring measurement results, uncertainty budgets and reference standard metadata according to the existing paper-based procedures. Both implementations reused the PTB schema structure while adapting quantity-specific data elements and were validated against the corresponding calibration procedures in use.

```
▼<ns0:column scope="reference" dataCategory="TargetValue" measurand="Measure.Frequency" metaDataCategory="Data">
    <ns0:heading lang="pt">Freq referencia</ns0:heading>
    <ns0:heading lang="en">Reference frequency</ns0:heading>
    <ns0:unit>\hertz</ns0:unit>
    <ns0:valueXMLList>100.000048815 999.999872960002 99999.987598 999999.8753999999 9999999.002 19999997.53</ns0:valueXMLList>
  </ns0:column>
▼<ns0:column scope="itemIndication" dataCategory="Value" measurand="Measure.frequency" metaDataCategory="Data">
    <ns0:heading lang="pt">Freq OUT</ns0:heading>
    <ns0:heading lang="en">OUT frequency</ns0:heading>
    <ns0:unit>\hertz</ns0:unit>
    <ns0:valueXMLList>99.99964855 999.9996280000001 99999.90039999998 999999.01 9999991.014 19999980.88</ns0:valueXMLList>
  </ns0:column>
▼<ns0:column scope="reference" dataCategory="Value" measurand="Measure.frequency" metaDataCategory="Data">
    <ns0:heading lang="pt">Erro</ns0:heading>
    <ns0:heading lang="en">Error</ns0:heading>
    <ns0:unit>\hertz</ns0:unit>
    <ns0:valueXMLList>-0.000392265000021122 -0.000244960000318242 -0.0071980000229086 7 -0.865399999893204 -7.987999999895692 -16.65000000223517</ns0:valueXMLList>
  </ns0:column>
    </ns0:statement>
▼<ns0:statement statementId="statement02">
    <ns0:category>method</ns0:category>
    <ns0:heading lang="pt"/>
    <ns0:heading lang="en"/>
    <ns0:body lang="pt">A incerteza expandida apresentada, está expressa pela incerteza-padrão multiplicada pelo fator de expansão k = 2,00, a qual para uma distribuição
    com nef = ∞ graus de liberdade efetivos corresponde a uma probabilidade de 95 %, aproximadamente. A incerteza foi calculada de acordo com o documento Evaluation of
    Measurement Data - Guide to the expression of uncertainty in measurement, JCGM 100:2008 (GUM 1995 with minor corrections).</ns0:body>
    <ns0:body lang="en">The reported expanded uncertainty is stated as the standard measurement uncertainty multiplied by the coverage factor k = 2,00, which for a t-
    distribution with nef = ∞ effective degrees of freedom corresponds to a probability of approximately 95 %. The uncertainty was calculated according to the document
    "Evaluation of Measurement Data - Guide to the expression of uncertainty in measurement, JCGM 100:2008 (GUM 1995 with minor corrections)". </ns0:body>
  </ns0:statement>
```


Figure-1: Example of the implementation of a DCC to a time interval measurement, machine readable part.

Beyond internal validation, a real-world calibration case was carried out involving equipment owned by ISQ. A DCC was generated for the calibration result, digitally signed, and issued to the Metrology Laboratory at ISQ through a practical operational workflow, from certificate creation to secure transfer and reception. This exercise allowed the assessment of data consistency between the source and receiving systems, interoperability with the digital infrastructure of ISQ Metrology Laboratory, and the overall usability of the certificate for the end user. It also surfaced practical limitations, including schema versioning, signature verification on the ISQ side, and the coexistence of digital and paper-based workflows, that will inform the wider roll-out of DCC to additional IPQ calibration services.

IV. CONCLUSIONS

The implementation of DCC at IPQ in time and frequency, and mass domains demonstrates the flexibility of the DCC concept across different measurement models while preserving compliance with ISO/IEC 17025:2017. The real-world exercise of issuing a DCC to ISQ as an external equipment owner confirmed the feasibility of the end-to-end workflow and identified practical points for improvement ahead of broader deployment.

REFERENCES

[1]ISO/IEC, ISO/IEC 17025:2017, “General requirements for the competence of testing and calibration laboratories”, 2017.

[2]EURAMET, TC-IM Project #1448, “Development of digital calibration certificates”.

[3]EURAMET, EMPIR Project 17IND02 “SmartCom - Industrial Metrology and Smart Manufacturing”.

[4]PTB, “Digital Calibration Certificate (DCC)”, www.ptb.de.

[5]EURAMET, EMPIR Project 21SCP01 “DCC2GO - Supporting the implementation of Digital Calibration Certificates in the European metrology community”.

[6]EURAMET, EPM FunSNM Project, “Fundamentals for trustworthy sensor network metrology”.

2026 IMEKO Joint Conference
TC7 - TC11 - TC13 - TC21 in association with TC6 and Mathmet
Faro, Portugal, 27–29 October, 2026

Virtual Knowledge Graph for Air Quality Measurement Datasets

Johannes Kombe[1], Keith Lines[1], Jean-Laurent Hippolyte[1]

[1]*National Physical Laboratory, Teddington TW11 0LW, United Kingdom*

***Abstract* – This paper demonstrates the integration of the UK Heavy Metals Network Database with semantic web technologies and the SI Reference Point. It establishes machine-actionable links between environmental measurements and authoritative SI concepts, and thereby illustrates a practical pathway for enhancing legacy measurement databases with semantic interoperability, FAIR data principles, and digital traceability.**

I. INTRODUCTION

The digital transformation of measurement science is increasing demand for measurement information that is interoperable, traceable and machine-actionable. While initiatives such as the SI Digital Framework (SIDF) [1] support this transition, many valuable scientific datasets remain stored in legacy systems that were not designed for semantic interoperability or FAIR (Findable, Accessible, Interoperable and Reusable) data stewardship [2]. As a result, important contextual information often remains embedded within database structures, software systems and expert knowledge. This work demonstrates how introducing purely additive semantic features can support interoperability, FAIRness and digital traceability in a real-world scientific database system, without disrupting operation.

II. BACKGROUND

The UK Heavy Metals Network Database (HMND) underpins the UK Heavy Metals Air Quality Monitoring Network, a national monitoring programme managed by the National Physical Laboratory (NPL) on behalf of the Department for Environment, Food, and Rural Affairs (Defra) [3]. It measures concentrations of heavy metals in ambient air, and supports compliance with the 2010 UK Air Quality Standards Regulations [4]. In 2024, the network operated 23 monitoring sites across the UK, generating quality-assured measurements, metadata, and provenance information for pollutants. Beyond regulatory reporting, it supports scientific research, air-quality modelling, emissions inventory validation and pollution trend analysis.

The SIDF supports machine-actionable metrology through authoritative digital representations of metrological concepts and resources. A central component is the SI Reference Point (SIRP), a machine-actionable knowledge model based on the SI Brochure that provides persistent identifiers and semantic descriptions for SI units, quantities, prefixes, defining constants and related concepts [5]. By linking measurements to SIRP concepts, units and quantities become explicit semantic entities, improving interoperability, digital traceability, and establishing a machine-readable connection between environmental observations and authoritative SI concepts.

III. SEMANTIC TECHNOLOGIES

The foundation of semantic technologies is the ontology: a formal representation of concepts, relationships, and properties that define a domain. By making domain knowledge explicit, ontologies enable consistent interpretation by both humans and machines, improving interoperability, discoverability, and reuse - thus providing a foundation for machine-actionable information systems [6].

```
<#HMDBLocalSiteOperatorsTriplesMap>
    a rr:TriplesMap ;
    rr:logicalTable [ rr:tableName "hmnd_local_site_operators" ] ;
    rr:subjectMap [
        rr:template "http://npl.co.uk/hmdb/resource/lso/{LSO_ID}" ;
        rr:class hmdb_ontology:LSO
    ] ;
    rr:predicateObjectMap [
        rr:predicate dcterms:identifier ;
        rr:objectMap [ rr:column "LSO_ID" ; rr:datatype xsd:string ]
    ] ;
    rr:predicateObjectMap [
        rr:predicate foaf:title ;
        rr:objectMap [ rr:column "LSO_TITLE" ; rr:datatype xsd:string ]
    ] ;
    rr:predicateObjectMap [
        rr:predicate foaf:firstName ;
        rr:objectMap [ rr:column "LSO_FIRST_NAME" ; rr:datatype xsd:string ]
    ] ;
    rr:predicateObjectMap [
        rr:predicate foaf:lastName ;
        rr:objectMap [ rr:column "LSO_LAST_NAME" ; rr:datatype xsd:string ]
    ] ;
```

Fig. 1. R2RML mappings for local site operators.

Ontology-Based Data Access (OBDA) enables semantic access to relational databases without moving or duplicating the underlying data. In this work, we used the OBDA platform Ontop to expose the HMND as a virtual knowledge graph (VKG), allowing users to query measurement information through domain-level concepts while preserving the existing database infrastructure [7]. R2RML mappings, with an example shown in Fig. 1, define how ontology concepts correspond to database tables and relationships, and link the semantic and relational layers [8].

The ontology developed for this work captures the principal entities and relationships represented within the UK HMND. Development was guided by competency questions, ensuring that the ontology reflects end-user require-

```
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX hmdb_ontology: <http://npl.co.uk/hmdb#>

# ----------------------------------------------------------------------
# CQ4-Sample
# Query area: Sample
# Competency question: Filter SAMPLE_ID is part of which analysis batch?
# ----------------------------------------------------------------------

SELECT DISTINCT ?sample_id ?analysis_batch_id ?batch_label
WHERE {
  VALUES ?input_sample_id { "100/13090" }

  ?sample a hmdb_ontology:Sample ;
          dcterms:identifier ?sample_id ;
          hmdb_ontology:inAnalysisBatch ?analysis_batch .

  FILTER(STR(?sample_id) = STR(?input_sample_id))

  OPTIONAL { ?analysis_batch dcterms:identifier ?analysis_batch_id . }
  OPTIONAL { ?analysis_batch hmdb_ontology:analysisBatchLabel ?batch_label . }
}
```

Table | Response | Pivot Table | Google Chart | Geo

Showing 1 to 2 of 2 entries (in 9.863 seconds)

Search: Show 50 entries

	sample_id	analysis_batch_id	batch_label
1	100/13090	47	100-13090-1
2	100/13090	49	100-13088-13091-4

Fig. 2. Query associating samples to analysis batches.

ments rather than the structure of the underlying database. To maximise interoperability, the ontology was aligned with established vocabularies where possible.

IV. RESULTS

To evaluate the demonstrator, the derived competency questions were implemented, and executed through the Ontop SPARQL endpoint. The purpose of this evaluation was to determine whether the ontology, mapping layer, and VKG could successfully expose information that would otherwise require detailed knowledge of the underlying database schema. Fig. 2 shows an example query and its result returned by the OBDA system. Two matching records were returned, demonstrating that the OBDA architecture successfully exposes relationships between a `Sample` and an `AnalysisBatch`, as represented in the underlying database. Importantly, the query is expressed entirely in terms of domain concepts, so the user is not required to understand database tables, foreign-key relationships, or SQL join operations. Instead, Ontop translates the semantic query into the corresponding SQL required to retrieve the information. To users, the system appears as a knowledge graph while the relational database remains the authoritative source of information.

V. DISCUSSION

The demonstrator shows how semantic technologies can enhance existing scientific databases while preserving their operational role, infrastructure, and investment.

A key advantage of the VKG is that the authoritative data remains within the original database, and is accessed only at query time. The information is hence always up-to-date, and the semantic layer can be developed independently from the underlying database. Semantic identifiers, standardised vocabularies, and SPARQL-based access improve the findability, and accessibility of information, while alignment with established ontologies support interoperability, and reuse. By linking quantities and units to authoritative digital SI resources, the demonstrator extends traditional metrological traceability into the digital domain, and illustrates how legacy measurement systems can participate in the broader SIDF.

Collectively, these findings demonstrate the potential of combining semantic web technologies and the SIRP as a practical pathway towards more interoperable, traceable, and trustworthy digital measurement infrastructures.

VI. ACKNOWLEDGEMENTS

This work was funded by the UK Government's Department for Science, Innovation and Technology through the UK's National Measurement System programmes. We thank Katie Williams, David Butterfield, and Sam Coulson for providing the HMND, establishing the end-user requirements, and supporting the semantic implementation.

REFERENCES

[1] International Bureau of Weights and Measures (BIPM). The SI digital framework, 2026. Available online: https://si-digital-framework.org/ (accessed on 23 July 2026).

[2] Mark D. Wilkinson et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1):160018, March 2016.

[3] K R Williams et al. The uk heavy metals monitoring network: Annual report 2024. NPL Report 10.47120/npl.ENV 60, August 2025.

[4] The air quality standards regulations 2010. https://legislation.gov.uk, 2010. (accessed on 27 July 2026).

[5] International Bureau of Weights and Measures (BIPM). The SI reference point. Digital representation of the International System of Units, 2026. Available online: https://si-digital-framework.org/SI (accessed on 23 July 2026).

[6] Tim Berners-Lee et al. The semantic web. *Scientific American*, 284(5):34–43, May 2001.

[7] Diego Calvanese et al. Ontop: Answering sparql queries over relational databases. *Semantic Web*, 8(3):471–487, 2016.

[8] Souripriya Das et al. R2RML: RDB to RDF mapping language, 2012. Available online: http://www.w3.org/TR/r2rml/ (accessed on 23 July 2026).

2026 IMEKO Joint Conference
TC7 - TC11 - TC13 - TC21 in association with TC6 and Mathmet
Faro, Portugal, 27–29 October, 2026

Validation of LiDAR sensors in virtual testing environments

Callum Whyte[1†], Bartlomiej Cieszynski[1], Russell Miller[1], Jack Clarke[1], Takuma Martin-Ito[1], Vivek Desai[1], Maurizio Bevilacqua[1]

[1] *Informatics, Data Science and AI Department, National Physical Laboratory (NPL), Teddington, Greater London, United Kingdom*
[†]*callum.whyte@npl.co.uk*

I. INTRODUCTION

Virtual testing is an essential part of the validation and verification process of autonomous vehicles. It provides a controlled and repeatable testing environment that enables validation of an autonomous system's operational envelope and edge cases while mitigating the risks, costs, and practical constraints associated with physical testing. Virtual testing environments (VTEs) are simulations which model an object of interest within a specific realistic environment for the purpose of testing the behaviour or properties of said object. Simulated scenarios can be repeated under controlled conditions within a virtual test environment, supporting systematic performance evaluation and the generation of safety-case evidence.

Autonomous vehicle's perception systems commonly combine LiDAR (Light Detection and Ranging), RADAR (Radio Detection and Ranging), and camera sensors to provide robust and adaptive situational awareness [1]. For the results of virtual tests to be trusted, the VTE must accurately represent both the vehicle and its sensors, as well as the operational envelope in which it is expected to function. Therefore, the corresponding sensor models within the VTE must be validated. This can be achieved by assessing whether the simulated sensor outputs are representative of those observed from physical sensors under the same conditions. However, representative outputs requires both accurate sensor models and a virtual environment with sufficient fidelity.

At the time of writing, limited attention has been given to assurance of VTE sensor models, with no widely accepted framework for demonstrating simulation fidelity and fitness for purpose. The purpose of this work is to introduce mathematical metrics to validate virtual sensor models by comparing synthetic sensor data with corresponding physical sensor data. This report uses an off-road testbed containing a range of targets as a proof of concept to evaluate both the proposed quality metrics and the validation framework. The aim is to develop generalisable methods for assessing the fidelity of virtual sensor models through comparison with real-world measurements. Although the work considered LiDAR, RADAR, and camera sensors, this report focuses solely on LiDAR for brevity. The overall validation workflow consists of three stages: data collection, sensor-specific processing and metric application/target identification.

