Metrics for Specification, Validation, and Uncertainty Prediction for Credibility in Simulation of Active Perception Sensor Systems

Vom Fachbereich Maschinenbau an der
Technischen Universität Darmstadt
zur Erlangung des Grades eines
Doktor-Ingenieurs (Dr.-Ing.)
genehmigte

Dissertation

vorgelegt von

Philipp Rosenberger, M. Sc.

aus Alzenau

Berichterstatter:	Prof. Dr. rer. nat. Hermann Winner
Mitberichterstatter:	Assoc.Prof. Dipl.-Ing. Dr.techn. Arno Eichberger

Tag der Einreichung:	30.06.2022
Tag der mündlichen Prüfung:	08.11.2022

Darmstadt 2022

D 17

Preface

“Nanos gigantum humeris insidentes”

“We all stand on the shoulders of giants”, as Newton already stated in 1676. In my case, clearly all the people cited in this work have paved the way for me to take the next step in understanding.

Special thanks go to my doctoral supervisor Hermann Winner, who was always available for fruitful discussions, got to the heart of difficult questions and helped me develop my solution. The term “corporate culture” is often used these days, but the way he has led FZD under the high pressure we all face has created a unique spirit that has helped us all to grow together.

I thank Arno Eichberger for showing interest in my research from the beginning of the ENABLE-S3 project, for never letting the discussions about the quality of the reference data and the model validation itself fade away, and for taking on the role of second investigator.

I would also like to thank all my friends and colleagues at FZD that always had an open ear and time for a beer when things weren’t going so well or when there was something to celebrate. Special thanks go to the Sensenheinis: Martin, Clemens, Lukas, and Nico. Countless joint reflections and brainstormings and innumerable hours of experiments at Griesheim Airfield are the basis for this work, without you guys I would not have been able to write it!

During my time at FZD, I had the pleasure of supervising many talented students (some are now doing their own PhDs) and had the help of very passionate HiWi students. A big thank you to all of you for the intense discussions that helped me find research gaps and your help with all the experiments that sharpened my ideas.

Additionally, I had a lot of fun to work with so many colleagues all around Germany and Europe in the research projects ENABLE-S3, PEGASUS, SET Level, and VVM. In the course of this, I would also like to thank the German Federal Ministry for Economic Affairs and Energy and the German Federal Ministry of Education and Research for funding these challenging and exciting research projects.

Last, but most importantly, I would like to thank my family for always being there for me. My beloved wife Laura and my dear son Anton always made me smile again when I was down and helped me getting this thing done.

Laura, *for thou art my rock and my fortress*! I will always be in your debt for what you have shouldered to help me get my Dr.-Ing.!

Darmstadt, June 2022.

Table of Contents

List of Symbols and Indices

Latin formula symbols

Symbol	Unit	Description
A	m^2	Area
c	$\frac{\mathrm{m}}{\mathrm{s}}$	Speed of light
C	m^2	Optical aperture constant
d	m	Distance
D	-	Divergence
F	-	CDF/EDF
h	m	Height
I	$[I]$	Intensity
H	-	Spatial impulse response
P	W	Power
r	m	Range
w	m	Width
$(x\,y\,z)$	m	Cartesian coordinates

Greek formula symbols

Symbol	Unit	Description
β	-	Differential reflectivity
η	-	Efficiency
γ	rad	Opening angle
Γ	-	Lambertian reflection characteristic
ψ	rad	Azimuth
ρ	m	Radius
τ	s	Duration
θ	rad	Elevation
χ	-	Crossover function
ζ	$[\zeta]$	Measurand
$\boldsymbol{\zeta}$	-	Vector of measurands
$\boldsymbol{Z}$	-	Grid matrix / array of (vectors of) measurands

Calligraphic symbols and fraktur characters

Symbol	Unit	Description
$\mathcal{N}$	-	Normal distribution
Ᵽ	-	Probability
$\mathcal{S}$	-	Safety Factor
$\mathcal{F}$	-	Set of CDFs or EDFs, a “p-box”
Ƶ	-	Set of measurand vectors
$\mathbb{R}$	-	Set of real numbers
$\mathbb{1}$	-	Matrix of ones

Indices

Symbol	Description
95	With 95 % confidence
AD	Anderson-Darling
BB	Bounding box
bias	Bias
c	Corrected
cnl	Channel
cog	Center of gravity
fov	Field of view
h	Horizontal
JS	Jensen-Shannon
KL	Kullback–Leibler
KS	Kolmogorov-Smirnov
L	Left
max	Maximal
medi	Median
min	Minimal
nom	Nominal
obj	Object
of	Of flight
P	Pulse
r	Received
R	Right
ref	Reference

Symbol	Description
stdv	Standard deviation
sym	Symmetrical
t	Transmitted
th	Threshold
v	Vertical

Accents and Operators

Symbol	Description
$\neg a$	Negated value
$\widetilde{a}$	Simulated value
$\widehat{a}$	Predicted/estimated value
$\overline{a}$	Arithmetic mean
$\overset{\vdash\!\dashv}{a}$	Standard deviation

List of Abbreviations

ADS	automated driving system
APSS	active perception sensor system
AVM	area validation metric
BSDF	bidirectional scattering distribution function
CAVM	corrected AVM
CDF	cumulative distribution function
CEPRA	Cause, Effect, and Phenomenon Relevance Analysis
DAS	driving automation system
DVM	double validation metric
EDF	empirical distribution function
EPW	echo pulse width
FMCW	frequency modulated continuous wave
FMEA	failure mode and effects analysis
FMI	Functional Mock-up Interface
FN	false negative
FoV	field of view
FP	false positive
FZD	Institute of Automotive Engineering
GPU	graphics processing unit
GT	ground truth
KPI	key performance indicator
MAVM	modified AVM
MEMS	microelectromechanical systems
MVC	metric validity criterion
ODD	operational design domain
OG	occupancy grid
OPA	optical phased array
OSI	Open Simulation Interface
OSPA	Optimal Sub-Pattern Assignment
PBA	probability bound analysis
p-box	probability box
PCA	principal component analysis
PCMM	predictive capability maturity model

PDF	probability density function
PerCollECT	Perception Sensor Collaborative Effect and Cause Tree
PIRT	Phenomena Identification and Ranking Table
R&R	repeatability and reproducibility
RCS	radar cross-section
RMSE	root mean squared error
SAE	Society of Automotive Engineers
SiPM	silicon photomultiplier
SNR	signal-to-noise ratio
SotA	state of the art
SPAD	single-photon avalanche diode
SRQ	system response quantity
SuT	system under test
TCSPC	time-correlated single-photon counting
ToF	time of flight
TTC	time-to-collision
UQ	uncertainty quantification
V&V	verification and validation
VV&UQ	verification, validation, and uncertainty quantification

List of Figures

List of Tables

Kurzzusammenfassung

Der immense Aufwand für die Sicherheitsvalidierung eines automatisierten Fahrsystems des SAE-Levels 3 oder höher ist bekanntermaßen nicht alleine durch reale Testfahrten darstellbar. Daher ist die Simulation selbst für begrenzte Betriebsbereiche die Lösung für die Homologation automatisierter Fahrfunktionen. Folglich müssen alle Simulationsmodelle vorher qualifiziert sein, die hierfür als Werkzeug verwendet werden. Dafür ist neben deren Verifizierung und Validierung auch die Unsicherheits-Quantifizierung (VV&UQ) und -Vorhersage in den Anwendungsbereich für die Glaubwürdigkeit des Simulationsmodells erforderlich. Um eine solche VV&UQ zu ermöglichen, werden am Beispiel einer eigens erarbeiteten Simulation von Lidar-Sensorsystemen neue Metriken vorgestellt, die ganzheitlich zum Nachweis der Modellglaubwürdigkeit und -reife für Simulationsmodelle aktiver Wahrnehmungssensorsysteme verwendet werden können.

Der ganzheitliche Prozess hin zur Modellglaubwürdigkeit beginnt bei der Formulierung der Anforderungen an die Modelle. Die Schwellwerte der Metriken als Akzeptanzkriterien sind dabei quantifizierbar durch die Relevanzanalyse der Ursache-Wirkungs-Ketten, die in verschiedenen Szenarien vorherrschen, und sollten dafür intuitiv in der gleichen Einheit wie die simulierte Messgröße vorliegen. Diese Zusammenhänge können über die vorgestellten, aufeinander abgestimmten Methoden "Perception Sensor Collaborative Effect and Cause Tree" (PerCollECT) und "Cause, Effect, and Phenomenon Relevance Analysis" (CEPRA) abgeleitet werden. Für die Stichprobenvalidierung muss jedes Experiment von Referenzmessungen begleitet werden, da diese dann als Simulations-Input dienen. Da die Referenzdatenerhebung neben aleatorischer auch epistemischer Unsicherheit unterliegt, welche in Form unterschiedlicher Eingangsdaten durch die Simulation durchpropagiert werden, führt dies zu mehreren leicht unterschiedlichen Simulationsergebnissen. Bei der hier betrachteten Simulation von Messsignalen und Daten über die Zeit lässt sich diese Kombination der Unsicherheiten am besten als übereinandergelegte kumulierte Wahrscheinlichkeitsfunktionen ausdrücken. Die Metrik muss daher in der Lage sein, solche sog. P-Boxen als Ergebnis der Massensimulationen zu verarbeiten.

Die Flächenvalidierungsmetrik (engl. AVM) wird durch eine detaillierte Analyse als beste der bereits genutzten Metriken ausgewählt und erweitert, um alle Voraussetzungen erfüllen zu können. Dabei entsteht die korrigierte AVM (engl. CAVM), die den Modellfehler in der Streuung der simulierten Messwerte quantifiziert. Schließlich wird die Doppelvalidierungsmetrik (DVM) herausgearbeitet als Doppelvektor aus dieser mit dem Schätzwert für den Modell-Mittelwert-Fehler. Die neuartige Metrik wird beispielhaft auf die empirischen kumulativen Verteilungsfunktionen von Lidar-Messungen und den P-Boxen aus deren Re-Simulationen angewendet. Dabei werden zum ersten Mal aleatorische und epistemische Unsicherheiten berücksichtigt und die neuartigen Metrik erfolgreich etabliert. Hierbei wird auch die Quantifizierung der Unsicherheiten und Fehlervorhersage für ein Sensormodell auf Basis der Stichprobenvalidierung erstmals demonstriert.

Abstract

The immense effort required for the safety validation of an automated driving system of SAE level 3 or higher is known not to be feasible by real test drives alone. Therefore, simulation is key even for limited operational design domains for homologation of automated driving functions. Consequently, all simulation models used as tools for this purpose must be qualified beforehand. For this, in addition to their verification and validation, uncertainty quantification (VV&UQ) and prediction for the application domain are required for the credibility of the simulation model. To enable such VV&UQ, a particularly developed lidar sensor system simulation is utilized to present new metrics that can be used holistically to demonstrate the model credibility and -maturity for simulation models of active perception sensor systems.

The holistic process towards model credibility starts with the formulation of the requirements for the models. In this context, the threshold values of the metrics as acceptance criteria are quantifiable by the relevance analysis of the cause-effect chains prevailing in different scenarios, and should intuitively be in the same unit as the simulated metric for this purpose. These relationships can be inferred via the presented aligned methods “Perception Sensor Collaborative Effect and Cause Tree” (PerCollECT) and “Cause, Effect, and Phenomenon Relevance Analysis” (CEPRA). For sample validation, each experiment must be accompanied by reference measurements, as these then serve as simulation input. Since the reference data collection is subject to epistemic as well as aleatory uncertainty, which are both propagated through the simulation in the form of input data variation, this leads to several slightly different simulation results. In the simulation of measured signals and data over time considered here, this combination of uncertainties is best expressed as superimposed cumulative distribution functions. The metric must therefore be able to handle such so-called p-boxes as a result of the large set of simulations.

In the present work, the area validation metric (AVM) is selected by a detailed analysis as the best of the metrics already used and extended to be able to fulfill all the requirements. This results in the corrected AVM (CAVM), which quantifies the model scattering error with respect to the real scatter. Finally, the double validation metric (DVM) is elaborated as a double-vector of the former metric with the estimate for the model bias. The novel metric is exemplarily applied to the empirical cumulative distribution functions of lidar measurements and the p-boxes from their re-simulations. In this regard, aleatory and epistemic uncertainties are taken into account for the first time and the novel metrics are successfully established. The quantification of the uncertainties and error prediction of a sensor model based on the sample validation is also demonstrated for the first time.

1 Motivation and Methodology for this Work

1.1 Active Perception Sensor System (APSS) Simulation and its Missing Credibility

As already stated by Emery in the journal of verification, validation, and uncertainty quantification (VV&UQ), *"we need to recognize that in today's world almost every real situation will sooner or later be modeled."* [1] Nevertheless, there cannot be any serious application of simulation (in contrast to e.g. gaming) without profound VV&UQ beforehand. In this context, *"Validation addresses the question of the adequacy of a model to represent a real situation"* [1] and *"Uncertainty Quantification is a recognition that different experiments will produce different results."* [1] While questioning simulation results and model credibility is as old as computer simulation itself and the terminology has been established for many years[2], *"VV&UQ methodology is an evolving field of research"* [3] until today.

Regardless of its level of driving automation, as defined by the Society of Automotive Engineers (SAE)[4a], there is a common sense on the impossible effort in safety validation of any driving automation system (DAS) by actual test-driving in real world.[5] In consequence, simulation is seen as the key enabler for safety validation before release of any DAS on public roads. This explains the need for simulation in the first place and the required high credibility of all tools and models, e.g. the active perception sensor system (APSS) models discussed here.

There is evidently a high demand in simulation-based testing and there is already e.g. a well described predictive capability maturity model (PCMM) for computer simulation[6]. However, to the knowledge of the author, until now there is no detailed report available for any APSS simulation regarding its PCMM level or similar reporting on any of the categories described by Oberkampf et al.[6] and depicted in Fig. 1-1. The PCMM overview emphasizes that for qualification or certification, as it is the case for safety validation of DAS, all elements of the simulation require maturity level 3. There are a few publications on APSS model verification and validation (V&V), such as that of Pliefke et al., but they lack the necessary details and do not provide metrics, a complete scenario catalog, or a sensitivity analysis, let alone a quantification of uncertainty[7].

[1] Emery, A. F.: Special Issue: Sandia V&V Challenge Problem (2016).

[2] Schlesinger, S. et al.: Terminology for model credibility (1979).

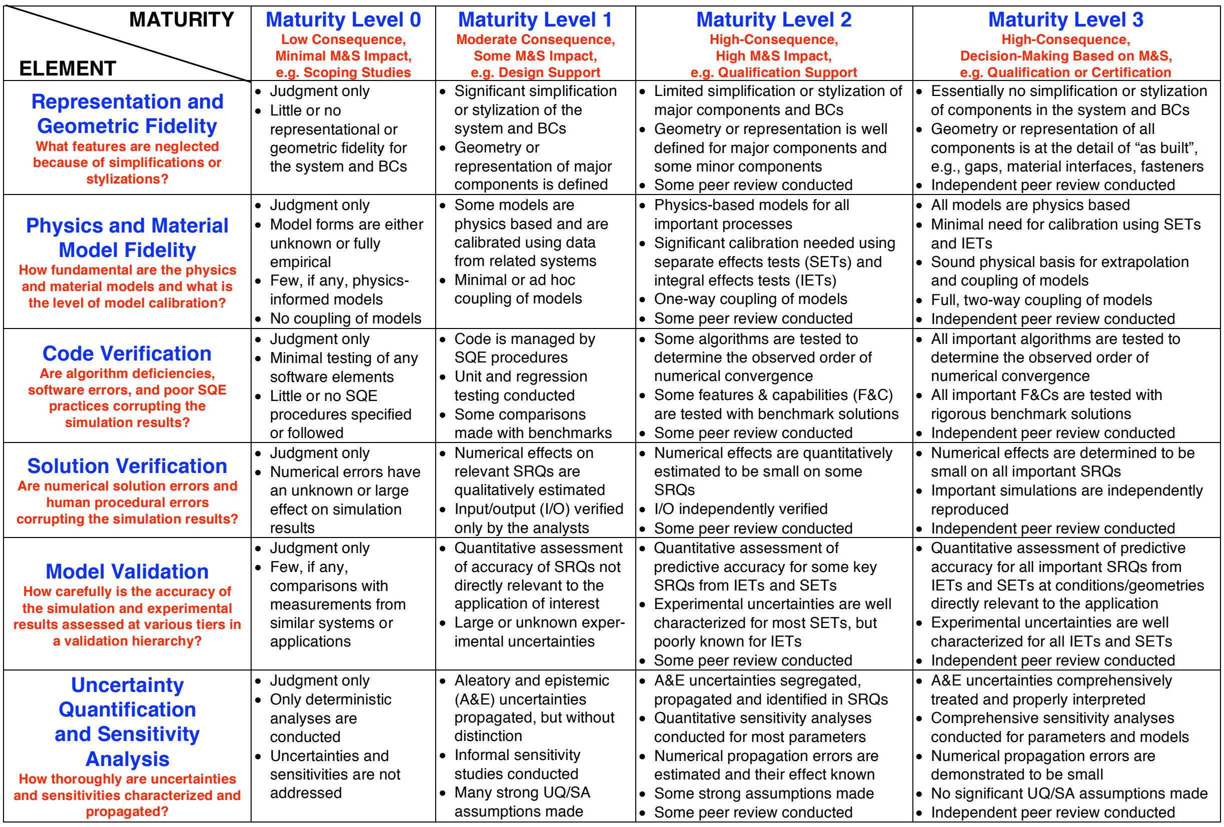
[3] Hu, K. T. et al.: Introduction: The 2014 Sandia Verification and Validation Challenge Workshop (2016).

[4] Society of Automotive Engineers: SAE-J3016 (2021). a: pp. 24-34.; b: p. 17.; c: p. 33.

[5] Wachenfeld, W.; Winner, H.: Die Freigabe des autonomen Fahrens (2015).

[6] Oberkampf, W. L. et al.: Predictive Capability Maturity Model for Modeling and Simulation. (2007), p. 38.

[7] Pliefke, S. et al.: Validation of a Ray-tracing-based Radar Sensor Model (2021).



MATURITY ELEMENT	Maturity Level 0 Low Consequence, Minimal M&S Impact, e.g. Scoping Studies	Maturity Level 1 Moderate Consequence, Some M&S Impact, e.g. Design Support	Maturity Level 2 High-Consequence, High M&S Impact, e.g. Qualification Support	Maturity Level 3 High-Consequence, Decision-Making Based on M&S, e.g. Qualification or Certification
Representation and Geometric Fidelity What features are neglected because of simplifications or stylizations?	• Judgment only • Little or no representational or geometric fidelity for the system and BCs	• Significant simplification or stylization of the system and BCs • Geometry or representation of major components is defined	• Limited simplification or stylization of major components and BCs • Geometry or representation is well defined for major components and some minor components • Some peer review conducted	• Essentially no simplification or stylization of components in the system and BCs • Geometry or representation of all components is at the detail of "as built", e.g., gaps, material interfaces, fasteners • Independent peer review conducted
Physics and Material Model Fidelity How fundamental are the physics and material models and what is the level of model calibration?	• Judgment only • Model forms are either unknown or fully empirical • Few, if any, physics-informed models • No coupling of models	• Some models are physics based and are calibrated using data from related systems • Minimal or ad hoc coupling of models	• Physics-based models for all important processes • Significant calibration needed using separate effects tests (SETs) and integral effects tests (IETs) • One-way coupling of models • Some peer review conducted	• All models are physics based • Minimal need for calibration using SETs and IETs • Sound physical basis for extrapolation and coupling of models • Full, two-way coupling of models • Independent peer review conducted
Code Verification Are algorithm deficiencies, software errors, and poor SQE practices corrupting the simulation results?	• Judgment only • Minimal testing of any software elements • Little or no SQE procedures specified or followed	• Code is managed by SQE procedures • Unit and regression testing conducted • Some comparisons made with benchmarks	• Some algorithms are tested to determine the observed order of numerical convergence • Some features & capabilities (F&C) are tested with benchmark solutions • Some peer review conducted	• All important algorithms are tested to determine the observed order of numerical convergence • All important F&Cs are tested with rigorous benchmark solutions • Independent peer review conducted
Solution Verification Are numerical solution errors and human procedural errors corrupting the simulation results?	• Judgment only • Numerical errors have an unknown or large effect on simulation results	• Numerical effects on relevant SRQs are qualitatively estimated • Input/output (I/O) verified only by the analysts	• Numerical effects are quantitatively estimated to be small on some SRQs • I/O independently verified • Some peer review conducted	• Numerical effects are determined to be small on all important SRQs • Important simulations are independently reproduced • Independent peer review conducted
Model Validation How carefully is the accuracy of the simulation and experimental results assessed at various tiers in a validation hierarchy?	• Judgment only • Few, if any, comparisons with measurements from similar systems or applications	• Quantitative assessment of accuracy of SRQs not directly relevant to the application of interest • Large or unknown experimental uncertainties	• Quantitative assessment of predictive accuracy for some key SRQs from IETs and SETs • Experimental uncertainties are well characterized for most SETs, but poorly known for IETs • Some peer review conducted	• Quantitative assessment of predictive accuracy for all important SRQs from IETs and SETs at conditions/geometries directly relevant to the application • Experimental uncertainties are well characterized for all IETs and SETs • Independent peer review conducted
Uncertainty Quantification and Sensitivity Analysis How thoroughly are uncertainties and sensitivities characterized and propagated?	• Judgment only • Only deterministic analyses are conducted • Uncertainties and sensitivities are not addressed	• Aleatory and epistemic (A&E) uncertainties propagated, but without distinction • Informal sensitivity studies conducted • Many strong UQ/SA assumptions made	• A&E uncertainties segregated, propagated and identified in SRQs • Quantitative sensitivity analyses conducted for most parameters • Numerical propagation errors are estimated and their effect known • Some strong assumptions made • Some peer review conducted	• A&E uncertainties comprehensively treated and properly interpreted • Comprehensive sensitivity analyses conducted for parameters and models • Numerical propagation errors are demonstrated to be small • No significant UQ/SA assumptions made • Independent peer review conducted

Figure 1-1: Overview of all elements and levels of the PCMM for simulation from Oberkampf et al.[6]

In combination with the lack of experience in APSS simulation for safety validation, there is no trust or confidence in such simulation at the moment, where actually highly (risk-)informed decision making should take place. To address this need for confidence, the following dissertation presents a holistic approach to requirements, testing, metrics, and uncertainty quantification (UQ) for actual trust in APSS models.

However, there is a tendency towards regulation and standardization of simulation-based safety validation of DAS, where credibility is seen as the key aspect for using simulation as validation tool. There is e.g. the United Nations Economic Commission for Europe (UNECE) intersecretariat working group for Validation Method for Automated Driving - SG 2 (Virtual testing) that aims for a UNECE regulation on the use of simulation for homologation. Additionally, the International Alliance for Mobility Testing and Standardization (IAMTS) has published a best practice[8] and as a *"leading global body of organizations in testing, standardization and verification of advanced mobility systems"* the IAMTS is expected to publish several follow-ups. Several standards exist for simulation-based testing, as e.g. the U.S. Department of Defense's MIL-STD-3022[9] or the U.S. National Aeronautics and Space Administration (NASA) standard NASA-STD-7009A[10].

[8] International Alliance for Mobility Testing and Standardization: IAMTS0001202104 (2021).

[9] U.S. Department of Defense: MIL-STD-3022 (2008).

[10] U.S. National Aeronautics and Space Administration: NASA-STD-7009A (2016). a: p. 57.

	M&S Development			**M&S Use (Operations)**			**Supporting Evidence**	
Level	**Data Pedigree**	**Verification**	**Validation**	**Input Pedigree**	**Uncertainty Characterization**	**Results Robustness**	**M&S History**	**M&S Process / Product Management**
4	All data known & traceable to RWS with acceptable accuracy, precision, & uncertainty.	Reliable practices applied to verify the end-to-end model; all model errors satisfy requirements.	All M&S outputs agree with data from the RWS over the full range of operation in its real operating environment.	All input data known & traceable to RWS with acceptable accuracy, precision, & uncertainty.	Statistical analysis of the output uncertainty after propagation of all known sources of uncertainty.	Sensitivities known for most parameters; most key sensitivities identified.	Nearly identical model and use.	Controlled processes are applied; measurements used for process improvement.
3	All data known & traced to sufficient referent. Significant data has acceptable accuracy, precision, & uncertainty.	Formal practices applied to verify the end-to-end model; all important errors satisfy requirements.	All key M&S outputs agree with data from the RWS operating in a representative environment.	All input data known & traced to sufficient referent. Significant input data has acceptable accuracy, precision, & uncertainty.	Uncertainty of results are provided quantitatively through propagation of all known uncertainty.	Sensitivities known for many parameters including many of the key sensitivities.	At most minor changes in model and at most minor differences in model use.	Controlled processes are applied; process compliance is measured.
2	Some data known & formally traceable with estimated uncertainties.	Documented practices applied to verify all model features; most important errors satisfy requirements.	Key M&S outputs agree with data from a sufficiently similar referent system.	Some input data known & formally traceable with estimated uncertainties.	Most sources of uncertainty identified, expressed quantitatively, and correctly classified. Propagation of the uncertainties is assessed.	Sensitivities known for a few parameters. Few or no key sensitivities identified.	At most moderate changes in model and at most moderate differences in model use.	Formal processes are applied.
1	Some data known and informally traceable.	Informal practices applied to verify some features of the model and assess errors.	Conceptual model addresses problem statement and agrees with available referents.	Some input data known and informally traceable.	Sources of uncertainty identified and qualitatively assessed.	Qualitative estimates only for sensitivities in M&S.	New model or major changes in model, or major differences in model use; but, model/changes/uses documented.	Informal processes are applied.

Figure 1-2: U.S. NASA Key Aspects of Credibility Assessment Levels for Modeling and Simulation[10a]

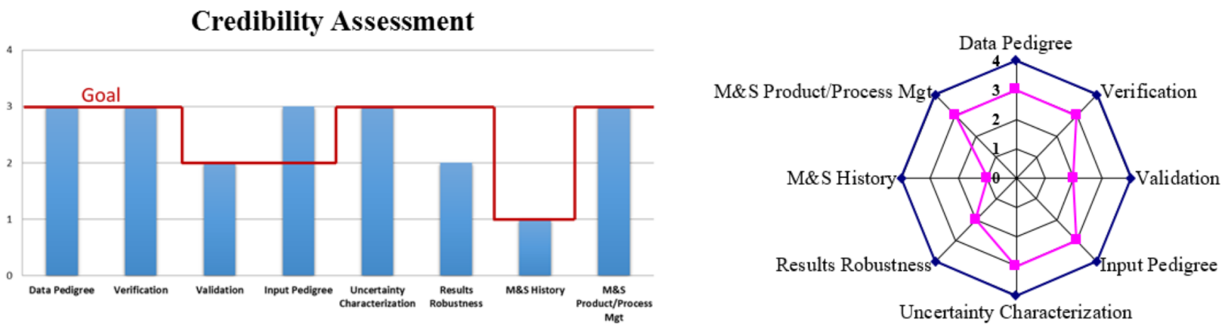


Figure 1-3: U.S. NASA Modeling and Simulation Credibility Assessment Synopsis[11a]

NASA-STD-7009A contains *"Key Aspects of Credibility Assessment Levels"* [10a] as depicted in Fig. 1-2 that are very related to the already presented PCMM. However, NASA concentrates more explicit on a profound data analysis that is used for modeling and model credibility assessment, while PCMM has this data analysis implicit within the UQ step. Therefore, maturity and credibility are slightly different, but both aim for the decision if the simulation results can be trusted for a given task. The NASA standard even has a handbook NASA-HDBK-7009A[11] that explains e.g. the credibility assessment and the final *"Credibility Assessment Synopsis"* [11a] that is providing a precise graphical visualization of the simulation credibility, as shown in Fig. 1-3.

In conclusion, PCMM and NASA-STD-7009A both show that random sample selection is simply not fulfilling the high requirements on simulation models for safety validation of any automated driving system (ADS) and model credibility demands not only comparison of measurements and simulation, but as Fig. 1-3 shows a sensitivity analysis, measurement and reference data analysis (pedigree), and UQ. Additionally, as Riedmayer et al. state, VV&UQ must be accompanied by overall maturity / credibility assessment procedures, such as the PCMM.[12] Consequently, simulation requires a lot of effort to reach an acceptable level of credibility and maturity for the usage as a serious tool for e.g. safety validation. The following dissertation will show the cornerstones along this way.

1.2 Introduction of Most Important Terms in APSS Simulation and Validation

The following section provides the definitions of terms used during this work to avoid misunderstandings in communication to the reader. Starting from operational design domain (ODD) and the respective parameter space, terms like scenario, effect, phenomenon, accuracy, and VV&UQ are explained and defined for the scope of this work in the context of DAS and ADS.

1.2.1 Operational Design Domain (ODD) and Parameter Space

Any DAS like adaptive cruise control (ACC), automatic emergency braking (AEB), or lane keeping assistance (LKA) should have a clearly specified ODD for its safety validation. Only an ADS of SAE Level 5 has an unlimited ODD that doesn't need any further restriction. Even active safety systems (ASS) like blind spot warning (BSW), lane departure warning (LDW), or forward collision warning (FCW) do have a specified ODD. Thereby, the exact definition of the term **operational design domain (ODD)** from the SAE is: *"Operating conditions under which a given driving automation system or feature thereof is specifically designed to function, including, but*

[11] U.S. National Aeronautics and Space Administration: NASA-HDBK-7009A (2019). a: p. 134.

[12] Riedmaier, S. et al.: Unified Framework and Survey for Model VV&UQ (2020), p. 27.

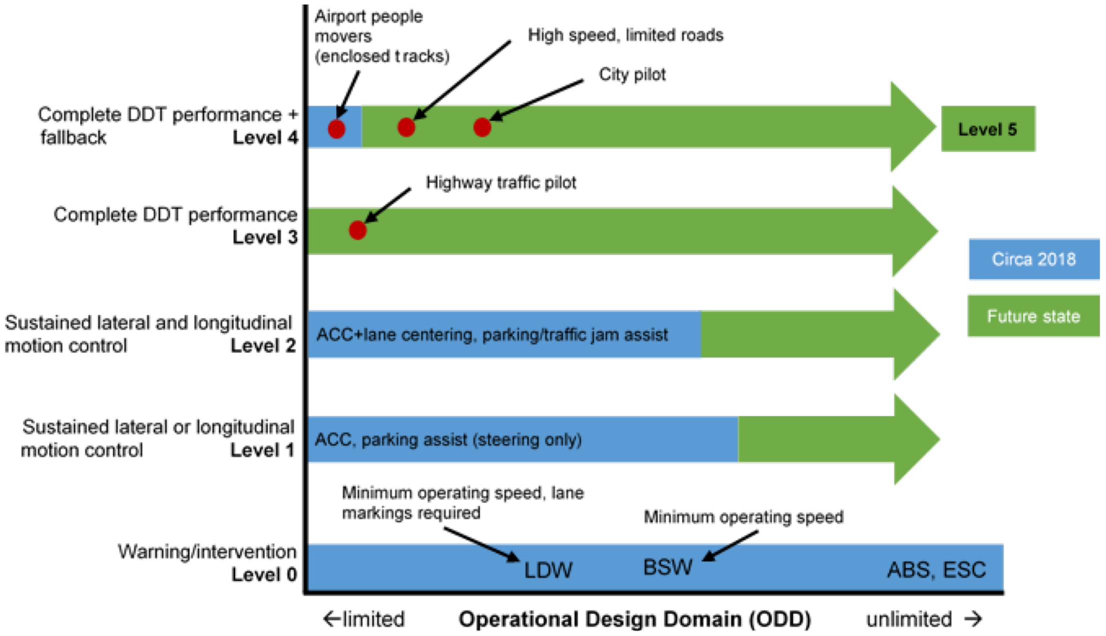

Airport people movers (enclosed tracks)
High speed, limited roads
City pilot
Complete DDT performance + fallback Level 4
Level 5
Highway traffic pilot
Complete DDT performance Level 3
Circa 2018
Future state
Sustained lateral and longitudinal motion control Level 2
ACC+lane centering, parking/traffic jam assist
Sustained lateral or longitudinal motion control Level 1
ACC, parking assist (steering only)
Minimum operating speed, lane markings required
Minimum operating speed
Warning/intervention Level 0
LDW
BSW
ABS, ESC
←limited
Operational Design Domain (ODD)
unlimited →

Figure 1-4: ODD relative to SAE driving automation levels from SAE-J3016[4c]

not limited to, environmental, geographical, and time-of-day restrictions, and / or the requisite presence or absence of certain traffic or roadway characteristics" [4b]

There is already the 6-layer model as an ontology for ODD by Scholtes et al., derived from originally four layers by Schuldt et al.[13] The 6-layer model considers already basic compatibility to standards like OpenDRIVE[14] and OpenSCENARIO[15], while still lacking a defined machine-readable format. Recently, ASAM e.V. started the project OpenODD[16] to overcome this issue. It has the goal to establish a commonly agreed machine-readable format for any ODD, while claiming that an *"ODD must be represented so it can easily be used within simulation and other machine processed environments."* [16]

Nevertheless, machine-readable does not necessarily include that the parameter space spanned by the ODD can be taken as-is for APSS model specification or even validation. For this purpose, parameters at least need to be interval scaled values, as defined by Stevens[17]. This is necessary for methods such as sensitivity analysis and, most importantly, for the application of metrics over the parameter space on the model outputs for model calibration, validation, and UQ. Therefore, in the context of the following work, the term **parameter space** stands for an at least interval scaled parameter space as prerequisite of the new methods and metrics for model specification and validation and the term ODD is therefore avoided later on and replaced. Still, the need of verification of the - at least - interval scaling of the parameters remains.

[13] Schuldt, F. et al.: Effiziente systematische Testgenerierung für Fahrerassistenzsysteme (2013).

[14] ASAM e.V.: ASAM OpenODD - Concept Paper (2021).

[15] ASAM e.V.: ASAM OpenSCENARIO® - User Guide (2021).

[16] ASAM e.V.: ASAM OpenDRIVE® - Specification (2021).

[17] Stevens, S. S.: On the Theory of Scales of Measurement (1946).

1.2.2 Scenario, Objects, and Experiment

In addition to the ODD, Ulbrich et al.[18] define the terms **scenery**, **scene**, **situation**, and **scenario** in the context of DAS with its APSS: While a scenery is the description of all static objects and the non-changeable environmental parameters, a scene is a snapshot of a scenario including all (possibly) moving and changing objects and circumstances. *"A scenario describes the temporal development between several scenes in a sequence of scenes."* [18a] The situation is the description of all rules and (planned) behavior of all objects within the scene, so a kind of another abstraction layer. Consequently, ODD is the unification of all possible scenarios in which DAS with its APSS is used.

Since, to the author's knowledge, there is no clear common understanding of the term **object**, at least in the scope of this work, an object could be any element in the scene that has a name and/or a unique ID. The ISO 23150, which standardizes data communication between perception sensors and the data fusion unit for automated driving functions, defines objects even more generally as *"representation of a real-world entity with defined boundaries and characteristics in the vehicle coordinate system"* [19a]. According to this definition, all actively moving scene elements such as vehicles, pedestrians, animals, etc. are objects, but passively moving elements such as plastic bags, etc. are also objects. Even more, all static scene elements like houses, road markings, traffic signs, traffic lights, and even object parts like windows are called objects as well.

Furthermore, Menzel et al. subdivide scenario descriptions into three abstraction layers called functional, logical, and concrete.[20] The authors already have simulation-based testing in mind and specifically design their layers to support e.g. OpenDRIVE and OpenSCENARIO. Still, APSS simulation is not in the focus of this fundamental work. The lack of deeper knowledge about sensors and especially APSS could lead to unnecessary scenarios for testing and safety validation of DAS, which should be avoided by concepts considering e.g. the dynamically changing perception ranges of APSS, as described in previous work of the author.[21] In the context of the following work, an experiment series for model calibration or validation is analogue to a logical scenario and a single **sample** of the experiment series is a concrete scenario. Consequently, the term scenario will be avoided in the following and replaced by experiment series and sample, which includes the implicit consequence that the parameter space of each sample and each experiment series is a subset of the full parameter space coming from the ODD description.

[18] Ulbrich, S. et al.: Defining and Substantiating the Terms Scene, Situation, and Scenario (2015). a: p. 986.

[19] International Organization for Standardization: ISO 23150:2021(E) (2021). a: p. 3.; b: p. 4.

[20] Menzel, T. et al.: Scenarios for Development, Test and Validation of Automated Vehicles (2018).

[21] Elster, L. et al.: Fundamental Design Criteria for Logical Scenarios (2021).

1.2.3 Signal, Cause, Effect, and Phenomenon

As the data at the interfaces is mentioned, it must be clarified what is propagated through the functional blocks in between and what happens to the propagated information during this process. Even if defined differently elsewhere, e.g. in the ISO 23150[19a], in this work the term **signal** is used in a physical way like in a previous work of the author as *"a quantity of energy influencing the sensor according to its measurement principle."* [22] Therefore, a signal is the basis for effects to appear, regardless if it has been sent actively by a perception sensor or is passively collected. The signal is received by the sensor's front-end and processed. After this signal processing, the produced data, now called detections, is handed over to the data processing part.

During signal propagation and processing, **effects** occur, defined as *"deviation from the originally existing information about the environment in the signal or data."* [22] By adding data at the definition's end, it is clarified that effects can still appear during data processing within the APSS. In addition to that, the mentioned originally existing information about the environment represents the so called *"ground truth (GT) under clean room conditions."* [22]

All effects have underlying **causes**, defined as the *"condition leading to a deviation in the information."* [22] Causes in this sense can be sensor hardware properties, weather conditions, object material properties, and so on.

After signal and data processing, **phenomena** can be observed at the output of the APSS. By this definition, phenomena are effects that are measurable and can influence subsequent functions like the automated driving functions. Consequently, mostly phenomena are listed in the specifications for APSS simulation, because they can be validated by comparison to real measurements.

1.2.4 Measurement, Reference, and Ground Truth

In this work, the term **measurement** is only used for the real data outputted by the APSS to be modeled. In contrast to measurements, all other data collected during real world experiments for calibration or validation of APSSs models is described by the term **reference data**. Reference data includes all information that can be collected during real world experiments, e.g. object poses, materials, trajectories, weather, etc. To clarify the term, it should be stated that the so called "reference sensor measurements" that try to come closer to the ideal value of the measurand with the same technology as the APSS itself could be a part of the collected reference data, but do not stand in contrast to it. When the replay-to-sim approach is used e.g. to generate simulation data for model validation, the reference data is used as parameters (e.g. weather, material) and input data (e.g. poses, trajectories) for the simulation.

In simulation, it can be named **ground truth (GT)**, but the term should be used with caution. To the understanding of the author, simulation is the only domain where GT exists when it

[22] Linnhoff, C. et al.: Towards Serious Sensor Simulation for Safety Validation of Automated Driving (2021), p. 2.

is used as a simulated reference sensor with total accuracy and any precision and trueness is possible and directly available as such. In contrast, when reference data is collected in real world, measurements are never error-free. They can be minimized but never eliminated completely, causing the reference data to be at least slightly different to the GT. The ideal value aimed for with reference sensors is often called GT, but for the already mentioned reasons the terms "ideal" and "GT" are avoided in further chapters of this work.

1.2.5 Error, Accuracy, and Measurement Uncertainty

ISO 5725-1 defines **accuracy** as combination of trueness and precision[23] and ISO 23150 applied in this work slightly refines the terms. It defines **trueness** as *"closeness of agreement between the average of an infinite number of replicated measured quantity values and a reference quantity value."* [19b]. Additionally, **precision** is the *"closeness of agreement between indications or measured quantity values obtained by replicate measurements on the same or similar measurands under specified conditions."* [19b] Neither trueness nor precision are quantities, but as depicted in Fig. 1-5 they are measured in form of **bias** for the former and **standard deviation** for the scattering error combined to the **measurement uncertainty**. To estimate these quantities, measurements are performed followed by re-simulations. During such campaigns, both **repeatability** of results under identical conditions and **reproducibility** of results by a different experimentalist, in a different laboratory, or at a different test track should be investigated.

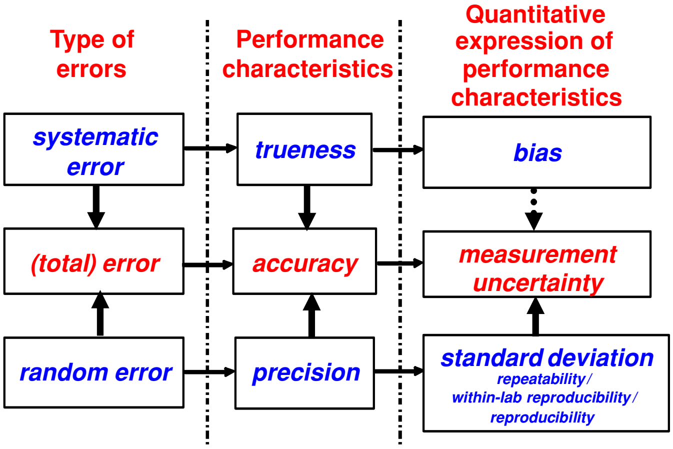


Figure 1-5: Relationships between type of error, qualitative performance characteristics and their quantitative expression from Menditto et al.[24] Reproduced with permission from Springer Nature.

When measurements are compared to a reference, or when simulated data is compared to real data, two types of errors are differentiated that form the overall error. **Systematic errors** lead to less trueness and appear as a bias in quantitative performance characteristics of a sensor or a sensor model. **Random errors** cause less precision and provoke a scattering error in quantitative performance characteristics. Menditto et al.[24] summarize this relationship vividly in Fig. 1-5.

[23] International Organization for Standardization: ISO 5725-1 (1994).

[24] Menditto, A. et al.: Understanding the meaning of accuracy, trueness and precision (2007), p. 46.

1.2.6 Model Verification and Validation (V&V)

According to Popper, model validity is only a temporary state with the prerequisite of an unsuccessful but profound attempt of falsification.[25] As computer model validation is a quite established field of research [26,27,28], Viehof and Winner have recently collected V&V definitions and methods and show that there is a common understanding in the field of automotive simulation.[29] The definitions written down by Oberkampf and Trucano[30], which are oriented on the definitions by the American Institute of Aeronautics and Astronautics (AIAA) and the American Society of Mechanical Engineers (ASME), are widely used in predecessor's works, e.g. by Viehof[31a], Schaermann[32], Riedmaier and Danquah[33], etc. so they will be used in this work as well:

- **Verification** is the *"process of determining that a model implementation accurately represents the developer's conceptual description of the model and the solution to the model."* [30]
- **Validation** is the *"process of determining the degree to which a model is an accurate representation of the real world from the perspective of the intended uses of the model."* [30]

Sargent discusses data validation as the basis for valid simulation modeling and then splits simulation validation into three parts:[34]

- **Data validity** is defined as *"ensuring that the data necessary for model building, model evaluation and testing, and conducting the model experiments to solve the problem are adequate and correct."* [34]
- **Conceptual model validation** is defined as *"determining that the theories and assumptions underlying the conceptual model are correct and that the model representation of the problem entity is 'reasonable' for the intended purpose of the model."* [34]
- **Computerized model verification** is defined as *"assuring that the computer programming and implementation of the conceptual model is correct."* [34]
- **Operational validation** is defined as *"determining that the model's output behavior has sufficient accuracy for the model's intended purpose over the domain of the model's intended applicability."* [34]

[25] Popper, K.: The Logic of Scientific Discovery (2002).

[26] Schlesinger, S. et al.: Terminology for model credibility (1979).

[27] Sargent, R. G.: Assessment Procedure and Set of Criteria in Evaluation of Computerized Models (1981).

[28] Balci, O.; Sargent, R. G.: Cost-Risk Analysis in the Statistical Validation of Simulation Models (1981).

[29] Viehof, M.; Winner, H.: Forschungsstand der Validierung (2017).

[30] Oberkampf, W. L.; Trucano, T. G.: Verification and validation benchmarks (2008), p. 719.

[31] Viehof, M.: Objektive Qualitätsbewertung von Fahrdynamiksimulationen (2018). a: pp. 13-14.; b: p. 46.

[32] Schaermann, A.: Systematische Bedatung und Bewertung umfelderf. Sensormodelle (2020), p. 17.

[33] Riedmaier, S. et al.: Unified Framework and Survey for Model VV&UQ (2020), p. 4.

[34] Sargent, R. G.: Verification and validation of simulation models (2007), p. 126.

Accordingly, Viehof finds that simulation validity (Sargent: Operational validity) is the superposition of model validity (Sargent: Conceptual model validity) and data / parameter validity, whereas singularities can happen, where invalid model and parameters lead to a valid simulation.[31a] Consequently, by comparison with real data, only the combination of model and parameterization is possible to validate.

Data / parameter validity refers to the already defined model input uncertainty. Model validity refers to the model form uncertainty defined above. Therefore, verification aims for a minimal numerical uncertainty during simulation, the last of the three sources of uncertainty.

Furthermore, in line with Popper, Viehof explains that global simulation validity does not exist, but only global falsification. However, he states that for singular samples, *"Stichprobenvalidität"* (English: sample-validity) is possible to show.[31b] In consequence, he defines **sample-validity** as the state that validity could not be falsified within an empirical series of sample experiments.

1.2.7 Model Uncertainty Quantification (UQ)

Terms are misleading if not used carefully when the topic is APSS model development and validation regarding its fidelity in replicating the measurement uncertainty. In this dissertation, uncertainty does not refer to the three often cited kinds of uncertainty in machine perception defined by Dietmayer[35], namely state, existence, and class uncertainty. Even if they seem to be independent in the first place on object level, they are connected at the thresholding step for detection identification from the signal.

In sensor modeling, especially on signal propagation and processing level, the key difference to classical physical modeling is to mimic actual measurement behavior of a device and not purely physics itself. The sensor's accuracy depends on extra influence parameters with respect to parameters that physically influence the system response quantity (SRQ). Consequently, the sensor's bias and scattering error must be modeled in addition to correct physical modeling.

Due to this work's overall context of modeling for simulation-based safety validation of automated driving, a risk assessment point of view is introduced here. Therefore, the model uncertainty must be quantified to enable model credibility assessment. The model uncertainty to replicate the measurement uncertainty is to be specified in the model requirements and estimated by model validation studies as basis of UQ. Measurement bias with respect to the reference that tries to capture the GT has to be replicated by a sensor model, while a model bias in form of a deviation to the untrue measurement itself should be avoided by the modeler. Fig. 1-6 illustrates the difference between the two types of bias, while indicating that there is a difference between the measurement's and the simulation's scattering error as well.

[35] Dietmayer, K.: Predicting of Machine Perception for Automated Driving (2016), pp. 412-413.

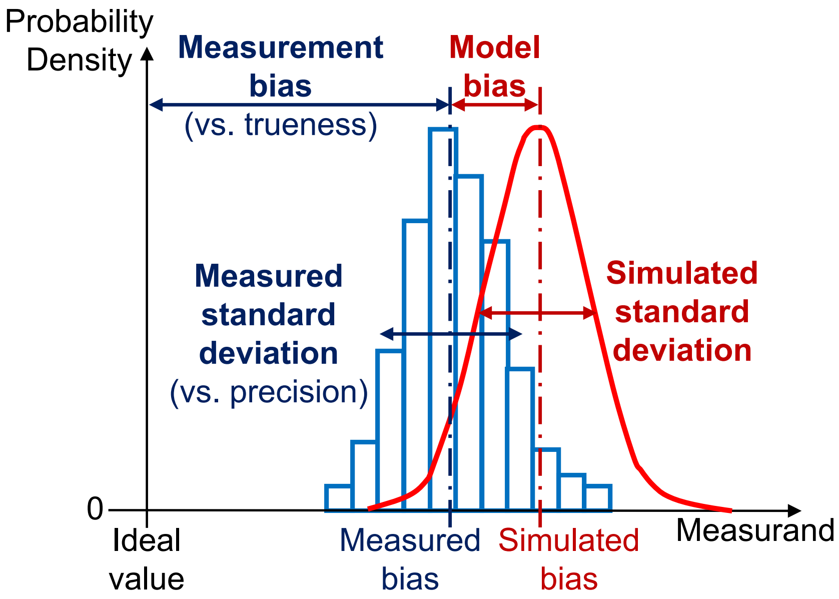


Figure 1-6: Relationship between measurement and model bias

For the mentioned risk assessment context, Roy and Oberkampf classify model uncertainty into epistemic and aleatory:

- **Epistemic model uncertainty** exists because of *"lack of knowledge by the modelers, analysts conducting the analysis, or experimentalists involved in validation."* [36]
- **Aleatory model uncertainty** means the *"inherent variation in a quantity."* [36]

Due to its character, epistemic uncertainty should be minimized, especially when reference data is collected for replay-to-sim model validation. Aleatory uncertainty is collected automatically during measurement and reference data recording over time and is expressed as the measurement scattering error.

Besides the two already mentioned types of uncertainty, Roy and Balch identify three sources of model uncertainty:

- **Model input uncertainty** originates in *"not only parameters used in the model of the system, but also data from the description of the surroundings (e.g., boundary conditions)."* [37a]
- **Numerical uncertainty** has its source in *"discretization error, iterative error, round-off error, and errors due to coding mistakes."* [37b]
- **Model form uncertainty** exists due to the overall modeling approach chosen[37c] and e.g. the physical equations used or the neuronal network design to learn a data-driven model.

[36] Roy, C. J.; Oberkampf, W. L.: Framework for verification, validation, and uncertainty (2011), p. 2132.

[37] Roy, C. J.; Balch, M.: A holistic approach to uncertainty quantification (2012). a: p. 366.; b: p. 368.; c: p. 369.

1.2.8 Model Credibility and Maturity

Although the terms **model credibility** and **model maturity** have already been mentioned in Sec. 1.1, they need to be defined for the rest of this dissertation. Maturity of scientific computing is based on credibility, considering several factors as e.g. in case of the PCMM shown in Fig. 1-1. They are in the order that is requested by the PCMM geometric fidelity, physics and material fidelity, code and solution verification, model validation and UQ and sensitivity analysis. While credibility is assessed as a single state in time, maturity considers the process that leads to this credibility. The technology readiness levels (TRL) pioneered by NASA in the late 1980s are very similar, but they only apply for hardware.[38a]

NASA defines model credibility shortly as the *"quality to elicit belief or trust in M&S results."* [39] Oberkampf and Roy discuss credibility in scientific computing starting from a broader view on the term.[38b] Starting from the motivation of an individual to trust a simulation, they switch the perspective to a credibility from a viewpoint of a whole project team or even the public.

On the examples of the NASA Space Shuttle and US Nuclear Regulatory Commission, Oberkampf & Roy define credibility as the degree to which a project manager would bet his career or company on the results or an analyst would bet public's safety and catastrophic damage to the environment on them. If one considers safety validation of ADS to be the same kind of high-consequence that has *"major consequences beyond the project itself"* [38c], the same degree of credibility is demanded for simulation-based safety validation as for Space Shuttles or nuclear power plants.

Therefore, credibility is used in the following work in that sense and has no alternative when using simulation as a serious tool from a scientist's point of view, as Roache already stated in the IEEE Journal of Computing in Science & Engineering in 2004 (cited later by Oberkampf & Roy in 2010[38b]) when asking *"Is Western Culture at Risk?"* [40]

"In an age of spreading pseudoscience and anti-rationalism, it behooves those of us who believe in the good of science and engineering to be above reproach whenever possible. Public confidence is further eroded with every error we make. Although many of society's problems can be solved with a simple change of values, major issues such as radioactive waste disposal and environmental modeling require technological solutions that necessarily involve computational physics. As Robert Laughlin noted in this magazine, "there is a serious danger of this power [of simulations] being misused, either by accident or through deliberate deception." Our intellectual and moral traditions will be served well by conscientious attention to verification of codes, verification of calculations, and validation, including the attention given to building new codes or modifying existing codes with specific features that enable these activities." [40]

[38] Oberkampf, W. L.; Roy, C. J.: Verification and Validation in Scientific Computing (2010). a: p. 698.; b: pp. 8-15.; c: p. 11.

[39] U.S. National Aeronautics and Space Administration: NASA-STD-7009A (2016), p. 11.

[40] Roache, P.: Building PDE codes to be verifiable and validatable (2004), p. 38.

1.3 Methodology towards Credibility in APSS Simulation

The holistic view in this work on VV&UQ to gain credibility in APSS simulation reflects the sequential nature of this extensive process. It has originally been stated by Trucano et al.[41] and was confirmed later by Oberkampf and Roy in their standard work on V&V[42]. The process is composed of the nine steps starting from specification and ending with documentation, as shown in Fig. 1-7. While being mainly linear, code (software) and overall solution verification with computation of SRQs are accompanying it from aside.

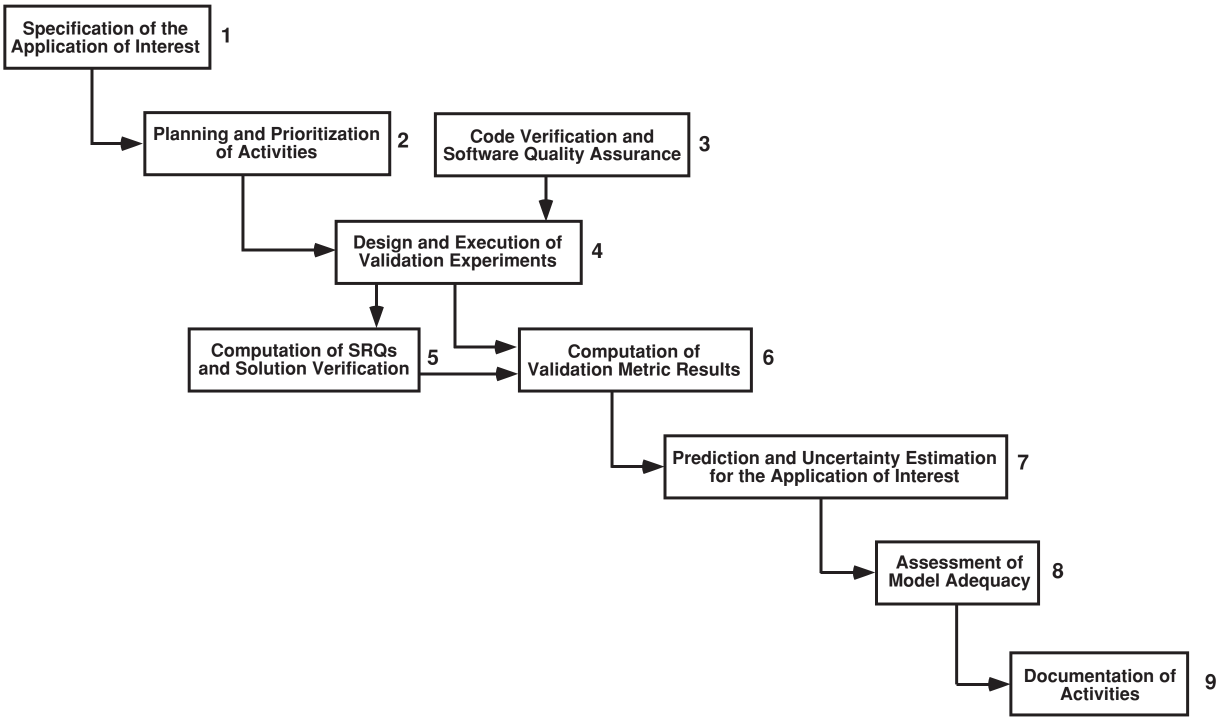

Specification of the Application of Interest
1
Planning and Prioritization of Activities
2
Code Verification and Software Quality Assurance
3
Design and Execution of Validation Experiments
4
Computation of SRQs and Solution Verification
5
Computation of Validation Metric Results
6
Prediction and Uncertainty Estimation for the Application of Interest
7
Assessment of Model Adequacy
8
Documentation of Activities
9

Figure 1-7: Integrated view of the elements of verification, validation, and prediction from Oberkampf and Roy[42] (adapted from Trucano et al.[41]). Reproduced with permission of The Licensor through PLSclear.

In the second chapter the components and categories of APSS models are described to ensure the basic knowledge on the systems to be validated later on. The third chapter describes the exemplary implementation of an APSS, the reflection-based lidar model that serves as application example. Afterwards, in Chap. 4 the state of the art (SotA) of model validation and metrics (step 6) is provided followed by an interim conclusion that leads to the actual challenges towards confidence in APSS simulation in Chap. 5. These challenges identify specification (step 1), experiments (step 4), and prediction (step 7) as major challenges, besides other important remarks. The SotA of APSS model validation and the further challenges lead to the central research questions in this work in Chap. 6 on metrics for specification and VV&UQ of APSS simulation.

[41] Trucano, T. G. et al.: General Concepts for Experimental Validation of ASCI Code Applications (2002), p. 17.

[42] Oberkampf, W. L.; Roy, C. J.: Verification and Validation in Scientific Computing (2010), p. 59.

To answer them and start filling the lack of confidence in APSS simulation, the main part of this dissertation consists of the corresponding answers:

1. A modular APSS model architecture and the two methods PerCollECT and CEPRA as basis of the fidelity requirements in Chap. 7.
2. A methodical evaluation of metrics for APSS model VV&UQ from the long list of metrics applied in literature collected in Chap. 8.
3. Further developed metrics tailored for sample validation of APSS simulation when real data repeatability and reproducibility (R&R) is limited to enable model error quantification, uncertainty aggregation and its prediction for credible simulation application in Chap. 9. metric.

The theory from these main chapters is then applied in Chap. 10 on lidar detections as exemplary APSS simulation to show its usage and value. Thereby, the steps 4 to 8 from the holistic view on VV&UQ in Fig. 1-7 are exemplary described for the novel Finally, in Chap. 11 a conclusion and outlook is provided to path the way for further progress towards credibility of APSS simulation and to finally enable actual safety validation of DAS or even ADS in simulation.

2 Components and Categories for APSS Simulation

Having the most important terms defined, the purpose of the following chapter is to present the functional structure of APSSs and the currently applied categories to prepare specification and validation later in this work. Additionally, some implementation details are described to provide the necessary background knowledge for the later described specification and validation concepts.

At first, functional blocks and interfaces are identified by functional decomposition and secondly, the categories for sensor models by approach, fidelity, and input are discussed. Thereby, as the chapter's title already suggests, in this work and in the following chapter, only APSSs are in focus. Of these, lidar sensors are explained in more detail, as they will later serve as application example for model specification and VV&UQ.

2.1 Functional Decomposition of APSS

The high-level principle of automated driving functions is the well-known sense-plan-act scheme, as e.g. summarized by Amersbach[43a]. This work mainly focuses on the sense task and its simulation as input for testing the remaining tasks. In automobiles, sensing can be fulfilled using APSSs like ultrasonic sensors, radars or lidars that send out signals into the environment. Besides, passive perception sensors like cameras or microphones can be used that collect signals originating elsewhere, but are out of scope in this work.

Amersbach decomposes automated driving functions into six functional layers.[43b] Layer 0 is named *"Information Access"*, describing he generally available information. Layer 1, called *"Information Reception"*, means everything that is perceived by sensors, ending with sensor *"raw data"*. Layer 2 is the *"Information Processing"* including not only object detection and tracking, but sensor fusion and building the overall environment model within the autonomous vehicle. Therefore, all three layers at least partly tackled by APSSs (simulation) and their interaction with the environment. This means that the actual driving function that is to be validated for its safety using simulation models actually starts with layer 3: *"Situational Understanding"* or in the middle of layer 2, depending on whether the sensor fusion is architecturally included in the driving function or not.

[43] Amersbach, C. T.: Functional Decomposition (2020). a: pp. 41-44.; b: pp. 52-60.

2.1.1 Functional Blocks of APSS

As already introduced in previous works of the author, the term “system” in APSS is used to include not only a sensor’s front-end, but some data processing steps as well.[44,45] So, the term **system** in active perception sensor system stresses the inclusion of some processing and the at least partial integration of layer 2. As lidar sensor systems are the application example to present and discuss the methods and metrics for sample validation and specification of accuracies and uncertainties in this work, Fig. 2-1 shows the decomposition of a lidar sensor system for object detection, as described in earlier work of the author.[44] There, the front-end is about emitting signals in the environment and measuring their echoes and the processing unit includes all steps to obtain a list of tracked and classified objects.

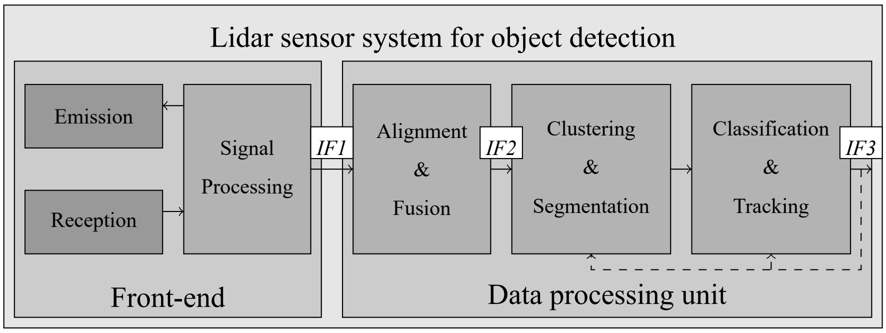

Lidar sensor system for object detection
Emission
Reception
Signal Processing
IF1
Alignment & Fusion
IF2
Clustering & Segmentation
Classification & Tracking
IF3
Front-end
Data processing unit

Figure 2-1: Lidar sensor system for object detection from Rosenberger et al.[44] ©IEEE 2019

Following the more generic functional decomposition from later work of the author[46] as depicted in Fig. 2-2, APSS models are divided into sensor **front-end**, **data extraction**, and **object tracking**. The front-end contains signal transmission, signal propagation channel, signal reception, and signal processing. Data extraction implements detection sensing and detection fusion. The last block represents all data processing to achieve a list of tracked and classified objects at the output of the sensor system including clustering of detections, classification, tracking, and prediction.

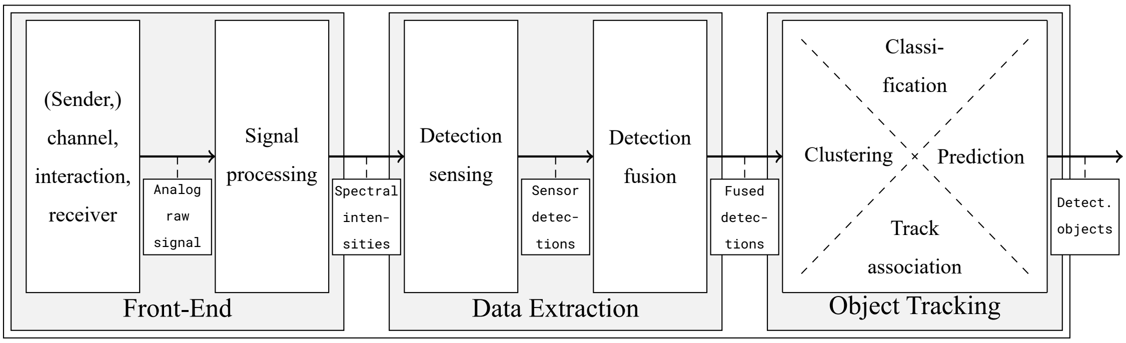


Figure 2-2: Generic functional decomposition of APSS from Linnhoff et al.[46]
Reproduced with permission from Springer Nature.

[44] Rosenberger, P. et al.: Benchmarking and Functional Decomposition of Lidar Sensor Models (2019), p. 2.

[45] Rosenberger, P. et al.: Functional Decomposition of Lidar Sensor Systems (2020), p. 139.

[46] Linnhoff, C. et al.: Highly Parameterizable Perception Sensor Model Architecture (2021), p. 6.

2.1.2 APSS (Model) Interfaces

While the term “raw data” has already been used in the previous section, this spot will be the last time mentioning it in this work (besides citations), as it is often used, but ambiguous. In lidar sensor context, some people mean analogue signals with it, others call data after analogue-to-digital conversion “raw”, while others have a point cloud in mind, so this confusing term will be avoided in the following. In Fig. 2-1 above and its source paper[44], *IF1* was called “raw scan” to differentiate it, but it still remains indeterminate. The term “analog raw signal” in Fig. 2-2 and its source paper[46] for data at the very beginning of any signal processing seems quite clear in the first place. As every receiver hardware already influences the signal and probably filters or amplifies it, “raw” is still a misleading term here.

It is not correct that every APSS front-end contains analogue information, as recent developments in lidar receiver units towards digital lidars with time-correlated single-photon counting (TCSPC) using single-photon avalanche diode (SPAD) arrays or silicon photomultiplier (SiPM) arrays for photon counting proof. Within the front-end, instead of the analog raw signal depicted in Fig. 2-2, a histogram of photons over time bins is obtained, as shown in Fig. 2-3 by Gupta et al.[47]

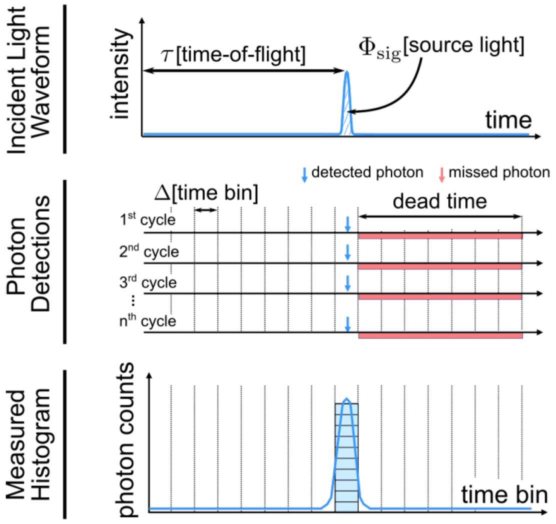


Figure 2-3: SPAD-based pulsed lidar histogram of the time-of arrival of incident photons over multiple laser pulse cycles from Gupta et al.[47] for idealized measurements without ambient light. © IEEE 2019

Even if one could think of a standardization of an interface for such histograms over time bins per beam, current data rates in real sensor systems do not allow that amount of data, so ISO 23150[48a] does not provide it. In its current release v3.4.0 - OSI “Gallant Glock”, ASAM Open Simulation Interface (OSI)[49] would allow to transfer such data in the `osi3::LidarSensorView::Reflection`, but here data rate prohibits its current usage as well.

[47] Gupta, A. et al.: Photon-Flooded Single-Photon 3D Cameras (2019), p. 3.

[48] International Organization for Standardization: ISO 23150:2021(E) (2021). a: p. 2.; b: p. 3.

[49] Hanke, T. et al.: A generic interface for the environment perception of automated driving functions (2017).

The term **detection** is defined in the international standard ISO 23150 as a *"sensor technology specific entity represented in the sensor coordinate system based on a single measurement of a sensor."* [48a] It is used for data at *IF1* from Fig. 2-1 between sensor front-end and data processing unit within the APSS in the following. In the generic decomposition shown in Fig. 2-2, detections are first available in between the data extraction block. Using the ISO 23150 term avoids potentially misleading terms like "low-level data" or "point cloud" and is well defined for radar, lidar, camera, and ultrasonic sensors.

Terms that will be avoided in the following, are "high-level data" or similar, which could mean object lists, but e.g. occupancy grids or voxels as well. Instead, it is clearly stated what data or interface is meant exactly. Here, again, ISO 23150 helps by setting internationally agreed definitions. As it is of interest for the following thesis, the term **object** is exemplary taken here, as has been done for *IF3* in Fig. 2-1 above and the corresponding paper[44]. ISO 23150 defines it as a *"representation of a real-world entity with defined boundaries and characteristics in the vehicle coordinate system."* [48b]

The ISO 23150 defines the term **feature** for data derived from detections that do not (yet) represent an object. It describes a *"sensor technology specific entity represented in the vehicle coordinate system (3.7.16) based on multiple measurements."*[48b] It fits quite well to the "point cloud" called data at *IF2* in Fig. 2-1 above and for the fused detections after the data extraction block in Fig. 2-2 and is therefore used for both in the following.

Besides detections and objects, several other interfaces within APSSs have been identified in previous work of the author for lidar[44,45] and radar[50]. Holder e.g. shows the information flow and the available data for radar from reflection to detection, target, and object.[51] Nevertheless, the focus of this work will be on simulation of detections and the validation of such, therefore there will be no further explanation on other interfaces here.

In its current release v3.4.0 - OSI "Gallant Glock", the field names for data in APSSs of the ASAM OSI[49] standard do not fit completely to ISO 23150. While there are `DetectedObjects` and `LidarDetectionData`, the lidar detections are part of `FeatureData`, which is in contrast to the ISO 23150 definitions. In addition, the features from ISO 23150 are called `LogicalDetectionData` in OSI. Nonetheless, due to the already ongoing alignment of OSI to the ISO 23150, this contradiction is expected to be eliminated in future OSI releases starting from v4.0 on.

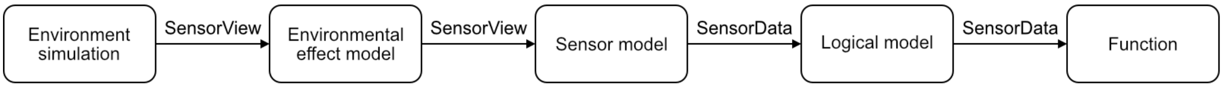


Figure 2-4: Model types and interfaces as defined in the official OSI documentation[52]

[50] Holder, M. F. et al.: Measurements revealing Challenges in Radar Sensor Modeling (2018).

[51] Holder, M. F.: Synthetic Generation of Radar Sensor Data for Virtual Validation (2021), p. 18.

It is possible in OSI to include environmental effects during signal propagation within a single sensor (system) model that consumes a `SensorView` and outputs `SensorData`, as depicted in Fig. 2-4. Still, the OSI standard's official documentation[52] names two other types of models: The environmental effect model that has a `SensorView` as in- and output and the logical model mimicking data processing algorithms that has `SensorData` as in- and output.

2.2 Categorization of APSS Models

Before different modeling approaches and implementations are presented, at first the models are to be categorized and their possible intended usage is presented.

2.2.1 Categorization by Input and Output Data

The first categorization results naturally from the input and output data of the model. As described in the previous section, the ASAM OSI standard clearly defines the interfaces, which supports a clear categorization in this case. E.g. for lidar, OSI contains in its current release 3.4 the data types `Object`, `Reflection`, `Detection`, and `LogicalDetection`.

Having these data fields in mind, the author recently introduced a naming scheme with Linnhoff et al.[53] categorizing by in- and outputs that has already been adopted within the simulation-based safety validation community as `<input>-based <technology> <output> model`. For example, **object-based** and **reflection-based object models** are differentiated. In tradition of the referenced work, the naming scheme is applied here as well.

Fig. 2-5 shows a screenshot of the Institute of Automotive Engineering (FZD) reflection-based lidar object model[54] in action connected to CarMaker from IPG Automotive as an example. The simulation tool, in this case IPG CarMaker, on the left side of the picture generates the reflections and sends them together with the GT object list to the APSS model. This model then computes detections (color scale), the actual lidar point cloud, from the reflections (gray scale) and finally identifies detected objects (blue) in this point cloud that are compared to the GT objects (green).

2.2.2 Categorization by Modeling Approach

There have been several attempts in the past, some motivated by different publicly funded research projects, to categorize APSSs simulation by modeling approach. Even though already stated by

[52] ASAM e.V.: ASAM OSI® (Open Simulation Interface) - Official Documentation (2022).

[53] Linnhoff, C. et al.: Refining Object-Based Lidar Sensor Modeling (2021), pp.24239-24240.

[54] Rosenberger, P. et al.: Reflection Based Lidar Object Model (2022).

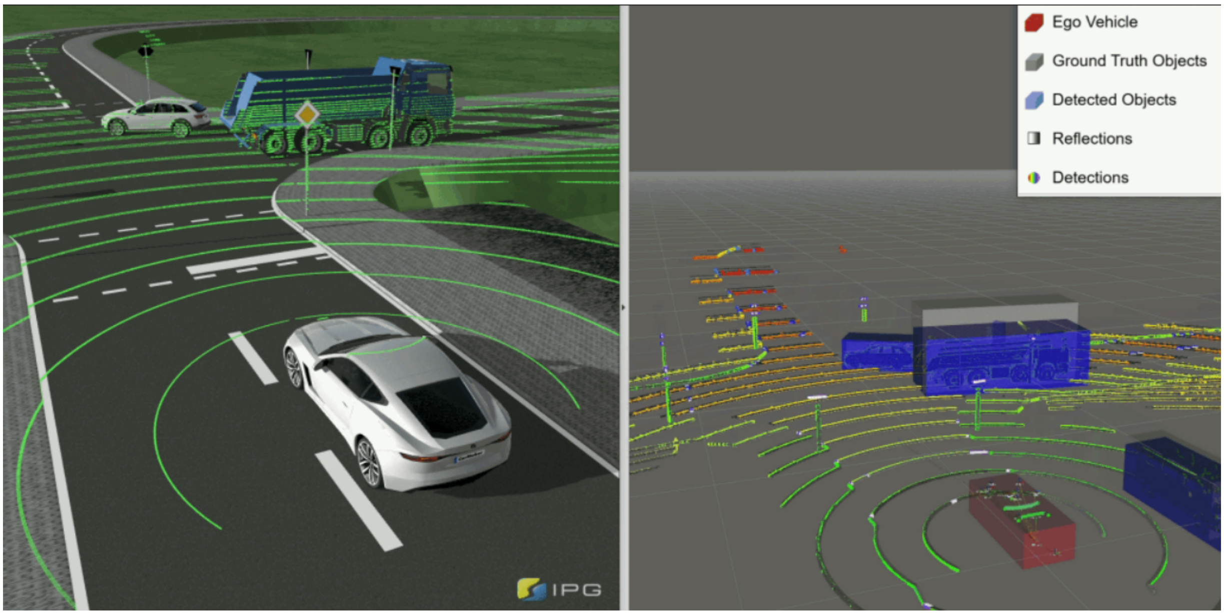


Figure 2-5: FZD Reflection-based lidar object model[54] in action
with GT objects and reflections from IPG Automotive's CarMaker

the author some years ago, still *"there is no clear separation when to call a model gray-box or black-box, stochastic, phenomenological, or data-based."* [55]

The often used terms "white-box", "gray-box" or "black-box" depict the level of provided knowledge about the models. In contrast, the terms **idealized**, **phenomenological**, **stochastic**, **data-driven**, or **physical** appear regularly when APSS models are categorized by modeling approach. The terms are widely accepted within the community, as shown in the survey by Schlager et al.[56] or in previous work of the author[57] as the source for Tab. 2-1. GT "models" that only transform all objects from world to sensor coordinates do not actually reflect any sensor and are therefore omitted in the following.

Before introducing the named model categories in more detail, it must be stated that very often APSS models consist of a mixture of modules from all kinds of these categories. E.g. on the example of Fig. 2-2, the front-end could be modeled physically, data extraction could be simulated ideally and object tracking could be done on a stochastic basis. While Tab. 2-1 already names phenomenological APSS models as combination of stochastic and physical model parts, the former example points out that there are even models of three types of modules.

Idealized APSS Models

In early development phases, if no fidelity requirements on the APSS model are present, **idealized APSS models** are implemented, neglecting any physical effects on the signal and any information

[55] Rosenberger, P. et al.: Benchmarking and Functional Decomposition of Lidar Sensor Models (2019), p. 4.

[56] Schlager, B. et al.: State-of-the-Art Sensor Models for Virtual Testing (2020), p. 238.

[57] Rosenberger, P. et al.: Towards a Generally Accepted Validation Methodology for Sensor Models (2019), p. 3.

