How predictable are macroscopic traffic states a perspective of uncertainty quantification

Li, Guopeng; Knoop, Victor L.; van Lint, Hans

DOI
10.1080/21680566.2024.2314766

Publication date
2024

Document Version
Final published version

Published in
Transportmetrica B: Transport Dynamics

Transportmetrica B: Transport Dynamics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ttrb20

How predictable are macroscopic traffic states: a perspective of uncertainty quantification

Guopeng Li, Victor L. Knoop & Hans van Lint

To cite this article: Guopeng Li, Victor L. Knoop & Hans van Lint (2024) How predictable are macroscopic traffic states: a perspective of uncertainty quantification, Transportmetrica B: Transport Dynamics, 12:1, 2314766, DOI: 10.1080/21680566.2024.2314766

To link to this article: https://doi.org/10.1080/21680566.2024.2314766

TRANSPORTMETRICA B: TRANSPORT DYNAMICS
2024, VOL. 12, NO. 1, 2314766
https://doi.org/10.1080/21680566.2024.2314766

How predictable are macroscopic traffic states: a perspective of uncertainty quantification

Guopeng Li, Victor L. Knoop and Hans van Lint

Department of Civil Engineering and Geoscience, Delft University of Technology, Delft, Netherlands

ABSTRACT

Traffic condition forecasting is fundamental for Intelligent Transportation Systems. Besides accuracy, many services require an estimate of uncertainty for each prediction. Uncertainty quantification must consider the inherent randomness in traffic dynamics, the so-called aleatoric uncertainty, and the additional distrust caused by data shortage, the so-called epistemic uncertainty. They together depict how predictable macroscopic traffic is. This study uses deep ensembles of graph neural networks to estimate both types of uncertainty in network-level speed forecasting. Experimental results given by the used model reveal that, although rare congestion patterns arise randomly, the short-term predictability of traffic states is mainly restricted by the irreducible stochasticity in traffic dynamics. The predicted future state bifurcates into congested or free-flowing cases. This study suggests that the potential for improving prediction models through expanding speed and flow data is limited while diversifying data types is crucial.

ARTICLE HISTORY
Received 2 December 2022
Accepted 31 January 2024

KEYWORDS
Traffic forecasting; uncertainty quantification; traffic dynamics; predictability

1. Introduction

The quality of many travel services and traffic management functions, such as arrival time estimation (ETA) (Van Lint, Hoogendoorn, and van Zuylen 2005), real-time route planning (Gehrke and Wojtusiak 2008; Liebig et al. 2017), and congestion control (Chen, Peng, and Wang 2000), relies on accurate short-term traffic state forecasting. The application value is a significant impetus for creating novel models that continuously improve prediction accuracy. These methods range from classical traffic-simulation-based approaches (Ben-Akiva 1998; Qiao, Yang, and Lam 2001; Wang, Papageorgiou, and Messmer 2006; Wang, Work, and Sowers 2016), regressive and time-series models (Castro-Neto et al. 2009; Davis and Nihan 1991), to recently popular deep learning models (Li et al. 2017; Ma et al. 2017; Van Lint, Hoogendoorn, and van Zuylen 2002) and generative models (Xu et al. 2022; Zhang et al. 2021). Now, there is significant interest in the research and application of machine learning and deep neural networks (DNNs) for predicting the traffic state of large road networks.

However, like other non-trivial prediction tasks, traffic forecasting is also associated with *uncertainty*. The evolution of traffic state in any road network is not deterministic but inherently stochastic for any observers. The unavailability of some critical information (to agencies performing the forecasting task), such as operational and tactical driving behaviours, demand (Gu et al. 2022), and route choices, induces randomness in traffic dynamics. This limited observability implies that the model's outcome is a (set of) random variables, which can be represented by either scenario-based quantities

CONTACT Guopeng Li G.Li-5@tudelft.nl Department of Civil Engineering and Geoscience, Delft University of Technology, Mekelweg 5, 2628CD Delft, Netherlands

(e.g. in policy analysis, Kim et al. 2013) or a probability distribution function (PDF). To construct predictive models that incorporate uncertainty quantification, it is necessary to clarify what the uncertainty represents and how to model it.

From the perspective of modelling and application, predictive uncertainty is typically categorized into two parts, the so-called *aleatoric* uncertainty and *epistemic* uncertainty (Der Kiureghian and Ditlevsen 2009; Kendall and Gal 2017). Aleatoric uncertainty (from *alea*, Latin for 'dice') represents the inherent randomness of a stochastic process and the measurement error in data collection, which cannot be reduced by expanding the dataset of the same type of data. This uncertainty draws a lower bound of predictive accuracy for any models using the same (types of) inputs (Li, Knoop, and van Lint 2022). For example, for a univariate Gaussian process, the variance gives the lower bound of the expectation of mean-square error for any predictors (otherwise the model is over-fitted). On the other hand, epistemic uncertainty mainly stems from the lack of data to learn the output distribution. Take the same example, the used predictor may be not good enough to model the Gaussian process or there are only a limited number of observations around some rare input points. This will induce additional uncertainty. The epistemic uncertainty is closely related to measuring the rareness of samples. For example, calibrating a Gaussian distribution from 500 samples gives consistent results. However, if only 3 samples are provided, the estimated parameters are highly uncertain. The idea is that–in principle–epistemic uncertainty can be reduced by expanding the training set, and/or adding (supposedly known) sophistication to the applied model.

Uncertainty Quantification (UQ) has been widely studied for many risk-sensitive systems with partially-observable information, such as nuclear safety (Helton 1993), hydrology (Beck 1987), meteorology (Deser et al. 2012), etc. Suppose a system is explicitly simulated through physical models. In that case, all uncertainty of the parameters, the model structure, and the data used to calibrate and identify these respectively, can influence the confidence of the output. Examples in the traffic domain include the heterogeneity of car-following parameters in microscopic traffic simulations (Sharma, Zheng, and Bhaskar 2019) and the distribution of fundamental diagram parameters (e.g. capacity, critical density) used in LWR-type models (Li et al. 2012). The output distribution for predictions using simulation-based models is typically obtained by running the *forward* propagation of such 'errors' (sources of uncertainty) (van Lint et al. 2012).

At the other end of the spectrum, many *inverse* data-driven UQ methods ignore how uncertainty propagates within the model but directly build the joint probability of input-output from collected datasets. With the development of deep learning techniques, this end-to-end approach is becoming popular for UQ. This comes at the cost of low interpretability and poor generalizability. The two approaches above are the same for distinguishing aleatoric and epistemic uncertainty. A distribution can measure the confidence of point predictions (aleatoric uncertainty). Further, Epistemic uncertainty measures how confident we are about the probabilistic prediction. Therefore, we need a 'distribution of distributions' to quantify epistemic uncertainty, either through getting an ensemble of output distributions or a distribution of the parameters of the output distribution (Amini et al. 2020).

Specific to network-level short-term traffic forecasting, we observe that most proposed models in the literature use those easily observable macroscopic variables (speed, flow etc.) as inputs (Li et al. 2017; Ma et al. 2017 etc.) and they specifically focus on accuracy improvement. Disentangling input-dependent aleatoric and epistemic uncertainty can address a critically important but in many cases ignored question:

> How predictable are macroscopic traffic states if only observations of traffic speed and flow are provided, and why?

The two types of uncertainty must be jointly considered to evaluate the 'predictability'. If aleatoric uncertainty dominates, then using better models or collecting more data of the same type cannot significantly improve the prediction accuracy. Conversely, epistemic uncertainty can be regarded as a measure of how much predictive accuracy can still be achieved by better modelling or collecting more data. If the epistemic uncertainty is higher, then investing in modelling techniques and

expanding datasets are worthwhile. In summary, quantifying and comparing these two sources of uncertainty is informative for whether analysts and scientists should focus on expanding the source of heterogeneous data, or on more sophisticated modelling techniques.

In this study, we employ probabilistic deep-learning models and Deep Ensembles (DE) (Lakshminarayanan, Pritzel, and Blundell 2016) to quantify both aleatoric and epistemic uncertainty associated with network-level, multi-step highway speed forecasting. The major contributions of this paper are summarized as follows:

- Use a deep ensemble of graph neural networks to quantify aleatoric and epistemic uncertainty in multistep highway speed forecasting.
- Illustrate that the predictability of traffic speed is mainly restricted by the inherent randomness of traffic dynamics, especially the bi-modality of long-term traffic state.
- Conclude that collecting more macroscopic traffic data or developing more complex correlation-based models cannot significantly improve the prediction accuracy, based on the employed uncertainty quantification method.

This paper is organized as follows. We first briefly overview the related works in the literature in Section 2. Section 3 presents the uncertainty quantification method and details of the proposed model. In Section 4, the model is tested on a real-world highway network and the analysis of the experimental results is presented in Section 5. Finally, Section 6 concludes and proposes several relevant research directions.

2. Overview

Quantifying uncertainty is an important topic in many domains. Various models and methods have been developed to address the uncertainty concerns in prediction and estimation tasks. Recently, deep neural networks (DNNs) have gained considerable prominence in the domains of traffic prediction and uncertainty quantification. They have demonstrated superior predictive accuracy and greater robustness in estimating uncertainty as compared to simulation-based models and other machine-learning techniques. In this subsection, we initially provide an overview of the DNN-based uncertainty quantification (UQ) methods. Next, a succinct overview of UQ techniques implemented for traffic state prediction and estimation is provided.

In principle, the key to quantifying aleatoric and epistemic uncertainty through DNNs involves directly getting an ensemble of output distributions or giving a set of parameters to describe the distribution ensemble. Bayesian neural networks (BNN) are amongst the most popular methods to this end. However, directly training a BNN is difficult. For a given training set $\mathcal{D}$, the exact form of the posterior distribution of model parameters $p(\theta \mid \mathcal{D})$ is intractable. So this term has to be approximated by a variational distribution. Monte-Carlo dropout (MC-dropout) (Kendall and Gal 2017) is one of the most widely-used methods. Dropout layers (Srivastava et al. 2014) are implemented in DNNs and the random dropout is also enabled during inference. One can acquire an ensemble of outcomes by executing the same model repeatedly. For example, Zechin et al. (Zechin and Cybis 2023) used MC-dropout LSTMs to forecast the probabilistic traffic breakdown. However, Foong et al. shows that MC-dropout is a relatively worse approximation for deeper BNNs (compared to shallow models), and not robust to the data collected from different situations (the so-called *dataset shift*). Other approximation techniques include Markov chain Monte-Carlo (MCMC) (Kupinski et al. 2003), variational inference (VI) (Swiatkowski et al. 2020), Taylor-expansion based Laplacian approximation (Ritter, Botev, and Barber 2018), to name a few.

Deep ensembles (Lakshminarayanan, Pritzel, and Blundell 2016) constitute another powerful approach for both improving predictive accuracy and estimating uncertainty. In particular, the utilization of random initialization has demonstrated unexpectedly strong performance and robustness (Ovadia et al. 2019), achieving outcomes comparable to state-of-the-art methods in many tasks. The

high performance of DE can be interpreted from the perspective of 'loss landscapes'. Fort, Hu, and Lakshminarayanan (Fort, Hu, Lakshminarayanan 2019) show that random-initialization DE can explore multiple local optima of loss functions whereas a single optimization trajectory often falls into one. During inference, we do not sample weights from one model but get multiple outputs from an ensemble of deterministic models (with different but fixed weights). The primary limitations of deep ensembles include the high computational resource demands and relatively slow inference speed. Some techniques, like DE distillation (Malinin, Mlodozeniec, and Gales 2019), are proposed to mitigate these problems. But to achieve comparable performance, pre-training a DE is still necessary.

Recently, some studies have attempted to achieve uncertainty quantification through a single deterministic model utilizing a sole forward pass, such as inducing radical-basis function (van Amersfoort et al. 2020) and deep evidential regression (Amini et al. 2020). However, the effectiveness and performance of these methods need to be tested on more tasks (Malinin et al. 2020). For a recent and comprehensive review of uncertainty quantification in deep learning models, we refer the readers to Abdar et al. (2021).

