Simulation Intelligence in the New Paradigm of Environmental Monitoring

Haruko Wainwright

Nuclear Science and Engineering; Civil and Environmental Engineering

Massachusetts Institute of Technology

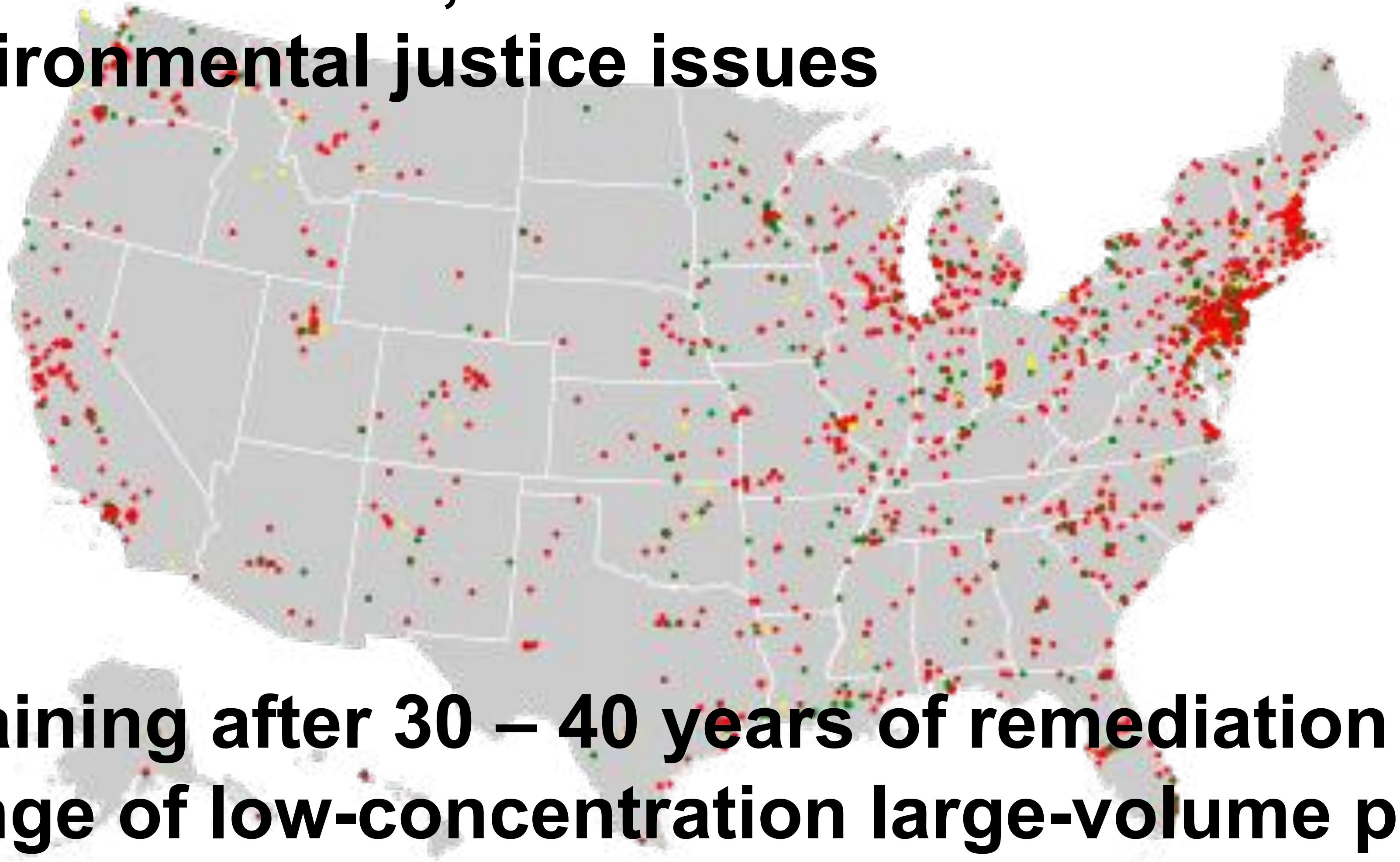
NSE
Nuclear Science & Engineering at MIT
science : systems : society

MIT

Civil and Environmental Engineering

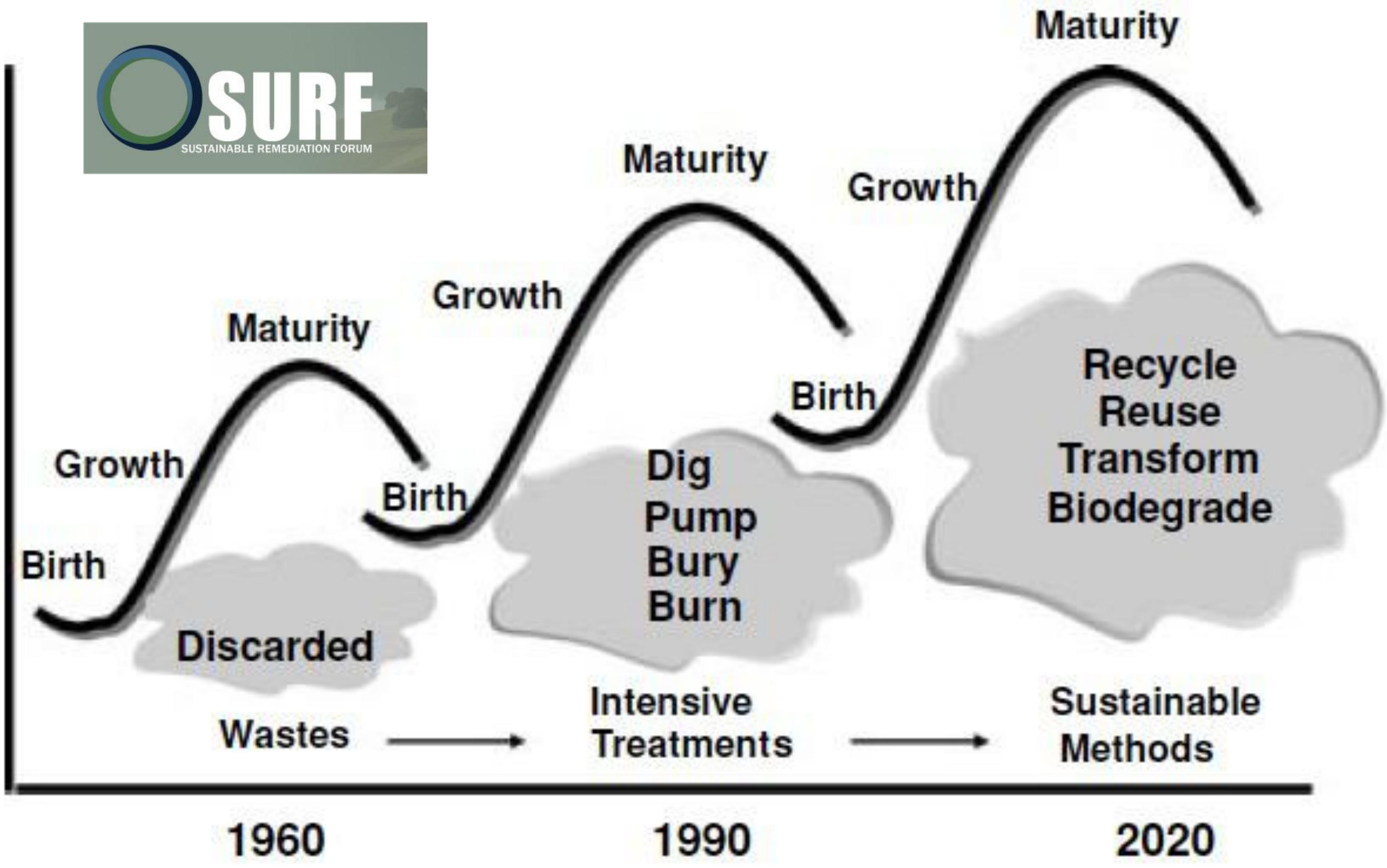
Soil and Groundwater Contamination

- **Superfund Sites: >1300 sites (organic/metal/radioactive)**
- **Brownfield Sites: ~450,000**
- **Major environmental justice issues**



>900 remaining after 30 – 40 years of remediation
→ Challenge of low-concentration large-volume plume

Environmental Remediation: Evolution



Trade offs:

Contaminant removal

vs

- **Waste**
- **CO2 emission**
- **Energy Use**
- **Ecological Impacts**
- **Noise, Air pollution**

Sustainable Remediation Forum (SURF), "Integrating sustainable principles, practices, and metrics into remediation projects", Remediation Journal, 19(3), pp 5 - 114, editors P. Hadley and D. Ellis, Summer 2009

Sustainable Remediation

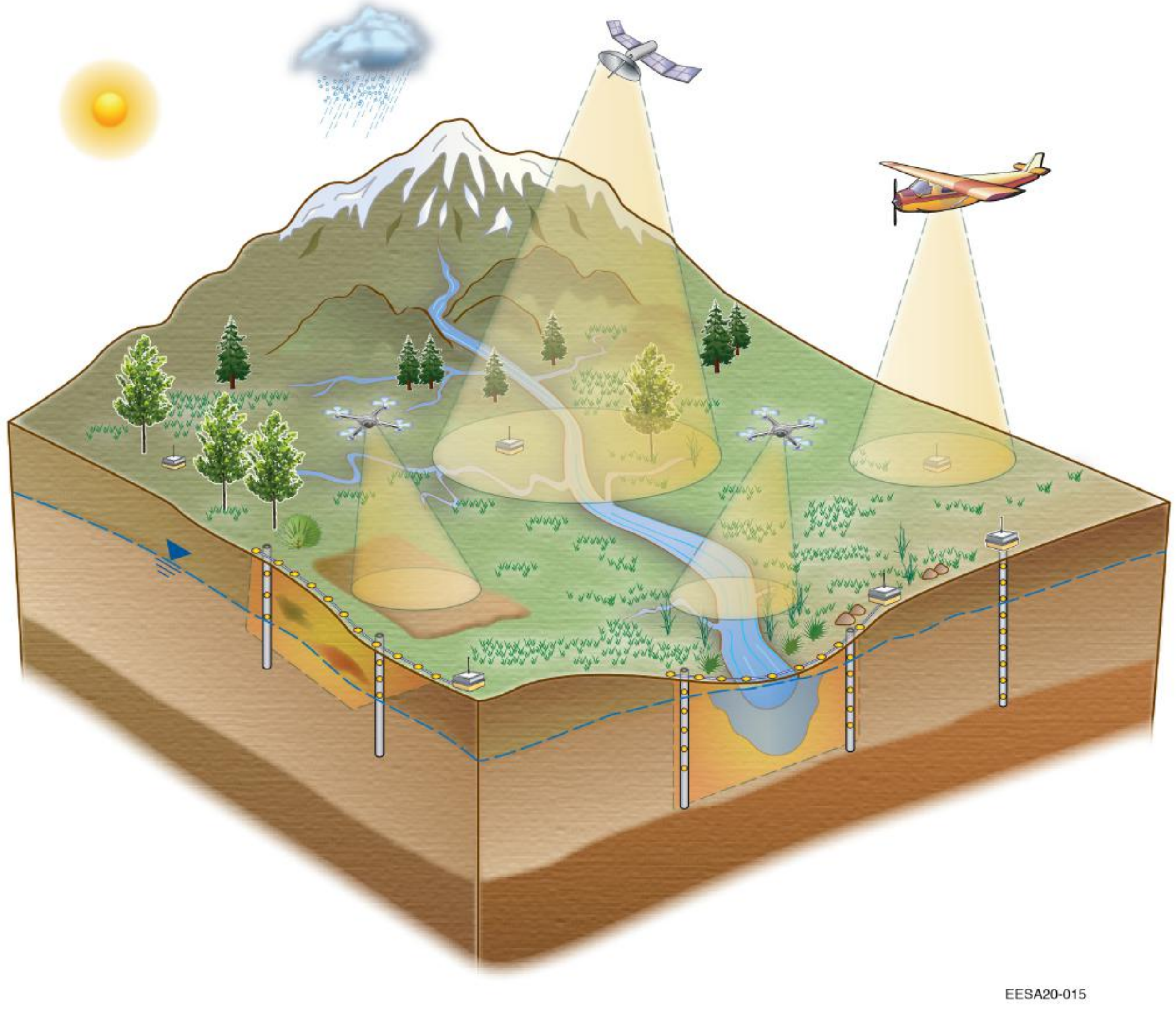
- Intensive/invasive clean up → Sustainable methods
- Minimize waste/pollution/energy-use/water-use/ecological damages
- Biodegradation, immobilization
- Monitored natural attenuation
- Longer institutional control with alternative/attractive end-use

