FDLPhysics-Informed Surrogate Modeling for Supporting Climate Resilience at Groundwater Contamination Sites – 23123

Aurelien Meray *, Lijing Wang **, Takuya Kurihana ***, Ilijana Mastilovic ****, Satyarth Praveen *****, Zexuan Xu *****, Alexander Lavin ******, Milad Memarzadeh *******, Haruko Wainwright ********
* Applied Research Center - FIU
** Stanford University
*** University of Chicago
**** University of Wisconsin - Milwaukee
***** Lawrence Berkeley National Laboratory
****** Pasteur Labs & ISI
******* NASA Ames Research Center
******** Massachusetts Institute of Technology

ABSTRACT

Soil and groundwater contamination is a widespread problem across the world. Contaminated sites often require decades to remediate or to monitor natural attenuation. Climate change exacerbates the problem because extreme precipitation and/or shifts in precipitation/evapotranspiration regimes can remobilize contaminants and proliferate affected groundwaters. With tools for fast and reliable contaminant plume prediction under future climate scenarios site managers and decision makers can evaluate the potential consequences and take rapid actions. Recent developments in machine learning introduce the Fourier Neural Operator (FNO) which has shown to be very effective in learning Partial Differential Equations (PDEs).

In this study, we utilize two versions of the FNO that are enhanced with U-Net architectures to model multiple resolutions: UFNO-3D and UFNO-2D. With these networks, we create surrogate flow and transport models under different Coupled Model Intercomparison Project 5 (CMIP5) climate scenarios. We use the Department of Energy's Savannah River Site F-Area – which has significant groundwater contamination – as a testbed for demonstrating this capability and evaluating the combined impact of uncertain subsurface properties and recharge rates from different climate projections. We train our UFNOs based on various loss terms that include both data-driven factors and physical boundary constraints. Results show that we can predict 1) contaminant concentration 2) hydraulic head 3) darcy's velocity from 1954 to 2100 accurately with different climate and subsurface inputs. Larger recharge rates have a complex impact on plumes with both remobilization and dilution of the contaminants.

In parallel, to scale such climate resilience assessment at any site, we develop an unsupervised approach to reduce the dimensionality of the vast historical and projected climate data by identifying similar climatic regions. We develop two convolutional autoencoders that are combined with 1) K-Means clustering or 2) an online clustering based on the Sinkhorn-Knopp algorithm, across the United States to capture unique climate patterns from the CMIP5 model. The unsupervised climate data clustering helps us return reliable future recharge rate projections immediately without querying large climate datasets. We hope this work can support the next level of environmental remediation modeling development under climate change.

INTRODUCTION

Contamination of groundwater poses a major health risk to millions of people across the world. The Global Atlas of Environmental Justice documented nearly 4,000 contaminated sites across the globe, yet many are undocumented and new sites continuously arise and propagate. These sites vary in size and significance, ranging from areas contaminated with toxic materials from past industrial or mining activities to nuclear waste storage, and can pose a variety of health and environmental hazards. Toxic contaminants can reach into nearby groundwater or surface water, then into a human drinking water supply, and they can also be taken up by plants and animals. They must be carefully managed to prevent hazardous materials from causing harm to humans, wildlife, and ecological systems. Additional challenges arise and exacerbate risks due to global climate change in that uncertain climate scenarios can impact the water balance in dynamic or disruptive ways and remobilize contaminants.

Physical simulation enables us to predict the spatiotemporal groundwater flow and contamination transport by various recharge rates (i.e., a net flow between precipitation and evapotranspiration) under warming climate scenarios. We incorporate multi-scale uncertainties in physical simulations: uncertain global climate projections are in the order of 10-100km while uncertain local soil and subsurface properties are limited in local-scale in 1-10km. We run groundwater flow and contaminant transport physics simulations using Amanzi [1] at the testbed: the Department of Energy's Savannah River Site F-Area [2], [3] with uncertain subsurface, soil, and climate factors. The Amanzi model takes soil and subsurface properties and recharge rates, and then simulates contaminant concentrations in 2D spatial cross-section. Here, temporal changes of recharge rates determined by a net flow from precipitation and evapotranspiration to aquifers, directly relate to climate factors. Recharge rates indicate how much water seeps into the ground to replenish underground aquifers. These time-varying recharge rates directly relate to the climate factors: precipitation and evapotranspiration. The Amanzi model outputs contaminant concentrations in 2D spatial cross-section from 1954 to 2100.

However, solving a complex groundwater flow and contaminant transport model such as Amanzi takes a long time to run even with supercomputers. Meanwhile, the global climate models [4] are on a different spatial scale than the local testbed. Therefore, this study leverages advances in machine learning (ML) to tackle the multi-scale problem via Multi-scale Digital Twin with two folds: 1) **ML-based surrogate model**: we develop a ML surrogate model using neural operator learning [5]–[7] to solve Partial Differential Equations (PDEs) for complex flow and transport physical simulations rapidly, 2) **Unsupervised climate data clustering**: we perform an unsupervised clustering of climate data from all different global climate models and classify the United States into a set of representative climate regions to immediately access the climate data for any target location. Climate data clustering quantifies uncertain projections of precipitation and evapotranspiration (ET), thereby providing climate inputs for groundwater systems without downloading large amounts of climate data.