Traffic state estimation and traffic prediction are closely related and they can benefit from the same uncertainty quantification (UQ) methods. Traffic state estimation involves interpolating state variables based on incomplete data, while traffic prediction extrapolates past traffic conditions to anticipate future outcomes. Seo et al. (2017) systematically reviews the traffic state estimation on highways. For traffic prediction, we refer the readers to Yuan and Li (2021) for a recent survey.

One of the most common UQ approaches is using stochastic differential equations (SDE) to estimate (Zheng et al. 2018) or predict traffic states (Chu et al. 2011; Tahmasbi and Mehdi Hashemi 2013). Building such a simulator on the desired level requires a thorough understanding of traffic dynamics. The simulation is explainable and controllable. However, some critical parameters and random terms are either unobservable or difficult to calibrate in real-world scenarios. Therefore, additional data or assumptions have to be induced. For example, the rate of vehicle arrival follows a Poisson distribution under free-flowing highway conditions (Daganzo 1997), which needs to be estimated from microscopic data. Modelling how the uncertainty propagates within the simulator and disentangling the epistemic uncertainty are also challenging (Punzo and Montanino 2020).

Data-driven approaches, on the other hand, prioritize learning the relationship between the input and output variables, thus requiring fewer assumptions. One of the most important methods is (extended) Kalman filter (Liu et al. 2006; Van Hinsbergen et al. 2011), in which total uncertainty is decomposed into process error and measurement error. It was applied to quantify the uncertainty of univariate traffic flow estimation (Wang and Papageorgiou 2005) and prediction (Guo, Huang, and Williams 2014), speed prediction (Guo and Williams 2010), and travel time estimation (Van Hinsbergen et al. 2011). Heteroscedastic Gaussian Process (HGP) (Rodrigues and Pereira 2018) is another representative example. The results show that large-scale crowd-sourced traffic data, specifically about speed, is subject to temporal uncertainty that varies over time. However, these models cannot evaluate the rareness of the current traffic state.

DNN-based UQ methods are also widely studied. Bayesian neural networks are used to give confidence intervals of univariate time series, such as travel time (van Hinsbergen, Van Lint, and Van Zuylen 2009) or traffic flow (Zheng, Lee, and Shi 2006). Graph neural networks (GNN) extend grid-like convolutional neural networks to graph structures (Kipf and Welling 2016) and have been applied to network-level traffic state estimation and forecasting because road networks can be naturally represented by a graph. In the literature, many GNN-based traffic forecasting models are proposed and some of them aim at quantifying predictive uncertainty, such as Bayesian GNN (Fu, Zhou, and Chen 2020), deep echo state networks (McDermott and Wikle 2019), and ensemble-based approaches (Chen et al. 2021; Del Ser et al. 2020). Recently, Mallick, Balaprakash, and Macfarlane (2022) and Qian et al. (2022) have employed the deep-ensemble method, with efforts to make it adaptive for traffic forecasting. However, previous papers primarily focussed on benchmarking accuracy and only presented the estimated uncertainty. The impact of aleatoric and epistemic uncertainties on traffic forecasting has not been thoroughly examined in these works.

Based on previous studies, this paper focuses on understanding the predictive uncertainty given by GNN ensembles from a traffic dynamics perspective. Particular emphasis is given to the implications for continuous data collection.

3. Method

This section describes the methodology employed to develop an uncertainty-aware traffic speed forecasting mode. We will present the deep-ensemble-based approach for quantifying uncertainty, the formulation of the problem, and the structure of the model in a sequential manner.

3.1. DE-based uncertainty quantification

In the overview, we mentioned that deep ensemble is one of the most robust UQ methods. In this section, we first briefly explain how DE works. Next, we will introduce the appropriate uncertainty metrics to be used and explain how uncertainty can be quantified.

Assume that a training dataset $\mathcal{D} = \{(\boldsymbol{x}_s, \boldsymbol{y}_s)\}_{s=1}^{S}$ is given. Here $(\boldsymbol{x}_s, \boldsymbol{y}_s)$ are observed input-output pairs. $\boldsymbol{X}$ and $\boldsymbol{Y}$ represent the input and the output random variables separately (such as traffic flow or speed). S is the number of samples. An ensemble of K randomly-initialized probabilistic models are independently trained by using this dataset, denoted as $\{\mathcal{M}_i\}_{i=1}^{K}$. Now for one specific test input $\boldsymbol{x}^*$, the DE is used to quantify the uncertainty of the output $\boldsymbol{y}^*$, whose probability distribution is denoted as $p(\boldsymbol{y}^*)$. Each model predicts the probability distribution of output so the K models give an ensemble of distributions $\{p_i(\boldsymbol{y}^* \mid \boldsymbol{X} = \boldsymbol{x}^*)\}_{i=1}^{K}$. We briefly note it as $\{p_i(\boldsymbol{y}^*)\}_{i=1}^{K}$ from now on.

Suppose there are a sufficient number of similar samples to $\boldsymbol{x}^*$ in the training set. In that case, the models can effectively learn and match the output distribution, regardless of how the parameters are initialized. The ensemble of output distributions is consistent. In this case, we say that the epistemic uncertainty is low and the estimation of $p(\boldsymbol{y}^*)$ is reliable. $p(\boldsymbol{y}^*)$ itself represents the irreducible aleatoric uncertainty. Conversely, if the test point $\boldsymbol{x}^*$ was rarely or almost never seen during training, the predictions will differ significantly due to the random initialization of parameters and the presence of unconstrained model behaviours in training. So the DE will give diverse distributions and thus the estimate of $p(\boldsymbol{y}^*)$ is unreliable. This is a so-called high epistemic uncertainty case. Figure 1 illustrates these concepts by a simple task, saying learning the Gaussian relationship $p(y \mid x) = \mathcal{N}(\mu(x), \sigma^2(x))$. The training set contains more samples in the centre but fewer samples around the boundary. The magnitude of noise σ^2 decays with $|x|$. $x^* \approx 0$ is a typical low epistemic and high aleatoric uncertainty test point. While $x^* \approx -3.5$ has significantly higher epistemic uncertainty due to the lack of data. We cannot give reliable predictions. In this example, gathering additional data or enhancing modelling techniques for the inputs located in the middle area, where the dataset already contains a sufficient number of samples, will not lead to an improvement in prediction accuracy. This is because the inherent noise within that region is irreducible and enough samples are already provided to learn the distribution. However, collecting more data around the borders can significantly reduce epistemic uncertainty and thus improve the total performance.

The discussion above gives a qualitative analysis of the DE-based UQ method. For quantitative estimation, three questions must be answered:

(1) What metrics should be used to measure uncertainty?
(2) How to quantitatively estimate them?
(3) What prior distributions should be used for a traffic speed forecasting task?

For the first question, there are two common metrics to represent the uncertainty of a continuous scalar random variable, *variance* and *differential entropy*. For a specific input, the mean and the variance of the output distribution given by the i-th model in the ensemble are denoted as μ_i and σ_i^2, respectively. Then the well-known *law of total variance* decomposes total uncertainty into epistemic

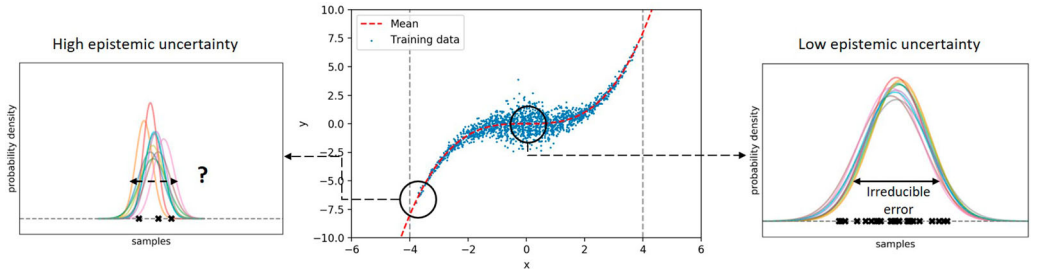

Figure 1. A 1D example of aleatoric uncertainty and epistemic uncertainty. Here 10 different models are used to learn $p(y \mid x)$ from the given training data. The left and right figure shows two ensembles of predicted distributions at two different test points, $x = -3.5$ (left) and $x = 0$ (right).

and aleatoric parts as follows (Lakshminarayanan, Pritzel, and Blundell 2016):

$$\mathrm{Var}(y^*) = \underbrace{\mathbb{E}(\sigma_i^2)}_{\text{aleatoric}} + \underbrace{\mathrm{Var}(\mu_i)}_{\text{epistemic}} \tag{1}$$

On the other side, differential entropy is defined as follows:

$$H(y^*) = -\int_Y p(y^*) \ln p(y^*)\, \mathrm{d}y^* \tag{2}$$

Here $p(y^*)$ is the posterior distribution of output, which can be approximated by the average distribution of the ensemble (if K is large enough):

$$p(y^*) \approx \frac{1}{K} \sum_{i=1}^{K} p_i(y^*) \tag{3}$$

Similarly, Ryabinin, Malinin, and Gales (2021) shows that the total entropy can be decomposed into the following terms:

$$H(y^*) = \underbrace{\mathbb{E}[H_i(y^*)]}_{\text{aleatoric}} + \underbrace{\mathbb{E}[D_{KL}(p_i(y^*) \| p(y^*))]}_{\text{epistemic}} \tag{4}$$

Here $H_i(y^*)$ is the differential entropy of each distribution in the ensemble and $D_{KL}(p_i(y^*) \| p(y^*))$ is the *Kullback–Leibler (KL) divergence* from each distribution to the (average) posterior distribution. KL divergence measures the directed 'distance' from the first distribution to the second one. It is non-negative. The definition is:

$$D_{KL}(p_i(y^*) \| p(y^*)) = \int_{p(y^*)>0} p_i(y^*) \ln \frac{p_i(y^*)}{p(y^*)}\, \mathrm{d}y^* \tag{5}$$

For the second question, Equations (1) and (4) give how to quantify both types of uncertainty. They have similar forms. The aleatoric uncertainty is measured by the average variance or differential entropy of the ensemble distribution, and the epistemic term measures the *diversity* of the distribution ensemble by the variance of their mean values (point estimate) or by their average distance to the posterior distribution. They are consistent with the example shown in Figure 1. Many articles rely on only one of the two metrics to assess predictive uncertainty. However, in this investigation, both metrics are assessed. The epistemic expressions presented in Equations (1) and (4) offer distinct advantages. Specifically, variance is scale-dependent but KL-divergence is not. Therefore, Equation (1) is suitable for estimating the remaining room for accuracy improvement and Equation (4) provides a more objective means of measuring the 'rareness' of input, which is a scale-independent concept.

For the third question, the answer depends on what quantity to predict. Since traffic flow [veh/h] is not a state variable (i.e. low flows may coincide with both free-flowing traffic and heavy congestion), and density [veh/km] is difficult to measure directly, speed [km/h] is an appropriate index of congestion. In the literature, there are two ways to learn the output distributions. The first one is the *parametric* approach, which means we assume that the prior distribution can be represented by a small set of parameters, such as Gaussian (Rodrigues and Pereira 2018; Yuan et al. 2021). Second, the non-parametric histogram regression used in motion prediction (Gu, Sun, and Zhao 2021) can assimilate the output distribution of traffic state variables directly. The output space is discretized and the model learns the probability that the prediction falls into each interval. Parametric probabilistic models are usually easier to train, but only some simple priors can be used. The histogram regression method, conversely, can approximate any distributions. Nevertheless, due to the distinctiveness of individual locations within a road network (they have different numbers of lanes, etc.), the histogram regression model cannot share weights across the network. This drawback substantially augments the complexity of the model. Therefore, non-parametric approaches are only appropriate for tasks that involve single or two scalar outputs, such as microscopic trajectory forecasting (Gilles et al. 2022) or dense object detection (Lin et al. 2017).

