
Taming Local Effects in Graph-based Spatiotemporal Forecasting

Andrea Cini ^{*1}, Ivan Marisca ^{*1}, Daniele Zambon ¹, Cesare Alippi ¹²

¹ The Swiss AI Lab IDSIA USI-SUPSI, Università della Svizzera italiana, ² Politecnico di Milano
{andrea.cini, ivan.marisca, daniele.zambon, cesare.alippi}@usi.ch

Abstract

Spatiotemporal graph neural networks have shown to be effective in time series forecasting applications, achieving better performance than standard univariate predictors in several settings. These architectures take advantage of a graph structure and relational inductive biases to learn a single (*global*) inductive model to predict any number of the input time series, each associated with a graph node. Despite the gain achieved in computational and data efficiency w.r.t. fitting a set of *local* models, relying on a single global model can be a limitation whenever some of the time series are generated by a different spatiotemporal stochastic process. The main objective of this paper is to understand the interplay between *globality* and *locality* in graph-based spatiotemporal forecasting, while contextually proposing a methodological framework to rationalize the practice of including trainable node embeddings in such architectures. We ascribe to trainable node embeddings the role of amortizing the learning of specialized components. Moreover, embeddings allow for 1) effectively combining the advantages of shared message-passing layers with node-specific parameters and 2) efficiently transferring the learned model to new node sets. Supported by strong empirical evidence, we provide insights and guidelines for specializing graph-based models to the dynamics of each time series and show how this aspect plays a crucial role in obtaining accurate predictions.

1 Introduction

Neural forecasting methods [1] take advantage of large databases of related time series to learn models for each individual process. If the time series in the database are not independent, functional dependencies among them can be exploited to obtain more accurate predictions, e.g., when the considered time series are observations collected from a network of sensors. In this setting, we use the term *spatiotemporal time series* to indicate the existence of relationships among subsets of time series (sensors) that span additional axes other than the temporal one, denoted here as *spatial* in a broad sense. Graph representations effectively model such dependencies and graph neural networks (GNNs) [2–4] can be included as modules in the forecasting architecture to propagate information along the spatial dimension. The resulting neural architectures are known in the literature as spatiotemporal graph neural networks (STGNNs) [5, 6] and have found widespread adoption in relevant applications ranging from traffic forecasting [6, 7] to energy analytics [8, 9]. These models embed inductive biases typical of graph deep learning and graph signal processing [10] and, thus, have several advantages over standard multivariate models; in fact, a shared set of learnable weights is used to obtain predictions for each time series by conditioning on observations at the neighboring nodes. Nonetheless, it has become more and more common to see node-specific trainable parameters being introduced as means to extract node(sensor)-level features then used as spatial identifiers within the processing [11, 12]. By doing so, the designer accepts a compromise in transferability that often

^{*}Equal contribution.

empirically leads to higher forecasting accuracy on the task at hand. We argue that the community has yet to find a proper explanatory framework for this phenomenon and, notably, has yet to design proper methodologies to deal with the root causes of observed empirical results, in practical applications.

In the broader context of time series forecasting, single models learned from a set of time series are categorized as *global* and are opposed to *local* models, which instead are specialized for any particular time series in the set [13, 14]. Global models are usually more robust than local ones, as they require a smaller total number of parameters and are fitted on more samples. Standard STGNNs, then, fall within the class of global models and, as a result, often have an advantage over local multivariate approaches; however, explicitly accounting for the behavior of individual time series might be problematic and require a large memory and model capacity. As an example, consider the problem of electric load forecasting: consumption patterns of single customers are influenced by shared factors, e.g., weather conditions and holidays, but are also determined by the daily routine of the individual users related by varying degrees of affinity. We refer to the dynamics proper to individual nodes as *local effects*.

Local effects can be accounted for by combining a global model with local components, e.g., by using encoding and/or decoding layers specialized for each input time series paired with a core global processing block. While such an approach to building hybrid global-local models can be effective, the added model complexity and specialization can negate the benefits of using a global component. In this paper, we propose to consider and interpret learned node embeddings as a mechanism to amortize the learning of local components; in fact, instead of learning a separate processing layer for each time series, node embeddings allow for learning a single global module conditioned on the learned (local) node features. Furthermore, we show that – within a proper methodological framework – node embeddings can be fitted to different node sets, thus enabling an effective and efficient transfer of the core processing modules.

Contributions In this paper, we analyze the effect of node-specific components in spatiotemporal time series models and assess how to incorporate them in the forecasting architecture, while providing an understanding of each design choice within a proper context. The major findings can be summarized in the following statements.

S1 Local components can be crucial to obtain accurate predictions in spatiotemporal forecasting.

S2 Node embeddings can amortize the learning of local components.

S3 Hybrid global-local STGNNs can capture local effects with contained model complexity and smaller input windows w.r.t. fully global approaches.

S4 Node embeddings for time series outside the training dataset can be obtained by fitting a relatively small number of observations and yield more effective transferability than fine-tuning global models.

S5 Giving structure to the embedding space provides an effective regularization, allowing for similarities among time series to emerge and shedding light on the role of local embeddings within a global architecture.

Throughout the paper, we reference the findings with their pointers. Against this backdrop, our main novel contributions reside in:

- A sound conceptual and methodological framework for dealing with local effects and designing node-specific components in STGNN architectures.
- An assessment of the role of learnable node embeddings in STGNNs and methods to obtain them.
- A comprehensive empirical analysis of the aforementioned phenomena in representative architectures across synthetic and real-world datasets.
- Methods to structure the embedding space, thus allowing for effective and efficient reuse of the global component in a transfer learning scenario.

We believe that our study constitutes an essential advancement toward the understanding of the interplay of different inductive biases in graph-based predictors, and argue that the methodologies and conceptual developments proposed in this work will constitute a foundational piece of know-how for the practitioner.

2 Preliminaries and problem statement

Consider a set of N time series with graph-side information where the i -th time series is composed by a sequence of d_x dimensional vectors $\mathbf{x}_t^i \in \mathbb{R}^{d_x}$ observed at time step t ; each time series $\{\mathbf{x}_t^i\}_t$ might be generated by a different stochastic process. Matrix $\mathbf{X}_t \in \mathbb{R}^{N \times d_x}$ encompasses the N observations at time t and, similarly, $\mathbf{X}_{t:t+T}$ indicates the sequence of observations within the time interval $[t, t+T)$. Relational information is encoded by a weighted adjacency matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$ that accounts for (soft) functional dependencies existing among the different time series. We use interchangeably the terms *node* and *sensor* to indicate the entities generating the time series and refer to the node set together with the relational information (graph) as *sensor network*. Eventual exogenous variables associated with each node are indicated by $\mathbf{U}_t \in \mathbb{R}^{N \times d_u}$; the tuple $\mathcal{G}_t = \langle \mathbf{X}_t, \mathbf{U}_t, \mathbf{A} \rangle$ indicates all the available information associated with time step t .

We address the multistep-ahead time-series forecasting problem, i.e., we are interested in predicting, for every time step t and some $H, W \geq 1$, the expected value of the next H observations $\mathbf{X}_{t:t+H}$ given a window $\mathcal{G}_{t-W:t}$ of W past measurements. In case data from multiple sensor networks are available, the problem can be formalized as learning from M disjoint collections of spatiotemporal time series $\mathcal{D} = \{\mathcal{G}_{t_1:t_1+T_1}^{(1)}, \mathcal{G}_{t_2:t_2+T_2}^{(2)}, \dots, \mathcal{G}_{t_m:t_m+T_m}^{(M)}\}$, potentially without overlapping time frames. In the latter case, we assume sensors to be homogeneous both within a single network and among different sets. Furthermore, we assume edges to indicate the same type of relational dependencies, e.g., physical proximity.

3 Forecasting with Spatiotemporal Graph Neural Networks

This section provides a taxonomy of the different components that constitute an STGNN. Based on the resulting archetypes, reference operators for this study are identified. The last paragraph of the section broadens the analysis to fit STGNNs within more general time series forecasting frameworks.

Spatiotemporal message-passing We consider STGNNs obtained by stacking spatiotemporal message-passing (STMP) layers s.t.

$$\mathbf{H}_t^{l+1} = \text{STMP}^l(\mathbf{H}_{\leq t}^l, \mathbf{A}), \quad (1)$$

where $\mathbf{H}_t^l \in \mathbb{R}^{N \times d_h}$ indicates the stack of node representations $\mathbf{h}_t^{i,l}$ at time step t at the l -th layer. The shorthand $\leq t$ indicates the sequence of all representations corresponding to the time steps up to t (included). Each $\text{STMP}^l(\cdot)$ layer is structured as follows

$$\mathbf{h}_t^{i,l+1} = \rho^l\left(\mathbf{h}_{\leq t}^{i,l}, \text{AGGR}_{j \in \mathcal{N}(i)}\left\{\gamma^l\left(\mathbf{h}_{\leq t}^{i,l}, \mathbf{h}_{\leq t}^{j,l}, a_{ji}\right)\right\}\right), \quad (2)$$

where ρ^l and γ^l are respectively the update and message functions, e.g., implemented by multilayer perceptrons (MLPs) or recurrent neural networks (RNNs). $\text{AGGR}\{\cdot\}$ indicates a generic permutation invariant aggregation function, while $\mathcal{N}(i)$ refers to the set of neighbors of node i , each associated with an edge with weight a_{ji} . Models of this type are fully *inductive*, in the sense that they can be used to make predictions for networks and time windows different from those they have been trained on, provided a certain level of similarity (e.g., homogenous sensors) between source and target node sets [15].

