Archives of Computational Methods in Engineering
https://doi.org/10.1007/s11831-020-09473-7

SURVEY ARTICLE

Unified Framework and Survey for Model Verification, Validation and Uncertainty Quantification

Stefan Riedmaier[1] · Benedikt Danquah[1] · Bernhard Schick[2] · Frank Diermeyer[1]

Received: 10 July 2020 / Accepted: 23 July 2020

Abstract
Simulation is becoming increasingly important in the development, testing and approval process in many areas of engineering, ranging from finite element models to highly complex cyber-physical systems such as autonomous cars. Simulation must be accompanied by model verification, validation and uncertainty quantification (VV&UQ) activities to assess the inherent errors and uncertainties of each simulation model. However, the VV&UQ methods differ greatly between the application areas. In general, a major challenge is the aggregation of uncertainties from calibration and validation experiments to the actual model predictions under new, untested conditions. This is especially relevant due to high extrapolation uncertainties, if the experimental conditions differ strongly from the prediction conditions, or if the output quantities required for prediction cannot be measured during the experiments. In this paper, both the heterogeneous VV&UQ landscape and the challenge of aggregation will be addressed with a novel modular and unified framework to enable credible decision making based on simulation models. This paper contains a comprehensive survey of over 200 literature sources from many application areas and embeds them into the unified framework. In addition, this paper analyzes and compares the VV&UQ methods and the application areas in order to identify strengths and weaknesses and to derive further research directions. The framework thus combines a variety of VV&UQ methods, so that different engineering areas can benefit from new methods and combinations. Finally, this paper presents a procedure to select a suitable method from the framework for the desired application.

1 Introduction

Simulation is getting more and more important in the development and testing process across many engineering fields. It has a long-standing history in numerical simulation areas such as finite element methods (FEM) and computational fluid dynamics (CFD). It is rooted in the development process of complex systems such as aircrafts, trains and vehicles in order to save time and costs and develop high-quality products [39]. Simulation has found its way into the testing process and regulatory approval, for example in the research field of safeguarding autonomous vehicles [149, 181].

Since simulation plays a key role in many engineering fields, accurate models of reality become a decisive factor for the credibility and trustworthiness of simulation, especially in safety-critical areas. Model-based activities such as model verification and validation (V&V) as well as uncertainty quantification (UQ), together referred to as VV&UQ, assess errors and uncertainties inherent in any simulation model. An excellent introduction into VV&UQ can be found in [19, 31, 129].

A large number of publications dealing with aspects of VV&UQ have been published in recent years. The following six heterogeneous examples represent main approaches in their respective fields. They will be shortly introduced here and explained later in more detail with an analysis of their strength and weaknesses. Oberkampf and Roy [133] present the so-called Probability Bound Analysis (PBA), which uses Frequentist statistics and imprecise probabilities [22] to bound model predictions. Sankararaman

✉ Stefan Riedmaier
riedmaier@ftm.mw.tum.de

Benedikt Danquah
danquah@ftm.mw.tum.de

Bernhard Schick
bernhard.schick@hs-kempten.de

Frank Diermeyer
diermeyer@ftm.mw.tum.de

[1] Technical University of Munich, Institute of Automotive Technology, Boltzmannstr. 15, 85748 Garching b. München, Germany

[2] Kempten University of Applied Sciences, Adrive Living Lab, Bahnhofstr. 61, 87435 Kempten, Germany

Published online: 07 August 2020

and Mahadevan [160] analogously describe a Bayesian approach. Both VV&UQ frameworks quantify and aggregate all sources of uncertainties and are frequently used in numerical simulations. Crespo et al. [35] propose Interval Predictor Models (IPMs) that directly predict interval-valued quantities to bound all future experiments within them. Hills [84] propose a meta-model approach to correct the erroneous model predictions through a data-driven model. In contrast to numerical simulations, complex systems usually rely on conventional deterministic V&V methods. For example, model validation in automotive and railway vehicle simulations is often based exclusively on a tolerance approach. In this approach, the deviation between simulation and experiment is determined and compared with an allowed tolerance. Eek et al. [56] describe an output uncertainty approach by making simplifications to PBA to enable an application to complex aircraft simulations.

In addition to the heterogeneous VV&UQ landscape across many engineering fields, there are still general challenges that need to be solved by the entire research community. Roy [154], Schroeder and Mullins [164], Mullins et al. [128] agree that one of the main current challenges is how to aggregate model uncertainties. The uncertainties are often identified in calibration and validation experiments, but the actual model predictions will take place under untested application scenarios, which may differ significantly from the experiments. In addition to such a challenging extrapolation in the scenario space, there are application fields such as nuclear reactor safety or spacecrafts, which require an extrapolation in the system hierarchy. They may only offer experiments at component-level due to safety or costs, but the entire model will be used for predictions at system-level. Furthermore, complex systems are often dynamic and show time-variant behavior, which poses difficulties for aggregation.

There are survey papers that provide an overview about certain aspects of VV&UQ. For example, Funfschilling and Perrin [66] summarize UQ methods in rail vehicle simulations and Durst et al. [47], Sargent and Balci [161] give an historical review of V&V methods. However, none of these papers gives an overview about multiple engineering fields, focuses on the aggregation of errors and uncertainties and develops a novel framework to unite the research activities across the different fields.

The main contributions of this publication are a

1. novel modular and unified framework for aggregation of errors and uncertainties shown in Fig. 1 with
2. formalization in a unified mathematical notation,
3. comprehensive survey and literature review about VV&UQ methods across many application fields with
4. integration into our unified framework, and
5. detailed comparison of main approaches and different application fields to derive future research directions.

In the following sections we describe our framework in Fig. 1 step-by-step. Section 2 begins with the domains of the framework, which reflect a model-based process including typical activities, errors and uncertainties. Section 3 describes continuously the columns of the framework with unique interface descriptions for each block. Section 4 zooms into the simulation model block of the framework to show various manifestations including deterministic, non-deterministic, hierarchical, dynamical and formal models. Section 5 extends the manifestations of the simulation model block to the entire framework with its manifestations. Section 6 focuses on aggregation methods of errors and uncertainties, which currently pose a great challenge. Section 7 gives a comprehensive survey with many references and examples across several engineering fields. Section 8 discusses and compares the strength and weaknesses of the different aggregation methods and draws conclusions. Finally, the conclusion in Sect. 9 summarizes the most important research findings.

In summary, we introduce a new framework that includes the majority of the individual approaches as a subset. Therefore, in Sects. 2 and 3 we will successively describe the framework first and then use it as a taxonomy in Sects. 4 and 6 to modularize main VV&UQ approaches and to classify them within it. This will enable a variety of new combinations in the future. In case approaches skip parts of the framework or in exceptional cases deviate from it, this will be mentioned in the text. Sections 7 and 9 contain examples, method comparisons and conclusions across different application fields. The paper thus offers comprehensive information on VV&UQ, including the latest developments, in one central location.

2 Model-Based Process

The framework in Fig. 1 represents a model-based testing process. The simulation model must be developed in advance by creating a conceptual model, deriving a mathematical model with equations from it, and implementing them as a computer code to obtain a computational model for simulation [133, p. 14]. Then, model verification assesses the numerical solution of the computer code, model calibration estimates the model parameters, model validation assesses the model-form by comparison with physical tests and finally the "validated" model can be used for prediction under new application conditions. The framework represents each of the four model-based activities with an own domain to reflect the model-based process from top to bottom. The activities address various

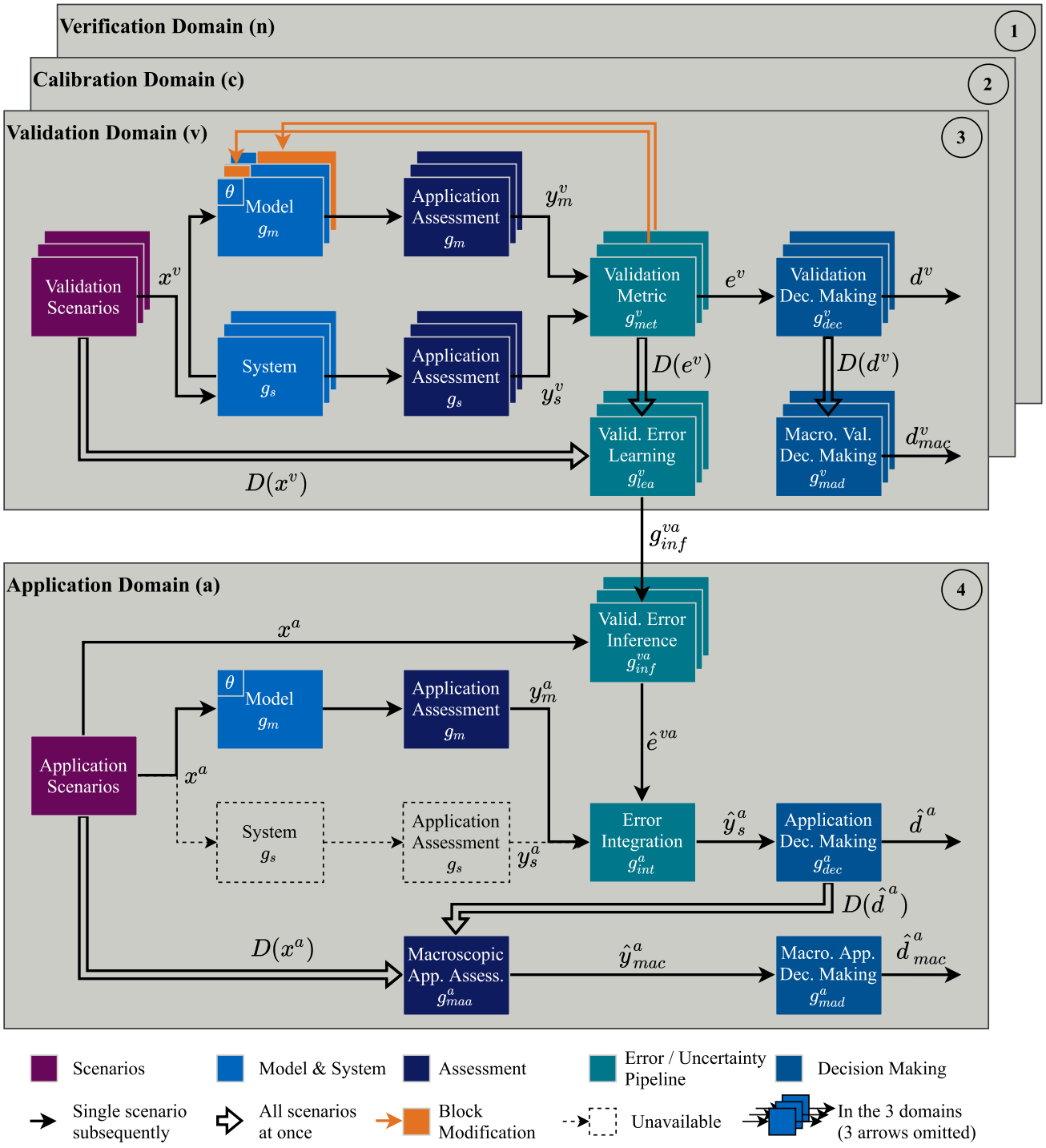


Fig. 1 Model-based framework. In principle, the process runs from 1 to 4. However, there may be loops, especially from the validation domain to the calibration domain, if the parameters need to be adjusted again. The inferred errors of the three stacked blocks can be merged in the error integration. The back-most of the three stacked system blocks, which refers to the verification domain, is actually not a system, but a mathematical model. (Color figure online)

types and sources of errors and uncertainties that are unavoidable in a computer simulation. This section introduces terminology and notation, types and sources of error and uncertainties as well as model-based activities with the corresponding domains. For more detailed information, the interested reader is referred to [133]. Some of the following explanations are based on this comprehensive introduction book.

2.1 Terminology

Oberkampf and Roy [133, Chap. 2.1–2.2] collect definitions for model-based activities across different communities and extend them. In the following we list the relevant ones for our framework.

- Model verification is "the process of determining that a model implementation accurately represents the developer's conceptual description of the model".
  - Code verification is "the process of determining that the numerical algorithms are correctly implemented in the computer code and of identifying errors in the software".
  - Solution verification is "the process of determining the correctness of the input data, the numerical accuracy of the solution obtained, and the correctness of the output data for a particular simulation".
- Model calibration is "the process of adjusting physical modeling parameters in the computational model to improve agreement with experimental data".
- Model validation is "the process of determining the degree to which a model is an accurate representation of the real world from the perspective of the intended uses of the model".
- Model prediction deals with "interpolating or extrapolating the model beyond the specific conditions tested in the validation domain to the conditions of the intended use of the model".

In accordance with [133, 135, 179, p. 88], we call the input conditions of a simulation or experiment a scenario. We summarize all conditions in a domain, sometimes also referred to as regime or space. This results in the application domain [133, Fig. 2.10] with application scenarios for model prediction and in the validation, calibration and verification domain with scenarios, respectively. We refer to the outputs of simulation and experiment as responses or, for emphasis, as response quantities of interest (QoIs) [135].

Various terms are used to reflect a deviation. Oberkampf and Roy [133] use the terms error, difference, bias and mismatch. Kennedy and O'Hagan [104] use the terms residual, inadequacy and discrepancy. These terms are all aimed at assessing the accuracy [133] or fidelity [15] of the simulation. We usually stick to the term error. In addition, VV&UQ approaches should also assess the robustness of the model's fidelity given uncertain parameters [15, 141, 150].

2.2 Types of Errors and Uncertainties

An error reflects the deviation to the true value. If errors can not be quantified exactly, uncertainties arise. They can be either aleatory, epistemic or mixed. Epistemic uncertainties arise from lack of knowledge and can be estimated or even reduced by gaining knowledge. Aleatory uncertainties refer to stochastic effects and variability and can only be quantified.

Deterministic, precisely known quantities without uncertainty are described by point values. Aleatory uncertainties can be described by probability distributions in form of probability density functions (PDFs) or cumulative distribution functions (CDFs), epistemic uncertainties by intervals and mixed uncertainties by a family of distributions or imprecise probabilities. Probability boxes (p-boxes) are a mixture of intervals and probabilities by adding a width to a CDF.

2.3 Mathematical Notation

We denote a point value as x, as usual a PDF as $f(x)$ and a CDF as $F(x)$, an interval as $I(x)$ and a p-box as $B(x)$. We usually omit the random variable X in $f_X(x)$ or $F_X(x)$ for the sake of consistency and readability. Strictly speaking, multiple repetitions of an experiment yield an empirical distribution function (EDF) for each quantity. The EDF has steps compared to the smooth true distribution function. This also applies to the simulation, but there the steps are very fine due to many samples.

Left and right limits $\underline{x}$ and $\overline{x}$ bound the interval

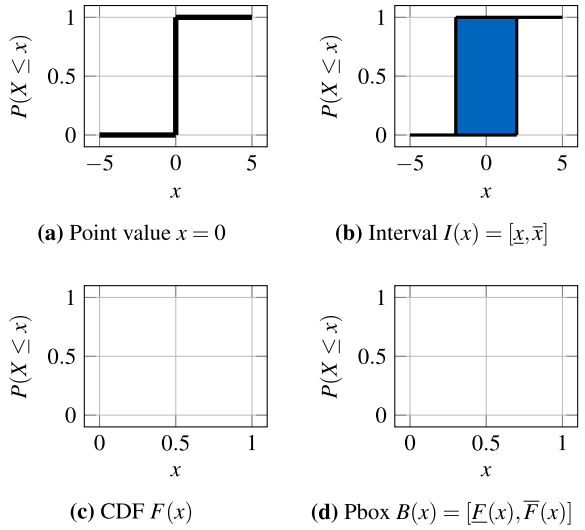
$$I(x) = [\underline{x}, \overline{x}] = \{x \mid \underline{x} \leq x \leq \overline{x}\} \tag{1}$$

in set-builder notation. In analogy, CDFs form the left limit $\underline{F}(x)$ and right limit $\overline{F}(x)$ to bound the p-box [22, Eq. (12)]

$$B(x) = [\underline{F}(x), \overline{F}(x)] = \{F(x) \mid \underline{F}(x) \leq F(x) \leq \overline{F}(x)\}. \tag{2}$$

We visualize the different mathematical structures in Fig. 2. A point value can also be interpreted as a degenerate interval or as a degenerate probability distribution and all of them as a degenerate p-box.

We introduce this uniform notation with x, $I(x)$, $F(x)$ and $B(x)$ so that the equations of the framework will be generically applicable. An overview about all quantities can be found directly in the framework in Fig. 1 for the deterministic manifestation (x) as representative and in the following figures for the remaining manifestations. Since identical quantities will occur in different domains, we will use a symbol in the upper index of a quantity for distinction (x^v, x^a). A collection of the quantities for all input scenarios is denoted by the data tuple $D(x)$.



(a) Point value $x = 0$ **(b)** Interval $I(x) = [\underline{x}, \overline{x}]$

(c) CDF $F(x)$ **(d)** Pbox $B(x) = [\underline{F}(x), \overline{F}(x)]$

Fig. 2 (Degenerate) Mathematical structures with cumulative probabilities $P(X \leq x)$ [63, Fig. 2]. (Color figure online)

2.4 Sources of Errors and Uncertainties

Simulations and experiments involve many errors e compared to nature's true behavior [7, Eq. (1-5-6)]

$$y_{true} = g_s(x) - e_{y,obs} \tag{3}$$

$$= g_m(x, \theta, h) - (e_m + e_x + e_\theta + e_h). \tag{4}$$

We denote the inputs as x, the model parameters as θ, the step size as h, the responses as y, the real system as g_s and the simulation model as g_m. Regarding the errors, $e_{y,obs}$ stands for the observation errors of the responses, e_m for model-form errors, e_x for input errors, e_θ for parameter errors and e_h for discretization errors on behalf of all kinds of numerical errors.

Equation (3) focuses on experimental errors, whereas (4) focuses on model errors. The responses of the system $y_s = g_s(x)$ and model $y_m = g_m(x, \theta, h)$ are known and result in a total error e. However, the individual sources of errors are unknown and might reinforce or compensate each other depending on the specific input conditions. The latter can cause a misleading trustworthiness in simulation. Therefore, it is crucial to separately quantify all sources of errors during the different model-based activities.

If the errors cannot be determined exactly, corresponding uncertainties must be considered. However, this requires a probabilistic representation. Assuming there are uncertain inputs and parameters and a total error e resulting from (3) and (4). Then a variance decomposition based on the law of total variance [48, Eq. (3, 4)]

$$\mathrm{Var}[y_s] = \mathrm{Var}[\mathrm{E}[y_s \mid x]] + \mathrm{E}[\mathrm{Var}[y_s \mid x]] \tag{5}$$

Table 1 Relationships between sources and types of uncertainties, model-based activities and domains

Sources	Type	Activity	Domains
Model solution	E	Model verification	n
Natural variability	A	Experiments	c, v
Measurement	A, M	Experiments	c, v
Inputs	E (A, M)	UQ	c, v, a
Model parameters	E (A, M)	Model calibration	c
Model-form	E	Model validation	v
–	–	Model prediction	a
Extrapolation	E	Extrapolation	n, c, v → a
All	E, A, M	Aggregation	n, c, v → a

Uncertainty types: E = epistemic, A = aleatory, M = mixed; Domains: n = verification (numerical), c = calibration, v = validation, a = application

$$= \mathrm{Var}[y_m + \mathrm{Mean}[e]] + \mathrm{E}[\mathrm{Var}[e]] \tag{6}$$

yields two summands. The first one contains the variance of the simulation, the second that of the total error. Input uncertainty quantification and propagation aims at quantifying the first summand, while model validation mainly aims at quantifying the second summand. Since both individually under-approximate the total prediction uncertainty, all uncertainties must be aggregated [48].