MULTI-SCALE DIGITAL TWIN

The primary component of our multi-scale digital twin are designed to 1) reduce the computational burden of flow and transport simulations approximately $O(100–1000)$ times faster by creating a surrogate model for groundwater system, and 2) reduce the dimensionality of the vast historical and projected climate data by grouping similar climate regions to help us quantify reliable future recharge rate projections immediately without downloading large climate datasets. Having the digital twin that evaluates spatial-temporal contaminant variations instantly with many possible combinations of climate and subsurface uncertainty, site managers and decision makers can evaluate the potential impact of groundwater contamination and take rapid actions.

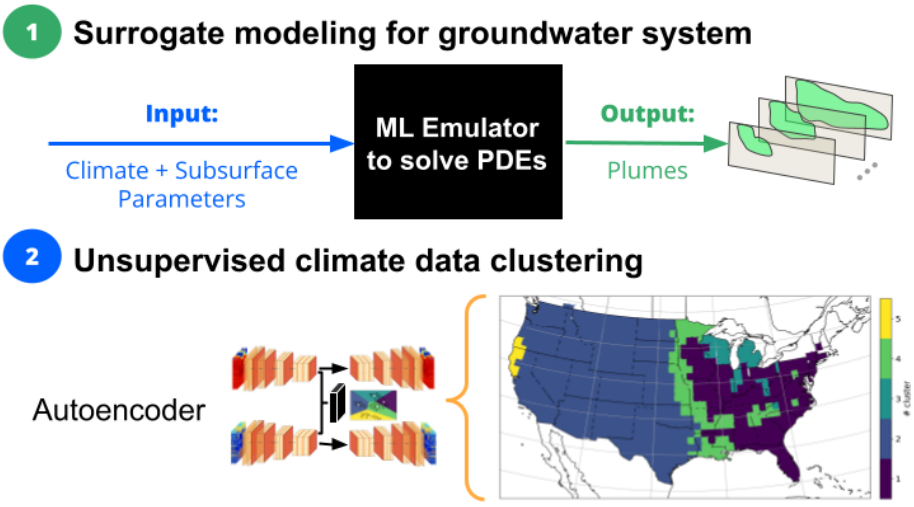


Figure 1. Digital Twin composed of two ML algorithms

ML-BASED SURROGATE MODELING FOR GROUNDWATER SYSTEM

The intersection of physics and ML provides a rich space to build models with the advantages of high-dimensional data-driven learning while maintaining (and even guiding) physical constraints and laws. One popular method is called neural operator learning [5], [6], [8], using neural networks to learn mesh-independent, resolution-invariant solution operators for PDEs. Using neural operator learning, we aim to learn a fast surrogate model for groundwater systems with the physical simulation datasets from the Amanzi model. The input of our datasets includes uncertain soil, subsurface, and climate properties (precipitation and ET): $m(\mathbf{x}, t)$ for a groundwater system, the output includes the flow and transport properties $y(\mathbf{x}, t)$ such as the spatiotemporal contaminant concentration. $\mathbf{x}$ is the location vector in 2D spatial cross-section, t is the time variable ranging from 1954 to 2100.

Architectures: U-FNO-3D and U-FNO-2D

We present two different neural operator architectures (U-FNO-3D and U-FNO-2D) to predict spatial-temporal contaminant concentration and groundwater flow properties. Both of these two architectures have Fourier Neural Operator (FNO) with enhanced U-Net architecture (U-FNO) [5], [7]. FNO has linear transformations in Fourier frequency modes so that FNO can represent any resolution with additional back transformation. U-FNO adds an additional U-Net architecture for each Fourier layer, achieving a lower test error for multi-phase flow predictions.