Among the different implementations of this general framework, we can distinguish between time-then-space (TTS) and time-and-space (T&S) models by following the terminology of previous works [16, 17]. Specifically, in TTS models the sequence of representations $\mathbf{h}_{\leq t}^{i,0}$ is encoded by a sequence model, e.g., an RNN, before propagating information along the spatial dimension through message passing (MP) [16]. Conversely, in T&S models time and space are processed in a more integrated fashion, e.g., by a recurrent GNN [5] or by spatiotemporal convolutional operators [7]. In the remainder of the paper, we take for TTS model an STGNN composed by a Gated Recurrent Unit (GRU) [18] followed by standard MP layers [19]:

$$\mathbf{H}_t^1 = \text{GRU}(\mathbf{H}_{\leq t}^0), \quad \mathbf{H}_t^{l+1} = \text{MP}^l(\mathbf{H}_t^l, \mathbf{A}), \quad (3)$$

where $l = 1, \dots, L-1$ and $\text{GRU}(\cdot)$ processes sequences node-wise. Similarly, we consider as reference T&S model a GRU with an MP network at its gates [5, 20], that process input data as

$$\mathbf{H}_t^{l+1} = \text{GRU}^l\left(\left\{\text{MP}^l(\mathbf{H}_t^l, \mathbf{A})\right\}_{\leq t}\right). \quad (4)$$

Moreover, STGNN models can be further categorized w.r.t. the implementation of message function; in particular, by loosely following Dwivedi et al. [21], we call *isotropic* those GNNs where the message function γ^l only depends on the features of the sender node $\mathbf{h}_{\leq t}^{j,l}$; conversely, we call *anisotropic* GNNs where γ^l takes both $\mathbf{h}_{\leq t}^{i,l}$ and $\mathbf{h}_{\leq t}^{j,l}$ as input. In the following, the case-study isotropic operator is

$$\mathbf{h}_t^{i,l+1} = \xi \left(\mathbf{W}_1^l \mathbf{h}_t^{i,l} + \text{MEAN}_{j \in \mathcal{N}(i)} \left\{ \mathbf{W}_2^l \mathbf{h}_t^{j,l} \right\} \right), \quad (5)$$

where \mathbf{W}_1^l and \mathbf{W}_2^l are matrices of learnable parameters and $\xi(\cdot)$ a generic activation function. Conversely, the operator of choice for the anisotropic case corresponds to

$$\mathbf{m}_t^{j \rightarrow i,l} = \mathbf{W}_2^l \xi \left(\mathbf{W}_1^l \left[\mathbf{h}_t^{i,l} \parallel \mathbf{h}_t^{j,l} \parallel a_{ji} \right] \right), \quad \alpha_t^{j \rightarrow i,l} = \sigma \left(\mathbf{W}_0^l \mathbf{m}_t^{j \rightarrow i,l} \right), \quad (6)$$

$$\mathbf{h}_t^{i,l+1} = \xi \left(\mathbf{W}_3^l \mathbf{h}_t^{i,l} + \text{SUM}_{j \in \mathcal{N}(i)} \left\{ \alpha_t^{j \rightarrow i,l} \mathbf{m}_t^{j \rightarrow i,l} \right\} \right), \quad (7)$$

where matrices $\mathbf{W}_0^l \in \mathbb{R}^{1 \times d_m}$, \mathbf{W}_1^l , \mathbf{W}_2^l and \mathbf{W}_3^l are learnable parameters, $\sigma(\cdot)$ is the sigmoid activation function and \parallel the concatenation operator applied along the feature dimension (see Appendix A.2 for a detailed description).

Global and local forecasting models Formally, a time series forecasting model is called *global* if its parameters are fitted to a group of time series (either univariate or multivariate), while *local* models are specific to a single (possibly multivariate) time series. The advantages of global models have been discussed at length in the time series forecasting literature [22, 14, 13, 1] and are mainly ascribable to the availability of large amounts of data that enable generalization and the use of models with higher capacity w.r.t. the single local models. As presented, and further detailed in the following section, STGNNs are global, yet have a peculiar position in this context, as they exploit spatial dependencies localizing predictions w.r.t. each node’s neighborhood. Furthermore, the transferability of GNNs makes these models distinctively different from local multivariate approaches enabling their use in cold-start scenarios [23] and making them inductive both temporally and spatially.

4 Locality and globality in Spatiotemporal Graph Neural Networks

We now focus on the impact of local effects in forecasting architectures based on STGNNs. The section starts by introducing a template to combine different processing layers within a global model, then continues by discussing how these can be turned into local. Contextually, we start to empirically probe the reference architectures.

4.1 A global processing template

STGNNs localize predictions in space, i.e., with respect to a single node, by exploiting an MP operator that contextualizes predictions by constraining the information flow within each node’s neighborhood. STGNNs are global forecasting models s.t.

$$\widehat{\mathbf{X}}_{t:t+H} = F_G(\mathcal{G}_{t-W:t}; \phi) \quad (8)$$

where ϕ are the learnable parameters shared among the time series and $\widehat{\mathbf{X}}_{t:t+H}$ indicate the H -step ahead predictions of the input time series given the window of (structured) observations $\mathbf{X}_{t-W:t}$. In particular, we consider forecasting architectures consisting of an encoding step followed by STMP layers and a final readout mapping representations to predictions; the corresponding sequence of operations composing F_G can be summarized as

$$\mathbf{h}_t^{i,0} = \text{ENCODER}(\mathbf{x}_{t-1}^i, \mathbf{u}_{t-1}^i), \quad (9)$$

$$\mathbf{H}_t^{l+1} = \text{STMP}^l(\mathbf{H}_{\leq t}^l, \mathbf{A}), \quad l = 0, \dots, L-1 \quad (10)$$

$$\hat{\mathbf{x}}_{t:t+H}^i = \text{DECODER}(\mathbf{h}_t^{i,L}). \quad (11)$$

ENCODER(\cdot) and DECODER(\cdot) indicate generic encoder and readout layers that could be implemented in several ways. In the following, the encoder is assumed to be a standard fully connected linear layer, while the decoder is implemented by an MLP with a single hidden layer followed by an output (linear) layer for each forecasting step.

4.2 Local effects

Differently from the global approach, local models are fitted to a single time series (e.g., see the standard Box-Jenkins approach) and, in our problem settings, can be indicated as

$$\hat{\mathbf{x}}_{t:t+H}^i = f_i(\mathbf{x}_{t-W:t}^i; \theta^i), \quad (12)$$

where $f_i(\cdot; \theta^i)$ is a sensor-specific model, e.g., an RNN with its dedicated parameters θ^i , fitted on the i -th time series. While the advantages of global models have already been discussed, local effects, i.e., the dynamics observed at the level of the single sensor, are potentially more easily captured by a local model. In fact, if local effects are present, global models might require an impractically large model capacity to account for all node-specific dynamics [13], thus losing some of the advantages of using a global approach (**S3**). In the STGNN case, then, increasing the input window for each node would result in a large computational overhead. Conversely, purely local approaches fail to exploit relational information among the time series and cannot reuse available knowledge efficiently in an inductive setting.

Combining global graph-based components with local node-level components has the potential for achieving a two-fold objective: 1) exploiting relational dependencies together with side information to learn flexible and efficient graph deep learning models and 2) making at the same time specialized and accurate predictions for each time series. In particular, we indicate global-local STGNNs as

$$\hat{\mathbf{x}}_{t:t+H}^i = F(\mathcal{G}_{t-W:t}; \phi, \theta^i) \quad (13)$$

where function F and parameter vector ϕ are shared across all nodes, whereas parameter vector θ^i is time-series dependent. Such a function $F(\cdot)$ could be implemented, for example, as a sum between a global model (Eq. 8) and a local one (Eq. 12):

$$\widehat{\mathbf{X}}_{t:t+H}^{(1)} = F_G(\mathcal{G}_{t-W:t}; \phi), \quad \hat{\mathbf{x}}_{t:t+H}^{i,(2)} = f_i(\mathbf{x}_{t-W:t}^i; \theta^i), \quad (14)$$

$$\hat{\mathbf{x}}_{t:t+H}^i = \hat{\mathbf{x}}_{t:t+H}^{i,(1)} + \hat{\mathbf{x}}_{t:t+H}^{i,(2)}, \quad (15)$$

or – with a more integrated approach – by using different weights for each time series at the encoding and/or decoding steps. The latter approach results in using a different encoder and/or decoder for each i -th node in the template STGNN (Eq. 9–11) to extract representations and, eventually, project them back into input space:

$$\mathbf{h}_t^{i,0} = \text{ENCODER}_i(\mathbf{x}_{t-1}^i, \mathbf{u}_{t-1}^i; \theta_{enc}^i), \quad (16) \quad \hat{\mathbf{x}}_{t:t+H}^i = \text{DECODER}_i(\mathbf{h}_t^{i,L}; \theta_{dec}^i). \quad (17)$$

MP layers could in principle be specialized as well, e.g., by using a different local update function $\gamma_i(\cdot)$ for each node. However, this would be impractical unless subsets of nodes are allowed to share parameters to some extent (e.g., by clustering them).

To support our arguments, Tab. 1 shows empirical results for the reference TTS models with isotropic message passing (TTS-IMP) on 2 popular traffic forecasting benchmarks (METR-LA and PEMS-BAY [6]). In particular, we compare the global approach with 3 hybrid global-local variants where local weights are used in the encoder, in the decoder, or in both of them (see Eq. 16-17 and the light brown block in Tab. 1). Notably, while fitting a separate RNN to each individual time series fails (LocalRNNs), exploiting a local encoder and/or decoder significantly improves performance w.r.t. the fully global model (**S1**). Note that the price of specialization is paid in terms of the number of learnable parameters which is an order of magnitude higher in global-local variants.

The table reports as a reference also results for FC-RNN, a multivariate RNN taking as input the concatenation of all time series. Indeed, having both encoder and decoder implemented as local layers leads to a large number of parameters and has a marginal impact on forecasting accuracy. The

Table 1: Performance (MAE) of TTS-IMP variants and number of associated trainable parameters in PEMS-BAY (5-run average).

