International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Calibrated Arrival-Time Prediction for Indian Road Freight

Mr. Saurabh Kumar

CTO, Co-Founder, Lead Researcher, Q-AI-ML

Abstract

Routing engines predict how long it takes to drive a road. For Indian long-haul freight they are wrong by a factor of three to four, because a vehicle is not driving most of the time. We present an arrival-time engine assembled from three measured layers: a per-road-class speed correction derived from map-matched fleet traces, a stop model derived from the observed distribution of halt durations, and a weather layer combining published speed-reduction coefficients with two locally trained models. The output is a calibrated arrival distribution rather than a point estimate.

Evaluated on 790 held-out journeys under a strict temporal split, the engine achieves a median absolute error of 2.85 h and a median absolute percentage error of 26.9%, against 7.79 h and 64.9% for the incumbent telematics estimate in production on the same fleet, and 3.50 h and 30.3% for a well-fitted constant-speed baseline. Nominal 50%, 80% and 90% quantiles achieve 44%, 77% and 88% empirical coverage.

We report five hypotheses that were measured and rejected, two of which were our own models that passed conventional validation while being materially broken. The most consequential result is methodological: the definition of *arrival* in commercial telematics data is unreliable enough to invalidate model evaluation, and a behaviour-derived definition is required before any accuracy figure can be trusted.

Keywords: arrival time estimation, travel time prediction, road freight, fleet telematics, quantile regression, probabilistic forecasting, weather impact on traffic, model calibration, India

1 Introduction

An estimated time of arrival is the primary output a freight control tower consumes. It determines dock scheduling, driver allocation, customer commitments and exception handling. For Indian road freight the estimates available in practice come from one of two sources: a routing engine, which answers a question nobody asked, or a telematics provider's own estimate, which in the system studied here is a constant speed divided into remaining distance.

The gap between the first of these and reality is large. A standard routing engine estimates 19 h 23 m for a 1,376 km trunk route that real vehicles complete in approximately 2.8 days. The engine is not wrong about the road. It is wrong about everything else.

Three properties distinguish the approach described here.

- **Every coefficient is measured.** Speed corrections, stop rates, drainage priors and calibration constants derive from observed vehicle behaviour on the network in question, not from literature values transplanted from other countries.

- **The output is a distribution.** On the evidence below, the point estimate is only modestly better than a constant, whereas the interval is a capability a constant does not possess at all.
- **Negative results are reported.** Section 7 documents five failed hypotheses. Two passed conventional validation while broken, and were caught only by checking outputs against physical plausibility.

2 Related work

2.1 Production ETA systems

The architecture of large-scale ETA systems has converged. Uber's DeepETA applies a learned residual correction atop a routing-engine baseline, reporting 7.83% relative MAE improvement and 7.99% at p95 at 3.25 ms median latency. Baidu's WDR and ConSTGAT, DiDi's HetETA, and the DeepMind and Google Maps graph neural network work , which reports over 40% reduction in negative ETA outcomes in one metropolitan area, all follow the same pattern: a physical routing baseline plus a learned correction, never a model trained from scratch on raw origin–destination pairs.

We adopt that pattern. The distinction in our setting is that the correction must account for a far larger non-driving component, and that the underlying road-speed assumptions require correction before any residual is meaningful.

Every one of these systems is evaluated on proprietary internal data. There is no reproducible comparison between them and there never has been. The aggregator that previously hosted travel-time-estimation leaderboards has been discontinued. Evaluation must therefore be constructed, and its construction defended on methodology.

2.2 Weather and travel time

The canonical coefficients are the 2010 Highway Capacity Manual Exhibit 10-15 and the SHRP 2-L08 adjustment factors . At 65 mi/h free-flow, heavy rain carries a speed adjustment factor of 0.93 and a capacity factor of 0.86; low visibility carries 0.93 and 0.88.

These are US freeway values. Indian evidence is thinner and points the same way but larger. Banik and Vanajakshi report a 457% spread in maximum Travel Time Index between dry and rainy conditions, with rainfall and day-of-week jointly significant. A study of 45 waterlogged urban segments in Shenzhen , the closest tropical-Asian analogue, reports speed reductions of 21.5% to 24.3% under waterlogging against 1.2% to 18.4% under rain alone.

