Predictions of Groundwater Arsenic Concentrations using Graph Neural Networks

Anonymous submission

Abstract

Arsenic contamination in groundwater presents a longstanding public health crisis in the United States, especially for households depending on private wells. Accurate and spatially informed prediction of arsenic concentration is vital to identify high-risk areas and focus mitigation efforts. However, there is a lack of generalizable models for representing continuous variation in arsenic concentrations across regions. In this work, we pose arsenic prediction as a regression task and construct a spatially integrated dataset to aggregate over 74,000 arsenic samples from the Water Quality Portal (WQP), Mineral Resources Data System (MRDS), and Gridded National Soil Survey Geographic Database (gNATSGO). Specifically, we use a variety of techniques including k-Nearest Neighbors (KNN) and Geographic Information Systems (GIS) to join arsenic measurement points from across the United States by location. Building on this dataset, we evaluate a diverse suite of machine learning models, including tree-based ensemble approaches, multilayer perceptrons, and spatially aware graph neural networks (GNN). Our findings show that while gradient-boosted trees are still considered state-of-the-art in the field of tab data, GNNs are able to further account for spatial dependence to match or outperform the results of gradient-boosted trees. These results demonstrate that graph-based and spatially informed learning can enhance environmental prediction and provide a foundation for improved groundwater risk mapping and monitoring.

1 Introduction

Arsenic contamination in groundwater is currently a major public health issue in the United States. It is estimated that 2.1 million Americans drink water from wells that have dangerously high arsenic levels (Ayotte et al. 2017). Both the World Health Organization (WHO) and the Environmental Protection Agency (EPA) have set the safe limit for arsenic at 10 µg/L, but numerous regions in America, particularly in states such as Nevada and Arizona, exceed this amount (Lombard et al. 2021). Chronic exposure to arsenic can lead to cardiovascular disease, diabetes, and cancer. Although direct chemical testing is the most reliable way to assess arsenic levels, this option is often infeasible due to financial or logistical constraints, especially for people who live in rural areas, where arsenic contamination is more prevalent and reliance on private wells is more common.

Recent studies have applied machine learning to address the prediction of arsenic contamination in groundwater, but several challenges remain to achieve robust generalization and real-world deployment. Very few prior studies approach arsenic prediction as a regression task, despite the fact that regression would allow for the creation of much more detailed prediction maps of areas with poor water quality (Hu et al. 2022). For example, two samples containing 9.9 µg/L and 10 µg/L of arsenic are nearly identical and both indicate unsafe conditions, yet a classification model would label one as "safe" and the other as "unsafe" while not providing further information.

While previous studies have reported very low loss metrics such as mean absolute error (MAE) or mean square error (MSE), many have focused on smaller or region-specific datasets. These approaches provide valuable localized insights but are often limited in their ability to generalize across diverse geologic settings. Moreover, there are studies that conduct experiments on a national or global scale, but their lack of data density remains a critical barrier, as arsenic values can vary greatly across short distances due to factors such as depth differences and geochemistry. Lastly, very few previous papers have attempted to predict arsenic concentrations in water using machine learning techniques in the United States, instead opting to focus more on various South American and Asian countries.

In this study, we aim to address this critical gap in the literature and introduce two novel advancements in the study the prediction of arsenic contamination. First, we provide a unified dataset for arsenic prediction which unifies location and geological datasets for a total of 74,706 data points. This dataset includes a diverse set of features that influence arsenic contamination and provides sufficient sample density to represent the corresponding geographic area. Our analysis focuses on the contiguous United States, which, compared to previous studies of localized or homogeneous areas, has not yet been extensively examined in detail by itself as an independent modeling region. Second, we train and evaluate more modern machine learning models such as graph neural network (GNN) variants for arsenic prediction alongside classical approaches such as random forests and gradient boosted trees. Although GNNs have not yet been applied to water contamination prediction, their ability to model spatial autocorrelation makes them particularly well-suited for this task, an important challenge highlighted in (Hu et al. 2022).

We are also concerned that the tree-based models widely used in prior studies may not fully capture the underlying relationships between contamination levels and environmental features, particularly due to their false assumption that each data point is independent of the others and their reliance on axis-aligned splits. Although tree-based ensemble models remain highly competitive, we found that advanced variations of graph neural networks (GNN) are able to outperform them, highlighting their growing potential for environmental science applications. Using a dense, geographically diverse dataset that spans the continental U.S, we demonstrate improved model performance and enhanced applicability for large-scale arsenic risk prediction.