		METR-LA	PEMS-BAY (# weights)		
Global TTS		3.35 ± 0.01	1.72 ± 0.00	4.71 × 10 ⁴	
Global-local TTS	Weights	Encoder	3.15 ± 0.01	1.66 ± 0.01	2.75 × 10 ⁵
		Decoder	3.09 ± 0.01	1.58 ± 0.00	3.00 × 10 ⁵
		Enc. + Dec.	3.16 ± 0.01	1.70 ± 0.01	5.28 × 10 ⁵
	Embed.	Encoder	3.08 ± 0.01	1.58 ± 0.00	5.96 × 10 ⁴
		Decoder	3.13 ± 0.00	1.60 ± 0.00	5.96 × 10 ⁴
		Enc. + Dec.	3.07 ± 0.01	1.58 ± 0.00	6.16 × 10 ⁴
FC-RNN		3.56 ± 0.03	2.32 ± 0.01	3.04 × 10 ⁵	
LocalRNNs		3.69 ± 0.00	1.91 ± 0.00	1.10 × 10 ⁷	

light gray block in Tab. 1 anticipates the effect of replacing the local layers with the use of learnable *node embeddings*, an approach discussed in depth in the next section. Additional results including TTS models with anisotropic message passing (TTS-AMP) are provided in Appendix B.

5 Node embeddings

This section introduces node embeddings as a mechanism to amortize the learning of local components and discusses the supporting empirical results. We then propose possible regularization techniques and discuss the advantages of embeddings in transfer learning scenarios.

Amortized specialization Static node features offer the opportunity to design and obtain node identification mechanisms across different time windows to tailor predictions to a specific node. However, in most settings, node features are either unavailable or insufficient to characterize the node dynamics. A possible solution consists of resorting to learnable node embeddings, i.e., a table of learnable parameters $\Theta = \mathbf{V} \in \mathbb{R}^{N \times d_v}$. Rather than interpret these learned representations as positional encodings, our proposal is to consider them as a way of amortizing the learning of node-level specialized models. More specifically, instead of learning a local model for each time series, embeddings fed into modules of a global STGNN and learned end-to-end with the forecasting architecture allow for specializing predictions by simply relying on gradient descent to find a suitable encoding.

The most straightforward option for feeding embeddings into the processing is to update the template model by changing the encoder and decoder as

$$\mathbf{h}_t^{i,0} = \text{ENCODER}(\mathbf{x}_{t-1}^i, \mathbf{u}_{t-1}^i, \mathbf{v}^i), \quad (18) \quad \hat{\mathbf{x}}_{t:t+H}^i = \text{DECODER}(\mathbf{h}_t^{i,L}, \mathbf{v}^i). \quad (19)$$

which can be seen as amortized versions of the encoder and decoder in Eq. 16-17. The encoding scheme of Eq. 18 also facilitates the propagation of relevant information by identifying nodes, an aspect that can be particularly significant as message-passing operators – in particular isotropic ones – can act as low-pass filters that smooth out node-level features [24, 25].

Tab. 1 (light gray block) reports empirical results that show the effectiveness of embeddings in amortizing the learning of local components (**S2**), with a negligible increase in the number of trainable parameters w.r.t. the base global model. In particular, feeding embeddings to the encoder, instead of conditioning the decoding step only, results in markedly better performance, hinting at the impact of providing node identification ahead of MP (additional empirical results provided in Sec. 7).

5.1 Structuring the embedding space

The latent space in which embeddings are learned can be structured and regularized to gather benefits in terms of interpretability, transferability, and generality (**S5**). In fact, accommodating new embeddings can be problematic, as they must fit in a region of the embedding space where the trained model can operate, and, at the same time, capture the local effects at the new nodes. In this setting, proper regularization can provide guidance and positive inductive biases to transfer the learned model to different node sets. As an example, if domain knowledge suggests that neighboring nodes have similar dynamics, Laplacian regularization [26, 27] can be added to the loss. In the following, we propose two general strategies based respectively on variational inference and on a clustering loss to impose soft constraints on the latent space. As shown in Sec. 7, the resulting structured space additionally allows us to gather insights into the features encoded in the embeddings.

Variational regularization As a probabilistic approach to structuring the latent space, we propose to consider learned embeddings as parameters of an approximate posterior distribution – given the training data – on the vector used to condition the predictions. In practice, we model each node embedding as a sample from a multivariate Gaussian $\mathbf{v}^i \sim q^i(\mathbf{v}^i|\mathcal{D}) = \mathcal{N}(\boldsymbol{\mu}_i, \text{diag}(\boldsymbol{\sigma}_i^2))$ where $(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i)$ are the learnable (local) parameters. Each node-level distribution is fitted on the training data by considering a standard Gaussian prior and exploiting the reparametrization trick [28] to minimize under the sampling of t

$$\delta_t \doteq \mathbb{E}_{\mathbf{V} \sim Q} \left[\left\| \widehat{\mathbf{X}}_{t:t+H} - \mathbf{X}_{t:t+H} \right\|_2^2 \right] + \beta D_{\text{KL}}(Q|P),$$

where $P = \mathcal{N}(\mathbf{0}, \mathbb{I})$ is the prior, D_{KL} the Kulback-Leibler divergence, and β controls the regularization strength. This regularization scheme results in a smooth latent space where it is easier to interpolate between representations, thus providing a principled way for accommodating different node embeddings.

Clustering regularization A different (and potentially complementary) approach to structuring the latent space is to incentivize node embeddings to form clusters and, consequently, to self-organize into different groups. We do so by introducing a regularization loss inspired by deep K -means algorithms [29]. In particular, besides the embedding table $\mathbf{V} \in \mathbb{R}^{N \times d_v}$, we equip the embedding module with a matrix $\mathbf{C} \in \mathbb{R}^{K \times d_v}$ of $K \ll N$ learnable centroids and a cluster assignment matrix $\mathbf{S} \in \mathbb{R}^{N \times K}$ encoding scores associated to each node-cluster pair. We consider scores as logits of a categorical (Boltzmann) distribution and learn them by minimizing the regularization term

$$\mathcal{L}_{reg} \doteq \mathbb{E}_{\mathbf{M}} [\|\mathbf{V} - \mathbf{M}\mathbf{C}\|_2], \quad p(\mathbf{M}_{ij} = 1) = \frac{e^{\mathbf{S}_{ij}/\tau}}{\sum_k e^{\mathbf{S}_{ik}/\tau}},$$

where τ is a hyperparameter. We minimize \mathcal{L}_{reg} – which corresponds to the embedding-to-centroid distance – jointly with the forecasting loss by relying on the Gumbel softmax trick [30]. Similarly to the variational inference approach, the clustering regularization gives structure to embedding space and allows for inspecting patterns in the learned local components (see Sec. 7).

Transferability of graph-based predictors Global models based on GNNs can make predictions for never-seen-before node sets, and handle graphs of different sizes and variable topology. In practice, this means that the graph-based predictors can easily handle new sensors being added to the network over time and be used for zero-shot transfer. Clearly, including in the forecasting architecture node-specific local components compromises these properties. Luckily, if local components are replaced by node embedding, adapting the specialized components is relatively cheap since the number of parameters to fit w.r.t. the new context is usually contained, and – eventually – both the graph topology and the structure of the embedding latent space can be exploited (**S4**). Experiments in Sec. 7 provide an in-depth empirical analysis of transferability within our framework and show that the discussed regularizations can be useful in this regard.

6 Related works

GNNs have been remarkably successful in modeling structured dynamical systems [31–33], temporal networks [34–36] and sequences of graphs [37, 38]. For what concerns time series processing, recurrent GNNs [5, 6] were among the first STGNNs being developed, followed by fully convolutional models [7, 39] and attention-based solutions [40–42]. Among the methods that focus on modeling node-specific dynamics, Bai et al. [11] use a factorization of the weight matrices in a recurrent STGNN to adapt the extracted representation to each node. Conversely, Chen et al. [43] use a model inspired by Wang et al. [44] consisting of a global GNN paired with a local model conditioned on the neighborhood of each node. Node embeddings have been mainly used in structure-learning modules to amortize the cost of learning the full adjacency matrix [39, 45, 12] and in attention-based approaches as positional encodings [46, 47, 41]. Shao et al. [48] observed how adding spatiotemporal identification mechanisms to the forecasting architecture can outperform several state-of-the-art STGNNs. Conversely, Yin et al. [49] used a cluster-based regularization to fine-tune an AGCRN-like model on different datasets. However, none of the previous works systematically addressed directly the problem of globality and locality in STGNNs, nor provided a comprehensive framework accounting for learnable node embeddings within different settings and architectures. Finally, besides STGNNs, there are several examples of hybrid global and local time series forecasting models. Wang et al. [44] propose an architecture where K global models extract dynamic global factors that are then weighted and integrated with probabilistic local models. Sen et al. [50] instead use a matrix factorization scheme paired with a temporal convolutional network [51] to learn a multivariate model then used to condition a second local predictor.

7 Experiments

This section reports salient results of an extensive empirical analysis of global and local models and combinations thereof in spatiotemporal forecasting benchmarks and different problem settings;

complete results of this systematic analysis can be found in Appendix B. Besides the reference architectures, we consider the following baselines and popular state-of-the-art architectures.

- RNN:** a global univariate RNN sharing the same parameters across the time series.
- FC-RNN:** a multivariate RNN taking as input the time series as if they were a multivariate one.
- LocalRNNs:** local univariate RNNs with different sets of parameters for each time series.
- DCRNN [6]:** a recurrent T&S model with the Diffusion Convolutional operator.
- AGCRN [11]:** the T&S global-local Adaptive Graph Convolutional Recurrent Network.
- GraphWaveNet:** the deep T&S spatiotemporal convolutional network by Wu et al. [39].

We also consider a global-local RNN, in which we specialize the model by using node embeddings in the encoder. Note that among the methods selected from the literature only DCRNN can be considered fully global (see Sec. 6). Performance is measured in terms of *mean absolute error* (MAE).