→ Long-term monitoring

Rocky Flats National Wildlife Refuge

Former Reilly Tar & Chemical Corporation Plant

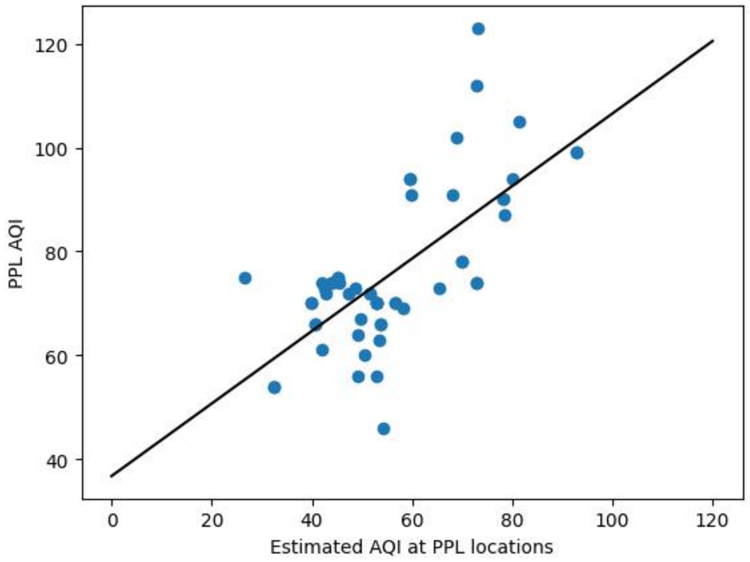
Earth Systems Monitoring



- **Multi-type multi-scale data**
  - Accuracy
  - Coverage
  - Footprints/resolution
- **"Proxy" information**
  - Plants/topography ~ soil
  - Electrical conductivity ~ contaminant concentration
- **Spatial-temporal correlation**
  - Data compression
  - Similar properties in vicinity

Challenges in ML/AI x Environmental Science

- Lack of training data
- Large uncertainty/variability



- Multiscale heterogeneity

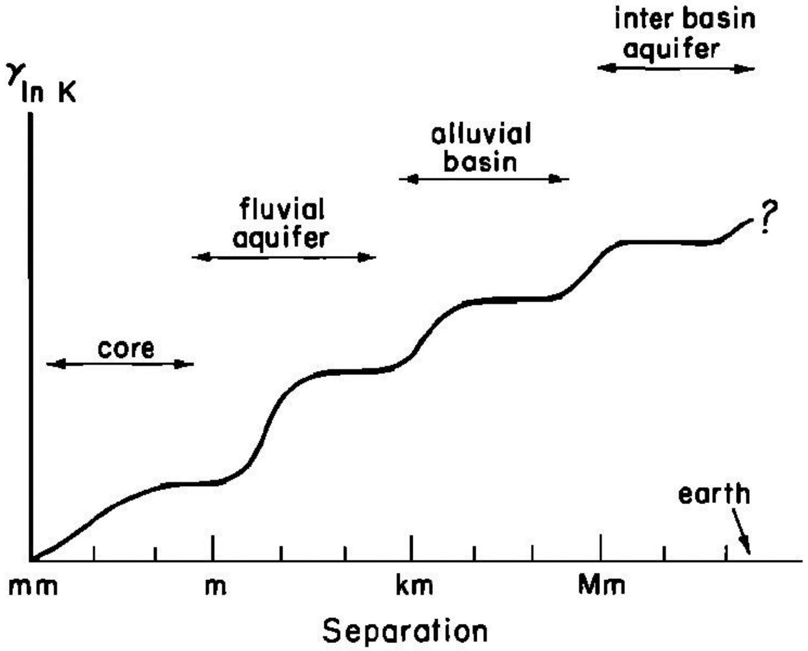
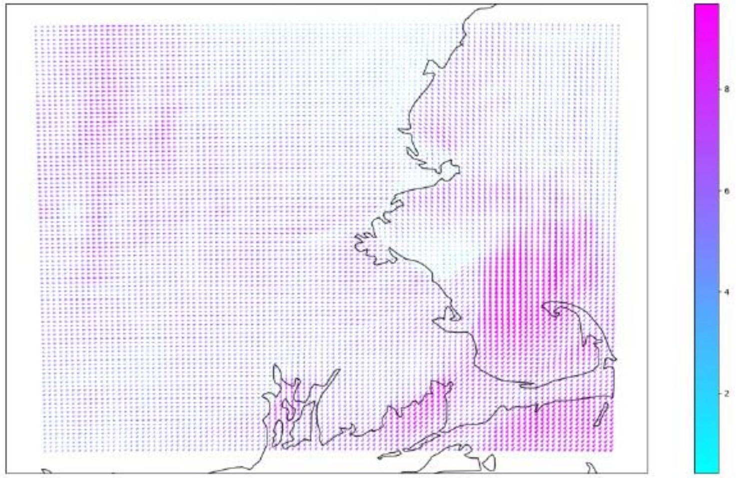
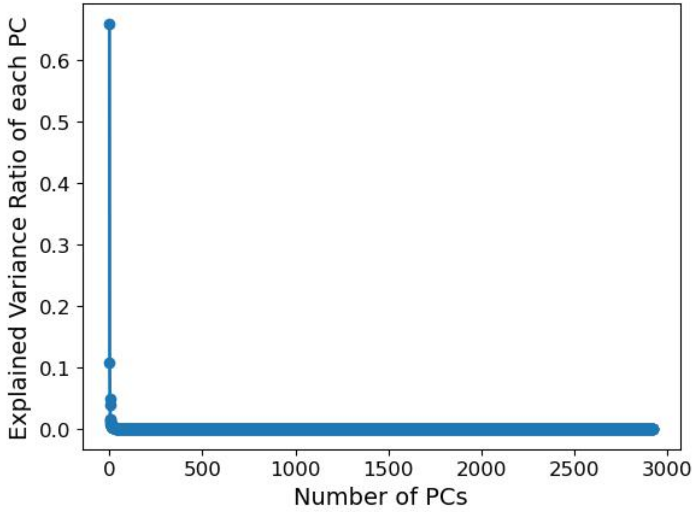


Fig. 8. Hypothetical ln *K* variogram illustrating the notion of scale-dependent correlation scales.

Gelhar, 1986

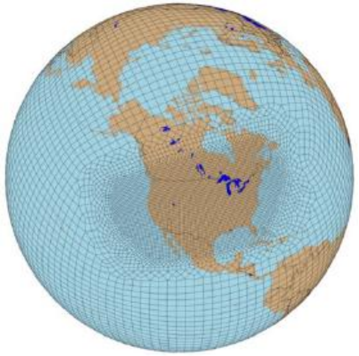
- Large data but little information content

