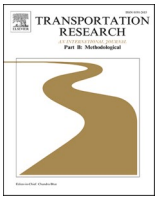
Transportation Research Part B 211 (2026) 103466

Contents lists available at ScienceDirect

Transportation Research Part B

journal homepage: www.elsevier.com/locate/trb



How do human factors affect the explanatory power of driver models? A methodology for comparative model assessment and validation

Vincenzo Punzo [a,*], Davide Iannelli [a], Andrea Saltelli [b,c], Marcello Montanino [a]

[a] Department of Civil, Environmental and Architectural Engineering, University of Naples Federico II, Via Claudio 21, 80123 Naples, Italy
[b] Barcelona School of Management, Pompeu Fabra University, Barcelona, Spain
[c] Centre for the Study of the Sciences and the Humanities, University of Bergen, Bergen, Norway

ARTICLE INFO

Keywords:
Driver models
Human factors
Adaptive driving
Verisimilitude
Calibration
Global sensitivity analysis
Validation

ABSTRACT

The evaluation of theories and models is key to the progress of science. This paper shows that current methodologies for assessing and comparing driver models – typically based on error distributions – are intrinsically flawed. A comprehensive methodology is proposed to evaluate driver models under both nominal and safety-critical driving conditions. Rooted in Popper's epistemology, the methodology relies on the execution of "risky tests" and on model "verisimilitude", i.e., the idea of a degree of better or worse correspondence to truth. The approach involves extensive testing of models via pair-wise calibration against individual trajectories and variance-based sensitivity analysis, enabling a systematic examination of how the trade-off between model accuracy and uncertainty evolves with increasing model complexity. The methodology is applied to 800 variants of the Intelligent Driver Model (IDM) and its improved formulations, augmented with human factors (HF) layers – namely perception errors and delays, temporal and spatial anticipation, and adaptive driving behaviours. A full factorial design isolates the explanatory contribution of each HF addon and its interaction effects. A novel formulation of the IDM, the M-IDM, is introduced and consistently outperforms both the original and all tested improved formulations across three diverse naturalistic trajectory datasets. The study, therefore, provides a falsifiability-oriented foundation for rigorous, transparent comparison and validation of HF-augmented driver models.

1. Introduction

1.1. Background and motivation

The safe and efficient coexistence of automated vehicles (AVs) and human-driven vehicles (HVs) poses profound interdisciplinary challenges to transportation science. As the virtual simulation of such mixed traffic has become a pillar in the road safety assessment of automated driving functions, credible modelling of human driving behaviour is at stake (see e.g., i4Driving, 2022, UN R157, 2023).

Credibility stems from the ability to capture human driving behaviour, dynamics, and interactions across different spatial and temporal scales in a way that is quantitatively corroborated by evidence. At the microscopic scale, models should reproduce realistic

* Corresponding author.
E-mail address: vinpunzo@unina.it (V. Punzo).

https://doi.org/10.1016/j.trb.2026.103466
Received 1 February 2026; Accepted 30 March 2026
Available online 28 May 2026

tactical manoeuvres and vehicle trajectories and plausibly describe behavioural mechanisms, such as driver adaptation to varying conditions. In the models, these mechanisms would eventually lead to inefficient or unsafe behaviour with a frequency of occurrence similar in magnitude to the real-world. At the macroscopic scale, realistic space-time congested patterns must emerge from the modelled microscopic dynamics.

Existing models in traffic flow theory, however, are inadequate to address these challenges. Most of the car-following models, for instance, assume perfect human perception and decision-making processes, and are inherently incapable of triggering safety-critical driving behaviours. To overcome these structural model limitations, several attempts have been made to embed Human Factors (HF) into the so-called "engineering" human driver models (Saifuzzaman and Zheng, 2014). Starting from theories about human perception and cognition in the context of driving behaviour (e.g., Endsley, 2000; Fuller, 2000, 2005, 2011; Engström et al., 2018; Markkula et al., 2018), existing car-following models were augmented in various ways, including the addition of perception delays and errors (e.g., Treiber et al., 2006; van Lint and Calvert, 2018), spatial and temporal anticipation (e.g., Treiber et al., 2006, Ngoduy, 2015) and adaptive driving behaviours (e.g., Saifuzzaman et al., 2015; Tian et al., 2016; van Lint and Calvert, 2018; Zheng et al., 2022, Durrani and Lee, 2024; Engström et al., 2024).

In the above-mentioned studies, most of the HF-augmented models were developed on the basis of sound theories of human behaviour. The outcomes of these models generally aligned with theoretical expectations in illustrative scenarios, typically under specific combinations of modelling assumptions and parameter values. Sometimes, models were calibrated against trajectory data, showing better performance than their original counterparts (see, e.g., Saifuzzaman et al., 2015).

All that said, can we consider the provided evidence sufficient to regard such HF-augmented models as credible in quantitative, high-stakes applications? And to what extent do these models elevate the explanatory power of their predecessors? Questions that lead back to a fundamental concern: are the methodologies currently applied to evaluate and compare models trustworthy?

The thesis of this study, grounded in Popper's epistemology (Popper, 1963), argues that little can be said about the credibility of most existing models. This is due in part to the nature of the models themselves and, in part, to the ways in which they have been developed and assessed – using weak evaluation methods that favour confirmation instead of falsification.

According to Popper: "*the method of science is criticism, i.e. attempted falsifications,* […] *the method of trial and error – conjectures and refutations: of boldly proposing theories; of try our best to show that these are erroneous, and of accept them tentatively, if our critical efforts are unsuccessful*". The rationale of this method is that "*only the falsity of the theory can be inferred from empirical evidence"* while no scientific theory can ever be proven to be true ("*no amount of experimentation can ever prove me right; a single experiment can prove me wrong"*, Einstein's letter to Bohr, 1926).

The falsification of a theory, then, consists in a genuine attempt to design a *risky test*. A test in which, if observation shows that the effect predicted by the theory is absent, the theory is simply refuted (an example of risky test by Popper is Eddington's expedition, aimed at measuring the gravitational deflection of light passing near the sun, which, according to Einstein's general relativity should have been twice than that predicted by Newton's gravitational law). Severe tests, however, are not always easy to design, since some theories are less exposed than others to refutation, as "*they take lower risks*" by providing less precise assertions and then lower "*empirical content"* – that is an informative content which can be verified or falsified by empirical observations. It is the case of social or behavioural science for which designing a severe test is much harder than e.g. for general relativity.

As traffic flow theory is an uncommon synthesis of behavioural science and physics, designing a severe test is inherently challenging. In addition, empirical evidence on driving behaviour offers a level of information that is significantly lower than what would be required by the complexity of human perceptual and cognitive processes. Consequently, these theories can often accommodate some empirical observations with relative ease – a situation reminiscent of Popper's critique of psychological theories by Adler, which he argued were unfalsifiable because they could explain virtually any observation. In this context, empirical tests rarely serve to decisively refute a theory but can readily provide apparent confirmation – particularly to those inclined to seek it.

While the nature of the theory and the limitations of empirical data pose one part of the challenge, the other lies in the methodologies used to evaluate models against such data. For instance, when vehicle trajectories are used to evaluate the accuracy of car-following models, the distribution of model errors across a set of trajectories (or simply their average) is commonly used to assess model performance and rank models. This approach, however, is fundamentally flawed. A visual inspection of Fig. 1, which compares calibration errors of five car-following models across three different trajectory datasets, would suggest that the models perform very similarly. In reality, as will be demonstrated in the following sections, a meaningful ranking of the models exists.

The reason the results presented in the figure are misleading is straightforward. Model performance varies significantly across individual trajectories; that is, each model tends to reproduce certain trajectories more accurately than others. However, the specific trajectories that are well captured by one model are not necessarily the same as those that are best reproduced by another. As a result, when aggregating all errors across trajectories and comparing overall error distributions important differences in model performance are obscured.

To complicate the picture and make model evaluation more challenging, substantial modelling uncertainty comes into play. Whenever the complexity of a model increases, so too does the uncertainty because of propagation errors. On the one hand, simple models are relatively easy to calibrate but may fail to capture critical mechanisms of the system. On the other hand, more complex and highly parameterised models may produce inferences that are too wide to be practically useful. The crux of the modelling challenge, therefore, lies in choosing the most appropriate level of complexity, which strikes a balance between model inadequacy and model output uncertainty according to O'Neill's conjecture (see Fig. 2; O'Neill, 1973; Turner and Gardner, 2015; Saltelli, 2019). In addition, whenever non-additive components are added to a model (i.e., uncertain factors whose effects are only felt in combinations – as human factors are expected to be) the contribution to the model explanatory power of each component remains unknown, if not investigated with appropriate techniques. Without applying these techniques, it becomes difficult to assess whether these additional components

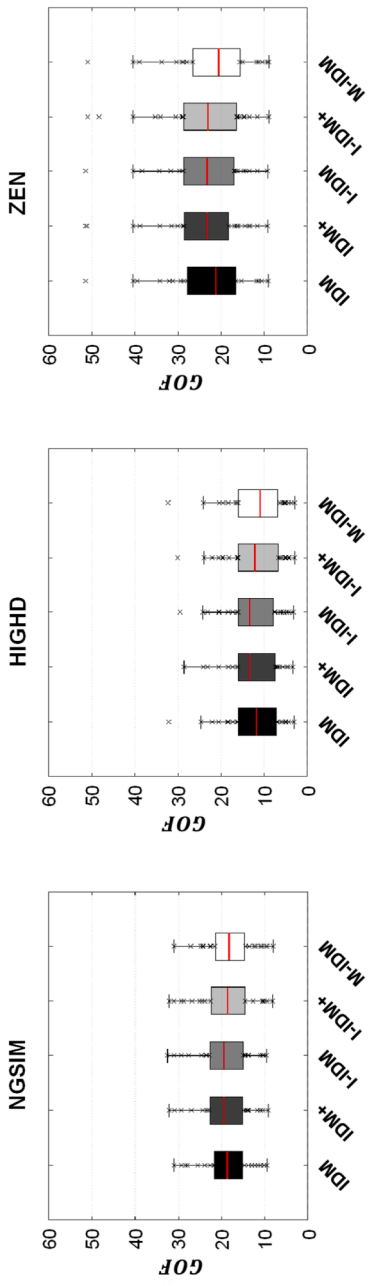


Fig. 1. Box plots of errors of five models across three trajectory datasets.

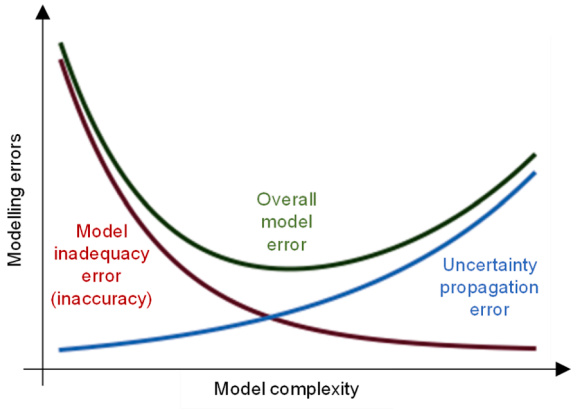


Fig. 2. Overall model error (green) as a combination of model inadequacy error (red) and uncertainty propagation error (blue). Source: Saltelli (2019).

meaningfully enhance the model or merely add complexity without improving explanatory power.

1.2. Contributions

To address the issues outlined in the Introduction, this study proposes a comprehensive methodology for the assessment and comparison of driver models, with particular emphasis on evaluating the impact of human factors addons on model explanatory power. The methodology itself constitutes a major contribution of this work and is built upon several original elements. The main methodological innovations and additional contributions are outlined below:

- To overcome the problem of indeterminacy in driver model comparisons based on error distributions, *pairwise calibration tests* are introduced.
- To ensure a fair and unbiased comparison of pairwise calibration tests across different trajectories or datasets, a *normalised relative error metric* is proposed.
- To draw a clear and meaningful ranking of driver model performance two *verisimilitude measures*, based on normalised relative errors from pairwise calibration tests, are defined.
- A novel *robustness metric* is introduced, reliant on variance-based Sobol's sensitivity indices, which quantifies the ability of a driver model to perform accurately regardless on the input trajectory i.e., independently of the specific driver styles and traffic conditions.
- An *evaluation framework for model prediction capability* is proposed, combining verisimilitude measures with Sobol's indices to captures the trade-off between model accuracy and uncertainty.
- Quasi-random Monte Carlo tests are developed to demonstrate a model's capability to produce collisions and to compute the corresponding collision rates.
- It is shown that calibrating a driver model to achieve high accuracy under both nominal and safety-critical conditions constitutes a Pareto optimisation problem.
- An *incremental augmentation framework* is introduced to systematically compare models and isolate the contribution of individual human factors components to the overall explanatory power of a model.
- *A new formulation of the IDM, the M-IDM*, is proposed. It is shown to outperform both the original IDM and several well-known improved formulations.
- The full methodology is demonstrated and discussed by applying it to *800 human-factors-augmented variants* of well-known improved IDM formulations. Results shed light on the overall driver modelling process and on the specific models and human factors addons.

2. Methodology

2.1. Verisimilitude measures

As outlined in the Introduction, within the framework of Popper's epistemology, theory evaluation relies on subjecting the model to *risky tests* – that is, tests which have the potential to falsify the theory if not passed. The higher the number of risky e.g., very precise affirmations or predictions a theory does and, therefore, of risky tests that can be built on those, the higher the theory *empirical content* – which consists in theory's predictions that can be falsified through observation (e.g., the statement "tomorrow either will rain or not", that is a tautology, has very low empirical content).

In the context of car-following modelling, a number of tests can be thought. For instance, a plausible model should satisfy fundamental rational driving constraints (Wilson, 2008), exhibit a non-decreasing speed-spacing relationship (except potentially at

high speeds), and capture driver relaxation behaviour following a disturbance.

Although these tests establish essential requirements for the plausibility of a model, they do not provide very precise assertions (for instance, a specific mathematical function for the speed-spacing function) and therefore are *not risky tests*. Any reasonably designed model will survive these tests, yet an analyst would remain blind about its explanatory and predictive power. Under such premises, any effective comparison among models and, consequently, any scientific advancement would be precluded. Luckily, more challenging tests are available, which consist in comparing vehicle trajectories predicted by a model with those observed. In the theoretical framework presented so far, these tests would apparently imply refuting every model, as there is clearly no model able to exactly reproduce a trajectory. However, Popper's theory combines the ideas of truth and of content into one, that is "*the idea of a degree of better (or worse) correspondence to truth*" or **verisimilitude**. It translates the idea that, though the absolute truth may never be fully attainable, we believe it is possible to get progressively closer to it and measure such progress. In this view, verisimilitude can be formally defined as the difference between the truth-content of a theory or a model *m, TC*(*m*) – the empirical content which is corroborated by observations – and its falsity-content, *FC*(*m*) – that contradicts observations:

$$VS(m) = TC(m) - FC(m) \tag{1}$$

According to this definition, a model m_j is closer to the truth than m_i, if it explains more phenomena and/or with higher accuracy than m_i, i.e., if and only if either:

- the truth-content but not the falsity-content of m_j exceeds that of m_i, or
- the falsity-content of m_i but not its true-content, exceeds that of m_j.

All that said, the distribution of residuals between predicted and observed trajectories would offer a straightforward way to gauge model proximity to truth, that is, to measure its degree of verisimilitude. However, as argued in the Introduction, error distributions are unsuitable to compare models, as errors obtained for different drivers and different traffic conditions mix and compensate for each other (see Fig. 1).

To overcome this limitation, we propose a different approach that relies on Popper's framework. First, we propose a ***pair-wise comparison*** of models on each single trajectory, that is every single trajectory becomes a test (in decision theory, pairwise comparison in a case of multi-attribute analysis is considered a best practice – see e.g. the method of Condorcet; Munda, G., 2008).

The corresponding basic statement is: the truth content of model m_i exceeds that of m_j on a trajectory k, if and only if the error of m_i is lower than that of m_j for the trajectory k, i.e.:

$$m_{ij}^k = 1 \Leftrightarrow \varepsilon_{m_i}^k < \varepsilon_{m_j}^k \tag{2}$$

In other words, $m_{ij}^k = 1$ means that model m_i is better than m_j on trajectory k. Conversely, if the error of m_i is greater than that of m_j its falsity content is greater, that is:

$$\neg m_{ij}^k = 1 \Leftrightarrow \varepsilon_{m_i}^k > \varepsilon_{m_j}^k \tag{3}$$

where $\neg m_{ij}^k$ is the negation of m_{ij}^k.

Given the heterogeneity of drivers and performances, this basic statement is not particularly useful to rank models. A much relevant information is when *a model is the best* among existing ones, that is when it consistently improves the existing modelling capabilities for a specific trajectory. Consequently, our truth statement becomes:

$$m_{i1}^k \ \wedge \ m_{i2}^k \wedge \cdots \wedge m_{iM}^k = 1 \Leftrightarrow \text{ model } m_i \text{ is the best among the } M \text{ models, where } \wedge \text{ is the logical 'and' symbol.} \tag{4}$$

Accordingly, the falsity statement is:

$$\neg m_{i1}^k \ \wedge \ \neg m_{i2}^k \wedge \ldots \wedge \neg m_{iM}^k = 1 \Leftrightarrow \text{ model } m_i \text{ is the worst.} \tag{5}$$

In the proposed approach, given a number of trajectories K, we test M models against every available trajectory k and, whenever the error of a model m_i, $\varepsilon_{m_i}^k$, is the lowest, m_i has passed the test. Therefore, the truth content of a model becomes the frequency of tests passed over a set of trajectories K, while its falsity content is the frequency of tests failed, i.e., when *model m_i is the worst* and degrades existing modelling capabilities:

$$TC(m_i|K) = \frac{1}{K}\sum_{k\in K}\left(m_{i1}^k \ \wedge \ m_{i2}^k \wedge \ldots \wedge m_{iM}^k\right) \tag{6}$$

$$FC(m_i|K) = \frac{1}{K}\sum_{k\in K}\left(\neg m_{i1}^k \ \wedge \ \neg m_{i2}^k \wedge \ldots \wedge \neg m_{iM}^k\right) \tag{7}$$

Recalling Eq. (1), a desired measure of model verisimilitude VS_1 is easily obtained for model m_i over a set of trajectories K:

$$VS_1(m_i|K) = \frac{1}{K}\sum_{k\in K}\left(m_{i1}^k \wedge \ m_{i2}^k \wedge \ldots \wedge m_{iM}^k\right) - \frac{1}{K}\sum_{k\in K}\left(\neg m_{i1}^k \ \wedge \ \neg m_{i2}^k \wedge \ldots \wedge \neg m_{iM}^k\right) \tag{8}$$

The higher the number of tests (trajectories) and the greater their heterogeneity, the more severe the whole test and the more meaningful the comparison, which in Popper's epistemology is intended as a struggle for the survival of the fittest theory. This measure, however, cannot tell *how much* a new model improves (or degrades) upon state-of-the-art models, being the basic statement in Eq. (3) a Boolean operator, and Eq. (8) expressing a logical probability (a similar critique can be found in Tichy, 1974). A second measure, VS_2, which explicitly accounts for the distance between models i.e., the difference in their errors, can therefore be adopted. Being $K^+(m_i) \subseteq K$ the subset of trajectories in which model m_i is the best, and $K^-(m_i) \subseteq K$, the one in which it is the worst (see Eqs. (4–5)), VS_2 can be expressed as:

$$VS_2(m_i|K) = \Delta^+_{m_i} - \left|\Delta^-_{m_i}\right| \tag{9}$$

where:

$$\Delta^+_{m_i} = \Delta(m_i|K^+) = \sum_{k\in K^+}\sum_{\substack{j\in M \\ j\neq i}}\left(\varepsilon^k_{m_i} - \varepsilon^k_{m_j}\right) \tag{10}$$

$$\Delta^-_{m_i} = \Delta(m_i|K^-) = \sum_{k\in K^-}\sum_{\substack{j\in M \\ j\neq i}}\left(\varepsilon^k_{m_i} - \varepsilon^k_{m_j}\right) \tag{11}$$

are the sums of the error differences between model m_i and all other models, for the trajectories in $K^+(m_i)$ and $K^-(m_i)$, respectively. When a model is neither the best nor the worst no error difference is calculated.