2 Related Works

Recent research has predominantly approached arsenic prediction as a classification problem, with one survey of 27 drinking water studies reporting that 70% of the reviewed papers adopted this approach (Hu et al. 2022). Furthermore, the majority of these works used tree-based ensemble models, presumably because of their robustness and well-established performance on tabular data. For example, (Podgorski and Berg 2020) attempts to predict global arsenic concentrations using around 55,000 data points, although their data is mostly concentrated in Asia, North America, and Europe. They aggregate all data points within 1-km^2, only keep samples from above 100 m depth, and use a random forest to make predictions. (Wu et al. 2021) models arsenic concentrations in groundwater across Uruguay, introducing a hybrid approach by combining data points at different depths or geological compositions. They apply spatial aggregation by aggregating arsenic measurements within the same pixel using the geometric mean of values. Various papers attempt to model arsenic contamination in India, a country known for its water pollution. The studies focused on India utilize boosted regression trees or random forest models for prediction and data processing techniques such as kriging or spatial analysis. (Chattopadhyay et al. 2023; Tan, Yang, and Zheng 2020; Podgorski et al. 2020; Wu et al. 2023).

Not many studies have applied regression models to predict water contamination, much less specifically for arsenic concentration. Tree based methods are still very popular for regression. (Subudhi et al. 2025) predicts the groundwater quality index in Odisha and also uses feature importance to find the most effectual contaminants to this value. Their dataset consists only of samples from 2019-2022 and includes features such as pH, electrical conductivity, and other substances found in groundwater. As for the specific machine learning algorithms, they use gradient boosted trees including XGBoost, LightGBM, and CatBoost along with random forests. (Zhao, Kumar, and Wang 2024) employs multiple linear regression (MLIR), multivariate logistic regression (MLOR), and random forest models to predict arsenic concentrations in groundwater in the Hetao Basin (China) and several regions of Bangladesh. Among these, the random forest model outperforms MLIR and MLOR.

Although GNNs have not been widely applied to arsenic prediction, they have been frequently used in other environmental science domains, where their capacity to capture spatial correlations makes them particularly effective. For example, (Zha and Yang 2024) uses GNN variants to predict the amount of heavy metals in the Pearl River Basin, China. Their GNN graph was created by first connecting nodes based on geographic distance and further refined by creating a reachability matrix to connect remaining nodes and avoid having a non-connected graph. They also found that GNNs are able to handle high-dimensional data and reveal complex spatial patterns of soil heavy metal distribution. (Truong et al. 2023) combines GNNs with physics-based modeling to estimate pressure and flow in water distribution networks. Their nodes represent the endpoints of the pipes (junctions, tanks, reservoirs, etc.) and their edges correspond with real-world pipe connections, modeling the relations between these endpoints. (Sun et al. 2022) similarly introduces a physics-guided GNN approach for basin-scale river network learning and streamflow forecasting.

3 Methodology

Data Sources

We created our data set by merging other datasets from a variety of sources. The Water Quality Portal (WQP) (Environmental Protection Agency and United States Geological Survey) is a reputable water contamination data source, integrating data from the United States Geological Survey (USGS), the Environmental Protection Agency (EPA), and other state, federal, tribal, and local agencies. In addition to nationwide water contamination measurements, the dataset includes predictive variables such as latitude, longitude, and well depth. To incorporate soil properties, we used the Gridded National Soil Survey Geographic Database (gNATSGO) (Staff), which provides high-resolution spatial data on soil horizons, components, and other attributes relevant to contaminant mobility and retention. Some of the predictive data provided by this dataset include the percentage of elements like clay and organic matter in the soil, the pH of the soil, and the slope gradient and elevation of the surrounding environment. Geological and mineralogical data were added from the Mineral Resources Data System (MRDS) (Mason and Arndt), which catalogs mineral deposit locations and a variety of associated parameters. Some examples of these parameters are the minerals present, geomorphic location, and the physiographic division.

Together, these datasets offer a complementary view of the environmental conditions that impact water contaminant levels. We provide the geographic distribution of each dataset in Figure 1.

Our dataset has 74,706 samples, each corresponding with a arsenic measurement, and 115 features describing the location of that samples. Approximately 87.69% of samples have a safe level of Arsenic, and about 12.31% of samples have an unsafe level of Arsenic. Figure 2 illustrates a detailed distribution arsenic concentrations across the country.