Waterlogging is a segment state, not a weather variable. It is driven by accumulated rainfall interacting with the drainage of a specific stretch of road, and therefore requires a per-location prior that no public dataset provides for India.

2.3 Weather forecasting benchmarks

WeatherBench 2 is the reference benchmark for global forecasting. Two properties matter here: it scores reanalysis against reanalysis rather than against station observations, and it contains no India or South Asia region. IndiaWeatherBench is the only India-native benchmark we could identify; its data carries a non-commercial licence and its code repository carries none.

For nowcasting, no public benchmark covers India or tropical convection. Indian monsoon convection is warm-rain dominated, so reflectivity-to-rain-rate relations calibrated on mid-latitude radar are inapplicable, and transfer should be assumed poor until measured.

3 Data

3.1 Fleet telematics

Behavioural measurements derive from a GPS position archive contributed by a commercial road freight operator in India: approximately 11.9 million records at nominal 30-second cadence, spanning roughly 17,000 vehicle-hours across national highway corridors and urban delivery areas. Positions originate from two device classes, on-vehicle GPS and SIM-derived location, which differ materially in accuracy and are never pooled without a source indicator.

Three data quality issues required handling and are likely generic to this class of data: stale positions from offline devices, ambiguous trip identity, and imprecise destination coordinates. The third proved decisive and is treated in Section 7.5.

3.2 Road network

Routing uses a self-hosted engine built on the OpenStreetMap India extract. The dominant network characteristic is tag sparsity: approximately 82% of the major road network carries no maxspeed tag, so speeds are assigned from defaults rather than data. Probe-derived national defaults were configured before any benchmarking; omitting that step produces a strawman comparison.

3.3 Weather

Six sources were evaluated and four adopted: a global open forecast API as primary, the national meteorological service for secondary forecasts and official warnings, its automatic weather stations as verification ground truth, and airport observations for measured visibility. Historical rainfall for drainage priors comes from reanalysis.

A *forecast-as-issued* archive was established at the outset, storing every forecast with issue and validity timestamps. No archive of past Indian forecasts is published, and without one forecast skill cannot be measured honestly: comparing a recent past value to a simultaneous observation is not verification.

4 Problem definition

The engine predicts **arrival at the destination gate**, given an origin, a destination, a departure time and a vehicle profile. Three distinct quantities are conflated in practice:

Quantity	Includes
Driving time	time in motion only
Gate arrival	plus en-route halts
Consignment closed	plus unloading and paperwork

The engine targets the middle quantity. Unloading is the facility's concern, in the same way a last-mile delivery application predicts arrival at a building rather than at a door. The incumbent targets the third, which explains a systematic positive bias when both are scored on gate arrival.

In deployment the endpoints are supplied by the operator, so no inference about intent is required. The endpoint-recovery problems of Section 7.5 are artefacts of retrospective evaluation and do not exist at prediction time.

5 Methods

5.1 Behavioural journey extraction

Trip records proved unusable as journey boundaries. Journeys are derived from movement: a journey *starts* when the vehicle begins moving after a sustained halt, *ends* when it stops and remains stationary for at least 90 minutes, and is *retained* only if the driven path is within 1.35 times the routed path

between endpoints. The last condition scopes evaluation to point-to-point journeys; approximately 13% of candidates were dropped on it.

Both endpoints are locations the vehicle demonstrably occupied, and the terminus is one it demonstrably remained at. No geocoding, status flag or closure timestamp enters the definition.

5.2 Road speed correction

Traces were map-matched to the routing graph using a hidden-Markov matcher, yielding per-edge observed speeds. Stationary intervals within each edge were excluded, so the correction represents road speed rather than stop incidence.

5.3 Stop model

The distribution of stationary intervals decomposes into two regimes with a boundary near ten minutes, modelled differently. Intervals under ten minutes are modelled as a *rate* per kilometre: signals, toll plazas and brief queues, too frequent and too small to predict individually. Intervals over ten minutes are modelled as *discrete events* with identifiable causes. A minority of geographic locations account for a disproportionate share of long-stop time and are predictable from route geometry alone.

5.4 Weather impact

Precipitation uses the FHWA speed adjustment factors directly. The central heavy-rain estimate is 0.93; the same source elsewhere reports a 3% to 17% range for that condition, and the range is treated as the uncertainty band feeding the upper quantiles rather than averaged away.