2.5 Model-Based Activities

Typical model-based activities focus on individual sources of errors and uncertainties in order to quantify them accurately. Table 1 summarizes the corresponding relationships. The sources of uncertainties have different uncertainty types and occur in different domains. Model verification, calibration, validation and prediction are represented each with an own domain. The other activities occur in multiple domains or transfer errors and uncertainties between domains.

2.5.1 Model Verification

Model verification should definitely be carried out by the tool manufacturer before model calibration, validation and prediction. Solution verification identifies numerical errors such as the discretization error [160, Eq. (1)]

$$e_h = e^n = g_m(x^n, \theta, h) - g_{mat}(x^n, \theta) \tag{7}$$

due to the solution y_m of the computational model g_m on a discrete computer with step size h compared to the exact solution y_{ex} of the mathematical model g_{mat}. Since for complex mathematical models it is often impossible to calculate the exact solution, Richardson extrapolation estimates the exact solution [133, Eq. (8.71)]

$$\hat{y}_{ex} = g_m(x^n, \theta, h_1) + \frac{g_m(x^n, \theta, h_1) - g_m(x^n, \theta, h_2)}{r^p - 1} \tag{8}$$

by performing code-to-code comparisons with three simulations of fine, medium and coarse step sizes $h_1 < h_2 < h_3$. The grid refinement factor describes the ratio $r = h_3/h_2 = h_2/h_1$ and yields the observed order of accuracy [160, Eq. (1)]

$$p = \frac{\ln\left(\frac{g_m(x^n,\theta,h_3) - g_m(x^n,\theta,h_2)}{g_m(x^n,\theta,h_2) - g_m(x^n,\theta,h_1)}\right)}{\ln(r)}. \tag{9}$$

Richardson extrapolation provides the numerical error [155, Sect. 5.1]

$$\hat{e}^n = g_m(x^n, \theta, h_3) - \hat{y}_{ex} \tag{10}$$

due to discretization for the coarsest step size. Based on Roache's Grid Convergence Index, it can be converted to a numerical uncertainty with a safety factor [155, Sect. 5.1]:

$$I(\hat{e}^n) = \pm F_{safety} |\hat{e}^n|. \tag{11}$$

2.5.2 Experiments

Experiments always contain a certain natural variability and aleatory uncertainty. Measurement systems are required to observe nature's true behavior, but introduce a bias error and aleatory uncertainty itself. The response quantities are observed during experiments for comparison with simulation. Regarding the observation of input quantities for re-simulation, Mullins et al. [126] distinguish fully, partially and un-characterized experiments. The former reports the inputs as precisely known point values, the latter reports at most the entire domain. Partially characterized experiments are placed in between and contain point values, probabilities or intervals. Observation errors such as $e_{y,obs}$ should be reduced as much as possible because they interfere with the quantification of input, parameter and response quantities and other sources of errors. Regarding the quantification of measurement uncertainties, the interested reader is referred to traditional standards [98] and more sophisticated methods based on statistical Design of Experiments (DoE) [147].

2.5.3 Input Uncertainty Quantification

Deterministic simulations often assume fully characterized experiments or neglect the resulting input and parameter errors e_x and e_θ. However, this assumption does not hold for partially and un-characterized experiments and also not for model predictions in the application domain, if the conditions are unknown in advance. Therefore, non-deterministic simulations rigorously quantify the input and parameter uncertainties in each domain and propagate them by the simulation model to obtain the resulting output uncertainties [48]. We dedicate Sect. 4.1.3 to UQ of non-deterministic simulations.

2.5.4 Model Calibration

In model calibration and validation, (3) and (4) must be compared with each other. Model calibration infers the parameters from the comparison so that the deviations are minimized:

$$\arg\min_{\theta}(g_m(x, \theta) - g_s(x)). \tag{12}$$

The framework in Fig. 1 shows an inverse (orange) arrow pointing back to the model parameters to emphasize that inverse methods are used for calibration. In a broader sense, direct measurements for individual parameters can also be performed for parameter identification in the "calibration" domain. This should even be preferred before inverse calibration methods are used [133, p. 45].

2.5.5 Model Validation

A possible combination with model calibration makes it very important that an independent authority assesses the model error

$$e^v = y_m^v - y_s^v \tag{13}$$

using a separate validation data set [133, p. 49]. The model validation of deterministic simulations quantifies the remaining total error e^v. The model validation of non-deterministic simulations considers input and parameter uncertainties to better isolate the epistemic model-form uncertainty in $I(e^v)$.

2.5.6 Model Prediction

Model prediction will occur in the application domain under conditions, which the real system will encounter after approval. The great advantage of the model-based process is that the required system characteristics can be extensively evaluated with the "validated" simulation model

$$g_m(x^a, \theta) \tag{14}$$

before approval and without having to perform further real tests in the application domain. For this reason, the system in the application domain is shown in gray and dashed in the framework. However, this is only possible, if the model prediction is accompanied by an aggregation of all errors and uncertainties.

2.5.7 Interpolation and Extrapolation Uncertainties

If the input conditions of a validation scenario x^v vary from a future application scenario x^a, that introduces interpolation and extrapolation uncertainties. Figure 3 illustrates this with an example. In general, different inter-domain constellations are conceivable. The validation and application domain may be congruent, one may be a subset of the other, there may be an intersection, or both may even be completely separate [133, Fig. 2.10]; [49, Fig. 2]. The latter causes a large extrapolation uncertainty and should be avoided whenever possible.

2.5.8 Aggregation of Errors and Uncertainties

The big challenge is that the sources of errors and uncertainties affect each other and should be quantified separately [59]. To enable the aggregation of all sources of errors and uncertainties for model prediction, connections from each domain to the application domain are required. A direct connection can be seen in the vertical error pipeline of the framework. In addition, there are also indirect connections to the application domain. Model verification can adjust the model-form and calibration the model parameters [160]. Both are indicated by the inverse arrows in the framework. Since the simulation model is part of each domain, this affects all subsequent domains in the model-based process including the actual application domain. We dedicate the separate Sect. 6 to the aggregation of errors and uncertainties.

3 Framework Blocks

This section describes each block of the framework. The sub-sections are ordered column-wise from left to right, since many blocks occur several times in different domains. In case of overlaps between the verification, calibration and validation domain, we use the validation domain as a representative example in the remaining paper to avoid unnecessary repetitions. For each block, we add mappings from inputs to outputs, since clear interface descriptions are a key requirement for a modular framework. We will start, as generically as possible, with the deterministic manifestation of the framework including errors and will extend it in the following main sections to non-deterministic models with uncertainties and to hierarchical, dynamical and formal models.

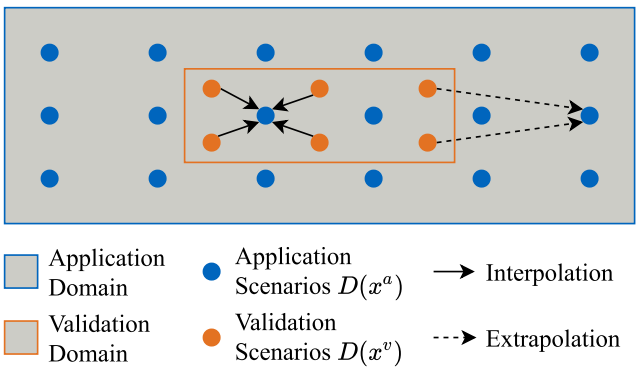


Fig. 3 Inter- and extrapolation in scenario space. (Color figure online)

3.1 Scenarios

Within the different domains, a test scenario x is input to the system and model and influences their behavior. We aggregate all N_D scenarios of a domain in a data tuple

$$D(x) = [x_1, \dots, x_i, \dots, x_{N_D}] \tag{15}$$

to preserve the order for assignment to the corresponding scenario results later. The N_D scenarios are composed of $N_{D,d}$ distinct ones to cover the scenario space and $N_{D,r}$ repetitive ones to consider the variability. The six orange validation scenarios and 18 blue application scenarios in Fig. 3 are examples for the former type, whereas the latter means to repeat the same orange validation scenario. There is one scenario tuple for each domain $D(x^n)$, $D(x^c)$, $D(x^v)$ and $D(x^a)$. The amount of scenarios N_D varies between the different domains. The number of application scenarios is often significantly larger than the number of calibration and validation scenarios to reduce the physical testing effort and legitimize the model-based process.

3.2 System and Model

Both the real system, denoted s, and the corresponding simulation model, denoted m, with model parameters θ encounter an input scenario x and show a certain response behavior y. Generally speaking, both perform a mapping g from inputs x to responses y:

$$g_s : x \mapsto y_s \tag{16}$$

$$g_m : (x, \theta) \mapsto y_m. \tag{17}$$

Strictly speaking, the experimentalist can just control a subset x_{con} of all relevant scenario inputs x, but the uncontrolled ones x_{unc} also affect the system behavior. This results in a natural variability in the repetition of real experiments. Besides, since a model is just a simplified abstraction of the real system, the modeler just considers a subset x_{mod} of the scenario quantities that actually affect the system behavior. The complementary quantities x_{unm} are either unknown or neglected by the modeler:

$$g_s : (x_{con}, x_{unc}) \mapsto y_s, x = x_{con} \cup x_{unc} \tag{18}$$

$$g_m : (x_{mod}, \theta) \mapsto y_m, x_{mod} \subseteq x. \tag{19}$$

In the calibration and validation domain, the physical tests are usually carried out first. During these tests, the input variables x^v_{mod} that the simulation model requires should be measured in addition to the response variables y^v_s. This enables a re-simulation under similar conditions as in the physical tests. In the application domain, the selected scenarios are directly forward to the simulation tool and an actual model prediction is performed.

3.3 Application Assessment

Many application fields perform a post-processing of model and system results after the actual simulations and tests and before decision making. The framework reflects this in the application assessment column.

3.3.1 Application Assessment in the Application Domain

The assessment depends on the respective application and can vary greatly. Typical examples are filtering of signals or extraction of characteristic values to obtain the actual QoIs. There are also applications that do not perform post-processing at all. Nevertheless, we want to offer it as an option in the framework. We therefore refrain from introducing a new symbol for the assessment and stick to the established y. This means that the assessment results are contained in the QoI y.

3.3.2 Application Assessment in the Validation Domain

We use the application assessment not only for the simulation model in the actual application domain, but also for model and system within the other domains. This is consistent with the nature of model validation, since there is no generic model validity, but always a validity with respect to a use case. In model validation, checking the identical QoI that will later be used for decision-making in the application domain is the best basis for trustworthiness of simulation [92, 180].

3.3.3 Macroscopic Application Assessment

Typically, the responses are only assessed for each individual scenario x independently. We call this a microscopic assessment and omit the term microscopic in text, figure and equations for simplicity, since it is the default case. Still, there are some applications that also make a macroscopic statement about all scenarios $D(x^a)$ in the application domain together [99]. The decomposition into a microscopic and macroscopic assessment has a big advantage, especially if the the macroscopic assessment

$$g_{mac} : (D(x^a), D(\hat{d}^a)) \mapsto y^a_{mac} \tag{20}$$

is based on all estimated binary microscopic decisions $D(\hat{d}^a)$ (Sect. 3.5). Then model validation can intervene at this interface and ensure that all binary decisions are correct, so that the macroscopic statements do not need to be validated directly. For example, a macroscopic assessment can weight all binary microscopic decisions with the importance of the corresponding scenarios and aggregate them to an overall risk measure.

3.4 Error Pipeline

In model calibration and validation, the simulation result y_m is compared with the experimental result y_s as a reference. If the system result y_s is unknown, the simulation result y_m and the error e must be known to determine it:

$$e = y_m - y_s \tag{21}$$

$$y_s = y_m - e. \tag{22}$$

The two very simple equations are visualized in Fig. 4, as they represent the basic building block of the error pipeline in the framework. Since the error can only be calculated in the (calibration or) validation domain, but the system result is unknown in the application domain, we need to separate both equations:

$$e^v = g_m(x^v, \theta) - g_s(x^v) \tag{23}$$

$$g_s(x^a) = g_m(x^a, \theta) - e^a. \tag{24}$$

It gets obvious that both errors appear in different domains and an interpolation and extrapolation from x^v to x^a is required. Thus, we extend the main principle in Fig. 4 with

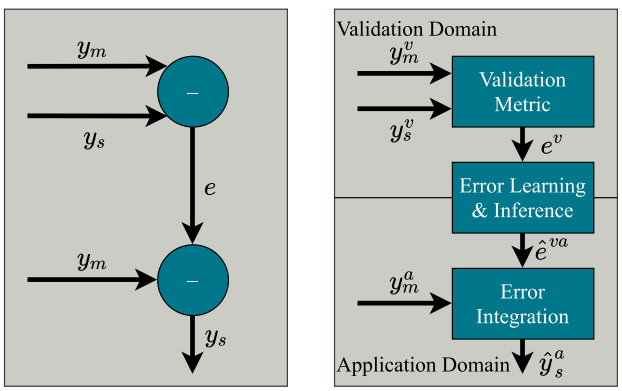


Fig. 4 Basic aggregation principle. (Color figure online)

the two framework blocks error learning and inference. The former tries to learn a data-driven behavior model of the errors within the validation domain so that the latter can perform a prediction in the application domain. The greater the difference between the domains and thus the greater the extrapolation uncertainties, the more important the two blocks are. In addition, we replace both subtractions with a validation metric (error calculation) and an error integration to get more flexibility.

3.4.1 Metric

Since the quantities can have different mathematical representations such as time signals or probabilities, the simple subtraction of point values will not be sufficient as metric in Sect. 5.2. Generally speaking, a calibration or validation metric

$$g^v_{met} : (y^v_s, y^v_m) \mapsto e^v \tag{25}$$

takes the responses from system and model and calculates an error. The verification metric uses the mathematical model or very accurate computational models as reference instead of the system:

$$g^n_{met} : (y^n_m, y^n_{exact}) \mapsto e^n. \tag{26}$$

3.4.2 Error Learning

From an abstract point of view we propose to decompose the modeling task in a combination of a physics-based modeling of the system behavior with a data-driven modeling of the error [133, p. 657]; [104]. This is more robust than the pure modeling of the system behavior. The errors typically fluctuate to a small extent around the nominal physics-based simulation model as stable baseline. The decomposition is especially beneficial if independent authorities perform the model validation and the intellectual property (IP) of the nominal model must be protected. The error learning

$$g^v_{lea} : (D(x^v, e^v)) \mapsto g^{va}_{inf} \tag{27}$$

intends to model the errors across the validation domain. It takes the data tuple $D(x^v, e^v)$ of scenario and error pairs as training data set to learn a data-driven model g^{va}_{inf}.

3.4.3 Error Inference

After learning the data-driven error model in the validation domain, it can be applied for inference (prediction)

$$g^{va}_{inf} : x^a \mapsto \hat{e}^{va} \tag{28}$$

of the error in the application domain. With $\hat{e}^{va}$ we emphasize that this is an estimate of the true error by extrapolation from the validation to the application domain.

3.4.4 Error Integration

In the application domain, the responses from model prediction are available as well as the inferred errors from model verification, calibration and validation. The uncertainty integration

$$g^v_{int} : (y^a_m, \hat{e}^{na}, \hat{e}^{ca}, \hat{e}^{va}) \mapsto \hat{y}^a_s \tag{29}$$

aggregates all of them to estimate the actual behavior of the real system in the application domain. In Sect. 6 we will distinguish different integration techniques.

3.5 Decision Making

In decision making, results are compared with thresholds t to obtain Boolean values that indicate whether the results are good enough. The decision making in the validation domain

$$g^v_{dec} : (e^v, t^v_e) \mapsto d^v = \begin{cases} 1 \text{ if } e^v \le t^v_e \\ 0 \text{ else} \end{cases} \tag{30}$$

compares an error with a permissible tolerance (threshold, model accuracy requirement) t^v_e. The comparison can be done as shown with the error and preliminary requirements in the validation domain or with the extrapolated error and final requirements in the application domain [133, p. 478]. The tolerance comparison offers the possibility to integrate expert knowledge into the model-based process. However, only using the tolerance approach without aggregation of errors and uncertainties in the application domain is very dangerous, since the tolerances are subjective and might insinuate misleading trustworthiness. The binary microscopic validation decisions can be combined to a macroscopic statement. For example, Viehof [185, Chap. 4.4.6.1] calculates two validation scores based on relative and absolute frequencies of the binary results.

The decision making in the application domain

$$g^a_{dec} : (\hat{y}^a_s, t^a_y) \mapsto \hat{d}^a = \begin{cases} 1 \text{ if } \hat{y}^a_s \le t^a_y \\ 0 \text{ else} \end{cases} \tag{31}$$

compares the assessment results with thresholds specified by standards, regulations or internal management. It is very important to mention that many references use the simulation results y^a_m for decision making without any model validation [149]. However, the estimated system behavior $\hat{y}^a_s$ considering errors and uncertainties should be used instead. Macroscopic application decision making works analogously.

4 Manifestations of Simulation Models

So far, we have presented the framework as generically as possible and provided clear interface descriptions to ensure the desired modularity. In this section we will introduce different manifestations of the simulation model block in the framework.

4.1 (Non-)Deterministic Models

At first we will distinguish deterministic point predictions from non-deterministic ones according to Fig. 2. The term non-deterministic includes probabilistic predictions, interval predictions or a mixture of both.

4.1.1 Deterministic Models

Deterministic simulations are point predictions that calculate a result for a single scenario without uncertainties [48]. The framework can be directly applied to the deterministic manifestation. Generally speaking, in a deterministic simulation

$$g_m : \begin{cases} \mathbb{R}^{N_x} \times \mathbb{R}^{N_\theta} \to \mathbb{R}^{N_y} \\ (x, \theta) \mapsto y_m \end{cases} \tag{32}$$

with multiple inputs, parameters and outputs, the quantities can be real values of a certain dimension $\mathbb{R}^N$. The deterministic inputs x are assumed fixed and precisely known. If physical tests are repeated several times to capture the natural variability, the observed input conditions are usually averaged and the mean value is used for re-simulation. If the model-form is strongly non-linear, the propagation of a mean value by the simulation model can cause significant deviations compared to propagating all values and taking the mean afterwards [133, p. 492].

4.1.2 Interval Predictor Models

Since it is unrealistic to get a deterministic point prediction that exactly matches the system behavior under various conditions, set-valued predictions try to enclose the system behavior. Interval Predictor Models (IPM) [35] perform a set-valued map [35, Eq. (2)]

$$g_m : (x, \Theta) \mapsto I(y_m) \tag{33}$$

from an input point x and a parameter set Θ to a response interval $I(y_m)$. An IPM can be represented as multiple executions of the deterministic model for several parameter samples [35, Eq. (3)]:

$$I(y_m) = \{y_m = g_m(x, \theta) \mid \theta \in \Theta\}. \tag{34}$$

IPMs were introduced in [29] and were extensively developed by Crespo et al. [35]. They represent the parameter set [35, Eq. (5)]

$$\Theta = I(\theta) = [\underline{\theta}, \overline{\theta}] \tag{35}$$

as an interval itself or as a hyper-rectangular set for multiple parameters. They show how IPMs can be constructed if both IPM boundaries can be described by polynomial or radial basis functions. An example can be seen in Fig. 5. In [112] they focus on polynomial function boundaries. Then the lower boundary function [112, Eq. (1)]

$$\underline{g}_m(x) : (x, \underline{\theta}) \mapsto \sum_{i=0}^{d_l} \underline{\theta}_i x^i. \tag{36}$$

is just dependent on the lower parameter coefficients $\underline{\theta}_i$ with degree d_l and the upper boundary function $\overline{g}_m(x)$ analogously on the upper coefficients. Finally, the IPM can be described by the upper and lower interval boundaries [112, Sect. III]:

$$g_m(x) = [\underline{g}_m(x), \overline{g}_m(x)]. \tag{37}$$

In [32] they introduce IPMs for dynamic systems and in [34] they compare IPMs as meta-modeling technique with linear regression, Gaussian processes and Bayesian credible intervals. In [33, 37] they extend IPMs to Random Predictor Models (RPMs) by replacing the interval-valued map with a random-variable-valued map to obtain probabilities within the IPM boundaries. In addition, Sadeghi et al. [158] describe how to propagate mixed uncertainties in the form of p-boxes by IPMs. In [157] they combine IPMs with a Frequentist probabilistic framework.

