Introducing the Double Validation Metric for Radar Sensor Models

Lukas Elster (✉ lukas.elster@tu-darmstadt.de)

Technical University of Darmstadt

Philipp Rosenberger

Persival GmbH

Martin Holder

Robert Bosch GmbH

Ken Mori

Technical University of Darmstadt

Jan Staab

Technical University of Darmstadt

Steven Peters

Technical University of Darmstadt

Research Article

Additional Declarations: No competing interests reported.

Introducing the Double Validation Metric for Radar Sensor Models

Lukas Elster[1*], Philipp Rosenberger[2*], Martin Holder[3], Ken Mori[1], Jan Staab[1]
and Steven Peters[1]

[1]Institute of Automotive Engineering, Technical University of Darmstadt,
Otto-Berndt-Str. 2, Darmstadt, 64287, Germany.
[2]Persival GmbH, Carlo-Mierendorff-Str. 2, Ober-Ramstadt, 64372, Germany.
[3]Robert Bosch GmbH, Postfach 10 60 50, Stuttgart, 70049, Germany.

*Corresponding author(s). E-mail(s): lukas.elster@tu-darmstadt.de;
philipp.rosenberger@persival.de;
Contributing authors: martin.holder@de.bosch.com; ken.mori@tu-darmstadt.de;
jan.staab@stud.tu-darmstadt.de; steven.peters@tu-darmstadt.de;

Abstract

In automated vehicles, environment perception is performed by various sensor types, such as cameras, radars, lidars, and ultrasonics. Simulation models of these, sensors as required in virtual validation methods, are available in various degrees of detail. However, proving the validity of such models is a subject of research. New metrics and methods for credibility assessment of simulation are needed to standardize the validation process in the future. The so-called double validation metric (DVM) has shown advantages and allows an intuitive interpretability of the validation results. The DVM has so far only been applied to lidar sensor models. In this paper, an extension to the DVM is introduced, which is called the DVM Map. A static measurement scenario is conducted in reality and transferred into simulation. The novel method is demonstrated on the obtained real and simulated radar sensor data. In this simple scenario special focus is put on the position accuracy of GNSS reference sensors. Therefore, their impact on the result of sensor model validation is discussed. The paper shows that the method provides a more detailed and accurate validation of a radar simulation, revealing previously undetected simulation errors. Errors due to the environment model, signal propagation, and signal processing are separated and satellite imagery is used for intuitive visualization of the results. This method is a complementary tool to existing validation techniques to improve the interpretability and judging the trustworthiness of radar simulations.

Keywords: Radar model validation, Double Area Validation Metric, DVM Map, Radar Cuboid Interface, Detections Interface

1

1 Introduction: What is so Special about Automotive Radar Modeling?

Simulation is playing an increasingly important role in proving the safety of automated vehicles. New procedures are envisaged in institutions such as UNECE [1, p. 5, 6], where simulation will be an integral part of the certification process. Automated vehicles (SAE level 3+ [2]) rely on robust environment perception with multiple sensor technologies. Radar ("radio detection and ranging") is one of these technologies, where a signal is actively transmitted to receive the echo instead of passively collecting it from other sources, like cameras. Automotive radar sensors commonly used in series production at different OEMs are based on the frequency modulated chirp-sequence principle, which is used to determine the range and the angular position of objects via antenna patches with respect to the sensor. What makes radars unique among current perception sensor technologies is their ability to also measure the radial velocity via the Doppler effect. This leads to an additional dimension of information and a radar "point cloud" therefore includes range r, azimuth angle ϕ, elevation angle θ, radar cross section (RCS) σ, and relative radial velocity v.

As shown in Fig. 1, a radar "point cloud", or more precise a radar "detection list" [4], appears more sparse and unstructured when compared to lidar. It becomes visible that modern digital beam forming radar sensors are not structuring the scans of the environment in a per layer order, as current time-of-flight lidar ("light detection and ranging") sensors do with their laser beams. The unstructured and more noisy appearance of radar detections is reasoned in the radar sensor's signal processing. Its front end, as shown in Fig. 2, entails the signal reception via the antennas and analog-to-digital conversion (ADC). After applying a fast Fourier transform (FFT) algorithm, the time-based signal is structured into the so-called radar cuboid. This term means a cubic multi-dimensional volume, often called radar cube for simplicity. It consists of multiple cells, so-called "bins", that can be divided into the dimensions range ι_r, relative radial velocity ι_v, azimuth ι_ϕ, and elevation ι_θ.

To enhance explicability, in the remainder, the content is limited to radars with range, azimuth, and radial velocity dimensions. However, the methodology can be applied to elevation without restriction.
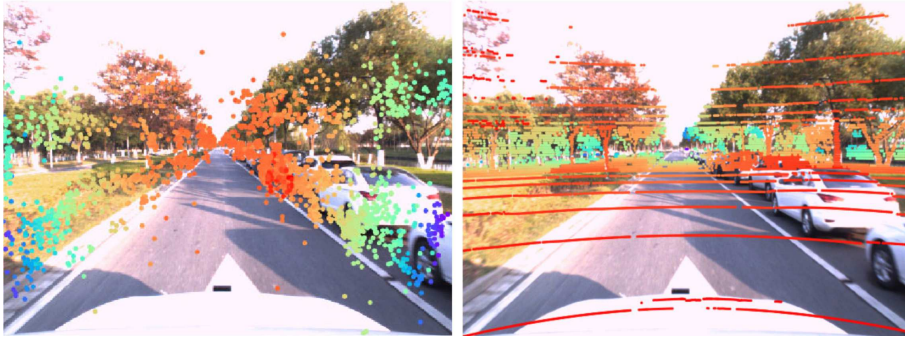

Fig. 1: Comparison of a radar (left) and lidar (right) "point clouds", projected into the captured scene with parking cars [3, p. 3]. The color represents the intensity-equivalent value per detection.
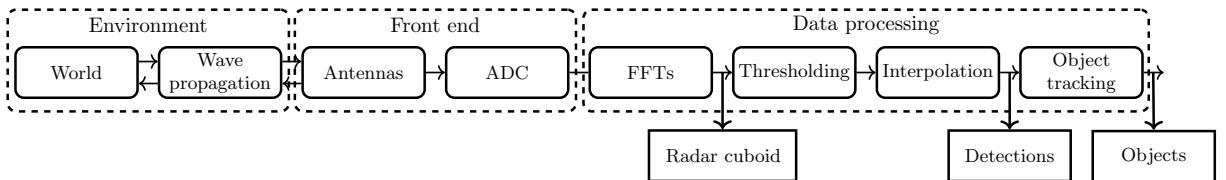

Fig. 2: An abstracted radar processing chain with elements visualized as blocks with rounded corners. The group within the processing chain is visualized as dashed rounded blocks and the sensor interfaces are marked as edged blocks.

The three dimensions of the radar cuboid are visualized in Fig. 3 with one single bin colored in orange and the teal color depicting the bins of the radar cuboid at a relative radial velocity of $0\,\frac{\text{m}}{\text{s}}$. Each bin contains a power ratio P in dB, which is calculated from the transmitted and the received signal. The number of bins $I_{r/v/\phi/\theta}$ of each dimension results from sensor design parameters like bandwidth, sampling rate and measurement time, as well as the configuration of the antennas. Except for digitization itself, only information due to windowing and noise is lost at radar cuboid level. Because of this condensation of all digital available information, it is reasonable to simulate the synthetic data with chirp sequence frequency modulated continuous wave radars with uniform array antennas at this very early interface, as shown in Fig. 3. For the required level of detail in simulation-based safety validation of automated driving, low-level interfaces must be taken into account to simulate specific perception tasks due to challenging environmental conditions or object constellations. Furthermore, simulated low level interfaces enable to enhance early signal processing.
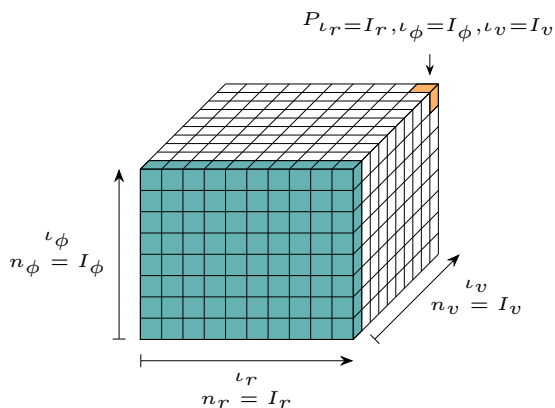
proven along with its delivery. Only in this case simulation models can also be used in a trustworthy manner for safety assessments, as already envisaged by the UNECE [1, p. 5, 6]. Therefore, the complexity of radar measurements is due to noise based on multi-path propagation and RCS sensitivities. Also the lack of public data sets and limitations in measurement repeatability is challenging [5]. Additionally, the number of detections depends on multiple causes and the noise of the sensor is complex due to the high frequency hardware components.

Fig. 3: Visualization of the radar cuboid with bins in range denoted as r, azimuth as ϕ and Doppler as v. The overall number of bins is I and one bin in the corresponding dimension is ι. In orange one bin with the power value P at the position I_r, I_ϕ, I_v is highlighted. The teal colored front marks the radar cuboid at the Doppler bin 0, which is of interest for static validation studies.

Nevertheless, it will not be sufficient to have only a simulation model of a sensor available, but the validity of the sensor model must be

2 Discussion of Metrics for Validation of Active Perception Sensor Simulation

In this chapter validation metrics from literature for radar model validation are listed and evaluated. Finally, an overview of the DVM and its application on lidar data is given.

2.1 Comprehensive Review of Already Used Validation Metrics

According to Oberkampf and Trucano, 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."* [6, p. 719] Viehof [7, p. 91] introduced the so-called sample-vise validity in the context of vehicle simulation validation. Therefore, radar sensor model validation is understood as the comparison of synthetic and real sensor data with decent metrics, sample-wise for a specific region of the desired parameter space of the application area. However, this means not only to be able to access the real sensor to simulate at the specific interface of interest, but also to design measurement campaigns according to the cause-effect chains that are modeled and investigated for a specific sample of the possible parameter space of the sensor model.

As described in the previous section, radar sensor modeling is a special case, as e.g. the relative velocity is directly measured. Validation of radar sensor models is a complex task due to the complexity of radar measurements, as outlined in the previous section. Besides experiment design to optimize repeatability and reproducibility of measurements, while minimizing epistemic and aleatory uncertainties in reference data, the metrics for comparing real and synthetic data play a crucial role in validation. Aleatory uncertainty is a statistical deviation based on probability distributions in data. Epistemic uncertainty means a lack of information regarding model structure, world knowledge and measurement errors. [8] While epistemic uncertainties can and should be reduced by enhanced reference data collection, aleatory uncertainties describe the inherent randomness of measurements.

A first decision guidance for metric selection is provided by the seven criteria for validation metrics as refined by Rosenberger [9, p. 99] and e.g. used by Magosi [10, p. 11], which were condensed by Schaermann [11, pp. 20-21], combining the original lists of six criteria by Oberkampf and Barone [12, pp. 11-12] and the seven features from Liu et al. [13, p. 2]:

1. Metrics meet the mathematical properties of a metric as defined by Fréchet [14]. (Unbounded results)
2. Metrics are intuitive. (Plausible & output in unit of measurand)
3. Metrics are applicable to both deterministic and non-deterministic data.
4. Metrics are quantitative and objective. (No manually tuned parameters)
5. Metrics do not include acceptance criteria. (No Boolean output)
6. Metrics consider uncertainties. (Epistemic and aleatory)
7. Metrics define a confidence interval with respect to the number of measurement data.

The state of the art in validation metrics for active perception sensor simulation is extensively discussed by Rosenberger [9, pp. 60ff.]. Multiple metrics in this collection of 34 options are only indirect metrics, where detections are sorted into occupancy grids first or object detection and tracking is applied. These cannot be applied on radar data at earlier interfaces like radar cuboid or detection level. Other metrics measure distances between points in space. These do not take into account intensity or power values. Therefore, they are not applicable to the radar cuboid, which is not a list of detections, but an equidistant distributed volume filled with power values.

Tab. 1 shows an excerpt of the remaining metric candidates for radar data. 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 table considers the data interfaces that the metrics are or could be applied to (**D**: Detections, **O**: Objects). Then, the ability to be applied to ($\bullet$/ $\rightsquigarrow$: (Quasi) static/dynamic) scenario is provided. Additionally, in Tab. 1 the scale of measurement it is able to process is considered (**M**: Metric (interval or ratio), **O**: Ordinal). The uncertainties it is able to process ($\int$ /$\rightmapsto$:

Aleatory/epistemic) are given per metric, as well. For the first four columns, it is marked if the metrics are applied without modification in literature (x), or if the metrics are applied in literature with moderate adaptions ($\star$). Additionally, the coverage of the seven criteria for validation metrics from the beginning of this section is marked in dark green.

