IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 19, 2022, accepted February 9, 2022, date of publication February 23, 2022, date of current version March 9, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3153722

Virtual Testing of Automated Driving Systems. A Survey on Validation Methods

RICCARDO DONÀ[1] AND BIAGIO CIUFFO[2], (Member, IEEE)
[1]Uni Systems Italy, 20145 Milan, Italy
[2]Joint Research Center (JRC), 21027 Ispra, Italy

Corresponding author: Biagio Ciuffo (biagio.ciuffo@ec.europa.eu)

This work was supported by the Joint Research Centre for the European Commission.

ABSTRACT This paper surveys the state-of-the-art contributions supporting the validation of virtual testing toolchains for Automated Driving System (ADS) verification. The work builds upon the well-known limitations of physical testing while conceiving the virtual counterpart as a fundamental ingredient for the type-approval of high automation level ADS. The purpose of the research effort is to summarize computational tools, validation methodologies, and the corresponding fidelity levels delivered by state-of-the-art simulation toolchains. The ultimate goal is to establish how effectively simulation can play the role of a ''virtual proving ground'' for ADS certification independently from any specific ADS implementation/effectiveness. The contribution includes classic high-level validation approaches and modern specific computational tools that can be adopted depending on the type of data under analysis. Moreover, the investigation covers approaches embraced both within the scientific community and in technical regulations for the sake of completeness. Ultimately, we identified two high-level validation schema: integrated environment and subsystem-based solutions. In addition, we found that modeling and validating virtual sensors for ADS is the most lacking area from a subsystem-level approach. On the other side, the closed-loop interaction between the ADS and other virtual traffic participants makes it difficult to directly compare the experimental results with simulated generated evidence as the emergent behaviors of the ADS may amplify minor discrepancies between the environments.

INDEX TERMS Automated driving, model validation, simulation, virtual testing.

I. INTRODUCTION

Virtual testing is gradually becoming part of the *certification* process of future Automated Driving System (ADS) [1], [2]. Virtual tests are starting to complement the traditionally performed proving ground and public road experiments after decades of simulation's utilization for new technology *development* purposes. Introducing a virtual component in the ADS certification pipeline brings all the well-known advantages of simulation with respect to physical testing. These include tests *repeatability* across different ADS/vehicle combinations, the possibility of *scaling* up the number of tests, a *safer* technology assessment, and a reduction in the *costs* associated with the certification process. Indeed, validating a high-automation level ADS by means of physical testing only would require traveling for decades, as already pointed out in several literature works [3], [4]. Simulation is thus the ideal testing environment to investigate how an ADS perform in high-mileage tests and in edge-case driving scenarios [5].

Naturally, the adoption of virtual testing as an ADS certification tool shall first investigate the appropriateness of the simulation environment for such a purpose. Despite the mentioned simulation's assets, suitable validation procedures have to be established before promoting a virtual testing environment as a certification environment. In particular, a *validation* activity shall ensure that the simulation-generated evidence is characterized by a fidelity level that serves the certification process's need. A virtual testing environment fulfilling such validation criteria could be regarded to as a ''virtual proving ground''. However, an acknowledged validation practice to accomplish the accreditation task is a research topic that is not yet regulated. Instead, a plethora of approaches exists which depend on the specific application.

Unlike widely acknowledged scientific contributions dealing with the implementation and validation of ADSs (see for instance [6]–[9]), our effort concentrates on the (simu-

The associate editor coordinating the review of this manuscript and approving it for publication was Tamas Tettamanti.

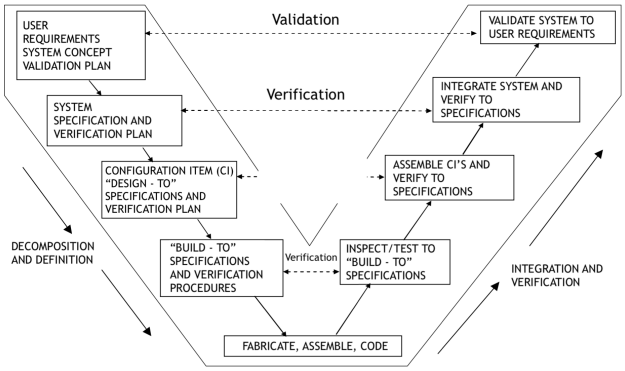


FIGURE 1. V-Model workflow, adapted from [11].

lated) environment used to support the ADS certification task. Indeed, the objective of this work is to survey the state-of-the-art approaches adopted to validate both complete ADS virtual testing toolchains and simulation submodels involved in the virtual experiments. Such a selection of the topics sets our work apart from the research stream focusing on the ADS development. In fact, the validation approaches presented here can be adopted *regardless* of the specific ADS implementation and its safety effectiveness. Despite some of the state-of-the-art surveys (for instance [7], [9]) also list simulation softwares, the fidelity of the tool is typically not discussed nor computational methods for a quantitative assessment are presented. Our contribution aims to fill the mentioned research gap by presenting a review of the approaches and computational tools adopted in actual simulation validation practices together with the corresponding acceptance correlation thresholds whenever applicable to the end of establishing the fidelity level that a virtual testing toolchain for ADS simulation can deliver.

A. STRUCTURE OF THE WORK

The work opens by recalling the general principles of validation in Section II, focusing on the validation of virtual models. The associated computational tools are presented in Section III. Then the discussion shifts to the ADS-specific validation approaches and challenges Section IV. Open issues as reported in Section V whereas conclusions are eventually drawn in Section VI.

II. VALIDATION IN ENGINEERING

The validation process, in general, aims at determining the *capability* of a *product* to fulfill its *purpose* and expectations for the required *application*. In contrast to the *verification* task, the process of validation takes place upon the completion of the product to ensure that the ''correct'' product was built. The verification, instead, is mainly concerned with the correct implementation of the conceptual (or mathematical) model [10] into the actual product. The considerations are graphically emphasized in Fig. 1, where the typical V-Model product development workflow is displayed.

The general definition for validation had been given a bespoke formulation for the field of Modeling and Simulation (M&S) in several literature works [12], [13]. In particular, the validation of a virtual model (product) can be defined as a procedure aimed at establishing the model's accuracy (capability) in representing the real-world (purpose) from the perspective of the intended use (application).

Ensuring suitable accuracy for a virtual model is a crucial task whenever the model plays a critical role. That is becoming more relevant in recent years due to the widespread adoption of simulation at any design and decision-making level [14].

One of the first attempts to build up a validation framework for virtual models was carried out by Carson in [15]. According to the author, the validation should be made up of a three steps procedure:

1) face validity (*i.e.*, answering the question: ''is the model returning reasonable results?'');
2) test the model over a range of input parameters (''stress test'');
3) comparison of the model's predictions with physical data whenever possible, in particular:
   a) collection of input data from real-world experiments;
   b) re-execution of the simulation model with the collected input;
   c) performance comparison with respect to the real-world;
   d) use statistical techniques to create confidence intervals in case multiple datasets are available.

Carson's scientific contribution also mentions a list of possible modeling errors, which include: project management (*e.g.*, missing key personnel or decision-makers at crucial meetings), data modeling (*e.g.*, use of incorrect data collection procedures), logic model (*e.g.*, modeling assumption not replicating physical behavior of the system), and experimentation errors (*e.g.*, too few executions of the model).

In [16], a three-steps validation approach is proposed as graphically reported in Fig. 2. More in detail, the virtual model is first compared against the real one in terms of predicted output. Secondly, an interpolation/extrapolation analysis over the required domain is carried out. Thirdly, the prediction uncertainty is characterized.

A widely recognized practical approach for the validation of simulation models has been proposed in [14] and depicted in Fig. 3. The approach foresees the distinction between the ''conceptualization'', *i.e.*, the mathematical/logical representation of the model and the ''computerization'', *i.e.*, the realization of the conceptual model in terms of a programming language. Validation then takes place by determining the adoption of reasonable theories/assumptions (''conceptual validation''), the correctness of the computer implementation (''computerized model verification'', *i.e.*, static and dynamic code testing), and the accuracy of the realized virtual model (''operational validation''). Eventually, a ''data validity'' procedure should assess that accurate data were available to forge

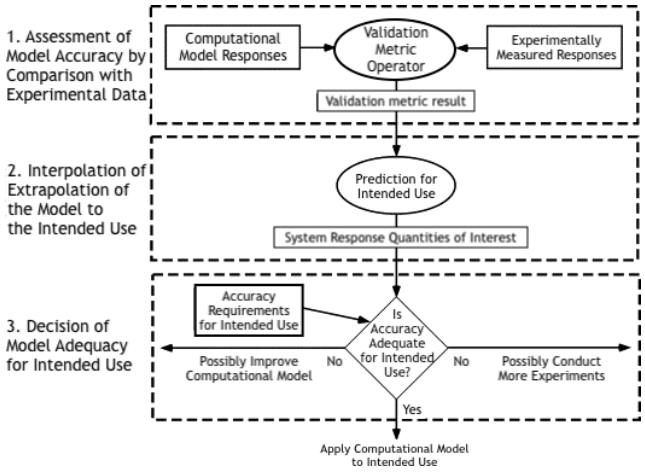


FIGURE 2. Validation workflow, adapted from [16].

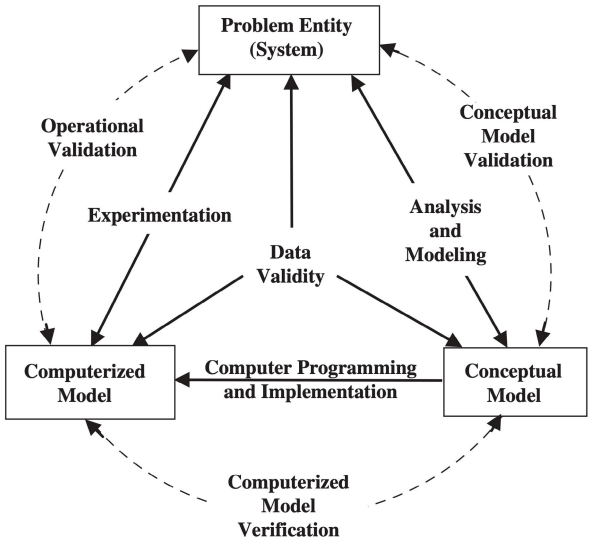


FIGURE 3. Validation workflow, as of [14].

the conceptual model, allow its computerization, and conduct a quantitative assessment of the truthfulness.

A. VALIDATION LIMITATIONS

The cited literature works highlight, however, inherent validation caveats. The first (and major) point is related to the *context-dependent* nature of the validity analysis: absolute validity is generally not achievable. Thus, a simulation model will always be an approximation of the real-world phenomena and will only serve the need of the specific application it aims at replicating.

The second point is the lack of globally defined user-acceptance criteria. Despite the outstanding cited contributions in terms of validation pipeline, in the real-world, the validation procedures are tailored to the specific application and Operative Design Domain (ODD). This limitation leads to a plethora of validation tools and corresponding compliance thresholds, some of which are presented in the following sections whenever available. Nevertheless, to the best of the authors' knowledge, an ultimate ''validation criterion'' for a complex virtual testing toolchain is not available.

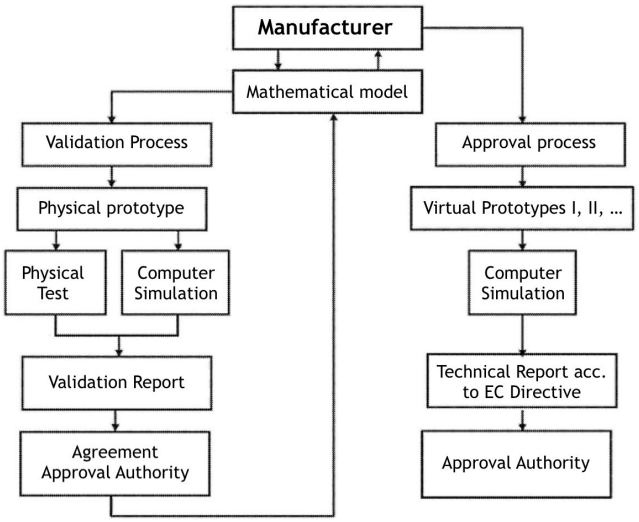


FIGURE 4. Type-approval based on virtual testing. EU regulation schematic process view, adapted from [19].

Eventually, on a practical level, obtaining validation-grade data might be very challenging, extremely costly, or even impossible (*e.g.*, when modeling a system not yet existing), thus limiting the applicability of the validation analysis.

In order to partially account for the limitations, a *credibility* analysis is typically carried out in parallel to the validation [17]. The credibility investigation aims at enhancing the confidence in the developed virtual toolchain [15]. A common framework for the credibility analysis of virtual models was developed at NASA and released publicly [18]. Similarly, a credibility framework is currently being discussed at the UN/ECE level[1] for ADS applications.

B. EXISTING REGULATIONS LEVERAGING ON VIRTUAL TESTING

Simulation models have recently started to be adopted in passenger vehicles regulations for type-approval purposes. For instance, the EU 2018/858 [19] regulating the type-approval of motor vehicle classes $M_{\{1,2,3\}}$, $N_{\{1,2,3\}}$, and $O_{\{1,2,3,4\}}$ according to the UNECE classification scheme,[2] allows the use of virtual testing for a list of regulatory acts which include superstructure's strength and interior visibility. Generic guidelines on how the model's validation should be carried out and its relationship to the approval process are also given and reported in the flowchart Fig. 4.

