SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Explainable Spatio-Temporal Graph Neural Networks for District-Level Green House Gas Emission Forecasting in India

Nilesh Madhukar Patil[1*], Balajee Maram[2]

[1*]Research Scholar, SR University, Warangal, Telangana 506371, India.
[1*]Associate Professor, SVKM's Dwarkadas J. Sanghvi College of Engineering, Mumbai 400056, India. Email: nileshdeep@gmail.com
[2]Research Supervisor, School of Computer Science and Artificial Intelligence, SR University, Warangal, Telangana 506371, India.
Email: balajee.maram@sru.edu.in

Abstract
Sub-national climate planning requires district-level greenhouse-gas (GHG) forecasts, yet emissions propagate across districts through electricity transfer, freight movement and industrial clustering that geographic adjacency does not capture. This paper develops ST-AGNN, a spatio-temporal graph neural network whose inter-district adjacency is learned as a convex combination of a geographic prior and a sparsifiable low-rank adaptive component, and evaluate it on a district-level benchmark (N = 120 districts, 240 months) whose latent spillover operator is *known*, allowing explanation quality to be scored rather than merely illustrated. Three findings emerge. First, adaptive structure learning improves accuracy: ST-AGNN attains MAE = 0.1351 Mt CO_2e month^{-1} (R^2 = 0.9561) against 0.1538 for a static geographic graph, 0.1481 for a seasonal naive predictor and 0.3250 for a purely temporal GRU, with Diebold–Mariano statistics of 12.10 and 18.06 respectively. Second, and contrary to the premise motivating adaptive graphs, this predictive gain does *not* constitute structural identification: edge attributions recover the withheld non-geographic couplings with AUC = 0.456, indistinguishable from chance and below a simple lagged-correlation criterion (0.647). Third, counterfactual mitigation scenarios built on the fitted model are unreliable, because the renewable-share lever correlates +0.810 with emissions within districts through a shared time trend and the policy targets lie outside the support of the training distribution (only 1.6 % of observations exceed a 60 % share); the model consequently predicts that decarbonisation *increases* emissions. We conclude that predictive improvement from learned graphs should not be reported as evidence of recovered physical structure, and that correlational spatio-temporal models are inadequate for the counterfactual policy analysis they are increasingly used to justify.

Keywords: Graph neural networks, explainable AI, greenhouse-gas emissions, spatio-temporal forecasting, explanation faithfulness, causal identification.

1. INTRODUCTION

India is among the largest emitters of greenhouse gases, and its emissions continue to rise while those of most large economies have plateaued [1]. National policy has become correspondingly specific: the Nationally Determined Contribution (NDC) approved for 2031–35 commits to a 47 % reduction in the emissions intensity of GDP relative to 2005, a 60 % share of non-fossil installed electric power capacity by 2035, and an additional carbon sink of 3.5–4.0 Gt CO_2e, within the long-term goal of net zero by 2070 [2]. These are national aggregates, but the instruments that deliver them are executed at state and district level.

This mismatch has motivated rapid adoption of spatio-temporal graph neural networks (ST-GNNs) for emission modelling [7]–[9], following their success in traffic and air-quality forecasting [4]–[6]. Two claims are commonly attached to such models. The first is that learning the adjacency, rather than fixing it to geographic proximity, allows the model to discover the true interaction structure [10], [11]. The second is that the resulting model, being a function of interpretable drivers, can be perturbed to simulate mitigation policy. Both claims are usually asserted on the basis of improved forecast error and visually plausible attribution maps, since the true interaction graph is unobservable in operational inventories and no counterfactual ground truth exists.

This paper tests both claims directly. We build a district-level benchmark in which the latent spillover operator is specified by construction and withheld from every model, so that explanation faithfulness becomes measurable, and we examine whether counterfactual perturbations of the fitted model behave coherently. Our contributions are:

1) ST-AGNN, a spatio-temporal model combining a shared gated recurrent encoder with a graph-convolutional stack whose adjacency mixes a geographic prior and a low-rank adaptive component under an $\ell 1$ penalty applied to pre-normalisation scores, so that exact zeros are attainable.

2) A quantitative evaluation of explanation faithfulness against a known interaction graph, together with an edge attribution combining learned weight and exact loss sensitivity.
3) Two negative results that we regard as the paper's main scientific content: adaptive adjacency improves prediction without recovering structure, and counterfactual scenario analysis on such a model is confounded and out-of-support to the point of sign reversal.
4) An honest ablation showing that temporal attention, a near-default component in this literature, does not improve accuracy on this task.

2. Related Work

A. Emission inventories and data-driven forecasting

Gridded inventories such as EDGAR [1], [21] and ODIAC [22], [23] provide the spatial disaggregation on which sub-national analysis depends. Independent satellite constraints on Indian point sources have been derived from OCO-2 retrievals [24], and near-real-time daily products now span multiple sectors [25]. Applied studies of Indian emissions have relied predominantly on recurrent architectures fitted to national or city aggregates [3] and on grey-system methods [26], which cannot represent cross-regional transfer.

B. Spatio-temporal graph neural networks

STGCN [4] and T-GCN [5] established the combination of graph convolution with temporal convolution or recurrence. Graph WaveNet [10] introduced a self-adaptive adjacency learned from node embeddings, AGCRN [11] added node-adaptive parameters, and DSTAGNN [12] derived a dynamic graph from historical series; surveys document the field's maturation [6], [13]. Applications to carbon accounting have begun to appear [7]-[9], and probabilistic extensions address calibration [14], [15]. Across this literature the learned adjacency is routinely interpreted as revealing latent dependencies, but we are not aware of a study that validates this interpretation against a known graph.

