OBJECTIVE FUNCTION DESIGN FOR TRAFFIC MICROSIMULATION CALIBRATION IN SMART-CITY DIGITAL TWINS: EVIDENCE FROM A DETECTOR-BASED STUDY OF THE ANTWERP R1 WEAVING SEGMENT

Yan Li
Xiaong Wie

Traffic microsimulation is central to smart-city transport control because digital twins and decision-support platforms depend on well-calibrated behavioural models before operational strategies are tested in the field. This paper examines how objective function design shapes calibration efficiency and parameter convergence consistency in a detector-based microsimulation setting. The study uses a VISSIM model of a 2.5 km weaving segment on the southbound Antwerp R1 motorway and calibrates 41 driving-behaviour parameters using detector-level speed and headway observations from the morning peak of 10 September 2019. A multifaceted objective function based on the 1-Wasserstein distance is evaluated against the Kolmogorov–Smirnov (K–S) distance and root mean squared relative error (RMSRE), under single-KPI and dual-KPI formulations. Two optimisers are considered: simultaneous perturbation stochastic approximation (SPSA) as the primary high-dimensional method and Bayesian optimisation as an external validation benchmark. In the synthetic experiment, the random-seed noise effect is 1.2%, and the optimisation trajectories approach attainable minima near 1.75% (SPSA) and 1.6% (Bayesian optimisation) when starting from dispersed initial points. In the real-data experiment, the strongest and most balanced behaviour is obtained when the Wasserstein distance is paired with a speed-plus-headway objective. Across the 32 dominant-class parameter instances used for convergence assessment, this setting yields 13 parameters meeting the consistency threshold under SPSA, compared with 5 under the K–S speed-plus-headway formulation. Although speed-only RMSRE also stabilises 13 parameters, it does so under a single-KPI setting that does not preserve balanced performance across KPIs. The results show that, for smart-city traffic digital twins, calibration quality depends not only on the optimiser but on whether the objective function preserves traffic-state heterogeneity and constrains the parameter search with sufficiently rich behavioural information.

Index Terms — smart cities; traffic microsimulation; digital twins; calibration; Wasserstein distance; stochastic optimisation; traffic management

https://doi.org/10.66033/judsc2025-228

INTRODUCTION

Traffic microsimulation models now occupy a central place in smart-city transport analytics because they provide a controlled digital environment in which complex traffic processes can be tested before interventions are implemented in the field. They are routinely used to support motorway and arterial operations, digital twins, pre-deployment assessment of intelligent transportation systems (ITS), and the ex ante evaluation of control strategies whose real-world testing would be costly, risky, or operationally disruptive. This role is especially important for smart-city governance, where traffic management increasingly depends on predictive and simulation-assisted decision support rather than on purely reactive control. In this broader context, microsimulation is not simply a visualisation tool; it is an operational model of urban mobility dynamics that can inform strategies such as coordinated ramp metering and variable speed limits under recurrent congestion and mixed traffic conditions [1, 2].

The usefulness of such models, however, depends fundamentally on calibration quality. A poorly calibrated model may reproduce selected aggregate outputs while still misrepresenting the behavioural mechanisms that actually generate congestion, lane-changing turbulence, shockwave propagation, queue spillback, and bottleneck discharge. This is a well-known problem in traffic flow analysis, where observed aggregate fit does not guarantee behavioural realism. Microsimulation models must therefore be judged not only by whether they match a few network-level indicators, but also by whether their internal behavioural logic remains plausible across the range of states that emerge during a peak-period traffic process. This requirement is especially demanding in weaving segments and merge-diverge systems, where localized interactions among vehicles strongly shape system-level performance and where small behavioural misspecifications can produce large errors in breakdown timing and queue development [7, 8].

The calibration problem is difficult for at least three closely connected reasons. First, traffic microsimulation models are stochastic and computationally expensive, which means that each function evaluation is noisy and each candidate parameter vector can be costly to assess. Second, the parameter spaces are often high-dimensional, with strong interactions among car-following, lane-changing, and route-choice components that make simple manual tuning unreliable. Third, the empirical objective function is itself a modelling decision rather than a neutral technical detail: it determines which discrepancies matter, how stochastic noise is filtered, how multiple key performance indicators (KPIs) are balanced, and which aspects of the traffic process are allowed to remain weakly identified. These difficulties place microsimulation calibration within the broader class of stochastic-search and simulation-based optimization problems, where algorithm choice and objective-function formulation interact strongly and where convergence quality depends on how informative the search signal is for the optimizer [3].

