WEATHERSAFE

SAFETY ENHANCEMENT UNDER ADVERSE WEATHER CONDITIONS THROUGH IMPROVED ADAS/AD ROBUSTNESS

Deliverable D_Simu.3

Guidelines for sensor model validation

Project Name	WeatherSafe	
Dissemination level	Consortium	
Written by	Dr. Philipp Rosenberger	Persival GmbH
Issue date	26/05/2026	
Reviewer	Pierre Duthon, CEREMA	
Keywords	Adverse Weather, Simulation, Lidar, Radar, Camera	
Version	1.1	

EXECUTIVE SUMMARY

This deliverable provides practical, step-by-step guidelines for validating perception sensor models under adverse weather. It consolidates the WeatherSafe WP Simulation results and describes how to plan and execute sensor model validation in a traceable and uncertainty-aware way, including (i) how to derive and prioritize weather-relevant requirements, (ii) which reference data must be collected to enable replay-to-sim validation, and (iii) which metrics and concepts can be used to quantify agreement between measured and simulated sensor outputs.

The methodology combines a staged validation process (including explicit separation of reference data validation, measurement data validation, scenario or parameter validation, and model validation) with a continuous verification and validation development workflow. Following (Hofrichter, et al. 2025), validation is treated as a lifecycle activity, with traceable requirements, test cases, acceptance criteria, and regression testing after each model update. This supports maintaining credibility as weather effects and sensor model components are iteratively added and refined.

To keep validation effort focused on WeatherSafe-relevant weather influences, the deliverable proposes PerCollECT[1] as a structured knowledge base of cause-effect chains and CEPRA from (Linnhoff, Rosenberger, et al. 2021) as a prioritization method that ranks chains by occurrence in the operational design domain and impact on the system under test. The result is a transparent pathway from weather phenomena and sensor effects to concrete, testable requirements and a scalable validation test portfolio.

A central prerequisite for credible validation is high-quality reference data for replay-to-sim. Based on the workflow demonstrated in deliverable D_Simu.2, this deliverable summarizes the minimum required data categories, including synchronized sensor outputs, target and geometry reference, weather reference sensing (for example precipitation, visibility or extinction coefficient, and illumination), timing and transforms, and structured metadata. Practical data-quality checks are proposed to ensure time synchronization, coordinate consistency, baseline availability, weather-category plausibility, and dataset completeness before performing any comparison between simulation and measurement.

For quantitative assessment, the deliverable adopts the DVM framework and the DVM map extension from (Elster, et al. 2024). These metrics enable uncertainty-aware comparison of measured and simulated sensor outputs and separate systematic bias from scattering errors. The deliverable also clarifies additional constraints that arise in dynamic scenarios, such as time windowing, Doppler handling for radar, sample-size comparability, and tighter requirements on time alignment. In addition, a complementary perspective based on (Linnhoff 2023) is included: expectation-value KPIs (for example atmospheric detections and attenuation trends as a function of quantified weather descriptors) support effect-level calibration and plausibility checks before applying sample-based DVM comparisons.

Overall, the deliverable provides a coherent guideline that supports partners in designing measurement campaigns, structuring validation datasets, selecting and applying metrics, and documenting applicability and limitations of weather-affected sensor simulations. This

[1] https://github.com/PerCollECT

enables more consistent and reproducible validation results across sensor types (lidar, radar, and camera) and provides a basis for acceptance-criteria discussions and future extension toward more complex and dynamic scenarios.

The following table provides an overview of WP Simulation and the planned tasks:

Task	Description
T_Simu.1	State of the art: review of existing European simulation means, studies and works dealing with adverse weather conditions in road safety based on PerCollECT.
T_Simu.2	Adverse weather conditions simulation: based on the work in WP1 and WP2, possibilities for implementing relevant weather effects in sensor simulation are demonstrated.
T_Simu.3	Validation of weather effect modelling: comparison of real and synthetic sensor data with weather influence based on a case study
T_Simu.4	Simulations' assessment guidelines: Constitute and promote a list of recommendations to evaluate different aspects of the simulation of adverse weather, such as physical-numerical correlation and methodologies as addon for WP3.

The work package results are summarized in three deliverables:

Task	Description
D_Simu.1	Report that summarizes the state of the art (T_Simu.1) and the prospect of simulation means able to reproduce adverse weather conditions.
D_Simu.2	Results from the comparison between real and synthetic sensor data with weather influence based on a case study. It reviews discussions, works and outputs of T_Simu.2 and T_Simu.3. Especially the methodologies, parameters and requirements needed to simulate adverse weather conditions. It will include the description of the variables usable to play a scenario and sensor perception levels.
D_Simu.3	Guidelines for sensor model validation including the important points to consider during data collection. It concludes on project guidelines (T_Simu.4) on the assessment of simulation and the parameters to be considered.

REVISION HISTORY

Revision	Date	Description, updates and changes	Status
1.0	31/03/2026	Consolidated version, open for final review	Draft
1.1	26/05/2026	Consolidated version after final review	Approved

ABBREVIATIONS

APSS	Active Perception Sensor Systems
AVM	Area Validation Metric
CAVM	Corrected Area Validation Metric
CDF	Cumulative Distribution Function
CEPRA	Cause, Effect, and Phenomenon Relevance Analysis
CI	Continuous Integration
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DVM	Double Validation Metric
EDF	Empirical Distribution Function
eCDF	Empirical Cumulative Distribution Function
KPI	Key Performance Indicator
LiDAR	Light Detection and Ranging
MOR	Meteorological Optical Range
MVC	Metric Validity Criterion
ODD	Operational Design Domain
OSI	Open Simulation Interface
p-box	Probability Box
PCMM	Predictive Capability Maturity Model
PerCollECT	Perception Sensor Collaborative Effect and Cause Tree
ROI	Region of Interest
SotA	State of the Art
SuT	System under Test
V&V	Verification and Validation
VAM	Validity Assessment Map
VV&UQ	Verification, Validation, and Uncertainty Quantification

TABLE OF CONTENTS

EXECUTIVE SUMMARY 2
REVISION HISTORY 3
ABBREVIATIONS 5
TABLE OF CONTENTS 6
1. INTRODUCTION 7
2. VALIDATION OF SENSOR MODELS IN GENERAL 8
3. METRICS FOR SENSOR MODEL VALIDATION 24
4. GUIDELINES FOR WEATHER EFFECT SIMULATION (ROLE-MODEL BASED ENVIRONMENTAL INFLUENCE MODELING) 27
5. CONCLUSION 31
6. ACKNOWLEDGEMENTS 33
7. REFERENCES 34
8. TABLE OF ILLUSTRATIONS 35

1. Introduction

The simulation of adverse weather effects in perception sensor data is a critical challenge in the development and validation of automated driving systems. Rain, snow, fog, glare, and other environmental influences introduce complex physical phenomena that significantly alter the behavior of sensors such as lidar, radar, and cameras. These effects impact signal propagation, reflection, and interpretation, leading to degraded performance, false positives and false negatives, and reduced detection ranges, among others. To ensure safety and reliability, it is essential to simulate these conditions with high fidelity and validate sensor models against real-world measurements.

Building on the state of the art consolidated in deliverable D_Simu.1 and the practical insights gained from the recording of real sensor data including weather reference sensors and the comparison with synthetic data in deliverable D_Simu.2, this deliverable provides guidelines for sensor model validation with a particular focus on adverse weather influence.

From a methodological perspective, sensor model validation in adverse weather must go beyond visually plausible rendering and instead demonstrate that the simulation reproduces measurable sensor behavior under defined environmental conditions. This includes defining a validation objective (what is considered "sufficiently accurate" for the intended use), carefully selecting reference measurements (ground truth, sensor outputs, and relevant environmental observations), and transparently defining the operational envelope (e.g., target types, distances, angles, and weather severity). Controlled or well-characterized test conditions are particularly valuable, as they allow repeatable exposure of sensors to rain, fog, snow, spray, and illumination effects and enable the derivation of quantitative performance curves. Such performance-oriented validation is aligned with the need for traceability and credibility in safety-related simulation, as discussed in the literature (e.g., (Holder 2021); (Linnhoff 2023))

A central component of validation is the use of appropriate metrics to compare simulated and measured sensor data (Rosenberger 2022). Depending on the sensor type and the intended use of the model, comparisons may be performed at different levels, for example at raw signal/point level, at detection/object level, or at higher-level key performance indicators (KPIs) such as effective range, detection probability, false alarm rate, or uncertainty/noise characteristics. The choice of metrics must reflect both the data structure (e.g., images, point clouds, radar detections) and the application context (e.g., perception algorithm input, sensor fusion, safety assessment). Where possible, metrics should be computed in a way that is robust to annotation uncertainty and sampling differences, and they should be interpretable to non-technical experts to facilitate discussions on acceptance criteria.

Finally, this deliverable addresses validation of weather influence itself, meaning whether the change in sensor behavior due to adverse weather is reproduced correctly by the simulation. This requires explicit parameterization and documentation of the weather condition (e.g., precipitation intensity, fog visibility range, etc.) and a validation approach that separates baseline sensor behavior (reference conditions) from weather-induced degradation. In practice, this can be achieved by stratifying results by weather severity levels and evaluating how selected metrics evolve across these levels. The result is not only an assessment of absolute agreement between simulation and measurement, but also an assessment of whether the simulated degradation envelope is consistent with measured trends.

