Appendix of Paper 36
Towards Learning in Grey Spatiotemporal Systems: A Prophet to Non-consecutive Spatiotemporal Dynamics

We first thank the anonymous reviewers for their dedicated efforts in reviewing our work. Due to the limited space of the main body, we here include a five-page Appendix, to further clarify some preliminaries, discuss the proposed model and its interpretations, as well as present detailed experimental implementations and result visualizations. We hope these Appendix will facilitate the understanding of our work.

1 Preliminaries of uncertainty quantification

Uncertainty quantification plays a critical role in understanding grey spatiotemporal system. In engineering modeling, uncertainty can be categorized into epistemic and aleatoric [3].

Epistemic uncertainty, raised by unseen samples and fitting capacity, is a model property that can be viewed as the model parameter variation during training period. Such kind of epistemic uncertainty can be reduced by incorporating more prior information, e.g., increasing the diversity and quantity of training samples. By imposing a distribution over learned parameters, epistemic uncertainty quantification aims to estimate the parameters of this potential distribution over weights. However, directly estimating this distribution is intractable, thus general solutions either impose dropout during both training and testing periods, or insert Bayesian Neural Networks, to derive multiple outputs from specific samples, and then compute the statistical variance of these predictions. Let $\widehat{y}_t^2$ be t-th sampled predictions and we repeat the sampling with T times. The estimated epistemic uncertainty $\widehat{u}_{epis}$ can be formulated as,

$$\widehat{u}_{epis} = var(y) = \frac{1}{T}\sum_{t=1}^{T}\widehat{y}_t^2 - (\frac{1}{T}\sum_{t=1}^{T}\widehat{y}_t)^2 \tag{1.1}$$

To this end, the epistemic uncertainty $\widehat{u}_{epis}$ describes the potential parameter variation of the specific model with trained samples, where the model encapsulates the knowledge of samples [7, 8]. Therefore, given the model and tested samples, it is rational and feasible to measure to what degree the model confident with the derived results.

Aleatoric uncertainty, stemming from inherent noise and unobservable factors in datasets, is mostly heteroscedastic and cannot be eliminated. Aleatoric uncertainty characterizes the intrinsic difficulty of learning tasks. Existing methods often construct mappings from input data to aleatoric uncertainty and maintain the consistency between errors and learnable aleatoric uncertainty with well-designed negative log likelihood loss [6]. We denote D as the number of samples, and the loss function $Loss_{au}(\theta)$ for predictive aleatoric uncertainty is,

$$Loss_{au}(\theta) = \frac{1}{D}\sum_{i}\frac{1}{2}\frac{||y_i - \widehat{y}_i||^2}{\widehat{u}_{a,i}^2(x)} + \frac{1}{2}\log \widehat{u}_{a,i}(x)^2 \tag{1.2}$$

where $\widehat{u}_{a,i}^2(x)$ is a function of input data point x and represents the aleatoric uncertainty. Minimizing this loss is equivalent to constraining the consistency of changes between both prediction errors and aleatoric uncertainty. Such operation can discourage the model from predicting low uncertainty for points with high residual error, inherently alleviating the influences of outliers in the probabilistic interpretation [6].

In this work, we will leverage these basic theories to realize our DisEUQ for providing responsible predictions in FDG2S.

2 Model discussions and interpretations

Benefits and insights of FoDGSL. The core idea of FoDGSL is decoupling and learning to re-aggregate, and the insights are two-fold. First, the disentangled proximity explains the compositions of factor-wise effects in a complex system, which enables factor-level flexible rearrangements and combinations of various conditions on arbitrary future predictions. Second, the ingenious combinations of GNN and CRF empower our framework to leverage informative exogenous factors to compensate for missing observations. We provide insights into coupling traditional machine learning or energy functions in other fields (e.g., physics) with DNNs to cooperatively tackle the challenges that cannot be well-addressed by neural networks only, such as data limitations.

Benefits and insights of DisEUQ. Regarding the epistemic one, the benefits are two-fold, 1) high inference efficiency without additional training process, and 2) interpretation of sample-specific confidence regarding models, which enables to moderate sample distributions and improves generalization by excluding or augmenting samples with high epistemic uncertainty. The aleatoric uncertainty quantification explains the task-oriented learning difficulty and cooperatively boosts the model robustness in two aspects. i) spatiotemporal autoencoder naturally possesses the property of anti-noise. ii) the error-uncertainty consistency allows the network to adapt the residual's weighting, and thus suppress high uncertainty points, enabling our model to ignore potential outliers.