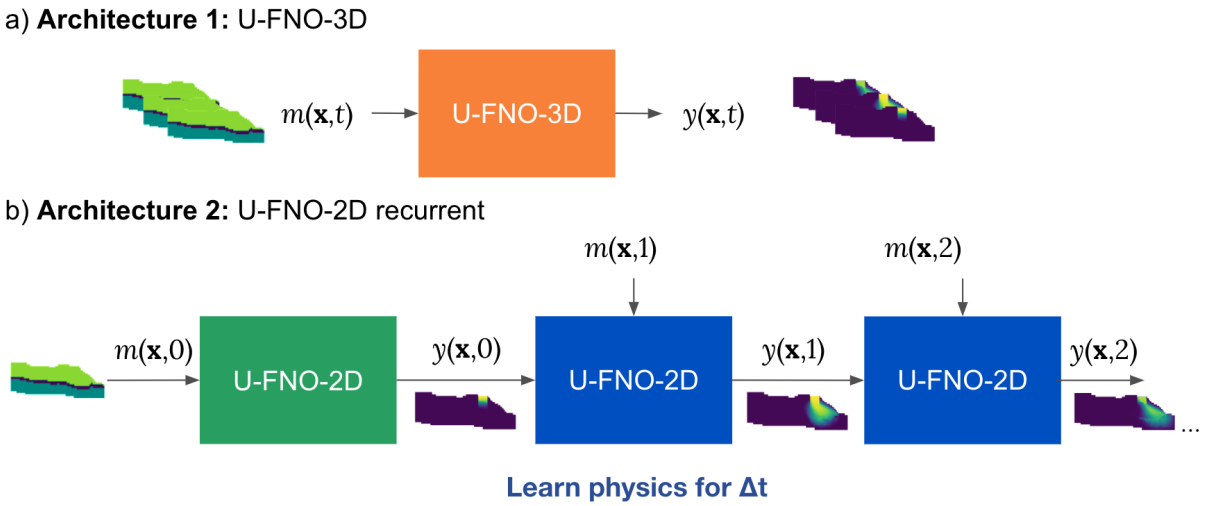


Figure 2. U-FNOs architectures

Our two architectures address the temporal dimension differently. U-FNO-3D takes all input time-series groundwater model parameters $m(\mathbf{x}, t)$ and predicts all contaminant concentrations for different time steps together $y(\mathbf{x}, t)$. U-FNO-2D uses a recurrent network: the input and output for each time t^* become $\{m(\mathbf{x}, t^*), y(\mathbf{x}, t^* - \Delta t)\}$ and $y(\mathbf{x}, t^*)$. U-FNO-2D learns physics between every fixed time interval Δt. The initial stage $t = 0$ for U-FNO-2D does not have the prediction from the previous step. Then, we train another U-FNO-2D model (the green block in Figure 2b for the initial stage $t = 0$ with the input $m(\mathbf{x}, 0)$ and the output $y(\mathbf{x}, 0)$. The advantage of U-FNO-2D is that we can preserve the time dependency, where the output $y(\mathbf{x}, t^*)$ is only determined by the input before and at time t^*. For example, if we want to investigate how future input recharge rates change output contaminant concentrations, then logically the previous output should stay the same, only the future concentration changes. U-FNO-3D cannot guarantee this time dependency. However, this additional time dependency makes the training of U-FNO-2D practically harder and hence the training takes longer time. There are also accumulated errors through the recurrent neural network.

Hybrid physics-based and data-driven loss functions

We introduce four different loss functions that include both data-driven factors and physical boundary constraints. Our surrogate model predicts transient flow $\hat{h}(\mathbf{x}, t)$ and transport $\hat{c}(\mathbf{x}, t)$ properties $\hat{y}(\mathbf{x}, t) = \{\hat{h}(\mathbf{x}, t), \hat{c}(\mathbf{x}, t)\}$ where the ground truth is from numerical solvers. Therefore the designed loss functions target the data-driven mismatch between predictions $\hat{y}(\mathbf{x}, t)$ and $y(\mathbf{x}, t)$, and more interestingly, the physical constraints for solving PDEs such as boundary conditions.

Mean Relative Error: We first quantify the data-driven mismatch using the mean relative error~(MRE) between ℓ_2 norm:

$$\mathcal{L}_{MRE}(y, \hat{y}) = \frac{\|y - \hat{y}\|_2}{\|y\|_2} \quad \text{(Eq. 1)}$$

Spatial derivatives: Additional mismatch on first derivatives in the horizontal direction x and the vertical direction z are also included.

$$\mathcal{L}_{der}(y,\hat{y}) = \frac{\|\partial y/\partial x - \partial \hat{y}/\partial x\|_2}{\|\partial y/\partial x\|_2} + \frac{\|\partial y/\partial z - \partial \hat{y}/\partial z\|_2}{\|\partial y/\partial z\|_2} \quad \text{(Eq. 2)}$$

Spatial derivatives on the contaminant boundary: Maximum Contaminant Level (MCL) is the highest level of a contaminant that is allowed in drinking water recommended by the Environmental Protection Agency (EPA) [2]. Therefore, predicting the boundary of contaminant with higher concentration than the MCL is essential for site managers to protect water supply. We add first derivatives on the contaminant boundary $c \geq MCL$.

$$\mathcal{L}_{conc}(y,\hat{y}) = \frac{\|\partial c'/\partial x - \partial \hat{c'}/\partial x\|_2}{\|\partial c'/\partial x\|_2} + \frac{\|\partial c'/\partial z - \partial \hat{c'}/\partial z\|_2}{\|\partial c'/\partial z\|_2},$$
$$\text{where } c' = \begin{cases} 0, & c < MCL \\ 1, & c \geq MCL \end{cases}, \quad \hat{c'} = \begin{cases} 0, & \hat{c} < MCL \\ 1, & \hat{c} \geq MCL \end{cases} \quad \text{(Eq. 3)}$$

Physics-informed boundary conditions: We add no flow boundary condition constraints in loss functions using physics-informed neural networks[9] to help solving the PDEs. The boundary of the spatial domain D is ∂D.

$$\mathcal{L}_{BC}(\hat{y}) = \|\hat{h}|_{\partial D}\|_2 \quad \text{(Eq. 4)}$$

The final loss function combines all above loss functions.

$$\mathcal{L}(y,\hat{y}) = \mathcal{L}_{MRE}(y,\hat{y}) + \beta_1 \mathcal{L}_{der}(y,\hat{y}) + \beta_2 \mathcal{L}_{conc}(y,\hat{y}) + \beta_3 \mathcal{L}_{BC}(\hat{y}) \quad \text{(Eq. 5)}$$

CLIMATE ANALYSIS

As a part of our ML workflow, we firstly perform a regional climate assessment over the South Carolina region to provide a realistic distribution of recharge rates, which serves as an input to our surrogate model. Then, we extend the climate analysis to the continental US (CONUS) and perform an unsupervised climate data clustering. By doing this, we reduced the dimensionality of the extensive climate data which would eventually help stakeholders to get a realistic recharge rate immediately in order to evaluate the groundwater flow anywhere across the CONUS.

