Towards Learning in Grey Spatiotemporal Systems: A Prophet to Non-consecutive Spatiotemporal Dynamics

Zhengyang Zhou[†] Kuo Yang[†] Wei Sun[†] Binwu Wang[†] Min Zhou [*] Yunan Zong [†]
Yang Wang[†] [‡]

Abstract

Spatiotemporal forecasting is an imperative topic in data science due to its critical applications in smart cities. Existing works mostly perform consecutive predictions of following steps with observations continuously obtained, where nearest observations can be exploited as the key knowledge for status estimation. However, the practical issues of early activity planning and sensor failures elicit a new task, non-consecutive forecasting. In this paper, we define spatiotemporal learning systems with missing observations as Grey Spatiotemporal Systems (G2S) and propose a Factor-Decoupled learning framework for G2S to hierarchically decouple multi-level factors, and enable flexible aggregations with uncertainty estimations. We especially select representative sequences to capture periodicity and instantaneous variations, and infer the non-consecutive future statuses under expected exogenous factors, compensating the missing observations. Given the inherent incompleteness and critical applications of G2S, a DisEntangled Uncertainty Quantification is put forward, to identify two types of uncertainty for model interpretations and robustness promotions. Experiments demonstrate that our solution can promote the performance by at least 8.50% on early planning and 2.01%-18.00% on sensor failures. The appendix of this paper can be found at https://github.com/zzyy0929/SDM-G2S.

1 Introduction

With the explosion of intelligent sensing devices, spatiotemporal learning, which supports various urban applications including intelligent transportation [7, 22], smart grid [11] and weather forecasting [17], has become a pivotal technique. Generally, traditional spatiotemporal forecasting mostly assumes that the information of urban systems is fully obtained, where the data integrity is an essential condition of their success.

However, in real scenarios, urban systems to us are not black or white, which means precise information is totally missed or fully obtained. Instead, urban systems are usually grey with incomplete information, and this can be an obstacle of achieving accurate and robust smart city scheduling. Specifically, there are two possible scenarios as illustrated in Figure 1: i) The early planning of both individual urban trips for vital events and citywide traffic scheduling for important urban activities, takes the urban perceptions of some days or even weeks in the future as a prerequisite, where the nearest observations are inherently unavailable. ii) With the expanding deployment of urban sensors, the probability of sensor failures increases and brings larger gaps to urban sensing datasets in temporal perspective. These two scenarios both point to a new unresolved issue, spatiotemporal forecasting with unobserved sequential information, i.e., non-consecutive spatiotemporal prediction. We define urban spatiotemporal systems with fragments of observation missing as Grey Spatiotemporal System (G2S). Considering the inherent property of data incompleteness in G2S, a key issue is effectively advancing next-step prediction to non-consecutive predictions with limited observations.

Given the conflict between the data integrity assumption and the real-world incomplete continuous data, the main **challenges** of learning grey spatiotemporal systems can be summarized as two aspects.

(1) Unavailability of nearest historical observations. Existing prediction methods in white urban systems involve the observations of nearest steps as features for training, hence accurately capturing the status evolutions towards near future [1, 22]. In these traditional solutions, the nearest statuses play a significant role in forecasting as they provide key knowledge to support status estimations on following consecutive steps. Even for those sparse spatiotemporal learning efforts where sensors are sparsely deployed, researchers take the status of spatially neighboring sensors as proxy or generate verisimilar real-time data by training discriminators [18]. Unfortunately, these sparse learning methods are inherently an interpolation in temporal and spatial domains, and are incapable of dealing with the long-

[*] M Zhou is with Huawei Technologies, Shenzhen, China.

[†] Z Zhou (zzy0929@mail.ustc.edu.cn), K Yang, W Sun, B Wang, Y Zong and Y Wang are with University of Science and Technology of China, Hefei, China.

[‡] Corresponding author: Yang Wang, angyan@ustc.edu.cn

term sequential observation missing issue.

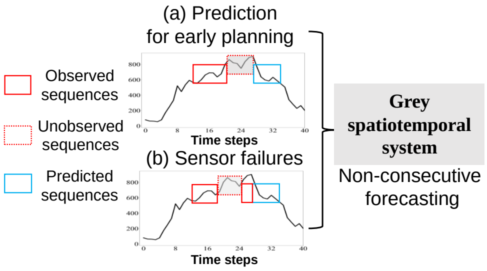


Figure 1: Illustration of grey spatiotemporal system and two typical scenarios of non-consecutive predictions

(2) Uncertainty quantification and disentanglement. Considering the pre-arrangements of crucial urban activities and events that served by G2S, the intrinsic data incompleteness deliver the uncertainty quantification (UQ) to be an indispensable issue for grey system learning. Also, literature has articulated that uncertainty in a learning system can be decomposed into epistemic and aleatoric. In particular, the epistemic one can identify out-of-distribution (OOD) samples and be reciprocal for model generalization [9], while the aleatoric one can aware the inherent challenge of tasks and alleviate the influences of outliers by regularization [8]. Unfortunately, pioneering spatiotemporal learning with uncertainty quantification either take uncertainty as a jointed value [14], or not tailored for spatiotemporal learning tasks [8, 9, 19]. Therefore, how to provide responsible predictions with effective and disentangled uncertainty measurement is another challenge in understanding grey systems.