Typical mathematical metrics that can be applied on object poses in space or detection coordinates like Manhattan distance (d_{Ma}) and overall error (OE) could be applied to the power values for a bin-wise comparison of a synthetic and a real radar cuboid. This also holds e.g. for mean error $\bar{d}$, root mean squared error (RMSE) and all other familiar error metrics from the collection that are not explicitly mentioned in this work, but listed in the original source collection. Still, none of these strictly mathematical metrics accounts for aleatory or epistemic uncertainties, which disqualifies them to be applied for radar sensor model validation on the detection or radar cuboid interface, due to the scattering characters. The machine learning-based Deep Evaluation Metric (DEM) as introduced by Ngo [15] is used to measure an overall simulation-to-reality gap, but it does not cover most of the seven criteria. Additionally, there is a lack of interpretability of the results, making it difficult to make safety arguments based on this metric. Also it does not provide insights on how to calibrate and enhance a sensor model and is therefore not considered in the following.

Consequently, only the Kullback–Leibler divergence D_{KL} as applied by Schaermann [11], the Jensen-Shannon distance d_{JS} as used by Magosi et al. [10], the area validation metric (AVM) d_{AVM} introduced by Ferson et al. [16], and the Frequency of positive Kolmogorov-Smirnov tests f_{KS}, as applied by Eder [17] remain as metric candidates. This means that besides the AVM as best candidate, two families of metrics should be evaluated further, namely divergences and hypothesis testing. Clearly, both are not intuitive concepts for most people and involve some more abstract thinking compared to just computing the area between two curves, as done for cumulative distribution functions (CDFs) or empirical cumulative distribution functions (EDFs) in case of the AVM. Rosenberger presents a detailed

analysis of both metric candidate families [9, pp. 105ff.], where the technique of manufactured universes [18] is used to compare EDFs from an artificial "simulation" against an EDF from an artificial "real" sensor.

After showing hypothesis testing results for the different compared EDFs and a short summary on the ongoing discussions on these kinds of tests in general, this metric family is dismissed for sensor model validation due to the sometimes misleading and above all counter-intuitive results. They are not available in the unit of the measurand, which makes it less user-friendly e.g. in model specification, especially for negotiations with people with non-technical background. The same difficulties in interpretation of the results from comparing the different EDFs are present when applying Kullback–Leibler divergence or Jensen-Shannon distance, leading to not further considering them for sensor model validation, too.

The remaining metric candidate is the AVM that is simply the integral of the absolute difference between two CDFs $F, \widetilde{F}$ over all real and simulated sensor measurements

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

Due to the fact that the cumulated probability $F(\zeta)$ for each measurand ζ is limited to $[0, 1]$ and unitless with m (e.g. 100) quantiles, the integral can be applied over the ordinate resulting in the mean error of all m quantiles of the CDF like

$$\begin{aligned} d_{\mathrm{AVM}}(F, \widetilde{F}) &= \int_0^1 |\zeta(F) - \widetilde{\zeta}(F)| \, \mathrm{d}F \\ &= \frac{1}{m} \sum_{i=1}^{m} |\zeta(F_i) - \widetilde{\zeta}(F_i)|. \end{aligned} \quad (2)$$

Therefore, the AVM is very similar to the mean error of all n measurements

$$\bar{d} = \frac{1}{n} \sum_{i=1}^{n} |\zeta_i - \widetilde{\zeta}_i|. \quad (3)$$

As visible in Tab. 1, the AVM is the only metric that handles aleatory and epistemic uncertainties. This ability is reasoned by the fact that it is not only applicable on EDFs describing aleatory uncertainties, but also on so-called probability

Table 1: Excerpt of the evaluation of metrics applied for active perception sensor simulation from Rosenberger [9, p. 99]. Green color: Metric is capable of a category, x/⋆: Metric applied in literature without/with adaptions.

Metric	Interf.		Scen.		Scale		Unc.		Covered criteria						
	D	**O**	**•**	**⤳**	**M**	**O**	**∫**	**⊢**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
d_{Ma}		x		x	x										
OE	⋆		x	x	x										
$\bar{d}$	x			x	x										
RMSE		x	x	x	x										
D_{KL}	⋆	x	x	x	x		x								
d_{JS}	x			x	x		x								
d_{AVM}	⋆	x		x	x		x								
f_{KS}	x			x	x		x								
DEM	x			x	x										

boxes (p-boxes). A p-box is expressed by the left and right boundaries of multiple EDFs. The width of the p-box at each quantile describes the epistemic uncertainties, as in Fig. 4 for the two simulation EDFs. First introduced by Williamson and Downs [19], a p-box gives the possible interval of cumulative probabilities for a specific measurand x and for a given cumulative probability it gives a possible interval of values, as discussed in detail e.g. by Ferson et al. [20].

As epistemic and aleatory uncertainties should always be minimized during measurements, but can never be eliminated, they must be propagated through the simulation to reflect these uncertainties when the model is validated. Practically, this means that e.g. every position of a sensed object must be captured with reference sensors during the measurements to collect the real sensor data to validate the model. The uncertainty of this reference position measurement device, e.g. +/-1.0 cm, is then input for multiple simulations per measurement, e.g. one with the exact reference position and two more for the edge cases of +/-1.0 cm. These multiple simulations result in several EDFs and a combination of all EDFs from simulation forms the p-box. Its boundaries are composed of the maximum and minimum x-values of the set of EDFs for each y-value.

The AVM for p-boxes is simply calculated by adding the two portions where the simulated p-box $\widetilde{\mathcal{F}}$ is higher (d^+) or lower (d^-) than the real p-box $\mathcal{F}$ as

$$d_{\mathrm{AVM}}(\mathcal{F}, \widetilde{\mathcal{F}}) = d^- + d^+ . \qquad (4)$$

For simplification, in Fig. 4 the EDF F is an infinitely thin p-box. Consequently, the AVM only considers the left and the right borders of the p-box, the original course of each EDF inside is irrelevant and the borders could actually originate from different EDFs.
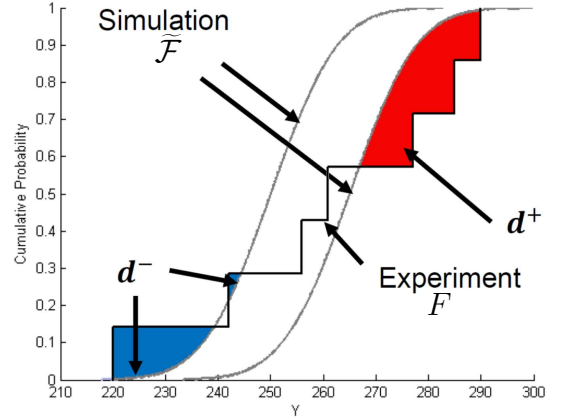

Fig. 4: Portions of the AVM, where the simulated p-box $\widetilde{\mathcal{F}}$ is higher (d^+) or lower (d^-) than the real EDF F, based on Voyles and Roy [21].

2.2 The DVM and its Application in Validation of Lidar Sensor Simulation

An additional requirement on validation metrics is the ability to distinguish model bias and model scattering error to enable the structured elimination of these two different modeling errors [9, p. 72]. Model bias is an approximation of the mean deviation and model scattering error the deviation in the distribution function's shape. Fig. 5 illustrates the difference between model and measurement bias and also shows the difference between the measurement standard deviation and the model scattering error. Indeed, measurement bias and scattering error are conceptually similar to the differences in mean and variance between a set of normal distributions However, the distribution functions of the measurand can deviate from normal distributions.
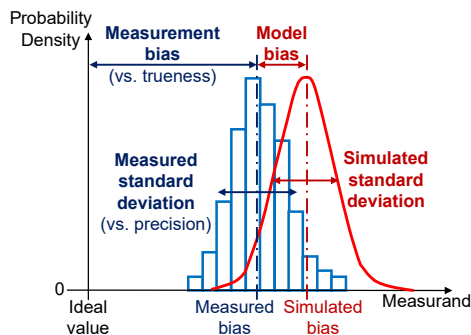

Fig. 5: Bias and scattering of measurement and model [9, p. 11].

Rosenberger therefore introduced the DVM that distinguishes the two components [9, p. 118ff.]. The first part is essentially the difference of d^+ and d^- in comparison to the original AVM as sum of these two portions. Voyles and Roy [21] introduced this difference that is proven to be a good estimate for the model bias as

$$d_{\text{bias}}(\boldsymbol{\mathcal{F}}, \widetilde{\boldsymbol{\mathcal{F}}}) = d^- - d^+ . \quad (5)$$

It eliminates symmetrically distributed area portions of the AVM, which reflect the model scattering error and therefore only keeps the model bias. Consequently, (5) can be used to estimate a

"corrected" [21] p-box as

$$\widetilde{\boldsymbol{\mathcal{F}}}_{\text{c}}(\zeta) = \widetilde{\boldsymbol{\mathcal{F}}}(\zeta - d_{\text{bias}}) = \widetilde{\boldsymbol{\mathcal{F}}}\Big(\zeta - (d^- - d^+)\Big). \quad (6)$$

Taking this idea of a corrected, bias-free simulated p-box further, a second-order AVM can be computed with $\widetilde{\boldsymbol{\mathcal{F}}}_{\text{c}}$ that now only entails the remaining model scattering error. This novel metric introduced by Rosenberger [9, p. 118] is called corrected AVM (CAVM) and formulated as

$$d_{\text{CAVM}}(\boldsymbol{\mathcal{F}}, \widetilde{\boldsymbol{\mathcal{F}}}) = d_{\text{AVM}}(\boldsymbol{\mathcal{F}}, \widetilde{\boldsymbol{\mathcal{F}}}_{\text{c}}) = d_{\text{c}}^- + d_{\text{c}}^+. \quad (7)$$

As illustrated in Fig. 6, the CAVM is a multi-step process that inherently includes the calculation of the model bias on its way. It starts with the calculation of d^+ and d^- for d_{bias} (5). Then the simulated p-box is corrected by d_{bias} to get $\widetilde{\boldsymbol{\mathcal{F}}}_{\text{c}}$ (6). Finally, d_{c}^+ and d_{c}^- are calculated, resulting in d_{CAVM} (7).

Consequently, the novel DVM for comparison of two p-boxes $\boldsymbol{\mathcal{F}}$, or EDF F as infinitely thin p-boxes, is achieved that distinguishes model bias and model scattering error with respect to the actual sensor bias and its real scattering behavior, as

$$d_{\text{DVM}}(\boldsymbol{\mathcal{F}}, \widetilde{\boldsymbol{\mathcal{F}}}) = \Big(d_{\text{bias}}(\boldsymbol{\mathcal{F}}, \widetilde{\boldsymbol{\mathcal{F}}}), \ d_{\text{CAVM}}(\boldsymbol{\mathcal{F}}, \widetilde{\boldsymbol{\mathcal{F}}}) \Big). \quad (8)$$

Aligned with distinguishing model bias and model scattering error, model validation should start simple to reach more complex scenarios later. Accordingly, Rosenberger starts with demonstrating the DVM for beam-wise model evaluation in static scenarios, like targets in different distances with no other effects taking place, to more complex and object-wise validation of synthetic lidar detections. Aside from the interpretability of the results in the unit of the measurand, the accuracies are considered by the reference tests using the p-boxes. Consequently, for radar model validation, the validation should follow this incremental approach. The experimenter should take special care to consider isolated cause-effect chains, which influence the radar signal propagation. A possible ontology to derive them is PerCollECT [22] as available on Github [23].
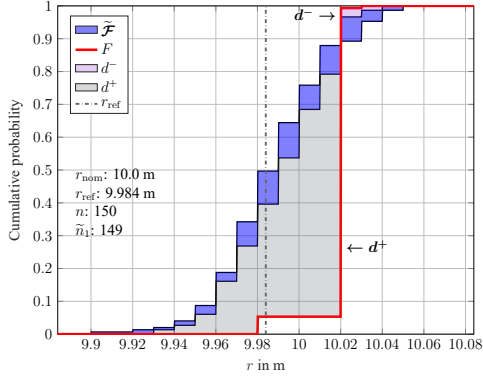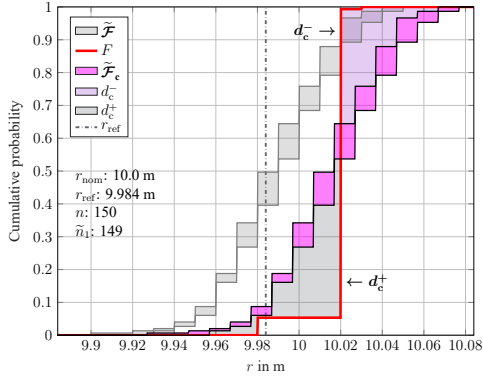
(a) Calculation of d^+ and d^- for d_{bias}.

(b) CAVM calculation with $\widetilde{\mathcal{F}}_{\mathbf{c}}$.