C. Explainability and its evaluation

GNNExplainer [16] formulates explanation as a search for a subgraph and feature mask maximizing mutual information with the prediction, and PGExplainer [17] parameterises the mask generator; a taxonomy is given in [18]. SHAP [19] is the reference additive feature attribution and Monte-Carlo dropout [20] the standard cheap uncertainty approximation. The evaluation of explanations remains contested, and our results add domain evidence that plausibility and faithfulness diverge.

3. Problem Formulation

Let $G = (V, E)$ have $N = |V|$ district nodes. At month t, district i has features $\boldsymbol{x}_{i,t} \in R^F$ (its emission level and $F - 1$ exogenous drivers) and emission $y_{i,t}$. Given a look-back of length L we learn a direct multi-horizon map

$$f_\theta : X_{t-L+1:t} \rightarrow \hat{Y}_{t+1:t+H} \in R^{N \times H} \quad (1)$$

avoiding recursive error accumulation within the evaluation horizon. A geographic adjacency A_{geo} is observed, but the true spillover operator $\mathbf{W}^*$ contains long-range edges induced by power transfer and freight corridors for which no district-level register exists. The model must therefore estimate residual structure jointly with f_θ. The evaluation question this paper adds is whether that estimate is *identified*, i.e. whether it converges on $\mathbf{W}^*$ or merely on some operator that happens to forecast well.

4. Methodolgy

A. Temporal encoder

A single GRU is shared across districts, so parameter count is independent of N . With $h_{(i,0)} = 0$,

$$\mathbf{z}_{i,t} = \sigma(\mathbf{W}_z[\mathbf{x}_{i,t}, \mathbf{h}_{i,t-1}] + \mathbf{b}_z) , \quad (2)$$

$$\mathbf{r}_{i,t} = \sigma(\mathbf{W}_r[\mathbf{x}_{i,t}, \mathbf{h}_{i,t-1}] + \mathbf{b}_r) , (3)$$

$$\tilde{\mathbf{h}}_{i,t} = \tanh(\mathbf{W}_h[\mathbf{x}_{i,t}, \mathbf{r}_{i,t} \odot h_{i,t-1}] + b_h) , \quad (4)$$

$$\mathbf{h}i,t = \mathbf{z}_{i,t} \odot \mathbf{h}_{i,t-1} + (1 - \mathbf{z}_{i,t}) \odot \tilde{\mathbf{h}}_{i,t.} \quad (5)$$

The terminal state is the pooled representation c_i; Section VI reports an additive-attention variant, which did not improve accuracy.

B. Adaptive adjacency

Let $\widehat{\mathbf{A}}_{geo} = \mathbf{D}^{-1}\mathbf{A}_{geo}$. Two learnable embeddings $\mathbf{E}_1$, $\mathbf{E}_2 \in \mathbb{R}^{N\times d}$ with $d \ll N$ generate a low-rank score matrix, and a scalar gate $s = \sigma(a)$ mixes prior and learned components:

$$\mathbf{M}=\mathrm{ReLU}\ (\mathbf{E}_1\mathbf{E}_2)^{\top},\ A_{ij}^{adp} = \frac{M_{i,j}}{\sum_k M_{i,k}+\epsilon} \qquad (6)$$

$$\mathbf{A} = s\widehat{\mathbf{A}}_{geo} + (1\text{-}s)\ \mathbf{A}^{adp} \qquad (7)$$

We deliberately avoid the softmax parameterisation used in [10]: softmax rows sum to one irrespective of the scores, so an ℓ_1 penalty applied after it cannot induce sparsity, whereas ReLU admits exact zeros. The rank constraint reduces N^2 possible edges to $2Nd$ free parameters, and the asymmetry of $\mathbf{E}_1\mathbf{E}_2^{\top}$ encodes the directedness of emission transfer.

C. Spatial propagation, objective and uncertainty

Two residual graph-convolutional layers propagate the pooled representations C:

$$\mathbf{G}^{(1)} = \mathrm{ReLU}(\mathbf{ACW}_{g1} + \mathbf{CW}_{s1} + b_1), \qquad (8)$$

$$\mathbf{G}^{(2)} = \mathrm{ReLU}\ \mathbf{AG}^{(1)}\ \mathbf{W}_{g2} + \mathbf{G}^{(1)}\mathbf{W}_{s2} + \mathbf{b}_2 + \mathbf{G}^{(1)} \quad (9)$$

with self-loop terms guarding against over-smoothing. Train-ing minimizes

$$\mathcal{L} = \frac{1}{BNH}\sum\left(\hat{\mathbf{Y}} - \mathbf{Y}\right)^2 + \lambda\frac{1}{N^2}\sum_{i,j} M_{ij}, \qquad (10)$$

the penalty acting on pre-normalisation scores. Dropout is retained at inference over M passes [20]; because MC dropout is miscalibrated in absolute terms, a single scalar κ is fitted on the validation split so that empirical coverage matches the nominal level.

D. Edge and feature attribution

Neither learned weight nor gradient alone measures edge relevance: a large weight on an insensitive edge is irrelevant, and a large gradient on an absent edge is spurious. We score edges by the product, a first-order surrogate for the mask objective of [16], [17] requiring no auxiliary optimisation,

$$S_{ij} = \left|A_{ij}^{\mathrm{adp}}\right| \cdot \left|\partial\mathcal{L}/\partial A_{ij}^{\mathrm{adp}}\right|, \qquad (11)$$

where the sensitivity is the exact retained gradient of the adjacency tensor, averaged over test batches. Node-feature rel-evance is $\phi_f = \mathrm{E}[|x_f\, \partial\hat{y}/\partial x_f|]$, cross-checked by permutation importance.