Visibility is not taken from the forecast, for reasons in Section 6.4. A classifier predicts the probability of visibility below 1,609 m from temperature, dewpoint depression, wind and seasonality, and the speed factor is applied in proportion to that probability rather than as a step, avoiding discontinuities driven by a forecast that is itself uncertain.

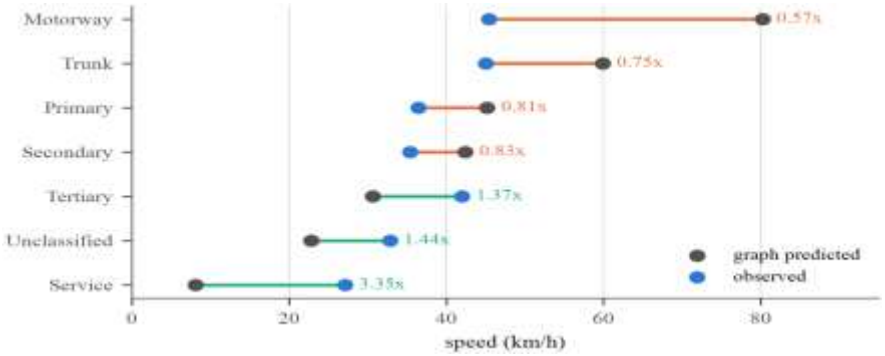
Waterlogging uses per-location drainage priors, applied only where precipitation is forecast and only to the upper quantiles, since it is a tail event.

5.4.0.1 Time alignment.

Weather is sampled at the time the vehicle is predicted to reach each segment, not at departure. This is circular, since arrival depends on speed which depends on weather which depends on arrival, and is solved by iteration. Multipliers are bounded in approximately [0.78,1.0], so each pass perturbs estimates by at most 22%; the iteration is a contraction and converges in two to three passes. Caching forecasts on a coarse spatial-temporal key reduced end-to-end latency on a 1,376 km route from 115 s to 23 s.

6 Results

6.1 Road speed correction



Graph-predicted against observed truck speed by road class, from 6,035 map-matched edges over 16,140 km. The graph errs in opposite directions depending on class. Ratios are observed over predicted.

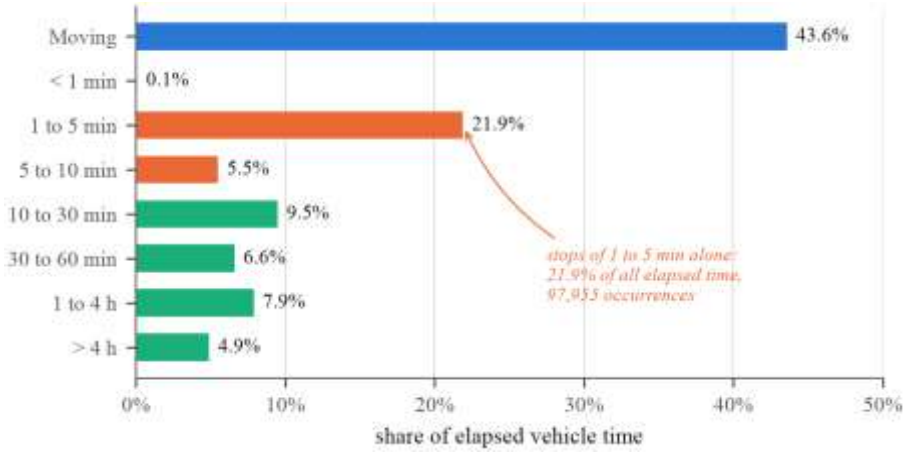
From 6,035 map-matched edges spanning 16,140 km and 2,339 distinct network ways (Figure 1), major roads are too fast and minor roads too slow.

The mechanism is understood rather than inferred. Major roads take their speed from maxspeed tags, which are *car* limits; Indian goods vehicles are capped at 80 km/h by regulation and run substantially below it. Minor roads fall back to probe-derived defaults, which already incorporate stopping and crawling, and therefore under-predict a vehicle genuinely in motion.

A single global correction factor would be actively harmful: it would fix highways and break everything else.

The per-edge correction independently reproduces the fleet-wide moving speed of 43.7 km/h derived by a wholly separate method. The two were never fitted to each other.