2.2. Error measures normalisation

The error measure $\varepsilon^k_{m_i}$ in Eqs. (10–11) must satisfy some requirements to be meaningful for model comparison and ranking across *a set of trajectories*:

a) it should not be an absolute error value, as errors sensibly vary among trajectories, so that using the absolute value would give extremely heterogeneous weights to different trajectories, e.g., the error of one trajectory could weigh as much as the sum of all other errors. Consequently, it is useful to anchor the error to the minimum error among models on a specific trajectory.
b) It should not be the percentage error variation relative to the minimum for a given trajectory, as it would overweigh trajectories whose minimum error is very low. A tiny improvement given by a model against the others in a trajectory which is quite easily reproducible (i.e., with a low minimum error), would count much more than a sensible improvement in a trajectory which is hardly captured by all models. To avoid this distortion, the error variation can be normalised over an interval.
c) The normalisation interval should be robust to outliers. Therefore, the min-max interval should be avoided, as a model that performs poorly on a specific trajectory (i.e., causing a high maximum error) would under-weigh the error distances between all other models. The min-median interval can be used instead, though it does not provide values in the range [0, 1].

Basing on the previous considerations a suitable normalised error metric for model m_i on a trajectory k is given by:

$$\varepsilon^k_{m_i} = \Delta GoF^k_{m_i,norm} = \frac{\Delta GoF^k_{m_i}}{\underset{j\in M}{\text{med}} GoF^k_{m_j} - \underset{j\in M}{min} GoF^k_{m_j}} \tag{12}$$

where $GoF^k_{m_i}$ is the chosen error statistic, e.g., *RMSE, NRMSE*, etc. (see Section 3.3) and $\Delta GoF^k_{m_i}$ is the difference between the model error and the minimum error, i.e.:

$$\Delta GoF^k_{m_i} = GoF^k_{m_i} - \underset{j\in M}{min} GoF^k_{m_j} \tag{13}$$

$\Delta GoF^k_{m_i,norm}$ compares the relative error of model m_i to the minimum, with the relative error of the median model to the minimum. A value greater than one identifies a model that performs worse than the median one, while a value lower than one a model which performs better (a zero-value meaning that the model is the best-performing).

As the normalised metric in Eq. (12) is fair among trajectories, it can be applied also to compare model performances across different trajectory datasets.

2.3. Methodology for selecting candidate models

In the previous sections a methodology to devise risky tests for car-following models has been proposed. A test relies on a pair-wise comparison of models against one single trajectory, and several of these tests are combined in a verisimilitude measure for a fair comparison and ranking of models. This methodology allows for a performance-based comparison of models, but leaves the analyst blind about the contribution of any individual model constituent to the performance of a model. Especially when human factors

theories are concerned, we do not only care about model comparison, but we would like to understand which of these perception and cognitive mechanisms are more relevant to the model's explanatory power. Furthermore, as many of these theories compete to describe similar behavioural mechanisms, we would like to discern the theories which are most effective. Eventually, we would like to understand the effect of the interaction of these model components on model performances.

For these reasons, we propose to apply the above methodology to a set of models of *incremental complexity*, built by progressively adding human-factors layers to base car-following models. Specifically, we apply a full factorial design to account for all interactions among base car-following models and competing state-of-the-art perception and cognitive theories. With this experimental design, it is possible to disentangle the contribution of each component (and of its interactions) while tracing the evolution of the trade-off between model accuracy and uncertainty. This approach led to an experimental design made of 800 model variants (see Section 3.1).

The model design phase represents the top block in the methodology workflow shown in Fig. 3. The final aim of the whole methodology is to select a set of *candidate models*, that is, models that satisfy the basic requirements and consistently outperform the other models. The proposed methodology relies on performance-based tests under nominal conditions, and on collision tests under safety-critical ones (see left and right branches in the "Model testing" box in Fig. 3).

If the performance-based approach relying on pairwise tests against trajectories and verisimilitude measures can be applied to nominal driving conditions, a different approach to testing models is needed for safety-critical ones. Therefore, as many base car-following models are structurally collision-free, their capability to produce rear-end collisions, once augmented with human factors, must be tested. Being such a capability a requirement for safety applications of human driver models, collision-free augmented models must be discarded (i.e., in the selection of candidate models). For the sake of the analysis, we propose to test models against a harsh braking manoeuvre of the leader vehicle at a constant deceleration rate of 9 m/s^2, from a steady state distance to a complete stop (this manoeuvre was shown to be the most dangerous for nonlinear string stability and rear-end collisions in Montanino et al., 2026). To guarantee a full coverage of the model input space and the significance of findings, a large quasi-random Monte Carlo simulation experiment for each model is needed. As this test is run for an infrequent manoeuvre (in real-world) and by sampling parameters from independent uniform distributions (an unrealistic but useful assumption just to test model collision capabilities) the resulting collision rates are not expected to be meaningful. Therefore, for models capable of producing collisions, a further quasi-Monte Carlo test is run to evaluate collision rates under reasonable parameter values – i.e., calibrated ones – and naturalistic driving conditions – i.e., feeding models with sampled naturalistic leader vehicle trajectories. In this test, to preserve correlation, parameter values are sampled from a Gaussian copula fitted to the calibrated parameter sets.

With regard to the selection of candidate models, it is crucial to evaluate each model's potential to overfit the data or, conversely, its parsimony. To this end, we rely on the concepts behind Fig. 2, that is, an evaluation of the trade-off between model inaccuracy and uncertainty.

Model *inaccuracy* can be quantified using calibration errors relative to naturalistic trajectories (see Section 3.3). As to simulation uncertainty, a more elaborate measure than the simulation error variance – which would be a straightforward proxy of simulation uncertainty – must be adopted. The unconditional error variance, in fact, is usually set to increase with the number of parameters. However, this variance increase does not necessarily imply a degradation of the model explanatory power (in a similar way as a good fitting after calibration does not necessarily imply an improvement of the model explanatory power, because of a potential data overfitting).

The consideration that the output error variance is a function of both model parameters and leader's input trajectory comes to our aid. In a sensitivity analysis setting in which both leader's trajectory and parameters are sampled, the separate contribution to the output error variance of these two types of factors can be quantified (see Punzo et al., 2015, for a detailed explanation of such a setting). Such contribution is measured by the "total sensitivity index", which in fact quantifies the portion of the output error variance explained by a factor, including both its first order effect and the effect of its interactions with all the other factors.

Ideally, we would like a model whose error variance is explained only by model parameters, that is a model in which the impact of parameters on the outputs does not change when the input trajectory changes. Such a model would be robust against the input leader's trajectory, i.e., it would be able to predict the follower trajectory with an accuracy which is independent on the leader's trajectory.

Therefore, we propose the total sensitivity index (Homma and Saltelli, 1996; Saltelli et al., 2010) of the input trajectory, computed following the algorithm in Saltelli et al., (2020), as the required proxy for the *robustness* of a model, that is, the ability to keep the model uncertainty bounded.[1] The lower this index, the higher the ability of a model to accurately capture heterogeneous driving behaviour. Notably, this proxy is independent of the number of model parameters (but depends on their influence). This is an advantage over existing criteria and methods, such as Bayesian information criterion, Akaike information criterion, likelihood ratio tests, or other Bayesian approaches (see van Hinsbergen et al., 2015), which penalise models proportionally to the number of parameters, independently of whether parameters are influential or not. These methods do not provide insight into the model's internal mechanisms. On the contrary, sensitivity analysis can reveal whether a module or a parameter, which is expected to be meaningful for a certain behaviour covered by the data applied in the analysis, is relevant for that behaviour, or can be simply neglected, simplifying or redesigning the model.

In summary, the overall methodology schematized in Fig. 3 relies heavily on numerical evaluation of models against observed data for both calibration and uncertainty/sensitivity analysis. Quantitative corroboration against observations is indeed essential to the process of crafting (and refuting) theories and models, as much as deductive reasoning. In truth, it is essential to deductive reasoning itself.

[1] While the index pertains to the trajectory, its values also depend on the interactions between trajectories and model parameters.

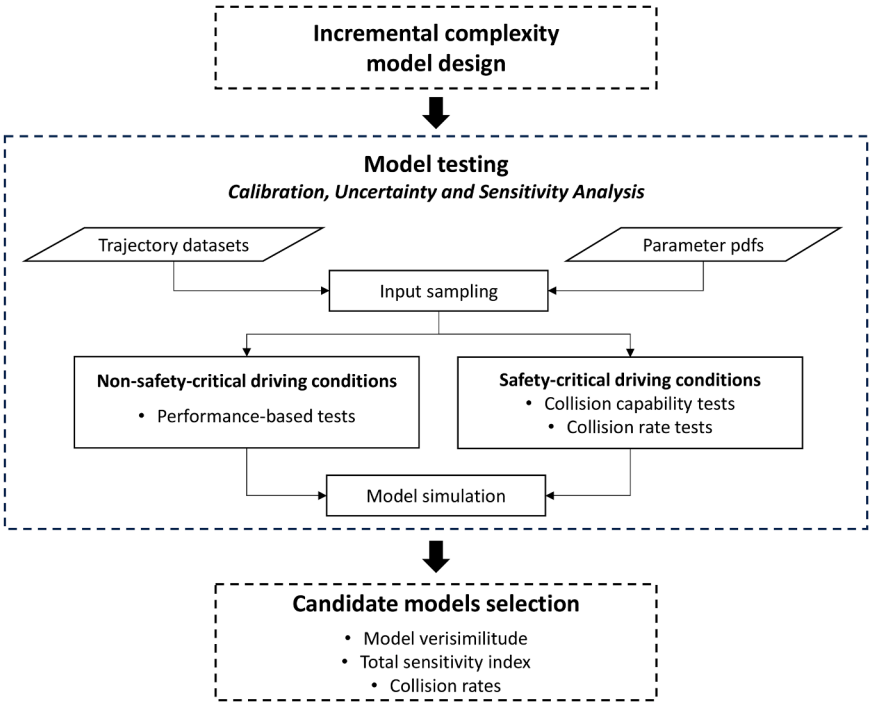


Fig. 3. Methodology workflow.

3. Application

3.1. Incremental complexity model design

Five base car-following models belonging to the family of the IDM were considered in this study: the IDM (Treiber et al., 2000), the IDM+ (Schakel et al., 2012), the improved IDM (I-IDM; Treiber and Kesting, 2013), a variation of the improved IDM (I-IDM+; Tian et al., 2016) and a new variant of the IDM proposed in this work (M-IDM), which combines formulations of the IDM and the I-IDM+.

Perception, anticipation, and adaptation layers, driven by HF theories, were incrementally added to the base models. In particular, perception capabilities were modelled in terms of perception delay and errors: *three perception-delay and four perception-error formulations* were considered. With regard to the perception delay: no delay, fixed perception delay (FPD) and time-varying perception delay (VPD) were modelled according to the formulation in van Lint and Calvert (2018). With regards to the perception errors: no error, fixed perception error for each stimulus (FPE), and two time-varying error formulations were modelled following van Lint and Calvert (2018) and Treiber et al. (2006), here referred to as VPE(FDTD) and VPE(HDM), respectively.

Two spatial anticipation and two temporal anticipation models were tested. Concerning the first: no anticipation and the multi-leader anticipation proposed in Ngoduy (2015) considering up to three preceding vehicles (MLTP) was included in the analysis. Concerning the temporal anticipation: no anticipation and the temporal anticipation proposed in Treiber et al. (2006) (ANT) was taken.

Five adaptive driving behaviour models were tested: no adaptive driving behaviour, the Task-Capability Interface model by Fuller (2000) in the variants by Saifuzzaman et al. (2015) and by van Lint and Calvert (2018) – here referred to as AD(TD) and AD(FDTD), respectively – and the 2D desired time-headway model proposed in Tian et al. (2016) and updated in Zheng et al. (2022) (2D+).

The modelling framework in which the human factors layers are combined, is shown in Fig. 4. At each time step the follower receives stimulus variables (gap, relative speed, ego speed) through the perception module (possibly delayed/noisy), applies the anticipation block to form an internal estimate of relevant quantities, computes the acceleration command via the car-following law augmented by cognition/adaption, and updates the kinematic state, which in turn changes the stimuli for the next time step.

A total of *800 car-following model variants* resulted from a full-factorial design of each base car-following model with alternative formulations of perception delay and error, spatial and temporal anticipation, and adaptive driving behaviour (i.e., 160 model variants were obtained for each base model[2]). The mathematical formulations of all tested model components are presented in Appendix A.

A graphical representation of the models design is shown in Fig. 5. The raster plot in the middle of the figure shows the activation pattern of the HF features listed on the y-axis. The activation pattern is the same for each base model (so that identifiers, per base

[2] Note that the temporal anticipation mechanism can be applied only in presence of a perception delay, which reduces the number of variants of each base model from 240 to 160.

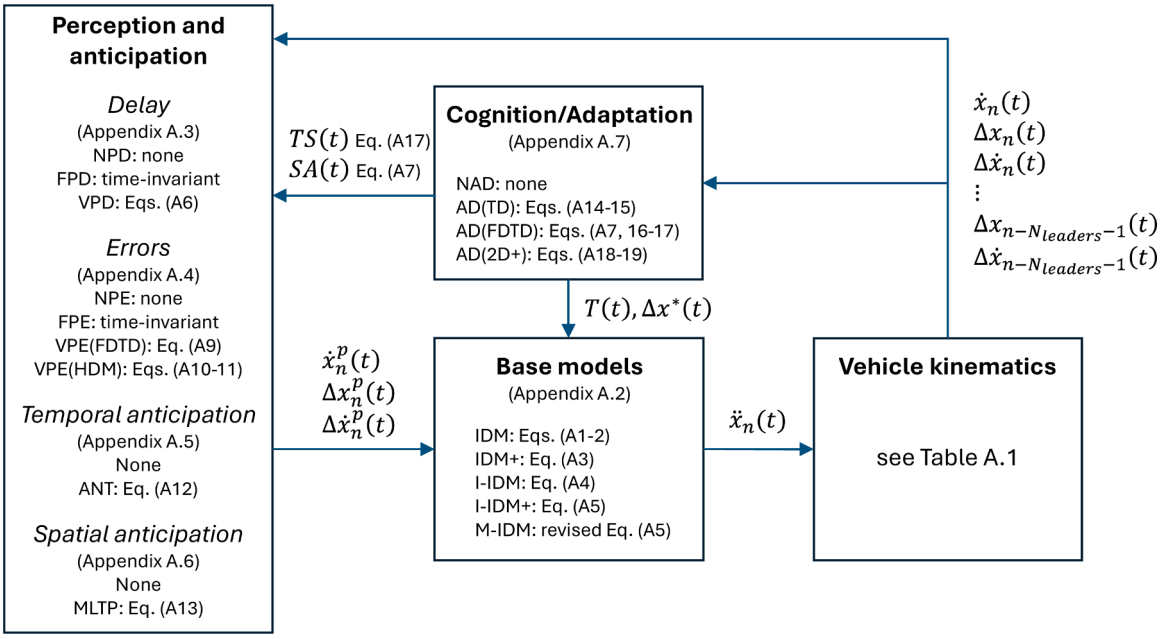


Fig. 4. Modelling framework.

model, are piled at the bottom). At the top of the figure, the number of parameters of each model variant is plotted. For the sake of recognisability, models are grouped by markers and colours. A group of models with the same colour share the same adaptive driving behaviour (blue for no adaptation, orange for TD, green for FDTD and violet for 2D+). Within the same colour, light and dark tones refer to models with and without spatial anticipation, respectively. Quintets of models with the same marker share the same perception error (circle for no error, star for FPE, square for VPE(FDTD) and triangle for VPE(HDM)). Within each quintet sharing a marker, models differ for the perception delay and the temporal anticipation (from left to right: no delay, FPD, FPD+ANT, VPD, VPD+ANT).

The model variants that include the VPE(HDM) or the AD(2D+) components are stochastic. To account for stochasticity, ten replications were performed for any run of Monte Carlo experiments or calibrations.

3.2. Vehicle trajectory data

Vehicle trajectories used in this study were drawn from three open-source datasets collected on three different continents: the "reconstructed NGSIM I80–1″ dataset from the USA (Montanino and Punzo, 2015), the highD dataset from Germany (Krajewski et al., 2018), and the Zen Traffic Data from Japan (Zen Traffic Data, 2024). For the purposes of this study, car-following trajectories were selected from each dataset according to the trajectory completeness requirements proposed by Sharma et al. (2019), complemented by additional criteria: only vehicles labelled as "car" were retained; trajectories shorter than 30 s were discarded; and any trajectory involving a lane change – either by the ego vehicle or by any of its three preceding vehicles – was excluded. Although these criteria necessarily reduce the variability of the retained trajectories, they are required to ensure parameter identifiability.

A total of 109 platoon trajectories satisfied these conditions and were selected for the analyses (35 from NGSIM, 40 from highD, and 34 from Zen Traffic Data). The selected trajectories are plotted in Fig. C1 in Appendix C.

3.3. Experimental setting

As mentioned in Section 2.1, the computation of model verisimilitude measures in pairwise tests is based on the evaluation of model errors against real trajectories. For the evaluation to be independent from parameter values, the minimum error of every model was calculated by calibrating the models against each selected trajectory.[3] 87,200 calibration experiments were executed (800 models x 109 trajectories) following the guidelines in Punzo et al. (2021). The sum of *Normalised Root Mean Square Error* (*NRMSE*) on spacing and speed was adopted as a goodness-of-fit function. For stochastic models, the 85th percentile of the NRMSE across 10 replications was chosen (we found the 85th percentile being more robust than the minimum, as suggested instead in Zhou et al., 2025). A Genetic Algorithm with a population size of 1000 was applied. The adopted population size was a compromise to balance computing time and solution robustness. The latter was tested by repeatedly calibrating the most parameterised model against the same trajectory. In the

[3] The short duration of publicly available naturalistic trajectories makes the split of a trajectory in two sufficient portions for model calibration and validation unfeasible. In fact, validation errors would be more informative than calibration ones in the proposed framework.

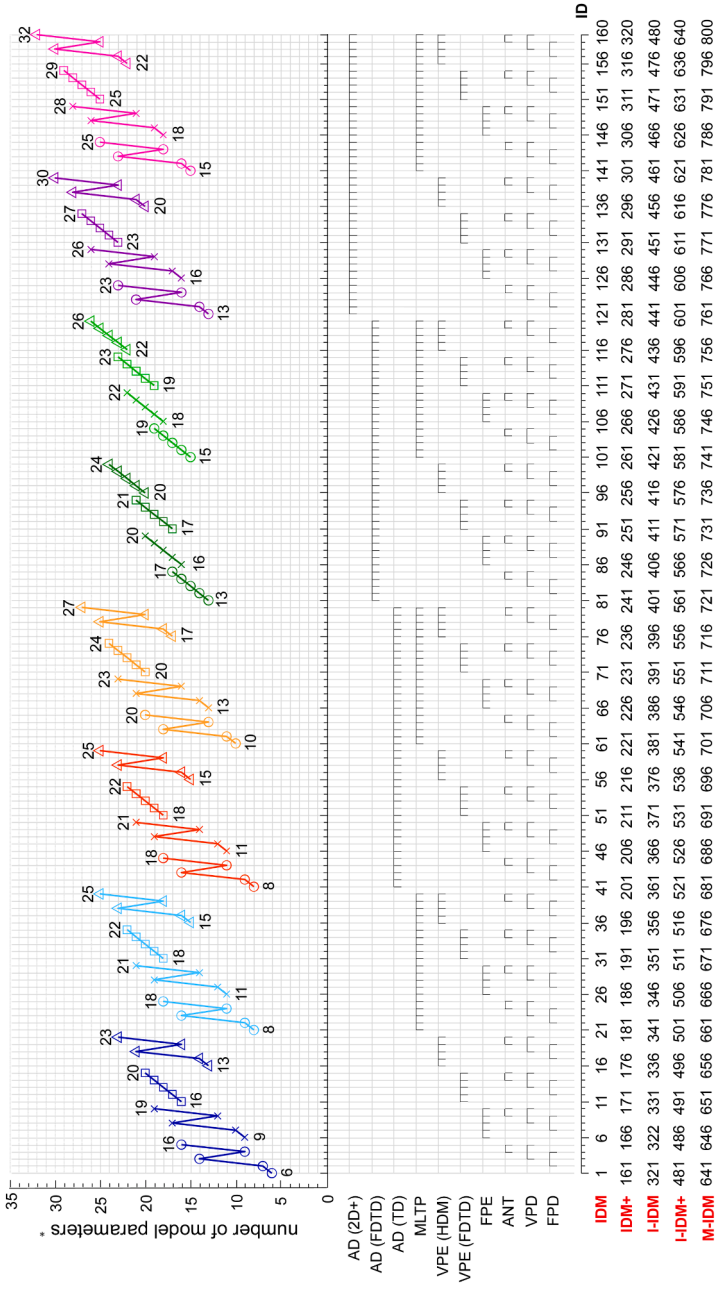


Fig. 5. Human factors activation pattern and degree of parameterization in the 800 model variants.