Synthetic data We start by assessing the performance of hybrid global-local spatiotemporal models in a controlled environment, considering a variation of GPVAR [52], a synthetic dataset based on a polynomial graph filter [53], that we modify to include local effects. In particular, data are generated from the spatiotemporal process

$$\mathbf{H}_t = \sum_{l=1}^L \sum_{q=1}^Q \Theta_{q,l} \mathbf{A}^{l-1} \mathbf{X}_{t-q},$$

$$\mathbf{X}_{t+1} = \mathbf{a} \odot \tanh(\mathbf{H}_t) + \mathbf{b} \odot \tanh(\mathbf{X}_{t-1}) + \eta_t, \quad (20)$$

where $\Theta \in \mathbb{R}^{Q \times L}$, $\mathbf{a} \in \mathbb{R}^N$, $\mathbf{b} \in \mathbb{R}^N$ and $\eta_t \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbb{I})$. We refer to this dataset as GPVAR-L: note that \mathbf{a} and \mathbf{b} are node-specific parameters that inject local effects into the spatiotemporal process. We indicate simply as GPVAR the process obtained by fixing $\mathbf{a} = \mathbf{b} = \mathbf{0.5}$, i.e., by removing local effects. We use as the adjacency matrix the community graph used in prior works, increasing the number of nodes to 120 (see Appendix A.1).

Tab. 2 shows forecasting accuracy for reference architectures with a 6-steps window on data generated from the processes. In the setting with no local effects, all STGNNs achieve performance close to the theoretical optimum, outperforming global and local univariate models but also the multivariate FC-RNN that – without any inductive bias – struggles to properly fit the data. In GPVAR-L, global and univariate models fail to match the performance of STGNNs that include local components (**S1**); interestingly, the global model with anisotropic MP outperforms the isotropic alternative, suggesting that the more advanced MP schemes can lead to more effective state identification.

Table 3: TTS-IMP one-step-ahead MAE on GPVAR-L with varying window length W and capacity d_h (5 runs).

W	Global			Embeddings		
	$d_h = 16$	$d_h = 32$	$d_h = 64$	$d_h = 16$	$d_h = 32$	$d_h = 64$
2	.5371±.0014	.4679±.0016	.4124±.0021	.3198±.0001	.3199±.0001	.3203±.0001
6	.4059±.0032	.3578±.0031	.3365±.0006	.3200±.0002	.3201±.0001	.3209±.0002
12	.3672±.0035	.3362±.0012	.3280±.0003	.3200±.0001	.3200±.0000	.3211±.0003
24	.3485±.0032	.3286±.0005	.3250±.0001	.3200±.0002	.3200±.0000	.3211±.0003

Benchmarks We then compare the performance of reference architectures and baselines with and without node embeddings at the encoding and decoding steps. Note that, while reference architectures and DCRNN are purely global models, GraphWaveNet and AGCRN use node embeddings to obtain an adjacency matrix for MP. AGCRN, furthermore, uses embeddings to make the convolutional filters

Table 2: One-step-ahead forecasting error (MAE) of the different models in GPVAR datasets (5 runs).

MODELS		GPVAR	GPVAR-L
FC-RNN		.4393±.0024	.5978±.0149
LocalRNNs		.4047±.0001	.4610±.0003
Global	RNN	.3999±.0000	.5440±.0003
	TTS-IMP	.3232±.0002	.4059±.0032
	TTS-AMP	.3193±.0000	.3587±.0049
Emb.	RNN	.3991±.0001	.4612±.0003
	TTS-IMP	.3195±.0000	.3200±.0002
	TTS-AMP	.3194±.0001	.3199±.0001
Optimal model		.3192	.3192

Finally, Tab. 3 shows that to match the performance of models with local components, in fully global models both window size and capacity should be increased (**S3**). This result is in line with what we should expect by considering the theory of local and global models and suggests that similar trade-offs might also happen in practical applications where both computational and sample complexity are a concern.

Table 4: Forecasting error (MAE) on 4 benchmark datasets (5 runs). The best result between each model and its variant with embeddings is in **bold**. N/A indicates runs exceeding resource capacity.

MODELS	METR-LA	PEMS-BAY	CER-E	AQI	METR-LA	PEMS-BAY	CER-E	AQI
Reference arch.	Global models				Global-local models (with embeddings)			
RNN	3.54±.00	1.77±.00	456.98±0.61	14.02±.04	3.15±.03	1.59±.00	421.50±1.78	13.73±.04
T&S-IMP	3.35±.01	1.70±.01	443.85±0.99	12.87±.02	3.10±.01	1.59±.00	417.71±1.28	12.48±.03
TTS-IMP	3.34±.01	1.72±.00	439.13±0.51	12.74±.02	3.08±.01	1.58±.00	412.44±2.80	12.33±.02
T&S-AMP	3.22±.02	1.65±.00	N/A	N/A	3.07±.02	1.59±.00	N/A	N/A
TTS-AMP	3.24±.01	1.66±.00	431.33±0.68	12.30±.02	3.06±.01	1.58±.01	412.95±1.28	12.15±.02
Baseline arch.	Original				Embeddings at Encoder and Decoder			
DCRNN	3.22±.01	1.64±.00	428.36±1.23	12.96±.03	3.07±.02	1.60±.00	412.87±1.51	12.53±.02
GraphWaveNet	3.05±.03	1.56±.01	397.17±0.67	12.08±.11	2.99±.02	1.58±.00	401.15±1.49	11.81±.04
AGCRN	3.16±.01	1.61±.00	444.80±1.25	13.33±.02	3.14±.00	1.62±.00	436.84±2.06	13.28±.03

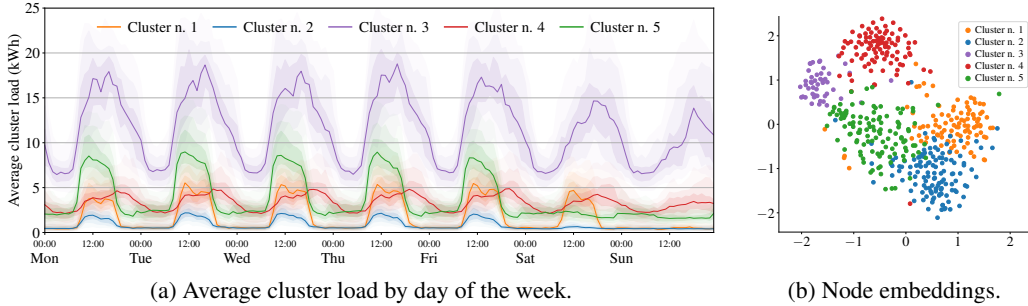


Figure 1: Time series clusters in CER-E, obtained by regularizing the embedding space. (a) Average load for each clusters. (b) t-SNE plot of the corresponding node embeddings.

adaptive w.r.t. the node being processed. We evaluate all models on real-world datasets from three different domains: traffic networks, energy analytics, and air quality monitoring. Besides the already introduced traffic forecasting benchmarks (**METR-LA** and **PEMS-BAY**), we run experiments on smart metering data from the **CER-E** dataset [54] and air quality measurements from **AQI** [55] by using the same pre-processing steps and data splits of previous works [20, 47]. The full experimental setup is reported in appendix A. Tab. 4 reports forecasting *mean absolute error* (MAE) averaged over the forecasting horizon. Global-local reference models outperform the fully global variants in every considered scenario (**S1**). A similar observation can be made for the state-of-art architectures, where the impact of node embeddings (at encoding and decoding) is large for the fully global DCRNN and more contained in models already equipped with local components. Note that hyperparameters were not tuned to account for the change in architecture. Surprisingly, the simple TTS-IMP model equipped with node embeddings achieves results comparable to that of state-of-the-art STGNNs with a significantly lower number of parameters and a streamlined architecture. Interestingly, while both global and local RNNs models fail, the hybrid global-local RNN ranks among the best-performing models, outperforming graph-based models without node embeddings in most settings.

Structured embeddings To test the hypothesis that structure in embedding space provides insights on the local effects at play (**S5**), we consider the clustering regularization method (Sec. 5.1) and the reference TTS-IMP model trained on the CER-E dataset. We set the number of learned centroids to $K = 5$ and train the cluster assignment mechanism end-to-end with the forecasting architecture. Then, we inspect the clustering assignment by looking at intra-cluster statistics. In particular, for each load profile, we compute the weekly average load curve, and, for each hour, we look at quantiles of the energy consumption within each cluster. Fig. 1a shows the results of the analysis by reporting the *median* load profile for each cluster; shaded areas correspond to quantiles with 10% increments. Results show that users in the different clusters have distinctly different consumption patterns. Fig. 1b

Table 5: Forecasting error (MAE) in the transfer learning setting (5 runs average). Results refer to a 1-week fine-tuning set size on all PEMS datasets.

TTS-IMP		PEMS03	PEMS04	PEMS07	PEMS08
Fine-tuning	Global	15.30 \pm 0.03	21.59 \pm 0.11	23.82 \pm 0.03	15.90 \pm 0.07
	Embeddings	14.64 \pm 0.05	20.27 \pm 0.11	22.23 \pm 0.08	15.45 \pm 0.06
	– Variational	14.56 \pm 0.03	20.19 \pm 0.05	22.43 \pm 0.02	15.41 \pm 0.06
	– Clustering	14.60 \pm 0.02	19.91 \pm 0.11	22.16 \pm 0.07	15.41 \pm 0.06
Zero-shot		18.20 \pm 0.09	23.88 \pm 0.08	32.76 \pm 0.69	20.41 \pm 0.07

Table 6: Forecasting error (MAE) on PEMS04 in the transfer learning setting by varying fine-tuning set size (5 runs average).

Model	Training set size			
	2 weeks	1 week	3 days	1 day
TTS-IMP				
Global	20.86 \pm 0.03	21.59 \pm 0.11	21.84 \pm 0.06	22.26 \pm 0.10
Embeddings	19.96 \pm 0.08	20.27 \pm 0.11	21.03 \pm 0.14	21.99 \pm 0.13
– Variational	19.94 \pm 0.08	20.19 \pm 0.05	20.71 \pm 0.12	21.20 \pm 0.15
– Clustering	19.69 \pm 0.06	19.91 \pm 0.11	20.48 \pm 0.09	21.91 \pm 0.21

shows a 2D t-SNE visualization of the learned node embeddings, providing a view of the latent space and the effects of the cluster-based regularization.