Table 2-1: Categorization of sensor data generation by modeling approach.
Source: Rosenberger et al.[57] (with slight adjustments)

	GT "models"	Idealized APSS models	Phenomenological APSS models	
			Stochastic	**Physical**
Principle	Transformation of global GT in sensor perspective GT	Perfect FoV (only values the sensor can actually measure)	Phenomena from data (probabilistics & statistics)	Cause-effect chains leading to phenomena (e.g. signal attenuation, reflection, absorption, transmission, etc.)
Possible specifi-cations	None (only transformation)	Position, orientation, idealized FoV	False detections (FP/FN), noise, pollution, manipulation, ...	Wave lengths, material properties, surfaces, signal processing, ...
Sensor accuracy	Not modeled	Not modeled	Realistic stochastic	Realistic single measurements
Complexity	None	Very small	Depends on parameter space	Very high to infinite

loss on the data. E.g. the term "idealized object-based APSS object model" means that the simulation's original GT object list on the model's input is just transformed from world to vehicle coordinate frame. As already described by von Neumann-Cosel[58], idealized APSS models simply select all objects inside their perfect field of view (FoV) with maximum measurement range and angular range from the real APSS's specification sheet, which means e.g. no occlusion is considered.

Nevertheless, idealized APSS models are not limited to object list output, but can output detections, too. E.g. the results of a simple lidar detection model using basic, "vanilla" ray casting as described in former work of the author[59] could be called idealized detections, as there are no signal propagation effects implemented besides of the real APSS's beam pattern as singular rays for basic hit-point calculation.

Stochastic APSS Models

Stochastic APSS models or in other words data-driven APSS models learn stochastic processes using statistics from real data, like false positive and false negative (FN) rates from real object lists, in different scenarios and replay them during simulation of similar scenarios. As a prominent example, noise models for object poses etc. or detection's positions or intensities for instance are typically built in a stochastic manner. Simpler models for manipulation, blockage, or failure of APSS are mostly of stochastic nature as well.

Still, parameter spaces in the field of APSS are typically huge having many dimensions, for instance for all different environmental conditions in which the APSS are operated. As bigger and

[58] Neumann-Cosel, K. von: Virtual Test Drive (2014), p. 88.

[59] Rosenberger, P. et al.: Sequential lidar sensor system simulation (2020), p. 192.

more complex the parameter space gets, the more data is needed to learn the stochastics, leading to a tipping point at which physical modeling could be the more efficient way of modeling.

Therefore, the complexity of the modeling approach highly depends on the parameter space it should cover and the entry in Tab. 2-1 has been changed compared to the original publication by the author[57].

Pure stochastic APSS models often are considered to be always object models, like the statistical object-based lidar object model from measurement runs by Hirsenkorn[60a] or the statistical object-based lidar object error model by Hanke[61a]. However, there are examples for stochastic APSS detection models like the statistical, non-parametric object-based lidar detection model by Hirsenkorn[60b] and the stochastic reflection-based radar detection model by Eder et al.[62]

While stochastic modeling is already data-driven and therefore often uses machine-learning to extract the statistics from data, there are models even using deep learning for mimicking the physics between real APSS's in- and outputs like the Deep Stochastic Radar Models by Wheeler et al.[63] In this particular case, the power loss following the radar range equation is learned with conditional variational autoencoders trained with both autoencoder loss and adversarial loss. It is remarkable that in this case not only the location of the detections from objects are reproduced, but also the detections' power and even roadside clutter, as well.

Phenomenological APSS Models

Physics are modeled using physical equations and relationships that are derived in a phenomenological manner. In consequence, there is a fuzzy interpretation and no clear distinction what is meant with physical or phenomenological modeling and how both terms are separated. Both terms stress the objective to formulate real world's behavior by inspection. To clarify the fuzziness at least for the scope of this work, as already stated in Tab. 2-1, the term **phenomenological APSS model** is taken as umbrella term for mixture models with physical and stochastic parts.

There is e.g. the often cited phenomenological radar object model by Bernsteiner et al.[64] It combines geometrical equations for modeling the detection's locations with stochastic noise models on these locations and the detections' signal strengths. To mention APSS models of other output besides detections and objects, there are even phenomenological models for data before detection level, as e.g. described by Slavik and Mishra[65].

[60] Hirsenkorn, N.: Modellbildung und Simulation der Fahrzeugumfeldsensorik (2018). a: pp. 22-51; b: pp. 52-91; c: pp. 92-110.

[61] Hanke, T.: Simulated Environmental Perception for Automated Driving Systems (2020). a: pp. 33-64; b: pp. 65-96.

[62] Eder, T. et al.: Data Driven Radar Detection Models (2019).

[63] Wheeler, T. A. et al.: Deep stochastic radar models (2017).

[64] Bernsteiner, S. et al.: Radar Sensor Model for the Virtual Development Process (2015).

[65] Slavik, Z.; Mishra, K. V.: Phenomenological Modeling of Millimeter-Wave Automotive Radar (2019).

Physical APSS Models

As stochastic APSS models are often confused with object models, physical APSS models are mainly considered to be detection models. Again, both and even more output data levels are possible. However, it can be stated that purely **physical APSS models** are rare and often concentrated on signal propagation modeling. In most cases, there are physical parts inside phenomenological models, as the geometry module by Bernsteiner et al.[64]

In other words, the rendering of the scene for the APSS wavelength, resolution, and sensitivity is physical modeling. However, there are several different approaches to implement such rendering. There is e.g. the so-called Z-buffer method that renders objects in optical simulation including geometrical occlusion calculation. It consists of projection of visible object geometries onto a cylindrical or spherical surface around the sensor and clipping of the shapes. The object-based lidar model from former work of the author[66] that is publicly available open source[67] could serve as an example in case of lidar simulation.

In radar modeling, there are approaches for complexity reduction and efficiency, called reflection center models. As described e.g. by Danielsson[68] or Cao[69], where objects and their complex $360°$ scattering profiles are simplified into one or multiple reflecting points. The reflecting points can be efficiently transformed into the sensor coordinate frame and further signal attenuation or even more complex effects can be superimposed on the radar's power attenuation equation[70].

Z-buffer and reflection center models are very efficient, as points just need to be transformed into the sensor's coordinate frame to calculate their spherical coordinates in this frame. In contrast, there are approaches where a direct or multiply-bouncing geometrical path is searched from the sensor's transmitter (the sensor frame's origin) onto the object. These approaches are called beam/ray casting/tracing depending on whether infinitesimally thin rays or volumes are "shot" and on whether the path is just cast until the hit point at the object or if it is traced back to hit the sensor front-end again. Meanwhile, especially ray casting is widely used in commercial and open source APSS simulation because of the ability of paralleling the calculation of the rays. This is supported by the advancements of graphics processing unit (GPU) technology for this exact purpose and for machine learning tasks in the last decade. Currently, even central processing units (CPUs) have adopted this trend for enhanced support of paralleling computations and are capable to support ray tracing, as well.

To name some physical APSS models using ray casting in chronological order without claim for completeness, there are detection models for radar by Gubelli et al.[71], Hirsenkorn[60c], Thiel-

[66] Linnhoff, C. et al.: Refining Object-Based Lidar Sensor Modeling (2021), pp. 24240-24242.

[67] Linnhoff, C. et al.: Object Based Generic Perception Object Model (2022).

[68] Danielsson, L.: Tracking and radar sensor modelling for automotive safety systems (2010).

[69] Cao, P.: Modeling Active Perception Sensors for Real-Time Virtual Validation (2018).

[70] Winner, H.: Automotive RADAR (2016), p. 331.

[71] Gubelli, D. et al.: Ray-Tracing Simulator for Radar Signals Propagation in Radar Networks (2013).

ing et al.[72], and Holder et al.[73]. The last is called Fouriertracing, based on the work of Linnhoff[74], as it outputs data before detection level, right after the fast Fourier transform (FFT).

Ray casting is not exclusive for radar, but in fact mostly applied for lidar simulation, as in the models by Hanke[61b], and Tamm-Morschel[75], while there are many more approaches as e.g. included in the benchmarking of lidar models in earlier work of the author[76].

Physical modeling is not limited to rendering, instead there are reflection-based models or model parts as described by Cao[69] or by Prinz et al.[77,78], where physical radar theory is superimposed on beforehand calculated radar reflections to model interference. In such reflection-based approaches, the way of rendering or reflection calculation does not matter for the subsequent calculations. As reflections or ray casting results are provided by many simulation engines, OSI already provides a standardized description for reflections and there are open source models like the ones provided by the research group of the author[79,80].

It should be stressed that in contrast to stochastic modeling, where only minimal comprehension about the underlying cause-effect chains leading to the modeled phenomena is necessary to extract these, physical modeling demands high efforts in comprehension and e.g. hypothesis testing to validate every single physical equation. While it is possible to use stochastic models and model parts in many use cases, physical modeling ensures higher maturity by design.

2.2.3 Categorization by Fidelity

The presented categorization of sensor models by approach transports a (possibly misleading) impression of model fidelity. Fidelity is the aim of the modeling approach and must be validated, but it is not the direct consequence. Especially when it comes to phenomenological approaches, it is not generally true that stochastic APSS models show lower fidelity than physical APSS models, even if it could be guessed in the first place.

To clarify fidelity levels, Schlager et al. divide in their survey into **low fidelity**, **medium fidelity**, and **high fidelity** APSS models listing different criteria from operating principles to v-model phases, as shown in Tab. 2-2[81a].

[72] Thieling, J. et al.: Scalable and Physical Radar Sensor Simulation for Interacting Digital Twins (2021).

[73] Holder, M. F. et al.: The Fourier Tracing Approach for Modeling Automotive Radar Sensors (2019).

[74] Linnhoff, C.: Entwicklung eines Radar-Sensormodells (2018).

[75] Tamm-Morschel, J. F.: Erweiterung eines Lidar-Sensormodells (2019).

[76] Rosenberger, P. et al.: Benchmarking and Functional Decomposition of Lidar Sensor Models (2019).

[77] Prinz, A. et al.: Validation Strategy for Radar-Based Assistance Systems (2020).

[78] Prinz, A. et al.: Automotive Radar Signal and Interference Simulation for Testing Autonomous Driving (2021).

[79] Rosenberger, P. et al.: Reflection Based Lidar Object Model (2022).

[80] Elster, L. et al.: Reflection Based Radar Object Model (2022).

[81] Schlager, B. et al.: State-of-the-Art Sensor Models for Virtual Testing (2020). a: p. 239.; b: p. 241.

[82] Hanke, T. et al.: Generic architecture for simulation of ADAS sensors (2015).

Table 2-2: Overview of the properties of low-, medium-, and high-fidelity sensor models by Schlager et al.[81a] Permission conveyed through Copyright Clearance Center, Inc.

	Low fidelity	Medium fidelity	High fidelity
Operating principles	Geometrical aspects	Physical aspects, detection probabilities	Rendering (rasterization, ray tracing, etc.)
Input	Object lists	Object lists	3D scene (meshes)
Output	Object lists	Object lists or "raw data"	"Raw data"
Pros	Low computational power needed	Trade-off btw. computing time and fidelity, a lot of effects can be considered	Most realistic output
Cons	High abstraction level, no realistic output	Lots of training data may be required	High computational power needed
V-model phases	First specification phases	Specification phases in the middle and integration phases	Component specification, implementation and integration phases
Design question	What point(s) or shape represents objects and which need to be in the line of sight for detection?	What point(s) or shape represents objects and which need to be in the line of sight for detection? What effects are considered?	What is the detection threshold? Which effects, material properties, and weather conditions are considered?

[83] Stolz, M.; Nestlinger, G.: Fast generic sensor models for testing highly automated vehicles in simulation (2018).

[84] Muckenhuber, S. et al.: Object-based sensor model for virtual testing of ADAS/AD functions (2019).

[85] Bühren, M.; Yang, B.: Simulation of Automotive Radar Target Lists: Novel Approach of Object (2006).

[86] Bühren, M.; Yang, B.: Automotive Radar Target List Simulation based on Reflection Centers (2006).

[87] Bühren, M.; Yang, B.: A Global Motion Model for Target Tracking in Automotive Applications (2007).

[88] Bühren, M.; Yang, B.: Extension of Automotive Radar Target List Simulation (2007).

[89] Bühren, M.; Yang, B.: Initialization Procedure for Radar Target Tracking (2007).

[90] Bühren, M.; Yang, B.: Simulation of Automotive Radar Target Lists: Clutter and Resolution (2007).

[91] Schneider, R.: Modellierung der Wellenausbreitung für ein bildgebendes Kfz-Radar (1998).

[92] Mesow, L.: Multisensorielle Datensimulation im Fahrzeugumfeld für die Bewertung von Sensorfusion (2007).

[93] Schuler, K.: Intelligente Antennensysteme für Kraftfahrzeug-Nahbereichs-Radar-Sensorik (2007).

[94] Schuler, K. et al.: Extraction of Virtual Scattering Centers of Vehicles by Ray-Tracing Simulations (2008).

[95] Hammarstrand, L. et al.: Adaptive Radar Sensor Model for Tracking Structured Extended Objects (2012).

[96] Hammarstrand, L. et al.: Extended Object Tracking using a Radar Resolution Model (2012).

[97] Cao, P. et al.: Perception sensor modeling for virtual validation of automated driving (2015).

[98] Li, Y. et al.: LiDAR Sensor Modeling for ADAS Applications under a Virtual Driving Environment (2016).

[99] Zhao, J. et al.: Method and Applications of Lidar Modeling for Virtual Testing of Intelligent Vehicles (2020).

[100] Peinecke, N. et al.: Lidar simulation using graphics hardware acceleration (2008).

[101] Hirsenkorn, N. et al.: A ray launching approach for modeling an FMCW radar system (2017).

[102] Maier, F. M. et al.: Environment perception simulation for radar stimulation (2018).

[103] Holder, M. F.: Synthetic Generation of Radar Sensor Data for Virtual Validation (2021).

[104] Eder, T.: Simulation of Automotive Radar Point Clouds in Standardized Frameworks (2021).

Table 2-3: Overview of radar and lidar sensor models for virtual testing of DAS/ADS by Schlager et al.[81b], classified into low-, medium-, and high-fidelity sensor models. (★: Additions by the author)

	Radar	Lidar
Low fidelity	Hanke et al.[82], Stolz and Nestlinger[83], Muckenhuber et al.[84], Hanke[61a]★	
Medium fidelity	Hirsenkorn et al.[60a,b]★	
	Bühren and Yang[85,86,87,88,89,90], Schneider[91]★, Mesow[92], Schuler[93], Schuler et al.[94], Danielsson[68], Hammarstrand et al.[95,96], Cao et al.[97], Bernsteiner et al.[64], Wheeler et al.[63], Cao[69], Eder et al.[62]★	Li et al.[98], Zhao et al.[99] Linnhoff et al.[66]★
High fidelity	Peinecke et al.[100]	
	Gubelli et al.[71]★, Hirsenkorn et al.[101], Hirsenkorn[60c]★, Maier et al.[102], Linnhoff[74]★, Holder et al.[73], Thieling et al.[72]★, Prinz et al.[77,78]★, Holder[103]★, Eder[104]★, Degen et al.[105]★, Maier[106]★	O'Brien and Fouche[107], Goodin et al.[108], Doria[109,110], Gschwandtner et al.[111], Rossmann et al.[112], Wang et al.[113], Gschwandtner[114], Wang[115], Bechtold et al.[116], Hanke et al.[117], Alldén et al.[118]★, Su et al.[119], Fang et al.[120], Woods[121], Hanke[61], Rott[122], Goodenough et al.[123]★

Nevertheless, these three fidelity categories from Schlager et al. are not always applicable for all kinds of models that strictly. Especially the definition of high fidelity starting with rendering as operating principle is up to discussion. The already introduced object-based lidar model from

[105] Degen, R. et al.: Methodical Approach to the Development of a Radar Sensor Model (2021).

[106] Maier, F. M.: Radar Perception Simulation for Automated Driving Tests (2022).

[107] O'Brien, M. E.; Fouche, D. G.: Simulation of 3D Laser Radar Systems (2005).

[108] Goodin, C. et al.: Sensor modeling for the Virtual Autonomous Navigation Environment (2009).

[109] Doria, D.: A Synthetic LiDAR Scanner for VTK (2009).

[110] Doria, D.: SyntheticLidarScanner (2021).

[111] Gschwandtner, M. et al.: BlenSor (2011).

[112] Rossmann, J. et al.: A Real-Time Optical Sensor Simulation Framework for Development and Testing (2012).

[113] Wang, S. et al.: Shader-based sensor simulation for autonomous car testing (2012).

[114] Gschwandtner, M.: Support framework for obstacle detection on autonomous trains (2013).

[115] Wang, S.: State Lattice-based Motion Planning for Autonomous On-Road Driving (2015).

[116] Bechtold, S.; Höfle, B.: HELIOS (2016).

[117] Hanke, T. et al.: Generation and validation of virtual point cloud data for automated driving systems (2017).

[118] Alldén, T. et al.: Virtual Generation of Lidar Data for Autonomous Vehicles (2017).

[119] Su, H. et al.: A Simulation Method for LIDAR of Autonomous Cars (2019).

[120] Fang, J. et al.: Augmented LiDAR Simulator for Autonomous Driving (2020).

[121] Woods, J. O.: GLIDAR (2021).

[122] Rott, R.: Dynamic Update of Stand-Alone Lidar Model (2022).

[123] Digital Imaging and Remote Sensing Lab: DIRSIG (2022).

former work of the author[124] uses geometrical and physical aspects in combination with refined bounding boxes and a modified Z-buffer method for rendering as well as stochastics.

The model has elements from all fidelity categories to achieve validity for modeled sensor effects in terms of FoV and occlusion calculation with lower computational power requirements. In its current state, as indicated in Tab. 2-3 one would call the model's fidelity "medium", but Z-buffer is a rendering method and therefore, the model should be called high fidelity, if the rules by Schlager et al. from Tab. 2-2 would be applied strictly. Still, the model shows the potential to rise to high fidelity and challenge ray tracing as the actual magic bullet in lidar rendering.

Finally, Schlager et al. presented a list of the state-of-the-art in sensor modeling, ordering radar, lidar, and camera models by their fidelity.[81b] Tab. 2-3 contains all models of radar and lidar sensor systems from that survey. To complete the survey of the SotA of modeling, the list is extended by the author (marked with *) with work published afterwards or not considered by Schlager et al.

2.3 Existing Tools for APSS Simulation

An extensive overview of 40 simulation tools was given in 2019 by Kang et al.[125]. Raju and Farah listed 32 tools in September 2021[126]. Recently, Salles et al. published a study on co-simulation with open source tools[127] that discusses CARLA and LG Silicon Valley Lab (SVL) besides listing further open source tools like Microsoft's AirSim, Gazebo, BeamNG or DeepDrive, as well as commercial simulation tools Siemens PreScan, IPG Automotive CarMaker, TESIS veDYNA, AVSimulation SCANeR, dSpace ASM, and NVIDIA DriveSim.

While the automotive simulation market is expected to grow significantly in the next few years, big players like Ansys and Hexagon have already incorporated tool manufacturers, like Hexagon with Vires VTD, or partner with them extensively, as is the case for the whole market these days. Those partnerships make it almost impossible to give a sophisticated actual overview, besides the high dynamic that is brought in by start up companies like AI Motive to name just one of many.

Due to the pressure on established companies through many new players, there is high dynamic under the hood of simulation tools regarding e.g. graphics engines, scene and scenario design, physics simulation, standardization, and many more. Therefore, as a snapshot of the simulation market in early 2022, market consolidation in automotive simulation is as far away as credibility and maturity of APSS simulation, as e.g. described for seven technical issues by El Mostadi et al.[128]

[124] Linnhoff, C. et al.: Refining Object-Based Lidar Sensor Modeling (2021).

[125] Kang, Y. et al.: Test Your Self-Driving Algorithm (2019).

[126] Raju, N.; Farah, H.: Evolution of Traffic Microsimulation and Its Use (2021).

[127] Salles, D. et al.: A Modular Co-Simulation Framework (2022).

[128] El Mostadi, M. et al.: Seven Technical Issues That May Ruin Your Virtual Tests for ADAS (2021).

3 Exemplary Implementation of a Lidar Sensor Simulation

After the different categorization principles and categories, the following chapter gives some implementation details of the exemplary implemented lidar simulation by the author. Besides all actually implemented effects, the most prominent effects and phenomena in lidar modeling that are not implemented are briefly introduced as well. Like the model itself, the explanations now are provided to support the methodological considerations later on. While some cause-effect chains are more suitable to physical modeling, others are favorable for stochastic modeling. As the goal is to obtain a phenomenological model at the end that serves as a reasonable example for model specification and VV&UQ later in this work, only a subset of all possible cause-effect chains has been implemented. The own development of this simulation model was necessary even if there are commercial and open source lidar simulations available, as own implementations are white-box, can be calibrated to available real sensors as necessary for validation, and do not cause unwanted effects in the data without knowledge about their origin.

However, the methodology provided in this work for model specification and VV&UQ is designed for all APSS simulations and the selection of lidar as one of many APSS is to some extent arbitrary at this point. The selection is necessary just to reduce explanation and application effort in this dissertation. Nevertheless, it would have been possible to select e.g. radar or ultrasonic sensor system models as application area for this work, as well. The implementation as described in the following, which has already been demonstrated in Fig. 2-5, has led to a reflection-based lidar object model that has been made publicly available by the author on GitLab[129] open source.

To adhere to the functional structure previously provided by Fig. 2-2, the following sections about implementation details are sorted by functional decomposition blocks. They start with the front-end including signal interaction with objects and the channel as well as beam divergence and temporal behavior, followed by data extraction and ending with object tracking.

3.1 Front-End Modeling

Lidar or “light detection and ranging” means that the sensor transmits (infrared) light pulses and measures the range r to reflecting objects by the signal’s time of flight until it is received t_{of} as $r = \frac{\mathrm{c}\, t_{\mathrm{of}}}{2}$. Therefore, implementation starts with modeling signal emission, interaction, and reception for the correct calculation of the received power, the so-called “laser radar equation”.

[129] Rosenberger, P. et al.: Reflection Based Lidar Object Model (2022).

For hit object ranges r_{obj} much longer than the pulse length in meter $c\tau_{\mathrm{p}}$, the laser radar equation for received laser power from a hit object is valid as e.g. shown by Rasshofer et al.[130a] and reads

$$P_{\mathrm{r}} = P_{\mathrm{t}} \cdot C_{\mathrm{r}} \cdot H_{\mathrm{cnl}}(r) \cdot H_{\mathrm{obj}}(r). \tag{3-1}$$

$$C_{\mathrm{r}} = \eta_{\mathrm{r}} A_{\mathrm{r}} \tag{3-2}$$

is the optical aperture constant including the area A_{r} and efficiency η_{r} of the receiver optic[130a].

$$H_{\mathrm{cnl}}(r) = \frac{\eta_{\mathrm{cnl}}^2(r)\chi_{\mathrm{fov}}(r)}{r^2} \tag{3-3}$$

denotes the range-dependent spatial impulse response of the channel with the total one-way transmission efficiency $\eta_{\mathrm{cnl}}(r)$ squared for out and back multiplied by the crossover function $\chi_{\mathrm{fov}}(r)$ and divided by the squared range, as it is traveled twice, as well[130b]. Channel is used in this context to describe the signal's interaction with fields and particles on its propagation path.

The crossover function $\chi_{\mathrm{fov}}(r)$ for the intersection of the area covered by the receiver's FoV $A_{\mathrm{fov}}(r)$ and the covered area of the transmitted beam $A_{\mathrm{t}}(r)$ at range r is defined as

$$\chi_{\mathrm{fov}}(r) = \frac{A_{\mathrm{fov}}(r) \cap A_{\mathrm{t}}(r)}{A_{\mathrm{t}}(r)} \tag{3-4}$$

and is constant for coaxial transmit/receive optics. However, it strongly depends on the range r for bistatic beam configurations, where transmitted beam and receiver FoV do not overlap completely ($r < R_2$), as shown in Fig. 3-1. In recent lidar sensors, where application-specific integrated circuits (ASICs) are used leading to very low displacement d between emitter and receiver, R_2 becomes relatively small and $\chi_{\mathrm{fov}}(r)$ plays a minor role.

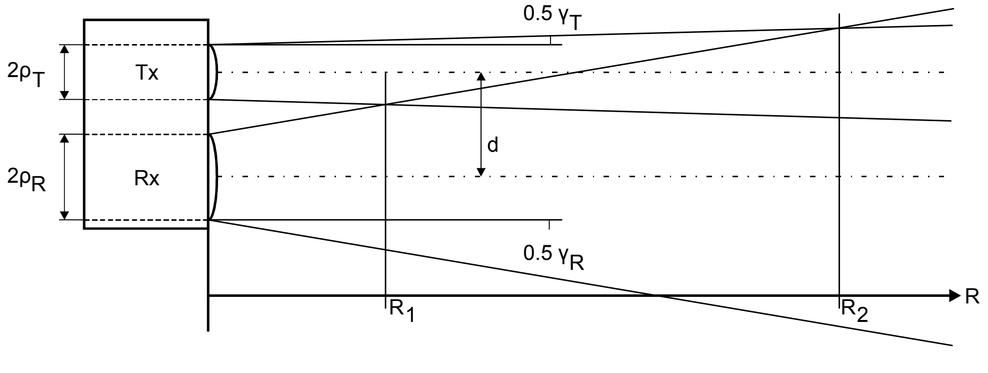


Figure 3-1: Crossover of beams for bistatic optic configuration with parallel optical axis from Rasshofer et al.[130b] d is the aperture displacement. R_1 is the range of first contact between emitted beam and receiver's FoV. R_2 determines where both overlap completely. ρ_{t} and ρ_{r} denote the radii of the transmission and reception aperture. γ_{t} and γ_{r} depict their vertical opening angles.

[130] Rasshofer, R. H. et al.: Influences of weather phenomena on automotive laser radar systems (2011). a: p. 51.; b: p.52.; c: p.53.

The range-dependent spatial impulse response of any hit object

$$H_{\mathrm{obj}}(r) = \beta_0 \chi_{\mathrm{obj}}(r) \tag{3-5}$$

is a product of the differential reflectivity β_0 of the object that e.g. equals Γ/π for Lambertian reflection characteristic $0 < \Gamma \leq 1$, and the crossover function

$$\chi_{\mathrm{obj}}(r) = \frac{A_{\mathrm{obj}} \cap A_{\mathrm{t}}(r)}{A_{\mathrm{t}}(r)} \tag{3-6}$$

of the area of the transmitted laser beam $A_{\mathrm{t}}(r)$ at range r and the object's area in the beams direction of sight A_{obj}.[130c]

Inserting (3-2) - (3-6) into (3-1) leads to the more detailed form of the laser radar equation for the received power

$$P_{\mathrm{r}} = P_{\mathrm{t}} \cdot \eta_{\mathrm{r}} A_{\mathrm{r}} \cdot \frac{\eta_{\mathrm{cnl}}^2(r) \chi_{\mathrm{fov}}(r)}{r^2} \cdot \beta_0 \chi_{\mathrm{obj}}(r) \tag{3-7}$$

However, in the current reflection-based lidar simulation, the ray tracing performed in a second simulation tool which provides options for changing reflectivities and shapes of objects but currently neglects the crossover function for possibly bistatic lidar front-ends (3-4).

3.1.1 Ray Casting / Tracing and Beam Super-Sampling

The SotA for lidar rendering is ray casting or tracing, as it is available in most game engines and modern GPUs are designed for parallel hit point calculation for each ray. While ray casting stops at the first hit point of the ray with bidirectional scattering distribution function (BSDF) calculation, ray tracing starts new rays from this hit point. This then enables to model signal propagation through transmissive objects and multi-path propagation over several reflections back to the receiver, but comes with way higher computation cost and needs high-performance GPUs for parallel computing of millions of rays. For this reason, often simpler ray casting is chosen and not ray tracing, as is the case for the exemplary lidar sensor simulation used here.

Modern lidar sensor simulations like the here exemplary implemented model reproduce the lidar beam pattern via super-sampling of the cone-shaped diverging lidar beams with multiple rays in a brute force manner or with advanced methods like Monte Carlo path tracing. In this work, the lidar sensor simulation as described by Tamm-Morschel [131] is used that contains the brute force super-sampling method. Fig. 3-2 from previous work of the author[132] illustrates the super-sampling for a single lidar beam. As shown, when hitting an edge or two objects behind

[131] Tamm-Morschel, J. F.: Erweiterung eines Lidar-Sensormodells (2019).

[132] Rosenberger, P. et al.: Sequential lidar sensor system simulation (2020), p. 192.

each other within one single beam (A-B and C-D), the real lidar sensor can identify two echoes in the signal, which can only be reproduced by shooting multiple rays per beam to produce hit points a.k.a reflections from different objects and ranges.

On the right side of Fig. 3-2, the green dots depict reflections from super-sampling over the exemplary analogue real world signal with its noise floor in blue. After application of the intensity threshold I_{th} and echo pulse width (EPW) identification, the information from simulation and real world could be identical detections. As marked in Fig. 3-2, EPW in this context depicts the width of a received pulse, as for (A-B) and (C-D), in contrast to the intensity that corresponds to the height of these pulses. The natural drawback of equidistant super-sampling rays is that the inter-ray distance grows with range r and the area corresponding to a single ray grows with r^2, while Monte-Carlo super-sampling or other ray distributions aim to improve this. It should be stated at this point that in contrast to the data from SPAD or SiPM arrays as shown in Fig. 2-3, ray super-sampling is spatial discretization, not time binning, while both seem very similar at first.

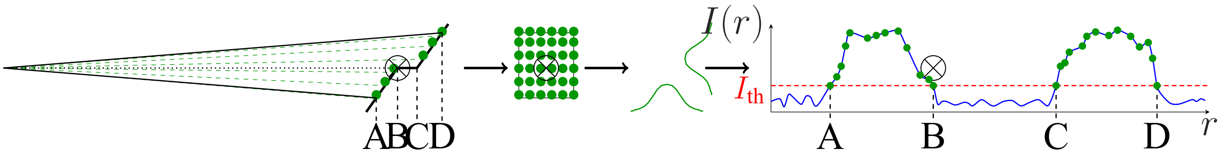


Figure 3-2: Super-sampling of a single lidar sensor beam and beam pattern application on the reflections for detection calculation within the reflection-based lidar sensor simulation from Rosenberger et al.[132]. (A-B) and (C-D) mark the two resulting echoes at the object (left) and in the intensity signal (right). ⊗ is the center of the beam that one would get by single-ray-per-beam ray casting without multi-echo capability.

3.1.2 Range Dependency of the Received Power

Combining the range independent terms in (3-7) leads to similar forms of the laser radar equation, as e.g. published by Gotzig[133a], Schmitt et al.[134a], and in former work of the author[135]. Such simplification, where the range dependency is made clearer, is ensured for the following considerations with the range considered to be long enough for full overlap of the transmitted beam and the receiver's FoV, so that $A_{\mathrm{fov}}(r) \cap A_{\mathrm{t}}(r) = A_{\mathrm{t}}(r)$ and $\chi_{\mathrm{fov}}(r) = 1$.

However, for the question of range dependency of the received power P_{r} from a transmitted lidar beam that gets reflected by a hit object, the non-linear range dependency of $\chi_{\mathrm{obj}}(r)$ from (3-6) is the key to its magnitude. In other words, the area that is covered by the object within the crossover of the transmitted beam and the receiver's FoV determines the range dependency fundamentally.

In the first case, when the hit object is relatively big compared to the beam, e.g. when illuminating a wall or vehicles in short ranges, $\chi_{\mathrm{obj}}(r) = 1$ and the received power $P_{\mathrm{r}} \propto 1/r^2$, as visible in

[133] Gotzig, H.; Geduld, G.: Automotive LIDAR (2016). a: p. 411.; b: p. 410.

[134] Schmitt, J. et al.: Phenomenological, Measurement Based LiDAR Sensor Model (2021). a: p. 427.; b: p.428.

[135] Rosenberger, P. et al.: Analysis of Real World Sensor Behavior (2018), p. 612.

(3-7) and shown e.g. by Schmitt et al.[134b] as

$$P_\mathrm{r} = P_\mathrm{t} \cdot \eta_\mathrm{r} A_\mathrm{r} \cdot \frac{\eta_\mathrm{cnl}^2(r)}{r^2} \cdot \beta_0 \Rightarrow P_\mathrm{r} \propto 1/r^2. \tag{3-8}$$

With higher range between sensor and object, $\chi_\mathrm{obj}(r)$ starts to be range-dependent. The horizontal $\gamma_\mathrm{t,h}$ and vertical beam divergence $\gamma_\mathrm{t,v}$ of the transmitted beam that are prominent content in specification sheets of real lidar sensor systems start to be the crucial factors together with the range r. At intermediate ranges, when only one side of the beam area $A_\mathrm{t} = r^2\,\gamma_\mathrm{t,h}\,\gamma_\mathrm{t,v}$ protrudes from the object area $A_\mathrm{obj} = w_\mathrm{obj}\,h_\mathrm{obj}$, e.g. the upper part for lidars with higher vertical beam divergence than in horizontal direction. In this case, $\chi_\mathrm{obj}(r)$ can be approximated to $\chi_\mathrm{obj}(r) \approx \dfrac{h_\mathrm{obj}}{r\,\gamma_\mathrm{t,v}}$ with $\chi_\mathrm{obj}(r) \propto 1/r$ leading to $P_\mathrm{r} \propto 1/r^3$, as shown by Gotzig[133b], yielding to

$$P_\mathrm{r} \approx P_\mathrm{t} \cdot \eta_\mathrm{r} A_\mathrm{r} \cdot \frac{\eta_\mathrm{cnl}^2(r)}{r^2} \cdot \beta_0 \frac{h_\mathrm{obj}}{r\,\gamma_\mathrm{t,v}} \Rightarrow P_\mathrm{r} \propto 1/r^3 \tag{3-9}$$

For relatively small objects, e.g. at high ranges, when the beam covers the whole object, the received power $P_\mathrm{r} \propto 1/r^4$. In this case, the numerator of $\chi_\mathrm{obj}(r)$ equals A_obj and stays range-independent. Consequently, $\chi_\mathrm{obj}(r) \propto 1/r^2$ leading to

$$P_\mathrm{r} = P_\mathrm{t} \cdot \eta_\mathrm{r} A_\mathrm{r} \cdot \frac{\eta_\mathrm{cnl}^2(r)}{r^2} \cdot \beta_0 \frac{A_\mathrm{obj}}{r^2\,\gamma_\mathrm{t,h}\,\gamma_\mathrm{t,v}} \Rightarrow P_\mathrm{r} \propto 1/r^4 \tag{3-10}$$

Muckenhuber et al.[136] also stress this range dependency to the reciprocal of the range r to the power of two, three or four depending on the relative size of the laser beam compared to the size of the illuminated object, and illustrate it as shown in Fig. 3-3.

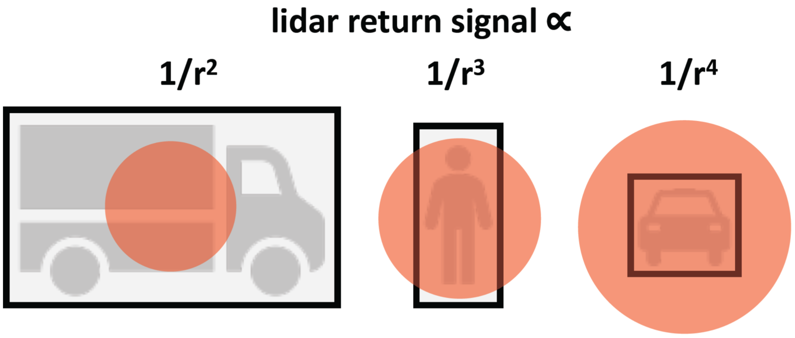


Figure 3-3: Schematic illustration of the range influence on the receivable lidar power depending on transmitted beam size (red circle) compared to the size of hit object (black rectangle) from Muckenhuber et al.[136]

[136] Muckenhuber, S. et al.: Automotive Lidar Modelling Approach Based on Material Properties and Lidar Capabilities (2020), p. 6.

3.1.3 Signal Interaction with Hit Objects

Following up on signal interaction with objects, the power reflected back to the sensor has to be computed to gather realistic data from simulation. In most cases, the absolute value is not of interest, but the proportion of the transmitted power that theoretically can be received. This refers to β_0 from (3-5) that could be described by a BSDF in the general case, or if just the reflection is calculated, by a bidirectional reflectance distribution function (BRDF) and is possible to tune with the exemplary implemented lidar simulation.

In a BSDF, the geometry of the hit surface is considered by the angle between incident ray and normal vector of the surface as interpolation between the mesh's points for the hit point. The surface roughness is considered by the reflection type. Most reflections are super positions of specular and diffuse portions, while often only one type is considered for specific materials for simplification. For correct BSDF calculations, reflectivities for different wavelengths need to be provided in simulation as e.g. lookup tables. While $905\,\mathrm{nm}$ are still most common for real lidar sensors, $885\,\mathrm{nm}$ in Ibeo NEXT, or $940\,\mathrm{nm}$ in Sense Photonics lidar, or even $1550\,\mathrm{nm}$ in AEye lidar sensors exist and should therefore be provided for their simulation.

In case of specular reflections and in case of transmissive materials, new rays should be shot into the simulated scene starting from the hit point in order to obtain all possible signal interactions. In the exemplary model, further tracing of the rays is avoided for shorter computation time on the available hardware. To some extend, every material has multi-path propagation potential, due to the high sensitivity of modern lidar receivers, when e.g. retro-reflective materials exist within the scene at small ranges to the sensor that reflect multiple magnitudes higher power than others. As multi-path propagation and multiple interactions per path could lead to infinite calculation loops, some sort of cut-off criteria must be given as e.g. a maximum number of interactions per path of rays.

While diffraction is considered to play a minor role in lidar measurements, it does have an effect on radar signal propagation and therefore enlarges the simulation challenge for these sensors. Refraction plays a minor role in radar sensor cause-effect chains, but has an impact on lidar measurements from transmissive objects. In all cases, the different absorption proportions for different wavelength and materials matter and are considered by actual BSDFs.

However, as every ray is computed independently and in parallel with ray casting/tracing there is no information for neighbor rays to be able to explicitly calculate a covered area on an object. Nonetheless, the reflected power per ray is always divided by r^2 and the higher-order dependency is implicitly given by super-sampling of the beam and more lost rays at higher ranges when the beam area rises.

3.1.4 Signal Attenuation within the Channel

Besides its general range dependency of the signal's energy during its propagation through the environment that is called channel, as described in Sec. 3.1.2, interactions with particles of the atmosphere take place, reflected by the term $\eta_{\text{cnl}}(r)$ from (3-3). As already stated in previous work of the author, this kind of attenuation includes weather effects like rain, snow, fog, and haze, as well as exhaust gases and pollution of the sensor itself.[137a]

When those interactions within the channel are modeled, the implementation approach for particles and especially particle clouds in the environment like dust or fog and spray becomes essential. When they are modeled as artificial cluster objects, the same kinds of equations as for solid objects can be applied. The already mentioned reflection-based lidar model from FZD contains a strategy for environmental effects by Linnhoff et al.[138] including a model for tire spray that produces lidar detections, while occluding others due to the already mentioned cause-effect chains.

The same modeling applies when each particle or drop is modeled individually within the particle cloud, as e.g. described by Hasirlioglu.[139] Alternatively, some sort of stochastic model for the path length within the particle cloud can be applied, needed to a probability distribution of simulated detections within the space covered by the particles.

All signal attenuation effects during its propagation through the channel have in common to produce a higher noise floor leading to a lower signal-to-noise ratio (SNR) which is easily reproduced in simulation in a stochastic manner. Additionally, occlusions of detections and consequently objects behind the particles in the atmosphere could occur leading to FN detections. Such missing detections need to be left out directly during ray casting/tracing of the virtual scene via absorbing particle clouds before them or they have to be eliminated after inserting the detections from the particle clouds before them in post-processing, as implemented by Linnhoff et al. in the reflection-based lidar model.[140] However, signal attenuation within the channel is out of scope of this dissertation, besides its general range dependency on its path.

Finally, more sparse point clouds and lower SNR with lower power or intensities of the measured detections lead to different and possibly wrong identification, tracking, and classification of objects or even FN objects, as e.g. described by Sebastian et al.[141]. This could be done stochastically in object-based models or it arises naturally during object identification when applied on sophisticated detection models.

[137] Rosenberger, P. et al.: Analysis of Real World Sensor Behavior (2018). a: p. 612.; b: p.615.; c: p.616.

[138] Linnhoff, C. et al.: Reflection Based Lidar Object Model · Environmental Effects (2022).

[139] Hasirlioglu, S.: Simulation-based Testing of Surround Sensors under Adverse Weather (2020), pp. 68ff.

[140] Linnhoff, C. et al.: Reflection Based Lidar Object Model · Environmental Effects (2022).

[141] Sebastian, G. et al.: RangeWeatherNet for LiDAR-Only Weather and Road Condition Classification (2021).

3.1.5 Temporal Lidar Behavior

There are several types of lidar sensors, each with different temporal data collection that should be considered for simulation and its VV&UQ. In the following subsection, the most prominent types are introduced, namely frequency modulated continuous wave (FMCW) lidar and time of flight (ToF) pulse lidar, divided into scanning and solid-state lidar.

FMCW chirp-sequence lidar sensors with included Doppler measurement are said to be a game changer for lidar, but are still under development and therefore expensive and not built by most lidar manufacturers. The continuously and coherently emitted light must also be detected continuously, with increased sensitivity, while even promising mm-precision (in contrast to several cm in current lidars). Coherent measurement means that the data is less affected by unwanted light influences from the sun or other lidar sensors, which is why FMCW results in a higher SNR ratio than pulse lidar. While modeling FMCW lidar is different to pulse lidar and refers more to radar modeling, besides modeling electromagnetic waves and not light pulses anymore, many cause-effect chains would stick the same. However, due to limited capacity in this dissertation, FMCW lidar simulation will not be further discussed in this work, just like radar modeling.

Scanning lidar on the other hand is the established SotA lidar approach and is used in series production cars by companies like AUDI[142]. In contrast to solid-state lidars, it involves mechanically moving parts, like a 360° turning mirror as illustrated by AUDI for the Valeo SCALA in Fig. 3-4 or a completely turning optical aperture[143] as within the Velodyne PUCK as shown by TechInsights Fig. 3-5, that spread the pulsed lidar beams into the environment and direct the received light into the receiver optic at (almost) the same time.

As the rotation of scanning lidars takes some time, effects well known from camera sensors occur in lidar, namely rolling shutter, motion blur, and aliasing. Rolling shutter is caused when different rows or columns of the lidar scan are recorded at different times leading to different positions of moving objects in a single scan, causing distorted point clouds.[137b] Motion blur means a wrong perceived size of an object (part) due to its movement during a single scan and capturing of it at slightly different positions.[137b] Aliasing happens when object parts with periodic movements like the wheels are recorded with a frequency that breaks the Nyquist-Shannon sampling theorem and therefore seem static or moving with a wrong frequency.[137c] While scanning lidars are very prone to the mentioned effects, they also occur in other lidar sensors depending on the receiver timing, albeit to a lesser extent.

Such sensor systems that are less exposed to temporal effects are called solid-state lidars, as they involve just very small or even no moving parts. As described by Li and Ibanez-Guzman[144] and depicted in Fig. 3-6, three types of solid-state lidars are differentiated from mechanical

[142] AUDI AG: Laserscanner (2017).

[143] TechInsights Inc.: Velodyne LiDAR Puck Teardown (2019).

[144] Li, Y.; Ibanez-Guzman, J.: Lidar for Autonomous Driving (2020), pp. 54-55.

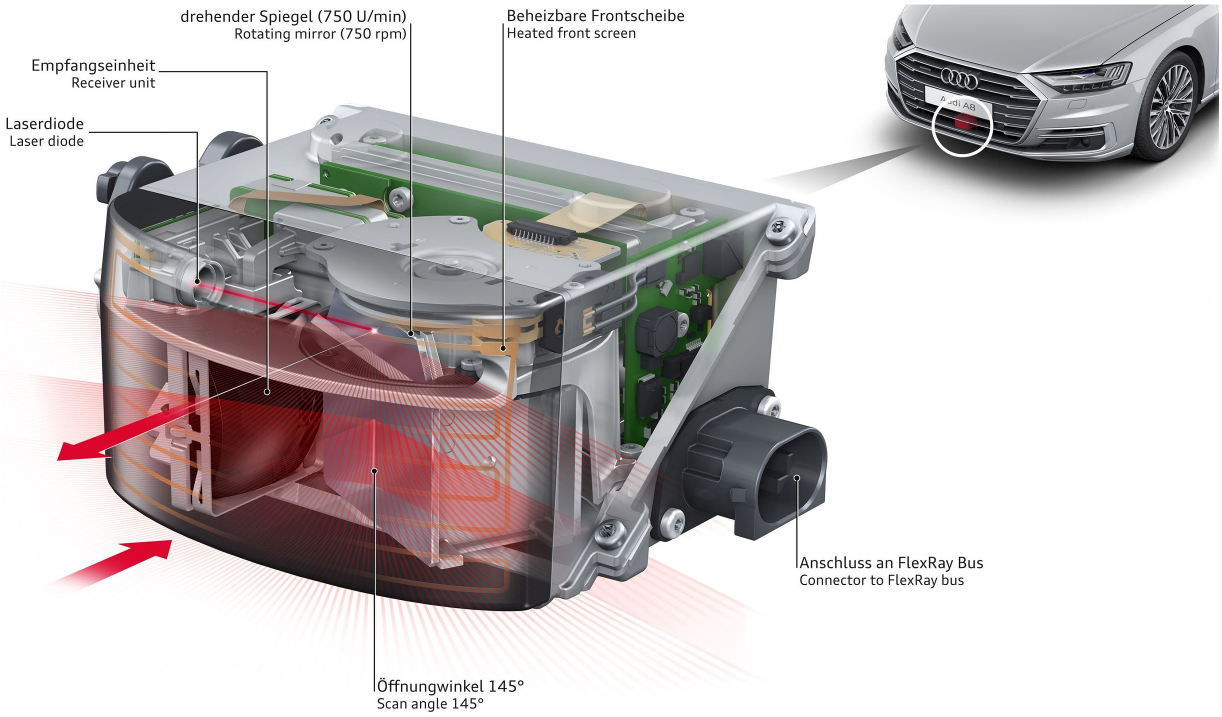


Figure 3-4: Valeo SCALA scanning lidar used in AUDI A8 from AUDI AG[142]

scanning (a): microelectromechanical systems (MEMS) (b), flash (c), and optical phased array (OPA) (d). The first solid-state lidars are called flash lidars and spread the light with an optical diffuser over a wide area in front of the sensor. Recent developments use sequential array-wise flashing with vertical-cavity surface-emitting laser (VCSEL) emitters like the Ibeo NEXT[145].

Others are using MEMS with multiple miniature mirrors on a chip, each steering one lidar beam, leading to highly flexible beam patterns while getting rid of the motors involved in scanning lidars. However, MEMS lidars are influenced by vibrations that regularly occur in vehicles caused by their engines or by the suspension induced by the rough ground it is driving on. Alternatively, completely solid-state sensors are achieved using OPA emitters, as promoted for the Quanergy S3 lidar[146]. OPA sensors use multiple coherent light emitters that are delayed by different numbers of phases, which leads to a directed beam in the far-field as depicted in Fig. 3-7

All three mentioned temporal effects leave two options for their simulation: Either simulating each transmitted beam in the same temporal manner as in reality by not stopping the virtual scene for the complete scan, but for each time beams are transmitted, which could lead to way higher computation time. Or it is implemented after calculating the whole scan at once in simulation and shifting each detection to the spot where it would actually been recorded in real life afterwards, which could lead to lower fidelity of the simulation. As both methods are theoretically possible, they allow to decide their implementation to be selected depending on the requirements, but left out of scope of this dissertation and the exemplary lidar model implementation.

[145] Ibeo Automotive Systems GmbH: Solid state LiDAR sensor (2022).

[146] Green Car Congress: Quanergy S3 Operation Principle (2016).

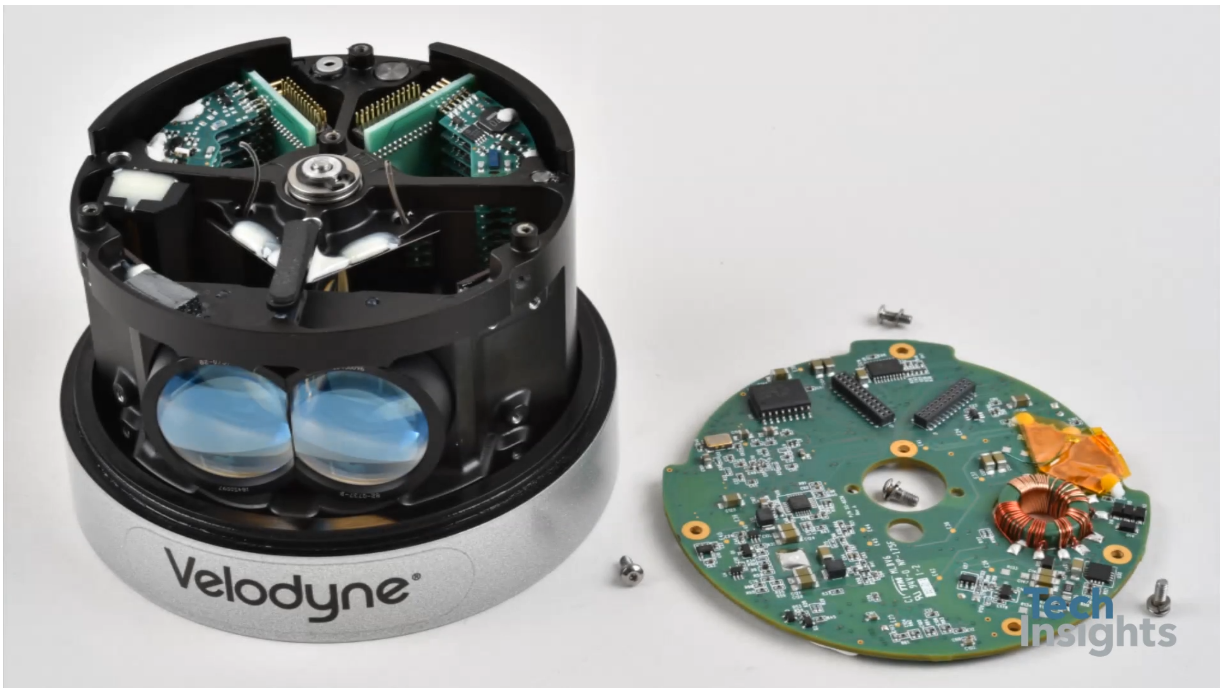


Figure 3-5: Velodyne Puck lidar teardown from TechInsights[143]

3.1.6 Receiver Effects

There are several receiver optics used in the industry within the still consolidating lidar market. All (except coherent FMCW lidars) are affected differently by glare from other light sources with the same wavelength like sun or other lidars, which should be modeled with different SNR ratios. However, there is a tendency lately to evolve to so-called digital lidars that use either SPAD arrays, d-SiPMs, or even SiPM arrays. Such digital lidar receivers using TCSPC directly provide digital signal consisting of histograms over range bins. Today's time-to-digital converter (TDC) resolution of $10\,\mathrm{ps}$[147] means $3.0\,\mathrm{mm}$ range resolution. Therefore, no cost intensive high performance analogue-to-digital converter is needed in such sensor systems anymore.

SPADs are single photon detectors that collect one photon at a time and then need to recover for some nanoseconds, thus limiting the maximum count rate to about $100\,\mathrm{MHz}$.[148] Therefore, the SPAD array receivers with many independent pixels, each one consisting of one SPAD[148] still suffers from the so-called "deadtime". A SiPM consists of multiple SPADs in parallel logical gating, but does not provide spatial information regarding which microcell got triggered.[148] (a-SiPM exist, as well.) The most advanced receivers formed by SiPM arrays are pixel-wise detectors where each is a SiPM (either a-SiPM or d-SiPM) and therefore provide spatial information, *"hence the imager spatial resolution is given by the number of SiPMs and not by the number of microcells"* [148] SiPM arrays today have a lower spatial resolution, but each pixel is photon number resolved almost without deadtime (being composed by many SPADs)[148]

[147] Sesta, V. et al.: A novel sub-10 ps resolution TDC for CMOS SPAD array (2018).

[148] Villa, F. et al.: SPADs and SiPMs Arrays for Long-Range High-Speed LiDAR (2021), pp. 9-10.

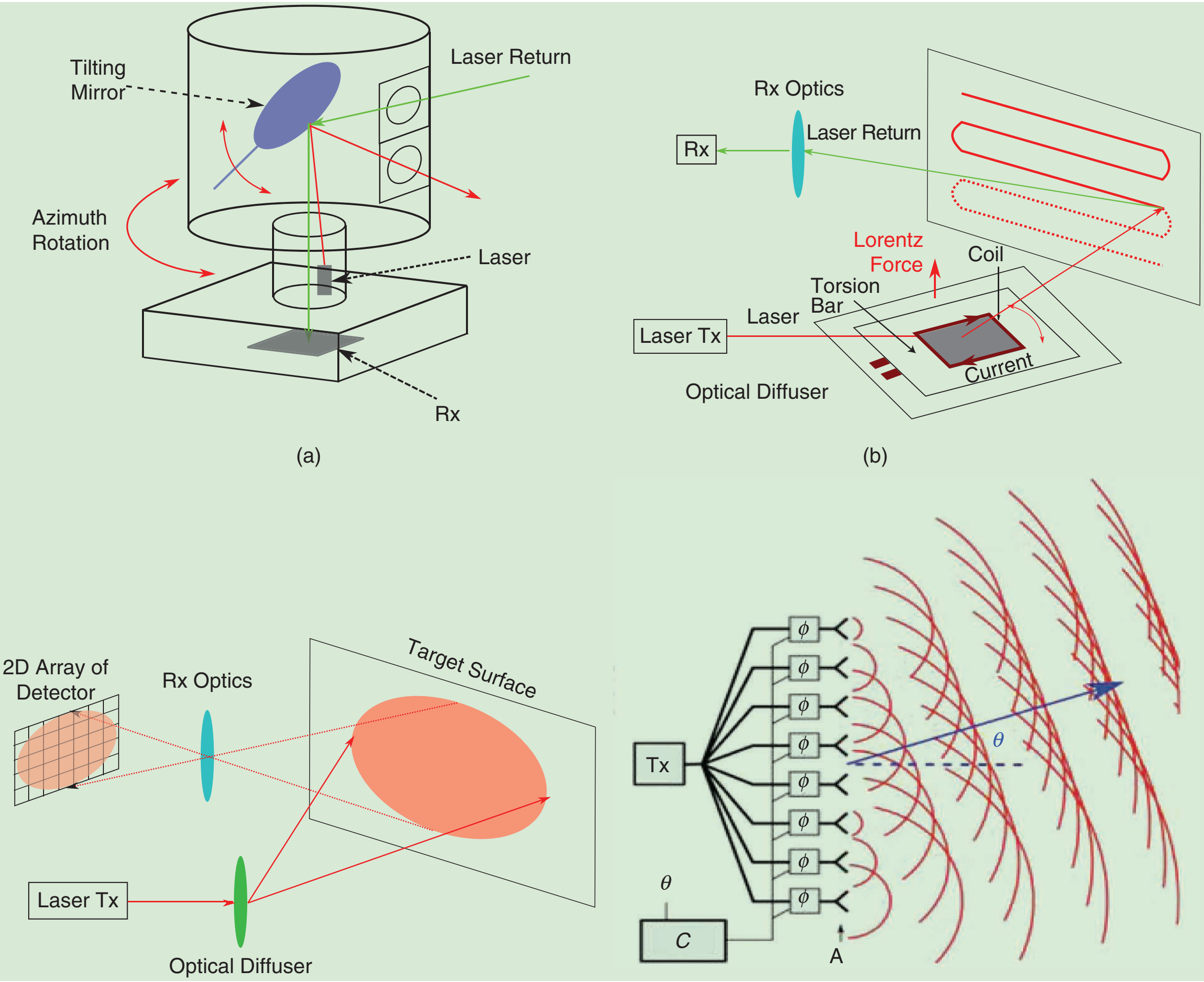


Figure 3-6: Lidar systems categorized by scanning approaches from Li and Ibanez-Guzman[144]. (a) Mechanical spinning lidar, (b) MEMS lidar,(c) flash lidar, and (d) OPA lidar. © IEEE 2020.

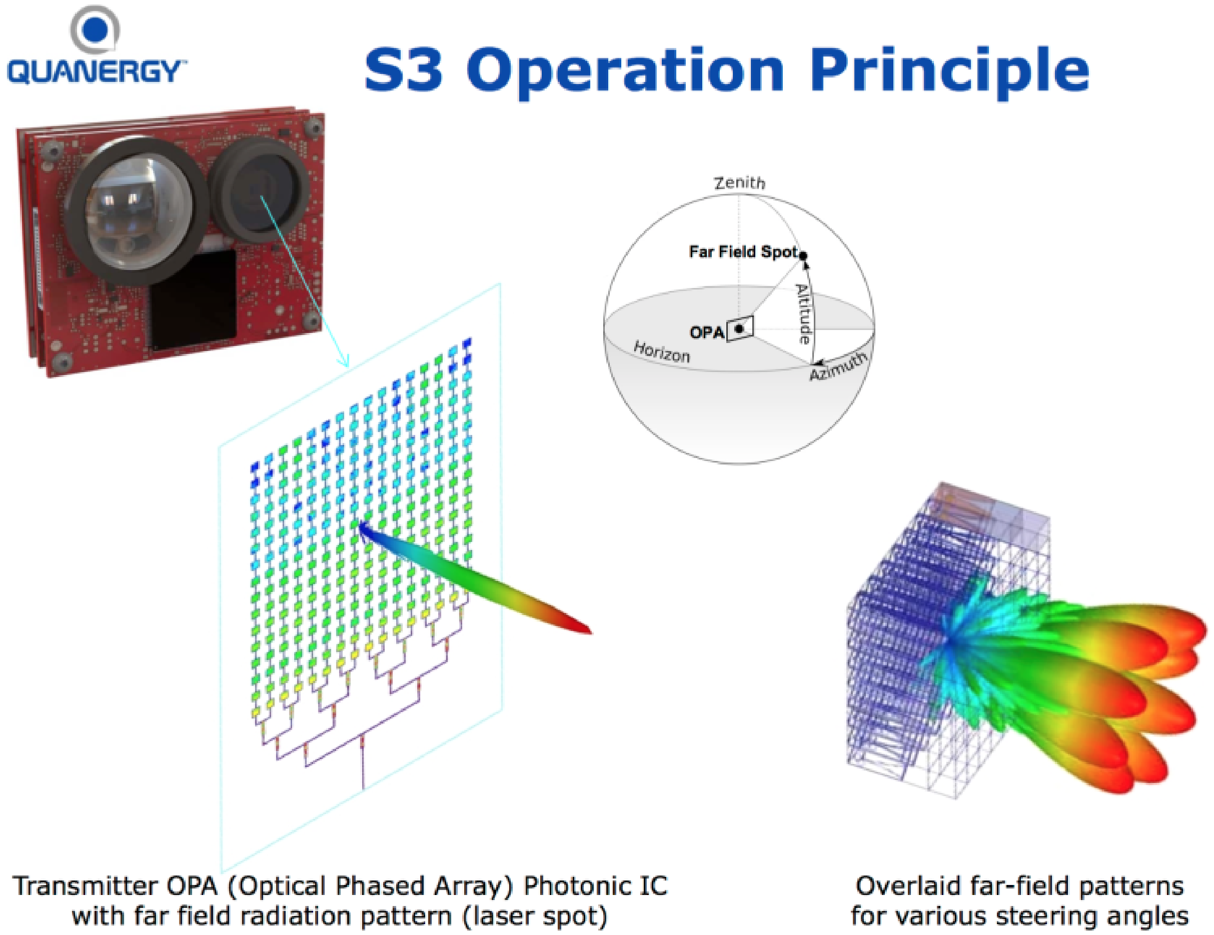
There is already a publicly available detailed SPAD/SiPM simulation by Tontini et al.[149] that could possibly be ported into modular simulation frameworks, like the one described in former work of the author and already mentioned in Sec. 2.1. However, as already mentioned in Sec. 3.1.1, modeling digital lidars is not done by beam super-sampling, even the computed reflections over range look very similar to the data output of such time-to-digital converters. In contrast, material or particle cloud transmissivity modeling is crucial for simulation of the first available data in digital lidars, the histograms over time/range per beam/pixel.

While PerCollECT - LidarLimbs[150] and the work from Hinsemann[151] give a condensed overview on lidar cause-effect chains, most are not yet included in today's lidar front-end simulation and are therefore out of scope in this work, too. This will have consequences for validation sample selection later in this work, while not limiting the applicability of the VV&UQ methods provided.

[149] Tontini, A. et al.: Numerical Model of SPAD-Based Direct Time-of-Flight Flash LIDAR (2020).

[150] Linnhoff, C. et al.: PerCollECT - LidarLimbs (2022).

[151] Hinsemann, T.: Analyse von Effekten in Lidardaten für die virtuelle Absicherung (2021).



S3 Operation Principle

Transmitter OPA (Optical Phased Array) Photonic IC with far field radiation pattern (laser spot)

Overlaid far-field patterns for various steering angles

Figure 3-7: Quanergy S3 operation principle.[146] © 2016 Quanergy Systems, Inc.

3.1.7 Detection Threshold Modeling

When it comes to actually simulating the thresholding step before detections exist at every time stamp, the correct simulation of the relative signal power or intensity becomes crucial. Considering analog signal processing as shown in Fig. 3-2 as well as digital lidars as already described, a list of spatially and temporally discretized data points exists in simulation at the point where thresholding takes place. Here, the modeling of the SNR is of significant importance.

It is common practice to use the time when the signal (digital or analogue) crosses the threshold (rising edge) as the point for computing the ToF, sometimes compensating the rise time of the signal using its steepness. While it would be possible to use the peak of the echo or its center between rising and falling edge for range calculation, for safety reasons selecting the range to the object as soon as visible over the noise seems reasonable. Nonetheless, actual lidar simulation models as the exemplary implemented lidar model provide the option to select it as a parameter.[152]

Thresholding is information reduction, but in some use cases, e.g. sensor fusion, confidence in detections becomes important. Therefore, some lidars provide an existence probability in

[152] Rosenberger, P. et al.: Reflection Based Lidar Object Model (2022).

addition to location and intensity. For this reason such values are part of ISO 23150 in case of lidar, providing an estimate of the “free space probability” of the area covered by the detection’s beam before it from sensor’s perspective.[153] More information can be provided from lidar data by values like the EPW (A-B and C-D in Fig. 3-2) as well as the mean or standard deviation that is used for background illumination or weather and road condition estimation.

3.2 Object Identification, Tracking, and Classification Modeling

Besides actually implementing object identification, tracking and even classification algorithms or neural networks for the same purpose to plug them onto the detection model, it is possible to mimic their behavior taking advantage of the available GT object list in simulation. Such an object model that does not require complex detection processing algorithms has been developed by Aust[154] under supervision of the author. It provides the option to select either idealized, fast but imprecise or sophisticated model behavior in case of the object’s position, orientation, velocity, acceleration, and dimensions.

Additionally, tracking behavior is modeled by proving parameters for the amount of frames in which the object must be visible before its actual listing at the output object list, and similar options are provided for the loss of objects from the output list. Special attention is payed for the actual reference point location of the object (either geometrical center, center of gravity of the points, or even the nearest corner of the L-shaped point cloud of the object) and its stability over time which directly influences track continuity.

It makes sense to filter out the ground reflections in a module that is upstream of the object model. A task not tackled by Aust is the modeling of the classification of objects. For this purpose, a feature-based classification is recommended, which assigns a class to the objects based on the most important properties such as length, width and speed. Another uncertainty that could be implemented in the model is the limited separation capability of the sensor due its limited spatial resolution and lack of velocity management especially in case of parked vehicles. Nevertheless, due to the complex task of VV&UQ, object modeling is left out of scope for the subsequent thesis.

[153] International Organization for Standardization: ISO 23150:2021(E) (2021), pp. 82-83.

[154] Aust, P.: Entwicklung eines lidartypischen Objektlisten-Sensormodells (2019).

3.3 Alternative Object-Based Modeling Approach

Former work of the author shows high potential of object-based lidar models to challenge reflection-based models using ray tracing.[155] E.g. in development phases where requirements on lidar sensor system simulation are not that high as for safety validation, the fast object-based approach is an option. Core of the proposed new object-based modeling method are refined bounding boxes for occlusion calculation instead of simple bounding box cuboids for each object.

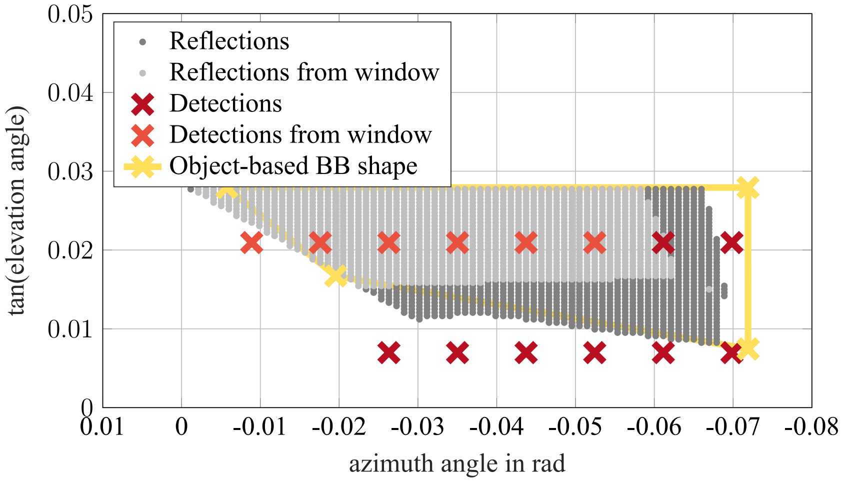


Figure 3-8: Rasterization of lidar detections at partly hidden objects over super-sampled reflections from former work of the author.[155a] © IEEE 2021.

As shown schematically in Fig. 3-8[155a], most lidars simply sort identified peaks from the signal onto the regularly rasterized beam pattern instead of measuring the incident angle. This induces high angular errors especially for hit objects at high ranges. Due to the beam divergence, the actual signal's peak often arises from object parts that are meters away from the beam's center line that is reflected by the rasterization.

When this behavior is simulated, it deprives precise ray casting and tracing techniques of most of their advantages. However, such rasterization of reflections onto the regular grid and the measurement bias it induces is already possible with the proposed object-based approach using the bounding boxes. However, it comes with much less effort than super-sampling with ray casting/tracing and therefore it is expected to have the potential to even outperform ray tracing for this simulation task.

[155] Linnhoff, C. et al.: Refining Object-Based Lidar Sensor Modeling (2021). a: p. 24244.

4 Existing Methodologies and Metrics for Verification, Validation, and Uncertainty Quantification (VV&UQ) of APSS Simulation

In the following chapter, existing VV&UQ approaches for simulation models of APSS are presented and the metrics involved are discussed. While there are no publicly available requirements for APSS simulation, the classical V-model, as e.g. described by Hakuli and Krug[156] cannot be applied directly straightforward in this work. However, inspecting the right side of the V reminds again that validation refers to the acceptance test against customer requirements and happens after several verifications and calibrations of the (simulation) solution have already taken place.

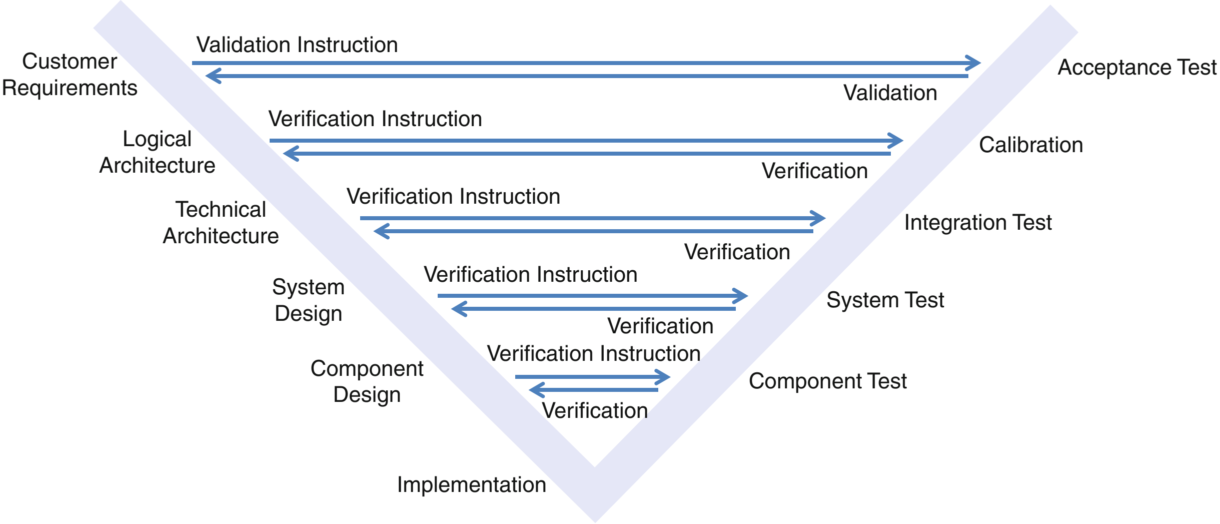


Figure 4-1: Development process according to the V-model from Hakuli and Krug.[156]
Reproduced with permission from Springer Nature.

At least, validation has already been addressed in earlier publications about automotive APSS simulation, like the ones from Roth et al.[157] or Bernsteiner et al.[158] Nevertheless, as stated by Viehof[159], the so-called validation is often limited to qualitative and subjective visual inspection of plots of the measured and simulated values over time as e.g. for object existence, positions and radial velocities.

[156] Hakuli, S.; Krug, M.: Virtual Integration in the Development Process of ADAS (2016), p. 5.

[157] Roth, E. et al.: Analysis and Validation of Perception Sensor Models (2011).

[158] Bernsteiner, S. et al.: Radar Sensor Model for the Virtual Development Process (2015).

[159] Viehof, M.: Objektive Qualitätsbewertung von Fahrdynamiksimulationen (2018), p. 24.

4.1 Methodologies for Simulation VV&UQ

Repeating the definition for validation by Oberkampf et al. from Sec. 1.2.6, it is the *"process of determining the degree to which a model is an accurate representation of the real world from the perspective of the intended uses of the model."* [160] Therefore, simulation validation involves real APSS data to compare against accompanied by reference data to reproduce the measurements in simulation, called replay-to-sim approach.

Model UQ, as defined in Sec. 1.2.5, is based on the V&V on multiple samples from the parameter space. It consists of an inter- and extrapolation (prediction) of the deviations determined for the chosen samples onto the whole parameter space including the confidence in this "model deviation model", as it could be named consequently. Model credibility is finally achieved only when all 9 steps of the linear process of VV&UQ in Fig. 1-7 have been performed and all categories of the predictive capability maturity model from Fig. 1-1 as described in Sec. 1.2.8 are checked including VV&UQ, but also physics, materials, geometries, numerical, and sensitivity analysis and a data analysis as demanded by NASA's key aspects for credibility from Fig. 1-2.

There are already methodologies applied in different simulation domains, which are designed to achieve credibility in simulation. To the knowledge of the author, there is currently no holistic and complete VV&UQ methodology specifically designed for APSS simulation. However, the methodology for model V&V from Viehof based on an exhaustive literature review has been discussed in former work of the author and used for APSS simulation thereby.[161] Schaermann proposes a methodology explicitly designed for V&V of APSS simulation and concentrates on efficient and precise data collection, while not considering UQ. Finally, again based on an exhaustive literature review, Riedmaier, Danquah et al. propose a holistic methodology for VV&UQ, which will be discussed and evaluated in the following.

4.1.1 Methodology for Model V&V by Viehof

The first methodology to be discussed in detail is the one published by Viehof in 2018[162]. He concludes after providing an exhaustive survey on existing model validation methodologies in 2017 that there is no universally valid strategy for model validation, while a tendency towards systematic and objective decisions is present in literature.[163] He derives from his literature research that model validation is often designed differently, the selection of validity criteria as well as the validation techniques are mostly subjective and validation mostly refers to single model, parameter, and data set combinations, which only generates confidence in the model in

[160] Oberkampf, W. L.; Trucano, T. G.: Verification and validation benchmarks (2008).

[161] Rosenberger, P. et al.: Towards a Generally Accepted Validation Methodology for Sensor Models (2019), p. 10.

[162] Viehof, M.: Objektive Qualitätsbewertung von Fahrdynamiksimulationen (2018). a: p. 47.; b: p. 102.; c: p. 106.

[163] Viehof, M.; Winner, H.: Forschungsstand der Validierung (2017), p. I.

combination with the respective parameter data set. Therefore, he identifies a need for objective assessment of sample-validity, as defined in Sec. 1.2.6, as the consequence of an unsuccessful falsification within an empirical series of sample experiments.