II. METHODS

A. Data collection

Physical sensor data were collected at a test site in Cardington, Bedfordshire. The site contains targets of different sizes and shapes, with their positions recorded as three-dimensional ground-truth coordinates. The wider data collection included LiDAR, RADAR and camera measurements, although the present analysis focuses on one physical LiDAR sensor. A digital representation of the Cardington site was produced using a LiDAR survey, drone imagery and known target locations. Corresponding synthetic sensor data were then generated using commercially available simulation software.

B. Data preprocessing

Raw LiDAR scans may contain thousands, if not millions, of points, including distant background structures, ground returns, measurement noise and sensor artefacts. A four-stage preprocessing pipeline was implemented:

1. Region-of-interest filtering isolates the relevant part of the scene and removes non-informative data.
2. Voxel downsampling divides the cloud into a regular three-dimensional grid and replaces the points within each voxel with a representative point. This reduces computational cost while maintaining the main scene geometry.
3. Statistical outlier removal identifies isolated points using their mean distance to neighbouring points and removes those that are inconsistent with the local structure.
4. Ground removal using iterative Random Sample Consensus plane fitting. A surface-normal constraint ensures that accepted planes have an orientation consistent with the expected ground geometry.

Together, these operations produce cleaner, more consistent and computationally manageable point clouds for subsequent registration, comparison and object detection.

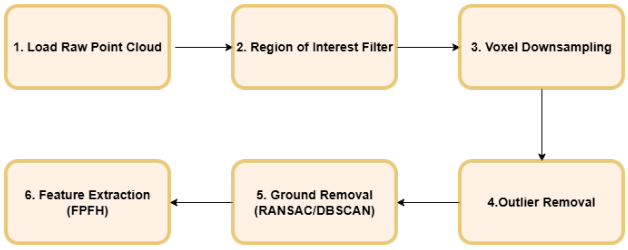


Fig. 1. Preprocessing workflow illustrating the sequence of operations applied to raw LiDAR point clouds

The preprocessing pipeline is visualised in Fig 1.

C. Point cloud registration and comparison

Meaningful point-by-point comparison requires physical and simulated clouds to be represented in a common coordinate frame. The registration pipeline begins with Principal Component Analysis-based coarse alignment. Fast Point Feature Histograms are then used to describe the local surface geometry around individual points. Correspondences between these descriptors support global registration using Random Sample Consensus (RANSAC), after which Iterative Closest Point (ICP) refinement minimises the remaining local alignment error.

No single metric can fully describe the agreement between two point clouds. Therefore a set of point cloud comparison metrics was chosen that assess structural, distributional, and density features of the LiDAR point clouds: Chamfer Distance [2], Earth Mover's Distance [2], Hausdorff Distance [3], Intersection over Union [4], and Point Cloud Density[5].

D. Target detection and classification

A rules-based target-detection pipeline was developed to demonstrate the complete LiDAR validation framework. Following ground removal, Density-Based Spatial Clustering of Applications with Noise (DBSCAN) identified candidate objects, and geometric features such as size, shape, planarity and point distribution were extracted for classification. Although a machine learning approach would be more robust to noise, occlusion and scene variation, the rules-based method was sufficient to demonstrate the end-to-end workflow. Developing and training a machine learning model was beyond the available project time, so rules-based classification was retained as a proof of concept for the end-to-end framework.

III. RESULTS AND DISCUSSION

Reliable registration between the real and simulated Cardington scenes was not achieved. The principal difficulty was a structural mismatch in the representation of the scene, leading to local surface normals and Fast Point Feature Histograms (FPFH) descriptors differing fundamentally between the two datasets, in addition to differences in ground representation and noise characteristics.

Because no valid real-to-simulated alignment was obtained, similarity metrics were not reported for this comparison, as their values would not be meaningful or reliably interpretable. Importantly, this limitation cannot be attributed solely to the LiDAR model. It demonstrates that sensor-model validation depends heavily on correct environmental alignment and on whether or not the virtual environment contains physically represented geometry with which the sensor model can interact.

Instead, a controlled perturbation was applied to a physical point cloud. Independent zero-mean Gaussian noise with a standard deviation of $0.5\,\mathrm{m}$ was added to each coordinate, 5 % of the points were removed, and a rotation of approximately $3.4\,\mathrm{deg}$ was applied about the vertical axis. Following registration, the applied transformation was recovered with a rotation error of $0.0047\,\mathrm{deg}$ and a translation error of $0.0034\,\mathrm{m}$, demonstrating that the pipeline can recover a known transformation despite moderate noise and point dropout.

The metrics described in section C. were applied to the full-resolution point clouds post-alignment. For the controlled example, the mean nearest-neighbour distance was $0.112\,\mathrm{m}$. This exceeds the approximately $0.080\,\mathrm{m}$ displacement expected from the applied Gaussian noise, likely because point dropout left some sparse, far-field points without close counterparts. The Hausdorff distance was $3.861\,\mathrm{m}$, demonstrating its sensitivity to the single worst-matched point. Voxel Intersection over Union was 0.534, indicating that even relatively small displacements can cause points to cross voxel boundaries. These findings show that metric parameters must be selected with respect to the expected noise level and point-cloud resolution.

REFERENCES

[1] Vargas, Jorge, et al. "An overview of autonomous vehicles sensors and their vulnerability to weather conditions." Sensors 21.16 (2021): 5397.

[2] Wu, Tong, et al. "Density-aware chamfer distance as a comprehensive metric for point cloud completion." arXiv preprint arXiv:2111.12702 (2021).

[3] Taha, Abdel Aziz, and Allan Hanbury. "An efficient algorithm for calculating the exact Hausdorff distance." IEEE transactions on pattern analysis and machine intelligence 37.11 (2015): 2153-2163.

[4] Yu, Jiahui, et al. "Unitbox: An advanced object detection network." Proceedings of the 24th ACM international conference on Multimedia. 2016.

[5] Li, Leihui, and Xuping Zhang. "A robust assessment method of point cloud quality for enhancing 3D robotic scanning." Robotics and Computer-Integrated Manufacturing 92 (2025): 102863.

2026 IMEKO Joint Conference
TC7 - TC11 - TC13 - TC21 in association with TC6 and Mathmet
Faro, Portugal, 27–29 October, 2026

Applying Data Quality Metrics to AIS Communications

Russell Miller*[1], Matthew Chan[1], Joao Gregorio[1], Maurizio Bevilacqua [1]

[1]*Informatics, Data Science & AI Department, National Physical Laboratory, Glasgow, UK*
**Russell.Miller@npl.co.uk*

I. BACKGROUND

Maritime autonomy is advancing rapidly, with uncrewed surface vessels (USVs) and remotely operated vehicles (ROVs) increasingly managed from Remote Operating Centres (ROCs). ROCs provide centralised oversight of multiple offshore assets, improving operational efficiency, reducing costs, and enhancing safety through real-time monitoring and decision-making. However, reliable communication between shore-based operators and offshore systems remains a key challenge. ROCs depend on continuous connectivity to receive telemetry data, maintain situational awareness, and issue command and control instructions. In maritime environments, communication links are often constrained by limited bandwidth, high latency, weather-related signal degradation, interference, and the inherent limitations of satellite communications. These constraints can reduce data availability and integrity, potentially affecting operational effectiveness and safety.

To address this challenge, data quality (DQ) metrics provide a systematic means of assessing and quantifying the reliability and trustworthiness of information exchanged between ROCs and maritime assets. This research demonstrates the practical application of a previously developed DQ framework [1] to maritime communications using Automatic Identification System (AIS) data as an example dataset. While AIS uses low-bandwidth VHF communications and does not fully replicate ROC communication systems, it exhibits many of the same data transmission challenges, including delays, data loss, and incomplete messages [2]. Its widespread adoption and standardised format make it a suitable use-case for demonstrating how DQ metrics can be used to assess communication performance and provide assurance in operational decision-making.

II. METHODS

AIS messages collected within the Plymouth maritime area were pre-processed and organised by MMMSI providing a unique identifier for each vessel. Messages were chronologically ordered to reconstruct vessel trajectories and identify discrete periods of presence within the study area. This allowed intervals associated with vessels leaving and later re-entering reception coverage to be separated from communication-induced reporting delays. Vessel observations were subsequently categorised according to speed and manoeuvring status, reflecting the reporting regimes specified in IMO Resolution A.1106(29) [3].

Data quality assessment focused on the dimensions of timeliness and compliance. Timeliness was evaluated through analysis of inter-message reporting intervals to identify delayed or irregular transmissions indicative of communication degradation. As AIS reporting rates vary according to vessel speed and manoeuvring state, messages were partitioned into groups representing distinct operating conditions, with a new group created whenever a vessel changed speed category, manoeuvring status, or course-alteration state. Gamma distributions and linear thresholding approaches were applied to the interval data within each group to characterise expected reporting behaviour and identify statistical outliers. In addition to quantifying the occurrence of delayed transmissions, the severity of latency events was assessed by measuring the extent to which reporting intervals exceeded expected values, enabling differentiation between minor delays of a few seconds and more substantial disruptions that may affect operational awareness.

Compliance was assessed independently of latency by comparing observed reporting intervals against the transmission schedules specified in IMO Resolution A.1106(29) for each vessel operating condition. Whereas timeliness characterises deviations from expected reporting behaviour within a dataset, compliance evaluates adherence to regulatory requirements. A transmission may therefore exhibit elevated latency relative to neighbouring messages while still remaining compliant with IMO reporting standards. Completeness was also assessed through the presence of mandatory AIS attributes, including vessel identity, position, speed, course, and timestamp information. Together, these metrics provide a structured assessment of communication performance and quantitative evidence of the reliability and trustworthiness of information available to Remote Operating Centre (ROC) operators.

III. RESULTS

Compliance analysis showed that approximately 44% of AIS transmissions met the reporting intervals specified in IMO Resolution A.1106(29). Allowing a tolerance of 2 s increased this value to 78%, indicating that many deviations from the prescribed transmission schedule were

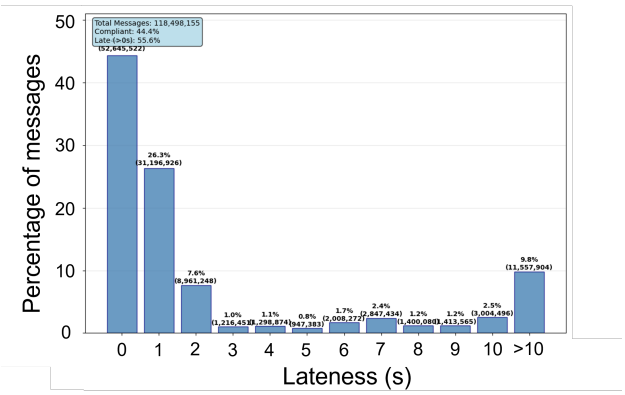


Fig. 1. AIS transmission compliance against IMO reporting requirements. Approximately 44% of messages were compliant, increasing to 78% within a 2 s tolerance. Around 10% of transmissions exhibited delays greater than 10 s.

relatively small. However, approximately 10% of transmissions exhibited delays exceeding 10 s, suggesting the presence of substantial reporting irregularities that cannot be explained by minor timing variations alone. Figure 1 shows the distribution of AIS message delays relative to the reporting intervals specified by the IMO.

Spatial analysis (shown in figure 2 of timeliness across the Plymouth study area revealed a general increase in transmission delays with distance from the receiving base station. Mapping latency metrics onto a 1 km² spatial grid enabled visualisation of areas exhibiting reduced communication performance, demonstrating how DQ metrics can be used to identify regions where communication reliability may be degraded.

IV. DISCUSSION

The results indicate that most AIS transmissions were either compliant with IMO requirements or deviated only marginally from the expected reporting interval. Nevertheless, the presence of a notable proportion of delays greater than 10 s suggests occasional communication disruption, message loss, transmission contention, or limitations associated with shore-side reception and message processing. As reception timestamps were used as a proxy for transmission time, some measured delays may also reflect infrastructure and processing effects rather than propagation delays alone.

Although AIS is not a direct representation of ROC communication architectures, the study demonstrates how DQ metrics can provide objective evidence of communication performance and regulatory compliance. The observed increase in latency with distance is consistent with expected characteristics of VHF-based communications,

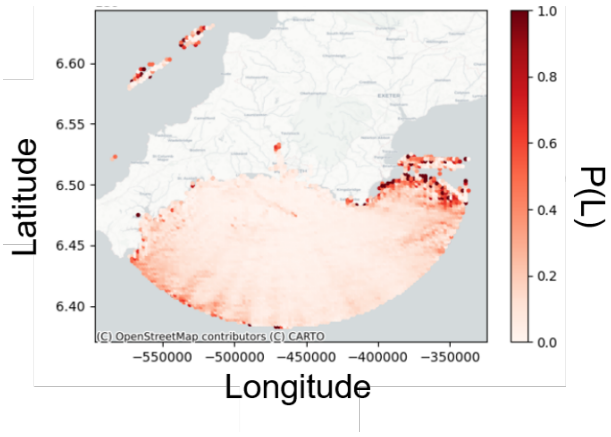


Fig. 2. Spatial distribution of latency probability across the Plymouth study area using a 1 km^2 grid. The map provides a geographic visualisation of AIS performance and indicates a general increase in latency occurrence with distance from the Plymouth receiving station.

although further statistical analysis is required to quantify this relationship. The spatial visualisation approach shows particular promise for operational use, enabling communication performance to be represented geographically and supporting the identification of areas where reduced data quality may affect situational awareness and decision-making. These findings demonstrate the practical application of DQ metrics as an assurance mechanism for maritime communications and provide a foundation for future assessment of higher-bandwidth ROC communication systems

REFERENCES

[1] R. Miller, H. Whelan, M. Chrubasik, D. Whittaker, P. Duncan, and J. Gregório, "A framework for current and new data quality dimensions: An overview," *Data*, vol. 9, no. 12, 2024.

[2] A. Veenstra and R. Harmelink, "Ais data quality assessment," Working Paper 25/001, MEL Centre for Maritime Economics and Logistics, Rotterdam, The Netherlands, Dec. 2025.

[3] International Maritime Organization, "Resolution A.1106(29): Revised Guidelines for the Onboard Operational Use of Shipborne Automatic Identification Systems (AIS)," tech. rep., International Maritime Organization, London, UK, Dec. 2015. Adopted 2 December 2015.

2026 IMEKO Joint Conference
TC7 - TC11 – TC13 - TC21 in conjunction with TC6 and Mathmet
Faro, Portugal, 27-29 October, 2026

Impact of Acquisition Parameters on the Metrological Accuracy of xCT Workflows

Teodor Tóth[1], Miroslav Dovica[1], Lucia Bednarčíková[1], Jakub Palenčár[2], Goran Jotic[3]

[1] *Technical University of Košice, Faculty of Mechanical Engineering, Slovakia, teodor.toth@tuke.sk, miroslav.dovica@tuke.sk, lucia.bednarcikova@tuke.sk*
[2] *Slovak University of Technology in Bratislava, Faculty of Mechanical Engineering, Slovakia, jakub.palencar@stuba.sk*
[3] *University of Banja Luka, Faculty of Mechanical Engineering, Bosnia and Herzegovina, goran.jotic@mf.unibl.org*

***Abstract* – This paper evaluates the influence of data acquisition parameters and system generations in industrial computed tomography against a tactile coordinate measuring machine (CMM) reference. A complex aluminum alloy sample was analyzed under varying projection counts, voxel sizes, and integration setups using two CT architectures. The evaluation demonstrates that linear dimensions remain stable under Gauss-fitting algorithms. Conversely, geometric form tolerances (GD&T) are highly sensitive to voxel mesh quality, high-frequency noise, and digital sampling density. The step width has a significant impact on the accuracy of measured features (plane, cylinder) and characteristics (flatness, cylindricity) by scanning, especially at lower resolution or fewer projections obtained from CT.**

I. INTRODUCTION

Industrial X-ray Computed Tomography (xCT) is rapidly becoming a standard method for the dimensional and geometrical verification of industrial components, as well as for non-destructive internal defect inspection [1]. However, establishing traceability and evaluating measurement uncertainty remains a challenge due to the complex interaction of acquisition settings (voltage, projection count, integration time) and post-processing parameters (surface extraction grid density). Kruth et al [1] also investigated techniques to improve spatial resolution during CT scanning and highlighted the critical components influencing the overall imaging outcomes.