Transfer In this experiment, we consider the scenario in which an STGNN for traffic forecasting is trained by using data from multiple traffic networks and then used to make predictions for a disjoint set of sensors sampled from the same region. We use the **PEMS03**, **PEMS04**, **PEMS07**, and **PEMS08** datasets [56], which contain measurements from 4 different districts in California. We train models on 3 of the datasets, fine-tune on 1 week of data from the target left-out dataset, validate on the following week, and test on the week thereafter. We compare variants of TTS-IMP with and without embeddings fed into encoder and decoder. Together with the unconstrained embeddings, we also consider the variational and clustering regularization approaches introduced in Sec. 5.1. At the fine-tuning stage, the global model updates all of its parameters, while in the hybrid global-local approaches only the embeddings are fitted to the new data. Tab. 5 reports results for the described scenario. The fully global approach is outperformed by the hybrid architectures in all target datasets (**S4**). Besides the significant improvement in performance, adjusting only node embeddings retains performance on the source datasets. Furthermore, results show the positive effects of regularizing the embedding space in the transfer setting (**S5**). This is further confirmed by results in Tab. 6, which report, for PEMS04, how forecasting error changes in relation to the length of the fine-tuning window. We refer to Appendix B for an in-depth analysis of several additional transfer learning scenarios.

8 Conclusions

We investigate the impact of locality and globality in graph-based spatiotemporal forecasting architectures. We propose a framework to explain empirical results associated with the use of trainable node embeddings and discuss different architectures and regularization techniques to account for local effects. The proposed methodologies are thoroughly empirically validated and, although not inductive, prove to be effective in a transfer learning context. We argue that our work provides necessary and key methodologies for the understanding and design of effective graph-based spatiotemporal forecasting architectures. Future works can build on the results presented here and study alternative, and even more transferable, methods to account for local effects.

Acknowledgements

This research was partly funded by the Swiss National Science Foundation under grant 204061: *High-Order Relations and Dynamics in Graph Neural Networks*.

References

- [1] Konstantinos Benidis, Syama Sundar Rangapuram, Valentin Flunkert, Yuyang Wang, Danielle Maddix, Caner Turkmen, Jan Gasthaus, Michael Bohlke-Schneider, David Salinas, Lorenzo Stella, François-Xavier Aubet, Laurent Callot, and Tim Januschowski. Deep learning for time series forecasting: Tutorial and literature survey. *ACM Comput. Surv.*, 55(6), dec 2022. ISSN 0360-0300. doi: 10.1145/3533382. URL <https://doi.org/10.1145/3533382>.
- [2] Davide Bacciu, Federico Errica, Alessio Micheli, and Marco Podda. A gentle introduction to deep learning for graphs. *Neural Networks*, 129:203–221, 2020.
- [3] Ljubiša Stanković, Danilo Mandić, Miloš Daković, Miloš Brajović, Bruno Scalzo, Shengxi Li, Anthony G Constantinides, et al. Data analytics on graphs part iii: machine learning on graphs, from graph topology to applications. *Foundations and Trends® in Machine Learning*, 13(4): 332–530, 2020.
- [4] Michael M Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.
- [5] Youngjoo Seo, Michaël Defferrard, Pierre Vandergheynst, and Xavier Bresson. Structured sequence modeling with graph convolutional recurrent networks. In *International Conference on Neural Information Processing*, pages 362–373. Springer, 2018.
- [6] Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *International Conference on Learning Representations*, 2018.
- [7] Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: a deep learning framework for traffic forecasting. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 3634–3640, 2018.
- [8] Simone Eandi, Andrea Cini, Slobodan Lukovic, and Cesare Alippi. Spatio-temporal graph neural networks for aggregate load forecasting. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2022.
- [9] Jonas Berg Hansen, Stian Normann Anfinsen, and Filippo Maria Bianchi. Power flow balancing with decentralized graph neural networks. *IEEE Transactions on Power Systems*, 2022.
- [10] Antonio Ortega, Pascal Frossard, Jelena Kovačević, José MF Moura, and Pierre Vandergheynst. Graph signal processing: Overview, challenges, and applications. *Proceedings of the IEEE*, 106(5):808–828, 2018.
- [11] Lei Bai, Lina Yao, Can Li, Xianzhi Wang, and Can Wang. Adaptive graph convolutional recurrent network for traffic forecasting. *Advances in Neural Information Processing Systems*, 33:17804–17815, 2020.
- [12] Ailin Deng and Bryan Hooi. Graph neural network-based anomaly detection in multivariate time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 4027–4035, 2021.
- [13] Pablo Montero-Manso and Rob J Hyndman. Principles and algorithms for forecasting groups of time series: Locality and globality. *International Journal of Forecasting*, 37(4):1632–1653, 2021.
- [14] Tim Januschowski, Jan Gasthaus, Yuyang Wang, David Salinas, Valentin Flunkert, Michael Bohlke-Schneider, and Laurent Callot. Criteria for classifying forecasting methods. *International Journal of Forecasting*, 36(1):167–177, 2020.
- [15] Luana Ruiz, Luiz Chamon, and Alejandro Ribeiro. Graphon neural networks and the transferability of graph neural networks. *Advances in Neural Information Processing Systems*, 33: 1702–1712, 2020.
- [16] Jianfei Gao and Bruno Ribeiro. On the equivalence between temporal and static equivariant graph representations. In *International Conference on Machine Learning*, pages 7052–7076. PMLR, 2022.

- [17] Andrea Cini, Daniele Zambon, and Cesare Alippi. Sparse graph learning from spatiotemporal time series. *Journal of Machine Learning Research*, 24(242):1–36, 2023.
- [18] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.
- [19] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pages 1263–1272. PMLR, 2017.
- [20] Andrea Cini, Ivan Marisca, and Cesare Alippi. Filling the g_ap_s: Multivariate time series imputation by graph neural networks. In *International Conference on Learning Representations*, 2022.
- [21] Vijay Prakash Dwivedi, Chaitanya K. Joshi, Anh Tuan Luu, Thomas Laurent, Yoshua Bengio, and Xavier Bresson. Benchmarking graph neural networks. *Journal of Machine Learning Research*, 24(43):1–48, 2023. URL <http://jmlr.org/papers/v24/22-0567.html>.
- [22] David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. Deepar: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3): 1181–1191, 2020.
- [23] Kenneth B Kahn and Charles W Chase. The state of new-product forecasting. *Foresight: The International Journal of Applied Forecasting*, (51), 2018.
- [24] Hoang Nt and Takanori Maehara. Revisiting graph neural networks: All we have is low-pass filters. *arXiv preprint arXiv:1905.09550*, 2019.
- [25] Kenta Oono and Taiji Suzuki. Graph neural networks exponentially lose expressive power for node classification. In *International Conference on Learning Representations*, 2020.
- [26] Dengyong Zhou, Olivier Bousquet, Thomas Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. *Advances in neural information processing systems*, 16, 2003.
- [27] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017.
- [28] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *Proceedings of International Conference on Learning Representations*, 2013.
- [29] Bo Yang, Xiao Fu, Nicholas D Sidiropoulos, and Mingyi Hong. Towards k-means-friendly spaces: Simultaneous deep learning and clustering. In *international conference on machine learning*, pages 3861–3870. PMLR, 2017.
- [30] C Maddison, A Mnih, and Y Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *International Conference on Learning Representations*, 2017.
- [31] Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter Battaglia. Learning to simulate complex physics with graph networks. In *International conference on machine learning*, pages 8459–8468. PMLR, 2020.
- [32] Zijie Huang, Yizhou Sun, and Wei Wang. Learning continuous system dynamics from irregularly-sampled partial observations. *Advances in Neural Information Processing Systems*, 33:16177–16187, 2020.
- [33] Song Wen, Hao Wang, and Dimitris Metaxas. Social ode: Multi-agent trajectory forecasting with neural ordinary differential equations. In *European Conference on Computer Vision*, pages 217–233. Springer, 2022.
- [34] Da Xu, Chuanwei Ruan, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. Inductive representation learning on temporal graphs. *International Conference on Learning Representations*, 2020.

- [35] Seyed Mehran Kazemi, Rishab Goel, Kshitij Jain, Ivan Kobyzev, Akshay Sethi, Peter Forsyth, and Pascal Poupart. Representation learning for dynamic graphs: A survey. *The Journal of Machine Learning Research*, 21(1):2648–2720, 2020.
- [36] Antonio Longa, Veronica Lachi, Gabriele Santin, Monica Bianchini, Bruno Lepri, Pietro Lio, franco scarselli, and Andrea Passerini. Graph neural networks for temporal graphs: State of the art, open challenges, and opportunities. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL <https://openreview.net/forum?id=pHCdMat0gI>.
- [37] Daniele Grattarola, Daniele Zambon, Lorenzo Livi, and Cesare Alippi. Change detection in graph streams by learning graph embeddings on constant-curvature manifolds. *IEEE Transactions on neural networks and learning systems*, 31(6):1856–1869, 2019.
- [38] Benjamin Paassen, Daniele Grattarola, Daniele Zambon, Cesare Alippi, and Barbara Eva Hammer. Graph edit networks. In *International Conference on Learning Representations*, 2020.
- [39] Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, and Chengqi Zhang. Graph wavenet for deep spatial-temporal graph modeling. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 1907–1913, 2019.
- [40] Jiani Zhang, Xingjian Shi, Junyuan Xie, Hao Ma, Irwin King, and Dit Yan Yeung. Gaan: Gated attention networks for learning on large and spatiotemporal graphs. In *34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018*, 2018.
- [41] Chuanpan Zheng, Xiaoliang Fan, Cheng Wang, and Jianzhong Qi. Gman: A graph multi-attention network for traffic prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 1234–1241, 2020.
- [42] Zonghan Wu, Da Zheng, Shirui Pan, Quan Gan, Guodong Long, and George Karypis. Traversenet: Unifying space and time in message passing for traffic forecasting. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.
- [43] Hongjie Chen, Ryan A. Rossi, Kanak Mahadik, Sungchul Kim, and Hoda Eldardiry. Graph deep factors for forecasting with applications to cloud resource allocation. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD '21*, page 106–116, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383325. doi: 10.1145/3447548.3467357. URL <https://doi.org/10.1145/3447548.3467357>.
- [44] Yuyang Wang, Alex Smola, Danielle Maddix, Jan Gasthaus, Dean Foster, and Tim Januschowski. Deep factors for forecasting. In *International conference on machine learning*, pages 6607–6617. PMLR, 2019.
- [45] Chao Shang and Jie Chen. Discrete graph structure learning for forecasting multiple time series. In *Proceedings of International Conference on Learning Representations*, 2021.
- [46] Victor Garcia Satorras, Syama Sundar Rangapuram, and Tim Januschowski. Multivariate time series forecasting with latent graph inference. *arXiv preprint arXiv:2203.03423*, 2022.
- [47] Ivan Marisca, Andrea Cini, and Cesare Alippi. Learning to reconstruct missing data from spatiotemporal graphs with sparse observations. *Advances in Neural Information Processing Systems*, 35:32069–32082, 2022.
- [48] Zezhi Shao, Zhao Zhang, Fei Wang, Wei Wei, and Yongjun Xu. Spatial-temporal identity: A simple yet effective baseline for multivariate time series forecasting. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, page 4454–4458, 2022.
- [49] Xueyan Yin, Feifan Li, Yanming Shen, Heng Qi, and Baocai Yin. Nodetrans: A graph transfer learning approach for traffic prediction. *arXiv preprint arXiv:2207.01301*, 2022.
- [50] Rajat Sen, Hsiang-Fu Yu, and Inderjit S Dhillon. Think globally, act locally: A deep neural network approach to high-dimensional time series forecasting. *Advances in neural information processing systems*, 32, 2019.