2. Validation of Sensor Models in General

2.1 SotA in Sensor Model Validation

This section summarizes the state of the art (SotA) in verification, validation, and uncertainty quantification (VV&UQ) for perception sensor simulation models, with a focus on active perception sensor systems (APSS) such as lidar and radar. The intent of this section is to provide a consistent foundation for the subsequent guideline parts in following sections.

2.1.1 Motivation: why APSS simulation needs credibility

For automated driving and advanced driver assistance systems, validation by real-world driving alone is not feasible for broad operational domains. Therefore, simulation is used as a key enabler for safety validation and homologation workflows. This shifts the burden of evidence: simulation models, including sensor models, must be qualified before they are used as decision-support tools in safety arguments. (Rosenberger 2022) emphasizes that credible simulation use requires systematic VV&UQ, not only ad hoc comparison runs or visual plausibility checks.

A useful conceptual anchor is the predictive capability maturity model (PCMM), which frames simulation credibility as a multi-dimensional topic. In this view, model validation and uncertainty quantification are necessary, but not sufficient; verification (code and solution), sensitivity analysis, and documentation of assumptions and limitations also contribute to whether a model is credible enough for its intended use. A related perspective is provided by NASA credibility assessment concepts, which emphasize data pedigree, traceability, uncertainty characterization, and robustness in addition to direct agreement with measured data.

This leads to three practical implications:

1. Validation must be explicitly tied to an intended use (for example, comparison of weather-induced degradation trends, parameter calibration, or safety argumentation).
2. Validation outcomes should be accompanied by uncertainty statements and a clear applicability domain description, rather than producing a single valid or invalid statement.
3. Validation should be organized as a continuous activity across the sensor simulation development lifecycle, using traceable requirements, test cases, acceptance criteria, and regression testing to prevent credibility loss after updates (Hofrichter, et al. 2025)

2.1.2 TERMINOLOGY FOUNDATION RELEVANT FOR SENSOR MODEL VALIDATION

To avoid misunderstandings when comparing approaches across projects and publications, (Rosenberger 2022) consolidates key terms that are central for sensor model VV&UQ. For the context of this deliverable, the following are particularly relevant:

Parameter space and operational domain. Automated driving functions have operating conditions under which they are designed to function (often discussed as an operational design domain, ODD). For quantitative validation and uncertainty prediction, parameters should be at least interval-scaled to support meaningful metric application, interpolation, and sensitivity reasoning. For that reason, (Rosenberger 2022) recommends focusing on a parameter-space representation that enables metric-based analysis.

Scenario and experiment. A scenario describes the temporal development of a sequence of scenes. In validation, an experiment refers to a concrete measurement campaign instance under specified conditions (targets, geometry, environment, and sensor configuration). Experiments bridge abstract parameter spaces and measurable sensor outputs.

Measurement, reference, and ground truth. In real-world campaigns, measurements are never error-free. The term "Ground Truth" is often used in practice, but strictly exists only in simulation. In real measurement campaigns, reference sensors and reference procedures only approximate an ideal value, and their uncertainty should be described rather than ignored. This is especially important under adverse weather because weather reference sensors (for example, precipitation and visibility sensors) are themselves uncertain and can become a dominant driver for validation uncertainty.

Accuracy, bias, and scatter. Accuracy is a combination of trueness and precision, which translate into systematic error (bias) and random error (scatter, often represented by a standard deviation) and together form measurement uncertainty. For sensor modeling, this distinction is critical because a sensor model should replicate the measurement behavior of a device, including bias and scatter characteristics, not only deterministic physics.

Verification, validation, and uncertainty quantification. Verification checks that the implementation represents the developer's conceptual description correctly, while validation assesses how accurately the model represents the real world for the intended uses. Data validity and conceptual model validity are separate concerns that can be intertwined in practice. (Rosenberger 2022) also highlights the idea of sample validity introduced by (Viehof 2018): validity cannot be proven globally, but it can fail to be falsified for a set of empirical samples.

These concepts guide the rest of the deliverable: the methodology in Section 2.2 formalizes how to validate, Sections 2.3 and 2.4 focus on relevant weather effects and reference data, and Section 3 details metrics aligned with the bias and scatter separation.

2.1.3 VV&UQ AS A PROCESS AND THE ROLE OF MATURITY FRAMEWORKS

A recurring theme in the literature is that VV&UQ is best understood as a process rather than a single test. (Rosenberger 2022) describes an integrated view that sequences activities from specification through validation experiments, metric computation, and application-domain prediction and documentation. This process view is important for APSS simulation because sensor behavior depends on many interacting effects (targets, geometry, sensor internals, environment, and processing), reference data is uncertain, and validation samples are limited. Therefore, a central practical question is how to justify model application beyond the validated points.

Maturity frameworks such as PCMM and NASA credibility assessment practices provide checklists and reporting structures that help prevent partial validation where only a subset of credibility aspects is addressed. For WeatherSafe, this suggests that validation reporting should capture at least:

- the intended use and the validated scope,
- the reference data pedigree and uncertainty characterization,
- the verification status and known numerical limitations, and
- the sensitivity of conclusions to scenario parameters and weather severity.

2.1.4 OVERVIEW OF EXISTING METHODOLOGIES FOR APSS MODEL VALIDATION

(Rosenberger 2022) reviews methodologies used in automotive and adjacent domains and highlights that no single, complete, APSS-specific holistic VV&UQ methodology has become a broadly accepted standard. However, several influential approaches contribute building blocks that are relevant to WeatherSafe.

Viehof (statistical assessment and labeling). (Viehof 2018) proposes a structured validation workflow using metric validity criteria, statistical reasoning, and a traffic-light labeling scheme for sample validity, including absolute and relative comparisons across samples. Such labeling can provide an efficient overview, but it can also remove information needed for later uncertainty aggregation and application-domain prediction. In addition, assumptions about distribution widths and acceptable dispersion may not hold for sensor signals, where a valid model should reproduce realistic variability rather than appearing artificially precise.

Schaermann (systematic data collection and multi-level evaluation). (Schaermann 2020) proposes an approach tailored to environment perception sensor models, emphasizing systematic data acquisition and evaluation on different interface levels (for example, detection, feature, and object level). This is valuable for practical test planning, but uncertainty handling and application-domain uncertainty prediction are typically not the main focus.

Ngo et al. (explicit vs. implicit evaluation and fidelity concepts). (Ngo, Bauer and Resch 2021) introduce concepts such as explicit versus implicit evaluation, where explicit evaluation targets sensor-level outputs directly and implicit evaluation evaluates downstream effects (for example, perception or fusion outputs). This is aligned with practical needs because requirements are often expressed at higher levels than raw sensor data. However, uncertainty treatment and intuitive metric interpretation for specification are commonly under-addressed in such approaches.

Riedmaier and Danquah (unified VV&UQ framework and survey). A major SotA contribution is the unified framework and survey by (Riedmaier, et al. 2021). Their work compares several VV&UQ approaches and proposes a generic framework that connects verification, calibration, validation, and application-domain inference, including mechanisms for error learning and integration of different uncertainty sources. Such frameworks address key gaps by moving beyond binary validation decisions and linking validation outcomes to application-domain uncertainty prediction.

Hofrichter et al. (continuous verification and validation integrated into simulation development). (Hofrichter, et al. 2025) propose a development method for active perception sensor simulations that integrates continuous verification and validation into an iterative implementation process. The method explicitly connects requirements definition, derivation of test cases with acceptance criteria, implementation, and repeated V&V in each iteration. Internal requirements are expressed as a structured set of sensor effects, causes, and sensor functions, building on PerCollECT and related specification approaches. From internal and external requirements, the authors derive primary, secondary, and external test cases and introduce a taxonomy that structures validation campaigns, validation suites, validation tests, and validation samples to support automated execution. The paper also proposes practical data validation steps (sanity and plausibility checks) for reference sensors, experimental setup, and real, reference, and simulated data before conducting re-simulation. For acceptance criteria, an empirical approach is discussed based on variation across multiple units of the same sensor, with a fallback to manufacturer specifications if the available sample size is insufficient. The overall approach is demonstrated on a ray tracing based lidar simulation architecture and highlights the need for automation, for example CI pipelines, as the number of test cases increases across iterations.

In summary, the SotA provides strong guidance on systematic experimentation and multi-level evaluation, and it shows an emerging consensus that uncertainty propagation and application-domain prediction are essential for credibility. Remaining gaps that are especially relevant for WeatherSafe are practical handling of uncertain weather references and selection of metrics that separate bias and scatter while remaining interpretable.

2.1.5 Uncertainty treatment and p-box representation (aleatory and epistemic)

A key contribution highlighted by (Rosenberger 2022) is the systematic separation of uncertainty types and sources, and the need to propagate them through sensor model validation workflows.

