REMPLEX
CENTER FOR THE REMEDIATION OF COMPLEX SITES
@PNNL

Enhancing Groundwater Contamination Monitoring at the Hanford Site with Sequential Deep Neural Network

1. Dr. Hardik Gohel, University of Houston-Victoria, Victoria, TX, 77901
2. Dr. Hilary Emerson, Pacific Northwest National Laboratory, Richland, WA 99354
3. Dr. Daniel I. Kaplan, Savannah River National Laboratory, Aiken, SC 29808

UHV University of Houston - Victoria

U.S. DEPARTMENT OF ENERGY | OFFICE OF ENVIRONMENTAL MANAGEMENT

AIM

The specific aims of this proposed research are to apply novel Machine Learning (ML) and Deep Learning (DL) techniques to modeling groundwater contaminant transport to aid remediation efforts at the Hanford Site, as seen in Figure 1.

OBJECTIVES

- ❖ Develop, test, and optimize ML/DL algorithms for spatiotemporal modeling of groundwater contamination.
- ❖ Establish a viable framework for normalizing, resampling, and dealing with sparsity in Hanford Site data.
- ❖ Automate the training and testing of ML/DL models.
- ❖ Use the developed methods to optimize groundwater sampling routines and pump and treat operations for efficient remediation of hexavalent chromium.

TECHNIQUES INVESTIGATED

Deep Learning-based models are critical for modeling Hanford site data due to the large number of contaminants and wells and can augment existing decision support systems. The complete framework for preprocessing and modeling groundwater data, as shown in Figure 3, will be used to identify patterns in training data that map the features to the target feature, Cr(VI) and features a sequential DNN architecture. The proposed framework can then be used in tandem with existing methods to further optimize treatment operations.

CONTACT

Hardik A. Gohel at GohelH@uhv.edu
Hilary Emerson at hilary.emerson@pnnl.gov
Daniel I Kaplan at Daniel.Kaplan@uga.edu