6.2 Stop structure



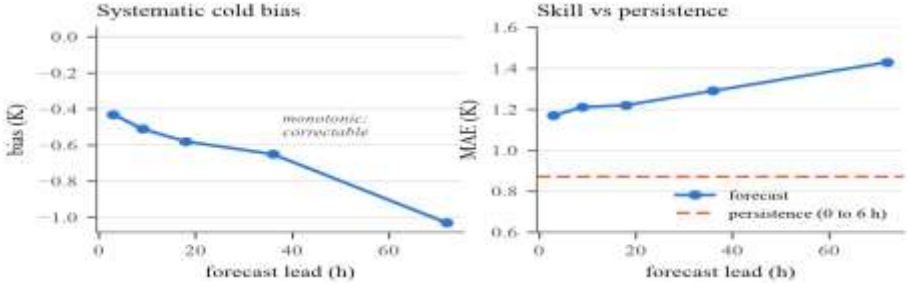
Where elapsed vehicle time goes, by stationary-interval duration. Vehicles are stationary approximately 56% of the time, and stops of one to five minutes alone consume 21.9%.

Decomposing approximately 17,000 vehicle-hours (Figure 2) yields three speeds, all legitimate and all different:

Moving	38.8 km/h	while in motion
In-transit	23.8 km/h	excluding stops >10 min
Door-to-door	16.9 km/h	everything included

Conflating these is the primary source of error in naive ETA construction. A routing engine predicts something close to the first; a consignment-level system reports something close to the third. A small number of locations account for 42% of all stopped hours and are predictable from route geometry before departure.

6.3 Forecast verification and bias correction



Left: systematic cold bias against forecast lead, from 27,966 verified pairs at 43 sites. Right: forecast MAE against the persistence baseline, which the forecast does not beat at short lead.

Forecasts were archived as issued and scored against station observations at their validity time (Figure 3). Two results deserve emphasis.

Neither forecast source beats persistence at 0 to 6 hours. At short lead the observation already in hand is a better temperature forecast than either NWP product.

The primary source carries a monotonic cold bias growing from -0.43 K to -1.03 K with lead. A nested correction was fitted, most specific first: site plus lead band plus hour block, falling back to coarser groupings where support is thin. The nesting is necessary, because a *constant* correction achieves nothing: the error sign flips by location, one major city at $+5.9$ K and another at -1.8 K. Validated on a temporal holdout, the correction reaches -24.0 % MAE overall and 28% to 32% at short lead, comparable to published station post-processing results obtained with a decade of data.

6.4 Visibility

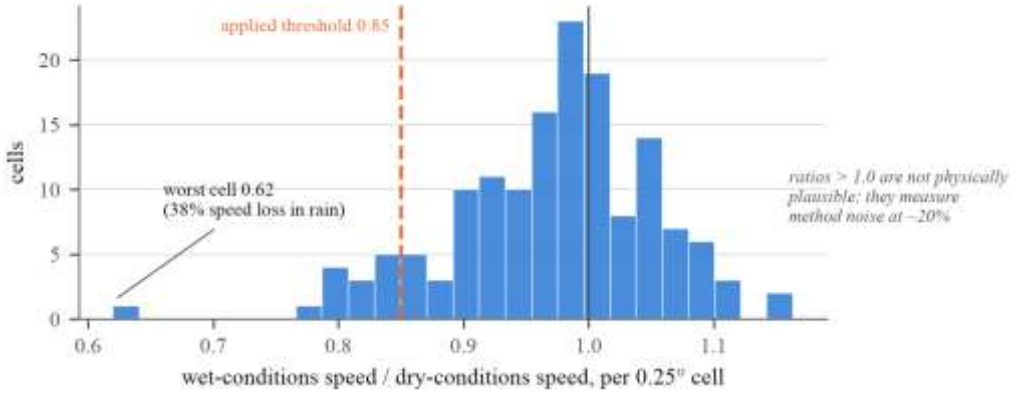
Modelled visibility was compared against 127 airport observations at matched times, giving a bias of $+5{,}613$ m and MAE of 6,293 m. Critically, two stations were genuinely below the 1,609 m threshold at which the speed factor applies; the model placed *both* above it.

Driving a fog multiplier from modelled visibility would mean it essentially never fires, in precisely the conditions the product exists to handle. The output stays confident and is wrong exactly when it matters.