In this study, both approaches are explored. First, the parametric approach is used for quantifying two types of uncertainty. Here we choose *Beta distribution* prior due to the boundness of speed:

$$b(v;\alpha,\beta) = \frac{1}{B(\alpha,\beta)} v^{\alpha-1}(1-v)^{\beta-1} \tag{6}$$

where $B(\alpha,\beta)$ is a Beta function that normalizes the density function. $\alpha > 1$ and $\beta > 1$ must be satisfied to make sure that the distribution is bounded everywhere. The mean and variance are:

$$\mu = \frac{\alpha}{\alpha+\beta} \tag{7}$$

$$\sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \tag{8}$$

and its differential entropy is given by:

$$H = \ln B(\alpha,\beta) - (\alpha-1)\psi(\alpha) - (\beta-1)\psi(\beta) + (\alpha+\beta-2)\psi(\alpha+\beta) \tag{9}$$

$\psi()$ is the di-gamma function. For the standard Beta distribution, the random variable is defined between 0 and 1. Therefore, when training the model, the output speed is also linearly normalized between 0 and 1. During inference, the prediction is re-scaled up between 0 and the speed limit. To summarize, the uncertainty quantification process is:

(1) Train an ensemble of probabilistic forecasting models that are randomly initialized.
(2) Calculate the mean, variance, and differential entropy of each output distribution.
(3) The average of the mean values gives the predicted speed.
(4) For variance-based uncertainty metrics, directly use Equation (1).
(5) For entropy-based metrics, first approximate the posterior distribution by Equation (3), then use numerical methods to calculate the integrals in Equation (4).

We also trained a histogram-regression model to explain what factors restrict predictability. The speed is discretized into 1 km h^{-1} intervals from 0 to the speed limit so the task is converted to a classification problem. The learnt probability of all classes (speed interval) together gives the approximated output distribution.

The description above outlines the general procedure for uncertainty quantification methods. However, it is crucial to emphasize that the result of epistemic and aleatoric uncertainty is conditional on the model design, problem formulation, and input variables. In the next section, we will introduce

the variables that are included in the input and how traffic flow theory is embedded in the design of the data-driven model.

3.2. Problem formulation

The problem formulation entails determining the desired output and selecting the variables to be included in the input. As previously discussed, speed is a suitable metric for predicting traffic congestion. To forecast traffic speed in the near future, traffic flow theory suggests that recent observations of at least two variables among speed, flow, and density should be provided to establish the fundamental diagram. These variables possess a physical and causal relationship with the output. In practice, it is often noticed that the flow tends to increase approximately 10–20 m before a speed drop near an on-ramp. This observation typically indicates that more vehicles are entering the traffic stream. Consequently, the input variables for prediction include both speed and flow. Some studies also consider the week-to-week or day-to-day periodicity, incorporating the average historical traffic state. However, we argue that this connection is merely a correlation or potentially confounded by unobservable factors such as traffic demand. There is no direct causal relationship with the current traffic state. Including traffic states from past days may cause overfitting issues in the models.

Given a highway network with N links represented by a graph $\mathcal{G}$. Its adjacency matrix is noted as $\mathbf{A}_{N\times N}$. The time interval between two adjacent observations (δt) is fixed. We aim to predict the *marginal* distribution of speed (V) at each link i and time stamp t in the next T steps from the observed speed and flow (Q) in the past P steps. The problem formulation is modelling the following conditional probability distribution:

$$p(V_{i,t} \mid \mathbf{V}_{P\times N}, \mathbf{Q}_{P\times N}, \mathbf{A}_{N\times N}) \quad \forall\, 1 \le i \le N; \quad 1 \le t \le T \tag{10}$$

3.3. Model structure

Given that the objective is to obtain the probability distribution of multi-step speed, a fully convolutional model is more desirable than an RNN-based model as it obviates the need for generating the traffic flow at each time step in the decoder. The backbone of the proposed model inherits the flexible structure of the attention-based spatial-temporal graph convolutional networks (ASTGCN) proposed by Guo et al. (2019). ASTGCN comprises a series of consecutive spatiotemporal blocks (ST-Blocks) as the main body for extracting features. Nevertheless, the model design must conform to the principles of traffic flow theory. This ensures that the model can accurately learn the appropriate physical (causal) relationships and provide reliable uncertainty quantification. Failing to do so may result in the repetition of erroneous correlations, such as similar congestion patterns occurring at different locations, which would lead to underestimated or exaggerated aleatoric uncertainty. To address this concern, various adjustments have been made to the original ASTGCN model.

One ST-block's structure is shown in Figure 2. It sequentially contains a Dynamic Graph Convolutional (DGC) module proposed in Li, Knoop, and van Lint (2021), a temporal attention layer, a normal convolutional layer along the time axis, and an extra skip connection to avoid gradient vanishing in model training (He et al. 2016). Batch normalization (Santurkar et al. 2018) is applied at the end of each ST-block. In contrast to the spatial attention layer employed in ASTGCN, DGC considers the constraints imposed by traffic flow theory. There are two key considerations. First, the learned impacts of adjacent locations are dynamic and asymmetric in relation to upstream and downstream links due to the directional nature of traffic flow. This acknowledges that the influence of adjacent locations may vary based on the direction of traffic flow. Second, the connectivity degrees of adjacent matrices are regulated by considering the maximum propagation speed of congestion and the time interval. This ensures that only causally influential locations are considered in the model, thereby preventing the learning of spurious correlations. We refer the readers to Li, Knoop, and van Lint (2021) for more detailed explanations.

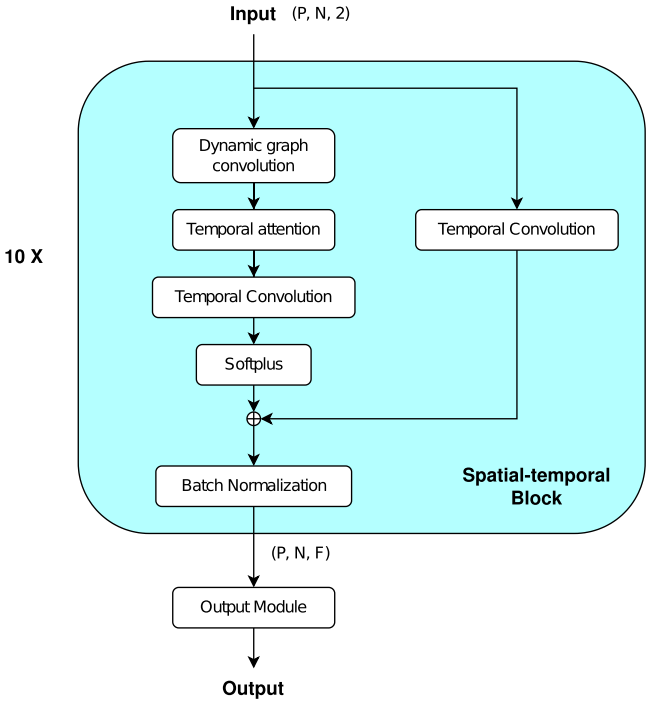

Figure 2. The structure of the proposed model.

The output shape of every ST-block is (P, N, F). Here F is the feature dimension. 10 ST blocks are stacked together to extract spatio-temporal features from inputs. More details of each layer in the ST-block can be found in the Appendix 1.

Following the stacked ST-blocks, the hidden representation is converted to the desired output shape by an output module. Different from the fully connected layers and the fusion module used in ASTGCN, the proposed model employs two new output modules. Their structures are shown in Figure 3. For modelling Beta distributions, we learn its mode ω and reduced concentration κ:

$$\begin{aligned} \kappa &= \alpha + \beta - 2 \\ \omega &= \frac{\alpha - 1}{\kappa} \end{aligned} \tag{11}$$

After applying two normal 2D convolutional layers, the hidden features are split into h_1 and h_2, and then they are activated by the following functions to give ω and κ, which are the target to learn:

$$\begin{aligned} \omega &= \text{sigmoid}(h_1) \\ \kappa &= \exp(h_2 + 3) \end{aligned} \tag{12}$$

Compared to directly learning α and β, this scheme is numerically more stable and converges faster (similar to the activation strategy used for Gaussian prior. See Kendall and Gal 2017 for more details). The loss function is negative-log-likelihood (NLL), defined as follows:

$$NLL = -\sum_{i,t} \ln b(V_{i,t}^{\text{label}};\ \omega_{i,t}, \kappa_{i,t}) \tag{13}$$

For histogram regression, assume that the speed is uniformly discretized into C intervals. The output module firstly adjusts the dimension to the desired shape by a 2D convolutional layer, then cross

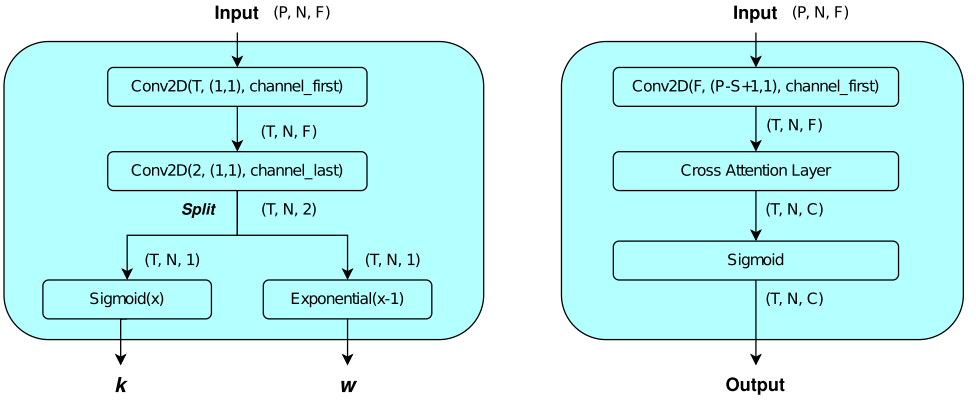

Figure 3. The structure of the output module. The left one is for Beta distributions and the right one is for histogram regression.

attention layers (Huang et al. 2019) map the features to each interval of each individual location. As mentioned before, the cross-attention layer shall not share parameters among different road links. Otherwise, the model will give similar predictions for all locations (the model fails). This non-shared strategy greatly increases the complexity of the model. Therefore, we only train one histogram-regression model for interpreting the estimated predictability instead of getting an ensemble of such models for uncertainty quantification.

More details of the cross-attention layer can be found in the Appendix 1. The final output is activated by the sigmoid function. Focal loss (Lin et al. 2017) is used to learn the unbalanced distribution of labels. Here Y_p is the ground truth of class (0 or 1) and $\hat{Y}_p$ is the predicted probability of each speed interval.

$$\text{Loss} = -\frac{1}{P}\sum_{p}(Y_p - \hat{Y}_p)^2 f(Y_p, \hat{Y}_p) \tag{14}$$

$$f(Y_p, \hat{Y}_p) = \begin{cases} \ln \hat{Y}_p & \text{if } Y_p = 1 \\ (1 - Y_p)^4 \ln(1 - \hat{Y}_p) & \text{else} \end{cases} \tag{15}$$

4. Experiments

In this section, experiments are carried out on a real-world highway network. All data used in this paper are collected and processed by the Dutch National Data Warehouse (NDW–www.ndw.nu). Dual inductive loop detectors on the freeways measure the speed and flow data. The raw data are aggregated to average speed and flow per lane for each loop detector. These loop detectors are not uniformly distributed on the target network. The Adaptive Smoothing Method (ASM) (Schreiter et al. 2010) is applied to project the aggregated data onto a uniform grid and to fill in missing values. In this study, we use the processed data.

