Self-Imputation and Cross-Variable Learning Improve Water Quality Prediction with Sparse Data

Xiaofeng Liu [1] **Xiaobo Xia** [2] **Xuechen Zhang** [1] **Mohna Chakraborty** [1] **Xiyuan Chang** [1] **Kuai Fang** [3]
William S. Currie [1] **Samet Oymak** [1]

Abstract

Accurate water quality prediction is essential for effective environmental management, yet infrequent sampling results in severe data sparsity, posing significant challenges for training traditional deep learning models. To address this, we propose a novel two-stage framework that leverages a tabular foundation model for multivariate time series prediction under sparse data conditions. In the first stage, the model self-imputes missing water quality values using hydroclimatic and calendar-based features; in the second stage, the imputed time series of all other water quality variables serve as augmented inputs to further improve prediction for each target variable. Evaluated on a continental-scale dataset, our proposed solution significantly outperforms both direct foundation models and traditional deep learning model baselines. We also demonstrate that explicit self-imputation for missing data yields more accurate predictions than relying on the model's internal mechanisms. To the best of our knowledge, this is the first study to demonstrate the effectiveness of tabular foundation models for sparse environmental time series prediction, providing a reliable and data-efficient alternative to traditional deep sequence models.

1. Introduction

Freshwater ecosystems, essential for their rich biodiversity and societal value, are experiencing increased water quality degradation globally (du Plessis, 2022; Bieroza et al., 2023). Poor water quality can lead to severe health and environmental consequences (Fazal-ur Rehman, 2019; Liu et al., 2025). Consequently, accurate water quality prediction is essential for timely interventions and effective environmental management (Zheng et al., 2025).

In real-world settings, accurate water quality prediction is severely hindered by the observational data sparsity. Monitoring budgets and logistics issues often result in infrequent sampling, with studies reporting up to 50-70% of observations missing (Rodríguez et al., 2021), and limited spatial coverage (Liu & Georgakakos, 2021). These data gaps pose a fundamental problem for most state-of-the-art deep learning-based models, which are typically "data-hungry" and require large amounts of data to train effectively from scratch. Furthermore, missing values introduce temporal discontinuities, and simple imputation methods (e.g., mean or median fillings) fail to capture the underlying process-driven variability. While strategies such as multi-site training or augmenting with additional variables have been proposed to mitigate these limitations (Heudorfer et al., 2025), these approaches remain grounded in traditional deep models and often require large datasets to outperform basic regression baselines (Fang et al., 2024), with limited reliability at individual sites (Xia et al., 2025).

In this study, we investigate the potential of tabular foundation models for addressing the challenges of water quality prediction under severe data sparsity. Foundation models have demonstrated remarkable adaptability across diverse domains (Vaghefi et al., 2023; Pai et al., 2024; Bodnar et al., 2025), yet their application to sparse, multivariate environmental time series remains underexplored. Our study investigates the usage of TabPFN-v2 (Hollmann et al., 2025), a pretrained tabular foundation model that has demonstrated strong performance on small structured datasets. TabPFN leverages a transformer-based architecture with attention mechanisms across both rows and columns, and it generates predictions for all test samples in a single forward pass. This allows TabPFN to make efficient use of limited training data without relying on auto-regressive forecasting. Recently, TabPFN has been adapted for univariate time-series forecasting (TabPFN-TS) with simple feature engineering (Hoo et al., 2025). However, the current TabPFN remains single-target and agnostic to temporal gaps, leaving two central

[1]University of Michigan [2]National University of Singapore [3]Stanford University. Correspondence to: Xiaofeng Liu <xiaofliu@umich.edu>, Samet Oymak <oymak@umich.edu>.

Proceedings of the 1st ICML Workshop on Foundation Models for Structured Data, Vancouver, Canada. 2025. Copyright 2025 by the author(s).

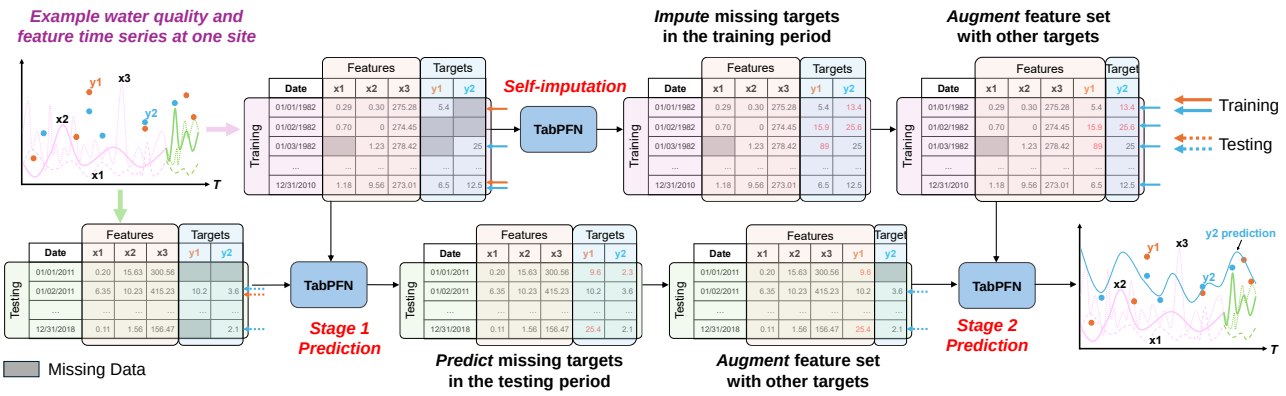

Figure 1. Overview of our proposed solution SIXL, using an example of three input features and two target water quality variables. Stage 1 performs single-variable in-context learning and explicit self-imputation to generate complete time series for each target using hydroclimatic and calendar-based features. Stage 2 refines the prediction by augmenting the original feature set with the imputed time series of all other target variables, enabling cross-variable learning.

challenges in water quality prediction—massive missingness and strong cross-variable dependencies unaddressed.