Liu et al. [2] conducted a comparative study between measurements obtained from xCT datasets and data from a coordinate measuring machine (CMM). Their analysis focused on how the mechanical filtering effect of the CMM probe, voxel size error, and beam hardening elimination in the CT workflow affect the final measurement accuracy. Pirillo et al. [3] addressed the challenges of surface determination, demonstrating that the results were largely independent of key acquisition parameters such as voltage, integration time, and sample orientation. Regarding data density, Villarraga-Gómez [4] evaluated the influence of the projection count on dimensional performance. Their results indicate that the number of projections heavily affects length parameters and form tolerances, while exhibiting negligible impact on hole diameters.

This contribution focuses on evaluating the influence of scan acquisition parameters on dimensional accuracy to optimize the trade-off between inspection time and data quality, utilizing tactile CMM measurements as the traceable reference. The experimental design systematically isolates three critical metrological variables. First, it investigates the impact of mechanical filtering caused by different CMM tip diameters (1.5 and 3.0 mm) and the impact of the measurement step width on the CMM measurement as well as on the generated CT data and their combination. Second, it evaluates the geometric degradation and tomographic reconstruction noise introduced by downsampling the angular projections. Finally, the study benchmarks the capabilities of a multi-purpose CT system operating under varying voxel resolutions against a legacy, dedicated metrological CT architecture.

II. MATERIALS AND METHODS

Data acquisition was performed using an industrial multi-purpose Phoenix v|tome|x L 240 (General Electric, USA) and a dedicated metrological Metrotom 1500 Gen 1 (Carl Zeiss, Germany) XCT system. To optimize time efficiency (t) on the GE system, projection averaging (Av), number of projection (N), and voxel sizes (Vx) were varied as key throughput parameters, while the Metrotom operated under a single stable metrological configuration. The comprehensive scanning parameters are detailed in Table 1. For all tomographic scans, the artifact was securely positioned inside a custom polystyrene fixture to maintain an identical orientation across all experiments.

A key hardware variance lies in detector specifications,

where the GE system utilizes a 2048 × 2048 layout (200 µm pitch) compared to the legacy Metrotom's 1024 × 1024 layout with a coarser 400 µm pixel pitch.

Table 1. Evaluated xCT scanning configurations

Strategy	N	Vx (µm)	Av	t (min)
Normal	2300	70	Y	91
Fast	2300	70	N	94
Projection	1150	70	Y	53
Resolution	1200	140	Y	55
Metrotom	1000	120	N	84

Reference measurements were performed on a ZEISS Contura G2 coordinate measuring machine (Carl Zeiss, Germany) equipped with a VAST XXT TL1 scanning probe mounted on an RDS articulating head, under a controlled temperature of 20°C (with a maximum drift of ±0.1°C during the measurements). To quantify the mechanical filtering effect of the stylus tip, data were captured using two different stylus diameters: an Ø 1.5 mm probe with a 20 mm length and a 20 mm extension (meeting the 40 mm minimum required probe length), and a standalone Ø 3.0 mm probe with a 50 mm length. The CMM reference values were obtained by averaging 5 consecutive measurement runs. The experimental object was a milled artifact machined from an EN AW-5083 (AlMg4.5Mn0.7) aluminium alloy (Fig. 1)

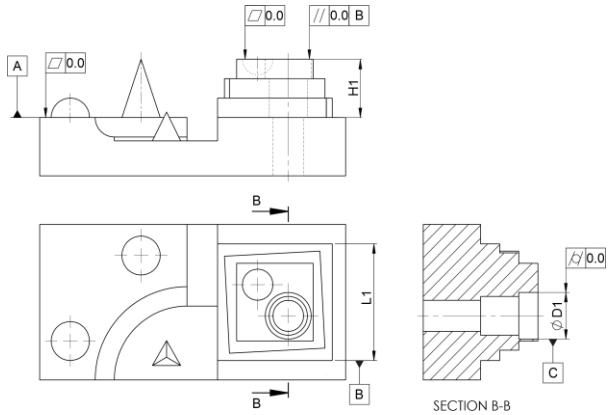


Fig. 1. Drawing with measured GD&T

The reconstructed CT volumetric datasets were processed in VGStudio MAX (Volume Graphics, Germany) utilizing Automatic Surface Determination in Advanced Mode with the Remove particles and all voids option enabled.

Dimensional and geometric evaluations for CMM and STL meshes were executed within ZEISS Calypso (Carl Zeiss, Germany) using an identical measurement program. All features were captured in continuous scanning mode at a speed of 3 mms^{-1} comparing two step width options (0.1 mm and 1.0 mm) to isolate the mathematical impact of data point density. Feature calculations utilized standard Gauss (least-squares) fitting models integrated with an extreme value elimination filter thresholded at 3σ. The metrological scope covers linear dimensions (L1, H1, D1) alongside GD&T evaluations of flatness, cylindricity, and parallelism (see Fig. 1).

III. CONCLUSION

Benchmarking against the tactile reference (CMM, D1.5, Step width 0.1) reveals distinct error sources in xCT metrology. The GE Phoenix system operating under high-resolution baseline parameters exhibits tight agreement with the CMM, with a diameter deviation (D1) of only 1.9 µm and flatness errors within 6.8 µm, demonstrating excellent laboratory process capability despite the absence of an OEM-specified MPE. Deviations in scanning with Reduced projections and fast scanning are driven by high-frequency voxel noise and projection aliasing, which degrade shape definitions. Conversely, the legacy Zeiss Metrotom 1500 confirms its robust metrological design the dimensions (D1, L1) are within the specified MPE (below 7.2 µm). However, the Metrotom's legacy 1K detector introduces severe edge blurring, inflating flatness deviations up to 61.5 µm. This form error affects the dependent orientation tolerances like parallelism (57.4 µm) and height H1.

IV. ACKNOWLEDGEMENT

The work was supported by the projects VEGA 1/0191/24, VEGA 1/0363/26 and KEGA 004TUKE-4/2024, KEGA 025TUKE-4/2025.

REFERENCES

[1]J.P.Kruth, M.Bartscher, S.Carmignato, R.Schmitt, L.De Chiffre, A.Weckenmann,"Computed tomography for dimensional metrology", CIRP Annals, Volume 60, Issue 2, 2011, pages 821-842, ISSN 0007-8506, https://doi.org/10.1016/j.cirp.2011.05.006.

[2] W.Liu, X.Chen, W.Zeng, W.Sun, D.Gorman, A.Wilson, Q.Qi, P.Scott, X.Jiang, S.Lou, "Comparison of X-ray computed tomography and coordinate-measuring system dimensional measurement for additive manufacturing parts using physical and simulation methods", Measurement, Volume 229, 2024, ISSN 0263-2241, https://doi.org/10.1016/j.measurement.2024.114414

[3] F.Pirillo, H.Shao, S.Petrò, G.Moroni, “A Surface Determination Technique for Dimensional and Geometrical Analysis in Industrial X-ray Computed Tomography”, Nanomanuf Metrol, vol.7, No.20, 2024, https://doi.org/10.1007/s41871-024-00239-4

[4] H.Villarraga-Gómez, S.T.Smith, "Effect of the number of projections on dimensional measurements with X-ray computed tomography", Precision Engineering, vol.66, 2020, Pages 445-456, https://doi.org/10.1016/j.precisioneng.2020.08.006.

2026 IMEKO Joint Conference
TC7 - TC11 – TC13 - TC24 in conjunction with TC6 and Mathmet
Faro, Portugal, 27-29 October 2026

Agent-Based Knowledge Graph Generation for Scientific Experiments

Moulham Alsuleman[1], Harvey Whelan[1], João Gregório[1], Michael Chrubasik[1]

[1] *1 National Physical Laboratory, Glasgow, G1 1RD, moulham.alsuleman@npl.co.uk*

I. INTRODUCTION

Scientific experiments generate far more information than the numerical values reported in a dataset [1]. Experimental outcomes are influenced by experimental methods, operating procedures, instrument identity, calibration status, configuration settings, and data-processing methods [2]. However, this information is typically distributed across instrument manuals, standard operating procedures (SOPs), calibration certificates, laboratory notebooks, spreadsheets and reports [1]. Consequently, an exported measurement dataset alone rarely provides sufficient information to understand how a measurement result was produced or whether it is trustworthy.

This challenge is particularly important in metrology, where confidence in a measurement depends on traceability, calibration history, measurement quantities and units, uncertainty information, method definitions, and observational context [3]. Although these metadata may be available, they are often represented using inconsistent terminology and metadata structures, making integration and automated interpretation difficult. As a result, reconstructing experimental context, instrument provenance, and measurement relationships is frequently a manual, time-consuming process that requires substantial domain expertise [1].

The need to improve the Findability, Accessibility, Interoperability, and Reusability (FAIR) of scientific data has increased interest in methods for integrating experimental metadata across heterogeneous sources [4]. Semantic interoperability is especially important for enabling measurements, instruments, calibration records, and experimental procedures to be understood consistently across systems and organisations [3]. Knowledge graphs have emerged as a promising approach because they can represent entities, relationships, and provenance in a machine-readable form, enabling traceability queries, calibration tracking, and cross-system integration [1].

Current approaches to knowledge graph generation typically rely on manual ontology population, rule-based extraction, ETL pipelines, or direct document-to-graph generation using large language models (LLMs) [2]. While these methods can produce useful results, they often struggle to scale across diverse scientific documentation. In particular, direct document-to-graph generation often combines evidence retrieval, interpretation, and semantic modelling within a tightly coupled workflow, making it difficult to trace how individual graph assertions were derived or independently validate intermediate decisions.

To address these limitations, this work presents ODINstrument, an ontology-guided multi-agent architecture that separates evidence retrieval, interpretation, formalisation, and validation into independently inspectable stages. The architecture is demonstrated using a pharmaceutical tablet-testing system as a representative case study, demonstrating how scientific knowledge can be reconstructed from heterogeneous documentation while preserving traceability and semantic context.

II. METHOD

Rather than generating a knowledge graph directly from source documents, ODINstrument decomposes knowledge graph generation into a sequence of specialised agent-driven transformations. This staged design improves provenance and transparency by allowing each transformation to be independently inspected and evaluated. Knowledge extraction is guided by a predefined ontology that serves as the semantic target for graph generation. The ontology is organised into modelling patterns representing experimental concepts such as instrument identity, methods, calibration activities, runtime information, observations, and data mappings. These patterns guide information extraction and provide constraints for subsequent validation.

The workflow consists of five specialised LLM-based agents, each named after a mythological character according to their role. *Alviss* performs retrieval planning, *Muninn* retrieves supporting evidence, *Huginn* converts evidence into atomic facts, *Mimir* formalises facts as ontology-compliant RDF fragments, and *Forseti* aggregates and validates RDF graph fragments . Together, these agents implement the workflow shown in Figure 1.

Each intermediate ODINstrument output remains traceable to its originating evidence, supporting transparency, auditability, and independent validation. The workflow was evaluated using documentation and experimental data from a SOTAX AT50 pharmaceutical tablet-testing system, providing a realistic case study involving fragmented information distributed across calibration records, methods, reports, and data exports.

III. RESULTS

Across all evaluated ontology patterns, the workflow successfully reconstructed a broad range of scientific and metrological knowledge, including instrument identities, system structure, calibration requirements, runtime metadata, observations, statistical summaries, uncertainty-related concepts, and provenance information. A consistent finding was that retrieval and evidence interpretation recovered substantially more information than was ultimately represented in ontology fragments. Consequently, the primary source of information loss occurred during ontology formalisation rather than evidence discovery or interpretation.

Alviss generated ontology-guided retrieval questions that adapted to document-specific terminology and improved recovery of information distributed across heterogeneous sources. Muninn provided a provenance-preserving retrieval layer that recovered evidence from manuals, calibration records, methods, and data exports. Huginn transformed retrieved evidence into structured atomic facts while explicitly preserving uncertainty and missing information. Mimir demonstrated that ontology-guided formalisation is feasible; however, not all recovered relationships were preserved in ontology form, making formalisation the stage with the greatest opportunity for improvement.

Pattern-level evaluation reinforced these findings. Provenance, calibration knowledge, runtime metadata, observations, uncertainty information, and statistical summaries were reconstructed successfully through explicit links between ontology patterns, retrieval questions, evidence, and atomic facts. However, only a subset was represented in valid RDF/Turtle fragments.

IV. DISCUSSION

The results show that staged knowledge reconstruction recovers information while maintaining provenance.

The evidence-interpretation stage proved particularly valuable because it created an intermediate layer of atomic facts with explicit links to supporting evidence, enabling independent inspection and audit. This approach preserved uncertainty and missing information rather than replacing gaps with assumptions, improving explainability and reducing the risk of unsupported assertions.

In contrast, ontology formalisation was the most challenging stage of the workflow. While substantial knowledge was recovered during retrieval and interpretation, only part of this information was ultimately represented within ontology fragments. These findings suggest that ontology generation and relationship modelling remain important areas for further development.

Provenance-preserving intermediate representations keep recovered knowledge transparent and reusable even when ontology generation is incomplete. The findings therefore suggest that future progress in AI-assisted scientific knowledge reconstruction will depend less on improving retrieval and more on developing robust, validation-aware ontology formalisation techniques. Future work will focus on SHACL-based validation, graph aggregation, and improved ontology generation.

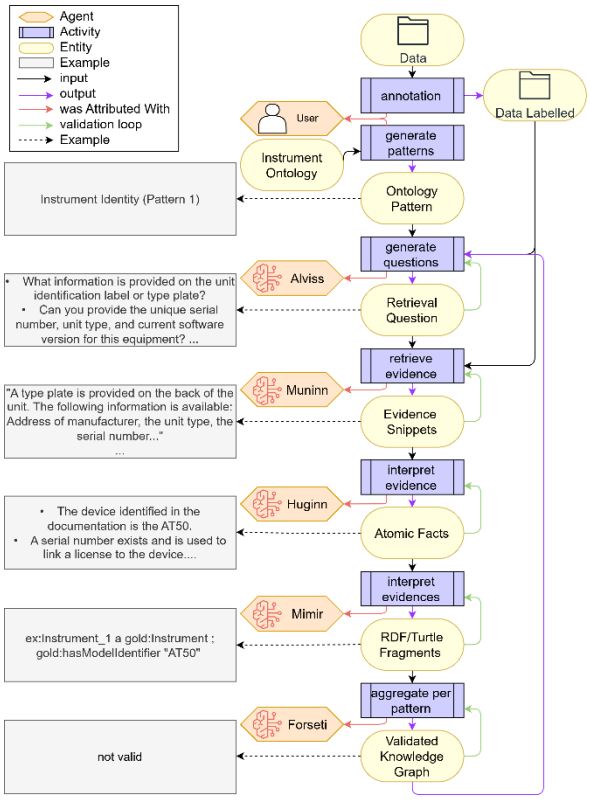


Figure 1. Example and illustration of the overall information flow through the architecture.