The highway network around Amsterdam (the Netherlands) is selected as a case study. The network is shown in Figure 4. It consists of 9 highways that connect the Amsterdam city centre, several smaller towns, and Schiphol International Airport. This busy network contains rich congestion patterns. All highways are uniformly partitioned into 400 m length links, and we consider specific driving directions only (marked in Figure 4), resulting in a network of 193 links ($N = 193$). Both speed and flow are aggregated every 4 m by averaging. The observed speed and flow in the past 60 m ($P = 15$) is used to predict the evolution of speed in the next 40 m ($T = 10$).

The data for the entire year of 2018 is chosen as the training set. To focus on predicting traffic congestion, only congested periods are considered. The data preparation method is as follows. For one

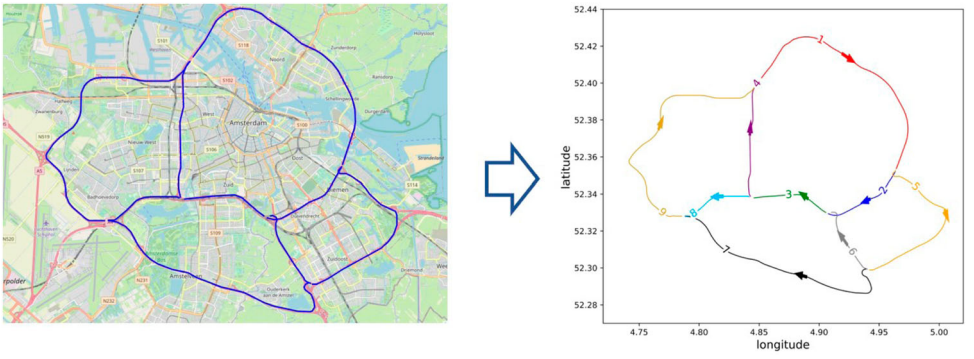

Figure 4. The selected highway network around Amsterdam. Arrows represent driving directions. The number of each road is also marked.

Table 1. Number of samples of used datasets.

	Training	Validation	2019-test	2022-test
Nb. of samples	12964	1830	4053	3163

moment t, if any position's speed is lower than 40 km h^{-1} within $(t - 20\,\text{m}, t + 20\,\text{m})$, then this sample is added in the data set. So the prepared data set includes diverse patterns ranging from pre- to post-congestion periods. We randomly select 15% samples from the training set as the validation set. To mimic the real-world model deployment and continuous data collection environment, two different test sets are prepared, Mar. 1st – May 31st of 2019 (noted as `2019-test`) and Mar. 1st – May 31st of 2022 (noted as `2022-test`). They are respectively before and after the lockdown measurements in the Netherlands. We expect that the congestion patterns may be different. The number of samples for all used data sets is listed in Table 1. `2022-test` has fewer samples than `2019-test` because it has less congestion.

Following the recommendation given in Lakshminarayanan, Pritzel, and Blundell (2016), we train a maximum of 15 randomly initialized Beta-based models for uncertainty quantification and 1 histogram-regression model for assimilating true distributions. The input speed and flow data are normalized by the z-score function for all datasets.

In addition, to validate that the proposed model is resilient to erroneous spatiotemporal correlations by design and can give reasonable uncertainty, we construct 3 imaginary and counterfactual datasets and carry out the corresponding extra experiments. The details and the results are provided in the Appendix 2.

5. Results and discussion

This section will present the experimental results of predictive accuracy, followed by an analysis of aleatoric uncertainty and epistemic uncertainty. Based on these results, we will address the primary research question: to what extent is macroscopic traffic state predictable, and what factors limit its predictability?

5.1. Accuracy

Here we consider three widely-used accuracy metrics, mean-absolute-error (MAE, which is L1 loss), mean-absolute-percentage-error (MAPE), and root-mean-square-error (RMSE, which is L2

Table 2. Performances of different ensemble sizes on both test sets.

Ensemble size K	MAE(km h^{-1})	MAPE(%)	RMSE(km h^{-1})	NLL
		2019-test		
1	5.95	15.96	11.77	−2.45
3	5.89	16.15	11.57	−2.46
5	5.83	16.27	11.55	−2.48
10	5.77	16.17	11.46	−2.49
15	5.76	16.12	11.46	−2.49
ASTGCN	5.89	16.25	11.31	–
		2022-test		
1	5.22	15.73	10.80	−2.51
3	5.07	12.15	10.37	−2.59
5	5.02	12.23	10.34	−2.60
10	4.96	12.12	10.25	−2.63
15	4.96	12.15	10.24	−2.64
ASTGCN	4.99	13.05	10.92	–

loss):

$$\text{MAE} = \frac{1}{TN}\sum_{i,t}|V_{i,t} - \hat{V}_{i,t}| \tag{16}$$

$$\text{MAPE} = \frac{1}{TN}\sum_{i,t}\frac{|V_{i,t} - \hat{V}_{i,t}|}{V_{i,t}} \tag{17}$$

$$\text{RMSE}^2 = \frac{1}{TN}\sum_{i,t}(V_{i,t} - \hat{V}_{i,t})^2 \tag{18}$$

Table 2 compares the prediction accuracy of the Beta-prior models used in this study with the original ASTGCN model that minimizes MSE (one model, not an ensemble). Overall, They have very close performances. We observe that the accuracy does not improve much for $K > 10$, which is consistent with the conclusion in Lakshminarayanan, Pritzel, and Blundell (2016). We emphasize that the focus of this study is quantifying uncertainty instead of the accuracy benchmark. We do not discuss deeply on the accuracy results.

Before analyzing uncertainty, it is necessary to evaluate how well the distributions are modelled. The quality of uncertainty quantification is evaluated by the method proposed in Kendall and Gal (2017) (explained in Appendix 3). The results of `2019-test` are presented in Figure 5. Figure 5(a) shows the relationship between the RMSE and the aleatoric/epistemic uncertainty thresholds. All curves are monotonically decreasing with the confidence level, which means that estimated uncertainty is strongly correlated with the confidence of predictions. Figure 5(b) further clarifies how well Beta distributions model the true distribution on average. The x-axis depicts the percentile of predicted distribution (expected confidence) and the y-axis is the percentage of observations that are indeed below this percentile (frequency). In the ideal case, the true distribution is perfectly modelled so the relationship between frequency and expected confidence should be $y = x$. In Figure 5(b), we see that the MAE distance to the ideal case increases with the prediction horizon. But the three curves are all close to the ideal line. It means that the uncertainty estimation is reliable on average.

The results above prove that the used Beta-prior model can well depict the output distributions in most cases. Next, we will focus on analyzing aleatoric and epistemic uncertainty.

5.2. Predictability of traffic congestion

Figure 6 shows the relationship between two types of uncertainty and the prediction horizon. On average, both epistemic and aleatoric uncertainty, and thus the total uncertainty, increases with the

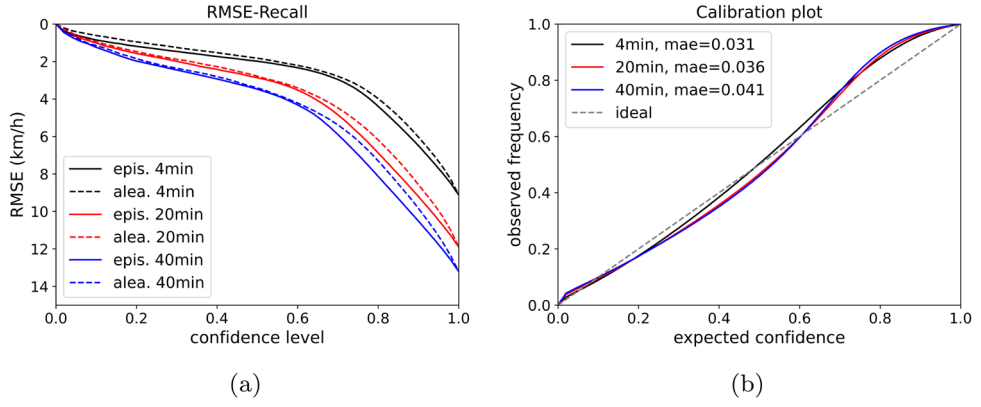

Figure 5. (a) RMSE-confidence threshold (recall) curves and (b) calibration plots for different single predictive steps on `2019-test`.

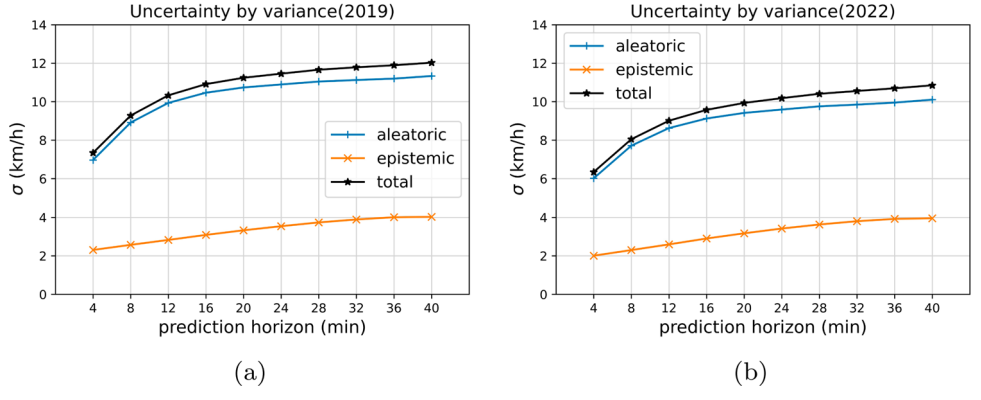

Figure 6. Relationship between the average aleatoric uncertainty, epistemic uncertainty, total uncertainty of each predictive step, and the prediction horizon. (a) `2019-test`; (b) `2022-test`. Notice that uncertainty is measured by σ here.

prediction horizon on both test sets. For all prediction steps, aleatoric uncertainty is significantly higher than epistemic uncertainty, which means that the inherent randomness of traffic dynamics overwhelmingly determines the total predictive error (RMSE). The remaining improvement room for better modelling and dataset expansion is limited.

Next, we compare different uncertainty metrics. In Figure 7, the distributions of two types of uncertainty measured by entropy metrics and variance metrics on both test sets are presented. We see that using either the variance or differential entropy, the distributions of aleatoric uncertainty are highly consistent. The `2022-test` set has lower average aleatoric uncertainty than `2019-test`. However, the distributions of epistemic uncertainty measured by variance or entropy are significantly different. Variance is scale-dependent so the epistemic uncertainty is positively correlated with the aleatoric uncertainty. The top right distributions are indeed very similar to those in the top left figure. There is no significant difference between the distributions of epistemic uncertainty measured by variance for the two test sets. The average epistemic uncertainty of `2022-test` is even slightly lower than `2019-test`. However, the entropy-based epistemic uncertainty metric is scale-independent. The bottom right figure clearly shows that `2022-test` has higher epistemic uncertainty than `2019-test` (see the two peaks). The difference originates from the fact that entropy-based metrics consider the diversity of distributions instead of only the diversity point estimates. For both test