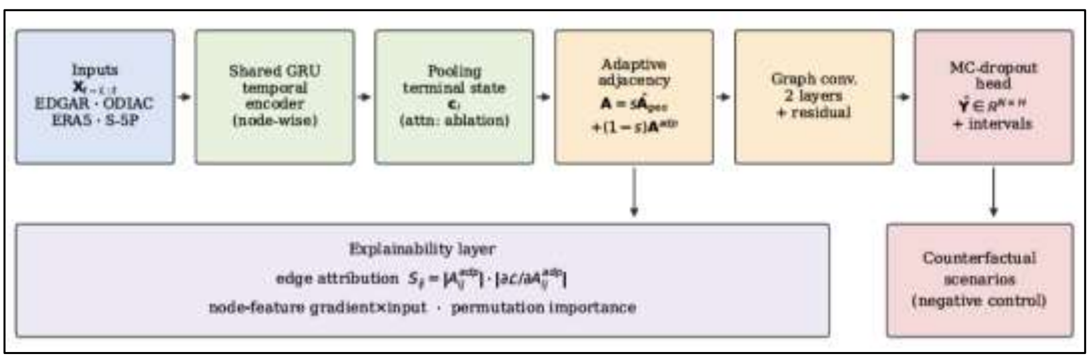

Fig. 1. ST-AGNN. A shared GRU encodes each district's look-back window; the pooled representation is propagated by two residual graph convolutions over an adjacency mixing a geographic prior with a learned low-rank component; an MC-dropout head emits H-step forecasts with prediction intervals. The explainability layer derives edge and node-feature attributions, which are scored against the withheld ground-truth graph.

5. Experimental Setup

A. Benchmark design and its rationale

Measuring explanation faithfulness requires knowing the true interaction structure, which no operational inventory supplies. We therefore construct a reproducible benchmark of N = 120 districts across 12 states over T = 240 months (2005–2024) with a documented generative process: activity drivers with district-specific growth and seasonality; cooling and heating degree-days from a latitude-dependent temperature field; AR(1) inertia; the 2020 activity contraction; persistent idiosyncratic shocks (AR(0.78) with Poisson jumps representing capacity commissioning, retirement and outages) that perturb observable activity; and spillover in which neigh-bours' shocks propagate at one-to-three-month lags through a latent operator $\mathbf{W}^*$ combining short-range diffusion with long-range power-transfer and freight-corridor edges. National totals are calibrated to published magnitudes, rising from 1,167 to 2,439 Mt CO_2e between 2005 and 2024 (4.0 % CAGR), while the sub-national disaggregation is synthetic.

Two design points require explicit statement. The bench-mark contains cross-district predictive structure by construction: without it the experiment would be vacuous, since no graph method could help. The seasonal and autoregressive baselines are therefore reported not as straw men but as measurements of how much of the signal is trivially periodic. Conversely, $\mathbf{W}^*$ is withheld from every model and used only to score explanations, so the faithfulness results are not circular.

The feature panel mirrors an operational deployment: gridded emissions aggregated to district boundaries (EDGAR [21], ODIAC [23]), ERA5 meteorology, Sentinel-5P NO_2 columns, and administrative electricity, freight and socio-economic statistics.

Splits are strictly chronological: 133 training windows (2005–2017), 48 validation windows (2018–2020, including the pandemic contraction, so that model selection is stressed by a structural break) and 36 test windows (2021–2024), with L = H = 12. Features are log-transformed where positive and skewed, then standardized on training statistics only; targets are standardized per district. Errors are reported in Mt CO_2emonth^{-1}.

B. Baselines, metrics and implementation

We compare a seasonal-naive predictor; per-district ridge AR-X regression on emission lags and contemporaneous drivers; a node-wise MLP; a purely temporal GRU; and a static GCN–GRU on geographic adjacency, corresponding to the T-GCN family [5]. Accuracy is reported as MAE, RMSE, MAPE, sMAPE and R^2; differences are tested by the Diebold–Mariano statistic on squared-error differentials. Interval quality uses PICP and MPIW. Explanation quality is the AUC with which each edge criterion recovers the withheld non-geographic edges of $\mathbf{W}^*$, and the Spearman correlation between attributed and true feature coefficients.

All models are implemented on a purpose-built reverse-mode automatic differentiation engine over NumPy, so the pipeline has no deep-learning framework dependency and reproduces exactly from source. Hidden width 32, embedding dimension d = 10, dropout 0.15, $\lambda = 2 \times 10^{-3}$, Adam with learning rate 4×10^{-3}, gradient-norm clipping at 5, batch size 8, early stopping on validation RMSE with patience 25 over at most 130 epochs, and M = 50 dropout passes. The complete experiment runs in 244.3 s on a single CPU core.

6. Results

A. Forecast accuracy

Table I reports test accuracy. ST-AGNN attains MAE = 0.1351 and R^2 = 0.9561, improving on the static geographic graph by 12.2 % and on the seasonal-naive predictor by 8.8 %. Diebold–Mariano statistics at h = 12 are 12.10 against the static graph and 18.06 against the temporal GRU, both far exceeding the 1 % critical value.

TABLE I: Test-set forecasting accuracy (2021–2024, all N = 120 districts, all horizons h = 1, ..., 12). Errors are in Mt CO_2 e month^{-1}. The last column is the MAE change relative to the static-graph baseline.