HANFORD NUCLEAR SITE

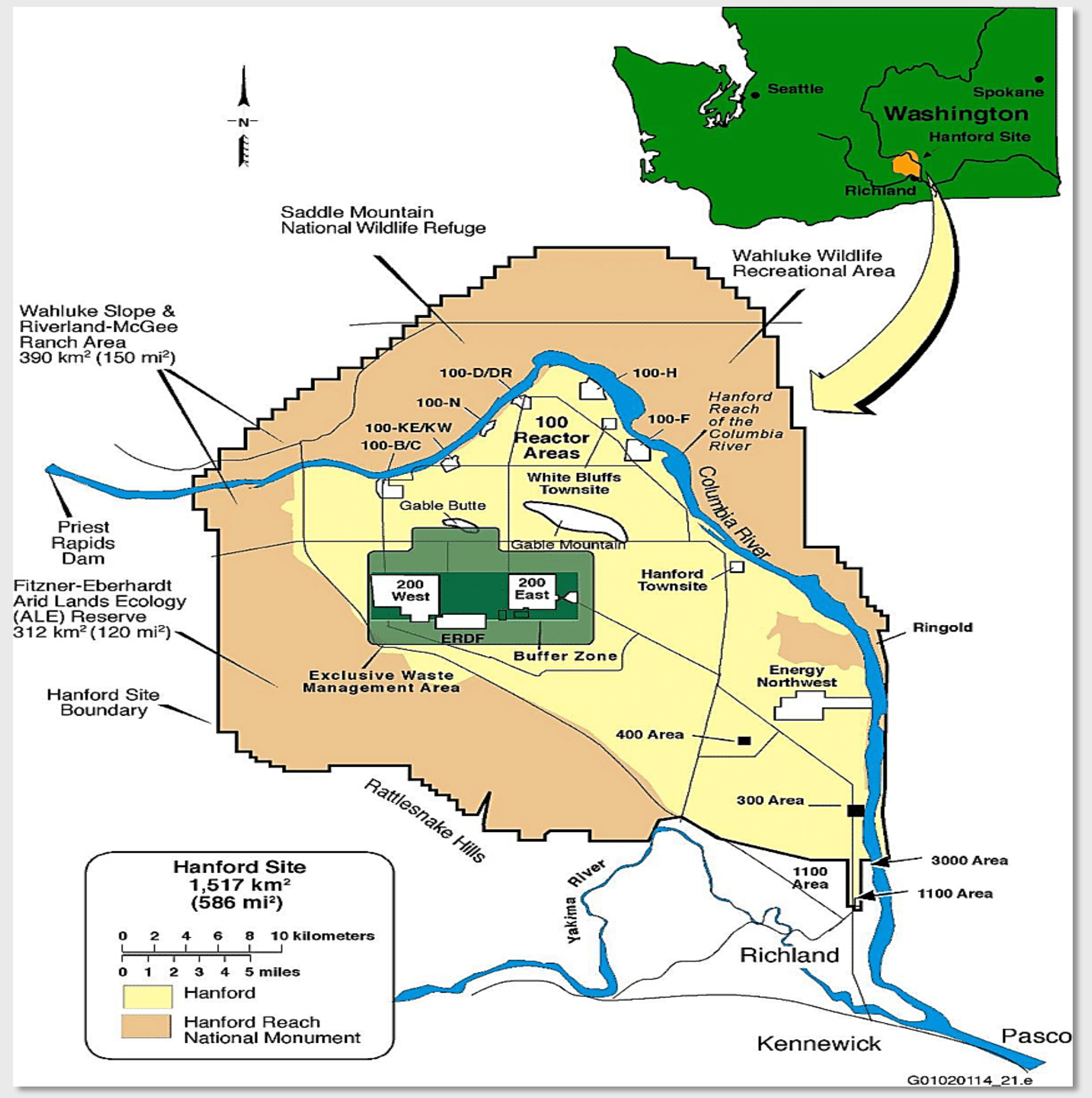

Figure 1. Map of the Hanford Site, including the 100 Reactor Area, where data from this study was derived.

INTRODUCTION

- ➢ ML/DL methods show promise as powerful support systems for risk assessment, forecasting, and policy optimization.
- ➢Traditional models have historically struggled to capture the variability in data for complex sites as observed in measurements recorded across time at the Hanford Site.

MATERIALS AND METHODS

1. Physical, hydraulic, and geochemical data are retrieved from the PNNL Phoenix database.
2. Missing data are imputed with the K-Nearest Neighbor algorithm by using positional encodings of each datapoint.
3. Outliers in the data are filtered out by computing the Z-score and removing any rows of the dataset containing a point with Z-score greater than 3.
4. Data are resampled to daily intervals based on the average distance between observations.
5. Linear interpolation is used to fill the gaps created by the resampling interval.
6. Data are split into training and testing sets for use with the autoencoder.
7. Bayesian Search Optimization is used to find the ideal hyperparameters for the sequential deep neural network (DNN) as well as identify ideal resampling period, scaling method, lag, and lookahead values.
8. The optimized DNN performance is compared to the default hyperparameter model as well as four common regression models on 5 different target features.