Feature Engineering

We only kept the most recently collected arsenic samples from each WQP station to avoid data leakage and temporal

(a) WQP Stations (b) MRDS Sites (c) gNATSGO Soil Polygons (Represent mapped soil units)

Figure 1: Spatial coverage of datasets used for arsenic prediction across the mainland United States.

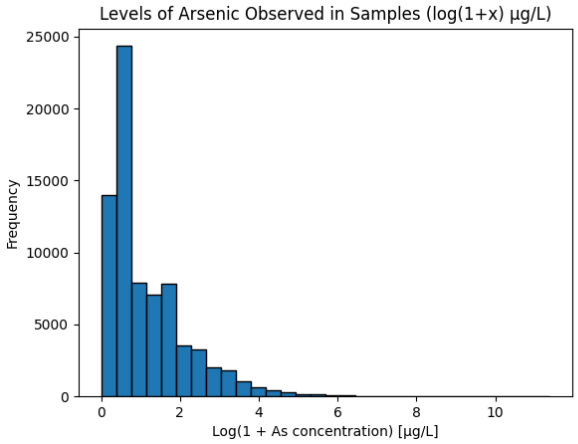


Figure 2: Distribution of arsenic levels in log scale

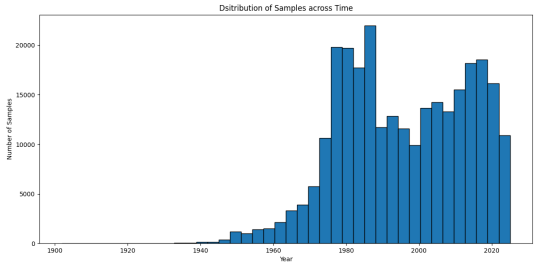


Figure 3: Distribution of samples across time

bias. Figure 3 represents the distribution of arsenic samples across time after filtering out older data samples. When arsenic values were missing, detection limits were typically reported, and when arsenic values were available, detection limits were usually absent. Because of this strong correlation between missing arsenic values and the presence of detection limits, we imputed the arsenic value with half the detection limit. This covered over 97% of missing values. We also filled in missing values in the "Well Depth" column by grouping together rows by their corresponding county code then imputing the median for each county. These two techniques boost the number of training values available, allowing for more accurate model performance.

For the MRDS dataset, we grouped together samples across time from the same MRDS station. One important technique we used here was word embeddings. MRDS contained many non-numeric columns, such as main commodity found and physiographic sector. We mapped these words to vectors of length 2, 4, or 8 using PyTorch embeddings. Lastly, we joined this physiographic information back to the original WQP dataset by averaging the values for rows within 5 km of a WQP station. In the case where no nearby stations within the set radius are found, the all values for that station are kept as 0. All coordinates were standardized to WGS84 (EPSG:4326), and distances were computed using the haversine metric. The low distance of 5 km ensures that the mineral information is relevant to the arsenic sampling location and the default value of 0 instead of assigning no values ensures that rows that are still useful for training.

gNATSGO is a bit different from the other datasets in that it divides the United States into different sectors instead of taking samples from specific locations. As a result, we used GIS to determine which gNATSGO sector contained each WQP station. Each sector has its own id, called a MUKEY, and each sector contains multiple components that also have an id, called a COKEY. Each COKEY corresponds to specific soil data such as pH levels and the concentration of different particles (sand, silt, etc.). Because gNATSGO divides the United States into polygonal map units rather than point-based sampling locations, we could not directly join this dataset to our WQP records using a k-nearest neighbors approach. Instead, we visualized both the WQP stations and gNATSGO polygons in a geographic information system (GIS) and used spatial overlay operations to assign each WQP station the corresponding MUKEY identifier. Because each station had one MUKEY but each MUKEY was mapped to multiple COKEYs, the rows were finalized by averaging the numeric values. For rows without a MUKEY, we decided to use a k-NN to impute the MUKEY and all the data of those rows to its nearest neighbor. One hot encoding was used for columns that did not contain too many different values and embeddings were used for columns that did contain a range of different values. Similarly to handling the WQP data set, we imputed the median for numeric columns with few missing values.