V. ACKNOWLEDGEMENTS

This work was funded by the UK Department for Science, Innovation and Technology (DSIT) through the National Measurement Service (NMS) programme. The authors thank CMAC, University of Strathclyde, for providing the data and industrial use case.

REFERENCES

1. M. Alsuleman et al., "Supporting Medicines Manufacturing through Semantic Technologies: Oligonucleotide Synthesis Use Case," Proc. HEALTHINF, pp. 408–418, 2026.
2. M. Hofer et al., "Construction of Knowledge Graphs: Current State and Challenges," Information, vol. 15, p. 509, 2024.
3. M. Compton et al., "The SSN Ontology of the W3C Semantic Sensor Network Incubator Group," J. Web Semant., vol. 17, pp. 25–32, 2012.
4. D. Garijo and M. Poveda-Villalón, "Best Practices for Implementing FAIR Vocabularies and Ontologies on the Web," Appl. Pract. Ontol. Des. Extr. Reason., 2020.

Abstracts accepted as ***poster presentations***

T1 - Measurement Theory and Foundations of Measurement Science

page

#4 Edita Raudienė, ***"Trusted Measurement Data through Optimized Uncertainty Evaluation for Science-Driven Decision-Making"*** *99*

#26 Kweonho Kang, Chang Hwa Lee, Jae Soo Ryu, ***"Standard Reference Data for Thermal Diffusivity of HANA-6"*** *101*

T2 - Testing, inspection and certification

#5 Asta Meškuotienė, ***"Technical challenges and interpretations: the interaction of legal metrology and accreditation requirements in laboratory practice"*** *103*

#47 Saowaluck Ukrisdawithid, ***"A Metrological Framework for the Comprehensive Annual Calibration of pH Meters"*** *106*

T3 - Measurements in Life Sciences

#21 Livio D'Alvia, venla korkko, ***"Experimental Microwave Array Sensing for Brain Tumor Detection: Sensitivity and Repeatability Evaluation in a Head Phantom Model"*** *108*

T4 - Mathematical and statistical modelling

#1 Salah BOUHOUCHE, ***"Extended robust algorithm (A) ISO 13528 to dynamic measurement for interlaboratory comparison"*** *110*

#32 Luís Alves, Dora Carinhas, Paulo Infante, José Rocha, ***"A Statistical Workflow for the Metrological Evaluation of a Thermoregulated Bath Used in Hydro-Oceanographic Sensor Calibration"*** *117*

2026 IMEKO Joint Conference
TC7 - TC11 – TC13 - TC24 in conjunction with TC6 and Mathmet
Faro, Portugal, 27-29 October, 2026

Trusted Measurement Data through Optimized Uncertainty Evaluation for Science-Driven Decision-Making

Edita Raudienė[1], Asta Meškuotienė[2]

[1] *Kaunas University of Technology, Faculty of Electrical and Electronics Engineering, Institute of Metrology, Kaunas, Lithuania,* edita.raudiene@ktu.lt

[2] *Kaunas University of Technology, Faculty of Electrical and Electronics Engineering, Institute of Metrology, Kaunas, Lithuania,* asta.meskuotiene@ktu.lt

Abstract – **Reliable scientific and engineering decisions require measurement results supported by transparent and traceable uncertainty evaluation. Although digital metrology software ensures computational consistency, the credibility of measurement results depends primarily on the quality of input data and the appropriate evaluation of uncertainty components. This work presents an integrated approach for improving the reliability of measurement data by optimizing uncertainty assessment in accordance with the Guide to the Expression of Uncertainty in Measurement (GUM). The proposed methodology analyses alternative Type A and Type B uncertainty evaluation methods and distinguishes the treatment of calibrated measuring instrument data for single and repeated measurements. A case study involving the volume determination of a steel cylinder demonstrates the practical implementation of the approach using a digital metrology workflow. The resulting uncertainty budget identifies the dominant contributors to the combined uncertainty, enabling evidence-based prioritization of measurement system improvements. The proposed methodology strengthens confidence in measurement results and supports trusted analysis for science-driven technological decision-making.**

I. INTRODUCTION

Modern science-driven engineering and manufacturing rely fundamentally on trusted data analysis for high-stakes decision-making. In metrology, a measurement result is legally and scientifically incomplete without an explicit quantitative statement of its uncertainty, governed internationally by the Guide to the Expression of Uncertainty in Measurement (GUM). Digital pipelines, such as GUM Workbench software, replace manual multi-variable equations prone to computational risks with traceable, automated processing chains. However, the overall reliability of measurement data largely depends not only on software modeling, but also on the strategic management of the initial data, the parameters of calibrated measuring instruments used in single or multiple measurements, and the evaluation of the final measurement result.

II. MATHEMATICAL MODELING AND CASE STUDY

To demonstrate the generation of trusted analytical data, the volume determination of a steel cylinder was modeled. The volume (V) is determined by its diameter (D) and its height (H), both measured repeatedly. To normalize the results to a baseline temperature (T_0 = 20 °C), environmental temperature (T) and the thermal expansion coefficient of steel (α_t) are integrated into the multi-variable model:

$$V = (\pi \cdot (D^2 + \Delta D)/4) \cdot (H + \Delta H) \cdot \left(1 + \alpha_t \cdot (T_0 - T)\right) \quad (1)$$

Where ΔD and ΔH represent dedicated correction factors mapping the calibrated tool's physical resolution thresholds.

III. STANDARD UNCERTAINTY TYPES AND COMPLIANCE

The reliability of the data depends on the accurate estimation of the standard uncertainty of the measurement based on the measurements of type A (statistical) and B (non-statistical). This is important when calculating the standard uncertainties of each input quantity. It is important to emphasize that in order to perform reliable measurements at the initial stage of data entry when calculating the standard uncertainty of the variance, it is important not only to choose the methods of type A and B, but also their variants, two cases of type A and three cases of type B estimation are analyzed in the article.

IV. CALIBRATED MEASURING INSTRUMENT USED FOR MULTIPLE AND SINGLE MEASUREMENTS

When a single measurement is carried out using a calibrated measuring instrument, a term reflecting the dispersion of the results, i.e. σ, must be used in the measurement uncertainty assonessment chain. Because systematic errors are identified during the calibration process, they can be estimated and accounted for in the measurement result, thereby reducing the instrument's error when required. As a result, the uncertainty associated with the same calibrated measuring instrument during

single measurements may be expressed in two different ways. If the corrected reading Xg is used as the final measurement result, the measuring instrument error is determined using the following formula:

$$X_g = X - \Delta; \quad \partial_1 = \alpha_\partial \cdot \sqrt{\left[\frac{U(\Delta)}{\alpha_k}\right]^2 + [u_G]^2 + \left[\frac{\sigma}{\sqrt{N}}\right]^2} \qquad (2)$$

where u_G is the standard uncertainty due to the resolution of the calibrated measuring instrument; X is the reading of the measuring instrument; ∂_1 is the uncertainty of the measured result if the corrections are evaluated for; α_∂ is the coverage factor, which depends on the distribution and the confidence probability.

If the reading of the measuring instrument X is taken as the result of a single measurement, then the error of the measuring instrument is calculated according to formula:

$$X_g = X; \quad \partial_1 = -\Delta \pm \alpha_\partial \cdot \sqrt{\left[\frac{U(\Delta)}{\alpha_k}\right]^2 + [u_G]^2 + \left[\frac{\sigma}{\sqrt{N}}\right]^2} \quad (3)$$

where ∂_2 is the uncertainty of the measured result if the corrections are not evaluated.

If the dispersion of the calibration results is not reported, it is not possible to determine the measurement uncertainty unless the measuring instrument error specified by the manufacturer is available. In many cases, however, calibration certificates do not include the dispersion value (σ).

In the case of multiple measurements (due to the processing of measurement results), the influence of the dispersion of the calibrated measuring instrument's readings the final measurement result is eliminated. In this way, the uncertainty of the final result of multiple measurements should not include the σ component.

V. METROLOGICAL BUDGETING AND DOWNSTREAM CORRECTIONS

The combined Type A and B inputs compile the finalized uncertainty budget detailed in Table 1.

The budget index isolates percentage variance contributions. Physical diameter (45.5%) and height (29.7%, ΔH: 24.7%) dominate systems, while thermal parameters contribute 0.0%. For science-driven decision-making, this establishes that to optimize product tolerances, capital expenditure must focus exclusively on upgrading dimension-measuring instruments rather than climate control units.

Uncertainty Budget:

Quantity	Value	Standard Uncertainty	Distribution	Sensitivity Coefficient	Uncertainty Contribution	Index
π	3.1415927	$28.9 \cdot 10^{-6}$	rectangular	250	$7.2 \cdot 10^{-3}$ cm^3	0.0 %
D	12.4200 cm	0.0374 cm	normal	130	4.7 cm^3	45.5 %
ΔD	0.0 cm	0.0289 cm	rectangular	5.1	0.15 cm^3	0.0 %
H	6.5000 cm	0.0316 cm	normal	120	3.8 cm^3	29.7 %
ΔH	0.0 cm	0.0289 cm	rectangular	120	3.5 cm^3	24.7 %
α	$11.5000 \cdot 10^{-6}$ 1/C	$28.9 \cdot 10^{-9}$ 1/C	rectangular	-2400	$-68 \cdot 10^{-6}$ cm^3	0.0 %
T	23.000 C	0.500 C	normal	$-9.1 \cdot 10^{-3}$	$-4.5 \cdot 10^{-3}$ cm^3	0.0 %
V	787.47 cm^3	7.03 cm^3				

Result:

Value:	Expanded Uncertainty:	Coverage Factor:	Coverage:
787 cm^3	± 14 cm^3	2.00	95% (normal)

Fig 1. Uncertainty budget table. The shadows represent the dominant components.

VI. CONCLUSIONS

In summary, trust analysis requires an integrated approach that combines traceable digital metrology workflows with optimized methods for standard uncertainty assessment at the initial data entry stage, calculating the standard uncertainty and assessing the difference between calibrated measuring instrument data for single and multiple measurements. Specialized software ensures strict GUM compliance and robust Monte Carlo validation, isolating the most important risk factors. By translating optimized input data into transparent uncertainty budgets, decision makers gain precise clarity about risk factors, allowing them to make fair, science-based technological decisions.

REFERENCES

[1] M. G. Cox, P. M. Harris, "Software specifications for uncertainty evaluation," Measurement Techniques, vol. 49, no. 4, Apr. 2006, pp. 311–318.

[2] L. R. Pendrill, "Using measurement uncertainty in decision-making and conformity assessment," Metrologia, vol. 51, no. 4, Aug. 2014, pp. S206–S211.

[3] L. de Vito, J. G. Rossi, "Modern trends in digital metrology and automated software verification for uncertainty evaluation," IEEE Transactions on Instrumentation and Measurement, vol. 73, no. 2, Feb. 2024, pp. 114–122.

2026 IMEKO Joint Conference
TC7 - TC11 – TC13 - TC21 in conjunction with TC6 and Mathmet
Faro, Portugal, 27-29 October, 2026

Standard Reference Data for Thermal Diffusivity of HANA-6

Kweonho Kang[1], Jaesoo Ryu[1], Chang Hwa Lee[1]

[1] *Korea Atomic Energy Research Institute, 1045 Daedeok-daero, Yuseong-gu, Daejeon, Korea, nghkang@kaeri.re.kr*

I. INTRODUCTION

The Nuclear Fuel and Materials Data Center (NFDC) was designated as a national data center under the National Standard Reference Data Program by the Ministry of Trade, Industry and Energy (MOTIE) of Korea on December 30, 2018. The designated fields include nuclear fuels and cladding materials. The target materials for the development of Standard Reference Data (SRD) include nuclear fuels (uranium metal, uranium dioxide, and spent fuel) and zirconium-based cladding materials such as Zircaloy-4, ZIRLO, and HANA alloys.

In this study, a systematic procedure for thermal diffusivity measurement, uncertainty evaluation, and technical assessment was established to develop Standard Reference Data for HANA-6 cladding alloy.

II. MEASUREMENT

A. Experimental procedure

Thermal diffusivity was measured using the laser flash analysis (LFA) method [1, 2]. The measurements were carried out with an LFA-427 (NETZSCH, Germany) over the temperature range from room temperature to 800 °C. The thermal diffusivity of HANA-6 (Zr–1.1Nb–0.05Cu), a zirconium alloy developed for nuclear fuel cladding, was measured.

B. System calibration

System calibration was performed using a certified reference material (CRM, NMIJ 5804-b graphite) to ensure the correct operation of LFA. The certified and measured data of the thermal diffusivity of SRM are plotted against temperature in Fig. 1. The correction factors ranged from 0.98 to 1.04, and the corresponding uncertainties were between 1.38% and 2.39%, confirming the proper performance of the measurement system.

III. RESULTS AND DISCUSSION

A. Thermal diffusivity of HANA-6

Fig. 2 shows the measured thermal diffusivity of HANA-6 as a function of temperature. The thermal diffusivity increased continuously with increasing temperature, which is a typical characteristic of zirconium alloys. The obtained data showed excellent repeatability throughout the entire temperature range.

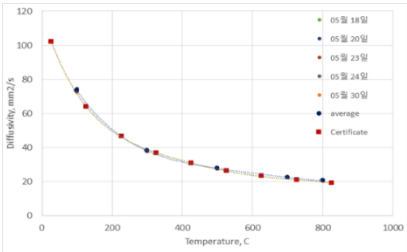


Fig. 1. Thermal diffusivity measurement data and certifed data of CRM

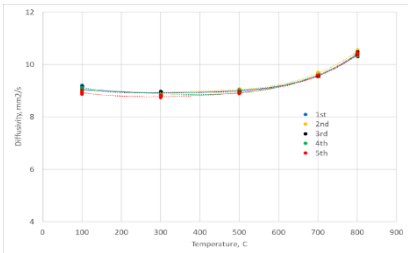


Figure 2 Thermal diffusivity measurement data of HANA-6

B. Uncertainty Evaluatuion

The uncertainty associated with the thermal diffusivity measurements was evaluated following the Guide to the Expression of Uncertainty in Measurement (GUM) [3]. A measurement model for thermal diffusivity was established, and the uncertainty sources were systematically identified and quantified. The standard uncertainty for each input quantity was evaluated, and the combined standard uncertainty was obtained using the corresponding sensitivity coefficients. The effective degrees of freedom were then determined to estimate the appropriate coverage factor, which was subsequently applied to the combined standard uncertainty to obtain the expanded uncertainty. The uncertainty evaluation results

for the thermal diffusivity measurements of HANA-6 are summarized in Table 1. The expanded uncertainty of the thermal diffusivity measurements was less than approximately 2–3% over the entire temperature range, demonstrating the high reliability of the measured data.