Fig. 6: Illustration of the CAVM [9, p. 118] in a lidar sensor measurement campaign. A plate is placed in front of the sensor at $10\,\text{m}$ in a static scenario. r_{nom} and r_{ref} are the nominal and measured reference range. n and $\widetilde{n}_1$ are the number of detections from real data and simulation sim1, which is a control factor for the comparison of the different EDFs [9, p. 103]. These F is the EDF from real data. $\widetilde{\mathcal{F}}$ is the p-box from simulation. $\widetilde{\mathcal{F}}_{\mathbf{c}}$ is the simulated p-box corrected with the estimated model bias d_{bias}. d^+ and d^- mark areas where the simulated p-box $\widetilde{\mathcal{F}}$ is higher (+) or lower (-) than the real EDF F. d_{c}^+ and d_{c}^- mark areas where the corrected simulated p-box $\widetilde{\mathcal{F}}_{\mathbf{c}}$ is higher (+) or lower (-) than the real EDF F.

Besides the DVM there are other applications of the AVM described in the literature. Brune et al. compare in their application the left and right edges of the measurement and simulation p-boxes with each other [24]. Compared to the DVM, this approach offers the advantage that the p-box size is included in the metric result. This aspect is missing in the DVM according to Rosenberger. However, Brune's AVM does not explicitly

consider the scattering error of the distribution function, which means that the information about the shape similarity of the p-box is lost. Fig. 7 shows the AVM according to Brune et al. and illustrates how the size of the p-boxes is incorporated into the result of the AVM.
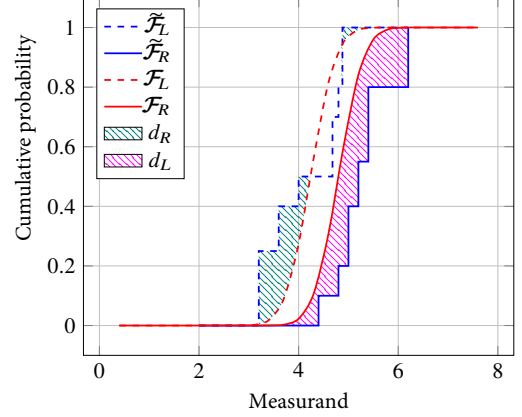

Fig. 7: Visualization of the AVM calculation d_L and d_R based on the left and right border comparison between measurement $\mathcal{F}_L$, $\mathcal{F}_R$ and simulation $\widetilde{\mathcal{F}}_L$, $\widetilde{\mathcal{F}}_R$, based on Brune et al. [24]

2.3 Double validation metric limitations

As mentioned in Sec. 1, the measurement of the radial relative component of the velocity is possible with radar sensors. This allows the discussion of the DVM to model validation in dynamic scenarios. As already shown by Holder, even simple measurement scenarios are subject to difficulties regarding repeatability of complex geometries that must be taken into account [25]. Fig. 8 shows the variation of the RCS denoted as $Q_{(\sigma)}$ over the distance in the radar sensor coordinate system denoted as $_S r$ of a retroreflector, which is a corner cube reflector (CCR), and a vehicle.

Therefore, in order to apply the metric, it is advisable to introduce p-boxes for the measurement data to consider the limited measurement repeatability and take them into account in the metric result. Based on Fig. 8 and the limited reproducibility the size of the measurement p-box and the distribution of EDFs must be part of the metric.

Additionally, the size of the simulation p-box is a factor influencing the quality of the DVM. These properties are missing in the DVM according to Rosenberger because these characteristics are lost during the transformation of EDFs to p-boxes. Theoretically constructed edge cases (ECs) substantiate the previous remarks. These ECs can appear in the application of the validation methodology to radar due to the sensitivity of the model to small changes in reference sensor measurement uncertainties, but also due to the problem of reproducibility of measurements.

Fig. 9 shows the first ECs with the results of the Rosenberger's DVM and the evaluation according to Brune et al. [24].

It addresses, on the one hand, the size of the p-box and the overlap of the simulation and measurement p-box. To proof the independence of the methodology regarding the number of simulation and measurement EDFs, the combination of EDFs is varied.

Fig. 9 illustrates EC 1 consisting of a large simulation p-box in comparison to the measurement. Additionally, the right side of the simulation p-box equals the left side of the measuremnt p-box. The AVM and the CAVM are both 0 and therefore the model is valid based on the DVM. The extension by Brune of the AVM covers this EC by resulting in a d_{L}, which has the same value as the simulation's p-box size.

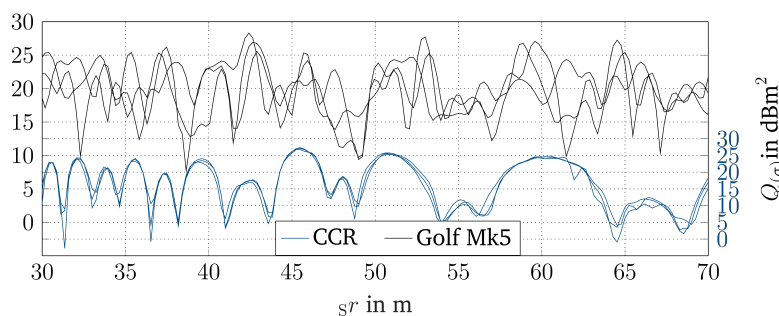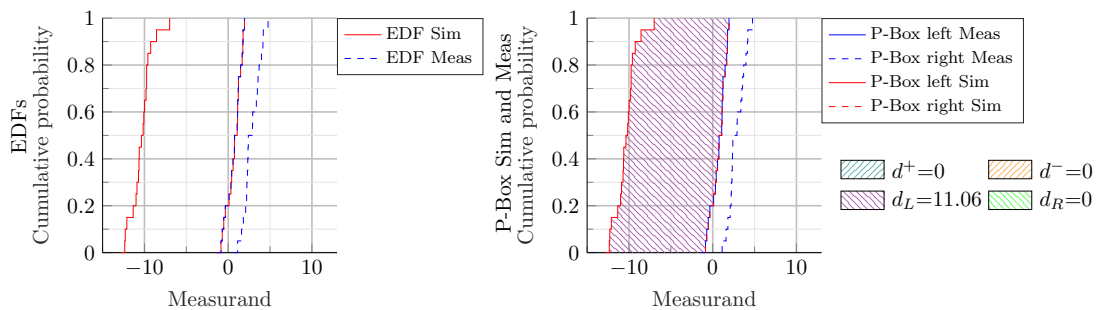**Fig. 8**: RCS in different experiment trials of a CCR in blue and a Golf Mk5 in black [25].

Fig. 9: EC 1 of the DVM definition by Rosenberger and AVM calculation by Brune. On the left side the EDFs are visualized. On the right side the p-boxes of measurement and simulation are shown with the corresponding colored area.
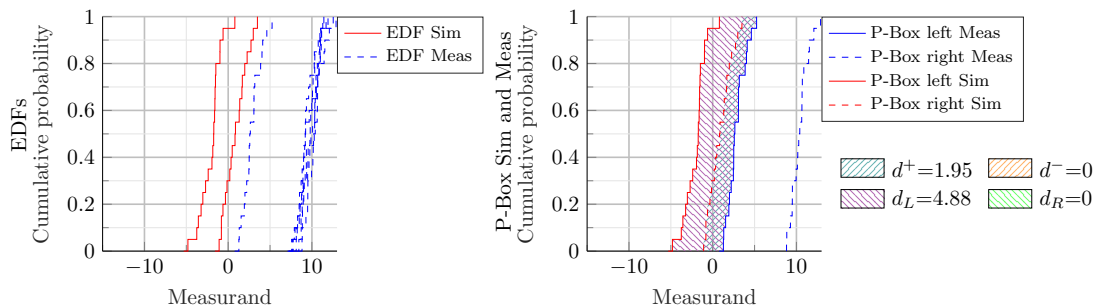

Fig. 10: EC 2 of the DVM definition by Rosenberger and AVM calculation by Brune.

Fig. 10 shows EC 2 with a large measurement p-box in comparison to the simulation. A concentration of measurements EDFs on the right side is also present. For Rosenberger's DVM the simulation results are only the deviation between the right simulation and left measurement EDF. The AVM is small and the CAVM is almost 0, because the shape of the right simulation and left measurement EDF are nearly identical. Also the AVM definition by Brune fails in this case. The value for d_L covers just the offset between the two p-boxes.

Fig. 11 visualizes EC 3 consisting of two simulation EDFs at the left and the right side, which are very similar in shape. Between the outer distributions there are several simulation EDFs that show a completely different distribution shape. By forming the simulation p-box, the information of the inner distributions is lost. This leads to the fact that the deviations in the shape of the distribution function are insufficiently considered in the DVM validity consideration. Brune's AVM covers the big size of the simulation p-box but fails with identifying the inner distribution functions.

Fig. 12 illustrates EC 4, which is similar to number 2, where the distribution of the simulation distribution functions has no effect on the validation result itself. In the case of Brune's AVM, only the total deviation of the p-boxes is quantified.

However, the important information of the accumulation of distribution functions at the right side of the simulation p-box is lost, which could be helpful for the modeler and the experimenter.

The presented ECs show the lack of the two validation metrics. Especially the distribution of the EDFs within the p-box borders are disregarded. Another disadvantage is that an assignment of the EDF outliers, which results from parameters varied in the simulation, to the validation result is excluded. Generally, the introduction of p-boxes obscures the validation result and leads to potential misinterpretation. Therefore, the methodology to apply the DVM has to be extended and modified.

In order to ensure comparability of simulated and real data based on DVM, it is essential to include the number of data points in the evaluation process. A deviation in the numbers between simulation and measurement of 10% is considered as an acceptable limit in the remainder of this paper by the authors. At the detection level, this is particularly problematic in the case of the radar sensor, as the number of detections can vary greatly between individual measurement cycles, for example due to clutter from vegetation or rain.
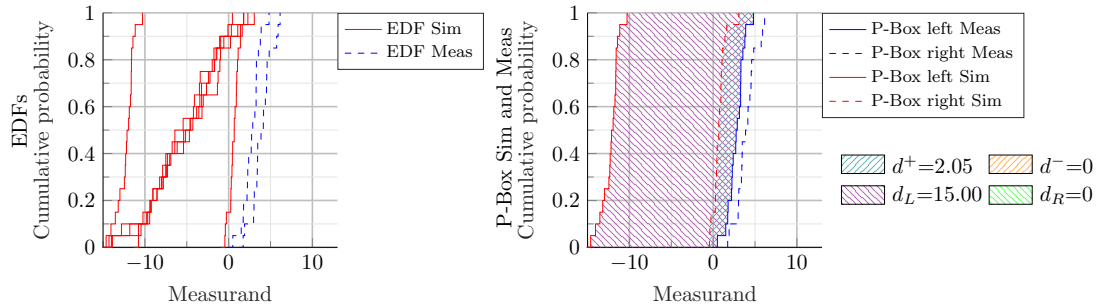

Fig. 11: EC 3 of the DVM definition by Rosenberger and AVM calculation by Brune.
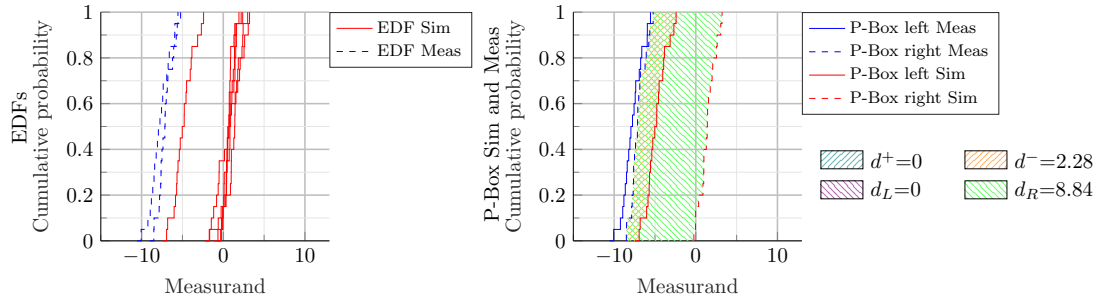

Fig. 12: EC 4 of the DVM definition by Rosenberger and AVM calculation by Brune.

3 Radar validation methodology

In this chapter, based on the theoretical ECs, the methodology for determining the DVM is adapted. Furthermore, it is shown how the new methodology can be applied to radar data.