For what concerns ADAS/ADS, the UN/ECE R140 [20] which regulates the Electronic Stability Control (ESC) system, allows the ''sine with dwell'' maneuver to be carried

[1]https://wiki.unece.org/pages/viewpage.action?pageId=117508578
[2]https://unece.org/fileadmin/DAM/trans/main/wp29/wp29resolutions/ECE-TRANS-WP29-78-r4e.pdf

out in a simulation set-up provided that the corresponding virtual model accounts for a list of parameters and complies with a selection of Key Performance Indicators (KPIs). Similarly, the recently released Automated Lane Keeping System (ALKS) regulation [2] permits the usage of simulation to validate the system for a lane-keeping maneuver. However, neither of the documents gives any quantitative specification on how the validation should be carried out. Instead, only a qualitative ''comparability'' of the simulated and physical test results is demanded. Moreover, no requirements are given for the used simulation environment.

III. VALIDATION METHODOLOGIES

From Section II, a validation effort is typically a multi-step approach embracing several techniques. A common starting point for the validation pipeline is the *conceptual* validation activity. The next tier in the validation process requires assessing the simulation model's degree of discrepancy with respect to the real-world system. Eventually, the simulation model shall be investigated from an uncertainty analysis perspective. The following paragraphs provide extensive details on how the three steps can be fulfilled. Whenever possible, the literature examples provided are concerned with vehicle dynamics or ADS virtual testing to promote methods that have already found their way into the case of study.

A. CONCEPTUAL VALIDATION OF THE VIRTUAL REALIZATION

Before any quantitative/qualitative assessment of the simulation results is undertaken, the validation task has to prove that the correct system theories were implemented while *virtualizing* the real process and that the simulation model is capable of representing the target device's structure, input/output relationships, and functioning logic [14], [15], [21], [22]. The conceptual model validation is thus mainly concerned with checking the consistency of the modeling hypotheses and level of detail with the target objective the model aims to accomplish [17].

Additionally, the conceptual validation may involve studying the validity of the data used to develop the model as in Fig. 3. At this stage, an assessor might question whether the dataset used to build the simulation model was sufficiently informative to capture all the nuances of the real-world relevant for the target application.

Concerning vehicle dynamics, an early attempt to provide a conceptual validation framework was proposed in both [23] and [24]. The works delineated the appropriateness of the modeling approach depending on the target maneuver the virtual realization is requested to replicate. The effort of establishing suitability is, however, mostly demanded to experienced engineers, thus making the overall assessment largely subjective.

Since then, the conceptual validation gradually started to recede from the scientific discussions as the increased computational capability allowed more extensive virtual experimentations and validation over a larger domain using techniques described in Section III-B.

Nonetheless, the concept was brought to new attention with the advent of complex cyber-physical systems [25], [26] and agent-based simulations [27], such as the case of study. In an ADAS/ADS application, the overall system to virtualize can be functionally decomposed in layers (vertical decomposition) or submodels (horizontal decomposition). A typical functional organization of a virtual testing toolchain for ADS is shown in Fig. 10. A similar modeling abstraction framework was also presented in [28, Fig. 5] and [4, Fig. 3]. The conceptual validation of a submodels-based pipeline shall assess that each component is a reasonable virtualization of the corresponding physical counterpart for the given purpose. For example, the level of detail needed to model an automotive powertrain for a fuel-consumption application study is substantially finer than what is demanded by microscopic traffic simulation frameworks. On the other side, the modeling philosophy of the virtual sensors plays a critical role for ADS. Thus, they require suitable conceptualization methods to ensure the required fidelity level as summarized in Section IV-C1.

B. MODEL VALIDATION VIA RESPONSE ANALYSIS

After assessing the theories employed to develop the model, the validation shall determine that the *degree of discrepancy* between the simulated model and the physical realization is contained below a prescribed threshold level. Such a phase is typically accomplished via defining a selection of KPIs and the appropriate computational method to compare the recorded signals. Selecting the suitable list of signals is not a trivial task, nor is it supported by established literature for the specific case of ADS. In general, the list of variables to investigate should be large enough to cover the phenomena to model attempt to replicate with sufficient confidence.

Once the eligible list is determined, one has to apply a *computational tool* to effectively contrast the model's output to the physical system's response. Several options are offered to the designers, which are described in the continuation of the section.

1) GRAPHICAL COMPARISONS

Graphical comparisons provide an intuitive and straightforward way to compare the results. Typical ways of visually representing data include 2/3D plots, histograms, and scatter charts. They can alternatively be supported by animations [14] to graphically represent of evolution in time of the virtual system.

In addition, the visual inspection of the model's output is a first step to accomplish the *face validity* task pointed out by Carson. That is particularly effective when exploring the response of the model to a combination of inputs that are not easily reproducible with the real system [14].

The drawback with graphical comparisons is the lack of objective criteria to validate the modeling assumptions despite being a quite widespread method in the actual

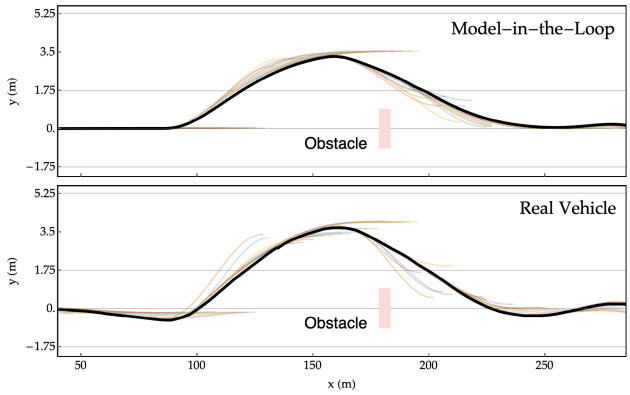


FIGURE 5. Graphical comparison of vehicle trajectories (black solid line), simulation (top) *vs.* proving ground (bottom), [31, Fig. 17].

KPI	EXPLANATION
Stops	number of times the car stopped in intersection
Time Stops	cumulative time of critical stops
Maximum Lateral Acceleration	Maximum lateral acceleration
Maximum Longitudinal Acceleration	Maximum longitudinal acceleration
Maximum Lateral Jerk	Maximum lateral jerk
Maximum Longitudinal Jerk	Maximum longitudinal jerk
Travel Time	time spent on intersection

FIGURE 6. List of scalar KPIs for a lane change application, adapted from [34, Table 3].

engineering practice [29]. The corroboration can nevertheless be carried out adopting Turing tests-like approaches [30], or by relying on Subject-Matter Experts (SMEs) [22], individuals who have a long experience in the discipline and are able to tell from the charts is the model shall be deemed valid.

An example of a virtual model validation using the graphical comparison of the recorded trajectories for an experimental automated vehicle is given in Fig. 5.

2) SCALAR QUANTITIES

Scalar quantities analysis affords the possibility of introducing a quantitative assessment despite being limited in terms of the provided information with respect to other methods described later.

Scalar quantities can derive from extracting a reference value from a time-history, for instance, by applying min or max operators. A widely established validation framework exploiting scalar quantities is the ISO standard 19364 [32], where the relative difference between the yaw-rate peaks (simulation Ω_{sim} *vs.* real-world Ω_{rw} experiment)

$$\frac{|\max \Omega_{\text{rw}} - \max \Omega_{\text{sim}}|}{\max \Omega_{\text{rw}}}, \tag{1}$$

is computed for a reference lane change maneuver and then compared to a threshold error. Eq. (1) is also known as the "Relative Error Criterion" (REC) and is adopted in other technical standards and fields, such as the EASA certification memorandum on structural mechanics [33] for a different set of KPI.

When studying higher-level automated driving functionalities, several scalar KPIs can be computed and clustered in a table such as in the summary of results of the European Project Enable-S3 [34], here adapted in Fig. 6. Additionally, Fig. 6 demonstrates how scalar metrics can account for high-level information, such as the number of times the vehicle stopped during the experiment and low-level information about the driving policy such as the maximum longitudinal jerk.

3) TIME-HISTORIES

Time-histories relationships give extensive information about the virtual model's fidelity but require extra care when carrying out the assessment with respect to the scalar quantities. In fact, the comparison of the M&S output with respect to the corresponding real-world experiment might need data conditioning procedures such as time synchronization and resampling.

A popular method for synchronization is the Time of Arrival (ToA) [35], [36] criterion. ToA is based on time-shifting the signals until the time-histories reach, for the first time, a reference amplitude. The time-occurrence when the reference amplitude is reached is compared for each signal making up the dataset and the computed time difference used to shift the signals in the time-domain.

Once the data are synchronized, a common method for validation is introducing an absolute or relative tolerance interval around the experimental reference data [37]. Tolerance intervals are representative of the measurement and model's uncertainty. For example, a validation assessment procedure can require the virtual model's output to stay within the $\pm 5\%$ amplitude interval of the corresponding real-world realization. A similar approach is adopted in the technical ISO standards 19364, 19365 [32], [38]. Alternatively, confidence levels can be used to define the validation intervals as displayed in Fig. 7 where the acceleration recorded on the proving ground (PG) is compared against multiple repetitions in a vehicle-hardware-in-the-loop (VeHIL) setup for a car-following study.

Time-histories can also be characterized in terms of their amplitude *distance* [36], [40], [41] via studying the residuals. Several options are available when assessing the magnitude discrepancy:

- residuals vector p-norms

$$\left(\sum_{i=1}^{N} \left| x_{\text{sim},i} - x_{\text{rw},i} \right|^{p} \right)^{\frac{1}{p}}, \tag{2}$$

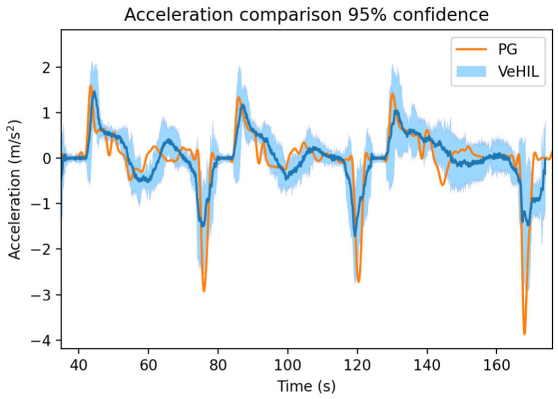


FIGURE 7. Example of relative tolerance intervals based on the standard deviation of the signal, [39, Fig. 8].

where N is total number of samples, $x_{\text{sim},i}$ is the generic KPI's time-sample deriving from the simulation environment, $x_{\text{rw},i}$ the corresponding real-world evidence, and p the order on norm. Typical solutions include the $\mathcal{L}_1$-norm ($p = 1$), $\mathcal{L}_2$-norm ($p = 2$) or $\mathcal{L}_{\text{inf}}$-norm ($p \to \text{inf}$);

- normalized vector norms according to N, such as the Mean Absolute Error (MAE)

$$\frac{1}{N}\sum_{i=1}^{N}|x_{\text{sim},i} - x_{\text{rw},i}|, \tag{3}$$

following the normalization of $\mathcal{L}_1$-norm. The normalization of the $\mathcal{L}_2$-norm yields instead the Root Means Square Error (RMSE)

$$\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(x_{\text{sim},i} - x_{\text{rw},i}\right)^2}; \tag{4}$$

- normalized distances, such as the Theil's inequality coefficient (TIC) [42]

$$\frac{\sqrt{\sum_{i=1}^{N}\left(x_{\text{sim},i} - x_{\text{rw},i}\right)^2}}{\sqrt{\sum_{i=1}^{N}x_{\text{sim},i}^2} + \sqrt{\sum_{i=1}^{N}x_{\text{rw},i}^2}}, \tag{5}$$

which lies between 0 (perfect agreement) and 1 (total incompatibility) and enables setting dimensionless validation thresholds.

The distance analysis can be further informed by exploiting more advanced tools capable of decoupling the mismatch in magnitude and phase (hence, the shape of the signal). Such techniques are known as Magnitude Phase Composite (MPC) [35]–[37]. Among the MPC methods, a widely acknowledged solution is represented by the Sprague and Geers (S&G) criterion, which combines the phase discrepancy information:

$$d_P = \frac{1}{\pi}\arccos\left(\frac{\sum_i^N x_{\text{sim},i}\,x_{\text{rw},i}}{\sqrt{\sum_i^N x_{\text{sim},i}^2\sum_i^N x_{\text{rw},i}^2}}\right), \tag{6}$$

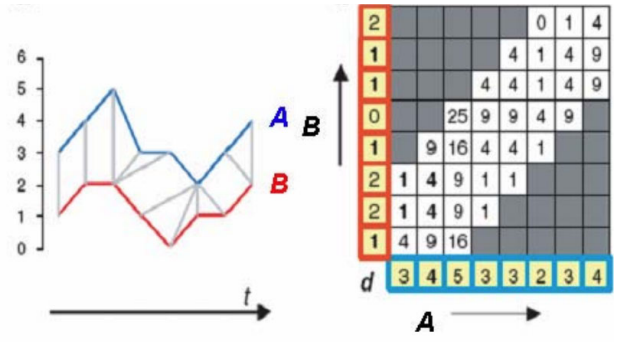


FIGURE 8. DTW graphical example with obtained warping path, [40].

and (integral) magnitude error

$$d_M = \sqrt{\frac{\sum_i^N x_{\text{sim},i}^2}{\sum_i^N x_{\text{rw},i}^2}} - 1, \tag{7}$$

into a "comprehensive factor"

$$d_{SG} = \sqrt{d_M^2 + d_P^2}. \tag{8}$$

A complementary solution to the S&G metric is Dynamic Time Warping (DTW) [43]. DTW provides a method to assess independently errors in phase, magnitude, and topology by aligning patterns (picks and valleys as in Fig. 8) in data through time axis scaling.