Regional climate analysis

For the regional climate analysis, we use climate models from CMIP5 Climate and Hydrology Projections[10]. Since our surrogate model evaluates the groundwater flow in the Savannah River site, the data for regional analysis are bound for the region from 32.0625N to 35.3125N and from -83.5652E to -78.8125E, covering most of South Carolina. We analyzed all four representative concentration pathways (RCP) projections from January 1950 to December 2099 for two monthly variables: precipitation and evapotranspiration (ET) and defined a recharge rate as a difference between monthly precipitation values and monthly evapotranspiration values. Four RCP pathways describe different climate projections, but all of them are considered possible and depend on the volume of greenhouse gases emission in the upcoming years.

Additionally, four Earth System Models (ESM) are used in this part of the study to incorporate more uncertainties of the future climate projections and improve our understanding of the impact on groundwater and contamination flow at the Savannah River site. As the recharge rate is a time-varying attribute in our surrogate model, we split our data into three-time windows: historical, mid-century, and late-century, each of which corresponds to 1950–2020 (hereafter history), 2021–2060 (mid-century), and 2061–2099 (late-century), to indicate how the recharge rate change by the end of this century.

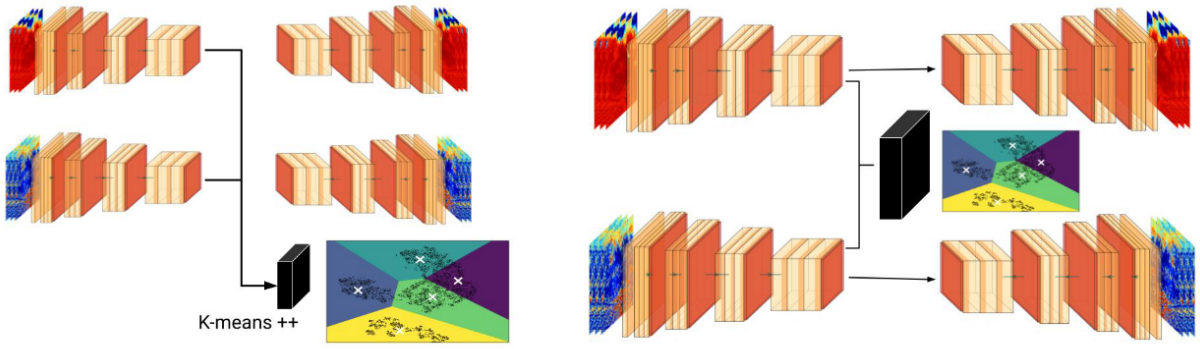
CONUS climate analysis

We revisit the analysis of distribution of precipitation and ET values for CONUS region, but we work with GFDL-ESM2G because the model represents the neural climatological trend based on literatures, so that our clusters generated from autoencoders, and clustering are less influenced by excessive extreme projections due to their model biases.

UNSUPERVISED CLIMATE DATA CLUSTERING

Unsupervised climate data clustering uses convolutional autoencoders to reduce the dimensionality of continental-scale climate simulation outputs and performs a spatial-temporal cluster analysis (see Figure. 3), which helps stakeholders to get a realistic recharge rate immediately in order to evaluate the groundwater flow anywhere across the continental US (CONUS). The median of precipitation and ET values from each resulting cluster will be used as inputs of the surrogate model.

We use monthly precipitation and ET values from GFDL-ESM2G[11], [12] that participated in CMIP5. We remove the 3-month-running mean and then standardize the monthly data so that our deseasonalizing approach removes the recurrent patterns but remains the climatological anomalies. We spatially subdivide each three months image into a 16 pixels $\times$ 16 pixels scale, $\approx 2^{\circ} \times 2^{\circ}$ area, giving a smaller geographical and temporal unit, *patch*.



(a) Convolutional autoencoder (b) End-to-end autoencoder.

Figure 3. Illustration of two autoencoders tested for unsupervised climate data clustering

We first build a convolutional autoencoder for precipitation and ET respectively (Figure 3a). The autoencoder[13] is a widely applied unsupervised learning techniques that learns a nonlinear mapping of input information to a low-dimensional through embedding (i.e., latent representation) via encoder and then reconstructing original information from the latent representation via decoder. Training of

autoencoder quantifies the difference between input image x and the reconstructed image $\hat{x}$ as following:

$$\mathcal{L}(\theta) = \sum_{x \in S} ||x - D_\theta\left(E_\theta(x)\right)||_p^p \qquad \text{(Eq.6)}$$

where S is a set of training images; θ represents trainable parameters in encoder and decode. The performance of image recognition improves with multiple layers of convolutional filters[14] to extract useful representations of spatiotemporal pattern of images and different data distributions among different groups of objects. In this study, we stuck 8 convolutional layers in the encoder and decoder has the miller structure of the encoder.