3.1 The DVM Map

To overcome the DVM limitations of simulation and measurement p-box size as well as distribution and shape of EDFs an adaption of the DVM is necessary. Fig. 13 shows the adapted validation methodology based on the AVM and CAVM metric. Compared to Rosenberger, the measurement and simulation p-boxes are resolved and each simulation is compared to each measurement, deriving the new "DVM Map" as a validity tool visualized as a heat map.
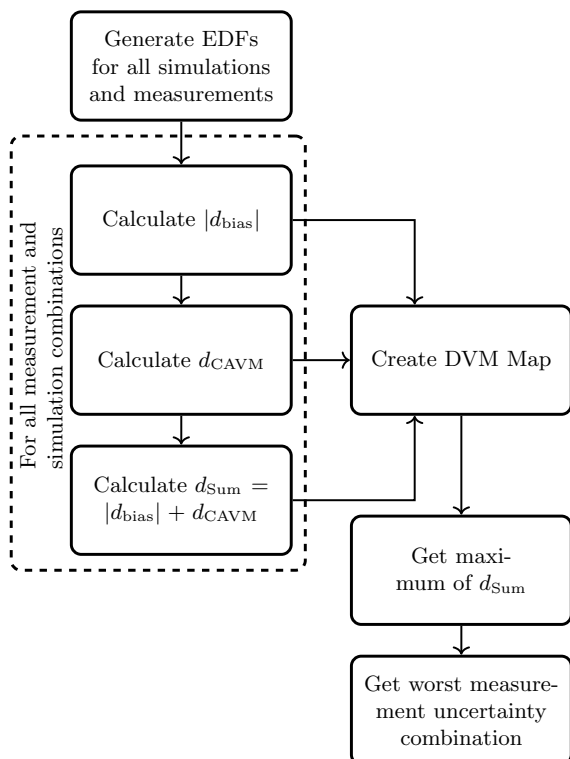

Fig. 13: This figure shows the adapted DVM methodology to address the different ECs with the new DVM Map as intermediate step.

In a first step, the AVM is formed for each simulation EDF in combination with each measurement and corrected by the determined d_{bias} according to (5). The absolute value of $|d_{\text{bias}}|$ is used for visualization in the DVM map so that for negative and positive model bias, the color value of the scale is unambiguous. Following the aforementioned procedure, the corresponding CAVM is formed and d_{CAVM} determined according to (7). This results in a value for the model bias and the scattering error for each simulation in comparison to each measurement. The scattering error finds an intuitive explanation in the shape deviation of the corrected simulation EDF in comparison to the measurement EDF. To calculate the overall comparison score d_{Sum} the absolute value of the model bias is added with the CAVM result as

$$d_{\text{Sum}} = |d_{\text{bias}}| + d_{\text{CAVM}}. \tag{9}$$

These deviations of the simulation can be quantified in comparison to all measurements. Therefore, the DVM Map shows the most critical measurement and simulation of the corresponding reference data uncertainty. When the results of the DVM Map need to be further processed, the maximum of d_{Sum} is formed, thus identifying the most critical combination of measured and simulation data for sample validation. To demonstrate the utility of the newly developed DVM Map, the ECs from Fig. 9 to Fig. 12 are re-examined.

Fig. 14 is sorted by the ECs from left to right and shows the results for $|d_{\text{bias}}|$, d_{CAVM} and d_{Sum} from top to bottom. The DVM Map of EC 1 shows that the large model bias of simulation 2 can be compensated. The CAVM shows minor deviations since the shape of all EDFs is similar to each other. It can be seen that the combination of measurement 2 and simulation 1 performs worst in terms of sample validity.

For EC 2, the focus is also on model bias examination. The accumulations of EDFs on the right-hand side are clearly evident, with simulation 1 deviating more than simulation 2. The result also shows up in d_{sum}, resolving the EC.

EC 3 shows that the distribution of the different simulation EDFs can be mapped using the d_{CAVM}. However, the influence of d_{bias} predominates for simulation 1.

EC 4 shows that the DVM Map can also reproduce high deviations of the simulation and
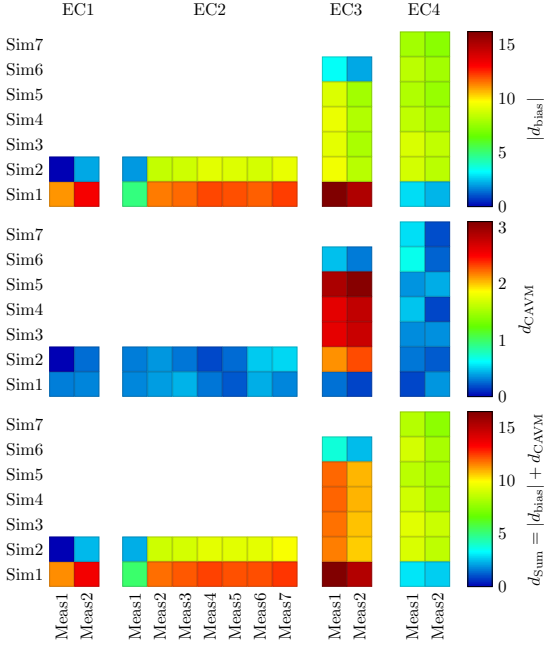
Fig. 14: DVM Map of the ECs defined in Fig. 9 to Fig. 12 processed with the newly introduced methodology. From left to right, the various ECs are shown. From top to bottom the DVM Map of $|d_{\mathrm{bias}}|$, d_{CAVM} and d_{Sum} is illustrated. The coloring corresponds to the quantitative deviation of the corresponding value, where blue means low and red means high. The upper and lower limit of the scale is determined by the minimum and maximum value in the corresponding DVM Map.

thus reflects valuable information regarding sample validity and the influence of various aleatory uncertainties.

3.2 DVM Map Application on Radar Sensor Interfaces

Fig. 15 shows the methodology application to radar cuboid and detections in a study of sample validity of a radar model. The first step is to run defined scenarios in the real world (bright red). In addition to the measurement data at detection and radar cuboid level, the measurement campaign yields the operational reference data, which is subject to epistemic and aleatory uncertainties. Operational reference data means is this context the ability to take a reference measurement with additional sensors independent from the radar sensor. The uncertainties are determined by means of reference sensors or reference sources.

The measured reference data is transferred to the simulation (light blue) in a further step. Here, in addition to the measured reference value, the epistemic uncertainties are propagated through the simulation. As a result, simulation data on detection and radar cuboid level are available, where the number of simulations depends on the number of simulated uncertainties.

Different variants of the new validation methodology can then be applied to the measurement and simulation data. First, a very rough consideration of all detections and all cells of the radar cuboid is advisable. From this, basic deviation of the sensor model from the measured data can be derived, as well as which uncertainty combination together with which measurement shows the largest measurand deviation. This allows conclusions to be drawn about gross modeling errors, as well as measurement outliers, provided the number of measurements is large enough to identify outliers.

Detections and the corresponding bins in the radar cuboid are of great interest of a validation study. This can be justified firstly by the fact that detections represent the input for all subsequent steps of radar processing, and secondly that here, either due to the environment or due to objects, power differences are present that a simulation model should represent. Therefore, the time aggregated detection data from all measurements are clustered using a Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm and thus a region of interest is defined, which in turn can be transformed into cells of the radar cuboid. By applying the new methodology, individual areas of particular interest in the radar measurement are examinable in a dedicated manner.

Finally, an application of the radar validation methodology to each individual cell of the radar cuboid is performed. Local effects and influences are the focus of the investigation and provide valuable information about the model sensitivity with respect to the reference sensor uncertainties. By matching the results to a satellite image detailed investigations of the influence of the environment are possible.
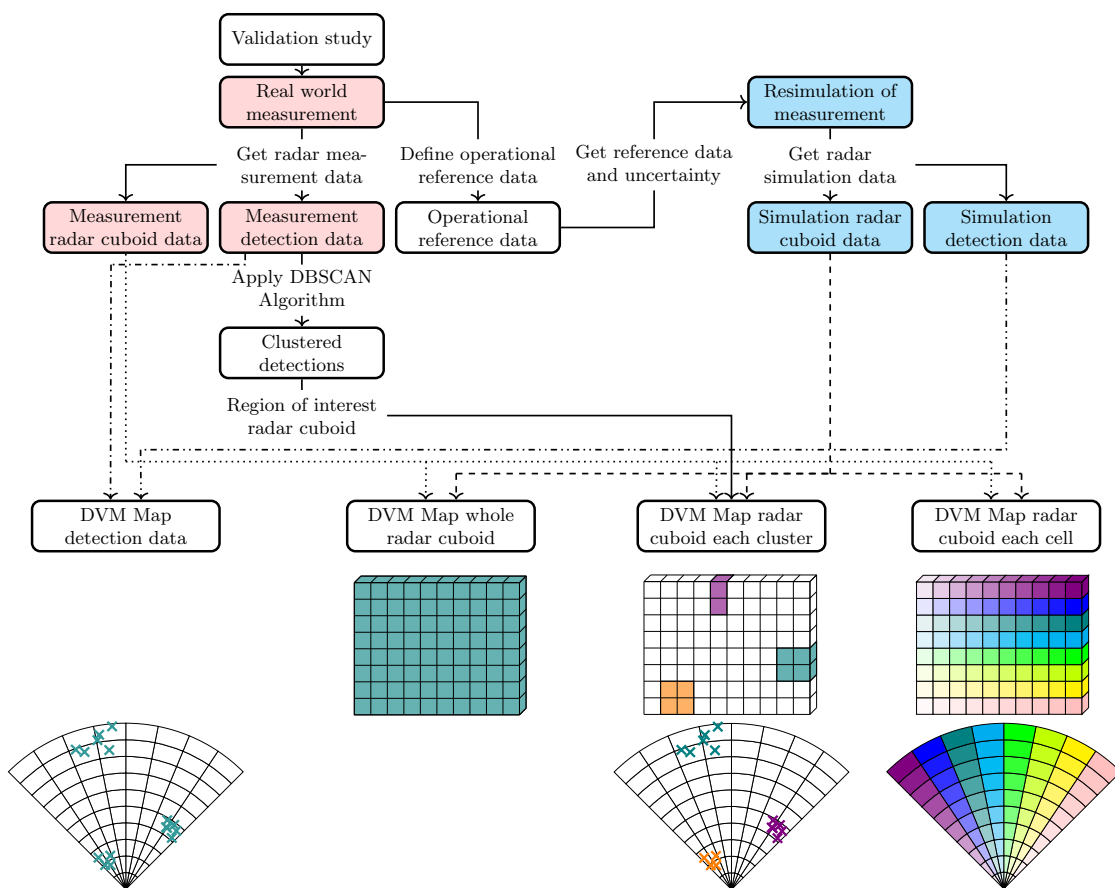
Fig. 15: Validation methodology to apply the DVM Map to different levels in the radar processing chain. From left to right the application of the DVM map to all detections, to the entire radar cuboid, the region of interests, which result from time aggregated, clustered detections, associated with the corresponding radar cuboid cells and each radar cuboid cell separately is visualized.

4 Application validation methodology

In this chapter the aspects considered theoretically so far are verified in the following by means of a validation study. For this purpose, a measurement campaign is carried out on the August-Euler airfield proving ground in Griesheim. The study is intended to serve as a proof of concept of the DVM Map, therefore a static scenario is measured.

4.1 Experimental setup

The object in the validation study is a CCR lying flat on the asphalt, which is placed on different positions in front of the radar sensor. The positions differ in the range r to the radar sensor and the azimuth angle ϕ. For r, 29.56 m and 48.33 m are chosen, respectively, due to the radial resolution, so that the CCR is located once close to the edge and once in the center of the range bins. For ϕ, positive as well as negative angles are defined so that the CCR is within the sensor's unambiguous azimuth measurement range and the width of the test site is sufficient. Thus, angles of -8 °, -4 °, 0 °, 4 °, 8 ° are obtained. Fig. 16 shows a sketch of the measurement setup with the different CCR positions, which are measured one after another, on the left side and the real world measurement setup for position 1 on the right side. The position of the CCR and the sensor is obtained using a real time kinematic (RTK)-based global navigation satellite system (GNSS) antenna. For each measuring position, 5 measurements of 60 seconds each are recorded with the radar sensor resulting in approx. 850 samples.

The measurement setup and the existing measurement technology result in the uncertain reference data listed in Tab. 2. Here, some parameters are determined directly and others result from the propagation of error of several measurement uncertainties.

4.2 Simulation model

The reference data uncertainties listed in the table are transferred to the simulation in a next step. Reference data uncertainty propagation through simulation is realized by defining separate scenarios with the upper $^+$ and lower limit $^-$ of this uncertainty. Additionally, one simulation with all uncertainty-free reference measurement data denoted as N is integrated. For the radar simulation, the output of a black box radar ray tracing algorithm by IPG CarMaker version 9.1.1 and an adapted open source radar signal processing model [26] is used.

The input from the ray tracing algorithm is interpreted in the radar model as a delta peak in frequency space and is called "Fourier tracing" [25]. The range and angle information of the ray is used to calculate the radar cuboid bin in the different dimensions in which the delta peak is located. A windowing function is used to smear the power of the delta peak into the neighbouring bins. This allows effects such as ambiguities, separation capabilities and interferences to be present in the radar model. In addition, a non-deterministic noise simulation is implemented for this model approach. Based on measurements, the mean and standard deviation are determined for each range-azimuth cell combination. A Gaussian distribution with the determined parameters is then imposed on cells whose minimum power is below the noise floor. Fig. 17 shows the determined mean and standard deviation as a range azimuth map. This result is conducted by placing the sensor on the asphalt with its front side pointing into the sky.

