Hydro-GRNNI: Hydrological Graph Recurrent Neural Network for Imputation

Shuo Han

Department of Statistics and Data Science

Northwestern

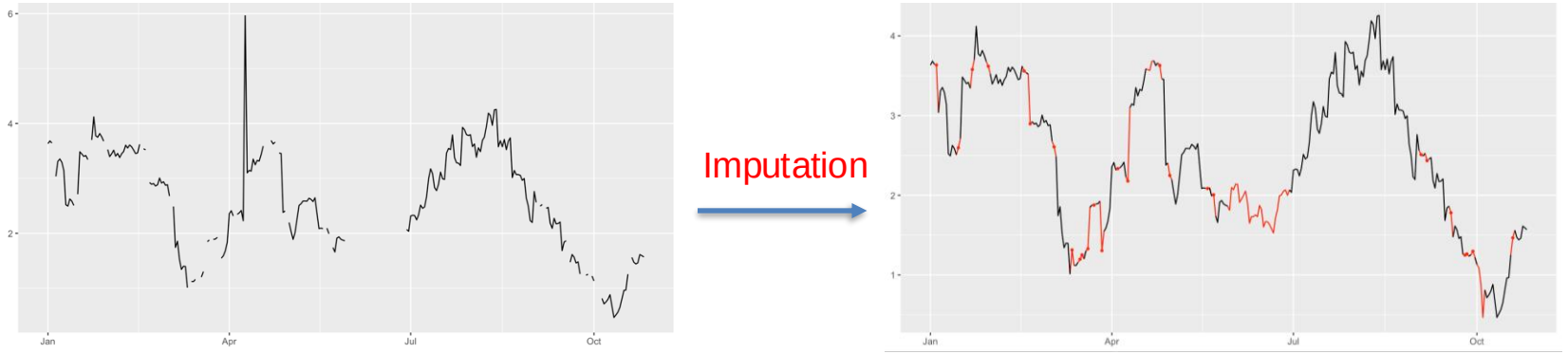
Problem Setting

Missing in Hydrological Data

- **Sparse Data**: Surface water quality observations in most US watersheds are spatiotemporally sparse, complicating water quality assessments and model calibration.
- **Limitations of Traditional Hydrological Tools**: Tools like USGS's EGRET and HydroClimATe overlook geographical relationships and struggle with missing data.

We need a spatio-temporal imputation method for hydrological data.