Table 1 Uncertainty evaluation of thermal diffusivity measurement data of HNA-6

Uncertainty Evaluation											
Temp.	measurement			Sensitivity factor			Standard uncertainty			Combind standard uncertainty	Extended uncertainty
T, C	d, mm	$t_{1/2}(T)$, s	$f_{cal}(T)$	C_d, mm/s	$C_{t1/2}$, mm^2/s^2	C_{fcal}, mm^2/s	$u(d)$, mm	$u(t_{1/2})$, s	$u(f_{cal})$	u_c, mm^2/s	U, mm^2/s
30	1.526	3.28E-02	1.02E+00	1.32E+01	-3.08E+02	9.86E+00	1.04E-03	1.80E-04	3.20E-04	5.72E-02	1.14E-01
100	1.526	3.41E-02	9.78E-01	1.21E+01	-2.71E+02	9.47E+00	1.04E-03	1.00E-04	3.47E-04	3.01E-02	6.02E-02
200	1.526	3.53E-02	9.85E-01	1.18E+01	-2.56E+02	9.17E+00	1.04E-03	1.50E-04	4.37E-04	4.06E-02	8.11E-02
300	1.526	3.57E-02	1.03E+00	1.22E+01	-2.61E+02	9.05E+00	1.04E-03	1.30E-04	5.78E-04	3.66E-02	7.33E-02
400	1.526	3.58E-02	1.04E+00	1.24E+01	-2.63E+02	9.03E+00	1.04E-03	8.00E-05	7.28E-04	2.56E-02	5.11E-02
500	1.526	3.56E-02	1.01E+00	1.21E+01	-2.58E+02	9.07E+00	1.04E-03	1.30E-04	8.61E-04	3.67E-02	7.33E-02
600	1.526	3.50E-02	9.84E-01	1.19E+01	-2.60E+02	9.23E+00	1.04E-03	1.50E-04	1.02E-03	4.19E-02	8.39E-02
700	1.526	3.35E-02	9.96E-01	1.26E+01	-2.87E+02	9.65E+00	1.04E-03	1.10E-04	1.26E-03	3.63E-02	7.25E-02
800	1.526	3.07E-02	9.83E-01	1.36E+01	-3.38E+02	1.05E+01	1.04E-03	1.00E-01	1.48E-03	3.99E-02	7.97E-02

C. Technical Evaluation

The thermal diffusivity measurement data were evaluated based on the technical evaluation criteria and subsequantly approved as Standard Reference Data (SRD) [4]. The technical evaluation criteria are presented below.

1. Whether the measurand and the purpose of the measurement are clearly defined.
2. Whether the measurement method is described in sufficient detail and is technically appropriate.
3. Whether the measurement procedure is described in sufficient detail and is technically appropriate.
4. Whether the traceability of the measurement results is established and adequately documented.
5. Whether the evaluation of the uncertainty associated with the measurement results is appropriate.
6. Whether the measurement results are consistent with the values calculated from the relevant theoretical relationships or equations.
7. Whether the measurement results are consistent with those reported in other independent measurements or published studies.
8. Whether the measurement results exhibit reasonable correlations or trends with the related physical properties or characteristics.

Based on these evaluation criteria, the thermal diffusivity data satisfied all technical requirements for Standard Reference Data (SRD).

D. Standard Reference Data of thermal diffusivity of HANA-6

The thermal diffusivity data of HANA-6 were finally approved as Standard Reference Data after completion of uncertainty evaluation and technical assessment. The approved SRD provide reliable thermophysical property values for fuel performance analysis, finite-element simulations, thermal analysis, and the design and safety assessment of nuclear fuel systems. In particular, these data can be utilized for heat transfer analysis, fuel temperature prediction, and safety evaluations of zirconium alloy cladding materials.

IV. CONCLUSIONS

- A systematic procedure for producing Standard Reference Data for the thermal diffusivity of zirconium cladding alloys was established.

- The thermal diffusivity of HANA-6 was successfully measured using the laser flash analysis method after calibration with certified reference materials.

- The measurement uncertainty was evaluated in accordance with the Guide to the Expression of Uncertainty in Measurement (GUM).

- The measured data satisfied all technical assessment criteria of the National Standard Reference Data Program and were approved as Standard Reference Data.

- The developed Standard Reference Data provide reliable thermophysical property values for nuclear fuel performance analysis, heat-transfer calculations, and safety evaluations.

REFERENCES

[1] Parker, W.J., Jenkins, R.J., Flash Method of Determining Thermal Diffusivity, Journal of Applied Physics, Vol.32, pp.1679–1684 (1961).

[2] ASTM E1461-13, Standard Test Method for Thermal Diffusivity by the Flash Method

[3] JCGM 100:2008, Evaluation of measurement data—Guide to the Expression of Uncertainty in Measurement (GUM).

[4] National Standard Reference Data Program, Technical Assessment Guidelines, Ministry of Trade, Industry and Energy, Republic of Korea.

2026 IMEKO Joint Conference
TC7 - TC11 – TC13 - TC21 in conjunction with TC6 and Mathmet
Faro, Portugal, 27-29 October, 2026

Technical challenges and interpretations: the interaction of legal metrology and accreditation requirements in laboratory practice

Asta Meškuotienė[1], Edita Raudienė[2], Paulius Kaškonas[3], Justina Dobilienė[4]

[1, 2, 3, 4] *Institute of Metrology of Kaunas University of Technology, asta.meskuotiene@ktu.lt, paulius.kaskonas@ktu.lt, edita.raudiene@ktu.lt, justina.dobiliene@ktu.lt*

I. INTRODUCTION

In today's metrological infrastructure, calibration laboratories and inspection bodies often find themselves caught between a rock and a hard place: they must comply with national legal metrology regulations while also meeting the requirements of international accreditation standards. This raises important questions: How should laboratories proceed in situations where the accreditation standard leaves room for interpretation- for example, regarding the validation and verification of methods, uncertainties in budgeting, calibration and measurement capabilities (CMC), or the selection and suitability of reference standards? Furthermore, how can legal metrology requirements be technically and accurately integrated into a laboratory's operating procedures?

The inconsistent interpretation of metrological requirements in various legal acts and standards creates serious technical obstacles to establishing and documenting specific criteria for measuring instruments and their supervision. This subsequently leads to systematic interpretation errors in day-to-day laboratory operations. Industry, laboratories, and regulatory authorities must share a common understanding of metrological supervision concepts. Achieving this requires not only a high level of technical competence among staff but also the ability to adapt assessment procedures flexibly and objectively to rapid technological advances.

Deliberately stepping away from general aspects of the management system, this presentation focuses primarily on technical requirements and practical issues as viewed through the eyes of a technical expert. The main focus will be on technical requirements covering staff competence, methodological consistency, and technical equipment management, as well as their interpretation and practical application.

Methods in Legal and Industrial Metrology: Mandates vs. Voluntary Practices

A clear distinction must be made between the methods used in legal and industrial metrology, considering what is mandatory and what is not:

- Standards: Voluntary, adopted by consensus, and developed by independent organizations. They are typically reviewed every 5 years and provide technical specifications and test methods, with a focus on safety, quality, and performance.
- Regulations: Mandatory and established by law or by government authorities. They are reviewed at the discretion of lawmakers (for example, when implementing EU directives) and set requirements designed to protect the public interest.

Metrological requirements in legal metrology for groups of measuring instruments should be established, taking into account their specific characteristics of metrological supervision: in areas regulated by the EU, at the national level, and for groups not subject to type approval.

	I. AREA REGULATED BY EU LEGAL ACTS	II. AREA REGULATED BY NATIONAL LEGAL ACTS	III. MI GROUP NOT SUBJECT TO TYPE APPROVAL
A. Requirements for Type Approval	Directives in the field of Metrology	OIML R, ISO standards for individual MI groups, metrological requirements of legal acts from other fields, and manufacturers' technical specifications	–
B. Requirements for Verification	General and specific verification methodologies (legislation), prepared according to directives, harmonised standards, OIML R, metrological requirements of legal acts from other fields, WELMEC guides, ISO standards, or the manufacturer's technical specifications		

Fig. 1. Documents regulating requirements for individual MI groups

Are only the currently valid standards applied?

It is important to note that there are various types of standards, such as laboratory accreditation standards (process-oriented) and product standards (specification-oriented). Over time, measurement technologies, equipment, and methodologies evolve, leading to new requirements. Standards must reflect this technological progress, and laboratories must transition smoothly to these updated criteria. Product standards, including standards for measuring instruments (MI), are closely linked to market access. Manufacturers must ensure that their products comply with the standards and directives (e.g., safety, health, environmental protection) in effect at the time the product is placed on the market. When a standard changes, newly manufactured products entering

the market must comply with the latest version. Importantly, products already on the market that complied with the standards in effect at the time of manufacture generally do not need to be modified to comply with newer regulations. This principle applies directly to measuring instruments in legal metrology. If the old device continues to operate in accordance with the original approval requirements and its metrological characteristics remain stable (e.g., test sieves), there is no urgent technical need to replace it. The only exceptions are when specific components are prohibited under updated safety or environmental standards, such as bans on mercury thermometers or hazardous substances.

How are methods selected, validated, and verified?

How should uncertainty be interpreted when legal metrology requires a strict "conforms/does not conform" assessment of compliance? This presentation will discuss the application of decision rules during verification or inspection.

Comparative standard methods are generally used for verification and calibration, which do not require validation. A consensus reached with metrology experts regarding the description of the method used and its validation may be considered validation. Verification of a method, as defined in the accreditation standard, refers to the readiness to perform work in accordance with the relevant procedure, based on criteria such as ensuring appropriate environmental conditions, selecting and using appropriate measuring instruments, and evaluating and presenting results.

The key result of the verification is the error, which is compared to the permissible error of the measuring instrument. If the criterion of an appropriate error level is applied to the selection of a reference device for the verification, the uncertainty of the measurement result is negligible in this case. Furthermore, in general, single measurements are performed during verification. For the reasons mentioned above, the "pass/fail" statement, defined in the ISO 17025 standard as a "decision rule, is not applied in the verification procedure.

In addition to method validation and verification, the assessment of a laboratory's effectiveness also includes other essential criteria related to the selection, maintenance (calibration), and management of reference measuring instruments. All these factors determine the laboratory's calibration and measurement capabilities (CMC).

Reference Measuring Instruments: Selection and Maintenance

Reference measuring instruments (MI) for verification are selected based on the required error ratio relative to the measuring instrument under test:

- If the permissible error of the reference MI is specified by the manufacturer or a standard, then this error is used to formulate the requirements.
- If the error of the reference MI is unknown, compliance with the requirements is assessed based on its calibration data.

The expanded calibration uncertainty stated in the reference MI calibration certificate does not reflect the error of this MI; it pertains only to the reliability of the systematic error determined during calibration and therefore cannot be used to assess compliance with the requirements for selecting a reference standard. To assess whether an MI is suitable for use as a reference instrument, its expanded uncertainty must be recalculated based on the data specified in the calibration certificate:

$$U = 2\sqrt{\left(\frac{U_r}{2}\right)^2 + u_{res}^2 + \left(\frac{\sigma_r}{\sqrt{N_p}}\right)^2}, \quad (1)$$

where U_r is the expanded uncertainty of the MI calibration, σ_r is the standard deviation of the sample representing the average dispersion of measurement results relative to the mean, and N_p is the number of measurements performed during the verification.

If the verification involves taking individual measurements, the deviation of which is compared to compliance with the permissible errors, then N_p=1. However, if the verification procedure calls for multiple measurements, then, when calculating the expanded uncertainty U, N_p corresponds to the sample size. The total error of the reference MI, whose value is compared to the permissible error, is estimated as follows (when a systematic error is included in the measurement result to obtain the corrected reading, then $\Delta_r = 0$):

$$\Delta_{total} = \Delta_r + U. \quad (2)$$

There are challenges in selecting the appropriate reference equipment and correctly interpreting the errors in measurement systems. There is no uniform system for calculating the total error when legislation specifies only general requirements for derived quantities, such as speed, volume, etc. The interpretation of the total error is presented in Figure 2.

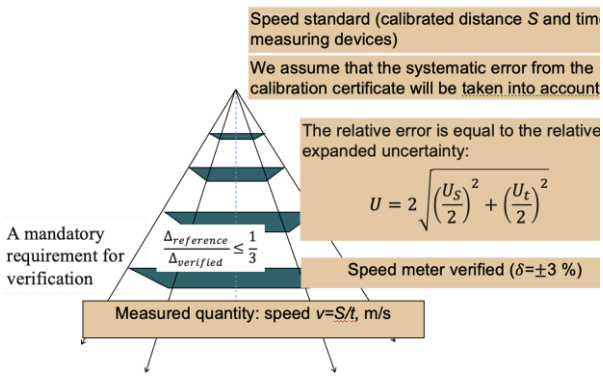


Fig. 2. The interpretation of the total error

However, if the above-mentioned criteria cannot be met, the test results can be approved alternatively by reducing the applied maximum permissible errors by the excess of the uncertainties:

$$the\ error\ of\ the\ measured\ quantity \leq \sqrt{\frac{8}{9} \cdot \Delta_{\mathrm{v}}^2 + \Delta_{\mathrm{r}}^2}, (3)$$

where Δ_v is the error of the verified MI, and Δ_r is the error of the reference MI.

In summary, effective assessment of compliance with metrological requirements requires not only the harmonization of standards and legislation, but also a unified technical approach to method validation and the management of reference equipment. This article provides practical recommendations for integrating these requirements into laboratory operations, ensuring technical accuracy and compliance with accreditation criteria.

2026 IMEKO Joint Conference
TC7 - TC11 – TC13 - TC21 in conjunction with TC6 and Mathmet
Faro, Portugal, 27-29 October, 2026

A Metrological Framework for the Comprehensive Annual Calibration of pH Meters

Saowaluck Ukrisdawithid[1], Thanundorn Pithakkiet[1]

[1] *Center for Calibration Reference Laboratory, Department of Science Service, Thailand, Saowaluck@dss.go.th, tpithakkiet@gmail.com*

***Abstract* – Although ISO/IEC 17025 mandates metrological traceability and regular equipment calibration, the lack of a standardized protocol for the annual calibration of pH meters remains a challenge for testing and calibration laboratories. Practical implementation varies significantly, particularly regarding electrode drift evaluation, multi-point calibration routines, and realistic uncertainty estimation. This paper proposes a structured annual calibration framework that integrates Certified Reference Materials covering a wide pH range (4.005 to 10.010) with multi-point linear regression analysis. The protocol systematically evaluates key electrode performance parameters, including practical slope, asymmetry potential, temperature effects, and time-dependent electrode drift. The proposed framework enhances measurement reliability, ensures technical validity, and provides a unified approach for annual calibration practices in accredited laboratories.**

I. INTRODUCTION

Electrochemical measurements, specifically pH determination, are vital across various sectors, including food safety, pharmaceuticals, environmental monitoring, and clinical analysis. Ensuring measurement reliability requires strict compliance with quality management systems such as ISO/IEC 17025. However, while guidelines exist for routine daily checks or initial buffer adjustments, there is a notable absence of a standardized international consensus protocol dedicated specifically to the annual calibration of pH meters[3].

In practice, laboratories often rely on manufacturer-specific auto-calibration routines or basic two-point buffer adjustments. These practices frequently ignore key metrological aspects, such as separate electrical meter validation, electrode slope degradation, asymmetric potential drift, and diaphragm clogging effects over an extended operational period. This paper presents a standardized annual calibration protocol designed to bridge these metrological gaps.

II. METHODOLOGY

The multi-point calibration procedure was performed on the integrated meter-glass electrode assembly using five Certified Reference Materials (CRMs) with nominal pH values of 4.005, 6.865, 7.000, 9.180, and 10.010 at a controlled temperature of (25±0.5) °C. Solution temperatures were continuously monitored using a calibrated digital thermometer equipped with a Resistance Temperature Detector placed in a temperature-controlled liquid bath.

A. Functional Setup & Calibration Scheme

The calibration framework was executed in two distinct stages:

1. Electrode Performance & Regression Analysis: The meter readings were recorded across the five certified buffer solutions. Linear regression analysis based on the least-squares method was applied to determine the practical slope (k') and the zero-point offset (E_0/ asymmetry potential).
2. Replicate Measurements and Temporal Drift Assessment: Replicate measurements of the certified buffer solutions were performed to evaluate repeatability. Short-term temporal stability was evaluated by continuously monitoring the electrode potential drift over a 1-hour period [2].

