

# Multi-level and Metrics Evaluation Approach for Data-Driven Based Sensor Models

Hexuan Li<sup>1</sup>  · Nadine Bammingner<sup>1</sup> · Li Wan<sup>2</sup> · Arno Eichberger<sup>1</sup>

Received: 22 August 2022 / Accepted: 6 September 2023 / Published online: 27 March 2024  
© The Author(s) 2024

## Abstract

Nowadays, with increased sensor perception performance for Advanced Driver Assistance Systems (ADAS), scenario-based simulation is becoming more frequent to manage the complexity of reality in terms of cost and time. The perception system provides the basis for the vehicle guidance algorithms calculation, but the simulation of ADAS sensors is a challenging task in virtual testing. Literature reports the magnitude of relevant modelling approaches and data-driven models becoming increasingly important. A basic method is to fit the sensor output in the virtual environment with high-fidelity measurements of real-world scenarios, thus a direct relation can be established between real and synthetic sensor data. To prove the suitability of a method, it is necessary to quantify the gap between simulation and reality to determine the performance of different models. In this work, authors address this problem and visualize the gap by introducing a multi-level evaluation approach that combines Model Generalization Ability Evaluation and Case Implicit Performance Evaluation. The former directly evaluates the model's overall performance, while the latter is used for specific cases in simulation. The study shows that this combined evaluation approach provides an in-depth framework for evaluating sensor models to make the differences apparent.

**Keywords** Sensor model evaluation · Uncertainty quantification · Estimation performance evaluation · Validation and verification · Virtual simulation

## 1 Introduction

Beginning with pioneering work in the late 1980s to the present [1], some of the most relevant technology demonstrations, competitions, and challenges have contributed significantly to the development of highly autonomous driving technology. However, reliable and accurate perception is the basis for driving automation. To accelerate and facilitate the development and verification of ADAS functions, scenario-based simulation is widely used [2]. For this reason, reliable sensor models are essential, represent the detection performance of the real sensor as closely as possible. However, all virtual sensors must be validated to quantify the difference between the synthetic data and the actual measurement data, thus ensuring fidelity for the intended use [3].

Given the differences in intended use, the requirements for sensor models can vary considerably. Sensor models must be validated based on balancing model realism and computation time. Schlager et al. inferred the fidelity of a sensor model by considering the inputs and outputs of the model as well as the modelling principles [4]. In addition, a more widely used approach is direct qualitative comparison of synthetic and real sensor data. For example, the work of [5–7] evaluates model fidelity by plotting comparisons between simulation results and real measurements, but this assessment is often based on expertise and development experience. To eliminate the drawbacks of empiricism, the focus of Ref. [8] is on comparing model outputs with measurements through cumulative distributions to assess model performance directly. Additionally, the study by Yang et al. [9] compared the distribution of model outputs in confidence space, thereby verifying the validity and effectiveness of the statistical model for representing real environment. In contrast to direct evaluation, sensor models are evaluated indirectly by feeding synthetic data to downstream algorithms to validate the behaviour of functions [10]. The indirect evaluation focuses on the performance of the perception system in functional applications and less on the realism

---

✉ Hexuan Li  
hexuan.li@tugraz.at

<sup>1</sup> Institute of Automotive Engineering, Graz University of Technology, Inffeldgasse 11, 8010 Graz, Steiermark, Austria

<sup>2</sup> IPG Automotive GmbH, Karlsruhe, Germany

of the sensor model itself. Therefore, by building a multi-layer evaluation, Ngo et al. combined direct and indirect evaluation methods to measure the gap between simulation and reality [11]. However, the explicit sensor model evaluation metrics should be designed relatively simple. In Refs. [12, 13], the same metrics are used to evaluate the point cloud consistency of the LiDAR model. These evaluation approaches are based on analysing key performance indicators and scenarios that reveal the sensor model's realism and performance.

Since the generation and validation of sensor data is a relatively new research area, no uniform ranking criteria have been given in previous studies. Therefore, there are challenges in selecting and designing metrics. Moreover, it is necessary to ensure that different models have a corresponding set of evaluation metrics. Among the various types of modelling mentioned above, simulation of physical models face the challenge of being extremely demanding hardware which requiring excellent computational power [14, 15]. Furthermore, the fidelity of ideal sensor models can not be guaranteed due to the lack of representation of detection uncertainties [5]. In order to balance computational time and fidelity, phenomenological sensor models could be the best choice, thus become the focus of this paper, which can better reflect the uncertainty and realism of detection through real data-driven modelling, and also can optimize the computational efficiency by using statistical algorithms [16, 17]. In this paper, the evaluation of a data-driven sensor model is investigated to refine the research area of sensor model evaluation. A series of evaluation metrics are proposed from the perspective of Model Generalization Ability Evaluation (MGAE) and Case Implicit Performance Evaluation (CIPE) for currently popular modelling approaches. In the MGAE phase, authors use the data collected in real scenarios to build sensor models and evaluate the generalization of the trained models. In addition, a digital twin-based simulation is used to replicate the real test scenario carried out in the CIPE phase, including road logic and test vehicle trajectories. Finally, the different modelling methods are ranked and visualized by Ranking Vector (RV).

The remainder of the paper is structured as follows: First, Sect. 2 introduces the generation of synthetic data using the data-driven modelling approach. Section 3 describes the proposed validation methodology with the different evaluation metrics in detail. Section 4 presents the conducted experiments and discusses the effectiveness of the method. Finally, Sect. 5 concludes the contribution to sensor model evaluation.