Time Series / Temporal Imputation

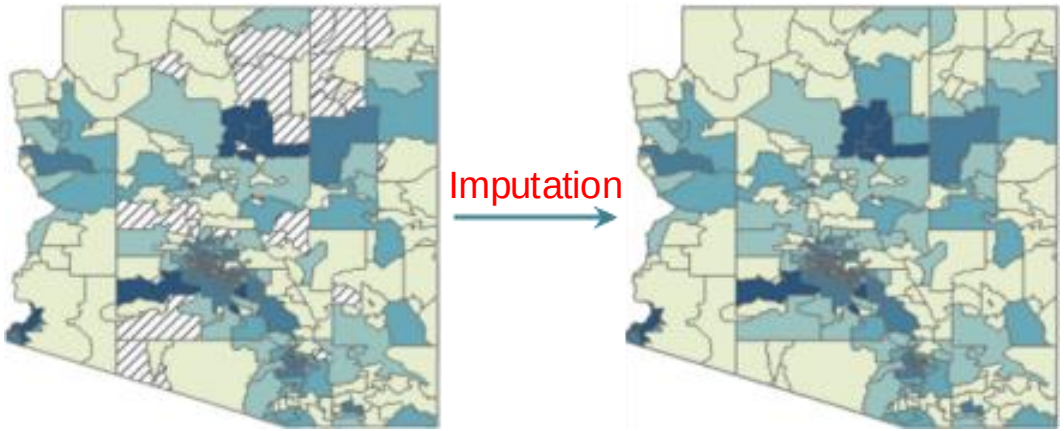


a. Original time series with missing values

b. Imputed time series

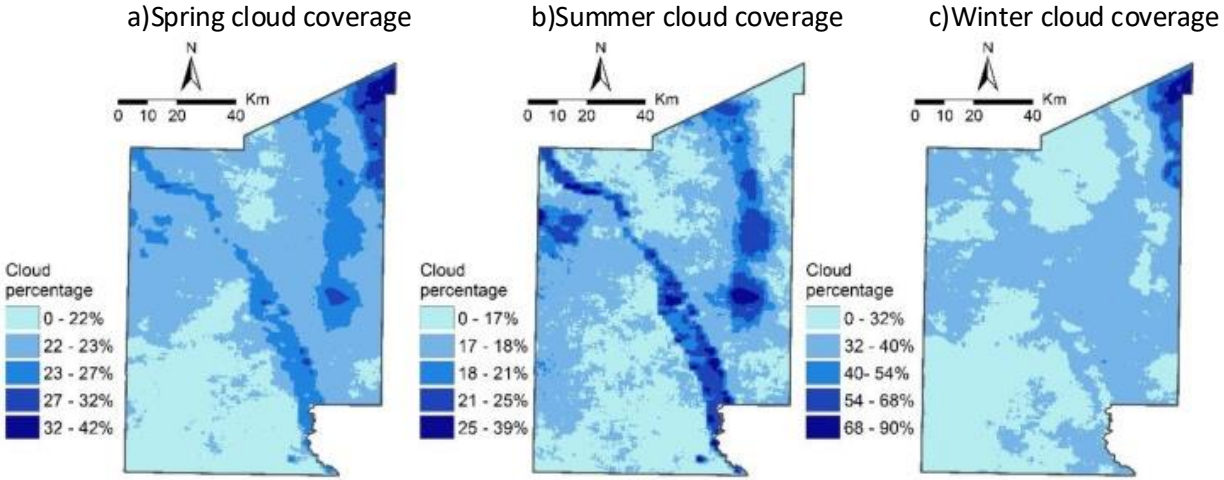
Time series imputation is the process of filling in missing values in time-ordered data.

Spatial Imputation



Spatial imputation refers to the process of estimating missing values in spatial data based on the information available from nearby locations or spatial relationships.

Spatio-temporal Imputation



Spatio-temporal imputation involves estimating missing values in datasets that have both spatial and temporal dimensions.

This approach is used when data is collected across different locations (spatial dimension) over time (temporal dimension), and some values are missing.

Methodology

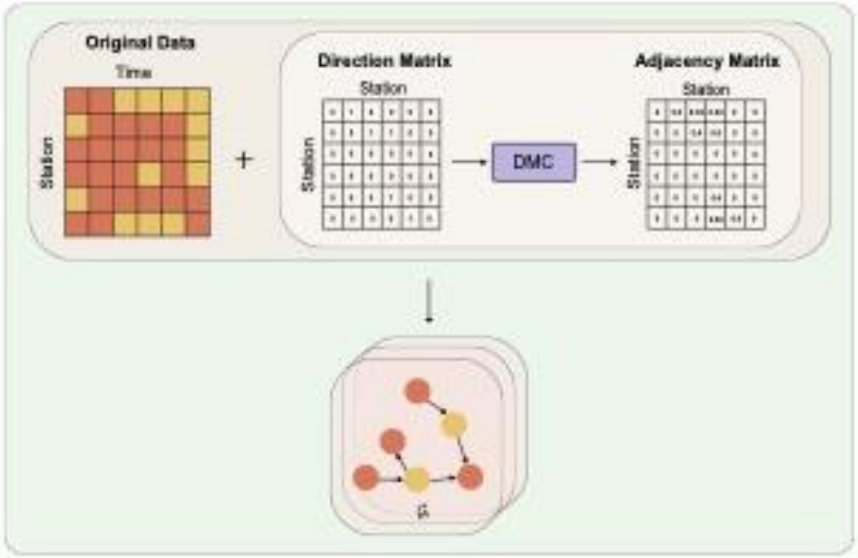
Graph Recurrent Neural Network

- **Recurrent Neural Network (RNN)**:
  - **Purpose**: Designed for sequence data processing.
  - **Mechanism**: Uses feedback loops to maintain information across time steps.