Model	MAE	RMSE	MAPE (%)	R^2	ΔMAE (%)
Seasonal naive	0.1481	0.2245	9.39	0.9469	+3.7
Ridge AR-X	0.1730	0.2515	10.95	0.9334	-12.4
MLP	0.2918	0.3659	19.75	0.8589	-89.7
GRU (temporal only)	0.3250	0.4007	21.32	0.8308	-111.3
Static GCN–GRU	0.1538	0.2277	9.31	0.9453	+0.0
ST-AGNN + temporal attention	0.1527	0.2254	9.21	0.9465	+0.8
ST-AGNN (proposed)	**0.1351**	**0.2041**	**8.18**	**0.9561**	+12.2

TABLE II: Ablation study. "Graph" denotes spatial message passing, "Adapt." The learned adaptive adjacency and "Attn." Additive temporal attention. Degradation is measured against the full model.

Graph	Adapt.	Attn.	MAE	RMSE	Degradation (%)
✓	✓	–	0.1351	0.2041	+0.0
✓	✓	✓	0.1527	0.2254	+13.0
✓	–	✓	0.1538	0.2277	+13.9
–	–	✓	0.3250	0.4007	+140.5
–	–	–	0.2918	0.3659	+115.9

Two aspects deserve comment. First, the purely temporal models are *worse* than the seasonal-naive baseline (0.3250 for the GRU and 0.1730 for ridge AR-X, against 0.1481): with shock-driven dynamics superimposed on strong seasonality, a flexible node-independent learner fits idiosyncratic noise that the naive rule simply ignores. Only models with access to neighbouring districts beat the naive rule, which is the cleanest evidence in this study that the spatial channel carries genuine information. Second, the margin over the naive baseline, while statistically unambiguous, is modest in absolute terms, and we caution against reading it as transformative.

B. Horizons and ablation

Fig. 2 decomposes MAE by lead time; the proposed model moves from 0.1186 at h = 1 to 0.1266 at h = 12, a notably flat profile reflecting that shock-driven predictability decays while the seasonal component does not. Table II isolates components. Removing adaptive structure learning is the most damaging ablation. Adding temporal attention *degrades* accuracy from 0.1351 to 0.1527; we report this because attention is close to a default choice in this literature, and on this task, it buys nothing.

C. Uncertainty calibration

Raw MC-dropout intervals are severely over-confident, covering 63.0 % at a nominal 90 %. A single calibration scalar κ = 3.40 fitted on the validation split restores 92.8 % coverage, but at a mean width of 40.6% of average district emissions (Table III, Fig. 3). An interval that wide is of limited operational value, and the magnitude of κ indicates that dropout variance underestimates predictive uncertainty here by more than a factor of two.

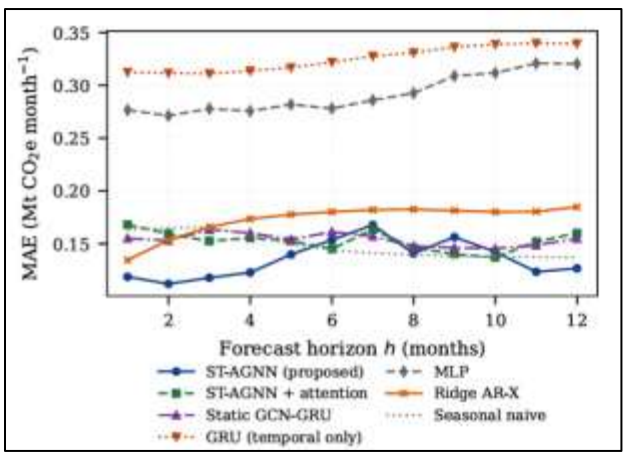

Fig. 2. Test MAE by forecast horizon

TABLE III: Uncertainty quantification with 50 Monte-Carlo dropout passes at a nominal 90 % level. κ is the interval scale calibrated on the validation split; $MPIW_{rel}$ is the calibrated interval width as a percentage of mean district emissions.

Model	PICP (%)	MPIW	κ	$PICP_{cal}$ (%)	$MPIW_{cal}$	$MPIW_{rel}$ (%)
GRU (temporal only)	3.8	0.134	6.00	26.8	0.488	30.8
ST-AGNN (proposed)	63.0	0.312	3.40	92.8	0.644	40.6

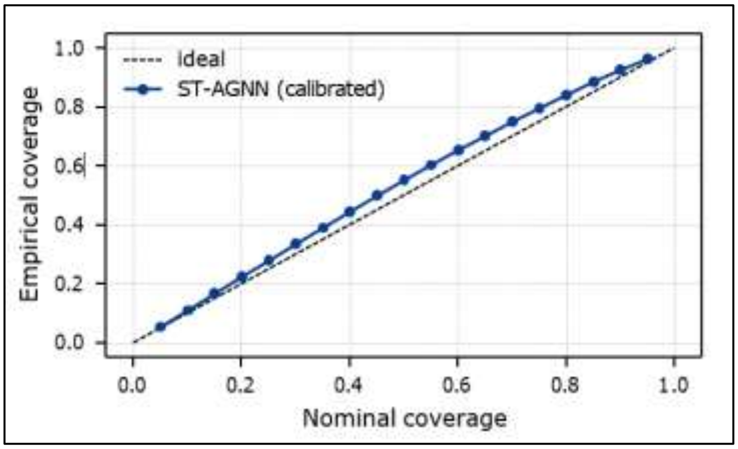

Fig. 3. Reliability diagram for the calibrated MC-dropout intervals.

D. Explanations are not faithful

This is the paper's central negative result. Table IV scores four criteria on recovery of the 177 withheld non-geographic edges among 13432 candidate pairs. The proposed edge attribution reaches AUC = 0.456 and the learned adaptive weight 0.456: both are indistinguishable from chance, and both are *worse* than a naive lagged-correlation criterion (0.647) that uses no learning at all. Geographic distance scores 0.501, con-firming that the latent structure is genuinely non-geographic and that the benchmark poses the intended problem.