Instead of aggregating data across sites, we leverage TabPFN's strength on small data by applying it independently for each monitoring site. We frame the time series water quality prediction task as a supervised tabular regression problem and introduce **SIXL** as a novel two-stage strategy integrating explicit **s**elf-**i**mputation and **cross**-variable dependency **l**earning within a *single foundation-model framework* to improve the model performance:

- In **Stage 1**, the model explicitly imputes each target water quality variable's missing values independently using hydroclimatic and calendar-based features.
- In **Stage 2**, it refines the prediction for each variable using the imputed time series of all other variables as additional inputs, thus capturing cross-variable relationships.

Our main contributions are as follows: (1) We propose a novel two-stage framework which utilizes a tabular foundation model to effectively predict sparse multivariate time series and demonstrate the effectiveness of our approach on a continental-scale water quality dataset with highly irregular sampling. (2) Comprehensive ablation studies validate the importance of key components of our proposed solution.

2. Methodology

2.1. Dataset

The dataset used in this study is sourced from (Xia et al., 2025). Water quality measurements were originally collected by the U.S. Geological Survey (USGS) at 482 river sites across the continental United States between 1/1/1982 and 12/31/2018. For each site, 20 variables were included in this dataset to characterize various aspects of water quality dynamics, including physical/chemical processes, geochemical weathering, and nutrient cycling. Water quality data are highly sporadic, with typically one or two observations per month or even fewer (as shown in Figure 3), and the proportion of missing values varies across variables. The average number of observations for each variable per site is summarized in Table 1.

Input features consist of 22 hydroclimatic time series variables, grouped into runoff, meteorological forcings, vegetation indices, and chemicals in the rain, which are relatively complete daily time series. Detailed descriptions of these variables are provided in Table 2. To enable the general tabular foundation model to capture temporal relationships, we derived calendar-based features (i.e., the day of the week, the day of the month, the day of the year, the week of the year, and the month of the year). We also used sine and cosine values of these features to capture cyclical patterns (Hoo et al., 2025). In addition, a running index was included as a temporal reference to preserve the sequential ordering of observations.

2.2. SIXL: Our Proposed Solution

Figure 1 presents an overview of our proposed two-stage solution, illustrated with three input features and two target water quality variables.

Stage 1: Single-Variable Training and Self-Imputation As the TabPFN supports only single-target prediction, we first train the model separately for each target variable, using only the days within the training period that have observed values (with the corresponding hydroclimatic and date-related features as inputs). Then we *explicitly impute*

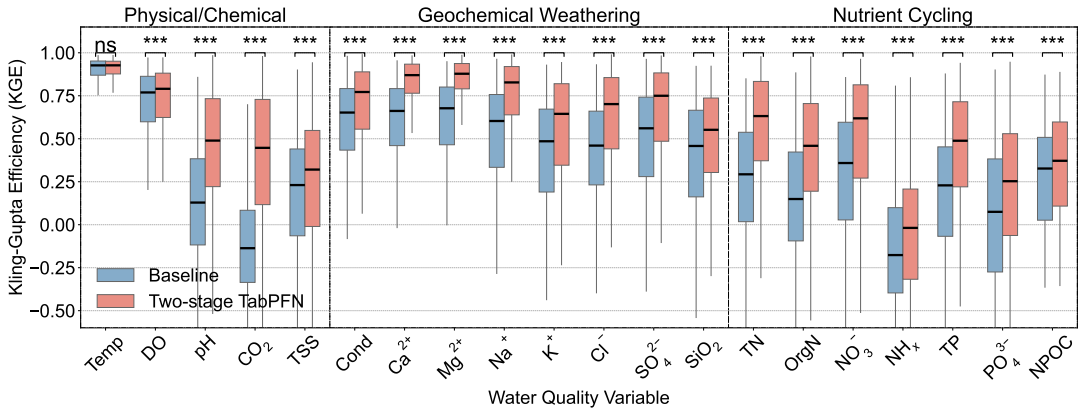

Figure 2. Comparison of Kling-Gupta Efficiency (KGE) values between baseline TabPFN (blue bars) and SIXL (red bars). The boxplots display the median (central line), interquartile range (IQR; box spans from the first quartile (Q1) to the third quartile (Q3)), and whiskers extending to Q1 $- 1.5 \times$ IQR and Q3 $+ 1.5 \times$ IQR. Wilcoxon signed-rank tests assess whether the two-stage TabPFN significantly outperforms the baseline. Significance levels are: $^{***}p < 0.001$, $^{**}p < 0.01$, and $^{*}p < 0.05$.

the target variable's missing values on all other days within the training period. The same trained model is also used to *predict* the target variable for all days in the testing period. This stage generates a complete daily time series for the given water quality variable, combining both actual observations on sampling days with model predictions on unsampled days.