Present works. We perform responsible learning in grey spatiotemporal systems by considering two non-consecutive forecasting settings, which conducts both point prediction and uncertainty estimation. To tackle above challenges, we propose a new ST-learning solution, Factor Decoupled Graph Learning framework for Grey Spatiotemporal System (FDG2S) by exploiting novel data organizations on environmental factors. FDG2S consists of two major components. To address the unavailability of nearest historical observations, we propose a Factor-Decoupled Graph Sequence Learning (FoDGSL) to progressively decouple multi-level factors and enable a dynamically learnable aggregation to vicariously estimate the statuses of future spatiotemporal elements. To disentangle the biases regarding two types of uncertainty and boost robustness of our G2S, a DisEntangled Uncertainty Quantification (DisEUQ) is proposed. It includes a post-explained sample density prober, which derives the epistemic uncertainty, and a weak-supervised aleatoric variation learner to approximate the aleatoric uncertainty and suppress the effects of outlier samples.

Main Contributions. Novel data organizations. To remedy lacking observations, a semantic-neighboring sequence sampling with factor-aware constraints is proposed to collect personalized sequences for pattern extractions. We also take exogenous factors as an intermediate proxy, and reorganize main observations by the types of exogenous factors, achieving pluggable factor-wise combinations for future target estimations. **FoDGSL** disentangles multi-level factors to enable dynamic aggregations. We design a sampling strategy to retrieve representative sequences, a factor-decoupled aggregation to disentangle complicated factor influences on aggregations and couple both endogenous and exogenous factors into a unified graph sequence learning architecture. **DisEUQ** identifies two sources of uncertainty. A post-explained sample density prober explores the epistemic uncertainty regarding learning experiences, while an intrinsic aleatoric variation learner quantifies aleatoric uncertainty and improve robustness.

2 Preliminaries

2.1 Notations and problem definitions

Definition 1. (**Urban regions and urban graph**) *The studied areas are discretized into N spatial regions $\mathcal{V} = \{v_i | i = 1, 2, ..., N\}$ by geographical divisions or natural observation stations, and the potential dependencies between two urban regions are denoted as the edge $\mathcal{E} = \{e_{ij} | 1 \leqslant i, j \leqslant N, \& \ i \neq j\}$. All regions and edges consist of an urban graph $G(\mathcal{V}, \mathcal{E})$.*

Definition 2. (**Endogenous spatiotemporal elements**) *Considering an interval set $\mathcal{T} = \{1, 2, 3..., T\}$, the endogenous spatiotemporal observations are defined as the task-specific primary elements, $\mathbf{X} = \mathbf{X}_{:,1}, \mathbf{X}_{:,2}, ..., \mathbf{X}_{:,T} \in \mathbb{R}^{N \times T}$. We also define T_d as the number of intervals in each day.*

Definition 3. (**Exogenous factors**) *Environmental factors that are not for predictions but beneficial to task optimization are defined as exogenous factors. Given M types of exogenous factors $\mathbb{C} = \{Cf_1, Cf_2, \cdots, Cf_M\}$. In this work, we consider the exogenous factors as Cf_d = day of week, Cf_s = daily timestamps, Cf_w = weather type, Cf_{nw} = numerical weather values, and Cf_l = location embedding, instantiating Cf_1 to Cf_5. Let $\mathbf{C}_t = \{\mathbf{c}_m(i, t)\}$ be the deterministic observations of exogenous factors Cf_m at region i during t, where d_m is the dimension of $\mathbf{c}_m(i, t)$.*

Definition 4. (**Grey spatiotemporal system**) *The combinations of data and algorithms focusing on*

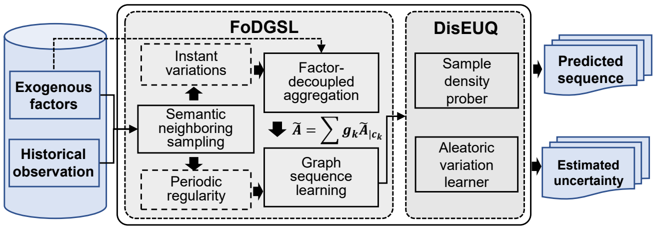


Figure 2: Framework overview of FDG2S

spatiotemporal learning tasks are defined as spatiotemporal systems. In this work, we define the spatiotemporal systems with fragment of observation missing as grey spatiotemporal systems, shown in Figure 1.

Problem 1. **(Non-consecutive forecasting in grey spatiotemporal system)** *Given incomplete endogenous spatiotemporal observations* $\mathbb{X} = \{\mathbf{X}_{:,1:s} \cup \mathbf{X}_{:,r:T}\}$ *and corresponding exogenous factors* $\mathbf{c}_{1:s}, \mathbf{c}_{r:T}$*, where* $\mathbf{X}_{:,s+1:r-1}$ *and* $\mathbf{c}_{s+1:r-1}$ *are the missing sequences with* $1 < s < r \leqslant T+1$*, we aim to design a learnable function* f_{ST} *to predict future h-step spatiotemporal elements* $\widehat{\mathbf{Y}}_t$*, and quantify both epistemic and aleatoric uncertainty* $((\widehat{\mathbf{u}}_e)_t, (\widehat{\mathbf{u}}_a)_t)$*, where* $t = \{t | T < t \leqslant T+h\}$*.*

3 Methodology

Our Factor Decoupled Graph Learning framework for Grey Spatiotemporal System (FDG2S) is illustrated in Figure 2, which consists of two major components, a Factor-Decoupled Graph Sequence Learning (FoDGSL) and a DisEntangled Uncertainty Quantification (DisEUQ).

3.1 FoDGSL for grey forecasting

3.1.1 Overview of FoDGSL. In spatiotemporal learning, the contributors of predicted targets are multi-level and can be decomposed into endogenous historical observations and exogenous environmental compositions. Further, historical observations can be distinguished by periodical regularity and instantaneous variations, while exogenous environmental factors usually consist of weather, time and locations, which are prone with complex interaction effects. Given the unavailability of partial observations in grey spatiotemporal systems, the key is to infer potential statuses under expected conditions. Hence, we propose an FoDGSL, which disentangles multi-level factors and exploits exogenous factors as intermediate proxy to enable flexible and dynamic aggregations for approaching targets. Our FoDGSL is unified with three sub-modules, i.e., semantic-neighboring sequence sampling, factor-decoupled aggregation, and a graph sequence learning.