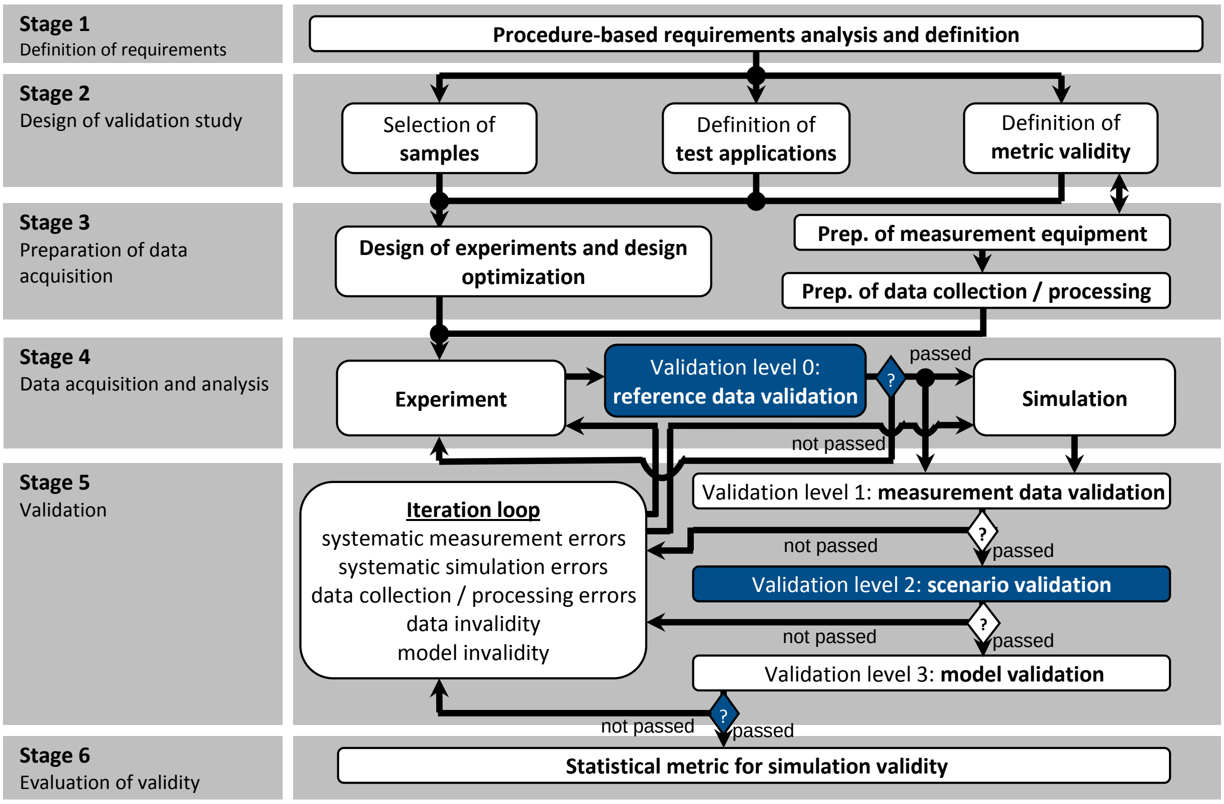
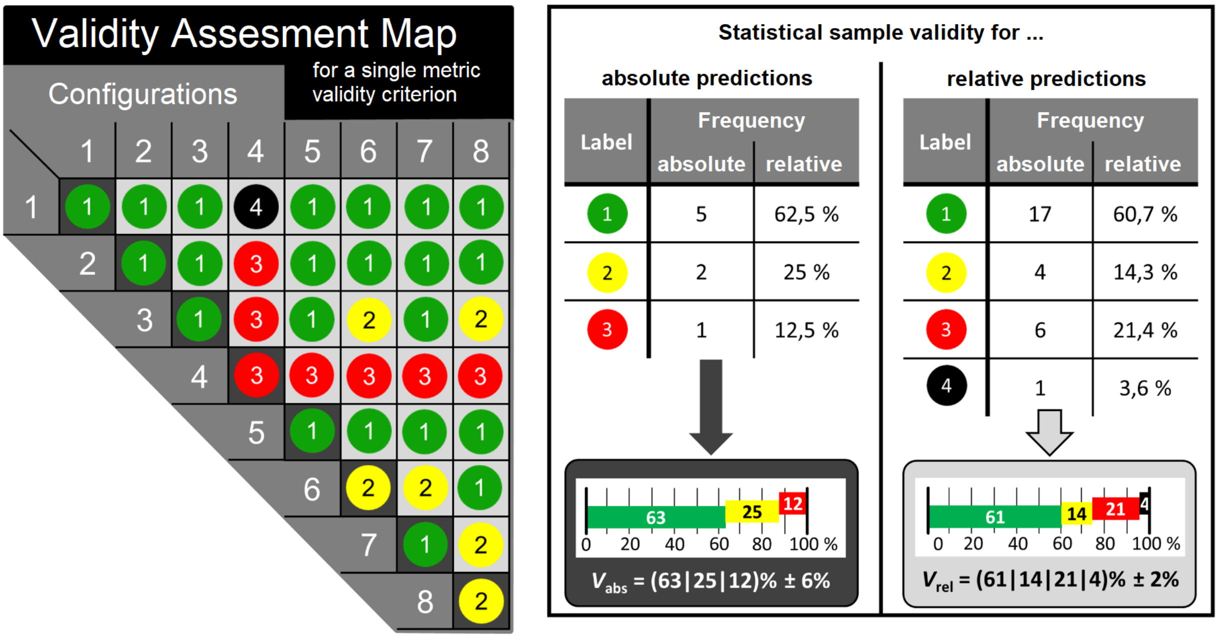

Stage 1
Definition of requirements
Procedure-based requirements analysis and definition
Stage 2
Design of validation study
Selection of samples
Definition of test applications
Definition of metric validity
Stage 3
Preparation of data acquisition
Design of experiments and design optimization
Prep. of measurement equipment
Prep. of data collection / processing
Stage 4
Data acquisition and analysis
Experiment
Validation level 0: reference data validation
?
passed
not passed
Simulation
Stage 5
Validation
Iteration loop
systematic measurement errors
systematic simulation errors
data collection / processing errors
data invalidity
model invalidity
Validation level 1: measurement data validation
?
not passed
passed
Validation level 2: scenario validation
?
not passed
passed
Validation level 3: model validation
?
not passed
passed
Stage 6
Evaluation of validity
Statistical metric for simulation validity

Figure 4-2: Objective quality assessment by statistical validation.
Blue: Changes to the methodology of Viehof[162a] in former work of the author[161].

In former work of the author[161] the methodology of Viehof[162a] is discussed in detail and questioned for its application on APSS models. As illustrated in Fig. 4-2, the methodology of Viehof consists of six subsequent stages, starting with definition of requirements, over design of the validation study, preparation of data acquisition, the data acquisition itself including its analysis, and validation as the central part, ending in a statistical evaluation. It further includes methods like a sensitivity analysis for sample selection for the validation study. Special focus is put on the different possible causes for iteration loops, e.g. due to systematic measurement or simulation errors.

Overall, the application study in former work of the author approves the methodology of Viehof. Only minor changes are demanded, as depicted in blue in Fig. 4-2. After experiments are performed in stage 4, it is necessary to inspect the recorded reference data leading to a first possible iteration loop. As the reference data is collected for re-simulation, this can only take place if accuracies of reference data in time and space are according to the requirements. In stage 5, three different iteration causes are present due to three levels of validation. It starts with measurement and simulation data validation, checking for systematic errors, e.g. significant differences between reference and simulated object trajectories, and data collection and processing failures.

Second comes scenario (originally: parameter) validation. It involves checking the selection of samples and if the sensitivity analysis estimations are fulfilled. Finally, the actual sample validation is performed, which leads to the validity assessment map, shown in Fig. 4-3, and its statistical evaluation, depicted in Fig. 4-6. As shown in Fig. 4-2, an additional iteration loop after sample validation is proposed to stress the option for new model calibration.



Validity Assessment Map

for a single metric validity criterion

Configurations

	1	2	3	4	5	6	7	8
1	1	1	1	4	1	1	1	1
2		1	1	3	1	1	1	1
3			1	3	1	2	1	2
4				3	3	3	3	3
5					1	1	1	1
6						2	2	1
7							1	2
8								2

Statistical sample validity for ...

absolute predictions

Label	Frequency	
	absolute	relative
1	5	62,5 %
2	2	25 %
3	1	12,5 %

relative predictions

Label	Frequency	
	absolute	relative
1	17	60,7 %
2	4	14,3 %
3	6	21,4 %
4	1	3,6 %

V_{abs} = (63|25|12)% ± 6%

V_{rel} = (61|14|21|4)% ± 2%

Figure 4-3: Validity assessment map of a single metric validity criterion from Viehof[162b], translated by the author

Viehof leaves it open, which metric validity criterion (MVC) are chosen for evaluation. It could be a deterministic simulated/real measurand, or characteristic values like mean or standard deviation extracted from a Gaussian aleatory process. The MVC should be defined in the preceding requirements definition in stage 1. They should cover the output variables of the simulation, but also the characteristic quantities of subsystem interfaces, when considered in the model specification. When multiple experiments are performed for each sample from the parameter space, the student's t-distribution is estimated from the MVCs for measurement and (re)-simulation of the experiments, as shown in Fig. 4-4. Then, from the estimated probability density function (PDF), a tolerance interval is extracted, symmetrical around the mean of the distribution and with a tolerance level, predefined in the model's specification.

The assumption of a normal distribution in the MVCs is questionable and should be justified beforehand. Another assumption for a t-distribution is that the scale applied to the collected data follows a continuous or ordinal scale. Furthermore, for calculation of an arithmetic mean from the MVCs, they must be interval scaled, which should be checked beforehand, too. Depending on the difference of the mean values, the overlap of the specified tolerance intervals from the estimated t-distributions, and the confidence levels demanded for the labeling, sample validity labels are given for each MVC, as illustrated in Fig. 4-5. While Viehof applies the commonly used, but subjective confidence level of 95 %, other confidence levels are possible as well.

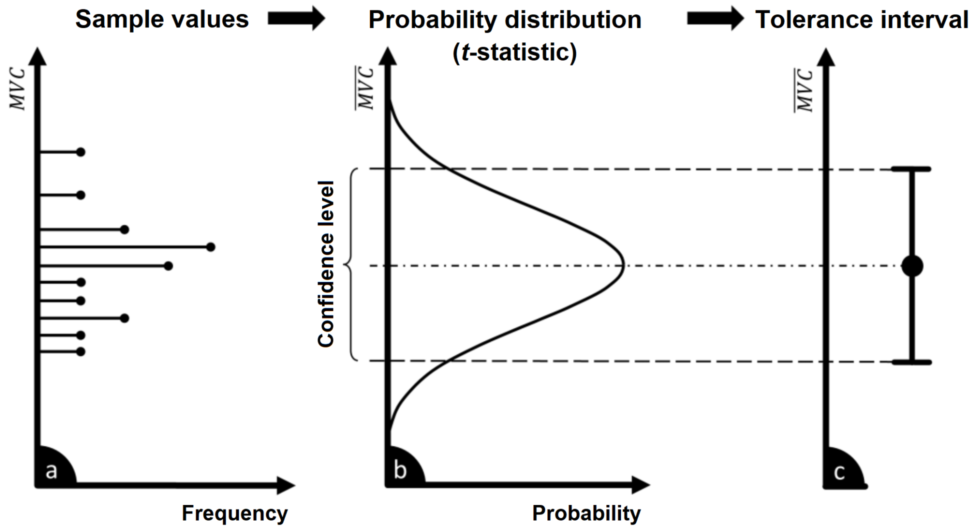


Figure 4-4: Elicitation of a tolerance interval from sample values by Viehof[162b], translated by the author

Fig. 4-5 shows on the top how the PDFs from simulation and measurement overlap for the five corner-cases (a - e) for labeling. A green label indicates that the estimated MVC distributions do not differ significantly, whereas a red label indicates that there is a significant difference. A yellow label states that the significance for a green (sample valid) or red (not sample valid) label is not given. For relative comparisons, a black label is introduced, when a change from one sample to another causes a different tendency of the MVC that is present in a different sign of the present minimum and maximum change in MVC. However, he implicitly assumes that the simulation's PDF is smaller as the measured counterpart. This is fine for models that predict single uncertain numbers like vehicle consumption or cornering stiffness. Simulation of a complete sensor signal and the sensor's behavior over time including its scattering, is only valid if the simulation reproduces a similar or same PDF, to the understanding of the author.

Only some metrics like difference in mean and standard deviation between simulation and measurement can form a MVC. Nevertheless, if the samples are sorted accordingly and the reflect the parameter space meaningfully, the validity assessment map already suggests in which parts of the parameter space the model performs fine (when every label is green), and where it might be necessary to start a new iteration on model calibration. Overall, no uncertainties on the labels are provided and no further credibility assessment is given, which would be beneficial for informed decision making.

The mentioned validity assessment map is basically a cross table for each MVC over all samples from the possible parameter space. It contains all sample validation results in form of labels for absolute comparison of simulation and measurement per sample on the main diagonal, and relative comparison for stepping from one sample to the next. For example, if the samples are lidar sensor measurements or simulations for different ranges, on the main diagonal are the results for each range, while the labels above it arise from comparison of the data deviations when stepping from one range to the next.

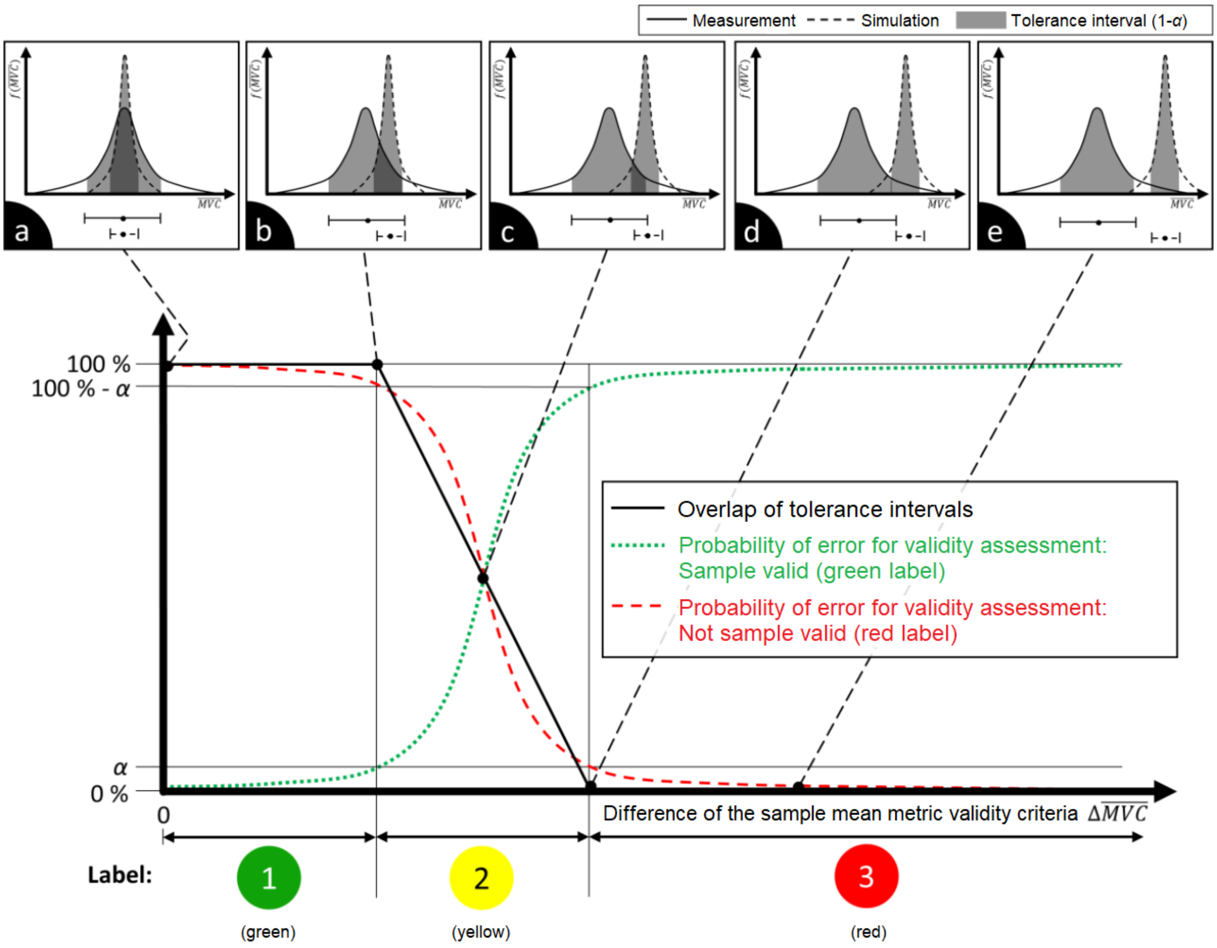

Measurement
Simulation
Tolerance interval (1-α)
a
b
c
d
e
f(MVC)
MVC
100 %
100 % - α
α
0 %
0
Overlap of tolerance intervals
Probability of error for validity assessment: Sample valid (green label)
Probability of error for validity assessment: Not sample valid (red label)
Difference of the sample mean metric validity criteria ΔMVC
Label:
1
2
3
(green)
(yellow)
(red)

Figure 4-5: Justification of the label assignment by Viehof based on the probability of error[162c], translated by the author

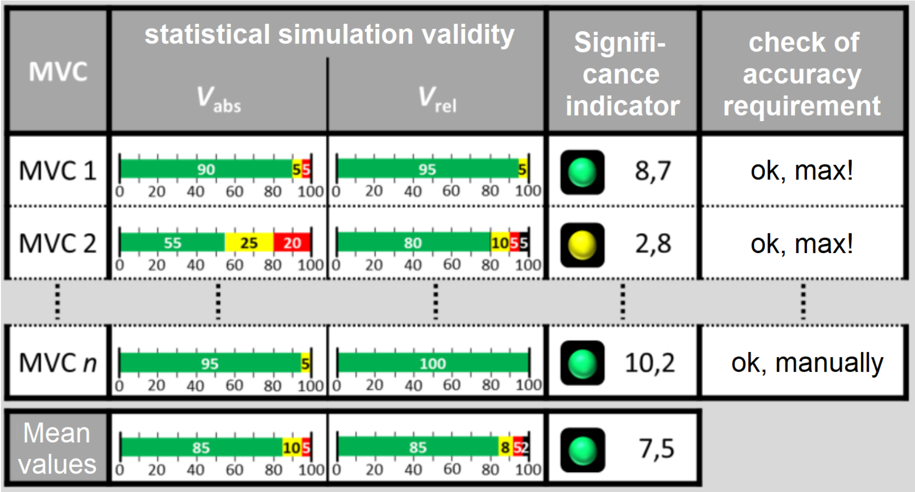
The labels from Viehof form a classical traffic light scheme (1: green, 2: yellow, 3: red) plus 4: black as worst label. While such a scheme provides a good visual overview about the absolute and relative sample validity, it is up to discussion, if the statistical evaluation at the end benefits from such intermediate information reduction. Viehof concludes that the labels seem to fulfill all needs of a validation study.[164b] Furthermore, he states that a continuous validity assessment instead of his Boolean labels, which would be based on similarity assessment of the real and simulated distributions, is only applicable if a high number of measurements is available / performable, while the additional effort is not justified after his research.[164b] Both conclusions are up to discussion and it is not justified, why such comparison should mean higher overall effort.

The impression from inspecting the validity assessment map can generate confidence in the model, but it is at least debatable where it comes from. At the very least, however, one would have to consider how to interpolate between the samples, which is probably straightforward for small parameter spaces and a few samples, but definitely not trivial in an N-dimensional parameter space. Therefore, the validity assessment map is only straightforward for simple examples. In

[164] Viehof, M.: Objektive Qualitätsbewertung von Fahrdynamiksimulationen (2018). a: p. 153.; b: p. 161.

addition, the number, selection, and concentration of samples must be taken into account if one wants to draw conclusions about the unprovable overall credibility for the given parameter space from the validity assessment map. However, Viehof states in his conclusion that the samples span a parameter space that allows interpolation when the system is step-free.[164a]

The objective statistical assessment proposed by Viehof as shown in Fig. 4-6 takes place after several MVCs are applied filling several validity assessment maps. It consists of a percentage preparation of the assigned labels for absolute and relative comparison and a significance indicator. The significance indicator is the ratio of the variation range of the MVC's expected value on the measurement data over the variance in the measurement data. Such statistical evaluation is only helpful for the application of the model under the beforehand formulated assumption that the samples included in the validity study span a fully-interpolatable parameter space and that the application area is completely covered. Both assumptions are highly critical and must be checked before every serious application.



MVC	statistical simulation validity		Signifi-cance indicator		check of accuracy requirement
	V_{abs}	V_{rel}			
MVC 1	🟩 90 🟨 5 🟥 5	🟩 95 🟨 5	🟢	8,7	ok, max!
MVC 2	🟩 55 🟨 25 🟥 20	🟩 80 🟨 10 🟥 5 ⬛ 5	🟡	2,8	ok, max!
⋮	⋮	⋮	⋮		⋮
MVC *n*	🟩 95 🟨 5	🟩 100	🟢	10,2	ok, manually
Mean values	🟩 85 🟨 10 🟥 5	🟩 85 🟨 8 🟥 5 ⬛ 2	🟢	7,5	

0
20
40
60
80
100

Figure 4-6: Statistical Assessment of VAMs from multiple metrics by Viehof[162c], translated by the author

4.1.2 Methodology for Model V&V by Schaermann

Schaermann derives a methodology for APSS Model V&V[165] from previous approaches from Sargent[166] and Oberkampf and Trucano[167]. He discusses his approach following Roth et al.[168] that mainly treats data collection and its re-simulation. Although aleatory and epistemic uncertainties are mentioned, they are not addressed separately in Schaermann's methodology.[165a] He concludes that all methodologies have different purposes and metrics must be chosen according to the specific use case.[165b] Therefore, Schaermann develops an own solution, which is illustrated in Fig. 4-7.[169] Similar to Viehof's stages 2-5, it includes a reference data check and it adds comparison on "raw data" and object level, which reflects its specific design for V&V of APSS simulation.

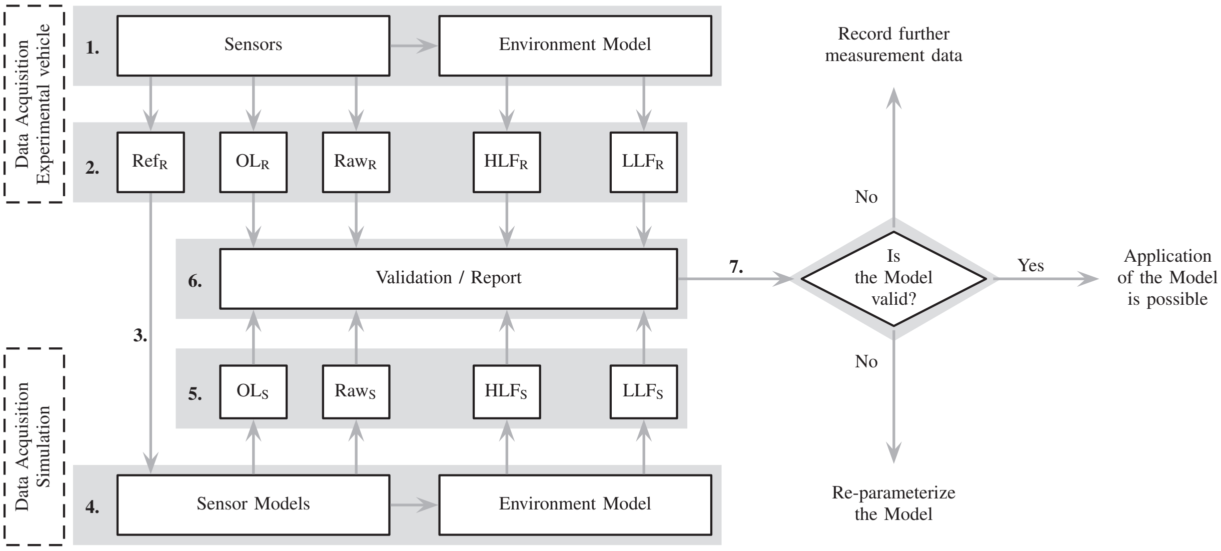


Figure 4-7: Validation methodology from Schaermann[169], i_{R}: Real-recorded data, i_{S}: Simulated data, Ref: Reference data, OL: Object list, Raw: Raw data. HLF: High level fusion, LLF: Low level fusion. © 2017 IEEE.

Schaermann lists requirements for V&V methodologies:[165c] They should be continuous for the whole model lifecycle, scalable for different model expansion stages, specifiable for its usage but holistic. He demands that V&V should be traceable, intuitive, objective and documented. Finally, he proposes to minimize measurement uncertainties and to avoid errors of type I, II, and III.[165c]

He provides a parameter reduction method for lower V&V complexity[165d] and proposes a reference data collection method[165e] to reduce aleatory and epistemic uncertainties. He applies several validation metrics on lidar detections, occupancy grids (OGs) filled with lidar detections and object lists.[165f] However, there is no error prediction towards the application parameter space or any UQ that would allow to call it a VV&UQ methodology for model credibility.

[165] Schaermann, A.: Systematische Bedatung und Bewertung umfelderf. Sensormodelle (2020). a: p. 19.; b: p. 25.; c: pp. 26-27.; d: pp. 40-50.; e: pp. 50-55.; f: pp. 59-100.

[166] Sargent, R. G.: Verification and validation of simulation models (2010).

[167] Oberkampf, W. L.; Trucano, T. G.: Verification and validation benchmarks (2008).

[168] Roth, E. et al.: Analysis and Validation of Perception Sensor Models (2011).

[169] Schaermann, A. et al.: Validation of vehicle environment sensor models (2017), p. 408.

4.1.3 Methodology for Model V&V by Ngo et al.

Ngo et al. follow the tradition of Schaermann, Viehof, Oberkampf, etc. and consequently present a two-step model validation process for a radar detection simulation that combines explicit and implicit sensor model evaluation, as depicted in Fig. 4-8.[170a] Again, a re-simulation approach of real world scenarios is proposed, where reference data like object trajectories are captured and used as input for re-simulation afterwards. With real and synthetic radar detections ("point clouds") at hand, the explicit model evaluation is performed with high level comparison of detection ranges to each other and low level evaluation of range, azimuth, and Doppler velocity separately. Subsequently, the same clustering and tracking algorithms are applied to both detection lists to obtain object lists and trajectories that are fed into the implicit sensor model evaluation. After both evaluations, the overall so-called "Simulation-to-Reality Gap G" is computed.

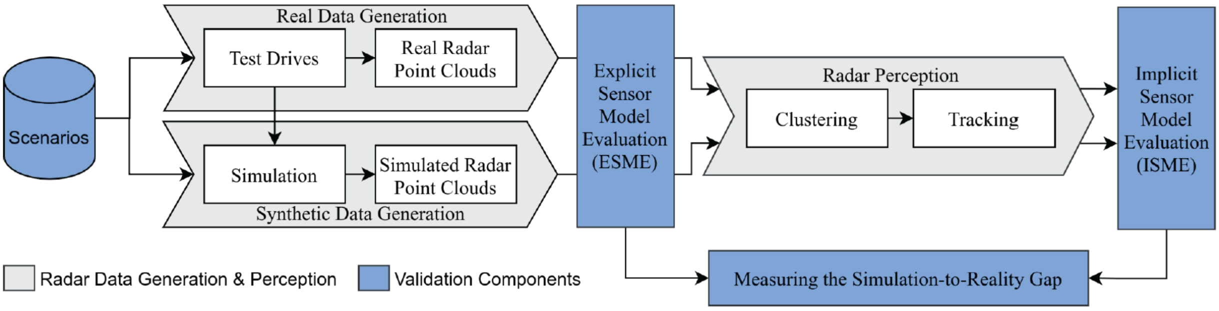
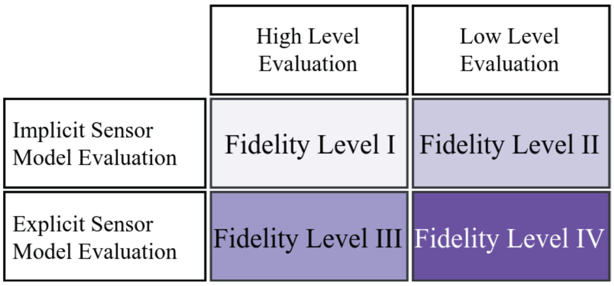

Real Data Generation
Test Drives
Real Radar Point Clouds
Scenarios
Simulation
Simulated Radar Point Clouds
Synthetic Data Generation
Explicit Sensor Model Evaluation (ESME)
Radar Perception
Clustering
Tracking
Implicit Sensor Model Evaluation (ISME)
Measuring the Simulation-to-Reality Gap
Radar Data Generation & Perception
Validation Components

Figure 4-8: Validation methodology from Ngo et al.[170a]. © 2021 IEEE.

In this regard, Ngo et al. divide the evaluation into four fidelity levels depending on the evaluated functional layer and the metrics involved, as shown in Fig. 4-9.[170b] Multiple metrics are applied in each level, normalized to the interval $[0, 1]$ and aggregated for each layer. At the end, the Simulation-to-Reality Gap G is computed by the average over all four fidelity levels.



	High Level Evaluation	Low Level Evaluation
Implicit Sensor Model Evaluation	Fidelity Level I	Fidelity Level II
Explicit Sensor Model Evaluation	Fidelity Level III	Fidelity Level IV

Figure 4-9: Fidelity levels for APSS simulation from Ngo et al.[170b]. © 2021 IEEE.

However, the metrics selection is crucial for the finally calculated gap G and averaging over the mutably chosen fidelity levels is up to discussion, as the authors state themselves. Still, having a final single score at the end that could possibly be predicted into application conditions of the simulation is a desirable objective, as also targeted for by Huch[171].

[170] Ngo, A. et al.: Multi-Layered Measuring the Simulation-to-Reality Gap for Radar (2021). a: p. 4009.; b: p. 4011.

[171] Huch, S.: Metrik zur Bewertung der Lidar-Sensor-Simulation (2018).

4.1.4 Methodology for Simulation VV&UQ by Riedmaier and Danquah

The third methodology further discussed in detail was published by Riedmaier, Danquah et al. in 2020[172]. Similar to Viehof, they provide an exhaustive survey and evaluation of model VV&UQ approaches at first.[173] They state that V&V must be accompanied by UQ and therefore only consider complete VV&UQ methodologies in their evaluation[172a] of six approaches in total:

1. The probability bound analysis (PBA) by Oberkampf and Roy[174a] that uses Frequentist statistics while propagating aleatory/epistemic uncertainties differently through the model.
2. The Bayesian approach by Sankararaman and Mahadevan[175] that includes subjective a-priori assumptions and therefore leads to lower uncertainties when compared to PBA.[174b]
3. The interval predictor model by Crespo et al.[176,177] that directly predict interval valued quantities to bound all future experiments within them.
4. The meta-model by Hills[178] that corrects model predictions with a data-driven model.
5. The output uncertainty integration by Eek et al.[179] that makes simplifications to the PBA.
6. The tolerance approach for deviations between simulation and experiment from ISO 19365.[180]

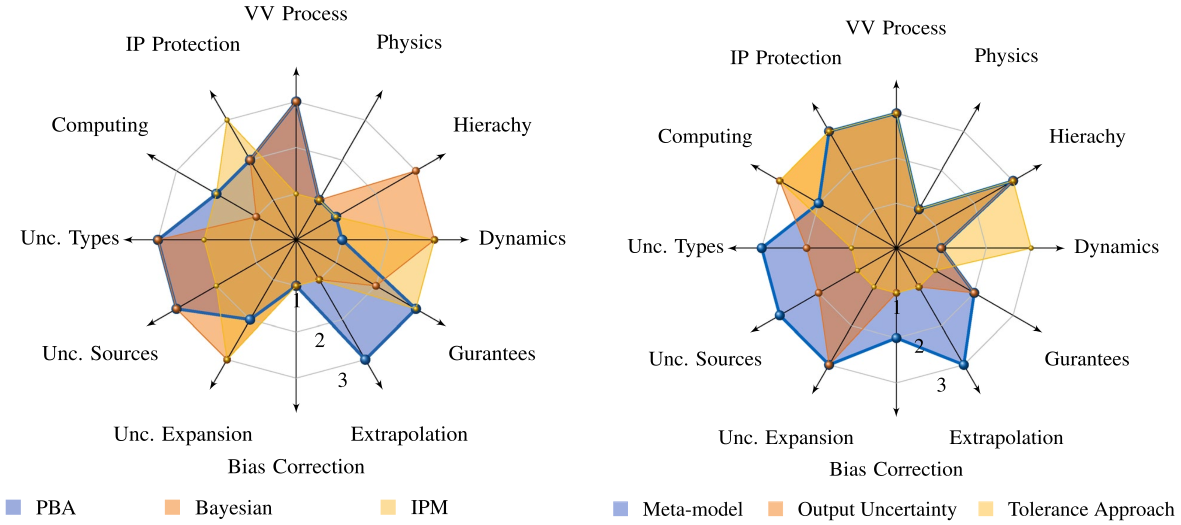


Figure 4-10: Comparison of VV&UQ approaches from Riedmaier, Danquah et al.[172b]

[172] Riedmaier, S. et al.: Unified Framework and Survey for Model VV&UQ (2020). a: p. 1.; b: p. 27.; c: p. 3.

[173] Danquah, B. et al.: Potential of statistical model VV&UQ in vehicle dynamics simulations (2020).

[174] Oberkampf, W. L.; Roy, C. J.: Verification and Validation in Scientific Computing (2010). a: p. 99.; b: p. 664.

[175] Sankararaman, S.; Mahadevan, S.: Integration of model V&V, and calibration for UQ (2015).

[176] Crespo, L. G. et al.: Interval predictor models with a formal characterization of uncertainty and reliability (2014).

[177] Lacerda, M. J.; Crespo, L. G.: Interval predictor models for data with measurement uncertainty (2017).

[178] Hills, R. G.: Roll-up of validation results to a target application. (2013).

[179] Eek, M. et al.: Definition and Implementation of a Method for Uncertainty Aggregation (2017).

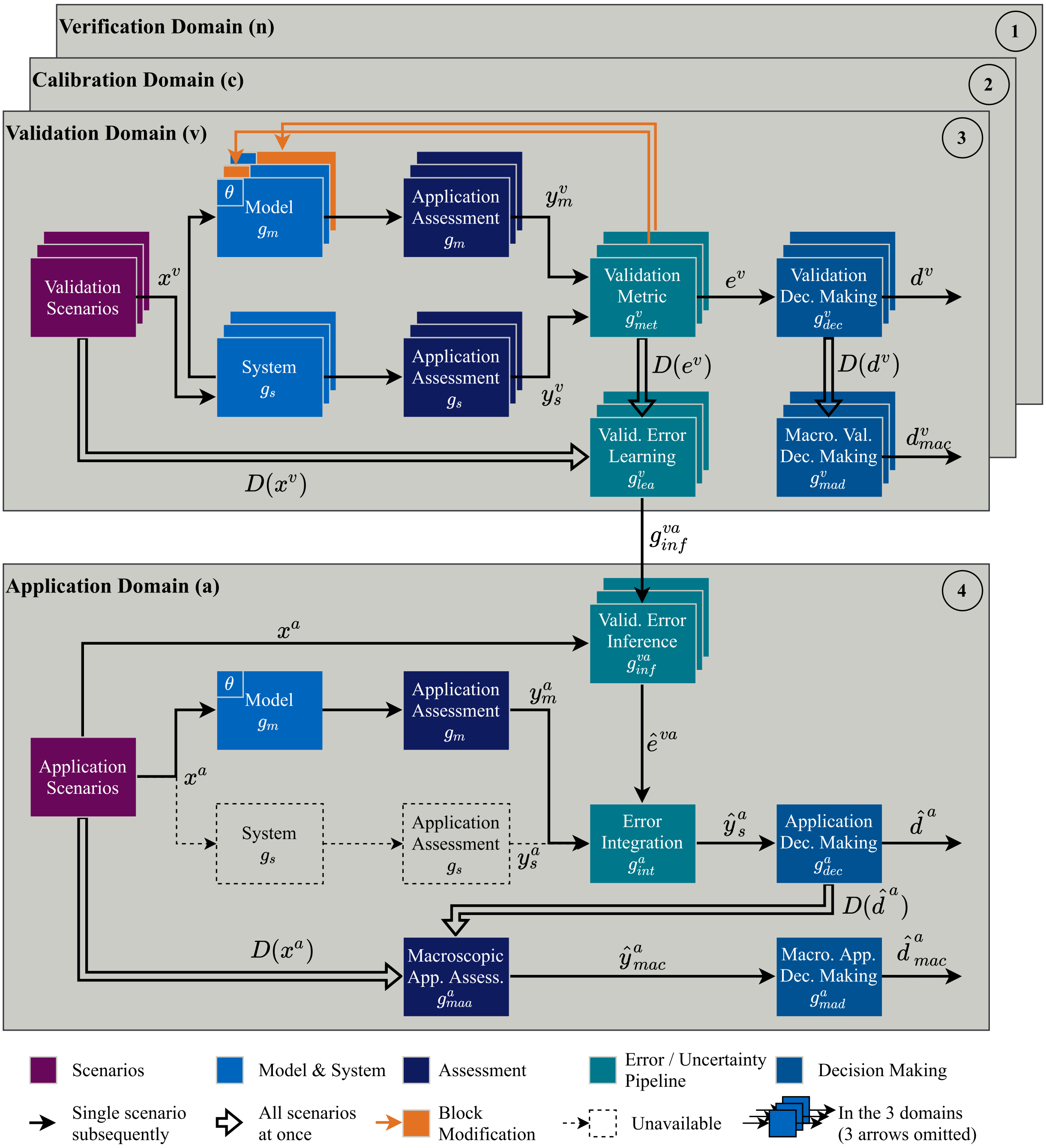
[180] International Organization for Standardization: ISO 19365:2016(E) (2016).

For the Kiviat diagrams in Fig. 4-10, Riedmaier, Danquah et al. evaluate the six selected VV&UQ approaches on twelve categories[172b] that are listed in Tab. 4-1.

Table 4-1: Criteria for the comparison of VV&UQ approaches from Riedmaier, Danquah et al.[172b]
IP: Intellectual property, PI: Prediction interval.

Criteria	**Description**	**Ratings**		
		1	**2**	**3**
V&V process	Model calibration accompanied by model verification & validation	Only cal.	Own verif.	Own valid.
Physics	Add extrapolative power instead of only extrapolating the error	None	Correction	Extrapol.
Hierarchy	Different types of architectures possible for hierarchical systems	None	One type	Many types
Dynamics	Different types of representations such as differential equations or discrete state-space equations	None	One type	Many types
Guarantees	Interval-based VV&UQ has abs., probabilistic has statistical, deterministic has no guarantees	Determ.	Probabil.	Interval
Extrapolation	Inter- and extrapolation uncert. at best combined with its inherent prediction uncert.	None	Without PI	With PI
Bias correction	Bias correction for the sim. model with a PI for thereby inherent prediction uncert.	None	Without PI	With PI
Uncert. expans.	Uncertainty expansion adding conservatism with uncertainty bounds as tight as possible	None	Wide bds.	Tight bds.
Uncert. sources	Separately quantify each source of errors and uncertainties	None	Jointly	Separately
Uncert. types	Explicit aggregation of epistemic (E) and aleatory (A) uncertainties	None	E or A	E and A
Computing	Computational complexity	Heavy	Medium	Light
IP protection	Handle gray-/black-box models for IP protection	White-b.	Gray-box	Black-box

As no approach covers everything, Riedmaier, Danquah et al. do not find a clear winner. The left Kiviat diagram in Fig. 4-10 shows that PBA handles all sources and types of uncertainties separately, but the strong uncertainty expansion provides very conservative extrapolation. The Bayesian approach *"is based on subjective probabilities with priors, cannot naturally represent epistemic and aleatory uncertainties, currently lacks extrapolation uncertainty and includes heavy inverse calculations."* [172b] Interval predictor models have no separate calibration and validation, but tight uncertainty bounds. The right Kiviat diagram in Fig. 4-10 gives the impression that the meta-model would win with its connection between the validation and extrapolation and by its bias correction with prediction interval, but the required linear dependency in the meta-model and the small application parameter space prevent this. The output uncertainty approach is widely applicable due to not quantifying all sources of uncertainty. The tolerance approach is even simpler, fast and flexible, but does not consider uncertainties and neglects aggregation.[172b]



Verification Domain (n)
1
Calibration Domain (c)
2
Validation Domain (v)
3
Validation Scenarios
x^v
θ
Model g_m
Application Assessment g_m
y_m^v
System g_s
Application Assessment g_s
y_s^v
Validation Metric g_{met}^v
e^v
Validation Dec. Making g_{dec}^v
d^v
$D(e^v)$
$D(d^v)$
Valid. Error Learning g_{lea}^v
Macro. Val. Dec. Making g_{mad}^v
d_{mac}^v
$D(x^v)$
g_{inf}^{va}
Application Domain (a)
4
x^a
Valid. Error Inference g_{inf}^{va}
θ
Model g_m
Application Assessment g_m
y_m^a
$\hat{e}^{va}$
Application Scenarios
x^a
System g_s
Application Assessment g_s
y_s^a
Error Integration g_{int}^a
$\hat{y}_s^a$
Application Dec. Making g_{dec}^a
$\hat{d}^a$
$D(\hat{d}^a)$
$D(x^a)$
Macroscopic App. Assess. g_{maa}^a
$\hat{y}_{mac}^a$
Macro. App. Dec. Making g_{mad}^a
$\hat{d}_{mac}^a$
Scenarios
Model & System
Assessment
Error / Uncertainty Pipeline
Decision Making
Single scenario subsequently
All scenarios at once
Block Modification
Unavailable
In the 3 domains (3 arrows omitted)

Figure 4-11: Generic model VV&UQ framework from Riedmaier, Danquah et al.[172c]

To be able to use modules from different approaches to combine them into a proper approach at the end, Riedmaier, Danquah et al. developed a generic, (almost) holistic framework for model VV&UQ, which is depicted in Fig. 4-11. The first three domains are verification of code and numerical solution, calibration of parameters, and validation of form by comparison with physical data. The basic assumption is that these three domains follow the same scheme, only differing in application details *“to address various types and sources of errors and uncertainties”* [172c]. The framework is however only almost holistic, as the model specification and requirements definition is not included, while being essential to start the model evaluation in the first place.

As an example, the methodology of Viehof as shown in Fig. 4-2 (stages 2-6, as requirements in stage 1 are excluded) is a concrete example for the validation domain. Validation scenarios are defined, MVCs are derived as application assessment for model and (real) system, validation metrics are applied and (sample) validation decision making is addressed as well as (statistical) macroscopic validation decision making. However, validation error learning in inference for the application domain is not addressed by Viehof, as already discussed earlier.

In the application domain of the model, model prediction is accompanied by an integration of all errors and uncertainties. This allows to assess the beforehand validated model for its application without collecting and processing new data in that application domain. The objective of considering errors and uncertainties from validation for possible applications after all helps, besides argumentation for the usage of simulation in the first place, to give concrete confidence values when reporting is needed for e.g. safety argumentation and homologation of ADS. Methods for model error learning and bias correction including the corresponding uncertainty aggregation will be discussed in the following in Sec. 4.1.5.

While the framework from Riedmaier, Danquah et al. is clear and generic, its practical value can only be shown by application in specific use cases. Therefore, Riedmaier, Danquah et al. show that it is capable of handling deterministic and non-deterministic VV&UQ approaches. Their results indicate that the non-deterministic PBA approach is more conservative, but outperforms the deterministic approach in precision with equal recall. However, the non-deterministic approach requires a higher effort, as the uncertainties need to be quantified and propagated through the simulations.[181a] The uncertainty treatment from identification, quantification, propagation, and aggregation within the framework from Danquah, Riedmaier et al.[182a] is illustrated in Fig. 4-12.

Nevertheless, Riedmaier, Danquah et al. decide to use PBA within their generic framework for safety validation of ADS, in their subsequent publications, as it is the *“main approach of Frequentist VV&UQ”* [181b]. They prefer it over Bayesian approaches, as it does not incorporate subjective beliefs as prior probabilities, nor distort the original model based on Bayes’ theorem. *“Therefore, it meets the requirements of an independent type approval.”* [181b]

Danquah, Riedmaier et al. apply the PBA-framework on a vehicle consumption simulation, quantifying all sources of uncertainty, just omitting the extrapolation uncertainty to the application domain at first.[182] Later, the application domain is addressed and the full PBA-framework is applied and validated while concluding advancements compared to the SotA of 2021.[183a]

[181] Riedmaier, S. et al.: Non-deterministic model validation methodology (2021). a: p. 16.; b: p. 6.

[182] Danquah, B. et al.: Statistical Model Verification and Validation Concept (2020). a: p. 265.

[183] Danquah, B. et al.: Statistical Validation Framework for Automotive Vehicle Simulations (2021). a: p. 19.; b: p. 6.

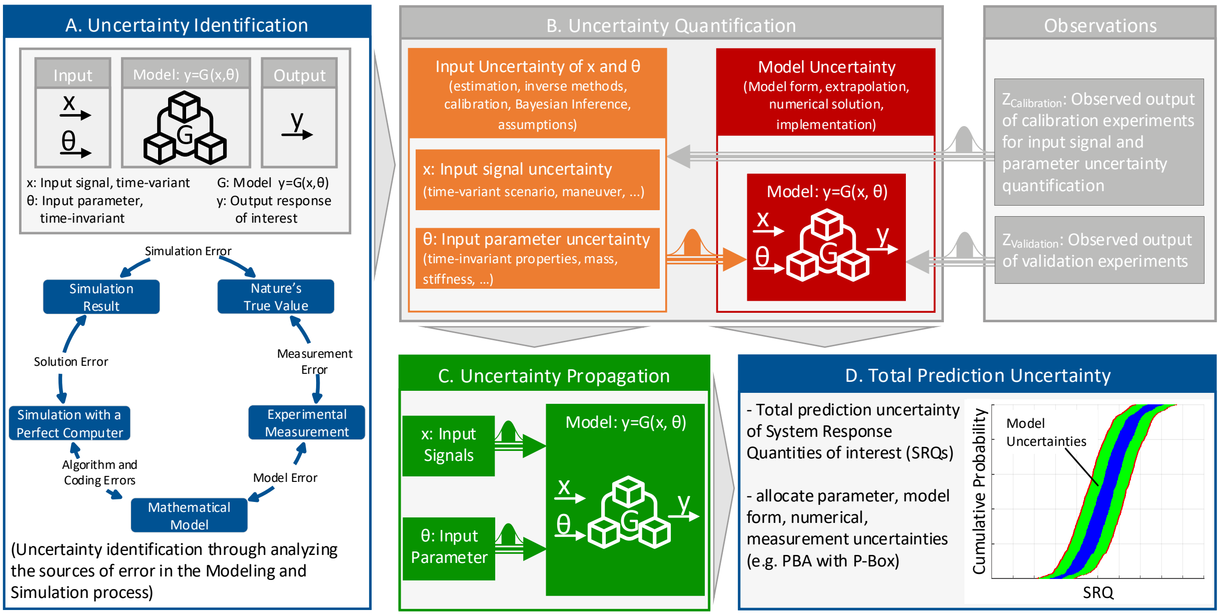

A. Uncertainty Identification
Input
Model: y=G(x,θ)
Output
x
θ
G
y
x: Input signal, time-variant
θ: Input parameter, time-invariant
G: Model y=G(x,θ)
y: Output response of interest
Simulation Error
Simulation Result
Nature's True Value
Solution Error
Measurement Error
Simulation with a Perfect Computer
Experimental Measurement
Algorithm and Coding Errors
Model Error
Mathematical Model
(Uncertainty identification through analyzing the sources of error in the Modeling and Simulation process)
B. Uncertainty Quantification
Input Uncertainty of x and θ (estimation, inverse methods, calibration, Bayesian Inference, assumptions)
x: Input signal uncertainty (time-variant scenario, maneuver, ...)
θ: Input parameter uncertainty (time-invariant properties, mass, stiffness, ...)
Model Uncertainty (Model form, extrapolation, numerical solution, implementation)
Model: y=G(x, θ)
Observations
Z_Calibration: Observed output of calibration experiments for input signal and parameter uncertainty quantification
Z_Validation: Observed output of validation experiments
C. Uncertainty Propagation
x: Input Signals
θ: Input Parameter
Model: y=G(x, θ)
D. Total Prediction Uncertainty
- Total prediction uncertainty of System Response Quantities of interest (SRQs)
- allocate parameter, model form, numerical, measurement uncertainties (e.g. PBA with P-Box)
Cumulative Probability
Model Uncertainties
SRQ

Figure 4-12: Uncertainty treatment within the framework from Danquah, Riedmaier et al.[182a].

They summarize that with the applied methodology binary, low-information validation results are superseded by high-information aggregated prediction uncertainty in the form of a p-box, uncertainties and non-deterministic simulations are considered, low extrapolation capability of model reliability is solved by uncertainty learning and the former small application domain is enlarged by uncertainty prediction.[183] Danquah finally summarizes the framework[184a], explains the findings from its application on vehicle dynamics[184b] and applies PBA in his dissertation on reliability determination of vehicle simulations[184c]. Special focus must be provided to the error prediction and uncertainty aggregation as inter- and extrapolation into application parameter space, shown in Fig. 4-13 and later discussed in this dissertation in Sec. 4.1.5, as this is the core of the overall model credibility.

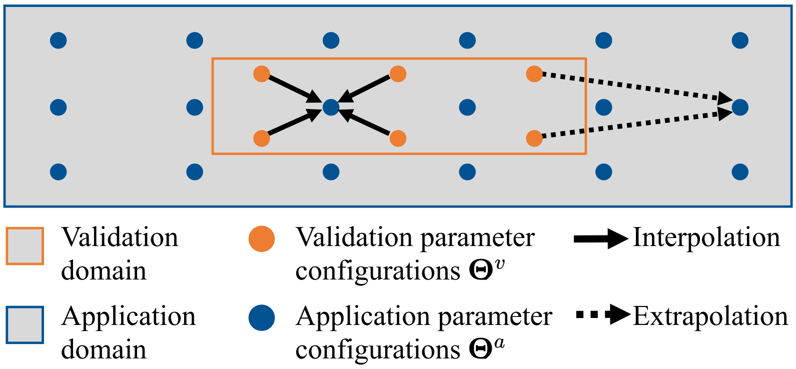

Validation domain
Validation parameter configurations $\mathbf{\Theta}^v$
Interpolation
Application domain
Application parameter configurations $\mathbf{\Theta}^a$
Extrapolation

Figure 4-13: Inter- and extrapolation into application parameter space from Danquah, Riedmaier et al.[183b].

[184] Danquah, B.: Zuverlässigkeitsbestimmung von Fahrzeugsimulationen durch statistische Validierung (2022). a: pp. 45-60.; b: pp. 61-78.; c: pp. 79-85.

4.1.5 Uncertainty Aggregation and Model Error Prediction

Re-simulation with reference data as input is the preferred method for validation of APSS simulation, starting with von Neumann-Cosel for an ideal sensor model and object (range) output against real lidar sensor object output[185] and continued by e.g. Schaermann or Ngo, as already described in Sec. 4.1.2 and Sec. 4.1.3. However, no re-simulation for validation of APSS simulation in literature considers epistemic uncertainty, so far.

Handling epistemic and aleatory uncertainties means combining an epistemic interval and an (estimated) aleatory cumulative distribution function (CDF). This cannot result in a new stretched CDF, as Roy and Balch already state: *"While it is common practice to treat epistemic uncertainties as random variables with uniform (or normal) distributions, when little is known about the value of an epistemic uncertainty, a probabilistic treatment is not justified and an interval characterization is appropriate."* [186a]

Williamson and Downs[187] therefore introduce the so-called probability box (p-box) that results from expanding (not stretching) an aleatory CDF (or empirical distribution function (EDF)) with an interval, as illustrated in Fig. 4-14. For a given parameter or measurand x, it gives the possible interval of cumulative probabilities and for a given cumulative probability, it gives a possible interval of values. P-boxes can be derived from credal sets and Dempster–Shafer structures [186b], as discussed in detail e.g. by Ferson et al.[188]

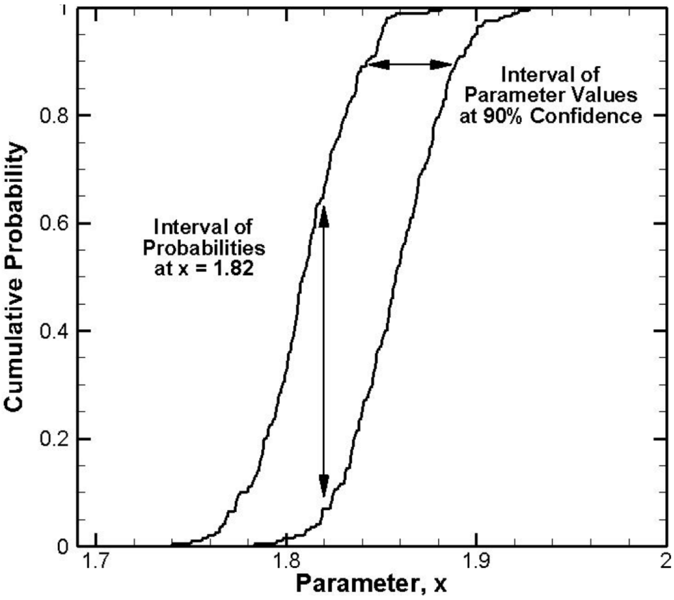


Figure 4-14: P-box for a mixture of aleatory and epistemic uncertainty from Roy and Balch.[186c]

[185] Neumann-Cosel, K. von: Virtual Test Drive (2014).

[186] Roy, C. J.; Balch, M.: A holistic approach to uncertainty quantification (2012). a: p. 380.; b: p. 367.; c: p. 365.; d: p. 370.; e: p. 377.; f: p. 379.

[187] Williamson, R. C.; Downs, T.: Probabilistic arithmetic (1990).

[188] Ferson, S. et al.: Constructing Probability Boxes and Dempster-Shafer Structures (2003).

In a first demonstration of a p-box for uncertainty aggregation, Roy and Balch validate nozzle thrust simulations for different stagnation pressures using the area validation metric (AVM), the area between the two compared CDFs, as can be seen for one sample validation in Fig. 4-15.

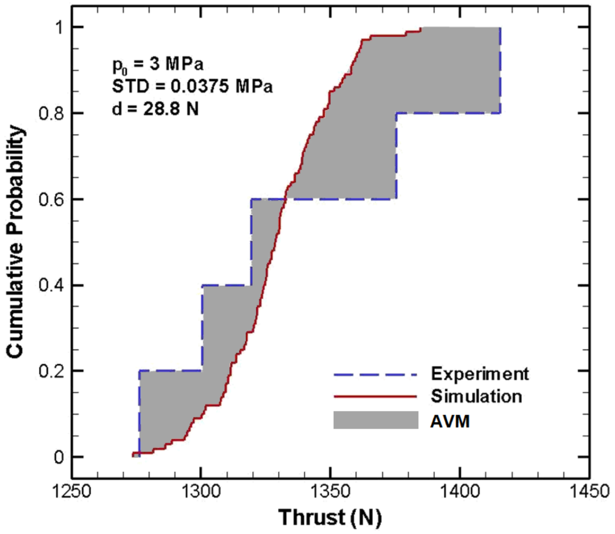
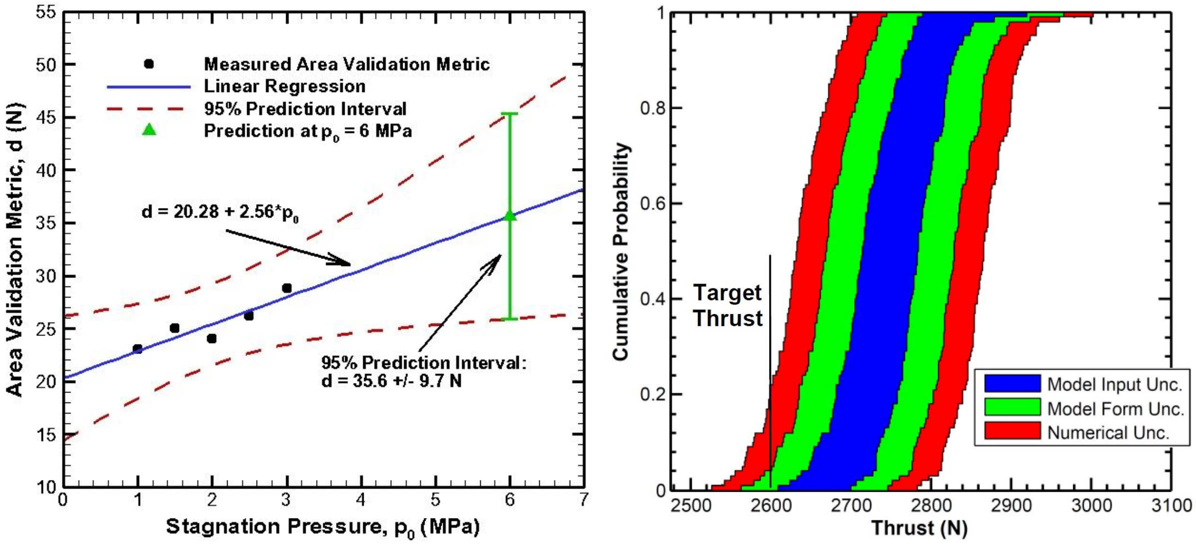


Figure 4-15: Simulated and real CDFs with the AVM in between for nozzle thrust from Roy and Balch.[186d]

They obtain metric results for different samples from their parameter space and apply linear regression for error propagation towards a new stagnation pressure that is far higher than all other pressures in their validation data, as shown in Fig. 4-16a. The $95\,\%$ confidence interval is provided by linear regression and adds (half) to the error estimate resulting in the overall model form uncertainty. Afterwards, Roy and Balch combine this model form uncertainty with the two other uncertainty sources already defined in Sec. 1.2.7: Model input and numerical uncertainty.



(a) AVM extrapolation to prediction conditions with uncertainty for nozzle thrust simulation over stagnation pressure from Roy and Balch.[186e]

(b) P-box from extrapolation of nozzle thrust simulation by Roy and Balch, separating model input, model form, and numerical uncertainty.[186f]

Figure 4-16: Prediction of validation sample errors towards application conditions by Roy and Balch.[186]

Two possibilities exist to obtain model input uncertainty from epistemic and aleatory uncertainty in reference data: Either starting with aleatory uncertainty around a first point from the epistemic interval and then shifting the center of the resulting CDF with the output from propagating epistemic uncertainty (E-outer), or the other way around (A-outer). However, Roy and Balch already find that the A-outer and E-outer propagation of an input p-boxes can differ significantly.[186d] For a fixed (unknown) bias uncertainty and a random uncertainty in an experimentally measured quantity, they suggest E-outer propagation.[186d] After model input uncertainty is propagated through the simulation, numerical uncertainty is obtained during verification of the implemented computer simulation and the hardware where it is running on. The summation of all three sources results in a p-box, as depicted in Fig. 4-16b. The inner blue p-box depicts model input uncertainty from reference data, the green part on both sides is the model form uncertainty from the predicted model error including a confidence interval, and the red outer parts arize from numerical uncertainty.

The evaluation of VV&UQ approaches by Riedmaier, Danquah et al. in Fig. 4-10 identifies bias correction as major gap in existing VV&UQ approaches besides the lack of extrapolative power instead of error extrapolation. Therefore, they provide a mechanism in their generic framework (Fig. 4-11) to enable bias correction, as shown in the detailed insight into the error learning, interference, and integration blocks from their framework in Fig. 4-17 [189a] Bias errors and uncertainties influence each other. Nevertheless, they are quantified separately.[189b]

Riedmaier, Danquah et al. generalize error learning and interference: It is possible to calculate either point errors or error intervals, depending on the metric's either deterministic point output or non-deterministic interval output. They state that an error metric with (symmetric) confidence interval extends a former deterministic simulation into non-deterministic. Furthermore, they find that an error estimate can be interpreted as an uncertainty, either as symmetrical safety factor around a simulated value or one sided, resulting in a corrected value or interval. It depends on the individual point of view, if the error prediction including its uncertainty from Roy and Balch in Fig. 4-16 resulting in model form uncertainty is uncertainty aggregation or model error correction. However, this interpretability is inherent due to the interdependency of bias error and uncertainties.

As choosing the right error model learning algorithm is crucial for confidence, it must be chosen very carefully. Danquah, Riedmaier et al. evaluate several approach regarding their applicability, amount of data learning points, over-fitting tendency, extrapolation capability and prediction reliability.[190a] Based on these requirements, basic surrogate models using the Gaussian process (GP), surrogate models using polynomial chaos expansion (PCE), and higher pronominal optimization are dismissed and simple linear regression is chosen, as also proposed by Oberkampf and Roy[191] and Roy and Balch[186].[190a]

[189] Riedmaier, S. et al.: Unified Framework and Survey for Model VV&UQ (2020). a: p. 27.; b: p. 7.; c: p. 19.

[190] Danquah, B. et al.: Statistical Validation Framework for Automotive Vehicle Simulations (2021). a: p. 13.; b: p. 15.

[191] Oberkampf, W. L.; Roy, C. J.: Verification and Validation in Scientific Computing (2010).

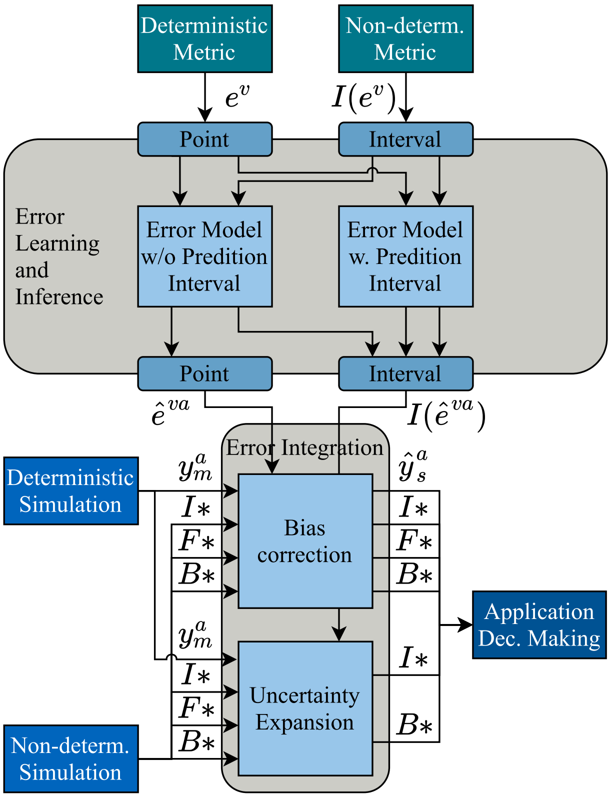


Figure 4-17: Generalization of the error pipeline by Riedmaier, Danquah et al.[189c]. The deterministic and non-deterministic simulation as well as metric are alternatives and just visualized in parallel for comparison. The stars are placeholders for measured or simulated, respectively.

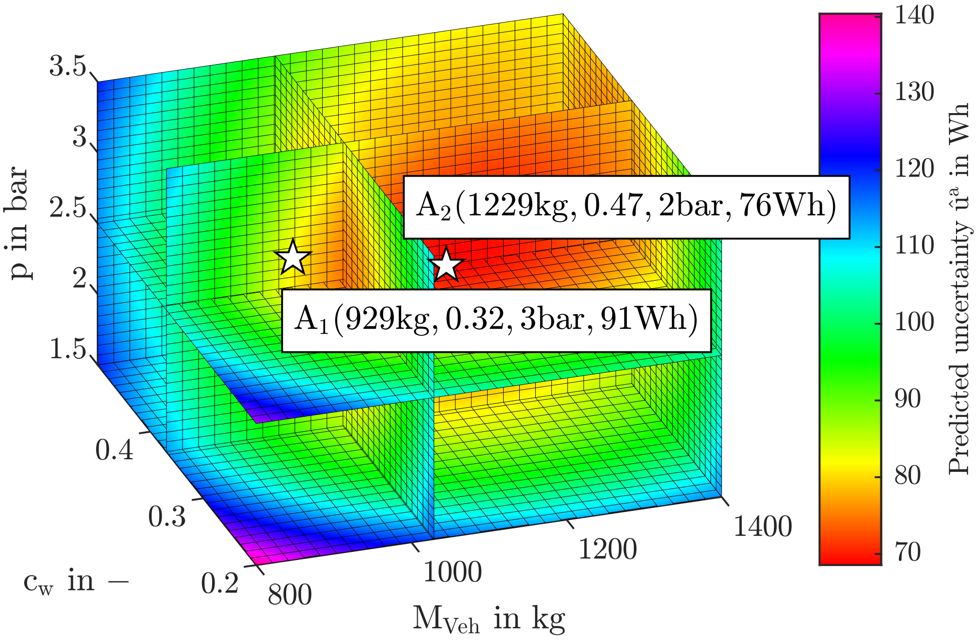


Figure 4-18: Uncertainty prediction of a vehicle consumption simulation from 10 validation samples in a 3-dimensional parameter space of vehicle mass, tire pressure, and aerodynamic drag coefficient with linear regression from Danquah, Riedmaier et al.[190b].

An exemplary result from Danquah, Riedmaier et al. is provided in Fig. 4-18, where linear regression is used to predict model form uncertainty of a vehicle consumption simulation from 10 validation samples in a 3-dimensional parameter space of vehicle mass, tire pressure, and aerodynamic drag coefficient. As expected, the plot shows lower predicted uncertainty where samples are validated and increasing uncertainty when moving away from those samples. However, selection and distribution of validation samples within the parameter space highly influences the uncertainty in the application parameter space.[190b]

4.2 Metrics for VV&UQ of APSS Simulation

Similar to sensor models themselves, also metrics for sensor model validation can be categorized by in- and output, by approach, by implementation, and by quality. Riedmaier, Danquah et al.[192] divide metric inputs into deterministic values and probability distributions, while metric outputs are separated into Boolean, probabilistic, or real-valued. However, both can be either static or dynamic, while dynamic metric outputs are neglected for the scope of this dissertation, which leads to Tab. 4-2 as excerpt from the list provided by Riedmaier, Danquah et al.[192]

As a first subsumption, deterministic metric inputs are mainly obtained when validation is performed on key performance indicators (KPIs) for feature or object level data processing like object classification or tracking. In this case, tolerance checks or bands output Boolean decisions as defined in the ISO standards for vehicle dynamics simulation[193,194,195,196], or differences/vector metrics[200] provide real values. Distributional metric inputs are gained by one or multiple measurements over time on all APSS interfaces and reflect inherent epistemic and aleatory uncertainties. While static inputs are the regular case in APSS simulation VV&UQ and evaluated with (Bayesian) hypothesis testing[197] or area validation metric[201], time series metric inputs need to be transformed to a simplified static form with approaches like KPIs[198], wavelets[199], or principal component analysis[202].

Table 4-2: Taxonomy of VV&UQ metrics including examples from Riedmaier, Danquah et al.[192]
HT: Hypothesis testing, KPI: Key performance indicator,
PCA: Principal component analysis, AVM: Area validation metric.

Metric outputs	**Deterministic metric inputs**		**Distributional metric inputs**	
	Static	**Dynamic**	**Static**	**Dynamic**
Boolean	Tolerance check[193]	Tolerance band[194,195,196]	HT[197]	HT with KPIs[198]
Probabilistic	-	-	Bayesian HT[197]	Bayesian HT with wavelets[199]
Real valued	Difference	Vector metric[200]	AVM[201]	AVM with PCA[202]

[192] Riedmaier, S. et al.: Unified Framework and Survey for Model VV&UQ (2020), pp. 13-14.

Schaermann recommends to perform APSS simulation V&V on subsequent interface levels like detection, feature, and object level.[203a] Huch[204] therefore collects and applies several metrics initially proposed e.g. by Schaermann[205] and Ackermann[206]. As already explained in Sec. 4.1.3, Ngo introduces the terms explicit and implicit sensor model evaluation to separate metrics directly applied on detections from others.[207] The collection of metrics in the following shows that (implicit) data processing output is often used as metric for sensor (detection) model validation as it is the source for requirements on sensor models.

4.2.1 Discussion of Visual Inspection of Plots as V&V

In some cases, as e.g. in recent radar simulation evaluation by Degen et al.[208], visual inspection is still chosen as validation tool. UNECE R157 requires that *"Manufacturers shall demonstrate the scope of the simulation tool, its validity for the scenario concerned as well as the validation performed for the simulation tool chain."* [209] However, it is unclear what exactly is required for the therefore demanded *"correlation of the outcome with physical tests."* [209] As this point is left open, e.g. German TÜV Süd together with dSPACE provide only some visual inspection of plots for comparison of real and simulated range, velocity, and acceleration of the tested vehicle as a validation of their simulation. Furthermore, a calculation of simulated and time-to-collision (TTC) criticality metrics is also just visually inspected in their recent report for virtual homologation of an automated lane keeping system (ALKS) according to UNECE R157.[210]

Viehof discusses CDF as intermediate processing step for the measured / simulated signal, as it considers all information contained in a signal. He finds that it has no absolute time reference and

[193] International Organization for Standardization: ISO 19365:2016(E) (2016)

[194] International Organization for Standardization: ISO 19364:2016(E) (2016)

[195] International Organization for Standardization: ISO 19585:2019(E) (2019)

[196] International Organization for Standardization: ISO 22140:2021(E) (2021)

[197] Rebba, R.; Mahadevan, S.: Computational methods for model reliability assessment (2008)

[198] Viehof, M.: Objektive Qualitätsbewertung von Fahrdynamiksimulationen (2018)

[199] Jiang, X.; Mahadevan, S.: Bayesian wavelet method for multivariate model assessment (2008)

[200] Sarin, H. et al.: Comparing Time Histories for Validation of Simulation Models (2010)

[201] Ferson, S. et al.: Model validation and predictive capability (2008). pp. 2416-2419.

[202] Xi, Z. et al.: Validation Metric for Dynamic System Responses under Uncertainty (2015)

[203] Schaermann, A.: Systematische Bedatung und Bewertung umfelderf. Sensormodelle (2020). a: p. 32.; b: pp. 20-21.

[204] Huch, S.: Metrik zur Bewertung der Lidar-Sensor-Simulation (2018).

[205] Schaermann, A. et al.: Validation of vehicle environment sensor models (2017).

[206] Ackermann, S. M.: Systematische Untersuchung von Radar Tracking (2017).

[207] Ngo, A. et al.: Multi-Layered Measuring the Simulation-to-Reality Gap for Radar (2021), pp. 4010-4011.

[208] Degen, R. et al.: Methodical Approach to the Development of a Radar Sensor Model (2021).

[209] United Nations Economic Commission for Europe: UNECE 157 (2021), p. 37.

[210] Miethaner, C.; Stavesand, J.-E.: Virtual homologation of an ALKS according to UNECE R157 (2022), p. 13.

therefore independent of the signal frequency and is thus not suitable for identifying frequency-dependent deviations between signals, under specific conditions. However, he states that the specific conditions are relatively unlikely in practical validation. In order to nevertheless be able to identify deviations of this type, it is proposed to supplement the CDF evaluation with a comparison of the power spectral density (PSD).[211a]

Holder et al. use CDFs / EDFs intensively for evaluation of synthetic radar data. After comparison of CDFs from simulated and real received power, Holder et al. use OGs and automatically derived free-space estimations for evaluation of synthetic radar detections by visual inspection of accumulated plots.[212] Additionally, decision making algorithm output regarding driveable paths on the road are given as KPI plot. Accordingly, the evaluation of a new approach for object simulation in former work of the author is performed by simple object position and dimension comparisons in a static scenario.[213]

In his dissertation, Holder evaluates his synthetic radar detections by comparison of real and simulated radar cross-section (RCS) and SNR. He starts with a visual inspection of RCS over $1/r$ and as amplitude characteristic over $1/r$ and compares EDFs with $\log(1\mathrm{EDF})$ scaling.[214a] Furthermore, signal propagation modeling is evaluated with range-Doppler plots and EDF plots of the SNR with regular scaling [214a]. Additionally, for evaluation of reflectivity modeling, $360°$-RCS plots and SNRs over range r are visually compared.[214b] Uncertainty modeling is assessed on noise in received power via box plots and on range accuracy with EDF[214c]. Finally, an object tracking algorithm is applied on synthetic and real radar detections. The resulting trajectories are visually compared accompanied by some KPIs like mean range to object or range at first sight.[214d] Overall, no objective metrics as listed in the following are used.

For checking systematic errors in measurement or reference data collection and or setup of the re-simulation, Viehof proposes the application of t-statistics and he chooses the full overlap of the $95\,\%$ confidence intervals on the CDFs from the arithmetic mean of several measurements and simulations as objective metric. His research proves that this specific metric is able to detect all errors like wrong virtual sensor mounting position.[211]

In his actual validation study, Viehof applies t-statistics for estimation of the PDF from MVCs and extracts a tolerance interval, e.g. of $95\,\%$, as shown in Fig. 4-4. Then he awards labels depending on the overlap of the estimated t-distributions and depending on the confidence levels, as depicted in Fig. 4-5. The discussion and the counterarguments about this approach are already contained in Sec. 4.1.1.

[211] Viehof, M.: Objektive Qualitätsbewertung von Fahrdynamiksimulationen (2018). a: pp.83-87.; b: p. 81.

[212] Holder, M. F. et al.: How to evaluate synthetic radar data? (2020).

[213] Linnhoff, C. et al.: Object Based Generic Perception Object Model (2022).

[214] Holder, M. F.: Synthetic Generation of Radar Sensor Data for Virtual Validation (2021). a: pp. 119-126.; b: pp. 127-136.; c: pp. 137-140.; d: p. 143-151.

In contrast to subjective visual inspections, as already explained in Chap. 1, a complete VV&UQ is inevitable for the credibility of APSS simulation used as a tool for safety validation of ADS. Therefore, to follow the overall goal in this dissertation to gain model credibility, in the following only publications that include at least actual metrics and sophisticated validation studies are considered.

4.2.2 Collection of VV&UQ Metrics Applied for APSS Simulation

In this section, actual metrics that have already been applied for APSS simulation VV&UQ in literature are collected, briefly described and listed in Tab. 4-3. A collection of all possible metrics for APSS simulation is out of the scope of this work. Nonetheless, a comprehensive overview is provided, that allows metric selection for further development and application in this dissertation. Besides, the application notes from literature are shortly discussed.

Ackermann provides a study on metrics for object tracking evaluation on synthetic data[215a]. He explains the progress from Hausdorff distance over Optimal Mass Transfer (OMAT) and Optimal Sub-Pattern Assignment (OSPA) metrics towards the proposal by Rahmathullah et al. For his own application, he chose OSPA-T as the most commonly used metric and the metric from Rahmathullah et al. as the SotA of tracking analysis at that time.[215b] Nevertheless, he finds that both tracking metrics have parameters themselves, which have to be chosen appropriately and documented.[215c]

Huch lists many possible metrics for detection, feature (OGs and segmentation), and (tracked and classified) objects and selects 18 of them for application and evaluation[216]. He concludes that the metrics on feature level like OGs (No. 13-20 in Tab. 4-3) highly rely on parameters like grid cell size.[216b] He finds that point-to-point distance (metric 10) in point clouds of lidar detections relies on a balanced number of detections and that Hausdorff distance (metric 8) is sensitive to outliers, which prevents its usage on point clouds. Metrics on object level, especially on tracking (No. 25-28 in Tab. 4-3) in his evaluation heavily rely on parameters, too.[216b] Intersection over union (IoU) suffers from its inability to penalize larger distances, when there is no overlap of the bounding boxes. Furthermore, he stresses that time synchronization of simulated and real data plays a critical role, when metrics are applied on tracking or for the mentioned OGs for time accumulated data. His overall goal however, to find a comprehensive metric that is build of metrics from different functional layers fails, while at least correlation between metrics on different layers is shown.[216a]

Huch suggests to extend the calculation of all presented OGs metrics by first optimizing the created simulated OGs.[216c] Possible editing steps are translating, rotating or scaling the OGs. After each editing step, the cross-correlation coefficient is calculated between the real and the

[215] Ackermann, S. M.: Systematische Untersuchung von Radar Tracking (2017). a: pp. 45-59.; b: p. 67.; c: p. 75.

[216] Huch, S.: Metrik zur Bewertung der Lidar-Sensor-Simulation (2018). a: pp. 89-92.; b: pp. 93-98.; c: p 38.

edited simulated OGs, aiming to maximize it e.g. regarding cross-correlation. The calculation of the metrics can also take into account how many processing steps are needed to maximize the cross-correlation coefficient. Furthermore, if OGs would be accumulated over time, the metrics could be used for dynamic and even distributional input data.

Schaermann and Hanke perform validation of lidar sensor simulation on detection and feature (OG) level with OE, $\mathrm{C_B}$, and $\mathrm{C_P}$.[217,218] To be able to use all three methods that are originally build for image comparison, data must be available in a map schematic. For this purpose, lidar detections are sorted into a matrix-shaped range view image, which is possible for regular (spherical) beam patterns, and further available information on the detections like range or intensity can be used as the (RGB)-values of such pixel-like fields. In mathematical sense, the detections must be projected into a matrix with azimuth angles as columns and elevation angles (layers) as rows to calculate their Euclidean distance matrix as illustrated in Fig. 4-19.[218]

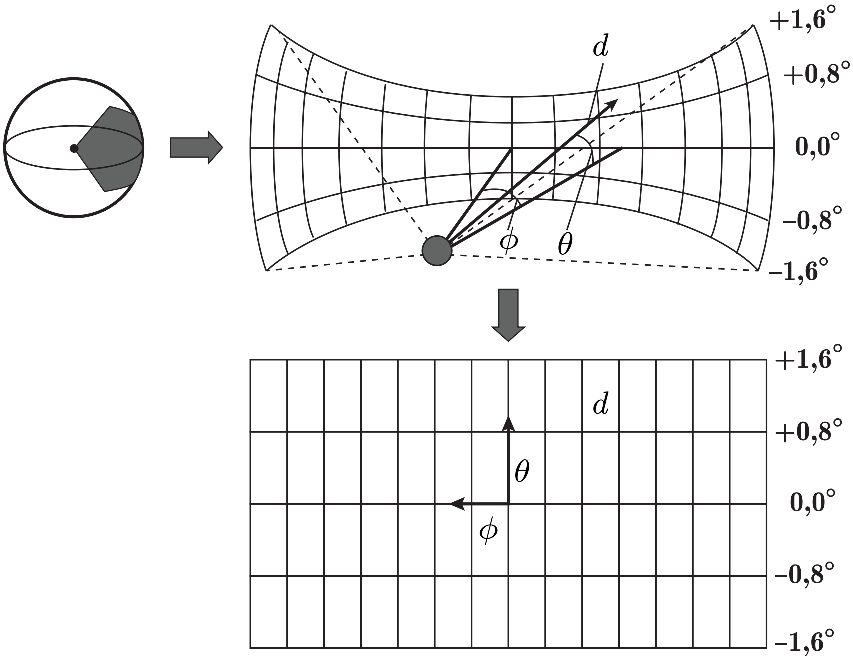


Figure 4-19: Projection of spherical coordinates into Euclidean distance matrix from Schaermann[219c].