A classifier was trained instead on 16 stations, 2021–2024, tested on 2025 ($n = 118{,}802$): ROC AUC 0.962, Brier 0.0300 against 0.0599 for climatology. Persistence scores 0.0205 and therefore beats the model; persistence also requires knowing visibility one hour ago, which is unavailable when forecasting hours ahead. Against what is available at forecast time, halving climatology's score is the operative result.

On a held-out December event the model predicts 99% at three northern stations with dewpoint depressions of 0.0 to 0.4 K, and 3% to 4% at a southern station with a comparable depression of 0.2 K. It has learned that saturated air at 14 °C in the Deccan does not behave as saturated air at 10 °C in the northern plain, which is correct.

6.5 Drainage priors



Distribution of per-cell wet/dry speed ratio across 151 cells, from 248,883 moving legs joined to hourly reanalysis rainfall. Ratios above 1.0 are not physically plausible and measure method noise.

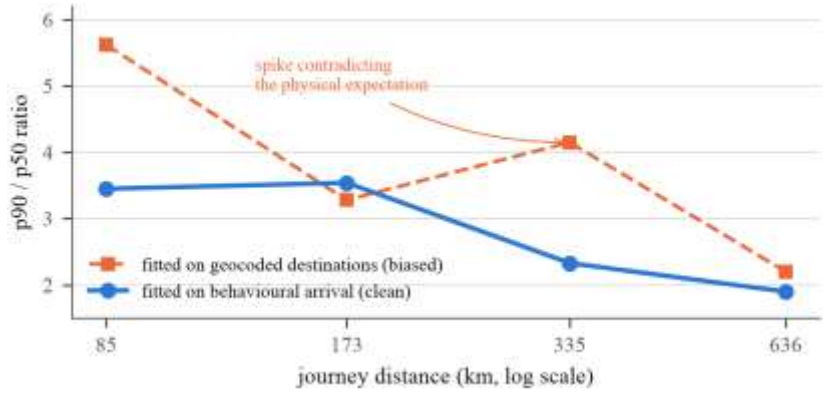
Waterlogging requires knowing which stretches flood, and no public dataset provides this for India. It proves to be latent in fleet GPS (Figure 4). Ping-level moving legs were bucketed into cells matching the reanalysis grid, joined to hourly historical precipitation, and wet-conditions speed compared against dry within each cell.

The median effect of rain is approximately 1%; the worst cell loses 38% of its speed. Five of the ten worst-draining cells fall within a single metropolitan corridor known for monsoon waterlogging, identified without any input describing drainage, topography or urban geography.

Some cells return ratios above 1.0, up to 1.27, which is not physically plausible as a rain effect and therefore measures method noise at roughly ±20%. Only cells below 0.85 are treated as findings.

Measured end to end, weather widens the distribution approximately 2.4 times more than it shifts the median. On a 400 km intercity leg, improving rainfall forecast skill from typical published levels to substantially better buys roughly 0.6 minutes of mean accuracy. The value of the weather layer is in the tail, and commercial claims should be framed accordingly.