3.1.2 Semantic-neighboring sequence sampling. To estimate unseen futures, we first exploit two compositions of both periodicity regularity and instantaneous variations as the sequences that are semantic-neighboring to targets [12, 22], and then select the representative periods on both aspects for pattern extractions. To simplify notations, we let $c(Cf_*|t)$ denote the values of exogenous factor Cf_* at step t, and organize each sequence by h steps for sequence learning. For periodic ones, we directly retrieve the largest k_p away from T that satisfies following principles,

$$\begin{cases} k_p = T - k * 7 * T_d, k \in \mathbb{N}^+ \\ c(Cf_s|k_p) = c(Cf_s|T) \\ \text{IsNull}(\mathbf{X}_{k_p-h:h}) = \text{False} \end{cases} \tag{3.1}$$

We can obtain the periodical sequence $\mathbf{X}_P = \mathbf{X}_{k_p-h:h}$, where IsNull$(\cdot)$ examines whether all elements are not null. Actually, we retrieve the intervals that satisfy three constraints, i.e., weekly periodical with the same day of week, same index of daily intervals, and observation availability. We keep the retrieved daily interval k_p be the same as T is with periodicity T_d, and organize h continuous steps to learn sequential evolution patterns from nearest to future ones.

Second, we utilize exogenous factors to construct the proxy of instantaneous variations. Modifications to the former one are that 1) adding the expected weather type as an additional retrieval index to find more similar contexts in historical observations, 2) removing the constraints of weekly periodicity to involve more recent available observations, but relaxing the same day of week/daily intervals constraints by tolerance of 1 day and $\widetilde{\varepsilon}$ intervals[1]. Thus, principles for retrieving instantaneous variations starting at k_h are obtained,

$$\begin{cases} ||c(Cf_d|k_h) - c(Cf_d|T)|| \leq 1 \\ ||c(Cf_s|k_h) - c(Cf_s|T)|| < \widetilde{\varepsilon} \\ c(Cf_w|k_h) = c(Cf_w|T) \\ \text{IsNull}(\mathbf{X}_{k_h-h:h}) = \text{False} \end{cases} \tag{3.2}$$

Then the proxy of instantaneous variations is achieved by $\mathbf{X}_H = \mathbf{X}_{k_h-h:h}$. So far, the input main observations is Concat$[\mathbf{X}_P; \mathbf{X}_H]$, which adaptively captures personalized semantic-neighboring sequences to support non-consecutive forecasting. This strategy will degenerate to the nearest consecutive sequences sampling when recent observations are available.

3.1.3 Factor-decoupled aggregation. Since exogenous factors influence instantaneous variations and

[1] The allowed maximum interval shifts for tolerance $\widetilde{\varepsilon}$ is 3.

thus aggregations, we are expected to disentangle both historical observation contributors and exogenous factor influences, and then re-aggregate factor-induced proximity to achieve learnable hybrid adjacencies.

Analysis on factor decoupling. In fact, the spatiotemporal elements are regular with their tidal patterns while it also reveals heterogeneous region-wise proximity patterns with interactions between time stamps and weather contexts [1, 12]. Therefore, the instantaneous variations influenced by exogenous factors can be further decoupled into a basic intensity and the factor-induced region-wise proximity. We resort our instantaneous aggregations to conditional random field (CRF). CRF is capable of capturing mappings between observable factors and targets as well as pairwise dependencies among reference data points [5]. Accordingly, we respectively take the exogenous factors as observable variables and predicted spatiotemporal observations as targeted variables. Then we can couple the CRF's energy functions, i.e., status feature function and transition feature function with graph representation learning, to realize factor-decoupled aggregations.

CRF-based factor-decouple modeling. Let $\mathbf{c}(i)$ and $\mathcal{N}_i$ denote the combined exogenous vector and the set of neighboring nodes of v_i. We can decouple the targeted y_i into a factor-induced basic intensity $\varphi_u(y_i, \mathbf{c}(i))$ and factor-specific influence $\varphi_p(y_i, y_j|\mathbf{c}(i))$,

$$E(y_i|\mathbf{c}(i)) = \varphi_u(y_i, \mathbf{c}(i)) + \sum_{j\in\mathcal{N}_i} \varphi_p(y_i, y_j|\mathbf{c}(i)) \tag{3.3}$$

φ_u is the status feature function capturing the direct basic intensity from exogenous factors to node values, and φ_p represents transition feature function modeling pairwise correlations conditioned on exogenous factors.

Status feature function for learning factor-induced intensity. Given H_i indicating the representation of i-th node, we construct the contributor of combined factors to spatiotemporal observations by a status feature function. In particular, φ_u learns to map the intensity by minimizing the following energy function,

$$\varphi_u(\underset{\min}{y_i}, \mathbf{c}(i)) = ||H_i - B_i(\mathbf{c}(i); w_B)||_2^2 \tag{3.4}$$

where $B_i(\mathbf{c}(i); w_B)$ is realized with a two-layer fully-connected network based on learnable parameters w_B.