Schaermann chooses in his dissertation eight different metrics for evaluation of a phenomenological lidar object model regarding real and simulated object positions over time.[219a] He applies the metrics 1, 4, 6, 20, 21, 23, 30, and 32 on object position vectors $\boldsymbol{\zeta}, \widetilde{\boldsymbol{\zeta}}$ and finds that Kullback–Leibler divergence D_{KL} and AVM outperform the other metrics in absolute error detection, while correlations and dynamic time warping are insensitive to phase shift. However, as the model object list output is more noisy than the real object trajectories due to low pass filtering from applied Kalman filters, the chosen metrics struggle to give reasonable evaluation results

[217] Schaermann, A. et al.: Validation of vehicle environment sensor models (2017).

[218] Hanke, T. et al.: Generation and validation of virtual point cloud data for automated driving systems (2017), pp. 4-6.

[219] Schaermann, A.: Systematische Bedatung und Bewertung umfelderf. Sensormodelle (2020). a: pp. 59-78.; b: pp. 79-101.; c: p. 86.

and he therefore decides to apply additional intensive auto-correlation (metric 22) analysis on the time signals, which is not included in Tab. 4-3.

In a second study, Schaermann chooses four metrics for a physical lidar detection model to evaluate it on detection (Euclidean distance matrix) and feature (OG) level.[219b] For the physical model and the metrics 2, 5, 13, and 18, he concludes that systematic model errors can be detected well, but can hardly be distinguished. Therefore, he recommends to use static scenarios in combination with test specimens for the detection of errors that may arise due to sensor displacement or rotation. Another finding is the necessity of unfiltered measurements for sensor model validation, otherwise non-systematic errors such as object losses can be overlooked, which represent critical scenarios and must therefore also be represented in sensor models.

Berghöfer stresses the challenges in data generation for APSS model validation, especially regarding the re-simulation approach and the objects and materials (not) available or (not) calibrated in simulation.[220] For his analysis, he applies the metrics 10 and 11 on detection level and metrics 2 (normed), 14, 15, 16, 17, and 19 on feature (OG) level. His first observation is that metrics 15 and 17 are useless for sparse point clouds, since most cells are free and both metrics tend towards unity. Still, they are used to help interpreting metrics 14 and 16, as both a simulated OG occupied on all cells and a simulated OG completely matching the real OG would return unity and only the former help distinguishing these cases.[220a] A second observation by Berghöfer is that all metrics applied to OGs in his work have similar trajectories and correlate over time of the performed and simulated parking scenario.[220b] A third observation from his work is that metrics 10 and 11 on detections show a similar tendency, whereas 10 highly depends on the overall number of detections, as it is not normalized.[220b]

The challenging selection and sometimes parametrization of metrics is discussed in former work of the author on the example of occupied cells ratio (metric 14).[221] Tamm-Morschel applies metrics 10 and 14 for the evaluation of his lidar detection model.[222] As both metrics are very sensitive for a difference in the number of detections, the new sensor model has difficulties to outperform an ideal model, as the modeling approach tends to produce less detections than the idealized model.

The same metrics are applied on simulated and real detections in former work of the author[223] together with metrics 24 and 27 on objects that are identified from the real / simulated detections with the same algorithm. All metrics seem to be equally suited for validation, as all show the tendency of the benchmarked lidar simulation to have lower fidelity for far detections/objects. On the other hand, they fail to distinguish the (similar) modeling approaches. However, the

[220] Berghöfer, M.: Generierung realer und synthetischer Sensordaten zur Simulations-Validierung (2019). a: p. 72.;b: p. 74.

[221] Rosenberger, P. et al.: Towards a Generally Accepted Validation Methodology for Sensor Models (2019).

[222] Tamm-Morschel, J. F.: Erweiterung eines Lidar-Sensormodells (2019).

[223] Rosenberger, P. et al.: Benchmarking and Functional Decomposition of Lidar Sensor Models (2019).

correlation between detection and object list level metrics is evident and further usage of such two-level evaluation is recommended.

Aust presents a new approach for APSS object output simulation by mimicking actual tracking algorithm behavior without performing the computational-expensive calculations involved.[224] He evaluates the fidelity of the approach by also applying the metrics 24 and 27 besides root mean squared error (RMSE) (metric 7) for comparison of simulated and real object positions, velocities, and dimensions. As object dimensions are besides their positions a key factor of the new object list simulation approach, intersection over union and RMSE are better suited in his case. RMSE for comparing object dimensions and trajectories/positions from simulated and real detections is also applied in former work of the author, when synthetic object list generation is evaluated on static and dynamic vehicles.[225]

Eder et al. apply machine learning to learn radar detection models with object list input from synthetic radar detection lists and GT object lists simulated with a phenomenological radar detection model that incorporates ray tracing. They compare their results from a kernel density estimator against four different neural networks.[226] For the comparison, they apply three different metrics: They start with the absolute mean error (metric 3), develop an own metric to obtain the mean of the mean detection to bounding box distance (metric 12), and finally use the Kullback–Leibler divergence (metric 30) to measure the differences of the spatial distributions within Cartesian bins of $0.1\,\mathrm{m}$. All metrics show that the kernel density estimator significantly outperforms the neural networks, while the metrics are no further discussed.

In a second publication, Eder et al. present a hybrid radar detection model that starts with ray casting towards object bounding boxes and then adds further radar-specific effects like range and angular dependent detection existence probabilities and range measurement accuracy.[227] Two validation studies are performed. The first one compares real and synthetic radar detections accumulated data for a full scenario, the second adds time and position reference to only compare synthetic and real data from the same aspect ratio towards the hit object. The validation of the radar detection simulation against real radar detections is based on testing the hypothesis that the simulated and real detection lists are from the same distribution. Therefore, the efficient multi-dimensional Kolmogorov-Smirnov test is chosen to test the hypothesis as described by Fasano and Franceschini.[228] Furthermore, Eder et al. use the frequency of positive K-S tests for multiple repetitions of experiments as validation metric (metric 33). During their validation study, they find that their radar detection model fails to replicate detections in the bounding box center that are present in the real data. Real data against itself accumulated from several measurement

[224] Aust, P.: Entwicklung eines lidartypischen Objektlisten-Sensormodells (2019).

[225] Rosenberger, P. et al.: Sequential lidar sensor system simulation (2020).

[226] Eder, T. et al.: Data Driven Radar Detection Models (2019), p. 4.

[227] Eder, T. et al.: Szenarienbasierte Validierung eines hybriden Radarmodells (2020).

[228] Fasano, G.; Franceschini, A.: A multidimensional version of the Kolmogorov–Smirnov test (1987).

runs of the scenario mostly gets positive test results. Still, the frequency is significantly lower when position and time stamps are fitted at first and only very few detection lists are validated.[227]

In his dissertation, Eder follows up on the two already presented publications.[229] At first, the same radar detection model and the same metrics (3, 12, 30) as in the first publication are presented and the same results as above are obtained when neural networks are compared against the kernel density estimator approach.[229a] Then special focus is laid on hypothesis testing for APSS model validation. Eder starts this section with stating that *"acceptance, intuitiveness and interpretability of validation methodologies is of utmost importance."* [229a] He furthermore states that for the frequency of positive K-S tests, which he uses as validation metric, several test runs should be accumulated to gather enough data. Additionally, he finds that only data from the same position and time within the re-simulated scenario should be compared. Therefore, he suggests and performs 100 repetitions of the same scenario in real world. Even if the simulation performs poorly in the validation, at least real against real data finally has higher frequency of positive tests to come from the same underlying distribution.[229b]

Ngo et al. state that radar detection model validation must be performed on different subsequent interfaces like detection and (tracked) object lists. They find that despite a *"sensor model might lack accuracy in a direct comparison, the results from a subsequent algorithm can still show a great consensus."* [230] Consequently, as already explained in Sec. 4.1.3, Ngo et al. follow a multi-layered approach for validation of radar detection simulations as shown in Fig. 4-8.[231] They obtain real and synthetic radar detections and reference data from eight different scenarios for the V&V of an idealized, a data-driven and a ray-tracing-based radar detection model. They split the evaluation into four different fidelity levels of the evaluation as depicted in Fig. 4-9 to obtain an overall simulation-to-reality gap G as the average over the four fidelity levels.

In this regard, Ngo et al. choose several metrics from literature without detailed reasoning for the different levels. Fidelity level III for explicit radar detection comparison consists of detection-list's Wasserstein and point-cloud-to-point-cloud distance (metrics 9, 10). Level IV consists of specific Wasserstein distances for range, azimuth, and Doppler velocity and a simple difference in the number of detections. In level I for implicit sensor model comparison, the OSPA (metric 25) is applied to obtained real and synthetic trajectories besides the intersection over union (metric 24). The fidelity level II applies RMSE on longitudinal and lateral object positions (metric 7) and the absolute cardinality error. As already mentioned, their metric selection and averaging of arbitrary fidelity levels is questionable. Nevertheless, the sheer amount of performed scenarios within the evaluation from Ngo et al. besides the scenario design itself and the obtained results set benchmarks for future work on radar detection simulation and its V&V.

[229] Eder, T.: Simulation of Automotive Radar Point Clouds in Standardized Frameworks (2021). a: p. 57.;b: p. 86.

[230] Ngo, A. et al.: A Sensitivity Analysis Approach for Evaluating a Radar Simulation (2020), p. 123.

[231] Ngo, A. et al.: Multi-Layered Measuring the Simulation-to-Reality Gap for Radar (2021).

In a follow-up publication, Ngo et al. present a new approach for finding a metric for APSS model V&V by using deep learning.[232] It is called "Deep Evaluation Metric" (metric 34) and uses the PointNet++ architecture. The training (and testing) data consists of real and synthetic radar detections. The model is optimized to distinguish the detection source. When applied to new (synthetic) data, its classification probability is used as metric. Ngo et al. compare the Deep Evaluation Metric results with conventional ones like d_{PP} (metric 10) and Wasserstein distance d_{Wa} (after normalizing all three with their respective maximum to the interval $[0, 1]$ and show that the trained neural network is able to distinguish data even where conventional metrics fail. The deep learning approach for metrics looks promising for falsification of models, but such a learned black-box for credibility assessment of simulations for safety critical simulations cannot be recommended for decision making, at least not alone.

Magosi et al. present a non-deterministic radar detection model and its validation.[233] They state that statistical evaluation methods are best suited for such highly stochastic simulation, as the detections can be treated as realizations of a PDF.[233a] They reference Maupin et al.[234] as source for their metric selection for comparing histograms and EDFs and decide for Jensen-Shannon distance d_{JS} (metric 31). It is the square-root of the Jensen-Shannon divergence, a symmetrized version of the D_{KL} and is an actual metric in mathematical sense. As it is the state of the art in validation, Magosi et al. also apply re-simulation of real measurements and reference data captured beforehand, while special attention is put on the reference data measurement equipment and its accuracy.[233b]

The following table Tab. 4-3 lists and summarizes all metrics applied in the mentioned publications. Besides metric number, name, and formula/description, the output value range from best to worst (B-W) is given and the applying publications are attached. While there cannot be a claim for completeness in this case, a comprehensive overview is provided about the sheer amount and diversity of metrics applied for APSS simulation in literature. While it is consensus to apply the method of re-simulation of real world experiments for model validation, the best suitable metric has not yet been found.

Table 4-3: List of metrics applied for APSS simulation. B-W: Best/worst results

#	Metric	Formula / Description	B-W	Applic.
1	Manhattan distance $\boldsymbol{d}_{\mathbf{Ma}}(\boldsymbol{\zeta}, \widetilde{\boldsymbol{\zeta}})$ for $\boldsymbol{\zeta}, \widetilde{\boldsymbol{\zeta}} \in \mathbb{R}^n$	$\lVert\boldsymbol{\zeta} - \widetilde{\boldsymbol{\zeta}}\rVert_1 = \sum_{i=1}^{n} \lvert\zeta_i - \widetilde{\zeta}_i\rvert$	$[0, \infty[$	219
2	Overall Error (**OE**) for $\boldsymbol{Z}, \widetilde{\boldsymbol{Z}} \in \mathbb{R}^{n \times m}$	$\lVert\boldsymbol{Z} - \widetilde{\boldsymbol{Z}}\rVert_1$ $= \sum_{i=1}^{n} \sum_{j=1}^{m} \lvert\zeta_{i,j} - \widetilde{\zeta}_{i,j}\rvert$	$[0, \infty[$	216, 217, 218, 219, 220

[232] Ngo, A. et al.: Deep Evaluation Metric (2021).

[233] Magosi, Z. F. et al.: Evaluation of Physical Radar Perception Sensor Models (2022). a: p. 10.; b: pp. 4-8.

[234] Maupin, K. A. et al.: Validation Metrics for Deterministic and Probabilistic Data (2019).

#	Metric	Formula / Description	B-W	Applic.
3	Mean Error $\overline{\boldsymbol{d}}(\boldsymbol{\zeta}, \widetilde{\boldsymbol{\zeta}})$ for $\boldsymbol{\zeta}, \widetilde{\boldsymbol{\zeta}} \in \mathbb{R}^n$	$\frac{1}{n} \sum_{i=1}^{n} \lvert \zeta_i - \widetilde{\zeta}_i \rvert$	$[0, \infty[$	226, 229
4	Euclidean distance $\boldsymbol{d}_{\mathbf{Eu}}(\boldsymbol{\zeta}, \widetilde{\boldsymbol{\zeta}})$ for $\boldsymbol{\zeta}, \widetilde{\boldsymbol{\zeta}} \in \mathbb{R}^n$	$\lVert \boldsymbol{\zeta} - \widetilde{\boldsymbol{\zeta}} \rVert_2 = \sqrt{\sum_{i=1}^{n} (\zeta_i - \widetilde{\zeta}_i)^2}$	$[0, \infty[$	216
5	Residual sum of squares (**RSS**) for $\boldsymbol{Z}, \widetilde{\boldsymbol{Z}} \in \mathbb{R}^{n \times m}$	$\sum_{i=1}^{n} \sum_{j=1}^{m} (\zeta_{i,j} - \widetilde{\zeta}_{i,j})^2$	$[0, \infty[$	219
6	Chebyshev distance $\boldsymbol{d}_{\mathbf{Ch}}(\boldsymbol{\zeta}, \widetilde{\boldsymbol{\zeta}})$ for $\boldsymbol{\zeta}, \widetilde{\boldsymbol{\zeta}} \in \mathbb{R}^n$	$\lVert \boldsymbol{\zeta} - \widetilde{\boldsymbol{\zeta}} \rVert_\infty = \max \lvert \zeta_i - \widetilde{\zeta}_i \rvert$	$[0, \infty[$	219
7	Root mean squared error (**RMSE**)	$\sqrt{\frac{1}{n} \sum_{i=1}^{n} (\zeta_i - \widetilde{\zeta}_i)^2}$	$[0, \infty[$	216, 224 225, 231
8	Hausdorff distance $\boldsymbol{d}_{\mathbf{Ha}}$ for compact subsets $\boldsymbol{\mathcal{Z}}, \widetilde{\boldsymbol{\mathcal{Z}}}$ of a metric space and a distance $d(\boldsymbol{\zeta}, \widetilde{\boldsymbol{\zeta}})$ (e.g. Euclidean)	$\max(\max_{\boldsymbol{\zeta} \in \boldsymbol{\mathcal{Z}}} \min_{\widetilde{\boldsymbol{\zeta}} \in \widetilde{\boldsymbol{\mathcal{Z}}}} d(\boldsymbol{\zeta}, \widetilde{\boldsymbol{\zeta}}),$ $\max_{\widetilde{\boldsymbol{\zeta}} \in \widetilde{\boldsymbol{\mathcal{Z}}}} \min_{\boldsymbol{\zeta} \in \boldsymbol{\mathcal{Z}}} d(\boldsymbol{\zeta}, \widetilde{\boldsymbol{\zeta}}))$	$[0, \infty[$	215, 216
9	Wasserstein distance $\boldsymbol{d}_{\mathbf{Wa}}$ for point sets $\boldsymbol{\mathcal{Z}}, \widetilde{\boldsymbol{\mathcal{Z}}}$ and a distance $d(\boldsymbol{\zeta}, \widetilde{\boldsymbol{\zeta}})$ (e.g. Euclidean)	$\frac{\sum_{i=1}^{n} \sum_{j=1}^{m} f_{m,n} d(\boldsymbol{\zeta}, \widetilde{\boldsymbol{\zeta}})}{\sum_{i=1}^{n} \sum_{j=1}^{m} f_{m,n}},$ $f_{m,n}$: optimal flow to rearrange the distributions	$[0, \infty[$	231, 232
10	Point cloud to point cloud distance $\boldsymbol{d}_{\mathbf{PP}}$ for $\boldsymbol{\zeta} \in \mathbb{R}^n$, $\widetilde{\boldsymbol{\zeta}} \in \mathbb{R}^m$	$\max(\sum_{i=1}^{n} \min \lVert \zeta_i - \widetilde{\boldsymbol{\zeta}} \rVert_1,$ $\sum_{j=1}^{m} \min \lVert \boldsymbol{\zeta} - \widetilde{\zeta}_j \rVert_1)$	$[0, \infty[$	216, 220, 222, 223, 231, 232
11	Point cloud center of gravity distance $\boldsymbol{d}_{\mathbf{PC}}$ for two subsets $\boldsymbol{\mathcal{Z}}, \widetilde{\boldsymbol{\mathcal{Z}}}$ of n, m detections $\boldsymbol{\zeta}, \widetilde{\boldsymbol{\zeta}} \in \mathbb{R}^3$	$\lVert \boldsymbol{\zeta}_{\text{cog}} - \widetilde{\boldsymbol{\zeta}}_{\text{cog}} \rVert_2,\ \boldsymbol{\zeta}_{\text{cog}} = \frac{1}{n} \lVert \boldsymbol{\mathcal{Z}} \rVert_1$	$[0, \infty[$	216, 220
12	Bounding box error (**BBE**) for two subsets $\boldsymbol{\mathcal{Z}}, \widetilde{\boldsymbol{\mathcal{Z}}}$ of n,m detections $\boldsymbol{\zeta}, \widetilde{\boldsymbol{\zeta}} \in \mathbb{R}^3$ and their bounding boxes $\boldsymbol{\mathcal{B}}, \widetilde{\boldsymbol{\mathcal{B}}}$	$\frac{1}{n} \sum_{i=1}^{n} \lVert d_{\text{BB}}(\boldsymbol{\zeta}, \boldsymbol{\mathcal{B}}) - d_{\text{BB}}(\widetilde{\boldsymbol{\zeta}}, \widetilde{\boldsymbol{\mathcal{B}}}) \rVert_2^2,$ $d_{\text{BB}}(\boldsymbol{\zeta}, \boldsymbol{\mathcal{B}}) = \frac{1}{n} \sum_{i=1}^{n} \lVert \zeta_i - \boldsymbol{\mathcal{B}} \rVert_1$	$[0, \infty[$	226, 229
13	Map score (**MS**) for $\boldsymbol{Z}, \widetilde{\boldsymbol{Z}} \in \mathbb{R}^{n \times m}$ with $\zeta_{i,j} \in \{0, 1\}$, 0: free, 1: occ.	$\sum_{i=1}^{n} \sum_{j=1}^{m} [1 + \log_2(\lVert \boldsymbol{Z} \rVert_1 \cdot \lVert \widetilde{\boldsymbol{Z}} \rVert_1$ $+ \lVert \neg \boldsymbol{Z} \rVert_1 \cdot \lVert \neg \widetilde{\boldsymbol{Z}} \rVert_1)],$ $\neg \boldsymbol{Z} = \mathbb{1} - \boldsymbol{Z}$	$[nm, 0]$	216, 219
14 / 15	Occupied / Free cells ratio (**OCR** / **FCR**) for $\boldsymbol{Z}, \widetilde{\boldsymbol{Z}} \in \mathbb{R}^{n \times m}$ with $\zeta_{i,j} \in \{0, 1\}$, 0: free, 1: occ.	OCR: $\frac{\lVert \boldsymbol{Z} \rVert_1}{\lVert \widetilde{\boldsymbol{Z}} \rVert_1}$, FCR: $\frac{\lVert \neg \boldsymbol{Z} \rVert_1}{\lVert \neg \widetilde{\boldsymbol{Z}} \rVert_1}$, $\neg \boldsymbol{Z} = \mathbb{1} - \boldsymbol{Z}$	$[1, 0/\infty]$	216, 220, 221, 222 223

#	Metric	Formula / Description	B-W	Applic.
16	Occupied picture distance function (**OPD**) for $\boldsymbol{Z}, \widetilde{\boldsymbol{Z}} \in \mathbb{R}^{n \times m}$ with $\zeta_{i,j} \in \{0,1\}$, 0: free, 1: occ.	$1 - \frac{\sum_{n=1}^{N_\mathrm{o}} \min(d_\mathrm{M}, r_\mathrm{sr})}{n_\mathrm{o} r_\mathrm{sr}}$, $d_\mathrm{M} = \min\limits_{\widetilde{\boldsymbol{Z}}} \Vert (i,j)_{n_\mathrm{o}} - (i,j)_{\widetilde{\boldsymbol{Z}}} \Vert_1$, n_o: No. of occ. cells in $\boldsymbol{Z}$, r_sr: Search radius in $\widetilde{\boldsymbol{Z}}$	$[1, 0]$	216, 220
17	Unoccupied picture distance function (**UPD**) for $\boldsymbol{Z}, \widetilde{\boldsymbol{Z}} \in \mathbb{R}^{n \times m}$ with $\zeta_{i,j} \in \{0,1\}$, 0: free, 1: occ.	$1 - \frac{\sum_{n=1}^{N_\mathrm{u}} \min(d_\mathrm{M}, r_\mathrm{sr})}{n_\mathrm{u} r_\mathrm{sr}}$, $d_\mathrm{M} = \min\limits_{\neg\widetilde{\boldsymbol{Z}}} \Vert (i,j)_{n_\mathrm{u}} - (i,j)_{\neg\widetilde{\boldsymbol{Z}}} \Vert_1$, n_u: No. of unocc. cells in $\boldsymbol{Z}$, r_sr: Search radius in $\neg\widetilde{\boldsymbol{Z}}$	$[1, 0]$	216, 220
18	Picture distance function (**PD**) for $\boldsymbol{Z}, \widetilde{\boldsymbol{Z}} \in \mathbb{R}^{n \times m}$ with $x = \text{Ᵽ}(x = 1)$	$\sum\limits_{f \in F} d_\mathrm{M}(\boldsymbol{\zeta}, \widetilde{\boldsymbol{\zeta}}, f) + d_\mathrm{M}(\widetilde{\boldsymbol{\zeta}}, \boldsymbol{\zeta}, f)$, $d_\mathrm{M}(\boldsymbol{\zeta}, \widetilde{\boldsymbol{\zeta}}, f) = (d_\mathrm{M} \mid \zeta_{i,j} = f) =$ $= \min\limits_{\widetilde{\boldsymbol{Z}}_f} \Vert (i,j)_{n_f} - (i,j)_{\widetilde{\boldsymbol{Z}}_f} \Vert_1$	$[0, \infty]$	219
19	Baron cross-correlation ($\boldsymbol{C}_\mathbf{B}$) for $\boldsymbol{Z}, \widetilde{\boldsymbol{Z}} \in \mathbb{R}^{n \times m}$ with $\zeta_{i,j} \in \{0,1\}$, 0: free, 1: occ.	$\frac{\overline{\boldsymbol{Z} \cdot \widetilde{\boldsymbol{Z}}} - \overline{\boldsymbol{Z}} \cdot \overline{\widetilde{\boldsymbol{Z}}}}{\overset{\vdash\!\dashv}{\boldsymbol{Z}} \cdot \overset{\vdash\!\dashv}{\widetilde{\boldsymbol{Z}}}}, \overline{\boldsymbol{Z}} = \frac{1}{nm} \Vert \boldsymbol{Z} \Vert_1$, $\overset{\vdash\!\dashv}{\boldsymbol{Z}} = \sqrt{\frac{1}{nm} \sum\limits_{i=1}^{n} \sum\limits_{j=1}^{m} (\zeta_{i,j} - \overline{\boldsymbol{Z}})^2}$	$[1, -1]$	216, 217, 218, 220
20	Pearson correlation ($\boldsymbol{C}_\mathbf{P}$) for $\boldsymbol{Z}, \widetilde{\boldsymbol{Z}} \in \mathbb{R}^{n \times m}$ with $\zeta_{i,j} \in \{0,1\}$, 0: free, 1: occ.	$\frac{(\boldsymbol{Z} - \overline{\boldsymbol{Z}}) \cdot (\widetilde{\boldsymbol{Z}} - \overline{\widetilde{\boldsymbol{Z}}})}{\Vert \boldsymbol{Z} - \overline{\boldsymbol{Z}} \Vert_2 \Vert \widetilde{\boldsymbol{Z}} - \overline{\widetilde{\boldsymbol{Z}}} \Vert_2}$	$[\pm 1, 0]$	216, 217, 218, 219
21	Cross-correlation ($\boldsymbol{C}_\mathbf{C}$) for $\boldsymbol{\zeta}, \widetilde{\boldsymbol{\zeta}} \in \mathbb{R}^n$, τ: shift	$\frac{(n-\tau) \sum\limits_{i=1}^{n-\tau} \zeta_i \widetilde{\zeta}_{i+\tau} - \sum\limits_{i=1}^{n-\tau} \zeta_i \sum\limits_{i=1}^{n-\tau} \widetilde{\zeta}_{i+\tau}}{\sqrt{(n-\tau) \sum\limits_{i=1}^{n-\tau} \zeta_i^2 - (\sum\limits_{i=1}^{n-\tau} \zeta_i)^2} \sqrt{(n-\tau) \sum\limits_{i=1}^{n-\tau} \widetilde{\zeta}_i^2 - (\sum\limits_{i=1}^{n-\tau} \widetilde{\zeta}_i)^2}}$	$[\pm 1, 0]$	219
22	Auto-correlation difference (**ACD**) for $\boldsymbol{\zeta}, \widetilde{\boldsymbol{\zeta}} \in \mathbb{R}^n$, τ: shift	$M_A(\tau) - \widetilde{M}_A(\tau)$, with $M_A(\tau) = \frac{\sum\limits_{t=0}^{n-\tau} (\boldsymbol{\zeta}(t) - \overline{\boldsymbol{\zeta}})(\boldsymbol{\zeta}(t+\tau) - \overline{\boldsymbol{\zeta}})}{\sum\limits_{t=0}^{n} (\boldsymbol{\zeta}(t) - \overline{\boldsymbol{\zeta}})^2}$	$[0, \pm 1]$	219
23	Dynamic Time Warping (**DTW**) for $\boldsymbol{\zeta}, \widetilde{\boldsymbol{\zeta}} \in \mathbb{R}^n$, τ: shift	DTW does not compare the time series by index, but calculates the min. dist. that can be found by shifting the two signals against each other.	$[0, \infty[$	219

#	Metric	Formula / Description	B-W	Applic.
24	**Jaccard** metric as arithm. mean of intersection over union for areas $A, \widetilde{A}$ as ground-projection of bounding boxes over time series with $i \in 1, ..., n$	$\overline{k_i}, k_i = \begin{cases} 1 - \dfrac{\lvert A_i \cap \widetilde{A}_i \rvert}{\lvert A_i \cup \widetilde{A}_i \rvert} \\ 0 \text{ if } \lvert A_i \cup \widetilde{A}_i \rvert = 0 \end{cases}$	$[0, 1]$	216, 219, 223, 224, 231
25	Optimal Sub-Pattern Assignment (**OSPA**) for two subsets of object positions ($\boldsymbol{\mathcal{Z}}, \widetilde{\boldsymbol{\mathcal{Z}}}$)	Find optimal assignment of all objects from $\boldsymbol{\mathcal{Z}}, \widetilde{\boldsymbol{\mathcal{Z}}}$ with Wasserstein dist., calculate their dist. with cutoff param. c, and calculate p-th order of metric by $(\overline{a^p})^{1/p}$	$[0, c]$	216, 231
26	OSPA for tracks (**OSPA-T**)	Like OSPA plus track continuity and timing by identification error	$[0, c]$	215, 216
27	OSPA for multiple tracks (**OSPA-MT**)	Like OSPA-T, but with tracks as vectors of traject. instead of single-frame states	$[0, c]$	216, 223, 224
28	Generalized OSPA (**GOSPA**)	Like OSPA, but with different calculation of the cardinality error and use of assignment optimization instead of traject. permutation	$[0, c]$	216
29	**Rahmathullah** et al.	Sum of localization errors for properly detected objects and a penalty for missed objects and objects with wrong object ID	$[0, \infty]$	215
30	Kullback–Leibler divergence $\boldsymbol{D}_{\mathbf{KL}}(Ᵽ(\zeta), \widetilde{Ᵽ}(\zeta))$ between simulated and real probability distributions	$\sum\limits_{\zeta \in \{\boldsymbol{\zeta} \cap \widetilde{\boldsymbol{\zeta}}\}} Ᵽ(\zeta) \cdot \log_2(\dfrac{Ᵽ(\zeta)}{\widetilde{Ᵽ}(\zeta)})$	$[0, \infty]$	219, 226 229
31	Jensen-Shannon distance $\boldsymbol{d}_{\mathbf{JS}}(Ᵽ(\zeta), \widetilde{Ᵽ}(\zeta))$ between simulated and real probability distributions	$d_{\text{JS}} = \sqrt{D_{\text{JS}}(Ᵽ(\zeta), \widetilde{Ᵽ}(\zeta))},$ $D_{\text{JS}} = \frac{1}{2} D_{\text{KL}}(Ᵽ(\zeta), \overline{Ᵽ}(\zeta))$ $+\frac{1}{2} D_{\text{KL}}(\widetilde{Ᵽ}(\zeta), \overline{Ᵽ}(\zeta)),$ $\overline{Ᵽ}(\zeta) = \frac{Ᵽ(\zeta)+\widetilde{Ᵽ}(\zeta)}{2}$	$[0, 1]$	233

#	Metric	Formula / Description	B-W	Applic.
32	Area validation metric **(AVM)** $\boldsymbol{d}_{\mathbf{AVM}}$ btw. simulated and real (empirical) CDF $F(\zeta)$ and $\zeta \in \{\boldsymbol{\zeta} \cup \widetilde{\boldsymbol{\zeta}}\}$	$\int\limits_{-\infty}^{\infty} \lvert F(\zeta) - \widetilde{F}(\zeta)\rvert \, \mathrm{d}\zeta$	$[0, \infty]$	219
33	Frequency $\boldsymbol{f}_{\mathbf{KS}}$ of positive Kolmogorov-Smirnov tests of the null hypothesis that two sets $\boldsymbol{\mathcal{Z}}, \widetilde{\boldsymbol{\mathcal{Z}}}$ of $n, m \geq 12$ observations $\boldsymbol{\zeta}, \widetilde{\boldsymbol{\zeta}}$ are from same distribution	$\sup\limits_{\zeta \in \{\boldsymbol{\zeta} \cup \widetilde{\boldsymbol{\zeta}}\}} \lvert F(\zeta) - \widetilde{F}(\zeta)\rvert \geq c_\alpha \sqrt{\frac{n+m}{nm}}$, with $c_\alpha = \sqrt{\frac{\ln 2 - \ln \alpha}{2}}$, and significance level α	$[0, 1]$	227, 229
34	Deep Evaluation Metric **(DEM)**	PointNet++ trained with real and synthetic radar detections to classify them.	$[0, 1]$	232

4.3 Interim Conclusion on the SotA in VV&UQ Methodologies and Metrics

This chapter provides a broad overview about methodologies and incorporated methods for V&V (Viehof, Schaermann, Ngo) and later VV&UQ Riedmaier and Danquah. Afterwards, all metrics applied for validation of APSS simulation are listed. In the course of the chapter, some key factors appear regularly: The ability to consider epistemic and aleatory uncertainty from reference data (model input) and the enabling of model bias and scattering error prediction from just a few validation samples (model form) for the whole application domain. Both topics are described in literature, but have not been performed for APSS simulation yet.

As mentioned in Sec. 4.1.1, Viehof introduces a practical concept for confidence-motivated labeling for sample validity, as shown in Fig. 4-5. However, the implicit assumption of smaller PDFs from simulation compared to measurements is only reasonable in case of simulated single uncertain numbers as e.g. for vehicle consumption modeling. In case of simulated sensor signals over time including its exact scattering behavior, the model should reproduce a similar or the same PDF, which would help the simulation to overlap in the first place. Here, introducing epistemic uncertainties and calculating the overlap from p-boxes could be beneficial.

Schaermann and Ngo both provide methodologies specifically for validation of APSS simulation. However, they focus on metrics and interfaces that could be used for detections, features (OGs) and objects, but also do not consider uncertainties in the data or the necessity of intuitive understanding of the metrics for their usage in specifications.

Riedmaier and Danquah also discuss the methodology for model validation from Viehof. They find his introduction of *"probabilistic simulations into automotive vehicle dynamics is an important contribution."* [172d] However, they miss in Viehofs validity assessment some aspects of experimental comparison described by Oberkampf and Roy[235]. Since they only have the simulation of scattering time signals in mind, they state that in this case *"the simulation PDF should have exactly the same width as the experimental PDF and not a smaller one."* [172d] Again, not considering parametric (epistemic) uncertainties *"under-approximates the true input uncertainties and leads to small simulation PDFs and erroneously to valid model hypothesis, even if the model-form might be inaccurate."* [172d] Also in alignment with the statements in Sec. 4.1.1, Riedmaier, Danquah et al. deny that the binary results from the hypothesis test could be used for aggregation of uncertainties to the application parameter space.[172d]

While presenting a generic framework that can be used for deterministic and non-deterministic simulation validation[236], as a result of the discussion of several VV&UQ methodologies, Riedmaier, Danquah et al. decide for application of PBA in their work on either safety validation of ADS (Riedmaier) or vehicle consumption and vehicle dynamics simulation (Danquah). PBA incorporates a separation of epistemic and aleatory uncertainties and the propagation of uncertainties through simulation before comparing simulated and real measurements. Furthermore, sources of uncertainty (input, model form, and numeric) are treated differently when the uncertainty is aggregated and inter-/extrapolated into the application domain. The uncertainty treatment within the framework from Danquah, Riedmaier et al. is illustrated in Fig. 4-12. Additionally, methods for bias correction are proposed to be included besides uncertainty aggregation, as visualized in Fig. 4-11.

The diversity and sheer amount of metrics applied for APSS validation in Sec. 4.2.2 and the summarizing Tab. 4-3 shows that there is no consensus on how to validate APSS simulation. Some prefer comparison of subsequent interfaces and combine them to an overall score like Ngo et al. or Huch, while others only validate on detection level. Some use strict mathematical metrics, while others use hypothesis testing or the comparison of distributions. However, there seems to be no difference in radar or lidar model evaluation, which leads to the first conclusion that all metrics and methodologies presented can be used for sensor modalities summarized as APSS.

[235] Oberkampf, W. L.; Roy, C. J.: Verification and Validation in Scientific Computing (2010), p. 490.

[236] Riedmaier, S. et al.: Non-deterministic model validation methodology (2021).

5 Further Challenges towards Credible APSS Simulation

There are several challenges that prevent APSS model credibility at the moment, some are discussed in literature, some are introduced in this dissertation. The lack of uncertainty distinction in metrics for APSS simulation VV&UQ has already been identified and the lack of UQ for APSS simulation credibility has already been discussed as well. Before the research questions for this dissertation can be deduced in Chap. 6, further identified challenges are presented in the following. This chapter serves as the bridge between the fundamentals in Chap. 2, the state of the art in the previous Chap. 4, and the innovations presented in the dissertation.

5.1 Lack of Requirements for APSS Simulation

Hirsenkorn presents a list of high-level requirements in his dissertation before presenting his developed sensor simulation method.[237] He states at first that concrete requirements demand a specific use case and as such an intended usage is not present in his case, no concrete values or acceptance criteria are provided. His list of shortly explained criteria reads as:

- *Completeness: All relevant behavior of APSS should be included like signal interactions.*
- *Representativeness: Effects appear with realistic incidence.*
- *Individuality: Every individual situation in driving simulation must be covered.*
- *Scalability: The model should be improvable with more effort.*
- *Robustness: The model should cover incompletely described environments.*
- *Interfaces: Ideally, standardized interfaces are used.*
- *Implementation: The model is implemented independent from the simulation tool.*
- *Computation time and storage: Real time capability depends on the intended usage.*
- *Parameterization: Calibration of parameters like materials, etc.*
- *Traceability: Errors in simulation should be traceable.*
- *Back-traceability: Simulation should be back-traceable to its calibration measurements.*
- *Mounting position: Validity should be ensured for different mounting.*

[237] Hirsenkorn, N.: Modellbildung und Simulation der Fahrzeugumfeldsensorik (2018), pp. 9-11.

- *Independence: Not (yet) existing sensors (under development) should be simulatable.*
- *Raw data: Sensor data should be generated as unprocessed as possible.*

Nevertheless, this list from Hirsenkorn is just a list of criteria and not a list of requirements. Some of the items even contradict each other, like computation time and generation of unprocessed data, while others currently seem unrealistic, like the coverage of every possible situation. Furthermore, validity, credibility, or maturity are missing from the list. In conclusion, it is obviously no concrete specification, as it lacks e.g. acceptance tests and criteria.

Former work of the author already identified and addressed the persistent lack of concrete requirements for APSS simulation to some extent[238] as well. In the mentioned work, features (in the sense of machine learning input parameters) in lidar sensor system data like standard deviation and mean value of detection locations, EPW of detections, as well as dimensions of objects identified by the sensor system were analyzed regarding their feature importance for object classification. This is a method to reduce features for machine learning by neglecting the less important ones and cross-validated. For derivation of important features, mutual information, principal component analysis, and a random forest are trained. As a result, a reduction from 89 features to 20 simple ones only reduced the classification accuracy of a random forest from $94.0\,\%$ to $93.7\,\%$ for the six classes for moving object in the OSI at that time. Therefore, the conclusion is drawn that these simple features are the most relevant to address when modeling the lidar sensor system and furthermore, it is proposed to start with simple sensor models, as only simple features are contained in the list. However, no method for specification or any further steps towards it are provided.

Additionally, as a lack of experience with APSS simulation and metrics for its validation is identified as the major cause for not having requirements for such simulation in contrast to e.g. vehicle dynamics simulation, first benchmarks for lidar detection models with different metrics on subsequent functional layers are derived in another former work.[239] In a subsequent work of the author the lack of requirements is further discussed and a guideline for formulating requirements for sensor system simulation is provided, for the first time.[240] Still, an exemplary list of requirements from a concrete intended function (e.g. an assistance system) is still not included and proposed to elaborate in future work.

A recent publication by the author is again dedicated to the lack of requirements for APSS simulation. It targets for functional requirements and starts with a comprehensive collection of known cause-effect chains and phenomena in APSS, as such lists are the SotA for requirements for simulation. As such lists are clearly no requirements in the strict sense, further methods are provided to derive them from such a collection. At first, the Perception Sensor Collaborative

[238] Holder, M. F. et al.: Data-driven Derivation of Requirements for a Lidar Sensor Model (2018).

[239] Rosenberger, P. et al.: Benchmarking and Functional Decomposition of Lidar Sensor Models (2019).

[240] Rosenberger, P. et al.: Towards a Generally Accepted Validation Methodology for Sensor Models (2019).

Effect and Cause Tree (PerCollECT) method is advised as important step to collaboratively draw tree-shaped graphs of cause-effect chains ending in phenomena and starting from design parameters of the APSS and environmental causes. Then the Cause, Effect, and Phenomenon Relevance Analysis (CEPRA) method is proposed to derive relevance of all cause-effect chains in an failure mode and effects analysis (FMEA)-like manner with scores for their occurrence in the specific sensor system and its impact on a specific function. At the end, and in analogy to the first publication discussed in this section, a methodically derived list of causes, effects, and phenomena is achieved, but still no actual APSS model specification. Nevertheless, there are no further attempts or publications regarding such requirements to the knowledge of the author and the lack is still present.

5.2 Selection of Validation Samples and Experiment Design

Before describing specific challenges for sample selection and experiment design for APSS simulation validation, some guidelines for such experiments from literature are presented at first. As a starting point and to stress the high requirements that exists for validation experiments, the six instructions from Oberkampf and Roy are:[241]

1. *A validation experiment should be jointly designed by experimentalists, model developers, code developers, and code users working closely together throughout the program, from inception to documentation, with complete candor about the strengths and weaknesses of each approach.*
2. *A validation experiment should be designed to capture the essential physics of interest, and measure all relevant physical modeling data, initial and boundary conditions, and system excitation information required by the model.*
3. *A validation experiment should strive to emphasize the inherent synergism that is attainable between computational and experimental approaches.*
4. *Although the experimental design should be developed cooperatively, independence must be maintained in obtaining the computational and experimental system response results.*
5. *Experimental measurements should be made of a hierarchy of system response quantities, for example, from globally integrated quantities to local quantities.*
6. *The experimental design should be constructed to analyze and estimate the components of random (precision) and systematic (bias) experimental uncertainties.*

[241] Oberkampf, W. L.; Roy, C. J.: Verification and Validation in Scientific Computing (2010), pp. 372-373.

Additionally, Saam states that *"validation experiments should not only be distinguished from scientific discovery experiments, model calibration experiments and experiments serving as system performance tests. They should also be designed, executed and analyzed separately."* [242] The guidelines from Oberkampf and Roy stress the high effort and the necessary diligence connected with validation experiments and the statement from Saam highlights the necessary independence of validation data from the data used for calibration and verification.

The model calibration experiments must be documented carefully regarding the samples taken from parameter space, to be able to take other samples for validation, analogue to train and test in machine learning. As one approach to distinguish between experiments, Amersbach addresses the question for sample independence that are taken from the parameter space by his definition in the context of safety validation scenarios: *"Two concrete scenarios are equal if their respective parameter combination is situated in the same volume cell of the common parameter space"* [243]

As already briefly discussed in Sec. 4.1.1, Viehof proposes to perform a diligent sensitivity analysis at first to find the parameters with highest impact on the simulation and applies extended Fourier amplitude sensitivity testing (eFAST) for his exemplary vehicle dynamics simulation validation.[244] Ngo et al. follow his advice and apply Fourier amplitude sensitivity testing (FAST) for radar detection simulation.[245] Besides, Viehof lists three concerns that are addressed with respective questions for finding relevant scenarios[244]:

1. *Relevance of changeable parameters in real experiments → Which parameters should be varied according to the sensitivity analysis and in which range?*
2. *Required degree of statistical validation → How granular should the parameter space be resolved? How many scenarios or configurations should be inspected?*
3. *Practicability → How many scenarios or configurations are feasible to be inspected or performed?*

From own experiment design experience, even more questions are to be asked for sample selection from the immense parameter space. With regard to the required degree of statistical validation, global and regional coverage should be discussed in addition to granularity in parameter space. This includes considering that some cause-effect chains and phenomena are visible in absolute values already in singular experiments, while others are only visible by relative changes in the course of one or several experiments like changing the range to a target object for the change in signal intensity.

For feasibility, in addition to the number of experiments, the number of their repetitions must also be taken into account. With respect to the parameters, there are some cause-effect chains

[242] Saam, N. J.: Validation Benchmarks and Related Metrics (2019), p. 438.

[243] Amersbach, C.; Winner, H.: Defining Required and Feasible Test Coverage (2019), p. 428.

[244] Viehof, M.: Objektive Qualitätsbewertung von Fahrdynamiksimulationen (2018), pp. 65-66

[245] Ngo, A. et al.: A Sensitivity Analysis Approach for Evaluating a Radar Simulation (2020).

in APSS sensors that are static, while others are dynamic. Either case reveals totally different spatial and temporal experiment designs. Obviously, there are different levels of complexity in modeling cause-effect chains and different gradations of effort when designing experiments for testing them, which gives some additional weights on the experiments when designing a complete validation study. Furthermore, the documentation of the validation experiments adds to the overall effort. This includes not only a semantic description and often a video of what is happening, but a searchable and logical cataloging of the measurement and reference data including a temporal synchronization of all measurements.

Hadelli performs a first rough estimate for the parameter space and its coverage with only static lidar sensor experiments.[246] He applies a decomposition on the complete set of parameters and divides it into five groups: Environment, surrounding, lidar sensor, target object, and sensor mounting. In total, he finds around 70 different parameters. and lists all possible values for each parameter. Furthermore, he gives estimates for the accuracy of measuring each of the value and estimates the duration of the complete measurement study where every parameter is varied once (not every point in the parameter space) to about 80 days netto (without preparation, waiting for weather effects, etc.) for a single person. Even if this is not a representative number of studies planned nor a very experienced engineer to plan the study, it gives a first educated guess and an impression to the effort needed for sample validation. Besides the estimation for the effort, Hadelli documents his experiments in detail and produces measurement and reference data that can serve as validation data set for lidar detection simulation. Additionally, there are some important findings on signal processing and especially thresholding that has high influence on real detections and therefore plays a crucial role in his exemplary experiments to a negative extend regarding repeatability of the data and especially on relative changes of signal intensity and EPW.

Riedmaier et al. summarize several approaches for sampling scenarios within parameter ranges, from parameter distributions, from accident data bases, criticality-based, and complexity-based in the field of driving scenarios for safety validation of ADS, which are applied in literature.[247] They further access the listed approaches regarding scenario representativeness, parameter compatibility, corner case identification, coverage, expansion, applicability, etc. As none of the beforehand found approaches solves all criteria and challenges, they advise to combine approaches for better results. Additionally, they state that model validation is not tackled in most research on scenario variation for safety validation and all scenario validation in simulation would have ”no credibility in terms of their use in decision making” without addressing that lack.

[246] Hadelli, A. A.: Messkampagne zur Parametrisierung und Validierung von Lidar-Sensor-Modellen (2020).

[247] Riedmaier, S. et al.: Survey on Scenario-Based Safety Assessment (2020), pp. 87464-87465.

5.3 Limited Repeatability and Reproducibility of APSS Measurement Data

R&R as already defined in Sec. 1.2.5 are generic challenges for experiment design and measurement data collection. Full repeatability of measurements does not mean an ideal and deterministic case, where all measurements constantly provide the ground truth value. In contrast, total reproducibility would mean to gather measurement data with the same distribution and same mean and variance, when the same experimentalist repeats an experiment under identical conditions (however defined). Reproducibility is broader conceived and allows different experimentalists and different laboratories or test tracks and is therefore not limited to (idealized) identical conditions. However, the limited R&R is the reason for some portion of aleatory and mostly epistemic uncertainty in measurement data.

Still, as Oberkampf and Roy already correctly mention, *"control and repeatability of the experiment are less important in a validation experiment than precisely measuring the conditions of an uncontrolled experiment. Variability in the surroundings of a validation experiment, for example due to weather conditions, is not critical, as long as the conditions of the surroundings are precisely measured."* [248a] Nevertheless, as reference data is never perfect, R&R is a major target for experiment design and of special importance for model credibility assessment. E.g. in the case when validation results are questioned, experiments are repeated to investigate aleatory and epistemic uncertainties of measurement and reference data.

Oberkampf and Roy exemplary list influence factors for systematic uncertainties in measurements to be overcome by randomization or blocking, if possible.[248b] All are true challenges especially for APSS measurements and read:

- *Experimental instrumentation*
- *Experimental procedures*
- *Experimental hardware*
- *Facility characteristics*
- *Data recording and reduction*
- *Experimental personnel*
- *Time of experiment*
- *Weather conditions*

[248] Oberkampf, W. L.; Roy, C. J.: Verification and Validation in Scientific Computing (2010). a: pp. 373-374.; b: pp. 441-442.

APSSs are very challenging regarding R&R, as they are prone to so many environmental and surrounding parameters that controllability outside of specially isolated laboratories and radar chambers simply does not exist. Radar is a very prominent example for very limited R&R of measurement data, as already presented in former work of the author.[249] In this work based on joint experiments within the ENABLE-S3 project[250], special focus is given to occlusion and separability of objects in radar detections as well as to the variation of RCS. In consequence of the efforts spend to perform and analyze the measurements, it is stated that *"radar sensor measurements are characterized by their stochastic behavior, which complicates reproducibility"* and that *"such behavior complicates sensor model validation, as the singular-comparison of measurement to simulation results is not meaningful."* [249a]

However, as already mentioned in the previous Sec. 5.2, this challenge in very low R&R of intensities of detections from objects is not limited to radar, but also appears in lidar. Hadelli investigates repeatability of measurements during his intensive exemplary validation study and describes the influence of dynamic thresholding in lidar on the detection's EPW and intensities that destroys R&R and the data's value for model V&V.[251]

APSS like radar and lidar are simply not designed for measurement R&R, as they do not provide calibrated absolute intensities, EPWs, or RCSs. Therefore, only a comparison of such measurands on a relative scale for changing samples is possible, but no absolute sample validation. For single samples, only the order of magnitude of the values can be checked for identity. Consequently, the importance of the test equipment for lidar validation is also explained by Gomes et al. [252]

When describing his carefully prepared validation study, Eder mentions the importance of reproducibility multiple times and stresses that *"the demands on data quality are significant. Even minor deviations during the reproduction of the maneuvers can affect the result. Therefore, an exact reproducibility of individual maneuvers is of great importance for the scenario selection. Moreover, it is important to determine potential limits of tolerance."* [253] As already described in an earlier publication[254], Eder tries to eliminate epistemic uncertainties and the influence of limited R&R by the law of large numbers and repeating experiments 100 times.

In consequence of the challenges explained after the mentioned experiments in the ENABLE-S3 project and from an immensely wealth of experience in conducting experiments for calibration, verification and validation of radar detection simulation, Holder gives a lot of sophisticated recommendations for data acquisition in his dissertation. He states that signal interactions cannot

[249] Holder, M. F. et al.: Measurements revealing Challenges in Radar Sensor Modeling (2018). a: p. 6

[250] AVL List GmbH: ENABLE-S3 Project (2019).

[251] Hadelli, A. A.: Messkampagne zur Parametrisierung und Validierung von Lidar-Sensor-Modellen (2020), pp. 84, 94.

[252] Gomes, T. et al.: Evaluation and Testing Platform for Automotive LiDAR Sensors (2021).

[253] Eder, T.: Simulation of Automotive Radar Point Clouds in Standardized Frameworks (2021).

[254] Eder, T. et al.: Szenarienbasierte Validierung eines hybriden Radarmodells (2020).

be isolated and can be seen in his measurement repetitions carried out with different vehicles when studying the shaking factor in the radar equation. Furthermore, he finds that small variations in the repeated execution of the experiment turn determinism into randomness.[255]

A recent publication from Holder et al., as already mentioned in Sec. 4.1.5, tackles the mostly neglected validation of the reference data measurement equipment. In all other publications on APSS simulation and its V&V listed in this dissertation like the recent Dynamic Ground Truth Sensor Model Validation Approach (DGT-SMV) and re-simulation by Magosi et al.[256], the re-simulation is conducted with taking reference data as perfect (almost GT) sensor data, which is simply not the case. Holder et al. proof that the uncertainties in systems like real time kinematic (RTK) GNSS for object localization and trajectory caption over time are not negligible by applying a so-called super-reference measurement to check the actual reference. Therefore, the challenge of R&R is always present and in the best case, as said by Oberkampf and Roy[248b], the uncertainties are known for validation and simply treated as either aleatory or epistemic uncertainties that are propagated through the simulation, as already described in Chap. 4.

5.4 Selection of Environment and Rendering Simulation Tool / Engine

Besides the explained academically valuable challenges explained in the previous sections, there exist a very practical challenge, namely the selection of the environment and rendering tool/engine when applying APSS simulation. The simulation of the environment and objects around the APSS is obviously influencing its performance. All parts of the simulation explained in Chap. 3 rely more or less on it. It starts with the implementation and parameterization of ray tracing in the rendering engine and e.g. the available techniques for beam super-sampling, as described in Sec. 3.1.1, followed by materials and shapes and the implementation of bidirectional scattering distribution function for their different kinds (Sec. 3.1.3) and the information provided for modeling of signal attenuation (Sec. 3.1.4), ending at the possibilities to model timing effects (Sec. 3.1.5).

While available commercial simulation solutions like IPG CarMaker, dSPACE ASM, Vires VTD, Siemens PreScan, and others provide their own sensor modeling, all have in common to be extendable with custom APSS models to various extent. However, open source simulation solutions exist that cover APSS simulation to a limited extent, and several startup companies are pushing into the growing market of simulation. Besides, the simulation solutions can select from several rendering engines like UNREAL, Unity, etc., which all provide similar but slightly different possibilities to implement ray tracing, materials, geometries, etc. Therefore, the solutions

[255] Holder, M. F.: Synthetic Generation of Radar Sensor Data for Virtual Validation (2021), pp. 153-156.

[256] Magosi, Z. F. et al.: Evaluation of Physical Radar Perception Sensor Models (2022).

are not consolidated and some kind of dynamic is present, which undermines credibility on the one hand, but promises a lot of progress in the near future on the other hand.

Besides fidelity, performance with regard to hardware support and demand for high performance GPUs and central processing units (CPUs) evolves as well. Furthermore, the tools are (slowly) extending their support for different operating systems like Microsoft Windows and Linux. Nevertheless, selecting a tool or engine is still challenging today. There is an initiative to tackle the selection challenge with objective criteria, called SimCert, founded by Dupuis[257]. He states that a function or system developer whose task is to ensure that the system under test (SuT) performs as intended within its ODD might require in-depth knowledge of simulation technology in general and the implementation of individual features in particular to perform an assessment of a simulation solution's fitness for a given task.[257a] Therefore, Dupuis proposes an expert-knowledge-based basic assessment of available solutions along criteria derived from the most common use cases as a short list of candidate-solutions, which a potential user might want to investigate further.[257b] The actual hierarchical list currently encompasses 1063 criteria.[257b] The objective of the initiative is that a user, an expert on the own use case, will not have to be an expert on simulation technology itself.[257c] However, the challenge is not eliminated and many simulation tools still struggle to create sophisticated synthetic data.[257e]

5.5 Standardization of APSS (Simulation) Interfaces

Consequently, after pointing out the challenge of simulation tool and engine selection, standardization of simulation interfaces comes into mind as possible solution or at least support in repeatable simulation results. However, the standards itself in the field of real APSS (e.g. ISO 23150[258] or AUTOSAR[259]) and their simulation (OSI[260]) are brand new and still massively evolving, when e.g. compared to vehicle dynamics simulation.

The interface at which the APSS simulation starts and where the environment simulation ends is crucial and up to discussion. Reflection-based models, like the one presented in Fig. 2-5, provide already some remarkable results, but are limited to some extend in performance (due to the amount of reflections from beam super-sampling that need to be transferred) and limited in fidelity or effect implementation, as ray tracing is not part of the model. Therefore, a more flexible solution for APSS simulation is to be object-based and create an own scene graph for own rendering/ray tracing from the input GT object list. However, this reveals the need for standardization of material descriptions and highlights the need for sophisticated standardization

[257] Dupuis, M.: Paving the way for certified performance (2022). a: p. 110-1.; b: p. 110-2; c: p. 110-3; d: p. 110-4.; e: p. 110-6.

[258] International Organization for Standardization: ISO 23150:2021(E) (2021).

[259] AUTOSAR development cooperation: AUTOSAR (AUTomotive Open System ARchitecture) (2022).

[260] Hanke, T. et al.: Open Simulation Interface (2017).

of describing environmental effects, and many more. Nevertheless, it provides the possibility to validate such models without any dependency to the simulation tool that is coupled to the sensor model when simulation data is computed for model V&V.

Some standards (e.g. ASAM OpenDRIVE) are not only new and dynamically evolving, but also leave room for interpretation through ambiguous definitions.[257d] OSI sees a lot of improvement and refinement of fields mainly revealed in the research projects ENABLE-S3[261], PEGASUS[262], and SET-Level[263], with strong engagement of the author in the project working groups and as an active member of the standards change control board (CCB). This engagement for the simulation standard is accompanied by continuous effort by the author for alignment of OSI with standards for real APSS's interfaces e.g. by participation in the working group (ISO 23150) or by establishing a regular exchange (AUTOSAR).

5.6 Interim Conclusion on Further Challenges

The first interim conclusion of all listed further challenges is that obviously they cannot be tackled all in a single dissertation. Consequently, an excerpt must be selected to address in this work, which is presented in the following Chap. 6. However, the presented challenges serve as a snapshot of the SotA of the whole field of APSS simulation.

Requirements engineering is a lot of effort and experience and carefulness are key for a specification sheet that is accepted by all parties involved. While carefulness mainly means applying agreed methods, this part of the specification is solvable for APSS simulation. Experience is missing in this quite new field of research and application and comes naturally with iterations and repetition over time, while it cannot be solved by this dissertation. While selection of samples from the parameter space could possibly be solved by iteration loops as well, experiment design and R&R are always present challenges and need carefulness during preparation and execution, but no new methods or theory development. Environment simulation tool selection and standardization are both evolving fields each for itself and also taken together. Here, a slightly optimistic prediction can be given as many talented people are involved in further development. However, this dissertation addresses such development and hopefully stresses the importance of credibility assessment at the end that is based on error prediction and uncertainty aggregation in the application portion of the parameter space.

[261] AVL List GmbH: ENABLE-S3 Project (2019).

[262] Deutsches Zentrum für Luft- und Raumfahrt e.V. (DLR): PEGASUS Project (2019).

[263] Deutsches Zentrum für Luft- und Raumfahrt e.V. (DLR): SET Level Project (2022).

6 Research Questions on Specification and VV&UQ of APSS Simulation

After the provided comprehensive overview of the SotA of APSS simulation, the discussed methodologies for its VV&UQ and the extensive collection of applied metrics in Chap. 4, and the explanation of further challenges in Chap. 5, the major research questions to tackle in this dissertation are to be selected in the following.

The interim conclusion of the remaining challenges in Sec. 5.6 already hints that a method for requirements definition is missing. There is simply no validation without requirements, as already depicted in Fig. 4-1. None of the presented publications contains such s specification sheet or an exemplary excerpt of such, as discussed in Sec. 5.1. Therefore, the first choice is to elaborate a method for model specification, as resulting in the first research question:

RQ. 1: How to specify an APSS simulation?

Metrics for VV&UQ of APSS models are selected as main topic for this dissertation. Tab. 4-3 provides an overview of the diverse and wide range of metrics used for APSS V&V. However, the major goal in this work is to provide methods and tools towards credibility in APSS models and their usage in simulation. Overall, such simulation should serve as a qualified and certified tool e.g. for safety validation of ADS, yielding to the highest possible requirements for the qualification of the tool (see Chap. 1), which consists of the VV&UQ of the simulation. The research question that follows from advancing from sample V&V (testing) to VV&UQ (qualification) therefore is:

RQ. 2: Which metrics are useful for sample validation of APSS simulation to enable model error quantification incl. uncertainty aggregation for credible simulation application?

Nonetheless, it can be estimated that only selecting a metric from a long list of already applied and evaluated examples is not enough. The high requirements for qualification of simulation as a tool for safety validation make a comprehensive discussion of the best metric option inevitable and some kind of further development and adaption to its usage in the field of APSS simulation qualification with the limited R&R (Sec. 5.3) is essential. This results in the last of the three major research questions:

RQ. 3: How to apply or further develop the selected metric(s) for VV&UQ with limited R&R of real APSS measurements?

7 Specification of APSS Simulation

The demand for methods to systematically derive requirements has already been identified and the SotA is part of Sec. 5.1. This chapter starts tackling the previously identified research questions by addressing RQ. 1, the lack of requirements for APSS simulation.

As briefly introduced in Sec. 5.1, the author already proposed a method for requirements definition in former work[264]. In detail, it reads as follows:

1. *At first, the SuT with inputs defines the outputs of the sensor system simulation.*
2. *Having these, the requirements engineer needs to have a catalog of possible phenomena that can be observed on real measurement data at the selected sensor system outputs.*
3. *Now, the SuT has to be analyzed in detail and its sensitivity with respect to the listed possible phenomena needs to be determined. As an example, object tracking and classification is mostly insensitive to noise on point clouds, but highly sensitive to simple features like length and width of L-shapes.*[265]
4. *The next step is to define whether a stochastic or physical approach should be used to describe the selected phenomena in particular. Here, the required fidelity and accuracy of the sensor data generation should be considered.*
5. *Finally, the actual accuracies of the different effects or phenomena should be determined.*

Even if these instructions already exist for quite a long time, no other publication has addressed the items in this order to gather a specific list of requirements, so far. The underlying chapter now tackles them step-by-step. It starts at first with simulation model interfaces and the necessity of a modular simulation framework, to secondly provide a method for collection and ordering of cause-effect chains including first comprehensive results. Third, a method for relevance analysis of the identified cause-effect chains is explained and finally a first exemplary list of specific requirements and acceptance tests and criteria is presented.

While the completeness of the methods is a novelty in this case, some of the items have been tackled in previous work of the author and are therefore marked as such. The exemplary results of the requirements and acceptance criteria is a product of a research work group within the VVM project[266], mainly together with consortium partners from Valeo, TÜV Süd, ZF, and dSPACE, but with the methods introduced by the author of this dissertation.

[264] Rosenberger, P. et al.: Towards a Generally Accepted Validation Methodology for Sensor Models (2019), p. 5.

[265] Holder, M. F. et al.: Data-driven Derivation of Requirements for a Lidar Sensor Model (2018).

[266] European Center for Information and Communication Technologies – EICT GmbH: VVM Project (2022).

7.1 Modular Framework as Basis for Specification, Development, and Validation

A popular approach to manage complex systems is to decompose them into subsystems and components that can be handled.[267] Therefore, modular simulation frameworks are popular, as e.g. the simulation by Danquat et al.[268], the approach for sensor error models by Hanke[269], or the highly-scalable sensor modeling approach by Thieling and Roßmann[270]. While it is popular in general, existing APSS approaches mainly address single interfaces like radar or lidar detections, but no subsequent interfaces. In other words, these modules do not reflect the functional blocks of real APSS as they are described in Sec. 2.1.1 and depicted in Fig. 2-1 and Fig. 2-2, but only sub-modules of only one or two of these blocks, like emission, reception, and signal processing. However, multiple validation approaches are based on the evaluation of subsequent interfaces of real APSS, e.g. Schaermann (Sec. 4.1.2) and Ngo et al. (Sec. 4.1.3). Additionally, earlier publications like the one from Hirsekorn et al.[271] demand that simulation should address multiple interfaces at best with standardized interfaces.

Accordingly, a modular simulation framework has been developed and presented in earlier work of the author at first for lidar sensor system simulation[272] and later as a general framework.[273] Functional decomposition of real APSS as briefly discussed in Sec. 2.1.1 and explained in earlier work of the author[274] is the basis for the modules. The interfaces are the result of different working groups in research projects like ENABLE-S3[275] and PEGASUS[276]. The usability of the framework has already been demonstrated in these research projects together with industry partners. It is designed to work in combination with multiple environment simulation tools and to test the full range of perception and sensor fusion functions from different parties. Co-simulation standards are met by implemented standardized interfaces like OSI and its sensor model packaging with Functional Mock-up Interface (FMI). Distributed simulations over e.g. TCP/IP connection are possible and ensured by the FMI standard.[277] The connection between the simulation tool and the framework is sketched in Fig. 7-1 from former work of the author.[272] Finally, the author has published the framework open source on the platforms GitLab and GitHub with FZD.[278]

[267] Liu, F.; Yang, M.: The Management of Simulation Validation (2019).

[268] Danquah, B. et al.: Modular, Open Source Simulation Approach (2019).

[269] Hanke, T.: Simulated Environmental Perception for Automated Driving Systems (2020), pp. 36-42.

[270] Thieling, J.; Rosmann, J.: Highly-Scalable and Generalized Sensor Structures (2018).

[271] Hirsenkorn, N.: Modellbildung und Simulation der Fahrzeugumfeldsensorik (2018), pp. 9-11.

[272] Rosenberger, P. et al.: Sequential lidar sensor system simulation (2020). a: p. 191.

[273] Linnhoff, C. et al.: Highly Parameterizable Perception Sensor Model Architecture (2021).

[274] Rosenberger, P. et al.: Functional Decomposition of Lidar Sensor Systems (2020).

[275] AVL List GmbH: ENABLE-S3 Project (2019).

[276] Deutsches Zentrum für Luft- und Raumfahrt e.V. (DLR): PEGASUS Project (2019).

[277] Modelica Association: Functional Mock-up Interface (2022).

[278] Rosenberger, P. et al.: Modular OSMP Framework (2022).

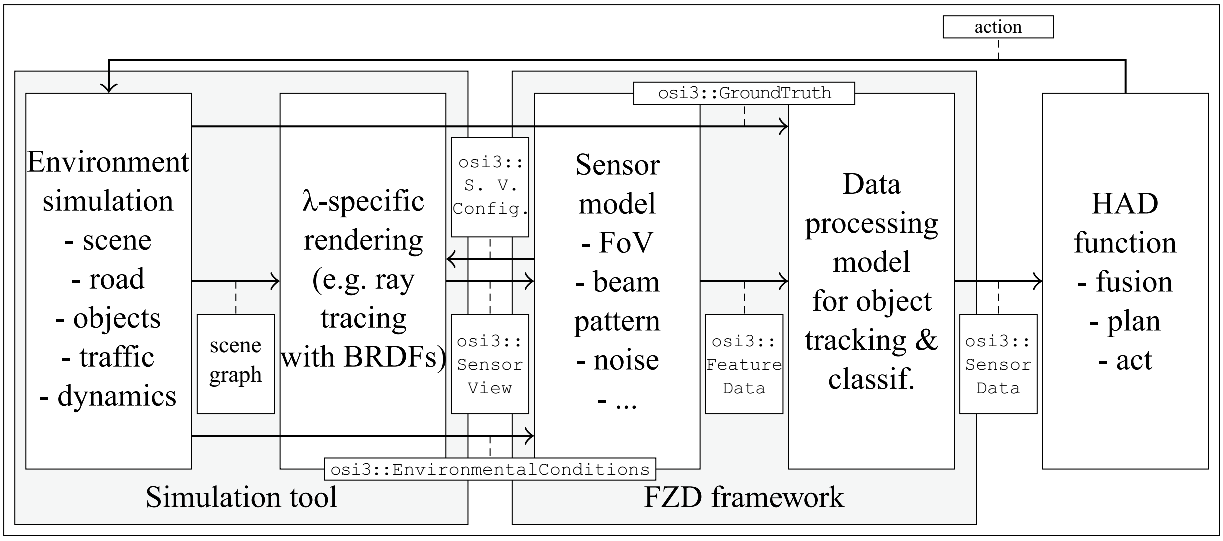


Figure 7-1: Interfaces of modular FZD simulation framework on the example of a reflection-based lidar object model in closed loop simulation from former work of the author[272a]

As shown in Fig. 7-1, the framework ensures the standard conformity to OSI and provides all necessary functions to ensure its packaging according to the OSI sensor model packaging specification[279] and the supported FMI standard. This includes the `SensorViewConfigurationRequest` mechanism, which happens as initialization of the co-simulation, where the sensor model is virtually mounted on the ego vehicle and both sensor model and simulation tool are parameterized. In case of the depicted reflection-based sensor model, e.g. the ray tracing and the super-sampling of rays (referring to Sec. 3.1.1) within the simulation tool is specified by the sensor model to serve its needs.

The actual sensor system simulation logic is contained in so-called strategies or strategy modules. The framework then takes the `SensorView` as input, wraps it into a more generic`SensorData` object and passes this to the first strategy. The strategies always take `SensorData` and output the same structure while manipulating information contained in the data fields like moving object positions and/or adding `FeatureData` with lidar detections, etc. Consequently, a singular data stream is implemented with subsequent strategies and the framework ensures the exchange of `SensorData` from one to another. At the end, the framework outputs the product of the last strategy.

In case of the reflection-based lidar object model, as explained in Chap. 2 and shown in action in Fig. 2-5, several such functional strategies are implemented, as drawn in Fig. 7-2 from its open source repository provided by the author with FZD on GitLab and GitHub[280]. The modules are detection sensing, lidar sensor fusion, segmentation, and tracking. They can be sorted to a real lidar system's front end(s), a possible fusion of data from several lidar front ends including a coordinate transformation, and detection clustering, segmentation, and tracking. Additional

[279] ASAM e.V.: ASAM OSI® (Open Simulation Interface) - Official Documentation (2022).

[280] Rosenberger, P. et al.: Reflection Based Lidar Object Model (2022).

debugging and output strategies are provided open source too and can be plugged in easily at the end of the functional strategies or in between. The figure shows several interfaces of real APSS and which strategy provides such data at which possible output interface.

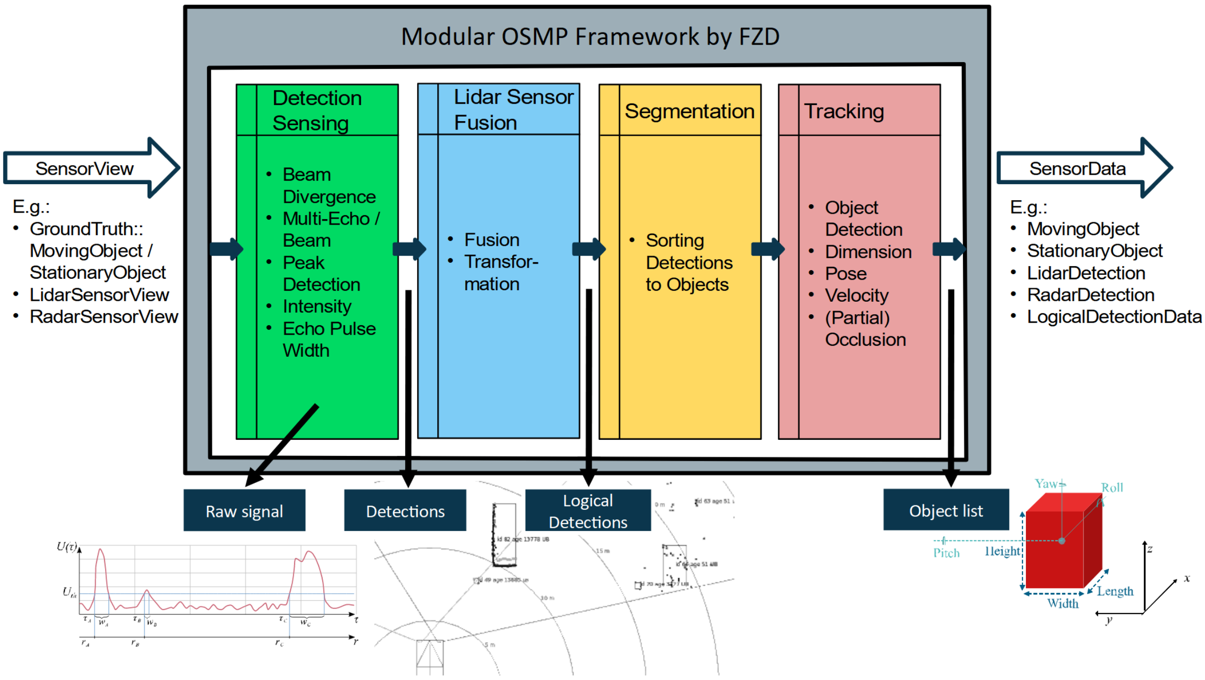


Figure 7-2: Strategy modules of the reflection-based lidar object model within FZD simulation framework[280]

Together with the framework comes a list of parameters of the APSS simulation in form of a profile structure and every strategy can bring further parameters to that profile structure. The profiles containing the values of the parameters are specific for different sensors in real world (e.g. Valeo SCALA, Fig. 3-4, or Velodyne Puck, Fig. 3-5). Such profiles contain values for e.g. mounting position and orientation, number of vertical/horizontal beams, number of super-sampling rays per beam, scanning frequency, and many more. In a FMI conform co-simulation, the simulation master can select and change profiles as a whole to set the APSS model as a different kind of sensor for easy exchange in simulation. Furthermore, such fixed parameter sets enable traceability of validity once the model is used by third parties.

However, other approaches exist that are totally granular like the configurable sensor model architecture by Schmidt et al. [281] It has a similar logical structure as the FZD framework with parameter sets instead of profiles and sequential manipulation of OSI `SensorData`. For every single effect, one module is introduced, instead of one strategy per functional layer. While providing clear instructions how to build up the overall model, the atomic structure neglects interconnections and correlations between effects and the needed isolation of effects for validation of each module is questionable. Therefore, the presented structure provides a good compromise between granularity and testability, while reflecting the real APSS interfaces as major advantage for V&V.

[281] Schmidt, S. et al.: Configurable Sensor Model Architecture (2021).

7.2 Perception Sensor Collaborative Effect and Cause Tree (PerCollECT)

As already briefly introduced in Sec. 5.1, a recent publication of the author[282] gives a method for solving the second task of the steps towards requirements: The catalog of cause-effect chains to choose from for modeling. The method is called Perception Sensor Collaborative Effect and Cause Tree (PerCollECT) and involves the collection of cause-effect chains and their sorting into a tree-shaped graph to visualize all interconnections between effects. The first contribution of the publication is the sorting of an immense list of around 70 effects in lidar, radar and camera sensor systems into a table[282a], already sorted by the generic functional layers, as described in Sec. 2.1.1. The existence of every effect is proven with one or multiple literature sources. Simply collecting effects and modeling them directly, is useful and produces respectable results[283], but it is not very systematic, so the relationships between effects must be explored in order to model them systematically.

Consequently, the listing table is taken as the basis for an ontology based on the functional layers. Every effect is put into a node and then the nodes are sorted to the respective layer. The effects are then connected to causing effects and caused effects, forming the cause effect chains. Finally, top-level effects are becoming phenomena, as they are visible on the output of the overall APSS and bottom-level effects are becoming environmental causes (marked green in Fig. 7-3) and design parameters (marked light blue in Fig. 7-3), depending on whether they are in the environment or can be caused by changes to the APSS. In this manner, a tree-shaped graph is formed that reflects a top-down structure from phenomena to causes.

The results are available open source on GitHub[284] and accompanied by an interactive visualization, where interconnections from and to every node can be highlighted and a short description and sources describing the existence of the effect and the cause-effect chains are provided. Fig. 7-3 shows an excerpt of PerCollECT for lidar, called ”LidarLimbs”.

Hinsemann provided the first but comprehensive collection of cause-effect chains in lidar sensor systems[285] and completed the first version of PerCollECT together with his Bachelor thesis supervisors from FZD including the author. Fig. 7-4 shows the massive amount of causes and effects and the bunch of interconnections of this first full scale of PerCollECT - LidarLimbs and the functional layers that are connected and selectable for the analysis of effects and phenomena on subsequent functional layers. Fig. 7-5 shows an exemplary zoom on a single effect-node of

[282] Linnhoff, C. et al.: Towards Serious Sensor Simulation for Safety Validation of Automated Driving (2021). a: p. 2690.; b: p. 2692.

[283] Holder, M. F. et al.: Modeling and Simulation of Radar Sensor Artifacts (2019).

[284] Linnhoff, C. et al.: PerCollECT - LidarLimbs (2022).

[285] Hinsemann, T.: Analyse von Effekten in Lidardaten für die virtuelle Absicherung (2021).