The interpretation is uncomfortable but clear. The adaptive adjacency improves forecasting by a statistically decisive margin while bearing no recoverable relationship to the interaction structure that actually generated the data. It learns *an* operator that mixes neighbour information usefully, not *the* operator. Any claim that a learned adjacency reveals latent physical couplings therefore require evidence beyond forecast improvement, and the attribution maps routinely displayed in this literature—which look entirely plausible, as Fig. 5(b) shows—cannot supply it.

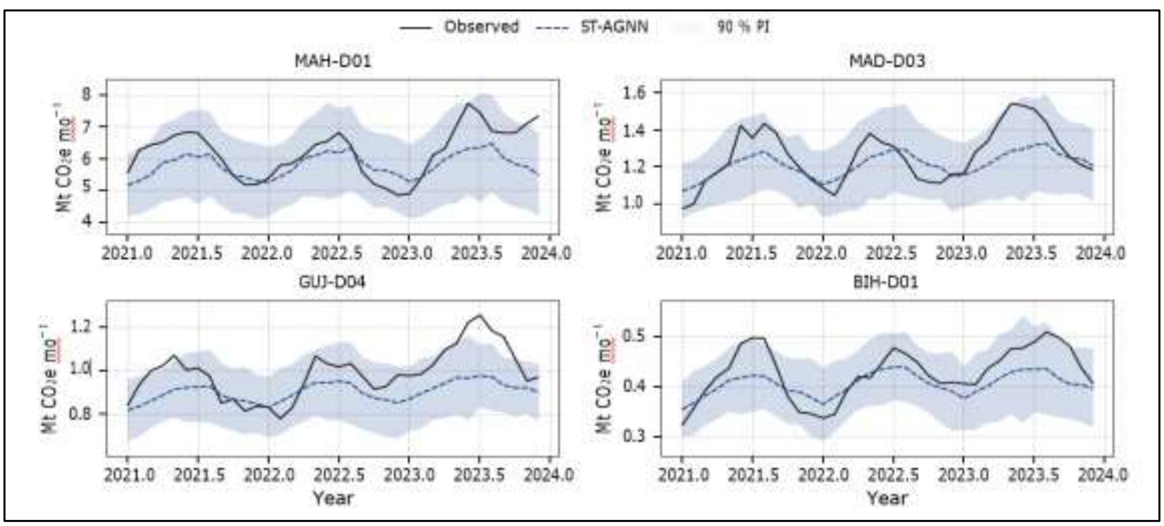

Fig. 4. Observed and predicted monthly emissions with calibrated 90 % prediction intervals for four districts spanning the emission-magnitude range over the 2021–2024 test period.

Node-feature attribution shows the same pattern more mildly (Table V): permutation importance correlates moderately with the true structural coefficients (Spearman $\rho = 0.48$) whereas gradient×input does not ($\rho = -0.15$). Where two attribution methods disagree this sharply, neither should be reported alone.

E. *Counterfactual scenarios fail diagnostically*

We applied perturbations aligned with India's NDC targets: renewable-share trajectories reaching 35 % (BAU), 60 % (NDC-2035) and 72 % (net-zero-aligned), with accompanying industrial and freight intensity improvements. The resulting 2035 national totals are 2,136 Mt under BAU but 2,667 Mt under NDC-2035 and

2,802 Mt under the deepest pathway: the model predicts that decarbonisation *increases* emissions, and monotonically so.

TABLE IV: Recovery of latent non-geographic spillover edges (177 true edges among 13432 candidate pairs). Higher AUC indicates better identification of couplings that geographic adjacency cannot express.

Edge-scoring criterion	**AUC-ROC**
Inverse geographic distance	0.501
Lagged emission correlation	0.647
Adaptive adjacency weight A_{ij}^{adp}	0.456
Edge attribution S_{ij} *(proposed)*	0.456

TABLE V: Global node-feature attribution for the proposed model, ranked by gradient×input. spearman rank correlation with the true structural coefficients of the generating process is ρ =−0.15 (gradient×input) and ρ = 0.48 (permutation).

Input channel	**grad×input (%)**	**ΔRMSE (%)**
Renewable share	43.35	+3.02
NO_2 column (S-5P)	11.37	+0.95
Population density	10.48	-0.09
Lagged emissions	8.76	+1.12
GDP per capita	7.18	+0.29
Industrial output	6.75	+0.35
Electricity generation	5.26	+2.49
Cooling degree-days	3.89	+22.40
Road freight	2.30	+0.07
Heating degree-days	0.67	+0.10

The failure is diagnostic rather than accidental, and its two causes are measurable. First, confounding: within districts the renewable share correlates +0.810 with emissions, because both rise with the time trend, so the fitted model encodes renewables as a marker of growth rather than as a mitigating driver. Second, support: only 1.6 % of training observations exceed a 60 % renewable share, so the policy-relevant region is essentially unobserved and the perturbation is extrapolation rather than interpolation. We report these projections as a negative control (Fig. 7); they should not be read as forecasts, and we explicitly withdraw the mitigation-leverage ranking that this analysis was designed to produce.