Predictive Models

We first evaluated simple and widely used models to establish reference performance, then expanded to include

advanced architectures. As a benchmark, we used random forests and multilayer perceptrons, two widely adopted ensemble methods known for performing pretty well on tab data. We further assessed many types of graph neural networks (GNN), including graph convolutional networks (GCN) (Kipf and Welling 2016), graph attention transformers (GAT) (Velickovic et al. 2017), GraphSAGE (Hamilton, Ying, and Leskovec 2017), and graph isomorphism networks (GIN) (Xu et al. 2018a). For additional comparison, we compared a range of gradient boosted trees, including XGBoost (Chen and Guestrin 2016), LightGBM (Ke et al. 2017), and CatBoost (Ostroumova et al. 2017). Although GNNs have been rarely applied in past water contamination studies, their successful use in related environmental science fields and their proven ability to capture spatial dependencies make them a highly promising approach.

At a high level, GNNs make inferences by collecting the feature vectors of each of its neighbors and aggregating this information with their own features. Neighbors are defined by the edges that are established during the creation of the graph. This kind of message passing allows for information to continuously flow further in the graph (i.e. after 2 passes a node's embedding will reflect information from its neighbors and its neighbors' neighbors). This allows for information to propagate through multiple passes and generalize to unseen regions, an important property for large-scale datasets. After this process is complete, all the embeddings are fed into a small neural network to produce a prediction. Different GNN variants mainly vary in how they aggregate neighbor information.

The way the graph is constructed plays a crucial role in determining GNN performance, as it controls how information propagates between nodes. In our setup, each node represents a sampling location defined by its latitude and longitude, and edges were constructed using the k-nearest neighbors approach based on spatial proximity. Because we only kept the most recent samples in the dataset, no two data points are from the same node. We experimented with multiple neighborhood sizes (e.g., 5, 10, 15, 20) and found that k=5 consistently yielded the best results across all models. The k-nearest neighbor graph captures local spatial dependencies in arsenic concentrations while preserving meaningful geographic structure. This design implicitly encodes spatial autocorrelation, a fundamental property of environmental processes, without introducing excessive noise from distant locations.

Each GNN variant was chosen for its unique strengths and modeling capabilities. GCNs update each node's representation by averaging and transforming the features of its neighboring nodes, effectively capturing local context. GAT introduces attention coefficients to weight each node differently, allowing this model to focus more on influential neighbors. GraphSAGE only samples a fixed amount of neighbors and uses a small neural network to aggregate them during training, allowing this model to better capture dependencies between neighbors. GIN also aggregates information from neighbors but does so using a much more expressive update rule, enabling it to capture subtle structural differences. Because environmental datasets often vary in spatial density, feature correlations, and noise structure, different GNN architectures will capture different aspects of the data. We therefore evaluate multiple variants to understand which modeling assumptions best align with the spatial and geochemical structure of arsenic contamination.

4 Experiments

We conducted experiments in both log(x + 1)-scale and real-value scale. The log(x + 1)-scale evaluation assesses the model's ability to predict the magnitude of arsenic contamination, while the real scale prediction generates results suitable for direct application. In many scenarios, being able to accurately predict the magnitude of arsenic contamination is extremely important due to how the safety limit (10 µg/l) coincides with a power of 10 and the fact that arsenic concentrations are extremely skewed. In many environmental tasks, relative error often matters more than absolute error. For example, predicting 20 µg/L instead of 10 µg/L is more serious than predicting 110 µg/L instead of 100 µg/L. We applied a log(x + 1) transformation instead of a standard log(x) transform to handle samples with zero or very low arsenic concentrations. This approach avoids the undefined value at log(0) and prevents small concentrations from being mapped to disproportionately large negative values. By shifting the data by 1 µg/L before taking the logarithm, we preserve the relative differences among low values while compressing higher concentrations, resulting in a smoother and more stable target distribution for regression. When testing in real-value scale, we chose to remove all outliers above 100 μg/L as these outliers only accounted for around 1% of the dataset and heavily inflated the loss. These few large values unnecessarily contribute disproportionately huge gradients to the loss, and their rarity means that they serve very little real world application.

Training Configuration. All neural networks were trained using a 70-15-15 train/validation/test split. Predictor columns were scaled using a StandardScaler, which standardizes each feature to have zero mean and unit variance. Training used a batch size of 256 and validation and testing used a batch size of 512. All neural networks were trained using the Mean Absolute Error (MAE) loss function and the Adam optimizer with a learning rate of 1e-3 for 150 epochs. All tree-based models were trained using an 80-20 train/validation split, and when possible, were also trained with the Mean Absolute Error (MAE) loss function. For each of our hyperparameters, we tested at most 5 different values.