Uncertainty types. Epistemic uncertainty stems from lack of knowledge and can be reduced with better information. Aleatory uncertainty reflects inherent variability. For sensor model validation, reference data acquisition is typically affected by both types: variability in measured signals (aleatory) and uncertainty in calibration, alignment, labeling, environmental measurement, and repeatability (epistemic).

Uncertainty sources. A common categorization for model uncertainty includes model input uncertainty (parameters and boundary conditions, including reference data uncertainty), model form uncertainty (limitations of the modeling approach), and numerical uncertainty (discretization, solver effects, and implementation issues).

P-boxes as a representation for mixed uncertainty. (Rosenberger 2022) discusses probability boxes (p-boxes) as a way to represent combined epistemic intervals and aleatory distributions without forcing an unjustified probabilistic assumption for epistemic effects. This is relevant for WeatherSafe because adverse weather validation often depends on uncertain environmental quantities (for example precipitation intensity, visibility, and road spray conditions) that may be better represented with bounded ranges rather than single values.

Uncertainty aggregation and application-domain prediction. (Rosenberger 2022) summarizes the idea of obtaining metric results at multiple sample points in parameter space and learning an error or uncertainty model to predict expected model deviation and confidence bounds in unvalidated application areas. For WeatherSafe, this implies that a validation campaign should not only compute metric values at validation samples, but should also support a documented reasoning of how those results generalize across weather severity and across the defined parameter space.

2.1.6 INTERIM CONCLUSIONS FOR WEATHERSAFE GUIDELINES

Based on the SotA summarized above, the following conclusions are used as guiding principles for the remainder of this deliverable:

1. **Validation is use-case dependent.** The same sensor model may be sufficient for one intended use (for example, qualitative trend reproduction) but insufficient for another (for example, safety validation with tight acceptance thresholds). Validation should state the intended use and the validated scope explicitly.
2. **Reference data uncertainty should be treated explicitly.** Real-world reference data, including weather reference sensing, is uncertain and can dominate conclusions if not characterized.
3. **Bias and scatter should be separated conceptually.** Sensor models should replicate both systematic and random measurement behavior, and validation metrics and reporting should support separate assessment of these components.
4. **Non-deterministic validation is often necessary for credible extrapolation.** Under adverse weather, uncertainty in environmental parameters and in reference data often motivates bounded representations (for example p-boxes) rather than single point estimates.
5. **Methodology and metrics should support aggregation and interpretation.** Binary validation outcomes alone are insufficient for later uncertainty aggregation and application reasoning. Metrics should remain interpretable and support extrapolation-oriented reasoning.

These principles are operationalized in Section 2.2 (Validation Methodology), which proposes a structured workflow for planning experiments, collecting data, choosing metric levels, and documenting applicability for sensor models under weather influence.

2.2 ITERATIVE VALIDATION METHODOLOGY

This section describes an iterative validation methodology that combines (1) the six-stage verification and validation (V&V) methodology by (Viehof 2018), including the perception-sensor-specific adaptations described by (Rosenberger 2022), and (2) the continuous verification and validation development method proposed by (Hofrichter, et al. 2025). The combined view is used because it covers both the methodological structure of validation from (Viehof 2018) and its practical implementation as a continuously applied development process from (Hofrichter, et al. 2025).

2.2.1 VIEHOF METHODOLOGY IN DETAIL, ADAPTED FOR PERCEPTION SENSOR MODEL VALIDATION

Viehof proposes a structured V&V methodology consisting of six stages, from requirements definition to an objective statistical assessment of validity, as shown in Figure 1. (Rosenberger, Wendler, et al. 2019) discuss this methodology and introduces small but important adjustments for perception sensor model validation, mainly to reflect replay-to-sim workflows and to separate typical iteration causes (for example invalid reference data versus an insufficient model).

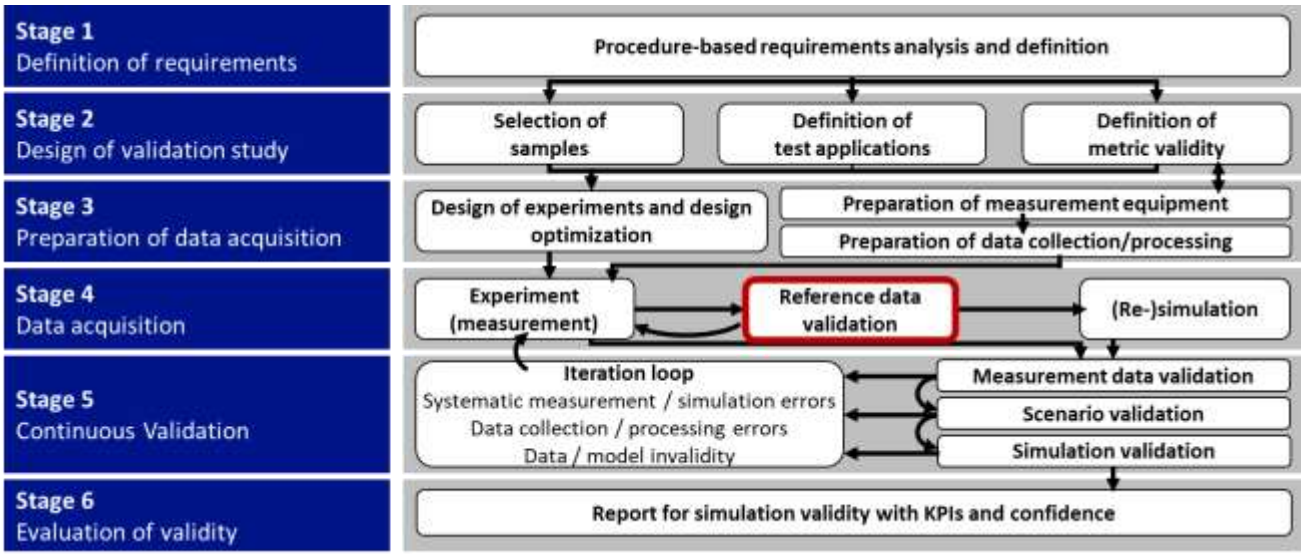

Figure 1: Step-wise validation methodology. Source: (Rosenberger, Wendler, et al. 2019)

1. **Stage 1: Definition of requirements.** Define the intended use, the validation scope, and metric validity criteria (MVCs) that can be evaluated on measured and simulated sensor outputs. For perception sensor models, MVCs should be tied to a defined sensor interface level (for example radar detections, radar cube, lidar point cloud, or object lists) and to measurable quantities.
2. **Stage 2: Design of validation study.** Define the parameter space and select validation samples, ideally supported by sensitivity analysis. For adverse weather, this should include weather severity parameters and the corresponding weather reference quantities.
3. **Stage 3: Preparation of data acquisition.** Prepare test setup, sensor configuration, and reference sensing so that replay-to-sim is possible and traceable.

4. **Stage 4: Data acquisition and analysis.** Conduct experiments and preprocess data. A key perception-sensor-specific point is to inspect the reference data early, because re-simulation is only meaningful if reference accuracy in time and space meets the requirements.
5. **Stage 5: Validation with explicit iteration causes.** Execute comparisons using MVCs and trigger iteration loops when issues are found. (Rosenberger 2022) refines this stage by separating validation levels (see below) so that failures can be traced to reference data, measurement processing, scenario parameterization, or model limitations.
6. **Stage 6: Evaluation of validity (objective statistical assessment).** Aggregate results over all samples and MVCs and document validity in a structured way. (Viehof 2018) introduces a validity assessment map (VAM) and a traffic-light labeling scheme to summarize absolute and relative validity.

For perception sensor model validation, (Rosenberger 2022) proposes separating Stage 5 into validation levels to make iteration causes explicit and to avoid attributing data problems to model deficiencies:

1. **Validation level 0 (reference data validation):** Check that reference measurements are suitable as simulation input (time synchronization, coordinate frames, completeness, plausibility, and uncertainty).
2. **Validation level 1 (measurement and simulation data validation):** Check for systematic issues in recorded and re-simulated data that would invalidate comparisons (for example preprocessing errors, misalignment, or corrupted recordings).
3. **Validation level 2 (scenario and parameter validation):** Check whether parameterization and sample selection represent the intended domain and whether assumptions from sensitivity reasoning are satisfied.
4. **Validation level 3 (model validation and sample validity):** Perform the quantitative comparison of measured versus simulated sensor outputs using the defined MVCs and metrics.

To summarize results across samples, Viehof's methodology uses the VAM concept, which organizes sample validity results per MVC and distinguishes absolute comparisons (measured vs. simulated for a given sample) from relative comparisons (how validity changes when stepping between samples). Labels (for example green, yellow, red) are then assigned based on statistical reasoning and predefined tolerance requirements. For perception sensors, (Rosenberger 2022) points out that distributional assumptions (for example normality and scale properties) must be justified and that it is often not appropriate to implicitly assume that simulated output variability is narrower than measured variability. Therefore, label maps are useful for overview, but should be complemented with continuous metrics and uncertainty statements when results are used for extrapolation or safety arguments.

2.2.2 HOFRICHTER ET AL.: CONTINUOUS VERIFICATION AND VALIDATION AS A DEVELOPMENT METHOD

(Hofrichter, et al. 2025) translate the idea of iterative validation into an engineering development method for active perception sensor simulations. The core contribution is explicit traceability from requirements to test cases and acceptance criteria, combined with continuous execution of verification and validation after each development iteration and regression testing to prevent credibility loss after updates. Figure 2 illustrates the iterative method.