The second step in the convolutional autoencoder approach (Figure 3a) is to cluster the latent representation produced by the trained autoencoder to identify unique climate patterns. We apply k–means++ as known for the probabilistic initialization to find an initial seed of K number of clusters, and the approach outperforms the native k-means algorithm [15]. We use k–means API provided by scikit-learn Python package[16] and apply the historical patches unseen in training of autoencoders. We then obtain a set of K cluster centroids, $\mu = \{\mu_1, \cdots, \mu_k\}$. From the range of $K \in \{2, \cdots, 9\}$, we determine the optimal number of clusters based on the elbow method[17]. The elbow method indicates that 5 clusters are our optimal number of clusters. For the rest of our study, we present clustering results working with 5 clusters.

While scalable clustering algorithms are widely available[18] an end-to-end autoencoder (see Figure 3b) and clustering training can further benefit the scalability. In particular, given that the disk amount of climate simulation outputs is increasing, clustering training could be limited to a subset of relatively smaller datasets. A joint loss function [19] formulates a combination of both reconstruction and clustering loss terms:

$$\mathcal{L}_{\text{joint}}(\theta) = \lambda_{\text{reconst}}\mathcal{L}_{\text{reconst}}(\theta) + \lambda_{\text{clustering}}\mathcal{L}_{\text{clustering}}(\theta) \qquad \text{(Eq. 7)}$$

Our $\mathcal{L}_{\text{clustering}}$ is motivated by an online method [20] with simplifying the cross entropy loss between two cluster assignment: ``codes" $q_c$ via the Sinkhorn-Knopp algorithm [21] and $p_c$ that is computed as ``prediction" obtained with a softmax of the dot product of K trainable prototype $\{c_1, \cdots, c_K\}$ and the latent representation $z_c = E_\theta(x)$. We calculate the dot product with an output layer from a single layer perceptron F_p such that $z_c^\top c = F_p(E_\theta(x))$. Thus the second loss term in Eq. 6 is

$$\mathcal{L}_{\text{clustering}}(\theta) = -\sum_{k \in K} q_c^{(k)} \log\left(p_c\right) \quad \text{where} \quad p_c^{(k)} = \frac{\exp\left(\frac{1}{\tau} z_c^\top c_k\right)}{\sum_{k'} \exp\left(\frac{1}{\tau} z_c^\top c'_k\right)} \qquad \text{(Eq. 8)}$$

In this study, we train our approach with 70% of only historical (1950 – 2020) patches (we leave 30% for inference) for 400 epochs on 4 K80 NVIDIA GPUs at a GCP instance to account for the future climate impact in inference.

RESULTS AND DISCUSSION

Formulating a ML-based surrogate model for groundwater system and an unsupervised clustering model for climate data, we evaluate the performance of the surrogate model in the combination of our newly developed loss terms and spatial-temporal physical properties of clusters.

U-FNO-3D vs U-FNO-2D and loss functions

We have in total 664 physical simulations from the Amanzi model with uncertain soil, subsurface and climate inputs. We split our dataset into training, validation, and testing (8:1:1) subsets. TABLE I (Index 1-4) shows the additional U-Net architecture gives us lower MRE and MSE on the validation dataset for both FNO-2D and FNO-3D architectures. In practice, U-FNO-3D trains around 4× faster (30 epochs, U-FNO-3D: 72 minutes, U-FNO-2D: 273 minutes). Therefore, we test our hybrid physics-based and data-driven loss functions on U-FNO-3D (Index 5-8 in TABLE I) . Every part of loss functions in Eq. 5, when $\beta_i \neq 0$, reduces the MRE and MSE error. We have the lowest validation error when adding all spatial derivatives and no flow boundaries ($\beta_1 = \beta_2 = \beta_3 = 0.1$), Index 9-10, in TABLE I. U-FNO-3D with all three combined loss functions achieves the lowest MRE and MSE after training 150 epochs on a A100 NVIDIA GPU at a GCP instance.

TABLE I: Mean relative errors (MRE) and mean squared errors (MSE) for all experiments

Index	Architectures	Epochs	Loss $(\beta_1, \beta_2, \beta_3)$	MRE	MSE	Dataset
1	FNO-2D	30	(0,0,0)	0.051	2.71e-4	Validation
2	FNO-3D	30	(0,0,0)	0.055	2.98e-4	
3	U-FNO-2D	30	(0,0,0)	**0.035**	**1.51e-4**	
4	U-FNO-3D	30	(0,0,0)	**0.037**	**1.29e-4**	
5	U-FNO-3D	30	(0.1,0,0)	0.029	8.83e-5	
6	U-FNO-3D	30	(0,0.1,0)	0.033	1.10e-4	
7	U-FNO-3D	30	(0,0,0.1)	0.034	1.27e-4	
8	U-FNO-3D	30	(0.1,0.1,0.1)	**0.028**	**8.14e-5**	
9	U-FNO-2D	150	(0.1,0.1,0.1)	0.020	4.49e-5	Test
10	U-FNO-3D	150	(0.1,0.1,0.1)	**0.014**	**2.44e-5**	