DTW-based validation metrics have compared favorably against SMEs' judgment in the field of simulation models for predicting the acceleration the human body is subjected to during a car accident in [40]. The proposed method, labeled Error Assessment of Response Time Histories (EARTH), exploits DTW to assess the magnitude, phase, and topology discrepancies independently. An enhanced version of EARTH (EEARTH) was recently developed, which uses linear regression to combine the three separate discrepancy components into a unique global error [44].

4) CORRELATION

Time-histories distance analysis can also be supported by the computation of the Pearson correlation coefficient $r(x_{\text{sim}}, x_{\text{rw}})$

$$\frac{\sum_{i=1}^{N}\left(x_{\text{sim},i} - \bar{x}_{\text{sim}}\right)\left(x_{\text{rw},i} - \bar{x}_{\text{rw}}\right)}{\sqrt{\sum_{i=1}^{N}\left(x_{\text{sim},i} - \bar{x}_{\text{sim}}\right)^2\sum_{i=1}^{N}\left(x_{\text{rw},i} - \bar{x}_{\text{rw}}\right)^2}}, \tag{9}$$

to detect the extent to which the model stays in a linear relationship with the real data. The squared value of r originates the coefficient of determination R^2, which is representative of the amount of variance predicted by the model. Both r and R^2 are widely adopted tools to assess the validity of the model as of several references [31], [39], [45].

5) FREQUENCY DOMAIN

Frequency domain approaches do not require time synchronization, and they are particularly suitable for the validation

of physics-based inspired models such as vehicle dynamics [46], [47]. Frequency analysis can help to disentangle the noise contribution to the actual true signal in evaluating the discrepancy between a real and a simulated dataset since the noise will mostly appear in the high-frequency region, whereas the true signal will be concentrated in the lower frequency area. Using frequency domain techniques, the distance analysis could thus account for the discrepancy affecting only a share of interest of the frequency spectrum. Fig. 13 depicts, for instance, the frequency responses of two virtual models for lateral vehicle dynamics with respect to the experimental transfer function.

The power spectrum of the residual is also particularly important to gain insights into the robustness of the calibrated models. Residuals mostly concentrated in the high-frequency domain are indicative of a robust calibration procedure which yielded a model grasping the real dynamics of the system without overfitting the noise (high-frequency) components [48].

6) STATISTICAL TESTING

Ultimately, the validation procedure shall enable an assessor to answer the question: ''is the simulation model an accurate representation of the physical phenomenon for the intended purpose?''. That is exactly the formulation of the null hypothesis (H_0) of a *hypothesis testing* [49], [50] problem. Conversely, the alternative hypothesis H_a reads instead as ''the simulation model is *not* an accurate representation of the physical phenomenon for the intended purpose''. The goal of validation can thus be seen as avoiding both Type I errors, *i.e.*, rejecting valid virtual models (model's builder risk), and Type II error, *i.e.*, accepting an invalid simulation model (model's user risk) [21]. The probability of Type I error equals the significance level α assumed in the hypothesis testing, where $\alpha = 0.5$ is a common assumption.

Statistical methods might also be particularly suitable to handle data generated by non-deterministic or hybrid (more on this in Section IV-A) virtual environments. Taking into account distribution functions allows, for instance, considering simulation stochasticity with no need of averaging over the repetitions, hence preserving a higher amount of information. Such an approach was exploited in [51] to derive validation metrics for a lane-keeping application.

Several works investigated which statistical tool to exploit for the sake of model validation, among them [13], [49], [52]–[56]. The cited contributions suggest the following methods as viable candidate tools to assess the validity of a model:

T-test: consists of checking whether two distributions have the same means (two-sample testing) or whether a population mean differs significantly from a sample;

F-test: similarly to the T-test, the F-test examines the consistency of the variances;

Kolmogorov-Smirnov test: measures the vertical distance between two Cumulative Distribution Functions (CDFs);

Anderson-Darling test: evaluates if a sample comes from a population characterized by a specific distribution [57].

A downside of statistical model validation is that statistical testing is mainly concerned with stationary distributions [58], which limits its effectiveness in investigating the transients behavior of the system. For example, one can straightforwardly utilize the T-test to validate a finite element method (FEM) model for a static structural analysis against experimental evidence. The same is not true in the case of the velocity profile of a vehicle dynamics simulation model, which typically exhibits time-varying components. Furthermore, handling transients requires the adoption of aggregation operators, which will inevitably drop relevant information about the signals [39]. Devising aggregation operators capable of retaining the amount of signal information necessary for the sake of validation is still, to the best of the authors' knowledge, an open point in the scientific literature that shall be addressed in the upcoming years.

7) STATISTICAL DISCREPANCY

Whenever multiple repetitions of the same driving scenario subjected to stochasticity are available (for either the simulated models and/or for the real-world), statistical tools can be exploited to appraise the distance between the generated distributions.

The statistical discrepancy can be computed exploiting one of the following tools:

z-score and its corresponding multidimensional generalization Mahalanobis distance [59]

$$\sqrt{(\mathbf{x}_{\text{rw/sim}} - \boldsymbol{\mu}_{\text{sim/rw}})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_{\text{rw/sim}} - \boldsymbol{\mu}_{\text{sim/rw}})}, \quad (10)$$

which assess how many standard deviations a point stands apart from a given distribution given the Σ covariance matrix;

Kullback-Leibler divergence (D_{KL}) measures the distance between two distributions [60], a method that can be helpful to analyze stochastic models such as probabilistic models for traffic behaviors [56].

C. SENSITIVITY ANALYSIS & UNCERTAINTY QUANTIFICATION

Beyond the model accuracy discussed in Section III-B, the validation should quantify the expected uncertainty from the model as resulting from input data error and modeling approximations. The combination of techniques capable of explicitly addressing the uncertainty content of a simulation model on top of the mentioned methods for validation in Section III-B, yields the so-called Verification, Validation and Uncertainty Quantification (VV&UQ) methodologies [61]. VV&UQ methods aim at improving conventional validation activities by specifically addressing some of the critical aspects associated with validation practices based on correlation thresholds only. In particular, VV&UQ allows:

1) accounting for the uncertainty in the calibration/validation dataset, which is especially critical for

applications where direct measurements are difficult or expensive to obtain;
2) surmounting the binary validation outcome, *i.e.*, the model can be either valid or invalid with an estimation of the *degree* of validity;
3) studying the extrapolation capabilities in the proximity of the validation domain.

1) SENSITIVITY ANALYSIS

Sensitivity analysis examines how small perturbations of the simulation model's input quantities, both in terms of the time-invariant model's parameters and time-varying model's inputs, propagate to the output. Such an analysis is particularly beneficial to support the validation activity. In fact, the sensitivity study affords the determination of the extent to which the model satisfies the validation thresholds when it is subjected to limited variations of the parameters. Thus, the robustness and the generalization capabilities of the simulation model can be established. As an additional outcome, the model's parameters which mostly contribute to the end result of the simulation can be identified. Potentially, the sensitivity analysis investigation of the virtual model output can be one of the few tools available to validate a model in the absence of real-world data together with the SMEs' judgment.

An overview of the sampling-based sensitivity analysis techniques for model validation can be found in [62]. The survey work identifies two main approaches: ''traditional'' techniques, which are best suited for a small number of parameters and operating points (also known as one-factor-at-a-time methods or first-order analysis), and ''modern'' techniques capable of supporting hundreds of parameters and their interaction through efficient sampling operations. An example of a traditional technique within the field of vehicle dynamics is proposed in [63] where a 20 parameters tire wear model is validated against real-world experiments and the individual individually parameters perturbed ($\pm 25\%$) to study the sensitivity. Overall, 60 model executions, which do not account for parameters' interdependence, are required using the traditional method for a relatively simple application compared to an ADS toolchain. Indeed, such an approach may turn out not to be a feasible option to globally (*i.e.*, first-order analysis plus interactions) assess complex ADS testing toolchain which might have thousands of parameters and efficient techniques shall be enforced such as Monte-Carlo or Latin Hypercube Sampling [64].

An alternative approach to sampling is to characterize a model based on differential equations using automatic differentiation [65] and first-order analysis given the higher computational efficiency of the method. An application of automatic differentiation is proposed in [66] for the sensitivity analysis of a complex multibody (18 degrees of freedom, DOFs) vehicle dynamics model.

The interdependency of parameters' variation was studied in [67] for a 9 DOFs, 12 parameters vehicle model using a two-steps approach. Firstly, the elementary effect of each parameter variation was investigated to reduce the model complexity. Secondly, a global sensitivity analysis (GSA) index was calculated using Sobol's method [68] based on the ANOVA theory. The GSA was found to be in agreement with the elementary analysis for most of the effects considered with expect of the vertical acceleration, which is underestimated for some elementary parameters' variations given the non-linear dynamics of the model. For the considered KPIs, the sprung mass and suspension damping coefficient turned out to be the most sensitive parameters. Moreover, the robustness of the study is further enhanced, given that two model parametrizations are analyzed with similar findings.

2) UNCERTAINTY ANALYSIS

Partially related to the sensitivity study is the assessment of the uncertainty. While the sensitivity analysis is mainly concerned with the establishing properties of the model in a local portion of the input space, the uncertainty examination explores the set of outcomes' distributions. Two sources of uncertainty are typically assumed for the simulation models: *aleatory* (random) components, *epistemic* (lack-of-knowledge) factors. Ultimately, examining how the model extrapolates for input perturbations provides sounder credibility to the virtual realization.

The uncertainty of simulation models can be established by specifying ranges for the model parameters as resulting from robust calibration procedures such as bootstrapping [69] or known aleatory properties of the modeled elements [70]. The virtual realization uncertainty can then be estimated by propagating through the model samples from the parameters' distributions. This can be done by exploiting Monte-Carlo techniques [71]. The Monte-Carlo method is based on executing deterministic simulations where, at every new simulation instantiation, a random set of parameters are drawn from the parameters' distributions. Such a methodology is graphically shown in Fig. 9. The uncertainty analysis's effectiveness depends on the total amount of virtual tests carried out: the higher the number of simulations performed, the more accurate the estimation of the output distributions. Convergence criteria based on the output variance have been proposed in [72].

From an industrial perspective, an exact quantification of the uncertainty is typically not pursued due to the high computational demand in executing a large number of simulations and collecting all the simulation model inputs' variance. Conversely, safety factors are adopted to investigate worst-case scenario parametrizations [33], [61]. This is also the case for the field of vehicle dynamics, where models are typically validated without resorting to an explicit formulation of the uncertainty [46]. Instead, a tolerance envelope is defined around the experimental data: only if the output is within the interval, the model is accepted.

Nonetheless, some recent scientific contributions dealing with vehicle models have started introducing uncertainty assessment [73]. For instance, in [74], the uncertainty of a simulation model for fuel consumption prediction is assessed

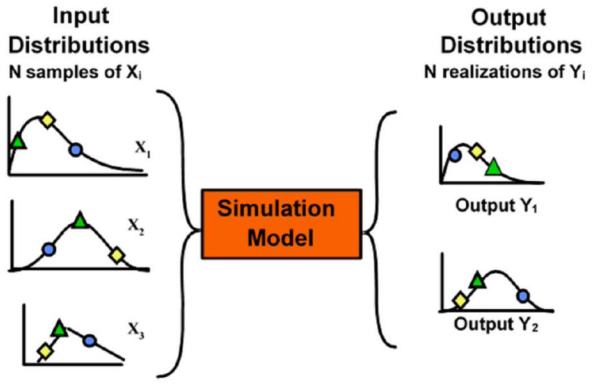


FIGURE 9. Uncertainty quantification through Monte-Carlo approaches, [33, Fig. 10].

via the propagation of input and parameters uncertainties following the approach visually represented in Fig. 9.

IV. VIRTUAL VALIDATION FOR ADS

ADS virtual testing can take advantage of a huge number of simulation tools that help the design and development of autonomous driving functionalities. However, the validation of the ADS technology is still largely based on physical testing, using either scenario-based approaches [75] or the number of miles before disengagements [76]. Given the discussed benefits of simulation in the introductory Section and the limitation of real-world testing alone in terms of scenario coverage and the unmanageable number of miles to be driven, it is in the authors' view that virtual testing will be a crucial pillar of future ADS certification.

This Section, thus, building up on the general concepts presented in Section II and on the specific methods presented in Section III, surveys the procedures empowered to support the virtual validation of ADS functionalities. The considerations here presented start from the underlying assumption that the validation of a virtual testing toolchain is pursued *regardless* of the specific ADS implementation. A virtual testing environment is here considered as the overall simulation framework that allows running a simulation once an ADS is plugged in. The validation of an ADS via virtual testing has first to demonstrate the validity of the toolchain involved, which is somehow implicitly done for the physical tests only.

From the literature analysis we carried out, we can split the validation of a virtual testing environment into two methodologies: ''integrated system'', where the overall simulation toolchain is tuned to replicate a distinct maneuver Section IV-B, and ''submodels-based'' Section IV-C, where each ingredient of the simulation pipeline is individually validated with respect to its physical counterpart. Hybrid solutions combining both integrated system and submodel-based validation in a cascade fashion are also foreseen as explained in [28]. However, no practical application of such an approach is available in the literature yet.