Efficiency issue. In our work, we turn the CRF estimation into computing the separated similarities organized by exogenous factors and re-aggregations. The solution contributes to the computation costs of $\mathcal{O}(M \cdot N^2)$ and $\mathcal{O}(M)$ where $M \ll N$. Fortunately, the former computation can be done only once during non-training phrase, and leads to very limited external workloads to training process. For UQ, our epistemic uncertainty is computed without any propagation process and achieves $\mathcal{O}(N)$ computation, while we limit the forward search times for constructing observation set of variances to $\pi_Q = 4$, which strives for a tradeoff between representativeness and efficiency. The whole complexity is generally identical to other GNNs and additional complexities are tolerable in our non-consecutive forecasting.

Limitations. First, even though our model can handle various patterns of sequence missing in non-consecutive forecasting, it still requires at least two complete periodicity-based observations, leading to data acquisition limitations. Second, we design our model based on two scenarios, however, for the arbitrary sequence missing patterns, it could be more efficient to build an adaptive model to accommodate different observation missing patterns without re-training. Fortunately, these two limitations can be potentially addressed by the emerging techniques of OOD learning and domain adaptations, which are left as our future works.

3 Additional experimental descriptions and results

In this section, we will provide more details regarding datasets, implementations, baselines and visualized uncertainty quantification results.

3.1 Dataset descriptions We collect three spatiotemporal datasets from different cities.

1) SIP: It is a private dataset that includes camera surveillance at interactions of Suzhou Industry Park (SIP);

2) NYC: Taxi trip records of New York City (NYC), which is available on NYC Government: https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page;

3) Metr-LA: highway loop detectors of Los Angeles, including traffic volumes, speeds and locations. It is available on https://github.com/liyaguang/DCRNN.

Table 1: Dataset statistics (m: million, k: thousand)

Dataset	Dataset Category	Records #	Time Span	Region #
SIP	Surveillance	2.7 m	01/01/2017-03/31/2017	108
	Weather	4.3k		
NYC	Taxi trips	7.5 m	01/01/2017-05/31/2017	354
	Weather	7.4k		
Metr-LA	Loop detectors	4.9 m	03/01/2012-06/30/2012	207
	Weather	5.7k		

The detailed descriptions and statistics of our datasets are figured in Table 1. Regarding exogenous context factors, the weather information is collected from an API: https://api.weather.com, and other factors including day of week, daily timestamps as well as holiday indicators are manually created from calendar.

3.2 Implementation details The implementation details are elaborated in four aspects, 1) how to organize exogenous factors, 2) the settings of non-consecutive forecasting, 3) how to estimate two types of uncertainty, and 4) evaluation metrics.

3.2.1 Organization of exogenous factors. We encode the exogenous contexts into fixed-length vectors and randomly initialize the location embeddings where the embeddings can be trainable during our end-to-end learning process. For fairness, we feed the same available exogenous factors into those baselines if they are with placeholders of context factors. Otherwise, we impose an additional fully-connected layer on exogenous factors and perform element-wise additions to mainstream outputs to incorporate exogeneous context factors.

3.2.2 Settings of non-consecutive predictions. We consider the grey spatiotemporal systems as non-consecutive predictions under two settings, which are illustrated in Figure 1. (a) Prediction for early planning, with one-day ahead and 1-week ahead. (b) Prediction under sensor failure interruption. To imitate scenario (b), we fix the nearest 3-day observations available, and respectively assume that the middle 3-day and 7-day observations are missing before the nearest available 3-day observations. They are both implemented by adjusting the spans and positions of unavailable observations. For fairness, we impose the same retrieval (semantic sampling) strategy as ours to enable their non-consecutive predictions unless specified, to evaluate the effectiveness of factor-decoupled learning. Regarding evaluations, we compare the errors of samples that have the same weather context with targets.

3.2.3 Estimation of uncertainty quantification.

First, the sample density prober produces multiple outputs for specific samples and further derives the variance of prediction results as the epistemic uncertainty. Second, the aleatoric uncertainty is quantified and integrated by identifying potential noise and factor-induced variations, and constrained with a consistency loss function. Third, only the aleatoric one is incorporated into interval-level evaluations, as the aleatoric is associated with the potential variations conditioned on factors, while the epistemic one is a relative value measuring model experiences and it can be evaluated by illustrating their variations during the training process.

3.2.4 Metrics

We respective introduce metrics for both spatiotemporal learning and uncertainty quantification.

Spatiotemporal learning: We employ **MAPE**, as it eliminates the influences of both magnitude orders across datasets and preprocessing strategies across baselines.

Uncertainty quantification: We exploit two metrics to jointly quantify the quality of uncertainty quantification.