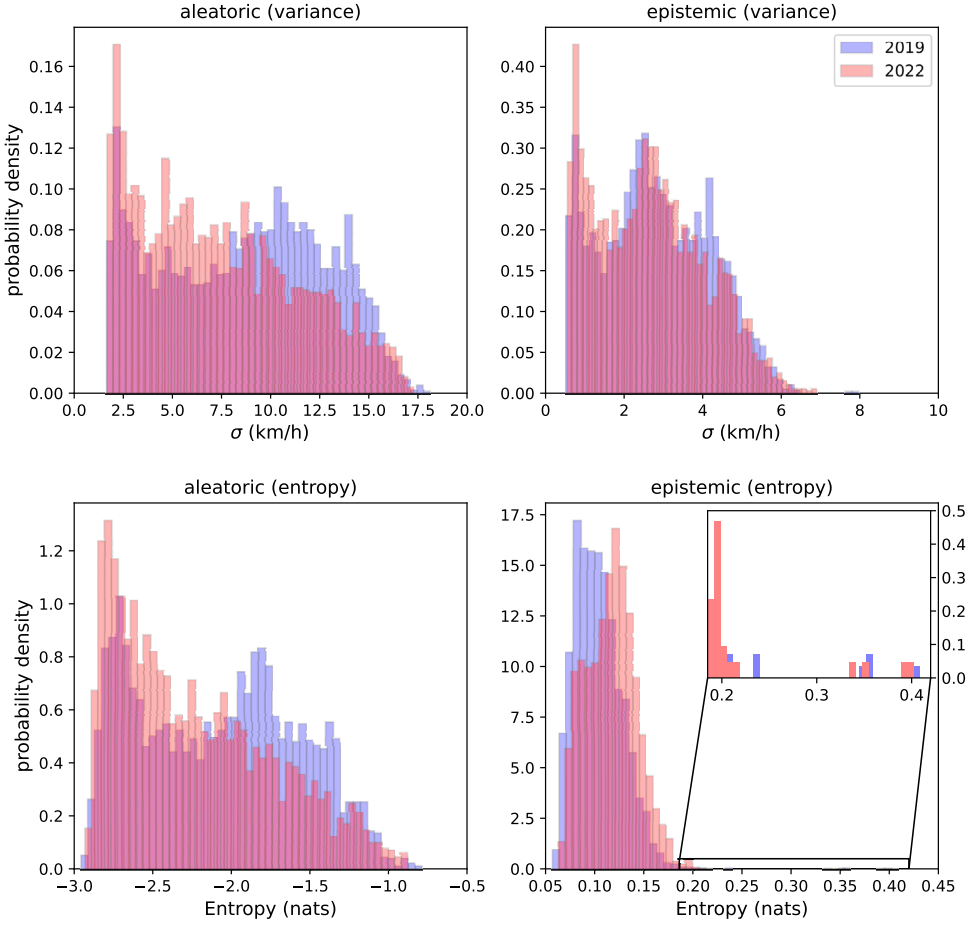

Figure 7. Distributions of prediction uncertainty on two test sets. Left column: aleatoric uncertainty; right column: epistemic uncertainty; top row: variance metrics; bottom row: entropy metrics.

sets, most samples are located at the low epistemic uncertainty end. Only 11 samples are identified as out-of-distribution cases due to large-scale loop detector failures.

In summary, the analysis above answers the first part of the question. The uncertainty quantified by the used model is reliable. Conditioned on this specific model, the predictability of the macroscopic traffic state (RMSE lower bound) is mainly determined by the irreducible aleatoric uncertainty. If using variance-based metrics, 2022 and 2019 have almost the same epistemic uncertainty. While using entropy-based metrics, we observe a significant shift between the 2019 and 2022 test sets (based on the 2018 training set). This result demonstrates that the variance-based metric can give the improvement room of RMSE accuracy. In contrast, the entropy-based metric is more suitable for measuring how 'rare' the current traffic state is.

5.3. Bi-modality of speed forecasting

This subsection answers the second part of the research question: why is aleatoric uncertainty so high, and what causes the low predictability? The key is using those concepts in traffic flow theory.

First of all, we visualize the statistical relationship between the estimated aleatoric uncertainty and the predicted speed. Considering that different locations on the highway network have different fundamental diagrams, it is reasonable to study this relationship for each specific link of the highway

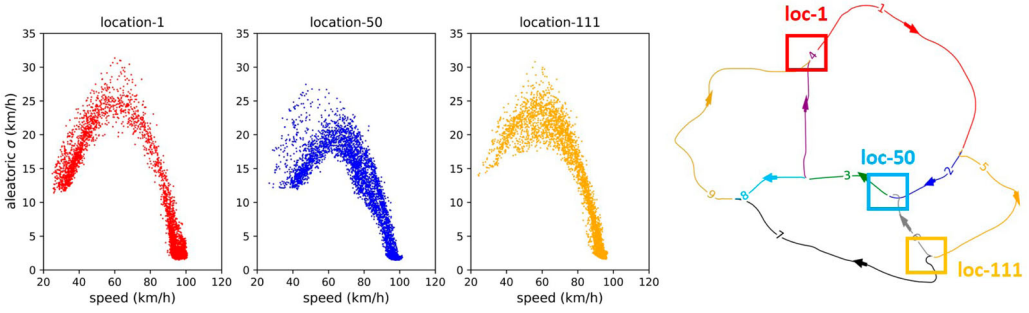

Figure 8. Relationships between the predicted speed and the aleatoric uncertainty at three positions around congestion bottlenecks. Here we only visualize the result of 20 m single-step predictions, but the conclusions hold for all prediction horizons.

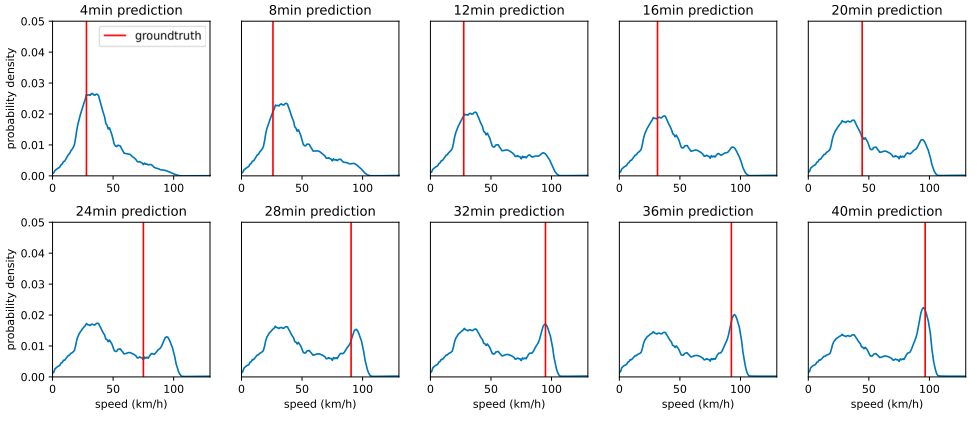

Figure 9. An example of predicted PDFs of speed given by histogram-regression model at link-55. The red lines are the evolution of the ground truth (labels).

network. We manually checked all 193 links. Three representative examples are given in Figure 8. These three links are around three frequently congested on-ramps (marked on the right figure). The aleatoric uncertainty-speed relationship has a consistently similar 'inverse U' shape. The model gives relatively lower aleatoric uncertainty for both congested (low speed) and free-flowing (high speed) predictions. Free-flowing prediction is even more certain. However, the inherent randomness is significantly higher around the transition (capacity) state (medium speed, 50 km h^{-1}–60 km h^{-1}), which represents the boundary between congested and free-flowing areas.

To explain this inverse-U relationship, we use the trained histogram-regression model to explicitly show the evolution of the approximated distribution of speed. Figure 9 presents an example at link-55. When the prediction horizon is short, the distribution is uni-modal (only one local maximum) and highly concentrated around the label, which means that it can be well approximated by a Beta distribution. With the increasing prediction horizon, the variance increases and the distribution gradually shows stronger bi-modality. One mode is the congested state and the other one is the free-flowing state. When such a bi-modal distribution is approximated by a uni-modal Beta distribution by minimizing NLL loss, the mean value (predicted point value) will be located at the middle (50 km h^{-1}–60 km h^{-1}, between two local maxima). The result directly interprets the inverse-U relationship observed in Figure 8.

Traffic flow theory explains the observed bi-modality in terms of the statistics of observed traffic states and the underlying explanatory mechanism. Figure 10(a) shows a typical speed-density relationship with 4-minute aggregate data from a typical location upstream of a major congested bottleneck,

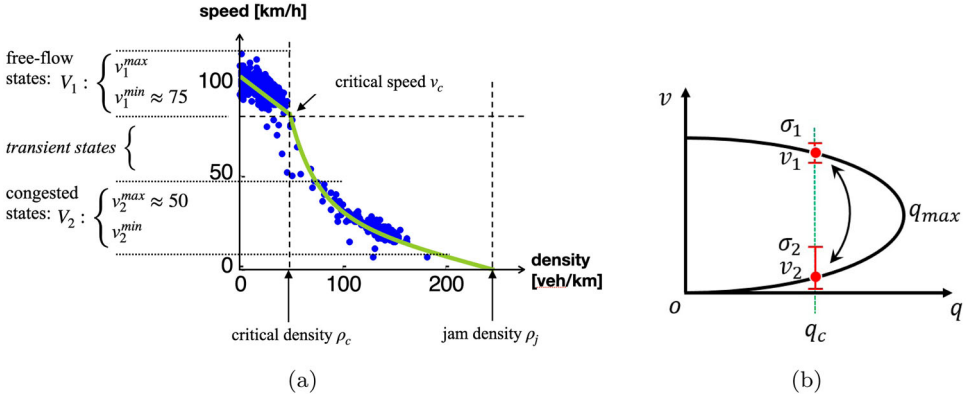

Figure 10. (a) A typical fundamental diagram measured at a location upstream of a major bottleneck. An approximate (Smulders) speed-density relationship is drawn over the measurements. Most observations fall into two intervals: free-flow states V_1 or (heavily) congested states V_2. (b) Fundamental diagram and capacity drop.

and a stylized Smulders fundamental diagram (FD) (Smulders 1988) drawn over the data. The data illustrate that one can roughly distinguish three types of observed traffic states: those related to freely-flowing traffic (speed interval V_1); (highly) congested conditions (speed interval V_2) and transient states between these. Clearly, transient states occur less frequently than either of the other two states. Transient states include passing shock waves between larger spatio-temporal regions of either free-flowing or congested traffic, in which due to acceleration and deceleration average speed is somewhere between V_1 and V_2. Most prolifically, this effect manifests as back-propagating stop-and-go wave patterns travelling upstream (often many kilometres). The precise frequency and onset of such waves are notoriously difficult to predict. Their emergence from microscopic disturbances (such as sudden braking or aggressive lane-changing behaviours) cannot be predicted based solely on aggregated traffic flow and speed information, no matter how fine the resolution of macroscopic data is.

Another explanation is based on the perspective of capacity drop, as shown in Figure 10(b). When congestion occurs, the capacity drops from $q_{\max}$ to q_c so traffic states may collapse to either intersected point on the fundamental diagram. The transient state between them is unstable and seldom observed. But which interval the traffic state will fall depends on microscopic driving behaviours that cannot be known from macroscopic data. We refer the readers to Helbing et al. (2009) for comprehensive analyses.

We now answer the second part of the question based on the non-parametric model. The long-term predictability (aleatoric uncertainty) of highway congestion is low because of the impossibility of predicting the precise time/frequency of congested wave patterns from speed and flow data, no matter what model is used and/or how big the flow/speed data set is. This uncertainty grows rapidly with the prediction horizon and will cause the bifurcation of future traffic states.

6. Conclusion and perspective

In this paper, we reconsider the network-level traffic speed forecasting problem from the perspective of uncertainty quantification. In this last section, we will first summarize the main findings and then propose several related research topics.

This study uses a deep ensemble of graph convolutional neural networks to quantify both aleatoric and epistemic uncertainty in traffic speed forecasting. The results from this specific model design give a crucial insight: when utilizing a full year of data and incorporating both speed and flow as inputs, the irreducible aleatoric uncertainty is significant and accounts for the majority of the total prediction error in highway speed forecasting. The stochastic nature of speed evolution on highways results in long-term congestion exhibiting substantially low predictability, given that only flow and speed

are considered as inputs. Additionally, in-depth analysis shows the bi-modality in the distribution of predicted speed, which can be explained by traffic flow theory. This bi-modality cannot be accurately predicted solely from averaged speed and flow data due to their limited information about the causal factors underlying the bifurcation. Regarding epistemic uncertainty, the results indicate that rare congestion patterns constitute only a small part of the data stream. Furthermore, compared to uncertainty-based metrics, entropy-based metrics are better indicators of the 'rareness' of samples.