Regional Climate Analysis

Analyzing regional climate data, we demonstrat how the recharge rate is expected to change over the time using probability density function (PDF). While we perform the regional climate analysis for four climate models from CMIP5: CCSM4, GFDL-ESM2G, IPSL-CM5A-MR , and MIROC-ESM for all four RCP scenarios, we only present two models here: CCSM4 and IPSL-CM5A-MR for RCP 2.6, as a scenario with the least amount of global warming and only limited climate change, and for RCP 8.5, as a scenario with more rapid warming and more climate change.

Figures 4 and 5 show the PDF of precipitation (Figure 4a and 4b) and ET (Figure 4c and 4d) and RCP 2.6 (left column) and RCP 8.5 (right column) for CCSM4 and IPSL-CM5A-MR models, respectively. Each plot shows the evolution of PDF of monthly spatial averaged data across the South Carolina area. We have historical PFD marked in blue against mid-century or late-century PDFs in orange and red, respectively. We also calculate the mean monthly precipitation and ET for each future time window. Different mean monthly values are noticeable between all four models and the most distinct models are CCSM4 and IPSL-CM5A-MR. As shown, the value of mean precipitation of CCSM4 (IPSL-CM5A-MR) model during the mid-century is 73.42 (67.31) mm/month, while a higher monthly value is expected during the late century and it is expected to reach 80.14 (65.27) mm/month. The monthly mean ET values, based on the four analyzed models, range between 67.31 to 73.41 mm/month during the mid-century and 65.27 to 80.14 mm/month during the late century. Moreover, it is important to notice that the most differences among analyzed models leaned toward the extremes of weather events. As it is known, extreme events are expected to be more severe in the future due to climate change which will probably have the strongest impact on the groundwater flow.

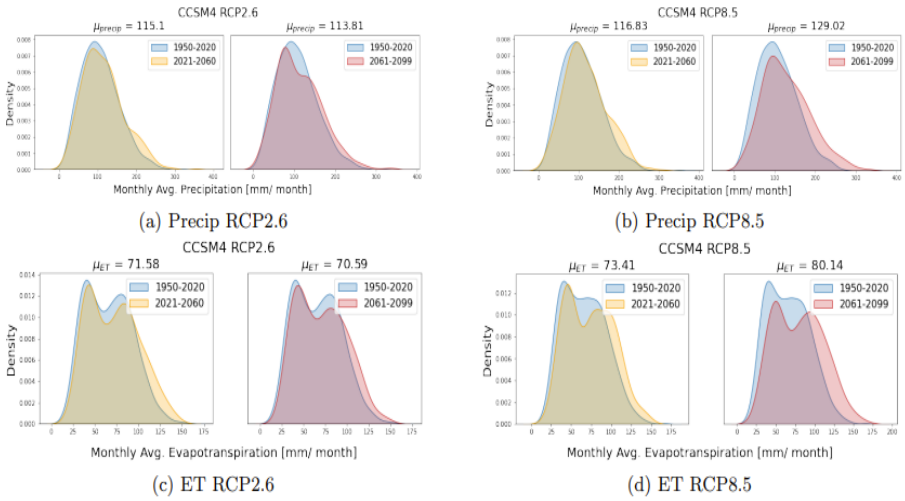


Figure 4. Probability density function for the spatial averaged monthly precipitation and ET values of RCP2.6 and RCP8.5 scenarios from CCSM4 simulation at the South Carolina area. The distribution from mid-century and late-century to be overlaid to that of historical.

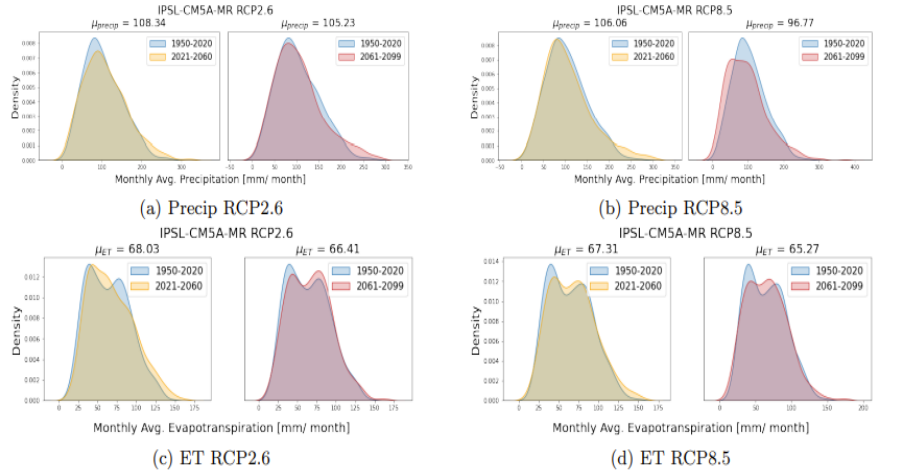
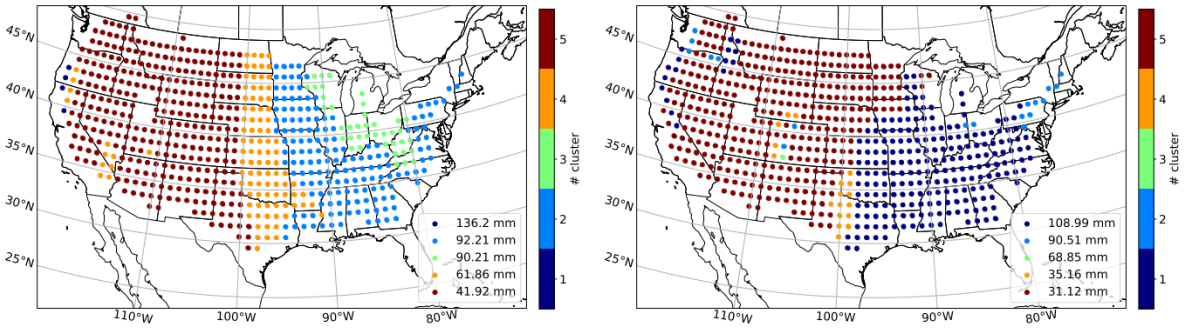