MODEL DESIGN

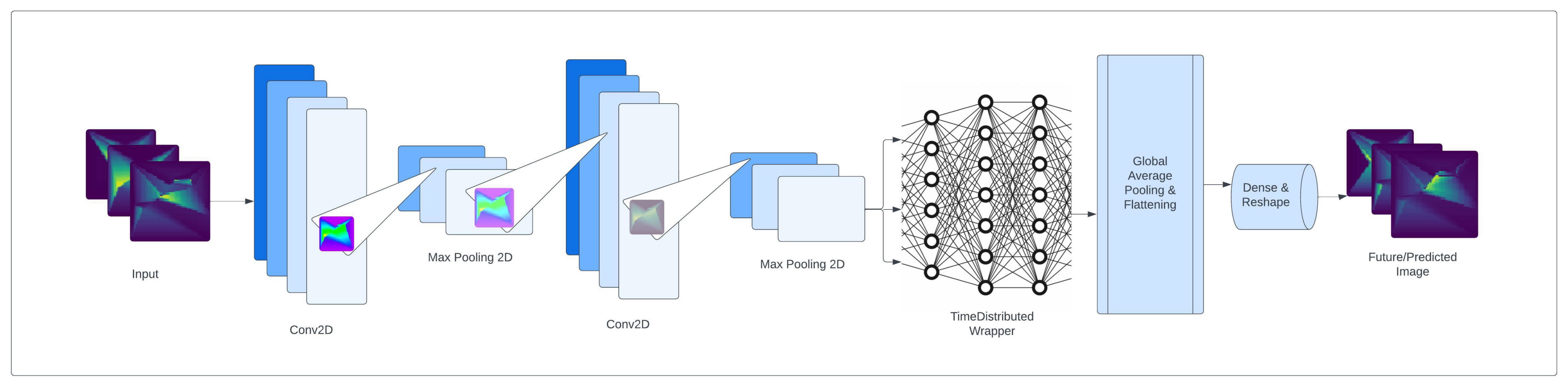
Input
Conv2D
Max Pooling 2D
Conv2D
Max Pooling 2D
TimeDistributed Wrapper
Global Average Pooling & Flattening
Dense & Reshape
Future/Predicted Image

Figure 2. Sequential VGG-16 hybrid architecture for Groundwater contaminant prediction at Hanford site.

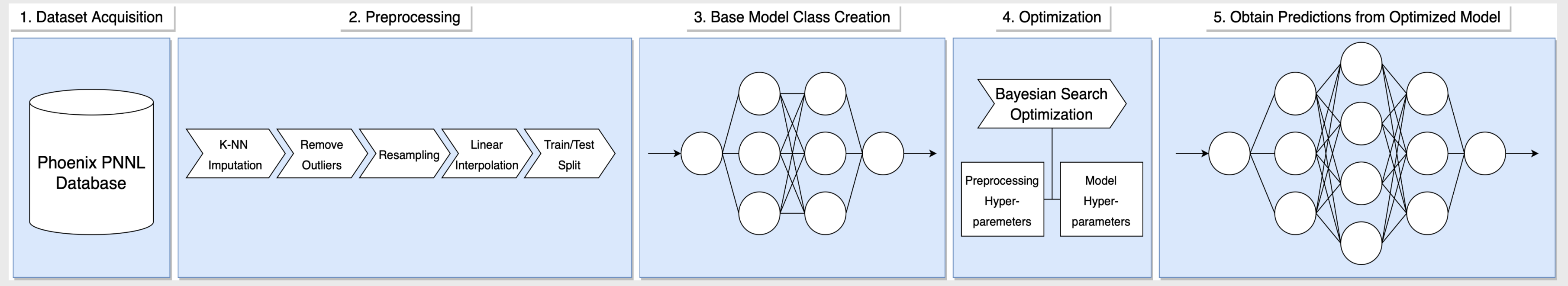
1. Dataset Acquisition
Phoenix PNNL Database
2. Preprocessing
K-NN Imputation
Remove Outliers
Resampling
Linear Interpolation
Train/Test Split
3. Base Model Class Creation
4. Optimization
Bayesian Search Optimization
Preprocessing Hyper-paremeters
Model Hyper-parameters
5. Obtain Predictions from Optimized Model

Figure 3. High-level data modeling pipeline showing retrieval of data from an online database, preprocessing steps, model creation, and optimization.