- [51] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.
- [52] Daniele Zambon and Cesare Alippi. AZ-whiteness test: a test for uncorrelated noise on spatio-temporal graphs. *Advances in Neural Information Processing Systems*, 35:11975–11986, 2022.
- [53] Elvin Isufi, Andreas Loukas, Nathanael Perraudin, and Geert Leus. Forecasting time series with VARMA recursions on graphs. *IEEE Transactions on Signal Processing*, 67(18):4870–4885, 2019.
- [54] Commission for Energy Regulation. CER Smart Metering Project - Electricity Customer Behaviour Trial, 2009-2010 [dataset]. *Irish Social Science Data Archive. SN: 0012-00*, 2016. URL <https://www.ucd.ie/issda/data/commissionforenergyregulationcer>.
- [55] Yu Zheng, Xiuwen Yi, Ming Li, Ruiyuan Li, Zhangqing Shan, Eric Chang, and Tianrui Li. Forecasting fine-grained air quality based on big data. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 2267–2276, 2015.
- [56] Shengnan Guo, Youfang Lin, Huaiyu Wan, Xiucheng Li, and Gao Cong. Learning dynamics and heterogeneity of spatial-temporal graph data for traffic forecasting. *IEEE Transactions on Knowledge and Data Engineering*, 34(11):5415–5428, 2021.
- [57] Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009. ISBN 1441412697.
- [58] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.
- [59] William Falcon and The PyTorch Lightning team. PyTorch Lightning, 3 2019. URL <https://github.com/PyTorchLightning/pytorch-lightning>.
- [60] Matthias Fey and Jan Eric Lenssen. Fast graph representation learning with pytorch geometric. *arXiv preprint arXiv:1903.02428*, 2019.
- [61] Andrea Cini and Ivan Marisca. Torch Spatiotemporal, 3 2022. URL <https://github.com/TorchSpatiotemporal/tsl>.
- [62] Charles R Harris, K Jarrod Millman, Stéfan J Van Der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J Smith, et al. Array programming with numpy. *Nature*, 585(7825):357–362, 2020.
- [63] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [64] Chao Chen, Karl Petty, Alexander Skabardonis, Pravin Varaiya, and Zhanfeng Jia. Freeway performance measurement system: mining loop detector data. *Transportation Research Record*, 1748(1):96–102, 2001.
- [65] Andrea Cini, Ivan Marisca, Filippo Maria Bianchi, and Cesare Alippi. Scalable spatiotemporal graph neural networks. In *NeurIPS 2022 Temporal Graph Learning Workshop*, 2022.
- [66] Xiuwen Yi, Yu Zheng, Junbo Zhang, and Tianrui Li. St-mvl: filling missing values in geosensory time series data. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, 2016.
- [67] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015*, 2015.

Appendix

The following sections provide additional details on the experimental setting and computational platform used to gather the results presented in the paper. Additionally, we also include supplementary empirical results and analysis.

A Experimental setup

Experimental setup and baselines have been developed with Python [57] by relying on the following open-source libraries:

- PyTorch [58];
- PyTorch Lightning [59];
- PyTorch Geometric [60]
- Torch Spatiotemporal [61];
- numpy [62];
- scikit-learn [63].

Experiments were run on a workstation equipped with AMD EPYC 7513 processors and four NVIDIA RTX A5000 GPUs. The code needed to reproduce the reported results is available online².

A.1 Datasets

Table 7: Statistics of datasets used in the experiments.

DATASETS	Type	Time steps	Nodes	Edges	Rate
GPVAR	Undirected	30,000	120	199	N/A
GPVAR-L	Undirected	30,000	120	199	N/A
METR-LA	Directed	34,272	207	1515	5 minutes
PEMS-BAY	Directed	52,128	325	2369	5 minutes
CER-E	Directed	25,728	485	4365	30 minutes
AQI	Undirected	8,760	437	2699	1 hour
PEMS03	Directed	26,208	358	546	5 minutes
PEMS04	Directed	16,992	307	340	5 minutes
PEMS07	Directed	28,224	883	866	5 minutes
PEMS08	Directed	17,856	170	277	5 minutes

In this section, we provide additional information regarding each one of the considered datasets. Tab. 7 reports a summary of statistics.

GPVAR(-L) For the GPVAR datasets we follow the procedure described in Sec. 7 to generate data and then partition the resulting time series in 70%/10%/20% splits for training, validation and testing, respectively. For GPVAR-L the parameters of the spatiotemporal process are set as

$$\Theta = \begin{bmatrix} 2.5 & -2.0 & -0.5 \\ 1.0 & 3.0 & 0.0 \end{bmatrix}, \quad \mathbf{a}, \mathbf{b} \sim \mathcal{U}(-2, 2),$$

$$\eta \sim \mathcal{N}(\mathbf{0}, \text{diag}(\sigma^2)), \quad \sigma = 0.4.$$

Fig. 2 shows the topology of the graph used as a support to the process. In particular, we considered a network with 120 nodes with 20 communities and added self-loops to the graph adjacency matrix.

²<https://github.com/Graph-Machine-Learning-Group/taming-local-effects-stgnns>

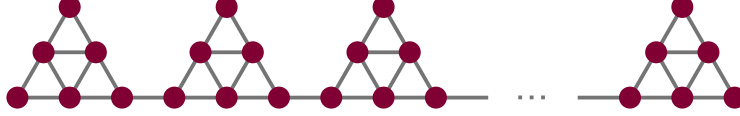


Figure 2: GPVAR community graph. We used a graph with 20 communities resulting in a network with 120 nodes.

Spatiotemporal forecasting benchmarks We considered several datasets coming from relevant application domains and different problem settings corresponding to real-world application scenarios.

Traffic forecasting We consider two popular traffic forecasting datasets, namely **METR-LA** and **PEMS-BAY** [6], containing measurements from loop detectors in the Los Angeles County Highway and San Francisco Bay Area, respectively. For the experiment on transfer learning, we use the **PEMS03**, **PEMS04**, **PEMS07**, and **PEMS08** datasets from Guo et al. [56] each collecting traffic flow readings, aggregated into 5-minutes intervals, from different areas in California provided by Caltrans Performance Measurement System (PeMS) [64].

Electric load forecasting We selected the **CER-E** dataset [54], a collection of energy consumption readings, aggregated into 30-minutes intervals, from 485 smart meters monitoring small and medium-sized enterprises.

Air quality monitoring The **AQI** [55] dataset collects hourly measurements of pollutant PM2.5 from 437 air quality monitoring stations in China. Note that all of these datasets have been previously used for spatiotemporal forecasting and imputation [47, 65].

For each dataset, we obtain the corresponding adjacency matrix by following previous works [20, 6, 56]. We use as exogenous variables sinusoidal functions encoding the time of the day and the one-hot encoding of the day of the week. For datasets with an excessive number of missing values, namely METR-LA and AQI, we add as an exogenous variable also a binary mask indicating if the corresponding value has been imputed.

We split datasets into windows of W time steps, and train the models to predict the next H observations. For the experiment in Tab.4, we set $W = 12$, $H = 12$ for the traffic datasets, $W = 48$, $H = 6$ for CER-E, and $W = 24$, $H = 3$ for AQI. Then, for all datasets except for AQI, we divide the obtained windows sequentially into 70%/10%/20% splits for training, validation, and testing, respectively. For AQI instead, we use as the test set the months of March, June, September, and December, following Yi et al. [66].

A.2 Architectures

In this section, we provide a detailed description of the reference architectures used in the study.