These findings conclude that neither investment in collecting more 'rare' (corner-case) speed and flow data, nor the development of more sophisticated models will lead to substantial improvement in traffic forecasting, since this may reduce epistemic uncertainty only. However, we also emphasize this conclusion needs to be substantiated with empirical evidence on different scales and different types of road networks.

We close with several ideas for further research. First, enriching the diversity of data types may be beneficial in reducing aleatoric uncertainty. For example, adding microscopic traffic quantities, such as vehicle trajectory data, may increase the predictability of congestion because they are the potential causation of the bi-modality of speed. How to effectively fuse these data from different levels into one predictive framework is a challenging topic. Second, in this paper, the rareness metric is directly computed from all predicted points. But in practice, we may be more interested in low-speed areas. So this definition could be combined with congestion extraction and classification techniques, such as Nguyen et al. (2019), to quantitatively study the recurrence of congestion patterns and have a bird-view of the long-term evolution of traffic. A third idea pertains to histogram regression models. This method can also be combined with deep ensembles to quantify uncertainty. One of the biggest advantages of this model is that we do not need to assume the prior form of output distribution so the uncertainty estimation may be more precise. However, this model has higher computational complexity and it is more difficult to train. How to address these challenges also needs more investigation.

Acknowledgments

This research is sponsored by the NWO/TTW project MiRRORS with grant agreement number 16270. We thank them for supporting this study.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

Abdar, Moloud, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, and Paul Fieguth. 2021. "A Review of Uncertainty Quantification in Deep Learning: Techniques, Applications and Challenges." *Information Fusion* 76:243–297. https://doi.org/10.1016/j.inffus.2021.05.008.

Amini, Alexander, Wilko Schwarting, Ava Soleimany, and Daniela Rus. 2020. "Deep Evidential Regression." In *Neural Information Processing Systems*. Vancouver.

Beck, M. Bruce. 1987. "Water Quality Modeling: A Review of the Analysis of Uncertainty." *Water Resources Research* 23 (8): 1393–1442. https://doi.org/10.1029/WR023i008p01393.

Ben-Akiva, Moshe. 1998. "Dynamit, A Simulation-Based System for Traffic Prediction and Guidance Generation."

Castro-Neto, Manoel, Young-Seon Jeong, Myong-Kee Jeong, and Lee D. Han. 2009. "Online-SVR for Short-term Traffic Flow Prediction Under Typical and Atypical Traffic Conditions." *Expert Systems with Applications* 36 (3): 6164–6173. https://doi.org/10.1016/j.eswa.2008.07.069.

Chen, Xinqiang, Huixing Chen, Yongsheng Yang, Huafeng Wu, Wenhui Zhang, Jiansen Zhao, and Yong Xiong. 2021. "Traffic Flow Prediction by An Ensemble Framework with Data Denoising and Deep Learning Model." *Physica A: Statistical Mechanics and Its Applications* 565:125574. https://doi.org/10.1016/j.physa.2020.125574.

Chen, Bor-Sen, Sen-Chueh Peng, and Ku-Chen Wang. 2000. "Traffic Modeling, Prediction, and Congestion Control for High-speed Networks: A Fuzzy AR Approach." *IEEE Transactions on Fuzzy Systems* 8 (5): 491–508. https://doi.org/10.1109/91.873574.

Chu, Kang-Ching, Li Yang, Romesh Saigal, and Kazuhiro Saitou. 2011. "Validation of Stochastic Traffic Flow Model with Microscopic Traffic Simulation." In *IEEE International Conference on Automation Science and Engineering*. Trieste.

Daganzo, Carlos F. 1997. *Fundamentals of Transportation and Traffic Operations*. 1st edition. Emerald Publishing.
Davis, Gary A., and Nancy L. Nihan. 1991. "Nonparametric Regression and Short-term Freeway Traffic Forecasting." *Journal of Transportation Engineering* 117 (2): 178–188. https://doi.org/10.1061/(ASCE)0733-947X(1991)117:2(178).
Del Ser, Javier, Ibai Lana, Eric L. Manibardo, Izaskun Oregi, Eneko Osaba, Jesus L. Lobo, Miren Nekane Bilbao, and Eleni I. Vlahogianni. 2020. "Deep Echo State Networks for Short-Term Traffic Forecasting: Performance Comparison and Statistical Assessment." In *IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*. Rhodes.
Der Kiureghian, Armen, and Ove Ditlevsen. 2009. "Aleatory Or Epistemic? Does it Matter." *Structural Safety* 31 (2): 105–112. https://doi.org/10.1016/j.strusafe.2008.06.020.
Deser, Clara, Adam Phillips, Vincent Bourdette, and Haiyan Teng. 2012. "Uncertainty in Climate Change Projections: The Role of Internal Variability." *Climate Dynamics* 38 (3): 527–546. https://doi.org/10.1007/s00382-010-0977-x.
Fort, Stanislav, Huiyi Hu, and Balaji Lakshminarayanan. 2019. "Deep Ensembles: A Loss Landscape Perspective." Preprint arXiv:1912.02757.
Fu, Jun, Wei Zhou, and Zhibo Chen. 2020. "Bayesian Spatio-Temporal Graph Convolutional Network for Traffic Forecasting." Preprint arXiv:2010.07498.
Gehrke, Jan D., and Janusz Wojtusiak. 2008. "Traffic Prediction for Agent Route Planning." In *International Conference on Computational Science*. Krakow.
Gilles, Thomas, Stefano Sabatini, Dzmitry Tsishkou, Bogdan Stanciulescu, and Fabien Moutarde. 2022. "Gohome: Graph-Oriented Heatmap Output for Future Motion Estimation." In *2022 International Conference on Robotics and Automation (ICRA)*. Philadelphia.
Gu, Junru, Chen Sun, and Hang Zhao. 2021. "Densetnt: End-to-end Trajectory Prediction From Dense Goal Sets." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. Montreal.
Gu, Chuanye, Changzhi Wu, Yonghong Wu, and Benchawan Wiwatanapataphee. 2022. "Distributionally Robust Ramp Metering Under Traffic Demand Uncertainty." *Transportmetrica B: Transport Dynamics*10:652–666.
Guo, Jianhua, Wei Huang, and Billy M. Williams. 2014. "Adaptive Kalman Filter Approach for Stochastic Short-term Traffic Flow Rate Prediction and Uncertainty Quantification." *Transportation Research Part C: Emerging Technologies* 43:50–64. https://doi.org/10.1016/j.trc.2014.02.006.
Guo, Shengnan, Youfang Lin, Ning Feng, Chao Song, and Huaiyu Wan. 2019. "Attention Based Spatial-Temporal Graph Convolutional Networks for Traffic Flow Forecasting." In *Proceedings of the AAAI Conference on Artificial Intelligence*. Honolulu.
Guo, Jianhua, and Billy M. Williams. 2010. "Real-time Short-term Traffic Speed Level Forecasting and Uncertainty Quantification Using Layered Kalman Filters." *Transportation Research Record* 2175 (1): 28–37. https://doi.org/10.3141/2175-04.
He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. "Identity Mappings in Deep Residual Networks." In *Computer Vision–ECCV 2016: 14th European Conference, Proceedings, Part IV 14*, 630–645. Amsterdam, The Netherlands: Springer.
Helbing, Dirk, Martin Treiber, Arne Kesting, and Martin Schönhof. 2009. "Theoretical Vs. Empirical Classification and Prediction of Congested Traffic States." *The European Physical Journal B* 69 (4): 583–598. https://doi.org/10.1140/epjb/e2009-00140-5.
Helton, Jon C. 1993. "Uncertainty and Sensitivity Analysis Techniques for Use in Performance Assessment for Radioactive Waste Disposal." *Reliability Engineering & System Safety* 42 (2-3): 327–367. https://doi.org/10.1016/0951-8320(93)90097-I.
Huang, Zilong, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. 2019. "Ccnet: Criss-Cross Attention for Semantic Segmentation." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. Seoul.
Kendall, Alex, and Yarin Gal. 2017. "What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?" Preprint arXiv:1703.04977.
Kim, Jiwon, Hani S. Mahmassani, Peter Vovsha, Yannis Stogios, and Jing Dong. 2013. "Scenario-based Approach to Analysis of Travel Time Reliability with Traffic Simulation Models." *Transportation Research Record* 2391 (1): 56–68. https://doi.org/10.3141/2391-06.
Kipf, Thomas N., and Max Welling. 2016. "Semi-Supervised Classification with Graph Convolutional Networks." Preprint arXiv:1609.02907.
Kupinski, Matthew A., John W. Hoppin, Eric Clarkson, and Harrison H. Barrett. 2003. "Ideal-observer Computation in Medical Imaging with Use of Markov-chain Monte Carlo Techniques." *JOSA A* 20 (3): 430–438. https://doi.org/10.1364/JOSAA.20.000430.
Lakshminarayanan, Balaji, Alexander Pritzel, and Charles Blundell. 2016. "Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles." Preprint arXiv:1612.01474.
Li, Jia, Qian-Yong Chen, Haizhong Wang, and Daiheng Ni. 2012. "Analysis of LWR Model with Fundamental Diagram Subject to Uncertainties." *Transportmetrica* 8 (6): 387–405. https://doi.org/10.1080/18128602.2010.521532.
Li, Guopeng, Victor L. Knoop, and Hans van Lint. 2021. "Multistep Traffic Forecasting by Dynamic Graph Convolution: Interpretations of Real-time Spatial Correlations." *Transportation Research Part C: Emerging Technologies* 128:103185. https://doi.org/10.1016/j.trc.2021.103185.

Li, Guopeng, Victor L. Knoop, and Hans van Lint. 2022. "Estimate the Limit of Predictability in Short-term Traffic Forecasting: An Entropy-based Approach." *Transportation Research Part C: Emerging Technologies* 138:103607. https://doi.org/10.1016/j.trc.2022.103607.

Li, Yaguang, Rose Yu, Cyrus Shahabi, and Yan Liu. 2017. "Diffusion Convolutional Recurrent Neural Network: Data-driven Traffic Forecasting." Preprint arXiv:1707.01926.

Liebig, Thomas, Nico Piatkowski, Christian Bockermann, and Katharina Morik. 2017. "Dynamic Route Planning with Real-time Traffic Predictions." *Information Systems* 64:258–265. https://doi.org/10.1016/j.is.2016.01.007.

Lin, Tsung-Yi, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. "Focal Loss for Dense Object Detection." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. Venice.

Liu, Hao, Henk Van Zuylen, Hans Van Lint, and Maria Salomons. 2006. "Predicting Urban Arterial Travel Time with State-space Neural Networks and Kalman Filters." *Transportation Research Record*1968 (1): 99–108. https://doi.org/10.1177/0361198106196800112.

Ma, Xiaolei, Zhuang Dai, Zhengbing He, Jihui Ma, Yong Wang, and Yunpeng Wang. 2017. "Learning Traffic As Images: A Deep Convolutional Neural Network for Large-scale Transportation Network Speed Prediction." *Sensors* 17 (4): 818. https://doi.org/10.3390/s17040818.

Malinin, Andrey, Sergey Chervontsev, Ivan Provilkov, and Mark Gales. 2020. "Regression Prior Networks." Preprint arXiv:2006.11590.

Malinin, Andrey, Bruno Mlodozeniec, and Mark Gales. 2019. "Ensemble Distribution Distillation." Preprint arXiv:1905.00076.

Mallick, Tanwi, Prasanna Balaprakash, and Jane Macfarlane. 2022. "Deep-Ensemble-Based Uncertainty Quantification in Spatiotemporal Graph Neural Networks for Traffic Forecasting." Preprint arXiv:2204.01618.

McDermott, Patrick L., and Christopher K. Wikle. 2019. "Deep Echo State Networks with Uncertainty Quantification for Spatio-temporal Forecasting." *Environmetrics* 30 (3): e2553. https://doi.org/10.1002/env.v30.3.