GNN Setup We added BatchNorm (Ioffe and Szegedy 2015) and residual connections (He et al. 2015) in between each layer to stabilize gradients and make training smoother. Furthermore, we used Jumping Knowledge (Xu et al. 2018b) to improve model performance. All the GNNs were trained for 500 epochs.

Evaluation Metrics We use MAE and MSE as the evaluation metrics. MSE is the standard loss function for regression-related tasks, although it can easily be skewed by outlier bad predictions. MAE places less weight on outlier bad predictions, which is desireable as we believe that there

will be cases where there are high levels of arsenic not easily explained by environment factors due to volatility of Arsenic levels.

Our experimental results in Table 1 show that GNN variants perform competitively with traditional tree-based models. In particular, GAT, GCN, and GraphSAGE achieve log-scale MAE values comparable to the widely adopted XGBoost baseline. Moreover, GAT, GraphSAGE, and GIN surpass XGBoost in real-scale MAE, with GraphSAGE improving performance by approximately 4.2%. Although their overall metrics are slightly lower in certain cases, GNN variants demonstrate strong robustness to outliers and superior stability in variance explanation. In contrast, both multilayer perceptrons and random forests exhibit noticeably weaker fits to the data, highlighting their limited capacity to model complex spatial and nonlinear relationships.

5 Analysis and Discussion

GNNs are particularly powerful for this problem because they explicitly model spatial autocorrelation through graph-based learning. Since arsenic contamination exhibits strong spatial dependence, with latitude and longitude already being highly informative features, the GNN leverages these relationships by propagating information across connected locations. This enables it to capture spatial structure and inter-site dependencies more effectively than traditional models, improving upon previous studies (Hu et al. 2022).

We chose XGBoost to represent the gradient boosted trees in error analysis. Figure 5 show that XGBoost tends to capture spatial and quantitative trends of arsenic pretty well, although its prediction errors do follow slight patterns. Because there are more points under the line y=x in Figure 5a, we know that XGBoost tends to underestimate the amount of arsenic concentration. Figure 4a further supports this observation, as the residual distribution extends more strongly into the negative range and exhibits a higher density of blue (under predicted) points compared to red (over predicted) ones. This figure also indicates that XGBoost tends to over-predict or under-predict in states in the upper Midwest, such as Minnesota, Iowa, and Illinois.

We chose GAT to represent the graph neural networks in error analysis. Figure 4 shows how GAT similarly tends to capture spatial and quantitative trends of arsenic pretty well, although patterns in its predictions are different from XGBoost. As shown in Figure 5b, GAT's predictions are divided more evenly along the line y=x, which is most likely one of the big reasons it is able to slightly outperform the gradient boosted trees. In this specific test set, there is a pretty big range of arsenic concentrations, and yet GAT is still able to estimate them pretty well. The spatial distribution error chart (Figure 4b is very similar to XGBoost, with a large majority of errors being concentrated in the north Midwest. Besides occasionally over predicting values a bit more than XGBoost, the range of error between these two models are very similar.

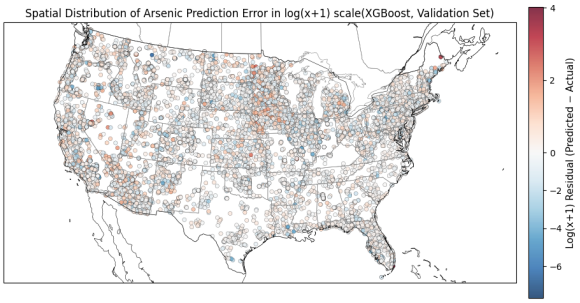
6 Conclusion and future work

Avenues for future work can include integrating more types of environmental datasets and testing the performance of modern tabular transformers. Additionally, using classification to predict arsenic concentrations in water is also beneficial as it directly models the exceedance of regulatory thresholds. However, this approach is difficult due to the heavy class imbalance and the skewed data. GNNs are relatively unexplored in environmental science, and applying them to key environmental challenges represents an exciting research direction.