Stage 2: Cross-Variable Learning and Prediction In this stage, we *retrain* the model for each target variable by also leveraging potential cross-variable dependencies. Specifically, we take the imputed daily time series of all other water quality variables obtained from Stage 1 and add these as additional input features for the target variable's prediction. We then train a new TabPFN model on the target variable using all days of the training period. During prediction for the testing period, when true observations of the auxiliary water quality variables are unavailable, their imputed values from Stage 1 are used as inputs. This process is repeated for all 20 water quality variables, treating each variable as the prediction target in turn.

3. Experiments

In this section, we present our results based on all 482 rivers datasets. We evaluate model performance on the testing samples using the Kling-Gupta Efficiency (KGE) (Gupta et al., 2009). KGE is a statistical goodness-of-fit measure widely used in hydrology and water quality modeling (Hunt et al., 2022; Xia et al., 2025; Fang et al., 2024), and the mathematical definition is provided in Section A.4. We also conduct the Wilcoxon signed-rank test (Conover, 1999) on the paired KGE values across sites to assess the statistical significance of performance differences between models.

3.1. Experimental Setup and Baseline

Details of the experimental setup are provided in Section A.2. Additional information on the underlying foundation model is presented in Section A.3. To evaluate the effectiveness of our proposed solution, we consider the model trained in Stage 1 as our baseline. Each trained model performs one-shot prediction across all days in the testing period.

3.2. Results

Figure 2 compares KGE values of the baseline TabPFN and our proposed two-stage approach across 20 water quality variables. The two-stage approach significantly improves prediction across all variables, with the exception of water temperature, which is already predicted very well by the baseline. The most notable improvements are observed for pH and CO_2.

The degree of performance improvement varies across variables, likely due to a combination of factors, including the extent and pattern of missing data in each variable, its relationship with hydroclimatic and temporal factors, its correlation with other water quality variables, as well as the inherent predictability of the variable itself.

3.3. Ablation Study

To understand the importance of different components within our proposed two-stage strategy, we conducted a series of ablation studies to specifically answer the following research questions:

RQ1: Does imputing the target variable alone improve

the model performance?

To isolate the effect of expanding the training set through target imputation alone, without introducing cross-variable information, we first imputed the missing values for each target variable in Stage 1. We then trained a new TabPFN model for each target using its fully imputed time series over the entire training period as labels, combined only with the hydroclimatic and calendar-based features.

Figure 4 shows that, despite the larger training set (10,811 examples compared to fewer than 300 in the original set), the model's testing performance does not improve significantly over the baseline.

While Li et al. (2025) recently demonstrated TabPFN can leverage unlabeled examples to enhance predictive performance when only a few labeled instances are available, this benefit is not evident here. One possible reason is that TabPFN is pretrained on datasets with fewer than 10,000 examples and excels in handling small to medium-sized datasets. Therefore, expanding the dataset beyond this scale may provide diminishing benefits.

RQ2: How does cross-variable learning improve performance?

To evaluate the benefit of cross-variable learning, we predict each target variable by augmenting the input features with the 19 other target variables, while using only their actual observations (i.e., no imputation for these auxiliary targets).

Compared to the target-imputed-only setting in RQ1, our approach substantially improves prediction accuracy across all 20 water quality variables, as shown in Figure 4. These results indicate that leveraging the relationships between water quality variables themselves (even sparsely observed) can compensate for limited temporal coverage and enhance model performance.

RQ3: Does explicit self-imputation of auxiliary variables outperform TabPFN's internal missing data handling?

To evaluate whether explicitly imputing the 19 auxiliary features provides an advantage over the model's internal handling of missing features, we compare the prediction performance among four configurations: (1) no imputation for augmented features, (2) imputation only in training, (3) imputation only in testing, and (4) imputation in both training and testing.

The comparison shown in Figure 5 reveals that the best performance is achieved when missing values in the auxiliary features are imputed in both training and testing samples. However, the model will not benefit from imputing only during testing if training samples remain missing and are handled internally by the model. This also suggests that consistent feature quality across both training and testing samples is essential. These results demonstrate that the explicit self-imputation strategy is more effective than TabPFN's internal mechanism, which simply fills missing values with training set means and flags them with binary indicators. This is likely because by explicitly training the model to predict the missing values for each water quality variable using hydroclimatic and date variables, the model learns the underlying process-driven variability, compared to using simple averages.

3.4. Comparison with the Time Series Model

To further evaluate the performance of TabPFN in the context of time-series modeling, we compare it with a long short-term memory (LSTM) network, which is widely recognized as state-of-the-art in hydrological and water quality modeling.

Following the setup in (Xia et al., 2025), a multi-task LSTM model is trained using data from all 482 river sites to perform one-day-ahead predictions for each target variable, based on the previous 365 days of input features. To enable a fairer comparison, we apply the foundation model in a rolling approach where, for each test sample, all prior days with observed values for the target variable are used as training samples. Both models are evaluated over the same testing period.

As shown in Figure 6, TabPFN significantly outperforms LSTM on 13 out of the 20 water quality variables. Notably, these results are based on the baseline TabPFN model trained only using a rolling approach, without applying our proposed two-stage strategy. Furthermore, the LSTM model includes 49 additional static features, such as basin characteristics, land use, and soil properties, alongside the shared hydroclimatic and calendar-based variables. Our findings demonstrate that a general-purpose tabular foundation model like TabPFN, even without sequence-specific architecture or static feature augmentation, can achieve state-of-the-art performance in water quality prediction tasks with appropriate feature engineering and training strategy.