The radar model is parameterized using the data sheet and calibrated with position 1 of the CCR on the detection interface. Additionally, a simulation model of the August-Euler airfield in Griesheim is used for the environment simulation.
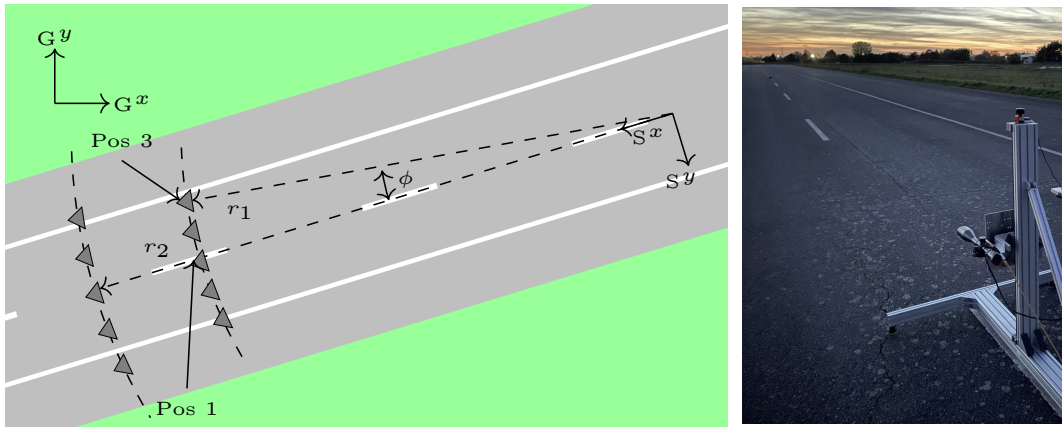
Fig. 16: Experimental setup of the validation study. CCR position 1 ($r_{\text{CCR,Pos1}} = 29.56\,\text{m}$, $\phi_{\text{CCR,Pos1}} = 0$) is used for simulation calibration purposes and position 3 ($r_{\text{CCR,Pos3}} = 29.56\,\text{m}$, $\phi_{\text{CCR,Pos3}} = \text{-8}\,°$) is analyzed based on the presented methodology,

Table 2: Measured reference data uncertainties for CCR position 3 defined in Fig. 16. The local cartesian coordinate system G is defined in East-North-Up direction with the origin located on the August-Euler airfield in Griesheim.

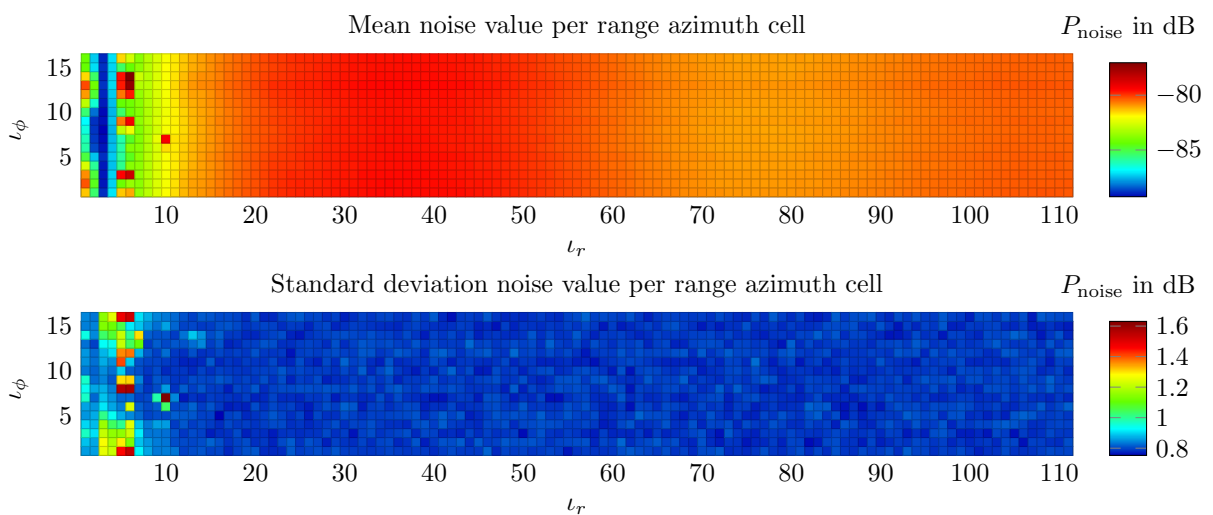| Reference data	Variable	Measurement	Uncertainty
Sensor azimuth orientation in G	$_{\text{G}}\phi_{\text{S}}$	$197.91\,°$	$\pm\,0.07\,°$
Sensor x position in G	$_{\text{G}}x_{\text{S}}$	$977.43\,\text{m}$	$\pm\,0.02\,\text{m}$
Sensor y position in G	$_{\text{G}}y_{\text{S}}$	$241.56\,\text{m}$	$\pm\,0.02\,\text{m}$
Sensor height in G	$_{\text{G}}h_{\text{S}}$	$12.89\,\text{m}$	$\pm\,0.02\,\text{m}$
CCR x position in G	$_{\text{G}}x_{\text{C}}$	$948.33\,\text{m}$	$\pm\,0.02\,\text{m}$
CCR y position in G	$_{\text{G}}y_{\text{C}}$	$216.46\,\text{m}$	$\pm\,0.02\,\text{m}$
Edge length CCR	l_{C}	$0.240\,\text{m}$	$\pm\,0.005\,\text{m}$

Fig. 17: Mean value and standard deviation illustrated as range azimuth map for a Gaussian distribution noise model.

5 Results and Discussion

In this chapter, the different results of the DVM Map are shown and discussed. Therefore, the methodology defined in Sec. 3 is applied to the static validation scenario. The DVM Map with the new validation methodology is applied to the different interfaces as described in Fig. 13 and Fig. 15. In Sec. 5.1 all detections are analyzed. Afterwards in Sec. 5.2 the results of the whole radar cuboid is discussed. The region of interest in the radar cuboid by means of clustered detections is described in Sec. 5.3. Finally, the results of each cell of the radar cuboid combined with a visualization on top of a satellite image are shown. As an example, the previously mentioned scenario of CCR position 3 at $r_{\text{CCR,Pos3}} = 29.56\,\text{m}$ and $\phi_{\text{CCR,Pos3}} = \text{-}8\,°$ is evaluated.

5.1 Whole Detections

This section contains the evaluation for all detections of the measurements and simulations. Fig. 18 shows all detections of all measurements combined in one diagram in Cartesian coordinates, where the origin is the sensor position. Therefore, the EDFs of the different distributions of the quantities r, ϕ and σ are calculated.
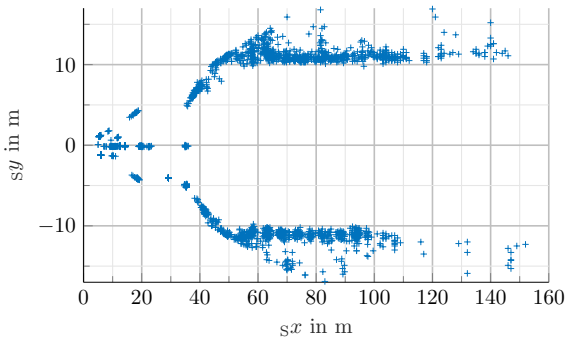

Fig. 18: Plot of all detections in the Cartesian sensor coordinate system. Especially in near range of the sensor as well as at the transition of the asphalt and vegetation clutter is present. The CCR is located at $_S x = 29.4\,\text{m}$ and $_S y = \text{-}4.2\,\text{m}$

The representation of the detections as distribution functions is visible the top row of Fig. 19. The EDFs of the different measurements are close to each other and the deviation of numbers of detections is less than $1\,\%$. This indicates a high

reproducibility of the measurement results. However, there is a significant difference in the number of detections between simulation and measurement, which prevents a valid evaluation based on the DVM.

Nevertheless, further analysis of the data will be conducted to describe and analyze the general methodology. The distributions of the detections from the simulation are almost at the same position in every cycle for the range as well as the azimuth angle. However, deviations of the simulations in RCS show up. The analysis of the RCS shows that the model reacts very sensitively to small changes of the scenario, which again emphasizes the necessity to consider the reference sensor uncertainties.

The basically different distribution functions between measured and simulated data result from the environment model of the August-Euler airfield as well as the ray tracing simulation. The ray tracer used in the CarMaker version has no reflections from the road surface in the given setup as well as no reflections from the vegetation next to the asphalt surface due to the lack of simulated vegetation. Therefore, only rays from the CCR are processed as radar signal processing model input. Thus, only detections at the object result, whereby besides the small number of detections also the distribution in range and azimuth are limited to the position of the CCR. However, the detections of the CCR are in the same distance in measurements and simulations. This can be seen particularly well in the second row at $r = 29.56\,\text{m}$ as well as $\phi = \text{-}8\,°$ by the step in the measurement EDFs highlighted by the red ellipse.

The second row of the Fig. 19 shows the simulation EDFs $\widetilde{F}_{\text{c}}$ corrected by the model bias d_{bias} based on the AVM calculation to measurement 1. This represents the second step of the methodology from Sec. 3. It is already evident, without a quantitative determination of the deviation by means of d_{CAVM}, the fundamental difference of the distribution functions.
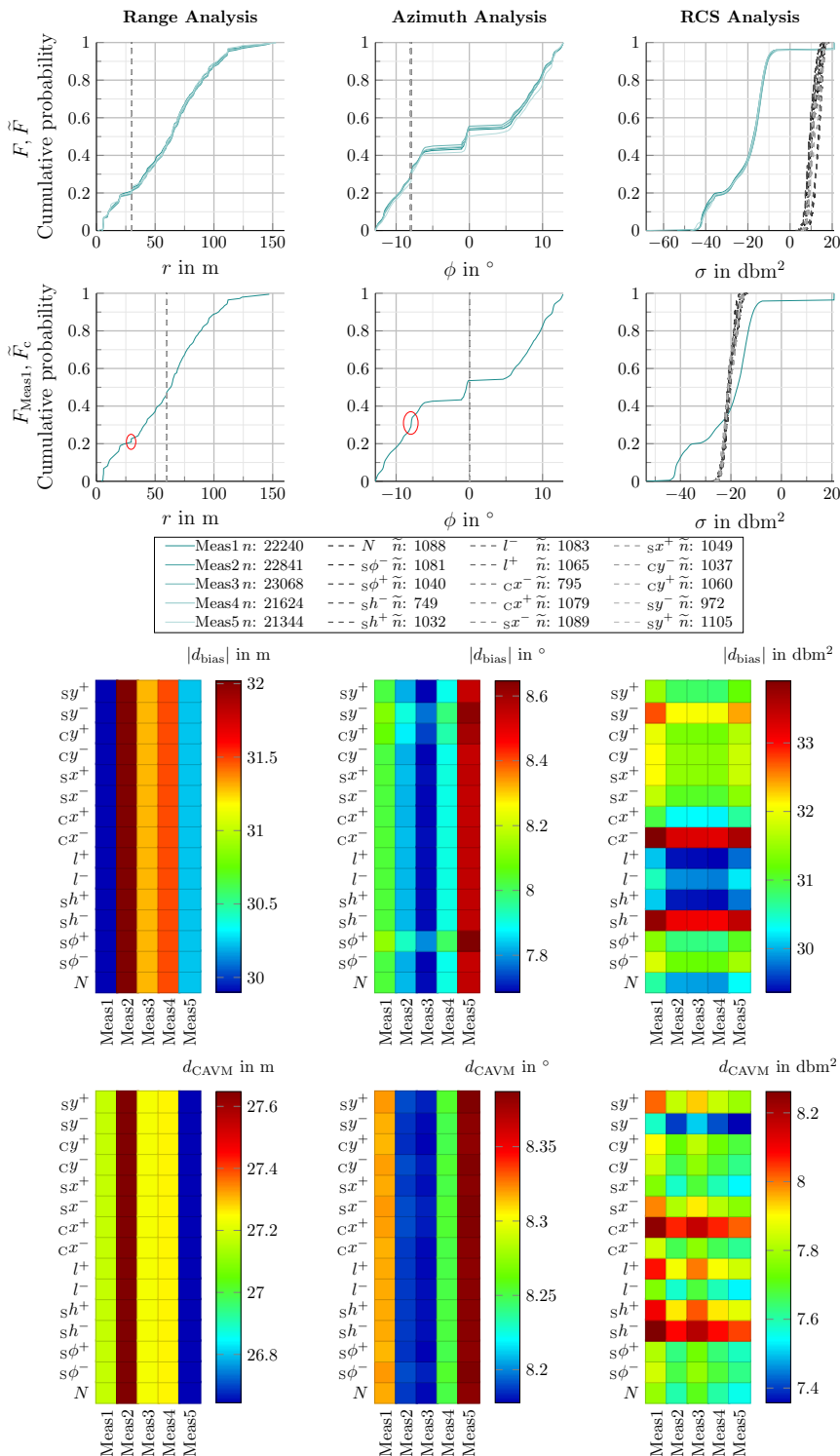
Fig. 19: From left to right are the range, azimuth, and RCS results of all detections. The first row visualizes the EDF of all measurements F and of all simulations $\widetilde{F}$. In the second row the EDF of measurement 1 F_{Meas1} and the corrected simulation EDFs $\widetilde{F}_{\text{c}}$ are shown. The red circles illustrate the data points in the measurement EDFs where the CCR is located. The last two rows show the DVM Map of the above mentioned quantities. The number of detections of the measurement n and simulation $\widetilde{n}$ is in the legend.