95% of cases, the algorithm found the same optimal solution, which was deemed an acceptable result given the uncertainties inherent in the whole evaluation framework. The parameter bounds used in the calibration are reported in Appendix A.8.

Concerning the tests to evaluate model collision capability and rates – i.e., in safety-critical conditions – two uncertainty propagation experiments were conducted for each model by means of a quasi-random design based on Sobol's low-discrepancy number sequences (Sobol', 1976). In the first experiment, aimed at verifying collision structural capability, each model was simulated in a deceleration scenario by sampling parameters from independent uniform pdfs (see Section 2.3). In the second experiment, the collision rate of each model was calculated after quasi-Monte Carlo simulations in which both the leader trajectory and the values of parameters were sampled – the latter, from a Gaussian copula fitted to calibrated parameters sets (see Section 2.3). The second experiment was run for each dataset. The size of the uncertainty propagation in both the experiments was $N = 2^{18} \cdot (P + 2)$, where P is the number of input factors i.e., the number of model parameters (plus one in the second experiment, as also the leader's trajectory is sampled). As a result, 17.9 billions of trajectory simulations were needed to cover the 3200 uncertainty propagations (i.e., 800 propagations for the first experiment, plus 2400 for the second one – 800 models x 3 datasets).

With regards to the study of the trade-off between accuracy and uncertainty – in nominal driving conditions – Sobol's sensitivity indices were calculated for all models after the uncertainty propagations for each trajectory dataset. Analysis factors were the input trajectory and the model parameters (as in Punzo et al., 2015), while the output was the same error statistic adopted for the calibrations i.e., the *NRMSE(s,v)*. The size of the quasi-Monte Carlo experiment was $N = 2^{18} \cdot (P + 2)$, as in propagations for the collision test, resulting in 13.6 billions of trajectory simulations.

4. Application results

4.1. Base models

In this section, we present the results of the pairwise tests against trajectories for the five base car-following models – reported as normalised model errors and the corresponding verisimilitude measures – together with the total sensitivity indices, which quantify the models' robustness to heterogeneous input trajectories (Section 4.1.1). These two strands of evidence are then jointly interpreted in Section 4.1.2 to discuss the trade-off between calibrated performance and robustness. The base models are the IDM (id 1), the IDM+ (id 161), the I-IDM (id 321), the I-IDM+ (id 481), the M-IDM (id 641).

4.1.1. Pairwise calibration tests and sensitivity analysis

In Fig. 6, the box plots of the relative calibration errors, ΔGoF, given by Eq. (13), and the normalized ones, ΔGoF_{norm}, by Eq. (12), are reported for each dataset. Contrary to Fig. 1 – where values of *GoF* were shown – **the differences among models are glaring**. A single model, the M-IDM, stands out from the plots as the best performing model.

The top row in the figure shows the box plot of the model error on a trajectory relative to the minimum on the same trajectory (i.e., ΔGoF), for each model and dataset. The M-IDM exhibits much lower relative errors than the other models, consistently across all datasets. The fact that its median relative error is zero in all datasets, means that it is the best performing model at least in 50% of all trajectories (this number is even higher, as the 75th percentile is close to zero, for all datasets).

As explained in Section 2.2, relative errors can introduce a bias when used to compare models across trajectories, though being very informative. Therefore, the bottom row in the same figure shows the box plots of the normalised relative errors, which provide a comparable quantitative performance measure across trajectories (and datasets). In addition, these errors are informative on the performance of a model relative to the median behaviour of all other models, which is consistent with the idea that a new model should improve the state-of-the-art of modelling. Findings about models' comparison are also confirmed by normalised errors.

Returning to the error box plots in Fig. 1, while they may not be useful for comparing models, they are valuable for comparing datasets. For instance, it can be observed that independently of the model, calibration errors for the trajectories in the highD dataset are lower, on average, than those in the other datasets, which would suggest a discrepancy in the quality of trajectories among the datasets. We rather suggest that this discrepancy is due to the difference in trajectory duration across the datasets (on average: NGSIM = 49.53 s, highD = 39.54 s, ZEN =131.08 s). Incidentally, the observed difference among datasets is another reason for adopting the normalised relative error in Section 2.2, as to have a fair comparison of models when using different datasets.

Fig. 7 details the results presented in the bottom row of Fig. 6. In the top diagrams, a black "x" represents how inaccurate a model is relative to the best performing model, on a specific trajectory. Red "x" are a subset of black ones representing errors when the model investigated (not only it is not the best, but) is the worst performing. Blue circles, instead, represent how much a model improves upon the second-best model (when it is the best), i.e., how much a model improves upon the state-of-the-art of modelling.

The M-MIDM results in the worst-performing model in only three trajectories, all within the highD dataset (3 out of 109 trajectories), and it is by far the model that most consistently improves upon the others. Among the remaining models, only one other model is never the worst in a dataset: the I-IDM+ in the highD dataset.

Pairwise tests of models against trajectories, based on the normalised relative errors, were used to compute the two verisimilitude measures introduced in Section 2.1 (Eqs. (8) and (9)). Table 1 presents the results. Core findings are consistent among datasets, though the highD one stands out for some peculiar results. Overall, the M-IDM outperforms the other models across all datasets, while the IDM and the I-IDM compete for being the worst performing.

While for *VS1*, these two models are equivalent overall, according to *VS2*, the IDM is by far the worst-performing model, especially given its very poor performance in the highD dataset. In this dataset, the IDM is, by far, both the most frequently worst model and the one with sensibly higher errors than the other models (measured by *VS2*).

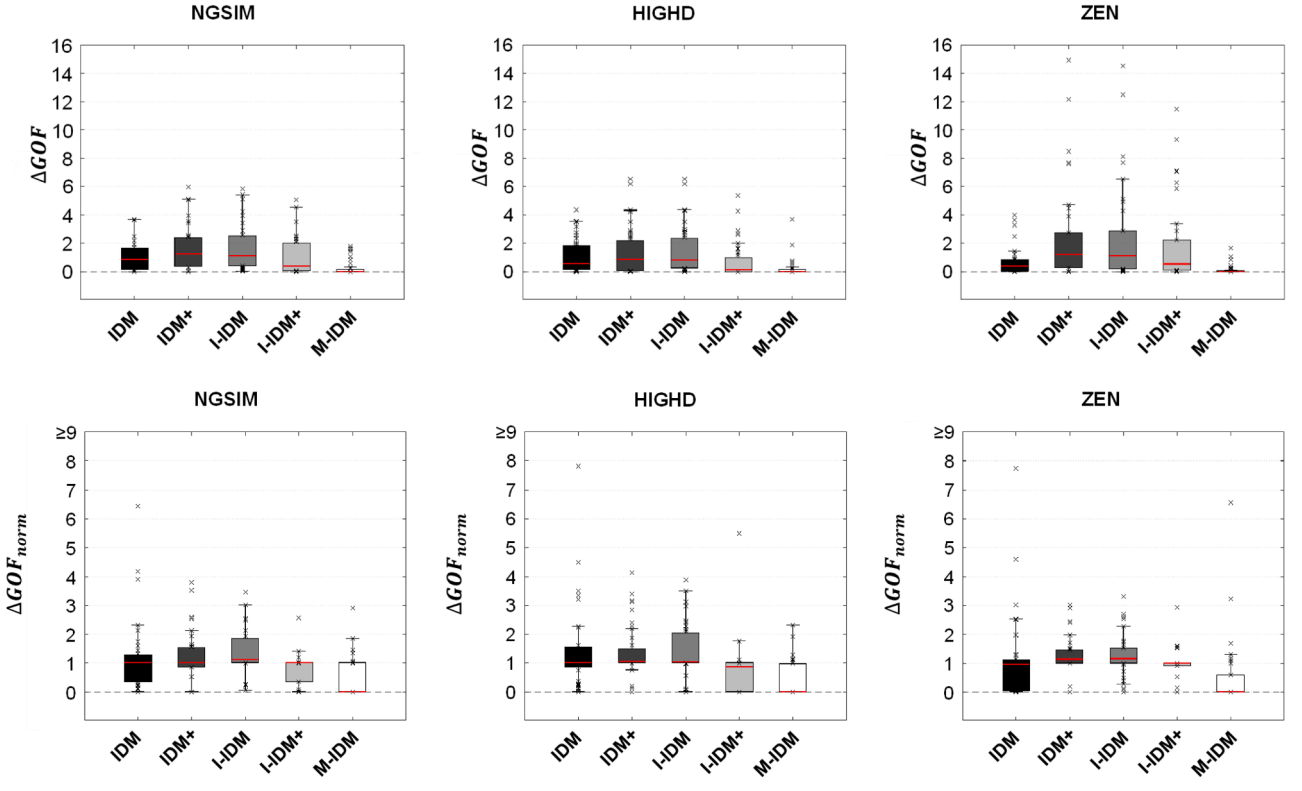


Fig. 6. Box plots of ΔGoF defined in Eq. (13) (top row) and ΔGoF_{norm} defined in Eq. (12) (bottom row) across the trajectories from NGSIM (left), HIGHD (middle) and ZEN (right) datasets, for the five base models.

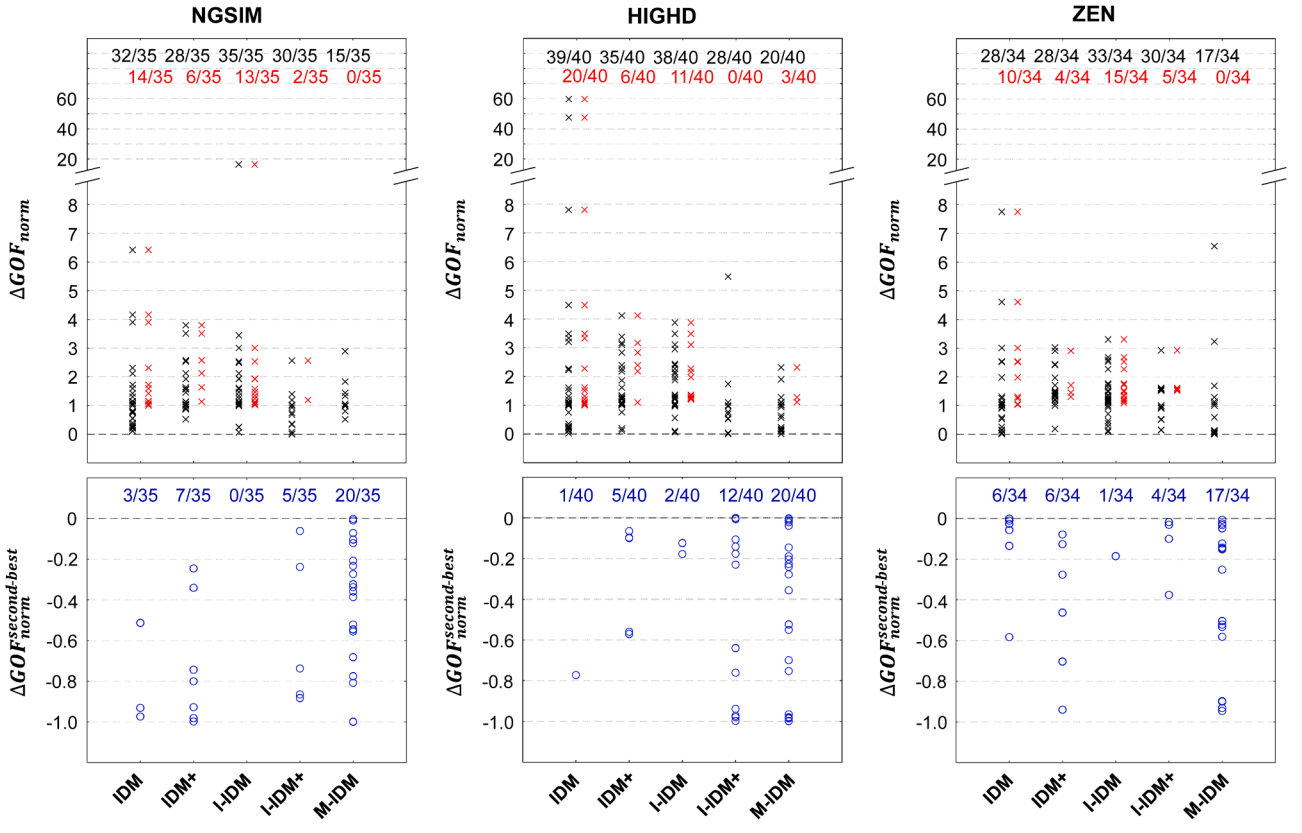


Fig. 7. Normalised relative errors of base models on each trajectory in the datasets. In the top diagrams, normalised relative errors of a model are represented for trajectories in which the model is not the best scoring (black crosses). Red crosses indicate errors when the model is the worst (they are a subset of black crosses). In the bottom diagrams, a blue circle represents the normalised error distance between the best model and the second-best. The fractions tell the frequencies of a model being the best one (blue fractions in bottom diagrams), not the best one (black fractions) and the worst one (red fractions) in top diagrams.

Table 1
Verisimilitude measures for the five base models computed on each trajectory dataset, and on their ensemble (*Eqs. (6–11)*). Values for the best and worst performing models are reported in bold and bold italics, respectively.

NGSIM								
Model	$\#_{best}$	$\#_{worst}$	*TC*	*FC*	***VS***1	Δ^{+}	Δ^{-}	***VS***2
IDM	3	14	0.09	0.40	−0.31	14.00	−74.32	−60.32
IDM+	7	6	0.20	0.17	0.03	55.34	−37.94	17.40
I-IDM	0	13	0.00	0.37	***−0.37***	0.00	−104.67	***−104.67***
I-IDM+	5	2	0.14	0.06	0.08	29.15	−10.06	19.09
M-IDM	20	0	0.57	0.00	**0.57**	90.49	0.00	**90.49**
HIGHD								
Model	$\#_{best}$	$\#_{worst}$	*TC*	*FC*	***VS***1	Δ^{+}	Δ^{-}	***VS***2
IDM	1	20	0.03	0.50	***−0.47***	7.38	−515.20	***−507.82***
IDM+	5	6	0.13	0.15	−0.02	130.36	−42.83	87.53
I-IDM	2	11	0.05	0.28	−0.23	11.44	−57.64	−46.20
I-IDM+	12	0	0.30	0.00	0.30	63.27	0.00	63.27
M-IDM	20	3	0.50	0.08	**0.42**	98.80	−9.88	**88.92**
ZEN								
Model	$\#_{best}$	$\#_{worst}$	*TC*	*FC*	***VS***1	Δ^{+}	Δ^{-}	***VS***2
IDM	6	10	0.18	0.29	−0.11	27.51	−75.40	−47.89
IDM+	6	4	0.18	0.12	0.06	42.62	−17.66	24.96
I-IDM	1	15	0.03	0.44	***−0.41***	5.20	−59.95	***−54.75***
I-IDM+	4	5	0.12	0.15	−0.03	15.74	−24.70	−8.96
M-IDM	17	0	0.50	0.00	**0.50**	79.18	0.00	**79.18**
ALL DATASETS								
Model	$\#_{best}$	$\#_{worst}$	*TC*	*FC*	***VS***1	Δ^{+}	Δ^{-}	***VS***2
IDM	10	44	0.09	0.40	−0.31	48.89	−664.92	***−616.03***
IDM+	18	16	0.17	0.15	0.02	228.32	−98.44	129.88
I-IDM	3	39	0.03	0.36	***−0.33***	16.64	−222.25	−205.61
I-IDM+	21	7	0.19	0.06	0.13	108.16	−34.76	73.40
M-IDM	57	3	0.52	0.03	**0.49**	268.47	−9.88	**258.59**

Focusing on the performance of IDM+ and M-IDM in the highD dataset, the comparison of the two measures provides an interesting insight into the models' behaviour. While the trade-off between "the number of best and worst occurrences" (i.e., *VS1*) clearly favours the M-IDM – with 20 best vs. 3 worst cases, compared to 5 best vs. 6 worst for IDM+ – the distance between the models almost vanishes when *VS2* is concerned (88.92 for IDM vs. 87.53). However, if the trade-off between the increase and the decrease upon the median models is independently considered (see Δ^{+} and Δ^{-}), results show that the IDM+ performs either very well or quite poorly, depending on whether it is the best or the worst model (Δ^{+}= 130.36, Δ^{-} = - 42.83). In contrast, the M-IDM appears more robust: it is most often the best model and rarely the worst, and in both cases its performance deviations from the median model are smaller than those of the IDM+ (Δ^{+}= 98.80, Δ^{-} = - 9.88). It is worth noting that these peculiar behaviours of models in the highD dataset may also be due to the shorter average duration of its trajectories, as previously reported.

If the results presented so far indicate the accuracy of a model, sensitivity indices in Fig. 8 measure the robustness of a model (see discussion in Section 2.3).

Based on the total sensitivity index of the input-trajectory factor, the IDM is the most robust model, followed by the M-IDM (dataset-specific results are reported in Appendix B.1, Fig. B1). Notably, although the M-IDM (as well as the I-IDM+) includes one additional parameter compared to the other base models (namely v_{crit}), this parameter has only a very limited influence on the error variance. This suggests that simplifying the model by fixing v_{crit} to an arbitrary value – or, since v_{crit} acts as a regime-activation threshold, by removing one of the two alternative formulations it enables – would have only a minor effect on the error variance. In this case, removing the formulation in the third row of Eq. (A5) in Appendix A worsens the 85th-percentile calibration error by about 8%.

4.1.2. Base model ranking

Combining results of the pairwise tests – as measured by the verisimilitude measures – and the sensitivity analysis, a clear ranking of the base models emerges. The M-IDM results the most accurate model by far (see *VS1* and *VS2* results in Table 1), and the second-most robust model (see the value of the trajectory factor index in Fig. 8). In fact, the most robust model i.e., the IDM, is the least accurate one. Therefore, as the M-IDM presents the best trade-off between model accuracy and uncertainty, it is deemed to be the model with the highest explanatory power among those tested.[4]

To qualitatively show the trade-off between model accuracy and uncertainty behind O'Neil's conjecture, a diagram inspired by

[4] Findings are obviously conditional on the trajectories applied in this study.

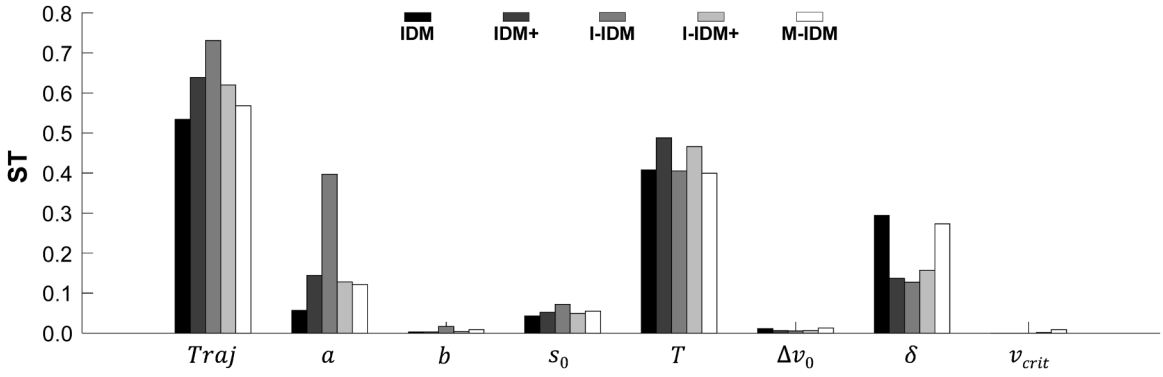


Fig. 8. Total sensitivity indices of the input trajectory factor and the model parameters for the five base models.