This issue is particularly relevant because much of the calibration literature still relies on pointwise error measures over aggregated time series, such as speed or flow deviations at fixed intervals. Although such measures are easy to compute and interpret, they may suppress the stochastic structure of traffic data and may also overweight one KPI while implicitly downgrading another [4]. In dense motorway environments, where detector observations reflect heterogeneous vehicle interactions and transient states rather than smooth deterministic processes, calibration targets based only on averaged time series can hide important behavioural information. A calibration framework that fits mean values while missing the underlying distribution of speeds or headways may therefore achieve numerical adequacy without fully capturing the operational character of congestion. Distribution-based dissimilarity measures provide a stronger alternative because they preserve more information about heterogeneity, variability, and overlap between simulated and observed traffic states, thereby exposing a richer and potentially more stable signal to the optimizer.

The present study develops that argument in a form appropriate for a smart-city transport journal. Rather than treating calibration as a purely technical preprocessing step, it examines calibration as a core component of

digital traffic-infrastructure reliability. This perspective aligns traffic microsimulation with a wider smart-city literature in which high-dimensional, sensor-informed, and computationally intensive models must be calibrated carefully if they are to support operational decisions [5, 6]. Comparable issues arise in building energy management, model-predictive control, HVAC learning systems, fault diagnosis, AI-based forecasting, retrofit decision support, demand-response optimization, IoT-enabled infrastructure management, and thermal-comfort coordination, where model usefulness depends not only on predictive accuracy but also on robustness, uncertainty handling, and the balance among multiple performance criteria [9]. In this sense, the paper situates traffic calibration within a broader family of smart-city model-governance problems: if the objective function is poorly designed, even a sophisticated optimizer may be guided toward unstable or weakly interpretable parameter values.

Accordingly, this paper presents a complete, detector-based analysis of how objective-function formulation influences calibration efficiency and parameter convergence consistency in a motorway weaving corridor that is operationally relevant to urban and metropolitan traffic management. The central question is therefore not only *which* optimizer to use, but *how to encode the calibration target* so that the optimizer receives enough behavioural information to guide parameters toward stable and balanced values. The manuscript is organised around three specific questions:

1. Can high-dimensional microsimulation models be calibrated effectively without prior global sensitivity screening?
2. Does calibrating multiple KPIs improve parameter convergence consistency?
3. Does a probabilistic distribution-based distance, particularly the 1-Wasserstein distance, produce a better and more balanced calibration than K–S or RMSRE?

STUDY CONTEXT AND RESEARCH DESIGN

Case-study corridor

The empirical setting is a VISSIM model of a 2.5 km complex weaving segment on the southbound Antwerp R1 motorway in Belgium. This corridor is a particularly demanding calibration environment because it combines recurrent peak-period congestion with intense merging, diverging, and lane-changing interactions in a relatively short spatial section. At the merge from the E313 motorway, the mainline consists of five lanes, with additional lanes allocated to on- and off-ramp movements [10]. Downstream, after the Berchem off-ramp, the corridor separates into two main directional branches: two lanes continue toward the E19 and A12 in the direction of Brussels, while three lanes continue on the R1 toward the Kennedy Tunnel. The section between the E313 on-ramp and the R1/E19–A12 divergence therefore operates as a recurrent bottleneck in which heavy demand and strong weaving pressure interact to create unstable traffic states [11].

From an analytical standpoint, this corridor is well suited to the study's objective because it is not a simple merge or uniform freeway segment. Instead, it represents a location where microsimulation calibration must capture both local behavioural turbulence and broader congestion formation. In such environments, detector outputs are highly sensitive to the interaction of car-following, discretionary and mandatory lane changes, and local class composition [12]. This makes the site a strong test case for examining whether richer objective functions produce more stable parameter convergence than conventional formulations based on aggregated errors alone [13].

Loop detectors are positioned across all lanes, especially before and after the principal merging and diverging influence zones. Their placement makes it possible to observe traffic behaviour at multiple points along the

corridor and to evaluate how well the simulation reproduces evolving conditions before, within, and after the weaving section [14]. The empirical reference period is the morning peak on 10 September 2019, from 07:00 to 11:00. During this period, congestion emerged endogenously, without reported accidents and without spillback from the off-ramp, which is analytically important because it allows the calibration exercise to focus on the corridor's intrinsic operational dynamics rather than on exogenous disruption. The available KPIs are the individual speeds and headways of vehicles passing the detectors, providing a comparatively rich behavioural dataset for calibration.