The DVM Map in Fig. 19 verifies the findings already made from the first visual impression. In the range domain, the constancy of the simulation of d_{bias} and d_{CAVM} stands out. In the range, the simulation model is insensitive to the simulated uncertainties. This is due to the fact, that the measurement uncertainties are very small compared to the radar's range resolution of 1.8 m. Measurement 1 and 5 show the smallest deviations in the model bias and measurement 2 the largest deviation. The scattering error of measurement 2 is largest and lowest for measurement 5 for the range dimension. Thus, the differences in the results of the DVM Map are due only to the differences in the measurements.

In the case of the azimuth dimension, the model is slightly more sensitive, as shown by the minor color changes in the perpendicular components of the DVM Map. These variations are small compared to the measurement influence shown in the horizontal of the various uncertainties. Here, the negative variations of the y-position and the rotation ϕ of the sensor are the uncertainties with the most notable influence on the result.

For the RCS, in turn, the negative x-position of the CCR variation and the negative sensor height variation have the largest influence. All parameters have a clear but different influence on the RCS distribution of detections. In addition, it is shown that the DVM Map is able to represent the different positions of the EDFs with respect to each other in a very intuitive and simple way.

The influence of the environment in combination with the ray tracing algorithm as stated above is simply too large to make a validity statement by means of the analysis of all detections. However, the analysis of all radar detections can be used to compare stochastic effects from the environment. This includes not only influences like vegetation but also weather influences like rain and snowfall.

5.2 Whole Radar Cuboid

In this section the DVM Map is applied to the whole radar cuboid. The only dimension considered in this case is the power distribution P of all cells of the radar cuboid, with the velocity dimension reduced to one cell $\iota_v = 0$. This reduces the radar cuboid to a range-azimuth map where each cell holds a spectral power value. Fig. 20 shows the EDFs of simulation and measurement as well as the corresponding DVM Map.

The course of the simulation EDFs is characterized by the noise simulation at the beginning. Only above -82 dB the effect of the radar cuboid cells filled by the CCR with a higher power is visible. In relation to the number, however, these are represented much less frequently, so that the gray curves just above -80 dB are just close to 100 % cumulative probability. All simulation runs are very close to each other and show only minor differences. This observation is again justified based on the sensor noise simulation introduced into the model. Nevertheless, the measurement EDFs deviate from the simulation course by a few dB up to -80 dB. Above -80 dB the model of the environment and the ray tracing algorithm become noticeable again. Due to reflections of the environment, cells of the radar cuboid in the measurement are filled with power up to -60 dB. Subsequently, the effect of the CCR is visible in the form of steps at -43 dB up to -20 dB in the EDFs.

The corrected EDFs are also close to each other, which means that only small deviations in the d_{CAVM} are to be expected. The number of cells still indicates that the cycle time of the simulation model does not yet match the real sensor. Nevertheless, the deviation in the number of data points allows a comparison because the deviation is less than 10 %. The problem of comparability, as evident in Sec. 5.1 when analyzing all detections, is less present in the radar cuboid.

The values of the d_{bias} are close to the real measurement due to the noise simulation and a deviation of 3.5 dB is tolerate able given the dynamic range of a radar sensor that spans over 80 dB. Measurement 2 and 3 show the largest deviations to all simulation parameters, with an increased sensitivity of the model to the uncertainties $_Cx^-$, $_Sy^-$ as well as $_Sh^-$.

d_{CAVM} is around 3.5 dB and can be justified by the aleatory uncertainties of the environment on the measurement result. Different areas on the test track produce higher powers in the measurement, which the environment simulation does not cover. In conclusion, the noise simulation distorts the influence of the environment model and the ray tracing algorithm. Thus, before integrating stochastic effects, it is recommended to analyze and optimize the whole simulation chain with ideal test objects and small region of interests.
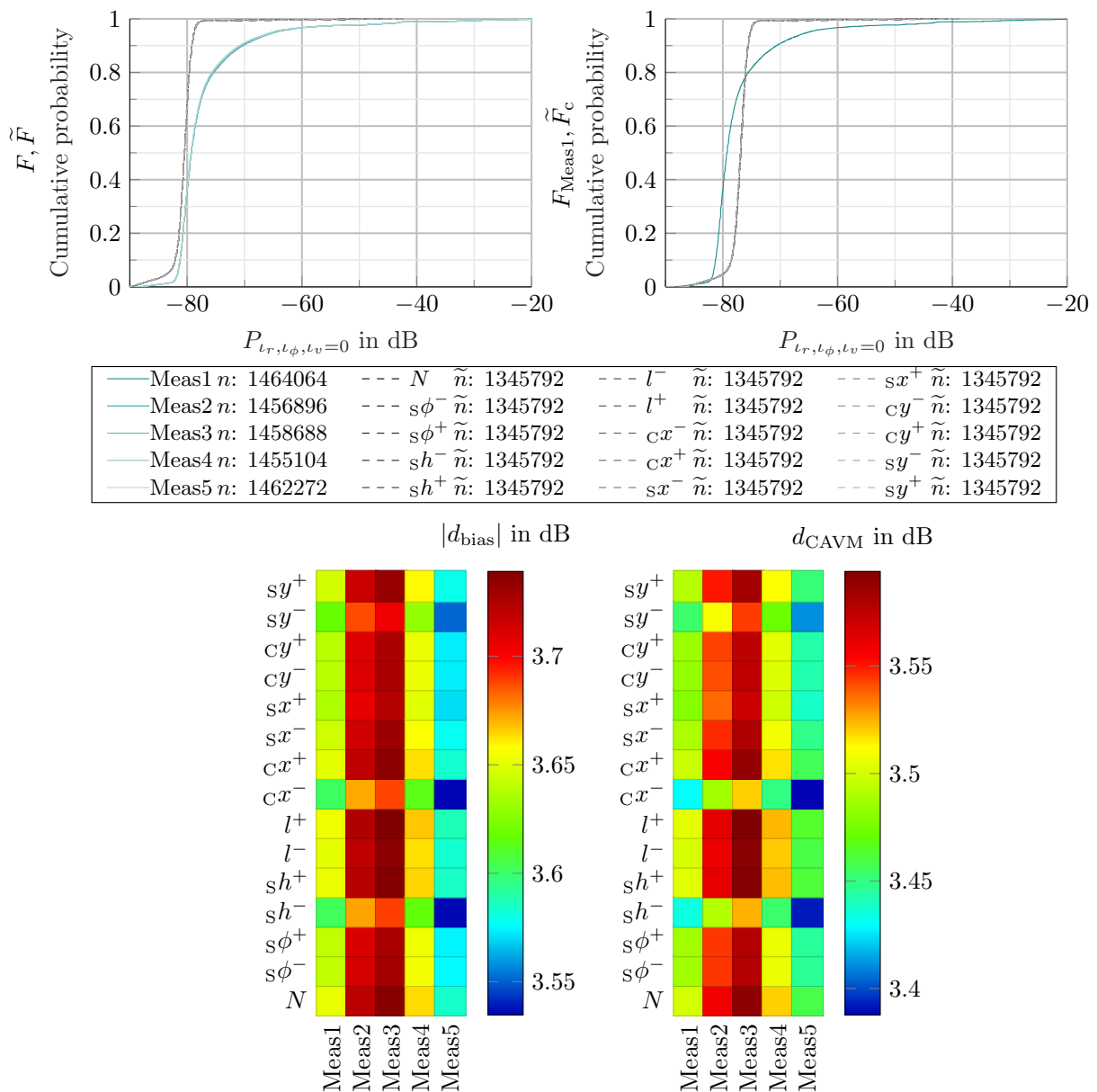
Fig. 20: The DVM Map of the whole radar cuboid is shown. The first diagram visualizes the EDF of all measurements F and of all simulations $\widetilde{F}$. In the second one the EDF of measurement 1 F_{Meas1} and the corrected simulation EDFs $\widetilde{F}_{\mathrm{c}}$ are shown. The second row shows the DVM Map of the above mentioned quantities. The number of analyzed radar cuboid cells of the measurement n and simulation $\widetilde{n}$ is in the legend.

5.3 Clustered Detections on Radar Cuboid Level

In this section, the results of clustered detections at radar cuboid level are presented and analyzed. Fig. 22 shows the position of the clustered detections in the Cartesian sensor coordinate system. The cluster of CCR is number 5 and highlighted in the figure with a red circle at $_Sx = 29.4\,\mathrm{m}$ and $_Sy = \text{-}4.2\,\mathrm{m}$. At close range of the sensor, some detections are visible due to the reflection of the road surface. Especially on the x-axis and in the edge region of the sensor more detections due to this effect show up. All detections from $_Sx = 40\,\mathrm{m}$ are located at the road border where the asphalt ends and the vegetation starts. This clutter is present in all measurements and differs only slightly between measurements.

In detail, the CCR and the corresponding bins of the radar cuboid are considered below. The upper part of Fig. 21 again represents the uncorrected and corrected EDFs of simulations and measurements and measurement 1, respectively.

The step shape visible in the measurements results from the 4 different range azimuth cells analyzed in the evaluation based on the CCR's position. The variations of a cell are within a few dB over time, which can be seen in the slope of the EDFs. Furthermore, the reproducibility of the measurements is exceptionally high, which is reflected in the overlap of the courses of the measurements.

In general, the simulations have a clear model bias. Here, the first modeling errors of the signal processing are evident. The model is calibrated to the RCS of the CCR at position 1 and therefore a difference in the calculation of the RCS from the radar cuboid power to the detections exists.

Furthermore, there is a clear influence of the uncertainties propagated by the simulation. In three simulations, the step shape is similar to the measurements, but the slope itself is more substantially smeared and not as steep. All other simulations have a much lower slope after the initial step. The beginning of the simulation slopes can be explained by the noise simulation. Two radar cuboid cells are considered here, which are not yet affected by the power increase due to the CCR. From this, a modeling error can again be identified. The window functions of the real sensor differ from those of the simulation model because the power increase of the CCR does not smear as far into neighboring bins as in the measurement. The discrepancy in the number of cell values considered indicates a sampling difference between the model and the real sensor.

The previously described findings from the EDFs are also reflected in the DVM Map. The simulations of $_Sy^-$, $_Cx^-$ as well as $_Sh^-$ show the lowest deviations in model bias as well as scattering error. This is in agreement with the findings from Sec. 5.2, where also the mentioned uncertainties represent the smallest deviation (see also Fig. 20).
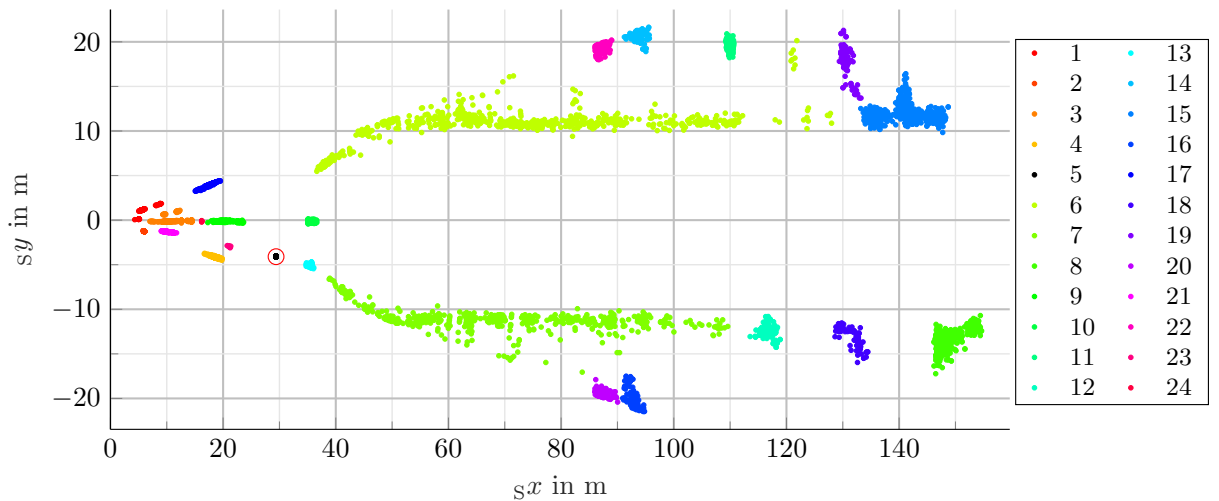