Fig. 2 is drawn for the models at stake, in Fig. 9. In the figure, the 85th percentile of the normalised relative errors after calibration is used as a statistic representing the model inadequacy error i.e., the inaccuracy. The total sensitivity index of the trajectory factor is instead applied as a proxy for the uncertainty propagation error. The green markers are therefore obtained by graphically[5] summing these two quantities and provide a qualitative representation of the overall model error being sought. The diagram confirms that the M-IDM has the best trade-off between accuracy and uncertainty and provides a visual representation of the base models' ranking.

Despite this evidence, for the sake of the analysis completeness all the base models were augmented with human factors layers and assessed in the rest of the study.

4.2. HF augmented models

4.2.1. Nominal driving conditions

In Fig. 10, the results of pairwise calibration experiments of the 800 model variants, for the ensemble of trajectories from the three datasets, are presented (corresponding figures for each dataset are reported in Appendix B.2, Fig. B2). Specifically, the 85th percentile of the normalised relative error in Eq. (12), and the number of parameters, are reported. Some general outcomes emerge from the figures.

As expected, the calibration error decreases with the model complexity, i.e., the number of model parameters. The overall impact of a specific HF addon on model performances, independently of the others, can be visually inspected in the figure. For instance, changes in the amplitude and position of the light-tone marker cloud relative to the dark marker one, tell the "overall impact" of the multi-anticipation on model errors – where "overall impact" means the impact of all the models which include the multi-anticipation. Similarly, looking at all the models with the same marker type, the impact of different perception errors on model errors can be inspected.

When the standalone contribution of an HF layer is concerned (i.e., when a single layer is added to the base model), and the best-performing model of each layer is considered, the spatial multi-anticipation layer presents the best marginal improvement of performance – i.e., the best trade-off between a model performance increase and the number of additional parameters required. Focusing on the specific results of the M-IDM models family, which are numerically reported in Table 2, spatial multi-anticipation improves performances by 20% (on the ensemble of datasets), with only two additional parameters (this can be visually appreciated also in Fig. 10 by comparing the shift of the light-tone markers relative to the dark-tone ones in calibration errors – top diagram – with the shift in the number of parameters – bottom diagram).

In fact, when the perception error is concerned, the best addon is the variable perception error given by the VPE(FDTD), which reduces the error by a 34% with 10 additional parameters. With regards to the perception delay, the variable perception delay combined with the temporal anticipation (VPD+ANT) is the best performing delay formulation, yielding a 32% error decrease, with 10 additional parameters. Among the adaptive driving layers, the 2D+, which requires 6 additional parameters, yields a 27% improvement.

When multiple layers are combined, the increase in performance is less than linear; that is, the model addons are not fully additive. The greatest improvement is obtained when four layers are added (81% vs. 109% which would be expected by summing the standalone improvement of each layer).

Considering three layers, the best combination is the one including VPE(FDTD), VPD+ANT and MLTP. In fact, when any of these three components is added to the other two, performances at the 85th percentile increase from 50–54% to 70%.

Looking back at Fig. 10 and focusing on the difference among base models, it is worth noting that the calibration error variances of

[5] The axis scales of the 85th percentile (red y-axis on the left) and the total sensitivity index (blue y-axis on the right) are drawn so that the projections of the minima of the two quantities over any of the two axes coincide, as well as those of the maxima. The green points are computed as the sum of the two quantities after they have been max-min normalised.

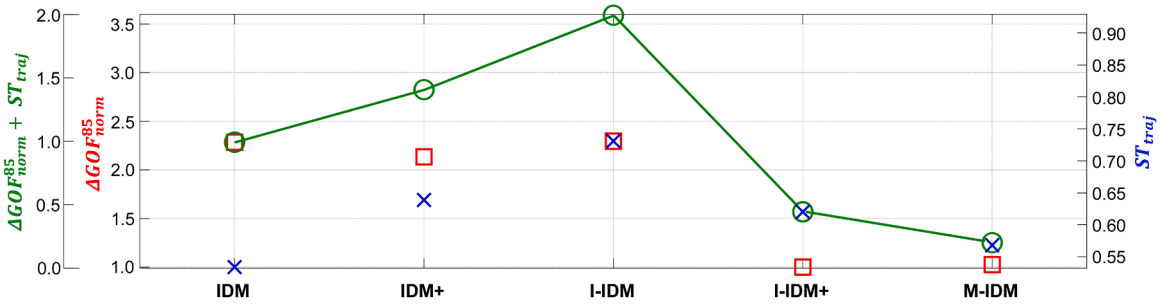


Fig. 9. Overall model error of the five base models (green point series) as the sum of model inadequacy error (red series) and uncertainty propagation error (blue series).

the IDM and the M-IDM model families are lower than those of the other base model families (across all datasets and within each one individually; see Appendix B.2). Finally, calibration errors sensibly vary from one dataset to another, with the highD dataset showing the highest variance of errors, although the pattern among models remains similar (see Appendix B.2).

The previous results allowed us to disentangle the impacts of different HF addons on model errors but provide an incomplete picture when a ranking of model performances is sought. The two verisimilitude measures introduced in Section 2.1 come to our aid, and their values for the 800 models are reported in Fig. 11 (measures per dataset are reported in Appendix B.3; Fig. B3 and Fig. B4). The figures clearly show that the model variants belonging to the families of the I-IDM+ and the M-IDM are those that perform consistently better than the others. Some models, in particular, stand out. These are all models with VPE(FDTD), VPD+ANT, MLTP and any of the three adaptive driving layers (555, 595, 635 in the I-IDM+ family; 715, 755, 795 in the M-IDM family). Regarding the three tested adaptive driving strategies, the choice of formulation is irrelevant for the I-IDM+ model, as all yield identical verisimilitude measures (which could be a symptom of data overfitting). In contrast, for the M-IDM model, the AD(FDTD) strategy looks preferable.

Overall, this result underscores an otherwise intuitive point: for a given HF layer, there is no universally optimal formulation. Rather, the effectiveness of a formulation might depend on its interplay with the chosen base model and the other integrated layers. We anticipate that this stems from the interaction effects among the various layers and parameters (see Section 4.2.3).

4.2.2. *Safety-critical driving conditions*

In Fig. 12, the results of the two tests for safety-critical driving conditions are presented for the M-IDM model family (very similar results for the other models are presented in Appendix B.4; Fig. B5). Models which result in collision-free conditions in the harsh deceleration test are indicated at the bottom of the figure. According to previous discussions, such models are inadequate for safety studies.

The rear-end collision rates per 100 million simulated kilometres are reported at the top of the figure. Most of the calibrated models show a collision rate equal to or greater than 10^5, which is four orders of magnitude greater than those observed (compare with Fig. 13, which reports the distribution of rear-end collision rates measured on the links of an Italian motorway network).

Surprisingly the highest collision rates are produced by models embedding a spatial anticipation mechanism, MLTP (see light-tone markers), which should help avoid collisions instead. Therefore, this result suggests that the tested layer formulation, i.e., MLTP, does not serve its main purpose, although it yields a meaningful improvement in model performance.[6]

The models that exhibit a zero collision rate but are not structurally collision-free – identified by comparing markers along the zero-collision rate axis with those on the collision-free axis – lack both perception error and delay, as well as temporal and spatial anticipation layers. They rely solely on an adaptive driving strategy. This suggests that while the tested adaptive strategies do not guarantee collision avoidance by design, they tend to prevent collisions under the applied parameter settings.

The fact that the remaining models – i.e., those incorporating perception and anticipation layers – produce unrealistic collision rates suggests two possible explanations: either the mechanisms introduced by these layers are not appropriate (as illustrated by the MLTP layer, which increases the collision rate by roughly an order of magnitude), or the calibration procedure is not adequate for this intended use. Additionally, calibration is made more difficult by the fact that the available data contain essentially no collisions and therefore provide little direct information on safety-critical responses. However, there is a further structural aspect to consider. As will be shown later (Section 4.2.3), the parameters most influential for nominal accuracy are also those that substantially affect collision rates. This indicates genuinely conflicting objectives: improving fit under nominal conditions can simultaneously worsen safety-critical behaviour unless the calibration problem is reformulated accordingly. Therefore, the issue is not only those previously mentioned, but also the calibration approach itself, which should be reframed as a Pareto (multi-objective) optimisation problem.

[6] An inspection of the model formulation suggests one possible mechanism that could explain this increase in collisions. Specifically, if the first leader in the platoon pulls away from the following vehicles, the model's computed mean spacing can become larger than the spacing to the ego vehicle's immediate leader. This can lead the ego vehicle to behave as if it had more available space than it actually does, thereby increasing the likelihood of a collision.

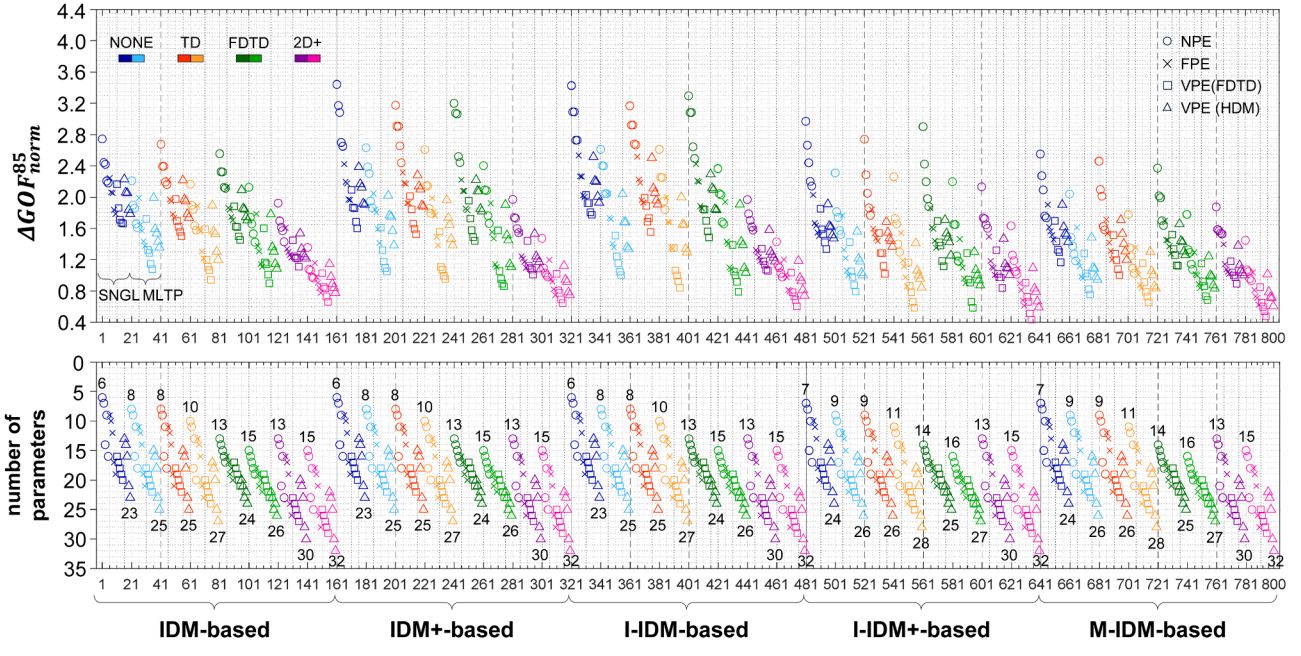


Fig. 10. 85th percentile of normalised relative calibration errors of the 800 model variants for the ensemble of the three datasets. At the bottom, the number of model parameters per model variant.

Table 2
Performance improvements in the M-IDM model family due to the addition of human factor (HF) layers. The 'best performing' formulation of each HF layer is considered. Models are grouped per increasing number of layers. Within each group, models are sorted from the best to the worst performing model (on the ensemble of all datasets). Cells are coloured according to the degree of improvement from white to dark green.

# of added layers	Model ID	HF layers added to the M-IDM	# of additional parameters	Improvements relative to the M-IDM			
				NGSIM	HIGHD	ZEN	ALL
1	*651*	*+ PE*	*10*	-15%	-41%	-41%	-34%
	645	*+ PD*	*10*	-18%	-41%	-18%	-32%
	761	*+ AD*	*6*	-11%	-16%	-46%	-27%
	661	*+ MLTP*	*2*	-12%	-32%	-11%	-20%
2	*655*	*+ PE + PD*	*14*	-35%	-63%	-56%	-54%
	771	*+ PE + AD*	*16*	-29%	-58%	-59%	-54%
	665	*+ PD + MLTP*	*12*	-43%	-65%	-45%	-54%
	671	*+ PE + MLTP*	*12*	-33%	-56%	-53%	-50%
	781	*+ AD + MLTP*	*8*	-26%	-45%	-53%	-43%
	765	*+ PD + AD*	*16*	-23%	-48%	-52%	-40%
3	*675*	*+ PE + PD + MLTP*	*16*	-65%	-76%	-62%	-70%
	791	*+ PE + AD + MLTP*	*18*	-51%	-68%	-73%	-67%
	775	*+ PE + PD + AD*	*20*	-50%	-74%	-68%	-65%
	785	*+ PD + AD + MLTP*	*18*	-52%	-72%	-63%	-64%
4	*795*	*+ PE + PD + MLTP + AD*	*22*	-73%	-87%	-78%	-81%

*For the sake of readability, the acronyms introduced in Section 3.1 were contracted as follows: PE = PE(FDTD), PD = VPD+ANT, and AD = AD (2D+). Therefore, results refer to the best formulation of each human factor layer (and not to the layer).

4.2.3. Candidate models selection

The analysis of model performances from pairwise tests allowed us to disentangle the contribution of each HF layer on the model performance and to draft a model ranking based on the ability of a model to improve upon the state of the art. Conversely, the analysis on safety-critical driving conditions highlighted the limitations of current formulations and parameterisations and allowed us to identify inadequate models.

Acknowledging the aforementioned limitations, the performance-based analysis (see Fig. 11) identifies a clear set of candidate models: 555, 595, and 635 from the I-IDM+ family, and 755 and 795 from the M-IDM family. These models share identical formulations for perception and anticipation (VPD+ANT+VPE(FDTD)+MLTP) and differ only in the adaptive driving strategy (AD(TD) for 555, AD(FDTD) for 595 and 755, and AD(2D+) for 635 and 795) and in the underlying base model. Regarding the former, we prefer the FDTD strategy over the 2D+ variant, as it consistently enhances the performance of both base models while using five fewer parameters. As for the base model, the results presented in Section 4.1 clearly support a preference for the M-IDM.

Accordingly, Fig. 13 presents, for the models in the M-IDM family, the same qualitative diagram shown earlier for the base models, illustrating the trade-off between accuracy and uncertainty (diagrams per dataset are presented in Appendix B.5; Fig. B6). The diagram indicates that models 755 and 795 exhibit the most favourable trade-off.

As the safety-critical analysis in Section 4.2.2 has revealed unrealistic collision rates for these models under the current parameterisation, a more in-depth investigation was conducted to assess the impact of individual parameters on both performance and collision rates. To this end, the top diagram in Fig. 14 presents the total sensitivity indices of each parameter of the two models calculated on the model errors. The bottom diagram shows the Kolmogorov-Smirnov statistics comparing the distributions of parameter values that lead to collisions versus those that do not (the parameter distributions resulting from the Monte Carlo Filtering are reported in Appendix B.6; Fig. B7). The comparison of these diagrams highlights that the parameters most influential on model performance are also those most strongly associated with collision occurrence. This finding suggests that calibrating a model to perform well under both safety-critical and nominal driving conditions constitutes a *Pareto optimisation problem*, as improving one objective inherently degrades the other.

4.2.4. Verification of candidate models on emerging macroscopic traffic characteristics

The M-IDM and the two variants mentioned above were used in a set of single-lane traffic simulation experiments to assess their emergent macroscopic behaviour.

Fig. 16 reports the fundamental diagrams of the three models across the three datasets, while Fig. 17 shows the distributions of shockwave speeds. To construct the fundamental diagrams, we simulated 20 equilibrium platoons, each consisting of 20 heterogeneous vehicles. Vehicle-specific car-following parameters were sampled from a Gaussian copula fitted to the calibrated parameter sets. For each equilibrium speed, the corresponding equilibrium density and flow are plotted (light points in Fig. 16). To reduce sampling noise due to platoon composition, we computed the median density at each speed (dark points) and fitted these median points with a fourth-order polynomial. For shockwave-speed estimation, experiments were initialised at equilibrium and a leader deceleration of 3 m/s^2 was imposed.

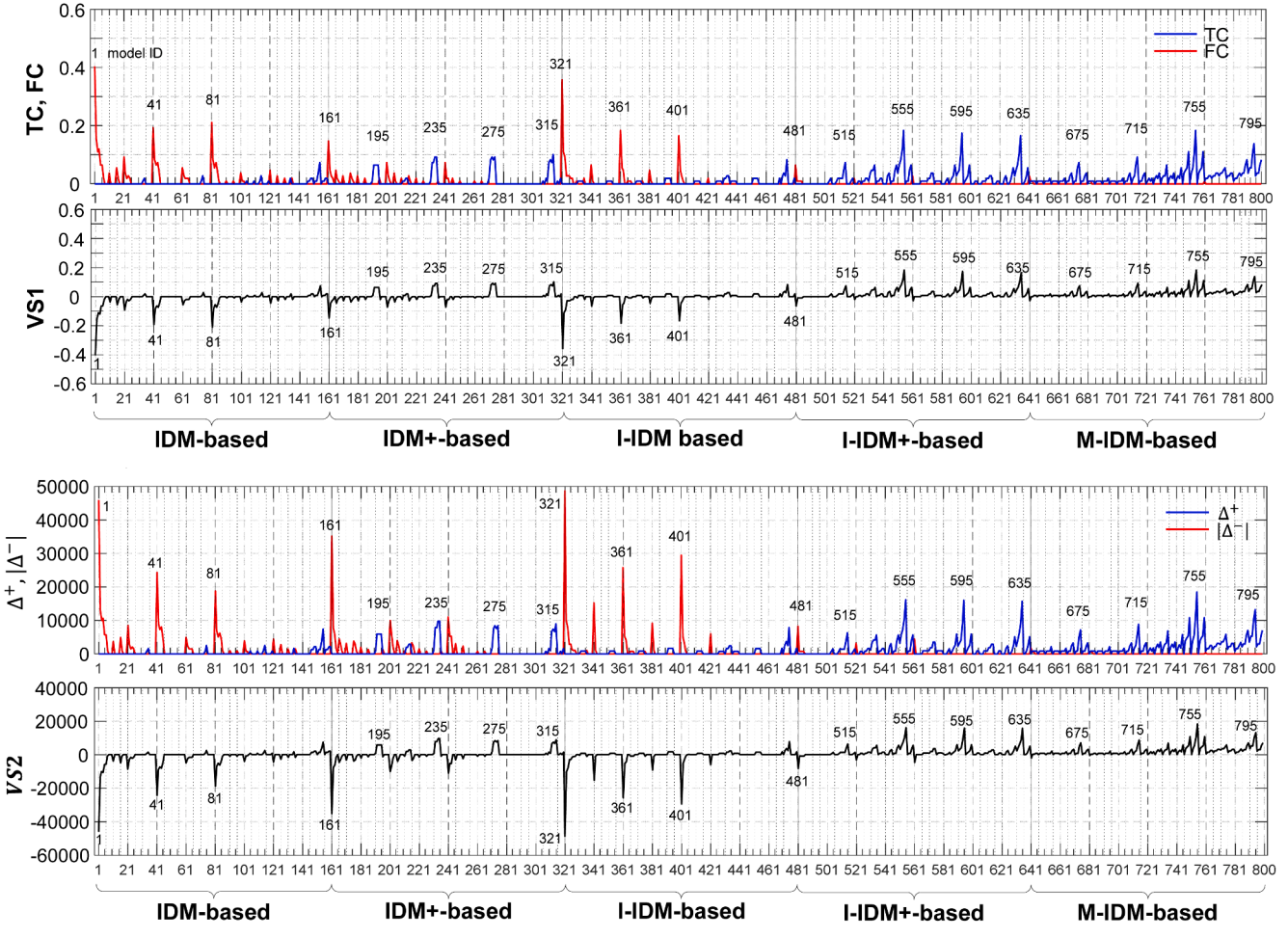


Fig. 11. Verisimilitude measures, VS1 and VS2 (black lines) and their positive and negative components (blue and red lines) for the 800 model variants across all datasets.