Vehicle classes and calibration scope

The model represents traffic heterogeneity through five vehicle classes: slow passenger cars, regular passenger cars, fast passenger cars, trucks, and trailers. This class structure is important because weaving operations are highly sensitive to differences in acceleration capability, desired speed, gap acceptance, and lane-changing behaviour across vehicle types. In the study corridor, regular passenger cars and trailers dominate the fleet composition, meaning that the behaviour of these classes is especially influential in shaping aggregate traffic conditions as well as detector-level distributions of speeds and headways [15].

Each class is described by 16 behavioural parameters, comprising 13 parameters associated with the Wiedemann 99 car-following and lane-changing logic and 3 connector-specific lane-changing distances. If all classes were calibrated independently, the resulting problem would involve 80 parameters, which would be computationally burdensome and statistically difficult to identify in a stochastic simulation setting. To preserve tractability while still maintaining behavioural realism for the most influential vehicle groups, the calibration strategy is deliberately selective. It updates the car-following and lane-changing parameters of the dominant classes, namely regular passenger cars and trailers, while calibrating the lane-changing distance parameters for all five classes. This yields a final calibration dimension of 41 parameters [16].

This design choice reflects a compromise between behavioural richness and optimization feasibility. The parameter space remains high-dimensional enough to test whether the calibration framework can handle realistic microsimulation complexity, yet it avoids an unnecessarily expansive formulation in which weakly influential classes consume computational effort without proportionate informational gain [17]. In practical terms, the study therefore asks not whether every possible behavioural degree of freedom can be estimated, but whether a strategically chosen high-dimensional subset can be calibrated consistently and efficiently under realistic detector-based conditions.

Two-stage calibration logic

The study follows a two-stage calibration logic. In the first stage, origin–destination flows are calibrated so that the model reproduces the corridor's demand structure and is capable of generating the observed breakdown phenomenon. This stage is essential because behavioural calibration cannot be meaningful if the underlying demand pattern is misrepresented. A model with incorrect route allocation or inflow structure may fit some detector outputs only by compensating through distorted behavioural parameters, thereby confounding demand representation with driving logic.

In the second stage, the driving-behaviour parameters are calibrated so that the simulated detector outputs reproduce observed traffic characteristics at the level of individual speeds and headways [18]. The present paper concentrates on this second stage and, more specifically, on the role of objective-function design within it. This focus is deliberate. Once the demand structure is fixed well enough to reproduce the corridor's general loading pattern, the remaining question is how behavioural calibration should be formulated so that the optimizer is guided by a target rich enough to identify plausible and stable parameter values [19].

The research design therefore treats calibration not as a one-dimensional minimization exercise, but as a structured comparison of alternative objective-function formulations under a common empirical setting [13]. By using the same corridor, the same detector data, the same behavioural parameter space, and the same staged calibration logic, the study isolates the influence of objective-function design on convergence consistency and calibration balance. In this way, the case-study corridor serves not only as an application context, but also as an experimental environment for testing how different calibration targets shape the behaviour of the optimization process itself.

METHODOLOGY

Optimisation problem

Let $\theta \in \mathbb{R}^P$ denote the calibration parameter vector, with $P = 41$. The calibration problem is written as

$$\theta^\star = \arg\min_{\theta} Z(\theta) \qquad \text{s.t.} \qquad \theta_p^{\min} \leq \theta_p \leq \theta_p^{\max}, \quad p = 1, \ldots, P. \tag{1}$$

The objective function $Z(\theta)$ measures the average dissimilarity between empirical and simulated KPI distributions across all detector locations.

Multifaceted detector-based objective function

For each detector d, let $D_{S,d}$ and $D_{H,d}$ denote the dissimilarity between the simulated and empirical distributions of speed and headway, respectively. The study evaluates objective functions built from one or both of these KPI facets:

$$Z(\theta) = \frac{\sum_{d=1}^{D} (\alpha_S D_{S,d} + \alpha_H D_{H,d})}{2D(\alpha_S + \alpha_H)}, \tag{2}$$

where α_S and α_H are KPI weights. Three practical formulations are used:

1. speed only (S),
2. headway only (H),
3. speed and headway jointly (S+H), weighted equally.