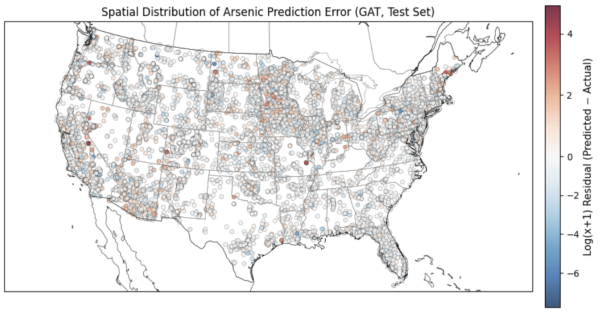
In this work, we assess a variety of machine learning models by investigating their ability to predict arsenic contamination in water across the contiguous United States. Our dataset also contains a variety of features aggregated from different public datasets, including substances in the soil and geologic data. Our findings suggest that GNN variants perform very well in environmental science tasks. By improving the robustness of arsenic prediction models while keeping maintaining low loss, this study contributes to more accurate and reliable arsenic prediction.

Table 1: Comparison of model performance across log-scale and real-scale metrics.

Model Type	**Model Name**	**MAE (µg/L)**		**MSE (µg/L^2)**		R^2	
		Log	**Real**	**Log**	**Real**	**Log**	**Real**
Simple Tree Models	Random Forest	0.538	3.966	0.605	70.032	0.439	0.298
Simple Neural Networks	Multilayer Perceptron	0.563	3.623	0.739	223.610	0.292	-1.187
Gradient Boosted Trees	XGBoost	0.527	3.399	0.675	83.958	0.374	0.158
	LightGBM	0.538	3.971	0.627	72.894	0.418	0.269
	CatBoost	0.533	3.433	0.680	87.133	0.369	0.126
GNN Variants	GCN	0.528	3.485	0.701	86.381	0.382	0.176
	GAT	0.522	3.319	0.682	75.039	0.370	0.215
	GraphSAGE	0.529	3.256	0.696	72.981	0.332	0.191
	GIN	0.546	3.282	0.759	76.801	0.326	0.201

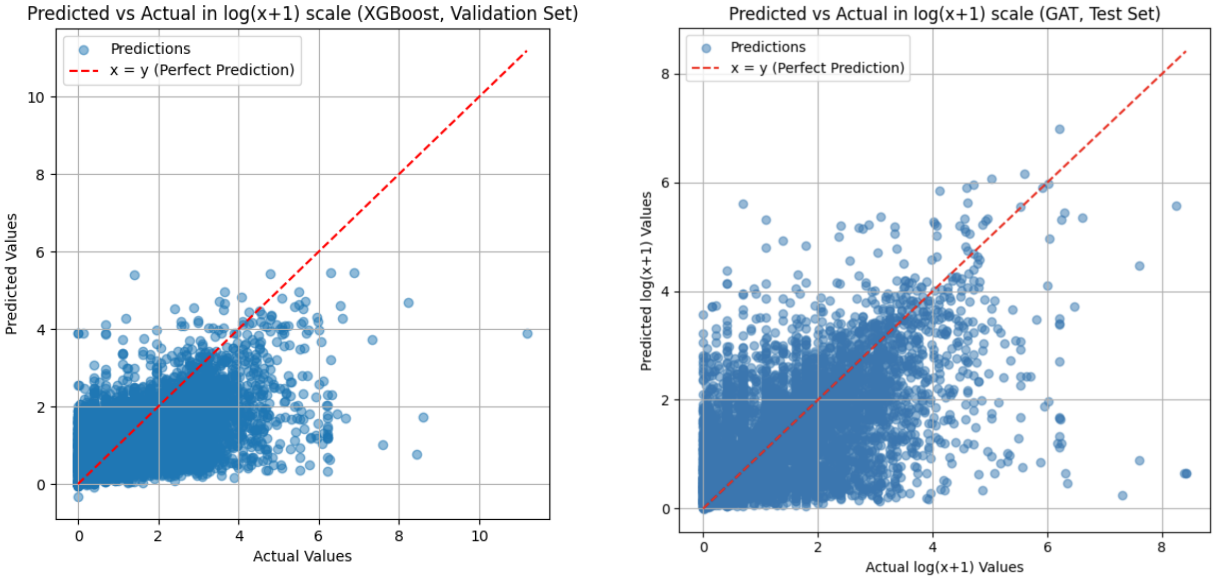


(a) Spatial distribution of arsenic prediction error (XGBoost).



(b) Spatial distribution of arsenic prediction error (GAT).