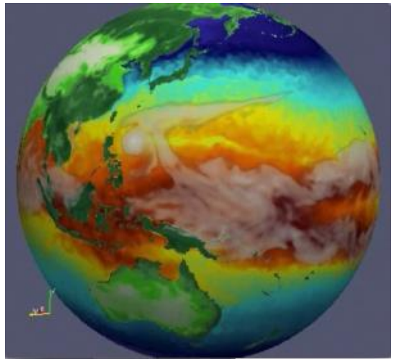




Challenges in Physical Models: Predictability

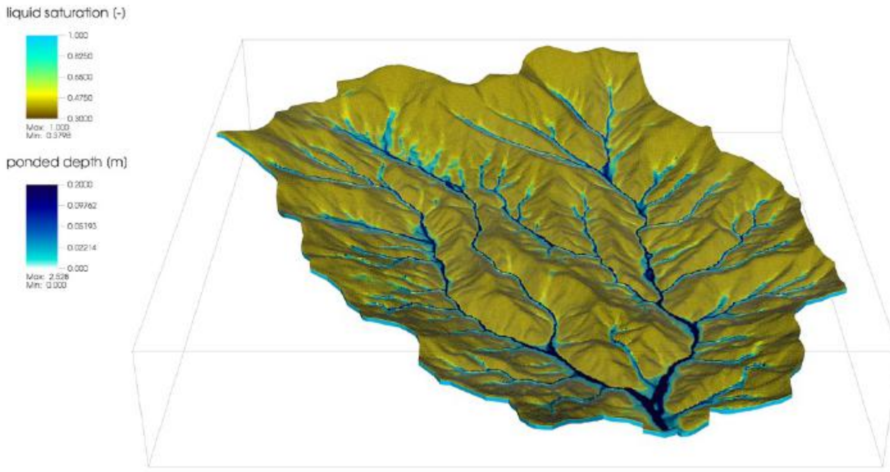
Global climate models





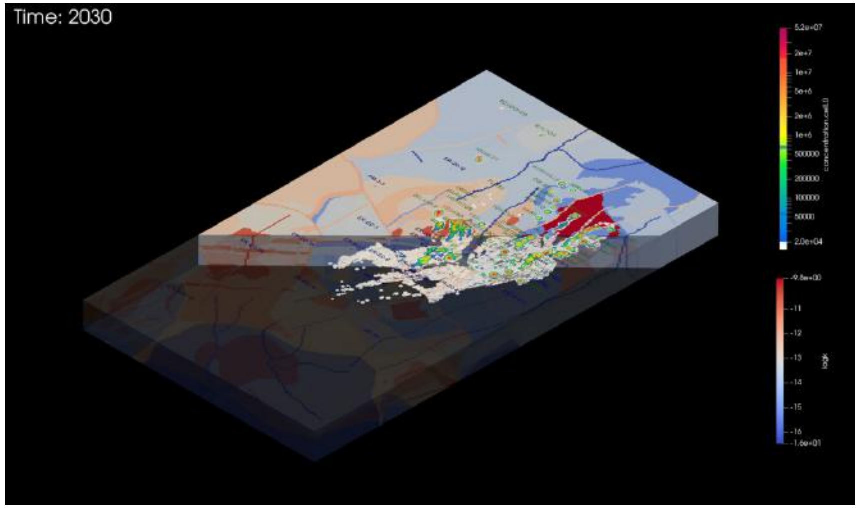
e3sm.org

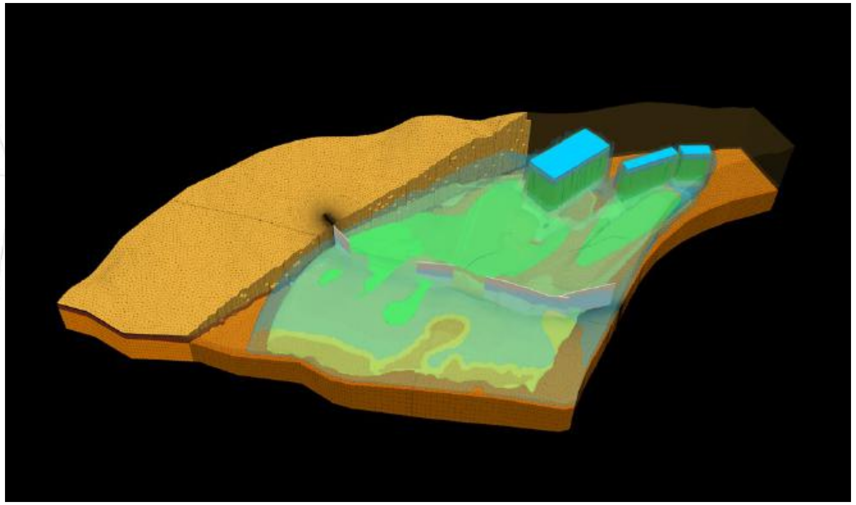
Watershed models



Amanzi-ATS

Contaminant transport models



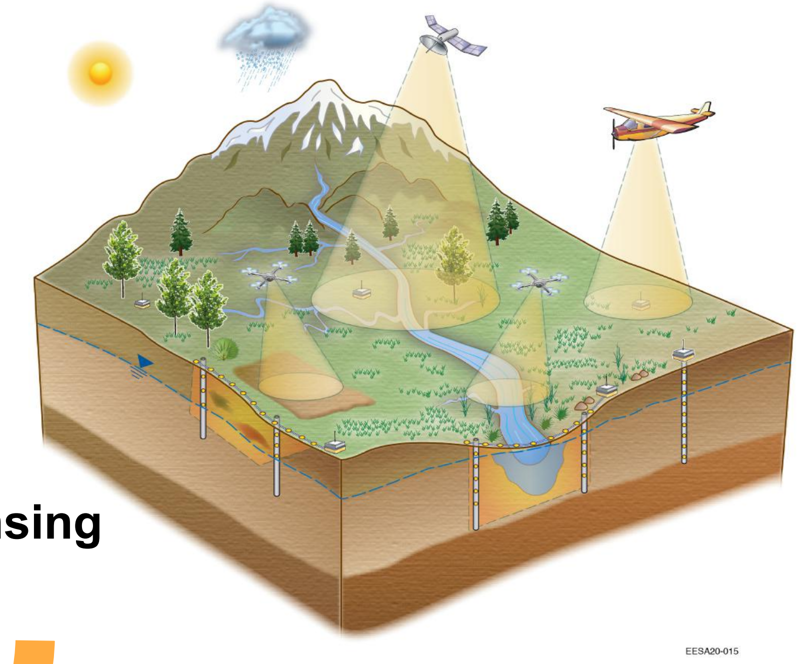


- Parameterization/heterogeneity
- Uncertainty quantification
- Inherit assumptions in models

Advanced Long-term Environmental Monitoring Systems

ML/AI

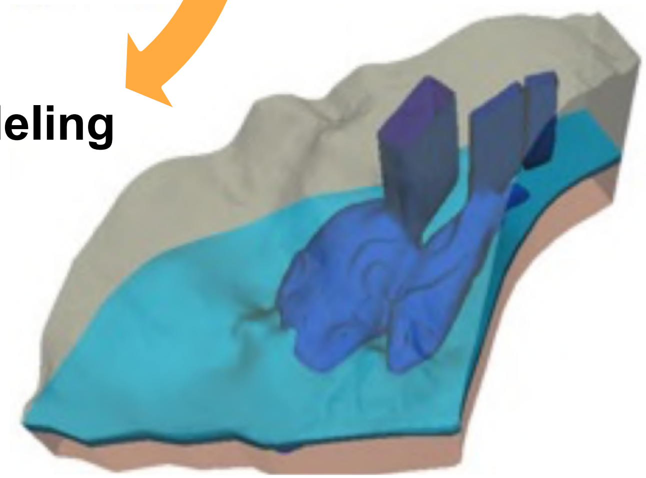
Sensing



ALTEMIS

altemis.lbl.gov

Modeling



New paradigm for long-term monitoring

ALTEMIS TEAM

Co-Leads

Carol Eddy-Dilek

Haruko Wainwright

AI for Soil and Groundwater

Himanshu Upadhyay

Styarth Praveen

Geochemical characterization

Hansell Gonzalez-Raymat

Miles Denham

Reactive Transport Modeling

Zexuan Xu

In situ real-time monitoring

Tom Danielson

Spatial monitoring: Geophysics

Tim Johnson

Baptiste Dafflon

Sebastian Uhlemann

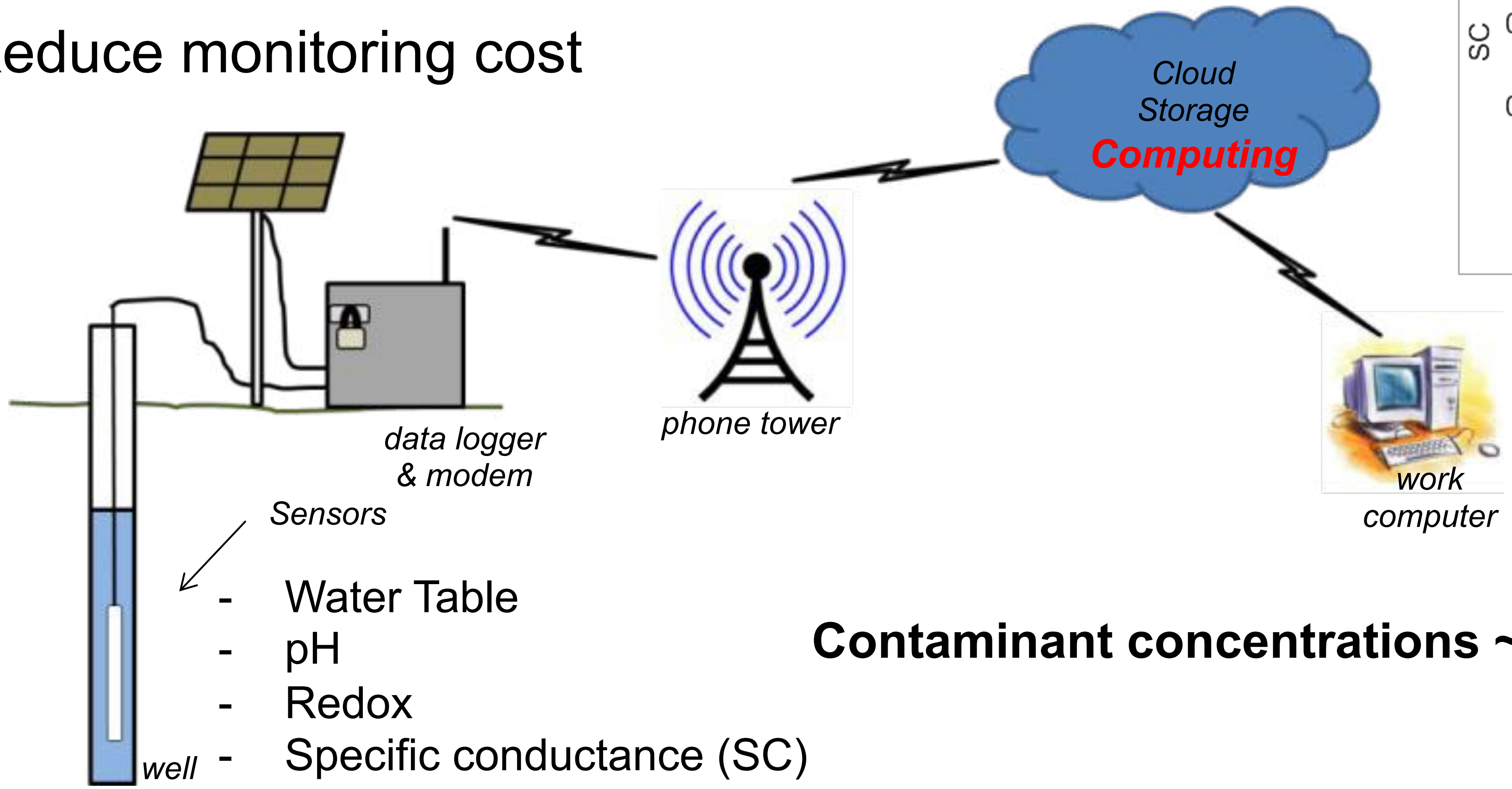
Spatial monitoring: Radiation

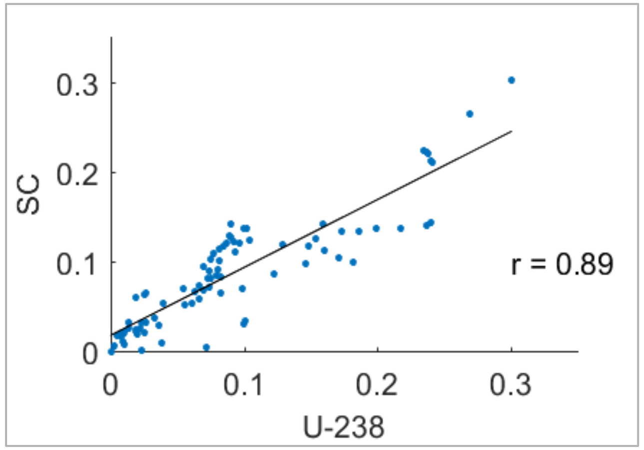
Kai Vetter

Brian Quiter

In situ Real-time Groundwater Monitoring

- **Low-cost in situ sensors, wireless network, cloud computing**
  - → Autonomous continuous monitoring
  - → Detect changes real-time = Early Warning
  - → Reduce monitoring cost