This design allows the study to isolate whether adding facets improves the identifiability of behaviourally meaningful parameters.

Dissimilarity metrics

Three dissimilarity metrics are compared.

1-Wasserstein distance. The Wasserstein distance measures the total area between empirical cumulative distribution functions. It therefore reflects discrepancies across the full support of the distribution and is well suited to detector data that combine free flow, transitional states, and congestion.

Kolmogorov–Smirnov distance. The K–S distance captures only the maximum vertical difference between the two cumulative distribution functions. It is therefore more local and can overemphasise a single region of the distribution.

Root mean squared relative error. RMSRE is computed on aggregated time-series outputs (five-minute averages). It is intuitive and commonly used, but it suppresses the stochastic structure of the underlying detector data.

Optimisation algorithms

Simultaneous perturbation stochastic approximation (SPSA). SPSA is used as the primary optimiser because the calibration problem is high-dimensional and noisy. It updates the current estimate recursively:

$$\hat{\theta}_{k+1} = \hat{\theta}_k - a_k \hat{g}_k(\hat{\theta}_k), \tag{3}$$

where the gradient estimate is formed from only two perturbed function evaluations:

$$\hat{g}_k(\hat{\theta}_k) = \frac{Z(\hat{\theta}_k + c_k\Delta_k) - Z(\hat{\theta}_k - c_k\Delta_k)}{2c_k\Delta_k}. \tag{4}$$

This makes the cost of each iteration effectively independent of the number of parameters.

Bayesian optimisation. Bayesian optimisation is used as a validation benchmark. It is better suited to lower- or moderate-dimensional settings, but it remains informative here as an external check on the central conclusions about objective-function design.

Synthetic and real-data experiments

The study uses two experimental settings.

Synthetic scenario. A simulated scenario is calibrated against another simulated scenario generated by the same model with different parameter values. This creates a setting in which the true optimum is known in principle, allowing the study to distinguish optimisation limitations from measurement uncertainty.

Real-world scenario. The 10 September 2019 detector observations are used as the reference target. Each objective-function configuration is run 10 times because the optimisation procedures are stochastic.

Convergence consistency criterion

For each parameter and each experiment, the calibration produces a sample of 10 estimated values. Let μ and σ be the sample mean and sample standard deviation. The study defines the ratio

$$\frac{\sigma}{\mu}$$

as a measure of convergence spread. A parameter is treated as consistently convergent when

$$\frac{\sigma}{\mu} \leq 0.10.$$

This threshold is used throughout the real-data analysis.

CALIBRATION PARAMETERS CONSIDERED

Table 1: Behavioural parameter set used for dominant-class calibration.

ID	*Parameter*	*Operational meaning*
1	Max. lookahead distance	Maximum forward distance over which drivers react to surrounding vehicles.
2	Standstill distance (CC0)	Desired average standstill spacing between vehicles.
3	Following variation (CC2)	Upper bound of desired safety distance before a driver intentionally closes in.
4	Threshold of entering following (CC3)	Time threshold governing entry into following mode.
5	Negative/positive following threshold (CC4/CC5)	Speed-difference threshold at which a driver accelerates during following.
6	Oscillation acceleration (CC7)	Acceleration oscillation in the Wiedemann process.
7	Minimum headway front/rear	Lower bound for post-lane-change safety distance.
8	Safety distance reduction factor	Reduction factor used to raise gap-acceptance probability.
9	Cooperative deceleration	Maximum cooperative deceleration to allow a preceding vehicle to merge.
10	Max. deceleration	Maximum deceleration accepted during lane changing.
11	Accepted deceleration	Accepted deceleration threshold during lane changing.
12	Deceleration reduction distance	Distance over which deceleration changes approaching the destination connector.
13	Mean headway distribution (CC1)	Mean of the speed-dependent desired safety-distance component.
14–16	Lane-changing distances (LC1–LC3)	Distances upstream of the Berchem, Brussels, and Kennedy Tunnel connectors at which drivers start searching for a suitable destination-lane gap.

RESULTS

Synthetic scenario: identifiable signal without prior sensitivity screening

The synthetic experiment establishes a controlled benchmark. Because the target output is generated by the same simulator, the theoretical global minimum is zero. In practice, this minimum is unattainable because of simulation noise. The random-seed noise effect (RSNE) is reported as 1.2%.