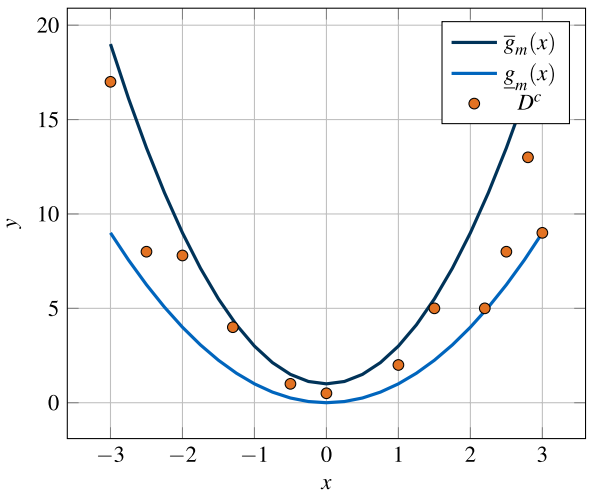


Fig. 5 Exemplary IPM bounding the calibration data. (Color figure online)

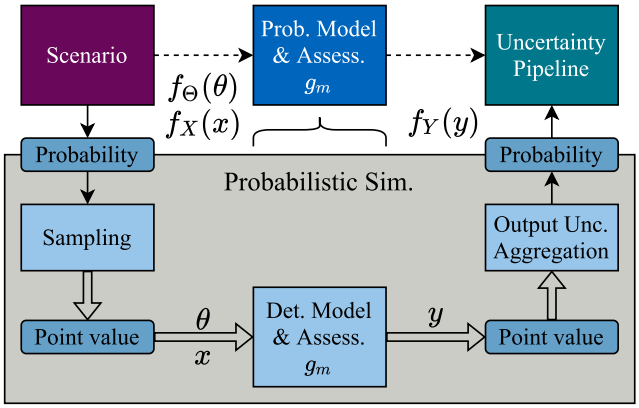


Fig. 6 Probabilistic simulation. The double arrow emphasizes multiple samples. (Color figure online)

4.1.3 Non-deterministic Models

A non-deterministic simulation model

$$g_m : (f_X(x), f_\Theta(\theta)) \mapsto f_Y(y) \tag{38}$$

maps uncertainties instead of points, but the interface quantities of the mapping are preserved compared to deterministic simulations. As shown in Fig. 6, several steps have to be performed inside the framework block of a non-deterministic simulation model to determine the uncertainties due to inputs. In advance, the inputs and model parameters that have uncertainties are identified. The uncertainties are classified as aleatory, epistemic or mixed and quantified. Since the simulation model occurs in all domains, care must be taken to quantify the input uncertainties for the conditions of the different domains. Subsequently, the input uncertainties are propagated by the deterministic simulation model [155]. To solve the forward uncertainty propagation for uncertain parameters and inputs, the integral [160, Eq. (8)]

$$f_Y(y) = \int f_Y(y \mid x, \theta) f_X(x) f_\Theta(\theta) d\theta dx \tag{39}$$

must be solved. Since this is often not possible, sampling methods are used for uncertainty propagation, so that the deterministic simulation model can be executed several times for the individual samples. Finally, all sample results are aggregated to quantify the output uncertainty due to inputs [160].

Frequentists often represent aleatory uncertainty as CDFs, epistemic as intervals and mixed as p-boxes. They treat the epistemic and aleatory uncertainties differently within the sampling block. Roy and Balch [155] propose a nested sampling with epistemic samples in the outer loop and for each of them aleatory samples in the inner loop. The order can also be reversed. The mathematical structure is preserved during propagation, if all input uncertainties are exclusively epistemic or aleatory. If different types of input uncertainties occur, the output uncertainty is always a p-box. Bayesians represent all uncertainties with probability distributions. The posterior distributions represent epistemic uncertainties and can be reduced with further calibration data [128]. Mahadevan [124] describes up to three loops for the distribution type in the outer loop, the distribution parameters in the middle loop and the actual quantity in the inner loop. He proposes to flatten the nested three-loop sampling with an auxiliary variable to get a single-loop sampling for aleatory and epistemic uncertainties. We avoid visualizing a certain number of loops in Fig. 6, since this depends on the selected approach.

Besides sampling, there are more sophisticated uncertainty propagation methods such as Polynomial Chaos Expansion (PCE) [87] or Taylor model approximations [57]. Recent trends in UQ are summarized in [58].

4.2 Hierarchical Models

So far, we have considered the system as a single entity. However, the internal system architecture might have a hierarchy with sub-systems and components. Mahadevan [124] distinguishes four different architectures. Besides a single-component system, a multi-level system is organized in a hierarchical fashion, a time-varying system includes a sequential processing of the components and a multi-physics system has simultaneous interactions between the components. Figure 7 shows an example of a multi-level simulation model.

There are certain applications, where calibration and validation experiments can just be performed on component-level, but the prediction in the application domain takes place on system-level [56]. This requires an extrapolation in the system hierarchy and must be taken into account by the aggregation in Sect. 6.6.

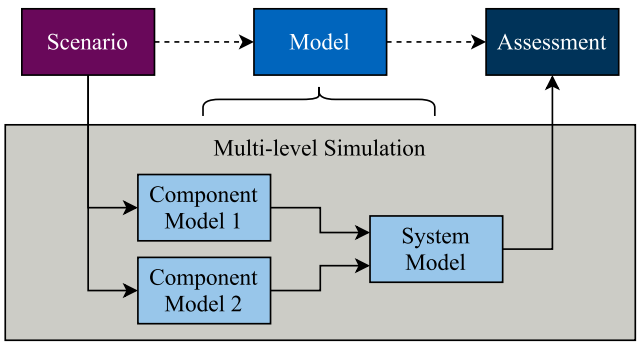


Fig. 7 Multi-level simulation model. (Color figure online)

4.3 Dynamic Models

Most of the VV&UQ literature addresses static systems with time-invariant quantities or stationary conditions [8, 89]. However, there are also dynamic systems with time varying signals. We denote a time signal with a bold symbol

$$\mathbf{y} = [y_1, \dots, y_k, \dots, y_{N_k}] \in \mathbb{R}^{N_k}. \tag{40}$$

In case of multiple outputs, it is a matrix $\mathrm{Y} \in \mathbb{R}^{N_y \times N_k}$ consisting of an output vector per time step. Nevertheless, we adhere to the vector representation to ensure readability without loss of validity. We represent the dynamic system and model as a vector-valued (or matrix-valued) map

$$g_m : (\mathbf{x}, \theta) \mapsto \mathbf{y}_m \tag{41}$$

to keep the interface consistent. Internally, a dynamic system depends on past inputs and system states to calculate the next output. This is usually described as a differential equation or in state-space form [89].

In a model-based testing process, the dynamic quantities are often transformed to a simplified static form according to Fig. 8 to enable the applicability of the classic VV&UQ methods [89]. This is a restriction compared to a generalized dynamic system, but is in line with many application fields. Regarding the input quantities, test scenarios are often parameterized [149] with typical signal characteristics like a sine wave. Regarding the response quantities, key performance indicators (KPIs) are often extracted from the data for clear assessment and decision making in the application domain. Here, a KPI function

$$g_{kpi} : \mathbf{y}_m = [y_{m,1}, \dots, y_{m,k}, \dots, y_{m,N_k}] \mapsto y_m \tag{42}$$

extracts one characteristic value out of the response time signal such as the rise time or an overshoot value [92]. Even if multiple KPIs are calculated per time signal, the result is still very low-dimensional. An alternative approach is to use data reduction techniques to automatically extract low-dimensional features. For example in VV&UQ, principal component analysis (PCA) [4, 193], possibly combined with singular value decomposition (PCA-SVD) [138], ensemble empirical mode decomposition (EEMD) [190], wavelet decomposition [96] or Karhunen–Loève (KL) expansion [189] have already been used.

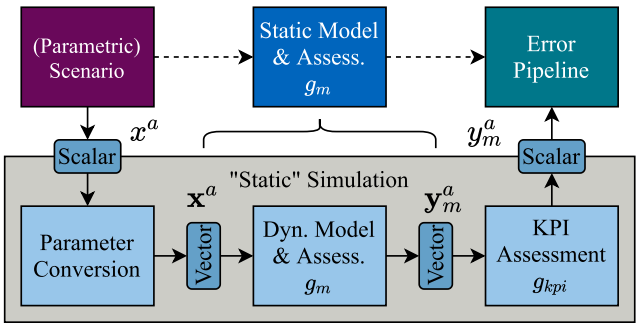


Fig. 8 Transformation of a dynamic simulation to a simplified static representation. (Color figure online)

4.4 Formal Models

Formal methods are used in computer science to obtain a proof of correctness of safety critical systems. The formal models of cyber-physical systems (CPS) are often represented as hybrid automata [151] or differential inclusions composed of a differential equation with set-based uncertainty for non-deterministic dynamic models [6]. They can integrate uncertainty via a so-called Minkowski addition into measurement and disturbance sets, into additional inputs, parameters or initial states [81]. Reachability analysis is a formal technique that can deal with set-based computations to determine the set of states a system can reach from given initial states and possible inputs and parameters [5]. Unlike sampling such as Monte Carlo, it can thus provide formal guarantees. However, similar to classic model validation, the so-called conformance of the formal model to the actual system must still be tested to ensure the transfer of the formal properties. The corresponding conformance testing approach of reachability analysis is to check behavior inclusion by ensuring that all observed data lies within the non-deterministic boundaries [6], similar to IPMs.

5 Manifestations of Scenarios, Metrics, Decision Making and Extrapolation

In the previous section we have extended the simulation model block to different manifestations. This also applies to the real system and the assessment. In this section we extend it to the resulting manifestations of the entire framework. We will address scenarios, metrics, decision making, error learning and inference for different manifestations where appropriate. We again choose the validation domain as representative example. The manifestations of the error aggregation including the error integration block will be addressed separately in the following main section.

5.1 Scenario Design

In the literature there is a variety of techniques to select test scenarios. The survey paper [149] gives an overview about scenario selection methods for autonomous vehicles. It includes conventional techniques such as combinatorial testing or statistical design of experiments (DoE), but also sophisticated techniques such as reinforcement learning or

rapidly exploring random trees that can generate entire trajectories to test dynamic and formal models. Furthermore, there are partitioning techniques to determine an optimal data split [125, 177]. Mullins et al. [127] select calibration and validation scenarios so that costs and prediction uncertainties are minimized. Böde et al. [27] select validation and application scenarios based on a similar concept.

5.2 Validation Metrics

We distinguish different types of validation metrics according to Table 2 and give some examples. The inputs of the validation metric can either be point values for deterministic simulations or probability distributions (, intervals or p-boxes) for non-deterministic simulations. The output of the validation metric can be a Boolean value, a probability or a real value in the unit of the response quantity. We devote the following three sub-sections to these three output types. For dynamic systems, there are different time series validation metrics. Feature-based time series validation metrics convert both input time signals from system and model to single scalar output values. Full time series validation metrics calculate a complete output time signal. We devote the last three sub-sections to time series validation metrics and to closeness notions for formal models.

5.2.1 Boolean Validation Metrics

Boolean results can be achieved by comparison with model accuracy requirements. This can be a simple tolerance comparison [92] or a statistical hypothesis test (HT) [185]. However, Oberkampf and Roy [133, p. 69] advise against integrating the tolerances into the validation metric, and instead advise to keep them separate, as shown in the validation decision making block of our framework.

5.2.2 Bayesian Validation Metrics

A Bayesian HT [146] fits well with the Bayesian approach, since the latter favors uncertainties as unit-less probability measure. Suppose the null hypothesis H_0 means that the model is correct and the alternate hypothesis H_1 that the model is incorrect. Starting with Bayes theorem [146, Eq. (4)]

$$\frac{P(H_0 \mid D^v)}{P(H_1 \mid D^v)} = \frac{P(D^v \mid H_0)P(H_0)}{P(D^v \mid H_1)P(H_1)} = B_F \frac{P(H_0)}{P(H_1)} \tag{43}$$

and a-priori knowledge $P(H_0)$ and $P(H_1)$, the Bayes factor B_F as ratio of likelihoods will be updated based on validation data D^v. Then the Bayes factor can be used to estimate the probability that the model is correct regarding the validation data [146, Sect. 2.2]

$$P(H_0 \mid D^v) = B_F/(B_F + 1). \tag{44}$$

Since the Bayesian HT requires subjective priors, is harder to interpret and since the Bayes factor is relative, Rebba and Mahadevan [146] also introduce an absolute so-called model reliability metric. The latter is the probability that the difference between experiment and simulation mean is smaller than a tolerance interval. They describe how to determine the reliability metric under aleatory uncertainty. Sankararaman and Mahadevan [159] extend the reliability metric to take aleatory and epistemic uncertainties into account by using the entire distributions.

5.2.3 Frequentist Validation Metrics

Frequentists favor errors and uncertainties in the unit of the response quantity. In the simplest deterministic case, an absolute deviation

$$e^v = y_m^v - y_s^v \tag{45}$$

might already be sufficient. Balci [20] summarize many subjective validation techniques such as face validation as well as objective validation metrics. In the probabilistic case, a wide variety of distributional metrics are available in mathematics that quantify the distance between two probability distributions. Gardner et al. [70] compare metrics and divergences such as the Kullback–Leibler divergence, the Hellinger distance or the Kolmogorov distance. Bi et al. [26] compare the Euclidian, Mahalanobis and Bhattacharyya

Table 2 Taxonomy of validation metrics including examples with input types in columns and outputs in rows

	Inputs	Deterministic		Distributional	
Outputs		Static	Dynamic	Static	Dynamic
Boolean	Static	Tolerance check [92]	Tolerance band [93]	Hypothesis Test (HT) [146]	HT with KPIs [185]
	Dynamic	–	–	–	–
Probabilistic	Static	–	–	Bayesian HT [146]	Bayesian HT with wavelets [96]
	Dynamic	–	–	–	Dynamic reliability metric [8]
Real-valued	Static	Difference	Vector metric [162]	Area metric [133]	Area metric with PCA [193]
	Dynamic	–	Difference vector	–	–

distance and analyze their impact on the calibration and validation results. Oberkampf and Roy [133, Chap. 12.8.2.2] favor the non-deterministic area validation metric (AVM) [133, Eq. (12.52)]

$$\overline{e}^v = -\underline{e}^v = \int_{-\infty}^{\infty} |B(y_m^v) - F(y_s^v)| dy \tag{46}$$

that quantifies the area between the observed CDF and the simulation p-box by a so-called Minkowski L_1 metric. AVM is a non-deterministic metric that describes the epistemic model-form uncertainty $I(e^v) = [\underline{e}^v, \overline{e}^v]$ symmetrically with $\overline{e}^v = -\underline{e}^v$. Voyles and Roy [186] propose the modified area validation metric (MAVM) as an asymmetric extension by distinguishing the area on the left and right side of the p-box and incorporate the number of experimental repetitions. Tanaka [175] also propose an asymmetric area validation metric (TAVM) by normalizing both distributions first and add tolerances for decision making. Wang et al. [188] define an interval area metric (IAM) that calculates a best- and worst-case model-form uncertainty and thus two possible intervals. An overview about validation metrics and a comparison of the classical HT, the Bayesian HT and the area metric among others is given in [119, 123, 126].

5.2.4 Feature-Based Time Series Validation Metrics

There is a wide variety of metrics available in mathematics to quantify the distance between two vectors. An overview about time series metrics can be found in [103, 162]. It starts with basics such as vector norms, average errors or correlations coefficients. Some metrics such as the Sprague and Geers metric separately quantify the phase information from the magnitude and combine both afterwards to avoid high metric values due to a shift of the correct signal-form. Other metrics typically based on dynamic time warping (DTW) address this issue in a direct calculation. Frequency metrics transform the time signal into the frequency domain before quantifying the distance [96].

To enable probabilistic techniques, KPIs are usually extracted from time signals or reduction techniques are applied as presented in Sect. 4.3. For example, Viehof [185] extracts characteristic values from vehicle dynamic signals or applies window functions to enable the application of hypothesis tests. Wang et al. [189] use a Karhunen–Loève (KL) expansion to apply the area metric to dynamic systems. Jiang and Mahadevan [96] perform a wavelet decomposition followed by Bayesian hypothesis testing.

5.2.5 Full Time Series Validation Metrics

Ao et al. [8] highlight disadvantages of the feature-based metrics such as the loss of time information in general and different principal components for simulation and experiment with possibly high numbers in PCA techniques. Instead, they extend the reliability metric to quantify the discrepancy over time. They propose and analyze three time domain metrics: instantaneous, first-passage and accumulated reliability. The first one performs the evaluation at each time step, the second over a time duration and the third aggregates the evidence over time.

5.2.6 Formal Closeness Notions

There are formal methods that try to falsify a deterministic formal model by means of optimization [1]. They use closeness notions similar to time series validation metrics. For example, a (τ, ϵ)-closeness between two time series is used in [2, 10, 23] and a Skorokhod metric in [41].

5.3 Validation Decision Making

After quantifying deviations between simulation and experiment with validation metrics, these can be compared with tolerance limits based on expert knowledge. As mentioned, there is an overlap with binary validation metrics such as the hypothesis test that include the tolerances.

5.3.1 Time-Series Validation Decision Making

The tolerance comparison can be easily extended from scalar values to time signals. For example, the vehicle dynamics standard [93] lays a tolerance band around the deterministic time signal of the simulation and checks whether all experimental repetitions fall within the band to obtain a binary result for model validity.

5.3.2 Formal Validation Decision Making

In robust control of dynamic systems, it is often favorable to have formal guarantees (certificates) about the validity of the model. Prajna [142] argues that absolute model validity can never be proven with finite resources, but only model invalidity. He uses barrier certificates, similar to Lyapunov functions in stability analysis, to separate possible model trajectories from experimental data. If a barrier can be found, the model's parameter set and the data are proven inconsistent. Harirchi et al. [80] use model invalidation for guaranteed model-based fault detection.

Many other references in the control community relax the hard invalidation for a wide variety of different system representations and consider probabilistic certificates for the model (in)validity instead [136, 171]. For example, Halder and Bhattacharya [75] address UQ of dynamic systems, compare the simulation and experimental PDF with a Wasserstein metric and calculate a probabilistically robust

validation certificate (PRVC) in each time step based on a required tolerance level. Karydis et al. [102] calculate the probability of violating a confidence region as tolerance value and use it with randomization techniques to expand stochastic models with uncertain parameters that capture the experiments.

There are different notions of conformance to check behavioral inclusion of the experiments within the non-deterministic simulation bounds determined by reachability analysis. Trace conformance [165, 173] is a strong relation that ensures that all state transitions are preserved. Reachset conformance [122, 151] is a weaker relation for safety that just checks the set of traces instead of each individual one. An overview about conformance notions can be found in [105].