## 2 Data-Driven Modelling Approaches

The current mainstream modelling approach is based on experimental on-road driving tests. This has the advantage of improving the efficiency of modelling and expressing

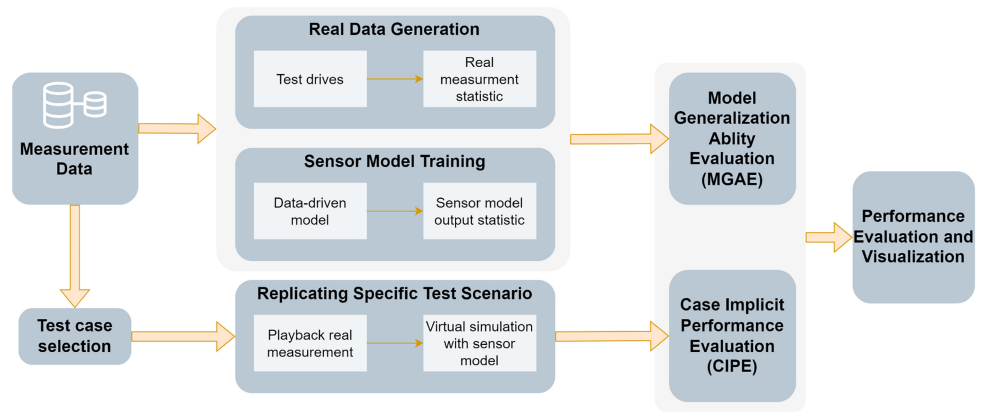
the perceived behaviours of sensors as much as possible with limited data. Although some fidelity is compromised compared to physical models, these models can achieve a better balance of simulation time and computing power. Data-driven approaches represent a significant improvement over traditional empirical models. Machine Learning (ML) algorithms or statistical distributions are used to determine the relationship between the inputs and outputs using a training dataset to represent all the behaviours in system. Once the model has been trained, it can be tested with independent datasets to determine how well it generalizes to unseen data [18]. Magosi et al. [19] introduced sensor modelling approaches from the system integrator's perspective. The V-model-based development process has different performance requirements for the sensors at different levels of abstraction, which inevitably leads to different development and verification strategies. Therefore, in present study, we will propose a data-driven model validation-oriented approach.

## 3 Model Evaluation Methodology

The methodology introduced in this section focuses on measuring this simulation-to-reality gap of a sensor model for a specific intended use. The complete evaluation and simulation architecture is illustrated in Fig. 1. The measurement dataset is the sensor data collected by the vehicle in test on a real road. The real data for the modelling target can be generated from the measurement dataset as ground truth data. The target is then modelled based on a data-driven approach, and a trained sensor model can be efficiently obtained in the sensor model training phase by ML [20]. The trained sensor model is then validated and simulated in the downstream phase for replicating specific test scenarios. An accelerated approach test case is selected from the collected dataset for comparison with real data. Meanwhile, the environment model provides a digital twin-based road logic and scenarios during the simulation.

To refine the evaluation of two modelling approaches in the validation phase, a multi-level testing method is proposed that consists of MGAE and CIPE illustrated in Fig. 1. In the MGAE phase, the modelling training performance is evaluated using pre-defined metrics for data-driven-based approaches. For the CIPE evaluation, some test cases are selected randomly from the database. Additionally, a digital twin-based scenario and vehicle trajectory are replicated in the simulation to compare the differences between the real measurements and the synthetic data generated by the sensor model. Finally, the proposed evaluation approach is used to calculate model performance rankings through comprehensive measures and visualize the results through polar charts to measure the gap between simulation and reality.

**Fig. 1** Overview of proposed validation approach



### 3.1 Model Generalization Ability Evaluation

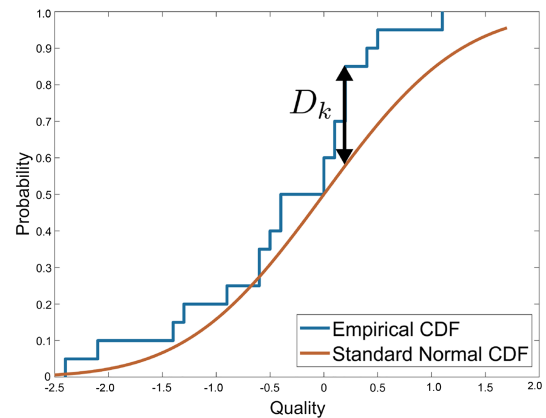
In this work, the MGAE focuses on the LiDAR model training performance. Namely, during model training, various problems affect the regression algorithms. For example, inefficient learning or too many interference factors can result in unsatisfactory analysis. Therefore, there is a need to evaluate the variability between the predicted and reference data. MGAE is designed to assess that the overall model predictions and the collected data can be matched as closely as possible. As the training time grows, the Prediction Error decreases, more accurate answers are obtained. Finally, the error converges to a certain level which satisfied the model accuracy as expected. The core concept is to verify whether the trained model achieves a better prediction performance.

In this work, we focus on understanding and quantifying uncertainty in the modelling process and providing better information to make more robust decisions from the estimator. The model output incorporating uncertainty provides more information than an average or deterministic prediction. Once the model is prepared, the average prediction performance and the uncertainty should be adequately assessed. The ideal validation metric reflects intuitive information about the key differences between model outputs and observed distributions, such as statistical distance or difference. Additionally, the reference data with the same inputs as the real distribution  $Q$  are calculated while comparing the variability between reference distribution  $Q$  and model outputs  $P$ . Here three metrics are used to evaluate the generalization ability of model output. The first considered is the Wasserstein distance  $D_{WS}$ , which measures the minimum effort required to reconfigure the probability mass of one distribution to recover another. It is defined in Eq. (1), where  $I$  and  $J$  respectively are number of points for two distributions in the data sets.  $f_{i,j}$  represents the optimal flow to rearrange the distributions and the Euclidean distance is given as  $d_{i,j}$ . The detailed derivation formula can be found in Ref. [21]. Subsequently, integral probability measures are introduced. In contrast to  $f$ -divergence, this category of metrics assesses the difference rather than the ratio in probability