Fig. 22: Clustered radar detections of all measurements in the sensor coordinate system, where the color represents the belonging to a cluster. The red circle at $_Sx = 29.4\,\mathrm{m}$ and $_Sy = \text{-}4.2\,\mathrm{m}$ shows cluster 5, where the CCR is located.
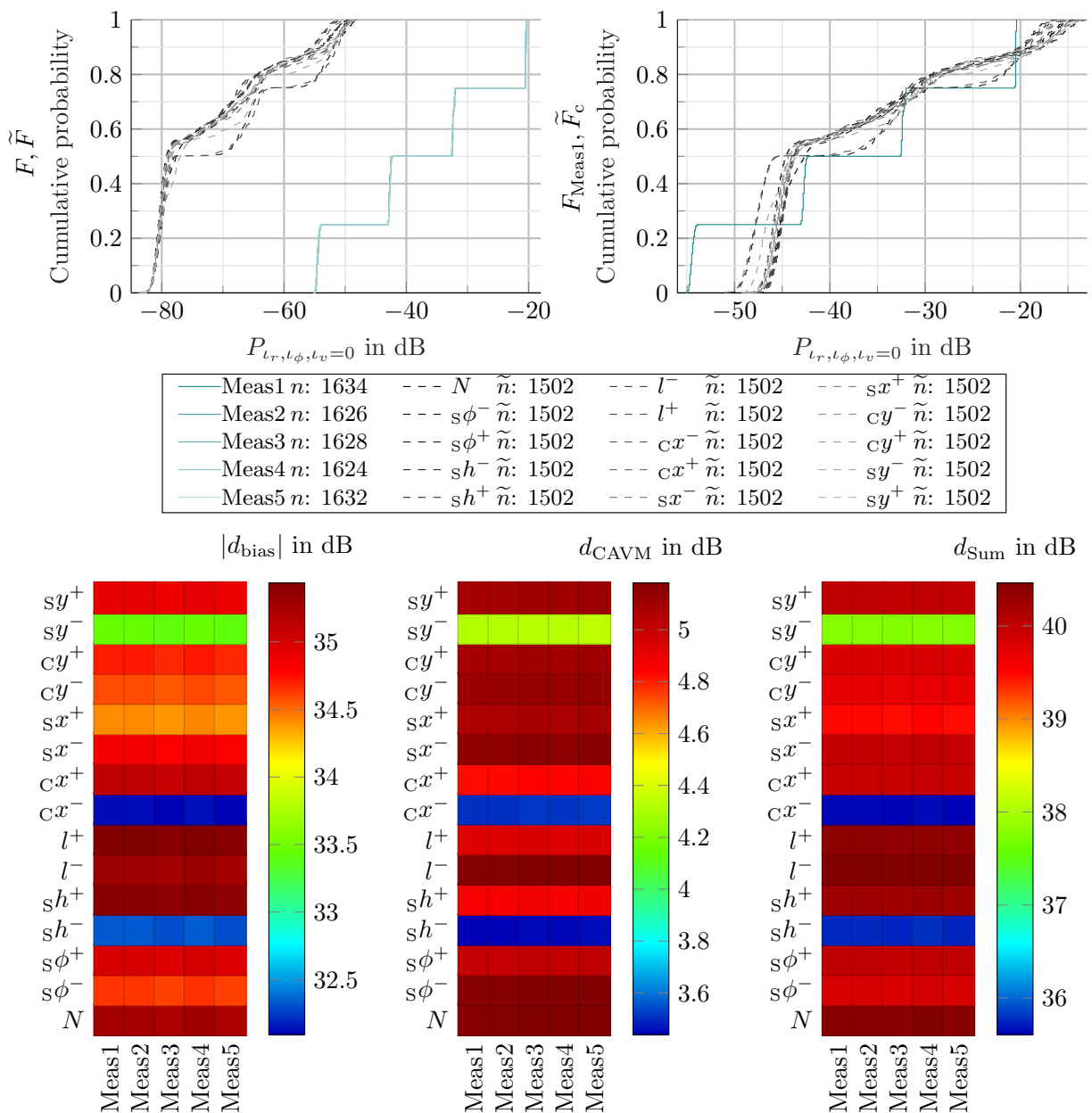
Fig. 21: The uncorrected and corrected EDFs as well as the DVM Map with d_{Sum} of the CCR cluster 5 is shown.

In comparison to the previous figures, the heat map of d_{Sum} is additionally shown, since the further considerations in Fig. 21 are based on these results. To compare the clusters with each other, the maximum value of d_{Sum} and its corresponding cell in the heat map is used. The values for $|d_{\text{bias}}|$ as well as d_{CAVM} of the uncertainty measurement combination are transferred to a separate bar diagram in Fig. 23.

Clusters 1 to 4, 10, 13, 17, 21 and 24 in Fig. 23 show the clear difference between simulation and measurement in the close range of the sensor. As already described, ground reflections are not further considered in the simulation model, which results in the visible difference between simulation and measurement. Clusters 6 and 7 represent the largest clusters with the main clutter due to vegetation. The influence of vegetation is not as large as the deviations in the near sensor range, since the distance is larger and thus the power in the radar cuboid approach the noise level. Nevertheless, a clear difference between simulation and measurement can be identified.

Among the clusters farthest from the sensor in the range, number 18 stands out. At this location, there is an intersection of runway and taxiway on the August-Euler airfield. The effects of the change in ground properties are thus detectable in the methodology using the DVM Map. Across all clusters, no trend in the measurements and uncertainty parameters can be detected, which on the one hand speaks for the good reproducibility of the measurements and on the other hand for the high sensitivity of the radar model.
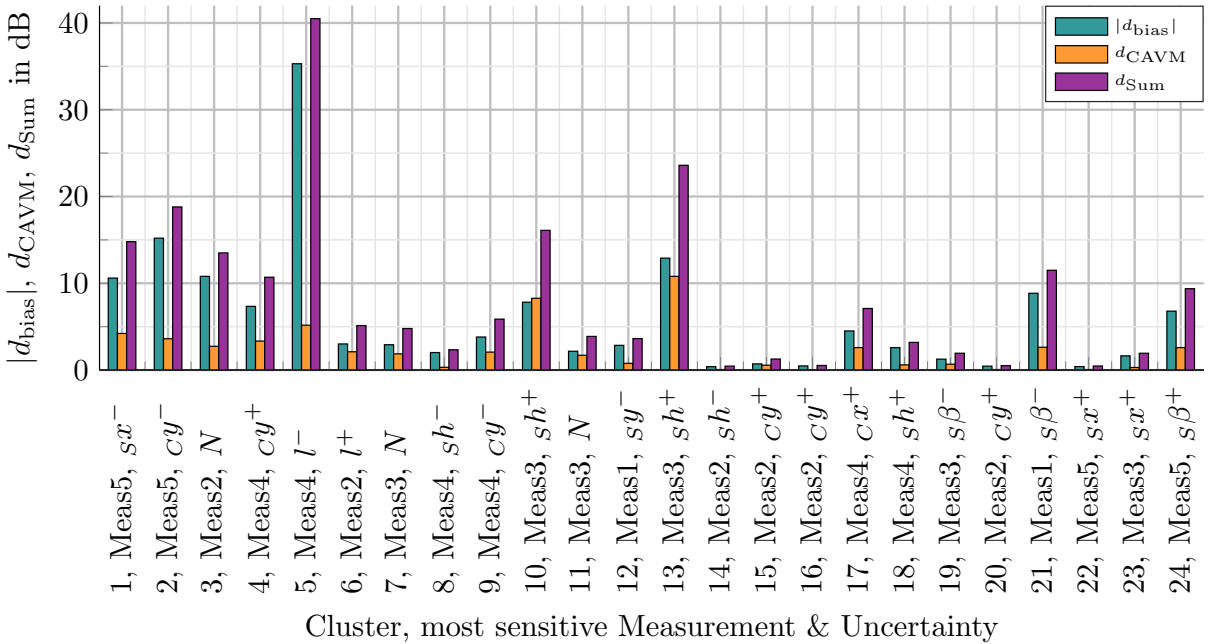

Fig. 23: Bar plot of all clusters of $|d_{\text{bias}}|$, d_{CAVM} and d_{Sum} with the corresponding cluster number as well as the biggest influence based on the measurement uncertainty parameter combination on the x-axis.

5.4 Each Range and Azimuth Radar Cuboid Cell

To increase the interpretability of the results, the outcome of the DVM Map are plotted on a satellite image. For this purpose, the measured positions of the sensor and the determined orientation are taken as origin and the range as well as azimuth resolution of the radar cuboid is used to distribute its cells over the satellite image. From top to bottom, $|d_{bias}|$, d_{CAVM} and d_{Sum} are visualized in Fig. 24. The coloring of the cells corresponds to the results of the DVM Map per radar cuboid cell. As an example, the DVM Map of cell $\iota_r = 18, \iota_\phi = 16$ is shown. As in Fig. 23 for further analysis, the maximum of d_{Sum} for the measurement and uncertainty combination is used. Therefore, this combination is used to color the cell in the value of $|d_{bias}|$, d_{CAVM} and d_{Sum}.

In the top plot it can be seen that there is a deviation of about 30 dB in the area where the CCR is located. The smearing of the power in neighbor bins due to the windowing function is included in the simulation model, but an assignment to the causal effect of the deviation is difficult. On the one hand, the power of the CCR is too low, which can be corrected by calibrating the model at radar cuboid level. On the other hand, the window function in the model is iteratively determined, which means that measurement and modeling errors may also be present here.

Directly next to the highlighted cell $\iota_r = 18, \iota_\phi = 16$ there is an area with the maximum deviation between simulation model and measurement, which is 55 dB. In the simulation model there is no input data from the raytracing algorithm and only the noise simulation fills these cells of the radar cuboid. During the measurements, no objects or asphalt peculiarities were noticed that justify this increase in power. For these reasons, there has to be an effect in the signal processing of the radar sensor, which is not considered in the radar model and is triggered by the CCR. Therefore, the method is able to identify a systematic model error at this point. Along the runway there are still increases in the model bias, ranging from 15 to 25 dB.

In the second satellite plot the scattering error represented by d_{CAVM} is visualized. The highest deviation is present at the CCR. The distribution

shape in the measurement looks like a step function. The simulation EDFs are not so steep and the propagated uncertainties have a big impact on the shape especially at lower powers. This proves the high sensitivity of the radar model chain with respect to the measurement data uncertainty

Along the runway and the transition to vegetation, notable deviations are evident, as already in $|d_{bias}|$. Of particular interest is the intersection of the taxiway with the runway highlighted by the red circle. Due to the transition between vegetation and asphalt, higher values in comparison to the surrounding cells for d_{CAVM} are shown.

In the d_{Sum} satellite plot, the differences now become even more apparent. In addition to the features of the runway, vegetation, and intersection already mentioned, the sensor's close range has significant discrepancies. This underlines the findings from the analysis of the clusters in the previous chapter.
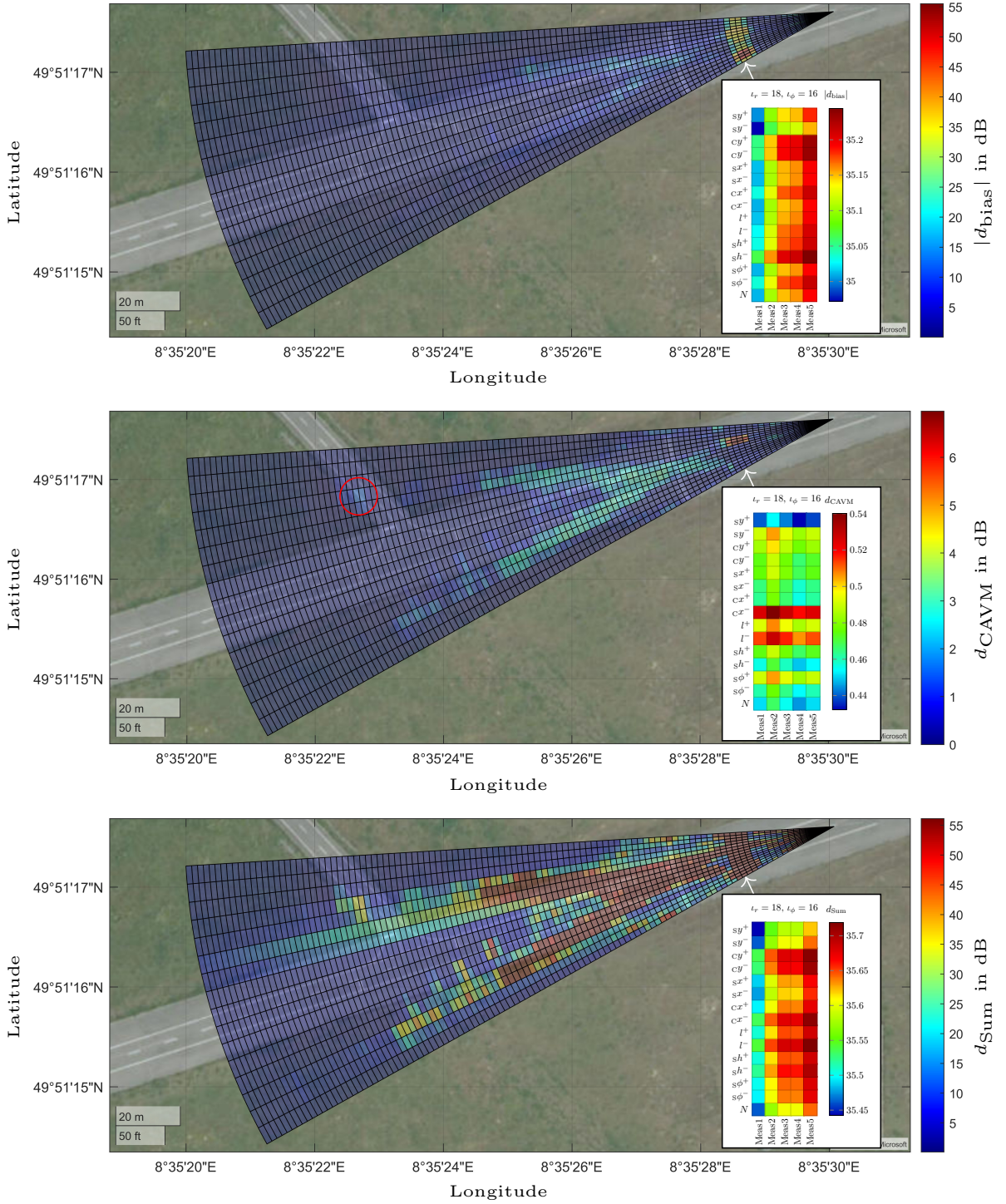
Fig. 24: Satellite image in which the results of the DVM Map are shown. From top to bottom, the validation results for $|d_{\text{bias}}|$, d_{CAVM} and d_{Sum} are illustrated. In the lower right corner of each plot, the DVM Map of cell $\iota_r = 18$, $\iota_\phi = 16$ is shown as an example on which the coloring in the satellite plot is based.