Transition feature function for capturing node-wise structural correlations. Recall that transition feature functions in CRF are capable of exploiting observable factors to quantify the node-wise proximity. We take this pairwise transition feature function for generating the region-wise proximity conditioned on exogenous factors. Specifically, we take sequence-level similarity of main observations as the region-wise proximity, where similar sequences tend to be aggregated to benefit accurate predictions. To disentangle the joint influences of multiple exogenous factors on region-wise correlations, inspired by the mean-field theory on decompositions [2], we design an interpretable influence decoupling mechanism. Concretely, we treat the hybrid factor influences on region-wise correlations as the linear combinations of single factor-induced region-wise similarity with an additional interactive factor. Hence, let c_k denote the k-th exogenous factor, our factor-decoupled pair-wise correlations can be written by,

$$\varphi_p(y_i, y_j|c_1, c_2..., c_M) = \sum_{k=1}^{M+1} g_k sim(H_{i|c_k}, H_{j|c_k}) \tag{3.5}$$

where the transition feature function φ_p describes the joint region-wise similarity conditioned on combined exogenous factor $\mathbf{c}$, $sim(p, q)$ measures the similarity between elements p and q, and $g_k(\cdot)$ weighs the importance of each factor-induced similarity. The $M+1$ category of similarity denotes the interactions of all exogenous factors. To achieve single factor-induced proximity, we reorganize historical observations by individual exogenous factors and take sequence-level similarity to surrogate node-wise similarity. Specifically, we compute node-wise sequence-level similarities conditioned on each exogenous factor, and average their sequence similarities,

$$sim(H_{i|c_k}, H_{j|c_k}) = \frac{1}{|\mathcal{T}_{c_k}|} \sum_{ts\in\mathcal{T}_{c_k}} sim(H_{i,ts|c_k}, H_{j,ts|c_k}) \tag{3.6}$$

where $\mathcal{T}_{c_k}$ refers to the set of timestamps satisfying that the exogenous factor category of statuses are c_k. sim can be instantiated as the average of Euclidean distance and Cosine similarity to preserve similarity on both numerical intensity and directional trend. Based on above, we can further derive the interactions of all exogenous factors by the continued product,

$$sim(H_{i|c_{M+1}}, H_{j|c_{M+1}}) = \prod_{k=1}^{M} sim(H_{i|c_k}, H_{j|c_k}) \tag{3.7}$$

Learning to re-aggregate. To accommodate various interaction effects induced by different factors, we learn to re-aggregate these influences via a learnable vector $\overrightarrow{g} = [g_1, g_2, ..., g_{M+1}] \in \mathbb{R}^{M+1}$, enabling single context influence to dynamically change with their varying combinations. Denoting the single context c_k-induced region-wise similarity $sim(H_{i|c_k}, H_{j|c_k})$ as $\widetilde{A}_{|c_k}$, the objective of energy function φ_p is equivalent to optimizing the maximum conditional probability of region-wise adjacencies, given available exogenous factors and historical observations. Thus, we can obtain the factor-decoupled adjacent matrix for message propagation as,

$$\widetilde{A}_{|\{c_1,\ldots,c_M\}} = \sum_{k=1}^{M} g_k \widetilde{A}_{|c_k} + g_{M+1} \prod_{k=1}^{M} \widetilde{A}_{|c_k} \tag{3.8}$$

Then g_k can be computed by imposing a series of learnable parameters S_k,

$$g_k = \frac{\exp({S_k}^T c_k)}{\sum\limits_{k=1}^{M+1} \exp({S_k}^T c_k)} \tag{3.9}$$

Here, c_{M+1} is the factor-wise interaction combined by $c_{M+1} = \underset{k}{\mathrm{Concat}}\{c_k\}(1 \leqslant k \leqslant M)$.

Our Factor-Decoupled Aggregation can be viewed as introducing two new additional objectives into the end-to-end graph representation learning by borrowing the idea of CRF. In particular, the factor-decoupled pair-wise correlation matrix $\widetilde{A}_{|\{c_1,\ldots,c_M\}}$ serves as the modified adjacency and the factor-induced intensity energy function $\varphi_u(y_i, \mathbf{c}(i))$ is leveraged to enhance context factor-related representations. The retrieved sequence periods $\boldsymbol{X}_H$ are expected to feed into our graph learning for information propagation, and we formulate l-th layer of message propagation by GNN,

(3.10)
$$H_i^{(l)} = \alpha B_i^{(l-1)}(\mathbf{c}(i)) + (1-\alpha) \sum_{j \in N_i} \widetilde{A}_{(i,j)|\{c_1,\ldots,c_M\}} \boldsymbol{X}_H(i) \omega_{G_i}^{(l-1)}$$

where $\omega_{G_i}^{(l)}$ are a series of layer-wise parameters for GNN aggregation, and α adjusts the importance between two contributors. We stack two GNN layers and obtain the h-step representation for the proxy of aggregated instantaneous variations $\mathbf{H}^{ins} \in \mathbb{R}^{N \times h}$.

3.1.4 Graph sequence learning.

Finally, we cascade LSTM layers with Factor-decoupled aggregation, to capture sequential patterns. We concatenate the $h-step$ periodicity sequence $\mathbf{X}_P$ and the after-aggregated instantaneous variations $\mathbf{H}^{ins}$ of h steps into an image-like sequence, and feed them into an LSTM,

$$\widehat{\mathbf{Y}}_{:,T+t} = \mathrm{LSTM}((\mathbf{X}_P, \mathbf{H}^{ins}), \mathbf{W}_{lstm}) \tag{3.11}$$

where the LSTM takes a $2h-$step sequence as inputs and outputs an $h-$step predictions for non-consecutive forecasting in G2S. $\mathbf{W}_{lstm}$ are learnable parameters. After that, we can obtain our f_{ST}.