4. Conclusion

This study presents SIXL as a novel two-stage strategy leveraging a tabular foundation model to improve stream water quality prediction under sparse data conditions. Our key finding is that explicitly using the model for self-imputation of missing values, followed by retraining on the completed dataset with additional water quality variables, substantially enhances predictive performance compared to direct application. Furthermore, we show that explicit self-imputation outperforms the model's internal mechanism for handling missing data, highlighting the importance of targeted imputation in sparse multivariate time series prediction.

Acknowledgments

This work was supported by the Schmidt Sciences, LLC.

References

Bieroza, M., Acharya, S., Benisch, J., Ter Borg, R. N., Hallberg, L., Negri, C., Pruitt, A., Pucher, M., Saavedra, F., Staniszewska, K., et al. Advances in catchment science, hydrochemistry, and aquatic ecology enabled by high-frequency water quality measurements. *Environmental Science & Technology*, 57(12):4701–4719, 2023.

Bodnar, C., Bruinsma, W. P., Lucic, A., Stanley, M., Allen, A., Brandstetter, J., Garvan, P., Riechert, M., Weyn, J. A., Dong, H., et al. A foundation model for the earth system. *Nature*, pp. 1–8, 2025.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33: 1877–1901, 2020.

Conover, W. J. *Practical nonparametric statistics*. john wiley & sons, 1999.

du Plessis, A. Persistent degradation: Global water quality challenges and required actions. *One Earth*, 5(2):129–131, 2022.

Fang, K., Caers, J., and Maher, K. Modeling continental us stream water quality using long-short term memory and weighted regressions on time, discharge, and season. *Frontiers in Water*, 6:1456647, 2024.

Fazal-ur Rehman, M. Polluted water borne diseases: Symptoms, causes, treatment and prevention. *J Med Chem Sci*, 2(1):21–26, 2019.

Feng, D., Fang, K., and Shen, C. Enhancing streamflow forecast and extracting insights using long-short term memory networks with data integration at continental scales. *Water Resources Research*, 56(9):e2019WR026793, 2020.

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F. Decomposition of the mean squared error and nse performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, 377(1-2):80–91, 2009.

Heudorfer, B., Gupta, H. V., and Loritz, R. Are deep learning models in hydrology entity aware? *Geophysical Research Letters*, 52(6):e2024GL113036, 2025.

Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Hollmann, N., Müller, S., Eggensperger, K., and Hutter, F. Tabpfn: A transformer that solves small tabular classification problems in a second. *arXiv preprint arXiv:2207.01848*, 2023.

Hollmann, N., Müller, S., Purucker, L., Krishnakumar, A., Körfer, M., Hoo, S. B., Schirrmeister, R. T., and Hutter, F. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025.

Hoo, S. B., Müller, S., Salinas, D., and Hutter, F. The tabular foundation model tabpfn outperforms specialized time series forecasting models based on simple features. *arXiv preprint arXiv:2501.02945*, 2025.

Huang, S., Xia, J., Wang, Y., Lei, J., and Wang, G. Water quality prediction based on sparse dataset using enhanced machine learning. *Environmental Science and Ecotechnology*, 20:100402, 2024.

Hunt, K. M., Matthews, G. R., Pappenberger, F., and Prudhomme, C. Using a long short-term memory (lstm) neural network to boost river streamflow forecasts over the western united states. *Hydrology and Earth System Sciences*, 26(21):5449–5472, 2022.

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., and Nearing, G. S. Toward improved predictions in ungauged basins: Exploiting the power of machine learning. *Water Resources Research*, 55(12): 11344–11354, 2019.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Li, Y., Chang, X., Kara, M., Liu, X., Roy-Chowdhury, A., and Oymak, S. When and how unlabeled data provably improve in-context learning. *arXiv preprint arXiv:2506.15329*, 2025.

Liu, X. and Georgakakos, A. P. Chlorophyll a estimation in lakes using multi-parameter sonde data. *Water Research*, 205:117661, 2021.

Liu, X., Zuo, C., Guan, J., Ma, Y., Liu, Y., Zhao, G., and Wang, R. Extreme rainfall disproportionately impacts e. coli concentrations in texas recreational waterbodies. *Science of The Total Environment*, 958:178062, 2025.

Nearing, G., Cohen, D., Dube, V., Gauch, M., Gilon, O., Harrigan, S., Hassidim, A., Klotz, D., Kratzert, F., Metzger, A., et al. Global prediction of extreme floods in ungauged watersheds. *Nature*, 627(8004):559–563, 2024.

Pai, S., Bontempi, D., Hadzic, I., Prudente, V., Sokač, M., Chaunzwa, T. L., Bernatz, S., Hosny, A., Mak, R. H.,

Birkbak, N. J., et al. Foundation model for cancer imaging biomarkers. *Nature machine intelligence*, 6(3):354–367, 2024.

Ramesh, G. et al. Enhancing water quality monitoring with explainable ai and wgan-based data augmentation. *Remote Sensing in Earth Systems Sciences*, pp. 1–12, 2025.