5.4 Extrapolation in Scenario Domain

A validation metric quantifies a deviation for a single validation scenario. Typically, multiple validation scenarios are executed across the entire validation domain. They differ from the application scenarios and cause interpolation and extrapolation uncertainties as visualized initially in Fig. 3. There are different techniques how to quantify the model-form error or uncertainty across the entire domain. Mullins et al. [126] distinguish ensemble and point-by-point validation. Ensemble validation aggregates the data of all scenarios before applying the validation metric. In contrast, point-by-point validation applies the validation metric to each scenario separately. They argue that the former is more suitable for uncharacterized experiments, whereas the latter is preferred for partially and fully characterized experiments. Point-by-point validation makes it possible to use interpolation or extrapolation techniques to transfer errors to untested application scenarios or to learn an error model across the entire application domain. Alternatively, all metric results of the point-by-point validations can be aggregated across the application domain to one macroscopic validation metric after applying the individual metrics. The ensemble validation can also be combined with point-by-point validation, if subsets of the data are aggregated to new scenarios and those are used for point-by-point comparisons [133, p. 656].

Strictly speaking, only the point-by-point validation with regression technique fits perfectly into the error learning and inference blocks of our framework. The ensemble validation is placed before the validation metric by already aggregating a tuple of scenarios (double arrow instead of single one) before applying the actual metric. The macroscopic validation metric can be interpreted as a degenerate case of the error learning, since it is not a function of the input scenarios anymore.

5.4.1 Ensemble Validation

If validation data is sparse, u-pooling [133, Chap. 12.8.3] is a technique that performs a transformation from the physical to a probability space. Suppose it is impossible to repeat experiments because the environment cannot be fully controlled in applications such as the safeguarding of automated vehicles with dynamic traffic objects. Then, u-pooling makes it possible to aggregate all experimental data from different scenarios to one point by switching to the probability space where the physical units do not matter anymore. This yields again a CDF for experiment and simulation so that the area validation metric can be used. However, the metric comparison is performed in the probability space. Therefore, a back-transformation is required afterwards to obtain the original physical units. He [82] extend u-pooling to p-boxes and Xi et al. [193] to dynamic responses by applying PCA in advance.

5.4.2 Error Learning and Inference

Oberkampf and Roy [133, p. 657] learn an error model in the validation domain via polynomial regression. They additionally calculate external prediction intervals to also consider the uncertainty of the regression model itself. A prediction interval is greater than a classic confidence interval [133, p. 657]. Kennedy and O'Hagan [104] learn an error term in the calibration domain with a Gaussian process. Farajpour and Atamturktur [59], Shinn et al. [167] use polynomial meta-models, Rutherford [156] a bilinear regression technique and Li et al. [116] a spatial random process. Crespo et al. [36] apply IPMs to learn error models with bounds instead of directly modeling the system behavior. Similar to IPMs, Feeley et al. [60] also aim to identify validated parameter sub-spaces. Romero [152] apply predictor-corrector techniques for extrapolation. Atamturktur et al. [12] extend the KOH equation by an additional application-specific extrapolation term. Hemez et al. [83] refers to this as predictive maturity index (PTI) and argues that it should be based on goodness-of-fit, the amount of calibration, the domain of applicability, etc. Therefore, Atamturktur et al. [14] define a metric that quantifies coverage of scenarios in the validation domain.

5.4.3 Meta-Model Approach

Hills [84], Hamilton and Hills [77, 78] propose a meta-model approach to connect the validation domain with the application domain. It differs from the error model in our framework, which models the errors between simulation and experiment as a function of the input conditions in the validation domain. Instead, they model the response behavior of the simulation model in the application domain as a function of its response

behavior in the validation domain. Both the error and meta-model will be used for aggregation of errors and uncertainties in model prediction. They define their simulation models [84, Eq. (5, 6)]

$$\Delta y_m^v = g_m^v(x_{nom}^v + \Delta x, \theta_{nom}^v + \Delta\theta) - \langle y_m^v \rangle \tag{47}$$

$$\Delta y_m^a = g_m^a(x_{nom}^a + \Delta x, \theta_{nom}^a + \Delta\theta) - \langle y_m^a \rangle \tag{48}$$

relative to defined nominal conditions. With a sensitivity analysis they determine the important inputs and parameters around the nominal conditions to reduce the data-driven model order. They assume having a constellation with a small application domain as a neighborhood around a nominal application condition and N experimental neighborhoods around N nominal validation conditions. Generating N_D samples for Δx and $\Delta\theta$ within each neighborhood yields a vector $\Delta\mathbf{y}_m^a \in \mathbb{R}^{N_D}$ for the application neighborhood and a matrix $\Delta\mathbf{Y}_m^v \in \mathbb{R}^{N_D \times N}$ for all validation neighborhoods. $\langle y_m^v \rangle$ and $\langle y_m^a \rangle$ are mean values over the N_D samples. They represent the meta-model as a linear combination [84, Eq. (7)]

$$\Delta\mathbf{y}_m^a \cong \Delta\mathbf{Y}_m^v \mathbf{w} \tag{49}$$

of the relative behaviors with the weighting vector $\mathbf{w} \in \mathbb{R}^N$. The learning of the meta-model can either be performed with a partial least square approach [182] or an objective function method [84, Eq. (9)]

$$\begin{aligned} L = {} & a\mathbf{w}^{\mathrm{T}}\mathrm{Var}(\mathbf{y}_m^v - \mathbf{y}_s^v)\mathbf{w} \\ & + (1-a)(\Delta\mathbf{y}_m^a - \Delta\mathbf{Y}_m^v\mathbf{w})^{\mathrm{T}}(\Delta\mathbf{y}_m^a - \Delta\mathbf{Y}_m^v\mathbf{w}) \end{aligned} \tag{50}$$

with a factor $0 \leq a \leq 1$. It reflects the trade-off between accuracy and robustness to variance in model parameters and measurement uncertainty.

5.4.4 Macroscopic Validation Metrics

Mullins et al. [126] aggregate the results e^v of point-by-point comparisons using the reliability metric into an overall macroscopic validation metric [126, Eq. (7)]

$$e_{mac} = \int e^v f(x^a) dx \tag{51}$$

by considering the joint probability density across the application domain. Eek et al. [56] determine deterministic errors for several experiments and calculate an error histogram $f(e^v)$. From this, they extract one uncertainty bound $I(e^v) = [\underline{e}^v, \overline{e}^v]$ for the entire scenario domain.

6 Manifestations of Error and Uncertainty Aggregation

It is one of the main current challenges in VV&UQ, how to aggregate errors and uncertainties for model prediction [128, 154, 164]. We present general aggregation techniques first. In the next three sub-sections we focus on the aggregation from the verification, calibration and validation domain to the application domain. We dedicate separate sub-sections to the aggregation in the scenario and parameter domain, to the system hierarchy domain from component- to system-level and to the time domain for dynamic systems.

6.1 Aggregation Techniques

Mullins et al. [128] distinguish three "classes of roll-up approaches:

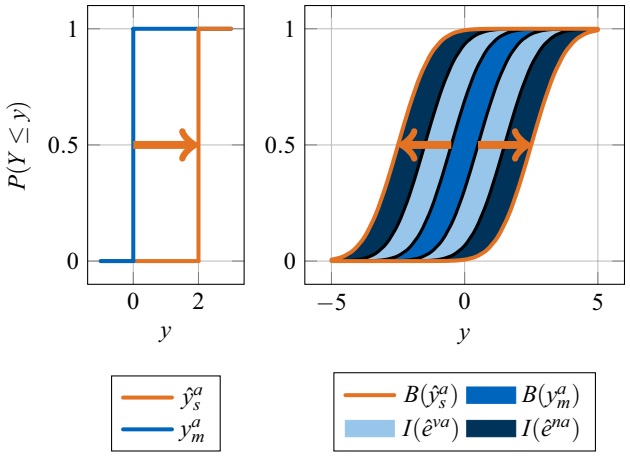
1. Apply bias correction to the model predictions based on errors observed in the validation assessment
2. Modify parameter uncertainty to add conservatism to prediction
3. Apply domain specific model form corrections to governing equations".

We extend this classification of aggregation (roll-up) approaches significantly through a hierarchical structure according to Table 3. We distinguish the two aggregation techniques bias correction and uncertainty expansion. A visual example of both aggregation techniques can be found in Fig. 9. The uncertainty expansion example refers to Sect. 6.4.5. If error estimates are available, they can be used for bias correction to bring the model responses $\hat{y}_s^a$ closer to the true value y_s^a by correcting the inferred error:

$$\hat{y}_s^a = y_m^a - \hat{e}^{va}. \tag{52}$$

Table 3 Aggregation categories

Aggregation techniques	Bias correction
	Uncertainty expansion
Aggregation stages	Internal model parameters
	Internal model-form
	Model responses
Aggregation source domain	Verification Domain
	Calibration Domain
	Validation Domain
Aggregation target domain	Application Domain
	All after the source domain



(a) Bias correction (b) Uncertainty expansion

Fig. 9 Aggregation techniques acc. to (52) and (65). (Color figure online)

Otherwise, uncertainties should be aggregated and integrated by expanding the model response with the uncertainty so that the true value lies within it:

$$y_s^a \in I(\hat{y}_s^a) = [\underline{y}_s^a, \overline{y}_s^a] = \{\hat{y}_s^a \mid \underline{y}_s^a \leq \hat{y}_s^a \leq \overline{y}_s^a\}. \tag{53}$$

Both techniques can be applied at multiple stages: either directly to the model responses as described so far or internally to the model parameters or model-form. Generally, numerical uncertainties can be aggregated from the verification domain, parametric uncertainties from the calibration domain or model-form uncertainties from the validation domain. For all these source domains, the target is always the application domain, either directly or internally via the domains in between. The direct aggregation uses the vertical error (uncertainty) pipeline in the framework consisting of metrics, error learning, inference and integration. The integration includes the actual implementation of the aggregation technique within the application domain. The internal aggregation uses the inverse (orange) arrows in the framework to apply the technique before reaching the application domain.

6.2 Aggregation from Verification Domain

The aggregation from the verification to the application domain aims to account for the numerical error or uncertainty in the model prediction.

6.2.1 Bias Correction of Model-Form

Sankararaman and Mahadevan [160] correct the bias of deterministic numerical errors such as discretization errors. They add an additional term to the original model-form so that the corrected model can be already applied in the subsequent calibration, validation and application domain. A correction of the model-form

$$g_m^a(x,\theta) = g_m^v(x,\theta) = g_m^c(x,\theta) = g_m^n(x,\theta) - \hat{e}^n \tag{54}$$

with an additional term can be interpreted as ensuring corrected model responses. Whereas Sankararaman and Mahadevan [160] correct deterministic errors, they sample from stochastic errors such as surrogate model uncertainty to add conservatism. Surrogate models are often used instead of complex high-fidelity simulation models to speed up computation.

6.2.2 PBA-Based Uncertainty Expansion of Model Responses

Oberkampf and Roy [133, Chap. 8.5] quantify the deterministic discretization error and convert it to a numerical uncertainty according to (11). They use a safety factor in the conversion to get one conservative value, instead of learning a model of the numerical uncertainty in the scenario domain. In Probability Bound Analysis (PBA) they use the numerical uncertainty in the application domain to shift the left and right boundary of the simulation p-box outwards [155, Fig. 12]:

$$B(\hat{y}_s^a) = \{F(\hat{y}_s^a) \mid \underline{F}(y_m^a - \underline{e}^{na}) \leq F(\hat{y}_s^a) \leq \overline{F}(y_m^a - \overline{e}^{na})\}. \tag{55}$$

However, they are not considering the numerical uncertainty in the validation domain so that it influences the determination of model-form uncertainty. The uncertainty expansion can be seen together with model-form uncertainty to obtain the whole p-box $B(\hat{y}_s^a)$ in Fig. 9 and generally in Sect. 6.4.5.

6.3 Aggregation from Calibration Domain

In classic model calibration, point values are estimated for selected model parameters. Especially with Bayesian approaches, it is also possible to integrate uncertainty into the model parameters to add conservatism.

6.3.1 Point Estimation of Model Parameters

In the simplest case, the parameter determination is understood as point estimation. Then a least square estimation [124] can be used to determine the parameters so that the errors are minimized

$$\arg\min_{\theta} \sum_{j=1}^{N_{D,d}^c} \sum_{i=1}^{N_{D,r}^c} (g_m(x_j,\theta) - g_s(x_{ij}))^2 \tag{56}$$

over the number $N^c_{D,d}$ of all distinct calibration scenarios and its replicates $N^c_{D,r}$. The errors are described here as squared deviation between simulation and experiment (calibration metric). Assuming that the errors follow an unbiased normal distribution $N(0, \sigma)$, a likelihood function [160, Eq. (6)]

$$L(\theta) \propto \sum_{j=1}^{N^c_{D,d}} \sum_{i=1}^{N^c_{D,r}} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(g_m(x_j, \theta) - g_s(x_{ij}))^2}{2\sigma^2} \right) \quad (57)$$

can also be constructed. In Maximum Likelihood Estimation (MLE), the parameters are determined so that the likelihood function is maximized

$$\arg\max_{\theta} L(\theta). \quad (58)$$

These methods do not consider parameter uncertainties. In some cases, however, separate confidence intervals can be calculated for the parameters [124].

6.3.2 Bayesian Uncertainty Expansion of Model Parameters

The Bayesian approach describes the model parameters as probability distributions to reflect uncertainties and to add conservatism. Bayes' Theorem can be used to calculate the posterior PDF [160, Eq. (5)]

$$f_\Theta(\theta|D^c) = \frac{L(\theta) f_\Theta(\theta)}{\int L(\theta) f_\Theta(\theta) d\theta} \quad (59)$$

based on the prior PDF $f_\Theta(\theta)$ and the likelihood function $L(\theta) = P(D^c|\theta)$ according to (57). We abbreviate the data pair $D^c = D(x^c_s, y^c_s)$ for readability, since it will occur several times. Since it is often hardly possible to solve the integral directly, sampling methods like Markov Chain Monte Carlo are used to determine the posterior PDF and Bayesian credible intervals for conversion to intervals [160].

6.3.3 Bayesian Bias Correction of Model Responses

For calibration, the error could previously be understood as a observation error e_{obs} with unbiased normal distribution. If the predictions are biased, Kennedy and O'Hagan [104] introduce a separate error model term $g^{ca}_{inf}(x)$ (often called model discrepancy term or model inadequacy function $\delta(x)$) to account for it [104, Eq. (5)]:

$$y^c_m = y^c_{s,i} + e^c_{obs,i} + g^{ca}_{inf}(x). \quad (60)$$

In the KOH framework, named after both authors, they learn the error model $g^{ca}_{inf}(x)$ with a Gaussian process dependent on the input conditions. They simultaneously calibrate the model parameters and the hyper-parameters of the Gaussian process and estimate the remaining model error. It can be used afterwards for a bias correction of the model responses in the application domain. However, this is risky, since it means to trust the data-driven regression model more than the original physics-based model [84, p. 15]. The error term enables no natural bias correction, since it reflects a model-form error in the calibration domain [128]. Instead, the error term is often omitted in calibration and instead the model-form error is determined in the validation domain with a separate validation data set using validation metrics [160]. This enables an easier aggregation to the application domain. Furthermore, the KOH framework often leads to an identifiability problem between the model parameters and the discrepancy term [11]. A six-step procedure is wrapped around the KOH framework in [21]. The selection of model discrepancy priors for Bayesian updating is described in [120, 121].

6.3.4 IPM-Based Uncertainty Expansion of Model Parameters

Crespo et al. [35] use data-driven IPMs to incorporate the uncertainties into intervals so that the IPM boundaries cover all system behaviors and no error

$$e^c = g_m(x, \theta) - y_s = 0 \quad \exists \theta \in \Theta \quad (61)$$

is left in the calibration domain by definition. Therefore, the IPM approach skips the entire verification and validation domain of the framework. After the internal integration of uncertainties into the IPM parameters using inverse calibration methods (orange arrow in the framework), they directly use the IPM for prediction in the application domain without any further bias correction or uncertainty expansion.

6.4 Aggregation from Validation Domain

The aggregation from the validation to the application domain seeks to correct model-form errors or to expand responses with model-form uncertainty.

6.4.1 Bias Correction of Model-Form

The VV&UQ approaches rely on the correctness of physics in the equations. Hills [84] distinguishes four cases regarding the implementation of physics. Mullins et al. [128] propose an additional approach that seeks to address the source of errors in the model-form rather than integrating uncertainty in model parameters or responses. The correction of the model-form depends strongly on the application and requires knowledge of the underlying physics.

6.4.2 Meta-Model-Based Bias Correction of Model Responses

Hills [84] uses the meta-model from Sect. 5.4.3 to perform a bias correction in the application domain [84, Eq. (32)]:

$$\hat{y}_s^a = \langle y_m^a \rangle + \mathbf{w}^{\mathrm{T}}(\mathbf{y}_s^v - \langle \mathbf{y}_m^v \rangle) \tag{62}$$

He uses the meta-model to aggregate both the errors between simulation and experiment and the uncertainties. He considers parametric and observation uncertainties via sampling and bootstrapping.

6.4.3 Bayesian Uncertainty Expansion of Model Responses

In the calibration domain in Sect. 6.3.2, Sankararaman and Mahadevan [160] integrated uncertainty into the model parameters by Bayesian model updating. They combine it with a Bayesian hypothesis test in the validation domain to incorporate model-form uncertainty. They use a Bayesian network to aggregate all uncertainties. According to the theorem of total probability, the uncertainties from calibration and validation (and verification through early correction) are combined to [160, Eq. (12)]

$$\begin{aligned} f(\hat{y}_s^a \mid D^c, D^v) = {} & P(H_0 \mid D^v) f(y_m^a \mid H_0, D^c) \\ & + P(H_1) f(y_m^a \mid H_1). \end{aligned} \tag{63}$$

6.4.4 Bayesian Uncertainty Expansion of Model Parameters

Similar to the last sub-section, Mullins et al. [127] also perform an uncertainty expansion with the Bayesian approach, but incorporate the uncertainties into the parameters instead of the responses to obtain an updated posterior distribution [127, Eq. (12)]

$$f(\theta \mid D^c, D^v) = e_{mac} f(\theta \mid D^c) + (1 - e_{mac}) f(\theta). \tag{64}$$

They use the macroscopic metric based on reliability according to (51) as a weighting factor of the prior distribution. To illustrate this approach in our framework, an additional arrow from the (degenerated) error model back to parameters would be necessary. We did not explicitly draw this due to readability reasons.

6.4.5 PBA-Based Uncertainty Expansion of Model Responses

Oberkampf and Roy [133, p. 657] use polynomial regression to learn a model for the epistemic model-form uncertainty $I(\hat{e}^{va}) = [\underline{e}^{va}, \overline{e}^{va}]$ and include the uncertainty of the regression itself via a prediction interval. They perform an uncertainty expansion with imprecise probabilities to aggregate all uncertainties within PBA. They obtain a final p-box [155, Fig. 12]

$$\begin{aligned} B(\hat{y}_s^a) = \{ & F(\hat{y}_s^a) \mid \underline{F}(y_m^a - (\underline{e}^{va} + \underline{e}^{na})) \\ & \leq F(\hat{y}_s^a) \leq \overline{F}(y_m^a - (\overline{e}^{va} + \overline{e}^{na})) \} \end{aligned} \tag{65}$$

by expanding the left side of the input uncertainty p-box to the left with the estimated model-form uncertainty and the numerical uncertainty and by expanding the right side, respectively. The uncertainty expansion is shown in Fig. 9.