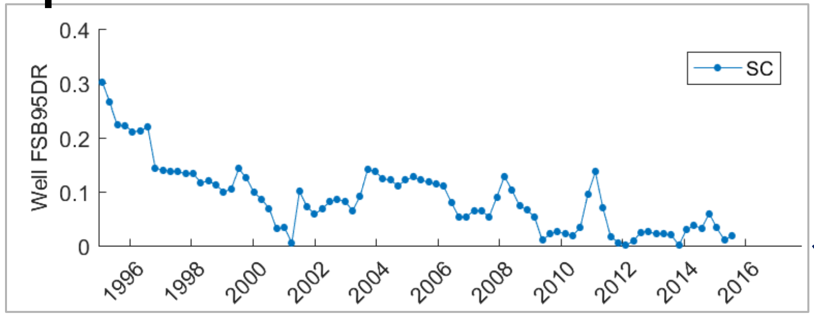


Contaminant concentrations ~ pH, WT, SC etc

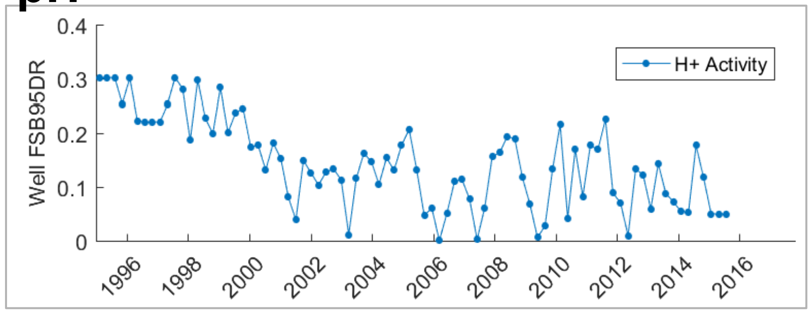
(Schmidt et al., 2018)

In situ Real-time Groundwater Monitoring

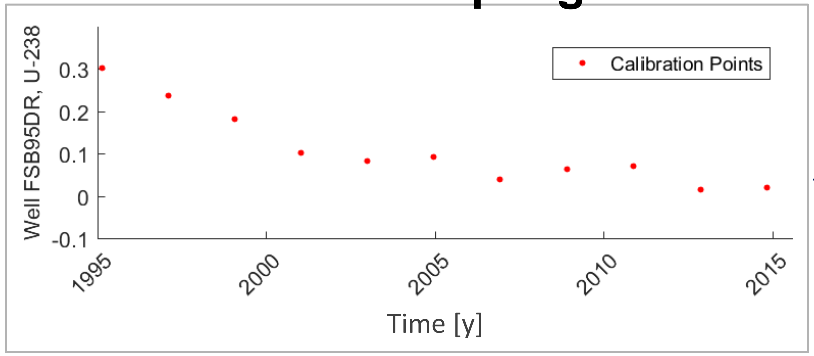
Specific Conductance



pH



U Groundwater Sampling Data



* Normalized concentrations

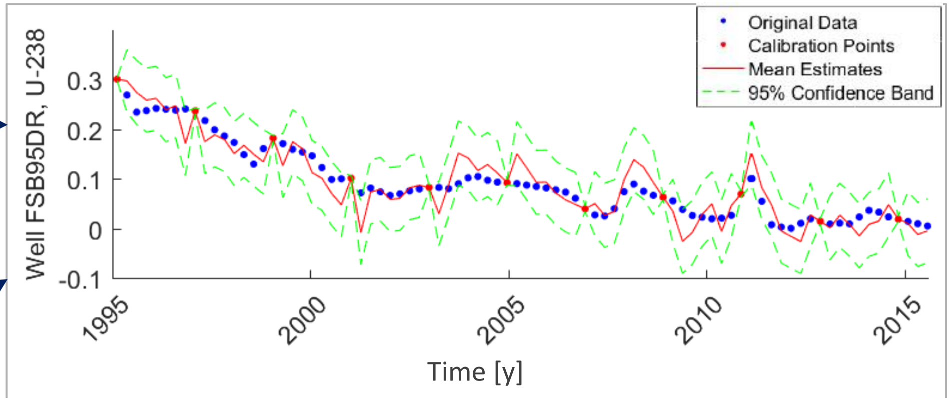
- Estimation of contaminant concentrations: $\{y_1, y_2, \ldots y_T\}$
  - Time evolution of contaminant concentration

$$\boldsymbol{y}_t = f(\boldsymbol{y}_{t-1}) + \tau \qquad \tau \sim N(0, \sigma^2)$$

  - Relationship to in situ datasets

$$\boldsymbol{z}_{pH,t} = g_{pH}(\boldsymbol{y}_t) + \varepsilon_{pH}$$

$$\boldsymbol{z}_{SC,t} = g_{SC}(\boldsymbol{y}_t) + \varepsilon_{SC}$$



$$\boldsymbol{p}(\boldsymbol{y}_t | \boldsymbol{y}_{t-1}, \boldsymbol{z})$$

- Confidence interval captures validation points
- Mean estimate captures the fluctuation
- Reduce #sampling from quarterly to every two years

(Schmidt et al., 2018)

Environmental Science & Technology

Search | Citation

Enter search text / DOI

Environ. Sci. Technol.

Environ. Sci. Technol. | Environ. Sci. Technol. Lett.

Browse the Journal | Articles ASAP | Current Issue | Submission & Review | Open Access | About the Journal

Article

Previous Ar

In Situ Monitoring of Groundwater Contamination Using the Kalman Filter

Franziska Schmidt[†], **Haruko M. Wainwright**[*,‡], **Boris Faybishenko**[§], **Miles Denham**[ǀ], and **Carol Eddy-Dilek**[⊥]

[†] Department of Nuclear Engineering, University of California Berkeley, Etcheverry Hall, 2521 Hearst Avenue, Berkeley, California 94709, United States

[‡] Climate and Ecosystem Sciences Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, MS 74R-316C, Berkeley, California 94720-8126, United States

[§] Energy Geosciences Division, Lawrence Berkeley Nat
8126, United States

[ǀ] Panoramic Environmental Consulting, LLC, P.O. Box 9

[⊥] Savannah River National Laboratory, Savannah River

Environ. Sci. Technol., **2018**, *52* (13), pp 7418–7425

DOI: 10.1021/acs.est.8b00017

Publication Date (Web): June 22, 2018

EurekAlert! The Global Source for Science News | AAAS

HOME | NEWS | MULTIMEDIA | MEETINGS | PORTALS | ABOUT

PUBLIC RELEASE: 13-AUG-2018

Algorithm provides early warning system for tracking groundwater contamination

Berkeley Lab researchers devise system to monitor contaminant plumes

DOE/LAWRENCE BERKELEY NATIONAL LABORATORY

SHARE | PRINT | E-MAIL

The University Net

News | Student Resources | Scholarships | Student Discounts

New Algorithm Provides Real-Time Monitoring Of Groundwater Pollutants

Energy Live News

NEWS | EVENTS | VIDEOS | TV & PODCASTS | INDUST

Efficiency & Environment

Scientists develop new meth to track groundwater pollutants in real-time

It is expected to reduce the frequency of manual groundw sampling and lab analysis and therefore cut the monitorin cost

GCN The Technology that Drives Government IT

AI & Automation | Cybersecurity | Cloud & Infrastructure | Data & Analytics | Smart Cities & IoT

Machine learning improves contamination monitoring

BY MATT LEONARD | AUG 14, 2018

Because groundwater is susceptible to pollution from automotive fuel, fe... or naturally occurring substances like iron, the Environmental Protection Agency and its state-level counterparts conduct annual or quarterly sampling and analysis.

Environmental Science & Technology

pubs.acs.org/est

Article

PyLEnM: A Machine Learning Framework for Long-Term Groundwater Contamination Monitoring Strategies

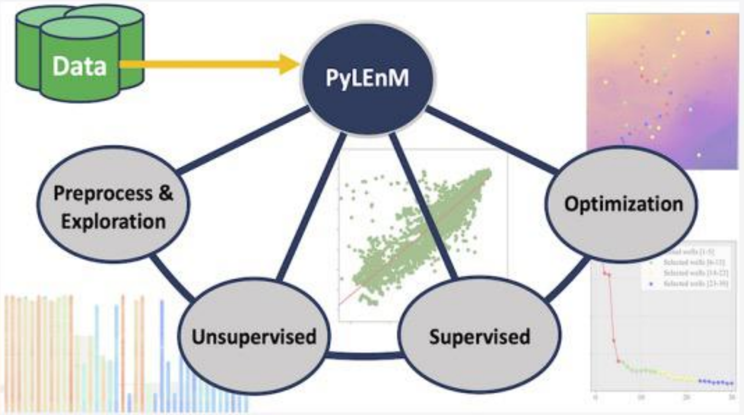
Aurelien O. Meray, Savannah Sturla, Masudur R. Siddiquee, Rebecca Serata, Sebastian Uhlemann, Hansell Gonzalez-Raymat, Miles Denham, Himanshu Upadhyay, Leonel E. Lagos, Carol Eddy-Dilek, and Haruko M. Wainwright*

Cite This: *Environ. Sci. Technol.* 2022, 56, 5973–5983

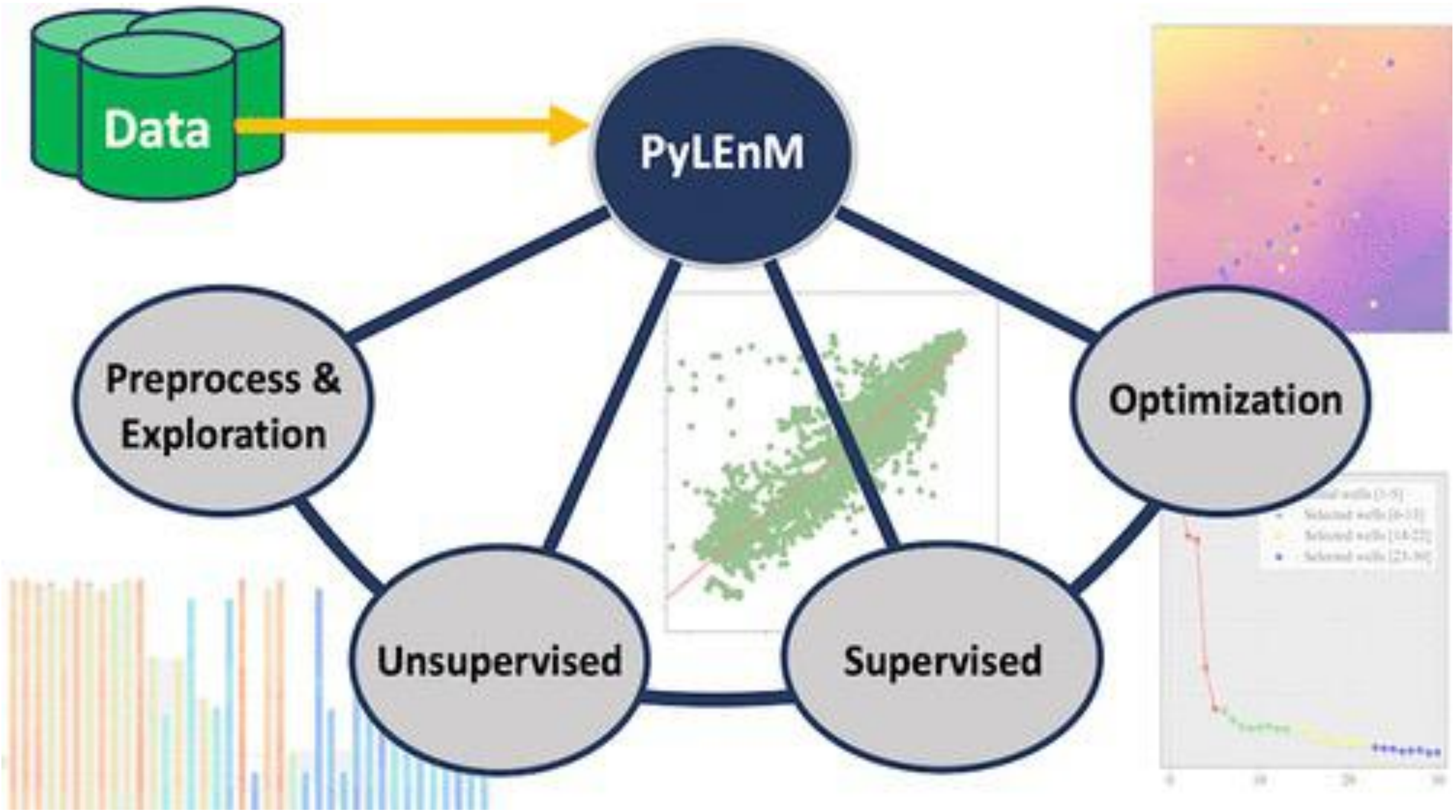
Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: In this study, we have developed a comprehensive machine learning (ML) framework for long-term groundwater contamination monitoring as the Python package PyLEnM (Python for Long-term Environmental Monitoring). PyLEnM aims to establish the seamless data-to-ML pipeline with various utility functions, such as quality assurance and quality control (QA/QC), coincident/colocated data identification, the automated ingestion and processing of publicly available spatial data layers, and novel data summarization/visualization. The key ML innovations include (1) time series/multianalyte clustering to find the well groups that have similar groundwater dynamics and to inform spatial interpolation and well optimization, (2) the automated model selection and parameter tuning, comparing multiple regression models for spatial interpolation, (3) the proxy-based spatial interpolation method by including spatial