Figure 5. Same as Figure 4 but calculated distribution from IPSL-CM5A-MR simulation at the South Carolina area.

Evaluation of climate data clustering

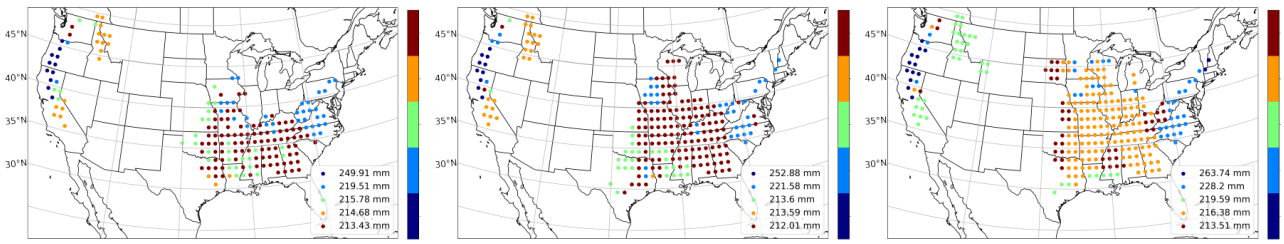
Figure. 6 shows a spatial distribution of the most frequent cluster at each patch location for 2061--2099. Cluster numbers are sorted in descending order by the mean precipitation per patch. Our cluster's climatological field is dominant by #5 (i.e., the driest climate cluster) over the CONUS from both autoencoder approaches. The major difference is seen in spatial distribution of cluster #1 (i.e., the wettest cluster) in that Figure 6a only locates the wettest climate pattern only over the West coast of CONUS, while Figure 6b (i.e., end-to-end autoencoder) classifies the pattern over the eastern area of 100W. We also observe that #2 from Figure 6a and #1 from Figure 6b overlap with humid subtropical climate zone (Cfa) and humid continental climate zone (Dfa) of a newly improved Koppen-Geiger climate classification [22] and #5 is spatially associated with their semi-arid (BS) and desert (BW) climate types, suggesting that our clusters capture physically meaningful spatial patterns.



(a) Convolutional autoencoder (b) End-to-end autoencoder.

Figure 6. The most frequent cluster patterns per patch resulting from two autoencoders for late-century

Moreover, to highlight the impact of warming trends in our clusters as the important concerns of the future climate are extreme weather events, we calculate the most frequent clusters at each grid cell point if a patch at a grid cell contains precipitation value that exceeds the 99% of precipitation value seen in historical time window for the convolutional autoencoder with k–means clustering approach. Here, extreme events are defined as events with monthly values higher than 99th percentiles based on historical data so that we have the same threshold over our time series. Figure 7 shows the locations with the extreme precipitation values during the historical (a), mid (b) and late (c) century based on autoencoders trained on precipitation and ET values.



(a) Historical (b) Mid-century (c) Late-century

Figure 7. The distribution of the locations which contains precip values that exceed the 99 percentile of historical data from convolutional autoencoder

Our results show that the locations with extreme precipitation events spatially expand over time. During the historical time window, most of the heavy precipitation events are concentrated over the southeast portion of the CONUS (Figure 7a) and gradually expand toward the Midwest as we approach the end of the century (Figure 7b and 7c), while the cluster mean precipitation values increase accordingly. The portion of the West coast of the US is also affected by the heavy precipitation, but here we see less of the spatial differences between three-time windows. Overall, the results show that we may expect higher monthly mean precipitation values for almost all locations.

CONCLUSION

In summary, we have successfully developed the ML-based multi-scale digital twin. Our ML-based surrogate model predicts spatiotemporal flow and transport properties immediately, saves computational times on solving PDEs, and thereby supports rapid decision-making for site managers.
Our proposed unsupervised approach reduces the dimensionality of the vast historical and projected climate data to capture five unique climate patterns and provides lightweight climate properties for surrogate modeling. We believe that more climate resilience analysis for other contamination sites can benefit from our method in this paper to develop the groundwater flow and contaminant transport surrogate model with climate uncertainty.