The synthetic runs yield two substantive findings. First, starting from the true parameter set with a deliberately small step size shows that the objective can drift from zero toward approximately 1.2%, demonstrating how simulation noise perturbs the search even at the optimum. Second, when the search begins from dispersed initial points, the algorithms converge toward the vicinity of the attainable global minimum: approximately 1.75% for SPSA and 1.6% for Bayesian optimisation. This is strong evidence that prior global sensitivity analysis is not a prerequisite for useful calibration in this setting. Parameters that matter can be detected through their local response to the objective function during optimisation itself.

The synthetic analysis also shows that not all parameters converge to the true optimum with equal stability. The clearest and most robust convergence is observed for LC3, while the cooperative, maximum, and accepted deceleration parameters, along with LC1 and LC2, tend to converge toward the neighbourhood of the optimum with moderate stability. Parameters such as the lookahead and several core following thresholds are more strongly affected by non-linearity and parameter interaction, and therefore exhibit multiple stable solutions.

Real-world scenario: KPI choice and metric choice both matter

The real-data experiment compares three metrics (Wasserstein, K–S, RMSRE) under three objective-function designs (S, H, S+H). The empirical pattern is consistent and operationally important:

- calibrating to a single KPI improves that KPI but can worsen the other;
- a joint speed-plus-headway objective reduces that trade-off;
- the Wasserstein distance provides the most balanced behaviour because it reflects the full KPI distributions rather than a single extreme discrepancy or a smoothed average.

In the source experiments, headway-only calibration increases speed error, while speed-only calibration increases headway error. This matters for smart-city digital twins, because traffic operations are not evaluated on one observable alone: robust corridor management depends on a model that captures both mobility and interaction quality.

The reported comparison further shows that when data exhibit substantial variability—particularly regular passenger-car headways—the Wasserstein distance is more reliable than K–S and RMSRE. The study also notes that speed-only RMSRE can achieve a speed fit comparable to the Wasserstein S+H formulation, but this occurs in a single-KPI design and therefore does not provide the same balanced constraint on the parameter space.

Average convergence stability in the SPSA experiments

Table 2 condenses the source paper's parameter-level convergence tables by reporting the average σ/μ ratio across the 16 parameter IDs for each dominant class. Lower values indicate stronger convergence consistency.

Table 2: Average σ/μ ratios across the 16 dominant-parameter IDs in the SPSA experiments.

Class	*Wasserstein*			*K–S*			*RMSRE*		
	S	*H*	*S+H*	*S*	*H*	*S+H*	*S*	*H*	*S+H*
Passenger cars	0.27	0.27	*0.17*	0.29	0.27	0.32	*0.17*	0.27	0.21
Trailers	0.25	0.22	*0.18*	0.25	0.26	0.25	*0.17*	0.21	0.21

The averages above are reproduced from the source study's reported parameter-consistency tables. Lower values indicate tighter clustering of repeated calibration outcomes.

Two points stand out. First, among the multifaceted (*S+H*) formulations, Wasserstein gives the best average consistency for both dominant classes (0.17 for passenger cars and 0.18 for trailers). Second, RMSRE is strongest only in the speed-only setting, which is precisely the configuration that fails to guarantee balanced performance across KPIs.

Threshold-based consistency counts

To make the stability pattern easier to interpret, Table 3 converts the source σ/μ tables into counts of parameters satisfying the study's consistency threshold ($\sigma/\mu \leq 0.10$), aggregated across the 32 dominant-class parameter instances considered in the convergence analysis.

Table 3: Number of dominant-class parameter instances meeting the consistency threshold ($\sigma/\mu \leq 0.10$).

Configuration	*S*	*H*	*S+H*
SPSA, Wasserstein	9	8	*13*
SPSA, K–S	6	6	5
SPSA, RMSRE	*13*	8	9
Bayesian optimisation, Wasserstein only	6	5	*7*

Counts are calculated directly from the source study's reported σ/μ values using the stated threshold of 0.10.

This table clarifies the central result. Under SPSA, the *Wasserstein S+H* configuration stabilises 13 parameter instances while preserving a dual-KPI formulation. The only configuration matching that count is *RMSRE S*, which does so under a speed-only objective. In practical terms, the source evidence supports the claim that the Wasserstein metric is the most suitable choice when the modeller needs both behavioural stability and a balanced trade-off across KPIs.