For application decision making based on the final p-box of the response QoI, it is possible to extract an interval of cumulative probabilities at the ordinate for a given response value or to extract a response interval at the abscissa for a given cumulative probability [30, Fig. 2] or even for a given interval of cumulative probabilities [39]. Suppose a probabilistic regularity exists, such as the probability of falling below a threshold must be less than one percent: $P(\hat{y}_s^a < t_y^a) < 0.01$. Then the point of intersection with the upper CDF boundary $\overline{F}(\hat{y}_s^a)$ can be read at the given threshold value. This represents the cumulative probability and

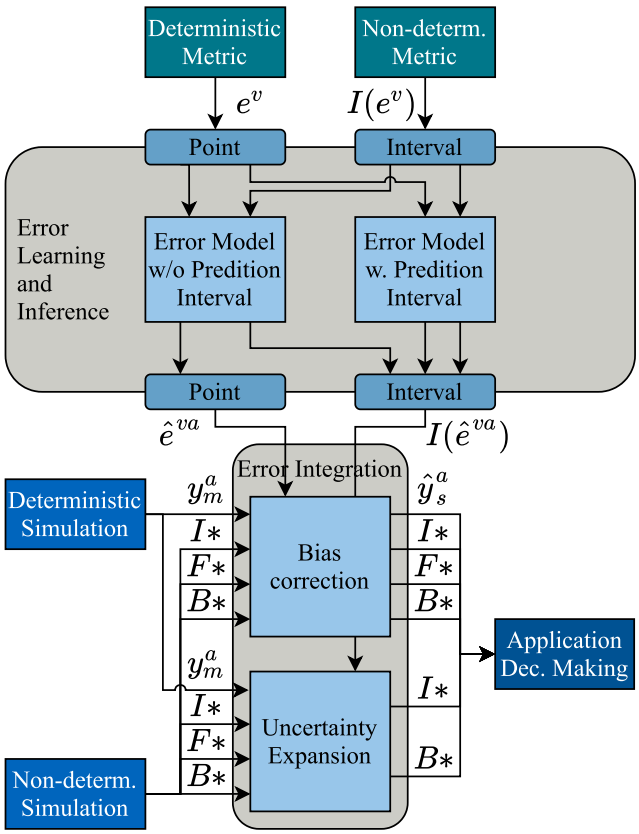


Fig. 10 Generalization of the error pipeline. The deterministic and non-deterministic simulation as well as metric are alternatives and just visualized in parallel for comparison. The stars are placeholders for y_m^a and $\hat{y}_s^a$, respectively. (Color figure online)

can be compared directly against the one percent. It works in reverse with exceedance probabilities [30, 64].

6.4.6 Generalization of the Error Pipeline

In general, there could be further manifestations of the error pipeline of our framework as shown in Fig. 10. It is possible to learn the deterministic error e^v or the epistemic model-form uncertainty $I(e^v)$, which in turn yields a point $\hat{e}^{va}$ or an interval estimate $I(\hat{e}^{va})$. Preferably, the uncertainty of the regression itself should also be considered, e.g. by means of a prediction interval. As a result, in the non-deterministic case the interval estimate becomes wider both on the positive and negative side or a conversion to an interval estimate is made around the error in the deterministic case. Usually the point estimate of the error is combined with a point estimate from the deterministic simulation by bias correction and the interval estimate of the model-form uncertainty is combined with the non-deterministic simulation by uncertainty expansion. However, bias correction can also be applied to non-deterministic simulations by shifting them, and uncertainty expansion can also be applied to point values by expanding them to intervals. However, according to (6), care must be taken that no relevant input and model-form uncertainties are neglected if they are not quantified separately. The bias correction always preserves the mathematical structure, while the uncertainty expansion yields an interval or p-box. Strictly speaking, expanding a deterministic simulation with an interval estimate resulting from a deterministic metric with a prediction interval is actually a combination of uncertainty expansion with bias correction, since the interval is centered around the error:

$$I(\hat{y}_s^a) = \{\hat{y}_s^a \mid y_m^a - \overline{e}^{va} \leq \hat{y}_s^a \leq y_m^a - \underline{e}^{va}\}. \tag{66}$$

An error estimate might also be converted to an uncertainty: symmetrically as with the area metric, including a safety factor as in (11) or one-sided from the nominal simulation to the corrected value:

$$I(\hat{y}_s^a) = \{\hat{y}_s^a \mid y_m^a - \overline{e}^{va} \leq \hat{y}_s^a \leq y_m^a\}. \tag{67}$$

6.5 Aggregation in Scenario and Parameter Domain

We devoted the last three sections to the aggregation from the verification, calibration and validation domain. Selected approaches within these sections used techniques from Sect. 5.4 to take into account the extrapolation of errors and uncertainties in the scenario and parameter domain. Oberkampf and Roy [133] used polynomial error learning and inference combined with external prediction intervals in (65) to obtain the final p-box of PBA. Hills [84] considered uncertainties of the partial least squares regression within the meta-model approach. Mullins et al. [127] incorporated uncertainty into model parameters in (64) considering the scenario domain in form of a macroscopic validation metric. However, many approaches do not address extrapolation uncertainties.

Oberkampf and Roy [133] treat model parameters and scenario inputs equally in PBA of CFD simulations. In such applications all equations can then be applied to parameters in the same way as described for the scenario inputs. However, in simulations of complex dynamic systems usually the scalar parameters and the discrete system configurations are distinguished from the continuous scenario domain. In the automotive sector, for example, it is important to validate different configurations of a vehicle series without conducting separate experiments for each of them [185]. Therefore we separated scenario inputs x and parameters θ in our framework, but these can be merged if desired.

Denham et al. [40] explicitly address the case, where experimental data is only available for a nominal configuration, but decision making focuses on a modified configuration. They consider an correction term from calibration and uncertainty bounds for both models. On the one hand, they describe a method to propagate the uncertainties from the nominal to the modified configuration by assuming that no additional calibration and update of the uncertainty bounds is required. On the other hand, they describe a method by interpreting the correction as an additional uncertainty.

6.6 Aggregation in System Hierarchy Domain

Roy [154] state that the extrapolation in the system hierarchy domain is even harder than in the scenario domain. Schroeder and Mullins [164] refer to the previous sub-section as interpolation and extrapolation type 1 in the input space. They refer to the current section as extrapolation type 2 in the output space. Simultaneous occurrence of both types is also possible. We give an overview about approaches that address this challenge.

6.6.1 Bayesian Network Approach

Sankararaman and Mahadevan [160] extend the Bayesian network to hierarchical systems. They focus on multi-level systems on the one hand and on reduced-order systems to replace a complex system for certain tests on the other. They state that they can convert multi-physics systems with simultaneous connections to the former type. They use the component outputs as linking variables for the multi-level type and the common parameters as linking variables for the reduced-order type. The Bayesian updating, Bayesian

hypothesis testing and Bayesian network integration work similar as shown previously. Li and Mahadevan [115] weight the relevance of the heterogeneous component data with a sensitivity analysis for aggregation on system-level. Neal et al. [130] formulates an optimization to select the most valuable calibration and validation data on component-level that yields the highest confidence on system-level.

Kwag et al. [111] define an overlapping coefficient (OC) between the simulation and experiment PDF, extrapolate it with a Bayesian network from component- to system-level and compare the system-level OC with a tolerance value.

6.6.2 Bayesian Re-Calibration Approach

Babuška et al. [17] present a re-calibration approach, that addresses the case where experimental quantities deviate from unobserved QoIs in prediction. In the first step, they infer unknown model parameters based on a calibration dataset of observed quantities. In the second step, they update the model parameters with Bayesian updating based on an independent dataset of observed quantities. In the third step, they use both the calibrated model and the updated model to predict the unobserved QoIs and check whether they lie within a rejection criterion. For the latter, they calculate the horizontal distance between both models' CDF and compare it with a tolerance value. In addition, they use bootstrapping or cross-validation to tackle insufficient data. Other approaches build on this re-calibration approach [137]. Oliver et al. [135] focus on physics-based composite models of high-fidelity models combined with low-fidelity embedded models. They calibrate the model parameters based on Bayesian updating and perform a validation decision making based on credible intervals. For prediction of unobserved QoIs, they build inadequacy models at the source of errors in the embedded models and perform a sensitivity analysis to analyze influences.

6.6.3 Meta-Model Approach

The meta-model approach [84] from Sect. 5.4.3 can be used not just in the scenario, but also the system hierarchy domain. The simulation models in (47) and (48) and their response quantities can vary between multiple components and the system.

6.6.4 Output Uncertainty Approach

Eek et al. [53, 54] are interested in the conceptual design of complex aircraft systems in an early development stage, where component-level data is available, but hardly any system data. At first, they verify, calibrate and validate the component models. In validation, they perform several experimental repetitions, re-simulate each repetition and calculate an error histogram for each component. They use intervals as uncertainty bounds on the error distribution. This component output uncertainty includes input and model-form uncertainties, but no strict separation as in PBA as stated by Eek et al. [56]. They assume constant input uncertainties across the input space [52]. Then, Eek et al. [53] use this component output uncertainty as input uncertainty on system-level and propagate it by the entire model. They solve this as an optimization problem with both interval boundaries assuming there is an approximately linear relationship. If one input is time-variant, they perform four additional simulations instead of two for both boundaries. They also perform a model verification on system-level. However, Eek et al. [56] state that it is not a complete VV&UQ framework. Model validation with system-level data is necessary to quantify the entire model-form uncertainty.

6.6.5 Multi-physics Coupling

Avramova and Ivanov [16] give an early overview about multi-physics coupling. van Buren et al. [184] focus on UQ to inform a covariance matrix from component-level to the coupled system-level. Stevens et al. [169] assume separate-effect experiments on component-level and integral-effect experiments on system level and use the component data to perform a bias correction of the system results. Stevens [170] additionally offers an inverse method to infer missing component knowledge empirically from system data. Stevens and Atamturktur [168] review recent literature from the last decade regarding strongly coupled models.

6.7 Aggregation in Time Domain

There are mainly two kinds of approaches that quantify the model error of dynamic models [89, 90]. The first type seeks to transform the dynamic responses to static ones as shown in Sect. 4.3. These approaches can build on all static VV&UQ methods presented so far, but are not applicable to generalized dynamic systems. The second type addresses this and predicts the model error in each time step. Since the first type uses the methods presented so far, we focus in this section on the second type.

6.7.1 Bayesian Bias Correction of Model States

Hu et al. [89, 90] focus on error quantification of discrete-time state-space models. They intend to create a single error model that performs a bias correction of the hidden state variables in each time step. The challenge is that the static methods, which compare the responses to construct the error model, do not work, since the error depends on all previous input values. Furthermore, the response error cannot infer the error of the hidden state variables. They show two

solution approaches. An estimation-modeling method first estimates the hidden states inversely from observations and then learns a non-parametric error model of the state variables. A modeling-estimation method first constructs a parametric error model for the errors of state variables and then estimates the error model parameters from observations. For both approaches they formulate the state-space model as a dynamic Bayesian network (DBN). They conclude that the second approach is preferred because of better accuracy, if appropriate parametric basis functions can be selected.

Subramanian and Mahadevan [174] combine a probabilistic Bayesian state estimation to update the states and estimate the error in known scenarios with a deterministic neural network for the transfer to untested scenarios in the application domain. In the first step, they estimate the error from validation data in general by particle filtering-based state estimators. If the model parameters are uncertain, a combined state and parameter estimation would be necessary. In the second step, they identify internal model-form errors in component models. In the third step, they learn a neural network in nonlinear auto-regressive exogenous (NARX) configuration as predictive error model. Either they learn an ensemble of networks that relates the updated Bayesian system state signals with the input signals or they learn one network that relates the observed system response with the original system state signal. The neural network and the simulation model generate data in untested scenarios so that the Bayesian state estimation can update the response. The trick is that they estimate the model-form errors in the equations rather than the response errors. Green [73] proposes to use Bayesian tracking algorithms and Gaussian processes in NARX configuration to predict the model error of dynamic systems. Wilkinson et al. [192] use particle filters to predict the error in the next time step.

6.7.2 Formal Uncertainty Expansion of Model-Form

Conformance Testing of formal models mainly consists of three steps [151]. The first step is to define a notion of conformance such as trace conformance or reachset conformance as described in Sect. 5.3.2. The second step means to establish a sound conformance check. This is formulated as an inverse optimization problem to get the set of model parameters that captivates by the tightest bounds on the calibration data. It includes heavy inverse calculations with cascading loops [81]. The third step defines the calibration scenarios to be tested.

7 Application Examples

After presenting our framework step by step in the last chapters, we now give a comprehensive overview of the literature in various technical areas. We present application examples of the methodology, classify the literature in our framework and present conclusions so that the different application areas can be developed further in a targeted manner. We select numerical simulations because they include the latest research on non-deterministic aggregation methods, and complex system simulations [47] from the automotive, railway and aircraft domain because they still rely heavily on conventional deterministic methods, but can benefit greatly from new methods.

7.1 Numerical Simulations

The explanations in the previous sections were very much based on advanced VV&UQ methods from numerical simulation fields such as FEM and CFD. We will therefore only give a brief summary of Probability Bound Analysis (PBA) and the Bayesian network approach, as both were spread over several sections. A comprehensive comparison between PBA and subjective probability approaches can be found in [65]. We supplement additional literature and frameworks, give an example and list application fields where those methods were already applied.

7.1.1 Probability Bound Analysis

Oberkampf and Roy [133] use PBA in CFD simulations to separately quantify input, numerical and model-form uncertainties as described in particular in Sect. 5.2.3, 6.2.2 and 6.4.5. Regarding the origins of PBA, the interested reader is referred to Ferson [61, 62, 64]. PBA is a complete non-deterministic VV&UQ framework. Regarding the classification into our framework, it addresses the verification, validation and application domain. It does not include model calibration, but only direct parameter estimation. The aggregation is done exclusively by uncertainty expansion of model responses.

7.1.2 Bayesian Network Approach

Sankararaman and Mahadevan [160] have developed the Bayesian network approach very far as described in Sect. 6.2.1, 6.3.2, 5.2.2 and 6.4.3, even with extensions to dynamic systems in Sect. 6.7.1 and hierarchical systems in Sect. 6.6.1. It addresses all domains and blocks of our framework, except the error learning and inference blocks for extrapolation. It uses bias correction for numerical errors and otherwise focuses on uncertainty expansion.

7.1.3 V&V 20 Standard

The standard V&V 20 [7] deals with V&V of CFD and thermal simulations. It distinguishes four cases regarding the observation of QoIs: direct observation, calculation from other observed variables with a simple equation either with or without joint error source and finally a model to relate the QoIs to observed variables. The standard describes extensively how to determine numerical, input and experimental uncertainties. These can then be used in each of the four cases to deduce the validation uncertainty of the model [7, Eq. (1-5-8, 1-5-10)]:

$$e^v \pm k u_{val}, \quad u_{val} = \sqrt{u_{num}^2 + u_{input}^2 + u_{obs}^2}. \tag{68}$$

An overview of six V&V standards for different numerical applications can be found in [166]. Tanaka [175] present the V&V plus UQ and Prediction (V2UP) procedure, which describes five steps for the assessment of thermal simulations, based on existing methods such as V&V20 and the AVM metric.

7.1.4 Numerical Application Fields

These VV&UQ methods were applied in many numerical simulations. For example, Choudhary et al. [30] applied PBA to a Sandia V&V challenge problem including 450 liquid tanks. The application decision making states that the quantity von Mises stress does not exceed the yield stress by a probability of 10^{-3}. They quantified and propagated input uncertainties for scenario inputs such as the liquid height and parameters such as the radius and material of the tank. In model verification they performed code-to-code comparisons with Richardson extrapolation. In model validation they applied u-pooling and the MAVM metric, since only sparse data of four tanks was available and not for the actual QoI. After the aggregation of all uncertainties, they estimated the maximum failure probability as 0.0034 and concluded that the tanks cannot be considered safe.

Further applications can be found in Reynolds-averaged Navier–Stokes equations [194], in manufacturing [44, 176], in civil engineering [13, 109, 118], in wind energy [183], in watershed modeling [197], in naval engineering [50], in power electronics [144, 145] and in nuclear reactor safety [18, 140]. In contrast to automotive system simulations in the following sub-section, probabilistic approaches and metrics have already been applied several times in automotive FEM crash simulations [195, 196, 199, 200].

7.2 Automotive System Simulations

Today, cars are in a transition to automated vehicles (AVs). AVs consist of multiple modules including environment sensors, a software stack (perception, planning and control) and vehicle dynamics and interact with other traffic participants. Regarding model validation, there is a rich history in classic vehicle dynamics and a few recent papers addressing AVs. Most of these papers present interesting validation metrics. Only in vehicle dynamics, tolerances are given for validation decision making. If we classify the automotive literature in our framework, it focuses mainly on the uppermost part. However, the aggregation of errors and uncertainties to the application decision making is missing. Future literature should therefore focus more on non-deterministic methods and aggregation in order to be able to make statements about the specific validity of the models in the application case.

7.2.1 Deterministic Simulations

Vehicle dynamics has a rich history in model validation with several standards and regulations. For assessment they describe test maneuvers such as sinusoidal or step steering inputs and define KPIs such as response times or overshoot values [91, 180]. Corresponding model validation standards [92, 94] use these assessment KPIs for comparison between simulation and experiments and define tolerances for validation decision making. Similarly, tolerance bands around the time signals are used instead of the scalar KPIs in [93]. In case of agreement, the generic validity of the vehicle dynamics model is concluded. Kutluay and Winner [110] give a comprehensive overview of the literature in vehicle dynamics model validation.

Sensor models are currently emerging for virtual safety assessment of AVs. Hanke et al. [79], Schaermann et al. [163] apply an overall error, Barons and Pearson correlation coefficients as three validation metrics to validate LIDAR sensor models. They compare multi-dimensional points clouds as raw sensor data as well as occupancy grids as processed sensor data in a parking lot scenario. Zec et al. [198] validate fused radar and camera sensor models. They use a log-likelihood and a root mean squared error (RMSE) for comparison of time signals of processed traffic objects. In addition, they convert the signals to histograms and perform a distributional comparison with a Jensen–Shannon divergence. Nentwig et al. [131] create a camera model and apply a classifier to the real and synthetic camera images to generate object hypothesis and bounding boxes around the objects. They compare the size of the bounding boxes and the final binary hypothesis with a confusion matrix. Gaidon et al. [69] present the very popular computer vision dataset Virtual KITTI. They also use algorithms on real and synthetic images and eight different metrics to assess the real-to-virtual performance gap. Abbas et al. [3] similarly assess image datasets of their driving simulator and additionally compare the visual complexity regarding color and spatial information. Similar to those camera approaches, Jasinski

[95] validates radar sensor models. Goodin et al. [71] qualitatively validate LIDAR sensor models compared to an analytical model, laboratory and field tests. Holder et al. [85], Rosenberger et al. [153] systematically derive requirements for LIDAR and radar sensor models from the developer's point of view, but do not perform actual model validation.

Model validation of the entire closed-loop AV is only very rarely addressed, although the safety assessment of AVs is mainly based on simulations in the literature [149]. A Model-in-the-Loop and a Vehicle-in-the-Loop simulation were validated against proving ground tests in multiple longitudinal scenarios in [148]. Graphical comparisons and validation metrics such as the correlation coefficient were used at different stages of the AV pipeline. The influence of ground truth measurements compared to environment sensor data on the model validation of the entire AV is analyzed in [74, 132, 187]. They use multiple box plots to visualize how simulation errors flow through the AV pipeline. They compare quantities such as the AV velocity, trajectories and an overall stochastic risk measure. Johnson et al. [97] apply a formal analysis based on correct-by-construction controller design and validate it by comparing the percentage of collision-free runs to a full-scale AV. Aramrattana et al. [9] evaluate the influence of modeling errors and uncertainties on the controller performance.

7.2.2 Non-deterministic Simulations

Viehof [185] performs several experimental repetitions of vehicle dynamics maneuvers and re-simulates each one to obtain two PDFs for comparison. For highest requirements, he uses a statistical t-test to accept the model if its PDF lies within the experimental one. For lower requirements, he checks whether the simulation PDF lies within conventional tolerance values around the experimental mean. Introducing probabilistic simulations into automotive vehicle dynamics is an important contribution. However, there are some aspects to be considered in the experimental comparison [133, p. 490]. Assuming a perfect model-form with true input uncertainties, the simulation PDF should have exactly the same width as the experimental PDF and not a smaller one. This is especially relevant since currently neither parametric uncertainties are taken into account nor extensive MC sampling. This under-approximates the true input uncertainties and leads to small simulation PDFs and erroneously to valid model hypothesis, even if the model-form might be inaccurate. In addition, the binary results from the hypothesis test cannot be used for aggregation of uncertainties to the application domain. These challenges were addressed in [39] by applying PBA to a longitudinal consumption simulation and by quantifying all sources of uncertainty, just omitting the extrapolation uncertainty to the application domain.