B. Measurement equation[1]

The linear relationship between the electromotive force (E) and pH is defined in eq (1):

.

$$E = E_0 - k'\mathrm{pH} \tag{1}$$

The unknown pH value of a solution (pH_x) is subsequently derived as eq (2):

$$\mathrm{pH}_x = \frac{E_x - E_0}{k'} \tag{2}$$

C. Uncertainty Evaluation Framework

The combined standard uncertainty for the pH measuring system was evaluated in accordance with the Guide to the Expression of Uncertainty in Measurement (GUM). Standard uncertainty contributions were identified and quantified from six key sources:

- Replicate measurement repeatability (δpH_{rep})
- Digital indication resolution (δpH_{res})
- Prediction interval from multi-point regression fit (δpH_{reg})
- Certified reference material uncertainty (δpH_{crm})
- Solution temperature variation effects (δpH_{t})
- Short-term electrode drift over 1 hour (δpH_{drift})

III. RESULTS AND DISSCUSSION

The primary calibration parameters—the practical slope and the asymmetry potential (zero-point offset) were determined from a five-point linear regression over the certified pH range (4.005 to 10.010). The observed slope efficiency (k'/k) and zero-point potential provide critical diagnostic indicators of glass electrode aging, membrane degradation, and junction clogging, which are evaluated against pre-defined laboratory acceptance criteria. The evaluation of measurement uncertainty was conducted in accordance with the Guide to the Expression of Uncertainty in Measurement (GUM). The functional measurement model for the pH deviation (ΔpH), the deviation of pH reading (pH_x) and certified buffer solutions (pH_{crm}), is expressed as eq. (3):

$$\Delta pH = pH_x - pH_{crm} - \delta pH_{rep} - \delta pH_{res} - \delta pH_{reg} - \delta pH_{drift} - \delta pH_{crm} - \delta pH_t \quad (3)$$

As shown in Table 1, the largest contributor to the combined standard uncertainty is the short-term electrode drift ($u_{\delta pHdrift}$ = 0.0058), followed by the CRM buffer uncertainty ($u_{\delta pHcrm}$ = 0.0030) and the regression analysis ($u_{\delta pHreg}$ = 0.0022). The regression analysis accounts for non-linear behavior across the wide pH scale (4.005–10.010), which conventional two-point calibration routines fail to quantify. Incorporating the temporal drift over a 1-hour operational window ensures that the resulting expanded uncertainty (U = 0.014, k=2) realistically reflects actual testing conditions throughout routine analytical operations.

Table 1. Uncertainty budget for pH measuring system calibration (at pH 7.000, T = 25 ℃).

Quantity	Source of Uncertainty	Standard Uncertainty	Divisor	c_i	u_i(pH)	υ_i or υ_{eff}
δpH_{rep}	repeatability	0	1	1	0	2
δpH_{res}	resolution	0.0005	$\sqrt{3}$	1	0.00029	∞
δpH_{reg}	regression	0.00223	1	1	0.00223	23
δpH_{drift}	drift	0.01	$\sqrt{3}$	1	0.0058	∞
δpH_{crm}	certificate	0.006	2	1	0.0030	∞
δpH_t	temperature	0.00165	$\sqrt{3}$	1	0.00095	∞
$u_c(pH_x)$					0.0070	
U			k=2		0.014	

IV. CONCLUSION

This study presents a comprehensive annual calibration framework that bridges critical metrological gaps in conventional pH measurement routines. By incorporating a multi-point calibration scheme with Certified Reference Materials (CRMs) and a systematic evaluation of short-term electrode drift, the proposed protocol decouples individual performance sources and provides a realistic basis for measurement uncertainty estimation.

Annual calibration and daily verification function synergistically within a laboratory quality system. While annual re-calibration anchors the measuring system to the international pH scale through an unbroken chain of traceability, daily operational checks ensure ongoing performance validity between calibration intervals. Implementing this standardized protocol enhances measurement reliability, supports compliance with ISO/IEC 17025 requirements, and provides testing laboratories with a technically robust foundation for electrochemical analysis.

REFERENCES

[1] A. K. Covington, R. G. Bates, R. A. Durst, "Definition of pH scales, standard reference values, measurement of pH and related terminology", Pure and Applied Chemistry, vol. 57, no. 3, 1985, pp. 531-542.

[2] IEC 60746-2:2003, Expression of performance of electrochemical analyzers – Part 2: pH value, International Electrotechnical Commission, 2003.

[3] V. Barwick, K. C. Tsimillis (Eds.), Eurachem/CITAC Guide: Guide to Quality in Analytical Chemistry: An Aid to Accreditation, 4th ed., Eurachem/CITAC, 2026. Available from www.eurachem.org.

2026 IMEKO Joint Conference
TC7 - TC11 – TC13 - TC21 in conjunction with TC6 and Mathmet
Faro, Portugal, 27-29 October, 2026

Experimental Microwave Array Sensing for Brain Tumor Detection: Sensitivity and Repeatability Evaluation in a Head Phantom Model

Livio D'Alvia[1], Venla Korkko[2], Emanuele Rizzuto[3], Mariella Särestöniemi[4], Teemo Myllylä[2]

[1] *Dept. of Well-being, Health and Environmental Sustainability, Sapienza university, Sapienza university of Rome, Italy Livio.dalvia@uniroma1.it*
[2] *Optoelectronics and Measurement Techniques Research Unit, University of Oulu, Finland, Venla.Korkko@student.oulu.fi, Teemu.Myllyla@oulu.fi*
[3] *Dept. of Mechanical and Aerospace Engineering, Sapienza university of Rome, Italy Emanuele.rizzuto@uniroma1.it, Zaccaria.delprete@uniroma1.it*
[4] *Research Unit of Health Sciences and Technology, University of Oulu,Finland, mariella.sarestoniemi@oulu.fi*

***Abstract* – Microwave imaging (MWI) represents a non-ionizing, cost-effective alternative for brain tumor monitoring. However, strong tissue attenuation and complex scattering require high-sensitivity array configurations and robust calibration techniques. This paper presents an experimental evaluation of a 12-antenna array arranged around an elliptical head phantom. By performing complex vector subtraction against the background medium, we evaluate the differential reflection (ΔS_{11}) and transmission (ΔS_{21}) signals across multiple target sizes (TS1–TS4 higher to smaller). Results demonstrate that short-path transverse pairs yield significantly higher signal-to-noise ratios (SNR) and monotonic sensitivity to tumor growth, providing a clear experimental benchmark for differential detection algorithms.**

I. INTRODUCTION

Brain tumor detection and post-treatment monitoring demand diagnostic modalities that are safe for frequent use. While Magnetic Resonance Imaging (MRI) and Computed Tomography (CT) remain the gold standards, microwave imaging (MWI) has emerged as a complementary, non-ionizing, and low-cost modality [1]. MWI relies on the contrast in complex dielectric permittivity between healthy brain tissues and malignant lesions [2].A persistent challenge in brain MWI is the high attenuation of microwave signals propagating through cranial structures. Since the human head exhibits an elliptical cross-section, the propagation path length—and consequently the baseline path loss—varies significantly depending on antenna placement.In this work, we evaluate a 6x6 multi-bistatic antenna setup operating on an elliptical head phantom model. We analyze the differential transmission (ΔS_{21}) and reflection (ΔS_{11}) responses, focusing on the short-path transverse channels across the right and left temporal/parietal lobes

II. METHODOLOGY & EXPERIMENTAL SETUP

A. Array Topology and Geometry

The measurement setup consists of a 12-element conformal array arranged sequentially along the outer perimeter of an elliptical head phantom model (Fig. 1A). The phantom shell was 3D-printed using Polylactic Acid (PLA) and filled with a homogeneous tissue-mimicking solution composed of distilled water and Ethylene Glycol in a 10:1 volume ratio. The 12 flexible patch antennas, A1 to A12, (Fig. 1C) are uniformly distributed in a counter-clockwise arrangement around the skull perimeter to cover 360° at 30° angular intervals. The spherical tumor targets used during testing were fabricated using agar-based gel recipes according to [3].

B. Automated Measurement and Switching System

The data acquisition setup relies on a 2-port PicoVNA Vector Network Analyzer (VNA) integrated with a custom-controlled dual 6 coaxial switching network to manage the 12-antenna conformal array (Fig. 1B). Port 1 connected to the first 6-channel switch, which routes the RF excitation signal sequentially to the right-side antenna cluster (P1A1 to P1A6). Port 2 connected to the second 6-channel switch, which routes the back-scattered RF signal sequentially to the left-side antenna cluster (P2A1 to P2A6).The switching system is controlled via an Arduino microcontroller board operating in a cyclic sequence. For each frequency sweep generated by the PicoVNA (spanning 0.3 GHz to 8.5 GHz), the Arduino triggers the multiplexers to scan all channel pairs in a structured cycle: while Port 1 fixes its transmission on a single channel, Port 2 sweeps through all its receiving channels. This produces a complete transmission/reflection matrix consisting of 36 unique cross-antenna combinations per measurement sweep.

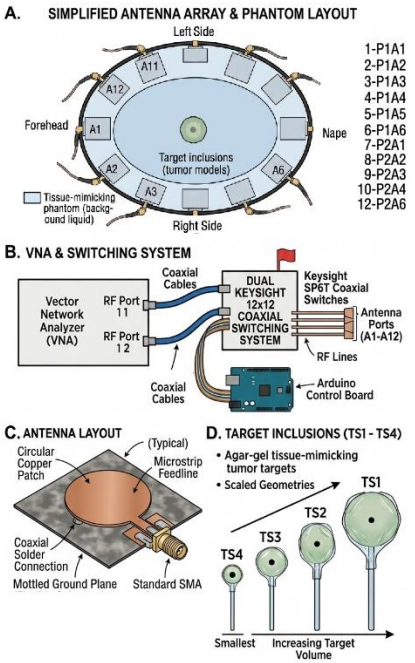


Fig. 1. Experimental setup diagram showing (A) the 12-antennas skull phantom layout, (B) the VNA and multiplexed switching network, (C) the flexible patch antenna geometry, and (D) tumor targets (TS1–TS4)

C. Measurement Protocol & Tumor Targets

Measurements were conducted across six distinct static phantom scenarios: (i) free-Space or Calibration benchmark, (ii) Background based on homogeneous tissue-mimicking liquid phantom, and (iii) tumor targets (TS1 to TS4) (Fig. 1D). To ensure statistical repeatability, each scenario was measured over $R = 5$ independent sweeps.

C. Complex Vector Differential

SubtractionRaw magnitude and phase data were averaged across 5 sweeps to form the mean complex

$$\bar{S}_{ij}(f) = \frac{1}{R}\sum_{r=1}^{R} |S_{ij}^{(r)}(f)| \qquad (1)$$

To isolate the scattering signature of the inclusion from the dominant background clutter, complex vector subtraction was applied:

$$\Delta S_{ij}(f) = \bar{S}_{ij}^{\text{TARGET}}(f) - \bar{S}_{ij}^{\text{BACKGROUND}}(f) \qquad (2)$$

III. EXPERIMENTAL RESULTS & DISCUSSION

A. Differential Transmission Analysis

Given the central placement of target inclusions in the head phantom, differential transmission ($|\Delta S_{21}|$) was evaluated across all 36 Tx-Rx channels. As expected from the elliptical geometry, short transverse paths exhibited lower baseline attenuation than longer anteroposterior or diagonal channels. As an example, Fig. 2 shows the

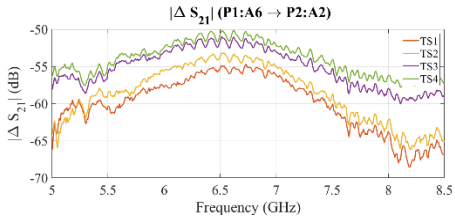


Fig. 2. Differential transmission response for the transverse antenna pair P1:A6 → P2:A2 across different tumor target sizes (TS4 to TS1).

response for the transverse pair $P_1A_6 \rightarrow P_2A_2$. The differential signal scaled monotonically with target volume, decreasing progressively in magnitude as tumor size increased from TS4 to TS1.

B. Differential Reflection Analysis

Monostatic differential reflection ($|\Delta S_{11}|$) was also evaluated across the array. Antennas closer to the target displayed a marked increase in magnitude and distinct frequency shifts in their resonant dips, whereas contralateral antennas showed minimal variations. This confirms that monostatic reflection effectively isolates localized dielectric perturbations within the phantom.

IV. CONCLUSION

This paper presented an experimental sensitivity evaluation of a 12-antenna microwave sensing array using an elliptical head phantom model. By performing complex vector background calibration, we demonstrated that short-path transverse channels (such as $P_1A_6 \rightarrow P_2A_2$) effectively exploit the head geometry to yield robust differential signals ($|\Delta S_{21}|$ and $|\Delta S_{11}|$).The experimental datasets confirm that target growth produces clear, monotonic changes in differential magnitude—exhibiting a progressive decrease in $|\Delta S_{21}|$ as the tumor volume increases from TS4 to TS1. These findings validate the proposed array topology and differential calibration approach for future radar-based and tomographic imaging algorithms.

REFERENCES

[1] D. Singh, et al., "Noninvasive and Quantitative Intracranial Pressure Estimation Using Microwave Technique," *IEEE Sensors Journal*, 2025, doi: 10.1109/JSEN.2025.3631129.

[2] L. D'Alvia, et al., "A Novel Microwave Resonant Sensor for Measuring Cancer Cell Line Aggressiveness," *Sensors*, 2022, doi: 10.3390/s22124383.

[3] M. Lazebnik,et al., "Tissue-mimicking phantom materials for narrowband and ultrawideband microwave applications," *Phys. Med. Biol.*, 2005, doi: 10.1088/0031-9155/50/18/001

2026 IMEKO Joint Conference
TC7 - TC11 – TC13 - TC21 in conjunction with TC6 and Mathmet
Faro, Portugal, 27-29 October, 2026

Extended robust algorithm (A) ISO 13528 to dynamic measurement for interlaboratory comparison

BOUHOUCHE Salah [1,3], AZZEDINE Abdel Hakim [2]

[1] *Eurl SERMA METRIC AND SCIENCE Ltd – R&D Support, Villa 138 RYM2 S/ACHOUR Annaba, 23000, Algeria, bouhouche11@yahoo.fr - ALGERIA*
[2] *Department of Mathematics, Faculty of Sciences, Benyoucef Benkhedda algiers 1 university, Algeria, abd.azzedine@univ-alger.dz*
[3] *Mathematical Modeling & Numerical Simulation Laboratory (LAM2SIN) at Badji Mokhtar University (UBMA) in Annaba-23000-ALGERIA.*

***Abstract* – Proficiency testing (PT) schemes are usually used as powerful tool in the performance evaluation in interlaboratory comparison to evaluate measurement results performed under equivalent conditions and assess the scientific and technical competence of participants to demonstrate the reliability of their measurement systems. For this purpose, ISO 13528:2022 robust algorithm A is used to compute different statistical performances.**

Robust algorithm (A) as developed by the standards ISO 13528 is well used to compute the assigned values i.e., reference value and robust standard deviation. This algorithm has shown its performance and it is widely applied for static measurement. In this paper, a same algorithm has been developed and numerically implemented by extension to dynamical systems which are generally characterized by dynamic variations of the characteristics of the artifacts.

The robust ISO 13528 algorithm (A) extended to dynamic systems is characterized by an additional loop relating to the variations over time. The assigned values i.e. the reference value and standard deviation - are instantly computed to provide the robust dynamic values. The evaluation consistently addresses prior dynamic behavior of the samples; recorded experimental observations are generated by a random dynamic model simulating a second order system. Comparative study between the main statistical characteristics obtained by the conventional approach based on the descriptive statistics and that obtained using the robust extended dynamic algorithm is considered. Comparative index like Zscore as defined is ISO 17043 is computed to compare performances of participants in dynamic way.