Reference architectures We consider 4 simple STGNN architectures that follow the template given in (Eq. 9–11). In all the models, we use as ENCODER (Eq. 9) a simple linear layer s.t.

$$\mathbf{h}_t^{i,0} = \mathbf{W}_{\text{enc}} [\mathbf{x}_{t-1}^i \parallel \mathbf{u}_{t-1}^i] \quad (21)$$

and as DECODER (Eq. 11) a 2 layer module structured as

$$\hat{\mathbf{x}}_{t:t+H}^i = \left\{ \mathbf{W}_h \xi \left(\mathbf{W}_{\text{dec}} \mathbf{h}_t^{i,L} \right) \right\}_{h=1, \dots, H} \quad (22)$$

where \mathbf{W}_{enc} , \mathbf{W}_{dec} , and $\mathbf{W}_1, \dots, \mathbf{W}_H$ are learnable parameters. In hybrid models with local ENCODER and/or DECODER, we use different parameter matrices for each node, i.e.,

$$\mathbf{h}_t^{i,0} = \mathbf{W}_{\text{enc}}^i [\mathbf{x}_{t-1}^i \parallel \mathbf{u}_{t-1}^i] \quad (23)$$

$$\hat{\mathbf{x}}_{t:t+H}^i = \left\{ \mathbf{W}_h^i \xi \left(\mathbf{W}_{\text{dec}} \mathbf{h}_t^{i,L} \right) \right\}_{h=1, \dots, H}. \quad (24)$$

When instead node embeddings $\mathbf{V} \in \mathbb{R}^{N \times d_v}$ are used to specialize the model, they are concatenated to the modules' input as

$$\mathbf{h}_t^{i,0} = \mathbf{W}_{\text{enc}} [\mathbf{x}_{t-1}^i \parallel \mathbf{u}_{t-1}^i \parallel \mathbf{v}^i] \quad (25)$$

$$\hat{\mathbf{x}}_{t:t+H}^i = \left\{ \mathbf{W}_h \xi \left(\mathbf{W}_{\text{dec}} \left[\mathbf{h}_t^{i,L} \parallel \mathbf{v}^i \right] \right) \right\}_{h=1, \dots, H}. \quad (26)$$

For the TTS architectures, we define the STMP module (Eq. 10) as a node-wise GRU [18]:

$$\mathbf{r}_t^i = \sigma(\mathbf{W}_1^l [\mathbf{h}_t^{i,0} \parallel \mathbf{h}_{t-1}^{i,1}]), \quad (27)$$

$$\mathbf{o}_t^i = \sigma(\mathbf{W}_2^l [\mathbf{h}_t^{i,0} \parallel \mathbf{h}_{t-1}^{i,1}]), \quad (28)$$

$$\mathbf{c}_t^i = \tanh(\mathbf{W}_3^l [\mathbf{h}_t^{i,0} \parallel \mathbf{r}_t^i \odot \mathbf{h}_{t-1}^{i,1}]), \quad (29)$$

$$\mathbf{h}_t^{i,1} = \mathbf{o}_t^i \odot \mathbf{h}_{t-1}^{i,1} + (1 - \mathbf{o}_t^i) \odot \mathbf{c}_t^i. \quad (30)$$

The GRU is then followed by L MP layers

$$\mathbf{H}_t^{l+1} = \text{MP}^l(\mathbf{H}_t^l, \mathbf{A}), \quad l = 1, \dots, L-1. \quad (31)$$

We consider two variants for this architecture: **TTS-IMP**, featuring the isotropic MP operator defined in Eq. 5, and **TTS-AMP**, using the anisotropic MP operator defined in Eq. 6-7. Note that all the parameters in the STMP blocks are shared among the nodes.

In T&S models, instead, we use a GRU where gates are implemented using MP layers:

$$\mathbf{r}_t^{i,l} = \sigma\left(\text{MP}_r^l\left(\left[\mathbf{h}_t^{i,l-1} \parallel \mathbf{h}_{t-1}^{i,l}\right], \mathbf{A}\right)\right), \quad (32)$$

$$\mathbf{o}_t^{i,l} = \sigma\left(\text{MP}_o^l\left(\left[\mathbf{h}_t^{i,l-1} \parallel \mathbf{h}_{t-1}^{i,l}\right], \mathbf{A}\right)\right), \quad (33)$$

$$\mathbf{c}_t^{i,l} = \tanh\left(\text{MP}_c^l\left(\left[\mathbf{h}_t^{i,l-1} \parallel \mathbf{r}_t^{i,l} \odot \mathbf{h}_{t-1}^{i,l}\right], \mathbf{A}\right)\right), \quad (34)$$

$$\mathbf{h}_t^{i,l} = \mathbf{o}_t^{i,l} \odot \mathbf{h}_{t-1}^{i,l} + (1 - \mathbf{o}_t^{i,l}) \odot \mathbf{c}_t^{i,l}. \quad (35)$$

Similarly to the TTS case, we indicate as **T&S-IMP** the reference architecture in which MP operators are isotropic and **T&S-AMP** the one featuring anisotropic MP.

Baselines For the RNN baselines we follow the same template of the reference architectures (Eq. 9-11) but use a single GRU cell as core processing module instead of the STMP block. The global-local **RNN**, then, uses node embeddings at encoding as shown in Eq. 25, whereas **LocalRNNs** have different sets of parameters for each time series. The **FC-RNN** architecture, instead, takes as input the concatenation all the time series as if they were a single multivariate one.

Hyperparameters For the reference TTS architectures we use a GRU with a single cell followed by 2 message-passing layers. In T&S case we use a single graph recurrent convolutional cell. The number of neurons in each layer is set to 64 and the embedding size to 32 for all the reference architectures in all the benchmark datasets. Analogous hyperparameters were used for the RNN baselines. For GPVAR instead, we use 16 and 8 as hidden and embedding sizes, respectively. For the baselines from the literature, we use the hyperparameters used in the original papers whenever possible.

For GPVAR experiments we use a batch size of 128 and train with early stopping for a maximum of 200 epochs with the Adam optimizer [67] and a learning rate of 0.01 halved every 50 epochs.

For experiments in Tab. 1 and 4, we instead trained the models with batch size 64 for a maximum of 300 epochs each consisting of maximum 300 batches. The initial learning rate was set to 0.003 and reduced every 50 epochs.

Transfer learning experiment In the transfer learning experiments we simulate the case where a pre-trained spatiotemporal model for traffic forecasting is tested on a new road network. We fine-tune on each of the PEMS03-08 datasets models previously trained on the left-out 3 datasets. Following previous works [56], we split the datasets into 60%/20%/20% for training, validation, and testing, respectively. We obtain the mini-batches during training by uniformly sampling 64 windows from the 3 training sets. To train the models we use a similar experimental setting of experiments in Tab. 4, decreasing the number of epochs to 150 and increasing the batches per epoch to 500. We then assume that 2 weeks of readings from the target dataset are available for fine-tuning, and use the first week for training and the second one as the validation set. Then, we test fine-tuned models on the immediately following week (Tab. 5). For the global model, we either tune all the parameters or none of them (zero-shot setting). For fine-tuning, we increase the maximum number of epochs to 2000 without limiting the batches processed per epoch and fixing the learning rate to 0.001. At the end of every

Table 8: Performance (MAE) of TTS-IMP and TTS-AMP variants and number of associated trainable parameters in PEMS-BAY (5-run average).

		METR-LA	PEMS-BAY (# weights)		METR-LA	PEMS-BAY (# weights)		
		TTS-IMP			TTS-AMP			
Global-local TTS	Global TTS	3.35 \pm 0.01	1.72 \pm 0.00	4.71 \times 10 ⁴	3.27 \pm 0.01	1.68 \pm 0.00	6.41 \times 10 ⁴	
	Weights	Encoder	3.15 \pm 0.01	1.66 \pm 0.01	2.75 \times 10 ⁵	3.12 \pm 0.01	1.64 \pm 0.00	2.92 \times 10 ⁵
		Decoder	3.09 \pm 0.01	1.58 \pm 0.00	3.00 \times 10 ⁵	3.09 \pm 0.01	1.58 \pm 0.00	3.17 \times 10 ⁵
		Enc. + Dec.	3.16 \pm 0.01	1.70 \pm 0.01	5.28 \times 10 ⁵	3.14 \pm 0.01	1.69 \pm 0.01	5.45 \times 10 ⁵
	Embed.	Encoder	3.08 \pm 0.01	1.58 \pm 0.00	5.96 \times 10 ⁴	3.05 \pm 0.02	1.58 \pm 0.01	7.66 \times 10 ⁴
		Decoder	3.13 \pm 0.00	1.60 \pm 0.00	5.96 \times 10 ⁴	3.12 \pm 0.01	1.60 \pm 0.00	7.66 \times 10 ⁴
Enc. + Dec.		3.07 \pm 0.01	1.58 \pm 0.00	6.16 \times 10 ⁴	3.04 \pm 0.01	1.59 \pm 0.01	7.86 \times 10 ⁴	
FC-RNN		3.56 \pm 0.03	2.32 \pm 0.01	3.04 \times 10 ⁵				
LocalRNNs		3.69 \pm 0.00	1.91 \pm 0.00	1.10 \times 10 ⁷				

training epoch, we compute the MAE on the validation set and stop training if it has not decreased in the last 100 epochs, restoring the model weights corresponding to the best-performing model. For the global-local models with variational regularization on the embedding space, during training, we set $\beta = 0.05$ and initialize the distribution parameters as

$$\boldsymbol{\mu}_i \sim \mathcal{U}(-0.01, 0.01), \quad \boldsymbol{\sigma}_i = \mathbf{0.2}.$$

For fine-tuning instead, we initialize the new embedding table $\mathbf{V}' \in \mathbb{R}^{N' \times d_v}$ as

$$\mathbf{V}' \sim \mathcal{U}(-\Delta, \Delta),$$

where $\Delta = \frac{1}{\sqrt{d_v}}$ and remove the regularization loss. For the clustering regularization, instead, we use $K = 10$ clusters and regularization trade-off weight $\lambda = 0.5$. We initialize embedding, centroid, and cluster assignment matrices as

$$\mathbf{V} \sim \mathcal{U}(-\Delta, \Delta) \quad \mathbf{C} \sim \mathcal{U}(-\Delta, \Delta) \quad \mathbf{S} \sim \mathcal{U}(0, 1)$$

respectively. For fine-tuning, we fix the centroid table \mathbf{C} and initialize the new embedding table $\mathbf{V}' \in \mathbb{R}^{N' \times d_v}$ and cluster assignment matrix $\mathbf{S}' \in \mathbb{R}^{N' \times K}$ following an analogous procedure. Finally, we increase the regularization weight λ to 10.