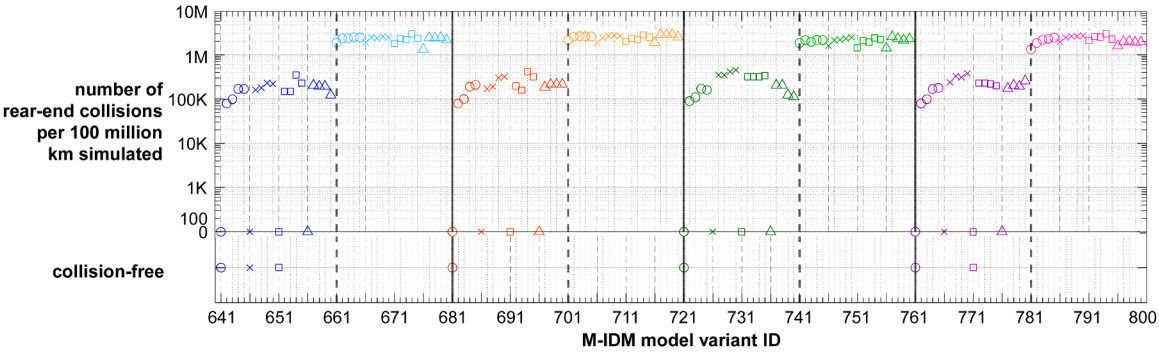


Fig. 12. Results of the two tests for safety-critical driving conditions. Top diagram: number of rear-end collisions per 100 million kilometres simulated, for the M-IDM model variants, by sampling the leader trajectory from the three datasets and the model parameter values from a Gaussian copula (estimated on calibrated parameter values). Bottom diagram: model variants resulting in collision-free conditions in the harsh deceleration experiment.

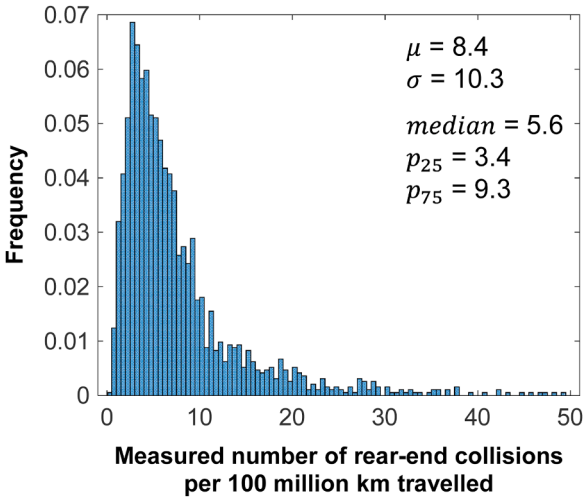


Fig. 13. Distribution across links of the number of rear-end collisions per 100 million km travelled measured between 2017 and 2021 on the largest Italian motorway network, operated by "Autostrade per l'Italia S.p.a." (total length: 23,319 km; average annual daily traffic: 23,324 veh/day).

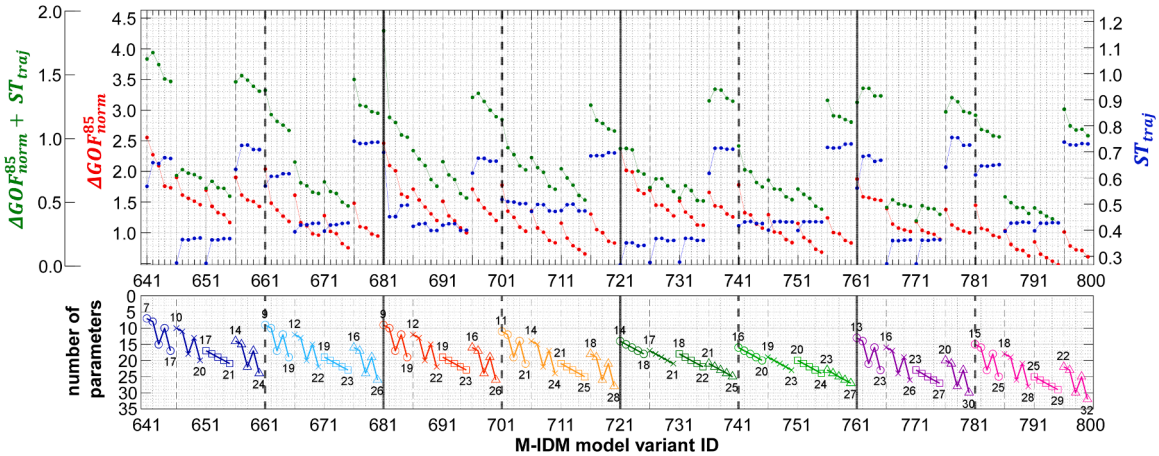


Fig. 14. Trade-off between model accuracy and uncertainty for the M-IDM model variants, on the ensemble of all datasets. In red, the 85th normalised relative calibration errors among all datasets. In blue, the average total sensitivity index of the trajectory factor among all datasets. In green, the resulting graphically combined measure.

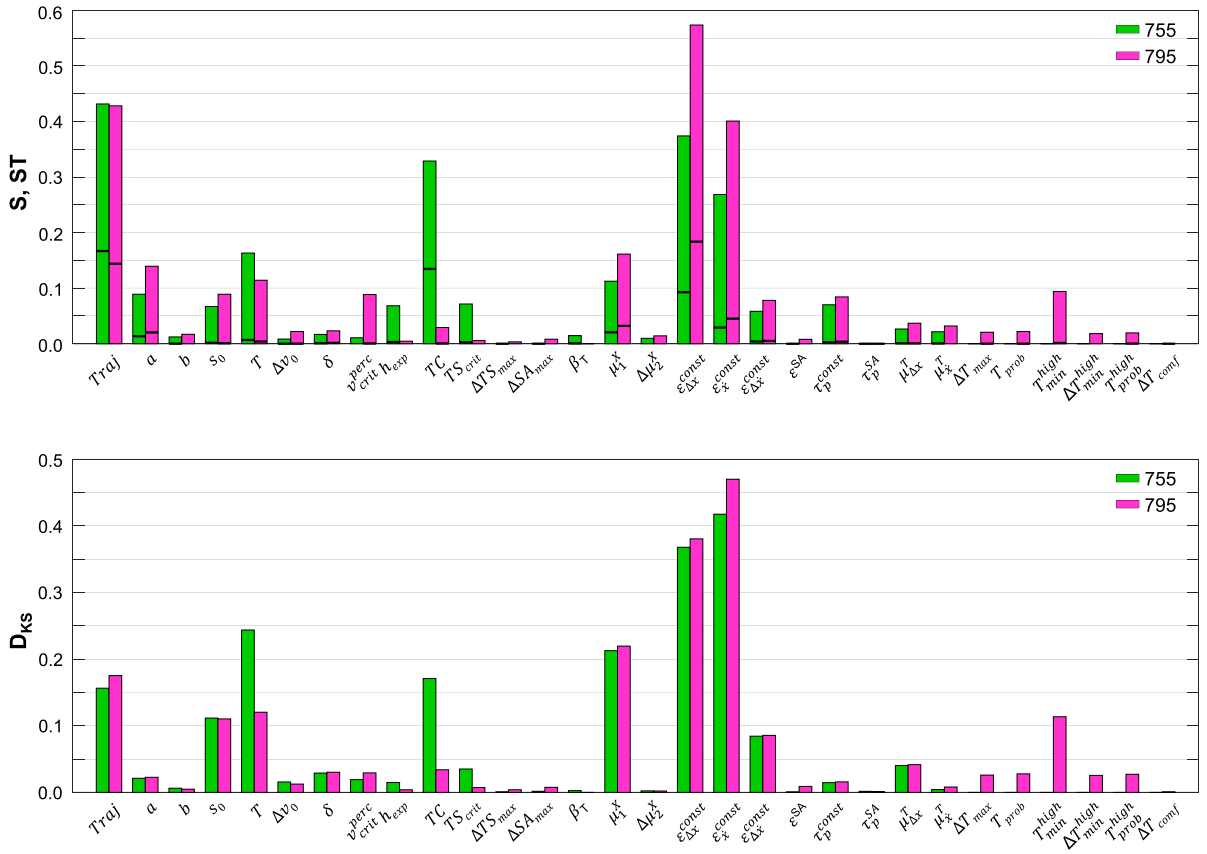


Fig. 15. Total sensitivity indices on model errors (top) and Kolmogorov-Smirnov distance measure on collisions (bottom) for each input factor of the two M-IDM model variants 755 and 795. In the top diagram the black horizontal mark in a bar denotes the first-order sensitivity index of that factor (755: M-IDM+VPD+ANT+VPE(FDTD)+MLTP+AD(FDTD), 795: M-IDM+VPD+ANT+VPE(FDTD)+MLTP+ AD(2D+)).

Regarding the fundamental diagrams, variants 755 and 795 show good agreement with the capacity values measured in the three datasets (capacity values estimated in the field literature for NGSIM, highD and Zen datasets are, respectively: 1700 veh/h, in Zhang et al., 2025; 1556 veh/h, in Chen et al., 2022; 1250 veh/h, in Dahiya et al., 2022). The M-IDM matches the experimental evidence only in the Zen data, while it yields capacity values above those measured in the NGSIM and highD datasets. Regarding shockwave propagation speeds, all three models produce median values comparable with those estimated in NGSIM (i.e., values between 14.5 to 18.9 km/h, see Chiabaut et al., 2009), whereas in Zen they underestimate the estimated values (i.e., 16.7 km/h, estimated in this study; see track 25, dir. 2, lane 3, first wave). In highD, results are heterogeneous, and only variant 755 shows values comparable to those observed (i.e., 18.6 km/h, estimated in this study;see Route 11, track 4, lane 2) .[7]

To investigate the reasons for these behaviours, Fig. 18 reports, for each of the three models, the distributions of the base-model parameters only (note that the two variants include, respectively, 17 and 22 human-factors parameters in addition to the base-model ones). In particular, ΔV_0 represents the increase in desired speed relative to the maximum speed measured in the data (i.e., an effective proxy for V_0 in calibration). Across all three datasets, the M-IDM exhibits a higher frequency of low values of the desired time headway T than its two variants. This is consistent with the higher capacity values produced by the M-IDM in the first two datasets (recall that, in the IDM family, the equilibrium spacing depends on T, s_0, δ, and V_0). The discrepancy in the distribution of T is evidently compensated by the behavioural mechanisms of models 755 and 795 in the Zen data, where capacity is nearly identical for the three models. Indeed, in the two variants the desired time gap is not a constant and varies as a function of driver's mental state. The question remains as to why the M-IDM shows the discrepancy in capacity in the first two datasets and not in Zen data. The only explanation we can offer is that the Zen trajectories are longer and more informative than the other two, as already noted.

[7] In this study, for the highD and Zen datasets, we estimated the propagation speed of the shockwave traversed by the largest number of selected trajectories within each dataset.

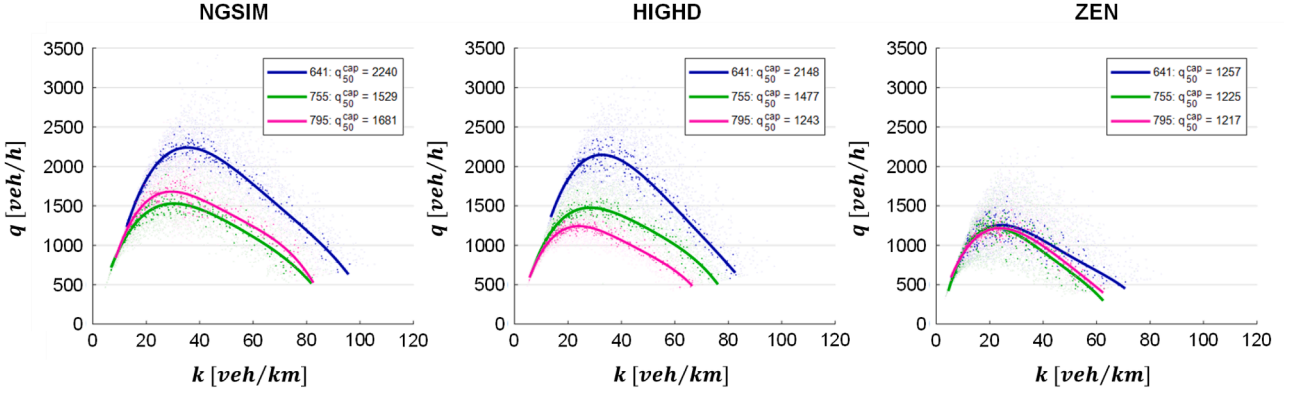


Fig. 16. Simulated fundamental diagrams for the M-IDM (641) and its variants 755 and 795 across the three datasets. For details on models 755 and 795, see Fig. 15 caption. In each diagram, light points denote all simulations, whereas darker points denote, for each equilibrium speed, the median density across simulations of 20 sampled platoons (with the corresponding flow). The curves are fourth-order polynomial fits to the median points.

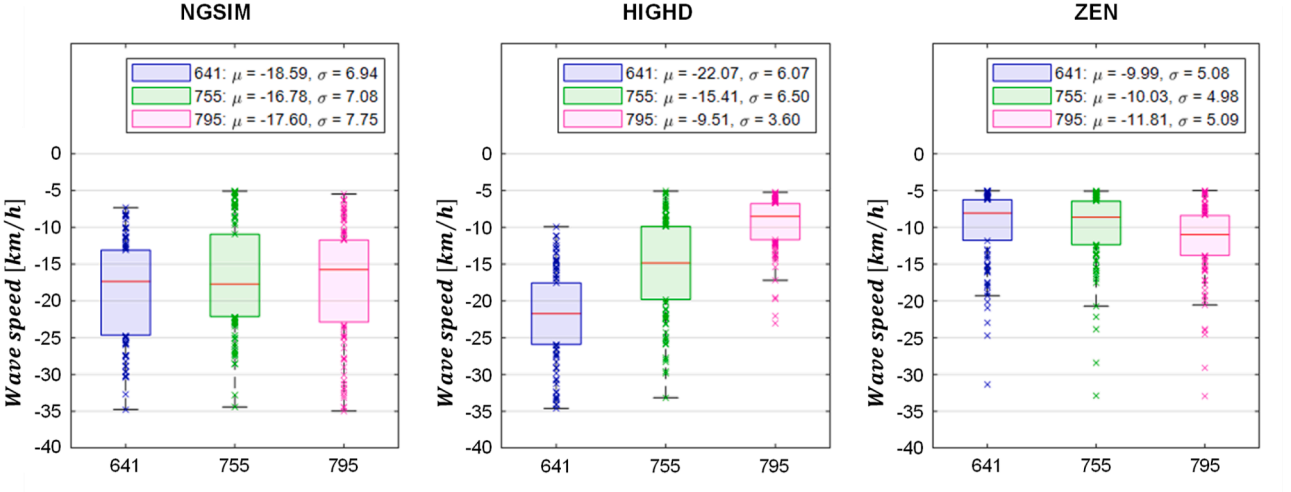


Fig. 17. Box plots of simulated shockwave speeds for the M-IDM (641) and its variants 755 and 795, across the three datasets (for the details of models 755 and 795, see Fig. 15 caption).

5. Summary and conclusion

The evaluation of theories and models is key to the progress of science. In the field of driver modelling – an inherently complex one – proper evaluation is hindered by the lack of appropriate methodologies. In fact, this paper shows that comparing models using error distributions is fundamentally flawed. Moreover, when these models are equipped with perception and cognitive modules to capture potentially unsafe driving mechanisms, a performance-based approach under nominal driving conditions is clearly incomplete.

For these reasons, the evaluation/validation approach proposed in this study incorporates assessment methods for both safety-critical and nominal conditions. Notably, evaluating a model across diverse and challenging conditions – through so-called risky tests – is a central tenet of Popper's philosophy of scientific progress, an epistemological stance that this paper fully embraces.

Popper's ideas of risky tests and of different degrees of model truthfulness have inspired the overall evaluation approach proposed here. With regards to the assessment of model performances in nominal conditions, we propose to carry on pairwise calibration tests to compare model performances, as risky tests. We suggest that a meaningful criterion to assess and compare models should be the ability of a new model to outperform existing ones under the same calibration tests. This idea materialises in two measures of model "verisimilitude" which allow us to rank models based on a meaningful interpretation of their performances. These measures rely on a relative and normalised error statistic, which is proposed in this study to enable the application of the methodology to heterogeneous trajectories (and datasets) without biasing results.

Since the accuracy of a model tells only half of the story, global sensitivity analysis is proposed to evaluate its robustness. By considering the input trajectory as a factor of the analysis – as well as model parameters – the robustness of a model against heterogeneous inputs can be evaluated. The total sensitivity index of the input trajectory factor, in particular, is recommended as a proxy of the model robustness to heterogeneous inputs i.e., a proxy for model's ability to equally capture heterogeneous behaviour irrespective of the increasing amount of uncertainty injected by additional parameters and model components.

The combination of model verisimilitude measures and total sensitivity indices is therefore proposed as a way to evaluate the trade-off between model inaccuracy and uncertainty – which are expected to, respectively, decrease and increase with model complexity and parameterisation – and as an indicator of the model explanatory power.

The proposed methodology has been applied to an extensive model evaluation exercise including 800 model variants, which allowed us to unveil the approach's potential. In this application, a further novel ingredient was proposed, i.e., an incremental augmentation of models, which allowed us to disentangle the individual impact of human factors components on model capability; as far as we know, this is the first time that such an investigation has been carried out.

With regard to base models, the methodology unveiled substantial differences. The M-IDM – a novel formulation proposed in this work – outperforms the IDM and its variants, consistently in all tested datasets. As these data were gathered across three different continents, the underlying driving mechanism better captured by this formulation seems not to be related to country-specific driving styles or attitudes.

With regard to the human factors addons, this study has quantified the contribution of each addon to model accuracy, both individually and in their mutual interactions. Results indicate that these human factors layers provide a substantial improvement to model accuracy – besides enabling the modelling of safety-critical driving conditions. However, when the safety outputs were investigated, driver models calibrated against nominal driving conditions (trajectories) exhibited unrealistic collision rates in simulation. Analyses also provided indications of which modelling components to refine to achieve more realistic rates. For instance, simulations showed that the tested spatial multi-anticipation mechanism could not prevent collisions, contrary to what was expected.

A more in-depth analysis of parameters capability to affect performances and collision rates (respectively, by means of sensitivity indices and Kolmogorov Smirnov distances) revealed that it is not possible to improve collision rates without deteriorating performances in nominal driving conditions, as the most influential parameters in the two driving conditions are the same (see, especially, the constant perception errors on speed and spacing) so that the two calibration objectives are conflicting. For a driver model of this type, therefore, future research should investigate calibration as a Pareto optimisation. This limitation is directly relevant for virtual safety assessment of automated driving systems, where driver models already enter regulatory/type-approval contexts (e.g., UN R157, 2023), and are actively developed as reference baselines in industry.