RESULTS AND DISCUSSION

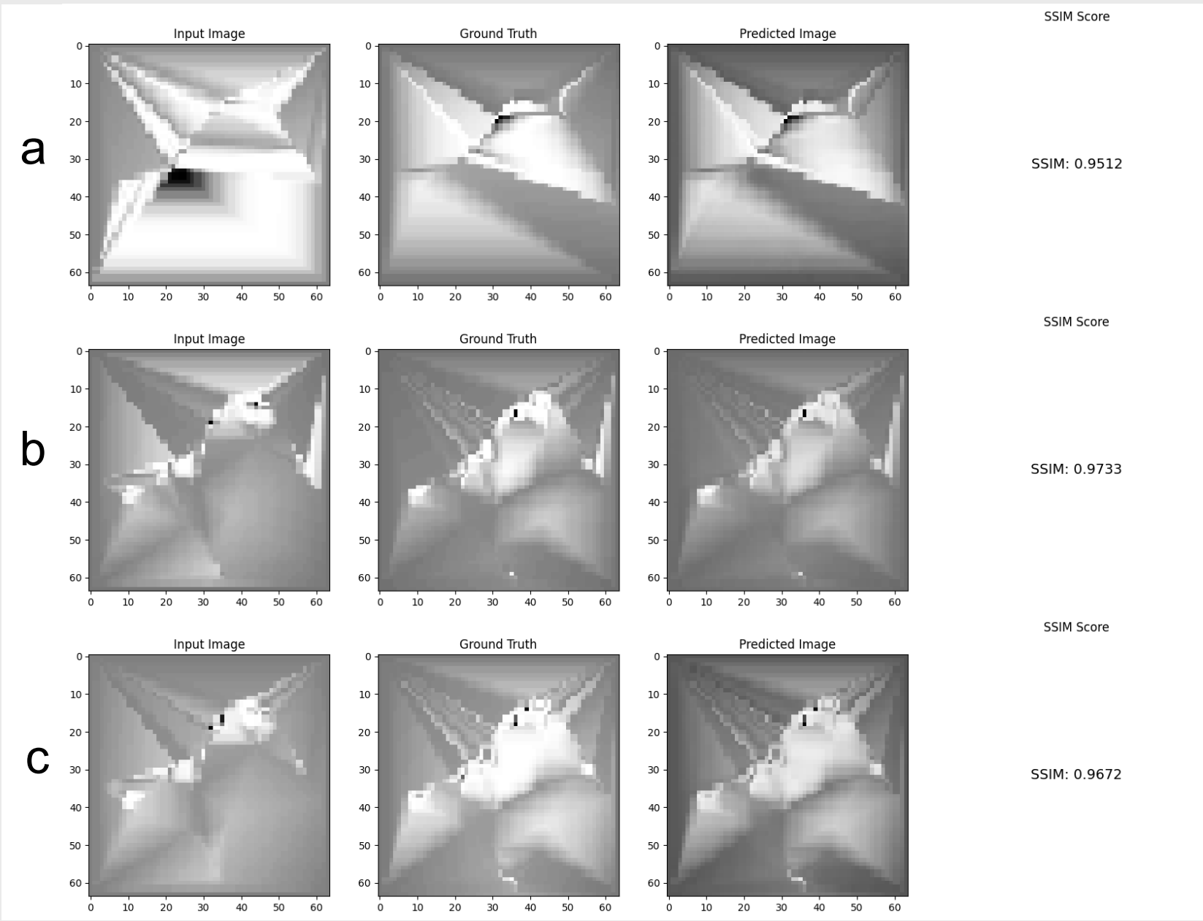

Figure 4. SeqVGG16 architecture with next frame prediction and structural similarity index (SSIM)