3.2 Disentangled uncertainty quantification

3.2.1 Overview of DisEUQ.

Given the distinctive roles and benefits of these two types of uncertainty, our DisEUQ consists of a post-explained sample density prober for exploring sample-specific epistemic uncertainty regarding knowledge learned from training samples, and an intrinsic aleatoric variation learner to quantify aleatoric one for suppressing outlier effects.

3.2.2 Post-explained sample density prober.

We argue that the epistemic uncertainty of a specific sample interprets the degree of the knowledge learned from similar samples by the model, and such quantification is equivalent to quantifying the density of samples that are similar with the tested one in the training set. In this way, we design a sample density prober to characterize it. Given the learned model f_{ST} and a specific tested sample X_0, we realize the sample density prober with a corruption-computation strategy. In particular, we first impose J times of corruptions on X_0 to generate the corrupted sample set with different perturbation coefficient $\varepsilon_j \ll X_j$ [2],

$$\widetilde{\mathbb{X}} = \{\widetilde{X_j} | \widetilde{X_j} = X_0 + \varepsilon_j, j = 1, 2, ...J\} \tag{3.12}$$

Then we derive the corrupted prediction set through f_{ST}, i.e., $\{\widetilde{Y_j}\} = f_{ST}(\{\widetilde{X_j}\})$. Intuitively, for a tested sample, the sparser of its similar training samples indicate fewer learning experiences of the model, where the fewer experiences can be reflected by greater variations of predictions even with small corruptions. Then, we can compute the variance of the corrupted prediction results as the epistemic uncertainty u_e,

$$u_e = \mathbb{E}(\{\widetilde{Y_j^2}\}) - \mathbb{E}^2(\{\widetilde{Y_j}\}). \tag{3.13}$$

Noted that $\mathbb{E}(\cdot)$ is the statistical expectation.

3.2.3 Aleatoric variation learner.

Aleatoric uncertainty can be explained by noise in input observations and interventions of unobservable factors where both of them are lacking explicit supervisions. To tackle this challenge, we propose an aleatoric variation learner, to estimate the noise and factor-induced variations.

Self-supervised noise detection. Considering the noise component, the observation X_0 can be decomposed as $X_0 = \widehat{X}_0 + n_0$, where $\widehat{X}_0$ is the ground-truth value, and n_0 is the noise we are expected to quantify. Inspired by the autoencoder-based solutions to video anomaly detection [20], we design a decoder connected with the factor-decoupled aggregations to reconstruct the observations of instantaneous variations $\mathbf{X}_H$. Once the general regularity is learned by the network, the reconstruction error can indicate the potential noise in

[2] As epistemic uncertainty is a relative value measuring model experiences during training process, the absolute value of ε_j is orthogonal to our results. We fix $J = 10$ to test the sensitivity.

inputs. We formulate the reconstruction loss as partial aleatoric uncertainty by,

$$\widehat{u_{as}} = L_{rec} = ||\mathbf{X}_H - \text{Recon}(\mathbf{X}_H; \mathbf{W}_{rec})||_2^2 \tag{3.14}$$

where Recon is a learnable function parameterized by $\mathbf{W}_{rec}$, and it can be instantiated as an LSTM.

Weakly supervised exogenous variation learning. As the uncertainty induced by unobservable factors can be reflected by different exogenous factors, we propose a factor-induced variation indicator, which summarizes variations among similar contexts and serves as a weak-supervised pseudo label of factor-induced variations. Concretely, by instantiating d_u, s_k, w_j as three exogenous factors regarding u-th day of week, k-th day timestamp and j-th weather type, the retrieved exogenous factor combinations for variation computation is $CF(d_u, s_k, w_j) = \{Cf_1 = d_u, Cf_2 = s_k, Cf_3 = w_j\}$. The context-similar observation set is constructed by finding a series of intervals $\mathbb{Q} = \{t_q\}$ that the statuses of these intervals share similar combined exogenous factors as (d_u, s_k, w_j). For location v_i and step t, we limit the maximum cardinality of $\mathbb{Q}$ to $\pi_Q = 3$ to avoid high computing costs, then the constructed set will be,

$$\begin{aligned} D(v_i, t)_{|d_u, s_k, w_j} = \{X_{i,t_q} | c_1(i, t_q) = d_u,\ c_2(i, t_q) = s_k, \\ c_3(i, t_q) = w_j, t_q \in \mathbb{Q}\} \end{aligned} \tag{3.15}$$

Thus, we can derive the weakly supervised variation specified by the factor combinations, through computing the standard variance (std) of set $D(v_i, t)$,

$$(u_{av})_{i,t|d_u,s_k,w_j} = \text{std}(D(v_i, t)) \tag{3.16}$$

After that, we can predict the variation by formulating the function of combined factors parameterized by ω_{av},

$$(\widehat{u}_{av})_{i,t} = \text{ReLU}(\omega_{av} * \underset{m \in \{1,2,\ldots,M\}}{\text{Concat}} \{c_m(i, t)\}) \tag{3.17}$$

The guidance for exogenous variance learning can be realized by minimizing the difference between the predicted $\widehat{u}_{av}$ and u_{av} as,

$$L_{av} = (u_{av} - \widehat{u}_{av})^2 \tag{3.18}$$

Hence, the above strategies enable our framework to learn the mappings from exogenous factors to potential factor-specific variations. We can finally obtain the learned aleatoric uncertainty from two perspectives by $\widehat{u}_a = \widehat{u}_{as} + \widehat{u}_{av}$. To strive a trade-off between the weak supervision indicator and the factual property of uncertainty, we further insert an uncertainty-error consistency constraint $L_{cons} = \frac{(y_i - \widehat{y}_i)^2}{((\widehat{u}_{as})_{i,t} + (\widehat{u}_{av})_{i,t})^2}$ as the third term. For one node v_i, the total aleatoric uncertainty learning is optimized by,

$$L_{Ale}(X_i, y_i, \widehat{y}_i) = \gamma_1 L_{rec}(v_i) + \gamma_2 L_{av}(v_i) + \gamma_3 L_{cons} \tag{3.19}$$

$\gamma_k (k = 1, 2, 3)$ are parameters balancing three losses. This joint constraint plays the role of preventing the uncertainty from unlimited increases and stabilizing the outlier effects, which can be viewed as the feedback-based optimization of uncertainty quantification.