With tools for fast and reliable contaminant plume prediction under future climate scenarios, site managers and decision makers can evaluate the potential consequences and take rapid actions. We believe that more climate resilience analysis for other contamination sites can benefit from the method utilized in this paper to develop the groundwater flow and transport surrogate model. The use of clustering in the latent space of autoencoders on climate data for building representative climate regions can be extended to other applications in addition to using AI to create surrogate models.

REFERENCES

[1] J. D. Moulton *et al.*, "Amanzi: An Open-Source Multi-process Simulator for Environmental Applications," in *AGU Fall Meeting Abstracts*, 2014, vol. 2014, pp. H51K–0758.

[2] A. Libera *et al.*, "Climate change impact on residual contaminants under sustainable remediation," *J Contam Hydrol*, vol. 226, Oct. 2019, doi: 10.1016/j.jconhyd.2019.103518.

[3] Z. Xu *et al.*, "Reactive transport modeling for supporting climate resilience at groundwater contamination sites," *Hydrol Earth Syst Sci*, vol. 26, no. 3, pp. 755–773, 2022.

[4] R. Knutti and J. Sedláček, "Robustness and uncertainties in the new CMIP5 climate model projections," *Nat Clim Chang*, vol. 3, no. 4, pp. 369–373, 2013.

[5] Z. Li *et al.*, "Fourier neural operator for parametric partial differential equations," *arXiv preprint arXiv:2010.08895*, 2020.

[6] A. Lavin *et al.*, "Simulation Intelligence: Towards a New Generation of Scientific Methods," Dec. 2021, [Online]. Available: http://arxiv.org/abs/2112.03235

[7] G. Wen, Z. Li, K. Azizzadenesheli, A. Anandkumar, and S. M. Benson, "U-FNO—An enhanced Fourier neural operator-based deep-learning model for multiphase flow," *Adv Water Resour*, vol. 163, p. 104180, 2022.

[8] L. Lu, P. Jin, and G. E. Karniadakis, "Deeponet: Learning nonlinear operators for identifying differential equations based on the universal approximation theorem of operators," *arXiv preprint arXiv:1910.03193*, 2019.

[9] M. Raissi, P. Perdikaris, and G. E. Karniadakis, "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations," *J Comput Phys*, vol. 378, pp. 686–707, 2019.

[10] E. P. Maurer, L. Brekke, T. Pruitt, and P. B. Duffy, "Fine-resolution climate projections enhance regional climate change impact studies." Wiley Online Library, 2007. doi: 10.1029/2007EO470006.

[11] J. P. Dunne *et al.*, "GFDL's ESM2 global coupled climate–carbon earth system models. Part I: Physical formulation and baseline simulation characteristics," *J Clim*, vol. 25, no. 19, pp. 6646–6665, 2012.

[12] J. P. Dunne *et al.*, "GFDL's ESM2 global coupled climate–carbon earth system models. Part II: carbon system formulation and baseline simulation characteristics," *J Clim*, vol. 26, no. 7, pp. 2247–2267, 2013.

[13] M. A. Kramer, "Nonlinear principal component analysis using autoassociative neural networks," *AIChE journal*, vol. 37, no. 2, pp. 233–243, 1991.

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[15] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," 2006.

[16] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[17] P. Bholowalia and A. Kumar, "EBK-means: A clustering technique based on elbow method and k-means in WSN," *Int J Comput Appl*, vol. 105, no. 9, 2014.

[18] B. Bahmani, B. Moseley, A. Vattani, R. Kumar, and S. Vassilvitskii, "Scalable k-means++," *arXiv preprint arXiv:1203.6402*, 2012.

[19] R. Aparna and S. M. Idicula, "Spatio-Temporal Data Clustering using Deep Learning: A Review," in *2022 IEEE International Conference on Evolving and Adaptive Intelligent Systems (EAIS)*, 2022, pp. 1–10. doi: 10.1109/EAIS51927.2022.9787701.

[20] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," *Adv Neural Inf Process Syst*, vol. 33, pp. 9912–9924, 2020.

[21] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," *Adv Neural Inf Process Syst*, vol. 26, 2013.

[22] H. E. Beck, N. E. Zimmermann, T. R. McVicar, N. Vergopolan, A. Berg, and E. F. Wood, "Present and future Köppen-Geiger climate classification maps at 1-km resolution," *Sci Data*, vol. 5, no. 1, pp. 1–12, 2018, doi: 10.1038/s41597-020-00616-w.

ACKNOWLEDGEMENTS

This work has been enabled by the Frontier Development Lab Program (FDL) USA. FDL is a collaboration between SETI Institute and Trillium Technologies Inc., in partnership with (Google Cloud, NVIDIA). This material is based upon work supported by the Department of Energy [National Nuclear Security Administration] under Award Number DE-AI0000001. Disclaimer: This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the US Government nor any agency thereof, nor any of their employees makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness or usefulness of any information, apparatus product or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the US Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the US Government or any agency thereof.