Figure 4: Spatial distribution of error in prediction of log-levels of Arsenic for the XGBoost and GAT models



(a) Predicted vs Actual in log(x+1) scale (XGBoost).

(b) Predicted vs Actual in log(x+1) scale (GAT).

Figure 5: Predicted vs actual log levels of Arsenic from XGBoost and GAT based models.

References

Ayotte, J. D.; Medalie, L.; Qi, S. L.; Backer, L. C.; and Nolan, B. T. 2017. Estimating the High-Arsenic Domestic-Well Population in the Conterminous United States. *Environmental Science & Technology*.

Chattopadhyay, A.; Singh, A. P.; Kumar, S.; Pati, J.; and Rakshit, A. 2023. The machine learning and geostatistical approach for assessment of arsenic contamination levels using physicochemical properties of water. *Water science and technology : a journal of the International Association on Water Pollution Research*, 88 3: 595–614.

Chen, T.; and Guestrin, C. 2016. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Environmental Protection Agency; and United States Geological Survey. ???? Water Quality Portal.

Hamilton, W. L.; Ying, Z.; and Leskovec, J. 2017. Inductive Representation Learning on Large Graphs. *ArXiv*, abs/1706.02216.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Hu, X. C.; Dai, M. Q.; Sun, J. M.; and Sunderland, E. M. 2022. The Utility of Machine Learning Models for Predicting Chemical Contaminants in Drinking Water: Promise, Challenges, and Opportunities. *Current Environmental Health Reports*, 10: 45 – 60.

Ioffe, S.; and Szegedy, C. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *ArXiv*, abs/1502.03167.

Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; and Liu, T.-Y. 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Neural Information Processing Systems*.

Kipf, T.; and Welling, M. 2016. Semi-Supervised Classification with Graph Convolutional Networks. *ArXiv*, abs/1609.02907.

Lombard, M. A.; Bryan, M. S.; Jones, D. K.; Bulka, C. M.; Bradley, P. M.; Backer, L. C.; Focazio, M. J.; Silverman, D. T.; Toccalino, P. L.; Argos, M.; Gribble, M. O.; and Ayotte, J. D. 2021. Machine Learning Models of Arsenic in Private Wells Throughout the Conterminous United States As a Tool for Exposure Assessment in Human Health Studies. *Environmental science & technology*.

Mason, G. T.; and Arndt, R. E. ???? Mineral Resources Data System (MRDS). ISBN: 9780607855081 ISSN: 2327-638X Publication Title: Data Series.

Ostroumova, L.; Gusev, G.; Vorobev, A.; Dorogush, A. V.; and Gulin, A. 2017. CatBoost: unbiased boosting with categorical features. In *Neural Information Processing Systems*.

Podgorski, J. E.; and Berg, M. 2020. Global threat of arsenic in groundwater. *Science*, 368: 845 – 850.

Podgorski, J. E.; Wu, R.; Chakravorty, B.; and Polya, D. A. 2020. Groundwater Arsenic Distribution in India by Machine Learning Geospatial Modeling. *International Journal of Environmental Research and Public Health*, 17.

Staff, S. S. ???? Gridded National Soil Survey Geographic (gNATSGO) Database for the Conterminous United States.

Subudhi, S.; Pati, A. K.; Bose, S.; Sahoo, S.; Pattanaik, A.; and Acharya, B. M. 2025. Integrating Boosted learning with Differential Evolution (DE) Optimizer: A Prediction of Groundwater Quality Risk Assessment in Odisha. *ArXiv*, abs/2502.17929.

Sun, A. Y.; Jiang, P.; Yang, Z.; Xie, Y.; and Chen, X. 2022. A graph neural network (GNN) approach to basin-scale river network learning: the role of physics-based connectivity and data fusion. *Hydrology and Earth System Sciences*.

Tan, Z.; Yang, Q.; and Zheng, Y. 2020. Machine Learning Models of Groundwater Arsenic Spatial Distribution in Bangladesh: Influence of Holocene Sediment Depositional History. *Environmental Science and Technology*.

Truong, H.; Tello, A.; Lazovik, A.; and Degeler, V. 2023. Graph Neural Networks for Pressure Estimation in Water Distribution Systems. *Water Resources Research*, 60.

Velickovic, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio', P.; and Bengio, Y. 2017. Graph Attention Networks. *ArXiv*, abs/1710.10903.