measures. Where Kolmogorov Distance  $D_K$  is the maximum  $L_1$ -norm between two Cumulative Distribution Functions (CDF) bounded  $[0, 1]$  [22] and calculated by Eq. (2), where  $F_P(x)$  and  $F_Q(x)$  is a CDF for the prediction probability distribution  $P$  and real probability distribution  $Q$  over the random variable  $x$  respectively. Figure 2 illustrates an example of the distance for a set of samples where an Empirical Cumulative Density Function (ECDF) is formed. Finally, the Area Metric  $D_{area}$ , proposed by Ferson et al. [23], is a popular validation metric in engineering for assessing the difference area between two CDFs ( $F_P(x)$  and  $F_Q(x)$ ) defined in Eq. (3). Figure 3 illustrates an example that this metric also represents the distance between the quantile function [24].

$$D_{WS}(P, Q) = \frac{\sum_{i=1}^I \sum_{j=1}^J f_{i,j} d_{i,j}}{\sum_{i=1}^I \sum_{j=1}^J f_{i,j}} \tag{1}$$

$$D_K(P, Q) = \sup_{x \in \mathbb{R}} |F_P(x) - F_Q(x)| \tag{2}$$



**Fig. 2** An example of Kolmogorov Distance

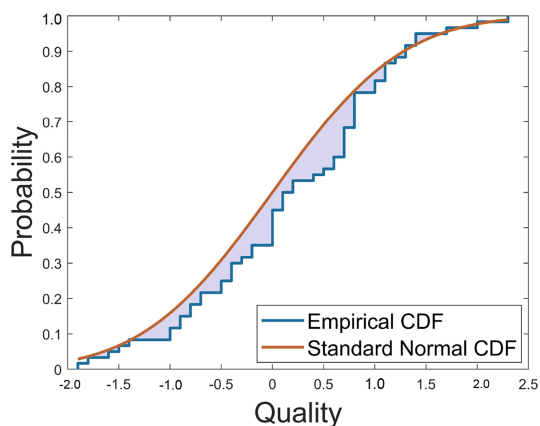


Fig. 3 An example of Area Metric

$$D_{\text{area}}(\mathbb{P}, \mathbb{Q}) = \int |F_P(x) - F_Q(x)| dx \tag{3}$$

### 3.2 Case Implicit Performance Evaluation

In previous sub-section, three metrics were used to evaluate the performance of the model in fitting the reference data. This subsection introduces the implicit or indirect evaluation of the sensor model by assessing the output of test cases fed with real vehicle trajectory and dynamic information in simulation. Therefore, CIPE should be executed combined with simulation.

Some test cases are randomly selected from the test set samples in established simulation platform. To evaluate the difference between simulation results and real measurement data, pattern similarity metrics are proposed including Chebyshev distance  $D_c$ , Pearson correlation-coefficient  $C_{pc}$  and cross-correlation coefficient  $C_{cc}$ . Because the model output is a deterministic prediction when applied in the simulation, it is more appropriate to use distance metrics to evaluate them. In addition, the similarity assessment of the overall model output also appears to be of interest.

Chebyshev distance  $D_c$  is a metric in vector space [25], where the distance between two points is defined as the maximum of the absolute value of the difference between their coordinate values. Equation (4) shows the definition of  $D_c$ , where  $p$  and  $q$  are model output and real measurement data, respectively.  $p_i$  and  $q_i$  are the coordinates in the dataset.

$$D_c(p, q) = \max_i (|p_i - q_i|) \tag{4}$$

Pearson correlation-coefficient  $C_{pc}$ , with value between -1 and 1 is used to measure the degree of correlation (linear correlation) between two variables [12]. When  $C_{pc}$  is 1, it

becomes a perfectly positive correlation. When  $C_{pc}$  is -1, it becomes a perfectly negative correlation; the larger the absolute value of the  $C_{pc}$ , the stronger the correlation. The closer the correlation coefficient is to 0, the weaker the correlation is. The mathematical expression is shown in Eq. (5).  $p$  and  $q$  respectively are the simulations and real data, with dimension of  $m$ .

$$C_{pc}(p, q) = \frac{\sum_{i=1}^m (q_i - \bar{q}) \cdot (p_i - \bar{p})}{\sqrt{\sum_{i=1}^m (q_i - \bar{q})^2 \cdot (p_i - \bar{p})^2}} \tag{5}$$

Cross-correlation coefficient  $C_{cc}$  is a measurement that tracks the movements of two or more sets of time series data relative to one another [26]. It is used to compare multiple time series and objectively determine match-up extent with each other, particularly, at which point the best match occurs. This metric is very suitable for comparing simulation and real measurement data because the cycle time between simulation and real measurement is usually different, resulting in the amount of collected data and time stamp alignment being inconsistent under the same test. Cross-correlation can be used when measuring information between two different time series. The possible range for the correlation coefficient of the time series data is from -1.0 to +1.0. The closer the cross-correlation value is to 1, the more closely the sets are identical. Equation (6) expresses the calculation, where  $\overline{p_i \cdot q_i}$  is the mean of  $p_i \cdot q_i$  and  $\sigma$  is the standard deviation.

$$C_{cc}(p, q) = \frac{\overline{p_i \cdot q_i} - \bar{p}_i \cdot \bar{q}_i}{\sigma(p) \cdot \sigma(q)} \tag{6}$$

To implement CIPE evaluation, as shown in Fig. 1, some random test cases are deployed in the simulation. In setting, the same trajectory and scenario as in the real measurement are reproduced in the simulation so that the implicit performance of the model in the specific test case can be analyzed.