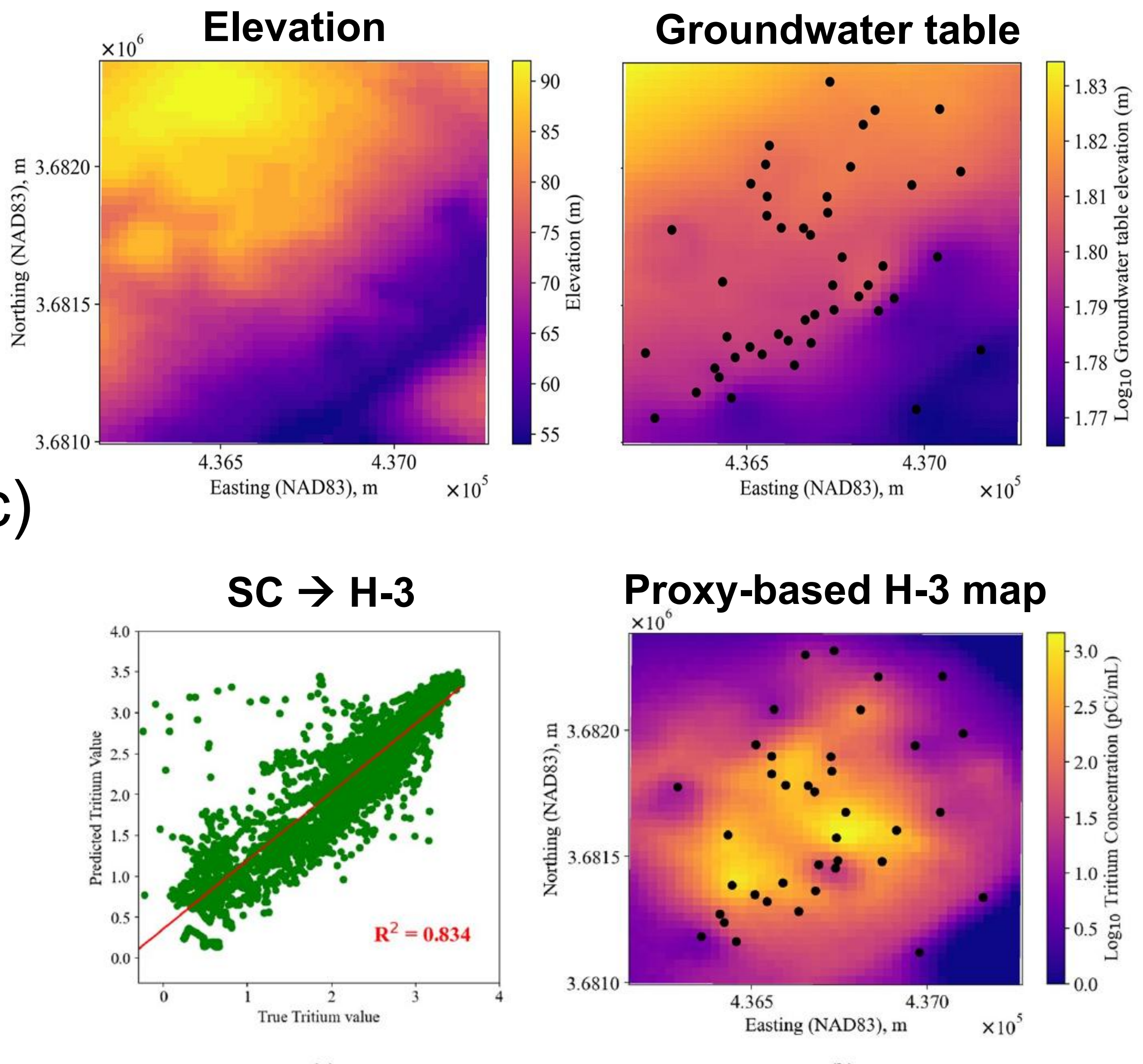


PyLenM: Python for Long-term Env. Monitoring



PyLenM: Supervised Learning

- **Spatiotemporal Interpolation**
  - Groundwater table
  - Contaminant concentration
- **Proxy variables**
  - LiDAR elevation data
  - Topographic metrics (slope etc)
  - Distance from the source
  - In situ measurable SC → tritium concentration
- **Comparison of multiple regression methods**



PyLenM: Well Placement Optimization

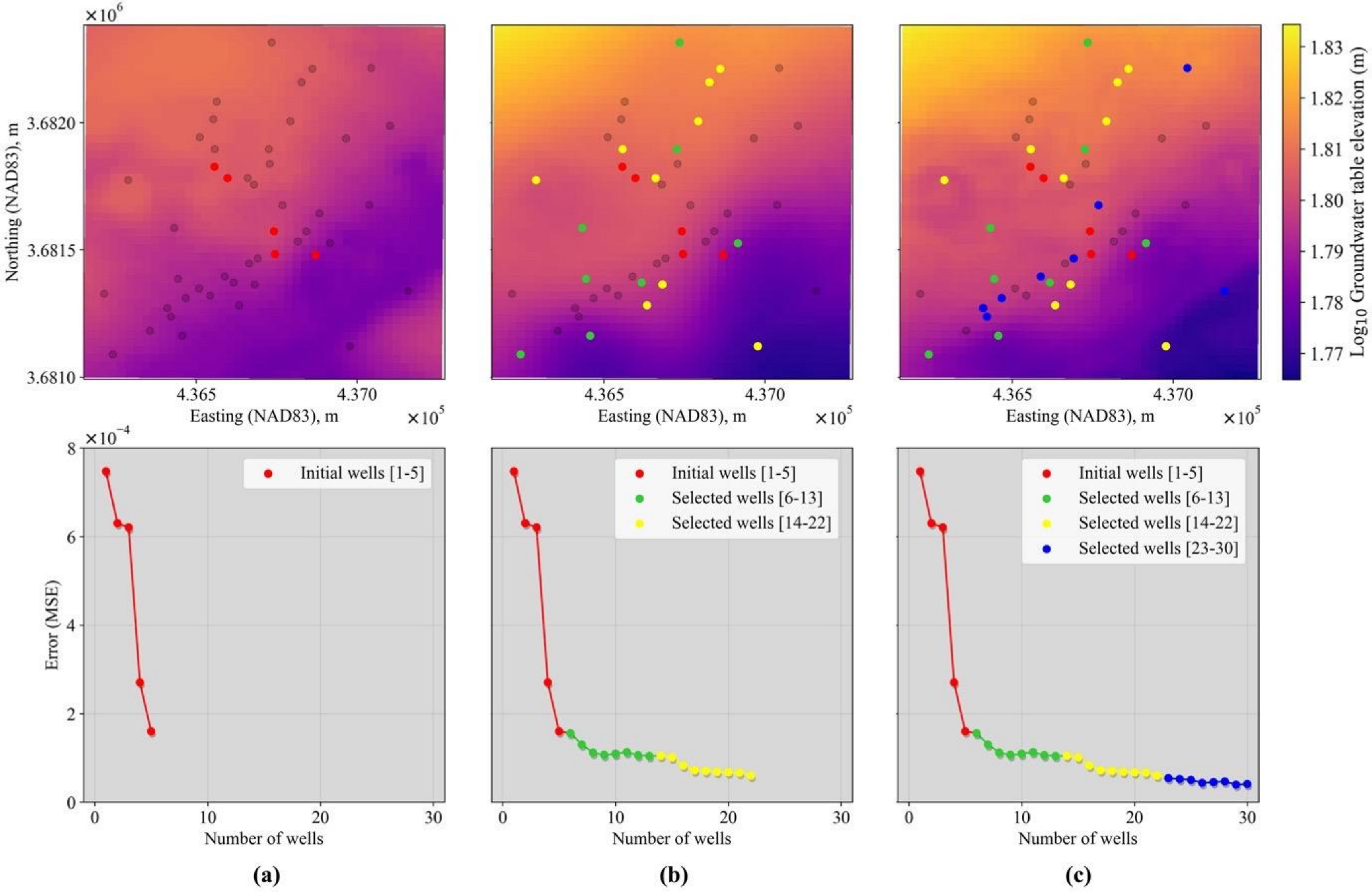
Sub-selection of wells for long-term monitoring