A. VIRTUAL ENVIRONMENTS FOR ADS SIMULATION

Before presenting the validation methodologies for ADS virtual testing, a brief explanation of the simulation environments' setups provides valuable insights for the later discussion. The flexibility of modern days simulation software enables the combination of hybrid simulation/real-hardware testing configurations known as 'X-in-the-Loop'' (XiL) approaches. A common classification in the scientific literature and industrial practice is to discriminate among:

Model-in-the-Loop: (MiL) full toolchain simulation on a general computing system;

Software-in-the-Loop: (SiL) full toolchain simulation using compiled code;

Hardware-in-the-Loop: (HiL) hybrid solution combining simulated models with real hardware components;

Vehicle-Hardware-in-the-Loop: (VEHiL) hybrid approach combining a real vehicle placed on a chassis dynamometer while environmental information is provided by either data injection or sensor stimulation;

Vehicle-in-the-Loop: (ViL) the vehicle can drive on a proving ground, however, virtual sensor information is still provided by the simulation environment via either signal injection or sensors stimulation.

Further details about testing using XiL-based approaches are outside the scope of the present paper and the reader may refer to the widely acknowledged survey works such as [77]–[79].

B. INTEGRATED SYSTEM VALIDATION

This set of validation methodologies are concerned with the definition of a reference maneuver and the tuning of a simulation environment to reproduce the driving task virtually. According to this methodology, the validation is carried out for a list of KPIs which are representative of the full (closed-loop) simulation environment and not of the models making up the toolchain, such as virtual sensors and virtual vehicle models. Additionally, these techniques are typically framed in a scenario-based approach [75].

From a legislation perspective, an approach following this philosophy is currently under discussion for the AEBS virtual test [80] based on an ad-hoc devised computational method. A similar method is already implemented in the UN/ECE R140 [20] for the Electronic Stability Control (ESC) type-approval, where generic guidelines are given on the simulation model's structure and which relevant KPIs to use for validation albeit the exact correlation threshold is not provided.

Several proposals are also found within the scientific literature, in particular in [45], [51], and [39]. In [45], the authors study the fidelity level that a Vehicle-Hardware-in-the-Loop (VeHIL) setup can deliver with respect to a MiL environment by comparing the real-world experiments against the evidence generated by simulation environments. Given the limitation of the VeHIL in not the allowing steering action, the chosen KPIs are representative of the longitudinal

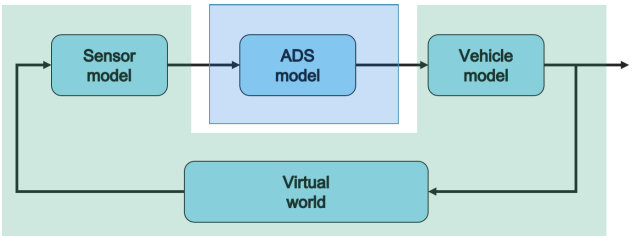


FIGURE 10. Sub-modules-based modeling framework for ADS simulation, [39, Fig. 1].

dynamics only. In particular, the relative vehicles spacing, ego vehicle velocity, and longitudinal acceleration time-histories are compared. The correlation is carried out by analyzing the mean values, standard deviations, and Pearson correlations. Overall, the VeHIL returned significantly better fidelity than the MiL. In [51], a method is suggested which firstly checks validation scenario coverage and then computes the correlation between the simulated and logged lateral acceleration for a lane-keeping application by also considering stochastic effects of the particular VeHIL environment. Similarly to [45], in [39] a VeHIL setup is investigated in terms of the achievable fidelity level for increasing complexity driving scenarios using camera stimulation rather than signal injection. The paper is complemented with statistical validation methods in addition to traditional discrepancy analysis techniques. The KPIs selection includes the ego vehicle velocity and longitudinal acceleration for which 95% confidence intervals are provided as resulting from the multiple repetitions carried out in the VeHIL environment.

On the positive side, this methodology does not foresee the validation of each simulation submodel making up the virtual testing environment, thus reducing the validation effort. The main focus is instead on providing high correlation with real-world data on a global level for the specific KPIs considered. On the downside, this philosophy provides little information on how the toolchain is going to extrapolate outside the validation domain, which affects the credibility of the developed solution [28].

C. VALIDATION OF VIRTUAL SUBMODELS

The validation of the testing environment can alternatively be carried out by validating each individual simulation model making up the toolchain. The methodology is based on the functional decomposition of the virtual testing toolchain into submodels that replicate the physical counterpart. In general, a widespread approach is to adopt the decomposition shown in Fig. 10, a common solution also depicted in [28, Fig. 5], [4, Fig. 3], and in the recently published ISO/TR document [81].

In Fig. 10, the toolchain is functionally divided into 4 main blocks: the sensors models, the vehicle model, the virtual world model and the ADS implementation which are discussed in the remaining of the Section.

1) SENSOR MODELS

A preliminary investigation of virtual *vs.* real sensors was qualitatively carried out in [82]. The work focused on presenting the simulation environment "Virtual Test Drive" (VTD) but also provided nuances on how to generate synthetic data and reusable (parametric) models for sensors. In the end, several charts are reported which compare the recorded position of a static traffic obstacle with respect to the same setup replicated by the virtual toolchain. The work does not provide acceptance thresholds nor validation methodologies that could be straightforwardly generalized to other case studies. However, it highlights the strong coupling that exists between models' fidelity level and the quality of the virtual environment-generated synthetic data.

In [83], the authors examine the modeling approaches for sensors from white-box modeling *i.e.*, based on the replication of the physical phenomena happening in the actual sensor device, to the black-box methods, which are concerned only in replicating the observed statistical I/O relationship. In addition, the authors propose a third-way grey-box modeling philosophy as a trade-off solution to combine the pros of white- and black-boxes approaches. Such a modeling approach defines a *functional* decomposition of the sensors' components rather than a physical reconstruction to allow a versatile and re-usable model parametrization.

The authors of [84] focus on the virtual validation for sensors and propose a two-steps approach where the virtual model is initially assessed in terms of a direct comparison with the real sensor's output. Secondly, the next tier in the ADS chain is fed the synthetic data, and its output is compared against the outcome originating from the real input. The authors' claim is that the direct comparison of the sensors' models is necessary but not sufficient as sensors are highly coupled with the consecutive layers in the ADS pipeline.

More recently, in [85], the lack of experience and clearly defined requirements with respect to other modeling fields (as vehicle dynamics for instance) is highlighted in addition to the difficulty in replicating trajectories in the simulation environment with a precision that matches the fidelity required by the sensor validation. The scientific effort is completed via repeating the metrics calculation in [86] for a different scenario and obtaining similar scores.

Overall, designers have several opportunities to model sensors despite the overall classification is still under discussion [83], [85]–[87]:

low-fidelity: also known as "black-box" models or "object list". Low-fidelity models retrieve the traffic objects' list and status directly from the simulation environment kernel. This modeling paradigm does not afford statistical aspects related to the perception, such as false positives/negatives rate. However, low fidelity models might account for basic sensor effects such as the Field of View (FoV) and occlusions to filter the whole object list;

medium-fidelity: also known as "grey-box" [83] or "phenomenological/data-driven" according to the classification presented in [85]. They share the basic working principle of low-fidelity models. Nonetheless,

they introduce the possibility of modeling false positives/false negatives rates, the effect of traffic objects' shape and texture on the detection, and environmental effects such as atmospheric degradation;

high-fidelity: also known as "white-box" or "physics-based". High-fidelity models try to replicate the physical phenomena regulating the interaction between the sensor and the external environment in simulation. Typically, advanced computer graphics techniques are adopted, such as ray-tracing and rasterization to render the 3D simulation environment.

Strongly linked to the modeling philosophy is the information level at which carrying out the calculation of the KPIs. The viable opportunities depend on modeling abstraction and can be summarized as:

objects detection level: the highest level information provided by the sensors' models, *e.g.* class and size of the object. This option is the only available when adopting the lowest fidelity virtual sensor models which cannot provide pixel-level detailed information;

occupancy grid: (OG) intermediate level sensor information which refers to the probability of a pixel to be occupied by an obstacle;

raw data: lowest level sensor information extracted before any tracking/classification algorithm is employed. It needs static obstacles (*e.g.* buildings, fences, ...) to be digitized and included in the simulation environment for the pixel-level comparability of the results.

Considering the object detection level, popular metrics to compare the generated bounding boxes (BB) are:

- the Intersection over Union (IoU), also knonw as the Jaccard distance [88]:

$$\frac{\text{Area}\left(\text{BB}_{\text{rw}} \cap \text{BB}_{\text{sim}}\right)}{\text{Area}\left(\text{BB}_{\text{rw}} \cup \text{BB}_{\text{sim}}\right)}; \tag{11}$$

- the multiple object tracking accuracy (MOTA) [89]:

$$1 - \frac{\sum_i^N \left(\text{FN}_i + \text{FP}_i + \text{MM}_i\right)}{\sum_i^N n_{\text{obj},i}}, \tag{12}$$

where FN_i are the false negatives at time-step i, FP the false positive, MM the misdetections and n_{obj} the total number of objects.

Based on the OG, some of the computational tools exploited in the literature to quantitatively evaluate the fidelity level are:

- OG cell-loss:

$$\sum_{x_c=0}^{\text{width}} \sum_{y_c=0}^{\text{height}} || \, \text{OG}_{\text{sim}}\left(x_c, y_c\right) - \text{OG}_{\text{rw}}\left(x_c, y_c\right) \, ||,$$

where:

$$\text{OG}_{\text{sim}}(x_c, y_c) = \begin{cases} 1, & \text{if } P\left(\text{obstacle}\right) > 0.5 \\ 0, & \text{else} \end{cases}$$

$$\text{OG}_{\text{rw}}(x_c, y_c) = \begin{cases} 1, & \text{if } P\left(\text{obstacle}\right) > 0.5 \\ 0, & \text{else;} \end{cases} \tag{13}$$

- OG Pearson correlation:

$$\frac{\sum_i^{N_c} \left(\text{OG}_{\text{sim,i}} - \overline{\text{OG}}_{\text{sim}}\right)\left(\text{OG}_{\text{rw,i}} - \overline{\text{OG}}_{\text{rw}}\right)}{\sqrt{\sum_i^{N_c} \left(\text{OG}_{\text{sim,i}} - \overline{\text{OG}}_{\text{sim}}\right)^2 \sum_i^{N_c} \left(\text{OG}_{\text{rw,i}} - \overline{\text{OG}}_{\text{rw}}\right)^2}}, \tag{14}$$

where N_c is the total number of cells;

- the occupied cells ratio (OCR):

$$\frac{\sum_{x_c=0}^{\text{width}} \sum_{y_c=0}^{\text{height}} \text{OG}_{\text{sim}}\left(x_c, y_c\right)}{\sum_{x_c=0}^{\text{width}} \sum_{y_c=0}^{\text{height}} \text{OG}_{\text{rw}}\left(x_c, y_c\right)}. \tag{15}$$

Assuming the availability of the point cloud (PC) raw data, the following assessment criteria can be adopted:

- normalized minimum Euclidean distance between point clouds [90]:

$$\frac{1}{N} \sum_i^N \min \left\| \text{PC}_{\text{sim,i}} - \text{PC}_{\text{rw,i}} \right\|, \tag{16}$$

where N is the total number of rays per scan;

- PC Pearson correlation [84]:

$$\frac{\sum_i^N \left(\text{PC}_{\text{sim,i}} - \overline{\text{PC}}_{\text{sim}}\right)\left(\text{PC}_{\text{rw,i}} - \overline{\text{PC}}_{\text{rw}}\right)}{\sqrt{\sum_i^N \left(\text{PC}_{\text{sim,i}} - \overline{\text{PC}}_{\text{sim}}\right)^2 \sum_i^N \left(\text{PC}_{\text{rw,i}} - \overline{\text{PC}}_{\text{rw}}\right)^2}}, \tag{17}$$

- PC RMSE [90]:

$$\sqrt{\frac{\sum_i^N \left(\text{PC}_{\text{sim,i}} - \text{PC}_{\text{rw,i}}\right)^2}{N}}. \tag{18}$$

Finally, sensors might be investigated by means of:

explicit open-loop simulations (E-OL): where only the sensors' output is obtained via the re-simulation of a previously recorded driving scenario and the validation is carried out on the point cloud level (IF1 interface in Fig. 11);

implicit open-loop simulations (I-OL): where only the sensors' output is obtained via the re-simulation of a previously recorded driving scenario and the validation is carried out after clustering and tracking (IF3 interface in Fig. 11);

closed-loop simulations (CL): where the actual virtual sensors generated information is fed to the ADS and the ego-vehicle motion is part of the validation process similarly to the process described in Section IV-B

Ultimately, the validation of the sensor models is an open topic in the scientific literature and industrial practice. Currently, no realistic correlation thresholds have been established to accept the models, and, secondly, no consolidated modeling framework has been adopted.

a: RADAR

RADAR sensors are particularly demanding when it comes to accurately replicate in simulation their working principle [85]. There are, in fact, several non-trivial factors that affect the performance of RADARs which include: multi-path

TABLE 1. Validation metrics and correlation levels for RADARs virtual models.