### 3.3 Evaluation Based on Multiple Performance Measures

Given various prediction models, the evaluation of individual estimators is presented in Sects. 3.1 and 3.2. In this section, the multiple estimators with the same category modelling approaches are summarized and ranked to evaluate comprehensive performance. If only one measure is used, the choice could be arbitrary, and the final results could be incomplete. In this case, it can lead to ranking differences in evaluation. Therefore, it is challenging to give a comprehensive and authoritative evaluation. To derive an overall conclusion, we propose a new visualization method and pairwise comparison rankings to address the above issues based on Pitman’s Closeness Measure (PCM).

PCM is based on the probability of the relative closeness of competing estimators  $\hat{x}$  to the estimate  $x$  [27]. Equation (7) measures the difference between two estimators  $s_1$  and  $s_2$  with respect to the  $i$ -th attribute  $a_i$ .

$$m(s_1, s_2, a_i) = \begin{cases} 1 & \text{if } s_1 > s_2 \\ 0.5 & \text{if } s_1 = s_2 \\ 0 & \text{if } s_1 < s_2 \end{cases} \quad (7)$$

where  $s_1 > s_2$  means  $s_1$  is preferred to  $s_2$ . With the same attribute  $a_i$ ,  $s_1$  is closer to the optimal solution. To use the comparison information, the Multiple-attribute Competition Measure (MCM) is given by Equation (8).

$$MCM(s_1, s_2, a) = \frac{1}{n} \sum_{i=1}^n m(s_1, s_2, a_i) \quad (8)$$

where  $a$  is the vector of attributes. The evaluation metrics used in MGAE and CIPE compose this vector for  $n$  elements.

Each estimator is compared to each other and summarized in a **MCM** matrix defined by Eq. (9) with  $m$  estimators. The **MCM** matrix  $X_{MCM}$  contains all cumulative results of all pairwise comparisons based on Eqs. (7) and (8). The pairwise comparisons reveal well the comparative information of the different estimators. Furthermore, to further calculate the eigenvectors, the eigenvalues  $r$  can be calculated by Perron-Forbenius theorem [28] and expressed by Eq. (10), where  $r$  is the only eigenvalue in the spectral circle of  $X_{MCM}$ . Particularly, if  $MCM(s_1, s_2, a) = 0$ , let  $MCM(s_1, s_2, a)$  equal to 0.0001. Because it is necessary to ensure that the calculated matrix is non-negative irreducible.

$$X_{MCM} = \begin{bmatrix} MCM(s_1, s_1, a) & \cdots & MCM(s_1, s_m, a) \\ \vdots & \ddots & \vdots \\ MCM(s_m, s_1, a) & \cdots & MCM(s_m, s_m, a) \end{bmatrix} \quad (9)$$

$$X_{MCM} \cdot r = \lambda \cdot r \quad (10)$$

For adapting the characteristics of the ranking problem and using the comparison information, Yin et al. [27] developed a ranking approach and define a RV as Eq. (11) for  $m$  estimator and  $r_i > 0$  ( $i = 1, \dots, n$ ) to get around the intransitivity problem.

$$r = [r_1, \dots, r_m]' \quad (11)$$

The elements of RV are all positive and represent how well-estimated the values are relative to each other. The ranking results provided by the RV explain the goodness of the estimator. This means that the larger the element, the better the corresponding estimator. Additionally, more information than ranking information is involved in an RV. Meanwhile,

its cardinal values also quantify the performance of the estimator. For example, the three estimators  $\hat{x}_1, \hat{x}_2, \hat{x}_3$  calculate RV result  $[1, 1.01, 5]'$ . The rank order can be expressed as  $\hat{x}_1 < \hat{x}_2 < \hat{x}_3$ . However, the results for  $\hat{x}_1$  and  $\hat{x}_2$  are very close, so the performance of both can be identified as similar. In contrast, the performance of  $\hat{x}_3$  is much better than the other two estimators. Thus the estimators can be ranked as  $\hat{x}_1 \approx \hat{x}_2 < \hat{x}_3$ .

The properties and advantages of the RV method based on linear mapping have been described and proofed in Ref. [27]. It can be expressed as follows:

- *Homogeneity* The ranking order calculated by performance measures will also be reflected in the RV.
- *Invariance* When a rank is given, it will not be influenced by adding a new performance metric that matches the rank.
- *Monotonicity* Suppose that  $\hat{x}_i > \hat{x}_j$  by the RV approach, if  $\hat{x}_i$  is better or  $\hat{x}_j$  is worse than before, then this ranking will still be maintained.
- *Decisiveness* The RV approach is deterministic in the sense that a unique RV is always available.

## 4 Experiments and Validation Results

In this section, the effectiveness of the proposed comparison approach will be examined in terms of its ability to measure the different sensor modelling approaches accurately. LiDAR models based on different approaches are presented here and created to reflect the longitudinal distance detection performance.

### 4.1 Driving Scenario and Measurement

Digitrans Proving Ground ([www.digitrans.expert/en/](http://www.digitrans.expert/en/)) in Sankt Valentin, Lower Austria, is a closed test terrain for facilitating the development of vehicle function. Additionally, one main research topic is concentrated on the ADAS function test. The test area consists of seven different zones, for example, a basic asphalt track, driving dynamics track, junction and arch, and off-road terrain, illustrated in Fig. 4. To cover a large area of individual scenarios, a unique outdoor rain plant with a length of 80 m and a width of 6 m is implemented in the dynamic driving area (dimensions of 450 m in length and 20 m in width). Three modes of rain intensities referring from light, moderate to heavy rain (10 mm/h to 100 mm/h) are possible. With an additional existing intelligent lighting system, thus a variety of scenarios can be performed. In order to investigate the behaviours of sensors, a complex test series was conducted where six test cases and two conditions (dry road and rain) were selected in this study. The test manoeuvres are illustrated in