- **Graph Neural Network (GNN)**:
  - **Purpose**: Handles data with graph structures.
  - **Mechanism**: Propagates and aggregates information over nodes and edges.

- Graph Recurrent Neural Network works well in prior imputation tasks.

Graph Construction



Graph Construction

Graph at time *t*:

- **Nodes**: Capture the original data points at specific monitoring stations and times.
- **Node Connections**: Show directional flow, highlighting spatial relationships.

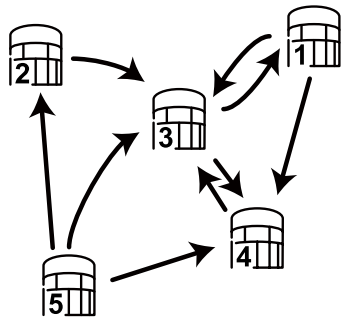
Sources of Adjacency Matrix Data

- **Traditional Imputation Method:** Geospatial coordinates of monitoring stations, including the latitude and longitude of monitoring stations.

- **Hydro-GRNNI:** Information about the upstream and downstream information between monitoring stations, capturing the flow relationships in the hydrological network.

Convert to Distance Matrix

Flows



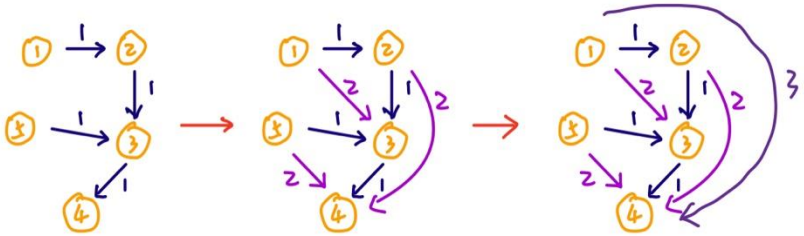
Flow Direction Matrix

origin \ destination	1	2	3	4	5
1	0	0	1	1	0
2	0	0	1	0	0
3	1	0	0	1	0
4	0	0	1	0	0
5	0	1	1	1	0

Distance Matrix

origin \ destination	1	2	3	4	5
1	0	∞	1	1	∞
2	2	0	1	2	∞
3	1	∞	0	1	∞
4	2	∞	1	0	∞
5	2	1	1	1	0

Distance Matrix Converter(DMC)



Initialization: Distance Matrix

- Set D[i][j] to F[i][j] for direct flows.
- Set D[i][j]=∞ if no direct flow and i≠j.
- Set D[i][j]=for i=j.

Iterative Update

- Intermediate Updates: Use: D[i][j]=min(D[i][j],D[i][k]+D[k][j])
- Path Evaluation: Update D[i][j] if a shorter path is found.