3.3 Optimization

The main objectives are three-fold, main loss for spatiotemporal forecasting, minimization of energy function $\varphi_u(y_i, \mathbf{c}(i))$ and aleatoric uncertainty loss for potential variation learning, i.e.,

$$Loss(X_i, y_i, \widehat{y}_i; t) = \text{MAPE}(y_i, \widehat{y}_i; t) + \varphi_u(y_i, \mathbf{c}(i)) + L_{Ale}(v_i, t) \tag{3.20}$$

where $t \in [T+1, T+h]$. For UQ, we have $(\widehat{u}_e)_i$ for epistemic and $(\widehat{u}_a)_i$ for aleatoric. During training process, to alleviate the intractable optimization of task-wise weights, we adopt an adaptive weighting strategy by computing the ratios of main loss (MAPE) to other auxiliary losses, e.g., we initialize $\gamma_1 = \text{MAPE}(y_i, \widehat{y}_i; t)/L_{rec}(v_i)$ in the first batch, and dynamically tune the weights according to above ratios in each following batch.

4 Experiments

4.1 Dataset descriptions

We collect three spatiotemporal datasets from different cities, including Suzhou Industry (SIP), taxi trip records of NYC, and highway loop detectors of Los Angeles (Metr-LA). We retrieve the weather information from APIs and prepare other factors including day of week, daily timestamps as well as holiday indicators from calendar. Detailed dataset descriptions are figured in Appendix 3.1.

4.2 Implementation details

The dataset is divided by 60%, 10% and 30% for training, validation and testing. We use a fixed 30-min time interval for each dataset, and organize samples to periods consisting of h=6 intervals. During training, we sample two semantic-neighboring sequences as input features. Regarding exogenous context factors, we sample the expected combined exogenous factors to feed into our framework. We consider the grey spatiotemporal systems with two non-consecutive prediction settings in Table 1, i.e., (a) p-day/week-ahead prediction for early planning and (b) m-day-missing prediction under sensor failures. Detailed implementations of exogenous factors, non-consecutive predictions, and UQ can be found in the Appendix 3.2.

4.3 Metrics. **MAPE** for spatiotemporal learning, which eliminates the influences of both magnitude orders across datasets and preprocessing across baselines.

PICP [17] and UP [23] for UQ to jointly measure the quantification quality.

4.4 Baselines. Spatiotemporal forecasting: (1) Traffic transformer [3], **(2) STFGNN** [12], **(3) STG2Seq** [1], **(4) MTGNN** [21], **(5) ASTGNN** [7], **(6) MTGNN-OSp:** Replace our sampling strategy with the originial sampling. **Uncertainty quantification:** We plug various UQ solutions into STG2Seq and our FDG2S, to illustrate the UQ quality. **(1) Dropout BNN** [8], **(2) DeepEnsembles** [10] [3], **(3) SDE** [9], **(4) MIS** [19], **(5) STUaNet** [23].

4.5 Experimental results

4.5.1 Results of spatiotemporal forecasting. The comparison performances on grey spatiotemporal systems are illustrated in Table 1. Our work consistently outperforms baseline methods under two nonconsecutive settings. We have the following observations. 1) Setting (b) achieves better overall performances than setting (a), and performances on 3-day missing beat those of 7-day missing due to the availability of nearest observations, verifying the vital role of nearest observations. 2) The degraded performances of MTGNN-OSp (by at least 6.73% on 1-week-ahead predictions) indicate the superiority of our designed semantic sampling strategy. 3) For detailed comparisons, our solution outperforms the best baselines by 8.50%∼20.76% on early planning setting, and 2.01%∼18.00% on sensor failure setting across three datasets. Traffic transformer and STFGNN achieve barely satisfactory performances on 3-day missing predictions as they are dedicated to traffic forecasting and 3-day-m is the most similar task to consecutive forecasting. STG2Seq and MTGNN perform better on nonconsecutive predictions as they are inherently designed to involve context vectors. Finally, we can summarize our FDG2S superiority as 1) sufficient exogenous factors to model basic intensity and aggregations, and 2) robustness brought by aleatoric uncertainty learning.

4.5.2 Results of uncertainty quantification. Quantitative uncertainty learning is shown in Table 2. We observe that our DisEUQ achieves satisfactory quality of interval predictions as it outperforms two state-of-the-art baselines on PICP and obtains comparable UP metric on all datasets. Specifically, dropout-based methods reveal narrower uncertainty intervals but cannot exactly capture the ground-truth in such intervals, while SDE-based solution achieves a relatively better trade-off. MIS reasonably captures the ground-truth with its natural interval-aware objectives but fails to restrict the intervals. In contrast, our DisEUQ, inheriting the advantages of both slight perturbation in SDE and interval objective in MIS, can concurrently capture the potential intervals around ground-truths, and prevent the unlimited growth of uncertainty intervals.