It is worth noting that the evaluation approach developed and applied in this study allowed us to inspect and interpret the behaviour of models and of their sub-components, thus providing key information for model design and refinement. Other available methods or criteria for model selection like the Bayesian or Akaike information criteria or other Bayesian approaches, only provide performance-based indicators for ranking models, but leave the analyst blind about the relevance of parameters and modules. Furthermore, the idea of penalising models with a higher number of parameters can be misleading, though sound in principle.

Overall, human factors, which are essential to use driver models in safety applications, have been shown in this study to sensibly improve model explanatory power. It was showed, however, that a considerate and responsible development of such models requires an extensive testing against observed data, which currently are not available to the required extent and complexity. Future efforts should be therefore devoted in gathering open datasets of naturalistic human driving – including in-cabin data and detailed surrounding traffic – and in the massive testing of any new theory against these data.

Finally, within the candidate set identified here, model 755 is the preferred variant (M-IDM+VPD+ANT+VPE(FDTD)+ MLTP+AD (FDTD)), as it remains among the best-performing models while being more parsimonious than others (e.g. 795) and, importantly, it is supported by the macroscopic traffic verification results (capacity and shockwave-speed consistency across datasets).

To the sake of reproducibility of results, all codes and data in this study are open source and are available at https://github.com/i4Driving.

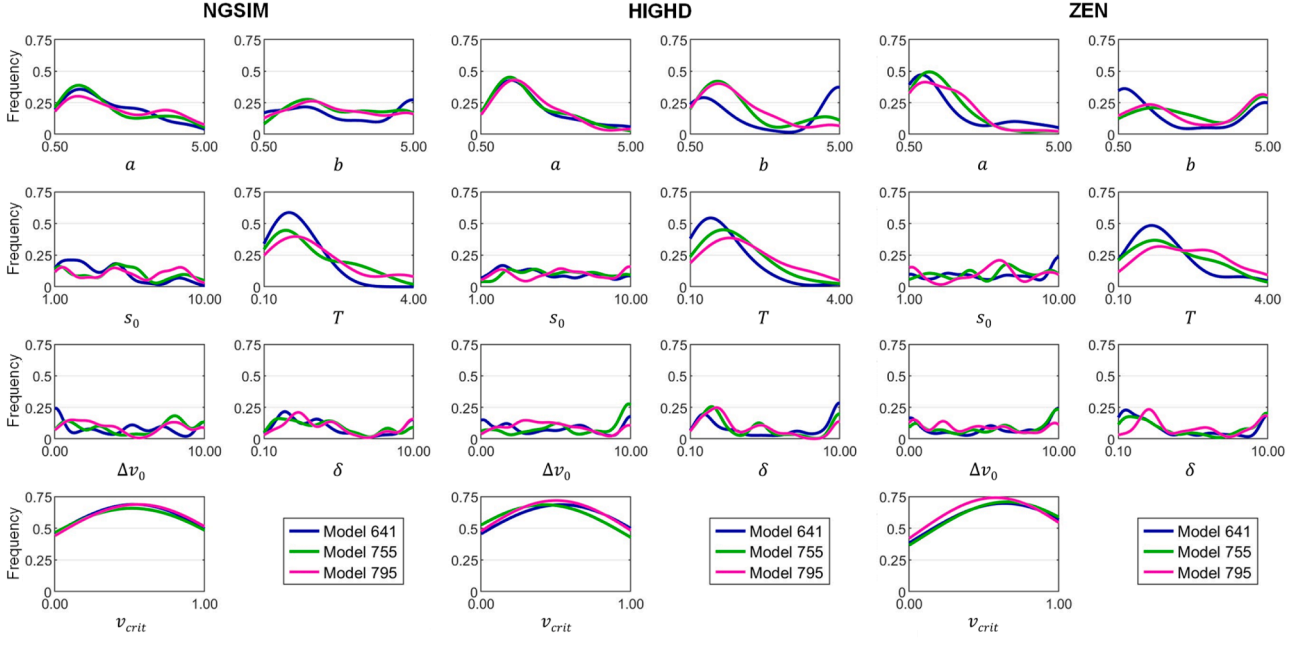


Fig. 18. Distributions of the calibrated values of the M-IDM parameters for the M-IDM (641) and its variants 755 and 795, across the three datasets (for the sake of the comparison only the distributions of the base model – the M-IDM – parameters have been reported for the variants 755 and 795). For details about models 755 and 795, see Fig. 15 caption.

CRediT authorship contribution statement

Vincenzo Punzo: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Writing – original draft, Writing – review & editing, Visualization. **Davide Iannelli:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization. **Andrea Saltelli:** Writing – review & editing, Conceptualization, Methodology. **Marcello Montanino:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

Acknowledgement

Research in this study was funded by the European Commission under the Horizon Europe Project 101076165 – "i4Driving: Integrated 4D driver modelling under uncertainty". The authors thank Dr. Filomena Mauriello for providing the statistics about rear-end collision rates on an Italian highway network.

Appendix A

A.1. General notation

Table A.1

Symbol	Description
$\ddot{x}_n(t)$	Acceleration of vehicle n at time t returned by the car-following model
$\dot{x}_n(t)$	Speed of vehicle n at time t: $\dot{x}_n(t) = \dot{x}_n(t-\Delta t) + \ddot{x}_n(t)\cdot\Delta t$
$x_n(t)$	Position of vehicle n at time t: $x_n(t) = x_n(t-\Delta t) + [\dot{x}_n(t-\Delta t) + \dot{x}_n(t)]\cdot\Delta t/2$
l_n	Length of vehicle n
$\Delta x_n(t)$	Net spacing between vehicles $n-1$ and n at time t: $\Delta x(t) = x_{n-1}(t) - l_{n-1} - x_n(t)$
$\Delta\dot{x}_n(t)$	Speed difference between vehicles $n-1$ and n at time t: $\Delta\dot{x}(t) = \dot{x}_{n-1}(t) - \dot{x}_n(t)$
$\dot{x}_n^p(t)$	Perceived speed of vehicle n at time t
$\Delta x_n^p(t)$	Perceived net spacing between vehicles $n-1$ and n at time t
$\Delta\dot{x}_n^p(t)$	Perceived speed difference between vehicles $n-1$ and n at time t
Δt	Numerical integration step

Please note that the perceived stimuli are obtained by sequentially applying mechanisms of perception and anticipation, which include perception delay and error, as well as temporal and spatial anticipation. These mechanisms are applied in cascade, following this specific order: (i) perception delay, (ii) perception error, (iii) temporal anticipation, and (iv) spatial anticipation (see Appendix A.3-A.6).

A.2. Base models

Five base car-following models, all belonging to the family of the Intelligent Driver Model (IDM), were tested:

- The original **IDM** whose formulation is (Treiber et al., 2000):

$$\ddot{x}_n(t) = a\left[1 - \left(\frac{\dot{x}_n^p(t)}{v_0}\right)^{\delta} - \left(\frac{\Delta x^*(t)}{\Delta x_n^p(t)}\right)^2\right] \tag{A1}$$

$$\Delta x^*(t) = s_0 + \max\left\{0,\ \dot{x}_n^p(t)\cdot T + \frac{\dot{x}_n^p(t)\cdot\Delta\dot{x}_n^p(t)}{2\sqrt{ab}}\right\} \tag{A2}$$

  where a is the maximum acceleration, v_0 is the maximum desired speed, δ is a constant, s_0 is the minimum net inter-vehicle spacing at stop, T is the desired time-headway, b is the comfort deceleration.
- The **IDM+** (Schakel et al., 2012), in which Eq. (A1) is modified as follows:

$$\ddot{x}_n(t) = a\cdot\min\left\{1 - \left(\frac{\dot{x}_n^p(t)}{v_0}\right)^{\delta}, 1 - \left(\frac{\Delta x^*(t)}{\Delta x_n^p(t)}\right)^2\right\} \tag{A3}$$

- An improved version of the original IDM, here referred to as **I-IDM**, in which Eq. (A1) is modified as follows (Treiber and Kesting, 2013):

$$\ddot{x}_n(t) = \begin{cases} \begin{cases} a\left[1-\left(\frac{\dot{x}_n^p(t)}{v_0}\right)^{\delta}\right]\left[1-\left(\frac{\Delta x^*(t)}{\Delta x_n^p(t)}\right)^{\overline{1-\left(\frac{\dot{x}_n^p(t)}{v_0}\right)^{\delta}}^{2}}\right] & \Delta x^*(t) \le \Delta x_n^p(t) \\ & \dot{x}_n^p(t) \le v_0 \\ a\left[1-\left(\frac{\Delta x^*(t)}{\Delta x_n^p(t)}\right)^2\right] & \Delta x^*(t) > \Delta x_n^p(t) \end{cases} \\ \begin{cases} -b\left[1-\left(\frac{v_0}{\dot{x}_n^p(t)}\right)^{\frac{a\delta}{b}}\right] & \Delta x^*(t) \le \Delta x_n^p(t) \\ & \dot{x}_n^p(t) > v_0 \\ -b\left[1-\left(\frac{v_0}{\dot{x}_n^p(t)}\right)^{\frac{a\delta}{b}}\right] + a\left[1-\left(\frac{\Delta x^*(t)}{\Delta x_n^p(t)}\right)^2\right] & \Delta x^*(t) > \Delta x_n^p(t) \end{cases} \end{cases} \tag{A4}$$

- A refinement of the I-IDM proposed by Tian et al. (2016), here referred to as **I-IDM+**, in which Eq. (A1) is modified as follows:

$$\ddot{x}_n(t) = \begin{cases} a\left[1-\left(\frac{\dot{x}_n^p(t)}{v_0}\right)^{\delta}\right]\cdot\left[1-\left(\frac{\Delta x^*(t)}{\Delta x_n^p(t)}\right)^2\right] & \Delta x^*(t) \le \Delta x_n^p(t) \\ \begin{cases} a\left[1-\left(\frac{\Delta x^*(t)}{\Delta x_n^p(t)}\right)^2\right] & \dot{x}_n^p(t) \le v_{crit} \\ min\left\{a\left[1-\left(\frac{\Delta x^*(t)}{\Delta x_n^p(t)}\right)^2\right], -b\right\} & \dot{x}_n^p(t) > v_{crit} \end{cases} & \Delta x^*(t) > \Delta x_n^p(t) \end{cases} \tag{A5}$$

where v_{crit} is the critical speed.
A new modification of the IDM – proposed in this work and named **M-IDM** – in which Eq. (A5) in the case of $\Delta x^*(t) \le \Delta x_n^p(t)$ is replaced by Eq. (A1).

A.3. Perception delay

Three perception delay models (PD) were tested:

- **NPD**: no perception delay is applied, i.e., $\tau_p(t) = 0$.
- **FPD**: a time-invariant, i.e., fixed perception delay is applied (Treiber et al., 2006), i.e., $\tau_p(t) = \tau_p^{const}$.
- **VPD**: a time-varying, i.e., variable perception delay modelled according to van Lint and Calvert (2018), is applied.

$$\tau_p(t) = \tau_p^{const} + (SA_{max} - SA(t)) \cdot \tau_p^{SA} \tag{A6}$$

where τ_p^{SA} is the attention-dependent time lag, and $SA(t)$ is the time-varying situational awareness level defined as follows:

$$SA(t) = \begin{cases} SA_{max} & TS(t) \le TS_{crit} \\ SA_{max} - \left(\frac{TS(t) - TS_{crit}}{TS_{max} - TS_{crit}}\right)\cdot(SA_{max} - SA_{min}) & TS(t) < TS_{max} \\ SA_{min} & otherwise \end{cases} \tag{A7}$$

where SA_{min} and SA_{max} are the minimum and maximum situational awareness level, $TS(t)$ is the time-varying task saturation level defined in Eq. (A17), TS_{crit} and TS_{max} being the critical and maximum task saturation values (see Appendix A.7 for more details on the task saturation model).

The stimuli affected by the perception delay are computed as $y_n^p(t) = y_n\left(t - \tau_p(t)\right)$, where $y = \{\dot{x}, \Delta x, \Delta\dot{x}\}$.

A.4. Perception error

Four perception error models (PE) were tested:

- **NPE**: no perception error is applied, i.e., $\varepsilon_{\dot{x}}(t) = \varepsilon_{\Delta x}(t) = \varepsilon_{\Delta\dot{x}}(t) = 0$.
- **FPE**: a time-invariant, i.e., fixed perception error is applied (Treiber et al., 2006), i.e., $\varepsilon_{\dot{x}}(t) = \varepsilon_{\dot{x}}^{const}$, $\varepsilon_{\Delta x}(t) = \varepsilon_{\Delta x}^{const}$, and $\varepsilon_{\Delta\dot{x}}(t) = \varepsilon_{\Delta\dot{x}}^{const}$.
- **VPE(FDTD)**: a time-varying, i.e., variable perception error modelled according to van Lint and Calvert (2018), is applied:

$$\varepsilon_y(t) = \varepsilon_y^{const} + (SA_{max} - SA(t)) \cdot \varepsilon^{SA}, \ \forall y \in \{\dot{x}, \Delta x, \Delta\dot{x}\} \tag{A9}$$

where $y = \{\dot{x}, \Delta x, \Delta\dot{x}\}$, and ε^{SA} is the attention-dependent perception error.

- **VPE(HDM)**: a variable perception error modelled according to Treiber et al. (2006), is applied.

$$w_y(t) = exp\left(-\frac{\Delta t}{\varepsilon_\tau}\right) \cdot w_y(t-\Delta t) + \eta_y(t)\sqrt{\frac{2\Delta t}{\varepsilon_\tau}} \tag{A10}$$

$$\begin{cases} \varepsilon_{\dot{x}}(t) = \varepsilon_{\dot{x}}^{const} + exp\left(\varepsilon_{\dot{x}}^{W} \cdot w_{\dot{x}}(t)\right) \\ \varepsilon_{\Delta x}(t) = \varepsilon_{\Delta x}^{const} + exp\left(\varepsilon_{\Delta x}^{W} \cdot w_{\Delta x}(t)\right) \\ \varepsilon_{\Delta\dot{x}}(t) = \varepsilon_{\Delta\dot{x}}^{const} - \dfrac{\Delta x(t)}{\Delta\dot{x}(t)} exp\left(\varepsilon_{\Delta\dot{x}}^{W} \cdot w_{\Delta\dot{x}}(t)\right) \end{cases} \tag{A11}$$

where $y = \{\dot{x}, \Delta x, \Delta\dot{x}\}$, $\{\eta_y(t)\} \sim$ *i.i.d.* $N(0,1)\ \forall t$, ε_τ is the correlation time of the Wiener process $w_y(t)$ and ε_y^W is the variation coefficient of the estimated stimulus.

The stimuli affected by perception errors are computed as $y_n^p(t) = \varepsilon_y(t) \cdot y_n(t)$, where $y = \{\dot{x}, \Delta x, \Delta\dot{x}\}$.

A.5. Temporal anticipation

In presence of a perception delay (see Appendix A.3), the following temporal anticipation mechanism (**ANT**) was tested (based on Treiber et al., 2006):

$$\begin{cases} \dot{x}_n^p(t) = \dot{x}_n\left(t - \tau_p(t)\right) + \mu_{\dot{x}}^T \cdot \ddot{x}_n\left(t - \tau_p(t)\right) \cdot \tau_p(t) \\ \Delta x_n^p(t) = \Delta x_n\left(t - \tau_p(t)\right) + \mu_{\Delta x}^T \cdot \Delta\dot{x}_n\left(t - \tau_p(t)\right) \cdot \tau_p(t) \end{cases} \tag{A12}$$

where $\mu_{\dot{x}}^T$ and $\mu_{\Delta x}^T$ – introduced in this work – are the driver's sensitivities to temporal anticipation of speed and spacing, respectively.

A.6. Spatial anticipation

The following spatial anticipation mechanism (**MLTP**) was tested (Ngoduy, 2015):

$$y_n^p(t) = \sum_{i=1}^{N_{leaders}} \mu_{y,i}^X \cdot y_{n-i+1}(t) \tag{A13}$$

where $y = \{\dot{x}, \Delta x\}$ and $\mu_{y,i}^X$ are the sensitivities to stimuli coming from $N_{leaders}$ preceding vehicles, such that: $\mu_{y,i}^X \geq \mu_{y,i+1}^X \geq \ldots \geq \mu_{y,N_{leaders}}^X$ and $\sum_{i=1}^{N_{leaders}} \mu_{y,i}^X = 1$ (in this work $N_{leaders} = 3$ was assumed, in accordance with evidence presented in Hoogendoorn et al., 2006, and $\mu_{\dot{x},i}^X = \mu_{\Delta x,i}^X$ as in Ngoduy, 2015).

A.7. Adaptive driving behaviour

Four adaptive driving behaviour mechanisms (AD) were tested:

- **NAD**: no adaptive driving behaviour is applied.
- **AD(TD)**: the Task-Capability Interface model proposed by Fuller (2000) and developed in Saifuzzaman et al. (2015), in which the desired inter-vehicle spacing $\Delta x^*(t)$ in Eqs. (A2) is modified as follows:

$$\Delta x^*(t) = \left[s_0 + \max\left\{0,\ \dot{x}_n^p(t) \cdot T + \frac{\dot{x}_n^p(t) \cdot \Delta\dot{x}_n^p(t)}{2\sqrt{ab}} \right\} \right] \cdot TD(t) \tag{A14}$$

with:

$$TD(t) = \frac{\dot{x}_n^p(t) \cdot T}{\left[(1 - \beta_{TD}) \cdot \Delta x(t)\right]^{\gamma_{TD}}} \tag{A15}$$

where β_{TD} and γ_{TD} additional model parameters. Please note that the modified desired spacing is applied in Eqs. (A1), (A3), (A4) and (A5).

- **AD(FDTD)**: the Fundamental Diagram Task Demand framework proposed in van Lint and Calvert (2018), in which the time-varying desired time-headway, $T^{var}(t)$, replaces the corresponding base model parameters in Eqs. (A1-A5).

$$T^{var}(t) = T \cdot \left[1 + max\{0,\ \beta_T \cdot (TS(t) - TS_{crit})\}\right] \tag{A16}$$

where T is the desired time-headway base model parameter, β_T is scaling parameter, TS_{crit} is the critical task saturation value and $TS(t)$ is the time-varying task saturation level due to car-following, proposed in Calvert et al. (2020):

$$TS(t) = exp\left(-\frac{h_n(t)}{h_{exp}}\right)\Big/ TC \tag{A17}$$

where $h_n(t) = \frac{\Delta x_n^p(t)}{\dot{x}_n^p(t)}$, TC is the driver task capacity, and h_{exp} is a parameter of the car-following task demand model.

- **AD(2D+)**: the desired time-headway model, 2D+, proposed in Tian et al. (2016) and updated in Zheng et al. (2022), in which a time-varying stochastic desired time-headway, $T^{var}(t)$ replace the corresponding base model parameters in Eqs. (A1-A5).

$$T^{var}(t) = \begin{cases} max\left\{\widetilde{T}^{var}(t), \widetilde{T}^{var}(t-\Delta t) - \Delta T_{comf}\right\} & \widetilde{T}^{var}(t) < T^{var}(t-\Delta t) \\ min\left\{\widetilde{T}^{var}(t), \widetilde{T}^{var}(t-\Delta t) + \Delta T_{comf}\right\} & otherwise \end{cases} \tag{A18}$$

$$\widetilde{T}^{var}(t) = \begin{cases} \begin{cases} T + \eta(t)\cdot(T_{max} - T) & \dot{x}_n^p(t) \le v_{crit} \text{ and } \eta(t) < T_{prob} \\ T_{min}^{high} + \eta(t)\cdot\left(T_{max}^{high} - T_{min}^{high}\right) & \dot{x}_n^p(t) > v_{crit} \text{ and } \eta(t) < T_{prob}^{high} \end{cases} \\ T^{var}(t-\Delta t) \quad otherwise \end{cases} \tag{A19}$$

where $\{\eta(t)\} \sim i.i.d.\ \mathrm{U}(0,1)\ \forall t$, T is the desired time-headway base model parameter, T_{max}, T_{max}^{high} and T_{min}^{high} are bounded values, T_{prob} and T_{prob}^{high} are probability threshold values. Eq. (A18) ensures smoother changes of $T^{var}(t)$ based on the comfort threshold, ΔT_{comf} (Zheng et al., 2022).