Wu, R.; Alvareda, E.; Polya, D. A.; Blanco, G.; and Gamazo, P. 2021. Distribution of Groundwater Arsenic in Uruguay Using Hybrid Machine Learning and Expert System Approaches. *Water*.

Wu, R.; Richards, L. A.; Roshan, A.; and Polya, D. A. 2023. Artificial Intelligence Modelling to Support the Groundwater Chemistry-Dependent Selection of Groundwater Arsenic Remediation Approaches in Bangladesh. *Water*.

Xu, K.; Hu, W.; Leskovec, J.; and Jegelka, S. 2018a. How Powerful are Graph Neural Networks? *ArXiv*, abs/1810.00826.

Xu, K.; Li, C.; Tian, Y.; Sonobe, T.; ichi Kawarabayashi, K.; and Jegelka, S. 2018b. Representation Learning on Graphs with Jumping Knowledge Networks. *ArXiv*, abs/1806.03536.

Zha, Y.; and Yang, Y. 2024. Innovative graph neural network approach for predicting soil heavy metal pollution in the Pearl River Basin, China. *Scientific Reports*, 14: 16505.

Zhao, Z.; Kumar, A.; and Wang, H. 2024. Predicting Arsenic Contamination in Groundwater: A Comparative Analysis of Machine Learning Models in Coastal Floodplains and Inland Basins. *Water*.

A Data Appendix

Water Quality Portal (WQP) The WQP (Environmental Protection Agency and United States Geological Survey) is a database of water quality measurements maintained by the National Water Quality Monitoring Council, which is part of the US Government. The dataset includes data from the United States Geological Survey (USGS), the Environmental Protection Agency (EPA), and other state, federal, tribal, and local agencies. It includes a wide variety of nation-wide measurements associated with water, but most importantly for this project, it includes a wide array of measurements of Arsenic quantities in water supplies.

Gridded National Soil Survey Geological Database (gNATSGO) The gNATSGO dataset (Staff) includes precise maps of the continental US for soil horizons and other geological properties that could capture or expel contaminants such as Arsenic. Information includes soil composition, soil pH, and soil topology.

Mineral Resources Data System (MRDS) The MRDS dataset (Mason and Arndt) includes information about mineral deposits in North America, but primarily the continental US. The data was collected in 1995, and there are more than 100k observations. Information includes the type and general location of minerals present.

B Code Appendix

Code Access

We will provide the full, public-access, open license code for the camera-ready version of the paper. This includes instructions for data sources, data aggregation and pre-processing, as well as all models.

Hardware Resources

All experimentation was completed using less than 5 non-continuous hours of time on an environment with the following specs accessed through the Runpod service.

Resource	Quantity
GPU	A100 SXM (x1)
vCPU	32
Memory	251 GB
Disk	30 GB

Table 2: Specs for computational environment

Software Resources

For this project we rely primarily on the most recent versions of numpy/pandas for data processing and torch/scikit for generating models. A full list of dependencies will be provided alongside the full source code in the camera-ready version of the paper.

C Reproducability Checklist

This paper:

1. Includes a conceptual outline and/or pseudocode description of AI methods introduced: **Yes**
2. Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results **Yes**
3. Provides well marked pedagogical references for less-familiar readers to gain background necessary to replicate the paper **Yes**

Does this paper make theoretical contributions? **No**
Does this paper rely on one or more datasets? **Yes**
If yes, please complete the list below.

1. A motivation is given for why the experiments are conducted on the selected datasets **Yes**
2. All novel datasets introduced in this paper are included in a data appendix. **Yes**
3. All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes. **Yes**
4. All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations. (yes/no/NA)
5. All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available. **Yes**
6. All datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisficing. **N/A**

Does this paper include computational experiments? **Yes**
If yes, please complete the list below.

1. This paper states the number and range of values tried per (hyper-) parameter during development of the paper, along with the criterion used for selecting the final parameter setting. **Partial**
2. Any code required for pre-processing data is included in the appendix. **Yes**
3. All source code required for conducting and analyzing the experiments is included in a code appendix. **Yes**
4. All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes. **Yes**
5. All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from **Yes**
6. If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results. **Yes**
7. This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks. **Yes**
8. This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics. **Yes**
9. This paper states the number of algorithm runs used to compute each reported result. **Yes**
10. Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information. **Yes**
11. The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank). **N/A**
12. This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments. **Partial**