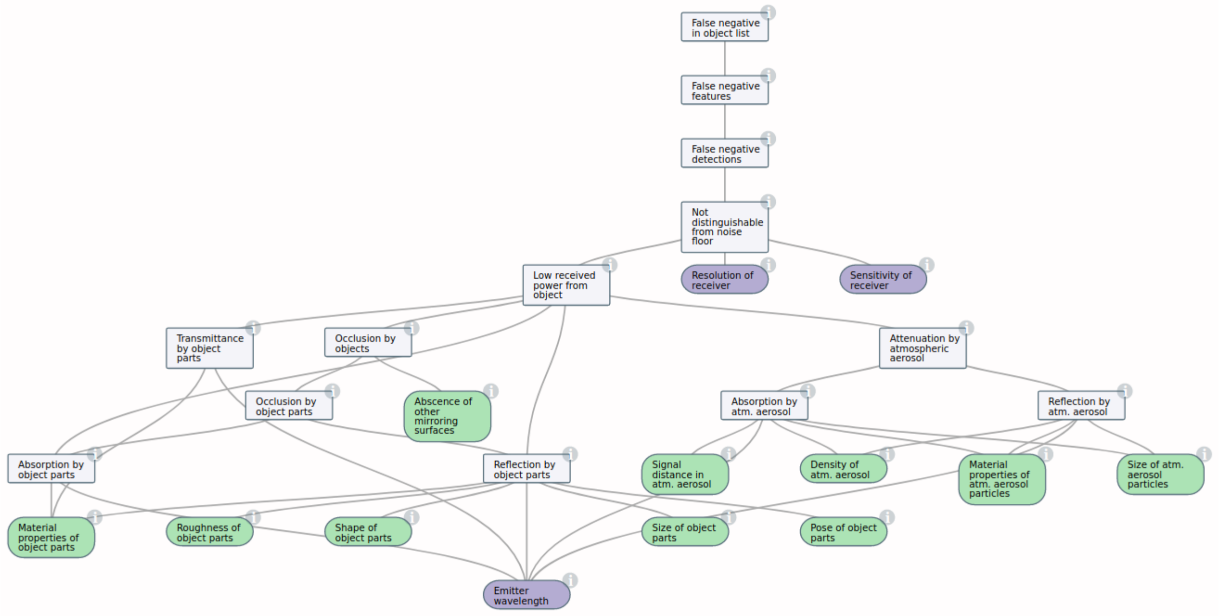

False negative in object list
False negative features
False negative detections
Not distinguishable from noise floor
Low received power from object
Resolution of receiver
Sensitivity of receiver
Transmittance by object parts
Occlusion by objects
Attenuation by atmospheric aerosol
Occlusion by object parts
Absence of other mirroring surfaces
Absorption by atm. aerosol
Reflection by atm. aerosol
Absorption by object parts
Reflection by object parts
Signal distance in atm. aerosol
Density of atm. aerosol
Material properties of atm. aerosol particles
Size of atm. aerosol particles
Material properties of object parts
Roughness of object parts
Shape of object parts
Size of object parts
Pose of object parts
Emitter wavelength

Figure 7-3: Exemplary excerpt of Perception Sensor Collaborative Effect and Cause Tree (PerCollECT) - LidarLimbs from former work of the author[284]

PerCollECT - LidarLimbs, the “low received power from object” and the highlighted interconnections (red) coming from other effects, environmental causes, or design parameters of the real sensor system, as well as the output-connections of the node to subsequent effect-nodes (blue).

As the term collaborative in the acronym of PerCollECT suggests, the method is targeted to be constantly developed further by the community through collaborative introduction of new nodes and intersections. However, the first impulse and collection of effects is provided and already successfully used in projects like VVM[286]. PerCollECT is *“one-way coupling”* [291a] and as generic as necessary to reflect all kinds of sensors per sensor technology, like scanning and solid-state (see Sec. 3.1.5) in case of lidar, and use case independent. It helps identifying inter-dependencies between effects and is therefore used when scenarios are designed, as demonstrated in a recent publication of the author.[287] Furthermore, a relevance analysis of cause-effect chains can be performed to derive requirements and decent acceptance tests, as will be presented in the next section.

[286] European Center for Information and Communication Technologies – EICT GmbH: VVM Project (2022).

[287] Elster, L. et al.: Fundamental Design Criteria for Logical Scenarios (2021).

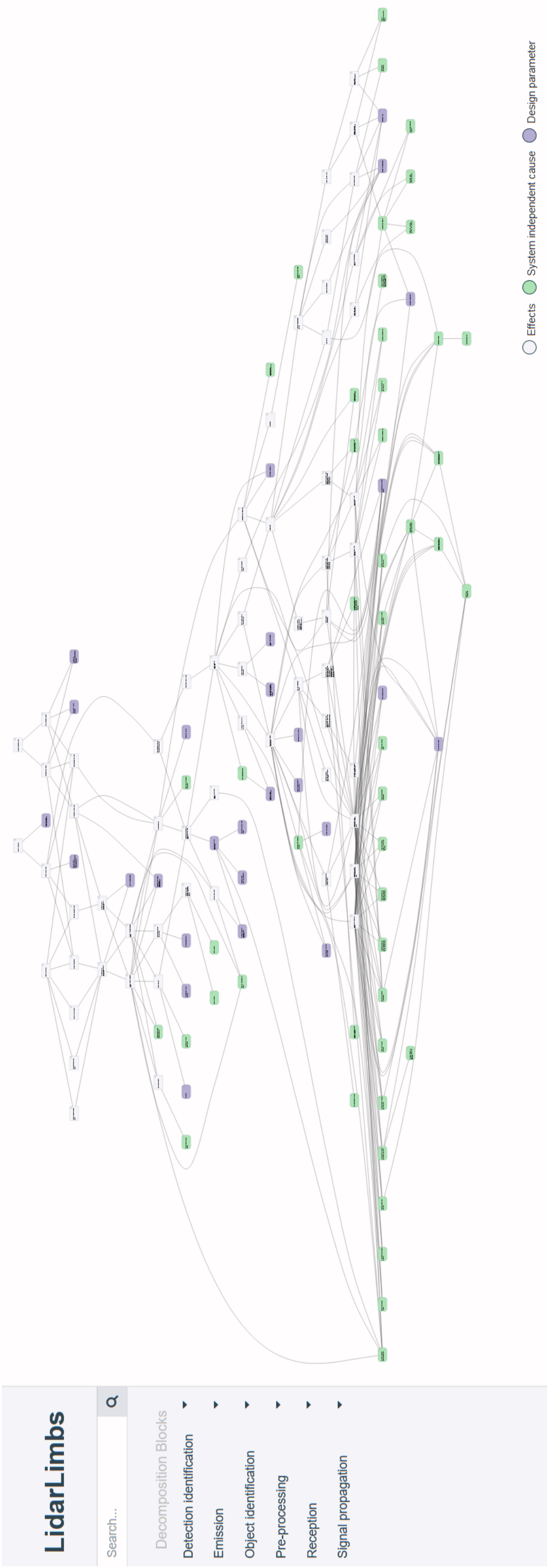
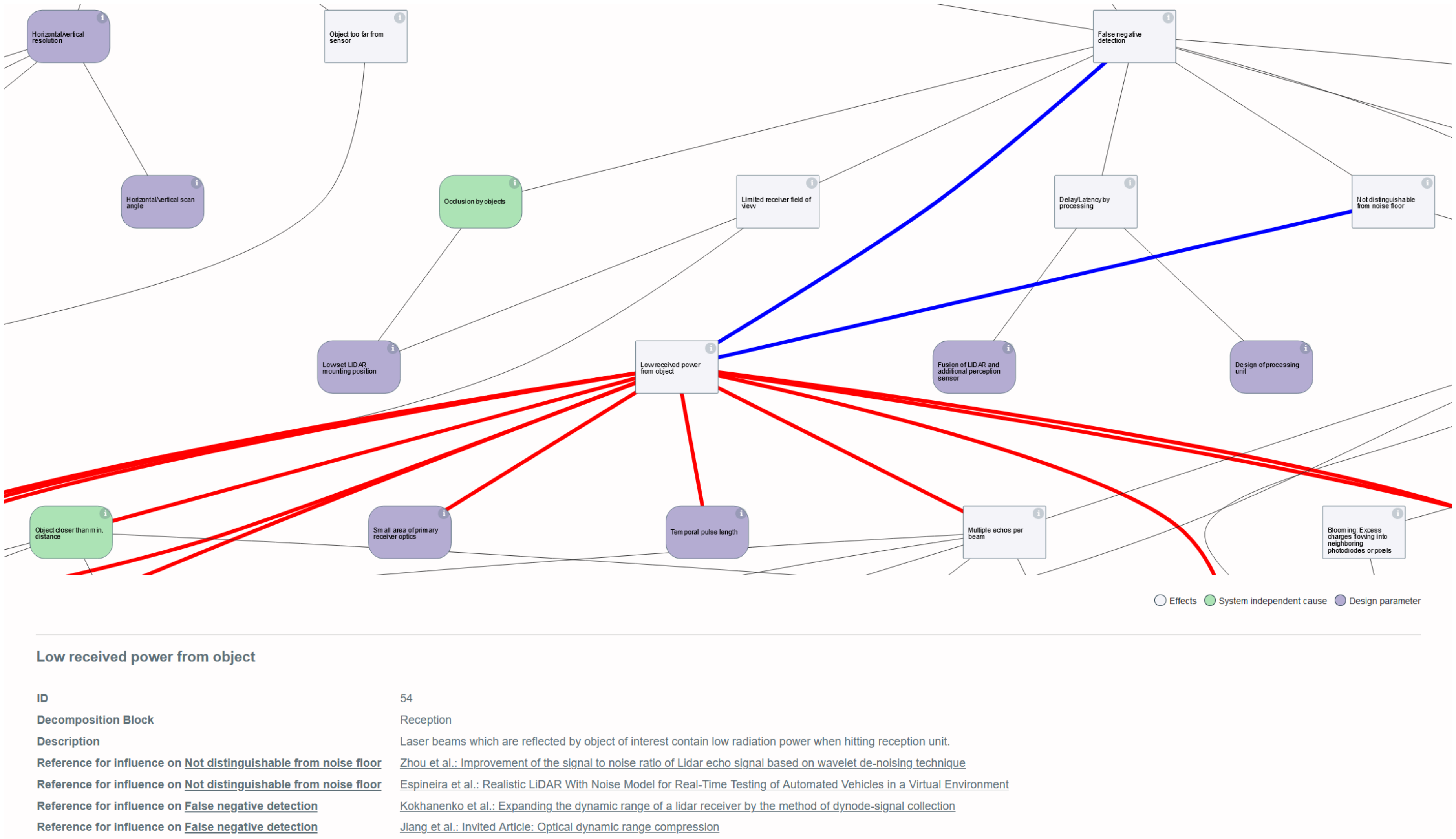


Figure 7-4: Full extent of PerCollECT - LidarLimbs, screenshot from its homepage[284]


Horizontal/vertical resolution
Object too far from sensor
False negative detection
Horizontal/vertical scan angle
Occlusion by objects
Limited receiver field of view
Delay/Latency by processing
Not distinguishable from noise floor
Lowset LIDAR mounting position
Low received power from object
Fusion of LIDAR and additional perception sensor
Design of processing unit
Object closer than min. distance
Small area of primary receiver optics
Temporal pulse length
Multiple echos per beam
Blooming: Excess charges flowing into neighboring photodiodes or pixels
○ Effects
● System independent cause
● Design parameter

Low received power from object

ID	54
Decomposition Block	Reception
Description	Laser beams which are reflected by object of interest contain low radiation power when hitting reception unit.
Reference for influence on Not distinguishable from noise floor	Zhou et al.: Improvement of the signal to noise ratio of Lidar echo signal based on wavelet de-noising technique
Reference for influence on Not distinguishable from noise floor	Espineira et al.: Realistic LiDAR With Noise Model for Real-Time Testing of Automated Vehicles in a Virtual Environment
Reference for influence on False negative detection	Kokhanenko et al.: Expanding the dynamic range of a lidar receiver by the method of dynode-signal collection
Reference for influence on False negative detection	Jiang et al.: Invited Article: Optical dynamic range compression

Figure 7-5: Zoom into one effect node and highlighted cause-effect chains of PerCollECT - LidarLimbs, screenshot from its homepage[284]

7.3 Cause, Effect, and Phenomenon Relevance Analysis (CEPRA)

As already shortly presented in Sec. 5.1, a method for relevance analysis of cause-effect chains is proposed in former work of the author[282b]. It is dedicated to any test engineer (team) that needs to specify an APSS simulation for development or safety validation. The prerequisites are that the ODD of the SuT is already clarified and PerCollECT is fully available for the sensor to be replicated in simulation as test tool within the test-suite for the SuT. It consequently addresses the third step of the instructions from the beginning of this chapter, the systematic, wise foreseeing analysis of the possible impacts of phenomena on the SuT. However, it extends this instruction to an analysis of the impact for the given ODD and an accompanying assessment of the occurrence of cause-effect chains up to different phenomena in the selected APSS to be modeled.

PerCollECT is built top-down, from the top phenomena over cause-effect chains down to the causes and design parameters, whereas CEPRA is flexible and can be performed either top-down or bottom-up. Anyways, all possible cause-effect chains for the selected APSS are to be listed in the first three left columns of CEPRA. Tab. 7-1 shows an exemplary provided excerpt of a CEPRA for a lidar sensor system simulation. Three rows are given for the phenomenon "false negative in object list", one for occluding objects, one for low received power from reflecting objects, and one for attenuation of the signal by an absorbing atmospheric aerosol. The first two reflect different ways through PerCollECT from the same causes to the same phenomenon, while the third one has different causes. Consequently, the first half of the table is filled from PerCollECT, which could be performed (semi) automatic from the provided Extensible Markup Language files on GitHub.

Table 7-1: Exemplary Cause, Effect, and Phenomenon Relevance Analysis (CEPRA) for APSS Modeling from former work of the author[282b]

Pheno-menon (P)	**Effect chain (EC) of phenomenon**	**Causes of effect chains**	**P&EC occurrence in ODD*** (*O*, filled by sensor expert)		**P&EC impact on SUT in ODD*** (*I*, filled by SUT expert)		**Relevance of P&EC**
			$[1,10]$	Rationale	$[1,10]$	Rationale	$O+I$
False negative in object list	→ FN features → FN detections → Not dist. from noise floor → Low rec. power from object → Occlusion by objects → Occlusion by object parts → Reflection by object parts	• Materials of reflect. obj. parts • Roughness of reflect. obj. parts • Shapes of reflect. obj. parts • Size of reflect. obj. parts • Emitter wavel.**	*4*	*FN objects caused by occluding reflecting objects occurs rarely in a front radar on a highway, because of multi-path propagation.*	*6*	*FN obj. occurring because of occlusion in a front radar have a moderate impact because mainly only direct neighbor objs. considered.*	***10***
	→ FN features → FN detections → Not dist. from noise floor → Low rec. power from object → Reflection by object parts		*2*	*FN objects caused by compl. away-reflecting obj. cannot be ruled out, but are not expected on highway.*	*9*	*FN objects occurring in a front radar have a very high impact on a highway pilot.*	***11***
	→ FN features → FN detections → Not dist. from noise floor → Low rec. power from object → Attenuation by atm. aerosol → Absorption by atm. aerosol	• Signal dist. in atm. aerosol • ... • Emitter wavel.**	*3*	*FN objects caused by completely absorbing atmospheric aerosol occur only in harsh weather in a front radar on a highway.*	*5*	*FN objects occuring in harsh weather conditions may be covered by safety concept with a moderate impact on the highway pilot.*	***8***
	• • •						
• • •							

Legend: Normal font: Automatically generated content from PerCollECT after sensor output definition; *Italic: Expert knowledge needed*

*Operational Design Domain (ODD) must be defined beforehand (here: a German highway with all its elements for a highway pilot as SUT).

**These causes are design parameters by the SUT (here: a highway pilot's radar at the front center) and must be defined beforehand.

At this point, the relevance analysis for all cause-effect chains starts. Two experts or expert groups need to be consulted by the mentioned test engineer. The first information source is needed to assess the occurrence of possible phenomena and cause-effect chains in the given ODD and the APSS to be simulated. The occurrence scores O in CEPRA have the same criteria as the frequency scores in a regular and standardized FMEA.[288a] The frequency scores are provided in Tab. A-2 in Annex A. They are selected for CEPRA, not only because they fit quite well, but also they are well established and promoted by Automotive Industry Action Group (AIAG) and Verband der Automobilindustrie (VDA). The score will be used to calculate the relevance score at the end, but the rationale is especially important for documentation of the whole requirement definition process and should be given carefully.

The same arguments hold for the impact scores I from the second expert (group) to be consulted, the SuT expert(s). The scores are in line with the severity scores from FMEA Handbook by AIAG and VDA.[288b] The severity scores are provided in Tab. A-1 in Annex A. Impact I means the expected impact on the SuT, which is not as negative in its meaning as the mentioned severity, but the criteria for the scores read very similar to the classical FMEA.

The overall score is the sum of impact and occurrence for each cause-effect chain. Using a sum for the overall score calculation instead of e.g. multiplication does not immediately eliminate cause-effect chains with extremely low or high individual scores for I or O. With respect to the usage of the simulation for safety validation, this is necessary to keep very rarely occurring phenomena with medium to high impact. This result is the basis of decision-making for the overall model specification.

Such expert-knowledge-based assessment of occurrence and impact is obviously subjective and requires a high level of oversight over all factors (ODD, sensor, SuT, etc.) However, FMEA like methods are always exposed to this uncertainty, but well established and professional teams can be trusted to fulfill such extensive challenges. Independent control instances can qualify the teams and can oversee the process to increase the overall credibility of the results. It should be clarified that the final relevance scores do not count as absolute values, but as relative scores against all other cause-effect chains listed for the given modeling-task defined by the given ODD, sensor, and SuT. In this regard, it is very well suited to select the most relevant phenomena and cause-effect chains from the immense number of possible tasks to model depending on the given monetary budget, test capacity, computer hardware, time budget, etc.

As a proof of concept, CEPRA is already successfully applied in the German publicly funded research project VVM[289] together with project-internal working group partners from the industry (simulation tool vendors, OEMs, TIER1s, testing organizations, etc.). In this project, it is derived from PerCollECT - LidarLimbs and used for specification of a simulation of the Valeo SCALA lidar sensor system, as illustrated in Fig. 3-4. The project has defined several functional use

[288] AIAG; VDA: FMEA Handbook - Failure Mode and Effects Analysis (2019). a: pp. 123-124.; b: p. 122.

[289] European Center for Information and Communication Technologies – EICT GmbH: VVM Project (2022).

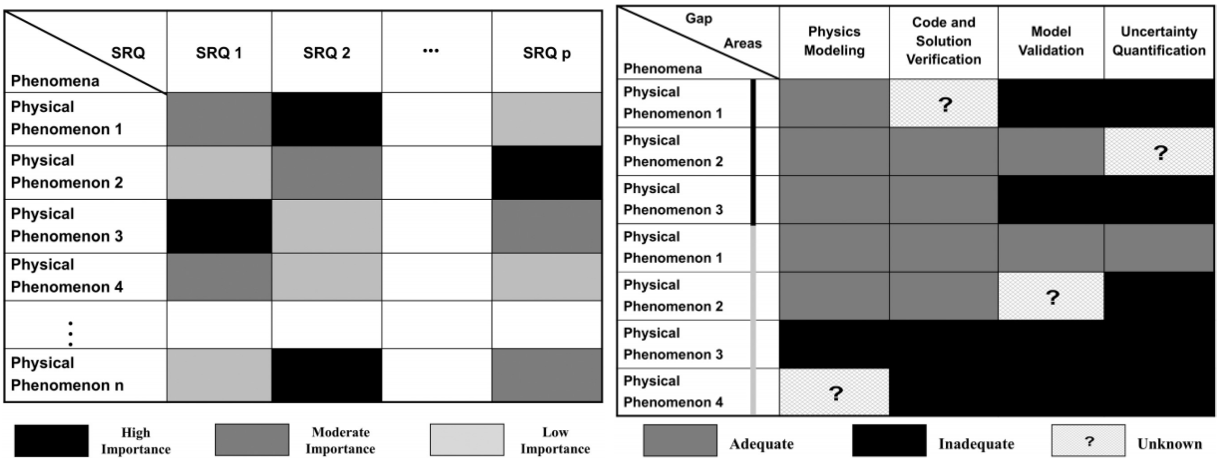
cases, where one is an occlusion scenario at an urban intersection with another vehicle and a crossing vulnerable road user. The ODD including possible weather influence and all possibly involved objects is defined to a certain extent as well. While the sensor expert column is filled out together with the sensor manufacturer, the SuT column is filled with the provider of the fusion and planning function that demands an object list as input. The results have already been shown by the author in a public presentation.[290] An exemplary excerpt of this CEPRA from the project is kindly provided and attached to this dissertation in Tab. A-3 in Annex A.

The novel method CEPRA is specifically designed for APSS simulation specification. However, there is one method in the field of simulation and its validation that is comparable, called Phenomena Identification and Ranking Table (PIRT)[291]. Oberkampf and Roy shortly present its historical development and state that it is a *"much more powerful tool than originally conceived when it was invented in the late 80s"* [291b]. As PIRT is an expert knowledge-based method like CEPRA, they stress the importance of the assembly of the team that performs the action. Besides definition of objectives and specification environments and scenarios, which is very similar to the definition of an ODD in the case of safety validation of ADS, the *"identification of plausible physical phenomena"* [291c] is a major step before the actual PIRT is constructed. This is very similar to the proposed steps in this dissertation and the proposed efforts resulting in PerCollECT. Oberkampf and Roy even propose a tree-shaped graph to derive relevant SRQs, called environment-scenario-SRQ tree. However, PerCollECT is totally generic and independent from concrete or logical scenarios, while Oberkampf and Roy try to solve two steps in one with their analysis. Furthermore, the method lacks the literature proof that is fundamental for PerCollECT.

The core of PIRT is a ranking table for phenomena that is very similar to CEPRA at first sight, as the rows of the table are formed by the physical phenomena. Nonetheless, the columns have a totally different meaning, making it a fundamentally different tool. There are two alternatives given from Oberkampf and Roy, which are displayed in Fig. 7-6 left and right. In its first version, the columns are the different SRQs and the importance of each SRQ per phenomenon is gray scaled in three levels. The second version is more or less a process diagram, which has the different steps towards model credibility as columns. They are formed in this case by modeling, verification, validation, and UQ and the fulfillment levels are adequate, inadequate, and unknown. While both tables help with different tasks during the overall process, they are not designed for requirements engineering, but for development. Therefore, no documentation is demanded inside the PIRT, in contrast to the rationale columns in CEPRA. However, they show that tables help structuring tasks a lot, as is the case for CEPRA.

[290] Rosenberger, P. et al.: Validation of Test Infrastructure (2022).

[291] Oberkampf, W. L.; Roy, C. J.: Verification and Validation in Scientific Computing (2010). a: p. 705.; b: p. 678.; c: p. 679.; d: p. 683.; e: p. 688.



SRQ / Phenomena	SRQ 1	SRQ 2	···	SRQ p
Physical Phenomenon 1	▩	■		░
Physical Phenomenon 2	░	▩		■
Physical Phenomenon 3	■	░		▩
Physical Phenomenon 4	▩	░		░
⋮				
Physical Phenomenon n	░	■		▩

■ High Importance ▩ Moderate Importance ░ Low Importance

Gap Areas / Phenomena	Physics Modeling	Code and Solution Verification	Model Validation	Uncertainty Quantification
Physical Phenomenon 1	▩	?	■	■
Physical Phenomenon 2	▩	▩	▩	?
Physical Phenomenon 3	▩	▩	■	■
Physical Phenomenon 1	▩	▩	▩	▩
Physical Phenomenon 2	▩	▩	?	■
Physical Phenomenon 3	■	■	■	■
Physical Phenomenon 4	?	■	■	■

▩ Adequate ■ Inadequate ? Unknown

Figure 7-6: Phenomena Identification and Ranking Table (PIRT) from Oberkampf and Roy.[291 d,e]
Reproduced with permission through PLSclear.

7.4 Graduated Definition of Requirements for Successive Interfaces

As defined in the fourth step of the instructions for model specification from former work of the author at the beginning of this chapter, the next step after having the prioritized cause-effect chains from CEPRA at hand is to define the modeling approach and its actual requirements. It includes to define the overall required fidelity and accuracy of the sensor data generation. Finally, the actual accuracies of the different cause-effect chains including acceptance tests is determined, completing the already given instructions. In specification, functional decomposition and the functional layers from Fig. 2-1 are needed, again. Besides specifying the overall APSS simulation, both intermediate interfaces, detection and object level, are specified as well. Such a graduated specification is best suited for complex systems like APSS and is a common approach.

As already described in Sec. 4.1.1, Viehof separates requirements in three priority levels as well:[292]

1. Global key requirements from intended use of the simulation
2. Subsystem requirements from functional decomposition for individual model parts
3. Statistical requirements on parameterization effort, sample selection, and parameter space coverage

There is no specification without intended usage of the APSS simulation. For this reason, the functional use case from the VVM project[293] is taken as working example, again. In the project, the mentioned working group defined requirements step wise, starting from general requirements for the lidar sensor system simulation, over the lidar object model requirements, down to the

[292] Viehof, M.: Objektive Qualitätsbewertung von Fahrdynamiksimulationen (2018), pp. 51-52.

[293] European Center for Information and Communication Technologies – EICT GmbH: VVM Project (2022).

lidar detection model requirements, as illustrated in Fig. 7-7 and already publicly presented by the author.[294]

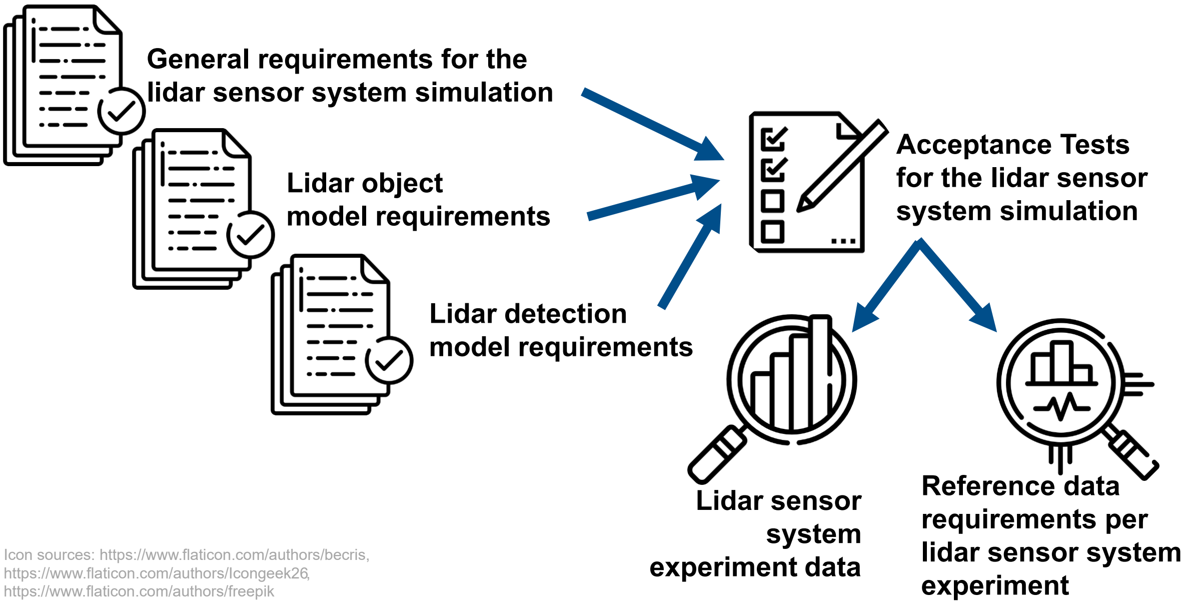

General requirements for the lidar sensor system simulation
Lidar object model requirements
Lidar detection model requirements
Acceptance Tests for the lidar sensor system simulation
Lidar sensor system experiment data
Reference data requirements per lidar sensor system experiment
Icon sources: https://www.flaticon.com/authors/becris, https://www.flaticon.com/authors/Icongeek26, https://www.flaticon.com/authors/freepik

Figure 7-7: Graduated definition of requirements for successive interfaces from recent presentation of the author[294]

An exemplary excerpt of the requirements on all all functional levels is kindly provided by the project in Annex B. To give an impression of such differentiated specification, the exemplary excerpt from the lidar detection model specification is provided in Tab. 7-2. The interconnections with the other lists is ensured by the IDs and the lists of related IDs from different tables in the respective columns. Due to the consistency of CEPRA and the specifications from the same working group in the same research project, there are even interconnections provided from the given specifications in Annex B to the filled CEPRA in Annex A.

To ensure credibility at the end, each requirement is accompanied by one or multiple acceptance tests, as depicted in Fig. 7-7 and listed e.g. in Tab. 7-2. Additionally, every acceptance test for the lidar sensor system simulation that uses real data has connected lidar sensor system experiment data requirements. As already explained several times, only specifying experiments and measurement data collection is not enough for re-simulation. Therefore, reference data requirements per lidar sensor system experiment are defined in the VVM project as well. Exemplary excerpts of all of the mentioned lists are provided in Annex B for completeness.

With this set of tables, for the first time ever, the APSS specifications list is complete. Of course, for the scope of this dissertation, only exemplary excerpts of the huge amount of tables are contained. So, the methodology for requirements definition from PerCollECT over CEPRA and the final specifications and acceptance tests is complete and a first application is provided.

[294] Rosenberger, P. et al.: Validation of Test Infrastructure (2022), p. 12.

Table 7-2: Lidar detection model requirements from the VVM project*

Lidar detection model requ. ID	Description	Type	Related lidar sensor system simulation requirement IDs	Related lidar object model requirement IDs	Acceptance test IDs	Status
Lid_ det_mod_ req_001	Lidar detection model shall not output any detections originating from an object completely occluded by another object.	Requirement	Lid_sys_sim_req_001	Lid_obj_mod_req_001	Lid_sys_sim_test_001	accepted
Lid_ det_mod_ req_002	Lidar detection model shall output less detections for partly occluded objects compared to the non-occluded case.	Requirement	Lid_sys_sim_req_002	Lid_obj_mod_req_001	Lid_sys_sim_test_002	accepted
Lid_ det_mod_ req_003	Lidar detection model shall output lower EPW for detections where lidar beams are partly occluded compared to the non-occluded case.	Requirement	Lid_sys_sim_req_002, Lid_sys_sim_req_009	Lid_obj_mod_req_002	Lid_sys_sim_test_002	accepted
Lid_ det_mod_ req_004	Lidar detection model shall contain signal scattering by atmospheric aerosols and therefore output lower EPW per detection and less detections for objects in or behind such aerosols compared to the aerosol-free case.	Requirement	Lid_sys_sim_req_009, Lid_sys_sim_req_010	Lid_obj_mod_req_001, Lid_obj_mod_req_002	Lid_sys_sim_test_003	accepted
Lid_ det_mod_ req_005	Lidar detection model shall contain absorption by atmospheric aerosols and therefore output lower EPW per detection and less detections for objects in or behind such aerosols compared to the aerosol-free case.	Requirement	Lid_sys_sim_req_009, Lid_sys_sim_req_011	Lid_obj_mod_req_001, Lid_obj_mod_req_002	Lid_sys_sim_test_004	accepted
• • •						

* Kindly provided, not yet published intermediate project result of publicly funded research project VVM

8 Metric Evaluation for Specification and VV&UQ of APSS Simulation

This chapter addresses RQ. 2, the evaluation of a metric for VV&UQ of APSS simulation, as identified in Sec. 4.3. Before this evaluation begins, the criteria for such metrics are to be clarified. For this purpose, Schaermann[203b] condensed seven criteria for validation metrics, combining the lists of six criteria by Oberkampf and Barone[295] and the seven desired features from Liu et al.[296]. As they are already used for metrics selection in literature e.g. by Magosi et al.[297], they are taken in this dissertation as well. They demand that:

1. Metrics are intuitive. (=> easily understandable & output in unit of measurand)*
2. Metrics are applicable to both deterministic and non-deterministic data.
3. Metrics are quantitative and objective. (=> no own parameters)*
4. Metrics do not include acceptance criteria. (=> no Boolean output)*
5. Metrics consider uncertainties. (=> epistemic and aleatory, as defined in Sec. 1.2.7)*
6. Metrics define a confidence interval with respect to the number of measurement data.
7. Metrics meet the mathematical properties of a metric. (=> unbounded results)*

Most criteria need further explaining, as it is originally left open what is meant by “intuitive” and how uncertainties should be considered (if and how epistemic and aleatory should be distinguished). Therefore, the metric criteria defined by Ferson et al. with a more practical point of view[298] are reflected by additions from the author in parenthesis behind some criteria and are applied in the following metric evaluation. For completeness, the four mathematical properties of a metric, as originally defined by Fréchet[299] in 1906, are:

1. Non-negativity: $d(\zeta, \widetilde{\zeta}) \geq 0$
2. Symmetry: $d(\zeta, \widetilde{\zeta}) = d(\widetilde{\zeta}, \zeta)$
3. Triangle inequality: $d(\zeta, \widetilde{\zeta}) + d(\widetilde{\zeta}, z) \geq d(\zeta, z)$
4. Identity of indiscernibles: $d(\zeta, \widetilde{\zeta}) = 0$ if and only if $\zeta = \widetilde{\zeta}$

[295] Oberkampf, W. L.; Barone, M. F.: Measures of agreement between computation and experiment (2006), pp. 11-12.

[296] Liu, Y. et al.: Toward a Better Understanding of Model Validation Metrics (2011), p. 2.

[297] Magosi, Z. F. et al.: Evaluation of Physical Radar Perception Sensor Models (2022), p. 11.

* Additions by the author

[298] Ferson, S. et al.: Model validation and predictive capability (2008), pp. 2415-2416.

[299] Fréchet, M. R.: Sur quelques points du calcul fonctionnel (1906).

8.1 Evaluation of Metrics Applied for APSS V&V

In Sec. 4.2.2, a comprehensive description and list of metrics applied for APSS V&V in literature is provided in Tab. 4-3. For metric selection later in this dissertation, the following section is the evaluation of that metric collection with Tab. 8-1 as the condensation of all information about the metrics from literature. If a metric is capable of a category given by the column title, it is marked in a specific shade of green, otherwise the cells stay blank.

The evaluation starts with the interfaces that the metrics are applied to and all possible interfaces it could have been applied to (**D**: Detections, **F**: Features/OGs, **O**: Objects). Then, the scenarios that are producing the data it is applied to ($\bullet/\rightsquigarrow$: (Quasi) static/dynamic) are provided, some have already been introduced in Sec. 5.2. Additionally, in Tab. 8-1 the scale of measurement it is able to process is considered (**M**: Metric (interval or ratio), **O**: Ordinal). The character of the uncertainties it is able to face ($\int/\mapsto$: Aleatory/epistemic) is given as well. For these first columns, indicators are inserted that mark if the metrics are applied without modification in literature (x), or if the metrics are applied in literature with adaptions like arithmetic mean, aggregation over time, or Euclidean distance matrix ($\star$). Furthermore, the output it is providing is considered for the evaluation (**R**: Real valued, **P**: Probabilistic). Special attention is given to covering the seven criteria for validation metrics from literature at the beginning of this chapter, which were expanded by the author to be able to objectively answer the binary assessment.

The scale of measurement plays a crucial role in statistics, when metrics are applied for comparison of measurements and simulations. In most cases, metric scales (interval or ratio) are necessary for application, as not only frequency (nominal scale) or rank (ordinal scale), but also e.g. the arithmetic mean of a set of values can be computed. For example for averaging given values, the arithmetic mean $\frac{1}{n}\sum_{i=1}^{n} x_i$ of n values $x_i \ldots x_n$ and the geometric mean $\sqrt[n]{\prod_{i=1}^{n} x_i}$ exist. However, the arithmetic mean can already be computed for interval-scaled values, whereas geometric mean needs ratio-scaled values. Therefore, before any metric is applied, the scaling must be checked and sometimes normalization or logarithms must be applied to avoid misleading evaluation of statistics, as explained in Sec. 8.2.5.

A first analysis of all metrics in Tab. 8-1 indicates that the usage of data processing like OGs and object tracking for validation is very subjective to a specific use case due to all parameters and embodiment of such algorithms. When the simulation is used to test an algorithm or a function with synthetic data, the comparison of results can be used for model falsification for that purpose. In the case of no difference between a function's reaction on real or synthetic data, sample validation for that specific function is possible, but to generalize this for other functions or function updates remains questionable. In other words: Using subsequent data processing is not a metric in the sense of measuring the distance between simulated and real detections, but mainly a falsification tool. Additionally, data processing cannot be generalized or used for benchmarking, as this would require standard algorithms to apply in the exact same form in every case.

Table 8-1: Evaluation of metrics applied for APSS simulation, see Tab. 4-3 for metric acronyms and calculation. Green color: Metric is capable of a category, x/⋆: Metric applied in literature without/with adaptions.

#	Metric	Interface			Scen.		Scale		Unc.		Out		Covered criteria						
		D	F	O	•	⇝	M	O	∫	⟼	R	P	1	2	3	4	5	6	7
1	d_{Ma}	🟩	🟩	x	🟩	x	x				🟩		🟩	🟩	🟩	🟩			🟩
2	OE	⋆	x	🟩	x	x	x				🟩		🟩	🟩	🟩	🟩			🟩
3	$\overline{d}$	x	🟩	🟩	🟩	x	x				🟩		🟩	🟩	🟩	🟩			🟩
4	d_{Eu}	🟩	🟩	x	🟩	x	x				🟩		🟩	🟩	🟩	🟩			🟩
5	RSS	⋆	x	🟩	🟩	x	x				🟩		🟩	🟩	🟩	🟩			🟩
6	d_{Ch}	🟩	🟩	x	🟩	x	x				🟩		🟩	🟩	🟩	🟩			🟩
7	RMSE	🟩	🟩	x	x	x	x				🟩		🟩	🟩	🟩	🟩			🟩
8	d_{Ha}	x	🟩	🟩	x	⋆	x				🟩			🟩	🟩	🟩			🟩
9	d_{Wa}	x	🟩	🟩	x	x	x		🟩		🟩			🟩	🟩	🟩			🟩
10	d_{PP}	x			x	⋆	x				🟩		🟩	🟩	🟩	🟩			🟩
11	d_{PC}	x			x	⋆	x				🟩		🟩	🟩	🟩	🟩			🟩
12	BBE	x			x	⋆	x				🟩			🟩		🟩			🟩
13	MS	⋆	x		x	⋆	x				🟩			🟩	🟩	🟩			
14	OCR	🟩	x		x	⋆	x				🟩		🟩	🟩	🟩	🟩			
15	FCR	🟩	x		x	⋆	x				🟩		🟩	🟩	🟩	🟩			
17	UPD	🟩	x		x	⋆	x				🟩		🟩	🟩	🟩	🟩			
16	OPD	🟩	x		x	⋆	x				🟩		🟩	🟩	🟩	🟩			
18	PD	⋆	x		x	⋆	x				🟩		🟩	🟩	🟩	🟩			🟩
19	$\mathrm{C_B}$	⋆	x		x	⋆	x				🟩			🟩	🟩	🟩			
20	$\mathrm{C_P}$	⋆	🟩	x	x	⋆	x				🟩			🟩	🟩	🟩			
21	$\mathrm{C_C}$	🟩	🟩	x	🟩	x	x				🟩			🟩		🟩			
22	ACD	🟩	🟩	x	🟩	x	x				🟩			🟩		🟩			
23	DTW			x	🟩	x	x				🟩					🟩			
24	Jaccard			x	x	x	x				🟩		🟩	🟩	🟩	🟩			
25	OSPA			x		x	x				🟩					🟩			
26	OSPA-T			x		x	x				🟩					🟩			
27	OSPA-MT			x		x	x				🟩					🟩			
28	GOSPA			x		x	x				🟩					🟩			
29	Rahmathulla			x		x	x				🟩					🟩			🟩
30	D_{KL}	⋆	🟩	x	x	x	x	🟩	x		🟩				🟩	🟩	🟩		
31	d_{JS}	x	🟩	🟩	🟩	x	x	🟩	x		🟩				🟩	🟩	🟩		
32	d_{AVM}	🟩	🟩	x	🟩	x	x	🟩	x	🟩	🟩		🟩	🟩	🟩	🟩	🟩		🟩
33	f_{KS}	x	🟩	🟩	🟩	x	x	🟩	x		🟩			🟩		🟩	🟩	🟩	
34	DEM	x	🟩	🟩	🟩	x	x					🟩				🟩			

As already stated, it is up to discussion, what is meant by “intuitively understandable” and how the uncertainties and their epistemic or aleatory character should be considered. However, there is only a single metric that outperforms all others regarding the coverage of the metric criteria, as visible in Tab. 8-1: The AVM (d_{AVM}, metric 32). Besides criterion 6 regarding the confidence interval that is only met by hypothesis testing in the derived list, it fulfills all criteria and is furthermore able to process all kinds of interfaces, data and scenarios. Furthermore, a confidence interval will be addressed by error prediction later on and is not necessarily part of a validation metric, from the experience of the author.

The AVM is intuitive, as calculating areas between two curves in a plot is very figurative to most people, either on CDFs, EDFs, or p-boxes. Besides, AVM results in a value in the unit of the measurand, which is very useful in specification of model acceptance and even practicable for project managers[300].

Additionally, AVM is the only already applied metric in literature that can theoretically handle p-boxes from epistemic and aleatory uncertainties[301], no other metric has this column filled green in Tab. 8-1. Therefore, it is the preferred metric after the evaluation of all already applied metrics in literature. Still, more investigations and a direct comparison of the AVM with the other possible metric candidates that can handle probabilistic data reflecting uncertainties (even if only aleatory) is provided in the next section. These candidates are Kolmogorov-Smirnov divergence, Jensen-Shannon distance/divergence, and frequency of positive Kolmogorov-Smirnov tests (metrics 30, 31, 33).

8.2 Assessment of Area Validation Metric (AVM) and Other Metric Candidates

For comparing the best metrics from the previous evaluation and some new candidates in detail, the method of manufactured universes is applied, as proposed for assessing VV&UQ approaches by Stripling et al.[302] or in the validation metric assessment by Liu et al.[303] In real world, uncertainty is always present and no GT exists, but an artificial, manufactured universe provides GT, as it is self-defined. Several EDFs of artificial range measurements are manufactured for the purpose of this section, as can be seen in Fig. 8-1. One serves as the measurement (real) EDF, seven other EDFs serve as simulation results to apply the metrics for measuring their difference to the real data. With that manufactured universe at hand, a detailed metric assessment and comparison is possible.

[300] Ferson, S. et al.: Model validation and predictive capability (2008), p. 2415.

[301] Oberkampf, W. L.; Ferson, S.: Validation Under Aleatory and Epistemic Uncertainty (2007).

[302] Stripling, H. F. et al.: The Method of Manufactured Universes (2011).

[303] Liu, Y. et al.: Toward a Better Understanding of Model Validation Metrics (2011).

All manufactured EDFs consist of $n = 100$ random samples from a normal distribution $\mathcal{N}(\overline{\zeta}, \overset{\vdash\!\dashv}{\zeta})$. The values of the respective means $\overline{\zeta}$ and standard deviations $\overset{\vdash\!\dashv}{\zeta}$ are given in Tab. 8-2. The first simulated EDF is sampled from a higher mean and doubled standard deviation than the real. The second simulated distribution has the same mean as the first one, but the same standard deviation as the real distribution. The third simulation produces data with half the standard deviation of the real data and a marginally smaller mean than the real data. Simulation 4 has a three times higher standard deviation, but the same mean as the real data. The fifth simulation's EDF is the exact same as the fourth, but shifted by $-0.50\,\mathrm{m}$. The sixth simulation's EDF is the exact same as the second, but shifted by $-0.65\,\mathrm{m}$. The seventh simulation has the exact same underlying distribution as the real data, but produces different samples.

Table 8-2: Manufactured real and simulated EDFs

EDF	**real**	**sim$_1$**	**sim$_2$**	**sim$_3$**	**sim$_4$**	**sim$_5$**	**sim$_6$**	**sim$_7$**
$\overline{\zeta}$	$22.00\,\mathrm{m}$	$22.55\,\mathrm{m}$	$22.55\,\mathrm{m}$	$21.91\,\mathrm{m}$	$22.00\,\mathrm{m}$	$21.50\,\mathrm{m}$	$21.91\,\mathrm{m}$	$22.00\,\mathrm{m}$
$\overset{\vdash\!\dashv}{\zeta}$	$0.20\,\mathrm{m}$	$0.40\,\mathrm{m}$	$0.20\,\mathrm{m}$	$0.10\,\mathrm{m}$	$0.60\,\mathrm{m}$	$0.60\,\mathrm{m}$	$0.20\,\mathrm{m}$	$0.20\,\mathrm{m}$

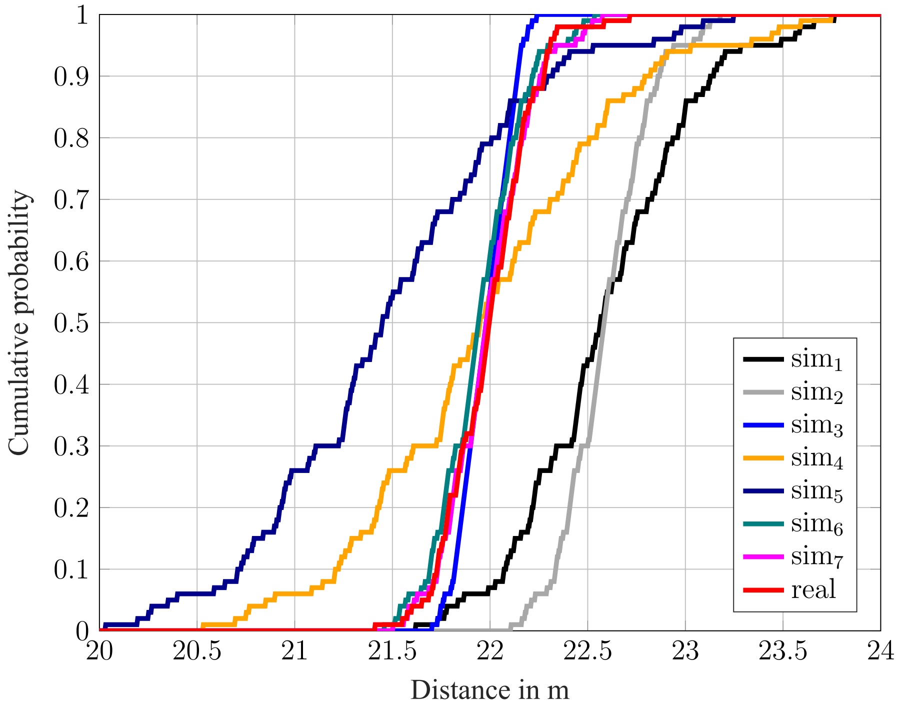


Figure 8-1: Manufactured EDFs for metric testing

The first important evaluation, however, is the comparison of the number of values from real data n and the number of values from simulation m. It seems trivial at first, but already contains information of detection or object existence uncertainty, which is based on the underlying signal thresholding for detection identification from the signal, as already mentioned in Sec. 1.2.7. Therefore, it is sometimes actually quite challenging to replicate a similar amount of detections, e.g. when environmental effects are considered. A comparable amount of values is a prerequisite when comparing CDFs or EDFs, as imbalanced data distorts the results.

8.2.1 AVM for Model Bias and Scattering Error

Quantiles like the median are characteristic values of distributions, essential in CDF and EDF plots, and can be compared even for ordinal data. Furthermore, as measurement data is mostly metric, as is the case for the range measurements of the manufactured universe, arithmetic mean and standard deviation can be computed. Both latter values together characterize (normal) distributions completely and are often the basis for stochastic model calibration. Therefore, median error (d_{medi}), model bias (mean error $\overline{d}$, metric 3), and scattering error (d_{stdv}) results are provided as a starting point for the metric evaluation in Tab. 8-3. Additionally, AVM results are listed there.

Table 8-3: First metric results for a (manufactured) real vs. seven simulated EDFs

real vs.	d_{medi}	$\overline{d}$	d_{stdv}	d_{AVM}
sim$_1$	$0.59\,\mathrm{m}$	$0.60\,\mathrm{m}$	$0.22\,\mathrm{m}$	$0.60\,\mathrm{m}$
sim$_2$	$0.60\,\mathrm{m}$	$0.60\,\mathrm{m}$	$0.00\,\mathrm{m}$	$0.60\,\mathrm{m}$
sim$_3$	$0.04\,\mathrm{m}$	$0.04\,\mathrm{m}$	$0.11\,\mathrm{m}$	$0.07\,\mathrm{m}$
sim$_4$	$0.03\,\mathrm{m}$	$0.00\,\mathrm{m}$	$0.44\,\mathrm{m}$	$0.34\,\mathrm{m}$
sim$_5$	$0.53\,\mathrm{m}$	$0.50\,\mathrm{m}$	$0.44\,\mathrm{m}$	$0.57\,\mathrm{m}$
sim$_6$	$0.04\,\mathrm{m}$	$0.04\,\mathrm{m}$	$0.00\,\mathrm{m}$	$0.05\,\mathrm{m}$
sim$_7$	$0.00\,\mathrm{m}$	$0.00\,\mathrm{m}$	$0.00\,\mathrm{m}$	$0.02\,\mathrm{m}$

A metric for probabilistic data should reflect if the underlying distribution is the same or just slightly different or even very different. At first sight, median and mean error provide almost the same information, while the median is a bit more sensitive to small differences. Both are insensitive to scattering errors, as the results for simulation 4 show. Nevertheless, no visual inspection of the seven simulated EDFs would say that the fourth simulation is as good as e.g. simulations 3, 6, and 7, so scattering error must be considered as well.

Here, AVM has its benefits, as it considers a combination of both. As can be taken from Tab. 8-3, not-overlapping EDFs (simulations 1 and 2) are penalized most, but simulations with different standard deviations get relatively high metric values, even if the mean is identical to the real data (simulation 4). It finds the very small difference between simulation 7 and the real data, so it is sensitive even for tiny deviations. And it provides the results in the unit of the measurand, as has been mentioned already earlier, which provides an intuitive feeling for the differences of the simulations.

A closer look at the initial results shows that when there is no overlap, as is the case with the first two simulations, AVM is very close to the median and mean error. This can be explained by AVM being an integral, so it can be computed either over the abscissa or over the ordinate. Therefore, on the one hand, as listed in Tab. 4-3 (metric 32), AVM is the integral of the absolute difference between the cumulative distribution functions over all real and simulated sensor measurements

$$d_{\mathrm{AVM}}(F,\widetilde{F}) = \int_{-\infty}^{\infty} |F(\zeta) - \widetilde{F}(\zeta)| \,\mathrm{d}\zeta\,. \tag{8-1}$$

On the other hand, due to the fact that cumulated probability $F(\zeta)$ is limited to $[0,1]$ and unitless with m (e.g. 100) quantiles, the integral over probability can be written as the mean error (metric 3) of all m quantiles of the CDF like

$$d_{\mathrm{AVM}}(F,\widetilde{F}) = \int_{0}^{1} |\zeta(F) - \widetilde{\zeta}(F)| \,\mathrm{d}F \;=\; \frac{1}{m}\sum_{i=1}^{m} |\zeta(F_i) - \widetilde{\zeta}(F_i)|\,. \tag{8-2}$$

In this form, the AVM is only slightly different to the mean error (metric 3 in Tab. 4-3) build over all n measurements

$$\overline{d} = \frac{1}{n}\sum_{i=1}^{n} |\zeta_i - \widetilde{\zeta}_i|\,. \tag{8-3}$$

8.2.2 Evaluation of Other Metric Candidates

There are other metric candidates found from Tab. 8-1 that can compare probabilistic data, namely Kolmogorov-Smirnov divergence, Jensen-Shannon distance/divergence, and frequency of positive Kolmogorov-Smirnov tests. Essentially, using the frequency of positive hypothesis testing as done by Eder et al. for Kolmogorov-Smirnov tests is a trick to not result in binary output, which would eliminate such a metric candidate according to criterion 4 from the beginning of this chapter. In the detailed metric evaluation by Liu et al., Bayes factor and the frequentist metric have also been evaluated and their results disqualified them for the scope of this dissertation due to being not objective (Bayes) or by not including all sources of uncertainty.[304]

Besides, there is a wide debate ongoing in the last decade in different fields of science on the excessive and often wrong usage of hypothesis testing, the p-value, and the significance level α of $5\,\%$.[305,306,307] Still, there are sound arguments for hypothesis testing, when applied correctly, and also using the frequency of such tests is somewhat permissible. A conclusion from the ongoing discussion is that when such tests are applied, at least all data and values should be given to a full extend and not only a binary statement on significance or not.[306] In this regard, in the following, all values produced by the tests are listed for the metric candidates. To obtain the listed results, for both tests, the widely used implementations by The MathWorks, Inc.[308,309] are applied.

[304] Liu, Y. et al.: Toward a Better Understanding of Model Validation Metrics (2011), p. 12.

[305] Nuzzo, R.: Scientific method (2014).

[306] Amrhein, V. et al.: Scientists rise up against statistical significance (2019).

[307] Matthews, R.: The p-value statement, five years on (2021).

[308] The MathWorks, Inc.: Two-sample Kolmogorov-Smirnov test (2022).

[309] The MathWorks, Inc.: Anderson-Darling test (2022).

As listed in Tab. 4-3 (metric 33), the Kolmogorov-Smirnov test computes the hypothesis result based on the test statistic

$$D_{\text{KS}} = \sup_{\zeta \in \{\boldsymbol{\zeta} \cup \widetilde{\boldsymbol{\zeta}}\}} |F(\zeta) - \widetilde{F}(\zeta)|, \tag{8-4}$$

with $H_{\text{KS}} = \left(D_{\text{KS}} \geq c_\alpha \sqrt{\frac{n+m}{nm}}\right)$, $c_\alpha = \sqrt{\frac{\ln 2 - \ln \alpha}{2}}$, and significance level α. Its results for the seven simulated EDFs and the real EDF are given in Tab. 8-4.

For a full impression on the metric candidates' capabilities, not only this test already used for validation of APSS simulation, but also Anderson-Darling[310] test results are computed with the test statistic computed as

$$D_{\text{AD}} = n \int_0^1 (F(\zeta) - \widetilde{F}(\zeta))^2 \, w(\zeta) \, \mathrm{d}F(\zeta), \tag{8-5}$$

with $w(\zeta) = [F(\zeta)(1 - F(\zeta))]^{(-1)}$ being provided in Tab. 8-4. While the former compares the supremum of the difference between two CDFs, the latter is a quadratic EDF statistic like the Cramér–von Mises test (that uses $w = 1$), but puts more weight on the tails of the EDF.

Table 8-4: Kolmogorov-Smirnov and Anderson-Darling test results, both with $\alpha = 5\,\%$, compared to AVM results for a (manufactured) real vs. seven simulated EDFs

real vs.	**Kolmogorov-Smirnov**			**Anderson-Darling**			d_{AVM}
	H_{KS}	D_{KS}	p_{KS}	H_{AD}	D_{AD}	p_{AD}	
sim$_1$	1	0.70	0.00	1	∞	0.00	0.60 m
sim$_2$	1	0.89	0.00	1	400	0.00	0.60 m
sim$_3$	1	0.27	0.00	1	11	0.00	0.07 m
sim$_4$	1	0.28	0.00	1	∞	0.00	0.34 m
sim$_5$	1	0.59	0.00	1	310	0.00	0.57 m
sim$_6$	0	0.16	0.14	1	4.3	0.01	0.05 m
sim$_7$	0	0.06	0.99	0	0.68	0.57	0.02 m

The results for the hypothesis tests show that the Anderson-Darling test is a much more sensitive test for comparing two distributions, as it only considers simulation 7 as originating from the same underlying distribution as the real data, which is very accurate as well. On the other hand, experts would probably vote like the Kolmogorov-Smirnov test and tell that simulation 6 is not distinguishable from the real data regarding the underlying distribution as well. Both tests somehow reflect the intuitive order of similarity between the simulations to some degree, while the Anderson-Darling test is again too extreme to be useful, as both simulation 1 and 4 get an ∞ statistic. Kolmogorov-Smirnov, however, makes a counter-intuitive statement for simulation 1 and 2, as both have a bias, but the first one has double the standard deviation than the real data and therefore should intuitively be penalized more than the second.

[310] Anderson, T. W.; Darling, D. A.: Asymptotic Theory of Certain "Goodness of Fit" Criteria (1952).

It should be noted that Ferson et al. find similar issues with counter-intuitive Kolmogorov-Smirnov test results for CDF comparison in contrast to the AVM matching intuition very well.[311] To conclude on hypothesis testing, the necessary trick to use their frequency, as Eder et al. did, the reported discussions on using them at all, and the sometimes misleading and counter-intuitive results are the reasons to be not considered further in this dissertation.

Finally, the last metric candidates left are the divergences defined by Kolmogorov-Smirnov and Jensen-Shannon, the metrics 30, 31 from Tab. 4-3. Their formula have already been written there, but are repeated here, as a third divergence is introduced for comparison with the AVM, which is a symmetrical version of the Kullback–Leibler divergence. In order, the Kullback–Leibler divergence of two distributions is obtained by

$$D_{\mathrm{KL}}(P(\zeta),\widetilde{P}(\zeta)) = \sum_{\zeta\in\{\boldsymbol{\zeta}\cap\widetilde{\boldsymbol{\zeta}}\}} P(\zeta)\cdot\log_2\Big(\frac{P(\zeta)}{\widetilde{P}(\zeta)}\Big). \tag{8-6}$$

The symmetrical version of the Kullback–Leibler divergence is the mean of both possible unsymmetrical Kullback–Leibler divergences that is given by

$$D_{\mathrm{sym}}(P(\zeta),\widetilde{P}(\zeta)) = \overline{P}(\zeta) = \frac{1}{2}[D_{\mathrm{KL}}(P(\zeta),\widetilde{P}(\zeta)) + D_{\mathrm{KL}}(\widetilde{P}(\zeta),P(\zeta))]. \tag{8-7}$$

The Jensen-Shannon divergence uses this mean and is itself symmetrical, calculated by

$$D_{\mathrm{JS}}(P(\zeta),\widetilde{P}(\zeta)) = \frac{1}{2}D_{\mathrm{KL}}(P(\zeta),\overline{P}(\zeta)) + \frac{1}{2}D_{\mathrm{KL}}(\widetilde{P}(\zeta),\overline{P}(\zeta)), \tag{8-8}$$

while being normalized to the interval $[0,1]$. The results from the three divergences together with the AVM are given in Tab. 8-5.

Table 8-5: Divergence results compared to AVM results for a (manufactured) real vs. seven simulated EDFs

real vs.	D_{KL}	D_{sym}	D_{JS}	d_{AVM}
$\textbf{sim}_1$	1.62	4.82	0.41	0.60 m
$\textbf{sim}_2$	7.11	5.70	0.66	0.60 m
$\textbf{sim}_3$	2.34	1.41	0.13	0.07 m
$\textbf{sim}_4$	0.95	4.39	0.27	0.34 m
$\textbf{sim}_5$	1.37	8.52	0.37	0.57 m
$\textbf{sim}_6$	0.06	0.05	0.01	0.05 m
$\textbf{sim}_7$	0.02	0.02	0.01	0.02 m

As with hypothesis testing, again counter-intuitive results are given for the first two simulation EDFs that have no overlap with the EDF from real data. The intuition tells that the second simulation with the same standard deviation as the real distribution should have less deviation scores than the first one with twice the standard deviation, but the opposite is shown in Tab. 8-5.

[311] Ferson, S. et al.: Model validation and predictive capability (2008), pp. 2418-2419.

The symmetrical Kullback–Leibler divergence in this example penalizes the simulation 5 with same mean but different standard deviation the most, while both other divergences are more focused on bias errors in this case. As it is up to discussion what type of error is more critical, again the main difference compared to the AVM is the missing output unit of the measurand, which is of major importance for interpreting the results, as e.g. stated by Ferson et al.[311]

8.2.3 AVM with Epistemic Uncertainty in P-Boxes

As already stated in Sec. 4.1.5, none of the reported re-simulations of reference data for validation of APSS simulation considers epistemic uncertainty. P-boxes are already introduced in the same chapter for reflecting aleatory and epistemic uncertainty and model error estimation based on sample validation with AVM as the metric and uncertainty aggregation including confidence from other simulation domains are given and visualized in Fig. 4-16b and Fig. 4-18.

Aleatory uncertainty is the omnipresent scatter of measurements around the GT value. Referring to Sec. 1.2.7, epistemic uncertainty is unfavorably called “uncertainty of the lazy experimentalist”, as it should be minimized as far as possible. Nonetheless, accuracy limitations exist even for so-called GT odometry sensors using global navigation satellite system (GNSS) with real time kinematics (RTK), as shown by Holder et al., which recently published their massive efforts for reference data calibration while determining and minimizing the inherent epistemic uncertainty[312]. Consequently, reducing epistemic uncertainty is associated with considerable expenses and its elimination is impossible. However, the immense effort is justified in the trivial observation that high epistemic uncertainty leads to wide p-boxes, which can make it impossible to falsify any simulation, as the EDF compared to the simulation is completely covered at some point. A validation metric can be viewed as the evidence for mismatch between real and simulated measurements. Consequently, when the uncertainty in simulation encompasses the real data, as written by Oberkampf and Roy, *“there is no evidence of mismatch because accuracy is distinct from precision.”* [313]

On the other hand, allowing epistemic uncertainties as simulation input is fairness to the simulation when it is validated against not completely reproducible measurement data, as the results have better chance to fit to the measurement values. When re-simulation is performed with p-box as input, it must be decided how the uncertainty is propagated through the simulation model(s), as discussed in Sec. 4.1.5. Either way, in the regular case, every (sample) validation task afterwards consists of a comparison of a EDF from real data with a p-box from simulation. Fig. 8-2 therefore contains seven simulated p-boxes that are expanded versions of the EDFs from Fig. 8-1, forming a new manufactured universe. The EDFs are expanded by $\pm 0.10\,\mathrm{m}$ reflecting the epistemic uncertainty in target location and measurement device positions during the fictional experiments. The EDF from real data is exactly the same as in the previous example.

[312] Holder, M. F. et al.: Digitalize the Twin (2022).

[313] Oberkampf, W. L.; Roy, C. J.: Verification and Validation in Scientific Computing (2010), p. 545.

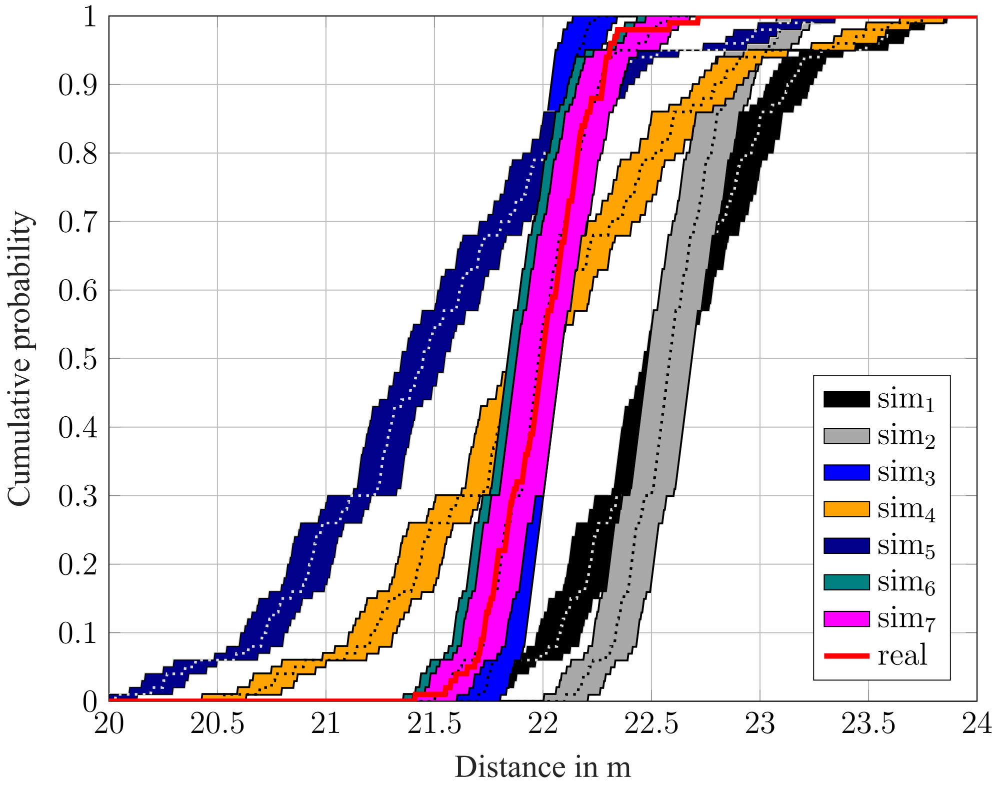


Figure 8-2: Manufactured p-boxes for metric testing

Ferson and Oberkampf [314a] give the more general form of the AVM from (8-1), when it is applied on two p-boxes or on a p-box and a EDF, which is written as

$$d_{\mathrm{AVM}}(\boldsymbol{\mathcal{F}},\widetilde{\boldsymbol{\mathcal{F}}}) = \int_{-\infty}^{\infty} d_{\mathrm{min}}\Big([F_{\mathrm{L}}(\zeta),F_{\mathrm{R}}(\zeta)],[\widetilde{F}_{\mathrm{L}}(\zeta),\widetilde{F}_{\mathrm{R}}(\zeta)]\Big)\,\mathrm{d}\zeta\,, \tag{8-9}$$

$$\text{with } \boldsymbol{\mathcal{F}} = [F_{\mathrm{L}}(\zeta),F_{\mathrm{R}}(\zeta)] \text{ and } d_{\mathrm{min}}(\boldsymbol{\mathcal{F}},\widetilde{\boldsymbol{\mathcal{F}}}) = \min_{\substack{\forall F\in\boldsymbol{\mathcal{F}}\\ \forall \widetilde{F}\in\widetilde{\boldsymbol{\mathcal{F}}}}} |F-\widetilde{F}|. \tag{8-10}$$

In this form, the smallest possible area is considered and not the mean or maximum of all possible AVMs that could be computed e.g. with the borders of the p-box. This is based on the statement from Oberkampf and Roy, who point out that *"a validation metric should not penalize the model for the empiricist's imprecision. [...] Thus, the validation metric between a point prediction and an interval datum is the shortest difference between the characterizations of the quantities."* [313]

Nonetheless, for a full impression, Tab. 8-6 provides the range of areas with $d_{\mathrm{AVM,min}} = d_{\mathrm{AVM}}$ and $d_{\mathrm{AVM,max}}$, which uses d_{max} instead of d_{min}. Thereby, the smallest possible area is straightforward computed, but the largest possible area is a complex optimization problem, as inside a p-box exists an infinite number of possible EDFs. Fig. 8-3 illustrates this problem, showing the smallest possible AVM on the left and largest AVM in the middle. The distribution leading to this largest possible AVM are depicted on the right.

[314] Ferson, S.; Oberkampf, W.: Validation of imprecise probability models (2009). a: p. 13.; b: p.17.; c: p.18.

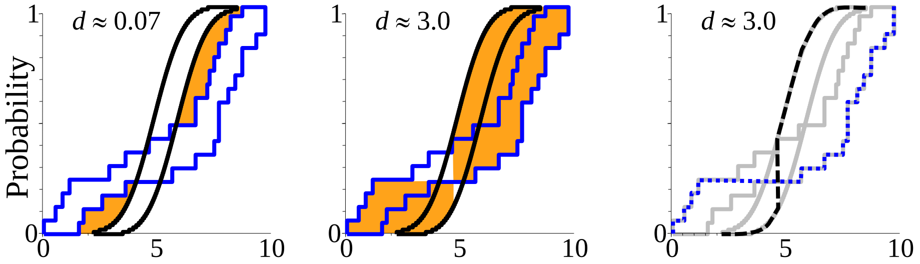


Figure 8-3: Exemplary computation of the range of AVMs for p-boxes, redrawn from Ferson and Oberkampf.[314b]

Besides, the results of the so-called “Double Metric” from Ferson and Oberkampf[314b] are given, that is a two-dimensional vector of the AVM results for the left and right border-EDFs, as

$$d_{\mathrm{AVM,2}} = \Big(d_{\mathrm{AVM,L}}, d_{\mathrm{AVM,R}}\Big) = \Big(d_{\mathrm{AVM}}(F_{\mathrm{L}}, \widetilde{F}_{\mathrm{L}}),\, d_{\mathrm{AVM}}(F_{\mathrm{R}}, \widetilde{F}_{\mathrm{R}})\Big). \tag{8-11}$$

Fig. 8-4 from Ferson and Oberkampf[314c] shows three examples of how the Double Metric is calculated in case of uncertain numbers and intervals of epistemic uncertainty.

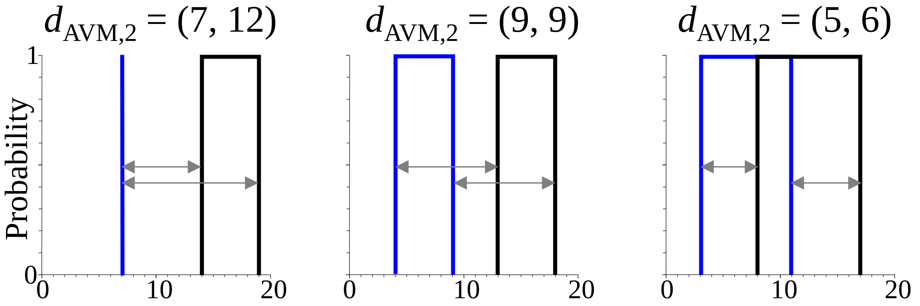


Figure 8-4: Example for two-dimensional Double Metric of AVMs for p-boxes, redrawn from Ferson and Oberkampf.[314c]

The values for the differently computed AVMs are given with two digits as all metric results before. The results show that again, the AVM scores of the two simulated p-boxes with no overlap with the real EDF have the same scores. In general, the smaller side of the Double Metric is often equal and always at least similar to the minimal AVM.

With the introduced epistemic uncertainty, the p-boxes from simulation 6 and 7 cover the EDF from real data almost completely and get a result for identity with the given two digits precision. Additionally, simulation 3, which is derived from data with half the standard deviation (see Tab. 8-2) also benefits from the epistemic uncertainty and the AVM of $0.02\,\mathrm{m}$ marks almost identity as well. Simulation 5 gets a metric result of $0.47\,\mathrm{m}$, which is almost as high as the results for simulation 1 and 2, whereas a small overlap with the real EDF is present. Finally, simulation 4, even if it has the same mean as the real data, still gets a metric result of $0.25\,\mathrm{m}$, due to its three times higher standard deviation. Therefore, the AVM has the same capabilities in the case of present epistemic uncertainty and the comparison of p-boxes, as shown for only aleatory uncertainty, when only EDFs are compared in Sec. 8.2.1.

Table 8-6: First metric results for a (manufactured) real vs. seven simulated p-boxes

real vs.	Range of AVMs		Double Metric	
	d_{AVM}	$d_{\mathrm{AVM,max}}$	$d_{\mathrm{AVM,L}}$	$d_{\mathrm{AVM,R}}$
sim$_1$	0.50 m	0.70 m	0.50 m	0.70 m
sim$_2$	0.50 m	0.70 m	0.50 m	0.70 m
sim$_3$	0.02 m	0.14 m	0.12 m	0.11 m
sim$_4$	0.25 m	0.45 m	0.36 m	0.35 m
sim$_5$	0.47 m	0.67 m	0.65 m	0.50 m
sim$_6$	0.00 m	0.15 m	0.14 m	0.06 m
sim$_7$	0.00 m	0.13 m	0.10 m	0.10 m

It should be noted that the AVM is no metric in the mathematical sense anymore, when applied to one or two p-boxes, as in the case of a full coverage of one p-box or EDF by the other p-box, the area in between falls to zero without the data becoming identical.[313] In contrast, the Double Metric is a mathematical metric and easily computed but too strict in assessing the equality of two distributions, as shown by Ferson and Oberkampf, while they *"expect that the shortest distance [AVM] will be most useful in many practical applications."*[314c]

8.2.4 AVM Insensitivity at Non-Overlapping Distribution Functions

All evaluations of the AVM have shown that it is the superior approach when comparing probabilistic data and in fact the only reported metric that can be applied to compare data with epistemic uncertainty successfully. Nevertheless, the results in Sec. 8.2.1 and Sec. 8.2.3 also reveal a limitation of the metric, namely the insensitivity to scattering error, when there is no overlap between the EDFs or p-boxes.

Oberkampf and Ferson already find that property of the AVM in their own study in 2007[315a], as shown in Fig. 8-5, which contains 6 different simulation results compared to a single uncertain number that serves as the real data in this case. The upper three plots show the ability of the AVM to distinguish smaller bias errors caused by higher epistemic uncertainty favoring the simulation results, even when there is no overlap. The lower three plots then show the insensitivity to different standard deviations, when there is no overlap or crossing point. Oberkampf and Ferson emphasize that this behavior shows an advantage of the AVM, namely the distinction between aleatory and epistemic uncertainty.[315a] Furthermore, they state that the insensitivity is intuitive in this shown example, as the three lower plots may have a smaller distance for higher quantiles, but have a higher distance for lower quantiles of the same amount and therefore should be penalized equally. However, the investigations in the previous sections of this dissertation show a different image that leads to the conclusion that in case of EDFs/p-boxes, the insensitivity is counter-intuitive and therefore a limitation and not a feature.

[315] Oberkampf, W. L.; Ferson, S.: Validation Under Aleatory and Epistemic Uncertainty (2007). a: p. 22.; b: p.7.

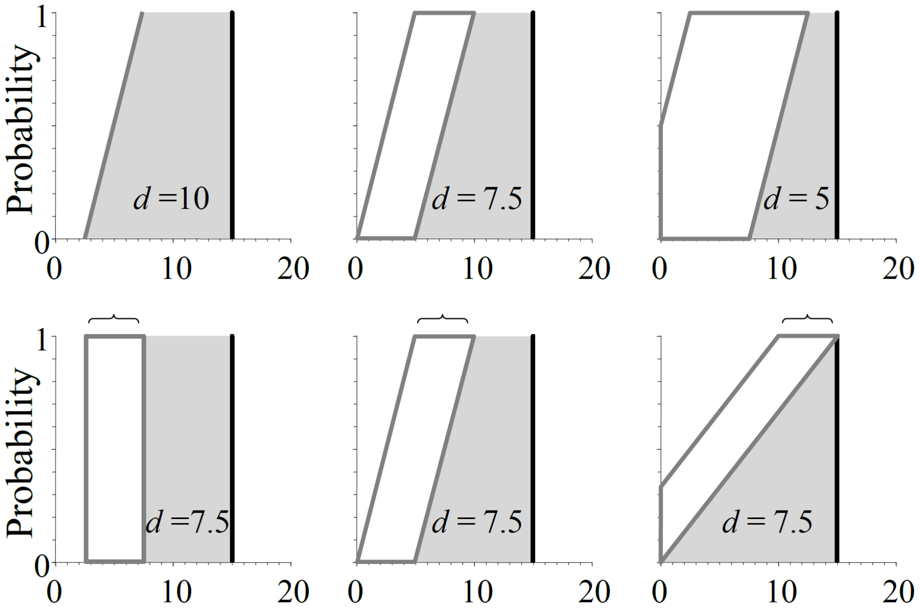


Figure 8-5: Insensitivity of AVM for increasing variance from Oberkampf and Ferson[315a]

8.2.5 AVM (not) for Ordinal Data

The scale of measurement is a quite important topic that has not yet been discussed in literature, to the knowledge of the author. It is common sense, that scaling of data, as defined by Stevens[316], matters and decides which metrics and statistical values can be obtained and when. While questioning the scale of measurement is not very popular, when sensor measurements are evaluated, it is known that already for calculating an arithmetic mean of data, strictly speaking, interval scaling (a.k.a. metric scaling) should be proven, beforehand. For measuring ranges, this seems quite trivial on first sight, but asking the question, “Does the range from $1.00\,\mathrm{m}$ to $2.00\,\mathrm{m}$ actually count the same as the range from $1000.00\,\mathrm{m}$ to $1001.00\,\mathrm{m}$?” can lead to some intense discussions and the answer is not that trivial anymore, especially, when criticality is influencing the discussion on range measurements to objects with APSS.