Hartung et al. [81] apply the conformance testing approach of checking behavioral inclusion presented in Sect. 6.7.2. They use a non-deterministic dynamics model with controller of the actual AV and non-deterministic models for the other traffic participants. With reachability analysis they can prove online during driving that both trajectory sets do not intersect and the AV behavior is safe. Therefore, they only require the mentioned sub-models, but no model of the entire AV. To the best of our knowledge, there is as yet no probabilistic model validation of the entire AV for offline safety assessment.

Whereas an individual AV is assessed in microscopic simulations, several agents interact in macroscopic traffic simulations. An overview about the calibration of traffic models and calibration metrics is given in [38, 86, 178]. Detering et al. [42] propose a measurement concept to determine parameter uncertainties for statistical MC simulations. Rao and Owen [143] perform an error analysis with autoregressive integrated moving average (ARIMA) models and Zheng et al. [201] validate statistical models using extreme value theory. Ferson and Sentz [63] introduce a concept how to extend PBA to agent-based simulations considering aleatory and epistemic uncertainty.

7.3 Railway System Simulation

In the railway field, different projects and research groups address deterministic validation methods on the one hand and UQ methods on the other hand. Both worlds must be combined to obtain a complete VV&UQ framework [117]. Similar to automotive vehicle dynamics, there are many KPIs, metrics and tolerances defined, but the overall aggregation part of our framework is missing.

7.3.1 Deterministic Simulations

The project DynoTrain led to several publications regarding virtual rail vehicle acceptance and resulted in the standard [43]. Polach and Böttcher [139] summarize stationary tests in the first stage and dynamic on-track tests in the second stage. Typically KPIs such as quasi-static, maximum or root mean square (RMS) values are defined. The mean and standard deviation of each quantity must not exceed certain tolerance values. They distinguish different types of tolerances such as relative, constant or decreasing ones. Götz and Polach [72] analyze the influence of certain input conditions on the validation results and give a short overview about validation metrics. Bezin et al. [25] summarize how to deal with cases such as a train running in new conditions or a slightly modified train after acceptance.

7.3.2 Non-deterministic Simulations

Funfschilling et al. [67] quantify uncertainty due to inputs and parameters in rail vehicle simulations. They show how the variability of the input conditions such as the track geometry can be modeled and how to propagate it by the simulation model. In [68] they perform a sensitivity analysis of the inputs and in [66] they present a survey on uncertainty quantification methods in rail vehicle dynamics. Lestoille et al. [114], Lestoille [113] create a stochastic model considering uncertainties and use advanced stochastic methods such as polynomial chaos expansion to calibrate it.

Bogojević and Lučanin [28] define a probabilistic validation metric comparing two CDFs by an f-test weighted with a mean value difference. Kraft et al. [107] use statistical indicators such as mean and standard deviation of validation errors as well as a so-called least-square misfit function. The latter quantifies the distance between the error time signals in least-square fashion and can be plotted as a CDF to compare different vehicles, responses or running conditions. In [108] they analyze the effect of measurement uncertainty on the validation results.

7.4 Aircraft System Simulation

In general, it is very difficult to apply probabilistic model validation to very complex system simulations with hundreds of parameters. To counteract this, the aircraft community developed simplified VV&UQ methods that focus in particular on the aggregation in the system hierarchy.

7.4.1 Deterministic Simulations

Hällqvist et al. [76] use field measurement data to validate the entire aircraft system. They automatically extract steady-state conditions in the data and compute relative errors. They also compute a coverage measure in the input space to find the model's domain of validity. They manually extract transient measures in the data such as overshoot values or rise times to quantify the model's dynamic validity.

Eek [51], Eek et al. [55] focus on the credibility assessment of entire simulators including a meta-model with credibility information and techniques for visualization.

7.4.2 Non-deterministic Simulations

Eek et al. [56] tried to apply PBA to aircraft simulation models, but found that the number of components and parameters of the aircraft is too complex for this. Instead they developed the output uncertainty approach presented in Sect. 6.6.4 that uses component validation data on system-level without strict separation of the uncertainty sources. Kammer et al. [101] focus on spacecraft models that can never be validated entirely. They also integrate component uncertainties on system-level.

Dorobantu et al. [45] combines the Theil Inequality Coefficient (TIC) as robust control metric with statistical model validation. They prove the relation for linear aircraft systems and use Monte Carlo simulations for non-linear systems. If the TIC of the MC simulations bounds the flight data TIC, the non-deterministic simulation can be used as an over-approximative representative. In [46] they generalize the TIC to a gap metric that takes time series input and response data instead of just response data.

8 Evaluation of VV&UQ Methods

In this section, we derive evaluation criteria and apply them for the analysis of important VV&UQ approaches presented in this paper.

8.1 Evaluation Criteria

We have derived twelve evaluation criteria based on expert judgment to assess and compare the capabilities of VV&UQ approaches. In order to ensure the traceability in this analysis process, we introduce a generic rating system in Table 4. The order of the following criteria is related to the model-based process.

8.1.1 Dynamics

For generalized dynamic systems according to Sect. 6.7.2, an approach shall consider different types of representations such as differential equations [174] or discrete state-space

Table 4 Rating system for the evaluation criteria

Criteria	Ratings		
	1	2	3
Dynamics	None	One type	Many types
Hierarchy	None	One type	Many types
Physics	None	Correction	Extrapolative
V&V process	Only calibr.	Own verif.	Own valid.
IP protection	White-box	Gray-box	Black-box
Computing	Heavy	Medium	Light
Unc. types	None	E or A	E and A
Unc. sources	None	Jointly	Separately
Unc. expansion	None	Wide bounds	Tight bounds
Bias correction	None	Without PI	With PI
Extrapolation	None	Without PI	With PI
Guarantees	Deterministic	Probabilistic	Interval

equations [89]. If a dynamic system can be simplified to a static one, the classic approaches can be applied instead.

8.1.2 Hierarchy

For hierarchical systems according to Sect. 4.2, different types of architectures [124] exist, which an aggregation approach should consider.

8.1.3 Physics

Most of the current VV&UQ approaches rely on the correctness of the physical equations in the simulation model. Thus, it is helpful to preserve the physical principles, but to add extrapolative power to the correction [128], instead of only extrapolating the error structure [106].

8.1.4 V&V Process

It is not sufficient to only calibrate parameters. Model calibration should be accompanied by a preceding model verification and a subsequent model validation.

8.1.5 IP Protection

In certain use cases such as a cooperation between system manufacturers, suppliers, tool manufacturers and technical services, the organizations must respect intellectual property (IP) so that the model equations or even the model parameters are unknown. In these cases, VV&UQ approaches that can deal with black-box models are inevitable.

8.1.6 Computing

The VV&UQ approaches differ in their computational complexity during forward simulation, often heavy inverse parameter updates or learning and inference of error models. In the end, it should be feasible to execute the approaches within reasonable time.

8.1.7 Uncertainty Types

Since epistemic (E) and aleatory (A) uncertainties are two inherently different types of uncertainty, aggregation approaches that can represent both are preferred over approaches that can only represent one type and eventually convert the other with some workaround.

8.1.8 Uncertainty Sources

Since there are several sources of errors and uncertainties that can potentially compensate each other, it is preferred to separately quantify each of them.

8.1.9 Uncertainty Expansion

Uncertainty expansion is an aggregation technique that adds conservatism to the simulation model. It is important since each simulation has uncertainties, but the uncertainty bounds should be as tight as possible compared to the true experimental value.

8.1.10 Bias Correction

Bias correction is an aggregation technique that uses the gained knowledge from the model-based activities to correct the simulation model. Since it is smart to use this knowledge, but also risky, it should be combined with a prediction interval (PI) that quantifies the inherent prediction uncertainty of the correction model.

8.1.11 Extrapolation

The more the application domain differs from the other domains, the more decisive inter- and extrapolation uncertainties become. Thus, it is important to take these uncertainties into account, at best combined with the inherent prediction uncertainty of the extrapolation model.

8.1.12 Guarantees

The reliability of the final statements is crucial for credible decision-making. Interval-based VV&UQ approaches can provide absolute guarantees, probabilistic approaches statistical guarantees and deterministic approaches without uncertainties no guarantees.

8.2 Strength and Weaknesses

We compare six presented VV&UQ approaches:

1. PBA [133] (Sect. 7.1.1),
2. Bayesian network approach [160] (Sect. 7.1.2),
3. IPMs [112] (Sect. 6.3.4),
4. Meta-model approach [84] (Sect. 6.4.2),
5. Output uncertainty approach [56] (Sect. 6.6.4) and
6. Tolerance approach [92] (Sect. 7.2.1).

The classification into our framework can be found in the respective sections. We choose PBA and the Bayesian approach because both are commonly used VV&UQ frameworks, IPMs and the meta-model approach because both deal with the sophisticated aggregation of errors and uncertainties, the output uncertainty approach because it represents a simplification of PBA for complex aircraft, and finally the tolerance approach because it represents a conventional, deterministic and commonly used baseline. The

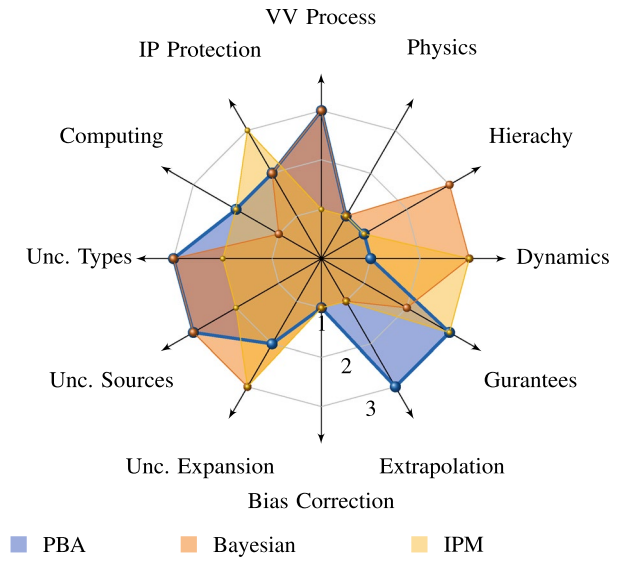


Fig. 11 Comparison of VV&UQ approaches. (Color figure online)

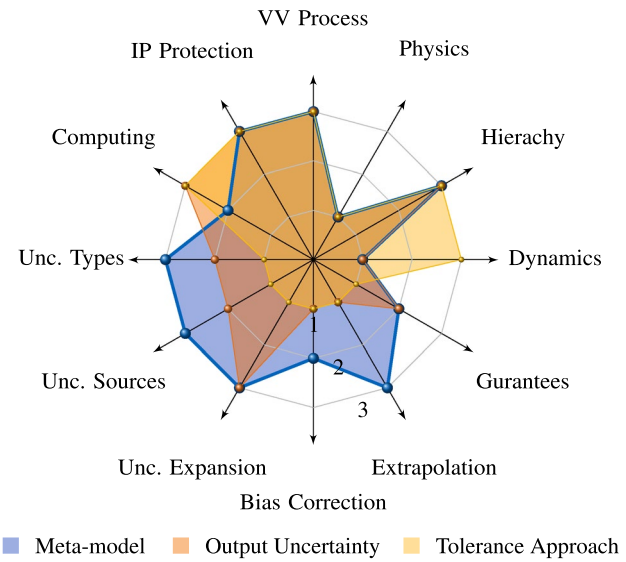


Fig. 12 Comparison of VV&UQ approaches. (Color figure online)

ratings can be seen in the Kiviat diagrams in Figs. 11 and 12. The approaches are divided into two diagrams for visualization purposes only. We have used our best judgment in choosing the rating that best represents the overall approach. Individual papers may differ from this.

As shown in the diagrams, all approaches have strength and weaknesses. PBA quantifies all sources of errors and uncertainties separately and can naturally represent all types of uncertainties, but is sometimes very conservative due to strong uncertainty expansion. The Bayesian approach has already been developed very far including extension to dynamical and hierarchical systems. It is based on subjective probabilities with priors, cannot naturally represent epistemic and aleatory uncertainties, currently lacks extrapolation uncertainty and includes heavy inverse calculations. IPMs impress with their tight bounds, but have no separate calibration and validation. Conformance testing via behavioral inclusion [81] is similar to IPMs and can also deal with dynamical systems. The meta-model approach captivates by the connection between the validation and application responses for extrapolation and by a bias correction with prediction uncertainty, but requires a linear dependency in the meta-model and a small application domain with many validation domains. The output uncertainty approach also does not quantify all sources of uncertainty, but is therefore applicable even to very complex hierarchical aircraft systems. The tolerance approach is very easy, fast and flexible in its extension to dynamical and hierarchical systems, but lacks the consideration of uncertainties and the entire aggregation.

8.3 Identification of Research Gaps

Since all aggregation approaches have strengths and weaknesses and none of them is superior with respect to all evaluation criteria, further research is needed. Each individual approach can be improved to eliminate weaknesses and extended to dynamic and hierarchical systems.

Furthermore, a combination of different methods could be interesting. With bias correction and uncertainty expansion we have presented two aggregation techniques. The former exploits the error estimation, but might be risky, since usually some uncertainties are neglected and even the estimation itself includes an uncertainty. The latter reflects all sources of uncertainties, but might be overly conservative due to uncertainty inflation [63]. The compromise between a risky bias correction and a conservative uncertainty expansion could be solved by a correction in combination with tight uncertainty bounds. Then the existing knowledge is exploited, but all remaining uncertainties of model inadequacy and the correction itself are still considered.

In general, the user needs to analyze his application in order to select the most appropriate approach and highlight the relevant weaknesses and assumptions. Therefore it is crucial that VV&UQ is accompanied by overall maturity assessment procedures, such as the Predictive Capability Maturity Model (PCMM), the Phenomenon Identification Ranking Table (PIRT) or the gap-analysis, to increase the credibility of simulation for decision-making. The interested reader is referred to [24, 88, 100, 134].

8.4 Method of Manufactured Universes

So far, we have highlighted the strengths and weaknesses of the different VV&UQ approaches, allowing the user to make a pre-selection of the most appropriate approaches at

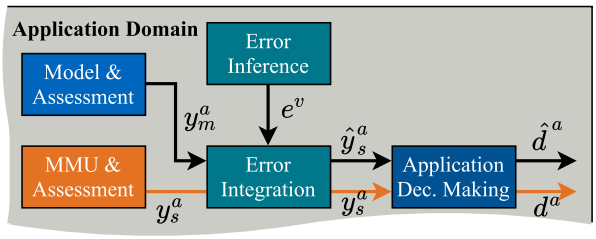


Fig. 13 Method of manufactured universes (MMU). (Color figure online)

an abstract level. In addition, we want to provide the user with a concrete procedure for analyzing and comparing in detail the effects of selected VV&UQ methods on his application. Therefore, we introduce the Method of Manufactured Universes (MMU) from Stripling et al. [172]. In contrast to the previous references, it is not a VV&UQ method to validate simulation models, but a method to validate the VV&UQ approaches themselves. It is also not a replacement for closing the identified research gaps, but rather a supportive procedure for selecting an appropriate approach from the available options.

The idea behind MMU is to replace the reality with a manufactured universe in which the true values are known and infinite simulations can be easily drawn. In reality, the true values of nature are typically unknown and experiments are very costly, whereas in the manufactured universe the true values are known and infinite simulations can be easily drawn. Thus, two simulation models are compared: the nominal model and the MMU model representing reality as a reference. The user should manufacture the universe so that the physics, model errors and experimental uncertainties are represented in the best possible way. This ensures a good transferability of the conclusions from the manufactured universe to reality. Stripling et al. [172] advise to use a high-fidelity model for MMU to get close to reality and show an example from a particle-transport universe.

Whiting et al. [191] use MMU to compare the following four approaches in a CFD universe: PBA with area metric [133], PBA with modified area metric [186], Bayesian KOH framework [104] and the V&V 20 standard [7]. They define two overall evaluation scores referred to as conservativeness and tightness. They conclude that if few data are available, PBA with modified area metric performs significantly well, whereas for many data the Bayesian approach becomes attractive.

The MMU method can be extended and embedded in our framework from Fig. 1 by replacing the system blocks of each domain with the MMU model and thus also "reviving" the dotted system block and the pipeline of the true values in the application domain. The principle is illustrated by the excerpt in Fig. 13 and highlighted in orange. Therefore, the results of an VV&UQ approach can be compared with the corresponding true value at the different evaluation stages (blocks and arrows) along the processing pipeline:

1. Responses after simulation: y_m^a versus y_s^a
2. Errors after inference: $\hat{e}^{va}$ versus e^a
3. Responses after error integration: $\hat{y}_s^a$ versus y_s^a
4. Buffer during decision-making: $\hat{y}_s^a - t_{\hat{y}}^a$ versus $y_s^a - t_y^a$
5. Binary results after decision-making: $\hat{d}^a$ versus d^a

In summary, MMU allows to inject modeling errors so that each VV&UQ approach can aggregate them for model prediction in untested application scenarios. Knowing the true values enables a detailed comparison and helps to select a suitable approach.

9 Conclusion

For credible decision-making based on simulation models, VV&UQ activities including an aggregation of all errors and uncertainties are crucial. However, the aggregation is challenging, since the errors and uncertainties are deeply interwoven. Therefore we introduced a novel modular and unified framework, which is based on the current state of science. Since the interfaces of all framework blocks are kept consistent, the framework can be used in different manifestations such as deterministic, probabilistic, interval, multi-level, dynamic and formal simulations. We gave a comprehensive survey of the VV&UQ literature across numerical simulations as well as automotive, railway and aircraft system simulations and integrated the approaches into our framework. Based on twelve evaluation criteria, we derived strength and weaknesses of the different approaches. Combined with the method of manufactured universes, those can support the user to select a suitable approach for his application.

The application fields show a very heterogeneous landscape. VV&UQ methods are comparatively advanced in numerical simulations, while complex systems present some challenges and often rely on conventional deterministic approaches for feasibility. With our framework we have created a novel, modular and uniform basis, which was a key enabler to integrate several VV&UQ methods. With our comprehensive cross-field survey, we brought approaches from several engineering fields together. This gives the possibility for new combinations and exchanges between the engineering fields, which promotes the improvement and development of new methods. Complex system simulations in particular can benefit from the transfer of new methods from numerical simulation, for example from Bayesian approaches or Probability Bound Analysis.

Through our analysis, we have found that no individual VV&UQ approach excels with regard to all criteria. The

identified weaknesses directly show the direction for individual improvements in the future. In addition, we emphasize that a combination of approaches, such as a risky bias correction of model predictions with a conservative expansion of uncertainties in untested application scenarios, can solve the trade-off and compensate weaknesses. We are convinced that both individual improvements and combinations are important to fill the research gaps.

Acknowledgements Open Access funding provided by Projekt DEAL. The authors want to thank TÜV SÜD Auto Service GmbH for the support and funding of this work. Additionally, the authors want to thank Thomas Ponn and Jakob Schneider for proofreading the article and for enhancing the content due to their critical remarks.

Author Contributions Stefan Riedmaier initiated and wrote this paper. He was involved in all stages of development and primarily developed the concept as well as the whole content of this work. Benedikt Danquah contributed to the structure of the paper and improved the content thanks to a close cooperation and many valuable discussions on VV&UQ methods. Frank Diermeyer and Bernhard Schick contributed to the conception of the research project and revised the paper critically for important intellectual content. Frank Diermeyer gave final approval of the version to be published and agrees to all aspects of the work. As a guarantor, he accepts responsibility for the overall integrity of the paper.