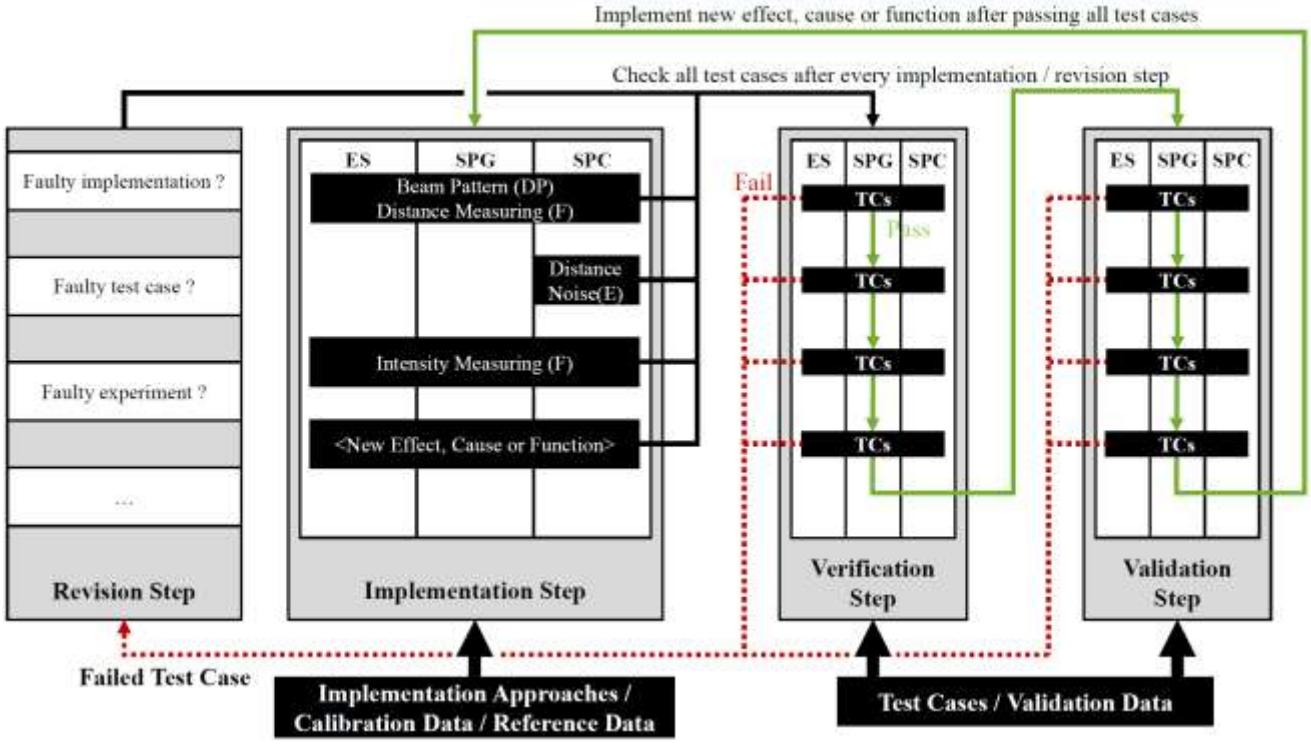

Figure 2: Iterative development method for sensor simulations with continuous V&V. The effects (E), design parameter causes (DPs), and functions (F) shown are to be considered examples. Test cases are abbreviated as TCs. Source: (Hofrichter, et al. 2025, 14)

Simulation architecture as context. The method is described for a modular sensor simulation architecture with environment simulation (ES), signal propagation (SPG), and signal processing components (SPC). This decomposition supports targeted verification and validation and helps localize failure causes.

Preparatory steps before iteration. Hofrichter et al. propose (i) defining external and internal requirements, where internal requirements are structured as effects, causes, and sensor functions, (ii) deriving primary, secondary, and external test cases from these requirements, (iii) structuring validation content using a taxonomy (validation campaign, validation suite, validation test, and validation sample) to support automation, (iv) defining and justifying acceptance criteria (for example empirically from unit-to-unit sensor variation, with fallback to manufacturer specifications), and (v) performing sanity and plausibility checks on reference sensors, experimental setup, and recorded data before replay-to-sim.

Iterative loop with continuous V&V. Each development iteration implements one isolated effect, cause, or function (or an inseparable set) and then executes verification and validation against the affected test cases. If a test fails, a revision step follows, and verification and validation are repeated. A defining feature is regression testing, meaning previously passed test cases are re-executed after each change. As the test portfolio grows, automated execution becomes essential, for example through CI-style pipelines.

2.2.3 Practical interpretation for WeatherSafe

For WeatherSafe, the Viehof methodology (with the perception-sensor-specific validation levels) provides the structure for planning experiments and documenting validity and

iteration causes. The Hofrichter et al. method complements this by providing a practical continuous V&V process that keeps validation traceable and repeatable as weather effects are added, parameterizations are refined, and datasets evolve. Section 2.3 builds on this by defining requirements and relevant weather effects and causes, and Section 2.4 specifies the reference data and data-quality checks required for replay-to-sim validation under adverse weather.

2.3 Requirements and Relevant Effects reg. Weather

This section describes how requirements for weather-related perception sensor simulation are derived in a structured and traceable way. The main idea is to avoid defining requirements as an ad hoc list of desired features. Instead, requirements should be derived from (i) a systematic catalog of sensor phenomena, effects, and causes, (ii) a relevance analysis for the operational design domain (ODD) and the system under test (SuT), and (iii) a graduated specification across sensor interfaces, including acceptance tests and data requirements. This approach follows the requirements workflow presented in Chapter 7 of (Rosenberger 2022), which is centered around PerCollECT[2] and CEPRA (Linnhoff, Rosenberger, et al. 2021) and results in concrete requirement and test artifacts.
For WeatherSafe, the weather focus primarily affects the selected causes and effect chains (for example precipitation, fog, spray, road wetness, illumination, and contamination) and can change the relevance ranking compared to clear-weather use cases. Therefore, the method in this section is used to identify which weather-induced effect chains should be simulated and validated first and which ones can be postponed due to limited project resources.

2.3.1 Requirements definition workflow (PerCollECT to CEPRA to specification)

The overall workflow to derive requirements for perception sensor simulation can be summarized as:

1. **Define the ODD and intended use of simulation.** Clarify where and how the simulation will be used (for example perception algorithm testing, sensor fusion testing, safety argumentation, or performance trend assessment under weather).
2. **Select the sensor technology and sensor interface level(s).** Determine whether requirements and validation are defined on detection level, point cloud level, radar cube level, object level, or multiple levels.
3. **Identify relevant phenomena, effects, and causes.** Use PerCollECT as a structured knowledge base of cause-effect chains to avoid missing relevant effects and to understand dependencies between effects.
4. **Prioritize cause-effect chains for the ODD using a relevance analysis.** Use CEPRA to rate occurrence in the ODD and impact on the SuT, producing a ranked list of effect chains to be addressed.
5. **Formulate requirements and acceptance tests.** Derive concrete, testable requirements and acceptance tests for the selected interface levels. For replay-to-sim validation, specify reference data requirements and checks so that re-simulation and comparison are feasible.

[2] https://github.com/PerCollECT

This workflow ensures traceability from the selected ODD and weather conditions to the list of modeled weather effects, and further to the validation test portfolio.
This workflow is also aligned with the preparatory steps in (Hofrichter, et al. 2025, Sec. 2.2.1), which emphasize the distinction between external requirements (constraints, interfaces, standards, fidelity, and available real sensor interfaces) and internal requirements that define which sensor and environment characteristics must be represented by the simulation.

2.3.2 PERCOLLECT AS THE BASIS FOR SYSTEMATIC IDENTIFICATION OF WEATHER-RELEVANT EFFECTS

A key challenge in perception sensor simulation is that weather-induced artifacts can result from many interacting effects across sensor functional layers. Modeling only a small subset without understanding interdependencies can lead to misleading results because an observed phenomenon can be caused by different combinations of effects, and effects can amplify or cancel each other.

PerCollECT[3], the Perception Sensor Collaborative Effect and Cause Tree, addresses this challenge by collecting and organizing cause-effect chains for lidar, radar, and camera sensor systems into a structured tree (graph) representation. In addition to the dissertation description (Rosenberger 2022), PerCollECT is maintained as a community-editable GitHub repository, which makes it suitable as a living knowledge base that can be extended as new weather-related findings emerge.

Following (Hofrichter, et al. 2025), PerCollECT distinguishes between sensor effects, design-parameter causes that are sensor-specific, and sensor-independent causes. PerCollECT itself does not provide a quantitative description of effects and causes or a specific definition of interfaces. Both must be added after selecting the individual effects and causes in order to apply the subsequent method. Quantitative descriptions allow the evaluation of effect-related and cause-related requirements via metrics, while interface definitions ensure that the effects and causes are detectable at the relevant sensor outputs (Hofrichter, et al. 2025).