I. INTRODUCTION

Inter laboratory comparison is a basic and an important activity needed by the quality assurance procedures, this is more critical during the accreditation according to the ISO 17025 standard. According to the importance of measurement errors between the true values and that given by the artifacts, the measurements results are classified satisfactory, acceptable or fails. Such decisions are made using statistical comparison based mainly on the computed indexes, like (z) score, Normalised index (En), more details are given in different references and standards [1] [2] [3] [4][5].

Also, proficiency testing implies the use of inter laboratory comparisons to determine the performance of participants based measurement provided by laboratories for specific analysis or measurements, and to monitor their performance. There are a number of typical purposes of proficiency testing, as described in the introduction to ISO/IEC 17043:2023. These include the evaluation of laboratory performance, the identification of problems in laboratories, establishing effectiveness and comparability of measurement methods, the provision of additional confidence to laboratory customers, validation of uncertainty, and the training of participating laboratories. The applied statistical design, analytical techniques and computerised procedure must be appropriate for the stated purpose(s). ISO 13528 provide mainly robust algorithms based on optimal computing scheme, it also allows requirements for the statistical design, validation of proficiency test items, review of results, and reporting summary statistics. Description of general statistical methods is also given in proficiency testing schemes like Annex B of ISO/IEC 17043:2023.

The ISO 13528 International Standard are intended to be complementary to ISO/IEC 17043, providing detailed computing note that is lacking in that document on particular statistical methods and numerical implementation of dynamic measurements for proficiency testing. Characterization of such material is made in a static way using the so-called robust algorithm A taking

into account of all variability of data connected to the static behaviour of the considered artefact.

The actual version of the standard ISO 13528 do not clearly specify if it is only applied for static measurement or it could be extended to dynamic way. But as presented on practical examples, we understand that is valid only in static measurement. In this work, we consider an extension of the above cited robust algorithm A to the dynamic way defined by unsteady state measurements. In the proposed method, ISO 13528 robust algorithm A will be implemented and tested, at every time which is considered as the independent variable, the algorithm is accurately executed – see details in the next sections.

ISO 13528 robust algorithm (A) and its allied methods have been widely considered as a promising for proficiency testing computing procedure. Its successful applications have been reported in numerous works. Traditionally, the above cited methods are used in static way; they constitute a part of the computing technique and take an important place in proficiency testing methods [4] [5] [7].

In this paper, the idea of extending the ISO 13528 robust algorithm (A) is somehow new. This approach consists in the following steps:

- The first step is a selection and preparation of test materials i.e. data provided by different participant are generated by a first order dynamic model perturbed by random normal distributed input.
- The second step is an implementation of the robust algorithm (A). Different proficiency items like robust assigned values are computed according to computing procedure based on the ISO 13528 algorithm (A).
- The third step an evaluation of the obtained performance of the recursive form of the robust algorithm (A) such as the convergence precision and other.
- The fourth step is about the evaluation of different participant, different index are computed like Zscore etc.

This paper is organized as follows. In section II, the proposed method is developed; the robust algorithm (A) extended to dynamic system is considered and a flow sheet of the computing procedure is implemented, particular importance to the solution based on the recursive form is given.

In section III, some applications will be considered to validate different computing procedures, first the recursive form of static model is implemented using data provided from table E.4 ISO 13528-2022 Page.76 and ISO 5725-5 – table31 ISO 5725-5 page 49.

In the second step the dynamic algorithm is considered, in this section the numerical model is implemented, performances are analysed and commented.

A. Static measurements

ISO13528 allow us guidance and methods to well establish the so-called robust algorithm using the measurement data provided by artefacts from different participants. Static measurement is characterised by unsteady state data, dynamic data is known to be not changed too much over time, the variations are generally limited around the mean values which are due to the repeatability of the measurement: There are any deep causality like the dynamics of the system for the proficiency testing. This revision of ISO 13528:2022 contains most of the statistical methods and guidance from the first edition, extended as necessary by the previously referenced documents and the extended scope of ISO/IEC 17043. ISO/IEC 17043 includes performances of proficiency testing for conformity assessment based measurements, after comparisons using statistical tools, decisions are done. The main items to be used are the robust assigned value and robust standard deviation, ISO 13528 gives the main guidance and basic elements of the computing procedure. More details will be given in the next sections.

B. Dynamic measurements

Dynamic measurements data are provided from dynamic systems, these data differ from static measurement by its stability.

In statistics, the z score is used for either standardization of data when they were obtained in different orders of magnitude or outlier detection procedure for univariate data set (if the data are normally distributed) [6]. Although ISO 13528:2022 note the normality assumption inherent in the z score, there is no mention about methods to handle heteroscedasticity. In PT schemes, heteroscedasticity is the inhomogeneity variance between laboratory participants. The PT participating laboratories may use different analytical methods to obtain their results [1], which may lead to inhomogeneous variance between the laboratory results. The ISO 13528:2022 standard does not provide statistical procedures to deal with heteroscedasticity.

We suppose that such data are generated by the Interlaboratory comparison is an activity required by the quality assurance of measurement results. This aspect is more required by the accreditation of laboratories according to the ISO 17025 standard.

II. ROBUST ALGORITHM EXTENDED TO DYNAMIC SYSTEM

A. Brief description of the robust algorithm (A) [ISO 13528 – 2022]

The robust algorithm (A) with iterated scale yields robust estimates of the mean and standard deviation of the data to which it is applied. The considered proficiency testing is based on measurement data provided by different participants.

Let p items of data, sorted into increasing order, by: $X_1, X_2, \ldots X_p$ and X*and S* denote the robust average and robust standard deviation of these data. Because we use an iterative approach for the optimization of the X* and S* parameters, to meet optimal solution, these parameters are initialised as follows.

X* = median of xi (i = 1, 2, ..., p)

$$S^* = 1.483.median\left|X_i - X^*\right| \text{ with (i = 1, 2, ..., p)} \quad (1)$$

In some cases more than half of the results X_i will be the identical. In these cases the initial value of S* will be zero and the robust procedure will not perform correctly. In the case that the initial S* = 0, it is acceptable to substitute the sample standard deviation, after checking for any gross outliers that could make the sample standard deviation unreasonably large. This substitution is made only for the initial S*, and after that the iterative algorithm can proceed as described.

Let δ = 1,5.S*

$$X_i^* = \begin{cases} X^* - \delta : X_i < X^* - \delta \\ X^* + \delta : X_i > X^* + \delta \\ X_i : Otherwise \end{cases} \quad (2)$$

X_i^* is robust new mean value of every measurement item.

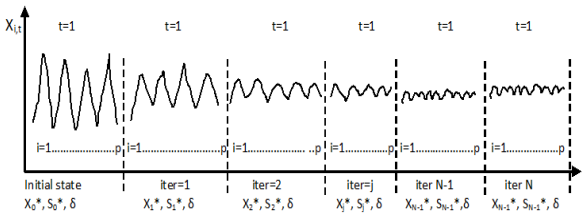
By running the algorithm iteratively, new items are computed as:

$$X_{new}^* = \frac{\sum_{i=1}^{p} X_i}{p} \quad (3)$$

X_{new}^* defines the new robust mean value and

$$S_{new}^* = 1.134.\sqrt{\frac{\sum_{i=1}^{p}(X_i^* - X_i)^2}{p-1}} \quad (4)$$

S_{new}^* is the new robust standard deviation.



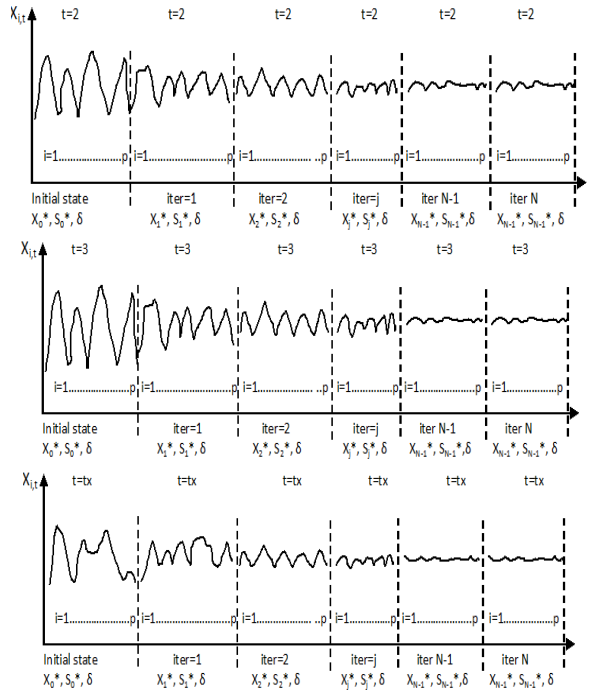


Fig. 1.Extended dynamic computing procedure ISO 13528 Algorithm (A)

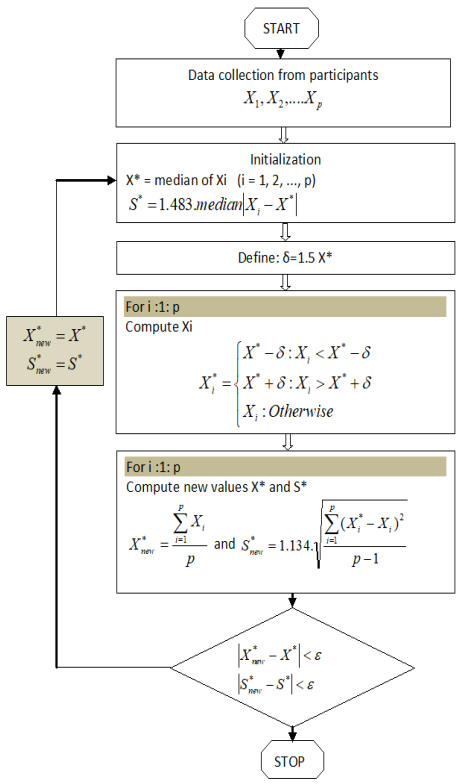


Fig. 2. general flow sheet of conventional robust algorithm (A) ISO 13528.

B. Proposed robust algorithm (A) of dynamic system

In this section, the algorithm (A) presented in the past

sub section A (section II) is extended to the dynamic measurement provided by different participants, the data of (p) participants are dynamic and sorted into increasing order and defined by:

$X_{1,t}, X_{2,t}, ...X_{p,t}$ are the measurement of (p) participants at very time (t), X_t^* and S_t^* denote the robust average and robust standard deviation of these data at every time (t).

By considering the dynamic change over time of measurements, the equations (1) to (4) can be written by equations (5) – (8) as follows.

$$X_t^* = median(X_{i,t}) \quad (5)$$

with X_t^* (i = 1, 2, ..., p and t=1:tmax)

$$S_t^* = 1.483.median\left|X_{i,t} - X_t^*\right| \quad (6)$$

with (i = 1, 2, ..., p and t=1:tmax)

The initial value of the robust standard deviation is defined as: $S_t^* = 1.5\delta$

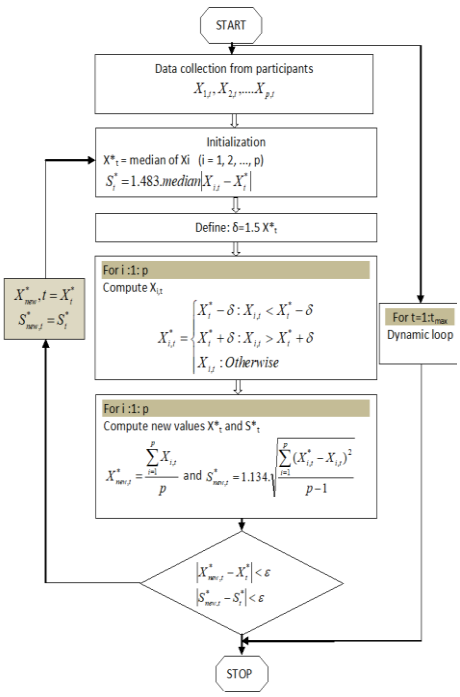


Fig. 3. General flow sheet of new dynamic robust algorithm.

III. APPLICATIONS

The validation of the implemented robust algorithm A in static way is obtained using published data as follows.

Headings

A. Validation of the robust algorithm A – part 1

The robust algorithm (A) as defined in ISO 13528, is implemented and the following results have been obtained.

Table 1. Typical presentation of Validation of robust algorithm (A) – see table E.4 ISO 13528-2022 Page.76 (Robust mean and robust standard deviation calculation for atrazine).

Iteration	0	1	2	3	4
x*-phi	0	0,2041	0,1997	0,2002	0,2003
x*+phi	0	0,3198	0,3160	0,3150	0,3148
phi	0	0,0578	0,0581	0,0574	0,0572
x1*	0,04	0,2042	0,2042	0,2041	0,2041
x2*	0,055	0,2042	0,2042	0,2041	0,2041
x3*	0,178	0,2042	0,2042	0,2041	0,2041
x4*	0,202	0,2042	0,2042	0,2041	0,2041
x5*	0,206	0,2060	0,2060	0,206	0,206
x6*	0,227	0,2270	0,2270	0,227	0,227
x7*	0,228	0,2280	0,2280	0,228	0,228
x8*	0,23	0,2300	0,2300	0,23	0,23
x9*	0,23	0,2300	0,2300	0,23	0,23
x10*	0,235	0,2350	0,2350	0,235	0,235
x11*	0,236	0,2360	0,2360	0,236	0,236
x12*	0,237	0,2370	0,2370	0,237	0,237
x13*	0,243	0,2430	0,2430	0,243	0,243
x14*	0,244	0,2440	0,2440	0,244	0,244
x15*	0,245	0,2450	0,2450	0,245	0,245
x16*	0,2555	0,2555	0,2555	0,2555	0,2555
x17*	0,26	0,2600	0,2600	0,26	0,2600
x18*	0,264	0,2640	0,2640	0,264	0,264
x19*	0,267	0,2670	0,2670	0,267	0,267
x20*	0,27	0,2700	0,2700	0,27	0,27
x21*	0,273	0,2730	0,2730	0,273	0,273
x22*	0,274	0,2740	0,2740	0,274	0,274
x23*	0,274	0,2740	0,2740	0,274	0,274
x24*	0,278	0,2780	0,2780	0,278	0,278
x25*	0,2811	0,2811	0,2811	0,2811	0,2811
x26*	0,287	0,2870	0,2870	0,287	0,287
x27*	0,287	0,2870	0,2870	0,287	0,287
x28*	0,288	0,2880	0,2880	0,288	0,288
x29*	0,289	0,2890	0,2890	0,289	0,289
x30*	0,295	0,2950	0,2950	0,295	0,295
x31*	0,296	0,2960	0,2960	0,296	0,296
x32*	0,311	0,3110	0,3110	0,311	0,311
x33*	0,331	0,3198	0,3160	0,3150	0,3148
x34*	0,4246	0,3198	0,3160	0,3150	0,3148
Mean	0,2512	0,2512	0,2512	0,25121	0,2512
Std	0,0672	0,2041	0,2041	0,2041	0,2041
(NEW) X*	0,2620	0,2579	0,2576	0,2576	0,2576

(NEW) S*	0,0386	0,0387	0,0383	0,0382	0,0381	Ecart type	5,0332	8,2398	8,2398	8,2398
						X*	18,2500	**19,0000**	**19,0000**	**19,0000**
						S*	6,6735	**5,7076**	**5,7076**	**5,7076**

The graphical form of such results are presented in the following figure.