The question on the underlying scale of measurement gets even more difficult, when real and simulated intensity measurements of APSS are to be compared. Some, like the RCS are given in logarithmic scale, others, like the intensities given by some lidar sensors are coded with $8\,\mathrm{bit}$ in $[0, 255]$, while $[0, 100]$ is reserved for linear scaling and $[101, 255]$ is a logarithmic scale.[317] Besides the lack of R&R, as already discussed in Sec. 5.3, these scaling problems could be the cause that no publication has described lidar detection model V&V on intensities, yet.

There is an ongoing debate in science on metric application for ordinal data, e.g. in medicine and psychology, when using the Likert scale.[318] Some authors are very strict and completely prohibit

[316] Stevens, S. S.: On the Theory of Scales of Measurement (1946).

[317] Velodyne LiDAR, Inc.: VLP-16 User Manual (2019), p. 32.

[318] Likert, R.: A technique for the measurement of attitudes (1932).

the usage of metrics on non-metric data. Others have a more practical viewpoint and make it dependent on the necessary discussion of the results.[319,320] At least, it is common sense, that the scale of measurement should be discussed, which is performed in the following.

For demonstration purpose, a manufactured universe is created, where the application of any metrics on ordinal data is allowed. The sensor output to be simulated is the output of an object size classification, which has classes ordered by road user size that are explicitly not interval scaled, namely: Bicycle-sized (1), motorbike-sized (2), coupe-sized (3), sedan-sized (4), SUV-sized (5), and truck-sized (6). Each set of simulated or real data consists of $n = 100$ values. The relative frequencies of the real data and both simulated EDFs are depicted in Fig. 8-6a. The same real data with respective simulated p-boxes is plotted in Fig. 8-6b. In this case, the epistemic uncertainty results in an interval width of 1 size class to the left for the first simulation and to the right for the second.

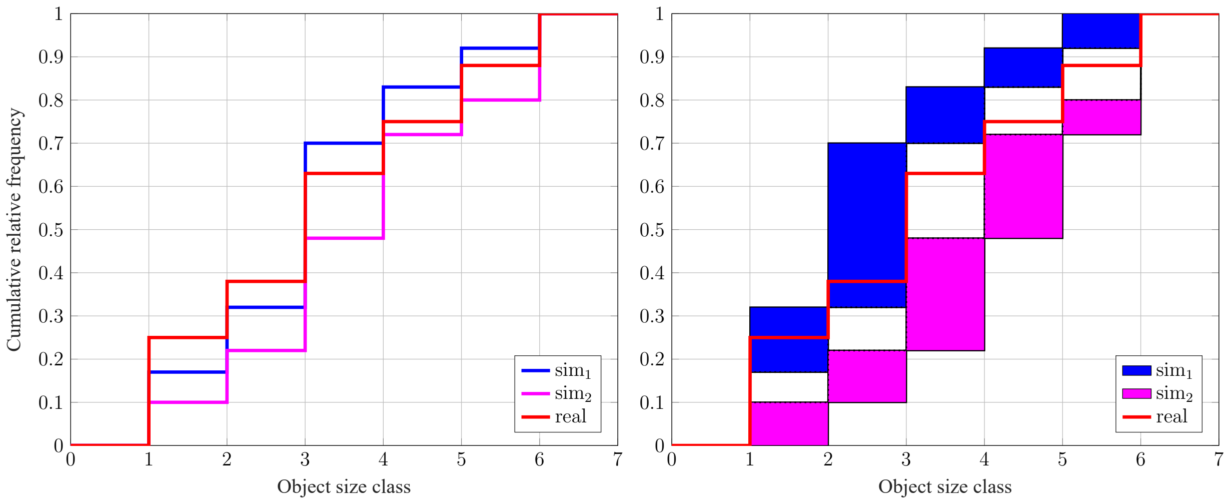


Figure 8-6: Manufactured EDFs (left, a) and p-boxes (right, b) for metric testing on ordinal scale

Table 8-7: Metric results for a (manufactured) real vs. two simulated ordinal EDFs

real vs.	d_{medi}	$\overline{d}$	d_{stdv}	D_{KS}	D_{AD}	D_{KL}	D_{sym}	D_{JS}	d_{AVM}
sim₁	0.00	0.05	0.26	0.08	80.6	0.06	0.06	0.01	0.33
sim₂	1.00	0.57	0.13	0.16	193	0.10	0.09	0.02	0.57

Table 8-8: AVM results for a (manufactured) real vs. two simulated ordinal p-boxes

real vs.	**Range of AVMs**		**Double Metric**	
	d_{AVM}	$d_{\text{AVM,max}}$	$d_{\text{AVM,L}}$	$d_{\text{AVM,R}}$
sim₁	0.19	0.89	0.88	0.33
sim₂	0.57	1.4	0.57	1.4

[319] Sauro, J.: Can You Take the Mean of Ordinal Data? (2016).

[320] Sullivan, G. M.; Artino, A. R.: Analyzing and Interpreting Data From Likert-Type Scales (2013).

While being very careful with such results that are treating ordinal data as metric, Tab. 8-7 shows that the first simulation has the same median and pseudo-mean size class as the real data, while the second simulation output is about half a size class to high on average. The pseudo standard deviation of the first simulation is way smaller than in the real measurements. The second simulation is not only biased, but has a slightly different pseudo standard deviation as well, even if the model scattering error is smaller. Considering this, all other metric results, be it from hypothesis testing, be it divergences, or the AVM, have the same message: The second simulation is double as different as the first one, when compared to the real data.

This exemplary demonstration is designed to stress the importance of checking the scale of measure. On purpose, it raises questions about the meaning of the intervals between the artificial size classes. In this case, the pseudo mean or pseudo scattering error do not have a meaning to anybody, as no one would say that the step from bicycle-sized to motorbike-sized is equal to SUV-sized to truck-sized. As already mentioned in Sec. 8.1, some time-based criticality metrics like the time-to-collision (TTC)[321] as discussed by Junietz[322] are prominent real world examples for values that are incorrectly treated as interval-scaled, while actually being only ordinal. Even if the TTC is measured in seconds, which suggest interval scaling, its meaning is the counterargument. A TTC interval of $0.5\,\mathrm{s}$ has a totally different meaning for TTC between $1.0\,\mathrm{s}$ and $1.5\,\mathrm{s}$ or TTC between $10.0\,\mathrm{s}$ and $10.5\,\mathrm{s}$.

Deprived of its capability to return values in the unit of the measurand, the AVM is as good or bad for ordinal EDFs as the other metrics in this evaluation. Nonetheless, it is still the only one handling epistemic uncertainty and p-boxes. In this case, as the first p-box has some overlap with the real EDF and the second simulation overlaps not at all, the second one is three times as bad as the first. Clearly, all comparisons in this section are of relative nature, and no absolute differences can be measured. This is a huge problem regarding model specification, as the simulation won't be specified with respect to a benchmark or reference simulation, but with absolute error thresholds, which is more or less impossible with simulations of ordinal data (and actually not allowed, with regard to strict statisticians).

8.3 Interim Conclusion on the Selected AVM

After all investigations on different metrics for data with and without epistemic uncertainty, the AVM is clearly the best metric candidate. Its capabilities are already used in automotive (vehicle consumption) simulation validation by Danquah et al.[323] While Schaermann neglects its usage on

[321] Hayward, J. C.: NEAR-MISS DETERMINATION THROUGH USE OF A SCALE OF DANGER (1972).

[322] Junietz, P. M.: Microscopic and Macroscopic Risk Metrics (2019).

[323] Danquah, B. et al.: Statistical Validation Framework for Automotive Vehicle Simulations (2021).

absolute APSS measurement values as *"its discrepancy values can result arbitrarily high"* [324], a bounded metric result is no criterion. Unboundedness is the case for most metrics evaluated, but there is no drawback on limiting the output, but quite the contrary. When there is a higher model bias, there should be a higher metric output.

Therefore, the AVM as defined by Ferson et al.[325] is unbounded and reflects differences in full distribution. Furthermore, it is intuitively calculated and gives results in the physical units of the measurand (and not some *"esoteric statistical units"*[315b]), and generalizes even to application on uncertain numbers and on deterministic values as well. Nonetheless, the insensitivity to variance, when there is no intersection, as shown in Fig. 8-5 is a known limitation. Finally, it has been shown that it can be used for cumulative relative frequency distribution of ordinal data, when the results are handled carefully as well.

[324] Schaermann, A.: Systematische Bedatung und Bewertung umfelderf. Sensormodelle (2020), p. 67.

[325] Ferson, S. et al.: Model validation and predictive capability (2008), pp. 2416-2419.

9 Tailored Metrics for Specification and VV&UQ of APSS Simulation

This chapter addresses RQ. 3, the application and further development (if necessary) of the best metric candidate. It should be tailored to the specifically high requirements in VV&UQ of APSS simulation, while being capable to handle epistemic uncertainty in the measurement and reference data arising from the limitations in R&R of APSS experiments.

9.1 Estimation of Model Bias and its Tendency

For validation, the sign of the metric does not matter, and the AVM consequently uses the absolute distance between two EDFs, as written in (8-1), or p-boxes as in (8-9). Nevertheless, for model development and calibration, the tendency of the simulation output compared to the real data is of high interest. In this regard, Voyles and Roy separate the AVM into two portions, d^+ and d^-, where the simulated EDF[326] or p-box[327] is higher (+) or lower (-) than the real EDF/p-box, which is illustrated for the latter in Fig. 9-1.

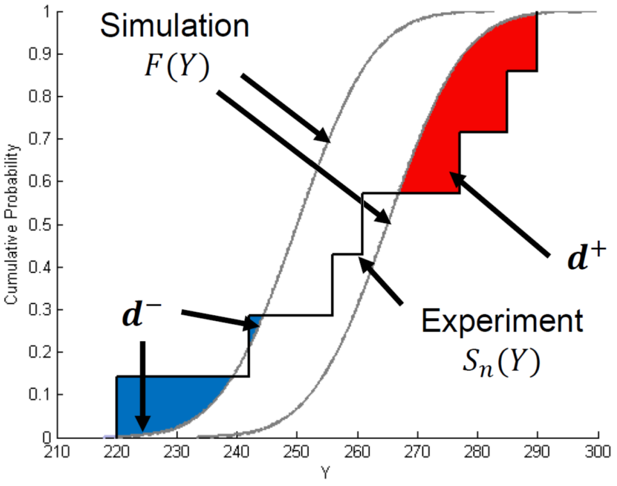


Figure 9-1: Portions of the AVM, where the simulated p-box is higher (d^+) or lower (d^-) than the real EDF, based on Voyles and Roy[327a]

With this distinction of its portions, the AVM computes as

$$d_{\mathrm{AVM}}(\boldsymbol{\mathcal{F}}, \widetilde{\boldsymbol{\mathcal{F}}}) = d^- + d^+ \,. \tag{9-1}$$

[326] Voyles, I. T.; Roy, C. J.: Model Validation in the Presence of Uncertainty (2014), p. 4.

[327] Voyles, I. T.; Roy, C. J.: Model Validation in the Presence of Aleatory and Epistemic Uncertainties (2015). a: p. 5.; b: p. 7; c: p. 4

Voyles and Roy conclude with the two portions at hand, an estimate for the model bias $\widehat{d}_{\text{bias}}$ can be computed[326] as

$$\widehat{d}_{\text{bias}}(\boldsymbol{\mathcal{F}},\widetilde{\boldsymbol{\mathcal{F}}}) = d^{-} - d^{+}\,. \tag{9-2}$$

It simply eliminates symmetrically distributed area portions, which reflect the model scattering error and therefore only keeps the model bias.

As already described in Sec. 8.2.1 with the results in Tab. 8-3 and also found by Voyles and Roy[327b], there is only a (slight) difference between AVM and model bias, when there is an overlap between the real and simulated EDFs/p-boxes. Now with the two portions of the AVM, it is clarified that this slight difference is a very good estimate for the model scattering error. In case there is no overlap and no epistemic uncertainty, the model bias estimate d_{bias} and the AVM are exactly the same as the mean error $\overline{d}$ (metric 3 in Tab. 4-3).

Furthermore, (9-2) can be used to estimate a *"corrected"*[326] simulated p-box as

$$\widetilde{\boldsymbol{\mathcal{F}}}_{\text{c}}(\zeta) = \widetilde{\boldsymbol{\mathcal{F}}}(\zeta - \widehat{d}_{\text{bias}}(\boldsymbol{\mathcal{F}},\widetilde{\boldsymbol{\mathcal{F}}})) = \widetilde{\boldsymbol{\mathcal{F}}}\Big(\zeta - (d^{-} - d^{+})\Big). \tag{9-3}$$

For better readability, the model bias estimate $\widehat{d}_{\text{bias}}$ will be written in the following without hat ($\widehat{\ }$) simply as d_{bias}.

9.2 Possible Modifications on the AVM

Voyles and Roy find that the AVM is limited in model bias elimination due to its symmetrical approach. Therefore, after introducing the two portions d^{+} and d^{-} and the directed model bias estimation (9-2), as described in the previous section, they construct a modified AVM (MAVM).[327c]

The MAVM is a two-dimensional metric like the Double Metric in (8-11) and calculated as

$$d_{\text{MAVM}}(\boldsymbol{\mathcal{F}},\widetilde{\boldsymbol{\mathcal{F}}}) = \Big(\big(-\frac{d^{-}-d^{+}}{2} - \mathcal{S}\frac{d^{-}+d^{+}}{2}\big),\big(-\frac{d^{-}-d^{+}}{2} + \mathcal{S}\frac{d^{-}+d^{+}}{2}\big)\Big)\,, \tag{9-4}$$

with a so-called safety factor $\mathcal{S}$, which depends on the number of available data n.

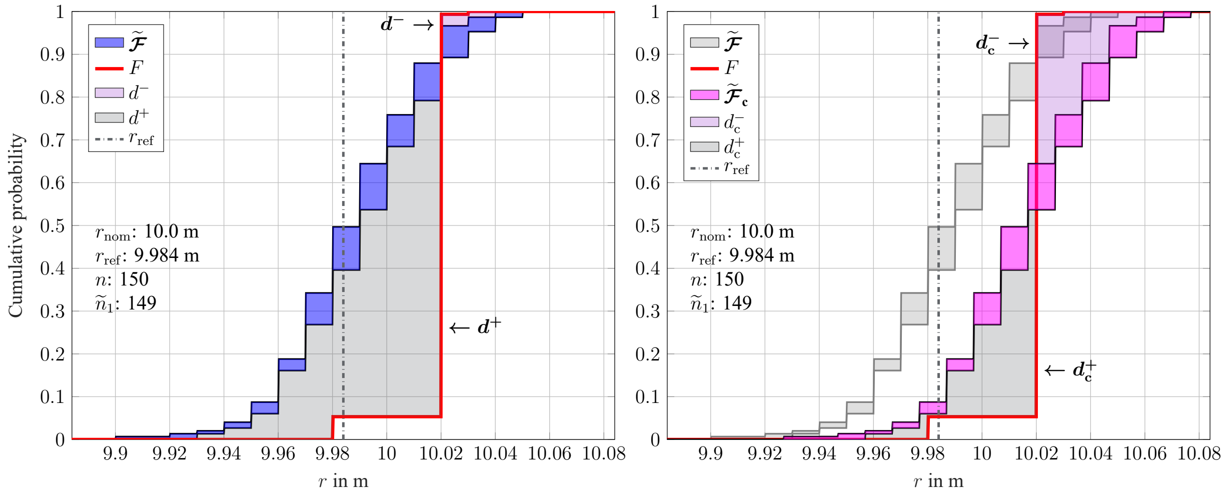
By aggregating the MAVM to the simulated EDF or p-box, the model form uncertainty is obtained, which is now directed towards the real data. While being less conservative than the original metric, the safety factor $\mathcal{S}$ is clearly not an objective parameter, but found by the authors in a subjective evaluation. Furthermore, the calculation of the MAVM is not intuitive anymore, which deprives the AVM of one of its key-features for the scope of this dissertation, where the metric is chosen to be used in a specification and possibly discussed with people without engineering background.

9.3 Corrected AVM (CAVM) and Double Validation Metric (DVM)

As concluded in Sec. 8.3, the AVM is the best metric candidate. However, it has the limitation of being insensitive to scattering errors, when there is no overlap between real and simulated EDFs/p-boxes. The MAVM from the previous section, as possible modification, does not bring any benefits needed, but complicates the computation, as already described. Therefore, a novel metric called **corrected AVM (CAVM)**, is introduced as an advancement of the original AVM to be able to validate model bias and model scattering error separately.

The CAVM is basically the AVM of the corrected simulated p-box $\widetilde{\boldsymbol{\mathcal{F}}}_{\text{c}}$ from (9-3). It therefore uses the model bias estimate d_{bias} from (9-2) and by that the two portions of the original AVM d^- and d^+ as its basis. Fig. 9-2 illustrates this two-step calculation. At first (a), d^+ and d^- as the areas where the simulated p-box is higher (+) or lower (-) than the real EDF are calculated. Afterwards, $d_{\text{bias}} = d^- - d^+$ is computed and $\widetilde{\boldsymbol{\mathcal{F}}}$ is shifted with this model bias estimate, resulting in the corrected p-box $\widetilde{\boldsymbol{\mathcal{F}}}_{\mathbf{c}}$ (b). Finally, the CAVM is the area between the real EDF F (in general: real p-box $\boldsymbol{\mathcal{F}}$) and $\widetilde{\boldsymbol{\mathcal{F}}}_{\mathbf{c}}$ with $d_{\text{c}}^- = d_{\text{c}}^+$. It is therefore formulated as

$$d_{\text{CAVM}}(\boldsymbol{\mathcal{F}}, \widetilde{\boldsymbol{\mathcal{F}}}) = d_{\text{AVM}}(\boldsymbol{\mathcal{F}}, \widetilde{\boldsymbol{\mathcal{F}}}_{\text{c}}) = d_{\text{c}}^- + d_{\text{c}}^+ . \tag{9-5}$$



(a) Calculation of d^+ and d^- for d_{bias} of $\widetilde{\boldsymbol{\mathcal{F}}}$

(b) CAVM calculation with corrected p-box $\widetilde{\boldsymbol{\mathcal{F}}}_{\mathbf{c}}$

Figure 9-2: Explanation of the two-step calculation of the CAVM

r_{nom} and r_{ref} are the nominal and reference range in Fig. 9-2. n and $\tilde{n}_1$ are the number of detections from real data and simulation sim1. F is the EDF from real data. $\widetilde{\boldsymbol{\mathcal{F}}}$ is the p-box from simulation. $\widetilde{\boldsymbol{\mathcal{F}}}_{\mathbf{c}}$ is the simulated p-box corrected with the estimated model bias d_{bias}. d^+ and d^- mark areas where the simulated p-box $\widetilde{\boldsymbol{\mathcal{F}}}$ is higher (+) or lower (-) than the real EDF F. d_{c}^+ and d_{c}^- mark areas where the corrected simulated p-box $\widetilde{\boldsymbol{\mathcal{F}}}_{\mathbf{c}}$ is higher (+) or lower (-) than the real EDF F.

This surprisingly simple concept for the CAVM follows the popular *KISS* principle from Kelly Johnson.[328] It is an intuitive and straightforward evolution of the AVM, corrected by the accurate model bias estimate. Therefore, the CAVM is an accurate approximation for the model scattering error.

To achieve a complete overview of a model's fidelity for sample validation and even model calibration, the author of this dissertation strongly advises to use the d_{CAVM} in combination with the bias error estimate d_{bias}. Consequently, all necessary information for an intuitive model assessment is provided with an universally applicable metric to be used when aleatory and epistemic uncertainty is present in the data. The proposed combination of CAVM and d_{bias} is therefore introduced as a two-dimensional metric called **double validation metric (DVM)**:

$$d_{\mathrm{DVM}}(\boldsymbol{\mathcal{F}}, \widetilde{\boldsymbol{\mathcal{F}}}) = \Big(\; d_{\mathrm{bias}}(\boldsymbol{\mathcal{F}}, \widetilde{\boldsymbol{\mathcal{F}}}), \;\; d_{\mathrm{CAVM}}(\boldsymbol{\mathcal{F}}, \widetilde{\boldsymbol{\mathcal{F}}}) \; \Big) \, . \tag{9-6}$$

According to Ferson et al., the DVM is a *"quasimetric"* [329], as it is non-symmetric, but satisfies the other criteria from Chap. 8. It fulfills the requirements for error prediction and uncertainty aggregation from Riedmaier, Danquah et al. being a *"correction in combination with tight uncertainty bounds."* [330] When several results for different samples are obtained using the DVM, they can be inter- and extrapolated with a confidence interval, as demonstrated by Roy and Balch[331] or Danquah, Riedmaier et al.[332], shown in Fig. 4-16 and Fig. 4-18 from Sec. 4.1.5.

[328] Rich, B. R.: Clarence Leonard (Kelly) Johnson (1995), p. 231.

[329] Ferson, S. et al.: Model validation and predictive capability (2008), p. 2416.

[330] Riedmaier, S. et al.: Unified Framework and Survey for Model VV&UQ (2020), p. 27.

[331] Roy, C. J.; Balch, M.: A holistic approach to uncertainty quantification (2012).

[332] Danquah, B. et al.: Statistical Validation Framework for Automotive Vehicle Simulations (2021).

10 Demonstration of the Novel DVM for VV&UQ of APSS Simulation

It is fundamentally different to show the capabilities of a metric instead of showing the capabilities of a simulation model. In the case of the latter, it is a challenge to assess only the APSS model fidelity without the influence of the simulated environment. In the case of the first, simulation serves as data generation as a whole and functional decomposition to find the origin of errors is not necessary.

Best case for metric assessment would be a kind of a reference metric, which is not present in the field of APSS simulation assessment, as discussed earlier. Consequently, only expert intuition can be used to deduce metric performance in this case. Such conclusions can be drawn best, if the metric is applied to non-epistemic EDFs first, where it can be compared to e.g. bias error $\overline{d}$ and scattering error d_{stdv}, which will be the metric assessment procedure in the following. Furthermore, it is beneficial, if the measurement and reference data and the simulation model are known in every detail, to draw conclusions about the applied metric. Additionally, model capabilities are assessed in an uncertain environment, where epistemic and aleatory uncertainties that must be tackled by the metric are known precisely, at best.

10.1 Specification of the Lidar Sensor System Simulation to Validate

As shown by Hadelli in his feasibility study, it is impossible to cover a complete validation study in a single thesis or dissertation and only an exemplary excerpt of it can be executed.[333] CEPRA based on PerCollECT presented earlier in Sec. 7.3 is the only systematic method to prioritize cause-effect chains for selecting the most relevant in a feasible exemplary sample validation. Here, the CEPRA excerpt in Tab. 7-1 is taken into account and the most relevant cause effect chain is chosen for validation. A lidar detection simulation is validated in this work, which means to end after detection sensing in Fig. 2-1 and validate on detection output. The selected phenomenon for demonstration is *False negative detection* with the cause-effect chain → *Not distinguishable from noise floor* → *Low received power from object & Distance between sensor and object* → *Reflection by object parts* and the causes *Materials of reflecting object parts, roughness of reflecting object parts, shapes of reflecting object parts, size of reflecting object parts, emitter wavelength, ...*

[333] Hadelli, A. A.: Messkampagne zur Parametrisierung und Validierung von Lidar-Sensor-Modellen (2020), pp. 46-54.

In line with the requirements for the lidar detection model in Tab. 7-2 and Annex B, two intuitive, short, and strict requirements are formulated. For simplicity, they are linearly range-dependent and independent from material/shape/orientation of hit object parts and read as:

1. *Lidar detection model shall output detections that only differ in range on average from real bias at maximum by 0.5 % of the reference range and in scattering error at maximum by 1.0 % of the reference range.*
2. *Lidar detection model shall contain range dependency of received power and therefore output lower EPW per detection at higher range. The relative descent of EPW should be the same ratio as in the real APSS with absolute deviation in relative descent lower than 1.0 % of the reference range. Absolute values of EPW should not differ in mean more than 1.0 m and the scattering should not differ more than 10.0 cm.*

Clearly, a single dissertation cannot cover complete APSS model specification and acceptance testing. In this regard, a single randomly chosen cause-effect chain and only two short requirements derived from it are only a very small excerpt of a full CEPRA and the long list of requirements and acceptance tests that is to be expected later on in real industrial projects. Nevertheless, this exemplary excerpt already allows to demonstrate the holistic approach of this dissertation and the application of the elaborated novel metric, called DVM.

10.1.1 Description of the Validated Lidar Detection Simulation

The validated APSS simulation in this dissertation is the reflection-based lidar detection model developed by the author. It is built with the modular framework described in Sec. 7.1 and has therefore the same architecture as the reflection-based lidar object model[334], but uses only a subset of strategies. Instead of the Velodyne VLP32 profile shown in Fig. 2-5 that is reproducing a 360° scanning lidar as depicted in Fig. 3-5, now the Ibeo LUX profile is used for the simulation. The model implements a ray casting approach where the nominal beam divergence is simulated with super-sampling, as described in Sec. 3.1.1.

As it is a reflection-based lidar model, the range dependency of the received power, as described in Sec. 3.1.2 and the signal interaction with hit objects in Sec. 3.1.3 is provided by the environment simulation tool, where the ray casting is performed. Sensor model and environment simulation tool are connected via FMI and the beam pattern is configured by the sensor model via OSI. The main part of the lidar detection model consists of reproducing the signal per beam from spatial super-sampled ray casting results, thresholding this signal (Sec. 3.1.7), finding one or more peaks in this thresholded signal per beam (Fig. 3-2) that form the detections, and calculating the correct ranges and intensities/EPWs for each detection. Signal interaction within the channel (Sec. 3.1.4), temporal lidar behavior (Sec. 3.1.5), and detailed receiver effects Sec. 3.1.6 are not implemented and therefore not validated in this dissertation.

[334] Rosenberger, P. et al.: Reflection Based Lidar Object Model (2022).

10.1.2 Description of the Simulated Lidar Sensor System

The real lidar sensor system simulated and used for measurement data collection is the Ibeo LUX 2010, a scanning lidar and the predecessor of the Valeo SCALA, which is pictured in Fig. 3-4. The sensor's specification is described by Hadelli[335a] himself, in previous work of the author[336,337], and by Tamm-Morschel[338a]. It has four layers of elevation angles and an angular range for the two lower reception angles (1, 2) of about $-50°$ to $50°$ and for the upper layers (3, 4) of about $-60°$ to $35°$. The two lower layers are shifted by half the angular distance in azimuth.

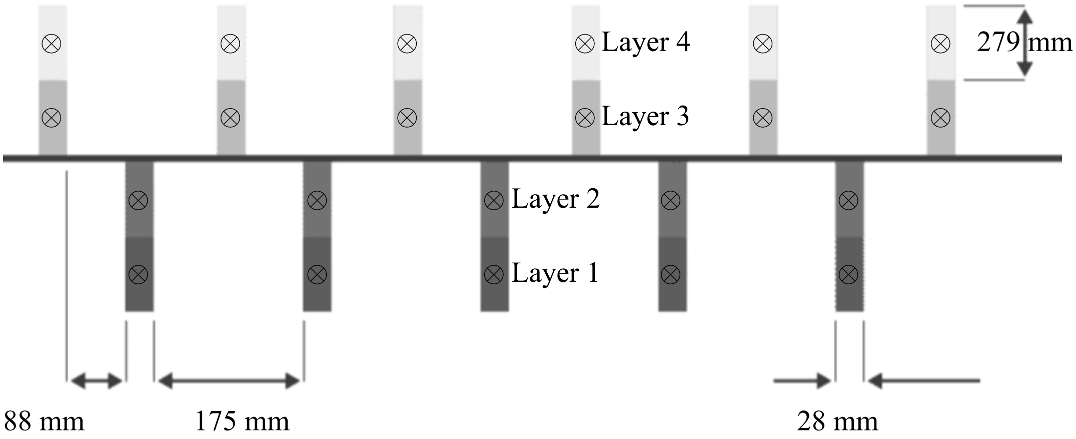


Figure 10-1: Nominal beam size at 20 m of the Ibeo LUX 2010 lidar sensor with marked beam centers points (⊗) from former work of the author[337]

Each beam has a divergence of $0.8°$ vertically and $0.08°$ horizontally, as illustrated in Fig. 10-1, which shows the nominal beam size at $r = 20\,\mathrm{m}$. Tamm-Morschel documented the actual illumination beam pattern of the Ibeo LUX with a modified consumer photo camera, where he removed the infrared filter[338c], as shown in Fig. 10-2. Comparing this photograph with the nominal beam pattern (Fig. 10-1) already reveals several differences that an accurate lidar detection model should reproduce.

However, available automotive scanning lidar sensor systems do not measure an incident angle of the light received from the shown imperfect illumination of the scene. They are simply clipping the detections that are identified from the back scattered signal to the nominal beam centers, as discussed in Sec. 2.3. Therefore, the lidar detection model mimics this behavior, as illustrated in Fig. 3-8. Consequently, the true illuminated area on the object has to be determined first, when measurements are performed for validation, to diminish this source of epistemic uncertainty. Fortunately, Hadelli has ensured to illuminate the targets as intended in his experiments.[335b]

[335] Hadelli, A. A.: Messkampagne zur Parametrisierung und Validierung von Lidar-Sensor-Modellen (2020). a: pp. 13-16.; b: p. 63.; c: pp. 66-67.; d: 83-85.

[336] Rosenberger, P. et al.: Analysis of Real World Sensor Behavior (2018).

[337] Rosenberger, P. et al.: Sequential lidar sensor system simulation (2020), p. 192.

[338] Tamm-Morschel, J. F.: Erweiterung eines Lidar-Sensormodells (2019). a: pp. 33-35.; b: p. 44.

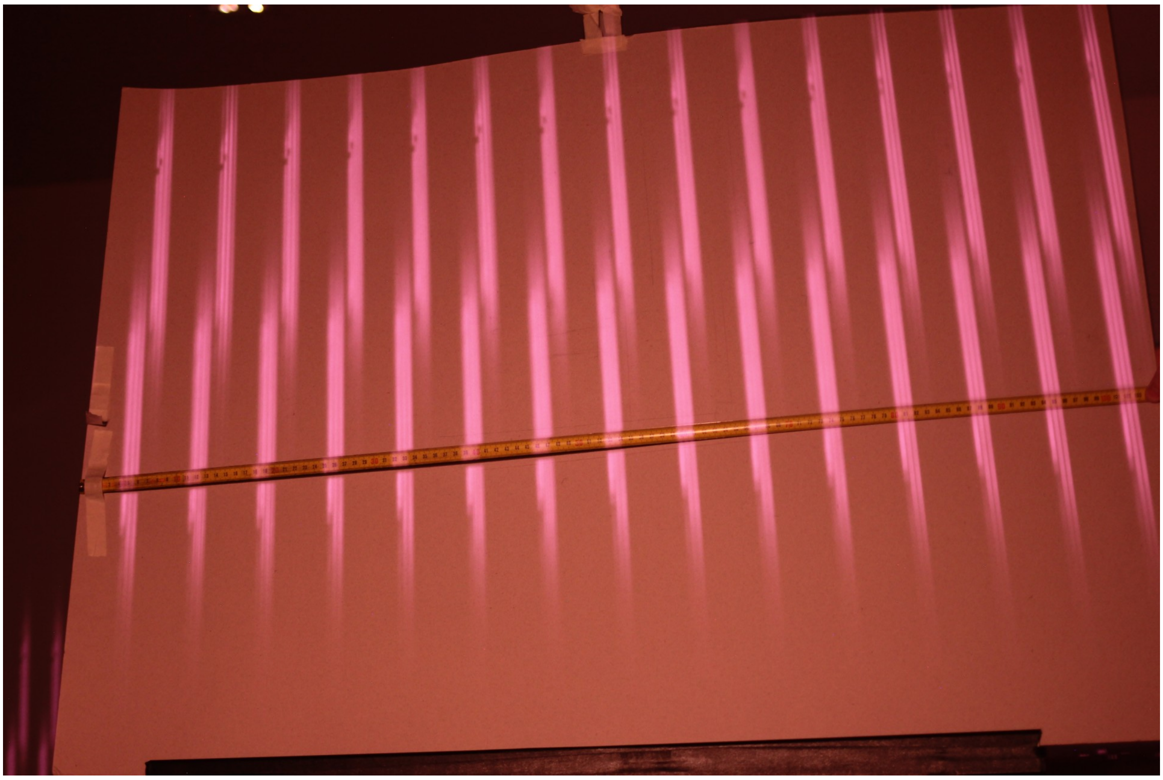


Figure 10-2: Photograph of illumination of a target by the Ibeo LUX lidar at 11 m with an exposure time of 8 s from Tamm-Morschel[338c]

10.2 DVM for Single Beam VV&UQ of APSS Simulation

Measurement data from Hadelli, more precisely from his fourth test setup[335c] is used for the single beam sample validation. Single beam comparison is a good starting point to evaluate the basic capabilities of a sensor system and its simulation, which is the case for range dependency investigations. Hadelli's experiment design aims to evaluate the expected tendency of EPW/intensity to fall with greater range, which perfectly serves as an acceptance test for the selected requirement from Sec. 10.1.[335d] He set up the experiments in a hangar at the August Euler Airfield near Darmstadt to eliminate weather and other atmospheric effects and sun glare, which is beneficial for validation experiments where validated effects should occur as isolated as possible. All experiment setups and all results were reviewed by the author of this dissertation as his supervisor.

In his thesis, Hadelli documented in detail the entire experimental design and the collection of measurement and reference data. His full report is publicly available (at least in German) for possible repeatability and reproducibility. The separation of personnel between the experimenter and the data evaluator/model validator is the general case in industry, so the practicality of the presented method and novel metric is demonstrated.

As a first sample, the central lidar beam is selected as single combination of azimuth and elevation to validate the lidar detection simulation. The respective range resolution for each dimension is determined by the sensor's counter resolution and the used digits and arithmetic precision of those digital output values. The central beams of a spherical measurement principle and especially 0° in azimuth is the same as $0\,\mathrm{m}$ in y-direction of a Cartesian coordinate frame with same origin and the range is equivalent to the distance in Cartesian x-direction. Furthermore, when objects are placed in front of the sensor that would be parallel to the y-axis with constant distance in x-direction produce non-constant range measurements that grow with the absolute azimuth angle. Therefore, as range matters for intensity, as shown in Sec. 3.1.2, the central azimuth angle is preferred for range-dependency evaluations.

Hadelli's experimental setup with gray cardboard as lidar target located centrally in front of the lidar sensor is shown in Fig. 10-3. Hadelli reports that gray cardboard is a good reference for a dirty car surface, as the measured EPW *"standard deviation is almost identical and only the mean value shows that the gray cardboard has a slightly higher echo pulse width of* $4\,\%$*"*[339a] compared to the dirty car surface. To obtain different validation samples, the measurements are performed with different ranges between sensor and target that are listed in Tab. 10-1. Nominally, the experiments start at $r_\mathrm{nom} = 10\,\mathrm{m}$ and with steps of $2.5\,\mathrm{m}$ end at $20\,\mathrm{m}$, which gives in total five samples. The reference ranges r_ref in Tab. 10-1 are provided reasonably in mm by Hadelli, while he reports an accuracy of the reference laser range measurement device of $1.5\,\mathrm{mm}$.

Table 10-1: Central nominal and reference ranges in experiments from Hadelli[339c]

#	1	2	3	4	5
r_nom	$10.000\,\mathrm{m}$	$12.500\,\mathrm{m}$	$15.000\,\mathrm{m}$	$17.500\,\mathrm{m}$	$20.000\,\mathrm{m}$
r_ref	$9.984\,\mathrm{m}$	$12.436\,\mathrm{m}$	$14.995\,\mathrm{m}$	$17.410\,\mathrm{m}$	$19.865\,\mathrm{m}$

In this exemplary sample validation study, the orientation of the cardboard target with respect to the lidar sensor's front-end is assumed to be perfectly parallel. With the photographs from the experimental setup in Fig. 10-3 at hand that indicate only small deviations from that assumption and for the investigated central lidar beam and the relatively small beam size at the selected nominal ranges, resulting from the known beam divergence, it is a reasonable assumption. It simplifies the sample validation study a lot, as it results in less parameters, less epistemic uncertainties to be considered for re-simulation, and finally less simulations to be performed.

Noise is present and responsible for most of the aleatory uncertainty in reference and measurement data. The SNR influences detection existence, but also location (range), and the intensity/EPW values, as can be seen in the cause-effect chain visualization PerCollECT.[340] Consequently, this effect can never be eliminated by experiment design. As mentioned in the previous section, due to

[339] Hadelli, A. A.: Messkampagne zur Parametrisierung und Validierung von Lidar-Sensor-Modellen (2020). a: p. 82.; b: pp. 74-80; c: pp. 46-54.

[340] Linnhoff, C. et al.: PerCollECT - LidarLimbs (2022).

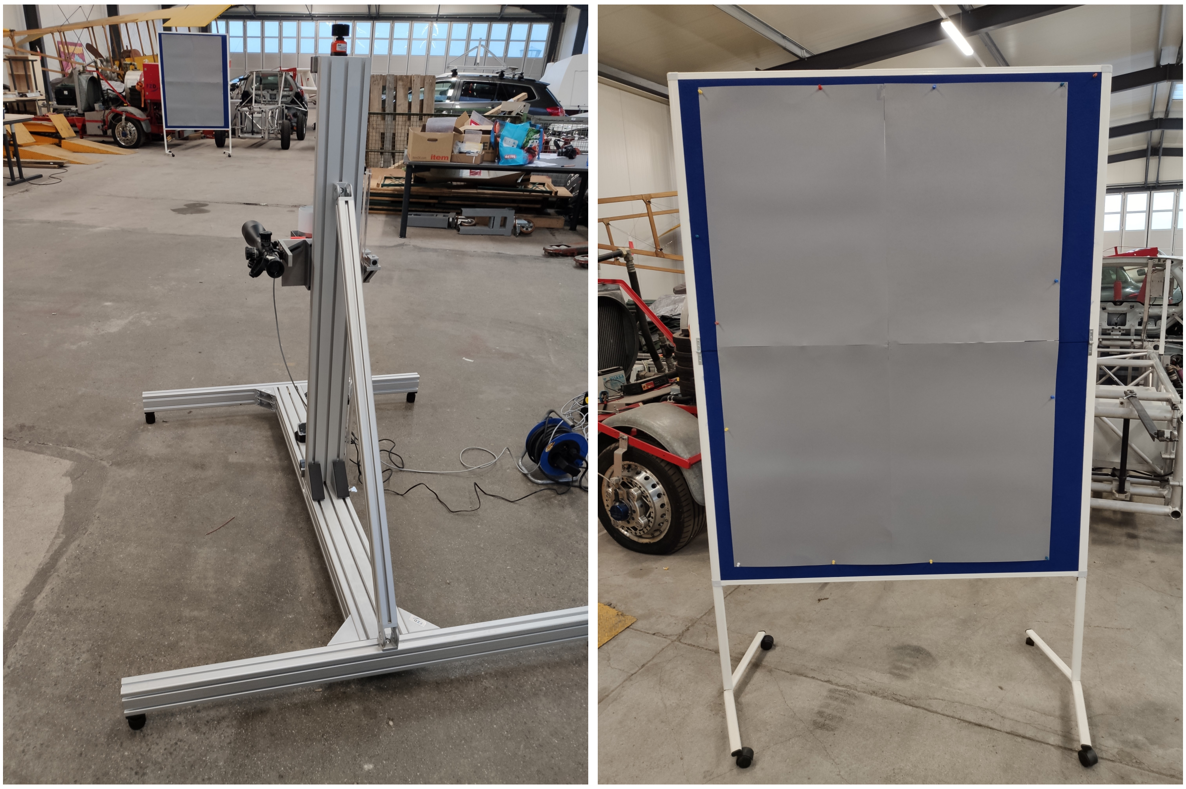


Figure 10-3: Experimental setup (left) and movable target covered with gray cardboard (right) from Hadelli[339c]

performing the validation experiments indoors, no other influences on the SNR should be present in the data except a global background illumination through the hangar's windows, the relative humidity of the air and the temperature. However, both influences have been investigated in- and outdoors by Hadelli and no measurable correlation on the lidar detections is reported.[339b]

10.2.1 DVM for Single Beam Validation of APSS Simulation

For single beam sample validation, only a central beam fits the reference range best, due to a higher range for higher azimuth angles by the cosine relation. As the simulated sensor has an alternating beam pattern, only the two upper layers of the four total layers actually illuminate the scene at nominally $0.0°$ azimuth. To obtain only a minimal vertical incident angle at the target, the lower layer of the two upper layers at nominal elevation $\theta = 0.8°$ is selected. Nonetheless, both real and simulated detections are facing the same elevation angle, so no deviations should arise from it. Three of the five sample ranges are taken for validation at first, the other two samples are reserved to evaluate the error prediction and uncertainty aggregation, later on. The selected nominal ranges r_{nom} are $10\,\mathrm{m}$, $15\,\mathrm{m}$, and $20\,\mathrm{m}$, while the reference ranges r_{ref} measured with the reference laser range measurement device slightly differ, as can be read in Tab. 10-1.

A virtual scene is generated with a three-dimensional model of a gray carton box placed at the respective ranges, big enough to cover the investigated central beam completely to obtain the desired data. In this example, all real experiments were repeated three times shortly after another leaving everything as is and keeping the lidar sensor on, with a measurement duration of around 50 frames each. Therefore, in total around $n \approx 150$ detections are obtained that are justifiably combined into a single EDF F from the measurements. Each simulation was run for a similar number of frames to get a comparable amount of detections $\widetilde{n} \approx 150$. The numbers of detections are plotted into each plot in Fig. 10-4 and also in all other plots, where relevant.

To reflect the epistemic uncertainty in reference measurements of the range between sensor and target, each simulation is performed with the exact reference range ($\mathrm{sim1}$, black) as measured and a variation of $\pm 2.0\,\mathrm{mm}$ in two more simulations ($\mathrm{sim2/sim3}$, blue). The relatively small epistemic uncertainty is only reasoned in the reported accuracy of the reference measurement device of $\pm 1.5\,\mathrm{mm}$ and a small tolerance on top for further uncertainty sources, while e.g. epistemic uncertainty in materials, etc. is ignored for simplification in this exemplary validation study.

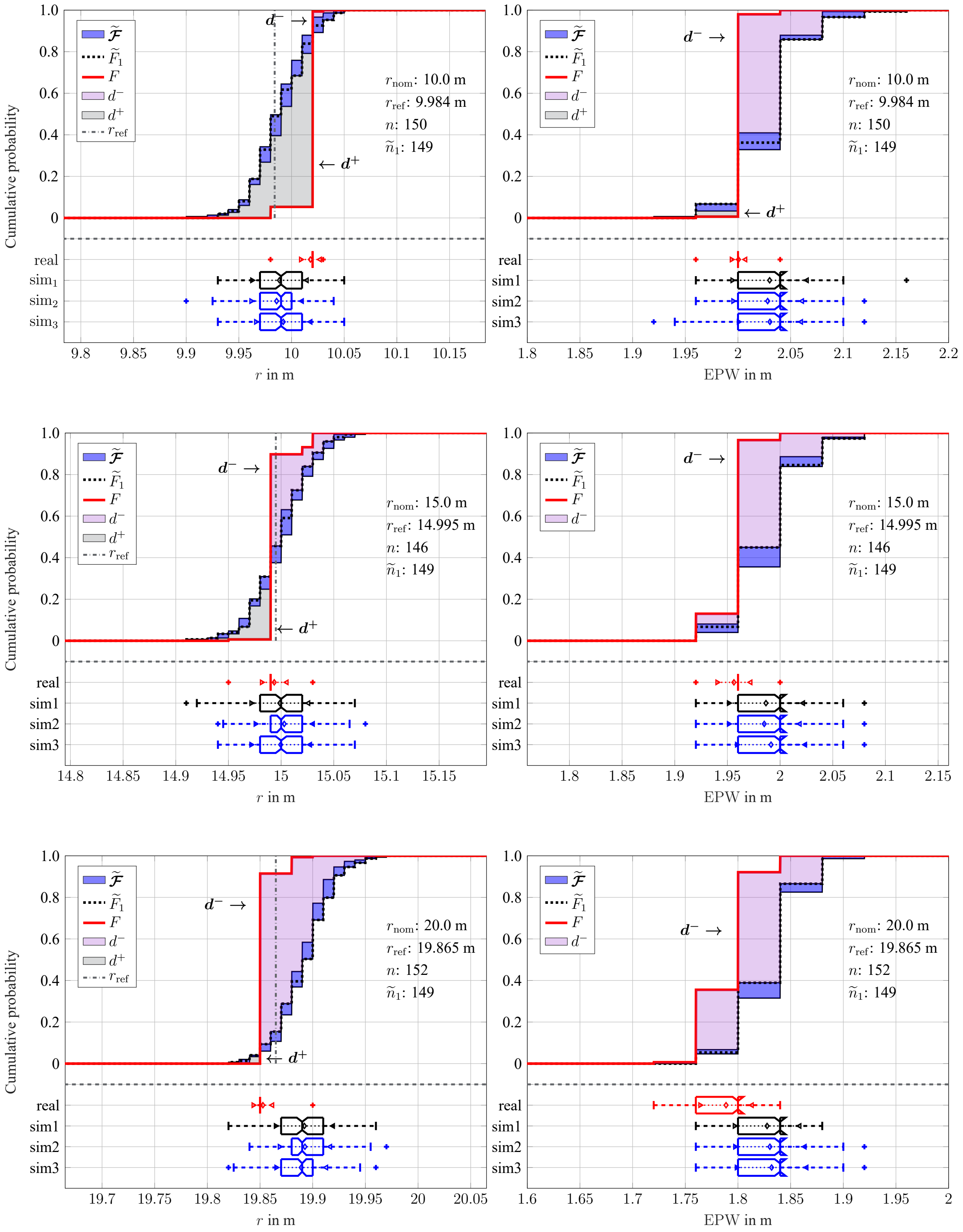
Due to noise on range and EPW measurements in reality and the respective noise models in simulation[341], aleatory uncertainty is present in both data. In this regard, the so called *"E-outer"* uncertainty propagation is performed, according to Roy and Balch[342], and leads to the blue p-boxes, while the real EDF is red. It should be noted that the obtained left and right side of each p-box is not necessarily an EDF from the same simulation, but the most left or right data at each quantile. Fig. 10-4 shows the resulting plots of the cumulated probabilities. On the left (a) is the simulated and real range r and on the right the simulated and real EPW (b). A first observation on both inspected measurands is that each step size or simulated counter resolution is fine. For range measurements this is $1.0\,\mathrm{cm}$, as visible in the middle plot, and for the EPW it is $4.0\,\mathrm{cm}$. Consequently, the simulation passes this possible falsification by simple visual inspection.

To support the initial visual inspection for possible intuitive falsification of the simulation, each sample validation plot should contain as much information as possible. Besides the respective detection numbers n for real data and $\widetilde{n}_1$ for simulation sim1 at r_{ref}, in Fig. 10-4 box plots for each real and simulated experiment are included at the bottom of each plot. These box plots show the median as a vertical line between the notches that depict its $95\,\%$ confidence interval. Thee mean is plotted as a diamond marker and the standard deviation to each side is visible as a triangle marker. Additionally, the $25\,\%$ and $75\,\%$ quantiles are depicted with the actual boxes and the whiskers to each side have a length of 1.5 times of the interquartile range at maximum, while plus sign markers represent outliers that are further away.

A closer look on the 6 plots in Fig. 10-4 gives the impression that the counter resolution for the simulated range is $1.0\,\mathrm{cm}$ and seems to small at first. Only the small step in the EDF from real data in the middle left plot for $r_{\mathrm{nom}} = 15.0\,\mathrm{m}$ shows that the real data has the same counter resolution as simulated, but the scattering happens in larger steps in most cases. The step size for EPW is correctly simulated with $4.0\,\mathrm{cm}$ and happens regularly in real data too.

[341] Rosenberger, P. et al.: Sequential lidar sensor system simulation (2020), p. 192.

[342] Roy, C. J.; Balch, M.: A holistic approach to uncertainty quantification (2012), p. 370.



(a) Simulated vs. real range r of central lidar detections

(b) Simulated vs. real EPW of central lidar detections

Figure 10-4: Comparison of simulated and measured central lidar detections for different nominal ranges r_{nom}. n and $\widetilde{n}_1$ are the number of detections from real data and simulation sim1. r_{ref} is the reference range. The nominal ranges are $r_{\text{nom}} = 10.0\,\text{m}$ (top), $r_{\text{nom}} = 15.0\,\text{m}$ (middle), and $r_{\text{nom}} = 20.0\,\text{m}$ (bottom). $\widetilde{\mathcal{F}}$ is the p-box from simulation with $\widetilde{F}_1$ being the EDF from sim$_1$ at r_{ref}. F is the EDF from real data. d^+ and d^- mark the areas where the simulated p-box is higher (+) or lower (-) than the real EDF.

The objective metric results are listed in Tab. 10-2. While the corrected simulated p-boxes $\widetilde{\mathcal{F}}_\mathbf{c}$ are not plotted in Fig. 10-4 for readability, they are calculated according to (9-3) from Sec. 9.1. With the respective $\widetilde{\mathcal{F}}_\mathbf{c}$ at hand, each CAVM is computed as described in (9-5) from Sec. 9.3 and illustrated in Fig. 9-2.

Table 10-2: DVM results for single beam sample validation with measurements from Hadelli

ζ	r_nom	F vs. $\widetilde{F}_1$		$d_\mathbf{DVM}(F, \widetilde{F}_1)$		$d_\mathbf{DVM}(F, \widetilde{\mathcal{F}})$	
		$\overline{d}$	$d_\mathbf{stdv}$	$d_\mathbf{bias}$	$d_\mathbf{CAVM}$	$d_\mathbf{bias}$	$d_\mathbf{CAVM}$
r	$10.00\,\mathrm{m}$	$0.029\,\mathrm{m}$	$0.017\,\mathrm{m}$	$-0.029\,\mathrm{m}$	$0.019\,\mathrm{m}$	$-0.027\,\mathrm{m}$	$0.015\,\mathrm{m}$
	$15.00\,\mathrm{m}$	$0.005\,\mathrm{m}$	$0.015\,\mathrm{m}$	$0.005\,\mathrm{m}$	$0.018\,\mathrm{m}$	$0.006\,\mathrm{m}$	$0.015\,\mathrm{m}$
	$20.00\,\mathrm{m}$	$0.039\,\mathrm{m}$	$0.018\,\mathrm{m}$	$0.039\,\mathrm{m}$	$0.019\,\mathrm{m}$	$0.036\,\mathrm{m}$	$0.015\,\mathrm{m}$
EPW	$10.00\,\mathrm{m}$	$0.030\,\mathrm{m}$	$0.029\,\mathrm{m}$	$0.030\,\mathrm{m}$	$0.026\,\mathrm{m}$	$0.027\,\mathrm{m}$	$0.022\,\mathrm{m}$
	$15.00\,\mathrm{m}$	$0.030\,\mathrm{m}$	$0.019\,\mathrm{m}$	$0.030\,\mathrm{m}$	$0.023\,\mathrm{m}$	$0.028\,\mathrm{m}$	$0.019\,\mathrm{m}$
	$20.00\,\mathrm{m}$	$0.039\,\mathrm{m}$	$0.006\,\mathrm{m}$	$0.039\,\mathrm{m}$	$0.006\,\mathrm{m}$	$0.038\,\mathrm{m}$	$0.005\,\mathrm{m}$

It is worth mentioning again that the goal here is to evaluate the metrics rather than the simulation. Therefore, the novel DVM is first applied only to simulation $\widetilde{F}_1$ with the target at exactly the reference range r_ref to be able to compare it with the ordinary bias and scattering error $\overline{d}$ and d_stdv that only can handle CDFs/EDFs. Here, the results are almost identical except that the estimated bias error as first part of the DVM is signed to enable to correct the bias error. Consequently, the novel DVM passes this practical test. Finally, the validation results considering the epistemic uncertainty in each p-box $\widetilde{\mathcal{F}}$ are provided. As can be seen in Tab. 10-2, the results considering the p-box from simulation instead of the singular EDF leads to equal or slightly better results. This is the expected behavior and reflects that considering epistemic model input uncertainty from reference data collection is fairness to the simulation for its VV&UQ.

In this case, the simulation model is sample valid, as its absolute biases for range r are always less than $0.5\,\%$ of r_ref, the associated scattering errors are less than $1.0\,\%$ of r_ref, and the relative errors and the orders of magnitude of the simulated EPW mean and scattering are as required, too.

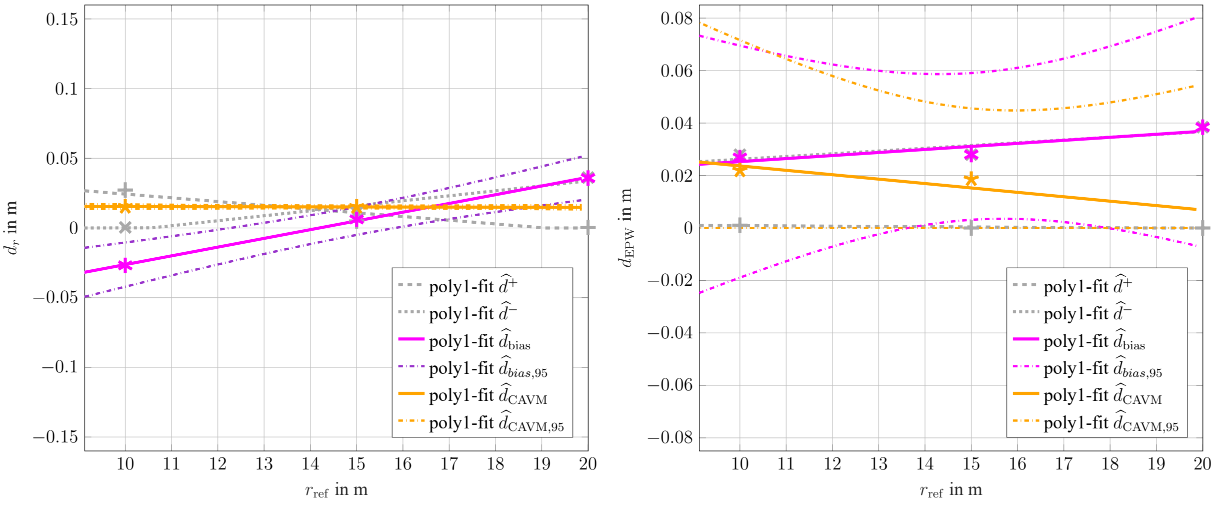
However, an important finding of the here presented small sample validation study is that all simulations tend to have higher scatter than in reality, which is very rare for most simulations of single uncertain numbers, e.g. for vehicle dynamics simulations discussed by Viehof.[343] The here presented metric for scattering error is unsigned and therefore does not provide information on whether the modeled scatter is too high or too low, which is also not initially of interest for just validating a specified performance. Nevertheless, it would be of great benefit for model calibration if this information could be obtained, but this is beyond the scope of this dissertation.

[343] Viehof, M.: Objektive Qualitätsbewertung von Fahrdynamiksimulationen (2018).

10.2.2 DVM for Single Beam and Error Prediction

As discussed, model credibility can only be assessed after error prediction and uncertainty aggregation for its application. To evaluate the previous predictions, the two remaining sample measurements from Hadelli at $12.5\,\mathrm{m}$ and $17.5\,\mathrm{m}$ serve as the application domain in this exemplary model credibility assessment. The linear interpolations between the results for the two area portions $\widehat{d}^+$ and $\widehat{d}^-$ and for the model bias $\widehat{d}_\text{bias}$ including $95\,\%$ confidence bounds $\widehat{d}_\text{bias,95}$ are calculated. Subsequently, the results for the CAVM are also linearly interpolated to obtain $\widehat{d}_\text{CAVM}$ and $\widehat{d}_\text{CAVM,95}$.

All interpolations are performed in this case with a linear polynomial curve with the poly-1 fit provided by The MathWorks, Inc.[344] and plotted in Fig. 10-5. Both measurands in this example, range r (left, a) and EPW (right, b), are treated equally and stay in acceptable value ranges with respect to the requirements from Sec. 10.1. The distance metric $\widehat{d}_\text{CAVM}$ and its confidence interval are bounded to not fall below $0.0\,\mathrm{m}$ during inter- and extrapolation.

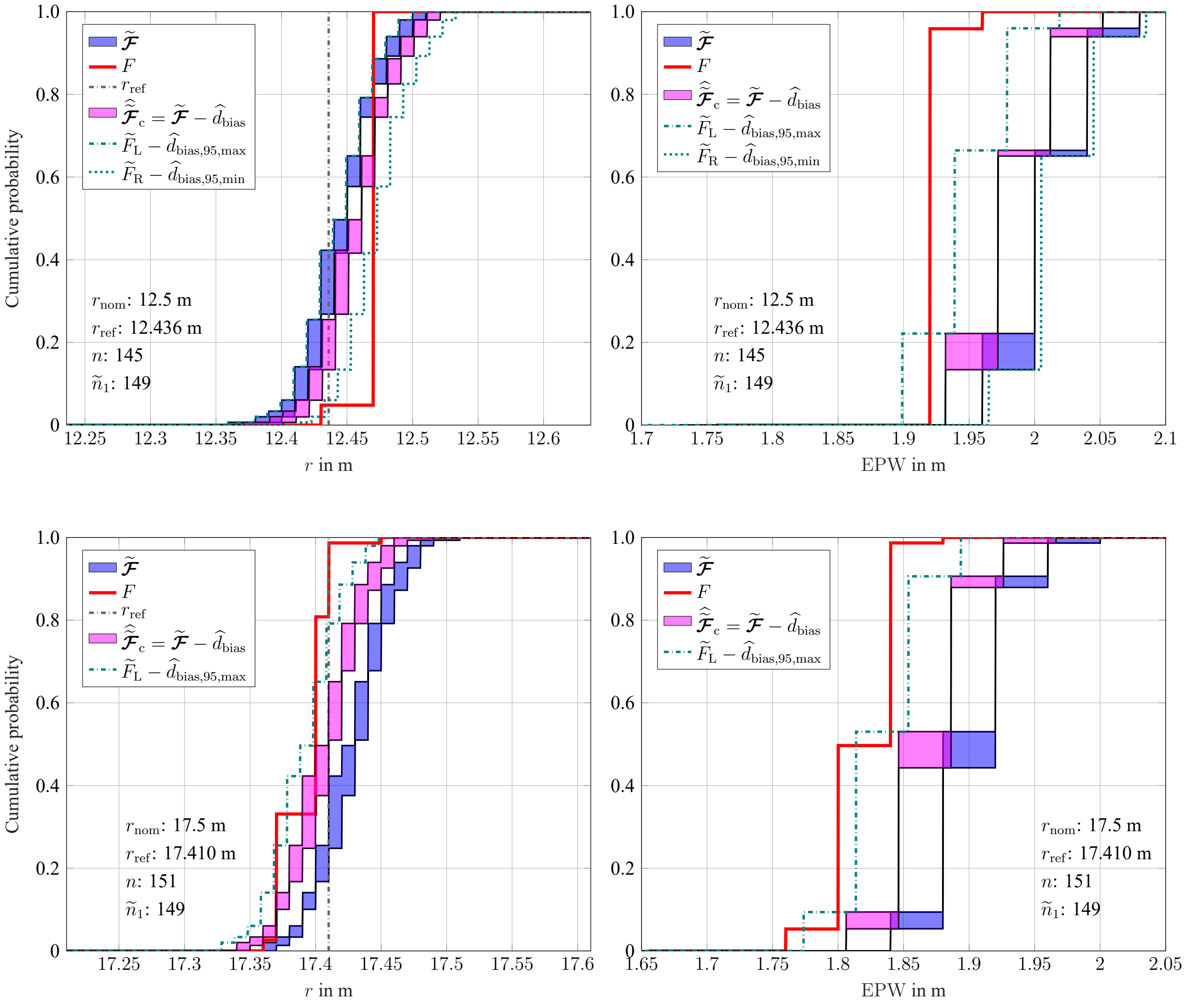


(a) DVM interpolation for range r of central detections (b) DVM interpolation for EPW of central detections

Figure 10-5: Interpolation of DVM results from single beam sample validation with +: d^+, ×: d^-, ∗: d_bias, ⋆: d_CAVM

Uncertainty aggregation is performed with these predicted curves and confidence intervals to validate the error predictions with the two beforehand reserved data sets for $r_\text{nom} = 12.5\,\mathrm{m}$ and $r_\text{nom} = 17.5\,\mathrm{m}$, as plotted in Fig. 10-6. At first, the simulated p-box $\widetilde{\mathcal{F}}$ (blue) and the EDF F from the real data (red) at the respective reference ranges r_ref are drawn. Additionally, the corrected p-box $\widehat{\widetilde{\mathcal{F}}}_\text{c}$ shifted by the predicted model bias $\widehat{d}_\text{bias}$ is given together with its $95\,\%$ confidence interval.

[344] The MathWorks, Inc.: Fit curve or surface to data (2022).



(a) Range r of detections from central lidar beam

(b) EPW of detections from central lidar beam

Figure 10-6: Validation of model bias prediction for a single beam with DVM

The two left plots show that in case of the simulated range, model bias predictions are accurate, as the predicted corrected p-box from simulation $\widehat{\widetilde{\boldsymbol{\mathcal{F}}}}_{\mathbf{c}}$ is very close (top left, $r_{\mathrm{nom}} = 12.5\,\mathrm{m}$) or overlaps at the median (bottom left, $r_{\mathrm{nom}} = 17.5\,\mathrm{m}$) with the real EDF F. In this case, the predicted model bias is very small, therefore, its $95\,\%$ confidence interval enlarges the predicted overall uncertainty only to a small extent, while it is still a convenient recommendation to add it too.

When investigating the scattering error of the simulated ranges in Fig. 10-6, the scatter is obviously a tiny bit too high in simulation, as already found for the three different sample ranges during sample validation shown in Fig. 10-4 from top to bottom. There, the box plots at the bottom of each plot stress the existing differences in scatter of real data and the data from each simulation run $\mathrm{sim1}$, $\mathrm{sim2}$ $\mathrm{sim3}$. However, the scattering error of the simulated ranges in absolute numbers is always around or lower than $1.5\,\mathrm{cm}$, which is significantly less than required.

The two plots on the right show the predicted model bias for the simulated EPW. It can be seen that the EDF from real data F sometimes falls even outside the more conservative uncertainty predictions with $95\,\%$ confidence. The reason for these too optimistic predictions for model bias is that in case of the evaluated EPW simulations from the investigated samples the model bias is very small, leading to too small model bias estimates in this case. Visual inspection indicates that the scattering error of the simulated EPWs is almost correct, being just a little too high.

The predicted scattering error is not part of the model uncertainty aggregation here and therefore not contained in Fig. 10-6. At first, this seems to be contrary to the reference literature, e.g. by Roy and Balch[345]. However, it is justified with the difference between the modeling of physical correlations leading to single uncertain numbers, as e.g. in vehicle dynamics simulation. In case of APSS simulation, the scattering error of the real sensor is an essential part to simulate. Nonetheless, simply adding the predicted scattering error on both sides of the shifted simulated p-box would mix up the evaluation of model bias and scattering error again and should be avoided consequently. In contrast to such plotting and visual inspection of EDFs and aggregated uncertainties, comparing actual and predicted results in a table is more clear and objective.

Therefore, after the initial visual inspections of the results for the predicted model bias $\widehat{d}_{\text{bias}}$, as first part of the DVM, the predictions are objectively compared to the actual sample validation results for both nominal ranges. As listed in Tab. 10-3, the two beforehand reserved data sets for $r_{\text{nom}} = 12.5\,\text{m}$ and $r_{\text{nom}} = 17.5\,\text{m}$ are used.

Table 10-3: DVM predictions and actual values for single beam sample VV&UQ of simulated lidar detections The confidence bounds $\widehat{d}_{\textbf{bias,95}}$ and $\widehat{d}_{\textbf{CAVM,95}}$ are the ones with highest absolute value.

ζ	r_{nom}	**Predicted DVM**				**Actual DVM**	
		$\widehat{d}_{\textbf{bias}}$	$\widehat{d}_{\textbf{bias,95}}$	$\widehat{d}_{\textbf{CAVM}}$	$\widehat{d}_{\textbf{CAVM,95}}$	$d_{\textbf{bias}}$	$d_{\textbf{CAVM}}$
r	$12.50\,\text{m}$	$-0.011\,\text{m}$	$-0.023\,\text{m}$	$0.015\,\text{m}$	$0.016\,\text{m}$	$-0.018\,\text{m}$	$0.015\,\text{m}$
	$17.50\,\text{m}$	$0.020\,\text{m}$	$0.032\,\text{m}$	$0.015\,\text{m}$	$0.016\,\text{m}$	$0.031\,\text{m}$	$0.007\,\text{m}$
EPW	$12.50\,\text{m}$	$0.028\,\text{m}$	$0.061\,\text{m}$	$0.020\,\text{m}$	$0.055\,\text{m}$	$0.085\,\text{m}$	$0.020\,\text{m}$
	$17.50\,\text{m}$	$0.034\,\text{m}$	$0.066\,\text{m}$	$0.011\,\text{m}$	$0.046\,\text{m}$	$0.080\,\text{m}$	$0.003\,\text{m}$

Tab. 10-3 enables an objective comparison of the predictions, their $95\,\%$ confidence bounds, and the actual values for both elements of the DVM. A first finding from this relatively small sample validation study and the here presented validation of the subsequent DVM result prediction is that in case of the predicted model bias $\widehat{d}_{\text{bias}}$ the more conservative $\widehat{d}_{\text{bias,95}}$ is the more accurate choice for both the range r and the EPW. Furthermore, it shows that the predicted CAVM is almost perfect in case of the simulated ranges, but too small in case of the simulated EPWs. It becomes evident that reserving some of the sample measurements from each sample validation study for validating the DVM result predictions is crucial and necessary for the targeted model credibility.

[345] Roy, C. J.; Balch, M.: A holistic approach to uncertainty quantification (2012).

Later, in practical application during actual VV&UQ studies, such (possibly automated) validation of the error predictions supports the model credibility in any case, but is expected to be performed only to a very limited extent due to the additional effort of extra experiments only for this purpose. Nonetheless, the fit options (linear, quadratic, etc.) for the DVM result inter- and extrapolation should be evaluated. During the investigations for this dissertation, higher order polynomials seem to lead to narrower/broader intervals compared to linear regression where many/little data is available, i.e. more conservative to estimate when the coverage of the parameter space is limited.

10.3 DVM for Full Scan Validation, and Error Prediction of APSS Simulation

The next consecutive step when validating a lidar sensor simulation after single beam investigation is a full scan evaluation. Therefore, a target is needed that is wide enough to cover the whole angular range of the full scan in reasonable distances. For the selected lidar sensor with $110°$ azimuth coverage, a large building or wall is sufficient, while for other sensors possibly scanning $360°$ cylindrical artificial targets or indoor areas with different sizes would be required. Alternatively, also half scan or thirds scan measurements could be combined to a full scan, while it would be necessary to investigate possible side effects. In the presented example, an airfield hangar's wall serves as the target and the sensor is placed in front at different ranges, which are listed in Tab. 10-4. Similar to the single beam validation in Sec. 10.2, four samples (no. 1, 3, 4, and 5) are taken as validation samples and two other samples (no. 2 and 6) are reserved for validation of the predicted DVM results.

Table 10-4: Central nominal and reference ranges in full scan experiments

#	1	2	3	4	5	6
r_{nom}	$1.000\,\text{m}$	$1.500\,\text{m}$	$2.000\,\text{m}$	$3.000\,\text{m}$	$5.000\,\text{m}$	$9.000\,\text{m}$
r_{ref}	$0.974\,\text{m}$	$1.478\,\text{m}$	$2.014\,\text{m}$	$2.997\,\text{m}$	$4.988\,\text{m}$	$8.938\,\text{m}$

The same material is chosen in simulation as for the single beam validation with the assumption that the wall has a similar reflectivity as the gray cardboard. While this assumption is not proven by reference measurements for the reflectivity of both materials here, this simulation serves as exemplary SuT and incorrect modeling of reflectivities supports demonstrating the capabilities to measure the model mean and scattering error. The geometry of the virtual object is enlarged to cover the simulated full scan. The reference measurement device in the case of the used real lidar sensor measurements has lower accuracy than in the indoor experiments from Hadelli before. To reflect this epistemic uncertainty exemplary as one of many more possible uncertainty sources, two simulations in this full scan evaluation are performed with a different range to the wall of $\pm 5\,\text{mm}$ besides the first simulation at exact reference distance (sim1, black).

With the described measurements and simulations at hand, the EDF F from real data and the p-box from all simulation runs are calculated as depicted in Fig. 10-8 for x and Fig. 10-9 for EPW. The reason for taking the Cartesian x coordinate in this case instead of the range measurements is that the spherical range rises with $1/(\cos(\psi)\cos(\theta))$ leading to indistinguishable EDFs, as demonstrated in Fig. 10-7. As the Cartesian x simply eliminates the influence of the incidence angles by $x = r\ \cos(\psi)\ \cos(\theta)$, it will be the investigated property of the real and simulated detections for full scan evaluation, besides the EPW. Nominal and reference range are measured at $\psi = 0$ and $\theta = 0$, so x and r are equivalent in this special case with $x_{\mathrm{nom}} = r_{\mathrm{nom}}$ and $x_{\mathrm{ref}} = r_{\mathrm{ref}}$.

The EDF F from real data is always plotted in red as already in Sec. 10.2, the p-box $\widetilde{\boldsymbol{\mathcal{F}}}$ is filled blue, with simulation sim1 at x_{ref} again denoted in dotted black lines. As before, the box plots of all data are given at the bottom of each plot to help interpreting the results and to stress possible differences in the probability distributions. Besides, the investigated number of detections from each real measurement n and simulation run sim1 $\widetilde{n}_1$ are given. Thereby, it becomes apparent that there are always less simulated detections, most likely because too less full scans are included in the data. Nonetheless, the validation is possible in this case, as the numbers are still comparable.

The plots for x in Fig. 10-8 indicate accurate simulation results with almost no model bias and a slightly too small scatter. They also indicate a tendency that higher ranges lead to wider p-boxes with higher probability for valid lidar detection simulations. However, when investigating the plots for EPW in Fig. 10-9 higher deviations occur. While the distribution of the measured EPW stays almost constant for the different range samples, the simulated mean is changing from being too high at $x_{\mathrm{nom}} = 1.0\,\mathrm{m}$ to being almost $1.0\,\mathrm{m}$ too small at $x_{\mathrm{nom}} = 5.0\,\mathrm{m}$.

The objective results for the DVM are given in Tab. 10-5 to clarify the visual impressions. The results for EPW show higher differences as for x, as these are almost zero and fulfill the requirements from Sec. 10.1. While the model bias for the selected samples is absolutely still below $1.0\,\mathrm{m}$ and therefore as specified, the scattering error especially of the first sample at $x_{\mathrm{nom}} = 1.0\,\mathrm{m}$ is slightly higher than specified. This deviation is caused by the known but wrong material and its scattering characteristic and shows the metric's sensitivity for such fidelity limitations of the simulation.

As the distribution's mean of the real EPW is almost constant, the relative descent of the EPW of the simulated detections is falsified, as well. Besides, Tab. 10-5 lists the results for the EDF F from real data with only the EDF $\widetilde{F}_1$ from simulation sim1 at x_{ref} and with the whole simulated p-box $\widetilde{\boldsymbol{\mathcal{F}}}$. As with the single beam sample VV&UQ, the p-box intentionally leads to slightly less modeling error, as discussed in Sec. 4.1.5.

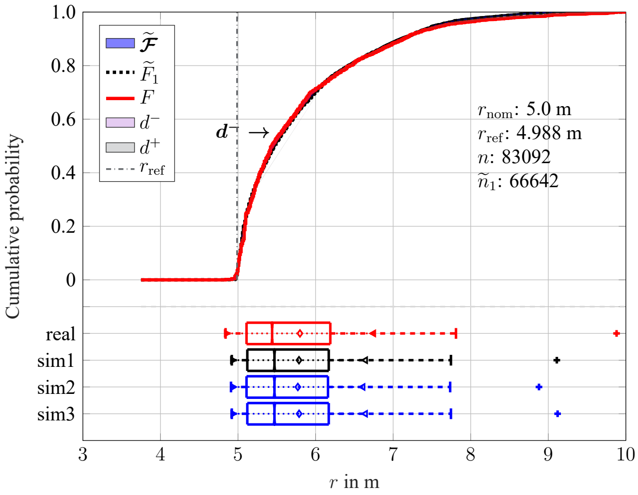


Figure 10-7: Indistinguishable EDFs for spherical range r of detections from full scan at $r_{\text{nom}} = 5.0\,\text{m}$

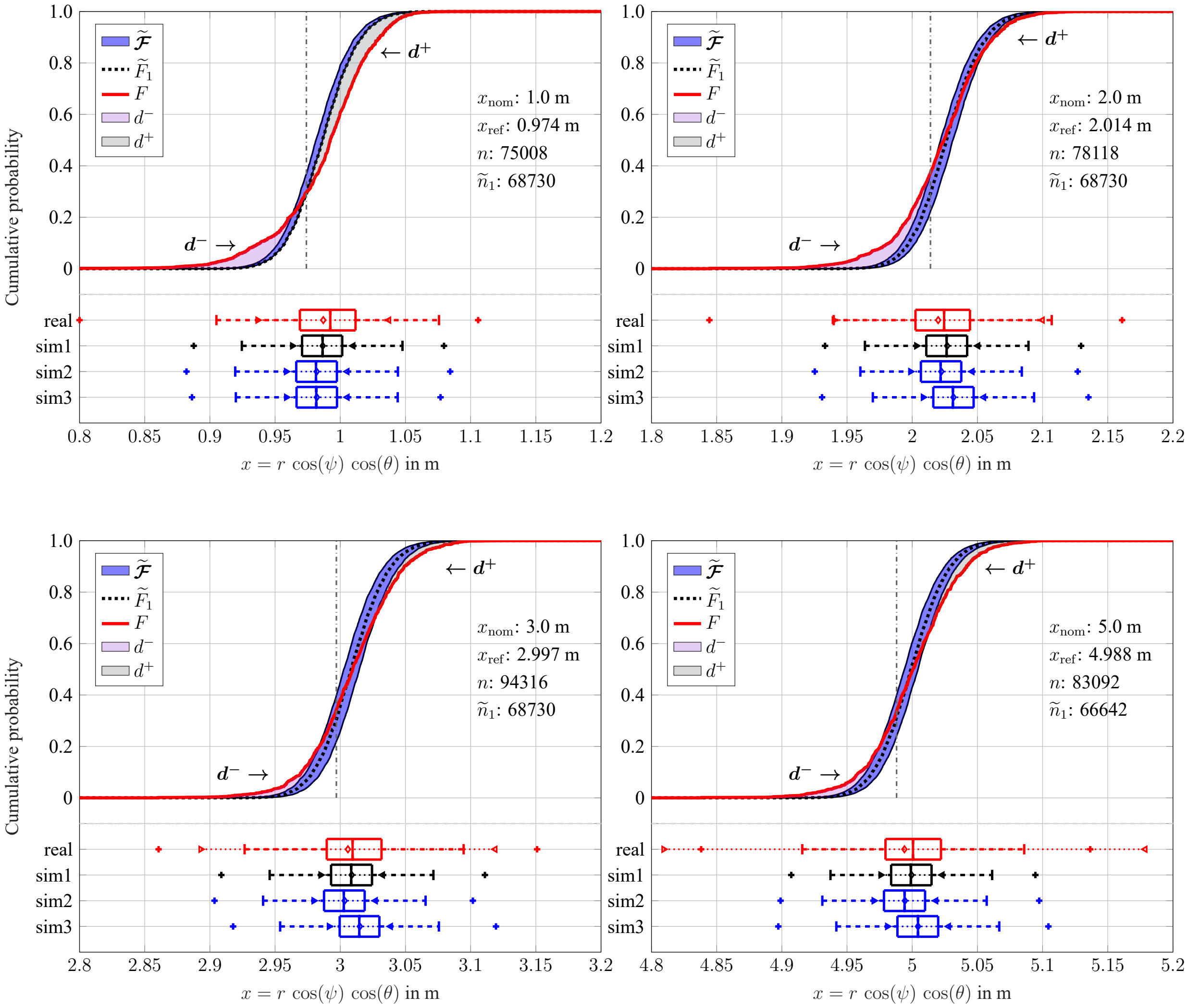


Figure 10-8: Comparison of simulated and measured lidar detections for $x = r\ \cos(\psi)\ \cos(\theta)$ from full scans. n and $\widetilde{n}_1$ are the number of detections from real data and simulation sim1. x_{ref} is the reference range. $\widetilde{\mathcal{F}}$ is the p-box from simulation with $\widetilde{F}_1$ being the EDF from sim_1 at x_{ref}. F is the EDF from real data. d^+ and d^- mark the areas where the simulated p-box is higher (+) or lower (-) than the real EDF. Top left: $x_{\text{nom}} = 1.0\,\text{m}$, Top right: $x_{\text{nom}} = 2.0\,\text{m}$, Bottom left: $x_{\text{nom}} = 3.0\,\text{m}$, Bottom right: $x_{\text{nom}} = 5.0\,\text{m}$.