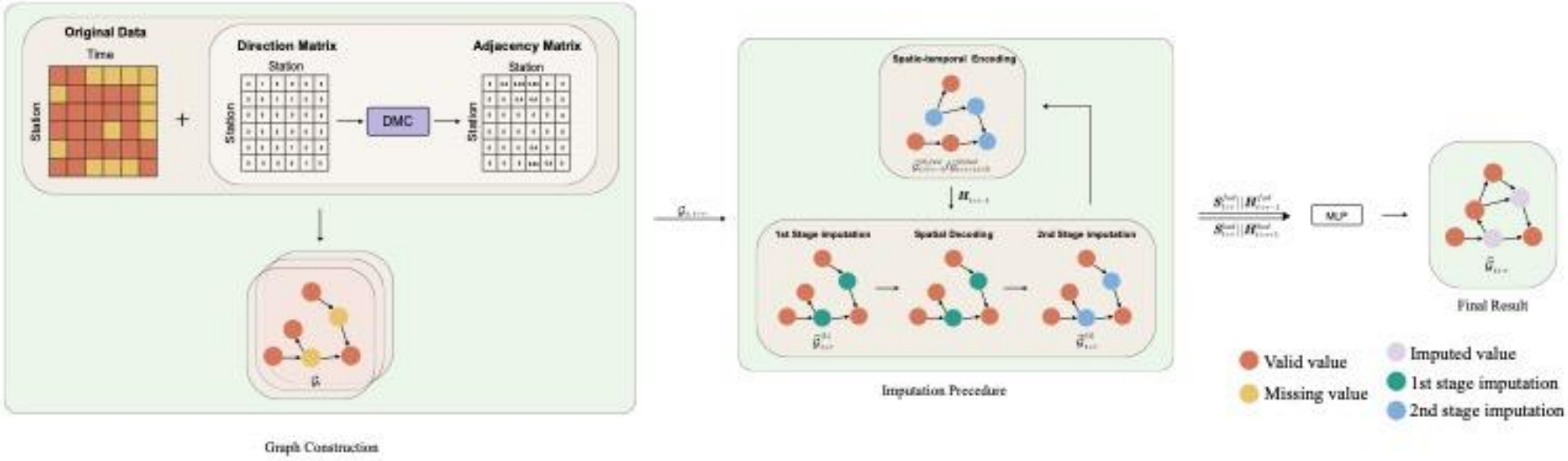
Convergence

- Stop when matrix D no longer changes.

Completion

- D represents the shortest paths between all stations.

Hydro-GRNNI(Hydrological Graph Recurrent Neural Network for Imputation)



Experiment

Hydrological Data

Dataset	Missing Rate(%)	# Stations	# Points in Time (Hourly)
Discharge	0	20	2811

Table 5.1: Details of the adopted dataset.

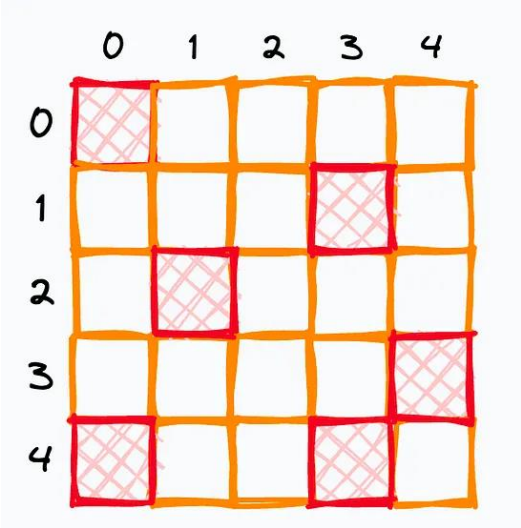
- The daily river flow sediment concentrations, measured in milligrams per liter (mg/l), were collected from 20 monitoring stations.
- These data were sourced from three organizations: the United States Geological Survey (USGS), the Water Quality Portal (WQP), and the National Center for Water Quality Research (NCWQR).
- The dataset spans from March 1, 2017, to September 30, 2022, providing comprehensive coverage over this period.

Adjacency Information Source

- **Flow Direction Information**
  - Derived from Hydro Network-Linked Data Index (NLDI) dataset.
  - used to represent the directional flow relationships between the stations.

- **Station Location Information**
  - From USGS, we have station's location information represented by its latitude and longitude coordinates