Ref.	Metric	Method	Correlation level	Unit	Optimum
[90]	(11)	I-OL	0.346	(−)	1
	(16)	E-OL	0.152	(m)	0

reflections, interference, ambiguities, clutter, ghost objects, and attenuation [91]. On one side, the replication of such phenomena in the virtual world is better accomplished using white-box modeling approaches. On the other side, the solution of the governing equations of a RADAR in real-time is an extremely challenging demand [92]. Indeed, the solution of Maxwell's equations for ≈77 GHz electromagnetic radiations (automotive RADAR operative frequency) using the Finite Difference Finite Time (FDFT) method [93] is not a viable option for most commercial automotive simulation environments and modeling assumptions have to be introduced. The most widespread solution to partially replicate the physics behind the RADAR is to adopt the *ray tracing* framework. Ray tracing is still a computational intensive alternative and requires a detailed 3D representation of the obstacles' geometry. Nonetheless, ray tracing can capture reflection, diffraction, and ghosts objects [94].

Regardless of the modeling framework adopted for the RADARs, there exists no generally accepted evaluation criteria to validate a virtual sensor model, nor unified testing procedure [95]. Additionally, from a modeling perspective, there is little difference in the virtual models for RADARs with respect to LiDARs, given the similar working principles as they both rely on similar modeling techniques.

In [96], a sensitivity analysis is performed to determine the most critical aspects that contribute to realistic RADAR simulation. The analysis is carried out by comparing real sensor data *vs.* synthetic ones generated using the open-source CARLA simulator [97]. Among the studied factors, the RADAR Cross Section (RCS) model and long-distance obstacles were found to be the main responsible for the dissimilarity with respect to the experimental data. A similar work is proposed in [95] which highlights the multipath propagation, separability, and sensitivity to the RCS as the major modeling challenges.

In [98], a neural network is trained to predict whether a point cloud is real or simulated. In addition, the classifier's confidence is used as a metric to determine the fidelity degree with respect to state-of-the-art methods. In a follow-up work, [90], multiple RADAR sensor models are compared by means of quantitative metrics in an explicit open-loop and implicit open-loop manner. The correlation level computed for the highest fidelity model is reported in Table 1.

b: LiDAR

LiDAR sensors are based on a working principle comparable to RADARs. However, their modeling effectiveness using ray tracing is considerably higher with respect to RADARs thanks to the higher frequency range where they operate (infrared spectrum region), which reduces interference [83].

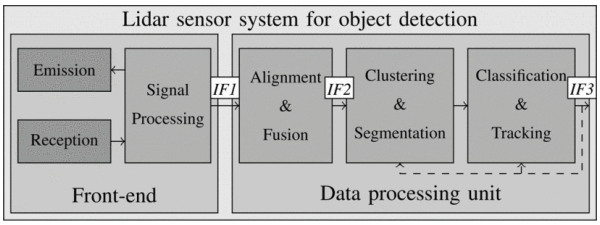


FIGURE 11. Functional decomposition of LiDAR sensors, [86].

Ray tracing can thus be safely exploited for white-box modeling approaches.

In [84], a ray tracing-based LiDAR model is studied and validated both explicitly and implicitly using Pearson correlation. Additionally, the authors investigated the robustness of the validation routine by manipulating the simulation scenario through the removal of a traffic vehicle present in the real-world and verifying the worsening of the validation KPIs.

In [99], the authors investigate LiDARs' sensitivity characteristics to determine which effects are prominent for the sake of defining requirements for LiDARs' virtual replication using high-fidelity models. The most critical aspects to reproduce were identified as the temporal scan order, noise figures, and the received signal's intensity. A first contribution comparing different LiDARs' models parametrization was published in 2019 in [86]. The work also proposes a functional decomposition that helps to standardize interfaces for metrics assessment as shown in Fig. 11.

The same work also proposes a validation method based on the occupied cells ratio (OCR) and point cloud distance criteria over 250 scans repetition. In [86], four high-fidelity sensor models are analyzed for the same set of inputs. OCR is shown to be dependent on the distance of the object, and in general poor performances were reported for the models. However, the contribution constitutes one of the first attempts to thoroughly compare different LiDAR modeling options given the same input and assessment criteria. The correlation levels obtained are reported in Table 2. From Table 2, a large variation of the fidelity is denoted where the highest correlation level corresponds to the portion of the validation scenario where the target obstacle is closer to the ego-vehicle.

A follow-up work, [100], a high-fidelity model is presented which is compatible with popular the popular Open Simulation Interface (OSI) [101] and the Functional Mock-up Interface (FMI). The model is then implicitly validated against both a real LiDAR's output and the ground-truth information derived from Real Time Kinematic (RTK) device by comparing the RMSE of the target vehicle trajectory.

c: CAMERA

Differently from RADARs and LiDARs, cameras are passive sensing devices. Therefore ad-hoc modeling techniques are required to craft high-fidelity models.

A first study involving the characterization of the fidelity level attainable with state-of-the-art camera models was proposed in [102]. The cited contribution introduces the modeling framework in Fig. 12, where the camera system is made

TABLE 2. Validation metrics and correlation level for LiDARs virtual models.

Ref.	Metric	Method	Correlation level	Unit	Optimum
[84]	(17)	E-OL	0.59	(−)	1
	(17)	I-OL	0.69	(−)	1
[101]	(17)	E-OL	0.824-0.832	(−)	1
	(17)	I-OL	0.677-0.703	(−)	1
[86]	(16)	E-OL	0.2 - 0.7	(m)	0
	(15)	E-OL	0.1 - 0.4	(−)	1
[100]	(4)	I-OL	0.014-0.139	(m)	0

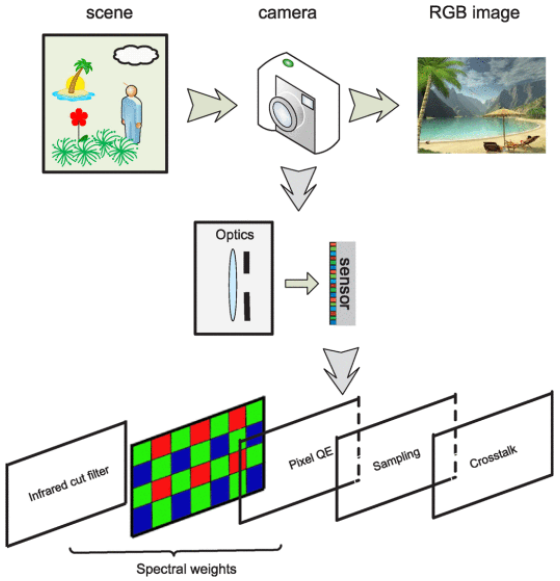


FIGURE 12. Camera model for simulation [102].

up of an optical module and a CCD sensor creating an RGB representation of the simulation engine rendered frame. The work details the implementation of the virtual optics and the virtual sensor and provides a preliminary validation activity using static reference scenarios.

Similarly to [102], in [103], the authors described the implementation of the main factors affecting cameras behavior before presenting an explicit validation of their realized camera model against three real camera systems. However, in [103], the focus is shifted for the first time towards ADAS/ADS applications.

In [104] a phenomenological camera model is presented and implicitly validated against real-world data and an ideal sensor model using a frequency-based approach. In [89] two camera models implemented in the simulation environments Vires VTD and IPG CarMaker are compared under different lighting conditions against both proving ground tests and the corresponding ground-truth. The authors of [89] defined several implicit validation metrics to ultimately compute an overarching index representative of the fidelity level obtained, the ''Simulation-to-Reality Gap''. In [89], weak points of the simulation environment under test are reported, such as the difficulties of replicating the real-world noise levels and the limited capabilities of reproducing the weather conditions, which affect how the Deep Neural Network (DNN)-based object tracking modules perform.

2) VEHICLE MODELS

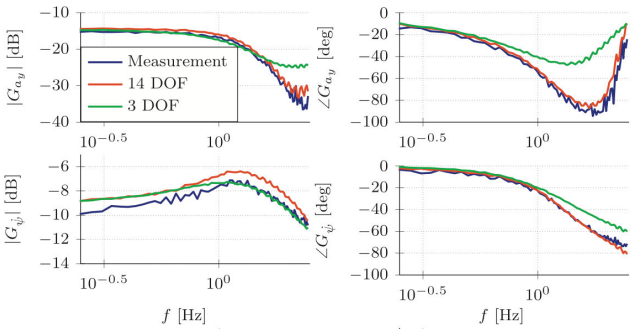
Modeling vehicle dynamics is an established activity supported by technical regulations and widely acknowledged scientific literature, including excellent textbooks.

Technical regulations such as [32], [38] provide specifications on how to virtualize and validate vehicle models for Electronic Stability Control (ESC) applications. Fidelity levels associated with the modeling approaches are also pointed out in the recent regulation [105].

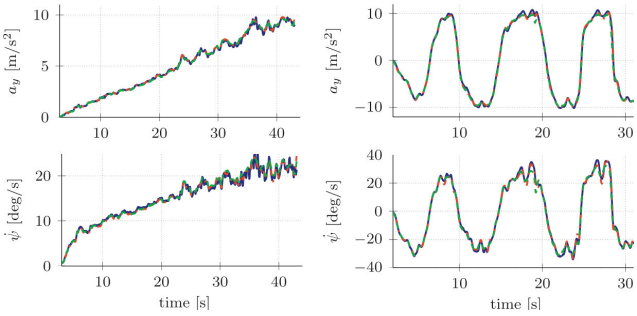
Within the scientific literature, a comprehensive review of modeling and validation methods for vehicle dynamics models is given in [46]. The work effectively summarizes the state-of-the-art contributions from the dawn of vehicle modeling up to contemporary scientific efforts. Two main applications for vehicle dynamics modeling are delineated: the crafting of artifacts for driving simulator platforms and the design of simulation models to support the development of vehicular technology. Concerning the first case-of-study, the topic has been widely discussed in [106] and goes beyond the scope of the present work given the importance that human feedback plays in validating the model for driving simulators, an aspect which is not relevant for an ADS application. Within the second class, models typically target a specific use case such as handling or riding studies, and each application is commonly backed by an ad-hoc devised validation procedure. Most of the works do not specify nor suggest correlation thresholds for validation. Instead, mainly subjective criteria based on a qualitative evaluation a selection of charts are reported. Moreover, statistical analysis and confidence intervals definition for the validity of the simulation models are only rarely found. On the other side, considerably broad literature is available to the end of defining maneuvers for the later characterization of the vehicle model's parameters, thus accomplishing the ''data validity'' objective suggested by Sargent [14]. The authors of [46], based on the extensive literature survey carried out, convey that a top-down approach should be established where:

1) the validation maneuvers have to be defined depending on the model requirements. In particular, the dataset should include both repeated steady-state and transient open-loop maneuvers with the aim of isolating the contribution of distinct vehicle parameters and building confidence intervals;
2) the validation procedures include both time-domain and frequency domain-metrics able to account for the uncertainty;
3) the validity domain of the simulation model is ultimately defined given no global validation is possible. For instance, a set of lateral accelerations and/or steering input frequency intervals can be used to limit the domain's validity.

In [47] two different models are validated with real-world trajectories to investigate the effect of order reduction on the fidelity delivered by the model. Interestingly, in the time domain metrics, no discernible discrepancy is observed



(a) Frequency domain.



(b) Time domain.

FIGURE 13. Effect of model reduction in the frequency domain (a), and time domain (b). Adapted from [47].

as shown in Fig. 13b. However, when comparing the transfer functions, an apparent deficit in replicating the high-frequency component can be detected for the low-order model as visible in Fig. 13a. Such a discrepancy is due to the missing poles and zeros of the low-order model which do not allow for replicating the real system's high-frequency behavior. Nonetheless, even a simple model proves to be faithful at reproducing the steady-state response and maneuvers which do not involve frequency components higher than 1 Hz.

More recently, in [107], a framework was proposed to introduce the model's uncertainty analysis into the validation activity. The foundation of the framework is that the candidate model shall satisfy the validation threshold for multiple vehicles (or different vehicle configurations) following the suitable calibration of the model. In other words, the authors try to decouple the validity of the model *structure* from the actual *parametrization* to isolate the model's inherent uncertainty and increase the credibility of the developed simulation framework. A graphical representation of the proposed solution is shown in Fig. 14.

Although the approach being particularly promising since it would allow solving some of the critical aspects of validation, namely the lack of sensitivity and uncertainty characterization in most of the validation procedures, only a formal description of the activity is presented, whereas an actual practical application is deferred to future work. Additionally, some aspects might still arise from the gradual introduction of ADAS/ADS technologies, given the different driving characteristics of artificial actuators with respect to

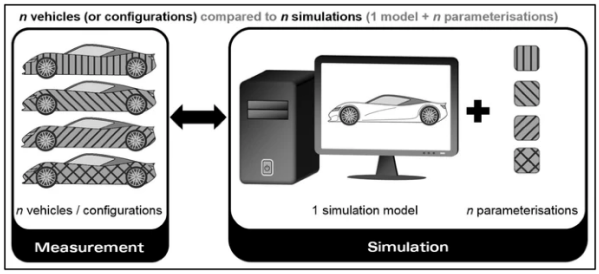


FIGURE 14. Viehof's proposed validation framework for vehicle dynamics models, [107].

human driving. For instance, the actuators' higher bandwidth might excite higher-frequency poles of the system [31] with respect to human driving.

Ultimately, in [108], the impact of different modeling approaches is investigated by comparing four classes of models having increasing complexity with real-world measurements on both icy and dry roads. It is reported that even the "simple" was able to predict surprisingly accurately the traveling velocity and path recorded on the dry road, whereas detailed modeling approaches were deemed necessary to predict transients at the beginning and the exit of curves, especially for the low-friction scenarios.