<u>Greedy algorithm</u>

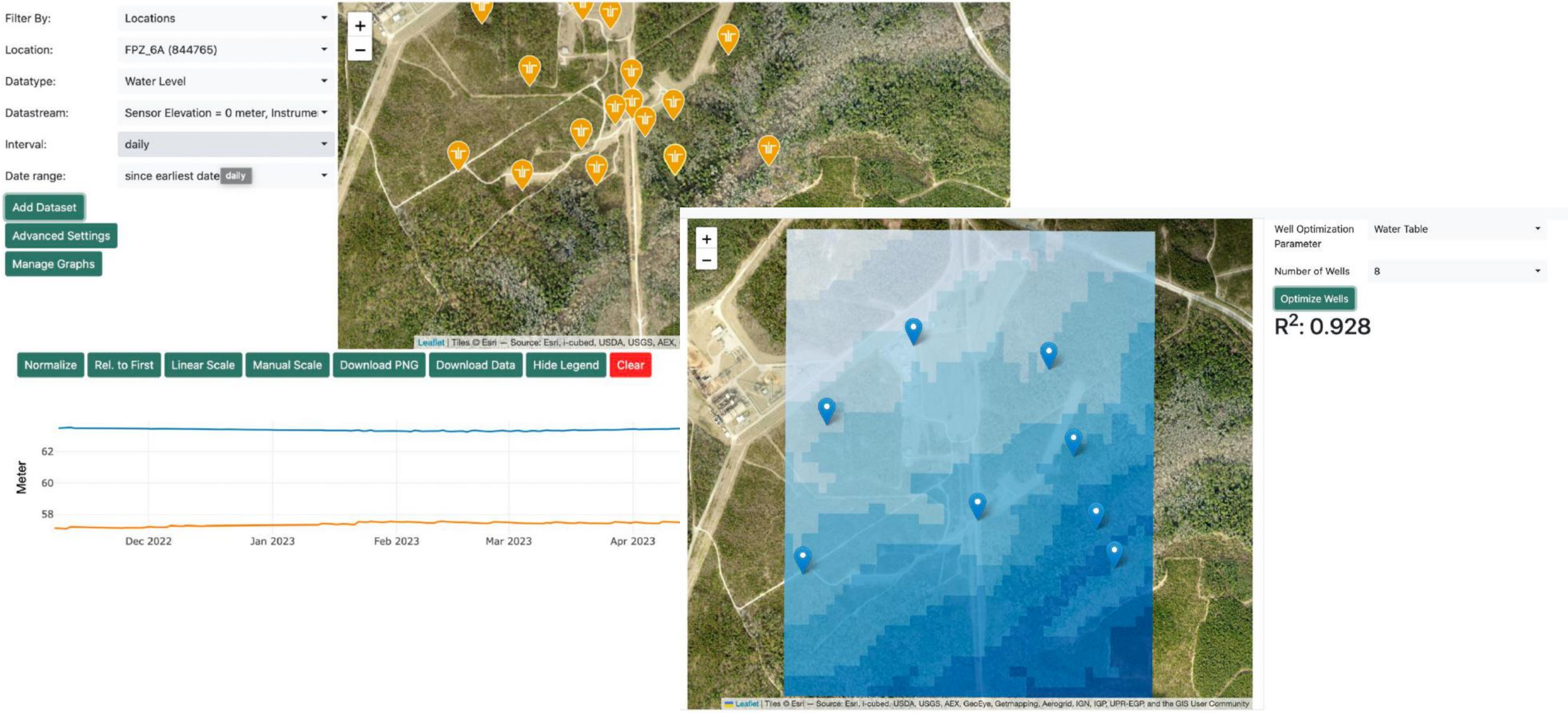
- Reference map created using all the wells
- Interpolation with one additional well at a time
- Find the well that minimize the overall error

<u>Minimum-but-sufficient # wells</u>

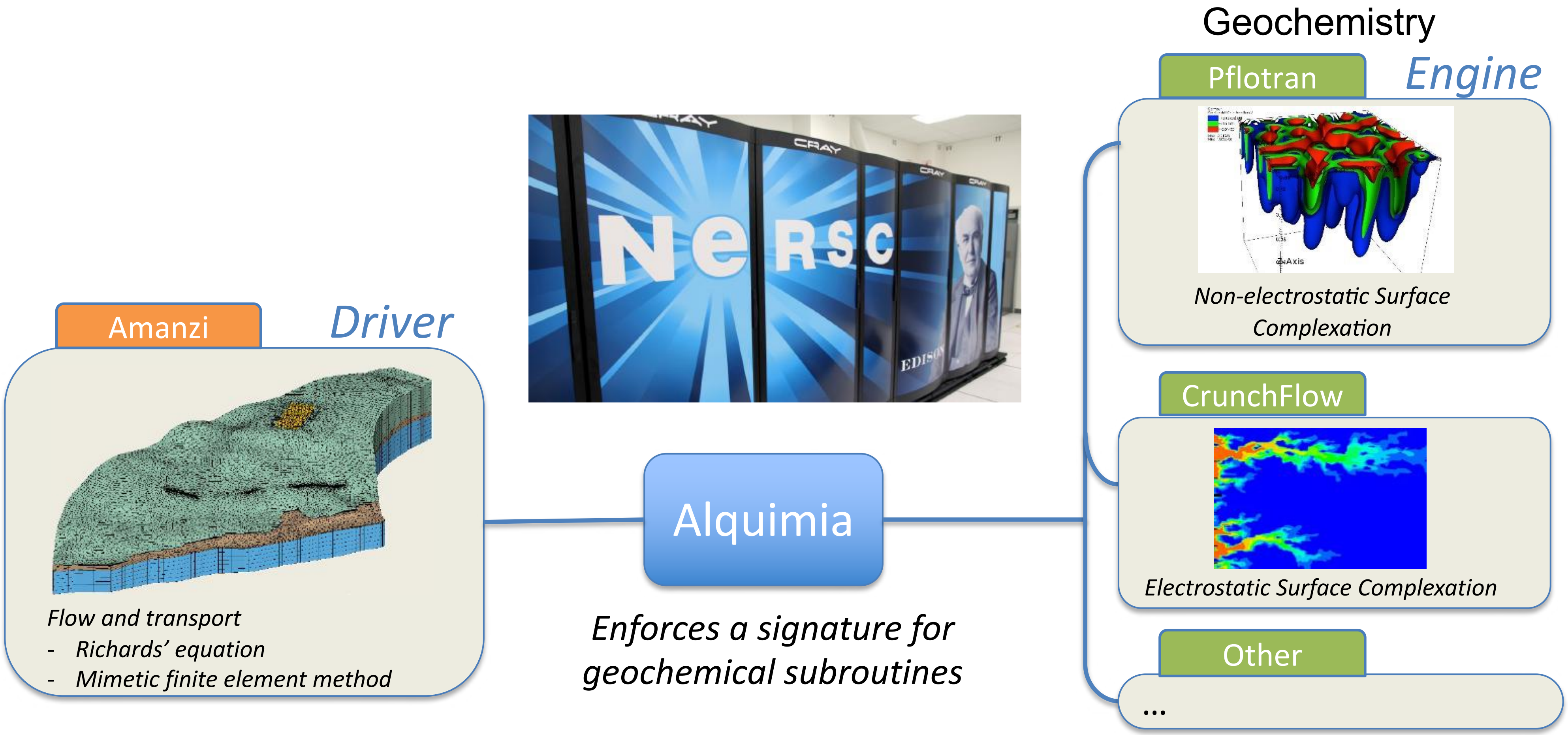
- Error convergence



Web-interface: Implementation



Contaminant Transport Modeling: Amanzi



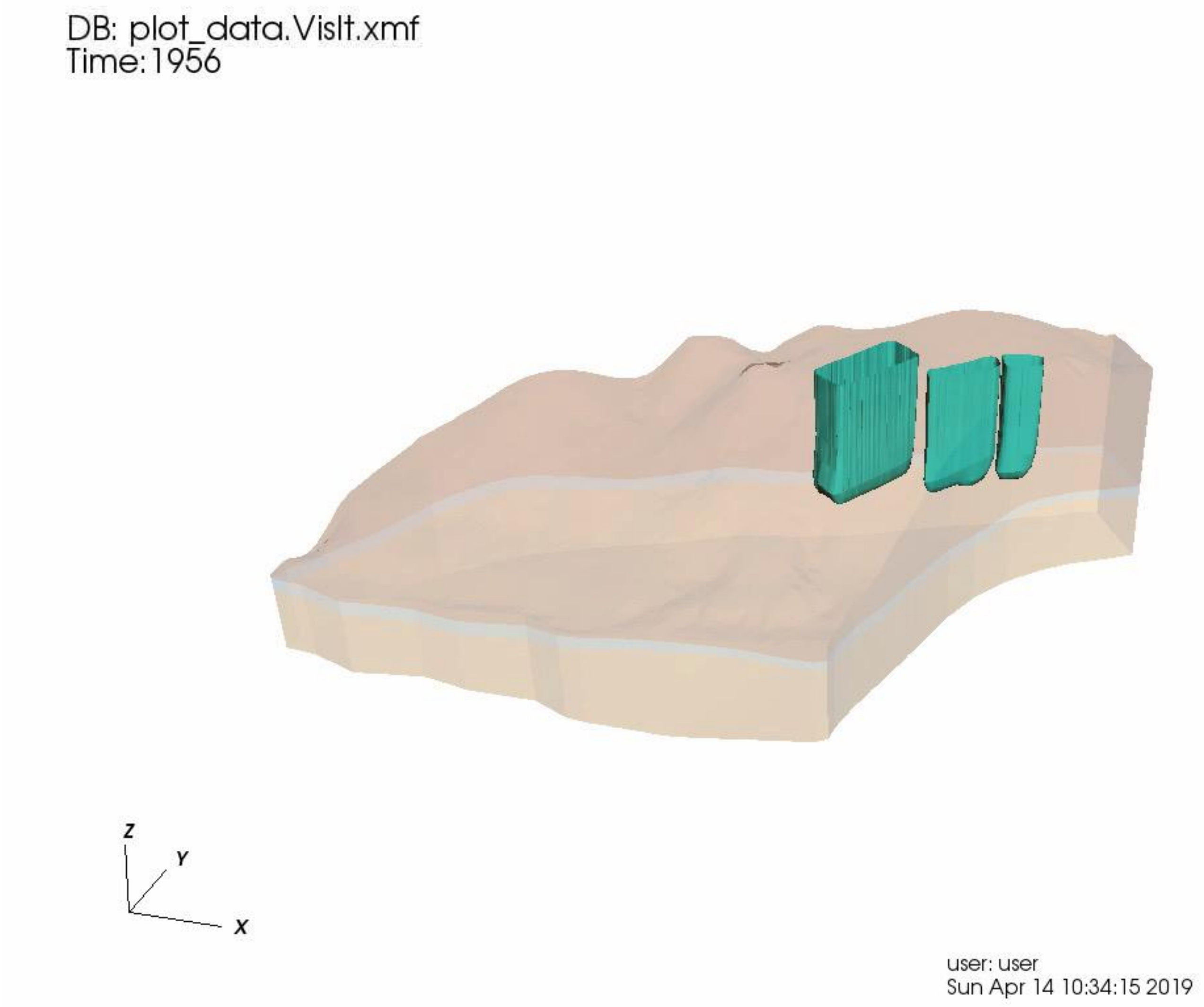
- **Complex flow**
- **Complex geochemical reactions**