Bayesian optimisation as a validation check

The Bayesian optimisation runs serve as a validation layer rather than the primary basis of inference. They confirm the same directional finding: within the Wasserstein experiments, the *S+H* configuration again yields the best average stability for passenger cars (0.19 versus 0.23 and 0.24 for S and H) and the highest threshold-based consistency count overall (7, versus 6 and 5 for S and H).

The study also reports that Bayesian optimisation shows comparable overall behaviour to SPSA in the Wasserstein experiments, but SPSA performs better for trailer speed improvement. Specifically, SPSA reduces trailer speed error in the speed-only and speed-plus-headway scenarios from 19.5% to an average of 16.3%, whereas Bayesian optimisation reduces it to 18.5%. This difference is consistent with the weaker trailer-parameter stability observed in the Bayesian optimisation validation table.

DISCUSSION

Why the multifaceted Wasserstein formulation is stronger

The principal advantage of the multifaceted Wasserstein formulation is not merely that it is mathematically elegant, but that it is substantially richer in the information it passes to the calibration algorithm. In a stochastic microsimulation setting, the optimizer does not observe the traffic process directly; it only receives a compressed signal through the objective function. The quality of calibration therefore depends heavily on whether that signal preserves the heterogeneity of the observed system or collapses it into a narrow summary statistic. By measuring the area between cumulative distributions, the 1-Wasserstein distance captures discrepancies across the full support of the detector observations, rather than focusing on a single local deviation or on an average trend alone. This makes it especially appropriate for motorway weaving

segments, where free flow, synchronized traffic, incipient breakdown, and queue discharge can coexist within the same observation window, and where different behavioural parameters are activated under different operating states [5, 6].

This broader informational content matters because high-dimensional microsimulation calibration is fundamentally a stochastic-search problem. In such problems, an optimizer can only converge reliably if the objective function provides gradients or directional cues that remain informative in the presence of simulation noise. The stochastic approximation literature has long emphasized that search efficiency depends not only on the algorithm itself, but also on the signal quality embedded in the loss function [3, 4]. When the calibration target is distributional rather than purely pointwise, the optimizer is exposed to more of the structure present in the empirical traffic data. In practical terms, this means that parameters related to car-following, lane changing, and local interaction are less likely to remain weakly identified merely because the objective function averages away the behavioural states in which they matter most.

By contrast, the K–S statistic concentrates only on the single maximum discrepancy between cumulative distributions. While useful as a diagnostic of whether two distributions diverge sharply at one point, it gives relatively little weight to the rest of the behavioural landscape represented in the data. In a weaving segment, where multiple traffic regimes coexist and where mismatches may be distributed across many parts of the speed or headway distribution, this can lead to an objective function that is overly selective in what it penalizes. RMSRE introduces a different but equally important limitation. When applied to smoothed time series, it suppresses the stochastic structure of the observations and privileges agreement in averaged trends over agreement in the underlying traffic-state variability. The result is that either K–S or RMSRE may provide an optimizer with a thinner and less behaviourally representative signal than the Wasserstein formulation. The empirical findings of the study are therefore consistent with what traffic-flow theory would lead one to expect: weaker convergence consistency, poorer cross-KPI balance, or both, when the calibration target does not preserve enough of the heterogeneity embedded in the detector data [7, 8].

The superiority of the multifaceted Wasserstein formulation is also important because it reduces the risk of compensatory miscalibration. In high-dimensional parameter spaces, a weak objective function may allow one set of parameters to offset the errors produced by another, producing acceptable fit for a selected KPI while distorting the underlying behavioural logic of the model. A richer distribution-based objective reduces this possibility by forcing the simulation to match the observed system across a wider range of traffic states [8]. This does not mean that the Wasserstein formulation fully resolves identifiability problems, but it does mean that it constrains the admissible parameter region more effectively than narrower targets. For smart-city transport modelling, that is a substantial methodological advantage, because it increases the likelihood that a calibrated digital model is not only numerically acceptable but behaviourally credible.

Implications for smart-city transport operations

For smart-city digital twins, the technical lesson is straightforward: calibration objectives must be aligned with the informational requirements of the intended operational use. If a digital twin is expected to support lane-management decisions, variable speed limits, coordinated ramp metering, or bottleneck-control strategies, then it must reproduce more than average speed or a single aggregate throughput indicator. It must also reproduce the interaction structure through which those interventions actually operate, including local headway formation, lane-changing turbulence, and the transitional dynamics that govern breakdown onset and recovery [1, 2]. A calibration target that ignores these behavioural signatures may yield a model that appears adequate in summary statistics while remaining unreliable for control-oriented policy testing.