(1) Prediction interval coverage probability (PICP), measures whether the predicted intervals maximally cover the ground-truth, where the higher the value is, the better performance [10]. It can be calculated by, $\mathrm{PICP}_{\mathrm{obj}} = \frac{1}{Nh}\sum_{t=1}^{h}\sum_{i=1}^{N} b^{i,t}$, where $b^{i,t} = 1$ only if the ground-truth is covered by the predicted uncertainty-incorporated intervals $[\widehat{y}_{i,t} - (\widehat{u}_a)_{i,t},\ \widehat{y}_{i,t} + (\widehat{u}_a)_{i,t}]$.

(2) Uncertainty percentage (UP), which measures whether the derived uncertainty is small enough to avoid infinitely increasing intervals. The lower this metric is, the better [13], i.e., $\mathrm{UP} = \frac{1}{Nh}\sum_{t=1}^{h}\sum_{i=1}^{N}\frac{(\widehat{u}_a)_{i,t}}{y_{i,t}}$. (3) Performance $\uparrow$: Error decreased ratios compared with best baseline. Note that only the aleatoric one is incorporated into interval-level evaluations, as the aleatoric is associated with the potential variations conditioned on factors, while the epistemic one is a relative value measuring model experiences and it can be evaluated by illustrating their variations during the training process.

3.3 Baselines

3.3.1 Spatiotemporal forecasting.

We carefully select a series of typical baselines that have the potential to perform non-consecutive predictions and simultaneously model exogenous context factors. For all baselines, we impose the same retrieval strategy of ours to enable their non-consecutive predictions unless specified, and feed the exogenous environmental factors into the model. Noted that as the selected methods have overwhelmingly beaten baselines in their corresponding papers, we will not repeatedly compare with the baselines appearing in their papers.

- **Traffic transformer:** An extension of Google's Transformer for traffic forecasting, to capture temporal continuity, periodicity and spatial dependency [2].
- **STFGNN:** A framework jointly learns localized heterogeneity and global homogeneity with data-driven graph generation [9].
- **STG2Seq:** A hierarchical convolutional framework to capture spatial and temporal dependencies of passenger demands with awareness of context factors [1].
- **MTGNN:** A state-of-the-art solution to multivariate time series predictions that can be learned without explicit graph structure [12].
- **ASTGNN:** A state-of-the-art fully-attention based solution to traffic forecasting, by considering contexts of observation themselves [5].
- **MTGNN-OSp:** We replace our sampling strategy with the originial sampling of MTGNN [12], to verify the superiority of semantic-neighboring sampling in non-consecutive predictions.

3.3.2 Uncertainty quantification task

We reproduce various UQ solutions into two representative forecasting frameworks, STG2Seq and our FDG2S, to illustrate the uncertainty quantification quality. The UQ baselines are as follows,

- **Dropout BNN:** We realize the Bayesian neural network with dropout [6], i.e., plugging dropout layers after graph convolutions in ST learning model.
- **DeepEnsembles:** It trains a series of neural networks with different initializations [8]. The number of ensembled networks is set as 5, according to corresponding literature [8].

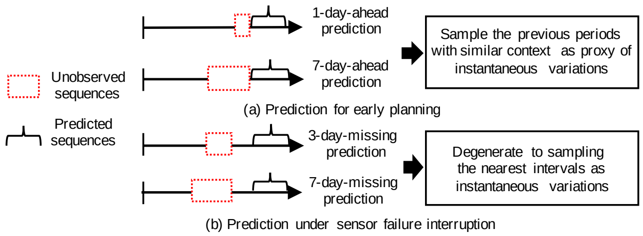

Figure 1: Illustration of non-consecutive prediction settings for grey spatiotemporal systems.

- **SDE:** It imitates Brownian motions and injects OOD samples [7] into the training process. We modify the loss functions and impose different perturbated samples to realize the SDE.
- **MIS:** It enables a scoring function that rewards narrower confidence or credible intervals and encourages intervals to include the targets. We implement it with two additional FC layers for upper and lower bound regressions, and combine MAPE with MIS score as an integrated loss [11].
- **STUaNet:** It is an uncertainty-aware spatiotemporal prediction model [13], and we improve it to multiple steps with LSTM.

3.4 Detailed analysis on two types of uncertainty quantification

3.4.1 Analysis of epistemic uncertainty. Here we perform a detailed evaluation on epistemic uncertainty. We present a series of graph-level epistemic uncertainty derived from sample density prober by increasing training epochs in Figure 2. Epistemic uncertainty results on three datasets vary consistently with errors and tend to decrease with fluctuations, verifying the rationality of our quantification solution and the inherent regularity of deep learning. These results can promote the interpretability for understanding the prediction reliability of each sample in grey systems, and help retrain or augment critical samples. In addition, we further demonstrate the consistency between the critical samples and their epistemic uncertainty. Thus, we filter 5%, 10% samples with the highest epistemic uncertainty and then re-train the model. The observed performances on SIP one-week predictions are 19.52%, 18.99%, leading promotions of 7% and 27%, which confirms that the epistemic uncertainty can exactly indicate the difficulty of model training for each sample.