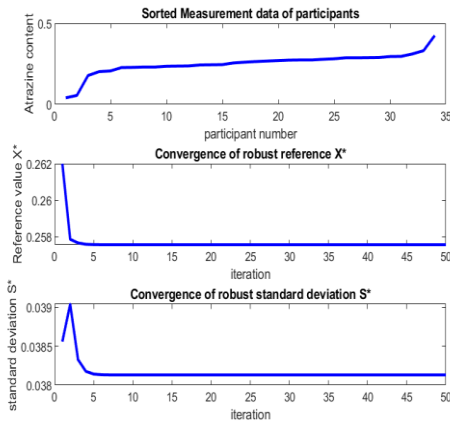


Fig. 4. Graphical representation of the validation of robust algorithm (A) – ISO 13528-2022.

B. Validation of the robust algorithm A – part 2

A second validation example of the robust algorithm (A) using typical data published in ISO 5725-5 is implemented and the following results have been obtained.

Table 2. Typical presentation of Validation of robust algorithm (A) – see table31 ISO 5725-5 page 49 (Example 6 mena cell %)

Iteration	0	1	2	3
x*+δ	"	8,2398	10,4385	10,4385
x*+δ	"	28,2603	27,5615	27,5615
δ	"	10,0103	8,5615	8,5615
x1*	13,4250	13,4250	13,4250	13,4250
x2*	13,4250	13,4250	13,4250	13,4250
x3*	13,7500	13,7500	13,7500	13,7500
x4*	14,4750	14,4750	14,4750	14,4750
x5*	17,0750	17,0750	17,0750	17,0750
x6*	18,2500	18,2500	18,2500	18,2500
x7*	21,0000	21,0000	21,0000	21,0000
x8*	21,2250	21,2250	21,2250	21,2250
x9*	23,6750	23,6750	23,6750	23,6750
x10*	26,2750	26,2750	26,2750	26,2750
x11*	26,4250	26,4250	26,4250	26,4250
Moyenne	19,0000	19,0000	19,0000	19,0000

C. Numerical implementation of the new dynamic robust algorithm

For simulation purpose, the data related to every participant i have been generated using random dynamic model defined as:

$Y(i,t) = a1.Y(i,t-1) – a2.Y(i,t-2) + b + \delta(t)$

Where a1, a2 and b are the model constant and δ(t) is a random normal distributed data. The index i defines the participant number. δ(t) is random normal distributed input with null mean value and with known standard deviation.

Fig.1 present the generated data i.e. the participant in conventional and sorted form.

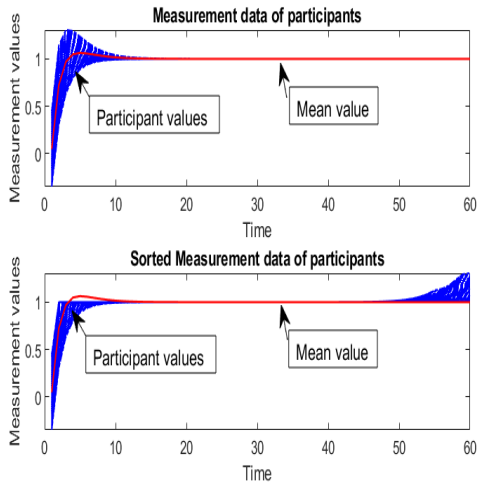


Fig. 5. Dynamic behavior data of participants

Sorted values are defined the p items of data into increasing order as defined in ISO 13528 standard.

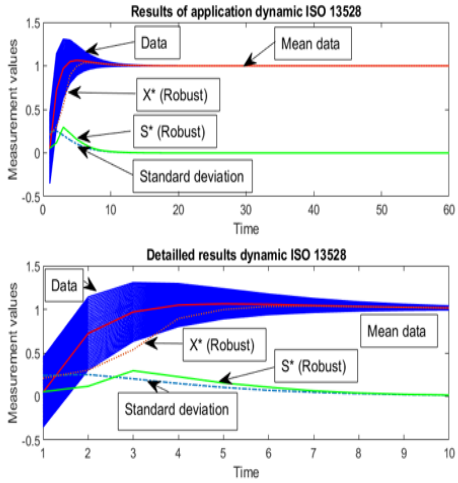


Fig. 6. detailed Dynamic robust x and s*.*

D. *Performance Analysis using Zscore*

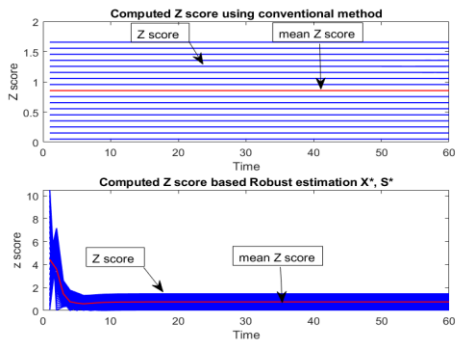


Fig. 7.Typical convergence of computed Z score.

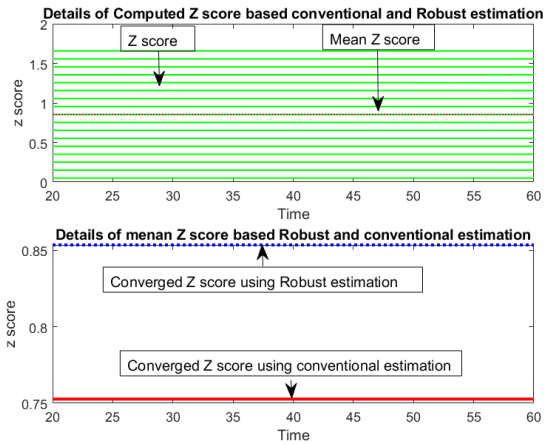


Fig. 8.Detailled values of performances after convergence (conventional and robust method)

IV. INTERPRETATION OF RESULTS

The obtained results show a good numerical implementation of the ISO 13528(A) algorithm for estimating intercomparison performance. The results are repeatable and consistent with the published values – see the results presented table 1 and table 2. This first step is used to validate the numerical implementation of the recusrsive form of ISO 13528.

The dynamic form of the ISO 13528 algorithm (A) was developed and implemented as defined in the followings figures. Indeed, the steps of the ISO 13528 algorithm (A) are applied at each instant (t) to the set of participants (p). A recursive loop related to time incrementation is thus added; the performance of the method given by the zscore, is also temporal (see Figures 5, 6, and 7). Figures 5 and 6 represent the dynamic behavior of the participants, while Figure 6 represents the data sorted in ascending order.

Figure 8 shows the calculation of performance, i.e., the z-score, using two methods. The first method calculates the z-score using robust values, i.e., S* and X*. The second method uses a conventional approach to calculate the z-score: i.e., the z-score is calculated by considering the reference value equal to the mean and the robust standard deviation equal to the standard deviation. According to Figure 6, obtaining robust values (X* and S*) requires convergence dynamics, i.e., stabilization of the solution. This is related to the properties of iterative algorithms.

V. CONCLUSIONS

The algorithm (A) ISO 13528 has been implemented and validated using example data provided from different reference standard particularly that given in ISO 13528 standard.

The results obtained after convergence are broadly similar. Its extension to the dynamic form was successfully achieved; the results show genuine convergence. The robust dynamic algorithm yields more severe performance results compared to the conventional method (in our case, the conventional z-score is lower than the robust z-score - 0.75 and 0.85).

References

[1] ISO 17025 General requirements for the competence of testing and calibration.

[2] ISO 13528 Statistical methods for use in proficiency testing by interlaboratory comparison.

[3] [3] ISO 5725-5 Accuracy (trueness and precision) of measurement methods and results — Part 5: Alternative methods for the determination of the precision of a standard measurement method.

[4] Kemal Kuş , Bülent Aydemir, A Comparative Analysis of Robust Statistical Methods by ISO 13528 for the Evaluation of Tensile Proficiency Testing Results, Celal Bayar University Journal of Science,Volume 22, Issue 1, 2026, p 32-41. Doi: 10.18466/cbayarfbe.1747154.

[5] Michèle Désenfant and al, Comparaison de

différentes stratégies d'évaluation de la performance dans les essais d'aptitude par comparaisons interlaboratoires, 17th International Congress of Metrology, 02003 (2015).

[6] ISO 17043:2023 Conformity assessment — General requirements for the competence of proficiency testing providers.

[7] Willem P. Cofino et al, Robustness comparison of three statistical methods commonly used in Proficiency Tests, Accreditation and Quality Assurance (2025) 30:507–519. 507–519. https://doi.org/10.1007/s00769-025-01669-3.

2026 IMEKO Joint Conference
TC7 - TC11 – TC13 - TC21 in conjunction with TC6 and Mathmet
Faro, Portugal, 27-29 October, 2026

A Statistical Workflow for the Metrological Evaluation of a Thermoregulated Bath Used in Hydro-Oceanographic Sensor Calibration

Dora Carinhas[1,2,3], José Rocha[1], Luís Alves[1], Paulo Infante[2]

[1] *Instituto Hidrográfico, Rua das Trinas, 49, 1249-093 Lisboa, Portugal; e-mail: dora.carinhas@hidrografico.pt*
[2] *CIMA/IIFA and DMAT/ECT, Universidade de Évora, 7000-671 Évora, Portugal*
[3] *CINAV, Escola Naval, Base Naval de Lisboa, Alfeite, 2810-001 Almada, Portugal*

I. INTRODUCTION

Reliable hydro-oceanographic observations depend on the metrological traceability of the sensors used to measure seawater temperature. In calibration laboratories, this traceability is influenced not only by the reference thermometer but also by the ability of the thermal medium to provide stable and spatially homogeneous conditions. Consequently, the performance of thermoregulated baths must be evaluated and periodically monitored to ensure that their contribution does not compromise calibration results, in accordance with ISO/IEC 17025 requirements [1].

Individual statistical techniques are commonly used to assess specific characteristics of thermal systems. However, isolated results do not necessarily provide a comprehensive basis for deciding whether a bath is suitable for its intended calibration use. This work therefore proposes an integrated statistical–metrological workflow combining spatial homogeneity, temporal stability, repeatability and reproducibility, statistical process control and process capability. The workflow is applied to a thermoregulated bath evaluated at three temperature set-points and three immersion heights, providing a structured approach for translating experimental measurements into objective evidence of metrological performance.

II. EXPERIMENTAL SETUP AND STATISTICAL WORKFLOW

The thermoregulated bath was evaluated at three temperature set-points (0 °C, 15 °C and 29 °C) and three immersion heights (70 mm, 250 mm and 345 mm), selected to characterise its performance throughout the working volume. Temperature measurements were obtained using a reference thermometer (SBE35) and two standard platinum resistance thermometers (SPRT252 and SPRT254). Data acquisition began after steady-state conditions had been verified, followed by consistency and completeness checks before statistical analysis.

The proposed workflow, presented in Fig. 1, integrates complementary statistical tools according to the metrological question addressed at each stage. Spatial homogeneity between immersion heights was assessed using one-way analysis of variance (ANOVA) and the non-parametric Kruskal–Wallis test. Levene's test was additionally applied to assess equality of variances. Temporal stability was investigated using the Mann–Kendall trend test and linear regression.

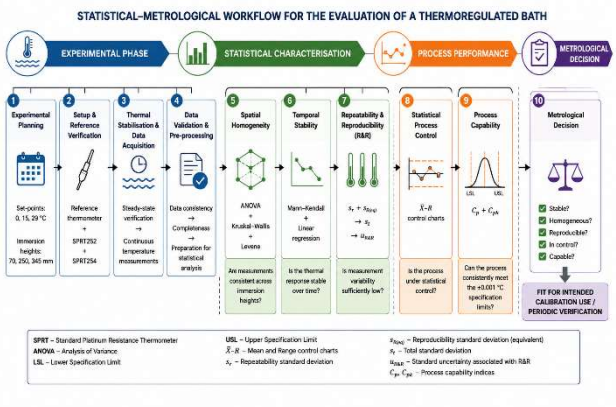


Fig. 1. Statistical–metrological workflow for the evaluation of a thermoregulated bath.

Repeatability and reproducibility were evaluated from the variability observed within and between the three thermometers. Statistical process control was subsequently performed using $\bar{X}$ and R control charts to identify potential special-cause variation [2], [3]. After confirming that the process remained under statistical control, its ability to satisfy the adopted specification limits of ±0.001°C was assessed using the capability indices *Cp* and *Cpk* [2], [4]. The combined evidence supported the final decision regarding the suitability of the bath for its intended calibration use and periodic verification.

III. RESULTS AND DISCUSSION

No statistically significant differences between immersion heights were detected at any of the evaluated set-points. ANOVA, Kruskal–Wallis and Levene's tests consistently returned $p>0.05$, providing no evidence of differences in mean temperature, distribution or variance across the working volume. Similarly, the Mann–Kendall test returned $p>0.05$ at all set-points, indicating no statistically significant monotonic temporal trend. These complementary results support the spatial homogeneity and temporal stability of the bath under the evaluated conditions (Table I).

Table 1. Statistical results for spatial homogeneity and temporal stability

Statistical test	**0 °C**	**15 °C**	**29 °C**
ANOVA, F (p)	0.49 (0.62)	0.13 (0.88)	0.65 (0.53)
Kruskal-Wallis, H (p)	1.04 (0.59)	0.33 (0.84)	1.42 (0.48)
Levene, W (p)	0.34 (0.71)	0.07 (0.93)	0.46 (0.69)
Mann–Kendall, Z (p)	0.46 (0.64)	0.29 (0.77)	0.36 (0.72)

The repeatability and reproducibility assessment showed low measurement variability. The total standard deviation ranged from 0.000260°C to 0.000329°C, while the associated standard uncertainty, $u_{R\&R}$, ranged from 0.00005°C to 0.00006°C. These values correspond to approximately 5–6% of the adopted ±0.001°C specification limits, indicating that variability associated with the measurement system remained small relative to the acceptance interval.

The control-chart analysis identified no points outside the control limits or systematic patterns, providing no evidence of special-cause variation under the evaluated conditions.

Process capability analysis subsequently showed that both *Cp* and *Cpk* exceeded the adopted criterion of 1.33 at all set-points. The agreement between inferential testing, *R&R* assessment, statistical process control and capability indices provides convergent evidence that the bath was stable, homogeneous and capable of maintaining the specified thermal conditions.

IV. CONCLUSIONS

The proposed statistical–metrological workflow provided a structured approach for evaluating the performance of a thermoregulated bath used in hydro-oceanographic sensor calibration. The combined results demonstrated spatial homogeneity, temporal stability, low repeatability and reproducibility variability, statistical control and adequate process capability at the three evaluated set-points. Under these experimental conditions, the bath was therefore considered suitable for its intended calibration use.

More generally, the sequential integration of statistical tests, *R&R* assessment, control charts and capability indices enables different sources of evidence to be translated into an objective metrological decision. The workflow is reproducible and may support the initial qualification and periodic verification of thermoregulated baths and other thermal systems used in calibration laboratories. Its application to additional temperature ranges and operating conditions should be considered before extending the conclusions beyond the conditions evaluated.

REFERENCES

[1] ISO/IEC 17025:2017, "General requirements for the competence of testing and calibration laboratories", International Organization for Standardization, Geneva, 2017.

[2] D. C. Montgomery, "Introduction to Statistical Quality Control", 7th ed. Hoboken, NJ, USA: John Wiley & Sons, 2013.

[3] ISO 7870-2:2013, "Control charts — Part 2: Shewhart control charts". Geneva, Switzerland: International Organization for Standardization, 2013.

[4] ISO 22514-1:2014, "Statistical methods in process management — Capability and performance — Part 1: General principles and concepts". Geneva, Switzerland: International Organization for Standardization, 2014.