Rodríguez, R., Pastorini, M., Etcheverry, L., Chreties, C., Fossati, M., Castro, A., and Gorgoglione, A. Water-quality data imputation with a high percentage of missing values: A machine learning approach. *Sustainability*, 13 (11), 2021.

Shen, C. A transdisciplinary review of deep learning research and its relevance for water resources scientists. *Water Resources Research*, 54(11):8558–8593, 2018.

Vaghefi, S. A., Stammbach, D., Muccione, V., Bingler, J., Ni, J., Kraus, M., Allen, S., Colesanti-Senni, C., Wekhof, T., Schimanski, T., et al. Chatclimate: Grounding conversational ai in climate science. *Communications Earth & Environment*, 4(1):480, 2023.

Wang, Z., Li, Y., Yang, C., Zhu, H., and Zhou, C. Graph-based active semi-supervised learning: Case study in water quality monitoring. *Adv. Eng. Informatics*, 62: 102902, 2024.

Xia, X., Liu, X., Liu, J., Fang, K., Lu, L., Oymak, S., Currie, W. S., and Liu, T. Identifying trustworthiness challenges in deep learning models for continental-scale water quality prediction. *arXiv preprint arXiv:2503.09947*, 2025.

Ye, H.-J., Liu, S.-Y., and Chao, W.-L. A closer look at tabpfn v2: Strength, limitation, and extension. *arXiv preprint arXiv:2502.17361*, 2025.

Zheng, Y., Zhang, X., Zhou, Y., Zhang, Y., Zhang, T., and Farmani, R. Deep representation learning enables cross-basin water quality prediction under data-scarce conditions. *npj Clean Water*, 8(1):33, 2025.

Zhi, W., Klingler, C., Liu, J., and Li, L. Widespread deoxygenation in warming rivers. *Nature Climate Change*, 13 (10):1105–1113, 2023.

A. Appendix

A.1. Related Work

Deep Learning and Small Data in Water Quality Research. Deep learning models, particularly long short-term memory (LSTM) networks (Hochreiter & Schmidhuber, 1997), have been widely applied in hydrologic and water quality prediction due to their ability to capture long-term temporal dependencies (Shen, 2018; Kratzert et al., 2019; Feng et al., 2020; Zhi et al., 2023; Fang et al., 2024; Nearing et al., 2024; Xia et al., 2025). However, traditional deep learning models typically require large volumes of labeled data, which is often unavailable in environmental domains where sampling is infrequent and costly. Missing values in time series further challenge predictive modeling, as naive imputation methods (e.g., mean or median filling) can distort temporal patterns. To overcome these limitations, recent studies have introduced several strategies to improve model generalization under data sparse constraints, including (i) cross-site transfer learning and meta-learning, which pretrain representations on data-rich basins before adapting to data-sparse ones (Zheng et al., 2025); (ii) synthetic data augmentation, using GAN generated or process-model simulated time series to expand the effective sample size and capture rare events (Ramesh et al., 2025); (iii) semi-/self-supervised and few-shot schemes that harness unlabeled sensor data through label propagation, active learning or contrastive pretraining (Wang et al., 2024); and (iv) knowledge-guided neural networks that embed mass balance or energy conservation constraints into LSTM or attention networks to reduce overfitting and improve interpretability with limited observations (Huang et al., 2024). While these strategies improve robustness, they often require extensive task-specific training. In addition, imputation and prediction are typically handled separately by different models. These limitations motivate exploring the application of foundation models that can generalize across tasks with minimal hyperparameter tuning and potentially unify both steps within a single model framework.

A.2. Experimental Setup

The proposed method (illustrated in Figure 1) was applied independently for each water quality monitoring site to account for site-specific differences in water quality dynamics. For each site-specific dataset, the time series is split chronologically, with the first 80% used for training and the remaining 20% held out for testing. All experiments were conducted using a single NVIDIA L40s GPU.

A.3. Tabular Foundation Model

TabPFN. The Tabular Prior-data Fitted Network (TabPFN) (Hollmann et al., 2023; 2025) is explicitly developed for small tabular datasets, typically those with fewer than 10,000 rows and 500 features. TabPFN leverages a Transformer architecture that is pre-trained offline once on millions of synthetic tabular datasets. This meta-learning approach allows TabPFN to approximate Bayesian inference and perform in-context learning (Brown et al., 2020) on new, unseen small datasets without requiring per-dataset training or hyperparameter tuning. This shifts the data requirement from task-specific labeled datasets to a one-time, large-scale synthetic pretraining process, enabling "off-the-shelf" application to small, sparse datasets (Ye et al., 2025). Recently, TabPFN has been successfully adapted for univariate time series forecasting (Hoo et al., 2025). However, it is unknown if it can effectively address the challenges of highly sparse multivariate time series that commonly exist in many scientific domains. In this work, we apply TabPFN to the domain of water quality prediction through a novel two-stage framework, addressing three key challenges: 1) incorporating hydroclimatic inputs on days with no water quality observations (unlabeled samples); 2) preserving temporal continuity by explicitly imputing missing target values in Stage 1, and 3) capturing cross-variable dependencies by using the imputed time series of other water quality variables as augmented features in Stage 2.