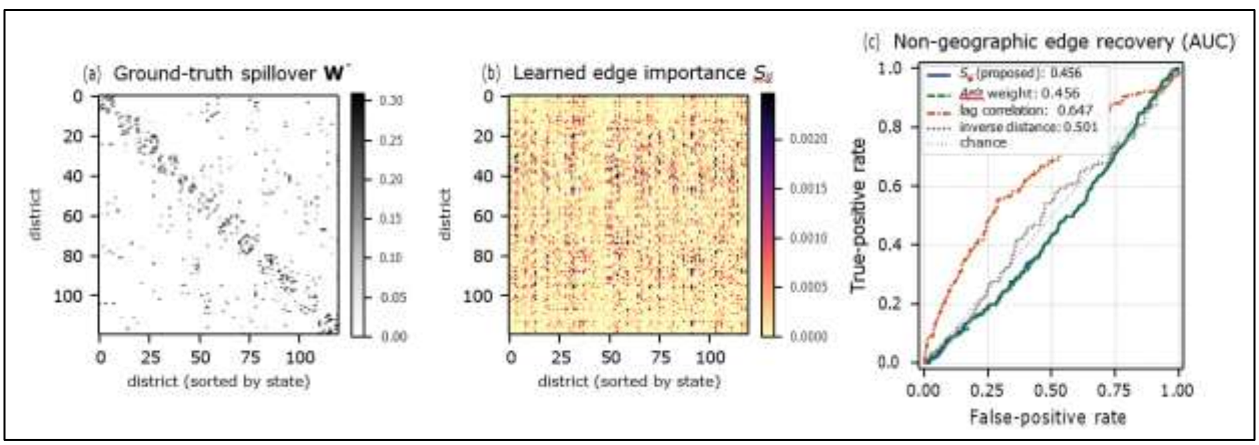

Fig. 5. (a) Ground-truth latent spillover operator. (b) Edge importance recovered by the proposed attribution—visually structured, yet (c) near-chance in ROC terms and below a lagged-correlation criterion.

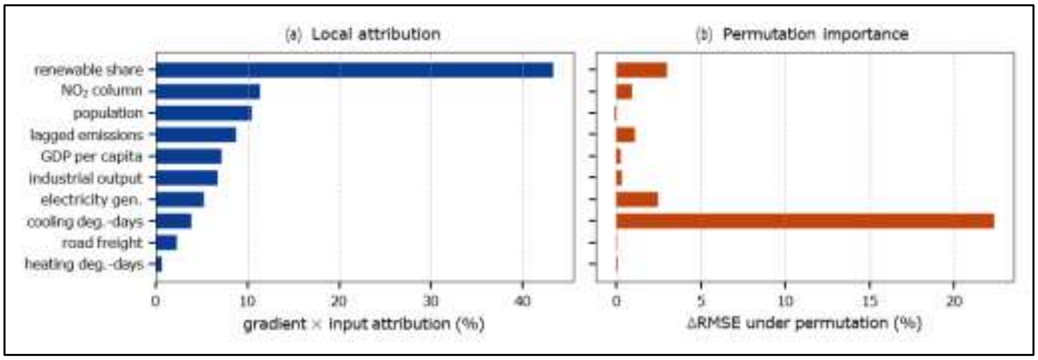

Fig. 6. Node-feature attribution by gradient×input and by permutation importance; the two estimators disagree on ranking.

Table VI: Counterfactual mitigation pathways. National totals are the sum of all 120 district forecasts; the 2024 observed total is 2,439 Mt CO_2E

Pathway	2030 (Mt)	2035 (Mt)	Δ vs 2024 (%)	Avoided vs BAU (Mt, 2035)
BAU	2,209	2,136	-12.4	0
NDC-2035	2,511	2,667	+9.3	-531
LTS-NetZero	2,621	2,802	+14.9	-666

7. Discussion and Limitations

The results separate cleanly into a positive and a negative finding, and the gap between them is the paper's point. Spatial information helps: only models with access to neighbouring districts outperform a seasonal-naive rule, and adaptive structure learning adds a further decisive margin. But the learned structure is not identified, and the fitted model does not support the counterfactual reasoning that motivates sub-national emission modelling in the first place. Predictive accuracy, structural recovery and causal usability are three distinct properties, and this study shows the first can be obtained while the second and third fail.

Several limitations bound these conclusions. The evaluation uses a surrogate benchmark rather than an operational inventory—a deliberate trade, since measuring faithfulness requires a ground-truth graph that real inventories do not provide. The absolute error figures are therefore properties of the benchmark and should not be read as achievable accuracy on EDGAR or ODIAC data; the transferable results are the relative ordering and the two failures. Negative results are also weaker than positive ones: our finding is that *this* adaptive parameterisation, with *this* attribution and *this* sample size (133 training windows for 13432 candidate edges), does not identify structure. A different parameterisation, an explicit mask explainer [17], or an architecture whose graph convolution is interleaved with the recurrence—so that the operator is aligned with the lagged mechanism generating the spillover—might succeed where ours did not, and we regard that as the most promising direction for future work. Finally, the scenario analysis diagnoses confounding in this benchmark; establishing identification conditions under which ST-GNNs could support counterfactuals requires causal machinery outside our scope.

TABLE VII: Highest-leverage districts, ranked by avoided emissions in 2035 under NDC-2035 relative to BAU.