This point is especially important because smart-city transport systems increasingly depend on simulation-

assisted experimentation before field deployment. Microsimulation models are often used as the analytic core of digital twins through which candidate policies are screened, compared, and stress-tested. In that setting, calibration is not a preliminary housekeeping task; it is a foundational condition for policy credibility. A corridor-management twin calibrated only to speed, for example, may retain too much freedom in the parts of the parameter space that actually govern local interactions. Such a model may still reproduce a speed profile reasonably well, but it may do so for the wrong behavioural reasons. That reduces confidence in any downstream experiment involving lane allocation, incident management, demand re-routing, or intelligent transport control [7, 8].

There is also a broader smart-city implication here that extends beyond transport. In many digitally coordinated urban systems, the objective function used for calibration or control shapes the behaviour of the entire decision-support architecture. This is evident in model-predictive control for building cooling, where the balance between cost and comfort depends on how the control objective is formulated [9]; in neural-network HVAC control, where control performance depends on the model learning the right operational relationships [10]; in explainable fault diagnosis, where interpretable model behaviour matters as much as raw predictive accuracy [11]; and in intelligent energy management, where optimization targets influence whether the system privileges robustness, flexibility, efficiency, or user comfort [12]. The present findings fit naturally within this wider literature: a model is only as operationally useful as the target through which it has been tuned.

The same logic applies to digitally networked infrastructure more generally. Smart-city platforms increasingly combine sensing, optimization, and control across transport, buildings, and energy systems, often under uncertainty and in real time. In such systems, richer objective formulations can improve the interpretability and trustworthiness of the resulting model behaviour, while overly narrow targets may produce brittle calibration that performs well only under the specific metric it was trained against. This is closely related to current work on retrofit decision support, IoT-enabled building infrastructure, and intelligent energy control, all of which emphasize that digital decision systems must be calibrated against objectives that reflect the true operational complexity of the environment they are intended to manage [14, 16, 17, 18]. The implication for smart-city transport is clear: calibration design should be treated as part of infrastructure governance, not merely as a technical optimization detail.

Methodological limits and future work

The study also identifies an important unresolved issue. Matching full-period distributions neglects the temporal ordering of traffic states. When both free flow and congestion occur within the same observation period, the mean, variance, and shape of the underlying process can change over time, and an apparently good distributional fit may still conceal temporal mismatches in the onset, duration, or dissipation of congestion. In other words, two simulations may generate similar full-period speed or headway distributions while differing materially in when those states occur. This limitation does not invalidate the present findings, because the study's results still show that distribution-based targets are stronger than narrower alternatives under otherwise identical conditions. It does, however, define a clear next step for methodological development [13].

A particularly important direction for future research is the design of time-conditioned or regime-aware objective functions layered on top of the present distributional framework. Instead of evaluating one distribution for an entire peak period, future work could compare distributions within specific congestion regimes, moving windows, or state-classified phases of the traffic process. This would preserve the informational benefits of the Wasserstein approach while restoring sensitivity to temporal sequencing. Such extensions could help determine whether the observed gains in convergence consistency persist when the objective function is required to match not only the statistical composition of traffic states, but also their timing and transitions [15].

Additional extensions should also broaden the empirical and operational scope of the analysis. First, multi-day validation across different demand conditions is necessary to determine whether the stronger convergence properties observed here are robust beyond a single peak period. Second, the incorporation of additional KPIs such as occupancy and flow would test whether the advantages of the multifaceted Wasserstein formulation scale to a more comprehensive detector-based calibration setting. Third, future work should directly assess how calibration choices affect the estimated performance of downstream smart-city control strategies [19]. A natural next step would be to compare whether models calibrated with RMSRE, K–S, and Wasserstein objectives produce systematically different conclusions about the benefits of ramp metering, variable speed limits, coordinated motorway management, or other ITS interventions in digital-twin policy experiments. Only through that final step can the field fully connect calibration methodology to the practical reliability of smart-city transport decision support.

CONCLUSION

This paper demonstrates that, in detector-based traffic microsimulation calibration for smart-city digital twins, objective function design is a first-order methodological choice.