A.4. Model Performance Metric

Kling-Gupta Efficiency (KGE) is defined as:

$$\text{KGE} = 1 - \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2}, \tag{1}$$

where r is the correlation coefficient between predictions and observations, α is the ratio of the standard deviation of predictions to that of observations (representing variability error), and β is the bias ratio (mean of predictions over mean of observations). The KGE ranges from $-\infty$ to 1, with 1 indicating a perfect match. We chose KGE because it provides a comprehensive evaluation of model performance by jointly accounting for correlation, variability, and bias, which is critical in water quality modeling where both the magnitude and temporal variability of variables must be accurately captured.

Table 1. Summary of the 20 predicted water quality variables and the average number of observations per basin based on data from 482 U.S. rivers collected between January 1, 1982 and December 31, 2018.

USGS Parameter Code	Description	Abbreviation	Unit	Ave. Obs.
00010	Water temperature	Temp	°C	331
00095	Specific conductance	Cond	uS/cm at 25°C	286
00300	Oxygen	DO	mg/L	198
00400	pH	pH	-	225
00405	Carbon dioxide	CO_2	mg/L	130
00600	Total nitrogen	TN	mg/L	193
00605	Organic nitrogen	OrgN	mg/L	172
00618	Nitrate	NO_3^-	mg/L as N	138
00660	Orthophosphate	PO_4^{3-}	mg/L as PO_4^{3-}	205
00665	Total phosphorus	TP	mg/L as P	267
00681	Organic carbon	NPOC	mg/L	60
00915	Calcium	Ca^{2+}	mg/L	132
00925	Magnesium	Mg^{2+}	mg/L	132
00930	Sodium	Na^+	mg/L	117
00935	Potassium	K^+	mg/L	115
00940	Chloride	Cl^-	mg/L	184
00945	Sulfate	SO_4^{2-}	mg/L	154
00955	Silica	SiO_2	mg/L	116
71846	Ammonia and ammonium	NH_x (NH_3 and NH_4^+)	mg/L as NH_4^+	184
80154	Suspended sediment concentration	TSS	mg/L	305

Table 2. List of input features used for water quality prediction, including 22 hydroclimatic variables and 16 derived calendar-based features.

Group	Feature	Description	Unit
Runoff	Q	Basin area normalized streamflow from USGS	m/y
Meteorological forcings	pr	Daily total precipitation	mm/day
	sph	Specific humidity	-
	srad	Surface downwelling solar radiation	W/m^2
	tmmn	Daily minimum 2-meter air temperature	K
	tmmx	Daily maximum 2-meter air temperature	K
	pet	Reference grass evapotranspiration	mm/day
	etr	Reference alfalfa evapotranspiration	mm/day
Rainfall chemistry	pH	Logarithm of the H ion activity	-
	Cond	Electrical conductivity of water	μS/cm
	Ca^{2+}	Ca ion concentration	mg/L
	Mg^{2+}	Mg ion concentration	mg/L
	K^+	K ion concentration	mg/L
	Na^+	Na ion concentration	mg/L
	NH_4^+	NH_4 concentration	mg/L
	NO_3^-	NO_3 concentration	mg/L
	Cl^-	Cl ion concentration	mg/L
	SO_4^{2-}	SO_4 concentration	mg/L
	distNTN	The distance to the nearest NTN sampling site	km
Vegetation indices	LAI	Leaf area index of vegetation	m^2/m^2
	FAPAR	Fraction of absorbed photosynthetically active radiation	unitless
	NPP	Net primary production	$gC/m^2/day$
Date features	DoW	Day of week	-
	DoM	Day of month	-
	DoY	Day of year	-
	WoY	Week of year	-
	MoY	Month of year	-
	sin	Sine values of DoW, DoM, DoY, WoY, and MoY	-
	cos	Cosine values of DoW, DoM, DoY, WoY, and MoY	-
	Run_Idx	Running index	-