Nguyen, Tin T., Panchamy Krishnakumari, Simeon C. Calvert, Hai L. Vu, and Hans Van Lint. 2019. "Feature Extraction and Clustering Analysis of Highway Congestion." *Transportation Research Part C: Emerging Technologies* 100:238–258. https://doi.org/10.1016/j.trc.2019.01.017.

Ovadia, Yaniv, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. 2019. "Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift." Preprint arXiv:1906.02530.

Punzo, Vincenzo, and Marcello Montanino. 2020. "A Two-level Probabilistic Approach for Validation of Stochastic Traffic Simulations: Impact of Drivers' Heterogeneity Models." *Transportation Research Part C: Emerging Technologies* 121:102843. https://doi.org/10.1016/j.trc.2020.102843.

Qian, Weizhu, Dalin Zhang, Yan Zhao, Kai Zheng, and James J. Q. Yu. 2022. "Uncertainty Quantification for Traffic Forecasting: A Unified Approach." Preprint arXiv:2208.05875.

Qiao, Fengxiang, Hai Yang, and William H. K. Lam. 2001. "Intelligent Simulation and Prediction of Traffic Flow Dispersion." *Transportation Research Part B: Methodological* 35 (9): 843–863. https://doi.org/10.1016/S0191-2615(00)00024-2.

Ritter, Hippolyt, Aleksandar Botev, and David Barber. 2018. "A Scalable Laplace Approximation for Neural Networks." In *6th International Conference on Learning Representations, ICLR*. Vancouver.

Rodrigues, Filipe, and Francisco C. Pereira. 2018. "Heteroscedastic Gaussian Processes for Uncertainty Modeling in Large-scale Crowdsourced Traffic Data." *Transportation Research Part C: Emerging Technologies* 95:636–651. https://doi.org/10.1016/j.trc.2018.08.007.

Ryabinin, Max, Andrey Malinin, and Mark Gales. 2021. "Scaling Ensemble Distribution Distillation to Many Classes with Proxy Targets." *Advances in Neural Information Processing Systems* 34:6023–6035.

Santurkar, Shibani, Dimitris Tsipras, Andrew Ilyas, and Aleksander Madry. 2018. "How Does Batch Normalization Help Optimization?." In *Advances in Neural Information Processing Systems*. Montreal.

Schreiter, Thomas, Hans van Lint, Martin Treiber, and Serge Hoogendoorn. 2010. "Two Fast Implementations of the Adaptive Smoothing Method Used in Highway Traffic State Estimation." In *13th International IEEE Conference on Intelligent Transportation Systems*. Madeira Island.

Seo, Toru, Alexandre M. Bayen, Takahiko Kusakabe, and Yasuo Asakura. 2017. "Traffic State Estimation on Highway: A Comprehensive Survey." *Annual Reviews in Control* 43:128–151. https://doi.org/10.1016/j.arcontrol.2017.03.005.

Sharma, Anshuman, Zuduo Zheng, and Ashish Bhaskar. 2019. "Is More Always Better? The Impact of Vehicular Trajectory Completeness on Car-following Model Calibration and Validation." *Transportation Research Part B: Methodological* 120:49–75. https://doi.org/10.1016/j.trb.2018.12.016.

Smulders, Stef A. 1988. "Control of Freeway Traffic Flow." Technical Report.

Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. "Dropout: a Simple Way to Prevent Neural Networks From Overfitting." *The Journal of Machine Learning Research* 15 (1): 1929–1958.

Swiatkowski, Jakub, Kevin Roth, Bastiaan Veeling, Linh Tran, Joshua Dillon, Jasper Snoek, Stephan Mandt. 2020. "The k-tied Normal Distribution: A Compact Parameterization of Gaussian Mean Field Posteriors in Bayesian Neural Networks." In *International Conference on Machine Learning*. Stanford.

Tahmasbi, Rasool, and S. Mehdi Hashemi. 2013. "Modeling and Forecasting the Urban Volume Using Stochastic Differential Equations." *IEEE Transactions on Intelligent Transportation Systems* 15 (1): 250–259. https://doi.org/10.1109/TITS.2013.2278614.

van Amersfoort, Joost, Lewis Smith, Yee Whye Teh, and Yarin Gal. 2020. "Simple and scalable epistemic uncertainty estimation using a single deep deterministic neural network."

Van Hinsbergen, Chris P. I. J., Thomas Schreiter, Frank S. Zuurbier, J. W. C. Van Lint, and Henk J. Van Zuylen. 2011. "Localized Extended Kalman Filter for Scalable Real-time Traffic State Estimation." *IEEE Transactions on Intelligent Transportation Systems* 13 (1): 385–394. https://doi.org/10.1109/TITS.2011.2175728.

van Hinsbergen, C. P. I. J., J. W. C. Van Lint, and H. J. Van Zuylen. 2009. "Bayesian Committee of Neural Networks to Predict Travel Times with Confidence Intervals." *Transportation Research Part C: Emerging Technologies* 17 (5): 498–509. https://doi.org/10.1016/j.trc.2009.04.007.

Van Lint, J. W. C., S. P. Hoogendoorn, and Henk J. van Zuylen. 2002. "Freeway Travel Time Prediction with State-space Neural Networks: Modelling State-space Dynamics with Recurrent Neural Networks." *Transportation Research Record* 1811 (1): 30–39. https://doi.org/10.3141/1811-04.

Van Lint, J. W. C., S. P. Hoogendoorn, and Henk J. van Zuylen. 2005. "Accurate Freeway Travel Time Prediction with State-space Neural Networks Under Missing Data." *Transportation Research Part C: Emerging Technologies* 13 (5-6): 347–369. https://doi.org/10.1016/j.trc.2005.03.001.

van Lint, H., O. Miete, H. Taale, and S. Hoogendoorn. 2012. "Systematic Framework for Assessing Traffic Measures and Policies on Reliability of Traffic Operations and Travel Time." *Transportation Research Record* 2302 (1): 92–101. https://doi.org/10.3141/2302-10.

Wang, Yibing, and Markos Papageorgiou. 2005. "Real-time Freeway Traffic State Estimation Based on Extended Kalman Filter: a General Approach." *Transportation Research Part B: Methodological* 39 (2): 141–167. https://doi.org/10.1016/j.trb.2004.03.003.

Wang, Y., M. Papageorgiou, and A. Messmer. 2006. "A Real Time Freeway Network Traffic Surveillance Tool." *IEEE Transactions on Control Systems Technology* 14 (1): 18–32. https://doi.org/10.1109/TCST.2005.859636.

Wang, R., D. B. Work, and R. Sowers. 2016. "Multiple Model Particle Filter for Traffic Estimation and Incident Detection." *IEEE Transactions on Intelligent Transportation Systems* 17 (12): 3461–3470. https://doi.org/10.1109/TITS.2016.2560769.

Xu, Dongwei, Zhenqian Lin, Lei Zhou, Haijian Li, and Ben Niu. 2022. "A GATs-GAN Framework for Road Traffic States Forecasting." *Transportmetrica B: Transport Dynamics* 10:718–730.

Yuan, Haitao, and Guoliang Li. 2021. "A Survey of Traffic Prediction: From Spatio-Temporal Data to Intelligent Transportation." *Data Science and Engineering*6 (1): 63–85. https://doi.org/10.1007/s41019-020-00151-z.

Yuan, Yun, Zhao Zhang, Xianfeng Terry Yang, and Shandian Zhe. 2021. "Macroscopic Traffic Flow Modeling with Physics Regularized Gaussian Process: A New Insight Into Machine Learning Applications in Transportation." *Transportation Research Part B: Methodological* 146:88–110. https://doi.org/10.1016/j.trb.2021.02.007.

Zechin, Douglas, and Helena Beatriz Bettella Cybis. 2023. "Probabilistic Traffic Breakdown Forecasting Through Bayesian Approximation Using Variational LSTMs." *Transportmetrica B: Transport Dynamics*11:1026–1044.

Zhang, Liang, Jianqing Wu, Jun Shen, Ming Chen, Rui Wang, Xinliang Zhou, Cankun Xu, Quankai Yao, and Qiang Wu. 2021. "SATP-GAN: Self-attention Based Generative Adversarial Network for Traffic Flow Prediction." *Transportmetrica B: Transport Dynamics* 9:552–568.

Zheng, Fangfang, Saif Eddin Jabari, Henry X. Liu, and DianChao Lin. 2018. "Traffic State Estimation Using Stochastic Lagrangian Dynamics." *Transportation Research Part B: Methodological* 115:143–165. https://doi.org/10.1016/j.trb.2018.07.004.

Zheng, Weizhong, Der-Horng Lee, and Qixin Shi. 2006. "Short-term Freeway Traffic Flow Prediction: Bayesian Combined Neural Network Approach." *Journal of Transportation Engineering* 132 (2): 114–121. https://doi.org/10.1061/(ASCE)0733-947X(2006)132:2(114).

Appendices

Appendix 1. Details about the model

The input tensor is concatenated speed and flow at N links in the past P time steps, so the input shape is $(P, N, 2)$. We first present the details of one ST-block. The first module in an ST-block is the dynamic graph convolutional (DGC) module proposed in Li, Knoop, and van Lint (2021). It learns input-dependent kernels instead of static kernels for applying graph convolution. We refer the readers to the paper for more details. This DGC module has two hyperparameters, output dimension F_{out} and the order of adjacent matrix k. The DGC module is applied to every time step. We briefly note it as:

$$\boldsymbol{H}_{(P,N,F_{out})} = \mathrm{DGC}(\boldsymbol{X}_{(P,N,F_{in})}; k, F_{out}) \tag{A1}$$

Table A1. Estimated uncertainty for both test sets.

Alea.(km h^{-1})	Epis.(km h^{-1})	Total(km h^{-1})
9.92	0.07	9.92

The temporal attention layer is a global attention layer along the time axis, it does not have any hyper-parameters and the output has the same shape as input. We note the input $\boldsymbol{X}_{(P,N,F_{in})}$ and its transpose $\boldsymbol{X}^T_{(P,F_{in},N)}$, then the layer writes:

$$\boldsymbol{Q} = \boldsymbol{X}^T\boldsymbol{W}^q, \quad \boldsymbol{K} = \boldsymbol{X}\boldsymbol{W}^k, \quad \boldsymbol{V} = \boldsymbol{X}\boldsymbol{W}^v \tag{A2}$$

$$\boldsymbol{H} = \text{softmax}(\boldsymbol{W}^c\boldsymbol{Q}\boldsymbol{K}^T)\boldsymbol{V} \tag{A3}$$

The trainable parameters are $\boldsymbol{W}^q \in \mathbb{R}^{F_{in}\times N\times N}, \boldsymbol{W}^k \in \mathbb{R}^{F_{in}}, \boldsymbol{W}^v \in \mathbb{R}^{F_{in}\times F_{in}}, \boldsymbol{W}^c \in \mathbb{R}^{P\times P}$. Then the output $\boldsymbol{H}$ has the same shape as $\boldsymbol{X}$. We briefly noted the process above as:

$$\boldsymbol{H}_{(P,N,F_{out}=F_{in})} = \text{TA}(\boldsymbol{X}_{(P,N,F_{in})}) \tag{A4}$$

The two temporal convolutional layers share the same hyperparameters, the length of the kernel L. Their number of channels is the same as the input and the zero-padding is used then these two layers do not change the tensor shapes. The activation, batch normalization, and residual connection are shown in Figure 2.

In summary, one ST-block has only three hyperparameters, k, F_{out}, L. In this study we choose $k = 5, F_{out} = 64, L = 5$. After applying 10 ST-blocks the output has the shape $(P, N, F_{out} = 64)$.