Three conclusions follow from the empirical analysis:

1. *Prior global sensitivity analysis is not required* for useful high-dimensional calibration in this setting. The synthetic experiment shows that influential parameters can be guided by the information embedded in the objective function itself.
2. *Single-KPI calibration is insufficient* when the goal is a credible operational model. It improves one KPI at the risk of degrading another.
3. *The 1-Wasserstein distance is the most robust basis for a multifaceted detector-level objective function.* In the real-data SPSA experiments, the Wasserstein speed-plus-headway formulation provides the strongest dual-KPI balance and stabilises 13 of the 32 dominant-class parameter instances under the study's consistency criterion, compared with only 5 for the K–S speed-plus-headway formulation.

For urban and metropolitan transport modelling, the broader implication is that digital twins should be calibrated with objective functions rich enough to represent traffic heterogeneity, not merely convenient enough to optimise quickly. In that sense, the study is not only about calibration efficiency; it is about building trustworthy smart-city simulation infrastructure.

REFERENCES

[1] Papamichail, I., Kotsialos, A., Margonis, I., Papageorgiou, M.: Coordinated ramp metering for freeway networks—a model-predictive hierarchical control approach. *Transportation Research Part C: Emerging Technologies 18*(3), 311–331 (2010)

[2] Li, D., Deng, L., Sun, Y., Song, Y.: Hybrid approach for variable speed limit implementation and application to mixed traffic conditions with connected autonomous vehicles. *IET Intelligent Transport Systems 12*(5), 327–334 (2018)

[3] Spall, J. C.: Implementation of the simultaneous perturbation algorithm for stochastic optimization. *IEEE Transactions on Aerospace and Electronic Systems 34*(3), 817–823 (1998)

[4] Spall, J. C.: *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*. Wiley, Hoboken (2005)

[5] Frazier, P. I.: A tutorial on Bayesian optimization. *arXiv preprint* arXiv:1807.02811 (2018)

[6] Rasmussen, C. E., Williams, C. K. I.: *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA (2006)

[7] Treiber, M., Kesting, A.: *Traffic Flow Dynamics: Data, Models and Simulation*. Springer, Berlin (2013)

[8] Wiedemann, R.: Modeling of RTI-elements on multi-lane roads. In: *Proceedings of the DRIVE Conference on Advanced Telematics in Road Transport*, Brussels (1991)

[9] Ascione, F., et al.: Optimizing space cooling of a nearly zero energy building via model predictive control: energy cost vs comfort. *Energy and Buildings* 278, 112664 (2023).

[10] Abida, A., Richter, P.: HVAC control in buildings using neural network. *Journal of Building Engineering* 65, 105558 (2023).

[11] Li, G., et al.: Interpretation of convolutional neural network-based building HVAC fault diagnosis model using improved layer-wise relevance propagation. *Energy and Buildings* 286, 112949 (2023).

[12] Zhao, H.: Intelligent management of industrial building energy saving based on artificial intelligence. *Sustainable Energy Technologies and Assessments* 56, 103087 (2023).

[13] Khan, S.U., et al.: Towards intelligent building energy management: AI-based framework for power consumption and generation forecasting. *Energy and Buildings* 279, 112705 (2023).

[14] Ma, D., et al.: A dynamic intelligent building retrofit decision-making model in response to climate change. *Energy and Buildings* 284, 112832 (2023).

[15] Ding, Y., et al.: Coordinated optimization of robustness and flexibility of building heating systems for demand response control considering prediction uncertainty. *Applied Thermal Engineering* 223, 120024 (2023).

[16] Moudgil, V., et al.: Integration of IoT in building energy infrastructure: a critical review on challenges and solutions. *Renewable and Sustainable Energy Reviews* 174, 113121 (2023).

[17] Jia, C., et al.: Intelligent decision optimization for energy control of direct current power distribution system with multi-port access for intelligent buildings. *Alexandria Engineering Journal* 63, 455–464 (2023).

[18] Baldi, S., et al.: Automating occupant-building interaction via smart zoning of thermostatic loads: a switched self-tuning approach. *Applied Energy* 231, 1246–1258 (2018).

[19] Korkas, C.D., et al.: Intelligent energy and thermal comfort management in grid-connected microgrids with heterogeneous occupancy schedule. *Applied Energy* 149, 194–203 (2015).

Yan Li, Anhui Vocational and Technical College, Hefei, Anhui, 230011, China

Xiaong Wie, Anhui Vocational and Technical College, Hefei, Anhui, 230011, China

Manuscript Published; 30 December 2025.