A.8. Model parameter bounds applied in calibration

Table A.2

Parameter	Lower bound	Upper bound	Parameter	Lower bound	Upper bound
a [m/s^2]	0.5	5.0	ε_x^W	0	1
b [m/s^2]	0.5	5.0	$\varepsilon_{\Delta x}^W$	0	5
s_0 [m]	1	10	$\varepsilon_{\dot{\Delta x}}^W$	0	2
T [s]	0.1	4.0	h_{exp}	0	10
Δv_0 [m/s]	0	10	TC	0.1	5.0
δ	0.1	10	TS_{crit}	0.5	1.5
v_{crit}^{perc}	0	1	ΔTS_{max}	0	2
τ_p^{const} [s]	0	1.5	ΔSA_{max}	0	1
τ_p^{SA} [s]	0	1.5	β_T	0	2
μ_x^T	0	2	γ_{TD}	0	2
$\mu_{\Delta x}^T$	0	2	β_{TD}	−2	0.7
$\mu_{\dot{x},1}^X = \mu_{\Delta x,1}^X$	1/3	1	ΔT_{max}	0	3
$\Delta\mu_{\dot{x},2}^X = \Delta\mu_{\Delta x,2}^X$	0.5	1	T_{prob}	0	1
$\varepsilon_{\dot{x}}^{const}$	0	2	T_{min}^{high}	0.1	4
$\varepsilon_{\Delta x}^{const}$	0	2	ΔT_{min}^{high}	0	3
$\varepsilon_{\dot{\Delta x}}^{const}$	0	2	T_{prob}^{high}	0	1
ε^{SA}	0	2	ΔT_{comf}	0	1
ε_τ	0	5			

Please note that, to avoid imposing constraints in the optimization problem, model parameter values that depend on the value of other parameters, are computed as follows:

$$v_0 = \max\left\{\dot{x}_n^{obs}(t)\right\} + \Delta v_0 \tag{A20}$$

$$v_{crit} = v_0 \cdot v_{crit}^{perc} \tag{A21}$$

$$T_{max} = T + \Delta T_{max} \tag{A22}$$

$$T_{\max}^{high} = T_{min}^{high} + \Delta T_{max}^{high} \tag{A23}$$

$$TS_{max} = TS_{min} + \Delta T_{max} \tag{A24}$$

$$SA_{max} = SA_{min} + \Delta SA_{max} \tag{A25}$$

$$\mu_{\dot{x},2}^X = \mu_{\Delta x,2}^X = \min\left\{\mu_{\dot{x},1}^X, \left(1 - \mu_{\dot{x},1}^X\right)\cdot\Delta\mu_{\dot{x},2}^X\right\} \tag{A26}$$

$$\mu^X_{\dot{x},3} = \mu^X_{\Delta x,3} = 1 - \mu^X_{\dot{x},1} - \mu^X_{\dot{x},2} \tag{A27}$$

Please also note that, in the optimization, $SA_{min} = 0$ was assumed.

Appendix B

B.1. Sensitivity indices of the base models per

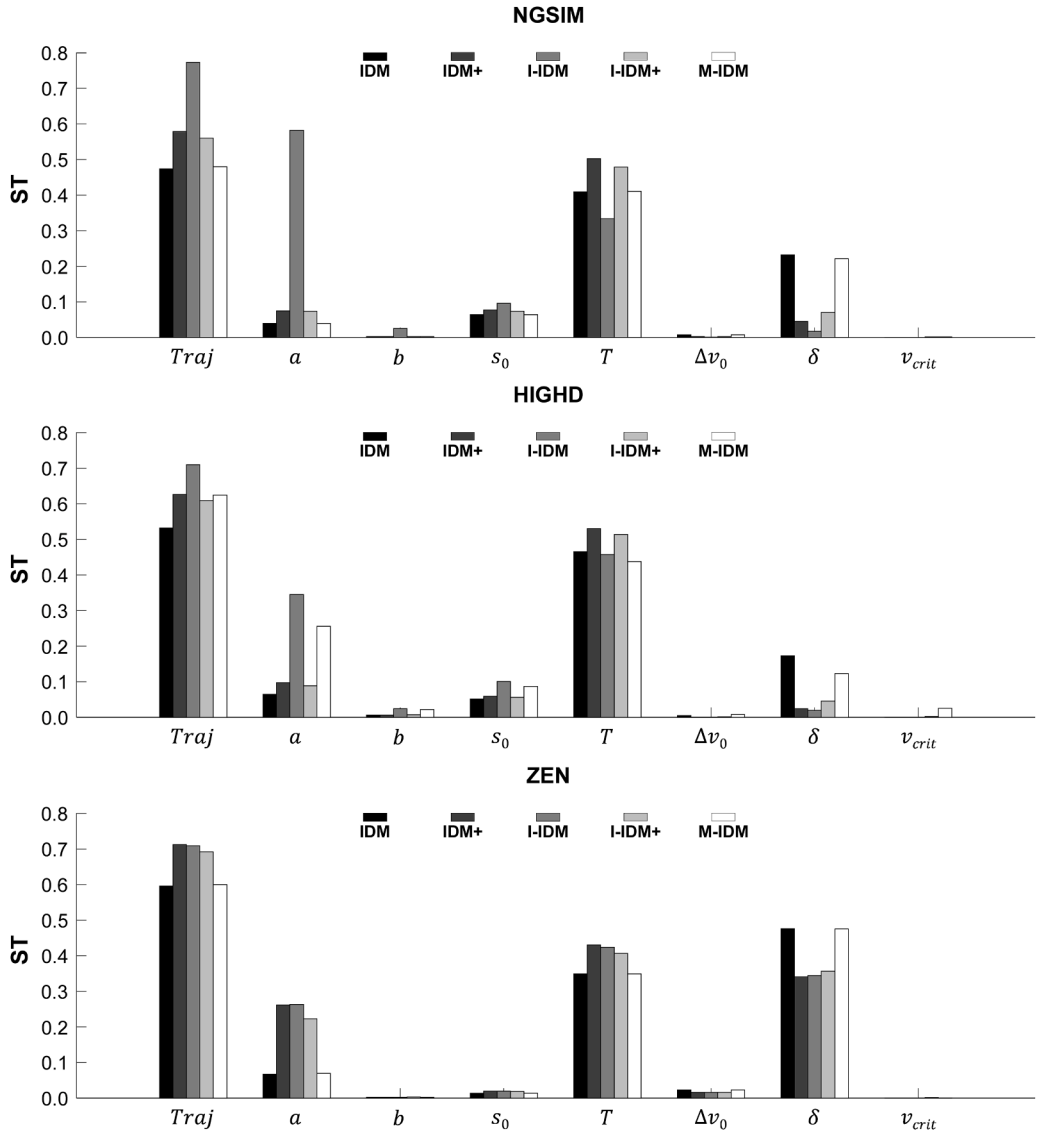


Fig. B1. Total sensitivity indices of each input factor for the five base models, for each dataset.

B.2. Normalised calibration error of the 800 model variants per dataset

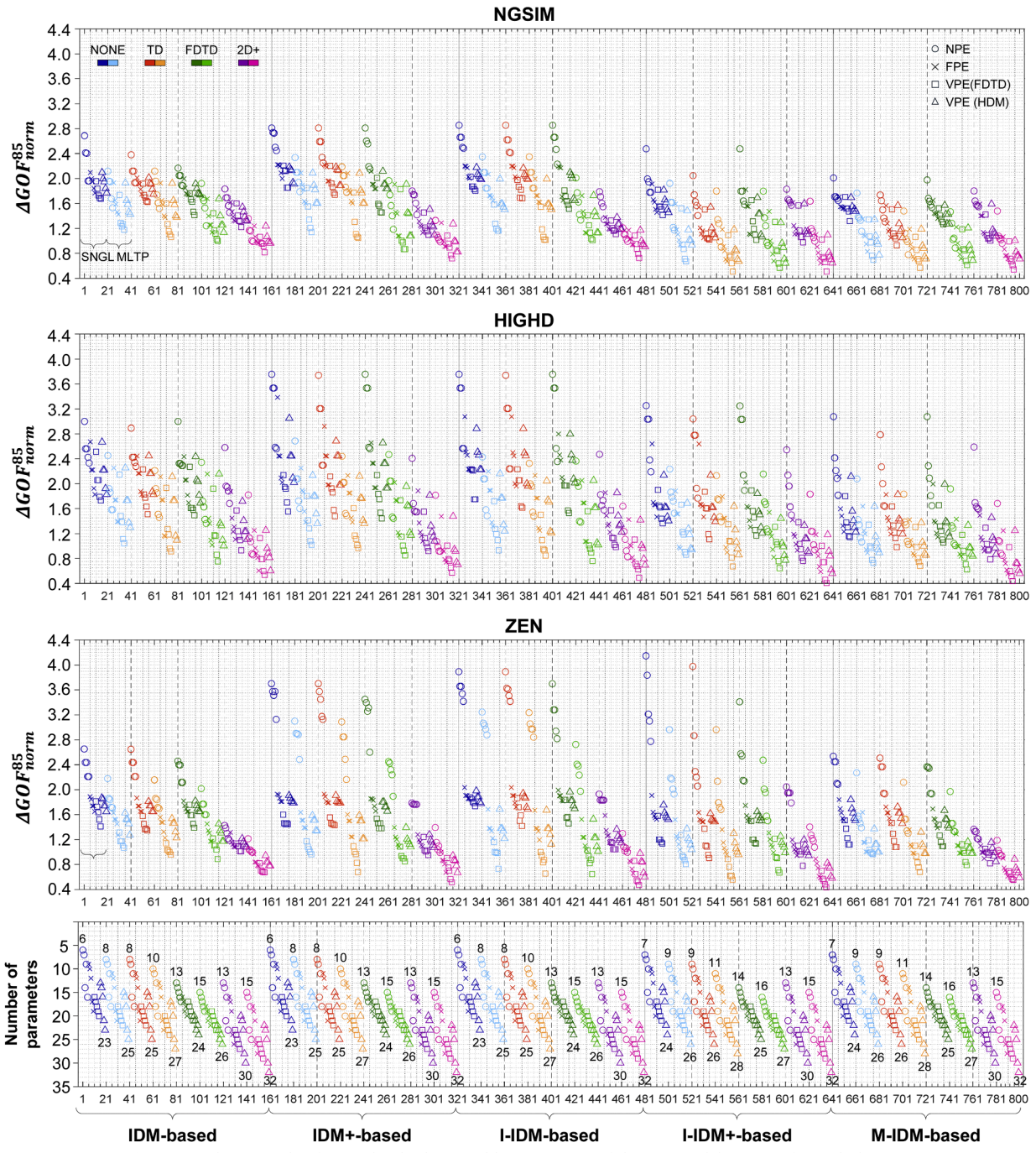


Fig. B2. 85th percentile of normalised relative calibration errors of the 800 model variants for each dataset.

B.3. Verisimilitude measures of the 800 model variants per dataset

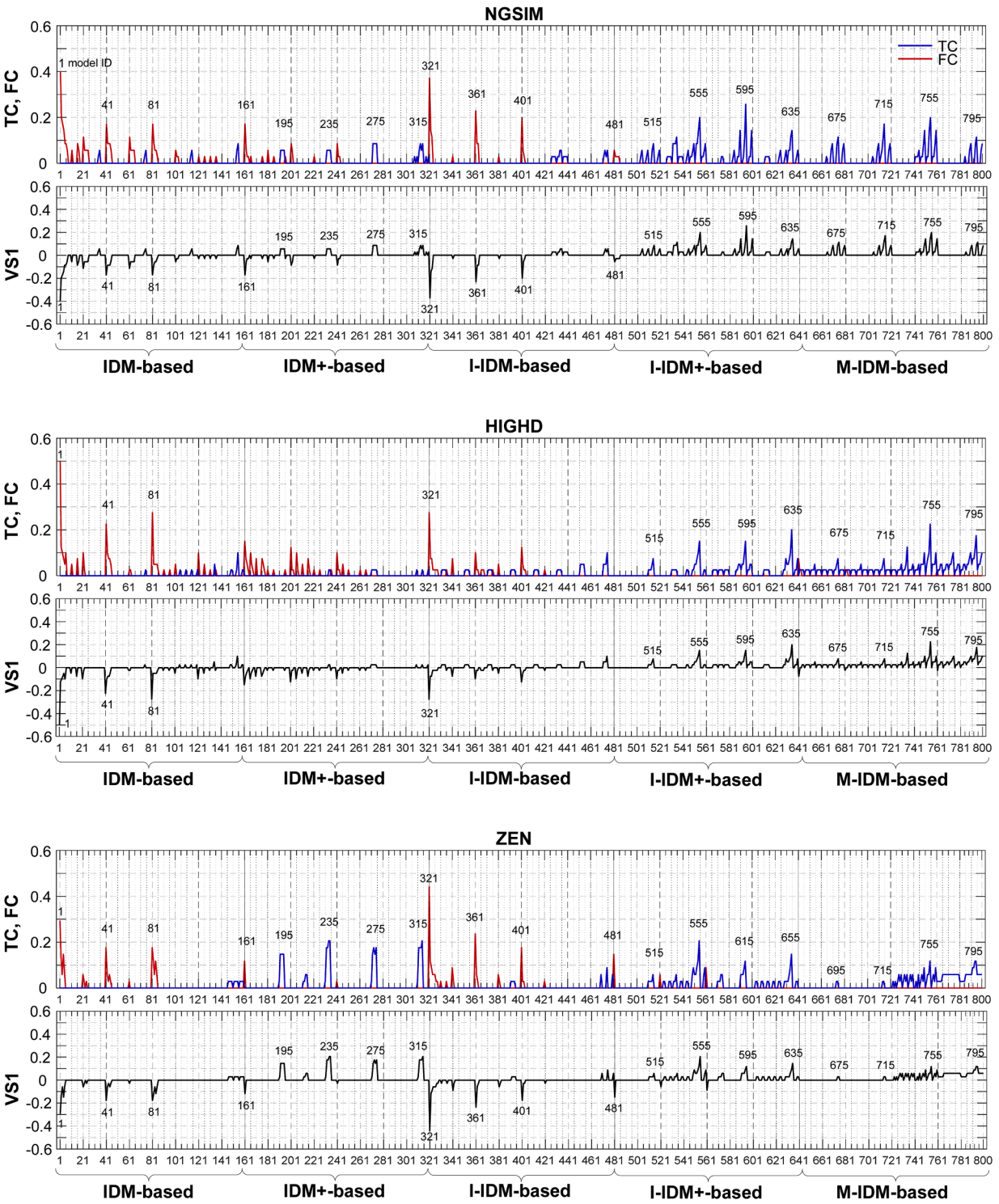


Fig. B3. Verisimilitude measure VS1 of the 800 model variants computed on each dataset.

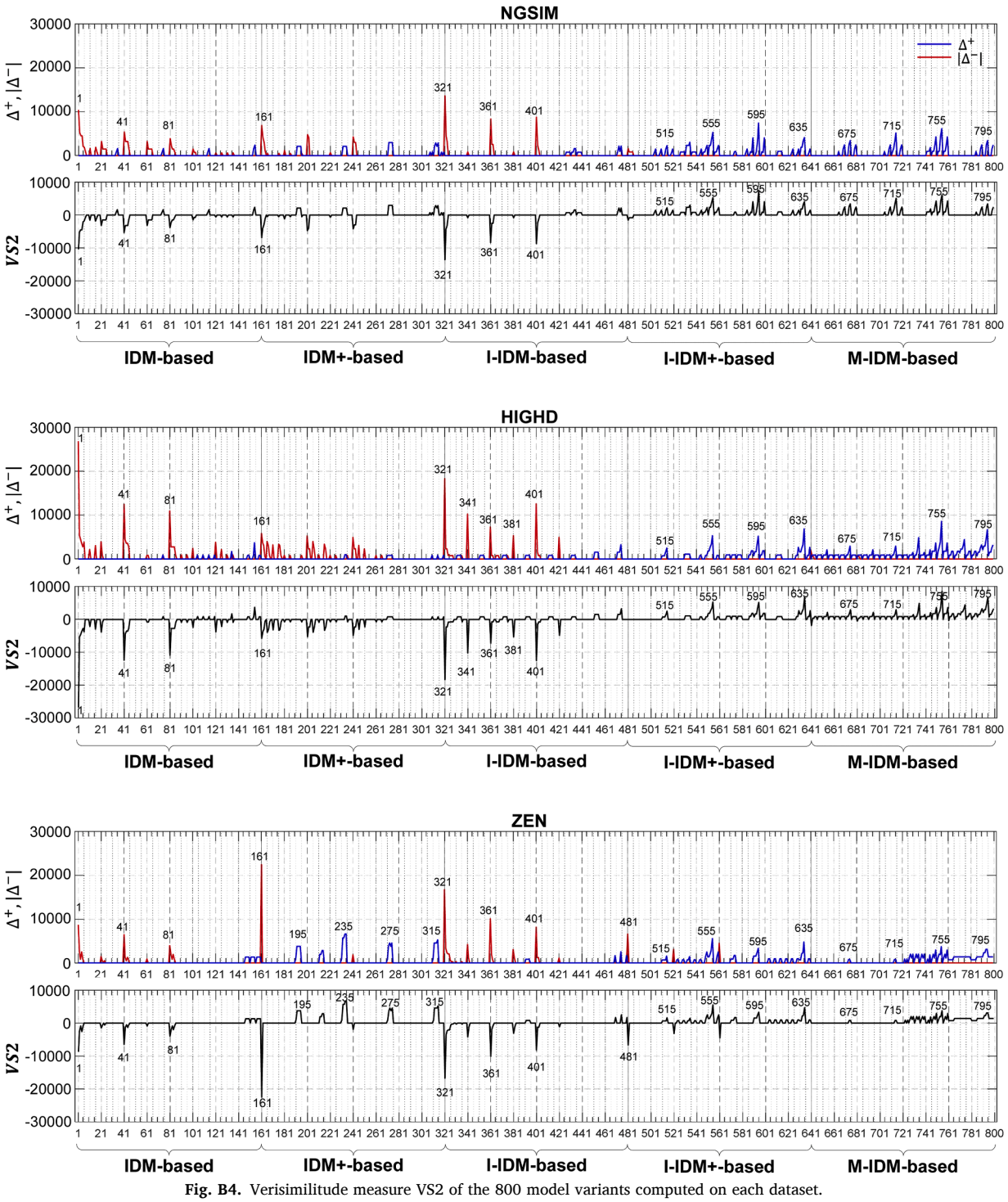


Fig. B4. Verisimilitude measure VS2 of the 800 model variants computed on each dataset.