For WeatherSafe, PerCollECT is especially valuable because it:

- supports completeness, meaning weather-related effects are less likely to be overlooked,
- makes interconnections explicit and helps identify when a weather phenomenon is a chain spanning propagation, receiver behavior, and processing,
- enables transparent communication of what is modeled and what is out of scope, which is essential for credibility, and
- provides a structured starting point for deriving requirements, test cases, and prioritization inputs.

In this deliverable, PerCollECT is used as the starting point for selecting weather-relevant cause-effect chains that should be addressed by simulation and validation activities.

2.3.3 CEPRA FOR PRIORITIZING WEATHER INFLUENCE (OCCURRENCE AND IMPACT)

PerCollECT provides the catalog and structure, but it does not decide what is relevant for a specific ODD and SuT. CEPRA, the Cause, Effect, and Phenomenon Relevance Analysis,

[3] https://github.com/PerCollECT

provides the prioritization step needed to derive requirements with a realistic scope (Rosenberger 2022).

CEPRA is performed by selecting relevant cause-effect chains and then rating each chain with two expert inputs:

1. **Occurrence in the ODD (O):** how frequently the phenomenon and its cause-effect chain is expected in the defined ODD.
2. **Impact on the SuT in the ODD (I):** how critical the phenomenon is for the SuT's performance or safety goals.

In line with (Hofrichter, et al. 2025), CEPRA can be performed by two experts who assess occurrence in the ODD and impact on the SuT. As an alternative or complement, a sensitivity analysis can be used to quantify the influence of individual effects and causes on a downstream SuT and prioritize those with the greatest influence. This can be especially relevant when the SuT includes machine learning components that are difficult to interpret by expert judgment alone (Hofrichter, et al. 2025).

The overall relevance can be computed as a combined score (for example a sum), which prevents rare but high-impact weather phenomena from being filtered out too early. CEPRA also captures rationale so that expert judgments remain transparent and can be reviewed and updated as new evidence becomes available.

For adverse weather, CEPRA enables explicit prioritization of chains such as atmospheric attenuation and scattering (fog, haze), rain and wet-environment scattering and multipath, reduced received power due to wet targets and changed reflectivity, contamination and occlusion of sensor apertures due to spray, snow, or mud, and illumination-related effects for cameras (glare, low light, precipitation on lens).

2.3.4 Graduated requirements and acceptance tests across sensor interfaces

After prioritization, requirements should be formulated in a graduated manner across sensor interfaces. The motivation is that different use cases require different interface levels, and weather effects often manifest first at lower levels (for example signal attenuation, increased noise, missing detections) before they propagate to higher levels (for example missed objects and degraded tracking).

For WeatherSafe, this means that each requirement should explicitly state:

- the weather condition and its severity definition used in the requirement,
- the sensor interface level where the requirement is evaluated,
- the metric and acceptance criterion, and
- the required reference data and uncertainty assumptions needed to validate the requirement.

2.3.5 Expected outputs for WeatherSafe (artifacts)

Applying the workflow shown in this section should produce at least the following concrete artifacts:

- a selected subset of weather-relevant PerCollECT cause-effect chains for each sensor modality (lidar, radar, camera),
- a CEPRA table with ranked chains, including documented rationale,
- a graduated list of requirements and acceptance tests per interface level, and
- a mapping from each acceptance test to the necessary experiment and reference data to support replay-to-sim validation.

2.4 Reference Data incl. Weather Sensors for Re-Simulation

This section summarizes the reference data required to enable replay-to-sim (re-simulation) validation of perception sensor models under adverse weather. The structure and practical recommendations follow the case-study workflow presented in deliverable D_Simu.2.
The core guideline is that weather-aware sensor model validation is only as credible as the reference data used to (a) parameterize the environmental condition and (b) reproduce the measurement setup in simulation. Therefore, reference data should be planned as a first-class validation artifact, including data-quality checks, metadata, and uncertainty considerations.

2.4.1 Purpose of reference data in replay-to-sim validation

Replay-to-sim validation compares measured sensor outputs against simulated outputs for the same experimental situation. In practice, this requires converting measured information into simulation inputs, including:

- environmental parameters, including weather severity descriptors and time stamps,
- scene layout and target definitions,
- sensor pose, mounting, and relevant configuration, and
- time synchronization across all recorded streams.

Deliverable D_Simu.2 demonstrates this by linking sensor KPIs to quantitative weather descriptors (for example illumination in lux, precipitation intensity in mm/h, and fog descriptors such as extinction coefficient or meteorological optical range). This yields performance curves and distributions that can support model calibration, plausibility checks, and definition of acceptance bands for validation.

2.4.2 Minimum reference data categories

For weather-aware sensor model validation, the required reference data can be grouped into the following categories:

- **Sensor under test (SuT) data:** raw or near-raw outputs at the chosen interface level (for example camera images, lidar point clouds, radar detections, radar cube), plus sensor configuration metadata needed for interpretation.
- **Scene and target reference:** geometry and relevant material properties of validation targets, plus a consistent coordinate-frame description and sensor mounting information for spatial alignment.
- **Weather reference sensing:** quantitative descriptors that parameterize weather severity, synchronized to SuT data, plus classification information to separate weather categories and detect mixed conditions.
- **Timing and synchronization:** a consistent time base across all recorded streams, including documented sampling intervals and any temporal averaging performed by weather sensors.
- **Metadata and provenance:** measurement identifiers, setup descriptors, timestamps, and structured summaries to support automated postprocessing and reproducibility.

2.4.3 Example reference sensing setup (case study from Deliverable D_Simu.2)

Deliverable D_Simu.2 provides a practical example of a stationary measurement setup that supports reproducible performance characterization under adverse weather. The Weather Test Bench from Persival[4] combines defined reference targets, synchronized perception sensors, and dedicated weather reference sensing.

Targets and geometry. Defined reference targets at known distances enable stable, repeatable regions of interest for sensor evaluation. Target choices can include multiple reflectance levels for optical sensors and a consistent target layout to support spatial filtering in lidar and radar data.

Weather reference sensing. The case study records quantitative weather descriptors that support parameterization of precipitation, illumination, and fog. Typical quantities include precipitation rate (mm/h), illumination (lux), classification information (for example present-weather codes), fog descriptors such as meteorological optical range and an extinction coefficient, and general atmospheric quantities such as temperature, humidity, pressure, and wind. Where available, a disdrometer-style measurement can add particle size and velocity information to interpret precipitation regimes beyond a single mm/h proxy.

Measurement automation and baselines. Deliverable D_Simu.2 describes using weather-threshold-triggered recordings for adverse conditions and maintaining clear-condition reference measurements as a baseline. Baselines are essential to detect sensor drift and configuration changes and to define a zero-severity reference for later degradation analysis.

2.4.4 Data structure and metadata for reproducibility

Replay-to-sim validation requires more than storing raw sensor streams. The dataset should preserve the context needed to reconstruct the experiment and to parameterize simulation runs. A practical approach is to store synchronized recordings of SuT outputs and weather reference sensors together with (i) a structured metadata file per measurement (setup, sensor configuration, and classification), (ii) a per-measurement summary of relevant weather quantities (for example min, mean, and max values), and (iii) an explicit representation of sensor mounting and static coordinate transforms to support consistent spatial filtering and region-of-interest evaluation.

[4] https://www.persival.de/weather-measurements

2.4.5 Data-quality checks and validity conditions before comparison

Before running a sensor model comparison, reference data validity should be established (compare Section 2.2). The following checks are recommended:

- **Time synchronization and sampling-rate checks:** verify alignment between SuT data and weather reference streams; document temporal averaging in weather sensors because it impacts alignment with short-lived sensor artifacts.
- **Coordinate frames and geometry checks:** verify target positions, sensor mounting transforms, and the coordinate frame used for filtering; confirm expected target appearance under baseline conditions.
- **Baseline checks:** ensure clear-condition reference measurements exist and are used as a zero-severity baseline; use them to detect drift, misalignment, or configuration changes.
- **Weather category plausibility:** verify consistency of weather categorization and flag mixed conditions; e.g. for fog prefer physically meaningful severity descriptors such as extinction coefficient or meteorological optical range.
- **Completeness and plausibility:** check that all required streams and metadata are present; screen for outages, saturation, and missing data in weather sensors.

2.4.6 Linking reference data to KPIs and acceptance bands

Deliverable D_Simu.2 illustrates how well-structured reference data enables KPIs that can later be compared between measurement and simulation. The recommended perspective is typically not a strict scan-by-scan match. Instead, comparisons should focus on trend agreement over weather severity and on agreement with variability envelopes, because weather-driven effects are often stochastic and spatially inhomogeneous.

Camera KPI mapping. Fixed regions of interest (ROIs) are defined on reference targets. Example KPIs include mean intensity per ROI, variance or signal-to-noise proxies, and inter-target contrast. These KPIs can then be analyzed as a function of illumination and weather descriptors to produce performance curves and distributions that support acceptance-band definition.

Lidar KPI mapping. A useful decomposition is to evaluate (i) an atmospheric region where no detections are expected under clear conditions and quantify weather-induced detections as a detection probability per beam, and (ii) target regions where spatial filters isolate detections belonging to defined targets. Target metrics can include detection counts and intensity or reflectivity statistics as a function of weather severity.