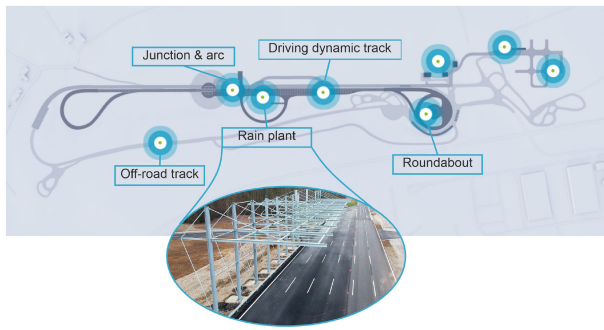


Fig. 4 Overview Digitrans proving ground with a detailed view of rain simulator

Fig 5, and all data collected during the measurements is used to build LiDAR model:

- *Manoeuvre 1* So-called “Accelerate Leaving” represents a drive-off at a traffic light. Both vehicles start with the same speed and a small initial distance between them. The target car in front accelerates until a desired speed beneath the rain plant, while the ego vehicle maintains its initial speed.
- *Manoeuvre 2* “Accelerate Approach” represents an approaching traffic jam scenario, meaning that the starting time of both vehicles is identical, but their speed are different. The ego car starts with a higher velocity and approaches the target car, which has a very low initial speed under the rain simulator.

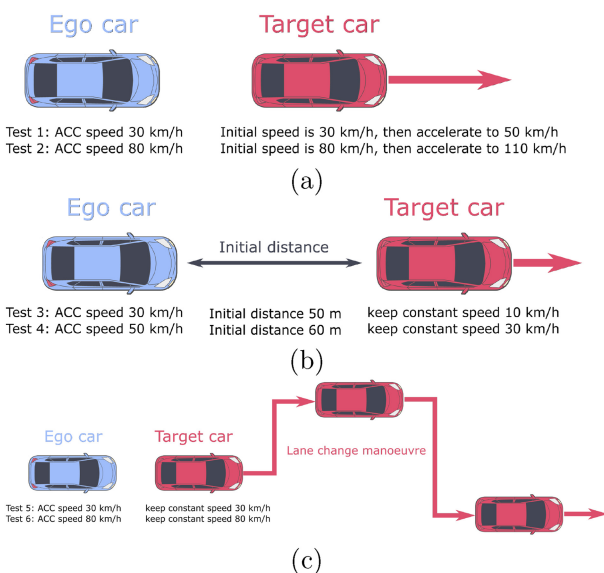


Fig. 5 Test manoeuvres. a Accelerate leaving b Accelerate approach c Lateral leaving

- *Manoeuvre 3* So-called “Lateral Leaving” depict a lateral movement, for example, a lane change manoeuvre. Both vehicles drive with the same constant velocity and the target in front evades to the left and then executes a double lane change to the left.

### 4.2 Digital Twin-Based Simulation

At the CIPE level, the real recorded manoeuvre is replicated in the simulation based on IPG CarMaker [29]. Therefore, the accuracy of the re-simulation result should be examined. The required digital twin-based high-definition map of the Digitrans proving ground is provided by Joanneum Research Forschungsgesellschaft GmbH. Figure 6 illustrates the digital twin-based map with a detailed 3D view of the driving dynamics track.

By converting the GPS data from the measurement system into relative metric coordinates, the necessary data for the trajectory can be extracted. Magosi et al. [3] enables the exact following of recorded trajectories by modifying the CarMaker C-code interface. However, the vehicle under test can only replicate the position and lacks dynamic information.

Along with the help of ScenarioRRR (Scenario Record, Replay, Re-arrange), an extended toolbox of IPG, the converted GPS routes can be converted into a ready-to-play test run in CarMaker. Table 1 shows the accuracy of re-simulation in CarMaker. We compare the difference of some basic kinematic information between real measurement and re-simulation in CarMaker. As a result, ScenarioRRR is a very precise tool for implementing real measurement into the simulation.

A test case is randomly generated from the database during the simulation phase. The CarMaker simulation platform configures the virtual vehicle with the same sensor parameters as the measurement vehicle. The trained model is deployed in Simulink to build a co-simulation with CarMaker. To make the traffic vehicle follow the previously recorded and post-transformed trajectory, the virtual vehicle is given the exact position in *X* and *Y* directions at each time

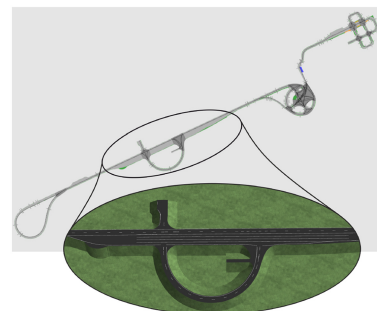


Fig. 6 Digital twin of Digitrans with detailed view

**Table 1** Offset between measurement data and simulation output

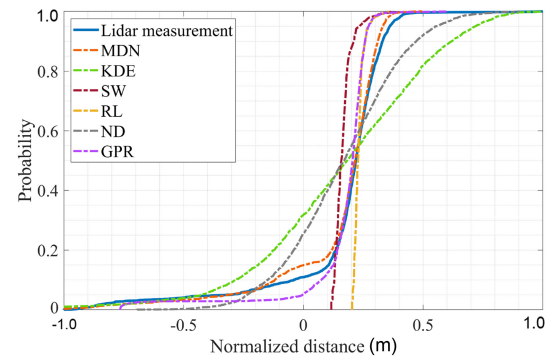
Parameter	Mean	Std
Heading error (rad)	4.9779e-04	9.8335e-04
Longitudinal distance (m)	-0.3042	0.0500
Lateral distance (m)	-0.1419	0.0259
Longitudinal velocity (m/s)	0.0099	0.0368
Lateral velocity (m/s)	-0.0019	0.0055

step. In this setup, the real measurement scenario can be play-back the simulation [3]. Ultimately, the corresponding sensor model is evaluated using CIPE metrics. The fidelity and performance of the different sensor models are evaluated and presented in Subsequent-sections 4.3 and 4.4.