B.4. Collision rates per base model family

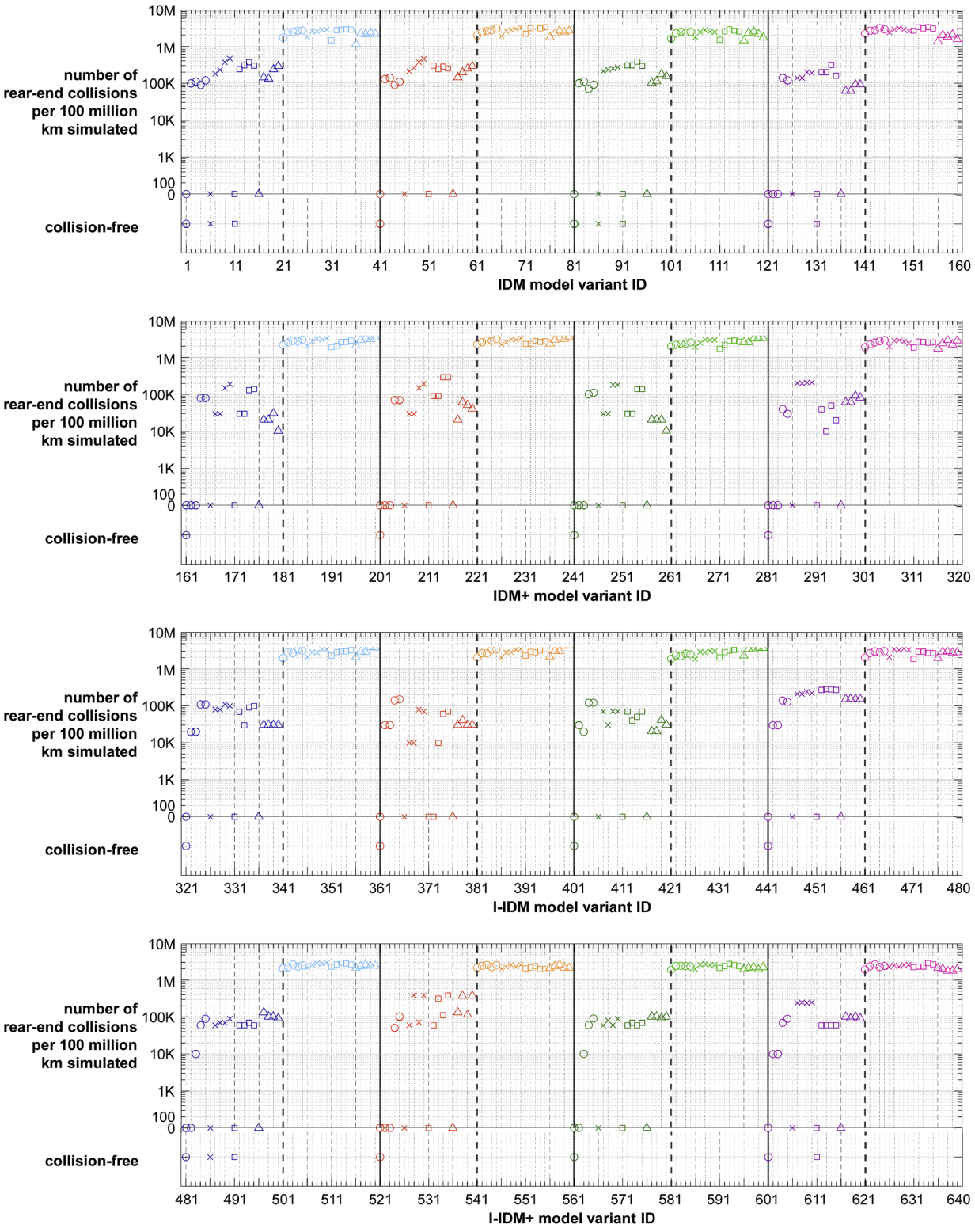


Fig. B5. Results of the two tests for safety-critical driving conditions for base mode variants.

B.5. Accuracy-uncertainty trade-off for the M-IDM model variants per dataset

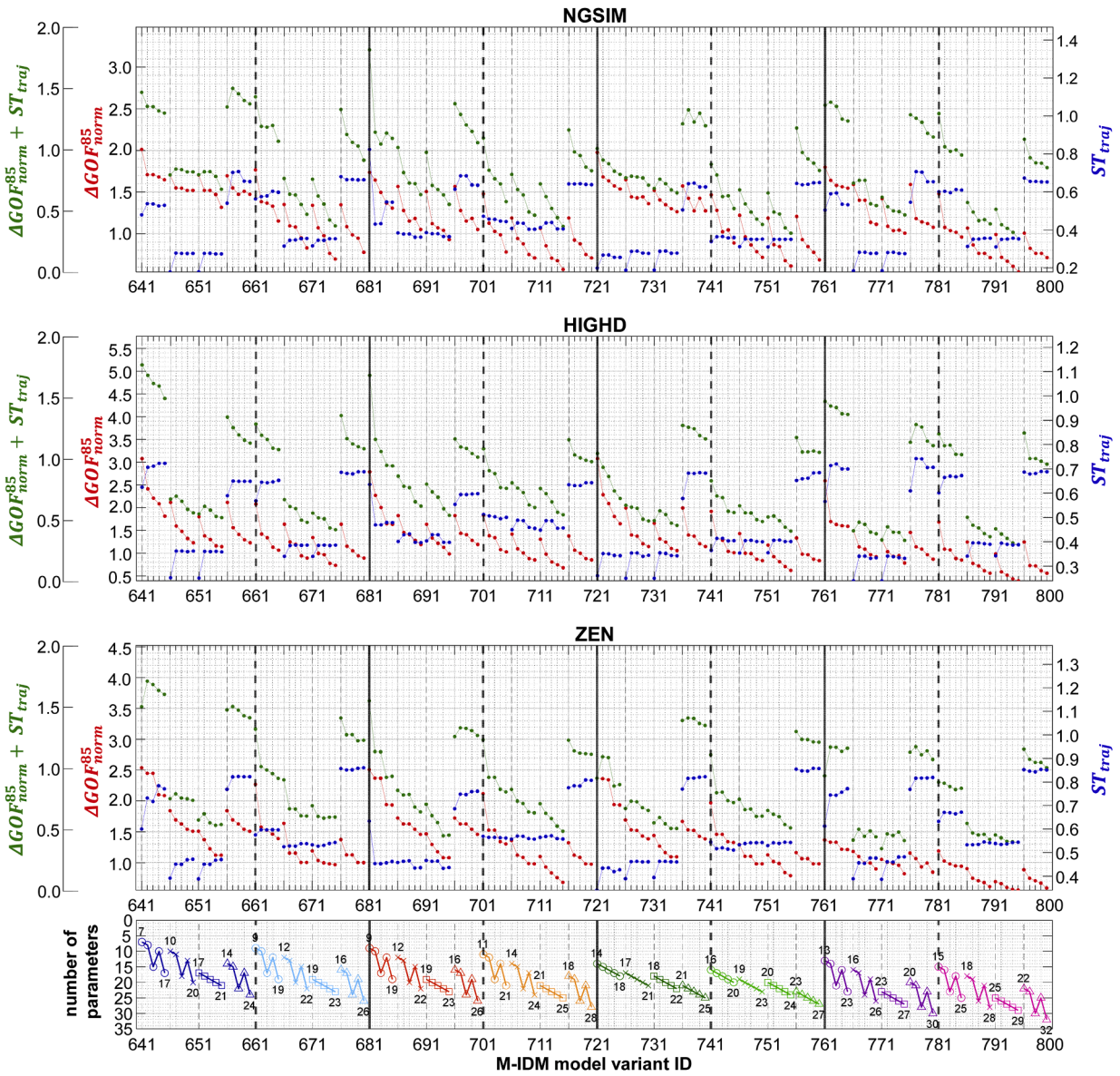


Fig. B6. Trade-off between model accuracy and uncertainty for the base model variants, on the ensemble of all datasets.

B.6. Distributions of model 755 and 795 parameter values producing collisions

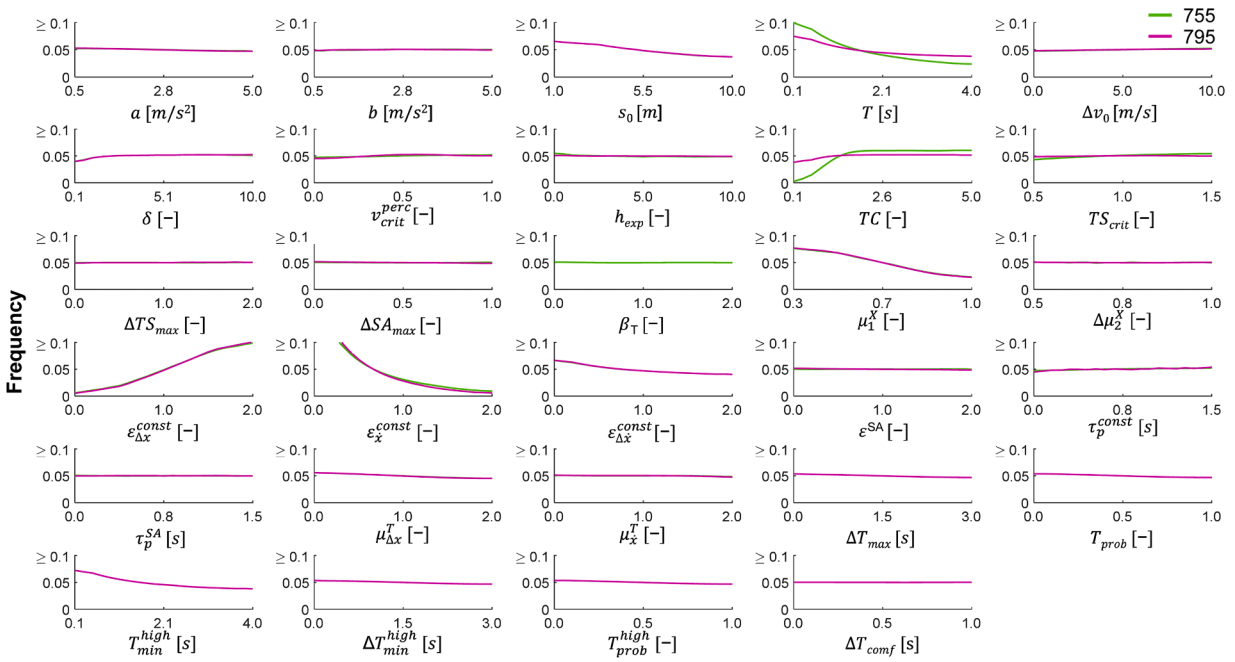


Fig. B7. Probability density functions of model 755 and model 795 parameter values that produce collisions in simulation with input leader trajectory sampled from NGSIM, highD and ZEN datasets.

Appendix C

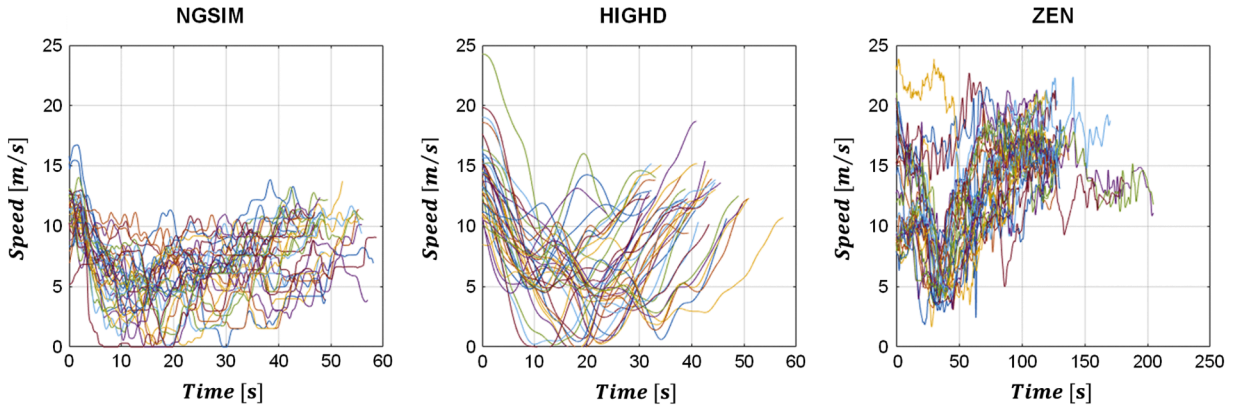


Fig. C1. Speed profiles of ego vehicle selected trajectories from the three datasets.

Data availability

Data and codes are available at the following link: https://github.com/i4Driving.

References

Calvert, S.C., Schakel, W.J., van Lint, J.W.C., 2020. A generic multi-scale framework for microscopic traffic simulation part II anticipation reliance as compensation mechanism for potential task overload. Transport. Res. Part B: Methodol. 140, 42–63.

Chen, X., Yin, J., Qin, G., Tang, K., Wang, Y., Sun, J., 2022. Integrated macro-micro modelling for individual vehicle trajectory reconstruction using fixed and mobile sensor data. Transport. Res. Part C: Emerg.Technol. 145 art. no. 103929.

Chiabaut, N., Buisson, C., Leclercq, L., 2009. Fundamental diagram estimation through passing rate measurements in congestion. IEEE Trans. Intell. Transport. Syst. 10 (2) art. no. 4914846.
Dahiya, G., Asakura, Y., Nakanishi, W., 2022. Analysis of the single-regime speed-density fundamental relationships for varying spatiotemporal resolution using Zen Traffic Data. Asian Transport Stud. 8 art. no. 100066.
Durrani, U., Lee, C., 2024. A new car-following model with incorporation of Markkula's framework of sensorimotor control in sustained motion tasks. Transport. Res. Part B: Methodol. 184, 102969.
Endsley, M.R., 2000. Theoretical underpinnings of situation awareness: a critical review. Situat. Awareness Anal. Measur. 1 (24), 3–32.
Engström, J., Bärgman, J., Nilsson, D., Seppelt, B., Markkula, G., Piccinini, G.B., Victor, T., 2018. Great expectations: a predictive processing account of automobile driving. Theor. Issues. Ergon. Sci. 19 (2), 156–194.
Engström, J., Wei, R., McDonald, A.D., Garcia, A., O'Kelly, M., Johnson, L., 2024. Resolving uncertainty on the fly: modeling adaptive driving behavior as active inference. Front. Neurorobot. 21 (18), 1341750.
Fuller, R., 2000. The task-capability interface model of the driving process. Recherche-Transports-Sécurité 66, 47–57.
Fuller, R., 2005. Towards a general theory of driver behavior. Accid. Anal. Prevent. 37, 461–472.
Fuller, R., 2011. Driver control theory: from task difficulty homeostasis to risk allostasis. Handbook of Traffic Psychology. Elsevier, pp. 13–26.
Homma, T., Saltelli, A., 1996. Importance measures in global sensitivity analysis of nonlinear models. Reliab. Eng. Syst. Saf. 52 (1), 1–17.
Hoogendoorn, S.P., Ossen, S., Schreuder, M., 2006. Empirics of multianticipative car-following behavior. Transp. Res. Rec. 1965, 112–120.
i4Driving, 2022. Integrated 4D driver modelling under uncertainty, https://i4driving.eu/, last accessed 31/07/2025.
Krajewski, R., Bock, J., Kloeker, L., Eckstein, L., 2018. The highD dataset: a drone dataset of naturalistic vehicle trajectories on German highways for validation of highly automated driving systems. 21st International Conference on Intelligent Transportation Systems (ITSC), Maui, HI, USA, 2118–2125.
Markkula, G., Boer, E., Romano, R., Merat, N., 2018. Sustained sensorimotor control as intermittent decisions about prediction errors: computational framework and application to ground vehicle steering. Biol. Cybern. 112 (3), 181–207.
Montanino, M., Punzo, V., 2015. Trajectory data reconstruction and simulation-based validation against macroscopic traffic patterns. Transport. Res. Part B: Methodol. 80, 82–106.
Montanino, M., Zaccaria, G., Punzo, V., 2026. The breakdown of linear string stability: Insights from nonlinear analysis of ACCs and car-following models, and implications on traffic safety. Transportation Research part B.
Munda, G., 2008. Social Multi-Criteria Evaluation For a Sustainable Economy. Springer.
Ngoduy, D., 2015. Linear stability of a generalized multi-anticipative car following model with time delays. Commun. Nonlinear Sci. Numer. Simul. 22 (1–3), 420–426.
O'Neill, R.V., 1973. Error analysis of ecological models. In: Radionuclides in Ecosystems. National Technical Information Service, Springfield, 898–908.
Popper, K., 1963. Conjectures and refutations. The growth of Scientific Knowledge. Routledge & Kegan Paul, New York.
Punzo, V., Montanino, M., Ciuffo, B., 2015. Do we really need to calibrate all the parameters variance-based sensitivity analysis to simplify microscopic traffic flow models. IEEE Trans. Intell. Transport. Syst. 16 (1), 184–193.
Punzo, V., Zheng, Z., Montanino, M., 2021. About calibration of car-following dynamics of automated and human-driven vehicles: methodology, guidelines and codes. Transport. Res. Part C: Emerg.Technol. 128, 103165.
Saifuzzaman, M., Zheng, Z., 2014. Incorporating human factors in car-following models: a review of recent developments and research needs. Transport. Res. Part C: Emerg.Technol. 48, 379–403.
Saifuzzaman, M., Zheng, Z., Mazharul Haque, Md., Washington, S., 2015. Revisiting the Task–Capability Interface model for incorporating human factors into car-following models. Transport. Res. Part B Methodol. 82, 1–19.
Saltelli, A., 2019. A short comment on statistical versus mathematical modelling. Nat Commun 10, 3870.
Saltelli, A., Annoni, P., Azzini, I., Campolongo, F., Ratto, M., Tarantola, S., 2010. Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index. Comput. Phys. Commun. 181 (2), 259–270.
Schakel, W., Knoop, V., Van Arem, B., 2012. Integrated lane change model with relaxation and synchronization. Transp Res Rec (2316), 47–57.
Sharma, A., Zheng, Z., Bhaskar, A., 2019. Is more always better? The impact of vehicular trajectory completeness on car-following model calibration and validation. Transport. Res. Part B: Methodol. 120, 49–75.
Sobol', I.M., 1976. Uniformly distributed sequences with an additional uniform property. U.S.S.R. Comput. Mathem. Mathem. Phys. 16, 236–242.
Tian, J., Jiang, R., Li, G., Treiber, M., Jia, B., Zhu, C., 2016. Improved 2D intelligent driver model in the framework of three-phase traffic theory simulating synchronized flow and concave growth pattern of traffic oscillations. Transport. Res. Part F: Traffic Psychol. Behav. 41, 55–65.
Tichy, P., 1974. On Popper's Definitions of Verisimilitude. British J. Philos. Sci. 25 (2), 155–160.
Treiber, M., Hennecke, A., Helbing, D., 2000. Congested traffic states in empirical observations and microscopic simulations. Phys. Rev. E 62 (2), 1805.
Treiber, M., Kesting, A., 2013. Traffic Flow Dynamics. Data, Models and Simulation. Springer.
Treiber, M., Kesting, A., Helbing, D., 2006. Delays, inaccuracies and anticipation in microscopic traffic models. Phys. A: Stat. Mech. Appl. 360 (1), 71–88.
Turner, M.G., Gardner, R.H., 2015. Introduction to models. In: Turner, M.G., Gardner, R.H. (Eds.), Landscape Ecology in Theory and Practice: Pattern and Process. Springer, New York, NY, pp. 63–95.
UN R157 Amendment 4, 2023, Uniform provisions concerning the approval of vehicles with regard to Automated Lane Keeping Systems. https://unece.org/sites/default/files/2024-07/R157am4e_1.pdf.
Van Hinsbergen, C.P.I.J., Schakel, W.J., Knoop, V.L., van Lint, J.W.C., Hoogendoorn, S.P., 2015. A general framework for calibrating and comparing car-following models. Transportmetrica A: Transport Sci. 11 (5), 420–440.
van Lint, J.W.C., Calvert, S.C., 2018. A generic multi-level framework for microscopic traffic simulation. Theory and an example case in modelling driver distraction. Transport. Res. Part B: Methodol. 117, 63–86.
Wilson, R.E., 2008. Mechanisms for spatio-temporal pattern formation in highway traffic models. Philosoph. Trans. R. Soc. A: Math., Phys. Eng. Sci. 366 (1872), 2017–2032.
Zhang, X., Sun, J., Sun, J., 2025. On the stochastic fundamental diagram: a general micro-macroscopic traffic flow modeling framework. Commun. Transport. Res. 5 art. no. 100163.
Zheng, S.T., Jiang, R., Tian, J., Li, X., Treiber, M., Li, Z.H., Gao, L.D., Jia, B., 2022. Empirical and experimental study on the growth pattern of traffic oscillations upstream of fixed bottleneck and model test. Transport. Res. Part C: Emerg.Technol. 140, 103729.
Zhou, S., Zheng, S., Xu, T., Treiber, M., Tian, J., Jiang, R., 2025. On the calibration of stochastic car following models. Transport. Res. Part B: Methodol. 196, 103224.
Zen Traffic Data, 2024. Hanshin Expressway Co. Ltd., Japan, https://zen-traffic-data.net, last accessed 31/07/2024.