From the literature analyzed, we can summarize the modeling approaches in three classes based on the fidelity level that they can provide:

low-fidelity: point-mass or kinematic models. Mainly used for controller synthesis and for simulations where detailed vehicle modeling is not required such as microscopic traffic studies;

medium-fidelity: chassis models such as single-track, double-track, and lumped mass [47] with linear or nonlinear tires. Their range of application spans from the synthesis of model predictive controllers [31], [109] to intermediate fidelity simulations involving the testing and prototyping of ADAS/ADS functionalities;

high-fidelity: multibody models [110], [111] including suspensions geometry, chassis compliance, engine mounting stiffness and damping characteristics, driveline dynamics, and tires contact points to provide the ultimate degree of faithfulness.

Concerning VV&UQ methods, a survey addressing the estimation of the uncertainty content in vehicle dynamics simulation models was proposed [73] which identifies four key aspects originating the uncertainty:

1) the reconstruction of the input signals used for executing the virtual tests;
2) the input model's parameters aleatory and epistemic uncertainty;
3) the parametrization of the model;
4) the model's output.

The work summarizes scientific contributions also discussed in [46] from the perspective of the "degree of statistical validation" (DSV). It highlights the potential beneficial effects of the statistical validation concept applied to the field of

vehicle dynamics in addressing the mentioned limitations of conventional validation. Nonetheless, the novelty of the approach is emphasized by the fact that no literature work could be found that fulfilled all the uncertainty sources.

3) VIRTUAL WORLD MODELS

The ''virtual world'' is representative of a broad set of models which include any virtual entity enabling the closed-loop interaction of the ego-vehicle with the virtual environment. These include: virtual road, traffic agents and weather models.

a: ROAD LAYOUT

The virtualization of the roads' layouts is well supported by the widespread ASAM modeling standards:

- OpenCRG[3]: which standardizes local properties of the lane in terms, for instance, of asphalt granularity and potholes location;
- OpenDRIVE[4]: which regulates the definition of the road geometry in terms of number of lanes, curvature radius, lane marking and slope/camber;
- OpenSCENARIO[5]: institutes a standard way to model the behaviors of road users and traffic agents.

Although no explicit validation activities have been pursued for the listed modeling approaches, they are widely recognized as faithfully replicating road geometry for the sake of ADAS/ADS virtual testing in the actual simulation practice [4], [112].

b: TRAFFIC AGENTS

Traffic models involve external agents to the ego vehicle which participate and/or interact in the simulation. These consist of pedestrians [113], drivers [114]. The validation of traffic models is an open point in the literature given the intrinsic stochasticity of human behavior, which makes it fundamentally impossible to obtain ground-truth information. Additionally, high-fidelity traffic models might even not be necessary in the case of scenario-based approaches where targets' trajectories are assigned beforehand [115].

c: STATIC OBJECTS

These include buildings, guardrails, and any other object making up the virtual environment. Currently, no modeling standard exists which aims at standardizing the modeling approaches for this class of components. Replicating the optical and reflectance properties of static obstacles is particularly important in the case of physics-based sensor models as false positives/false negatives might arise due to the complex phenomena involved.

d: COLLISION MODELS

Whenever collisions occur, the rigid body assumptions underlying the traffic participants and traffic objects might not hold anymore. Thus collision models might be established and assigned to the objects acting the virtual test depending on the use-cases. Based on [81], several approaches can be pursued:

low-fidelity: using relative velocity and heading angle to directly estimate damage of the collision;

medium-fidelity: use the involved agent's kinematic to determine acceleration and classify the damage consequently;

high-fidelity: use FEM to precisely compute forces exerted during the collision.

e: WEATHER

Adverse weather conditions are known to blur and darken cameras' output [116]; absorb and scatter RADARs' pulses [117]; back-scatter and reduce surface reflectivity in the case of LiDARs [118]. Replicating such complex phenomena in the virtual environment is, at the moment of writing, an extremely challenging task due to the need of providing *sensor-grade* realism. Indeed, despite widespread metrics exist to evaluate the quality of generated images, such as the structural similarity (SSIM) [119], the actual focus is on the human perception and a dual solution capable of grasping sensors' perception properties is still missing.

In [120], a model-based approach to introduce sensor-specific rain effect via the post-processing manipulation of the rendered frames is presented. The approach proposed in [120] uses ray tracing to model rain-induced noise for virtual cameras and an equivalent RCS for RADAR sensors. The preliminary validation analysis performed demonstrates that the proposed methodology enables achieving a higher fidelity level for all the sensors' types.

4) OTHER SOFTWARE COMPONENTS & VERIFICATION

The sub-components division proposed overarches most of the ingredients making up the virtual toolchain. There are, however, additional tools that need to be introduced in order to carry out the virtual test. Among them: time steps, solvers, and coupling algorithms beyond the craftsmanship required to set up a simulation or a co-simulation framework.

Despite the adoption of validated sub-components, the overall virtual test's outcome might be far from reality due to integration and software implementation issues which shall be addressed using software verification techniques. Namely [121],

code verification: concerned with the execution of test demonstrating that no numerical/logical flaws affect the virtual models;

calculation verification: deals with the estimation of numerical errors affecting the M&S toolchain.

Code and calculation verification methods are a well-established practice in many engineering applications. For instance, in FEM/CFD (computational fluid dynamics) simulations, the spatial and time discretization are typically iterated until the result converges to an equilibrium value [122]. Nonetheless, the same concept is commonly

[3]https://www.asam.net/standards/detail/opencrg/
[4]https://www.asam.net/standards/detail/opendrive/
[5]https://www.asam.net/standards/detail/openscenario/

overlooked for ADAS/ADS simulations, where time-step is rarely discussed.

V. DISCUSSION

Despite the substantial scientific effort in the last years concerning the definition of validation methodologies for simulation models, several open research questions remain unsolved.

Integrated testing procedures are exceptionally delicate to handle as the ADS agent is interacting in a closed-loop fashion with other traffic participants within the virtual environment. Such a dynamic multi-agent modeling framework is indeed noticeably different from a traditional engineering simulation applications where the validation of the virtual submodels would ensure the accuracy of the overall output [123]. Moreover, most of the validation frameworks are conceived for open-loop tests where the model under analysis shows limited or null closed-loop interaction with internal state variables of the environment, such as in FEM/CFD simulation. That is not the case for ADS (especially for high-automation level technology), where the driving agents are capable of reasoning and producing emergent behaviors based on the sensor data. From a sensitivity perspective, it is also particularly hard to judge how small deficiencies in reconstructing the virtual scenarios propagate throughout the toolchain given the role played by the ADS, which can either dampen or amplify the discrepancies [39], [124]. Unfortunately, to the best of the authors' knowledge, this aspect is rarely discussed given the extremely high computational cost that a sensitivity study over such a large parameter space would imply.

Considering vehicle dynamics, most of the validation procedures are tailored to judge the validity of replicating a specific dynamical effect, *e.g.* vertical dynamics, comfort or handling. This limitation results from the narrow application vehicle dynamical models have played in the technical regulations and scientific literature in comparison to what is expected from an ADS application where the full vehicle is simulated, albeit with a degree of fidelity which is yet to be established. Nonetheless, vehicle dynamics models benefit from the support of conceptual model validity techniques [24], statistical testing [74], and sensitivity analysis methodologies [67] which increase the credibility of the developed model with respect to other simulation modules contributing to the ADS virtual tests.

Sensor models, on the other side, are still immature. Currently, no standard modeling framework has been adopted (only an ''informal'' classification is available as in Section IV-C1), no validation framework has been standardized both in terms of KPIs and via setting realistic correlation thresholds. This lack of standardization results in disconnected modeling validation methodologies, which complicate the comparison of virtual sensor models fidelity. Some specific aspects are reported to be particularly critical to replicate, which include noise figures [89], RCS, and real-world weather parametrization/replication capabilities of the simulation environments. That is mainly true because human-eye realism is commonly pursued with simulation environments with respect to the *sensor-grade* realism that should be pursued.

Eventually, the survey work has emphasized the infancy of the validation approaches adopted for validation ADS-related models: simple techniques and limited statistical assessment characterize most of the approaches presented in Section IV with respect to the state-of-the-art methods in Section III. This is also unsurprisingly accentuated by the novelty of the publications cited in our work. Most of the references are indeed a few years old, which contrasts with the widely recognized ASME/AIAA validation standards for FEM/CFD applications grounded on decades of utilization.

VI. CONCLUSION

This work surveyed the most relevant contributions supporting the virtual validation of simulation models for ADS certification. The scientific effort pursued is agnostic to any ADS structure and functionality and was primarily aimed at establishing procedures to assess how a virtual testing toolchain could be awarded the ''virtual proving ground'' trait for the sake of ADS certification.

Our contribution started by summarizing the state-of-the-art traditional simulation models validation to build up a benchmark of validation techniques across different engineering fields. We clustered the methodologies in three classes in Section III: conceptual validation, validation via response analysis, and sensitivity & uncertainty. We spent particular effort on the quantitative methods within the response analysis family, given the large variety of computational tools available and their most widespread adoption as validation tools both in the scientific community and in technical regulations. We found conceptual analysis to have been largely superseded by other validation tools for the field of ADAS/ADS. On the other side, VV&UQ approaches are still in an early stage of development. Nonetheless, they have huge potential to increase the overall M&S credibility, especially for complex simulation toolchains whose robustness is hard to establish via conventional threshold-based validation criteria.

We then investigated the literature contributions dealing specifically with simulation models for ADS virtual testing after briefly presenting what is meant with ''simulation environment'' in Section IV. We found that two main approaches can be outlined: integrated tests and submodels-based solutions. For the integrated test category, we analyzed the most relevant scientific efforts, and we outlined the authors' selection of KPIs and validation methodologies. Similarly, for the submodel category, we proposed a modeling framework in Fig. 10, and we analyzed the literature concerning virtual sensors, virtual vehicles, and virtual world models with their corresponding validation strategy. Special emphasis was dedicated to the sensors' models validation given the novelty of the approaches proposed.

Eventually, we summarized the open challenges in Section V. Overall, the field of model validation for ADS virtual certification is an emerging technology with enormous potential but still relatively immature with respect to other simulation disciplines where widely acknowledged validation procedures have been established.

ACKNOWLEDGMENT

The authors are grateful to the JRC colleagues Maria Cristina Galassi, Sandor Vass, Kostantinos Mattas for the exchanges within the project and to Barnaby Simkin, Tobias Duser, Gil Amid, Siddartha Khastgir, Espedito Rusciano, and all the experts participating to the VMAD SG2 working group on Virtual Testing and Simulation for the insightful discussions had on the subject. The work of Riccardo Donà has been carried out for the European Commission Joint Research Centre.

REFERENCES

[1] M. C. Galassi and A. Lagrange, "New approaches for automated vehicles certification," JRC, Ispra, Italy, Tech. Rep. 1, 2020. Accessed: Aug. 23, 2021. [Online]. Available: file:///Users/riccardodona/Desktop/new_approaches_for_automated_vehicles_certification.pdf

[2] United Nations Economic Commission for Europe, "Proposal for a new un regulation on uniform provisions concerning the approval of vehicles with regards to automated lane keeping system (ECE/TRANS/WP.29/2020/81)," Geneva, Switzerland, Tech. Rep. 1, 2020.

[3] N. Kalra and S. M. Paddock, "Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability?" *Transp. Res. A, Policy Pract.*, vol. 94, pp. 182–193, Dec. 2016.

[4] H. Abdellatif and C. Gnandt, "Use of simulation for the homologation of automated driving functions," *ATZelectronics Worldwide*, vol. 14, no. 12, pp. 68–71, Dec. 2019.

[5] OICA. (Mar. 2019). *OICA Views on the Certification of Automated/Autonomous Vehicle*. Accessed: Aug. 24, 2021. [Online]. Available: https://unece.org/DAM/trans/doc/2019/wp29/WP.29-177-20e.pdf

[6] S. Riedmaier, T. Ponn, D. Ludwig, B. Schick, and F. Diermeyer, "Survey on scenario-based safety assessment of automated vehicles," *IEEE Access*, vol. 8, pp. 87456–87477, 2020.

[7] E. Yurtsever, J. Lambert, A. Carballo, and K. Takeda, "A survey of autonomous driving: Common practices and emerging technologies," *IEEE Access*, vol. 8, pp. 58443–58469, 2020.

[8] T. Ersal, I. Kolmanovsky, N. Masoud, N. Ozay, J. Scruggs, R. Vasudevan, and G. Orosz, "Connected and automated road vehicles: State of the art and future challenges," *Vehicle Syst. Dyn.*, vol. 58, no. 5, pp. 672–704, May 2020.

[9] N. Rajabli, F. Flammini, R. Nardone, and V. Vittorini, "Software verification and validation of safe autonomous cars: A systematic literature review," *IEEE Access*, vol. 9, pp. 4797–4819, 2021.

[10] I. Babuska and J. T. Oden, "Verification and validation in computational engineering and science: Basic concepts," *Comput. Methods Appl. Mech. Eng.*, vol. 193, nos. 36–38, pp. 4057–4066, Sep. 2004.

[11] K. Forsberg and H. Mooz, *Proceedings of the First Annual NCOSE Conference*. NY, USA: Wiley, 2000.

[12] *Guide: Guide for the Verification and Validation of Computational Fluid Dynamics Simulations (AIAA G-077-1998 (2002))*, Comput. Fluid Dyn. Committee, VA, USA, 1998.