### 4.3 Modelling Approaches

In order to compare the different approaches, several modelling approaches are selected to build sensor models using the same dataset. The modelling process is shown in Fig. 1 in Sect. 3. For the current example, the modelling target is the relative distance of the obstacle from the LiDAR. Therefore, during the data processing phase, some irrelevant information is removed (e.g., temperature, altitude, etc.), retaining the parameters that impact the vehicle dynamics and perception performance as much as possible. Once the modelling is complete, MGAE evaluates the model's overall performance to verify the model's goodness to the ground truth data. Some of the current mainstream ML methods have been chosen in creating the sensor model, which has received popularity in many studies. The focus of Refs. [9, 30] is to train a LiDAR point cloud model using Gaussian Process Regression (GPR). Li et al. [20] create a Radar model with the help of a Mixture Density Network (MDN) to present detection uncertainty. Meanwhile, Genser et al. [31] introduce a Kernel Density Estimation (KDE) modelling and focus on the camera's position measurement error. Other modelling methods, such as Robust Linear (RL) regression [32], are useful when expecting to repeat fitting a model multiple times in a loop. Moreover, the study of Ref. [33] introduces a sampling from a Normal Distribution (ND) which allows a radar model being deployed efficiently in a real-time system. Finally, although Stepwise (SW) regression [34] does not have any advantage in predicting new data, it has a low prediction error for large data sets. Hence, these different modelling approaches are compared as below.

At the MGAE level, we need to ensure that the model has been treated correctly and trained successfully. In this work, we prepared 2791 data to train the model and split the data set into a training set, a validation set and a test set according to the ratio of 70%, 15% and 15%, respectively. To illustrate the distribution of the statistics more clearly, the

**Fig. 7** Comparison of CDFs measured by different modelling methods and LiDAR

CDF is used here to intuitively compare the gap between the real LiDAR sensor measurement data (target data) and the predicted data from the other models. From Fig 7, we can observe that the MDN is closer to the target data. However, the intuitive conclusion is insufficient to prove the generalizability of the model. Therefore, we continue to calculate the final generalisation evaluation from Eqs. (1)–(3). The final statistics MGAE are presented in Table 2.

As introduced in Sect. 4.2, the digital twin-based simulation can be replicated in CarMaker. Namely, re-simulation is a prerequisite for CIPE. All trajectory points of the ego and target vehicle are reproduced in the simulation with high accuracy when the corresponding test is carried out on the test track. Moreover, a CarMaker co-simulation system for use with Simulink is built to allow all models to be deployed. To intuitively visualise the simulation-to-reality gap, the predicted results of all models are compared with the real LiDAR measurements in the simulation shown in Fig. 8. The boxplot indicates the accuracy of the different models and the range of uncertainty fluctuations. In light of the current simulation results, the MDN remains much closer to the actual performance of LiDAR, and the error fluctuations are relatively less. While, SW, KDE and GPR perform not differently and within acceptable error ranges. However, RL and ND have a larger error fluctuation range, which indicating that the model predictions are unstable and more random. Further analysis of the simulation results can be calculated by using Eqs. (4)–(6) and the detailed conclusions are shown in Table 2. The  $C_{pc}$  values for all models show a strong positive correlation. However, giving an accurate assessment based on one evaluation metric alone is difficult. Therefore, the introducing  $D_c$  and  $C_{cc}$  allows a more comprehensive evaluation of the simulation results. By comparison, it is easy to see that MDN is superior in terms of correlation and distance error.

Finally, we collate the MGAE and CIPE results in Table 3. Also, the RV is calculated by Eqs. (7)–(11), expressed in Table 4. By comparing the scores of each

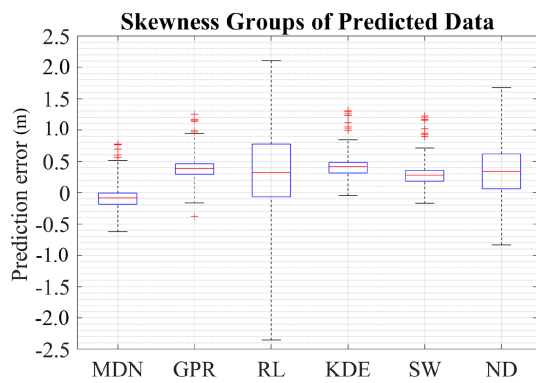


Fig. 8 Comparison of prediction errors in simulations

provides a very good theoretical basis for evaluating the generalizability of models. Meanwhile, CDF plots can be plotted for the goodness of fit after training. However, we believe this is insufficient, as the final model will be applied to the simulation to emulate the detection behaviour of the sensor. As a result, CIPE can draw more accurate conclusions about the effects of the simulation. The most crucial aspect is replicating the measurement scenario in a virtual environment. The real vehicle trajectories, kinematic information and road network logic are all implemented with the help of the CarMaker add-on. Accordingly, different models can be further validated in the same scenario, and the RV summarizes the overall performance eventually. By comparison, MDN demonstrates

Table 2 Performance evaluation summary for non-parametric-based approaches

Approach	MGAE			CIPE		
	$D_{ws}$	$D_K$	$D_{area}$	$D_c$	$C_{pc}$	$C_{cc}$
MDN	0.0212	0.0450	0.0197	0.7748	0.9988	0.4328
KDE	0.1818	0.2856	0.1813	2.3527	0.9988	0.3160
SW	0.1223	0.4937	0.1121	1.2243	0.9959	0.3181
RL	0.1112	0.4242	0.1793	2.3527	0.9978	0.2794
ND	0.1147	0.2474	0.1067	1.6774	0.9988	0.3087
GPR	0.0547	0.2280	0.0506	1.2489	0.9988	0.2906