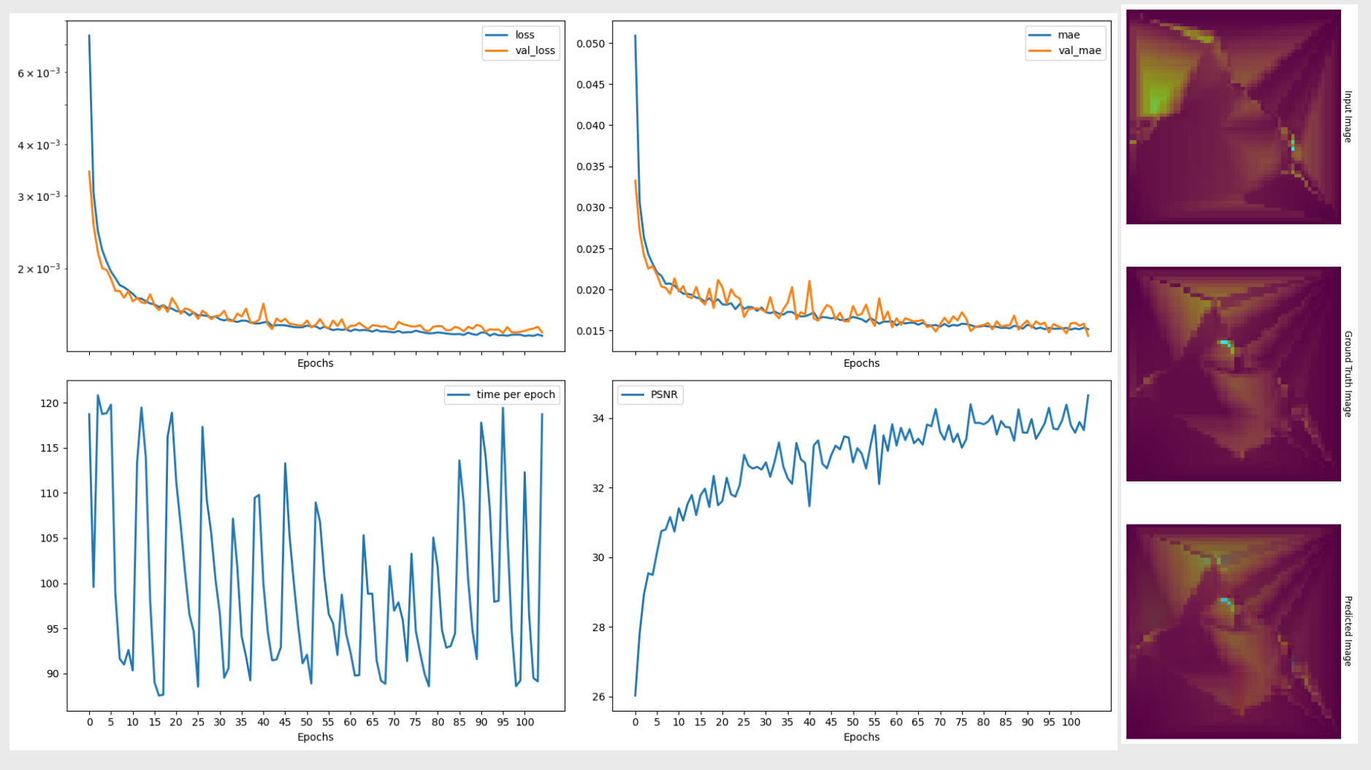

Figure 5. SeqVGG16 architecture implementation loss (MSE), MAE, PSNR plots; 10th frame Prediction vs. GroundTruth

Table 1

Prediction Metrics on Target Features (Testing Data)

Model	Conductivity		pH		Turbidity	
	MSE, RMSE	R^2	MSE, RMSE	R^2	MSE, RMSE	R^2
RCNN (Opt.)	0.041, 0.20	0.94	0.031, 0.18	0.94	0.0020, 0.045	0.95
RCNN (Base)	0.17, 0.41	0.70	0.13, 0.36	0.64	0.011, 0.11	0.38
MLR	1.0, 1.0	0.027	0.33, 0.58	0.27	0.046, 0.22	0.36
RF	0.60, 0.78	0.40	0.27, 0.52	0.41	0.032, 0.18	0.049
CB	0.53, 0.73	0.47	0.24, 0.49	0.48	0.030, 0.17	0.11
ET	0.75, 0.86	0.26	0.33, 0.58	0.27	0.039, 0.20	0.14