Table 10-5: DVM results for sample validation for full scan

ζ	x_{nom}	$d_{\text{DVM}}(F, \widetilde{F}_1)$		$d_{\text{DVM}}(F, \widetilde{\mathcal{F}})$	
		d_{bias}	d_{CAVM}	d_{bias}	d_{CAVM}
x	1.00 m	−0.001 m	0.010 m	−0.003 m	0.008 m
	2.00 m	0.005 m	0.008 m	0.003 m	0.004 m
	3.00 m	−0.001 m	0.007 m	0.000 m	0.003 m
	5.00 m	−0.001 m	0.007 m	0.000 m	0.003 m
EPW	1.00 m	0.456 m	0.121 m	0.456 m	0.119 m
	2.00 m	−0.129 m	0.162 m	−0.127 m	0.160 m
	3.00 m	−0.451 m	0.118 m	−0.446 m	0.111 m
	5.00 m	−0.920 m	0.099 m	−0.915 m	0.095 m

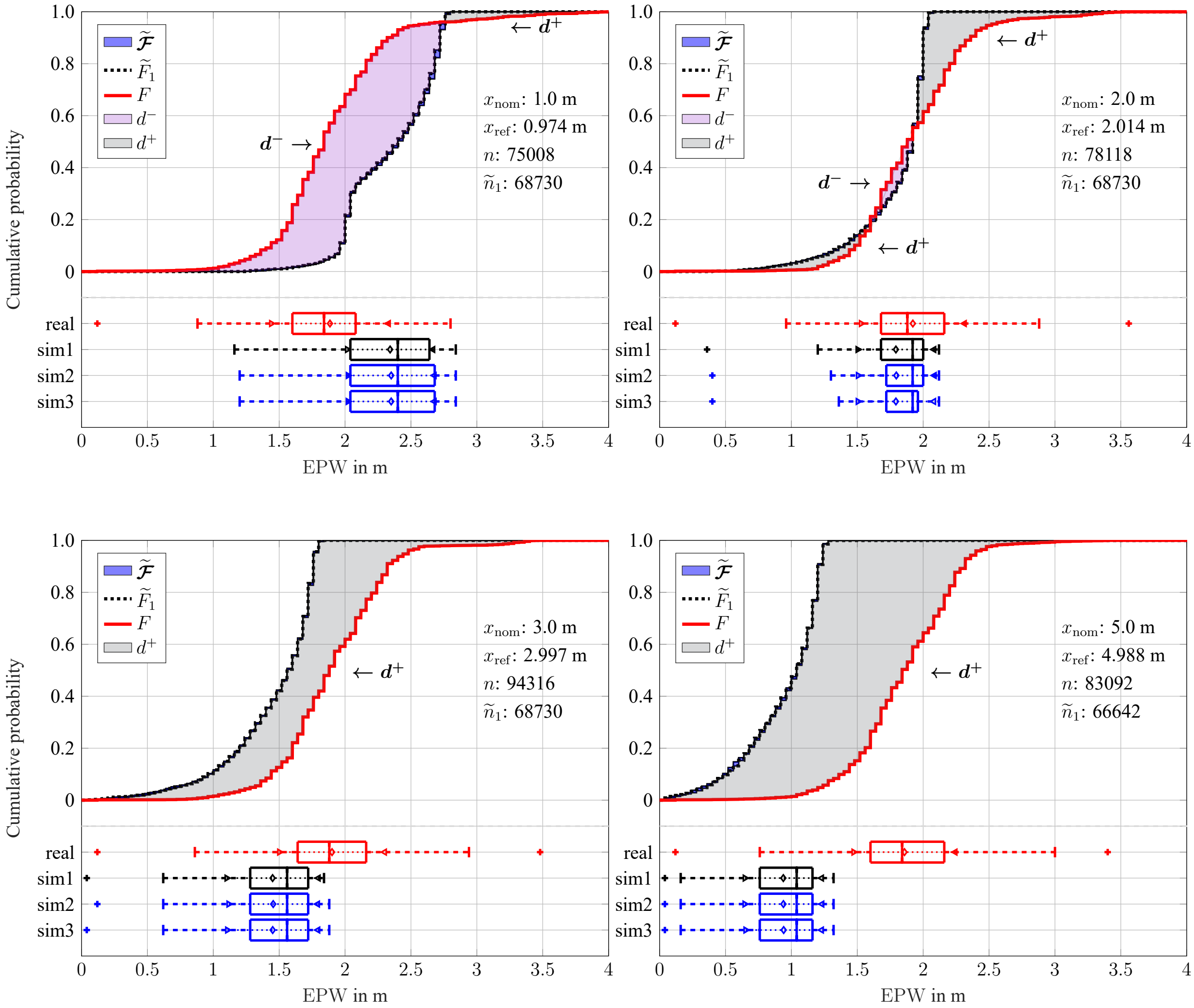
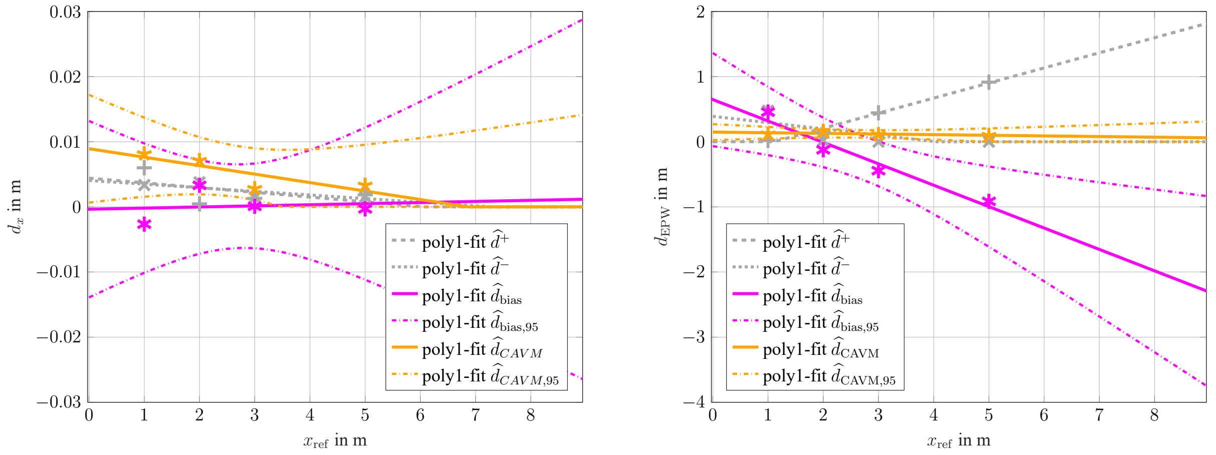


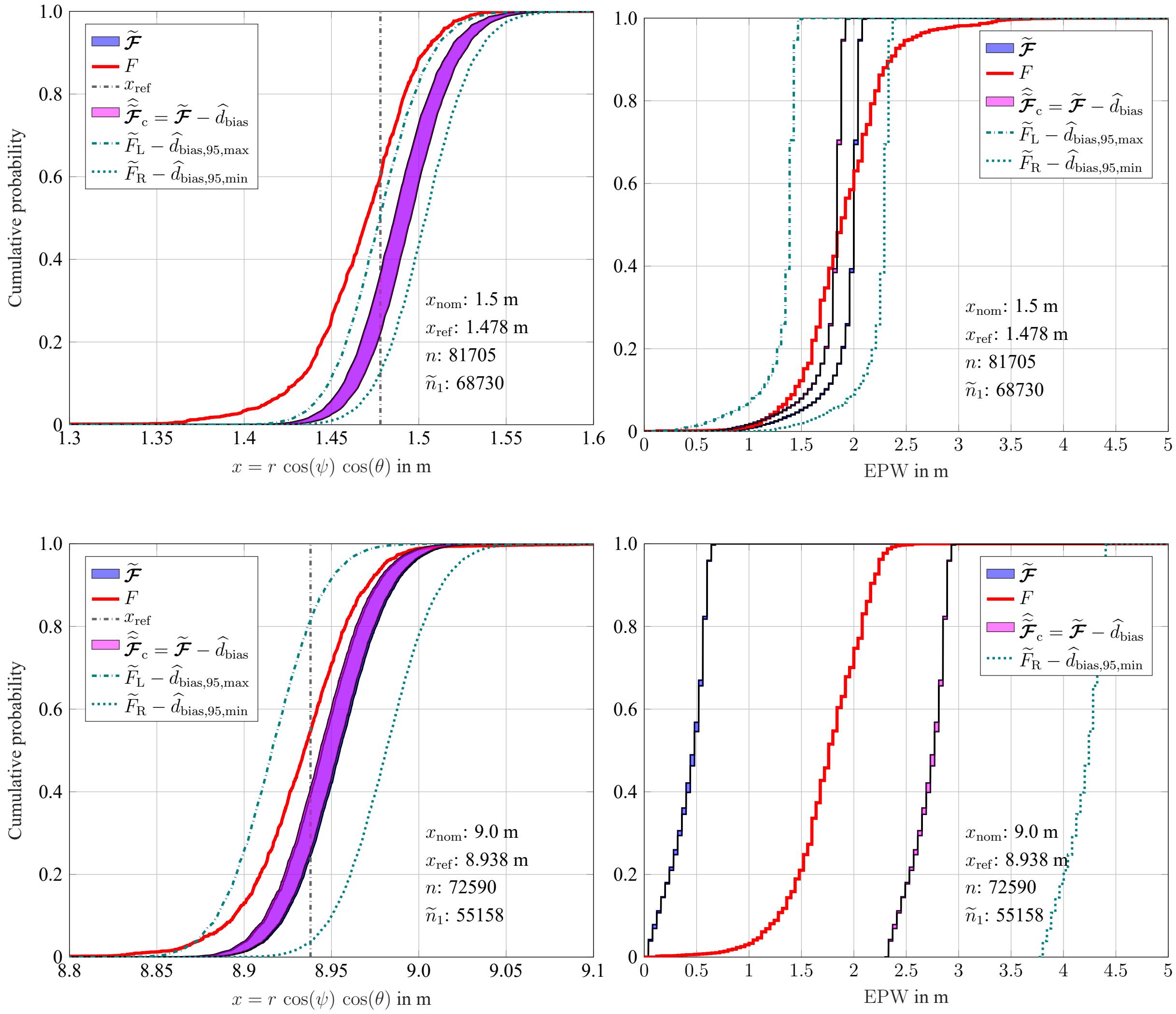
Figure 10-9: Comparison of simulated and measured lidar detections for EPW from full scans. n and $\widetilde{n}_1$ are the number of detections from real data and simulation sim1. x_{ref} is the reference range. $\widetilde{\mathcal{F}}$ is the p-box from simulation with $\widetilde{F}_1$ being the EDF from sim$_1$ at x_{ref}. F is the EDF from real data. d^+ and d^- mark the areas where the simulated p-box is higher (+) or lower (-) than the real EDF. Top left: $x_{\text{nom}} = 1.0\,\text{m}$, Top right: $x_{\text{nom}} = 2.0\,\text{m}$, Bottom left: $x_{\text{nom}} = 3.0\,\text{m}$, Bottom right: $x_{\text{nom}} = 5.0\,\text{m}$.



(a) DVM interpolation for range r of full scan

(b) DVM interpolation for EPW of full scan

Figure 10-10: Inter- and extrapolation of DVM results from full scan sample validation with +: d^+, ×: d^-, ∗: d_{bias}, ⋆: d_{CAVM}



(a) Cartesian x coordinate of detections from full scan

(b) EPW of detections from full scan

Figure 10-11: Validation of model bias prediction for a full scan with DVM

Linear regression of the DVM results is performed as already described for single beam VV&UQ and depicted in Fig. 10-10. This is necessary, as without such interpolation between metric results for the validation samples no argumentation would be possible for the application of the lidar detection simulation in scenarios that do not exactly fit to one of the investigated samples. It is worth noting that with the outlier sample at $r_\text{nom} = 9.0\,\text{m}$ an extrapolation of the model form uncertainty is assessed besides the interpolation for $r_\text{nom} = 1.5\,\text{m}$.

The provided plots show that the confidence intervals are increasing towards higher ranges, as there are no validation samples above $5.0\,\text{m}$, which could bound it otherwise. Nevertheless, the x coordinate of the simulated detections stays fine in absolute numbers in the given interval. The model bias of the EPW is expected to grow over $1.0\,\text{m}$ at ranges higher than $5.0\,\text{m}$, which is finally not fulfilling the last remaining requirement anymore.

To be able to check the interpolation of the sample validation results, Fig. 10-11 shows the predicted model bias and its $95\,\%$ confidence interval for x and EPW at $x_\text{nom} = 1.5\,\text{m}$ and $x_\text{nom} = 1.5\,\text{m}$ together with the EDF from the real measurements and the simulated p-box, similar to Sec. 10.2.

This investigation of the predictions again proves the former claim that the effort does not end with the inter- and extrapolation of the sample validation results and that they must be checked. While there was almost no model bias or model scattering error in all four validation samples, the interpolation for the model bias at $x_\text{nom} = 1.5\,\text{m}$ for the x coordinate in the upper left plot of Fig. 10-11 shows that there is evidently a model bias. The predicted p-box $\widehat{\widetilde{\boldsymbol{\mathcal{F}}}}_\text{c}$ completely overlaps with the simulated one $\widetilde{\boldsymbol{\mathcal{F}}}$, as the estimated model bias $\widehat{d}_\text{bias}$ is $0.000\,\text{m}$. However, the EDF from the measured detections F locates even fully outside of the $95\,\%$ confidence interval for $\widehat{d}_\text{bias}$. Still, the bias prediction error $\widehat{d}_\text{bias} - d_\text{bias}$ is not big in absolute values, as can be seen for this sample in Tab. 10-6.

Table 10-6: DVM predictions and actual values for full scan sample validation of simulated lidar detections. The confidence bounds $\widehat{d}_\textbf{bias,95}$ and $\widehat{d}_\textbf{CAVM,95}$ are the ones with highest absolute value.

$\boldsymbol{\zeta}$	$\boldsymbol{x}_\text{nom}$	**Predicted DVM**				**Actual DVM**	
		$\widehat{d}_\textbf{bias}$	$\widehat{d}_\textbf{bias,95}$	$\widehat{d}_\textbf{CAVM}$	$\widehat{d}_\textbf{CAVM,95}$	$d_\textbf{bias}$	$d_\textbf{CAVM}$
$\boldsymbol{x}$	$1.50\,\text{m}$	$0.000\,\text{m}$	$0.008\,\text{m}$	$0.006\,\text{m}$	$0.012\,\text{m}$	$0.019\,\text{m}$	$0.003\,\text{m}$
	$9.00\,\text{m}$	$0.001\,\text{m}$	$0.029\,\text{m}$	$0.000\,\text{m}$	$0.018\,\text{m}$	$0.012\,\text{m}$	$0.003\,\text{m}$
EPW	$1.50\,\text{m}$	$0.160\,\text{m}$	$0.613\,\text{m}$	$0.134\,\text{m}$	$0.210\,\text{m}$	$0.025\,\text{m}$	$0.218\,\text{m}$
	$9.00\,\text{m}$	$-2.290\,\text{m}$	$-3.763\,\text{m}$	$0.061\,\text{m}$	$0.310\,\text{m}$	$-1.326\,\text{m}$	$0.140\,\text{m}$

In case of the extrapolation for $x_\text{nom} = 9.0\,\text{m}$, the predicted model bias for x is very small, but the $95\,\%$ confidence bounds cover the real EDF in this case. Clearly, the reason for the accurate prediction and the higher bounds in this case is a lack of samples nearby and not a well-balanced sample selection leading to it.

The scattering error prediction based on the CAVM results shown in Tab. 10-6 is also too small for x, but its more conservative $95\,\%$ confidence bounds slightly underestimate the actual value at $x_{\mathrm{nom}} = 1.5\,\mathrm{m}$, while slightly overestimating it for $x_{\mathrm{nom}} = 9.0\,\mathrm{m}$. Similar to the validation of the model error predictions for the single beam evaluation in Sec. 10.2, using the $95\,\%$ confidence bounds is a reasonable choice to not underestimate the model errors for the range simulation.

Despite the already falsified simulation results for EPW, the interpolated results are investigated as well. For both samples at the different ranges / x coordinates of the wall in front of the sensor, the measured EPW is still almost not changing, as already found during sample validation. The linear interpolation of the model bias is expected to cover the model's gain error in case of the EPW, but the predicted bias is overestimated by more than $1.0\,\mathrm{m}$ at $x_{\mathrm{nom}} = 9.0\,\mathrm{m}$. The prediction of the scattering error with the CAVM results for the EPW is in the correct magnitude and most importantly correctly falsify the simulation with respect to its specification in this case.

10.4 DVM for Object Oriented Validation of APSS Simulation

Gating is a widely used detection tracking technique[346] and has already been applied for radar model validation by Magosi et al.[347]. It means that a gating area, as depicted in Fig. 10-12, is drawn around a vehicle to cluster detections that mainly arise from actual reflections caused by it.

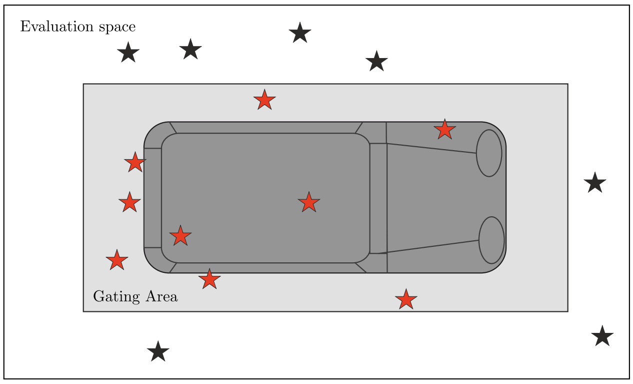


Figure 10-12: Gating area around a vehicle including associated detections (★) and not associated detections (★) from Magosi et al.[347]

Nevertheless, spatial gating of the scene reveals several side-effects. It means that the model is possibly validated against detections that do not necessarily origin in the object reflecting in simulation, but are received from e.g. ground reflections. In this regard, it is hard to distinguish where metric results are mainly caused from, which is at least a drawback for simulation calibration.

[346] Wang, X. et al.: Gating techniques for maneuvering target tracking in clutter (2002).

[347] Magosi, Z. F. et al.: Evaluation of Physical Radar Perception Sensor Models (2022), p. 10.

For validation purposes only, it means that the digital twin of the environment must be validated beforehand. If the differences between the real and simulated environment and objects are well known, however, validation of measurand distributions in gating areas is possible using the proposed DVM and gating can be applied even in dynamic scenarios, as demonstrated by Magosi et al.[347]

To be able to use the novel DVM, a frame-per-frame comparison of dynamic scenarios is proposed, as it is equal to comparing several static scenes, especially when adverse temporal effects can be neglected due to low velocities. When the movements of all objects involved are tracked with high accuracy, a correct re-simulation would lead to comparable lidar scans that can be processed like the static scenario in the full scan evaluation in the previous section. When high velocities of objects are to be investigated, e.g. for VV&UQ of Doppler measurement simulation, as is the case for radar sensors or frequency modulated continuous wave (FMCW) lidars, reference sensor requirements rise even more, while not being impossible to fulfill, as shown by Holder et al.[348]. Additionally, time stamping and the synchronization of all measurement devices during each experiment must be ensured for frame-per-frame comparison afterwards, leading to immense efforts but supporting repeatability and reproducibility.

Each frame actually becomes a new sample from the parameter space due to e.g. variation of range or aspect ratio of the objects and a dynamic scenario reflects a subspace of the overall parameter space resulting in research demand how to tackle all mentioned challenges. Because of all the mentioned drawbacks, dynamic scenarios are left out of scope of this dissertation without restricting the applicability of the novel DVM, as explained.

Berghöfer[349] highlights another aspect of object oriented validation of APSS simulation in his thesis. His comparison of a real and virtual scene in Fig. 10-13 serves as a realistic example for possible differences between simulation and reality concerning sensor/object positions and the more or less accurate virtual representations e.g. of the objects, the ground, and the vegetation.

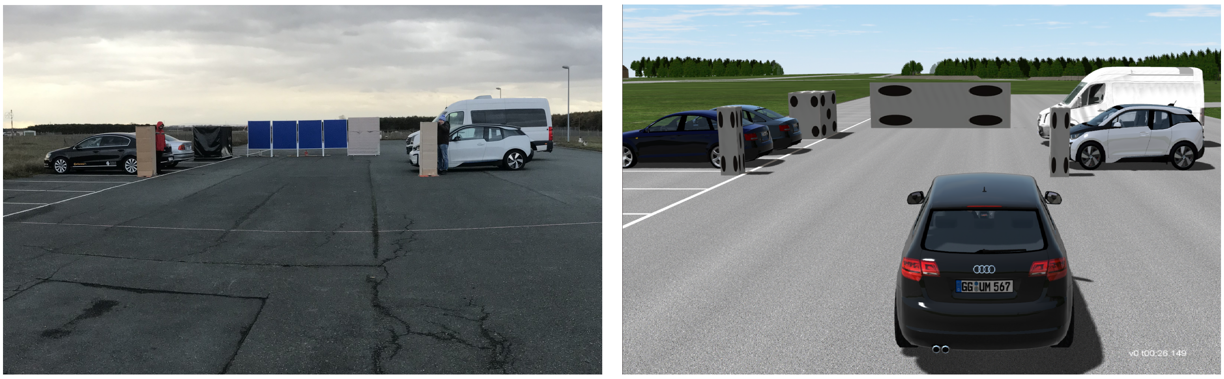


Figure 10-13: Comparison of real and virtual scene in actual model validation studies from Berghöfer[349]

[348] Holder, M. F. et al.: Digitalize the Twin (2022).

[349] Berghöfer, M.: Generierung realer und synthetischer Sensordaten zur Simulations-Validierung (2019), pp. 60 & 62.

In the following object oriented sample validation of simulated lidar detections, data from an experiment already described in former work of the author[350] is used. The scene consists of a mid-size car in front and a partly occluded delivery van in the back. A sketch of the scene together with a photograph are shown in Fig. 10-14. However, the depicted removable bounding box mock-up was not present during measurements or simulations used in this dissertation. The spatial gating tolerance applied in this case is $10\,\mathrm{cm}$ around each vehicle. Tab. 10-7 lists the nominal and reference ranges from the sensor up to each vehicle.

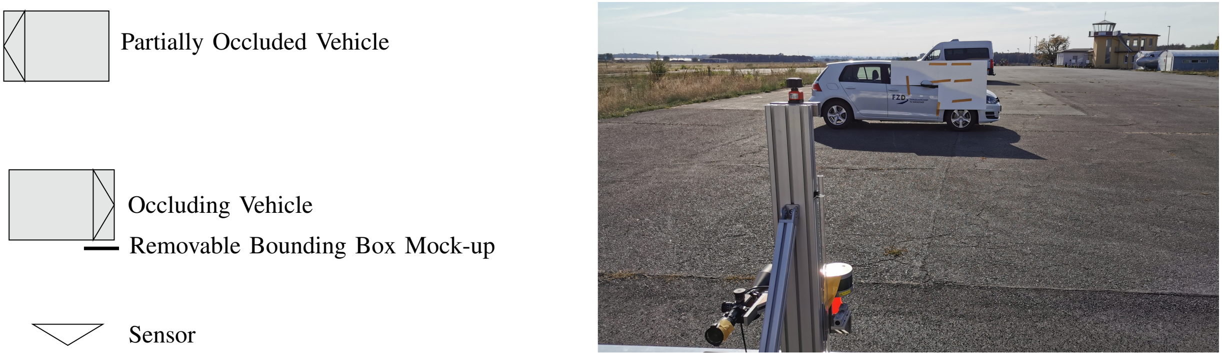


Figure 10-14: Experimental setup with partially occluded object as sketch and photograph from former work of the author.[350] © IEEE 2021.

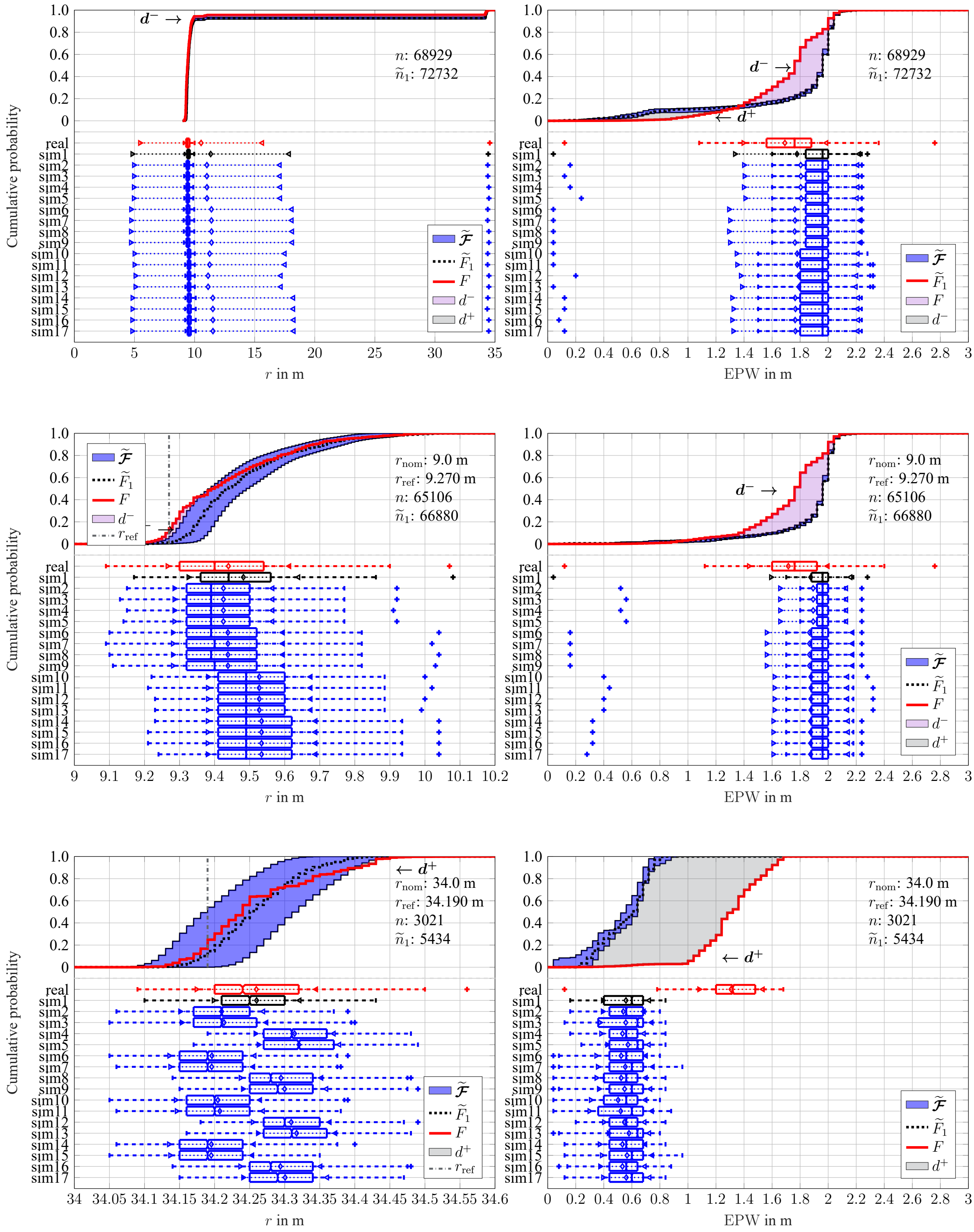
Table 10-7: Central nominal and reference ranges in object occlusion experiments

Object	**Front, occluding**	**Back, partially occluded**
$r_{\mathbf{nom}}$	$9.00\,\mathrm{m}$	$34.00\,\mathrm{m}$
$r_{\mathbf{ref}}$	$9.27\,\mathrm{m}$	$34.19\,\mathrm{m}$

Fig. 10-15 shows three plots for each measurand. The range r is shown on the left (a) and the EPW is plotted on the right (b). On top are full scan evaluations with the two objects, the middle plots are from detections inside the gating area around the occluding object in front, and the plots at the bottom show the results for gating around the partially occluded object in the back. In simulation, besides simulating both cars at the positions measured with the reference measurement device, several simulations are performed varying lateral and longitudinal positions by $\pm 5.0\,\mathrm{cm}$. The multiple simulations result in several box plots below each EDF plot in Fig. 10-15 and lead to obviously wider p-boxes compared to the two sample validation campaigns before.

In contrast to full scan validation, the object oriented validation is performed in this case on range r measurements and not with the Cartesian x coordinate of the detections. Therefore, the influence of azimuth and elevation of the detections on the range measurements is not neglected in this case. This is allowed as on the one hand, the influence influences measurement and simulation results in the same way and on the other hand, both objects are positioned in central azimuth range and in relatively far range from the sensor. However, this should be checked for other scenarios and sometimes validating on Cartesian coordinates instead or additionally has to be considered.

[350] Linnhoff, C. et al.: Refining Object-Based Lidar Sensor Modeling (2021), p. 24240.



(a) Range r of detections from occluding/occluded objects (b) EPW of detections from occluding/occluded objects

Figure 10-15: Comparison of simulated and measured lidar detections from occluding/occluded objects. Top: Full scan, Middle: occluding vehicle, Bottom: occluded vehicle
n and $\widetilde{n}_1$ are the number of detections from real data and simulation sim1. is the reference range.
$\widetilde{\mathcal{F}}$ is the p-box from simulation with $\widetilde{F}_1$ being the EDF from sim_1 at r_{ref}. F is the EDF from real data.
d^+ and d^- mark the areas where the simulated p-box is higher (+) or lower (-) than the real EDF.

As indicated by the numbers of real (n) and simulated ($\widetilde{n}_1$) detections in Fig. 10-15, there is a significantly smaller number of detections from simulation than in real data for the partially occluded object due to a imprecise 3d model of this object and its windows in simulation. In contrast, the numbers of detections for the occluding object and even for the full scan are comparable. However, the higher number of detections from the partially occluded car in simulation id the only reason for the area between the real EDF F and the p-box from the 17 simulations, as it pulls the cumulative probability of the detections at the occluding object (r_{nom}=$9.00\,\text{m}$).

This example of an object oriented sample validation and especially the plots for the range simulation at the left of Fig. 10-15 show how a wider p-box eliminates any measurable error with DVM. One can imagine how many reference measurement devices are necessary for complex scenarios including e.g. weather influence. Therefore, the reference sensors must be highly accurate to not automatically prove any simulation as valid simply by the high epistemic uncertainty. On the other hand, the width of the p-box correctly is fairness for the simulation, as it should not be compared to a single simulated EDF, when there evidently exists a significant epistemic uncertainty.

In case of simulated range r of each detection, due to the relatively wide p-boxes, both EDFs for each vehicle are almost fully covered. The objective results in Tab. 10-8 for model bias and model scattering error prove the visual impression of sample valid simulation for the range measurements and the simulation pass the specifications from Sec. 10.1. The results for the full scan are there for completeness and the wrong virtual car model for the partially occluded vehicle is the reason for the too high values in this case, as explained.

However, the simulated EPW is not as accurate as the range. For the occluding object in front at the relatively short range of r_{nom}=$9.00\,\text{m}$, the simulated values are slightly too high, as can be seen in the plot proven by the results in Tab. 10-8. For the higher range, the simulated EPW is far too low. Basically, the same model gain error is found as in the full scan evaluation in the previous section. For EPW the different number of detections from simulation does not effect the results in contrast to the range.

Table 10-8: DVM results for object gating sample validation

ζ	Gating	$d_{\text{DVM}}(F, \widetilde{\mathcal{F}})$	
		d_{bias}	d_{CAVM}
r	Full scan	$0.484\,\text{m}$	$0.784\,\text{m}$
	Occluding obj.	$0.007\,\text{m}$	$0.004\,\text{m}$
	Occluded obj.	$-0.001\,\text{m}$	$0.001\,\text{m}$
EPW	Full scan	$0.096\,\text{m}$	$0.116\,\text{m}$
	Occluding obj.	$0.154\,\text{m}$	$0.073\,\text{m}$
	Occluded obj.	$-0.716\,\text{m}$	$0.034\,\text{m}$

It has been shown that EDF/p-box plots incl. the number of values from measurement and simulation $n, \widetilde{n}$ are a valuable tool for validation. The object oriented sample validation also shows that it supports finding systematic errors in reference data e.g. for object positions or even object geometry errors. Finally, the plot for cumulative probabilities of the real and simulated azimuth ψ of the detections in Fig. 10-16 indicates a slightly misplaced virtual car in lateral direction. Consequently, systematic simulation errors can be found very well with the presented method too.

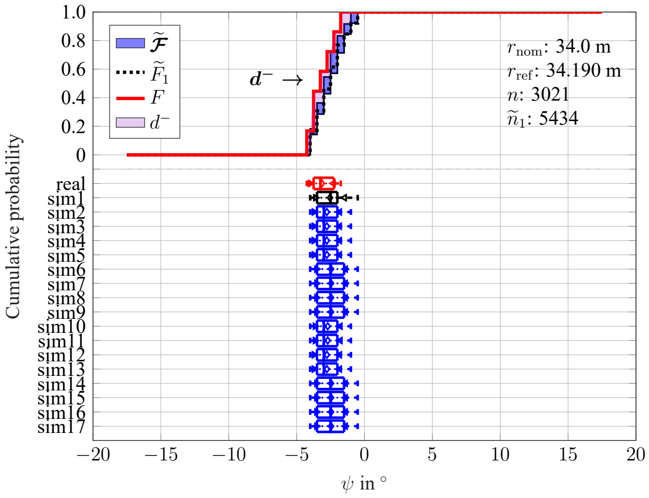


Figure 10-16: EDF/p-box plot for partially occluded object

11 Conclusion and Outlook towards APSS Simulation Credibility

As Trucano et al. state in their conclusion of their description of the Sandia Validation Metrics Project in 2001, there *"is not a unique set of steps one follows to establish model validity for a complex application. Validation is a continuous process that must adapt as it progresses."* [351] Nonetheless, the whole VV&UQ process for APSS simulation is addressed in this dissertation, alongside the exemplary development of a reflection-based lidar model as application example. In the following final chapter, a conclusion is given for the novel specification method and the novel DVM and an outlook towards further work following up on the presented findings is provided.

11.1 Conclusion

The treatise has a holistic aspiration for credible simulation of APSS. Thereby, credibility of APSS simulation is named for the first time as overall goal. To be able to demonstrate the novel metrics for VV&UQ, at first an APSS model is needed. Therefore, the reflection-based lidar model is exemplary developed to have a white-box lidar simulation that can be calibrated and validated. However, it relies on reflections that are computed in a separate simulation environment. Therefore, the validation of the model output detections is always in combination with a ray tracing tool and only true for this specific investigated tool, even if standardized interfaces like OSI and standardized material libraries are hopefully covering this issue in the future.

After presenting the SotA of APSS modeling and the implemented reflection-based lidar model approach, the SotA of APSS modeling and VV&UQ (APSS) simulation is recorded and discussed. An extensive collection of all known metrics applied in the field of APSS is presented, resulting in a list of 34 possible metric candidates.

The VV&UQ discussion starts after describing the most urging challenges in this field of science with a novel requirements methodology for APSS simulation, based on the methods PerCollECT and CEPRA, both elaborated by the author. Furthermore, every element of simulation credibility assessment from Fig. 1-7 is discussed.

From the collection of 34 SotA metrics, the most promising candidates to handle probabilistic data, namely hypothesis testing, divergences like the Kullback-Leibler, and the AVM are selected to be evaluated in more detail. From this pre-selection, the AVM, which is the area between two CDFs/EDFs/p-boxes is elaborated as best possible candidate and analyzed in more detail. One of

[351] Trucano, T. G. et al.: Description of the Sandia Validation Metrics Project (2001), p. 55.

the most beneficial properties of the AVM is that it provides intuitive results in the unit of the measurand. Computing the area between two cumulative probability distributions is an intuitive step, while this is equivalent to calculating the mean of the differences between the quantiles of the two CDFs/EDFs/p-boxes. Nonetheless, the drawbacks of the AVM are not ignored, like the insensitivity to scattering differences, when there is no overlap of the curves.

The novel metrics are based on the AVM, but improve it by not mixing up model bias and scattering error. The CAVM measures the scattering error of the model with respect to the real sensor and the DVM, which combines the CAVM with the model bias with respect to the real sensor's bias, is the finally recommended and demonstrated metric. CAVM and DVM are novel metrics introduced here for the first time, while their approaches follow Einstein's principle, where he stated that *"Everything should be as simple as it can be, but not simpler."*[352]

It can be applied to probabilistic data in any form and origin as single uncertain number, CDF, EDF, or p-box. In other words, it is able to process real and simulated data considering aleatory and epistemic uncertainty in sensor measurements to model and in reference measurements as simulation input, as both parts of the DVM are derived from the AVM. Additionally, the DVM is not only useful for VV&UQ, but also for model calibration and for model verification, as described and demonstrated.

After the elaboration of the novel DVM, its applicability for VV&UQ of lidar detection simulation is demonstrated. Validation of lidar simulation has already been tackled in literature by other authors with the listed metrics. However, in this dissertation UQ and its inter- and extrapolation to the application domain within the possible parameter space spanned by the ODD of the SuT is shown for the first time. To be able to predict the errors, the separation of model bias and scattering error with the novel DVM is of special importance for the confidence bounds of the model error prediction.

It should be stressed again that it is a crucial step to investigate the validity of predicted model error for application. Even if some investigated validation samples do not fulfill the requirements, one could still use the simulation, if the model error predictions would be accurate, as they could be used to eliminate the invalidity during application. Without interpolation between the metric results for the validation samples, they just stay samples. However, as linear regression is applied for inter- and extrapolation, the model needs to be sample valid at first to have a chance to stay below the specified error limits with the predictions and the uncertainties on them reflected by the $95\,\%$ confidence bounds.

While one could argue that instead of reserving some sample validation results for validating the error prediction, they better would have been used for more accurate inter- and extrapolation in the first place. Clearly, the author of this dissertation strongly recommends to at least reserve a small portion for a final cross-check. However, a good balance of prediction and test at the

[352] Calaprice, A.: The Ultimate Quotable Einstein (2010), pp. 384-385.

end relies on the personal experience of the team performing the VV&UQ and is expected to get progressively better balanced over time.

Three different use cases are demonstrated to stress the applicability of the novel DVM and its prediction towards the application domain. A single-beam evaluation of the simulation, a full-scan evaluation, and an object oriented validation. The first two model assessments include a validation of the model error prediction, the final step of every consequent VV&UQ. All three use cases include a VV&UQ not only of the range value of the simulated detections, but also their echo pulse width (EPW), which is similar to intensity and a specialty of the exemplary simulated lidar sensor. Neither lidar detection intensity nor EPW simulations have been validated in literature yet, to the knowledge of the author.

Nonetheless, it has been shown and explained that simulating lidar detection intensities or EPWs is very complex and relies on many parameters and influences to simulate. Besides, it is hard to calibrate due to huge differences between sensors of the same margin. This is analog to the radar cross-section (RCS) simulation challenge described in former work[353].

The concentration on static scenarios might seem as a limitation at first sight, but they equal a frame-by-frame comparison of dynamic scenarios. The usability of the metric is based on its property to provide results in the unit of the measurand that is simulated and validated. Therefore, it is very intuitive for writing and understanding requirements, even for non-experts, closing the loop on the holistic approach in this dissertation.

Overall, the novel metric, the DVM, is not meant to be used exclusively. Instead, other metrics are expected to be used, as well. Especially in case of model calibration, a hypothesis test for same distribution like Kolmogorov-Smirnov test or Anderson-Darling test is expected to be useful, among others.

In conclusion, the combination of the individual parts of the holistic process towards simulation credibility and making adding a systematic process for requirements engineering to the methodology is a major contribution besides the novel metrics CAVM and DVM and their demonstration. Additionally, and in consequence of the holistic claim of this dissertation, even small, but possibly painful conditions like scaling of axes for metric application are discussed.

11.2 Outlook towards APSS Simulation Credibility

One aspect that has not yet been explicitly demonstrated regarding the applicability of the presented DVM is its application on FMCW lidar or radar data including Doppler velocities. Nevertheless, radar detections are similar to lidar detections in general, which leads to the assumption that the presented metrics apply, with the slight difference of RCS instead of intensity and additional

[353] Holder, M. F. et al.: Measurements revealing Challenges in Radar Sensor Modeling (2018).

velocity measurements that should be validated like the other measurands. Still, the applicability to radar and other sensor simulations is to be demonstrated in the future.

Additionally, investigations for the novel DVM for quasi-static scenarios with constant velocities, possibly reflecting only a single sample in parameter space, or even dynamic scenarios covering a set of samples from parameter space and how repeatability and reproducibility of covering the same cells of that specific parameter space should follow.

Furthermore, a possible extension to the presented approach would be to reflect the epistemic uncertainty in measurement data due to missing calibration of intensity / EPW / RCS output in real APSS as a p-box as well. This would lead to a comparison of two p-boxes for validation, which is possible with the presented metrics without limitation.

In some cases, the p-box from simulation can become large due to high epistemic uncertainty or many parameters to sum up for it. Then, calculating the area of the p-box besides the values in the DVM is necessary to support the sample validation.

The chosen linear interpolation of the DVM results for error prediction is up to discussion, as it is explicitly performed for the first time in this dissertation for APSS simulation. Alternatively, exponential functions could be the better choice for CAVM interpolation, as they do not fall below 0, so clipping negatively interpolated values to 0 would not be necessary anymore. Sigmoid functions should also be considered for an investigation, as they are constrained for high values.

While experiment design and sample selection with sensitivity analysis has been discussed in this dissertation, sample representativeness is still a research topic for the future. Representativeness in this context is the objective to find the right samples from the parameter space e.g. by sensitivity analysis, but could alternatively be motivated by criticality of scenarios and the so-called corner-cases. In this regard, the question of effect chain isolation for validation of each modeled effect chain separately should be investigated in the future as well. While it is obviously not possible to reach a complete isolation, at least the experiment design is to be optimized in this regard.

An automated process or a software tool for full VV&UQ and also for the validation of the model error and uncertainty prediction seems possible and reasonable for the future. It should get the data from simulation and the real measurements and analyze it at first for systematic errors in re-simulation, as e.g. shown for lateral shift of virtual objects or imbalanced amounts of values. Then the DVM results are to be calculated and interpolated, if an application domain is specified. Eventually, a VV&UQ report should be provided.

The final goal and a consequent next step after VV&UQ and model error prediction would be a (self-)conscious simulation based on the predictions, where each model reports its current predicted error and confidence bounds live during the simulation run. While most parameters of the parameter space do not change during a scenario, when scenario-based safety validation is performed, some still vary during simulation, so the actual error predictions must be calculated for each simulation step.

A Exemplary Excerpt of CEPRA

Table A-1: Severity S between 1 and 10 based on the FMEA Handbook by Automotive Industry Action Group (AIAG) and Verband der Automobilindustrie (VDA)[354a]

S	Severity	Criteria
10	Very High	Affects safe operation of the vehicle and/or other vehicles, the health of driver or passenger(s) or road users or pedestrians.
9		Noncompliance with regulations.
8	High	Loss of primary vehicle function necessary for normal driving during expected service life.
7		Degradation of primary vehicle function necessary for normal driving during expected service life.
6	Moderate	Loss of secondary vehicle function.
5		Degradation of secondary vehicle function.
4		Very objectionable appearance, sound, vibration, harshness, or haptics.
3	Low	Moderate objectionable appearance, sound, vibration, harshness, or haptics.
2		Slightly objectionable appearance, sound, vibration, harshness, or haptics.
1	Very Low	No discernible Failure Effect.

Table A-2: Frequency F between 1 and 10 based on the FMEA Handbook by Automotive Industry Action Group (AIAG) and Verband der Automobilindustrie (VDA)[354b]

F	Frequency	Criteria
10	Extremely high or cannot be determined	Frequency of occurrence of the Failure Cause is unknown or known to be unacceptably high during the intended service life of the vehicle.
9	High	Failure Cause is likely to occur during the intended service life of the vehicle.
8		Failure Cause may occur often in the field during the intended service life of the vehicle.
7	Medium	Failure Cause may occur frequently in the field during the intended service life of the vehicle.
6		Failure Cause may occur somewhat frequently in the field during the intended service life of the vehicle.
5		Failure Cause may occur occasionally in the field during the intended service life of the vehicle.
4	Low	Failure Cause is predicted to occur rarely in the field during the intended service life of the vehicle. At least ten occurrences in the field are predicted.
3	Very low	Failure Cause is predicted to occur in isolated cases in the field during the intended service life of the vehicle. At least one occurrence in the field are predicted.
2		Failure Cause is predicted not to occur in the field during the intended service life of the vehicle based on prevention and detection controls and field experience with similar parts. Isolated cases cannot be ruled out. No proof it will not happen.
1	Cannot Occur	Failure Cause cannot occur in the field during the intended service life of the vehicle or is virtually eliminated. Evidence that Failure Cause cannot occur. Rationale is documented.

[354] AIAG; VDA: FMEA Handbook - Failure Mode and Effects Analysis (2019). a: p. 122.; b: pp. 123-124.

Table A-3: Exemplary excerpt of CEPRA*

CE-PRA-ID	Pheno-menon (P)	Effect chain (EC) of phe-nomenon	Causes of effect chains		P&EC occurrence in ODD* (O, filled by sensor expert)		P&EC impact on SUT in ODD* (I, filled by SUT expert)		Rele-vance of P&EC
			Environ-mental causes	**Design parameters**	***[1, 10]***	***Rationale***	***[1, 10]***	***Rationale***	***O + I***
Lid_ CEPRA _001	False nega-tive in object list	→ FN features → FN detections → Not dist. from noise floor → Low rec. power from object → Transmittance by object parts	Materials of reflect. obj. parts	Emitter wavelength	3	*Materials with high transmittance are common but usually come in a combination with objects of normal reflectivity*	9	*False Negatives are likely to occur in other sensors as well, causing the ego vehicle to completely ignore the object.*	***12***
Lid_ CEPRA _002	False nega-tive in object list	→ FN features → FN detections → Not dist. from noise floor → Low rec. power from object → Absorption by object parts	Materials of reflect. obj. parts	Emitter wavelength	2	*Materials with high absorption (such as Vantablack) cannot be excluded completely, but usually come in a combination with objects of normal reflectivity.*	4	*Cameras will likely detect objects with high absorption, but with less precise spatial information. Object fusion will cover the false negative. Jerky maneuvers can be a result.*	***6***

* Kindly provided, not yet published intermediate project result of publicly funded research project VVM

Lid_ CEPRA _003	False negative in object list	→ FN features → FN detections → Not dist. from noise floor → Low rec. power from object → Reflection by object parts	Materials of reflect. obj. parts Roughness of reflect. obj. parts Shapes of reflect. obj. parts Size of reflect. obj. parts	Emitter wavelength	3	*Materials with total reflection are common, but usually occur in combination with objects with normal reflection*	2	*Likely, only single sensors are affected by this, so sensor fusion in combination with a multi sensor setup can cover the false negative.*	***5***
Lid_ CEPRA _004	False negative in object list	→ FN features → FN detections → Not dist. from noise floor → Low rec. power from object → Occlusion by objects → Occlusion by object parts → Absorption by object parts	Materials of reflect. obj. parts	Emitter wavelength	2	*FN objects caused by occluding absorbing objects is predicted not to occur, as materials with almost total absorption (such as Vantablack) still reflect a portion of the signal that the lidar sensor is able to detect.*	9	*False Negatives of objects occluded by absorbing objects will cause a lack of awareness of possible occlusions by the environment model. Depending on the traffic situation, rule violations or even collisions can be a cause.*	***11***

Lid_ CEPRA _005	False nega- tive in object list	→ FN features → FN detections → Not dist. from noise floor → Low rec. power from object → Occlusion by objects → Occlusion by object parts → Reflection by object parts	Materials of reflect. obj. parts Roughness of reflect. obj. parts Shapes of reflect. obj. parts Size of reflect. obj. parts	Emitter wavelength	9	*FN objects caused by occluding reflecting objects occur often in a front lidar e.g. for situations at the intersections caused by parking cars, trees, billboards and other objects.*	4	*Occluded objects can cause emergency brake maneuvers, if they come to interact with the ego vehicle. Collisions are unlikely, since the occlusion can be detected by the environment model.*	***13***
Lid_ CEPRA _006	False nega- tive in object list	→ FN features → FN detections → Not dist. from noise floor		Resolution of receiver	3	*Small objects (wire fence) cannot be detected by lidar, mostly occur in a combination with larger objects*	7	*Small objects are typically not located on the road, but beside. Moreover, they cause only minor damage and threat to the passengers in case of an unlikely collision. Vulnerable road users are not among those objects.*	***10***

Lid_CEPRA_007	False negative in object list	→ FN features → FN detections → Not dist. from noise floor		Sensitivity of receiver	2	*Calibration of sensor sensitivity is consistently performed and adjusted while on the road.*	4	*Likely, only single sensors are affected by this, so sensor fusion in combination with a multi sensor setup can cover the false negative. Camera sensors will not be affected.*	***6***
Lid_CEPRA_008	False negative in object list	→ FN features → FN detections → Not dist. from noise floor → Low rec. power from object → Attenuation by atm. aerosol → Absorption by atm. aerosol	Signal dist. in atm. aerosol Density of atm. aerosol Material prop. of atm. aerosol particles Size of atm. aerosol particles	Emitter wavelength	8	*FN objects due to fully absorbing atmospheric aerosol occur at strong intensity of rain, snow, or heavy fog and dust. Absorption is particularly relevant for object detection at long distances*	9	*Other sensors will likely be affected as well. Thus, sensor fusion can not cover false negatives, resulting in traffic rule violation or collisions.*	***17***
Lid_CEPRA_009	False negative in object list	→ FN features → FN detections → Not dist. from noise floor → Low rec. power from object → Attenuation by atm. aerosol → Scattering by atm. aerosol	Density of atm. aerosol Material prop. of atm. aerosol particles Size of atm. aerosol particles	Emitter wavelength	9	*FN objects due to scattering atmospheric aerosol occur very often in lidar measurements even in light rain.*	9	*Other sensors will likely be affected as well. Thus, sensor fusion can not cover false negatives, resulting in traffic rule violation or collisions.*	***18***

B Exemplary Lidar Sensor System Simulation Requirements

Important information:

- Status can be ”accepted”, ”rejected”, or ”under review”
- Scheme for requ. IDs: <Phenomenon>_<Causes>_<No.>
- The lidar sensor system simulation consists of two subsequent models that are specified here:
  - Reflection based lidar detection model
  - Detection based lidar object model.
- If values are given e.g. for ODD description, the measurement device to be used should be listed as well.
- Tests can be performed in two ways depending on the requirement:
  - Simulation-only tests:
    - Regression tests (is there any change?)
    - Against specific KPIs
    - Against GT from simulation
    - Against results from another simulation model
  - As re-simulation with an experiment performed at first, collecting reference data, as input for the simulation later, and real sensor data to compare the simulated sensor data against after re-simulation.
- The experiment can be performed in the field, in the lab (sensor, subsystem, full ADS), ...

Table B-1: General requirements for lidar sensor system simulation*

Lidar sensor system simulation requ. ID	Description	Type	Source CEPRA ID	Related lidar sensor system simulation requ. IDs	Related lidar object model requ. IDs	Related lidar detection model requ. IDs	Accep-tance test IDs	Status
Lid_ sys_sim_ req_001	Lidar sensor system simulation shall not detect GT objects that are completely occluded in the sense of geometric optics.	Requirement	Lid_ CEPRA _005	Lid_ sys_sim_ req_003, Lid_ sys_sim_ req_004, Lid_ sys_sim_ req_005,	Lid_ obj_mod_ req_001	Lid_ det_mod_ req_001	Lid_ sys_sim_ test_001	accepted
Lid_ sys_sim_ req_002	Lidar sensor system simulation shall not detect GT objects that are partly occluded. These GT objects can not be recognized by he system due to a low number of detections and/or low echo-pulse width (EPW) and/or false object clustering.	Requirement	Lid_ CEPRA _005	Lid_ sys_sim_ req_003, Lid_ sys_sim_ req_004, Lid_ sys_sim_ req_005, Lid_ sys_sim_ req_009	Lid_ obj_mod_ req_001, Lid_ obj_mod_ req_002	Lid_ det_mod_ req_002, Lid_ det_mod_ req_003	Lid_ sys_sim_ test_002	under review

[*] Kindly provided, not yet published intermediate project result of publicly funded research project VVM

Lid_sys_sim_req_003	Occlusion in the sense of geometric optics always involves occluding (GT) objects in the line-of-sight of the sensor towards the occluded object.	Definition	/	Lid_sys_sim_req_001, Lid_sys_sim_req_002, Lid_sys_sim_req_004, Lid_sys_sim_req_005	Lid_obj_mod_req_001, Lid_obj_mod_req_002	/	/	under review
Lid_sys_sim_req_004	GT objects include houses, cars, pedestrians, and cyclists.	Definition	/	Lid_sys_sim_req_001, Lid_sys_sim_req_002, Lid_sys_sim_req_003, Lid_sys_sim_req_005	Lid_obj_mod_req_001, Lid_obj_mod_req_002	/	/	under review

Lid_sys_sim_req_005	Completely occluding GT objects in the sense of geometric optics are completely non-transparent.	Definition	/	Lid_sys_sim_req_001, Lid_sys_sim_req_002, Lid_sys_sim_req_003, Lid_sys_sim_req_004	Lid_obj_mod_req_001, Lid_obj_mod_req_002	/	/	under review
Lid_sys_sim_req_006	The Cartesian coordinate system for sensors and objects for x-y-z is defined front-left-up according to DIN ISO 8855:2013-11[355]	Definition	/	/	/	/	Lid_sys_sim_test_001, Lid_sys_sim_test_002, Lid_sys_sim_test_003, Lid_sys_sim_test_004	under review

[355] Normenausschuss Automobiltechnik (NAAutomobil) im DIN: DIN ISO 8855:2013-11 (2013)

Lid_ sys_sim_ req_007	Reference point of all objects is the geometric center of their bounding box.	Definition	/	/	/	/	Lid_ sys_sim_ test_001, Lid_ sys_sim_ test_002, Lid_ sys_sim_ test_003, Lid_ sys_sim_ test_004	under review
Lid_ sys_sim_ req_008	Reference point of the sensors is the eye-point as defined by the manufactorers specification.	Definition	/	/	/	/	Lid_ sys_sim_ test_001, Lid_ sys_sim_ test_002, Lid_ sys_sim_ test_003, Lid_ sys_sim_ test_004	under review

Lid_sys_sim_req_009	The echo-pulse width (EPW) of the measured and simulated detections is given in m.	Definition	/	Lid_sys_sim_req_002, Lid_sys_sim_req_010, Lid_sys_sim_req_011	Lid_obj_mod_req_002	Lid_det_mod_req_003, Lid_det_mod_req_004, Lid_det_mod_req_005	Lid_sys_sim_test_002	under review
Lid_sys_sim_req_010	Lidar sensor system simulation shall not detect GT objects that are not recognized due to lower EPW per detection and a low number of detections because of signal scattering by atmospheric aerosols.	Requirement	Lid_CEPRA_009	Lid_sys_sim_req_009, Lid_sys_sim_req_012, Lid_sys_sim_req_013	Lid_obj_mod_req_001, Lid_obj_mod_req_002	Lid_det_mod_req_004	Lid_sys_sim_test_003	under review
Lid_sys_sim_req_011	Lidar sensor system simulation shall not detect GT objects that are not visible due to lower EPW per detection and a low number of detections because of signal absorption by atmospheric aerosols.	Requirement	Lid_CEPRA_008	Lid_sys_sim_req_009, Lid_sys_sim_req_012, Lid_sys_sim_req_013	Lid_obj_mod_req_001, Lid_obj_mod_req_002	Lid_det_mod_req_005	Lid_sys_sim_test_004	under review

Lid_sys_sim_req_012	Atmospheric aerosol includes rain, spray, fog and snow.	Definition	/	Lid_sys_sim_req_010, Lid_sys_sim_req_011	/	/	/	under review
Lid_sys_sim_req_013	Rain intensity shall be from 0 mm/h to 42 mm/h as measured by the reference device Thies Clima 5.4110.10.	Definition	/	Lid_sys_sim_req_010, Lid_sys_sim_req_011	/	/	/	under review
• • •								

Table B-2: Lidar object model requirements*

Lidar object model requ. ID	**Description**	**Type**	**Related Lidar sensor system simulation requirement IDs**	**Related Lidar detection model requirement IDs**	**Acceptance test IDs**	**Status**
Lid_ obj_mod_ req_001	The lidar object model shall have a parameter that represents the minimum number of detections within the GT bounding box to detect an object build from these detections.	Requirement	Lid_sys_sim_req_001, Lid_sys_sim_req_002, Lid_sys_sim_req_003, Lid_sys_sim_req_004, Lid_sys_sim_req_005, Lid_sys_sim_req_010, Lid_sys_sim_req_011	Lid_det_mod_req_001, Lid_det_mod_req_002, Lid_det_mod_req_004, Lid_det_mod_req_005	Lid_sys_sim_test_001, Lid_sys_sim_test_002, Lid_sys_sim_test_003, Lid_sys_sim_test_004	under review
Lid_ obj_mod_ req_002	The lidar object model shall have a parameter that represents the minimum sum of EPW of detections within the GT bounding box to detec an object build from these detections.	Requirement	Lid_sys_sim_req_002, Lid_sys_sim_req_003, Lid_sys_sim_req_004, Lid_sys_sim_req_005, Lid_sys_sim_req_009, Lid_sys_sim_req_010, Lid_sys_sim_req_011	Lid_det_mod_req_003, Lid_det_mod_req_004, Lid_det_mod_req_005	Lid_sys_sim_test_002, Lid_sys_sim_test_003, Lid_sys_sim_test_004	under review
• • •						

* Kindly provided, not yet published intermediate project result of publicly funded research project VVM

Table B-3: Acceptance Tests for lidar sensor system simulation*

Acceptance test ID	Description	Lidar sensor system data ID	Reference data ID	Metrics and corresponding acceptance ranges	Related lidar sensor system simulation requ. IDs	Related lidar object model requ. IDs	Related lidar detection model requ. IDs	Status
Lid_sys_ sim_test_ 001	Simulation-only test with an occluded car behind a wall. Specification: The wall should cover the hole part of the sensor's field of view that would hit the occluded object by a combination of its dimensions, orientation and position. Expected results: No detections outputed by the lidar detections model on the occluded object. Object behind the wall is not detected by the lidar object model.	/	/	Lidar object model: Intersection over Union (IoU) of simulated (not detected) object and GT object: IoU = 0 Lidar detection model: Number of detections within GT bounding box N_det = 0	Lid_sys_sim _req_001, Lid_sys_sim _req_006, Lid_sys_sim _req_007, Lid_sys_sim _req_008	Lid_obj_ mod_req_ 001	Lid_det_ mod_req_ 001	accepted

* Kindly provided, not yet published intermediate project result of publicly funded research project VVM

Lid_sys_sim_test_002	Experiment with a partly occluded car partly behind a wall that is not identified by the object clustering while the sensor outputs some, but not many, lidar detections on it. Specification: The wall should be placed 20 m in front of the sensor, 90° degree turned w.r.t. the sensor. The car should be placed 10 m behind the wall, 90° degree turned w.r.t. the sensor. The back of the car should be visible in the sensor with a small number of detections, but there should not be detected an object by the sensor system at the car's back. Expected results: Less detections on the object compared to the non-occluded case. Detections from single beams that are only partly hitting the object (while partly hitting the wall) have lower echo-pulse width compared to the fully hitting case. The object partly behind the wall is not detected.	Lid_sys_data_001	Lid_sys_ref_data_001	Lidar object model: Intersection over Union (IoU) of simulated (not detected) object and GT object: IoU = 0 Intersection over Union (IoU) of measured (not detected) object and GT object: IoU = 0 Lidar detection model: Number of simulated detections within GT object's bounding box: N_det_GT_bb <10 Echo-pulse width of detections within GT object's bounding box, where only half of the beam hits the partly occluded object: EPW_det_half_beam <1 m	Lid_sys_sim_req_002, Lid_sys_sim_req_006, Lid_sys_sim_req_007, Lid_sys_sim_req_008, Lid_sys_sim_req_009	Lid_obj_mod_req_001, Lid_obj_mod_req_002	Lid_det_mod_req_002, Lid_det_mod_req_003	under review

Lid_sys_sim_test_003	Experiment with a car in heavy rain. Specification: The car should be placed 20 m in front of the sensor, 90° degree turned w.r.t. the sensor. Due to heavy rain, the car should be visible in the sensor with a small number of detections, but there should not be detected an object by the sensor system at the car. Expected results: Less detections on the object compared to the non-occluded case. The object is not detected.	Lid_sys_data_002	Lid_sys_ref_data_002	Lidar object model: Intersection over Union (IoU) of simulated (not detected) object and GT object: IoU = 0 Intersection over Union (IoU) of measured (not detected) object and GT object: IoU = 0 Lidar detection model: Number of simulated detections within GT object's bounding box: N_det_GT_bb <10	Lid_sys_sim_req_006, Lid_sys_sim_req_007, Lid_sys_sim_req_008, Lid_sys_sim_req_009, Lid_sys_sim_req_010	Lid_obj_mod_req_001, Lid_obj_mod_req_002	Lid_det_mod_req_004	under review
Lid_sys_sim_test_004	Experiment with a car in dense fog. Specification: The car should be placed 20 m in front of the sensor, 90° degree turned w.r.t. the sensor. Due to dense fog, the car should be visible in the sensor with a small number of detections, but there should not be detected an object by the sensor system at the car. Expected results: Less detections on the object compared to the non-occluded case. The object is not detected.	Lid_sys_data_003	Lid_sys_ref_data_003	Lidar object model: Intersection over Union (IoU) of simulated (not detected) object and GT object: IoU = 0 Intersection over Union (IoU) of measured (not detected) object and GT object: IoU = 0 Lidar detection model: Number of simulated detections within GT object's bounding box: N_det_GT_bb <10	Lid_sys_sim_req_006, Lid_sys_sim_req_007, Lid_sys_sim_req_008, Lid_sys_sim_req_009, Lid_sys_sim_req_011	Lid_obj_mod_req_001, Lid_obj_mod_req_002	Lid_det_mod_req_005	under review
• • •								

Table B-4: Lidar sensor system experiment data*

Lidar sensor system data ID	**Description**	**Reference data ID**	**Acceptance test ID**	**Status**
Lid_sys_data_001	Object list and detection list from an experiment with a partly occluded car behind a wall.	Lid_sys_ref_data_001	Lid_sys_sim_test_002	under review
Lid_sys_data_002	Object list and detection list from an experiment with a car in heavy rain.	Lid_sys_ref_data_002	Lid_sys_sim_test_003	under review
Lid_sys_data_003	Object list and detection list from an experiment with a car in dense fog.	Lid_sys_ref_data_003	Lid_sys_sim_test_004	under review
• • •				

[*] Kindly provided, not yet published intermediate project result of publicly funded research project VVM

Table B-5: Reference data requirements per lidar experiment*

Reference data ID	Measured values	Measurement device, its precision and frequency	Lidar sensor system data ID	Acceptance test ID	Status
Lid_sys_ref_data_001	Relative position of the car behind the wall Relative orientation of the car behind the wall Dimensions of the car behind the wall Relative position of the wall Relative orientation of the wall Dimensions of the wall	Genesys ADMA, 1.0 cm, 1.0 kHz Genesys ADMA, 0.1°, 1.0 kHz Laser rangefinder, 1.0 cm, - Laser rangefinder, 1.0 cm, - Laser rangefinder, 1.0 cm, - Laser rangefinder, 1.0 cm, -	Lid_sys_data_001	Lid_sys_sim_test_002	under review
Lid_sys_ref_data_002	Relative position of the car behind the wall Relative orientation of the car behind the wall Dimensions of the car behind the wall Rain intensity	Genesys ADMA, 1.0 cm, 1 kHz Genesys ADMA, 0,1°, 1 kHz Laser rangefinder, 1.0 cm, - t.b.d.	Lid_sys_data_002	Lid_sys_sim_test_003	under review
Lid_sys_ref_data_003	Relative position of the car behind the wall Relative orientation of the car behind the wall Dimensions of the car behind the wall Density of the fog	Genesys ADMA, 1.0 cm, 1 kHz Genesys ADMA, 0.1°, 1 kHz Laser rangefinder, 1.0 cm, - t.b.d.	Lid_sys_data_003	Lid_sys_sim_test_004	under review
• • •					

* Kindly provided, not yet published intermediate project result of publicly funded research project VVM

Bibliography

Ackermann, S. M.: Systematische Untersuchung von Radar Tracking (2017)
Ackermann, Stefan Martin: Systematische Untersuchung von Radar Tracking-Algorithmen, Master Thesis, Technische Universität Darmstadt, 2017

AIAG; VDA: FMEA Handbook - Failure Mode and Effects Analysis (2019)
AIAG; VDA: FMEA Handbook - Failure Mode and Effects Analysis, AIAG, 2019

Alldén, T. et al.: Virtual Generation of Lidar Data for Autonomous Vehicles (2017)
Alldén, Tobias; Chemander, Martin; Davar, Sherry; Jansson, Jonathan; Laurenius, Rickard; Tibom, Philip: Virtual Generation of Lidar Data for Autonomous Vehicles, Bachelor Thesis, University of Gothenburg, Chalmers University of Technology, 2017

Amersbach, C. et al.: Defining Required and Feasible Test Coverage (2019)
Amersbach, Christian; Winner, Hermann: Defining Required and Feasible Test Coverage for Scenario-Based Validation of Highly Automated Vehicles, in: 2019 IEEE 22nd International Conference on Intelligent Transportation Systems (ITSC), pp. 425–430, 2019

Amersbach, C. T.: Functional Decomposition (2020)
Amersbach, Christian Thomas: Functional Decomposition Approach - Reducing the Safety Validation Effort for Highly Automated Driving, PhD Thesis, Technische Universität Darmstadt, 2020

Amrhein, V. et al.: Scientists rise up against statistical significance (2019)
Amrhein, Valentin; Greenland, Sander; McShane, Blake: Scientists rise up against statistical significance, in: Nature, Vol. 567, pp. 305–307, 2019

Anderson, T. W. et al.: Asymptotic Theory of Certain ”Goodness of Fit” Criteria (1952)
Anderson, T. W.; Darling, D. A.: Asymptotic Theory of Certain ”Goodness of Fit” Criteria Based on Stochastic Processes, in: The Annals of Mathematical Statistics, Vol. 23, pp. 193–212, 1952

ASAM e.V.: ASAM OpenDRIVE® - Specification (2021)
ASAM e.V.: ASAM OpenDRIVE® - Specification, URL: `https://www.asam.net/standards/detail/opendrive/`, 2021, visited on 02/28/2022

ASAM e.V.: ASAM OpenODD - Concept Paper (2021)
ASAM e.V.: ASAM OpenODD - Concept Paper, URL: `https://www.asam.net/standards/detail/openodd/`, 2021, visited on 02/28/2022

ASAM e.V.: ASAM OpenSCENARIO® - User Guide (2021)
ASAM e.V.: ASAM OpenSCENARIO® - User Guide, URL: `https://www.asam.net/standards/detail/openscenario/`, 2021, visited on 02/28/2022

ASAM e.V.: ASAM OSI® (Open Simulation Interface) - Official Documentation (2022)
ASAM e.V.: ASAM OSI® (Open Simulation Interface) - Official Documentation, URL: `https://opensimulationinterface.github.io/osi-documentation/`, 2022, visited on 02/28/2022

AUDI AG: Laserscanner (2017)
AUDI AG: Laserscanner, in: Audi MediaCenter, URL: `https://www.audi-mediacenter.com/de/fotos/detail/laserscanner-52974`, 2017, visited on 03/23/2022

Aust, P.: Entwicklung eines lidartypischen Objektlisten-Sensormodells (2019)
Aust, Philip: Entwicklung eines lidartypischen Objektlisten-Sensormodells, Master Thesis, Technische Universität Darmstadt, 2019

AUTOSAR development cooperation: AUTOSAR (AUTomotive Open System ARchitecture) (2022)
AUTOSAR development cooperation: AUTOSAR (AUTomotive Open System ARchitecture), URL: `https://www.autosar.org/standards/`, 2022, visited on 04/27/2022

AVL List GmbH: ENABLE-S3 Project (2019)
AVL List GmbH: ENABLE-S3 - European Initiative to Enable Validation for Highly Automated Safe and Secure Systems, URL: `https://enable-s3.eu/`, 2019, visited on 04/05/2022

Balci, O. et al.: Cost-Risk Analysis in the Statistical Validation of Simulation Models (1981)
Balci, Osman; Sargent, Robert G.: A Methodology for Cost-Risk Analysis in the Statistical Validation of Simulation Models, in: Communications of the ACM, Vol. 24, pp. 190–197, 1981

Bechtold, S. et al.: HELIOS (2016)
Bechtold, S.; Höfle, B.: HELIOS: A MULTI-PURPOSE LIDAR SIMULATION FRAMEWORK FOR RESEARCH, PLANNING AND TRAINING OF LASER SCANNING OPERATIONS WITH AIRBORNE, GROUND-BASED MOBILE AND STATIONARY PLATFORMS, in: ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. III-3, pp. 161–168, 2016

Berghöfer, M.: Generierung realer und synthetischer Sensordaten zur Simulations-Validierung (2019)
Berghöfer, Moritz: Generierung realer und synthetischer Sensordaten zur Validierung von Sensormodellen für die simulationsbasierte Absicherung der Valet Parking Funktion, Bachelor Thesis, Technische Universität Darmstadt, 2019

Bernsteiner, S. et al.: Radar Sensor Model for the Virtual Development Process (2015)
Bernsteiner, Stefan; Magosi, Zoltan; Lindvai-Soos, Daniel; Eichberger, Arno: Radar Sensor Model for the Virtual Development Process, in: ATZelektronik worldwide, Vol. 10, pp. 46–52, 2015

Bühren, M. et al.: A Global Motion Model for Target Tracking in Automotive Applications (2007)

Bühren, Markus; Yang, Bin: A Global Motion Model for Target Tracking in Automotive Applications, in: 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07, vol. 2, pp. II–313–II–316, 2007

Bühren, M. et al.: Automotive Radar Target List Simulation based on Reflection Centers (2006)

Bühren, Markus; Yang, Bin: Automotive Radar Target List Simulation based on Reflection Center Representation of Objects, in: Proc. Intern. Workshop on Intelligent Transportation (WIT), pp. 161–166, 2006

Bühren, M. et al.: Extension of Automotive Radar Target List Simulation (2007)

Bühren, Markus; Yang, Bin: Extension of Automotive Radar Target List Simulation to consider further Physical Aspects, in: 2007 7th International Conference on ITS Telecommunications, pp. 1–6, 2007

Bühren, M. et al.: Initialization Procedure for Radar Target Tracking (2007)

Bühren, Markus; Yang, Bin: Initialization Procedure for Radar Target Tracking without Object Movement Constraints, in: 2007 7th International Conference on ITS Telecommunications, pp. 1–6, 2007

Bühren, M. et al.: Simulation of Automotive Radar Target Lists: Clutter and Resolution (2007)

Bühren, Markus; Yang, Bin: Simulation of Automotive Radar Target Lists considering Clutter and Limited Resolution, in: 2007 8th International Radar Symposium (IRS), pp. 195–200, 2007

Bühren, M. et al.: Simulation of Automotive Radar Target Lists: Novel Approach of Object (2006)

Bühren, Markus; Yang, Bin: Simulation of Automotive Radar Target Lists using a Novel Approach of Object Representation, in: 2006 IEEE Intelligent Vehicles Symposium (IV), pp. 314–319, 2006

Calaprice, A.: The Ultimate Quotable Einstein (2010)

Calaprice, Alice: The Ultimate Quotable Einstein, Princeton University Press, 2010

Cao, P.: Modeling Active Perception Sensors for Real-Time Virtual Validation (2018)

Cao, Peng: Modeling Active Perception Sensors for Real-Time Virtual Validation of Automated Driving Systems, PhD Thesis, Technische Universität Darmstadt, 2018

Cao, P. et al.: Perception sensor modeling for virtual validation of automated driving (2015)

Cao, Peng; Wachenfeld, Walther; Winner, Hermann: Perception sensor modeling for virtual validation of automated driving, in: it - Information Technology, Vol. 57, 2015

Crespo, L. G. et al.: Interval predictor models with a formal characterization of uncertainty and reliability (2014)
Crespo, Luis G.; Giesy, Daniel P.; Kenny, Sean P.: Interval predictor models with a formal characterization of uncertainty and reliability, in: 53rd IEEE Conference on Decision and Control, pp. 5991–5996, 2014

Danielsson, L.: Tracking and radar sensor modelling for automotive safety systems (2010)
Danielsson, Lars: Tracking and radar sensor modelling for automotive safety systems, PhD Thesis, Chalmers University of Technology, 2010

Danquah, B.: Zuverlässigkeitsbestimmung von Fahrzeugsimulationen durch statistische Validierung (2022)
Danquah, Benedikt: Zuverlässigkeitsbestimmung von Gesamtfahrzeugsimulationen durch statistische Validierung, PhD Thesis, Technische Universität München, 2022

Danquah, B. et al.: Modular, Open Source Simulation Approach (2019)
Danquah, Benedikt; Koch, Alexander; Weiß, Tony; Lienkamp, Markus; Pinnel, André: Modular, Open Source Simulation Approach: Application to Design and Analyze Electric Vehicles, in: 2019 Fourteenth International Conference on Ecological Vehicles and Renewable Energies (EVER), pp. 1–8, 2019