Point missing

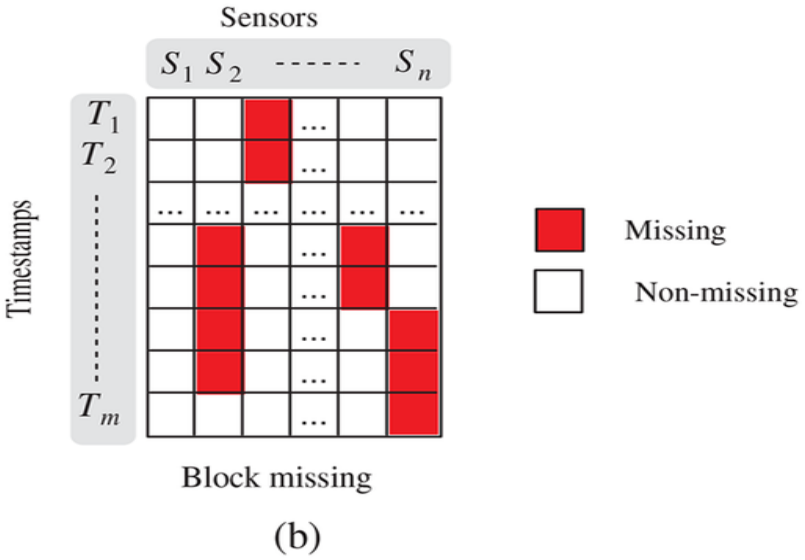


Data Masking Strategy: Implement random dropout of individual data points.

Point Missing Patterns: Introduce isolated missing data points throughout the dataset.

Objective: Simulate random, isolated instances of missing data to reflect real-world scenarios, such as brief sensor glitches or momentary communication disruptions.

Block missing



Data Masking Strategy: Simulate random dropout of data blocks.

Block Missing Patterns: Introduces contiguous missing sequences. sequence length: 12 to 48 time steps.

Objective: Simulate contiguous sequences of missing data to reflect real-world scenarios, such as prolonged sensor failures or systematic data transmission interruptions.

Data Setting

- **In-Sample Imputation:**
  Train the model on the full sequence, reflecting cases where the entire dataset is available for handling missing values.

- **Out-of-Sample Imputation:**
  Train and evaluate the model on separate sequences, simulating imputation for new target sequences using data it hasn’t seen before.

Baselines Included

Basic Statistical Methods:

- **Mean Imputation:** Fill missing values with the average value of the series.
- **K-Nearest Neighbors (KNN):** Impute missing values by averaging values from nearby observations based on an adjacency matrix.

Dynamic Approaches:

- **Vector AutoRegressive (VAR) Model:** Capture linear relationships between multiple time series.
- **rGAIN:** Use adversarial training with bidirectional Recurrent Neural Networks (RNNs) for robust imputation in an unsupervised setting.

Advanced Neural Network Models:

- **BRITS:** Model sequential dependencies in time series data using RNNs.
- **MPGRU:** Combine Graph Neural Networks (GNNs) and Gated Recurrent Units to capture spatial and temporal patterns for one-step-ahead predictions.
- **GRIN:** Use GNNs and RNNs to impute missing data in multivariate time series by reconstructing relationships and employing bidirectional processing for better temporal understanding.

Out-of-Sample Evaluation

		Point Missing			Block Missing		
D	M	MAE	RMSE	MRE(%)	MAE	RMSE	MRE(%)
Discharge	Mean	998.34	2469.02	1.12	1264.21	3781.88	0.93
	KNN	1013.43	2732.56	1.14	1540.31	4369.41	1.13
	VAR	746.56	2596.45	0.57	597.58	1977.11	0.46
	rGAIN	607.34	2088.86	0.47	578.59	2126.00	0.44
	BRITS	505.68	1646.39	0.39	494.53	1,627.71	0.38
	MPGRU	562.66	2339.57	0.43	625.12	2381.28	0.48
	GRIN	384.11	1639.11	0.29	504.77	1753.84	0.39
	Hydro-GRNNI	**246.01**	**910.06**	**0.19**	**327.84**	**1185.10**	**0.25**