Table 2

Model	Turbidity		Oxidation-Reduction Potential		Cr(VI) Concentration	
	MSE, RMSE	R^2	MSE, RMSE	R^2	MSE, RMSE	R^2
RCNN (Opt.)	1.3, 1.1	0.97	0.0030, 0.055	0.91	0.046, 0.21	0.95
RCNN (Base)	1.1, 1.1	0.97	0.0090, 0.095	0.54	0.16, 0.40	0.83
MLR	32, 5.7	0.060	0.060, 0.25	0.063	0.97, 0.98	0.033
RF	11, 3.3	0.68	0.049, 0.22	0.025	0.65, 0.81	0.35
CB	11, 3.3	0.68	0.044, 0.21	0.17	0.61, 0.78	0.39
ET	12, 3.5	0.64	0.058, 0.24	0.12	0.79, 0.89	0.21

- ➢ The optimized sequential DNN outperforms all other models including the base model and Multiple Linear Regression, Random Forest, CatBoost, and Extra Trees Regressor.
- ➢ Bayesian Search Optimization successfully tunes model hyperparameters and finds ideal preprocessing techniques for each target feature.
- ➢ The addition of 2D CNN layers improves the performance of the sequential DNN.

BENEFITS

- ➢ Enhancement of long-term groundwater contamination monitoring
- ➢ Rectify sparse data using machine learning techniques in groundwater datasets
- ➢ Establish scalable framework for modeling contaminant behavior in larger areas of concern
- ➢ The proposed method can be used to enhance groundwater contaminant monitoring and optimize sampling routines for the U.S. Department of Energy Office of Environmental Management sites

ANTICIPATED OUTCOMES

The goal of this research is to develop models that accurately model groundwater contaminant fate and transport over time to assist in the creation of more preemptive monitoring and treatment operations and; therefore, reduce the costs related to long-term monitoring of subsurface contamination at DOE sites. The use of Deep Learning models, especially those based on artificial neural networks, have the capability to further enhance contaminant modeling by overcoming many of the limitations associated with traditional statistical models.

CONCLUSIONS

Deep Learning-based methods that use an autoencoder architecture are complementary to traditional regression-based and tree-based models such as standard linear regression, gradient-boosted regression, decision tree regression, or random forest regression. The proposed preprocessing framework provides a means to transform the data into a format suitable for time-series based models by regularizing, normalizing, and handling sparse data and shows that areas of concern such as the Hanford Site may benefit from regular sampling routines to increase the predictive performance of existing and future models.

ACKNOWLEDGEMENT

The research is supported by the Department of Energy (DoE) Minority Serving Institution Partnership Program under BOA #875/TOA No. 0000525185. DoE-Environmental Management also supported this work under Award Number DE-EM0005228 to the University of Georgia Research Foundation.

REFERENCES

1. Tao, H., Hameed, M. M., Marhoon, H. A., Zounemat-Kermani, M., Heddam, S., Sungwon, K., Sulaiman, S. O., Tan, M. L., Sa'adi, Z., Mehr, A. D., Allawi, M. F., Abba, S. I., Zain, J. M., Falah, M. W., Jamei, M., Bokde, N. D., Bayatvarkeshi, M., Al-Mukhtar, M., Bhagat, S. K. Yaseen, Z. M. (2022). Groundwater level prediction using machine learning models: A comprehensive review. In *Neurocomputing* (Vol. 489, pp. 271–308).
2. Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., & Tabona, O. (2021). A survey on missing data in machine learning. *Journal of Big Data*, *8*(1).
3. Turner, R., Eriksson, D., McCourt, M., Kiili, J., Laaksonen, E., Xu, Z., & Guyon, I. (2021). *Bayesian Optimization is Superior to Random Search for Machine Learning Hyperparameter Tuning: Analysis of the Black-Box Optimization Challenge 2020*.
4. Ahsan, M., Mahmud, M., Saha, P., Gupta, K., & Siddique, Z. (2021). Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance. *Technologies*, *9*(3), 52.
5. Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. ArXiv, 1409.1556.