2.4.7 Limitations and transferability notes

- **Fog:** extinction coefficient or meteorological optical range are physically meaningful severity descriptors and typically support interpretable performance trends for optical sensors.
- **Precipitation:** precipitation rate in mm/h is often an imperfect proxy because particle size distributions and spatial inhomogeneity can dominate sensor response. Validation should therefore emphasize trends and variability envelopes rather than purely deterministic attenuation curves.
- **Spatial mismatch and inhomogeneity:** weather reference sensors do not measure exactly the same volume as the sensor field of view. Patchy fog or localized spray can

cause sensor artifacts that are not captured by global descriptors, so acceptance criteria should account for this mismatch.

The specific sensor list, file formats, and thresholds should be adapted to the concrete WeatherSafe measurement setup. However, the data categories, quality checks, and the KPI-to-weather linkage concept from deliverable D_Simu.2 remain applicable across setups.

3. Metrics for Sensor Model Validation

Metrics are the link between experimental data and a defensible validity statement. In the context of active perception sensor systems (APSS), metrics should be selected so that they match the sensor interface used in the development and integration chain, remain interpretable for acceptance criteria discussions, and support uncertainty-aware evaluation under imperfect reference sensing and limited repeatability. While (Rosenberger 2022) provides an extensive analysis of metric candidates and introduces the double validation metric (DVM) concept, (Elster, et al. 2024) extend this work for radar sensor models and introduce the DVM map.

Despite the extensive analysis by Rosenberger, as mentioned before, the metrics discussed and used in the following are not the only ones possible. There are other metrics used for sensor model validation in literature, e.g. by (Ngo, Bauer and Resch 2021) or (Magosi, et al. 2022).

The DVM map increases diagnostic power by resolving limitations that occur when multiple empirical distributions are aggregated into probability boxes (p-boxes), and it provides a structured way to identify worst-case combinations of measurement and simulation uncertainty. This chapter uses (Elster, et al. 2024) as the primary guideline for metric selection and application, and it highlights implications for dynamic scenarios.

3.1 Requirements for validation metrics and the role of sensor interfaces

A validation metric should support quantitative comparison and engineering interpretability at the same time. For radar, (Elster, et al. 2024) highlight that validation is challenging because radar outputs are shaped strongly by signal processing (for example FFT processing and thresholding), and results can be affected by clutter, multipath propagation, and radar cross section sensitivities. Radar also measures relative radial velocity via Doppler, which enables dynamic validation but introduces additional comparability and time-dependent effects.

- **Detections interface:** sparse output that is strongly coupled to thresholds and internal processing. It supports evaluation close to downstream perception use, but comparability can break if the number of detections differs substantially between measurement and simulation.
- **Radar cuboid interface:** dense grid of power values after FFT processing (range, azimuth, Doppler bins). It supports distribution-based comparisons of power and noise behavior and is often more stable with respect to sample size because it is not restricted to detected objects only.
- **Higher-level object interfaces:** close to function evaluation, but dependent on tracking and association logic and can hide earlier modeling errors.

Across all interfaces, WeatherSafe-relevant validation should be uncertainty-aware because reference data (sensor pose, target pose, calibration, and weather descriptors) is never exact. This becomes especially important under adverse weather, where weather descriptors and reference sensing may be spatially mismatched to the sensor field of view.

3.2 AVM, CAVM, and DVM as the core metric family (bias and scattering)

(Elster, et al. 2024) build on the area validation metric (AVM), which quantifies the area between cumulative distribution functions (CDFs) or between probability boxes (p-boxes). In this framework, aleatory uncertainty is represented by empirical distributions, while epistemic uncertainty introduced by uncertain reference inputs is reflected by the p-box width. A key requirement for active perception sensor simulation validation is to separate two different modeling-error types: model bias error (systematic deviation, comparable to a shift in distribution location) and model scattering error (deviation in distribution shape and spread).

The DVM framework preserves this separation by reporting a bias-related component derived from the AVM split and a scattering-related component based on a corrected AVM (CAVM) after compensating the estimated bias.

3.3 DVM LIMITATIONS AND WHY THE DVM MAP IS NEEDED

(Elster, et al. 2024) discuss limitations that can arise when uncertainty is handled only via p-box aggregation:

- **Loss of information inside the p-box.** Aggregating multiple empirical distributions into a p-box retains only the envelope, while the internal structure of the distribution set is lost.
- **Edge cases driven by p-box size.** Certain combinations of large simulation p-box and small measurement p-box (or vice versa) can lead to misleadingly small AVM and CAVM values.
- **Missing attribution to uncertainty combinations.** P-box aggregation makes it difficult to identify which specific uncertainty combination causes the worst deviation.

To address these issues, (Elster, et al. 2024) introduce the DVM map. Instead of comparing aggregated p-boxes only, the DVM map compares each simulation empirical distribution against each measurement empirical distribution, yielding a matrix (heatmap) that visualizes (i) the absolute bias magnitude, (ii) the scattering error (CAVM), and (iii) a combined score (absolute bias plus CAVM) to identify worst-case combinations. For WeatherSafe, this is especially relevant because uncertain reference inputs are expected under adverse weather, and the DVM map makes sensitivity to uncertainty combinations visible.

3.4 ADDED ASPECTS FOR DYNAMIC SCENARIOS (ESPECIALLY IMPORTANT FOR WEATHERSAFE)

(Elster, et al. 2024) emphasize that Doppler information opens the path toward dynamic scenario validation especially for radar and FMCW-lidar sensors, but it also introduces new challenges. For WeatherSafe, dynamics matter because weather influence often occurs during motion (ego motion and object motion) and because weather phenomena can be intermittent (spray, transient fog patches, wiper cycles, changing glare). When moving from static to dynamic scenarios, the metric workflow requires additional care:

- **Time aggregation and windowing:** define whether distributions are computed per frame, per fixed time window, or per scenario segment. The decision affects stability and interpretability.
- **Sample-size comparability:** empirical-distribution metrics require consistent sampling logic. At detection level, variable detection counts can invalidate direct comparisons, which motivates increased use of radar cuboid evaluation or robust ROI logic.
- **Doppler-dimension handling:** define how Doppler slices are selected or aggregated and how Doppler is aligned across measurement and simulation. In dynamic scenarios, misalignment can occur due to timing offsets, ego-motion uncertainty, and processing differences.

- **Time alignment and reference uncertainty:** synchronization between radar, reference sensors, and simulation time becomes more critical. Uncertainty in reference pose, velocity, and timing can dominate results and should be propagated and analyzed using DVM maps.
- **Separating weather influence from motion artifacts:** motion-induced effects (for example changing aspect angle and radar cross section behavior) can alter returns independent of weather. Dynamic validation designs should either control motion patterns or treat motion parameters explicitly as part of the validation parameter space.

3.5 PRACTICAL RECOMMENDATIONS FOR METRIC SELECTION AND REPORTING IN WEATHERSAFE

- **Define metrics per interface level, not only per sensor type.** For radar, prioritize radar cuboid evaluation for stable distribution comparisons. Use detections primarily to define ROIs and to support localized analysis.
- **Report bias and scattering separately.** Use at least one bias-related and one scattering-related value to avoid hiding systematic shifts behind distribution spread.
- **Use DVM maps when uncertain reference inputs are relevant.** Under adverse weather, uncertainty in weather descriptors and reference data is expected. DVM maps provide a structured way to visualize sensitivity to uncertainty combinations and identify worst cases.
- **Prefer trend and envelope validation over strict frame-to-frame matching.** Weather-driven effects are often stochastic and spatially inhomogeneous. Focus on trend agreement over weather severity and on agreement with variability envelopes.
- **Keep results interpretable.** Use visualizations such as heatmaps, ROI-level results, and localized cuboid projections to support acceptance criteria discussions and model refinement.

4. Guidelines for Weather Effect Simulation (Role-Model Based Environmental Influence Modeling)

Chapter 3 focuses on quantitative metric-based validation by comparing measured and simulated sensor outputs at selected samples (or scenario segments) in parameter space, using DVM and the DVM map (Elster, et al. 2024). This is essential for uncertainty-aware, distribution-based comparisons and for extending validation toward dynamic scenarios.

However, aside from direct sample-to-sample comparisons using DVM, credible weather effect simulation also requires a second, complementary perspective: understanding and modeling environmental influences as measurable, parameterized role-model behavior. In this perspective, weather influences are characterized by expectation values derived from real-world datasets with synchronized weather reference sensing and then used to calibrate, sanity-check, and modularize effect implementations. This approach is described in detail by (Linnhoff 2023), and it was already used implicitly as a role model in deliverable D_Simu.2 when deriving KPIs and trends from stationary measurements with weather reference sensors.

Therefore, this chapter complements Chapter 3 by describing how role-model based environmental influence modeling supports model development, calibration, debugging, and plausibility validation of weather effects. In practice, it helps answer whether a model reproduces the expected scaling of weather-induced behavior with respect to quantified environmental descriptors, even before applying uncertainty-aware DVM comparisons across selected parameter-space samples.