Compliance with Ethical Standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Abbas H (2015) Test-based falsification and conformance testing for cyber-physical systems. Ph.D. thesis, Arizona State University
2. Abbas H, Hoxha B, Fainekos G, Deshmukh JV, Kapinski J, Ueda K (2014) Conformance testing as falsification for cyber-physical systems. In: 2014 ACM/IEEE international conference on cyber-physical systems (ICCPS). IEEE, p 211
3. Abbas H, O'Kelly M, Rodionova A, Mangharam R (2017) Safe at any speed: a simulation-based test harness for autonomous vehicles. In: Seventh workshop on design, modeling and evaluation of cyber physical systems (CyPhy'17)
4. Allemang R, Spottswood M, Eason T (2014) A principal component analysis (pca) decomposition based validation metric for use with full field measurement situations. In: Atamturktur HS, Moaveni B, Papadimitriou C, Schoenherr T (eds) Model Validation and Uncertainty Quantification, vol 3. Springer International Publishing, Cham, pp 249–264
5. Althoff M (2010) Reachability analysis and its application to the safety assessment of autonomous cars. Ph.D. thesis, Technical University of Munich, Munich
6. Althoff M, Dolan JM (2012) Reachability computation of low-order models for the safety verification of high-order road vehicle models. In: 2012 American control conference (ACC). IEEE, pp 3559–3566
7. American Society of Mechanical Engineers (2009) Standard for verification and validation in computational fluid dynamics and heat transfer: an American national standard, ASME V&V, vol 20-2009, reaffirmed 2016 edn. The American Society of Mechanical Engineers, New York, NY
8. Ao D, Hu Z, Mahadevan S (2017) Dynamics model validation using time-domain metrics. J Verif Valid Uncertain Quantif 2(1):011004
9. Aramrattana M, Patel RH, Englund C, Härri J, Jansson J, Bonnet C (2018) Evaluating model mismatch impacting cacc controllers in mixed traffic using a driving simulator. In: 2018 IEEE intelligent vehicles symposium (IV). IEEE
10. Araujo H, Carvalho G, Mohaqeqi M, Mousavi MR, Sampaio A (2018) Sound conformance testing for cyber-physical systems: theory and implementation. Sci Comput Program 162:35–54
11. Arendt PD, Apley DW, Chen W (2012) Quantification of model uncertainty: calibration, model discrepancy, and identifiability. J Mech Des 134(10):100908
12. Atamturktur S, Hemez F, Williams B, Tome C, Unal C (2011) A forecasting metric for predictive modeling. Comput Struct 89(23–24):2377–2387
13. Atamturktur S, Hemez FM, Laman JA (2012) Uncertainty quantification in model verification and validation as applied to large scale historic masonry monuments. Eng Struct 43:221–234
14. Atamturktur S, Egeberg MC, Hemez FM, Stevens GN (2015a) Defining coverage of an operational domain using a modified nearest-neighbor metric. Mech Syst Signal Process 50–51:349–361
15. Atamturktur S, Stevens GN, Cheng Y (2015b) Clustered parameters of calibrated models when considering both fidelity and robustness. In: Atamturktur HS, Moaveni B, Papadimitriou C, Schoenherr T (eds) Model Valid Uncertain Quantif, vol 3. Springer International Publishing, Cham, pp 215–224
16. Avramova MN, Ivanov KN (2010) Verification, validation and uncertainty quantification in multi-physics modeling for nuclear reactor design and safety analysis. Prog Nucl Energy 52(7):601–614
17. Babuška I, Nobile F, Tempone R (2008) A systematic approach to model validation based on Bayesian updates and prediction related rejection criteria. Comput Methods Appl Mech Eng 197(29–32):2517–2539
18. Baccou J, Zhang J, Nouy E (2017) Towards a systematic approach to input uncertainty quantification methodology. In: The 17th international topical meeting on nuclear reactor thermal hydraulics (NURETH-17)
19. Baker A (2014) Summary -vvuq: verification, validation, uncertainty quantification. In: Baker A (ed) Optimal modified continuous Galerkin CFD, vol 9. Wiley, Chichester, pp 459–474
20. Balci O (1998) Verification, validation, and accreditation. In: 1998 winter simulation conference. IEEE, pp 41–48
21. Bayarri MJ, Berger JO, Paulo R, Sacks J, Cafeo JA, Cavendish J, Lin CH, Tu J (2007) A framework for validation of computer models. Technometrics 49(2):138–154

22. Beer M, Ferson S, Kreinovich V (2013) Imprecise probabilities in engineering analyses. Mech Syst Signal Process 37(1–2):4–29
23. Beg OA, Abbas H, Johnson TT, Davoudi A (2017) Model validation of pwm dc–dc converters. IEEE Trans Ind Electron 64(9):7049–7059
24. Beghini LL, Hough PD (2016) Sandia verification and validation challenge problem: a pcmm-based approach to assessing prediction credibility. J Verif Valid Uncertain Quantif 1(1):011002
25. Bezin Y, Funfschilling C, Kraft S, Mazzola L (2015) Virtual testing environment tools for railway vehicle certification. Proc Inst Mech Eng Part F J Rail Rapid Transit 229(6):755–769
26. Bi S, Prabhu S, Cogan S, Atamturktur S (2017) Uncertainty quantification metrics with varying statistical information in model calibration and validation. AIAA J 55(10):3570–3583
27. Böde E, Büker M, Ulrich E, Fränzle M, Gerwinn S, Kramer B (2018) Efficient splitting of test and simulation cases for the verification of highly automated driving functions. In: Gallina B, Skavhaug A, Bitsch F (eds) Computer safety, reliability, and security. Springer International Publishing, pp 139–153
28. Bogojević N, Lučanin V (2014) The proposal of validation metrics for the assessment of the quality of simulations of the dynamic behaviour of railway vehicles. Proc Inst Mech Eng Part F J Rail Rapid Transit 230(2):585–597
29. Campi MC, Calafiore G, Garatti S (2009) Interval predictor models: identification and reliability. Automatica 45(2):382–392
30. Choudhary A, Voyles IT, Roy CJ, Oberkampf WL, Patil M (2016) Probability bounds analysis applied to the Sandia verification and validation challenge problem. J Verif Valid Uncertain Quantif 1(1):011003
31. Coleman HW, Steele WG (2009) Experimentation, validation, and uncertainty analysis for engineers. Wiley, Hoboken
32. Crespo LG, Morelli EA, Kenny SP, Giesy DP (2014) A formal approach to empirical dynamic model optimization and validation. In: AIAA guidance, navigation, and control conference
33. Crespo LG, Kenny SP, Giesy DP (2015) Random predictor models for rigorous uncertainty quantification. Int J Uncertain Quantif 5(5):469–489
34. Crespo LG, Kenny SP, Giesy DP (2016a) A comparison of metamodeling techniques via numerical experiments. In: 18th AIAA non-deterministic approaches conference. American Institute of Aeronautics and Astronautics, p 1
35. Crespo LG, Kenny SP, Giesy DP (2016b) Interval predictor models with a linear parameter dependency. J Verif Valid Uncertain Quantif 1(2):021007
36. Crespo LG, Kenny SP, Giesy DP, Norman RB, Blattnig SR (2016c) Application of interval predictor models to space radiation shielding. In: 18th AIAA non-deterministic approaches conference. AIAA SciTech Forum
37. Crespo LG, Kenny SP, Giesy DP (2018) Staircase predictor models for reliability and risk analysis. Struct Saf 75:35–44
38. Daamen W (ed) (2015) Traffic simulation and data: validation methods and applications, [elektronische ressource] edn. Taylor and Francis and CRC Press, Hoboken and Boca Raton
39. Danquah B, Riedmaier S, Rühm J, Kalt S, Lienkamp M (2020) Statistical model verification and validation concept in automotive vehicle design. In: 30th CIRP design 2020
40. Denham CL, Patil M, Roy CJ (2018) Estimating uncertainty bounds for modified configurations from an aerodynamic model of a nominal configuration. In: 2018 AIAA atmospheric flight mechanics conference
41. Deshmukh JV, Majumdar R, Prabhu VS (2017) Quantifying conformance using the Skorokhod metric. Formal Methods Syst Des 50(2–3):168–206
42. Detering S, Schnieder L, Schnieder E (2010) Two-level validation and data acquisition for microscopic traffic simulation models. Int J Adv Syst Meas 3(1–2)
43. Deutsches Institut für Normung, European Committee for Standardization (2019) Railway applications—testing and simulation for the acceptance of running characteristics of railway vehicles—running behaviour and stationary tests
44. Díaz-Ibarra OH, Spinti J, Fry A, Isaac B, Thornock JN, Hradisky M, Smith S, Smith PJ (2018) A validation/uncertainty quantification analysis for a 1.5 mw oxy-coal fired furnace: sensitivity analysis. J Verif Valid Uncertain Quantif 3(1):011004
45. Dorobantu A, Seiler PJ, Balas GJ (2013) Validating uncertain aircraft simulation models using flight test data. In: AIAA atmospheric flight mechanics (AFM) conference. American Institute of Aeronautics and Astronautics
46. Dorobantu A, Balas GJ, Georgiou TT (2014) Validating aircraft models in the gap metric. J Aircr 51(6):1665–1672
47. Durst PJ, Anderson DT, Bethel CL (2017) A historical review of the development of verification and validation theories for simulation models. Int J Model Simul Sci Comput 08(02):1730001
48. Easterling RG (2001) Measuring the predictive capability of computational models: principles and methods, issues and illustrations
49. Easterling RG, Berger JO (2003) Statistical foundations for the validation of computer models
50. Eça L, Vaz G, Koop A, Pereira F, Abreu H (2016) Validation: What, why and how. In: Volume 2: CFD and VIV, ASME
51. Eek M (2016) On credibility assessment in aircraft system simulation. Ph.D. thesis, Linköping University, Linköping, Sweden
52. Eek M, Steinkeller S, Gavel H, Ölvander J (2013) Enabling uncertainty quantification of large aircraft system simulation models. In: 4th CEAS conference, CEAS2013: "Innovative Europe", Air & Space conference
53. Eek M, Karlén J, Ölvander J (2015a) A framework for early and approximate uncertainty quantification of large system simulation models. In: Proceedings of the 56th conference on simulation and modelling (SIMS 56), Linköping University Electronic Press, Linköping Electronic Conference Proceedings, pp 91–104
54. Eek M, Kharrazi S, Gavel H, Ölvander J (2015b) Study of industrially applied methods for verification, validation and uncertainty quantification of simulator models. Int J Model Simul Sci Comput 06(02):1550014
55. Eek M, Hällqvist R, Gavel H, Ölvander J (2016) A concept for credibility assessment of aircraft system simulators. J Aerosp Inf Syst 13(6):219–233
56. Eek M, Gavel H, Ölvander J (2017) Definition and implementation of a method for uncertainty aggregation in component-based system simulation models. J Verif Valid Uncertain Quantif 2(1):011006
57. Enszer JA, Lin Y, Ferson S, Corliss GF, Stadtherr MA (2011) Probability bounds analysis for nonlinear dynamic process models. AIChE J 57(2):404–422
58. Faes M, Moens D (2019) Recent trends in the modeling and quantification of non-probabilistic uncertainty. Arch Comput Methods Eng 179(3–4):327
59. Farajpour I, Atamturktur S (2013) Error and uncertainty analysis of inexact and imprecise computer models. J Comput Civ Eng 27(4):407–418
60. Feeley R, Seiler P, Packard A, Frenklach M (2004) Consistency of a reaction dataset. J Phys Chem A 108(44):9573–9583
61. Ferson S, Moore JDR, van den Brink JP, Estes LT, Gallagher K, O'Connor R, Verdonck F (2010) Bounding uncertainty analyses. In: Hart A (ed) Application of uncertainty analysis to ecological risks of pesticides. CRC Press, Boca Raton

62. Ferson S, Oberkampf WL (2009) Validation of imprecise probability models. Int J Reliab Saf 3(1/2/3):3
63. Ferson S, Sentz K (2016) Epistemic uncertainty in agent-based modeling. In: 7th international workshop on reliable engineering computing, pp 65–82
64. Ferson S, Oberkampf WL, Ginzburg L (2008) Model validation and predictive capability for the thermal challenge problem. Comput Methods Appl Mech Eng 197(29–32):2408–2430
65. Flage R, Aven T, Berner CL (2018) A comparison between a probability bounds analysis and a subjective probability approach to express epistemic uncertainties in a risk assessment context - a simple illustrative example. Reliab Eng Syst Saf 169:1–10
66. Funfschilling C, Perrin G (2019) Uncertainty quantification in vehicle dynamics. Veh Syst Dyn 229(6):1–25
67. Funfschilling C, Perrin G, Kraft Sönke (2012) Propagation of variability in railway dynamic simulations: application to virtual homologation. Veh Sys Dyn 50(sup1):245–261
68. Funfschilling C, Perrin G, Sebes M, Bezin Y, Mazzola L, Nguyen-Tajan ML (2015) Probabilistic simulation for the certification of railway vehicles. Proc Inst Mech Eng Part F J Rail Rapid Transit 229(6):770–781
69. Gaidon A, Wang Q, Cabon Y, Vig E (2016) Virtual worlds as proxy for multi-object tracking analysis. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR). IEEE, pp 4340–4349
70. Gardner P, Lord C, Barthorpe RJ (2018) An evaluation of validation metrics for probabilistic model outputs. In: ASME 2018 verification and validation symposium. ASME, p V001T06A001
71. Goodin C, Doude M, Hudson C, Carruth D (2018) Enabling off-road autonomous navigation-simulation of Lidar in dense vegetation. Electronics 7(9):154
72. Götz G, Polach O (2017) Verification and validation of simulations in a rail vehicle certification context. Int J Rail Transp 6(2):83–100
73. Green PL (2016) Towards the diagnosis and simulation of discrepancies in dynamical models. In: Atamturktur S, Schoenherr T, Moaveni B, Papadimitriou C (eds) Model validation and uncertainty quantification, vol 3. Springer International Publishing, Cham, pp 271–277
74. Groh K, Wagner S, Kuehbeck T, Knoll A (2019) Simulation and its contribution to evaluate highly automated driving functions. In: WCX SAE world congress experience, SAE International 400 Commonwealth Drive, Warrendale, PA, United States, SAE Technical Paper Series
75. Halder A, Bhattacharya R (2014) Probabilistic model validation for uncertain nonlinear systems. Automatica 50(8):2038–2050
76. Hällqvist R, Eek M, Lind I, Gavel H (2015) Validation techniques applied on the saab gripen fighter environmental control system model. In: Proceedings of the 56th conference on simulation and modelling (SIMS 56), Linköping University Electronic Press, Linköping Electronic Conference Proceedings, pp 199–210
77. Hamilton JR, Hills RG (2010a) Relation of validation experiments to applications. Numer Heat Transf Part B Fundam 57(5):307–332
78. Hamilton JR, Hills RG (2010b) Relation of validation experiments to applications: a nonlinear approach. Numer Heat Transf Part B Fundam 57(6):373–395
79. Hanke T, Schaermann A, Geiger M, Weiler K, Hirsenkorn N, Rauch A, Schneider SA, Biebl E (2017) Generation and validation of virtual point cloud data for automated driving systems. In: 2017 IEEE 20th international conference on intelligent transportation systems (ITSC). IEEE, pp 1–6
80. Harirchi F, Yong SZ, Ozay N (2018) Passive diagnosis of hidden-mode switched affine models with detection guarantees via model invalidation. In: Sayed-Mouchaweh M (ed) Diagnosability. Security and safety of hybrid dynamic and cyber-physical systems. Springer International Publishing, Cham, pp 227–251
81. Hartung M, Hess D, Lattarulo R, Oehlerking J, Perez J, Rausch A (2017) Report on conformance testing of application models
82. He Q (2019) Model validation based on probability boxes under mixed uncertainties. Adv Mech Eng 11(5):168781401984741
83. Hemez F, Atamturktur HS, Unal C (2010) Defining predictive maturity for validated numerical simulations. Comput Struct 88(7–8):497–505
84. Hills RG (2013) Roll-up of validation results to a target application
85. Holder M, Rosenberger P, Winner H, Makkapati VP, Maier M, Schreiber H, Magosi Z, D'hondt T, Slavik Z, Bringmann O, Rosenstiel W (2018) Measurements revealing challenges in radar sensor modeling for virtual validation of autonomous driving. In: 2018 IEEE 21th international conference on intelligent transportation systems (ITSC). IEEE pp 2616–2622
86. Hollander Y, Liu R (2008) The principles of calibrating traffic micro-simulation models. Transportation 35(3):347–362
87. Hosder S, Walters R (2010) Non-intrusive polynomial chaos methods for uncertainty quantification in fluid dynamics. In: 48th AIAA aerospace sciences meeting including the new horizons forum and aerospace exposition. [American Institute of Aeronautics and Astronautics], p 5047
88. Hu KT, Paez TL (2016) Why do verification and validation? J Verif Valid Uncertain Quantif 1(1):011008
89. Hu Z, Hu C, Mourelatos ZP, Mahadevan S (2018) Dynamic model discrepancy quantification in simulation-based design of dynamical systems. In: Volume 2B: 44th Design automation conference. ASME, p V02BT03A052
90. Hu Z, Hu C, Mourelatos ZP, Mahadevan S (2019) Model discrepancy quantification in simulation-based design of dynamical systems. J Mech Des 141(1):011401
91. International Organization for Standardization (2011) Road vehicles—lateral transient response test methods—open-loop test methods
92. International Organization for Standardization (2016a) Passenger cars—validation of vehicle dynamic simulation—sine with dwell stability control testing
93. International Organization for Standardization (2016b) Passenger cars—vehicle dynamic simulation and validation—steady-state circular driving behaviour
94. International Organization for Standardization (2020) Road vehicles—passenger cars—vehicle dynamic simulation and validation—lateral transient response test methods
95. Jasinski M (2019) A generic validation scheme for real-time capable automotive radar sensor models integrated into an autonomous driving simulator. In: 2019 24th International conference on methods and models in automation and robotics (MMAR). IEEE, pp 612–617
96. Jiang X, Mahadevan S (2008) Bayesian wavelet method for multivariate model assessment of dynamic systems. J Sound Vib 312(4–5):694–712
97. Johnson B, Havlak F, Kress-Gazit H, Campbell M (2017) Experimental evaluation and formal analysis of high-level tasks with dynamic obstacle anticipation on a full-sized autonomous vehicle. J Field Robot 34(5):897–911
98. Joint Committee for Guides in Metrology (JCGM (2008) Evaluation of measurement data—guide to the expression of uncertainty in measurement (gum)
99. Junietz P (2019) Microscopic and macroscopic risk metrics for the safety validation of automated driving. Ph.D. thesis, TU Darmstadt, Darmstadt
100. Kaizer JS, Heller AK, Oberkampf WL (2015) Scientific computer simulation review. Reliab Eng Syst Saf 138:210–218