4.5.3 Ablation study. Ablation studies are conducted on one-week-ahead predictions, and we name the ablative variants below. **(1) w/o SF**: remove Status Feature function ϕ_u, **(2) w/o TF**: replace the learnable factor-decoupled aggregation, i.e., Transition Feature function ϕ_p with a static distance-based adjacent matrix, **(3) w/o STA**: remove the spatiotemporal autoencoder for noise detection, **(4) w/o Var**: remove the spatiotemporal variance of exogenous factor-induced variation. Table 3 demonstrates the results. Empirically, the performances of 1-week-ahead predictions deteriorate when two energy functions and Autoencoders are removed. Particularly, the factor-decoupled aggregation plays a most important role with performance drops of 15.76%, 29.16% and 15.02% on three datasets. Further, without ST Variance guidance, the performances can be slightly improved, because the introduced uncertainty objective may distract the main objective.

4.6 Detailed model analysis

4.6.1 Prediction stability on farther horizons. We extend the prediction horizons to next 10 days, 14 days, 20 days on three datasets and compare their performance with STG2Seq and MTGNN in Figure 3. Our FDG2S not only reveals lower errors when predicted horizons become longer, but also performs more stable than others. Such stable behaviors show the effectiveness of our well-designed modules for tackling grey system incompleteness. Also interestingly, relatively better performances on 7 and 14 days are achieved, probably because of the weekly periodical assistances, which also verifies the intuition of our periodical sampling.

4.6.2 Hyperparameter settings. The main hyperparameters are three-fold, α concerning two target contributors of CRF energy functions, the dimensions of learnable aggregation kernels and LSTM hidden dimension. We run the hyperparameter searching on one-week predictions to achieve the best performance on each dataset and describe the process in Figure 4. Since the energy function weights and aggregation kernel dimen-

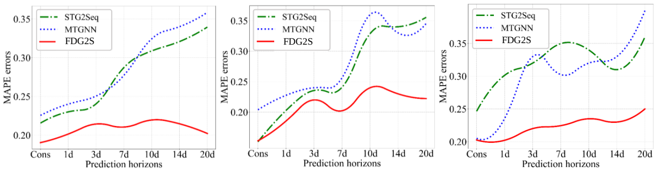
[3] The number of ensembled networks is set as 5.

Table 1: Performances on different settings of grey spatiotemporal systems

	SIP				NYC				Metr-LA			
	1-day-a	1-week-a	3-day-m	7-day-m	1-day-a	1-week-a	3-day-m	7-day-m	1-day-a	1-week-a	3-day-m	7-day-m
Transformer	0.257	0.301	0.245	0.250	0.554	0.542	0.457	0.475	0.245	0.343	0.227	0.326
STFGNN	0.258	0.309	0.238	0.263	0.223	0.248	0.231	0.252	0.266	0.365	0.245	0.313
STG2Seq	0.230	0.287	0.210	0.236	0.203	0.241	0.199	0.203	0.301	0.350	0.289	0.321
MTGNN	0.245	0.273	0.225	0.240	0.227	0.256	0.213	0.220	0.236	0.302	0.224	0.310
ASTGNN	0.250	0.279	0.231	0.264	0.235	0.265	0.225	0.235	0.243	0.285	0.230	0.245
MTGNN-OSp	0.280	0.292	0.244	0.230	0.252	0.273	0.202	0.212	0.246	0.331	0.246	0.273
Ours	**0.201**	**0.216**	**0.173**	**0.195**	**0.185**	**0.202**	**0.184**	**0.198**	**0.202**	**0.224**	**0.211**	**0.240**
Performance ↑	12.42%	20.76%	18.00%	15.20%	8.50%	16.02%	7.20%	2.31%	14.18%	21.22%	14.10%	2.01%

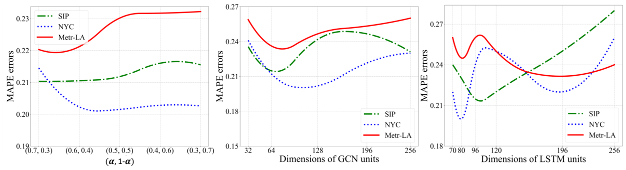
Table 2: Uncertainty quantification comparisons on three datasets

	SIP (PICP ↑, UP ↓)		NYC (PICP ↑, UP ↓)		Metr-LA (PICP ↑, UP ↓)	
	STG2Seq	Our Net	STG2Seq	Our Net	STG2Seq	Our Net
Dropout BNN	(0.606,0.368)	(0.618,**0.419**)	(0.549,0.402)	(0.744,0.460)	(0.502,0.335)	(0.754,0.394)
DeepEnsembles	(0.582,2.580)	(0.697,1.381)	(0.524,2.310)	(0.742,2.710)	(0.628,1.522)	(**0.876**,1.937)
SDE	(0.605,**0.257**)	(0.609,0.587)	(0.615,**0.233**)	(0.679,0.588)	(0.677,**0.235**)	(0.791,0.495)
MIS	(0.652,0.890)	(0.640,1.200)	(**0.705**,0.965)	(0.714,1.050)	(0.653,0.645)	(0.720,0.595)
STUaNet (Multi-step)	(**0.657**,0.356)	(0.659,0.450)	(0.695,0.365)	(0.723,0.468)	(0.684,0.425)	(0.708,0.423)
Our UQ	(0.627,0.386)	(**0.703**,0.494)	(0.507,0.325)	(**0.754**,**0.450**)	(**0.688**,0.402)	(0.766,**0.379**)



(a) SIP (b) NYC (c) Metr-LA

Figure 3: Longer predictions comparisons. 'Cons' refers to predictions on the next consecutive 6 steps.



(a) Energy function weights (α, 1-α) (b) Dimensions of GCN units (c) Dimensions of LSTM units

Figure 4: Hyperparameter studies on all datasets

sions have few influences on performances, we respectively set α to 0.5 and aggregation kernel dimensions to 64 across all datasets. For LSTM hidden dimensions, according to Figure 4(c), we set to 96, 80, 196 on SIP, NYC and Metr-LA. This analysis reveals the trade-off between efficiency and performances.