Table 3 Six attributes for sensor models

	MDN	KDE	SW	RL	ND	GPR
$D_{ws}$	0.0212	0.1818	0.1223	0.1112	0.1147	0.0547
$D_K$	0.0450	0.2856	0.4937	0.4242	0.2474	0.2280
$D_{area}$	0.0197	0.1813	0.1121	0.1793	0.1067	0.0506
$D_c$	0.7748	2.3527	1.2243	2.3527	1.6774	1.2489
$C_{pc}$	0.9988	0.9988	0.9959	0.9978	0.9988	0.9988
$D_{cc}$	0.4328	0.3160	0.3181	0.2794	0.3087	0.2906

Table 4 Ranking vector for sensor models

	MDN	KDE	SW	RL	ND	GPR
RV	0.2850	0.1803	0.1161	0.0851	0.1497	0.1839

model, MDN achieves the best overall rating. Besides, both GPR and KDE have comparable performance, which is the reason why many sensor models in the state-of-the-art have been built based on these two approaches. However, SW, RL and ND models perform weakly and are not recommended.

### 4.4 Discussion

A multi-level model evaluation framework can be applied to the evaluation of data-driven models, as the purpose of the model is to fit the distribution of the reference data via a specific probability distribution function. MGAE

the performance since this model takes into account the influence of the input on the output, and the multi-modal output can cover more extreme cases. Although KDE and GPR show comparable performance, KDE can be a histogram statistic for all data and outputs a generic distribution function. Hence, this approach ignores the influence of relative distances on the model objectives. In addition, the improvement of the prediction performance by GPR is significant because different inputs affect the probabilistic prediction of the model, thus improving the results of MGAE. Similarly to the KDE, the ND outputs a fixed

distribution function and therefore has large error fluctuations in the simulation. While, SW and RL models are not suitable as sensor modelling methods.

## 5 Conclusion

Currently, virtual sensor perception is always a basis in ADAS simulation. However, existing evaluation methods lack uniformity as they focus on specific modelling methods or models. Additionally, a single metric could make the evaluation results incomprehensive. To overcome these shortcomings, we propose a multi-level and metrics approach for estimating performance rankings based on pairwise comparisons. This approach can comprehensively consider different performance metrics and give a complete evaluation result. The RV method is homogeneous, invariant, monotonic, and decisively. Moreover, it can enrich the evaluation system by introducing more evaluation metrics based on expertise, experience or understanding of performance measures.

To investigate the validity of the proposed approach, we have implemented and evaluated a LiDAR longitudinal distance detection model based on different modelling approaches. The results have shown that by introducing different evaluation metrics and perspectives (MGAE and CIPE). The fidelities of the sensor models under different modelling approaches are compared, and it is possible to make existing biases visible in detail by using a boxplot and CDF plot. This objective and quantitative evaluation avoids subjectivity based on experience and the incomprehensiveness of a few metrics.

**Funding** Open access funding provided by Graz University of Technology. Funding by the Graz University of Technology. This activity is part of the research project InVADE (FFG nr. 889349) and has received funding from the program Mobility of the Future, operated by the Austrian research funding agency FFG. Mobility of the Future is a mission-oriented research and development program to help Austria create a transport system designed to meet future mobility and social challenges. This paper is supported by TU Graz Open Access Publishing Fund.

## Declarations

**Conflicts of interest** The authors declare no conflicts of interest.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will

need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

1. Marti, E., De Miguel, M.A., Garcia, F., Perez, J.: A review of sensor technologies for perception in automated driving. *IEEE Intell. Transp. Syst. Mag.* **11**(4), 94–108 (2019)
2. Amersbach, C., Winner, H.: Defining required and feasible test coverage for scenario-based validation of highly automated vehicles. In: 2019 IEEE Intelligent Transportation Systems Conference (ITSC), pp. 425–430. IEEE (2019)
3. 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. *Energies* **15**(7), 2545 (2022)
4. Schlager, B., Muckenhuber, S., Schmidt, S., Holzer, H., Rott, R., Maier, F.M., Saad, K., Kirchengast, M., Stettinger, G., Watzenig, D., et al.: State-of-the-art sensor models for virtual testing of advanced driver assistance systems/autonomous driving functions. *SAE Int. J. Connect. Autom. Veh.* **3**, 233–261 (2020)
5. Roth, E., Dirndorfer, T., Neumann-Cosel, K.V., Fischer, M.-O., Ganslmeier, T., Kern, A., Knoll, A.: Analysis and validation of perception sensor models in an integrated vehicle and environment simulation. In: Proceedings of the 22nd Enhanced Safety of Vehicles Conference (2011)
6. Chipengo, U., Sligar, A., Carpenter, S.: High fidelity physics simulation of 128 channel MIMO sensor for 77ghz automotive radar. *IEEE Access* **8**, 160643–160652 (2020)
7. Holder, M., Rosenberger, P., Winner, H., D'hondt, T., Makkapati, V.P., Maier, M., Schreiber, H., Magosi, Z., Slavik, Z., Bringmann, O. et al.: Measurements revealing challenges in radar sensor modeling for virtual validation of autonomous driving, In: 2018 21st International Conference on Intelligent Transportation Systems (ITSC), pp. 2616–2622. IEEE (2018)
8. Holder, M.F., Thielmann, J.R., Rosenberger, P., Linnhoff, C., Winner, H.: How to evaluate synthetic radar data? Lessons learned from finding driveable space in virtual environments. In: 13. UNIDAS eV Workshop Fahrerassistenz und automatisiertes Fahren (2020)
9. Yang, T., Li, Y., Ruichek, Y., Yan, Z.: Performance modeling a near-infrared TOF lidar under fog: A data-driven approach. *IEEE Transactions on Intelligent Transportation Systems* (2021)
10. Bernsteiner, S., Magosi, Z., Lindvai-Soos, D., Eichberger, A.: Radar sensor model for the virtual development process. *ATZelektronik Worldw.* **10**(2), 46–52 (2015)
11. Ngo, A., Bauer, M.P., Resch, M.: A multi-layered approach for measuring the simulation-to-reality gap of radar perception for autonomous driving. In: 2021 IEEE International Intelligent Transportation Systems Conference (ITSC), pp. 4008–4014. IEEE (2021)
12. Schaermann, A., Rauch, A., Hirsenkorn, N., Hanke, T., Rasshofer, R., Biebl, E.: Validation of vehicle environment sensor models. In: 2017 IEEE Intelligent Vehicles Symposium (IV), pp. 405–411. IEEE (2017)
13. Hanke, T., Schaermann, A., Geiger, M., Weiler, K., Hirsenkorn, N., Rauch, A., Schneider, S.-A., Biebl, E.: Generation and validation of virtual point cloud data for automated driving systems, In: 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC), pp. 1–6. IEEE (2017)
14. Hirsenkorn, N., Subkowski, P., Hanke, T., Schaermann, A., Rauch, A., Rasshofer, R., Biebl, E.: A ray launching approach for