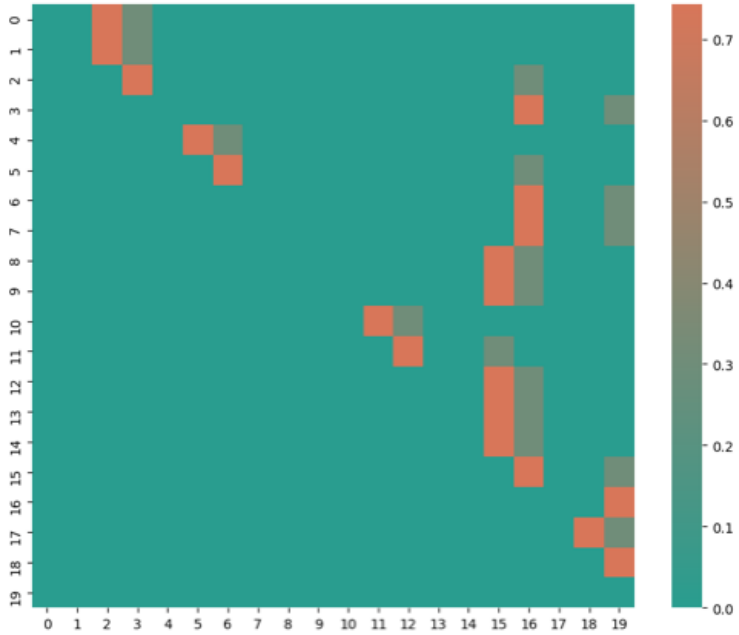
Table 6.1: Performance Results of Imputation Methods on Discharge Dataset for Out-of-Sample Evaluation

In-Sample Evaluation

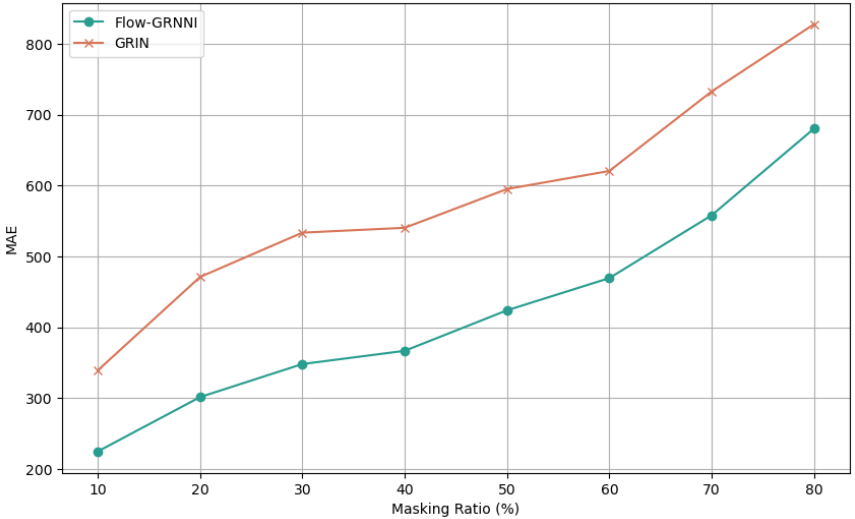
		Point Missing			Block Missing		
D	M	MAE	RMSE	MRE(%)	MAE	RMSE	MRE(%)
Discharge	Mean	967.04	2454.02	1.08	1235.57	3799.04	0.91
	KNN	1013.43	2732.56	1.14	1540.31	4369.41	1.13
	VAR	244.71	1040.65	0.18	257.19	1069.80	0.19
	rGAIN	171.43	576.50	0.13	199.94	687.44	0.15
	BRITS	85.86	236.36	0.07	113.02	322.21	0.08
	MPGRU	225.42	1160.06	0.17	273.73	1237.21	0.21
	GRIN	99.16	566.98	0.07	135.90	417.89	0.10
	Hydro-GRNNI	**79.46**	**273.01**	**0.06**	**110.13**	**335.47**	**0.08**

Table 6.2: Performance Results of Imputation Methods on Discharge Dataset for In-Sample Evaluation