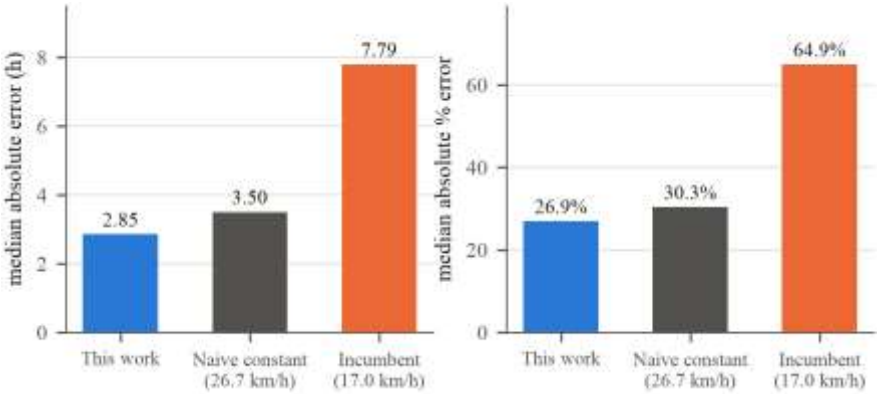
6.6 Interval width



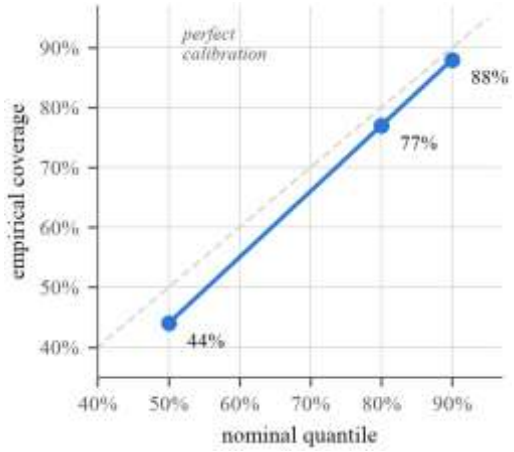
Ratio of the 90th to 50th percentile against journey distance. Bands fitted on geocoded destinations show a spike contradicting the expectation that longer journeys average out; bands fitted on behavioural arrival decrease monotonically apart from one bump inside noise.

Figure 5 illustrates the effect of the arrival-definition problem on a derived quantity. Under the biased sample, interval width against distance is non-monotonic in a way that contradicts the physical expectation that longer journeys average out many independent perturbations. Under the behavioural definition the relationship recovers.

6.7 End-to-end evaluation



Held-out error on 790 journeys under a temporal split. The demanding comparison is the naive constant, which is fitted on the training split and lands within 1% of the observed test median.



Empirical coverage against nominal quantile, held out. All three quantiles fall within tolerance of the identity line.

The evaluation uses 2,631 behaviourally extracted point-to-point journeys, split temporally with 1,841 for calibration and 790 held out. A temporal split is required rather than a random one: journeys in the same week share weather, vehicles and consignments.

Two baselines are used (Figure 6). The incumbent telematics estimate in production on the same fleet, empirically a constant near 17 km/h. And a naive constant fitted on the training split at 26.7 km/h, against an observed test median of 27.0 km/h, included because outperforming a poorly chosen constant while tying a well-chosen one is not a meaningful result.

Model	Bias	MAE	MAPE
This work	−0.58 h	**2.85 h**	**26.9%**
Naive (26.7 km/h)	+0.10 h	3.50 h	30.3%
Incumbent (17.0 km/h)	+5.19 h	7.79 h	64.9%

The engine reduces error against the incumbent by 63% and against a well-fitted constant by 19%, with all three quantiles calibrated (Figure 7). The incumbent comparison should be read with care: it targets consignment closure while being scored on gate arrival, which accounts for its positive bias. The naive comparison is the demanding one, and the 19% margin is the evidence that route-level structure contributes.

7 Negative results

These are reported because they were expensive to obtain, because two passed conventional validation while broken, and because the mechanisms are likely to recur.

7.1 No fatigue signal in stop behaviour

Stop duration was measured against accumulated preceding driving time ($n = 1{,}835$). Median duration was 13 to 18 minutes across every band from under one hour to over eight hours of preceding driving. No relationship. The stop model is therefore geographic rather than driver-state, which is both simpler and better supported than the design originally proposed.

7.2 Time of day does not affect road speed

Moving speed by hour is flat: peak-to-trough is 1.08× for urban cells and 1.12× for highway cells, and this holds after separating them rather than dissolving under it. A related hypothesis, that journey duration per kilometre varies 1.30× by *departure* hour, failed a temporal holdout; the likely mechanism is confounding with journey type, which the route already encodes.

7.3 Target leakage in the visibility classifier

The first visibility model included visibility one hour prior as a feature. It scored ROC AUC 0.991 and beat persistence by 1.6× on Brier score: the strongest apparent result in the work.

It was leakage. Holding conditions fixed at saturated, calm, December-dawn values, moving the lagged feature from 1,600 m to 3,000 m dropped the predicted probability from 96.3% to 1.3%, while sweeping dewpoint depression across its entire physical range moved it only from 2.1% to 0.1%.

The feature does not exist when forecasting ahead. The neutral value supplied at inference placed every prediction on the clear side of the decision boundary, so the model returned approximately 2% on textbook radiation-fog mornings. It would have failed silently for an entire fog season while its headline metric remained excellent.

Detected by evaluating a known meteorological case, not by any validation metric.