3.4.2 Analysis of aleatoric uncertainty. The spatial maps of aleatoric uncertainty and uncertainty-aware prediction intervals are characterized in Figure 3(a) and (b). As illustrated, the aleatoric uncertainty is spatiotemporal heterogeneous, and the uncertainty maps on rainy Sunday are found with similar spatial patterns but both are different from those on fair-weather workdays, verifying the intuition of factor-varying uncertainty. In addition, our predictive intervals tend to be wider when errors become larger, and the widths can exactly capture ground-truth with predicted aleatoric intervals. As a result, these uncertainties can make great sense for urban perceptions to pre-arrange urban traffic controls and contingency plans during critical activities such as sport competitions and big concerts, avoiding unexpected city emergencies.

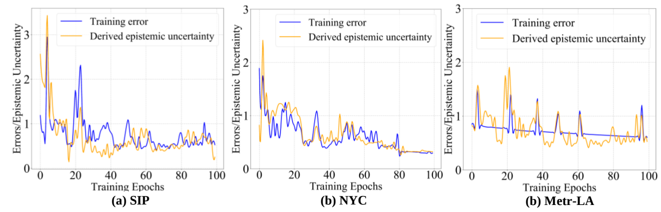

Figure 2: Epistemic uncertainty quantification along with training epochs on all datasets.

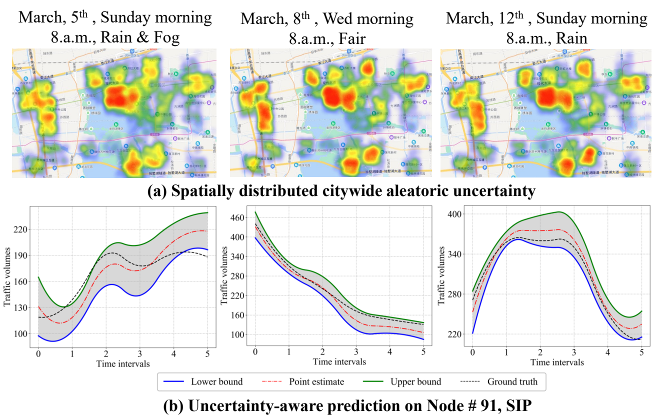

Figure 3: Detailed model analysis and case study

These promising results demonstrate the robustness and interpretability of our factor-decoupled solution towards learning grey spatiotemporal systems.

References

[1] Lei Bai, Lina Yao, Salil S Kanhere, Xianzhi Wang, and Quan Z Sheng. Stg2seq: spatial-temporal graph to sequence model for multi-step passenger demand forecasting. In *IJCAI*, pages 1981–1987, 2019.

[2] Ling Cai, Krzysztof Janowicz, Gengchen Mai, Bo Yan, and Rui Zhu. Traffic transformer: Capturing the

continuity and periodicity of time series for traffic forecasting. *Transactions in GIS*, 24(3):736–755, 2020.

[3] Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? does it matter? *Structural safety*, 31(2):105–112, 2009.

[4] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICLR*, pages 1050–1059, 2016.

[5] Shengnan Guo, Youfang Lin, Ning Feng, Chao Song, and Huaiyu Wan. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. In *AAAI*, volume 33, pages 922–929, 2019.

[6] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *NIPS*, pages 5574–5584, 2017.

[7] Lingkai Kong, Jimeng Sun, and Chao Zhang. Sde-net: Equipping deep neural networks with uncertainty estimates. 2020.

[8] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NIPS*, pages 6402–6413, 2017.

[9] Mengzhang Li and Zhanxing Zhu. Spatial-temporal fusion graph neural networks for traffic flow forecasting. In *AAAI*, volume 35, pages 4189–4196, 2021.

[10] Bin Wang, Jie Lu, Zheng Yan, Huaishao Luo, Tianrui Li, Yu Zheng, and Guangquan Zhang. Deep uncertainty quantification: A machine learning approach for weather forecasting. In *KDD*, pages 2087–2095, 2019.

[11] Dongxia Wu, Liyao Gao, Xinyue Xiong, and Matteo Chinazzi. Quantifying uncertainty in deep spatiotemporal forecasting. In *KDD*, 2021.

[12] Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, Xiaojun Chang, and Chengqi Zhang. Connecting the dots: Multivariate time series forecasting with graph neural networks. In *KDD*, pages 753–763, 2020.

[13] Zhengyang Zhou, Yang Wang, Xike Xie, Lei Qiao, and Yuantao Li. Stuanet: Understanding uncertainty in spatiotemporal collective human mobility. In *The Web Conference*, pages 1868–1879, 2021.