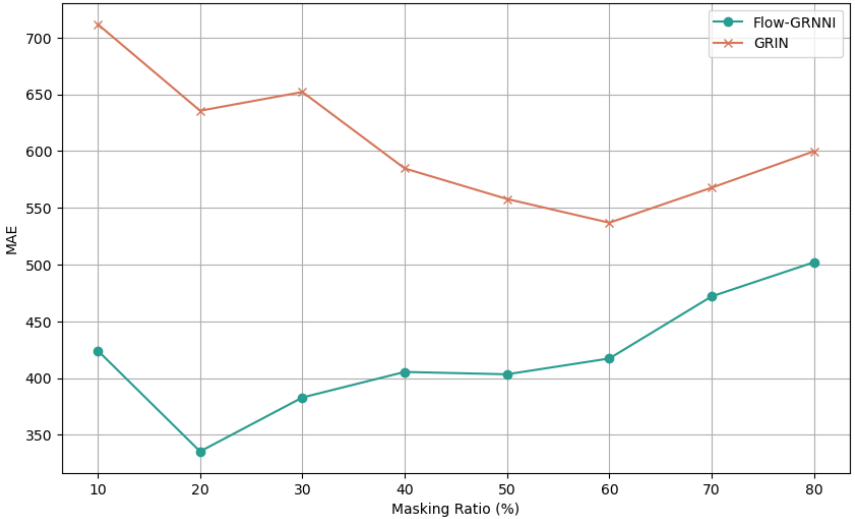
Visualization of Adjacency Matrix



Sensitivity Analysis



a. Point Missing



b. Block Missing

Conclusion

Summary

- Hydro-GRNNI achieved lowest MAE in sediment concentration predictions.
  - Out-of-sample: 246.01 (point), 327.84 (block)
  - In-sample: 79.46 (point), 110.13 (block)
- Outperforms other baseline models.
- Effective integration of flow direction with GNNs.

Limitations

- Sensitive to data quality and availability.
- Needs validation for larger or more complex regions.

Future Research

- Test across various watersheds.
- Explore advanced ML techniques.
- Investigate real-time monitoring and adaptive modeling.

Thank you!!!

Reference

- Kuppannagari, S.R., Fu, Y., Chueng, C.M., & Prasanna, V.K. (2021). *Spatio-Temporal Missing Data Imputation for Smart Power Grids.* ACM e-Energy '21, 458-465. DOI: 10.1145/3447555.3466586
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., et al. (2014). *Generative Adversarial Nets.* NeurIPS, 27.
- Che, Z., Purushotham, S., Cho, K., Sontag, D., & Liu, Y. (2018). *Recurrent Neural Networks for Multivariate Time Series with Missing Values.* Scientific Reports, 8(1), 1-12.
- Cho, K., Van Merrienboer, B., Gulcehre, C., et al. (2014). *Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation.* arXiv:1406.1078.
- Gilmer, J., Schoenholz, S.S., Riley, P.F., et al. (2017). *Neural Message Passing for Quantum Chemistry.* ICML, 1263-1272.
- Rubin, D.B. (1976). *Inference and Missing Data.* Biometrika, 63(3), 581-592.
- Dickinson, J.E., Hanson, R.T., & Predmore, S.K. (2014). *HydroClimATe—Hydrologic and Climatic Analysis Toolkit.* USGS Techniques and Methods, 4-A9.
- Sleckman, M.J., Hinman, E.D., Hamshaw, S.D., & Stanish, L. (2024). *surface-water-geospatial-data-assembly.* USGS Software Release. DOI: 10.5066/P165UIYY
- Stein, M.L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging.* Springer.
- DeCicco, L., Prinos, S., Eslick-Huff, P., et al. (2022). *HASP: Hydrologic AnalySis Package.* USGS. DOI: 10.5066/P9BUN5GV
- Hirsch, R., DeCicco, L., & Murphy, J. (2023). *Exploration and Graphics for RivEr Trends (EGRET).* USGS.
- Cini, A., Marisca, I., & Alippi, C. (2022). *Filling the Gaps: Multivariate Time Series Imputation by Graph Neural Networks.* ICLR. URL: https://openreview.net/forum?id=kOu3-S3wJ7