Danquah, B. et al.: Potential of statistical model VV&UQ in vehicle dynamics simulations (2020)
Danquah, Benedikt; Riedmaier, Stefan; Lienkamp, Markus: Potential of statistical model verification, validation and uncertainty quantification in automotive vehicle dynamics simulations: a review, in: Vehicle System Dynamics, Vol. 60, pp. 1292–1321, 2020

Danquah, B. et al.: Statistical Validation Framework for Automotive Vehicle Simulations (2021)
Danquah, Benedikt; Riedmaier, Stefan; Meral, Yasin; Lienkamp, Markus: Statistical Validation Framework for Automotive Vehicle Simulations Using Uncertainty Learning, in: MDPI Applied Sciences, Vol. 11, p. 1983, 2021

Danquah, B. et al.: Statistical Model Verification and Validation Concept (2020)
Danquah, Benedikt; Riedmaier, Stefan; Rühm, Johannes; Kalt, Svenja; Lienkamp, Markus: Statistical Model Verification and Validation Concept in Automotive Vehicle Design, in: Procedia CIRP, Vol. 91, pp. 261–270, 2020

Degen, R. et al.: Methodical Approach to the Development of a Radar Sensor Model (2021)
Degen, Rene; Ott, Harry; Overath, Fabian; Schyr, Christian; Leijon, Mats; Ruschitzka, Margot: Methodical Approach to the Development of a Radar Sensor Model for the Detection of Urban Traffic Participants Using a Virtual Reality Engine, in: Journal of Transportation Technologies, Vol. 11, pp. 179–195, 2021

Deutsches Zentrum für Luft- und Raumfahrt e.V. (DLR): PEGASUS Project (2019)
Deutsches Zentrum für Luft- und Raumfahrt e.V. (DLR): PEGASUS - Projekt zur Etablierung von generell akzeptierten Gütekriterien, Werkzeugen und Methoden sowie Szenarien und Situationen zur Freigabe hochautomatisierter Fahrfunktionen, URL: `https://www.pegasusprojekt.de/en`, 2019, visited on 04/05/2022

Deutsches Zentrum für Luft- und Raumfahrt e.V. (DLR): SET Level Project (2022)
Deutsches Zentrum für Luft- und Raumfahrt e.V. (DLR): SET Level - Simulationsbasiertes Entwickeln und Testen von automatisiertem Fahren, URL: `https://setlevel.de/en`, 2022, visited on 04/05/2022

Dietmayer, K.: Predicting of Machine Perception for Automated Driving (2016)
Dietmayer, Klaus: Predicting of Machine Perception for Automated Driving, in: Maurer, Markus; Gerdes, J. Christian; Lenz, Barbara; Winner, Hermann (Hrsg.): Autonomous Driving, Springer, 2016

Digital Imaging and Remote Sensing Lab: DIRSIG (2022)
Digital Imaging and Remote Sensing Lab: DIRSIG - Digital Imaging and Remote Sensing Image Generation, URL: `http://dirsig.org/index.html`, 2022, visited on 01/09/2022

Doria, D.: A Synthetic LiDAR Scanner for VTK (2009)
Doria, David: A Synthetic LiDAR Scanner for VTK, in: The VTK Journal, 2009

Doria, D.: SyntheticLidarScanner (2021)
Doria, David: SyntheticLidarScanner, URL: `https://github.com/daviddoria/SyntheticLidarScanner`, 2021, visited on 01/09/2022

Dupuis, M.: Paving the way for certified performance (2022)
Dupuis, Marius: Paving the way for certified performance: Quality assessment and rating of simulation solutions for ADAS and autonomous driving, in: Electronic Imaging, Vol. 34, pp. 1–6, 2022

Eder, T.: Simulation of Automotive Radar Point Clouds in Standardized Frameworks (2021)
Eder, Thomas: Simulation of Automotive Radar Point Clouds in Standardized Frameworks, PhD Thesis, Technische Universität München, 2021

Eder, T. et al.: Data Driven Radar Detection Models (2019)
Eder, Thomas; Hachicha, Rami; Sellami, Houssem; Driesten, Carlo van; Biebl, Erwin: Data Driven Radar Detection Models: A Comparison of Artificial Neural Networks and Non Parametric Density Estimators on Synthetically Generated Radar Data, in: 2019 Kleinheubach Conference, pp. 1–4, 2019

Eder, T. et al.: Szenarienbasierte Validierung eines hybriden Radarmodells (2020)
Eder, Thomas; Prinz, Alexander; Brabetz, Ludwig; Biebl, Erwin: Szenarienbasierte Validierung eines hybriden Radarmodells für Test und Absicherung automatisierter Fahrfunktionen, in: Tille, Thomas (Hrsg.): Automobil-Sensorik 3: Prinzipien, Technologien und Anwendungen, Springer Berlin Heidelberg, 2020

Eek, M. et al.: Definition and Implementation of a Method for Uncertainty Aggregation (2017)
Eek, Magnus; Gavel, Hampus; Ölvander, Johan: Definition and Implementation of a Method for Uncertainty Aggregation in Component-Based System Simulation Models, in: Journal of Verification, Validation and Uncertainty Quantification, Vol. 2, 2017

El Mostadi, M. et al.: Seven Technical Issues That May Ruin Your Virtual Tests for ADAS (2021)
El Mostadi, Mohamed; Waeselynck, Helene; Gabriel, Jean-Marc: Seven Technical Issues That May Ruin Your Virtual Tests for ADAS, in: 2021 IEEE Intelligent Vehicles Symposium (IV), p. 6, 2021

Elster, L. et al.: Reflection Based Radar Object Model (2022)
Elster, Lukas; Linnhoff, Clemens; Rosenberger, Philipp: Reflection Based Radar Object Model, URL: `https://gitlab.com/tuda-fzd/perception-sensor-modeling/reflection-based-radar-object-model`, 2022, visited on 03/02/2022

Elster, L. et al.: Fundamental Design Criteria for Logical Scenarios (2021)
Elster, Lukas; Linnhoff, Clemens; Rosenberger, Philipp; Schmidt, Simon; Stark, Rainer; Winner, Hermann: Fundamental Design Criteria for Logical Scenarios in Simulation-based Safety Validation of Automated Driving Using Sensor Model Knowledge, in: IV Workshop on Ensuring and Validating Safety for Automated Vehicles, 2021

Emery, A. F.: Special Issue: Sandia V&V Challenge Problem (2016)
Emery, Ashley F.: Special Issue: Sandia V&V Challenge Problem, in: Journal of Verification, Validation and Uncertainty Quantification, Vol. 1, 2016

European Center for Information and Communication Technologies – EICT GmbH: VVM Project (2022)
European Center for Information and Communication Technologies – EICT GmbH: VVM - Verification & Validation Methods, URL: `https://www.vvm-projekt.de/en`, 2022, visited on 04/05/2022

Fang, J. et al.: Augmented LiDAR Simulator for Autonomous Driving (2020)
Fang, Jin; Zhou, Dingfu; Yan, Feilong; Zhao, Tongtong; Zhang, Feihu; Ma, Yu; Wang, Liang; Yang, Ruigang: Augmented LiDAR Simulator for Autonomous Driving, in: IEEE Robotics and Automation Letters, Vol. 5, pp. 1931–1938, 2020

Fasano, G. et al.: A multidimensional version of the Kolmogorov–Smirnov test (1987)
Fasano, G.; Franceschini, A.: A multidimensional version of the Kolmogorov–Smirnov test, in: Monthly Notices of the Royal Astronomical Society, Vol. 225, pp. 155–170, 1987

Ferson, S. et al.: Constructing Probability Boxes and Dempster-Shafer Structures (2003)
Ferson, Scott; Kreinovick, Vladik; Ginzburg, Lev; Myers, Davis S.; Sentz, Kari: Constructing Probability Boxes and Dempster-Shafer Structures, URL: `https://www.semanticscholar.org/paper/Constructing-Probability-Boxes-and-Dempster-Shafer-Ferson-Kreinovick/8eff743341521cca30f6d2a48df50bf6977c96b2`, 2003

Ferson, S. et al.: Validation of imprecise probability models (2009)
Ferson, Scott; Oberkampf, William: Validation of imprecise probability models, in: International Journal of Reliability and Safety, Vol. 3, 2009

Ferson, S. et al.: Model validation and predictive capability (2008)
Ferson, Scott; Oberkampf, William L.; Ginzburg, Lev: Model validation and predictive capability for the thermal challenge problem, in: Computer Methods in Applied Mechanics and Engineering, Vol. 197, pp. 2408–2430, 2008

Fréchet, M. R.: Sur quelques points du calcul fonctionnel (1906)
Fréchet, Maurice René: Sur quelques points du calcul fonctionnel, in: Rendiconti del Circolo Matematico di Palermo (1884-1940), Vol. 22, pp. 1–72, 1906

Gomes, T. et al.: Evaluation and Testing Platform for Automotive LiDAR Sensors (2021)
Gomes, Tiago; Roriz, Ricardo; Cunha, Luís; Ganal, Andreas; Soares, Narciso; Araújo, Teresa; Monteiro, João: Evaluation and Testing Platform for Automotive LiDAR Sensors, in: Preprints, 2021

Goodin, C. et al.: Sensor modeling for the Virtual Autonomous Navigation Environment (2009)
Goodin, Chris; Kala, Raju; Carrrillo, Alex; Liu, Linda Y.: Sensor modeling for the Virtual Autonomous Navigation Environment, in: 2009 IEEE SENSORS Conference, pp. 1588–1592, 2009

Gotzig, H. et al.: Automotive LIDAR (2016)
Gotzig, Heinrich; Geduld, Georg: Automotive LIDAR, in: Winner, Hermann; Hakuli, Stephan; Lotz, Felix; Singer, Christina (Hrsg.): Handbook of Driver Assistance Systems: Basic Information, Components and Systems for Active Safety and Comfort, Springer International Publishing, 2016

Green Car Congress: Quanergy S3 Operation Principle (2016)
Green Car Congress: Quanergy S3 Operation Principle, in: Green Car Congress, URL: `https://www.greencarcongress.com/2016/08/20160823-quanergy.html`, 2016, visited on 03/24/2022

Gschwandtner, M.: Support framework for obstacle detection on autonomous trains (2013)
Gschwandtner, Michael: Support framework for obstacle detection on autonomous trains, PhD Thesis, Universität Salzburg, 2013

Gschwandtner, M. et al.: BlenSor (2011)
Gschwandtner, Michael; Kwitt, Roland; Uhl, Andreas; Pree, Wolfgang: BlenSor: Blender Sensor Simulation Toolbox, in: Advances in Visual Computing, pp. 199–208, 2011

Gubelli, D. et al.: Ray-Tracing Simulator for Radar Signals Propagation in Radar Networks (2013)
Gubelli, Demetrio; Krasnov, Oleg A; Yarovyi, Olexander: Ray-Tracing Simulator for Radar Signals Propagation in Radar Networks, in: Proceedings of the 10th European Radar Conference, pp. 73–76, 2013

Gupta, A. et al.: Photon-Flooded Single-Photon 3D Cameras (2019)
Gupta, Anant; Ingle, Atul; Velten, Andreas; Gupta, Mohit: Photon-Flooded Single-Photon 3D Cameras, URL: `http://arxiv.org/abs/1903.08347`, 2019, visited on 03/14/2022

Hadelli, A. A.: Messkampagne zur Parametrisierung und Validierung von Lidar-Sensor-Modellen (2020)
Hadelli, Ali Adel: Messkampagne zur Parametrisierung und Validierung von Lidar-Sensor-Modellen, Bachelor Thesis, Technische Universität Darmstadt, 2020

Hakuli, S. et al.: Virtual Integration in the Development Process of ADAS (2016)
Hakuli, Stephan; Krug, Markus: Virtual Integration in the Development Process of ADAS, in: Winner, Hermann; Hakuli, Stephan; Lotz, Felix; Singer, Christina (Hrsg.): Handbook of Driver Assistance Systems: Basic Information, Components and Systems for Active Safety and Comfort, Springer International Publishing, 2016

Hammarstrand, L. et al.: Adaptive Radar Sensor Model for Tracking Structured Extended Objects (2012)
Hammarstrand, Lars; Lundgren, Malin; Svensson, Lennart: Adaptive Radar Sensor Model for Tracking Structured Extended Objects, in: IEEE Transactions on Aerospace and Electronic Systems, Vol. 48, pp. 1975–1995, 2012

Hammarstrand, L. et al.: Extended Object Tracking using a Radar Resolution Model (2012)
Hammarstrand, Lars; Svensson, Lennart; Sandblom, Fredrik; Sorstedt, Joakim: Extended Object Tracking using a Radar Resolution Model, in: IEEE Transactions on Aerospace and Electronic Systems, Vol. 48, pp. 2371–2386, 2012

Hanke, T. et al.: Generic architecture for simulation of ADAS sensors (2015)
Hanke, T.; Hirsenkorn, N.; Dehlink, B.; Rauch, A.; Rasshofer, R.; Biebl, E.: Generic architecture for simulation of ADAS sensors, in: 2015 16th International Radar Symposium (IRS), pp. 125–130, 2015

Hanke, T.: Simulated Environmental Perception for Automated Driving Systems (2020)
Hanke, Timo: Virtual Sensorics: Simulated Environmental Perception for Automated Driving Systems, PhD Thesis, Technische Universität München, 2020

Hanke, T. et al.: Open Simulation Interface (2017)
Hanke, Timo; Hirsenkorn, Nils; Driesten, Carlo van; Garcia Ramos, Pilar; Schiementz, Mark; Schneider, Sebastian; Biebl, Erwin: Open Simulation Interface: A generic interface for the environment perception of automated driving functions in virtual scenarios. URL: `https://www.hot.ei.tum.de/forschung/automotive-veroeffentlichungen/`, 2017

Hanke, T. et al.: A generic interface for the environment perception of automated driving functions (2017)
Hanke, Timo; Hirsenkorn, Nils; Driesten, Carlo van; Garcia-Ramos, Pilar; Schiementz, Mark; Biebl, Erwin: A generic interface for the environment perception of automated driving functions in virtual scenarios, URL: `https://www.ei.tum.de/hot/forschung/automotive-veroeffentlichungen/`, 2017, visited on 01/25/2020

Hanke, T. et al.: Generation and validation of virtual point cloud data for automated driving systems (2017)
Hanke, Timo; Schaermann, Alexander; Geiger, Matthias; Weiler, Konstantin; Hirsenkorn, Nils; Rauch, Andreas; Schneider, Stefan-Alexander; Biebl, Erwin: Generation and validation of virtual point cloud data for automated driving systems, in: 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC), 2017

Hasirlioglu, S.: Simulation-based Testing of Surround Sensors under Adverse Weather (2020)
Hasirlioglu, Sinan: A Novel Method for Simulation-based Testing and Validation of Automotive Surround Sensors under Adverse Weather Conditions, PhD Thesis, Universität Linz, 2020

Hayward, J. C.: NEAR-MISS DETERMINATION THROUGH USE OF A SCALE OF DANGER (1972)
Hayward, John C: NEAR-MISS DETERMINATION THROUGH USE OF A SCALE OF DANGER, in: Highway Research Record, p. 12, 1972

Hills, R. G.: Roll-up of validation results to a target application. (2013)
Hills, Richard Guy: Roll-up of validation results to a target application. URL: `https://www.osti.gov/biblio/1096465`, 2013, visited on 04/07/2022

Hinsemann, T.: Analyse von Effekten in Lidardaten für die virtuelle Absicherung (2021)
Hinsemann, Timo: Analyse von Effekten in Lidardaten für die virtuelle Absicherung automatisierter Fahrfunktionen, Bachelor Thesis, Technische Universität Darmstadt, 2021

Hirsenkorn, N.: Modellbildung und Simulation der Fahrzeugumfeldsensorik (2018)
Hirsenkorn, Nils: Modellbildung und Simulation der Fahrzeugumfeldsensorik, PhD Thesis, Technische Universität München, 2018

Hirsenkorn, N. et al.: A ray launching approach for modeling an FMCW radar system (2017)
Hirsenkorn, Nils; Subkowski, Paul; Hanke, Timo; Schaermann, Alexander; Rauch, Andreas; Rasshofer, Ralph; Biebl, Erwin: A ray launching approach for modeling an FMCW radar system, in: 2017 18th International Radar Symposium (IRS), pp. 1–10, 2017

Holder, M. F.: Synthetic Generation of Radar Sensor Data for Virtual Validation (2021)
Holder, Martin Friedrich: Synthetic Generation of Radar Sensor Data for Virtual Validation of Autonomous Driving, PhD Thesis, Technische Universität Darmstadt, 2021

Holder, M. F. et al.: Digitalize the Twin (2022)
Holder, Martin Friedrich; Elster, Lukas; Winner, Hermann: Digitalize the Twin: A Method for Calibration of Reference Data for Transfer Real-World Test Drives into Simulation, in: MDPI Energies, Vol. 15, p. 989, 2022

Holder, M. F. et al.: Modeling and Simulation of Radar Sensor Artifacts (2019)
Holder, Martin Friedrich; Linnhoff, Clemens; Rosenberger, Philipp; Popp, Christoph; Winner, Hermann: Modeling and Simulation of Radar Sensor Artifacts for Virtual Testing of Autonomous Driving, in: 9. Tagung Automatisiertes Fahren, 2019

Holder, M. F. et al.: The Fourier Tracing Approach for Modeling Automotive Radar Sensors (2019)
Holder, Martin Friedrich; Linnhoff, Clemens; Rosenberger, Philipp; Winner, Hermann: The Fourier Tracing Approach for Modeling Automotive Radar Sensors, in: 2019 20th International Radar Symposium (IRS), p. 8, 2019

Holder, M. F. et al.: Data-driven Derivation of Requirements for a Lidar Sensor Model (2018)
Holder, Martin Friedrich; Rosenberger, Philipp; Bert, Felix; Winner, Hermann: Data-driven Derivation of Requirements for a Lidar Sensor Model, in: 11th Grazer Symposium Virtuelles Fahrzeug (GSVF), 2018

Holder, M. F. et al.: Measurements revealing Challenges in Radar Sensor Modeling (2018)
Holder, Martin Friedrich; Rosenberger, Philipp; Winner, Hermann; D'hondt, Thomas; Makkapati, Vamsi Prakash; Maier, Michael; Schreiber, Helmut; Magosi, Zoltan; Slavik, Zora; Bringmann, Oliver; Rosenstiel, Wolfgang: Measurements revealing Challenges in Radar Sensor Modeling for Virtual Validation of Autonomous Driving, in: 2018 IEEE 21st International Conference on Intelligent Transportation Systems (ITSC), pp. 2616–2622, 2018

Holder, M. F. et al.: How to evaluate synthetic radar data? (2020)
Holder, Martin Friedrich; Thielmann, Jan; Rosenberger, Philipp; Linnhoff, Clemens; Winner, Hermann: How to evaluate synthetic radar data? Lessons learned from finding driveable space in virtual environments, in: 13. Uni-DAS e.V. Workshop Fahrerassistenz und automatisiertes Fahren 2020, 2020

Hu, K. T. et al.: Introduction: The 2014 Sandia Verification and Validation Challenge Workshop (2016)
Hu, Kenneth T.; Carnes, Brian; Romero, Vicente: Introduction: The 2014 Sandia Verification and Validation Challenge Workshop, in: Journal of Verification, Validation and Uncertainty Quantification, Vol. 1, 2016

Huch, S.: Metrik zur Bewertung der Lidar-Sensor-Simulation (2018)
Huch, Sebastian: Entwicklung einer umfassenden Metrik für die Bewertung einer Lidar-Sensor-Simulation durch Betrachtung mehrerer aufeinander folgender Verarbeitungsebenen, Master Thesis, Technische Universität Darmstadt, 2018

Ibeo Automotive Systems GmbH: Solid state LiDAR sensor (2022)
Ibeo Automotive Systems GmbH: ibeoNEXT Solid-State Lidar, in: LiDAR sensor technology for autonomous driving - Ibeo, URL: `https://www.ibeo-as.com/en/products/sensors/ibeoNEXT`, 2022, visited on 03/23/2022

International Alliance for Mobility Testing and Standardization: IAMTS0001202104 (2021)
International Alliance for Mobility Testing and Standardization: IAMTS0001202104 Best Practice - A Comprehensive Approach for the Validation of Virtual Testing Toolchains, URL: `https://iamts.sae-itc.com/binaries/content/assets/itc/content/iamts/iamts0001202104.pdf`, 2021, visited on 03/14/2022

International Organization for Standardization: ISO 19364:2016(E) (2016)
International Organization for Standardization: ISO 19364:2016(E): Passenger cars — Vehicle dynamic simulation and validation — Steady-state circular driving behaviour, 2016

International Organization for Standardization: ISO 19365:2016(E) (2016)
International Organization for Standardization: ISO 19365:2016(E): Passenger cars — Validation of vehicle dynamic simulation — Sine with dwell stability control testing, 2016

International Organization for Standardization: ISO 19585:2019(E) (2019)
International Organization for Standardization: ISO 19585:2019(E): Heavy commercial vehicles and buses — Vehicle dynamics simulation and validation — Steady-state circular driving behavior, 2019

International Organization for Standardization: ISO 22140:2021(E) (2021)
International Organization for Standardization: ISO 22140:2021(E): Passenger cars — Validation of vehicle dynamics simulation — Lateral transient response test methods, 2021

International Organization for Standardization: ISO 23150:2021(E) (2021)
International Organization for Standardization: ISO 23150:2021(E): Road vehicles - Data communication between sensors and data fusion unit for automated driving functions - Logical interface, 2021

International Organization for Standardization: ISO 5725-1 (1994)
International Organization for Standardization: ISO 5725-1:1994: Accuracy (trueness and precision) of measurement methods and results — Part 1: General principles and definitions, 1994

Jiang, X. et al.: Bayesian wavelet method for multivariate model assessment (2008)
Jiang, Xiaomo; Mahadevan, Sankaran: Bayesian wavelet method for multivariate model assessment of dynamic systems, in: Journal of Sound and Vibration, Vol. 312, pp. 694–712, 2008

Junietz, P. M.: Microscopic and Macroscopic Risk Metrics (2019)
Junietz, Philipp Matthias: Microscopic and Macroscopic Risk Metrics for the Safety Validation of Automated Driving, PhD Thesis, Technische Universität Darmstadt, 2019

Kang, Y. et al.: Test Your Self-Driving Algorithm (2019)
Kang, Yue; Yin, Hang; Berger, Christian: Test Your Self-Driving Algorithm: An Overview of Publicly Available Driving Datasets and Virtual Testing Environments, in: IEEE Transactions on Intelligent Vehicles, Vol. 4, pp. 171–185, 2019

Lacerda, M. J. et al.: Interval predictor models for data with measurement uncertainty (2017)
Lacerda, Márcio J.; Crespo, Luis G.: Interval predictor models for data with measurement uncertainty, in: 2017 American Control Conference (ACC), pp. 1487–1492, 2017

Li, Y. et al.: LiDAR Sensor Modeling for ADAS Applications under a Virtual Driving Environment (2016)
Li, Yaxin; Wang, Ying; Deng, Weiwen; Li, Xin; liu, Zhenyi; Jiang, Lijun: LiDAR Sensor Modeling for ADAS Applications under a Virtual Driving Environment, in: SAE-TONGJI 2016 Driving Technology of Intelligent Vehicle Symposium, pp. 2016–01–1907, 2016

Li, Y. et al.: Lidar for Autonomous Driving (2020)
Li, You; Ibanez-Guzman, Javier: Lidar for Autonomous Driving: The Principles, Challenges, and Trends for Automotive Lidar and Perception Systems, in: IEEE Signal Processing Magazine, Vol. 37, pp. 50–61, 2020

Likert, R.: A technique for the measurement of attitudes (1932)
Likert, Rensis: A technique for the measurement of attitudes, in: Archives of Psychology, Vol. 22, p. 55, 1932

Linnhoff, C.: Entwicklung eines Radar-Sensormodells (2018)
Linnhoff, Clemens: Entwicklung eines Radar-Sensormodells, Master Thesis, Technische Universität Darmstadt, 2018

Linnhoff, C. et al.: PerCollECT - LidarLimbs (2022)
Linnhoff, Clemens; Hinsemann, Timo; Rosenberger, Philipp; Elster, Lukas: PerCollECT - LidarLimbs, URL: `https://github.com/PerCollECT/LidarLimbs`, 2022, visited on 03/10/2022

Linnhoff, C. et al.: Object Based Generic Perception Object Model (2022)
Linnhoff, Clemens; Rosenberger, Philipp; Elster, Lukas: Object Based Generic Perception Object Model, URL: `https://gitlab.com/tuda-fzd/perception-sensor-modeling/object-based-generic-perception-object-model`, 2022, visited on 03/02/2022

Linnhoff, C. et al.: Reflection Based Lidar Object Model · Environmental Effects (2022)
Linnhoff, Clemens; Rosenberger, Philipp; Elster, Lukas; Hofrichter, Kristof: Reflection Based Lidar Object Model · Environmental Effects, in: GitLab, URL: `https://gitlab.com/tuda-fzd/perception-sensor-modeling/reflection-based-lidar-object-model/-/tree/master/src/model/strategies/lidar-environmental-effects-strategy`, 2022, visited on 06/15/2022

Linnhoff, C. et al.: Highly Parameterizable Perception Sensor Model Architecture (2021)
Linnhoff, Clemens; Rosenberger, Philipp; Holder, Martin Friedrich; Cianciaruso, Nicodemo; Winner, Hermann: Highly Parameterizable and Generic Perception Sensor Model Architecture - A Modular Approach for Simulation Based Safety Validation of Automated Driving, in: Automatisiertes Fahren 2020, pp. 195–206, 2021

Linnhoff, C. et al.: Towards Serious Sensor Simulation for Safety Validation of Automated Driving (2021)
Linnhoff, Clemens; Rosenberger, Philipp; Schmidt, Simon; Elster, Lukas; Stark, Rainer; Winner, Hermann: Towards Serious Perception Sensor Simulation for Safety Validation of Automated Driving - A Collaborative Method to Specify Sensor Models, in: 2021 IEEE 24th International Conference on Intelligent Transportation Systems (ITSC), pp. 2688–2695, 2021

Linnhoff, C. et al.: Refining Object-Based Lidar Sensor Modeling (2021)
Linnhoff, Clemens; Rosenberger, Philipp; Winner, Hermann: Refining Object-Based Lidar Sensor Modeling - Challenging Ray Tracing as the Magic Bullet, in: IEEE Sensors Journal, Vol. 21, pp. 24238–24245, 2021

Liu, F. et al.: The Management of Simulation Validation (2019)
Liu, Fei; Yang, Ming: The Management of Simulation Validation, in: Beisbart, Claus; Saam, Nicole J. (Hrsg.): Computer Simulation Validation: Fundamental Concepts, Methodological Frameworks, and Philosophical Perspectives, Springer International Publishing, 2019

Liu, Y. et al.: Toward a Better Understanding of Model Validation Metrics (2011)
Liu, Yu; Chen, Wei; Arendt, Paul; Huang, Hong-Zhong: Toward a Better Understanding of Model Validation Metrics, in: Journal of Mechanical Design, Vol. 133, 2011

Magosi, Z. F. et al.: Evaluation of Physical Radar Perception Sensor Models (2022)
Magosi, Zoltan Ferenc; Wellershaus, Christoph; Tihanyi, Viktor Roland; Luley, Patrick; Eichberger, Arno: Evaluation Methodology for Physical Radar Perception Sensor Models Based on On-Road Measurements for the Testing and Validation of Automated Driving, in: MDPI Energies, Vol. 15, p. 20, 2022

Maier, F. M. et al.: Environment perception simulation for radar stimulation (2018)
Maier, F. Michael; Makkapati, Vamsi P.; Horn, Martin: Environment perception simulation for radar stimulation in automated driving function testing, in: e & i Elektrotechnik und Informationstechnik, Vol. 135, pp. 309–315, 2018

Maier, F. M.: Radar Perception Simulation for Automated Driving Tests (2022)
Maier, Franz Michael: Radar Perception Simulation for Automated Driving Tests, PhD Thesis, Graz University of Technology, 2022

Matthews, R.: The p-value statement, five years on (2021)
Matthews, Robert: The p-value statement, five years on, in: Significance, Vol. 18, pp. 16–19, 2021

Maupin, K. A. et al.: Validation Metrics for Deterministic and Probabilistic Data (2019)
Maupin, Kathryn A.; Swiler, Laura P.; Porter, Nathan W.: Validation Metrics for Deterministic and Probabilistic Data, in: Journal of Verification, Validation and Uncertainty Quantification, Vol. 3, 2019

Menditto, A. et al.: Understanding the meaning of accuracy, trueness and precision (2007)
Menditto, Antonio; Patriarca, Marina; Magnusson, Bertil: Understanding the meaning of accuracy, trueness and precision, in: Accreditation and Quality Assurance, Vol. 12, pp. 45–47, 2007

Menzel, T. et al.: Scenarios for Development, Test and Validation of Automated Vehicles (2018)
Menzel, Till; Bagschik, Gerrit; Maurer, Markus: Scenarios for Development, Test and Validation of Automated Vehicles, in: 2018 IEEE Intelligent Vehicles Symposium (IV), pp. 1821–1827, 2018

Mesow, L.: Multisensorielle Datensimulation im Fahrzeugumfeld für die Bewertung von Sensorfusion (2007)
Mesow, Lars: Multisensorielle Datensimulation im Fahrzeugumfeld für die Bewertung von Sensorfusionsalgorithmen, PhD Thesis, Technische Universität Chemnitz, 2007

Miethaner, C. et al.: Virtual homologation of an ALKS according to UNECE R157 (2022)
Miethaner, Christoph; Stavesand, Jann-Eve: Virtual homologation of an ALKS according to UNECE R157, URL: `https://www.tuvsud.com/en/-/media/global/pdf-files/whitepaper-report-e-books/tuvsud_virtual-homologation-of-an-alks-according-to-unece-r157.pdf`, 2022, visited on 04/08/2022

Modelica Association: Functional Mock-up Interface (2022)
Modelica Association: Functional Mock-up Interface, URL: `https://fmi-standard.org/`, 2022, visited on 04/28/2022

Muckenhuber, S. et al.: Automotive Lidar Modelling Approach Based on Material Properties and Lidar Capabilities (2020)
Muckenhuber, Stefan; Holzer, Hannes; Bockaj, Zrinka: Automotive Lidar Modelling Approach Based on Material Properties and Lidar Capabilities, in: MDPI Sensors, Vol. 20, p. 3309, 2020

Muckenhuber, S. et al.: Object-based sensor model for virtual testing of ADAS/AD functions (2019)
Muckenhuber, Stefan; Holzer, Hannes; Rubsam, Jonas; Stettinger, Georg: Object-based sensor model for virtual testing of ADAS/AD functions, in: 2019 IEEE International Conference on Connected Vehicles and Expo (ICCVE), 2019

Neumann-Cosel, K. von: Virtual Test Drive (2014)
Neumann-Cosel, Kilian von: Virtual Test Drive - Simulation umfeldbasierter Fahrzeugfunktionen, PhD Thesis, Technische Universität München, 2014

Ngo, A. et al.: Multi-Layered Measuring the Simulation-to-Reality Gap for Radar (2021)
Ngo, Anthony; Bauer, Max Paul; Resch, Michael: A Multi-Layered Approach for Measuring the Simulation-to-Reality Gap of Radar Perception for Autonomous Driving, in: 2021 IEEE 24th International Conference on Intelligent Transportation Systems (ITSC), pp. 4008–4014, 2021

Ngo, A. et al.: A Sensitivity Analysis Approach for Evaluating a Radar Simulation (2020)
Ngo, Anthony; Bauer, Max Paul; Resch, Michael: A Sensitivity Analysis Approach for Evaluating a Radar Simulation for Virtual Testing of Autonomous Driving Functions, in: 2020 5th Asia-Pacific Conference on Intelligent Robot Systems (ACIRS), pp. 122–128, 2020

Ngo, A. et al.: Deep Evaluation Metric (2021)
Ngo, Anthony; Paul Bauer, Max; Resch, Michael: Deep Evaluation Metric: Learning to Evaluate Simulated Radar Point Clouds for Virtual Testing of Autonomous Driving, in: 2021 IEEE Radar Conference (RadarConf), pp. 1–6, 2021

Normenausschuss Automobiltechnik (NAAutomobil) im DIN: DIN ISO 8855:2013-11 (2013)
Normenausschuss Automobiltechnik (NAAutomobil) im DIN: DIN ISO 8855:2013-11: Straßenfahrzeuge – Fahrzeugdynamik und Fahrverhalten – Begriffe, 2013, visited on 05/01/2022

Nuzzo, R.: Scientific method (2014)
Nuzzo, Regina: Scientific method: Statistical errors, in: Nature, Vol. 506, pp. 150–152, 2014

O'Brien, M. E. et al.: Simulation of 3D Laser Radar Systems (2005)
O'Brien, Michael E.; Fouche, Daniel G.: Simulation of 3D Laser Radar Systems, in: Lincoln Laboratory Journal, Vol. 15, pp. 37–60, 2005

Oberkampf, W. L. et al.: Validation Under Aleatory and Epistemic Uncertainty (2007)
Oberkampf, William L.; Ferson, Scott: Model Validation Under Both Aleatory and Epistemic Uncertainty, in: NATO/RTO Symposium on Computational Uncertainty in Military Vehicle Design, vol. Paper No. AVT-147/RSY-022. P. 26, 2007

Oberkampf, W. L.; Roy, C. J.: Verification and Validation in Scientific Computing (2010)
Oberkampf, William L.; Roy, Christopher J.: Verification and Validation in Scientific Computing, Cambridge University Press, 2010

Oberkampf, W. L. et al.: Verification and validation benchmarks (2008)
Oberkampf, William L.; Trucano, Timothy G.: Verification and validation benchmarks, in: Nuclear Engineering and Design, Vol. 238, pp. 716–743, 2008

Oberkampf, W. L. et al.: Measures of agreement between computation and experiment (2006)
Oberkampf, William Louis; Barone, Matthew F.: Measures of agreement between computation and experiment: Validation metrics, in: Journal of Computational Physics, Vol. 217, pp. 5–36, 2006

Oberkampf, W. L. et al.: Predictive Capability Maturity Model for Modeling and Simulation. (2007)
Oberkampf, William Louis; Pilch, Martin M.; Trucano, Timothy Guy: Predictive Capability Maturity Model for computational modeling and simulation. URL: `http://www.osti.gov/servlets/purl/976951-meC28s/`, 2007, visited on 07/26/2021

Peinecke, N. et al.: Lidar simulation using graphics hardware acceleration (2008)
Peinecke, Niklas; Lueken, Thomas; Korn, Bernd R.: Lidar simulation using graphics hardware acceleration, in: 2008 IEEE/AIAA 27th Digital Avionics Systems Conference, pp. 4.D.4–1–4.D.4–8, 2008

Pliefke, S. et al.: Validation of a Ray-tracing-based Radar Sensor Model (2021)
Pliefke, Sebastian; Germer, Max; Höfer, Andreas; Groner, Achim: Validation of a Ray-tracing-based Radar Sensor Model, in: ATZelectronics worldwide, Vol. 16, pp. 40–43, 2021

Popper, K.: The Logic of Scientific Discovery (2002)
Popper, Karl: The Logic of Scientific Discovery, 2. Edition. Edition, Routledge, 2002

Prinz, A. et al.: Automotive Radar Signal and Interference Simulation for Testing Autonomous Driving (2021)
Prinz, Alexander; Peters, Leo-Tassilo; Schwendner, Johannes; Ayeb, Mohamed; Brabetz, Ludwig: Automotive Radar Signal and Interference Simulation for Testing Autonomous Driving, in: Intelligent Transport Systems, From Research and Development to the Market Uptake, vol. 364, pp. 223–240, 2021

Prinz, A. et al.: Validation Strategy for Radar-Based Assistance Systems (2020)
Prinz, Alexander; Roth, Jonathan; Schwendner, Johannes; Ayeb, Mohamed; Brabetz, Ludwig: Validation Strategy for Radar-Based Assistance Systems under the Influence of Interference, in: 2020 German Microwave Conference (GeMiC), pp. 252–255, 2020

Raju, N. et al.: Evolution of Traffic Microsimulation and Its Use (2021)
Raju, Narayana; Farah, Haneen: Evolution of Traffic Microsimulation and Its Use for Modeling Connected and Automated Vehicles, in: Journal of Advanced Transportation, Vol. 2021, p. 29, 2021

Rasshofer, R. H. et al.: Influences of weather phenomena on automotive laser radar systems (2011)
Rasshofer, R. H.; Spies, M.; Spies, H.: Influences of weather phenomena on automotive laser radar systems, in: Advances in Radio Science, Vol. 9, pp. 49–60, 2011

Rebba, R. et al.: Computational methods for model reliability assessment (2008)
Rebba, Ramesh; Mahadevan, Sankaran: Computational methods for model reliability assessment, in: Reliability Engineering & System Safety, Vol. 93, pp. 1197–1207, 2008

Rich, B. R.: Clarence Leonard (Kelly) Johnson (1995)
Rich, Ben R.: Clarence Leonard (Kelly) Johnson, in: National Academy of Sciences (Hrsg.): Biographical Memoirs, The National Academies Press, 1995

Riedmaier, S. et al.: Unified Framework and Survey for Model VV&UQ (2020)
Riedmaier, Stefan; Danquah, Benedikt; Schick, Bernhard; Diermeyer, Frank: Unified Framework and Survey for Model Verification, Validation and Uncertainty Quantification, in: Archives of Computational Methods in Engineering, 2020

Riedmaier, S. et al.: Survey on Scenario-Based Safety Assessment (2020)
Riedmaier, Stefan; Ponn, Thomas; Ludwig, Dieter; Schick, Bernhard; Diermeyer, Frank: Survey on Scenario-Based Safety Assessment of Automated Vehicles, in: IEEE Access, Vol. 8, pp. 87456–87477, 2020

Riedmaier, S. et al.: Non-deterministic model validation methodology (2021)
Riedmaier, Stefan; Schneider, Jakob; Danquah, Benedikt; Schick, Bernhard; Diermeyer, Frank: Non-deterministic model validation methodology for simulation-based safety assessment of automated vehicles, in: Simulation Modelling Practice and Theory, Vol. 109, p. 102274, 2021

Roache, P.: Building PDE codes to be verifiable and validatable (2004)
Roache, P.J.: Building PDE codes to be verifiable and validatable, in: Computing in Science Engineering, Vol. 6, pp. 30–38, 2004

Rosenberger, P. et al.: Modular OSMP Framework (2022)
Rosenberger, Philipp; Cianciaruso, Nicodemo; Linnhoff, Clemens; Elster, Lukas: Modular OSMP Framework, URL: `https://gitlab.com/tuda-fzd/perception-sensor-modeling/modular-osmp-framework`, 2022, visited on 03/02/2022

Rosenberger, P. et al.: Sequential lidar sensor system simulation (2020)
Rosenberger, Philipp; Holder, Martin Friedrich; Cianciaruso, Nicodemo; Aust, Philip; Tamm-Morschel, Jonas Franz; Linnhoff, Clemens; Winner, Hermann: Sequential lidar sensor system simulation: a modular approach for simulation-based safety validation of automated driving, in: Automotive and Engine Technology, Vol. 5, pp. 187–197, 2020

Rosenberger, P. et al.: Benchmarking and Functional Decomposition of Lidar Sensor Models (2019)
Rosenberger, Philipp; Holder, Martin Friedrich; Huch, Sebastian; Winner, Hermann; Fleck, Tobias; Zofka, Marc René; Zöllner, J Marius; D'hondt, Thomas; Wassermann, Benjamin: Benchmarking and Functional Decomposition of Automotive Lidar Sensor Models, in: 2019 IEEE Intelligent Vehicles Symposium (IV), 2019

Rosenberger, P. et al.: Analysis of Real World Sensor Behavior (2018)
Rosenberger, Philipp; Holder, Martin Friedrich; Zirulnik, Marina; Winner, Hermann: Analysis of Real World Sensor Behavior for Rising Fidelity of Physically Based Lidar Sensor Models, in: 2018 IEEE Intelligent Vehicles Symposium (IV), pp. 611–616, 2018

Rosenberger, P. et al.: Functional Decomposition of Lidar Sensor Systems (2020)
Rosenberger, Philipp; Holder, Martin Friedrich; Zofka, Marc René; Fleck, Tobias; D’hondt, Thomas; Wassermann, Benjamin; Prstek, Juraj: Functional Decomposition of Lidar Sensor Systems for Model Development, in: Leitner, Andrea; Watzenig, Daniel; Ibanez-Guzman, Javier (Hrsg.): Validation and Verification of Automated Systems, Springer International Publishing, 2020

Rosenberger, P. et al.: Reflection Based Lidar Object Model (2022)
Rosenberger, Philipp; Linnhoff, Clemens; Elster, Lukas: Reflection Based Lidar Object Model, URL: `https://gitlab.com/tuda-fzd/perception-sensor-modeling/reflection-based-lidar-object-model`, 2022, visited on 03/02/2022

Rosenberger, P. et al.: Validation of Test Infrastructure (2022)
Rosenberger, Philipp; Schunk, Gerhard; Ikemeyer, Frederik; Duong Quang, Tuan: Validation of Test Infrastructure - from cause trees to a validated system simulation, URL: `https://www.vvm-projekt.de/fileadmin/user_upload/Mid-Term/Presentations/VVM_HZE_S2_P6_20220315_ValidationTestInfrastructure.pdf`, 2022, visited on 05/01/2022

Rosenberger, P. et al.: Towards a Generally Accepted Validation Methodology for Sensor Models (2019)
Rosenberger, Philipp; Wendler, Jan Timo; Holder, Martin Friedrich; Linnhoff, Clemens; Berghöfer, Moritz; Winner, Hermann; Maurer, Markus: Towards a Generally Accepted Validation Methodology for Sensor Models - Challenges, Metrics, and First Results, in: 12th Grazer Symposium Virtuelles Fahrzeug (GSVF), 2019

Rossmann, J. et al.: A Real-Time Optical Sensor Simulation Framework for Development and Testing (2012)
Rossmann, Jürgen; Hempe, Nico; Emde, Markus; Steil, Thomas: A Real-Time Optical Sensor Simulation Framework for Development and Testing of Industrial and Mobile Robot Applications, in: ROBOTIK 2012; 7th German Conference on Robotics, pp. 1–6, 2012

Roth, E. et al.: Analysis and Validation of Perception Sensor Models (2011)
Roth, Erwin; Dirndorfer, Tobias; Knoll, A.; Neumann-Cosel, Kilian von; Ganslmeier, T.; Kern, Andreas; Fischer, Marc: Analysis and Validation of Perception Sensor Models in an Integrated Vehicle and Environment Simulation, in: 22nd Enhanced Safety of Vehicle Conference (ESV), 2011

Rott, R.: Dynamic Update of Stand-Alone Lidar Model (2022)
Rott, Relindis: Dynamic Update of Stand-Alone Lidar Model Based on Ray Tracing Using the Nvidia Optix Engine, in: 2022 IEEE International Conference on Connected Vehicles and Expo (ICCVE), pp. 1–6, 2022

Roy, C. J. et al.: A holistic approach to uncertainty quantification (2012)
Roy, C. J.; Balch, M.: A HOLISTIC APPROACH TO UNCERTAINTY QUANTIFICATION WITH APPLICATION TO SUPERSONIC NOZZLE THRUST, in: International Journal for Uncertainty Quantification, Vol. 2, pp. 363–381, 2012

Roy, C. J. et al.: Framework for verification, validation, and uncertainty (2011)
Roy, Christopher J.; Oberkampf, William L.: A comprehensive framework for verification, validation, and uncertainty quantification in scientific computing, in: Computer Methods in Applied Mechanics and Engineering, Vol. 200, pp. 2131–2144, 2011

Saam, N. J.: Validation Benchmarks and Related Metrics (2019)
Saam, Nicole J.: Validation Benchmarks and Related Metrics, in: Beisbart, Claus; Saam, Nicole J. (Hrsg.): Computer Simulation Validation: Fundamental Concepts, Methodological Frameworks, and Philosophical Perspectives, Springer International Publishing, 2019

Salles, D. et al.: A Modular Co-Simulation Framework (2022)
Salles, Dominik; Lang, Lukas; Kehrer, Martin; Reuss, Hans-Christian: A Modular Co-Simulation Framework with Open Source Software and Automotive Standards, in: Bargende, Michael; Reuss, Hans-Christian; Wagner, Andreas (Hrsg.): 22. Internationales Stuttgarter Symposium, Springer Fachmedien Wiesbaden, 2022

Sankararaman, S. et al.: Integration of model V&V, and calibration for UQ (2015)
Sankararaman, Shankar; Mahadevan, Sankaran: Integration of model verification, validation, and calibration for uncertainty quantification in engineering systems, in: Reliability Engineering & System Safety, Vol. 138, pp. 194–209, 2015

Sargent, R. G.: Assessment Procedure and Set of Criteria in Evaluation of Computerized Models (1981)
Sargent, Robert G.: An Assessment Procedure and a Set of Criteria for Use in the Evaluation of Computerized Models and Computer-Based Modelling Tools, URL: `https://apps.dtic.mil/sti/citations/ADA098785`, 1981, visited on 01/29/2022

Sargent, R. G.: Verification and validation of simulation models (2007)
Sargent, Robert G.: Verification and validation of simulation models, in: 2007 Winter Simulation Conference, pp. 124–137, 2007

Sargent, R. G.: Verification and validation of simulation models (2010)
Sargent, Robert G.: Verification and validation of simulation models, in: 2010 Winter Simulation Conference, pp. 166–183, 2010

Sarin, H. et al.: Comparing Time Histories for Validation of Simulation Models (2010)
Sarin, H.; Kokkolaras, M.; Hulbert, G.; Papalambros, P.; Barbat, S.; Yang, R.-J.: Comparing Time Histories for Validation of Simulation Models: Error Measures and Metrics, in: Journal of Dynamic Systems, Measurement, and Control, Vol. 132, p. 061401, 2010

Sauro, J.: Can You Take the Mean of Ordinal Data? (2016)
Sauro, Jeff: Can You Take the Mean of Ordinal Data?, in: MeasuringU, URL: `https://measuringu.com/mean-ordinal/`, 2016, visited on 05/08/2022

Schaermann, A.: Systematische Bedatung und Bewertung umfelderf. Sensormodelle (2020)
Schaermann, Alexander: Systematische Bedatung und Bewertung umfelderfassender Sensormodelle, PhD Thesis, Technische Universität München, 2020

Schaermann, A. et al.: Validation of vehicle environment sensor models (2017)
Schaermann, Alexander; Rauch, Andreas; Hirsenkorn, Nils; Hanke, Timo; Rasshofer, Ralph; Biebl, Erwin: Validation of vehicle environment sensor models, in: 2017 IEEE Intelligent Vehicles Symposium (IV), pp. 405–411, 2017

Schlager, B. et al.: State-of-the-Art Sensor Models for Virtual Testing (2020)
Schlager, Birgit; Muckenhuber, Stefan; Schmidt, Simon; Holzer, Hannes; Rott, Relindis; Maier, Franz Michael; Saad, Kmeid; Kirchengast, Martin; Stettinger, Georg; Watzenig, Daniel; Ruebsam, Jonas: State-of-the-Art Sensor Models for Virtual Testing of Advanced Driver Assistance Systems/Autonomous Driving Functions, in: SAE International Journal of Connected and Automated Vehicles, Vol. 3, pp. 233–261, 2020

Schlesinger, S. et al.: Terminology for model credibility (1979)
Schlesinger, Stewart; Crosbie, Roy E.; Gagné, Roland E.; Innis, George S.; Lalwani, C.S.; Loch, Joseph; Sylvester, Richard J.; Wright, Richard D.; Kheir, Naim; Bartos, Dale: Terminology for model credibility, in: SIMULATION, Vol. 32, pp. 103–104, 1979

Schmidt, S. et al.: Configurable Sensor Model Architecture (2021)
Schmidt, Simon; Schlager, Birgit; Muckenhuber, Stefan; Stark, Rainer: Configurable Sensor Model Architecture for the Development of Automated Driving Systems, in: MDPI Sensors, Vol. 21, p. 4687, 2021

Schmitt, J. et al.: Phenomenological, Measurement Based LiDAR Sensor Model (2021)
Schmitt, Jakob; Robel, Christopher; Bäker, Bernard: Phenomenological, Measurement Based LiDAR Sensor Model, in: 21. Internationales Stuttgarter Symposium, pp. 424–435, 2021

Schneider, R.: Modellierung der Wellenausbreitung für ein bildgebendes Kfz-Radar (1998)
Schneider, Robert: Modellierung der Wellenausbreitung für ein bildgebendes Kfz-Radar, PhD Thesis, Universität Fridericana Karlsruhe, 1998

Schuldt, F. et al.: Effiziente systematische Testgenerierung für Fahrerassistenzsysteme (2013)
Schuldt, Fabian; Saust, Falko; Lichte, Bernd; Maurer, Markus; Scholz, Stephan: Effiziente systematische Testgenerierung für Fahrerassistenzsysteme in virtuellen Umgebungen, in: Beiträge zum gleichnamigen 14. Braunschweiger Symposium am Forschungsflughafen, Braunschweig, 2013

Schuler, K.: Intelligente Antennensysteme für Kraftfahrzeug-Nahbereichs-Radar-Sensorik (2007)

Schuler, Karin: Intelligente Antennensysteme für Kraftfahrzeug-Nahbereichs-Radar-Sensorik, PhD Thesis, Universität Karlsruhe (TH), 2007

Schuler, K. et al.: Extraction of Virtual Scattering Centers of Vehicles by Ray-Tracing Simulations (2008)

Schuler, Karin; Becker, Denis; Wiesbeck, Werner: Extraction of Virtual Scattering Centers of Vehicles by Ray-Tracing Simulations, in: IEEE Transactions on Antennas and Propagation, Vol. 56, pp. 3543–3551, 2008

Sebastian, G. et al.: RangeWeatherNet for LiDAR-Only Weather and Road Condition Classification (2021)

Sebastian, George; Vattem, Teja; Lukic, Luka; Bürgy, Christian; Schumann, Thomas: RangeWeatherNet for LiDAR-Only Weather and Road Condition Classification, in: 2021 IEEE Intelligent Vehicles Symposium (IV), p. 8, 2021

Sesta, V. et al.: A novel sub-10 ps resolution TDC for CMOS SPAD array (2018)

Sesta, Vincenzo; Villa, Federica; Conca, Enrico; Tosi, Alberto: A novel sub-10 ps resolution TDC for CMOS SPAD array, in: 2018 25th IEEE International Conference on Electronics, Circuits and Systems (ICECS), pp. 5–8, 2018

Slavik, Z. et al.: Phenomenological Modeling of Millimeter-Wave Automotive Radar (2019)

Slavik, Zora; Mishra, Kumar Vijay: Phenomenological Modeling of Millimeter-Wave Automotive Radar, in: 2019 URSI Asia-Pacific Radio Science Conference (AP-RASC), pp. 1–4, 2019

Society of Automotive Engineers: SAE-J3016 (2021)

Society of Automotive Engineers: SAE-J3016: Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles, URL: `https://www.sae.org/standards/content/j3016_202104/`, 2021, visited on 01/06/2022

Stevens, S. S.: On the Theory of Scales of Measurement (1946)

Stevens, S. S.: On the Theory of Scales of Measurement, in: Science, Vol. 103, pp. 677–680, 1946

Stolz, M. et al.: Fast generic sensor models for testing highly automated vehicles in simulation (2018)

Stolz, Michael; Nestlinger, Georg: Fast generic sensor models for testing highly automated vehicles in simulation, in: e & i Elektrotechnik und Informationstechnik, Vol. 135, pp. 365–369, 2018

Stripling, H. F. et al.: The Method of Manufactured Universes (2011)

Stripling, H. F.; Adams, M. L.; McClarren, R. G.; Mallick, B. K.: The Method of Manufactured Universes for validating uncertainty quantification methods, in: Reliability Engineering & System Safety, Vol. 96, pp. 1242–1256, 2011

Su, H. et al.: A Simulation Method for LIDAR of Autonomous Cars (2019)
Su, Hu; Wang, Rui; Chen, Kaixin; Chen, Yizhan: A Simulation Method for LIDAR of Autonomous Cars, in: IOP Conference Series: Earth and Environmental Science, Vol. 234, p. 012055, 2019

Sullivan, G. M. et al.: Analyzing and Interpreting Data From Likert-Type Scales (2013)
Sullivan, Gail M.; Artino, Anthony R.: Analyzing and Interpreting Data From Likert-Type Scales, in: Journal of Graduate Medical Education, Vol. 5, pp. 541–542, 2013

Tamm-Morschel, J. F.: Erweiterung eines Lidar-Sensormodells (2019)
Tamm-Morschel, Jonas Franz: Erweiterung eines phänomenologischen Lidar-Sensormodells durch identifizierte physikalische Effekte, Master Thesis, Technische Universität Darmstadt, 2019

TechInsights Inc.: Velodyne LiDAR Puck Teardown (2019)
TechInsights Inc.: Velodyne LiDAR Puck Teardown, URL: `https://www.techinsights.com/featured-reports/velodyne-lidar-puck`, 2019, visited on 03/23/2022

The MathWorks, Inc.: Anderson-Darling test (2022)
The MathWorks, Inc.: Anderson-Darling test - MATLAB adtest, URL: `https://de.mathworks.com/help/stats/adtest.html`, 2022, visited on 05/07/2022

The MathWorks, Inc.: Fit curve or surface to data (2022)
The MathWorks, Inc.: Fit curve or surface to data - MATLAB fit, URL: `https://de.mathworks.com/help/curvefit/fit.html`, 2022, visited on 06/15/2022

The MathWorks, Inc.: Two-sample Kolmogorov-Smirnov test (2022)
The MathWorks, Inc.: Two-sample Kolmogorov-Smirnov test - MATLAB kstest2, URL: `https://de.mathworks.com/help/stats/kstest2.html`, 2022, visited on 05/07/2022

Thieling, J. et al.: Scalable and Physical Radar Sensor Simulation for Interacting Digital Twins (2021)
Thieling, Jörn; Frese, Susanne; Roßmann, Jürgen: Scalable and Physical Radar Sensor Simulation for Interacting Digital Twins, in: IEEE Sensors Journal, Vol. 21, 2021

Thieling, J. et al.: Highly-Scalable and Generalized Sensor Structures (2018)
Thieling, Jörn; Rosmann, Jürgen: Highly-Scalable and Generalized Sensor Structures for Efficient Physically-Based Simulation of Multi-Modal Sensor Networks, in: 2018 IEEE International Conference on Sensing Technology (ICST 2018), pp. 202–207, 2018

Tontini, A. et al.: Numerical Model of SPAD-Based Direct Time-of-Flight Flash LIDAR (2020)
Tontini, Alessandro; Gasparini, Leonardo; Perenzoni, Matteo: Numerical Model of SPAD-Based Direct Time-of-Flight Flash LIDAR CMOS Image Sensors, in: MDPI Sensors, Vol. 20, p. 5203, 2020

Trucano, T. G. et al.: Description of the Sandia Validation Metrics Project (2001)
Trucano, Timothy Guy; Easterling, Robert G.; Dowding, Kevin J.; Paez, Thomas L.; Urbina, Angel; Romero, Vicente J.; Rutherford, Brian M.; Hills, Richard G.: Description of the Sandia Validation Metrics Project, 2001

Trucano, T. G. et al.: General Concepts for Experimental Validation of ASCI Code Applications (2002)
Trucano, Timothy Guy; Pilch, Martin; Oberkampf, William Louis: General Concepts for Experimental Validation of ASCI Code Applications, URL: `https://www.osti.gov/biblio/800777`, 2002, visited on 04/03/2022

U.S. Department of Defense: MIL-STD-3022 (2008)
U.S. Department of Defense: MIL-STD-3022 DOCUMENTATION VERIFICATION VALIDATION, URL: `http://everyspec.com/MIL-STD/MIL-STD-3000-9999/MIL-STD-3022_4197/`, 2008, visited on 03/14/2022

U.S. National Aeronautics and Space Administration: NASA-HDBK-7009A (2019)
U.S. National Aeronautics and Space Administration: NASA HANDBOOK FOR MODELS AND SIMULATIONS: AN IMPLEMENTATION GUIDE FOR NASA-STD-7009A, 2019, visited on 03/14/2022

U.S. National Aeronautics and Space Administration: NASA-STD-7009A (2016)
U.S. National Aeronautics and Space Administration: NASA TECHNICAL STANDARD NASA-STD-7009A - STANDARD FOR MODELS AND SIMULATIONS, 2016, visited on 03/14/2022

Ulbrich, S. et al.: Defining and Substantiating the Terms Scene, Situation, and Scenario (2015)
Ulbrich, Simon; Menzel, Till; Reschka, Andreas; Schuldt, Fabian; Maurer, Markus: Defining and Substantiating the Terms Scene, Situation, and Scenario for Automated Driving, in: 2015 IEEE 18th International Conference on Intelligent Transportation Systems (ITSC), vol. 18, pp. 982–988, 2015

United Nations Economic Commission for Europe: UNECE 157 (2021)
United Nations Economic Commission for Europe: UNECE 157: Uniform provisions concerning the approval of vehicles with regard to Automated Lane Keeping Systems, URL: `https://unece.org/sites/default/files/2021-03/R157e.pdf`, 2021

Velodyne LiDAR, Inc.: VLP-16 User Manual (2019)
Velodyne LiDAR, Inc.: VLP-16 User Manual, URL: `https://velodynelidar.com/wp-content/uploads/2019/12/63-9243-Rev-E-VLP-16-User-Manual.pdf`, 2019, visited on 06/15/2022

Viehof, M.: Objektive Qualitätsbewertung von Fahrdynamiksimulationen (2018)
Viehof, Michael: Objektive Qualitätsbewertung von Fahrdynamiksimulationen durch statistische Validierung, PhD Thesis, Technische Universität Darmstadt, 2018

Viehof, M. et al.: Forschungsstand der Validierung (2017)
Viehof, Michael; Winner, Hermann: Stand der Technik und der Wissenschaft: Modellvalidierung im Anwendungsbereich der Fahrdynamiksimulation, URL: `http://tuprints.ulb.tu-darmstadt.de/6662/`, 2017

Villa, F. et al.: SPADs and SiPMs Arrays for Long-Range High-Speed LiDAR (2021)
Villa, Federica; Severini, Fabio; Madonini, Francesca; Zappa, Franco: SPADs and SiPMs Arrays for Long-Range High-Speed Light Detection and Ranging (LiDAR), in: MDPI Sensors, Vol. 21, p. 3839, 2021

Voyles, I. T. et al.: Model Validation in the Presence of Aleatory and Epistemic Uncertainties (2015)
Voyles, Ian T.; Roy, Christopher J.: Evaluation of Model Validation Techniques in the Presence of Aleatory and Epistemic Input Uncertainties, in: 17th AIAA Non-Deterministic Approaches Conference, 2015

Voyles, I. T. et al.: Model Validation in the Presence of Uncertainty (2014)
Voyles, Ian T.; Roy, Christopher J.: Evaluation of Model Validation Techniques in the Presence of Uncertainty, in: 16th AIAA Non-Deterministic Approaches Conference, 2014

Wachenfeld, W. et al.: Die Freigabe des autonomen Fahrens (2015)
Wachenfeld, Walther; Winner, Hermann: Die Freigabe des autonomen Fahrens, in: Maurer, Markus; Gerdes, J. Christian; Lenz, Barbara; Winner, Hermann (Hrsg.): Autonomes Fahren: Technische, rechtliche und gesellschaftliche Aspekte, Springer, 2015

Wang, S.: State Lattice-based Motion Planning for Autonomous On-Road Driving (2015)
Wang, Shuiying: State Lattice-based Motion Planning for Autonomous On-Road Driving, PhD Thesis, Freie Universität Berlin, 2015

Wang, S. et al.: Shader-based sensor simulation for autonomous car testing (2012)
Wang, Shuiying; Heinrich, Steffen; Wang, Miao; Rojas, Raúl: Shader-based sensor simulation for autonomous car testing, in: 2012 IEEE 15th International Conference on Intelligent Transportation Systems (ITSC), pp. 224–229, 2012

Wang, X. et al.: Gating techniques for maneuvering target tracking in clutter (2002)
Wang, Xuezhi; Challa, S.; Evans, R.: Gating techniques for maneuvering target tracking in clutter, in: IEEE Transactions on Aerospace and Electronic Systems, Vol. 38, pp. 1087–1097, 2002

Wheeler, T. A. et al.: Deep stochastic radar models (2017)
Wheeler, Tim Allen; Holder, Martin Friedrich; Winner, Hermann; Kochenderfer, Mykel J.: Deep stochastic radar models, in: 2017 IEEE Intelligent Vehicles Symposium (IV), pp. 47–53, 2017

Williamson, R. C. et al.: Probabilistic arithmetic (1990)
Williamson, Robert C.; Downs, Tom: Probabilistic arithmetic. I. Numerical methods for calculating convolutions and dependency bounds, in: International Journal of Approximate Reasoning, Vol. 4, pp. 89–158, 1990

Winner, H.: Automotive RADAR (2016)

Winner, Hermann: Automotive RADAR, in: Winner, Hermann; Hakuli, Stephan; Lotz, Felix; Singer, Christina (Hrsg.): Handbook of Driver Assistance Systems: Basic Information, Components and Systems for Active Safety and Comfort, Springer International Publishing, 2016

Woods, J. O.: GLIDAR (2021)

Woods, John O.: GLIDAR, URL: `https://github.com/WVU-ASEL/glidar`, 2021, visited on 01/09/2022

Xi, Z. et al.: Validation Metric for Dynamic System Responses under Uncertainty (2015)

Xi, Zhimin; Pan, Hao; Fu, Yan; Yang, Ren-Jye: Validation Metric for Dynamic System Responses under Uncertainty, in: SAE International Journal of Materials and Manufacturing, Vol. 8, pp. 309–314, 2015

Zhao, J. et al.: Method and Applications of Lidar Modeling for Virtual Testing of Intelligent Vehicles (2020)

Zhao, Jian; Li, Yaxin; Zhu, Bing; Deng, Weiwen; Sun, Bohua: Method and Applications of Lidar Modeling for Virtual Testing of Intelligent Vehicles, in: IEEE Transactions on Intelligent Transportation Systems, 2020

Own Publications

Rosenberger, Philipp; Schunk, Gerhard; Ikemeyer, Frederik; Duong Quang, Tuan: *Validation of Test Infrastructure - from cause trees to a validated system simulation*, URL: `https://www.vvm-projekt.de/fileadmin/user_upload/Mid-Term/Presentations/VVM_HZE_S2_P6_20220315_ValidationTestInfrastructure.pdf`, 2022, visited on 05/01/2022

Linnhoff, Clemens; **Rosenberger, Philipp**; Winner, Hermann: *Refining Object-Based Lidar Sensor Modeling - Challenging Ray Tracing as the Magic Bullet*, in: IEEE Sensors Journal, Vol. 21,S. 24238–24245, 2021

Elster, Lukas; Linnhoff, Clemens; **Rosenberger, Philipp**; Schmidt, Simon; Stark, Rainer; Winner, Hermann: *Fundamental Design Criteria for Logical Scenarios in Simulation-based Safety Validation of Automated Driving Using Sensor Model Knowledge*, in: 2021

Linnhoff, Clemens; **Rosenberger, Philipp**; Schmidt, Simon; Elster, Lukas; Stark, Rainer; Winner, Hermann: *Towards Serious Perception Sensor Simulation for Safety Validation of Automated Driving - A Collaborative Method to Specify Sensor Models*, in: 2021

Linnhoff, Clemens; **Rosenberger, Philipp**; Holder, Martin Friedrich; Cianciaruso, Nicodemo; Winner, Hermann: *Highly Parameterizable and Generic Perception Sensor Model Architecture*, in: Automatisiertes Fahren 2020, S. 195–206, 2021

Rosenberger, Philipp; Holder, Martin Friedrich; Cianciaruso, Nicodemo; Aust, Philip; Tamm-Morschel, Jonas Franz; Linnhoff, Clemens; Winner, Hermann: *Sequential lidar sensor system simulation: a modular approach for simulation-based safety validation of automated driving*, in: Automotive and Engine Technology, Vol. 5, S. 187–197, 2020

Holder, Martin Friedrich; Thielmann, Jan; **Rosenberger, Philipp**; Linnhoff, Clemens; Winner, Hermann: *How to evaluate synthetic radar data? Lessons learned from finding driveable space in virtual environments*, in: 13. Uni-DAS e.V. Workshop Fahrerassistenz und automatisiertes Fahren 2020, 2020

Holder, Martin Friedrich; Linnhoff, Clemens; **Rosenberger, Philipp**; Popp, Christoph; Winner, Hermann: *Modeling and Simulation of Radar Sensor Artifacts for Virtual Testing of Autonomous Driving*, in: 9. Tagung Automatisiertes Fahren, 2019

Rosenberger, Philipp; Holder, Martin Friedrich; Zofka, Marc René; Fleck, Tobias; D'hondt, Thomas; Wassermann, Benjamin; Prstek, Juraj: *Functional Decomposition of Lidar Sensor Systems for Model Development*, in: Leitner, Andrea; Watzenig, Daniel; Ibanez-Guzman, Javier (Hrsg.): Validation and Verification of Automated Systems, Springer International Publishing, 2020

Holder, Martin Friedrich; Linnhoff, Clemens; **Rosenberger, Philipp**; Winner, Hermann: *The Fourier Tracing Approach for Modeling Automotive Radar Sensors*, in: 20th International Radar Symposium (IRS), S. 8, 2019

Rosenberger, Philipp; Holder, Martin Friedrich; Huch, Sebastian; Winner, Hermann; Fleck, Tobias; Zofka, Marc René; Zöllner, J Marius; D'hondt, Thomas; Wassermann, Benjamin: *Benchmarking and Functional Decomposition of Automotive Lidar Sensor Models*, in: IEEE Intelligent Vehicles Symposium, 2019

Rosenberger, Philipp; Wendler, Jan Timo; Holder, Martin Friedrich; Linnhoff, Clemens; Berghöfer, Moritz; Winner, Hermann; Maurer, Markus: *Towards a Generally Accepted Validation Methodology for Sensor Models - Challenges, Metrics, and First Results*, in: 2019 Grazer Symposium Virtuelles Fahrzeug (GSVF 2019), 2019

Rosenberger, Philipp; Holder, Martin Friedrich; Zirulnik, Marina; Winner, Hermann: *Analysis of Real World Sensor Behavior for Rising Fidelity of Physically Based Lidar Sensor Models*, in: 2018 IEEE Intelligent Vehicles Symposium (IV), S. 611–616, 2018

Holder, Martin Friedrich; **Rosenberger, Philipp**; Bert, Felix; Winner, Hermann: *Data-driven Derivation of Requirements for a Lidar Sensor Model*, in: Graz Symposium Virtual Vehicle, 2018

Holder, Martin Friedrich; **Rosenberger, Philipp**; Winner, Hermann; D'hondt, Thomas; Makkapati, Vamsi Prakash; Maier, Michael; Schreiber, Helmut; Magosi, Zoltan; Slavik, Zora; Bringmann, Oliver; Rosenstiel, Wolfgang: *Measurements revealing Challenges in Radar Sensor Modeling for Virtual Validation of Autonomous Driving*, in: 2018 21st International Conference on Intelligent Transportation Systems (ITSC), S. 2616–2622, 2018

Own Open Source Content

Rosenberger, Philipp; Cianciaruso, Nicodemo; Linnhoff, Clemens; Elster, Lukas: *Modular OSMP Framework*, `https://gitlab.com/tuda-fzd/perception-sensor-modeling/modular-osmp-framework`, 2022, Accessed 02.03.2022

Rosenberger, Philipp; Linnhoff, Clemens; Elster, Lukas: *Reflection Based Lidar Object Model*, `https://gitlab.com/tuda-fzd/perception-sensor-modeling/reflection-based-lidar-object-model`, 2022, Accessed 02.03.2022

Linnhoff, Clemens; **Rosenberger, Philipp**; Elster, Lukas: *Object Based Generic Perception Object Model*, `https://gitlab.com/tuda-fzd/perception-sensor-modeling/object-based-generic-perception-object-model`, 2022, Accessed 02.03.2022

Elster, Lukas; Linnhoff, Clemens; **Rosenberger, Philipp**: *Reflection Based Radar Object Model*, `https://gitlab.com/tuda-fzd/perception-sensor-modeling/reflection-based-radar-object-model`, 2022, Accessed 02.03.2022

Rosenberger, Philipp; Hinsemann, Timo; Linnhoff, Clemens; Elster, Lukas: *PerCollECT - LidarLimbs*, `https://github.com/PerCollECT/LidarLimbs`, 2022, Accessed 10.03.2022

Linnhoff, Clemens; Hinsemann, Timo; Elster, Lukas; **Rosenberger, Philipp**: *PerCollECT - RadarRami*, `https://github.com/PerCollECT/RadarRami`, 2022, Accessed 10.03.2022

Elster, Lukas; Hinsemann, Timo; Linnhoff, Clemens; **Rosenberger, Philipp**: *PerCollECT - CameraCopse*, `https://github.com/PerCollECT/CameraCopse`, 2022, Accessed 10.03.2022

Elster, Lukas; Hinsemann, Timo; Linnhoff, Clemens; **Rosenberger, Philipp**: *PerCollECT - UltrasonicUnderwood*, `https://github.com/PerCollECT/UltrasonicUnderwood`, 2022, Accessed 10.03.2022

Supervised Theses

Ackermann, Stefan Martin: *Systematische Untersuchung von Radar Tracking-Algorithmen*, Master Thesis, Technische Universität Darmstadt, 2017

Aust, Philip: *Entwicklung eines lidartypischen Objektlisten-Sensormodells*, Master Thesis, Technische Universität Darmstadt, 2019

Bangalore Vijayendra, Vishwas: *Refinement of a Virtual Environment Representation to Conduct Validation Tests for Automated Driving*, Master Thesis, Technische Universität Darmstadt, 2018

Berghöfer, Moritz: *Generierung realer und synthetischer Sensordaten zur Validierung von Sensormodellen für die simulationsbasierte Absicherung der Valet Parking Funktion*, Bachelor Thesis, Technische Universität Darmstadt, 2019

Bert, Felix: *Identifikation von Merkmalen für die Objekterkennung mit lernenden LiDAR-Algorithmen*, Independent research assignment, Technische Universität Darmstadt, 2017

Chen, Xing: *Enhancement of a Software Tool for Annotation of Sensor Data*, Master Thesis, Technische Universität Darmstadt, 2018

Domhardt, Kai: *Retrospektive Korrektur von Objektexistenzfehlern in der Umfelderfassung*, Bachelor Thesis, Technische Universität Darmstadt, 2016

Fu, Junfeng: *Auswahl und Implementation eines Ultraschall-Sensormodells für die Fahrzeuganwendung*, Master Thesis, Technische Universität Darmstadt, 2018

Gomes-Martins, Samuel: *Entwicklung einer Regelung für eine Pkw-Drehbühne zur Zeitsynchronen Messung des RADAR-Querschnitts von Pkw*, Master Thesis, Technische Universität Darmstadt, 2018

Gu, Penguri: *Sensordatenfusion für eine Lehrplattform für Autonomes Fahren*, Master Thesis, Technische Universität Darmstadt, 2018

Guridi, Arturo: *Selection and Implementation of an angle and turning rate control for a car turntable*, Bachelor Thesis, Technische Universität Darmstadt, 2018

Hadelli, Ali Adel: *Messkampagne zur Parametrisierung und Validierung von Lidar-Sensor-Modellen*, Bachelor Thesis, Technische Universität Darmstadt, 2020

Hellwig, Sven: *Development of a Radar SLAM Algorithm*, Master Thesis, Technische Universität Darmstadt, 2018

Hinsemann, Timo: *Analyse von Effekten in Radar- und Lidardaten für die virtuelle Absicherung automatisierter Fahrfunktionen*, Bachelor Thesis, Technische Universität Darmstadt, 2020

Huch, Sebastian: *Entwicklung einer umfassenden Metrik für die Bewertung einer Lidar-Sensor-Simulation durch Betrachtung mehrerer aufeinander folgender Verarbeitungsebenen*, Master Thesis, Technische Universität Darmstadt, 2018

Hofrichter, Kristof: *Sammlung und Aufbereitung der Einflüsse von Wetterbedingungen auf Radar und Lidar für die Modellbildung*, Bachelor Thesis, Technische Universität Darmstadt, 2020

Hoyer, Tobias: *Vergleich von gemessenen und simulierten RCS Profilen von Fahrzeugen*, Bachelor Thesis, Technische Universität Darmstadt, 2020

Jiao, Yifei: *Implementierung eines generischen Simulationsmodells für die aktive Umfeldsensorik zur realistischen Objektlisten-Ausgabe*, Master Thesis, Technische Universität Darmstadt, 2020

Jin, Geng: *Enhancement of an ideal Lidar sensor model with identified physical effects*, Master Thesis, Technische Universität Darmstadt, 2018

Knerr, Jonathan: *Development of a Tracking Algorithm for a Lidar Sensor Model*, Master Thesis, Technische Universität Darmstadt, 2017

Krüger, Jonas: *Lernende Algorithmen zur Objekterkennung mit Radar*, Master Thesis, Technische Universität Darmstadt, 2018

Linares Arellano, Alberto: *Implementation and Integration of an Experimental Vehicle Sensor Setup for Automated Parking*, Bachelor Thesis, Technische Universität Darmstadt, 2018

Linnhoff, Clemens: *Entwicklung eines Radar-Sensormodells*, Master Thesis, Technische Universität Darmstadt, 2018

Liu, Tienan: *Auswahl und Implementation eines Ultraschall-Sensormodells für die Fahrzeuganwendung*, Master Thesis, Technische Universität Darmstadt, 2017

Mei, Xiaobo: *Entwicklung geeigneter Metriken zum Vergleich von Lidar-Punktewolken*, Master Thesis, Technische Universität Darmstadt, 2018

Müller, Jochen: *Construction and Development of a Test Stand for Identification of Radar Reflection Properties of Vehicles*, Master Thesis, Technische Universität Darmstadt, 2017

Müller, Till Mathis: *Kombination von Leistungs- und Geschwindigkeitsschätzung in einem Deep-Learning Radar-Modell*, Bachelor Thesis, Technische Universität Darmstadt, 2018

Ngo, Anthony: *Enhancement of a Lidar Sensor Model for Simulation-based Development and Testing of Object Detection Algorithms*, Master Thesis, Technische Universität Darmstadt, 2018

Popp, Christoph: *Entwicklung einer Methode zur automatischen Erkennung von Schein- und Spiegelzielen im Automobil-Radar*, Master Thesis, Technische Universität Darmstadt, 2019

Scheiwe, Gunnar: *Charakterisierung von Radarsensoren für automatisiertes Fahren*, Bachelor Thesis, Technische Universität Darmstadt, 2018

Struck, Merten: *Objektklassifizierung mit Radar*, Master Thesis, Technische Universität Darmstadt, 2019

Sun, Bo: *Integration einer Sicherheitszone anhand eines bestehenden Sensormodells für das vollautomatisierte Valet-Parken*, Bachelor Thesis, Technische Universität Darmstadt, 2018

Tamm-Morschel, Jonas Franz: *Erweiterung eines phänomenologischen Lidar-Sensormodells durch identifizierte physikalische Effekte*, Master Thesis, Technische Universität Darmstadt, 2019

Zhang, Yanni: *Implementation of an enhanced Lidar Tracking algorithm for automated driving*, Master Thesis, Technische Universität Darmstadt, 2018

Zirulnik, Marina: *Untersuchung der Charakteristik der Mehrzielfähigkeit sowie des Rauschverhaltens eines Automotive LIDAR Sensors*, Master Thesis, Technische Universität Darmstadt, 2017