[13] D. Murray-Smith, "Methods for the external validation of continuous system simulation models: A review," *Math. Comput. Model. Dyn. Syst.*, vol. 4, no. 1, pp. 5–31, 1998.

[14] R. G. Sargent, "Verification and validation of simulation models," *J. Simul.*, vol. 7, no. 1, pp. 12–24, Feb. 2013.

[15] J. S. Carson, "Model verification and validation," in *Proc. Winter Simulation Conf.*, vol. 1, 2002, pp. 52–58.

[16] W. L. Oberkampf and T. G. Trucano, "Verification and validation benchmarks," *Nucl. Eng. Des.*, vol. 238, no. 3, pp. 716–743, Mar. 2008.

[17] S. Robinson, "Simulation model verification and validation: Increasing the users' confidence," in *Proc. 29th Conf. Winter Simulation (WSC)*, 1997, pp. 53–59.

[18] National Aeronautics and Space Administration, "NASA handbook for models and simulations: An implementation guide for NASA-STD-7009a," Washington, DC, USA, Tech. Rep. 1, 2019.

[19] European Union. (May 2018). *Regulation (EU) 2018/858 of the European Parliament and of the Council*. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32018R0858

[20] UNECE. (Jan. 2017). *Uniform Provisions Concerning the Approval of Passenger Cars With Regard to Electronic Stability Control (ESC) Systems*. [Online]. Available: https://unece.org/fileadmin/DAM/trans/main/wp29/wp29regs/2017/R140e.pdf

[21] R. G. Sargent, "Verification and validation of simulation models," in *Proc. Winter Simulation Conf.*, 2010, pp. 166–183.

[22] O. Balci, "Verification, validation, and certification of modeling and simulation applications," in *Proc. Winter Simulation Conf.*, vol. 1, 2003, pp. 150–158.

[23] R. W. Allen, T. J. Rosenthal, D. H. Klyde, K. J. Owens, and H. T. Szostak, "Validation of ground vehicle computer simulations developed for dynamics stability analysis," SAE Tech. Paper 920054, 1992.

[24] J. E. Bernard and C. Clover, "Validation of computer simulations of vehicle dynamics," *SAE Trans.*, vol. 103, pp. 162–170, Jan. 1994.

[25] J. Liu, Y. Yu, L. Zhang, and C. Nie, "An overview of conceptual model for simulation and its validation," *Proc. Eng.*, vol. 24, pp. 152–158, Jan. 2011.

[26] S. Robinson, "Conceptual modeling for simulation: Issues and research requirements," in *Proc. Winter Simulation Conf.*, Dec. 2006, pp. 792–800.

[27] B. Heath, R. Hill, and F. Ciarallo, "A survey of agent-based modeling practices (January 1998 to July 2008)," *J. Artif. Soc. Social Simul.*, vol. 12, no. 4, p. 9, 2009.

[28] T. Dueser and C. Gutenkuns, "A comprehensive approach for the validation of virtual testing toolchains," IAMTS, USA, Tech. Rep. 0001202104, Apr. 2021.

[29] Y. Ling and S. Mahadevan, "Quantitative model validation techniques: New insights," *Rel. Eng. Syst. Saf.*, vol. 111, pp. 217–231, Mar. 2013.

[30] G. S. Spring, "Validating expert system prototypes using the Turing test," *Transp. Res. C, Emerg. Technol.*, vol. 1, no. 4, pp. 293–301, Dec. 1993.

[31] M. Da Lio, R. Dona, G. P. R. Papini, F. Biral, and H. Svensson, "A mental simulation approach for learning neural-network predictive control (in self-driving cars)," *IEEE Access*, vol. 8, pp. 192041–192064, 2020.

[32] *Passenger Cars—Vehicle Dynamic Simulation and Validation—Steady-State Circular Driving Behaviour*, Standard ISO 19364, ISO, 2016.

[33] EASA. (Jul. 2020). *Proposed CM-S-014 Modelling & Simulation*. Accessed: Aug. 3, 2021. [Online]. Available: https://www.easa.europa.eu/sites/default/files/dfu/proposed_cm-s-014_modelling_simulation_-_for_consultation.pdf

[34] A. Leitner and M. Paulweber. (2019). *Enable-S3 Summary of Results*. [Online]. Available: https://drive.google.com/file/d/15c1Oe69dpvW5dma8-uS8hev17x-6V3zU/view

[35] M. H. Ray, M. Mongiardini, and C. Plaxico, "Quantitative methods for assessing similarity between computational results and full-scale crash tests," in *Proc. 91th Annu. Meeting Transp. Res. Board*, 2012, pp. 1–21.

[36] L. E. Schwer, "Validation metrics for response histories: Perspectives and case studies," *Eng. With Comput.*, vol. 23, no. 4, pp. 295–309, Oct. 2007.

[37] K. A. Maupin, L. P. Swiler, and N. W. Porter, "Validation metrics for deterministic and probabilistic data," *J. Verification, Validation Uncertainty Quantification*, vol. 3, no. 3, pp. 1–10, Sep. 2018.

[38] *Passenger Cars—Validation of Vehicle Dynamic Simulation—Sine With Dwell Stability Control Testing*, Standard ISO 19365, ISO, 2016.

[39] R. Dona, S. Vass, K. Mattas, M. C. Galassi, and B. Ciuffo, "Introducing virtual testing in ads certification. A validation example," *IEEE Trans. Intell. Vehicles*, to be published.

[40] H. Sarin, M. Kokkolaras, G. Hulbert, P. Papalambros, S. Barbat, and R.-J. Yang, "A comprehensive metric for comparing time histories in validation of simulation models with emphasis on vehicle safety applications," in *Proc. Int. Design Eng. Tech. Conf. Comput. Inf. Eng. Conf.*, vol. 43253, 2008, pp. 1275–1286.

[41] D. Ao, Z. Hu, and S. Mahadevan, "Dynamics model validation using time-domain metrics," *J. Verification, Validation Uncertainty Quantification*, vol. 2, no. 1, pp. 1–15, Mar. 2017.

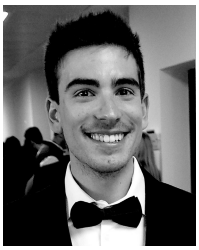
[42] H. Theil, *Economic Forecasts and Policy*. Amsterdam, The Netherlands: North-Holland, 1961.

[43] L. Rabiner, "Fundamentals of speech recognition," in *Fundamentals of Speech Recognition*. USA: Pearson College Div, 1993.

[44] Z. Zhan, Y. Fu, and R.-J. Yang, "Enhanced error assessment of response time histories (EEARTH) metric and calibration process," SAE Tech. Paper 2011-01-0245, 2011.

[45] S. Riedmaier, J. Nesensohn, C. Gutenkunst, T. Düser, B. Schick, and H. Abdellatif, "Validation of x-in-the-loop approaches for virtual homologation of automated driving functions," in *Proc. 11th Graz Symp. Virtual Vehicle*, 2018, pp. 1–12.
[46] E. Kutluay and H. Winner, "Validation of vehicle dynamics simulation models—A review," *Vehicle Syst. Dyn.*, vol. 52, no. 2, pp. 186–200, Feb. 2014.
[47] K.-U. Henning and O. Sawodny, "Vehicle dynamics modelling and validation for online applications and controller synthesis," *Mechatronics*, vol. 39, pp. 113–126, Nov. 2016.
[48] A. Papoulis and H. Saunders, *Probability, Random Variables and Stochastic Processes*. U.K.: McGraw-Hill, 1989.
[49] R. Rebba, S. Huang, Y. Liu, and S. Mahadevan, "Statistical validation of simulation models," *Int. J. Mater. Product Technol.*, vol. 25, nos. 1–3, pp. 164–181, 2006.
[50] W. L. Oberkampf and M. F. Barone, "Measures of agreement between computation and experiment: Validation metrics," *J. Comput. Phys.*, vol. 217, no. 1, pp. 5–36, Sep. 2006.
[51] S. Riedmaier, D. Schneider, D. Watzenig, F. Diermeyer, and B. Schick, "Model validation and scenario selection for virtual-based homologation of automated vehicles," *Appl. Sci.*, vol. 11, no. 1, p. 35, Dec. 2020.
[52] J. P. C. Kleijnen, "Statistical validation of simulation models," *Eur. J. Oper. Res.*, vol. 87, no. 1, pp. 21–34, Nov. 1995.
[53] J. P. C. Kleijen, "Validation of models: Statistical techniques and data availability," in *Proc. 31st Winter Simulation Conf., Simulation Bridge Future*, vol. 1, 1999, pp. 647–654.
[54] Y. Liu, W. Chen, P. Arendt, and H.-Z. Huang, "Toward a better understanding of model validation metrics," *J. Mech. Des.*, vol. 133, no. 7, pp. 1–13, Jul. 2011.
[55] D. Lamb, M. Castanier, H. Pan, M. Kokkolaras, and G. Hulbert, "Model validation for simulations of vehicle systems," Dept. Mech. Eng., Michigan Univ., Ann Arbor, MI, USA, Tech. Rep. ADA566037, 2012.
[56] T. A. Wheeler, P. Robbel, and M. J. Kochenderfer, "Analysis of microscopic behavior models for probabilistic modeling of driver behavior," in *Proc. IEEE 19th Int. Conf. Intell. Transp. Syst. (ITSC)*, Nov. 2016, pp. 1604–1609.
[57] T. W. Anderson and D. A. Darling, "Asymptotic theory of certain 'goodness of fit' criteria based on stochastic processes," *Ann. Math. Statist.*, vol. 23, pp. 193–212, Jun. 1952.
[58] G. Lee, W. Kim, H. Oh, B. D. Youn, and N. H. Kim, "Review of statistical model calibration and validation—From the perspective of uncertainty structures," *Struct. Multidisciplinary Optim.*, vol. 60, no. 4, pp. 1–26, 2019.
[59] P. C. Mahalanobis, "On the generalised distance in statistics," *Proc. Nat. Inst. Sci. India*, vol. 2, no. 1, pp. 49–55, 1936.
[60] S. Kullback, *Information Theory and Statistics*. Chelmsford, MA, USA: Courier Corporation, 1997.
[61] S. Riedmaier, B. Danquah, B. Schick, and F. Diermeyer, "Unified framework and survey for model verification, validation and uncertainty quantification," *Arch. Comput. Methods Eng.*, vol. 28, pp. 1–34, Aug. 2020.
[62] J. P. C. Kleijnen, "An overview of the design and analysis of simulation experiments for sensitivity analysis," *Eur. J. Oper. Res.*, vol. 164, no. 2, pp. 287–300, Jul. 2005.
[63] F. Braghin, F. Cheli, S. Melzi, and F. Resta, "Tyre wear model: Validation and sensitivity analysis," *Meccanica*, vol. 41, no. 2, pp. 143–156, Apr. 2006.
[64] M. D. McKay, R. J. Beckman, and W. J. Conover, "A comparison of three methods for selecting values of input variables in the analysis of output from a computer code," *Technometrics*, vol. 42, no. 1, pp. 55–61, Feb. 2000.
[65] H. C. Frey and S. R. Patil, "Identification and review of sensitivity analysis methods," *Risk Anal.*, vol. 22, no. 3, pp. 553–578, Jun. 2002.
[66] A. Callejo and D. Dopico, "Direct sensitivity analysis of multibody systems: A vehicle dynamics benchmark," *J. Comput. Nonlinear Dyn.*, vol. 14, no. 2, pp. 1–9, Feb. 2019.
[67] Y. Qin, Z. Wang, C. Xiang, M. Dong, C. Hu, and R. Wang, "A novel global sensitivity analysis on the observation accuracy of the coupled vehicle model," *Vehicle Syst. Dyn.*, vol. 57, no. 10, pp. 1445–1466, Oct. 2019.
[68] I. M. Sobol, "On sensitivity estimation for nonlinear mathematical models," *Matematicheskoe modelirovanie*, vol. 2, no. 1, pp. 112–118, 1990.
[69] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. Boca Raton, FL, USA: CRC Press, 1994.
[70] N. R. Council, *Assessing the Reliability of Complex Models: Mathematical and Statistical Foundations of Verification, Validation, and Uncertainty Quantification*. Washington, DC, USA: National Academies Press, 2012.
[71] H. Janssen, "Monte-Carlo based uncertainty analysis: Sampling efficiency and sampling convergence," *Rel. Eng. Syst. Saf.*, vol. 109, pp. 123–132, Jan. 2013.
[72] M. J. Gilman, "A brief survey of stopping rules in Monte Carlo simulations," Inst. Elect. Electron. Eng., Piscataway, NJ, USA, Tech. Rep. 1, 1968.
[73] B. Danquah, S. Riedmaier, and M. Lienkamp, "Potential of statistical model verification, validation and uncertainty quantification in automotive vehicle dynamics simulations: A review," *Vehicle Syst. Dyn.*, pp. 1–30, Dec. 2020.
[74] B. Danquah, S. Riedmaier, J. Rühm, S. Kalt, and M. Lienkamp, "Statistical model verification and validation concept in automotive vehicle design," *Proc. CIRP*, vol. 91, pp. 261–270, Jan. 2020.
[75] T. Menzel, G. Bagschik, and M. Maurer, "Scenarios for development, test and validation of automated vehicles," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2018, pp. 1821–1827.
[76] F. Favarò, S. Eurich, and N. Nader, "Autonomous vehicles' disengagements: Trends, triggers, and regulatory limitations," *Accident Anal. Prevention*, vol. 110, pp. 136–148, Jan. 2018.
[77] W. Huang, K. Wang, Y. Lv, and F. Zhu, "Autonomous vehicles testing methods review," in *Proc. IEEE 19th Int. Conf. Intell. Transp. Syst. (ITSC)*, Nov. 2016, pp. 163–168.
[78] C. Ebert and M. Weyrich, "Validation of autonomous systems," *IEEE Softw.*, vol. 36, no. 5, pp. 15–23, Aug. 2019.
[79] F. Rosique, P. J. Navarro, C. Fernández, and A. Padilla, "A systematic review of perception system and simulators for autonomous vehicles research," *Sensors*, vol. 19, no. 3, p. 648, 2019.
[80] IWG AEBS-UTAC. *Validation Method: Virtual Testing*. Accessed: Feb. 5, 2021. [Online]. Available: https://wiki.unece.org/download/attachments/101554586/AEBS-12-07%20%28UTAC%29%20Virtual%20testing%20AEBS.pdf?api=v2
[81] *Road Vehicles—Prospective Safety Performance Assessment of Pre-Crash Technology by Virtual Simulation—Part 1: State-of-the-Art and General Method Overview*, Standard ISO/TR 21934, ISO, 2021.
[82] E. Roth, T. Dirndorfer, K. V. Neumann-Cosel, M.-O. Fischer, T. Ganslmeier, A. Kern, and A. Knoll, "Analysis and validation of perception sensor models in an integrated vehicle and environment simulation," in *Proc. 22nd Enhanced Saf. Vehicles Conf.*, 2011, pp. 1–9.
[83] P. Cao, W. Wachenfeld, and H. Winner, "Perception sensor modeling for virtual validation of automated driving," *it-Inf. Technol.*, vol. 57, no. 4, pp. 243–251, Aug. 2015.
[84] A. Schaermann, A. Rauch, N. Hirsenkorn, T. Hanke, R. Rasshofer, and E. Biebl, "Validation of vehicle environment sensor models," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2017, pp. 405–411.
[85] P. Rosenberger, J. T. Wendler, M. F. Holder, C. Linnhoff, M. Berghöfer, H. Winner, and M. Maurer, "Towards a generally accepted validation methodology for sensor models-challenges, metrics, and first results," Tech. Univ. Graz, Graz, Austria, Tech. Rep. 8653, 2019.
[86] P. Rosenberger, M. Holder, S. Huch, H. Winner, T. Fleck, M. R. Zofka, J. M. Zollner, T. D'hondt, and B. Wassermann, "Benchmarking and functional decomposition of automotive lidar sensor models," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2019, pp. 632–639.
[87] B. Schlager, S. Muckenhuber, S. Schmidt, H. Holzer, R. Rott, F. M. Maier, K. Saad, M. Kirchengast, G. Stettinger, D. Watzenig, and J. Ruebsam, "State-of-the-art sensor models for virtual testing of advanced driver assistance systems/autonomous driving functions," *SAE Int. J. Connected Automated Vehicles*, vol. 3, no. 3, pp. 233–261, Oct. 2020.
[88] P. Jaccard, "The distribution of the flora in the Alpine zone. 1," *New Phytologist*, vol. 11, no. 2, pp. 37–50, Feb. 1912.
[89] F. Reway, A. Hoffmann, D. Wachtel, W. Huber, A. Knoll, and E. Ribeiro, "Test method for measuring the simulation-to-reality gap of camera-based object detection algorithms for autonomous driving," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Oct. 2020, pp. 1249–1256.
[90] A. Ngo, M. P. Bauer, and M. Resch, "A multi-layered approach for measuring the simulation-to-reality gap of radar perception for autonomous driving," 2021, *arXiv:2106.08372*.
[91] M. I. Skolnik, *Radar Handbook*. New York, NY, USA: McGraw-Hill, 2008.
[92] U. Chipengo, P. M. Krenz, and S. Carpenter, "From antenna design to high fidelity, full physics automotive radar sensor corner case simulation," *Model. Simul. Eng.*, vol. 2018, pp. 1–19, Dec. 2018.
[93] C. Warren and A. Giannopoulos, "Creating finite-difference time-domain models of commercial ground-penetrating radar antennas using Taguchi's optimization method," *Geophysics*, vol. 76, no. 2, pp. G37–G47, Mar. 2011.