The output module for learning Beta distribution is easy to understand. Next, we present the cross-attention layer used in the histogram regression module. The input tensor's shape is $\boldsymbol{X}_{T,N,F_{in}}$. The speed range is uniformly discretized into C intervals, noted as $\boldsymbol{V}_{C,1}$. Considering that different locations have different numbers of lanes and speed limits, the cross attention to each interval should be location-dependent. So the query speed tensor $\boldsymbol{V}$ must be duplicated N times, noted as $\boldsymbol{Z}_{N,C,1}$. Then the cross-attention layer writes:

$$\boldsymbol{Q} = \boldsymbol{Z}\boldsymbol{W}^q, \quad \boldsymbol{K} = \boldsymbol{X}\boldsymbol{W}^k \tag{A5}$$

$$\boldsymbol{H} = \boldsymbol{Q}\boldsymbol{K}^T \tag{A6}$$

The trainable parameters are $\boldsymbol{W}^q \in \mathbb{R}^{1\times F_{in}}$ and $\boldsymbol{W}^q \in \mathbb{R}^{F_{in}\times F_{in}}$. Then the output tensor has the shape (T, N, C).

Appendix 2. Counterfactual experiments

A.1 Aleatoric uncertainty only

The first test is constructing a dataset that exclusively includes aleatoric uncertainty. In this case, we achieve this by replacing the labels in such a way that, regardless of the input's historical traffic state, the output is always white noise (Gaussian) centred at 80 km h^{-1} with a standard variance of 10 km h^{-1}. Consequently, the output becomes independent of the input.

We proceed to retrain the models using this dataset and provide an overview of the estimated uncertainty metrics in Table A1. As expected, we observe that the learned aleatoric uncertainty is close to 10 km h^{-1}, while the epistemic uncertainty approaches zero. However, it is important to note that attaining a strictly zero value for epistemic uncertainty is not feasible in practice due to the inherent fluctuations when training neural networks.

A.2 High epistemic uncertainty

The second task involves constructing a training set consisting solely of rare congestion patterns to investigate whether the given epistemic uncertainty can be increased. For this purpose, we execute the pre-trained deep ensemble on the training set and retain only the samples whose epistemic uncertainty ranks in the top 20%. As a result, we obtain a smaller training set comprising 2600 samples. Subsequently, we train another deep ensemble using the same methodology on this reduced dataset and examine the estimated uncertainty on the test sets. The results are depicted in Figure A1.

From the findings, it becomes evident that due to the smaller size yet diverse patterns of the training set, the proportion of epistemic uncertainty is significantly higher compared to the results illustrated in Figure 6. In fact, it is almost on par with the aleatoric uncertainty in this scenario, leading to an overall increase in total uncertainty. With such high epistemic uncertainty, the estimates of aleatoric uncertainty are not reliable either. These outcomes demonstrate that epistemic uncertainty can effectively reflect the rarity of a test sample based on the specific training set employed (whether the model has seen similar cases enough times during training).

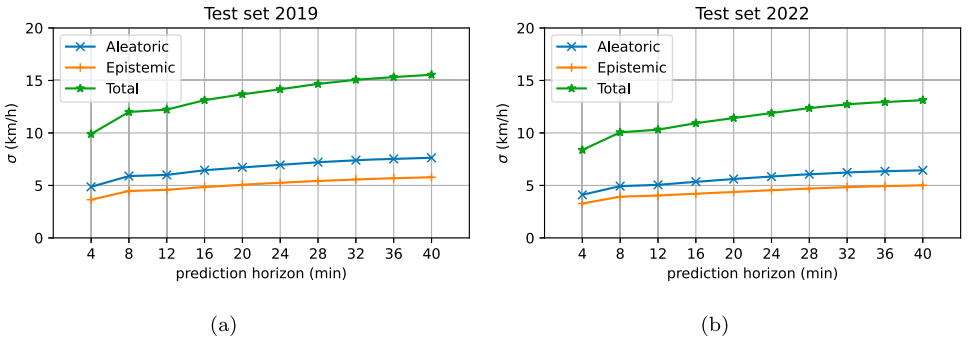

Figure A1. Uncertainty-horizon relationships on both test sets.

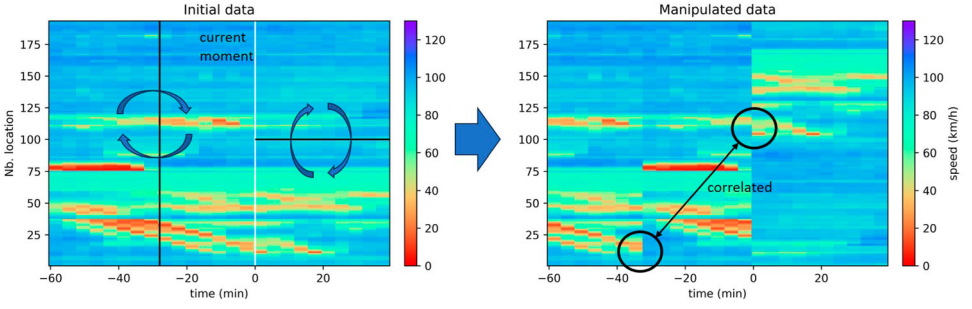

Figure A2. Manipulation of the traffic data. 0 min means the current moment.

A.3 Mismatched spatiotemporal correlation

The third experiment is designed to demonstrate the robustness of the method and its ability to avoid being influenced by spurious spatiotemporal correlations. To achieve this, we construct a counterfactual dataset that deliberately disrupts the spatiotemporal congestion features. This manipulation involves making changes to both the input and output as follows:

- Input: We alter the order of times by placing the most recent 7 steps at the beginning, effectively rearranging the temporal sequence.
- Output: We swap the first 100 locations with the remaining locations, thereby disrupting the spatial correlation.

The processing steps for this manipulation are illustrated in Figure A2. By creating the mutated dataset as described, we ensure that all correlations within this dataset are inconsistent with traffic flow theory. If the ensemble is still able to provide meaningful predictions despite these non-physical correlations, it suggests that the method is not resilient to such correlations. On the other hand, if the method fails to make accurate predictions, it indicates robustness against non-physical correlations.

In Figure A3, we present three examples of predictions obtained from this experiment. The output does not exhibit clear congestion patterns along the time axis, further emphasizing the lack of meaningful correlations. Additionally, Figure A4 illustrates how the estimated aleatoric uncertainty and RMSE (Root Mean Square Error) change with the prediction horizon. The curves are nearly horizontal, indicating consistently high values. When compared to Figure 6, we observe minimal differences between the two test sets in this counterfactual experiment.

These results demonstrate that the proposed method fails to capture any informative correlations in this mutated dataset. Therefore, we can conclude that the model design prevents the learning of non-physical relationships, reinforcing its robustness and adherence to capturing meaningful uncertainty.

Appendix 3. RMSE-Recall curve and calibration plotting

The RSME-recall curve shown in Figure 5(a) is obtained by the following steps:

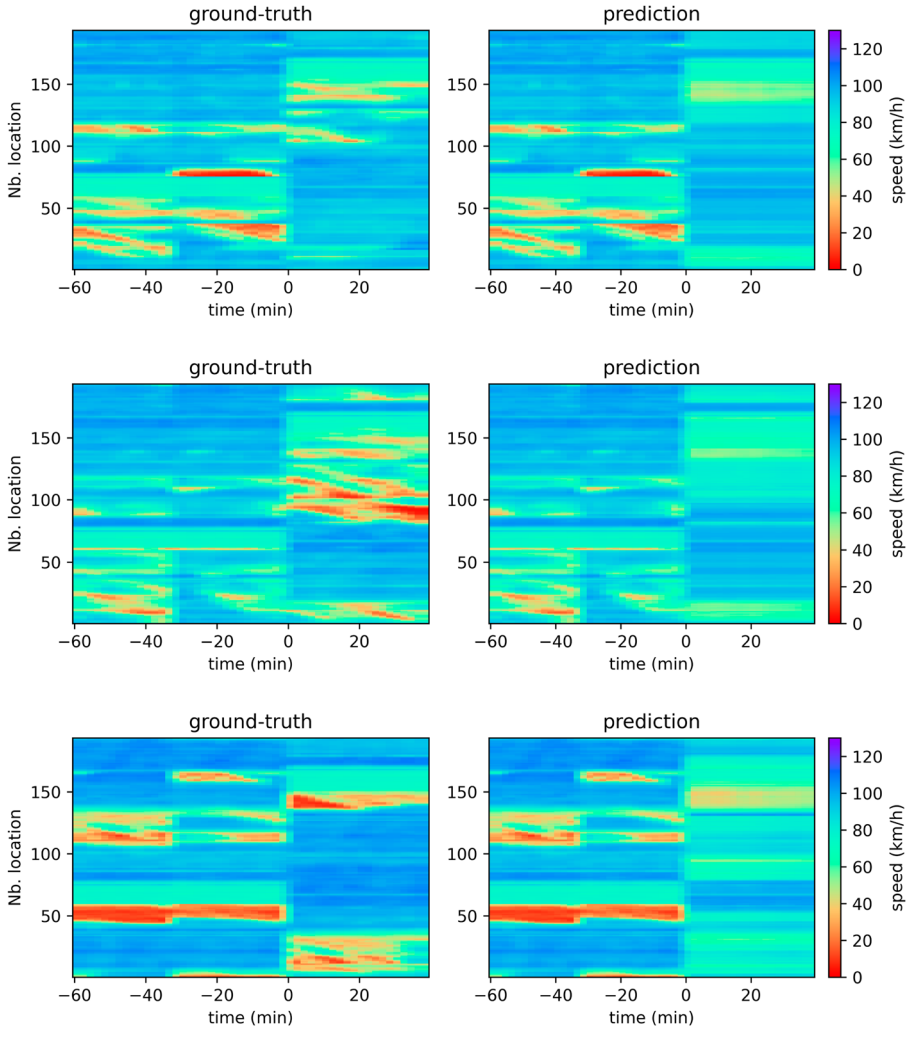

Figure A3. 3 examples of predicted traffic state. To show consistency, the prediction is refined to 2 m time interval by linear interpolation. The predictions are compared with the ground truth.

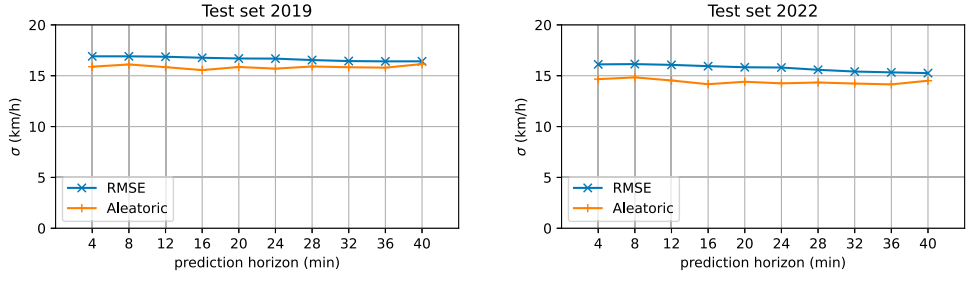

Figure A4. RMSE/aleatoric uncertainty and prediction horizon relationships on the two test sets.

(1) Sort the predictions according to the quantified aleatoric or epistemic uncertainty in descending order.
(2) Define a range of threshold values between 0 and 1. Here we use a uniform interval of 0.02.
(3) For each threshold value, compute the average RMSE of all predictions whose RMSE are below that threshold.
(4) Plot the RMSE and the confidence level values (Recall) on a 2D plot.
(5) Connect the plotted points to form the RMSE-Recall curve.

The calibration plot shown in Figure 5(b) is obtained by the following steps:

(1) Define a range of threshold values between 0 and 1. Here we use a uniform interval of 0.02.
(2) For each predicted posterior distribution and the corresponding ground truth, compute the percentile of the ground truth.
(3) For each threshold value (expected confidence), calculate the percentage of the percentiles below the threshold (observed frequency).
(4) Present the scatters on a 2D plot.
(5) Connect the plotted points to form the calibration plot.