4.1 Two complementary validation perspectives (how this chapter relates to Chapter 3)

For WeatherSafe, validation of weather effect modeling benefits from two complementary perspectives:

1. **Sample-based comparison in parameter space (metric-first, Chapter 3).** Measured and simulated sensor outputs are compared using uncertainty-aware metrics (DVM and DVM map). This perspective supports quantitative statements about bias and scattering errors, identifies worst-case uncertainty combinations, and supports dynamic scenario analysis when time windowing, Doppler handling, and comparability constraints are defined consistently (Elster et al. 2024).
2. **Role-model based environmental influence modeling (effect-first, this chapter).** Environmental influences are characterized by measurable changes in sensor behavior (for example atmospheric false detections and detection loss due to attenuation) as a function of quantified environmental descriptors (for example visibility, precipitation rate, and illumination). This perspective, shown in Figure 4, emphasizes experimental design, reference sensing quality, expectation-value extraction, and modular effect implementation and calibration (Linnhoff 2023).

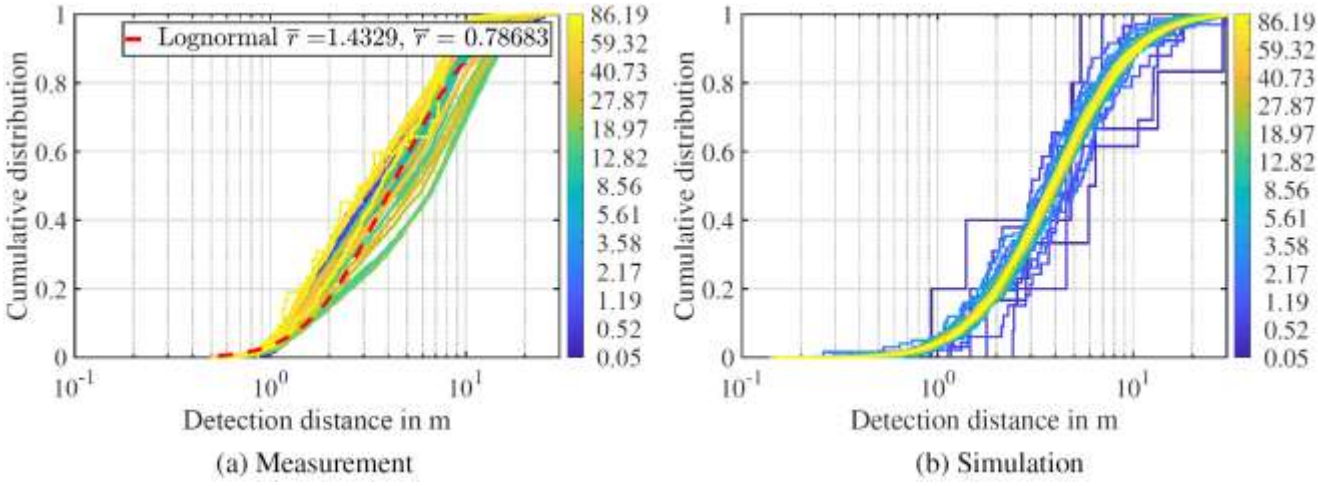

Figure 3: Comparison of detection distributions in atmosphere for rain. (Linnhoff 2023, 124).

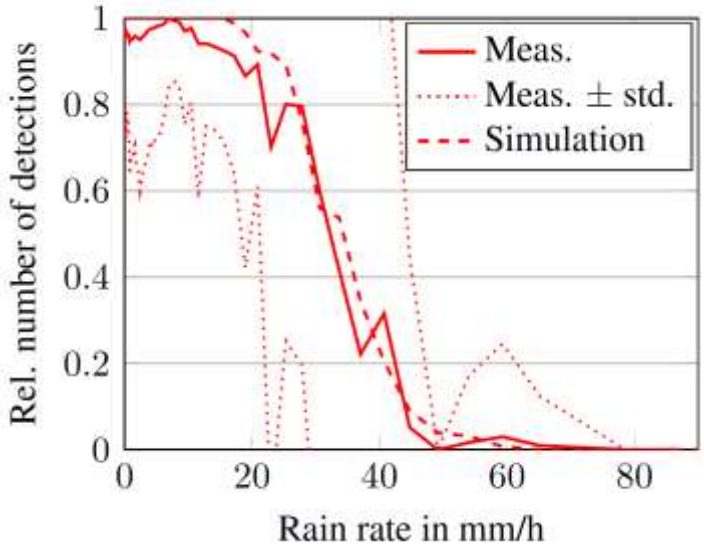

Figure 4: Comparison of target detection probability for measurement and simulation in rain (Linnhoff 2023, 125). The dotted lines denote the measurement with added and subtracted standard deviation as an upper and a lower boundary.

These perspectives serve different purposes and should be used together: role-model based influence analysis provides physically and statistically plausible targets for effect behavior and parameterization, while DVM-based comparison provides the uncertainty-aware quantitative comparison and acceptance discussion for selected samples and scenarios.

4.2 CATEGORIZATION OF ENVIRONMENTAL CONDITIONS AND IMPLICATIONS FOR MODELING SCOPE

A core concept from (Linnhoff 2023) is to categorize environmental conditions in a way that directly supports experiment design and effect modeling:

1. **Object-independent environmental conditions**, such as fog, rain, snow, and direct sunlight. These can be parameterized as environment descriptors in simulation, but their impact is still influenced by scene interactions (for example rain absent under bridges, sunlight blocked by buildings).
2. **Object-dependent environmental conditions**, such as road spray. These arise through interaction between objects and environment (for example spray requires a moving vehicle, wet pavement, and turbulent aerodynamics).

For WeatherSafe, this categorization clarifies why a single measurement strategy is insufficient. Object-independent effects can often be analyzed with stationary long-term datasets and weather reference sensing, while object-dependent effects typically require controlled dynamic experiments and additional reference information about the generating object and interaction.

4.3 REFERENCE SENSING AND PARAMETERIZATION: WHAT MUST BE MEASURED (AND AT WHAT RATE)

A key takeaway from (Linnhoff 2023) is that qualitative weather labels are not sufficient for model calibration and validation. Environmental conditions should be parameterized with quantitative descriptors that are compatible with operational design domain definitions and simulation interfaces.

(Linnhoff 2023) also shows that environmental conditions can change rapidly. If weather reference sensing is too slow or too strongly averaged, the pairing of sensor behavior to environmental descriptors becomes noisy and reduces calibration and validation value. This directly connects to Section 2.4: weather reference sensing should be treated as a primary validation artifact, and its update rate and averaging behavior should be documented.

4.4 EXTRACTING EXPECTATION VALUES AS ROLE-MODEL KPIS FOR WEATHER EFFECTS

(Linnhoff 2023) proposes extracting expectation values from real-world datasets to quantify dominant weather impacts visible in lidar data, most notably (i) atmospheric false positive detections due to scattering and backscattering by droplets or particles and (ii) attenuation and detection loss due to increased extinction along the propagation path.

A robust KPI for atmospheric detections is the detection probability per beam in an atmospheric region where no detections are expected in clear conditions. This normalization supports comparability across sensors with different angular resolutions. In addition to existence probability, atmospheric detections can be characterized by distance and intensity distributions.

For attenuation effects, defined reference targets at multiple distances enable evaluating how target detection probability decreases as a function of visibility (fog) or precipitation rate (rain). A key design element is a stationary setup with an unobstructed atmospheric

region and defined targets, so that both atmospheric clutter and detection loss trends can be extracted as expectation-value curves.

These expectation values do not replace DVM. Instead, they serve as role-model KPIs that calibrate and sanity-check individual weather effect modules, reduce ambiguity when interpreting later DVM results, and support model debugging by indicating which effect behavior is missing or mis-parameterized.

4.5 IMPLICATIONS FOR SIMULATION ARCHITECTURES: MODULAR EFFECT IMPLEMENTATION AND CALIBRATION

(Linnhoff 2023) emphasizes that modeling environmental influence is not only a physics problem, because sensor-specific signal processing and thresholding can dominate observable behavior. Even sensors operating at similar wavelengths can show substantially different responses to the same weather condition. Therefore, calibration and validation should be sensor-specific.

For implementation, (Linnhoff 2023) describes a modular simulation approach aligned with OSI and OSMP, where effects can be inserted at different entry points. In many practical workflows, effect modules are implemented at detection level, where they can add atmospheric detections (parameterized by existence probability and state distributions) and attenuate or remove detections generated by a base model to represent extinction and detection loss. This modular structure also supports continuous V&V workflows (Hofrichter, et al. 2025)) because individual effects can be implemented, calibrated, and revalidated incrementally.
(Linnhoff 2023) also highlights limitations that should be made explicit in WeatherSafe guidelines because they can dominate discrepancies between measurement and simulation:

- reference sensing mismatch, because weather sensors do not measure the same volume as the sensor field of view,
- strong variability in snow, different regimes can occur at similar precipitation rates,
- direct-sun effects that are strongly angle dependent and can be influenced by cloud structure, and
- lidar-to-lidar interference in multi-sensor experiments.