- modeling an FMCW radar system. In: 2017 18th International Radar Symposium (IRS), pp. 1–10. IEEE (2017)
15. Schneider, S.-A., Saad, K.: Camera behavior models for ADAS and ad functions with open simulation interface and functional mockup interface. *Center Model-Based Cyber-Phys. Product Dev.* **20**(12), 19–19 (2018)
  16. Li, H., Tarik, K., Arefnezhad, S., Magosi, Z.F., Wellershaus, C., Babic, D., Babic, D., Tihanyi, V., Eichberger, A., Baunach, M.C.: Phenomenological modelling of camera performance for road marking detection. *Energies* **15**(1), 194 (2021)
  17. Hirsenkorn, N., Hanke, T., Rauch, A., Dehlink, B., Rasshofer, R., Biebl, E.: A non-parametric approach for modeling sensor behavior. In: 2015 16th International Radar Symposium (IRS), pp. 131–136. IEEE (2015)
  18. Solomatine, D., See, L.M., Abrahart, R.: Data-driven modelling: concepts, approaches and experiences. *Practical Hydroinformatics* 17–30 (2009)
  19. Magosi, Z.F., Li, H., Rosenberger, P., Wan, L., Eichberger, A.: A survey on modelling of automotive radar sensors for virtual test and validation of automated driving. *Sensors* **22**(15), 5693 (2022)
  20. Li, H., Kanuric, T., Eichberger, A.: Automotive radar modeling for virtual simulation based on mixture density network. *IEEE Sens. J.* (2022)
  21. Rubner, Y., Tomasi, C., Guibas, L.J.: The earth mover's distance as a metric for image retrieval. *Int. J. Comput. Vis.* **40**(2), 99–121 (2000)
  22. Andrews, D.F., Mallows, C.L.: Scale mixtures of normal distributions. *J. R. Stat. Soc. Ser. B (Methodol.)* **36**(1), 99–102 (1974)
  23. Ferson, S., Oberkampf, W.L., Ginzburg, L.: Model validation and predictive capability for the thermal challenge problem. *Comput. Methods Appl. Mech. Eng.* **197**(29–32), 2408–2430 (2008)
  24. Rosenberger, P.: Metrics for specification, validation, and uncertainty prediction for credibility in simulation of active perception sensor systems
  25. Gultom, S., Sriadhi, S., Martiano, M., Simarmata, J.: Comparison analysis of k-means and k-medoid with euclidian distance algorithm, chaberra distance, and chebyshev distance for big data clustering. In: *IOP Conference Series: Materials Science and Engineering*, vol. 420, pp. 012092. IOP Publishing (2018)
  26. Yoo, J.-C., Han, T.H.: Fast normalized cross-correlation. *Circuits Syst. Signal Process.* **28**(6), 819–843 (2009)
  27. Yin, H., Li, X.R., Lan, J.: Pairwise comparison based ranking vector approach to estimation performance ranking. *IEEE Trans. Syst. Man Cybern. Syst.* **48**(6), 942–953 (2017)
  28. Bapat, R., Raghavan, T.: *Nonnegative Matrices and Applications*. Cambridge University Press (1997)
  29. IPG CarMaker, Reference manual (v 8.1.1). In: *IPG Automotive GmbH* (2019)
  30. Li, Y., Duthon, P., Colomb, M., Ibanez-Guzman, J.: What happens for a ToF LiDAR in fog? *IEEE Trans. Intell. Transp. Syst.* **22**(11), 6670–6681 (2020)
  31. Genser, S., Muckenhuber, S., Solmaz, S., Reckenzaun, J.: Development and experimental validation of an intelligent camera model for automated driving. *Sensors* **21**(22), 7583 (2021)
  32. Dumouchel, W., O'Brien, F. et al.: Integrating a robust option into a multiple regression computing environment. In: *Computer Science and Statistics: Proceedings of the 21st Symposium on the Interface*, pp. 297–302. American Statistical Association Alexandria (1989)
  33. Hirsenkorn, N., Hanke, T., Rauch, A., Dehlink, B., Rasshofer, R., Biebl, E.: Virtual sensor models for real-time applications. *Adv. Radio Sci.* **14**(B), 31–37 (2016)
  34. Schulz-Stellenfleth, J., König, T., Lehner, S.: An empirical approach for the retrieval of integral ocean wave parameters from synthetic aperture radar data. *J. Geophys. Res. Oceans* **112**(C3), (2007)