B Additional experimental results

Local components Table 8, an extended version of Table 1, shows the performance of reference architecture with and without local components. Here we consider TTS-IMP and TTS-AMP, together with FC-RNN (a multivariate RNN) and LocalRNNs (local univariate RNNs with a different set of parameters for each time series). For the STGNNs, we consider a global variant (without any local component) and global-local alternatives, where we insert node-specific components within the architecture by means of (1) different sets of weights for each time series (light brown block) or (2) node embeddings (light gray block). More precisely, we show how performance varies when the local components are added in the encoding and decoding steps together, or uniquely in one of the two steps. Results for TTS-AMP are consistent with the observations made for TTS-IMP, showing that the use of local components enables improvements up to 7% w.r.t. the fully global variant.

Transfer learning Tables 9 to 12 show additional results for the transfer learning experiments in all the target datasets. In particular, each table shows results for the reference architectures w.r.t. different training set sizes (from 1 day to 2 weeks) and considers the settings where embeddings are fed to both encoder and decoder or decoder only. We report results on the test data corresponding to the week after the validation set but also on the original test split used in the literature. In the last columns of the table, we also show the performance that one would have obtained 100 epochs after the minimum in the validation error curve; the purpose of showing these results is to hint at the performance that one would have obtained without holding out 1 week of data for validation. The results indeed suggest that fine-tuning the full global model is more prone to overfitting.

Table 9: Forecasting error (MAE) on PEMS03 in the transfer learning setting (5 runs average).

Model		Testing on 1 subsequent week				Testing on the standard split				Testing 100 epochs after validation min.			
TTS-IMP		2 weeks	1 week	3 days	1 day	2 weeks	1 week	3 days	1 day	2 weeks	1 week	3 days	1 day
Global		14.86±0.02	15.30±0.03	16.26±0.08	16.65±0.07	16.11±0.05	16.36±0.05	16.95±0.04	17.39±0.12	16.30±0.10	16.58±0.07	17.62±0.23	18.33±0.32
ENC+DEC.	Embeddings	14.53±0.02	14.64±0.05	15.87±0.08	16.78±0.12	16.03±0.05	16.12±0.05	17.18±0.16	17.82±0.15	16.09±0.06	16.16±0.05	17.28±0.18	17.94±0.17
	– Variational	14.50±0.04	14.56±0.03	15.40±0.06	15.65±0.11	15.69±0.10	15.70±0.12	16.33±0.06	16.52±0.10	15.70±0.10	15.70±0.12	16.35±0.07	16.54±0.10
	– Clustering	14.58±0.02	14.60±0.02	15.67±0.08	16.53±0.13	15.65±0.07	15.71±0.08	16.76±0.15	17.76±0.15	15.70±0.05	15.74±0.07	16.80±0.14	17.78±0.14
DEC.	Embeddings	14.79±0.02	14.84±0.03	15.49±0.03	16.01±0.08	16.06±0.05	16.12±0.07	16.74±0.04	17.25±0.07	16.08±0.05	16.13±0.07	16.75±0.04	17.29±0.06
	– Variational	15.33±0.03	15.39±0.02	15.83±0.04	16.03±0.04	16.15±0.02	16.20±0.02	16.60±0.04	16.75±0.06	16.15±0.02	16.20±0.02	16.60±0.04	16.76±0.06
	– Clustering	14.96±0.06	15.09±0.06	15.88±0.07	15.81±0.03	16.25±0.08	16.29±0.06	16.72±0.08	16.87±0.07	16.28±0.07	16.19±0.05	16.91±0.10	16.93±0.07

Table 10: Forecasting error (MAE) on PEMS04 in the transfer learning setting (5 runs average).

Model		Testing on 1 subsequent week.				Testing on the standard split				Testing 100 epochs after validation min.			
TTS-IMP		2 weeks	1 week	3 days	1 day	2 weeks	1 week	3 days	1 day	2 weeks	1 week	3 days	1 day
Global		20.86±0.03	21.59±0.11	21.84±0.06	22.26±0.10	20.79±0.02	21.68±0.10	22.10±0.10	22.59±0.11	20.89±0.05	21.97±0.13	22.88±0.17	23.85±0.22
ENC+DEC.	Embeddings	19.96±0.08	20.27±0.11	21.03±0.14	21.99±0.13	19.87±0.07	20.27±0.07	21.20±0.15	22.38±0.14	19.88±0.07	20.29±0.07	21.28±0.14	22.46±0.14
	– Variational	19.94±0.08	20.19±0.05	20.71±0.12	21.20±0.15	19.92±0.06	20.23±0.05	20.82±0.08	21.46±0.13	19.92±0.06	20.23±0.05	20.82±0.08	21.47±0.12
	– Clustering	19.69±0.06	19.91±0.11	20.48±0.09	21.91±0.21	19.70±0.06	19.96±0.11	20.62±0.09	22.28±0.21	19.72±0.07	19.97±0.10	20.65±0.08	22.29±0.21
DEC.	Embeddings	20.10±0.06	20.27±0.04	20.87±0.08	21.44±0.09	20.18±0.07	20.39±0.05	21.01±0.09	21.70±0.08	20.19±0.07	20.40±0.05	21.03±0.08	21.74±0.08
	– Variational	20.79±0.06	20.94±0.05	21.23±0.07	21.51±0.06	20.94±0.06	21.10±0.05	21.40±0.08	21.76±0.05	20.94±0.06	21.10±0.05	21.40±0.08	21.77±0.05
	– Clustering	20.19±0.09	20.45±0.10	20.63±0.06	21.03±0.05	20.27±0.09	20.56±0.10	20.81±0.06	21.28±0.06	20.75±0.05	20.78±0.05	20.89±0.05	21.35±0.05

Table 11: Forecasting error (MAE) on PEMS07 in the transfer learning setting (5 runs average).

Model		Testing on 1 subsequent week				Testing on the standard split				Testing 100 epochs after validation min.			
TTS-IMP		2 weeks	1 week	3 days	1 day	2 weeks	1 week	3 days	1 day	2 weeks	1 week	3 days	1 day
Global		22.87±0.05	23.82±0.03	24.52±0.06	25.40±0.06	22.64±0.04	23.58±0.02	24.20±0.06	25.04±0.06	22.72±0.05	23.72±0.02	24.45±0.11	25.48±0.05
ENC+DEC.	Embeddings	21.68±0.07	22.23±0.08	23.54±0.19	26.11±0.61	22.10±0.09	22.77±0.05	24.17±0.24	26.79±0.63	22.11±0.09	22.78±0.05	24.21±0.22	26.82±0.63
	– Variational	22.05±0.05	22.43±0.02	23.23±0.08	24.40±0.13	22.18±0.04	22.59±0.04	23.44±0.11	24.62±0.13	22.18±0.04	22.59±0.04	23.44±0.10	24.62±0.13
	– Clustering	21.75±0.05	22.16±0.07	23.36±0.20	26.44±0.26	22.03±0.08	22.52±0.10	23.85±0.24	27.12±0.27	22.03±0.09	22.55±0.11	23.85±0.24	27.13±0.28
DEC.	Embeddings	22.50±0.14	22.83±0.13	23.59±0.12	24.89±0.19	22.68±0.12	23.13±0.10	24.04±0.09	25.41±0.20	22.69±0.12	23.14±0.10	24.04±0.09	25.43±0.20
	– Variational	24.32±0.16	24.60±0.16	25.12±0.17	25.50±0.15	24.25±0.14	24.60±0.13	25.16±0.13	25.56±0.12	24.25±0.14	24.60±0.13	25.16±0.13	25.57±0.12
	– Clustering	23.02±0.09	23.53±0.09	24.42±0.15	24.87±0.13	23.18±0.09	23.77±0.08	24.66±0.12	25.24±0.09	23.91±0.16	24.10±0.15	24.73±0.10	25.27±0.09

Table 12: Forecasting error (MAE) on PEMS08 in the transfer learning setting (5 runs average).

Model		Testing on 1 subsequent week				Testing on the standard split				Testing 100 epochs after validation min.			
TTS-IMP		2 weeks	1 week	3 days	1 day	2 weeks	1 week	3 days	1 day	2 weeks	1 week	3 days	1 day
Global		15.51±0.03	15.90±0.07	16.87±0.05	17.59±0.04	15.36±0.03	15.71±0.06	16.71±0.06	17.41±0.03	15.47±0.06	15.87±0.04	17.58±0.16	18.46±0.12
ENC+DEC.	Embeddings	15.45±0.08	15.45±0.06	16.34±0.07	17.15±0.08	15.34±0.07	15.32±0.04	16.27±0.07	17.11±0.08	15.32±0.09	15.30±0.03	16.27±0.05	17.13±0.08
	– Variational	15.34±0.04	15.41±0.06	15.83±0.07	16.32±0.11	15.21±0.03	15.27±0.06	15.70±0.06	16.19±0.12	15.21±0.03	15.27±0.05	15.69±0.06	16.19±0.12
	– Clustering	15.41±0.06	15.41±0.06	15.96±0.04	16.99±0.07	15.27±0.07	15.27±0.07	15.90±0.05	16.99±0.08	15.30±0.08	15.28±0.07	15.90±0.05	17.02±0.10
DEC.	Embeddings	15.72±0.06	15.74±0.06	16.41±0.07	16.97±0.08	15.61±0.06	15.61±0.06	16.33±0.08	16.90±0.09	15.61±0.06	15.61±0.06	16.35±0.08	16.92±0.10
	– Variational	16.31±0.10	16.33±0.15	16.53±0.15	16.74±0.12	16.13±0.09	16.14±0.14	16.34±0.13	16.55±0.11	16.13±0.09	16.14±0.13	16.35±0.14	16.56±0.11
	– Clustering	15.81±0.11	15.92±0.14	16.11±0.08	16.55±0.10	15.70±0.10	15.77±0.11	15.97±0.08	16.43±0.10	15.98±0.06	15.90±0.06	16.01±0.08	16.45±0.11