4.6 HOW WEATHERSAFE SHOULD USE THE ROLE-MODEL PERSPECTIVE TOGETHER WITH DVM

Based on (Linnhoff 2023), WeatherSafe should treat role-model based environmental influence modeling as an enabling layer that makes DVM-based validation more effective:

- Use expectation-value KPIs (for example atmospheric detection probability per beam, distance and intensity distributions, and target detection loss curves) to calibrate and sanity-check individual weather effect modules.
- Apply DVM and DVM maps (Chapter 3) afterward for uncertainty-aware, quantitative validation across selected parameter-space samples and scenario segments, especially when moving toward dynamic scenarios.

In short, DVM answers how far measurement and simulation differ at selected samples, while the role-model approach clarifies which weather effect behavior should exist and how it should scale with quantified environmental descriptors. Both are needed for credible adverse-weather sensor simulation.

5. Conclusion

This deliverable consolidates a practical guideline for validating perception sensor models under adverse weather, with an explicit focus on traceability, uncertainty awareness, and applicability for continuous simulation development. The guideline builds on three complementary pillars: (i) a structured VV&UQ process for planning and executing validation activities, (ii) a reference-data and replay-to-sim view that treats weather sensing and metadata as first-class validation artifacts, and (iii) metric and role-model concepts that enable quantitative and interpretable statements about model validity under weather influence.

Chapter 2 established the validation foundation. First, it summarized the state of the art and clarified terminology and credibility expectations for active perception sensor simulations. Second, it described an iterative validation methodology based on (Viehof 2018)'s staged V&V structure, augmented with perception-sensor-specific validation levels to separate reference data issues from actual model deficiencies. Third, it embedded this methodology into a continuous verification and validation development process following (Hofrichter, et al. 2025), emphasizing requirements traceability, acceptance criteria, and regression testing after each model update. Finally, it defined how weather-relevant requirements should be derived systematically using PerCollECT and prioritized using CEPRA, and it summarized the reference data needed for replay-to-sim validation based on the measurement workflow described in deliverable D_Simu.2.

Chapter 3 focused on metrics for sensor model validation, using (Elster, et al. 2024) as the primary reference. The DVM framework supports uncertainty-aware comparison of measured and simulated sensor outputs and separates systematic bias from scattering errors. The DVM map extends this concept by resolving the influence of specific uncertainty combinations, which is especially important when weather descriptors and other reference inputs are uncertain. The chapter also highlighted key considerations when moving from static measurements toward dynamic scenarios, including time windowing, Doppler handling for dynamic scenarios, sample-size comparability at detection interfaces, and tighter requirements on time alignment and reference uncertainty characterization.

Chapter 4 added the complementary role-model perspective from (Linnhoff 2023). While DVM-based comparisons quantify differences at selected samples, role-model based environmental influence modeling provides calibration targets and plausibility checks for individual weather effects. Expectation-value KPIs such as atmospheric detection probability per beam, as well as distance and intensity distributions and target detection-loss trends as a function of quantified weather descriptors, help ensure that effect modules scale realistically before performing sample-to-sample DVM validation.

In WeatherSafe practice, this combination improves diagnostic capability: role-model KPIs support effect-level debugging and calibration, while DVM and DVM maps provide uncertainty-aware acceptance discussions on selected samples and scenario segments.
Based on the consolidated guideline, the following takeaways and next steps are recommended:

- **Treat validation as a lifecycle process.** Apply continuous verification and validation with traceable requirements, test cases, and regression testing so that credibility is maintained across model updates.

- **Derive scope from PerCollECT and prioritize with CEPRA.** Use a transparent cause effect chain selection and ranking to keep modeling and validation effort focused on the most ODD-relevant weather influences.
- **Invest in reference data quality.** Replay-to-sim validation requires synchronized sensor outputs, weather reference sensing, transforms, metadata, and documented averaging and uncertainty characteristics, as reflected in the deliverable D_Simu.2 workflow.
- **Use role-model KPIs for effect-level calibration.** Establish expectation-value trends (for example atmospheric detections and attenuation curves) as sanity checks for individual weather effect modules before running broader metric-based validation.
- **Use DVM and DVM maps for quantitative acceptance discussions.** Report bias and scattering separately and use DVM maps to identify worst-case uncertainty combinations, especially when weather descriptors are uncertain.
- **Plan explicitly for dynamic scenarios.** Define time windows, alignment tolerances, Doppler handling, and interface comparability conditions up front to avoid ambiguous results when evaluating moving scenes.

Despite these guidelines, validation under adverse weather remains constrained by limited repeatability of real weather, spatial mismatch between weather reference sensors and the perception sensor field of view, and the stochastic nature of precipitation and spray. For this reason, WeatherSafe validation should generally emphasize trend and envelope agreement over strict frame-to-frame matching, and it should document uncertainty sources and known limitations explicitly. As more measurement data becomes available and additional weather effects are implemented, the guideline can be applied iteratively by extending PerCollECT cause-effect coverage, updating CEPRA prioritizations, and expanding the continuous V&V test portfolio and DVM map analyses toward increasingly complex and dynamic scenarios.

6. Acknowledgements

The WeatherSafe consortium would like to acknowledge for their support and work all the Partners and Third Parties involved:



7. References

Elster, Lukas, Philipp Rosenberger, Martin Holder, Ken Mori, Jan Staab, and Steven Peters. 2024. "Introducing the double validation metric for radar sensor models." *Automotive and Engine Technology* 9 (1): 6. doi:10.1007/s41104-024-00143-5.

Hofrichter, Kristof, Lukas Elster, Clemens Linnhoff, Timm Ruppert, and Steven Peters. 2025. "Introducing a Development Method for Active Perception Sensor Simulations Using Continuous Verification and Validation." (Sensors) (25): 7642. doi:10.3390/s25247642.

Holder, Martin Friedrich. 2021. *Synthetic Generation of Radar Sensor Data for Virtual Validation of Autonomous Driving.* Darmstadt: PhD Thesis, Technische Universität Darmstadt. doi:10.26083/tuprints-00017545.

Linnhoff, Clemens. 2023. *Analysis of Environmental Influences for Simulation of Active Perception Sensors.* Darmstadt: PhD Thesis, Technische Universität Darmstadt.

Linnhoff, Clemens, Philipp Rosenberger, Simon Schmidt, Lukas Elster, Rainer Stark, and Hermann Winner. 2021. "Towards Serious Perception Sensor Simulation for Safety Validation of Automated Driving - A Collaborative Method to Specify Sensor Models." *International Conference on Intelligent Transportation Systems (ITSC).* Indianapolis, IN, USA. 2688-2695. doi:10.1109/ITSC48978.2021.9564661.

Magosi, Zoltan Ferenc, Christoph Wellershaus, Viktor Roland Tihanyi, Patrick Luley, and Arno Eichberger. 2022. "Evaluation Methodology for Physical Radar Perception Sensor Models Based on On-Road Measurements for the Testing and Validation of Automated Driving." *Energies* 15: 2545.

Ngo, Anthony, Max Paul Bauer, and Michael M. Resch. 2021. "A Multi-Layered Approach for Measuring the Simulation-to-Reality Gap of Radar Perception for Autonomous Driving." *IEEE International Intelligent Transportation Systems Conference (ITSC).* Indianapolis, IN, USA. 4008 - 4014.

Riedmaier, Stefan, Benedikt Danquah, Bernhard Schick, and Frank Diermeyer. 2021. "Unified Framework and Survey for Model Verification, Validation and Uncertainty Quantification." *Arch Computat Methods Eng* 28: 2655–2688.

Rosenberger, Philipp. 2022. *Metrics for Specification, Validation, and Uncertainty Prediction for Credibility in Simulation of Active Perception Sensor Systems.* Darmstadt: PhD Thesis, Technische Universität Darmstadt. doi:10.26083/tuprints-00023034.

Rosenberger, Philipp, Jan Timo Wendler, Martin Friedrich Holder, Clemens Linnhoff, Moritz Berghöfer, Hermann Winner, and Markus Maurer. 2019. "Towards a Generally Accepted Validation Methodology for Sensor Models - Challenges, Metrics, and First Results." *Grazer Symposium Virtuelles Fahrzeug.* Graz, Austria.

Schaermann, Alexander. 2020. *Systematische Bedatung und Bewertung umfelderfassender Sensormodelle.* München: PhD Thesis, Technische Universität München.

Viehof, Michael. 2018. *Objektive Qualitätsbewertung von Fahrdynamiksimulationen durch statistische Validierung.* Darmstadt: PhD Thesis, Technische Universität Darmstadt.

8. Table of illustrations

Figure 1: Step-wise validation methodology. Source: (Rosenberger, Wendler, et al. 2019) 13
Figure 2: Iterative development method for sensor simulations with continuous V&V. The effects (E), design parameter causes (DPs), and functions (F) shown are to be considered examples. Test cases are abbreviated as TCs. Source: (Hofrichter, et al. 2025, 14) 15
Figure 3: Comparison of detection distributions in atmosphere for rain. (Linnhoff 2023, 124). 28
Figure 4: Comparison of target detection probability for measurement and simulation in rain (Linnhoff 2023, 125). The dotted lines denote the measurement with added and subtracted standard deviation as an upper and a lower boundary..... 28