SRS F-Area: Geochemical Model

Reaction	log *K* (25 °C)
Aqueous species[1]	
$(UO_2)_2(OH)_2^{+2} \leftrightarrow 2UO_2^{+2} + 2H_2O - 2H^+$	5.62
$(UO_2)_2CO_3(OH)_3^- \leftrightarrow 2UO_2^{+2} + 3H_2O + HCO_3^- - 4H^+$	11.18
$(UO_2)_2OH^{+3} \leftrightarrow 2UO_2^{+2} + H_2O - H^+$	2.7
$(UO_2)_3(CO_3)_6^{-6} \leftrightarrow 3UO_2^{+2} + 6HCO_3^- - 6H^+$	7.97
$(UO_2)_3(OH)_4^{+2} \leftrightarrow 3UO_2^{+2} + 4H_2O - 4H^+$	11.9
$UO_2(OH)_4^{-2} \leftrightarrow UO_2^{+2} + 4H_2O - 4H^+$	32.4
$(UO_2)_3(OH)_5^+ \leftrightarrow 3UO_2^{+2} + 5H_2O - 5H^+$	15.55
$(UO_2)_3(OH)_7^- \leftrightarrow 3UO_2^{+2} + 7H_2O - 7H^+$	32.2
$(UO_2)_3O(OH)_2(HCO_3)^+ \leftrightarrow 3UO_2^{+2} + 3H_2O + HCO_3^- -$	9.68
$(UO_2)_4(OH)_7^+ \rightarrow 4UO_2^{+2} + 7H_2O - 7H^+$	21.9
$UO_2NO_3^+ \leftrightarrow UO_2^{+2} + NO_3^-$	−0.3
$UO_2(OH)^+ \leftrightarrow UO_2^{+2} + H_2O$	5.25
$UO_2(OH)_2(aq) \leftrightarrow UO_2^{+2} + 2H_2O - 2H^+$	12.15
$UO_2(OH)_3^- \leftrightarrow UO_2^{+2} + 3H_2O - 3H^+$	20.25
$UO_2CO_3(aq) \leftrightarrow UO_2^{+2} + HCO_3^- - H^+$	0.39
$UO_2(CO_3)_2^{-2} \leftrightarrow UO_2^{+2} + 2HCO_3^- - 2H^+$	4.05
$UO_2(CO_3)_3^{-4} \rightarrow UO_2^{+2} + 3HCO_3^- - 3H^+$	9.14
$CaUO_2(CO_3)_3^{-2} \rightarrow Ca^{+2} + UO_2^{+2} + 3HCO_3^- - 3H^+$	3.8
$Ca_2UO_2(CO_3)_3(aq) \leftrightarrow 2Ca^{+2} + UO_2^{+2} + 3HCO_3^- - 3H^+$	0.29
$MgUO_2(CO_3)_3^{-2} \leftrightarrow Mg^{+2} + UO_2^{+2} + 3HCO_3^- - 3H^+$	5.19
$UO_2SiO(OH)_3^+ \leftrightarrow SiO_2(aq) + UO_2^{+2} + 2H_2O - H^+$	2.48
Surface and exchange species[2]	
$(>k\text{-}OH)_2UO_2^+ \leftrightarrow 2>k\text{-}OH^{-0.5} + UO_2^{+2}$	−5.3
$(>k\text{-}OH)_2UO_2CO_3^- \leftrightarrow 2>k\text{-}OH^{-0.5} + UO_2^{+2} + HCO_3^- - H^+$	−6.2
$>k\text{-}OH_2^{+0.5} \leftrightarrow >k\text{-}OH^{-0.5} + H^+$	−4.9
$>k\text{-}OHNa^{+0.5} \rightarrow >k\text{-}OH^{0.5} + Na^+$	2.1
$>k\text{-}OH_2NO_3^{-0.5} \leftrightarrow >k\text{-}OH^{-0.5} + H^+ + NO_3^-$	−4.9
$>k_2UO_2 \leftrightarrow 2>k^- + UO_2^{+2}$	−7.1
$>kNa \leftrightarrow >k^- + Na^+$	−2.9
$>kH \leftrightarrow >k^- + H^+$	−4.5
$>k_2Ca \rightarrow 2>k^- + Ca^{+2}$	−6.8
$>k_3Al \rightarrow 3>k^- + Al^{+3}$	−8
$(>Fe\text{-}OH)_2UO_2^+ \leftrightarrow 2>Fe\text{-}OH^{-0.5} + UO_2^{+2}$	−14.11
$(>Fe\text{-}OH)_2UO_2CO_3^- \leftrightarrow 2>Fe\text{-}OH^{-0.5} + UO_2^{+2} + HCO_3^- - H$	−4.35
$>Fe\text{-}OH_2^{+0.5} \leftrightarrow >Fe\text{-}OH^{-0.5} + H^+$	−9.18
$(>Fe\text{-}OH)_2CO_2^- \leftrightarrow 2>Fe\text{-}OH^{-0.5} + H^+ + HCO_3^- - 2H_2O$	−12.23
$>Fe\text{-}OCO_2Na^{-0.5} \leftrightarrow >Fe\text{-}OH^{-0.5} + Na^+ + HCO_3^- - H_2O$	−3.28
$>qz\text{-}OH_2^+ \leftrightarrow >qz\text{-}OH + H^+$	1.1[3]
$>qz\text{-}O^- \leftrightarrow >qz\text{-}OH - H^+$	8.1[3]
$>qz\text{-}ONa \leftrightarrow >qz\text{-}OH - H^+ + Na^+$	6.8[4]

Arora et al, 2017

3D Plume Modeling and Simulations



Helpful for understanding
- Residual contaminants under the basins and within the clay layer
- Climate change impact (Libera et al., 2019; Xu et al., 2022)

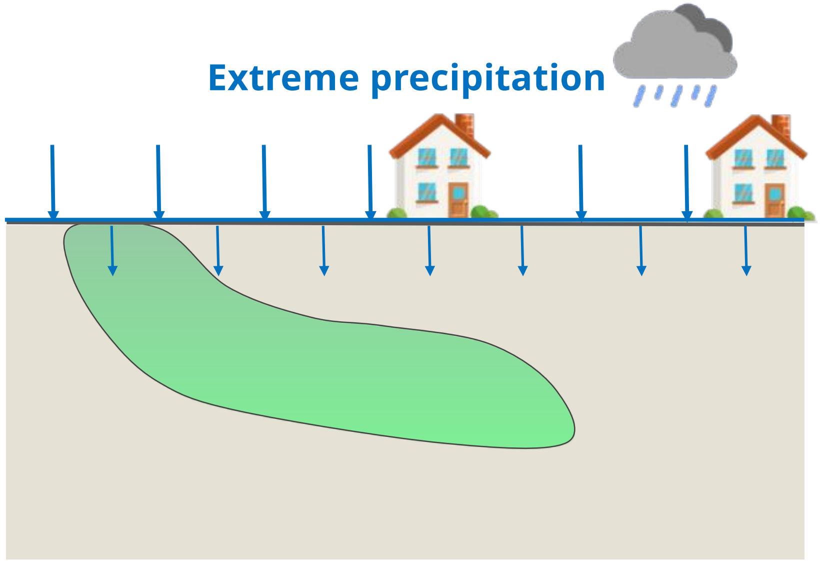
High computational burden
- >1M grid blocks
- Complex geochemical reactions
- Up to 100s of simulations for UQ

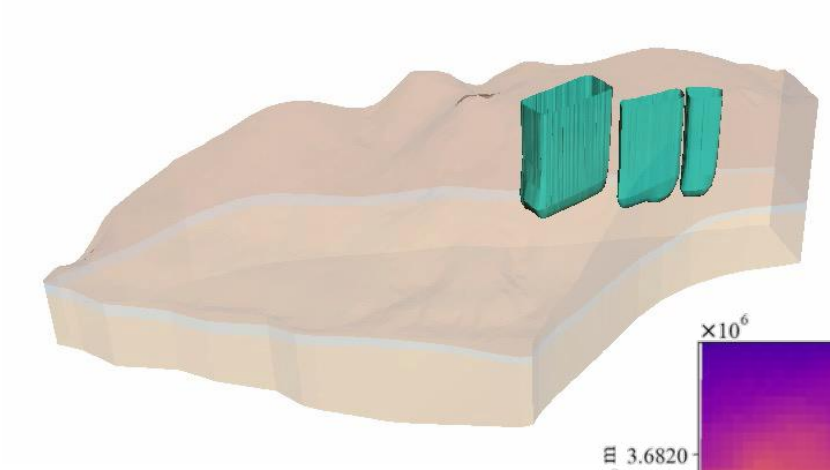
Challenging to fit at all the points
- Geological heterogeneity
- Conceptual model error

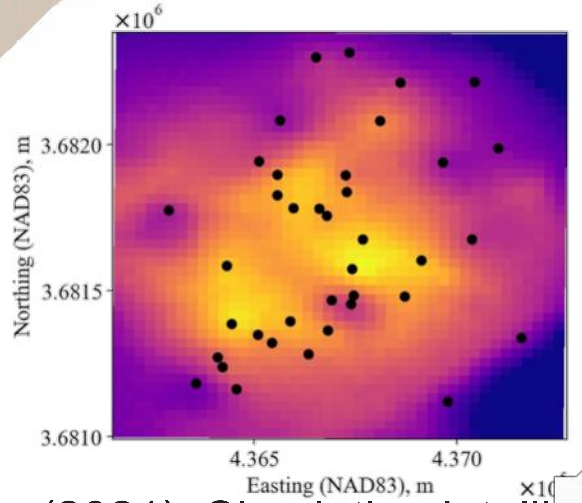
Simulation Intelligence: Simulations x ML/AI

Climate Change Impact on Groundwater contamination
→ Emulator with Fourier Neural Operator

Physics-informed interpolation
→ Model-data integration with Bayesian hierarchical model