5 Related Work

Spatiotemporal learning methods usually devise various spatial and temporal aggregations, i.e., multi-view GCN [6], attention [7] and adaptive GCN-TCN [21] to achieve forecasting. However, majority of them assume the availability of sufficient historical observations, especially recent observations. Recently, by assuming the incompleteness, pioneering works investigate the imputation tasks [4, 16, 18] with GNN-based message passing [4], and coalition learning with designing discriminators [16, 18]. Actually, in practical scenarios, early planning of both urban activities and individual travelings require the models to foresee non-consecutive series, where fragments of observations are unavailable. Unfortunately, these solutions mentioned above only consider the sporadic observation missing, which is inherently incapable of dealing with lacking continuous fragments. To this end, techniques for non-consecutive spatiotemporal forecasting are urgently desired to develop more robust urban learning systems.

Uncertainty quantification (UQ). Uncertainty can be categorized into epistemic and aleatoric types [8, 9]. Regarding epistemic UQ, existing literature aims to capture distributions of model parameters. They devise various techniques, including dropout Bayesian networks [14, 17], Ensemble methods [10, 15] and Brownian motions [9], to collect multiple outputs from specific inputs and leverage sampling-computing strategy to achieve prediction variance as epistemic uncertainty. Aleatoric UQ usually exploits loss functions to maintain the consistency between errors and learnable aleatoric uncertainty [9]. More recently, a state-of-art work [19] provides comprehensive baselines and benchmarks on spatiotemporal uncertainty, while [23] devises a variation indicator as a guidance for uncertainty learning. However, these above works either neglect the model-induced epistemic uncertainty [23], or fail to internalize the context into dynamic spatiotemporal uncertainty [14, 17, 19], posing challenges to adapting them to uncertainty quantification of our G2S.

6 Conclusion

In this paper, we define a grey system where non-consecutive spatiotemporal forecasting is performed. Technically, we propose a FDG2S by exploiting exogenous factors for imitating the evolution patterns of un-

Table 3: Performances on ablative studies

Variants	MAPE		
	SIP	NYC	Metr-LA
w/o SF	0.2437	0.2642	0.2856
w/o TF	0.2499	0.2856	0.2642
w/o STA	0.2107	0.2305	0.2305
w/o Var	0.2034	0.2128	0.2058
FDG2S	0.2105	0.2023	0.2245

seen future status and providing responsible uncertainty estimations. We evaluate our solutions on two nonconsecutive settings in G2S. Extensive experiments verify the effectiveness and stability of our FDG2S, while case studies investigate the interpretability and robustness brought by UQ. For future works, we will further explore a unified model adaptive to various nonconsecutive settings and study learning on OOD samples in real-time data streams.

Acknowledgement

This work is supported by National Natural Science Foundation of China (No.62072427, No.12227901), Project of Stable Support for Youth Team in Basic Research Field, CAS (No.YSBR-005), and Academic Leaders Cultivation Program, USTC.

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] Albert-László Barabási, Réka Albert, and Hawoong Jeong. Mean-field theory for scale-free random networks. *Physica A: Statistical Mechanics and its Applications*, 272(1-2):173–187, 1999.

[3] 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.

[4] Andrea Cini, Ivan Marisca, and Cesare Alippi. Filling the g_ap_s: Multivariate time series imputation by graph neural networks. In *ICLR*, 2021.

[5] Hongchang Gao, Jian Pei, and Heng Huang. Conditional random field enhanced graph convolutional neural networks. In *KDD*, pages 276–284, 2019.

[6] Xu Geng, Yaguang Li, Leye Wang, Lingyu Zhang, Qiang Yang, Jieping Ye, and Yan Liu. Spatiotemporal multi-graph convolution network for ride-hailing demand forecasting. In *AAAI*, volume 33, pages 3656–3663, 2019.

[7] 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.

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

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

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

[11] Vincent Le Guen and Nicolas Thome. Probabilistic time series forecasting with structured shape and temporal diversity. In *NIPS*, 2019.

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

[13] Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *ICLR*, 2018.

[14] Yongqi Liu, Hui Qin, Zhendong Zhang, Shaoqian Pei, Zhiqiang Jiang, and al et. Probabilistic spatiotemporal wind speed forecasting based on a variational bayesian deep learning model. *Applied Energy*, 260:114259, 2020.

[15] Weizhu Qian, Dalin Zhang, Yan Zhao, Kai Zheng, and James JQ Yu. Uncertainty quantification for traffic forecasting: A unified approach. *arXiv preprint arXiv:2208.05875*, 2022.

[16] Huiling Qin, Xianyuan Zhan, Yuanxun Li, Xiaodu Yang, and Yu Zheng. Network-wide traffic states imputation using self-interested coalitional learning. In *KDD*, pages 1370–1378, 2021.

[17] 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.

[18] Pengkun Wang, Chaochao Zhu, Xu Wang, Zhengyang Zhou, Guang Wang, and Yang Wang. Inferring intersection traffic patterns with sparse video surveillance information: An st-gan method. *IEEE TVT*, 2022.

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

[20] Jie Wu, Wei Zhang, Guanbin Li, Wenhao Wu, Xiao Tan, Yingying Li, Errui Ding, and Liang Lin. Weakly-supervised spatio-temporal anomaly detection in surveillance video. In *IJCAI*, 2021.

[21] 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.

[22] Junbo Zhang, Yu Zheng, and Dekang Qi. Deep spatio-temporal residual networks for citywide crowd flows prediction. In *AAAI*, volume 31, 2017.

[23] 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.