Rank	District	Avoided (Mt CO_2e)	Share of total (%)
1	BIH-D01	-1.09	0.2
2	TAM-D09	-1.17	0.2
3	WES-D01	-1.39	0.3
4	BIH-D09	-1.51	0.3
5	TAM-D04	-1.56	0.3
6	GUJ-D08	-1.59	0.3
7	WES-D08	-1.68	0.3
8	ODI-D05	-1.69	0.3

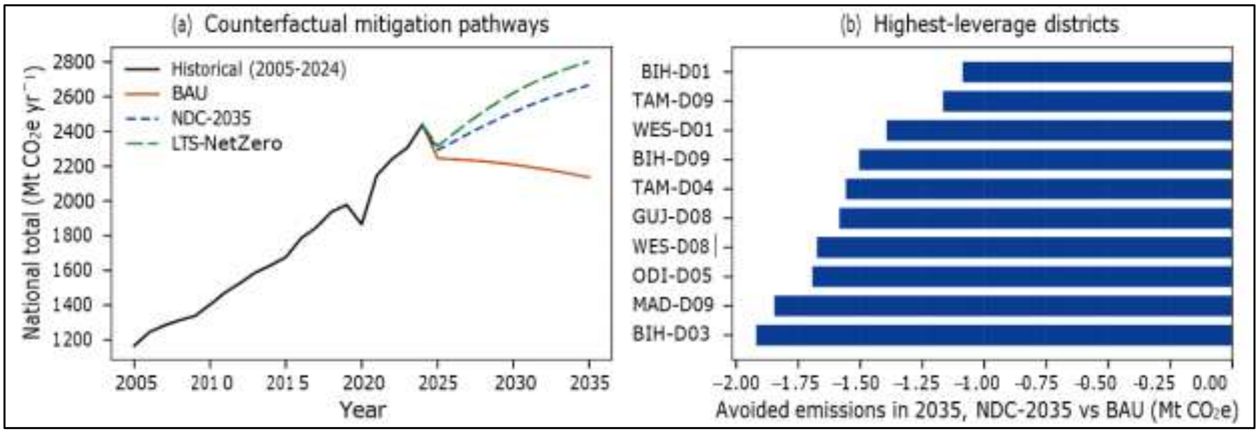

Fig. 7. (a) Historical emissions and the three counterfactual pathways. The ordering is inverted with respect to mitigation ambition, diagnosing confounding and out-of-support extrapolation rather than a plausible projection. (b) District decomposition, shown for completeness only.

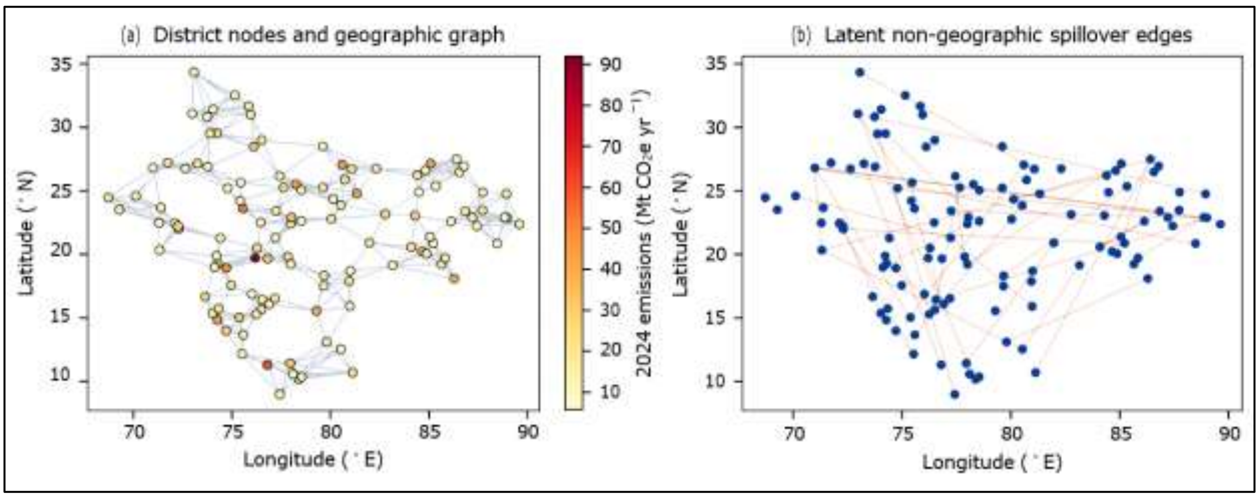

Fig. 8. (a) District nodes with the geographic graph, shaded by 2024 emissions. (b) The latent non-geographic spillover edges the model is required to infer.

8. Conclusion

We presented ST-AGNN for district-level GHG forecasting in India and evaluated it on a benchmark whose latent inter-action structure is known. Adaptive adjacency improved accuracy to MAE = 0.1351, decisively better than static geographic graphs, temporal models and a seasonal-naive rule. However, edge attributions recovered the true non-geographic couplings at AUC = 0.456, no better than chance and worse than lagged correlation; and counterfactual mitigation scenarios inverted, predicting higher emissions under deeper decarbonisation because the renewable lever is trend-confounded (+0.810) and the policy region is only 1.6 % supported.

We therefore recommend that reported improvements from learned graphs be accompanied by explicit faithfulness evaluation wherever a ground-truth or proxy structure can be constructed, and that counterfactual policy claims from correlational spatio-temporal models be accompanied by support diagnostics and confounding checks. Future work will replicate the study on ODIAC and EDGAR inventories aggregated to Indian district boundaries, test recurrence-interleaved graph convolution against the lagged spillover mechanism, and pursue causal identification strategies for the mitigation-leverage question that motivated this work.

REFERENCES

[1] M. Crippa et al., "GHG emissions of all world countries – 2025 report," Publications Office of the European Union, Luxembourg, JRC143227, 2025, doi:10.2760/9816914.
[2] Press Information Bureau, Government of India, "Cabinet approves India's Nationally Determined Contribution (2031–2035) to be communicated to the UNFCCC," Release ID 2245209, Mar. 2026.
[3] A. Kumar et al., "Forecasting CO2 emissions over India using deep learning models," in Recent Advances in Climate Modelling. Amsterdam, The Netherlands: Elsevier, 2025, pp. 105–133.