101. Kammer DC, Blelloch PA, Sills J (2019) Test-based uncertainty quantification and propagation using hurty/craig-bampton substructure representations. In: Proceedings of the IMAC-XXXVII
102. Karydis K, Poulakakis I, Sun J, Tanner HG (2015) Probabilistically valid stochastic extensions of deterministic models for systems with uncertainty. Int J Robot Res 34(10):1278–1295
103. Kat CJ, Els PS (2012) Validation metric based on relative error. Math Comput Model Dyn Syst 18(5):487–520
104. Kennedy MC, O'Hagan A (2001) Bayesian calibration of computer models. J R Stat Soc Ser B (Stat Methodol) 63(3):425–464
105. Khakpour N, Mousavi MR (2015) Notions of conformance testing for cyber-physical systems: overview and roadmap. In: Aceto L, Frutos Escrig Dd (eds) 26th International conference on concurrency theory, Leibniz international proceedings in informatics, Schloss Dagstuhl - Leibniz-Zentrum für Informatik GmbH Dagstuhl Publishing, Saarbrücken/Wadern, Germany, pp 18–40
106. King WE, Arsenlis A, Tong C, Oberkampf WL (2012) Uncertainties in predictions of material performance using experimental data that is only distantly related to the system of interest. In: Dienstfrey AM, Boisvert RF (eds) Uncertainty quantification in scientific computing, IFIP Advances in Information and Communication Technology, vol 377. Springer, Berlin, Heidelberg, pp 294–311
107. Kraft S, Causse J, Coudert F (2015) An approach for the validation of railway vehicle models based on on-track measurements. Veh Syst Dyn 53(10):1480–1499
108. Kraft S, van Clooster Q, Causse J (2017) Validation of railway vehicle models considering measurement uncertainty. In: 19th International conference on railway engineering (ICRE 2017)
109. Kumar M, Whittaker AS (2018) Cross-platform implementation, verification and validation of advanced mathematical models of elastomeric seismic isolation bearings. Eng Struct 175:926–943
110. Kutluay E, Winner H (2014) Validation of vehicle dynamics simulation models—a review. Veh Syst Dyn 52(2):186–200
111. Kwag S, Gupta A, Dinh N (2018) Probabilistic risk assessment based model validation method using Bayesian network. Reliab Eng Syst Saf 169:380–393
112. Lacerda MJ, Crespo LG (2017) Interval predictor models for data with measurement uncertainty. In: 2017 American control conference (ACC). IEEE, pp 1487–1492
113. Lestoille N (2015) Stochastic model of high-speed train dynamics for the prediction of long-term evolution of the track irregularities. Ph.D. thesis, Université Paris-Est, Paris, France
114. Lestoille N, Soize C, Funfschilling C (2016) Stochastic prediction of high-speed train dynamics to long-term evolution of track irregularities. Mech Res Commun 75:29–39
115. Li C, Mahadevan S (2014) Uncertainty quantification and integration in multi-level problems. In: Atamturktur HS, Moaveni B, Papadimitriou C, Schoenherr T (eds) Model validation and uncertainty quantification, vol 3. Springer International Publishing, Cham, pp 89–98
116. Li W, Chen S, Jiang Z, Apley DW, Lu Z, Chen W (2016) Integrating Bayesian calibration, bias correction, and machine learning for the 2014 Sandia verification and validation challenge problem. J Verif Valid Uncertain Quantif 1(1):011004
117. Licciardello R, Funfschilling C, Malavasi G (2016) Accuracy of the experimental assessment of running dynamics characteristics quantified through an uncertainty framework. Proc Inst Mech Eng Part F J Rail Rapid Transit 231(8):945–960
118. Lin X, Zong Z, Niu J (2015) Finite element model validation of bridge based on structural health monitoring–part ii: uncertainty propagation and model validation. J Traffic Transp Eng (Engl Ed) 2(4):279–289
119. Ling Y, Mahadevan S (2013) Quantitative model validation techniques: new insights. Reliab Eng Syst Saf 111:217–231
120. Ling Y, Mullins J, Mahadevan S (2014a) Options for the inclusion of model discrepancy in bayesian calibration. In: 16th AIAA non-deterministic approaches conference. American Institute of Aeronautics and Astronautics
121. Ling Y, Mullins J, Mahadevan S (2014b) Selection of model discrepancy priors in Bayesian calibration. J Comput Phys 276:665–680
122. Liu SB, Althoff M (2018) Reachset conformance of forward dynamic models for the formal analysis of robots. In: 2018 IEEE/RSJ international conference on intelligent robots and systems (IROS). IEEE, pp 370–376
123. Liu Y, Chen W, Arendt P, Huang HZ (2011) Toward a better understanding of model validation metrics. J Mech Des 133(7):071005
124. Mahadevan S (2018) Uncertainty aggregation variability, statistical uncertainty, and model uncertainty. In: École Thématique sur les Incertitudes en Calcul Scientifique (ETICS)
125. Morrison RE, Bryant CM, Terejanu G, Prudhomme S, Miki K (2013) Data partition methodology for validation of predictive models. Comput Math Appl 66(10):2114–2125
126. Mullins J, Ling Y, Mahadevan S, Sun L, Strachan A (2016a) Separation of aleatory and epistemic uncertainty in probabilistic model validation. Reliab Eng Syst Saf 147:49–59
127. Mullins J, Mahadevan S, Urbina A (2016b) Optimal test selection for prediction uncertainty reduction. J Verif Valid Uncertain Quantif 1(4):041002
128. Mullins J, Schroeder B, Hills R, Crespo L (2016c) A survey of methods for integration of uncertainty and model form error in prediction. In: Probabilistic mechanics & reliability conference (PMC)
129. National Research Council (2012) Assessing the reliability of complex models: mathematical and statistical foundations of verification, validation, and uncertainty quantification. National Academies Press, Washington, DC
130. Neal K, Li C, Hu Z, Mahadevan S, Mullins J, Schroeder B, Subramanian A (2019) Confidence in the prediction of unmeasured system output using roll-up methodology. In: Barthorpe R (ed) Model validation and uncertainty quantification, vol 3. Springer International Publishing, Cham, pp 105–107
131. Nentwig M, Miegler M, Stamminger M (2012) Concerning the applicability of computer graphics for the evaluation of image processing algorithms. In: 2012 IEEE international conference on vehicular electronics and safety (ICVES 2012). IEEE, pp 205–210
132. Notz D, Sigl M, Kühbeck T, Wagner S, Groh K, Schütz C, Watzenig D (2019) Methods for improving the accuracy of the virtual assessment of autonomous driving. In: 2019 IEEE international conference on connected vehicles and expo (ICCVE) proceedings
133. Oberkampf WL, Roy CJ (2010) Verification and validation in scientific computing. Cambridge University Press, Cambridge
134. Oberkampf WL, Smith BL (2017) Assessment criteria for computational fluid dynamics model validation experiments. J Verif Valid Uncertain Quantif 2(3):031002
135. Oliver TA, Terejanu G, Simmons CS, Moser RD (2015) Validating predictions of unobserved quantities. Comput Methods Appl Mech Eng 283:1310–1335
136. Ozay N, Sznaier M, Lagoa C (2014) Convex certificates for model (in)validation of switched affine systems with unknown switches. IEEE Trans Autom Control 59(11):2921–2932
137. Panesi M, Miki K, Prudhomme S, Brandis A (2012) On the assessment of a Bayesian validation methodology for data reduction models relevant to shock tube experiments. Comput Methods Appl Mech Eng 213–216:383–398
138. Pasha HG, Allemang RJ, Agarkar M (2016) Application of pca-svd validation metric to develop calibrated and validated

structural dynamic models. In: Atamturktur S, Schoenherr T, Moaveni B, Papadimitriou C (eds) Model validation and uncertainty quantification, vol 3. Springer International Publishing, Cham, pp 213–226
139. Polach O, Böttcher A (2014) A new approach to define criteria for rail vehicle model validation. Veh Syst Dyn 52(sup1):125–141
140. Porter NW, Mousseau VA, Avramova MN (2018) Quantified validation with uncertainty analysis for turbulent single-phase friction models. Nucl Technol 2008(5):1–11
141. Prabhu S, Atamturktur S, Cogan S (2017) Model assessment in scientific computing: considering robustness to uncertainty in input parameters. Eng Comput 34(5):1700–1723
142. Prajna S (2006) Barrier certificates for nonlinear model validation. Automatica 42(1):117–126
143. Rao L, Owen L (2000) Validation of high-fidelity traffic simulation models. Transp Res Rec J Transp Res Board 1710(1):69–78
144. Rashidi Mehrabadi N, Wen B, Burgos R, Boroyevich D, Roy C, (2014) Verification, validation and uncertainty quantification (vv & uq) framework applicable to power electronics systems. In: SAE, (2014) Aerospace systems and technology conference, SAE International 400 Commonwealth Drive. Warrendale, PA, United States, SAE Technical Paper Series
145. Rashidi Mehrabadi N, Burgos R, Boroyevich D, Roy C (2017) Modeling and design of the modular multilevel converter with parametric and model-form uncertainty quantification. In: 2017 IEEE energy conversion congress and exposition (ECCE). IEEE, pp 1513–1520
146. Rebba R, Mahadevan S (2008) Computational methods for model reliability assessment. Reliab Eng Syst Saf 93(8):1197–1207
147. Rhode MN, Oberkampf WL (2017) Estimation of uncertainties for a model validation experiment in a wind tunnel. J Spacecr Rockets 54(1):155–168
148. Riedmaier S, Nesensohn J, Gutenkunst C, Düser T, Schick B, Abdellatif H (2018) Validation of x-in-the-loop approaches for virtual homologation of automated driving functions. In: 11th Graz symposium virtual vehicle (GSVF)
149. Riedmaier S, Ponn T, Ludwig D, Schick B, Diermeyer F (2020) Survey on scenario-based safety assessment of automated vehicles. IEEE Open Access
150. Roche G, Prabhu S, Shields P, Atamturktur S (2015) Model validation in scientific computing: considering robustness to non-probabilistic uncertainty in the input parameters. In: Atamturktur HS, Moaveni B, Papadimitriou C, Schoenherr T (eds) Model validation and uncertainty quantification, vol 3. Springer International Publishing, Cham, pp 189–198
151. Roehm H, Oehlerking J, Woehrle M, Althoff M (2016) Reach-set conformance testing of hybrid automata. In: Abate A, Fainekos G (eds) Proceedings of the 19th international conference on hybrid systems: computation and control—HSCC '16. ACM Press, pp 277–286
152. Romero V (2019) Real-space model validation and predictor–corrector extrapolation applied to the Sandia cantilever beam end-to-end uq problem. In: AIAA Scitech 2019 Forum, American Institute of Aeronautics and Astronautics
153. Rosenberger P, Holder M, Zirulnik M, Winner H (2018) Analysis of real world sensor behavior for rising fidelity of physically based Lidar sensor models. In: 2018 IEEE intelligent vehicles symposium (IV). IEEE
154. Roy CJ (2018) Unanswered questions in 1) verification, 2) validation and 3) uncertainty quantification. In: ASME 2018 verification and validation symposium. ASME
155. Roy CJ, Balch MS (2012) A holistic approach to uncertainty quantification with application to supersonic nozzle thrust. Int J Uncertain Quantif 2(4):363–381
156. Rutherford BM (2008) Computational modeling issues and methods for the "regulatory problem" in engineering—solution to the thermal problem. Comput Methods Appl Mech Eng 197(29–32):2480–2489
157. Sadeghi J, Angelis Md, Patelli E (2018a) Frequentist history matching with interval predictor models. Appl Math Model 61:29–48
158. Sadeghi J, Angelis Md, Patelli E (2018b) Robust propagation of probability boxes by interval predictor models. In: Proceedings of the joint ICVRAM ISUMA UNCERTAINTIES conference
159. Sankararaman S, Mahadevan S (2013) Assessing the reliability of computational models under uncertainty. In: 54th AIAA/ASME/ASCE/AHS/ASC structures, atructural dynamics and materials conference
160. Sankararaman S, Mahadevan S (2015) Integration of model verification, validation, and calibration for uncertainty quantification in engineering systems. Reliab Eng Syst Saf 138:194–209
161. Sargent RG, Balci O (2017) History of verification and validation of simulation models. In: 2017 Winter simulation conference (WSC). IEEE, pp 292–307
162. Sarin H, Kokkolaras M, Hulbert G, Papalambros P, Barbat S, Yang RJ (2010) Comparing time histories for validation of simulation models: error measures and metrics. J Dyn Syst Meas Control 132(6):061401
163. Schaermann A, Rauch A, Hirsenkorn N, Hanke T, Rasshofer R, Biebl E (2017) Validation of vehicle environment sensor models. In: 2017 IEEE intelligent vehicles symposium (IV). IEEE, pp 405–411
164. Schroeder BB, Mullins JG (2016) Exploring model form uncertainty approaches with a burgers' equation example. In: ASME 2016 verification & validation symposium
165. Schürmann B, Heß D, Eilbrecht J, Stursberg O, Koster F, Althoff M (2017) Ensuring drivability of planned motions using formal methods. In: 2017 IEEE 20th international conference on intelligent transportation systems (ITSC). IEEE, pp 1–8
166. Sharma V, Freitas CJ, Kim M, Bell J (2018) Verification and validation of computational modeling in energy systems. In: Offshore technology conference, Offshore Technology Conference
167. Shinn R, Hemez FM, Doebling SW (2003) Estimating the error in simulation prediction over the design space. In: Proceedings of the 44th AIAA/ASME/ASCE/AHS structures, structural dynamics, and materials conference
168. Stevens G, Atamturktur S (2017) Mitigating error and uncertainty in partitioned analysis: a review of verification, calibration and validation methods for coupled simulations. Arch Comput Methods Eng 24(3):557–571
169. Stevens G, Atamturktur S, Lebensohn R, Kaschner G (2016) Experiment-based validation and uncertainty quantification of coupled multi-scale plasticity models. Multidiscip Model Mater Struct 12(1):151–176
170. Stevens GN (2016) Experiment-based validation and uncertainty quantification of partitioned models: improving predictive capability of multi-scale plasticity models. Ph.D. thesis, Clemson University, Clemson, South Carolina, USA
171. Streif S, Henrion D, Findeisen R (2014) Probabilistic and set-based model invalidation and estimation using lmis. IFAC Proc Vol 47(3):4110–4115
172. Stripling HF, Adams ML, McClarren RG, Mallick BK (2011) The method of manufactured universes for validating uncertainty quantification methods. Reliab Eng Syst Saf 96(9):1242–1256
173. Stursberg O, Kontny D, Liu Z, Rausch A, Oehlerking J, Prandini M, Frehse G (2017) Report on modelling of networked cyber-physical system for verification and control
174. Subramanian A, Mahadevan S (2019) Bayesian estimation of discrepancy in dynamics model prediction. Mech Syst Signal Process 123:351–368

175. Tanaka M (2016) Application of area validation methods for uncertainty quantification in validation process of thermal-hydraulic code for thermal fatigue issue in sodium-cooled fast reactors. In: ASME 2016 verification & validation symposium
176. Tapia G, King W, Johnson L, Arroyave R, Karaman I, Elwany A (2018) Uncertainty propagation analysis of computational models in laser powder bed fusion additive manufacturing using polynomial chaos expansions. J Manuf Sci Eng 140(12):121006
177. Terejanu G (2015) Predictive validation of dispersion models using a data partitioning methodology. In: Atamturktur HS, Moaveni B, Papadimitriou C, Schoenherr T (eds) Model validation and uncertainty quantification, vol 3. Springer International Publishing, Cham, pp 151–156
178. Toledo T, Koutsopoulos HN (2004) Statistical validation of traffic simulation models. Transp Res Rec J Transp Res Board 1876(1):142–150
179. Ulbrich S, Menzel T, Reschka A, Schuldt F, Maurer M (2015) Defining and substantiating the terms scene, situation, and scenario for automated driving. In: 2015 IEEE 18th international conference on intelligent transportation systems. pp 982–988
180. United Nations Economic Commission for Europe (UNECE) (2017) Addendum 139—regulation no. 140—uniform provisions concerning the approval of passenger cars with regard to electronic stability control (esc) systems
181. United Nations Economic Commission for Europe (UNECE) (2020) Proposal for a new un regulation on: uniform provisions concerning the approval of vehicles with regard to automated lane keeping systems: Grva-06-02-rev.4
182. Urbina A, Hills RG, Hetzler AC (2014) On the aggregation and extrapolation of uncertainty from component to system level models. In: Atamturktur HS, Moaveni B, Papadimitriou C, Schoenherr T (eds) Model validation and uncertainty quantification, vol 3. Springer International Publishing, Cham, pp 11–23
183. van Buren KL, Mollineaux MG, Hemez FM, Atamturktur S (2013) Simulating the dynamics of wind turbine blades: Part ii, model validation and uncertainty quantification. Wind Energy 16(5):741–758
184. van Buren KL, Ouisse M, Cogan S, Sadoulet-Reboul E, Maxit L (2017) Effect of model-form definition on uncertainty quantification in coupled models of mid-frequency range simulations. Mech Syst Signal Process 93:351–367
185. Viehof M (2018) Objektive qualitätsbewertung von fahrdynamiksimulationen durch statistische validierung. Ph.D. thesis, Technische Universität Darmstadt, Darmstadt
186. Voyles IT, Roy CJ (2015) Evaluation of model validation techniques in the presence of aleatory and epistemic input uncertainties. In: 17th AIAA non-deterministic approaches conference. American Institute of Aeronautics and Astronautics
187. Wagner S, Groh K, Kuhbeck T, Knoll A (2019) Towards cross-verification and use of simulation in the assessment of automated driving. In: 2019 IEEE intelligent vehicles symposium (IV). IEEE, pp 1589–1596
188. Wang N, Yao W, Zhao Y, Chen X, Zhang X, Li L (2018) A new interval area metric for model validation with limited experimental data. J Mech Des 140(6)
189. Wang Z, Fu Y, Yang RJ, Barbat S, Chen W (2016) Validating dynamic engineering models under uncertainty. J Mech Des 138(11):111402
190. Wei Z, Robbersmyr KG, Karimi HR (2017) An eemd aided comparison of time histories and its application in vehicle safety. IEEE Access 5:519–528
191. Whiting NW, Roy CJ, Duque EP, Lawrence S (2019) Assessment of model validation and calibration approaches in the presence of uncertainty. In: AIAA Scitech 2019 Forum, American Institute of Aeronautics and Astronautics
192. Wilkinson RD, Vrettas M, Cornford D, Oakley JE (2011) Quantifying simulator discrepancy in discrete-time dynamical simulators. J Agric Biol Environ Stat 16(4):554–570
193. Xi Z, Pan H, Fu Y, Yang RJ (2015) Validation metric for dynamic system responses under uncertainty. SAE Int J Mater Manuf 8(2):309–314
194. Xiao H, Cinnella P (2019) Quantification of model uncertainty in rans simulations: a review. Prog Aerosp Sci 108:1–31
195. Yang J, Zhan Z, Chen C, Shu Y, Zheng L, Yang RJ, Fu Y, Barbat S (2015) Development of a comprehensive validation method for dynamic systems and its application on vehicle design. SAE Int J Mater Manuf 8(3)
196. Yang X, Zhan Z, Wang Q, Wang P, Fang Y, Zheng L (2018) An integrated deformed surfaces comparison based validation framework for simplified vehicular cae models. In: WCX world congress experience, SAE International 400 Commonwealth Drive, Warrendale, PA, United States, SAE Technical Paper Series
197. Yen H, Wang X, Fontane DG, Harmel RD, Arabi M (2014) A framework for propagation of uncertainty contributed by parameterization, input data, model structure, and calibration/validation data in watershed modeling. Environ Model Softw 54:211–221
198. Zec EL, Mohammadiha N, Schliep A (2018) Statistical sensor modelling for autonomous driving using autoregressive input–output hmms. In: 2018 IEEE 21th international conference on intelligent transportation systems (ITSC). IEEE, pp 1331–1336
199. Zhan Z, Yang J, Fu Y, Yang RJ, Barbat S, Zheng L (2015) Research on validation metrics for multiple dynamic response comparison under uncertainty. SAE Int J Mater Manuf 8(2)
200. Zhan Z, Yang J, Chen X, Shen Z (2016) An integrated validation method for nonlinear multiple curve comparisons. SAE Int J Mater Manuf 9(2)
201. Zheng L, Sayed T, Essa M, Guo Y (2019) Do simulated traffic conflicts predict crashes? An investigation using the extreme value approach. In: 2019 IEEE 22th international conference on intelligent transportation systems (ITSC). IEEE