6 Conclusion

This paper introduces the concept of the DVM Map and its application to radar data using a static scenario. Based on four ECs, the need for an extension of the existing DVM definition is presented. A methodology which allows to apply the DVM Map on radar cuboid and detection level is described. A validation study is exemplified using the described experimental setup and an adapted radar simulation model.

It is evident that looking at all detections only makes sense if the environment simulation is matched with the ray tracing algorithm. It can be seen that the different number of detections is a fundamental problem of the simulation model with all its components.

The evaluation of the entire radar cuboid has the advantage that the number of data depends only on the correct model parameterization. The comparison reveals whether aleatory uncertainties such as noise are modeled correctly.

The clustered detections are used to analyze the areas that are particularly affected by power differences. In the case of the validation study, the analysis of the CCR shows high deviations between measurement and simulation. Therefore, objects can be identified and evaluated particularly well using this approach.

Finally, all cells of the radar cuboid are analyzed. Local aleatory and epistemic uncertainties of the environment model, e.g. vegetation and asphalt, are visible. Additionally, effects of the radar signal processing model can be separated from environment model and ray tracing algorithm. In general, it can be seen that validation of a sensor model and its signal processing is only possible if the environment simulation is qualified regarding physical effects and aleatory uncertainties.

Overall, it can be seen that the DVM Map with its application to the different levels significantly increases the interpretability of scenarios in the following manner:

- The DVM Map gives the output score in units of the analyzed size.
- The DVM Map gives information about the sensitivity of each modeled reference data inaccuracy in the simulation with each measurement.
- Clustering gives dedicated information about object model and modeling errors.
- Using the DVM Map in combination with the satellite plot, errors can be spatially localized and thus the environment model can be examined.

The DVM Map can also be used to compare measurements with each other and thus investigate stochastic effects such as rain and compare the similarity of rain conditions between measurements. Furthermore, measurement setups that have to be dismantled and reassembled can be examined and comnpared by a reference measurement. Furthermore, an analysis of the signal-to-noise ratio offers further potential to improve the understanding of the underlying effects in the future. So far, the consideration of uncertainties is limited to the upper and lower bounds, which does not take into account mutual influences of the uncertainties. Therefore, it is recommended in a next step to combine the uncertainties with each other and to limit the parameter space in the process. As soon as further uncertainties are added and not only the upper and lower bounds of the uncertainties are varied, this inevitably leads to an explosion of the parameter space. Assuming a parameter number in the present scope with five instead of three variations is examined full factorially, the number of necessary simulations for just one scenario with a CCR is

$$n_{\text{sim}} = n_{\text{var}}^{n_{\text{param}}} = 5^7 = 78125. \qquad (10)$$

This estimate does not include material properties or complex geometries and each simulation has to be repeated for a change in the model itself. Therefore, the parameter sensitivity of the model must be determined in advance and thus reduce the parameter space.

In the future, we will extend the methodology developed here for static scenarios to dynamic scenarios. However, this poses challenges specifically with respect to temporal aggregation of data. It is imperative that these challenges be resolved in order to qualify the methodology for model validation.

Declarations

Conflict of interest The corresponding authors declare no conflict of interest.

References

[1] Devision, U.T.: New Assessment/Test Method for Automated Driving (NATM) Guidelines for Validating Automated Driving System (ADS) – amendments to ECE/TRANS/WP.29/2022/58. UNECE (2022). https://wiki.unece.org/download/attachments/184615115/VMAD-SG2-39-01%20SG2%20solved%20UK%20comments%20on%20NATM%20Guidelines.docx?api=v2 Accessed 2023-04-17

[2] Society of Automotive Engineers: SAE-J3016: Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles (2021). https://www.sae.org/standards/content/j3016_202104/ Accessed 2022-01-06

[3] Zheng, L., Ma, Z., Zhu, X., Tan, B., Li, S., Long, K., Sun, W., Chen, S., Zhang, L., Wan, M., Huang, L., Bai, J.: TJ4DRadSet: A 4D Radar Dataset for Autonomous Driving. In: 2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC), pp. 493–498. IEEE, Macau, China (2022). https://doi.org/10.1109/ITSC55140.2022.9922539. https://ieeexplore.ieee.org/document/9922539/ Accessed 2023-04-03

[4] 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)

[5] Holder, M.F., Rosenberger, P., Winner, H., D'hondt, T., Makkapati, V.P., Maier, M., Schreiber, H., Magosi, Z., Slavik, Z., Bringmann, O., Rosenstiel, W.: 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. IEEE, Maui, HI (2018). https://doi.org/10.1109/ITSC.2018.8569423. https://ieeexplore.ieee.org/document/8569423/ Accessed 2020-11-10

[6] Oberkampf, W.L., Trucano, T.G.: Verification and validation benchmarks. Nuclear Engineering and Design **238**(3), 716–743 (2008). https://doi.org/10.1016/j.nucengdes.2007.02.032

[7] Viehof, M.: Objektive Qualitätsbewertung von Fahrdynamiksimulationen durch statistische Validierung. PhD thesis, Technische Universität Darmstadt, Darmstadt (2018). http://tuprints.ulb.tu-darmstadt.de/7457/

[8] Benke, K.K., Norng, S., Robinson, N.J., Benke, L.R., Peterson, T.J.: Error propagation in computer models: analytic approaches, advantages, disadvantages and constraints. Stochastic Environmental Research and Risk Assessment **32**(10), 2971–2985 (2018). https://doi.org/10.1007/s00477-018-1555-8. Accessed 2023-04-18

[9] Rosenberger, P.: Metrics for specification, validation, and uncertainty prediction for credibility in simulation of active perception sensor systems. PhD thesis, Technische Universität Darmstadt, Darmstadt (2023). https://doi.org/10.26083/tuprints-00023034. http://tuprints.ulb.tu-darmstadt.de/23034/

[10] Magosi, Z.F., Wellershaus, C., Tihanyi, V.R., Luley, P., Eichberger, A.: Evaluation Methodology for Physical Radar Perception Sensor Models Based on On-Road Measurements

for the Testing and Validation of Automated Driving. MDPI Energies **15**(7), 20 (2022). https://doi.org/10.3390/en15072545. Number: 7 Publisher: Multidisciplinary Digital Publishing Institute. Accessed 2022-03-31

[11] Schaermann, A.: Systematische Bedatung und Bewertung umfelderfassender Sensormodelle. PhD thesis, Technische Universität München, München (May 2020). http://nbn-resolving.de/urn/resolver.pl?urn:nbn:de:bvb:91-diss-20200526-1518611-1-9

[12] Oberkampf, W.L., Barone, M.F.: Measures of agreement between computation and experiment: Validation metrics. Journal of Computational Physics **217**(1), 5–36 (2006). https://doi.org/10.1016/j.jcp.2006.03.037. Accessed 2022-04-08

[13] Liu, Y., Chen, W., Arendt, P., Huang, H.-Z.: Toward a Better Understanding of Model Validation Metrics. Journal of Mechanical Design **133**(7) (2011). https://doi.org/10.1115/1.4004223. Accessed 2021-08-23

[14] Fréchet, M.R.: Sur quelques points du calcul fonctionnel. Rendiconti del Circolo Matematico di Palermo (1884-1940) **22**, 1–72 (1906). https://doi.org/10.1007/BF03018603

[15] Ngo, A.: A methodology for validation of a radar simulation for virtual testing of autonomous driving. PhD thesis (2023). https://doi.org/10.18419/opus-12703. Accepted: 2023-02-07T13:58:14Z ISBN: 9781833425246. http://elib.uni-stuttgart.de/handle/11682/12722 Accessed 2023-02-11

[16] Ferson, S., Oberkampf, W.L., Ginzburg, L.: Model validation and predictive capability for the thermal challenge problem. Computer Methods in Applied Mechanics and Engineering **197**(29), 2408–2430 (2008). https://doi.org/10.1016/j.cma.2007.07.030. Accessed 2021-08-17

[17] Eder, T.: Simulation of Automotive Radar Point Clouds in Standardized Frameworks. PhD thesis, Technische Universität München, München (November 2021). Google-Books-ID: QYJSEAAAQBAJ

[18] Stripling, H.F., Adams, M.L., McClarren, R.G., Mallick, B.K.: The Method of Manufactured Universes for validating uncertainty quantification methods. Reliability Engineering & System Safety **96**(9), 1242–1256 (2011). https://doi.org/10.1016/j.ress.2010.11.012. Accessed 2022-05-02

[19] Williamson, R.C., Downs, T.: Probabilistic arithmetic. I. Numerical methods for calculating convolutions and dependency bounds. International Journal of Approximate Reasoning **4**(2), 89–158 (1990). https://doi.org/10.1016/0888-613X(90)90022-T. Accessed 2022-05-04

[20] Ferson, S., Kreinovick, V., Ginzburg, L., Myers, D.S., Sentz, K.: Constructing Probability Boxes and Dempster-Shafer Structures. Technical Report SAND2002-401 (January 2003). https://www.semanticscholar.org/paper/Constructing-Probability-Boxes-and-Dempster-Shafer-8eff743341521cca30f6d2a48df50bf6977c96b2

[21] Voyles, I.T., Roy, C.J.: Evaluation of Model Validation Techniques in the Presence of Aleatory and Epistemic Input Uncertainties. In: 17th AIAA Non-Deterministic Approaches Conference. American Institute of Aeronautics and Astronautics, Kissimmee, Florida (2015). https://doi.org/10.2514/6.2015-1374. https://arc.aiaa.org/doi/10.2514/6.2015-1374 Accessed 2021-08-18

[22] Linnhoff, C., Rosenberger, P., Schmidt, S., Elster, L., Stark, R., Winner, H.: 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), Indianapolis, IN, USA, pp. 2688–2695 (2021). https://doi.org/10.1109/ITSC48978.2021.9564661

[23] Linnhoff, C., Hinsemann, T., Rosenberger, P., Elster, L.: PerCollECT - LidarLimbs. PerCollECT. original-date: 2021-03-19T07:34:25Z (2022). https://github.com/PerCollECT/LidarLimbs Accessed 2023-04-03

[24] Brune, A.J., West, T.K., White, L.M.: Calibration Probe Uncertainty and Validation for the Hypersonic Material Environmental Test System. NTRS Author Affiliations: NASA Langley Research Center NTRS Report/-Patent Number: NF1676L-33286 NTRS Document ID: 20200002798 NTRS Research Center: Langley Research Center (LaRC) (2020). https://doi.org/10.2514/1.T5839

[25] Holder, M.F.: Synthetic generation of radar sensor data for virtual validation of autonomous driving. PhD thesis, Technische Universität Darmstadt, Darmstadt (2021). https://doi.org/10.26083/tuprints-00017545. http://tuprints.ulb.tu-darmstadt.de/17545/

[26] Github Open MSL: Reflection Based Radar Model. https://github.com/openMSL/sl-1-1-reflection-based-radar-object-model, accessed on 23.03.2023 (2023)