7.4 Learned quantile regression on route features

A gradient-boosted quantile model on route features initially improved pinball loss by 8.0%, 12.8% and 20.4% at p50, p80 and p90, with the gain growing with the quantile exactly as hypothesised.

It was contaminated: features used routed origin–destination distance while the target was elapsed time for the distance actually covered, which on multi-drop journeys differ by up to 7×. Detected when the model predicted 13 hours for a 49 km route, longer than its own prediction for a 340 km route. Refitted on clean data five times larger, it scored 3.47 h MAE against 2.85 h for the existing engine and 3.50 h for a plain constant.

The interpretation is instructive. The engine is *already* route-dependent, through per-edge routing and per-class speed correction. A model consuming eleven aggregate features is a cruder encoding of information the router already holds at higher resolution.

7.5 Arrival cannot be reliably determined from telematics records

This invalidated three evaluation designs before being resolved.

Signal	Failure mode
Proximity to recorded destination	24% of vehicles within 1 km; median closest approach 3.58 km
Trip closure timestamp	batch entry, after the vehicle departs
Final position before closure	median 35.5 km from destination
Trip record boundaries	median record spans 42.8 h, endpoints 56 km apart

Each proxy fails differently. Selecting evaluation trips on a status flag already documented as unreliable produced a test in which four of eight vehicles never approached their recorded destination and two ended further away than they began.

A noisy target can conceal a genuine effect. The 19% margin over a constant was undetectable under any of the earlier arrival definitions.

8 Limitations

All behavioural coefficients derive from one fleet on one set of corridors, so generalisation is untested. Drainage priors rest on 66 monsoon days with no dry-season comparison and cannot currently separate flooding from seasonal traffic variation; this is the largest caveat on the tail argument. The visibility classifier trains on a source whose non-US records carry redistribution restrictions that may or may not have been superseded by a later policy; an independent replacement from the national service's own public feed is accumulating. Evaluation excludes multi-drop and circuitous journeys. Upper-quantile bands rest on a few hundred journeys per distance band, with p90 estimates resting on tens of tail

observations. Finally, no reproducible benchmark exists for travel-time estimation, and none exists for weather-conditioned travel-time estimation; we regard this as a genuine gap in the field.

9. Conclusions

Four conclusions we consider transferable.

The non-driving component dominates. Vehicles are stationary 56% of elapsed time and short stops alone consume 21.9%. Any ETA system for this domain that models road speed without modelling stop structure will be wrong by a large multiple regardless of routing quality.

Routing graph error is signed by road class. Major roads are too fast because maxspeed encodes car limits; minor roads too slow because probe defaults embed stopping. A scalar correction would be actively harmful.

Weather belongs in the interval, not the median. The mean-case effect is commercially uninteresting. The tail effect is large and location-specific, and per-location drainage priors recoverable from fleet GPS are the mechanism by which it becomes predictable.

Validate the target before validating the model. Two models passed conventional metrics while materially broken, and both were caught by checking outputs against physical plausibility rather than by any score.

References

1. Banik, S., Vanajakshi, L. (2024). Impact of Rainfall on Traffic Mobility and Reliability Under Indian Traffic Conditions. *Transportation in Developing Economies*.
2. Derrow-Pinion, A. et al. (2021). ETA Prediction with Graph Neural Networks in Google Maps. *CIKM*. arXiv:2108.11482.
3. Federal Highway Administration. Analytical Procedures for Determining the Impacts of Reliability Mitigation Strategies, Appendix A.
4. Hu, X. et al. (2022). DeepETA: A Spatial-Temporal Sequential Neural Network Model for Estimating Time of Arrival. arXiv:2206.02127.
5. IndiaWeatherBench: A Dataset and Benchmark for Data-Driven Regional Weather Forecasting over India (2025). arXiv:2509.00653.
6. Rasp, S., Lerch, S. (2018). Neural Networks for Postprocessing Ensemble Weather Forecasts. *Monthly Weather Review* 146(11).
7. Rasp, S. et al. (2023). WeatherBench 2. arXiv:2308.15560.
8. Wang, Z., Fu, K., Ye, J. (2018). Learning to Estimate the Travel Time. *KDD*.
9. Impacts of Rain and Waterlogging on Traffic Speed and Volume on Urban Roads (2018). *IEEE ITSC*.