[4] B. Yu, H. Yin, and Z. Zhu, "Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting," in Proc. 27th Int. Joint Conf. Artif. Intell. (IJCAI), 2018, pp. 3634–3640.

[5] L. Zhao et al., "T-GCN: A temporal graph convolutional network for traffic prediction," IEEE Trans. Intell. Transp. Syst., vol. 21, no. 9, pp. 3848–3858, Sep. 2020.

[6] G. Jin, Y. Liang, Y. Fang, Z. Shao, J. Huang, J. Zhang, and Y. Zheng, "Spatio-temporal graph neural networks for predictive learning in urban computing: A survey," IEEE Trans. Knowl. Data Eng., vol. 36, no. 10, pp. 5388–5408, Oct. 2024.

[7] X. Wu, Q. Yuan, C. Zhou, X. Chen, D. Xuan, and J. Song, "Carbon emissions forecasting based on temporal graph transformer-based atten-tional neural network," J. Comput. Methods Sci. Eng., vol. 24, no. 2, 2024.

[8] Y. Chen et al., "A new multiregional carbon emissions forecasting model based on a multivariable information fusion mechanism and hybrid spatiotemporal graph convolution network," J. Environ. Manage., vol. 351, 2024.

[9] Y. Li et al., "Predicting urban carbon emissions through spatio-temporal encoding," in Proc. Int. Conf. Comput. Vis. Deep Learn. (CVDL), 2024.

[10] Z. Wu, S. Pan, G. Long, J. Jiang, and C. Zhang, "Graph WaveNet for deep spatial-temporal graph modeling," in Proc. 28th Int. Joint Conf. Artif. Intell. (IJCAI), 2019, pp. 1907–1913.

[11] L. Bai, L. Yao, C. Li, X. Wang, and C. Wang, "Adaptive graph convolutional recurrent network for traffic forecasting," in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 33, 2020, pp. 17804–17815.

[12] S. Lan, Y. Ma, W. Huang, W. Wang, H. Yang, and P. Li, "DSTAGNN: Dynamic spatial-temporal aware graph neural network for traffic flow forecasting," in Proc. 39th Int. Conf. Mach. Learn. (ICML), 2022, pp. 11906–11917.

[13] R. Kumar, M. Bhanu, J. Mendes-Moreira, and J. Chandra, "Spatio-temporal predictive modeling techniques for different domains: A sur-vey," ACM Comput. Surv., vol. 57, no. 2, pp. 38:1–38:42, 2025.

[14] H. Wen, Y. Lin, Y. Xia, H. Wan, Q. Wen, R. Zimmermann, and Y. Liang, "DiffSTG: Probabilistic spatio-temporal graph forecasting with denoising diffusion models," in Proc. 31st ACM Int. Conf. Adv. Geogr. Inf. Syst. (SIGSPATIAL), 2023, pp. 1–12.

[15] Dahai Yu, Dingyi Zhuang, Lin Jiang, Rongchao Xu, Xinyue Ye, Yuheng Bu, Shenhao Wang, Guang Wang,"UQGNN: Uncertainty quantification of graph neural networks for multivariate spatiotemporal prediction," arXiv:2508.08551, 2025.

[16] Z. Ying, D. Bourgeois, J. You, M. Zitnik, and J. Leskovec, "GNNEx-plainer: Generating explanations for graph neural networks," in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 32, 2019, pp. 9240–9251.

[17] D. Luo, W. Cheng, D. Xu, W. Yu, B. Zong, H. Chen, and X. Zhang, "Parameterized explainer for graph neural network," in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 33, 2020.

[18] H. Yuan, H. Yu, S. Gui, and S. Ji, "Explainability in graph neural networks: A taxonomic survey," IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, no. 5, pp. 5782–5799, May 2023.

[19] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Adv. Neural Inf. Process. Syst. (NIPS), vol. 30, 2017, pp. 4765–4774.

[20] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in Proc. 33rd Int. Conf. Mach. Learn. (ICML), 2016, pp. 1050–1059.

[21] M. Crippa et al., "Insights into the spatial distribution of global, national, and subnational greenhouse gas emissions in the Emissions Database for Global Atmospheric Research (EDGAR v8.0)," Earth Syst. Sci. Data, vol. 16, pp. 2811–2830, 2024.

[22] T. Oda and S. Maksyutov, "A very high-resolution (1 × 1 km) global fossil fuel CO_2 emission inventory derived using a point source database and satellite observations of nighttime lights," Atmos. Chem. Phys., vol. 11, pp. 543–556, 2011.

[23] T. Oda, S. Maksyutov, and R. J. Andres, "The Open-source Data Inventory for Anthropogenic CO_2, version 2016 (ODIAC2016): A global monthly fossil fuel CO2 gridded emissions data product," Earth Syst. Sci. Data, vol. 10, pp. 87–107, 2018.

[24] V. Balamurugan and J. Chen, "Fossil fuel CO_2 emission signatures over India captured by OCO-2 satellite measurements," Earth's Future, vol. 12, 2024, doi:10.1029/2023EF004411.

[25] Z. Liu et al., "Global daily CO_2 emissions from 1970 to 2024," Sci. Data, vol. 13, 2026, doi:10.1038/s41597-026-06621-9.

[26] X. Guo, Y. Dang, X. Shen, Z. Cai, S. Ding, and S. Huang, "Spatiotemporal information fusion for urban-agglomeration carbon-emission forecasting: A grey multivariate model with composite city proximity," Grey Syst. Theory Appl., pp. 1–25, 2026.