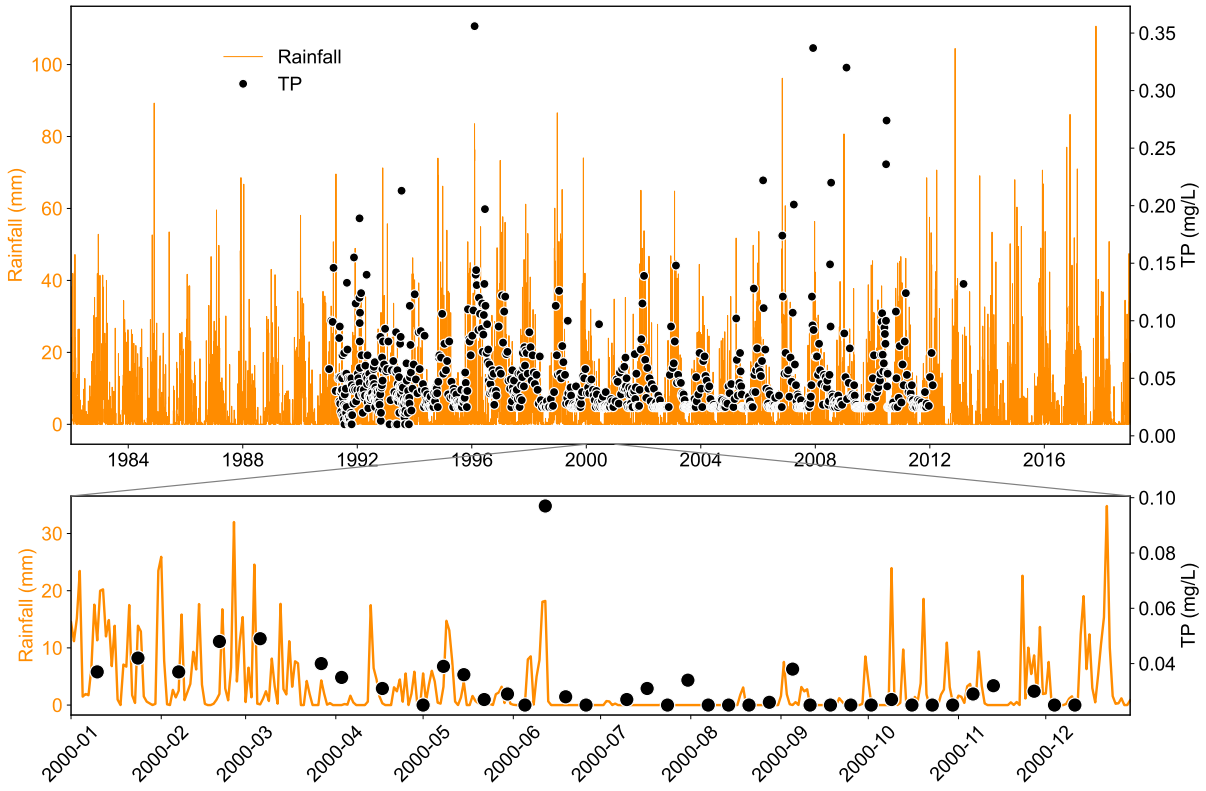

Figure 3. Example time series of rainfall (a hydroclimatic feature) and total phosphorus (a target water quality variable). Hydroclimatic variables are recorded as relatively complete daily time series, whereas water quality variables are highly sparse, with observations typically collected biweekly, monthly, or less.

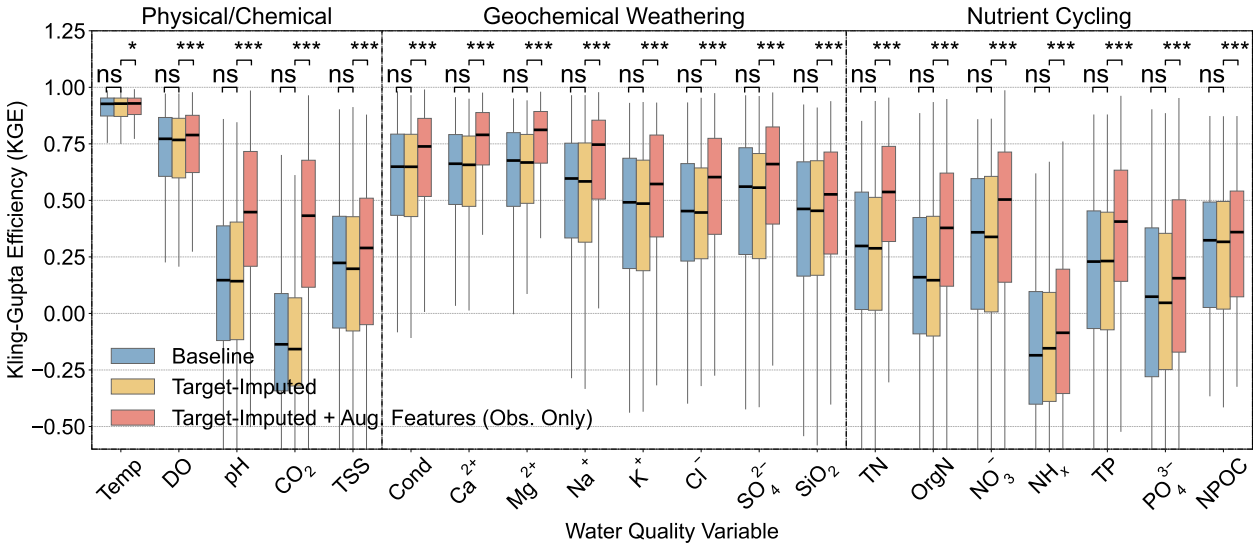

Figure 4. Comparison of Kling-Gupta Efficiency (KGE) values across three configurations: (1) baseline, (2) target-imputed only, and (3) target-imputed with augmented features using only observed values, while leaving missing values to be handled by TabPFN's internal missing data mechanism. The boxplots show the median (central line), interquartile range (IQR, represented by the boxes spanning the first (Q1) to the third quartile (Q3)), and whiskers extending to Q1 $- 1.5 \times$ IQR and Q3 $+ 1.5 \times$ IQR; Wilcoxon Signed-Rank tests assess whether (2) significantly outperforms (1), and whether (3) significantly outperforms (2). Significance levels are: $^{***}p < 0.001$, $^{**}p < 0.01$, and $^{*}p < 0.05$.