[94] Z. Yun and M. F. Iskander, "Ray tracing for radio propagation modeling: Principles and applications," *IEEE Access*, vol. 3, pp. 1089–1100, 2015.
[95] M. Holder, P. Rosenberger, H. Winner, T. Dhondt, V. P. Makkapati, M. Maier, H. Schreiber, Z. Magosi, Z. Slavik, O. Bringmann, and W. Rosenstiel, "Measurements revealing challenges in radar sensor modeling for virtual validation of autonomous driving," in *Proc. 21st Int. Conf. Intell. Transp. Syst. (ITSC)*, Nov. 2018, pp. 2616–2622.
[96] A. Ngo, M. P. Bauer, and M. Resch, "A sensitivity analysis approach for evaluating a radar simulation for virtual testing of autonomous driving functions," in *Proc. 5th Asia–Pacific Conf. Intell. Robot Syst. (ACIRS)*, Jul. 2020, pp. 122–128.
[97] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *Proc. 1st Annu. Conf. Robot Learn.*, 2017, pp. 1–16.
[98] A. Ngo, M. P. Bauer, and M. Resch, "Deep evaluation metric: Learning to evaluate simulated radar point clouds for virtual testing of autonomous driving," in *Proc. IEEE Radar Conf. (RadarConf)*, May 2021, pp. 1–6.
[99] P. Rosenberger, M. Holder, M. Zirulnik, and H. Winner, "Analysis of real world sensor behavior for rising fidelity of physically based LiDAR sensor models," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2018, pp. 611–616.
[100] P. Rosenberger, M. F. Holder, N. Cianciaruso, P. Aust, J. F. Tamm-Morschel, C. Linnhoff, and H. Winner, "Sequential LiDAR sensor system simulation: A modular approach for simulation-based safety validation of automated driving," *Automot. Engine Technol.*, vol. 5, nos. 3–4, pp. 187–197, Dec. 2020.
[101] T. Hanke, N. Hirsenkorn, C. van Driesten, P. Garcia-Ramos, M. Schiementz, S. Schneider, and E. Biebl, "Open simulation interface: A generic interface for the environment perception of automated driving functions in virtual scenarios," Technische Univ. München, Munich, Germany, Tech. Rep. 1, 2017.
[102] J. Chen, K. Venkataraman, D. Bakin, B. Rodricks, R. Gravelle, P. Rao, and Y. Ni, "Digital camera imaging system simulation," *IEEE Trans. Electron Devices*, vol. 56, no. 11, pp. 2496–2505, Nov. 2009.
[103] D. Gruyer, M. Grapinet, and P. De Souza, "Modeling and validation of a new generic virtual optical sensor for ADAS prototyping," in *Proc. IEEE Intell. Vehicles Symp.*, Jun. 2012, pp. 969–974.
[104] M. Höber, D. Nalic, A. Eichberger, S. Samiee, Z. Magosi, and C. Payerl, "Phenomenological modelling of lane detection sensors for validating performance of lane keeping assist systems," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Oct. 2020, pp. 899–905.
[105] *Passenger Cars—Simulation Model Classification—Part 1: Vehicle Dynamics*, Standard ISO/DIS11010-1, ISO, 2022.
[106] A. Hoskins and M. El-Gindy, "Technical report: Literature survey on driving simulator validation studies," *Int. J. Heavy Vehicle Syst.*, vol. 13, no. 3, pp. 241–252, 2006.
[107] M. Viehof and H. Winner, "Research methodology for a new validation concept in vehicle dynamics," *Automot. Engine Technol.*, vol. 3, nos. 1–2, pp. 21–27, Aug. 2018.
[108] J. K. Subosits and J. C. Gerdes, "Impacts of model fidelity on trajectory optimization for autonomous vehicles in extreme maneuvers," *IEEE Trans. Intell. Vehicles*, vol. 6, no. 3, pp. 546–558, Sep. 2021.
[109] P. Falcone, F. Borrelli, J. Asgari, H. E. Tseng, and D. Hrovat, "Predictive active steering control for autonomous vehicle systems," *IEEE Trans. Control Syst. Technol.*, vol. 15, no. 3, pp. 566–580, May 2007.
[110] M. Blundell and D. Harty, *Multibody Systems Approach to Vehicle Dynamics*. Amsterdam, The Netherlands: Elsevier, 2004.
[111] M. Dempsey, G. Fish, and J. G. D. Beltran, "High fidelity multibody vehicle dynamics models for driver-in-the-loop simulators," in *Proc. 11th Int. Modelica Conf.* Versailles, France: Linköping Univ. Electronic Press, Sep. 2015, pp. 273–280, no. 118.
[112] J. Wishart, S. Como, U. Forgione, J. Weast, L. Weston, A. Smart, G. Nicols, and S. Ramesh, "Literature review of verification and validation activities of automated driving systems," *SAE Int. J. Connected Automated Vehicles*, vol. 3, no. 4, pp. 267–323, Feb. 2021.
[113] F. Camara, N. Bellotto, S. Cosar, D. Nathanael, M. Althoff, J. Wu, J. Ruenz, A. Dietrich, and C. W. Fox, "Pedestrian models for autonomous driving part I: Low-level models, from sensing to tracking," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 10, pp. 6131–6151, Oct. 2021.
[114] T. Gindele, S. Brechtel, and R. Dillmann, "Learning driver behavior models from traffic observations for decision making and planning," *IEEE Intell. Transp. Syst. Mag.*, vol. 7, no. 1, pp. 69–79, Jan. 2015.
[115] S. Riedmaier, J. Schneider, B. Danquah, B. Schick, and F. Diermeyer, "Non-deterministic model validation methodology for simulation-based safety assessment of automated vehicles," *Simul. Model. Pract. Theory*, vol. 109, May 2021, Art. no. 102274.
[116] S. Starik and M. Werman, "Simulation of rain in videos," in *Proc. Texture Workshop, ICCV*, vol. 2, 2003, pp. 406–409.
[117] A. Arage, W. M. Steffens, G. Kuehnle, and R. Jakoby, "Effects of water and ice layer on automotive radar," in *Proc. German Microw. Conf*, 2006, pp. 1–10.
[118] R. H. Rasshofer, M. Spies, and H. Spies, "Influences of weather phenomena on automotive laser radar systems," *Adv. Radio Sci.*, vol. 9, no. 2, pp. 49–60, Jul. 2011.
[119] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[120] S. Hasirlioglu and A. Riener, "A model-based approach to simulate rain effects on automotive surround sensor data," in *Proc. 21st Int. Conf. Intell. Transp. Syst. (ITSC)*, Nov. 2018, pp. 2609–2615.
[121] C. J. Roy, "Review of code and solution verification procedures for computational simulation," *J. Comput. Phys.*, vol. 205, no. 1, pp. 131–156, May 2005.
[122] B. H. Thacker, "ASME standards committee on verification and validation in computational solid mechanics," Amer. Soc. Mech. Eng. (ASME) Council Codes Standards, USA, Tech. Rep., Sep. 2001, vol. 8.
[123] P. J. Durst, D. T. Anderson, and C. L. Bethel, "A historical review of the development of verification and validation theories for simulation models," *Int. J. Model., Simul., Sci. Comput.*, vol. 8, no. 2, Jun. 2017, Art. no. 1730001.
[124] K. Groh, S. Wagner, T. Kuehbeck, and A. Knoll, "Simulation and its contribution to evaluate highly automated driving functions," *SAE Int. J. Adv. Current Pract. Mobility*, vol. 1, no. 2, pp. 539–549, Apr. 2019.



RICCARDO DONÀ received the B.Sc. and M.Sc. degrees in mechatronics engineering and the Ph.D. degree in materials, mechatronics, and systems engineering from the University of Trento, in 2014, 2017, and 2021, respectively, as part of the EU research project Dreams4Cars. He is currently working with Uni Systems Italy. He is also serving as an External Consultant for the Joint Research Centre for the European Commission in the role of Virtual Testing Expert. In particular, he supports the development of traffic models and virtual testing environments. His main scientific interests include simulation and control algorithms for autonomous driving vehicles.

BIAGIO CIUFFO (Member, IEEE) received the Ph.D. degree in transportation engineering from the Department of Transportation Engineering, University of Napoli Federico II, in 2008. He held a three-year postdoctoral position at the European Commission Joint Research Centre (JRC), Ispra, Italy, working on the sustainability assessment of traffic and transport-related measures and policies. He is currently an Official of the European Commission, working for the Directorate for Energy, Transport, and Climate of the JRC. In the past, he has led different projects concerning the analysis of the environmental and economic impacts of different transport policies. He is currently leading the JRC Project focusing on the wide implications of connected and automated mobility. He has published more than 100 scientific papers in peer-reviewed journals and conference proceedings in transportation and traffic engineering. He is also one of the main authors of the *JRC Report on the Future of Road Transport*, which analyzes the wide implications of connected, automated, low-carbon, and shared mobility. He has been awarded the 2012 Greenshields Prize from the Traffic Flow Theory and Characteristics Committee and the 2013 and 2020 Prizes of the SimSub Committee of the Transportation Research Board of the U.S. National Academy of Science, for his research activities on traffic simulation. He is an Associate Editor of the IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS and serves as a reviewer for the most important journals in the transportation field.