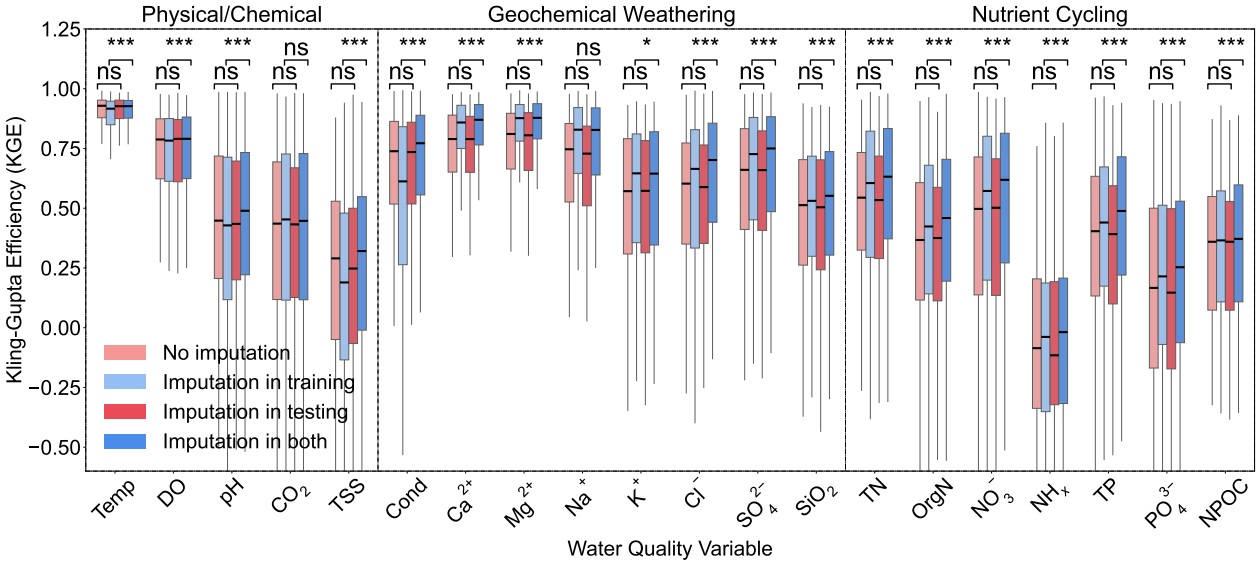

Figure 5. Comparison of Kling-Gupta Efficiency (KGE) values across four configurations to evaluate whether explicit self-imputation of auxiliary variables outperforms TabPFN's internal missing data handling: (1) no imputation, (2) imputation in training only, (3) imputation in testing only, and (4) imputation in both training and testing. The boxplots display the median (central line), interquartile range (IQR; box spans from the first quartile (Q1) to the third quartile (Q3)), and whiskers extending to Q1 $- 1.5 \times$ IQR and Q3 $+ 1.5 \times$ IQR. Wilcoxon signed-rank tests assess whether imputation in testing significantly improves performance over no imputation, and whether imputing in both training and testing significantly outperforms training-only imputation. Significance levels are: $^{***}p < 0.001$, $^{**}p < 0.01$, and $^{*}p < 0.05$.

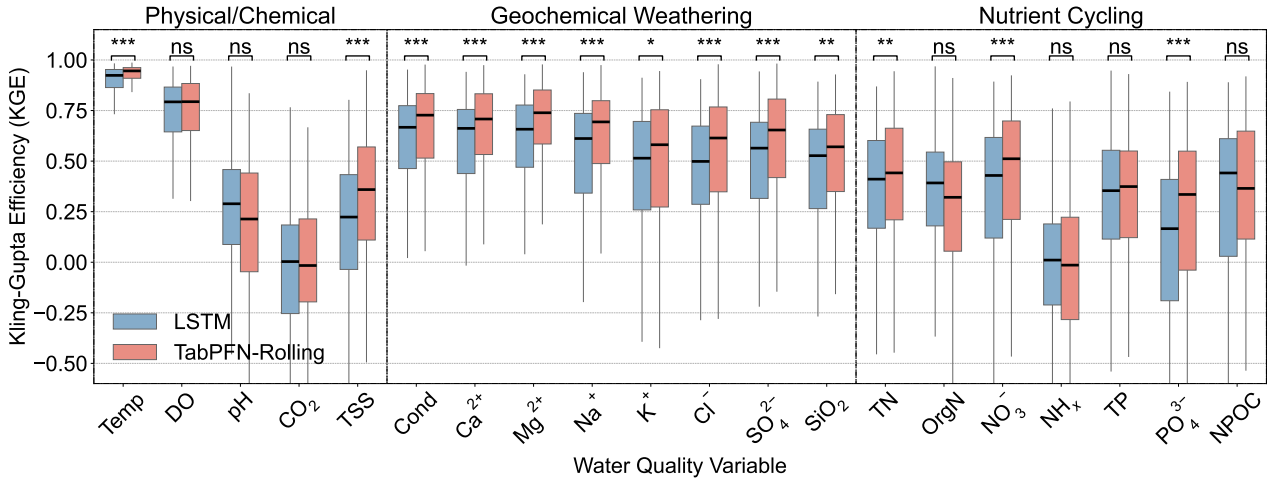

Figure 6. Comparison of Kling-Gupta Efficiency (KGE) values between LSTM and TabPFN using in a rolling prediction approach. The boxplots display the median (central line), interquartile range (IQR; box spans from the first quartile (Q1) to the third quartile (Q3)), and whiskers extending to Q1 $- 1.5 \times$ IQR and Q3 $+ 1.5 \times$ IQR. Wilcoxon signed-rank tests assess whether TabPFN significantly outperforms LSTM. Significance levels are: $^{***}p < 0.001$, $^{**}p < 0.01$, and $^{*}p < 0.05$.