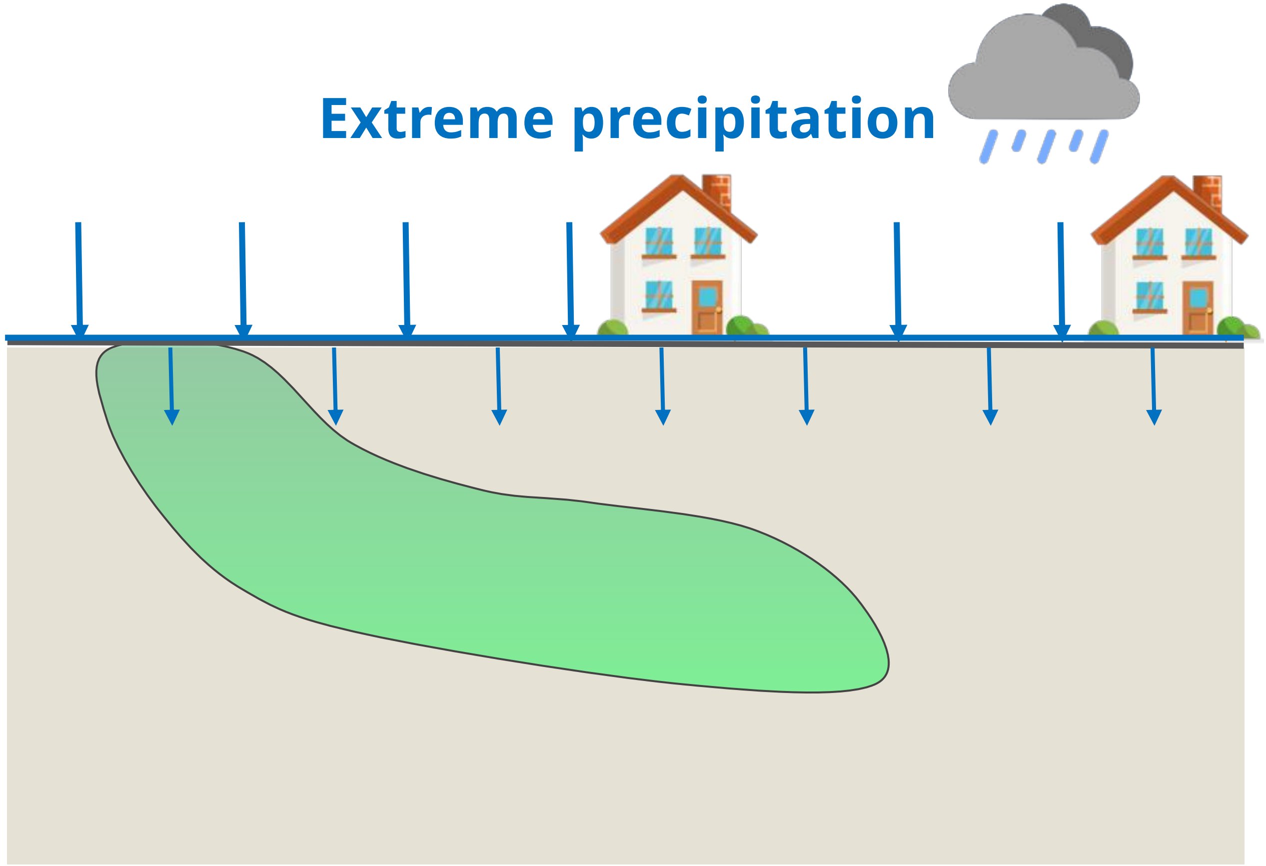




In collaboration with NASA Frontier Development Lab

Lavin, A., Zenil, H., Paige, B., Krakauer, D., Gottschlich, J., Mattson, T., ... & Pfeffer, A. (2021). Simulation intelligence: Towards a new generation of scientific methods. *arXiv preprint arXiv:2112.03235*.

Climate Change Impacts on Contamination



Higher precipitation

- Re-mobilize residual contaminants?
- Dilute concentrations?

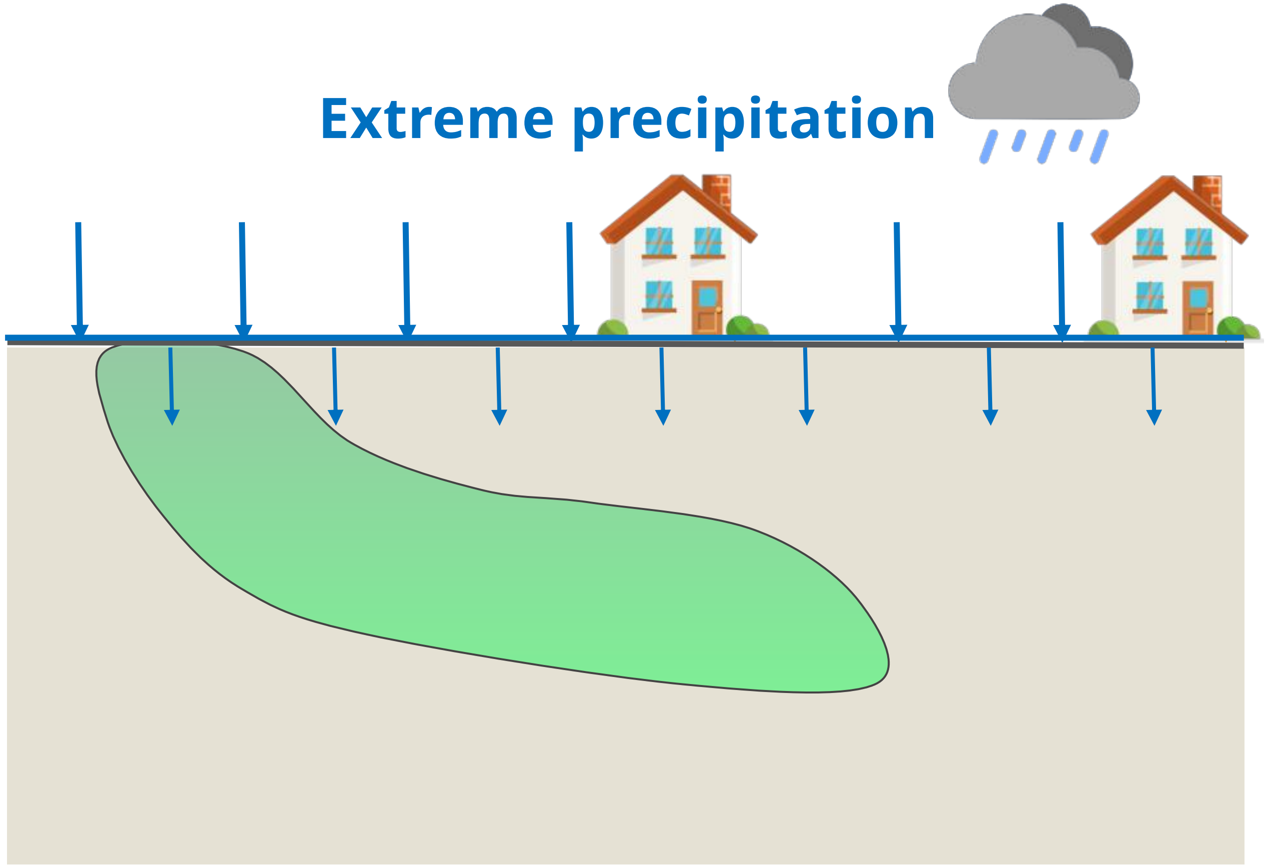
→ Change management strategies?

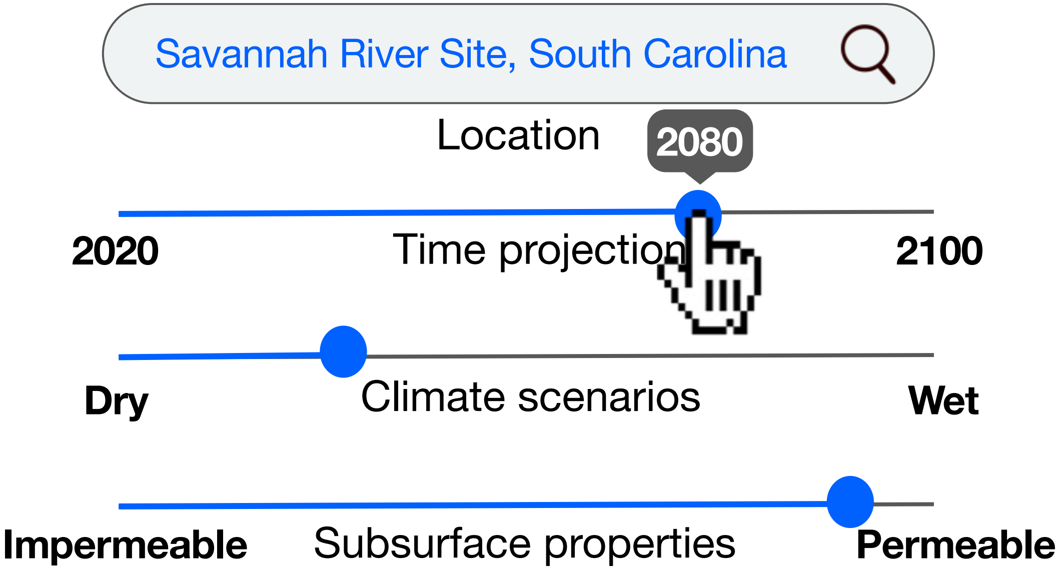
→ Change monitoring well configuration?

Wang, L., Kurihana, T., Meray, A., Mastilovic, I., Praveen, S., Xu, Z., ... & Wainwright, H. (2022). Multi-scale Digital Twin: Developing a fast and physics-informed surrogate model for groundwater contamination with uncertain climate models. *arXiv preprint arXiv:2211.10884*.

Climate Change Impacts on Contamination

When and where to make modification?

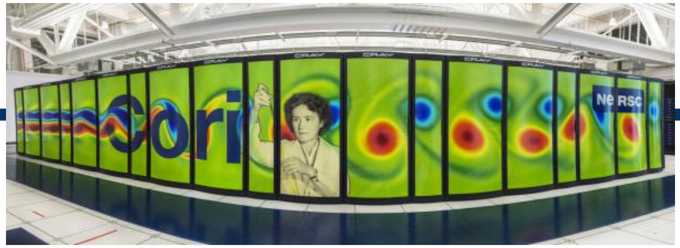




- But computation is pretty heavy
- We can’t run simulation on laptops

Wang, L., Kurihana, T., Meray, A., Mastilovic, I., Praveen, S., Xu, Z., ... & Wainwright, H. (2022). Multi-scale Digital Twin: Developing a fast and physics-informed surrogate model for groundwater contamination with uncertain climate models. *arXiv preprint arXiv:2211.10884*.

Emulator/Surrogate Modeling



HPC Clusters

Parameters from PDF

$\{p_1, p_2 \ldots . p_N\}$

Simulations:

$\phi = f(p)$

$\{\phi_1, \phi_2, .. \phi_N\}$

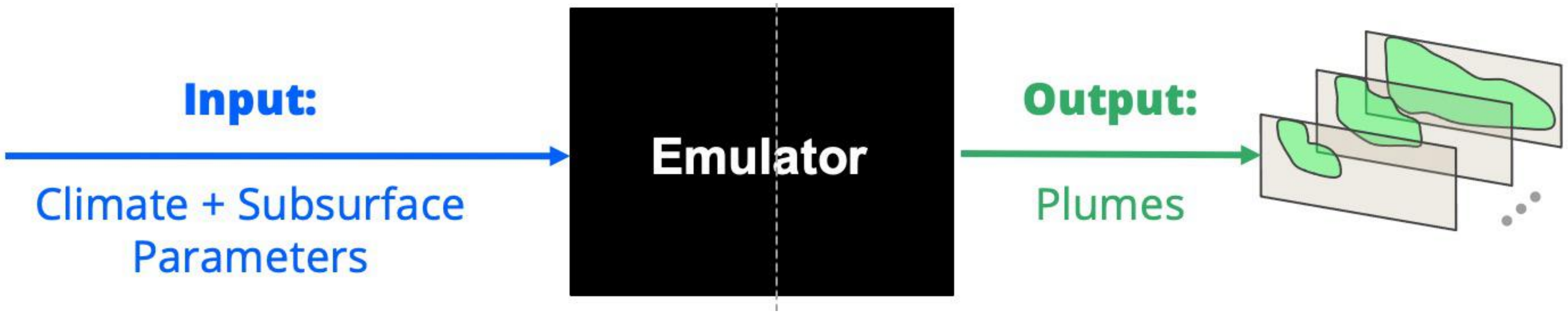
Regressions:

$\phi \sim f(p)$

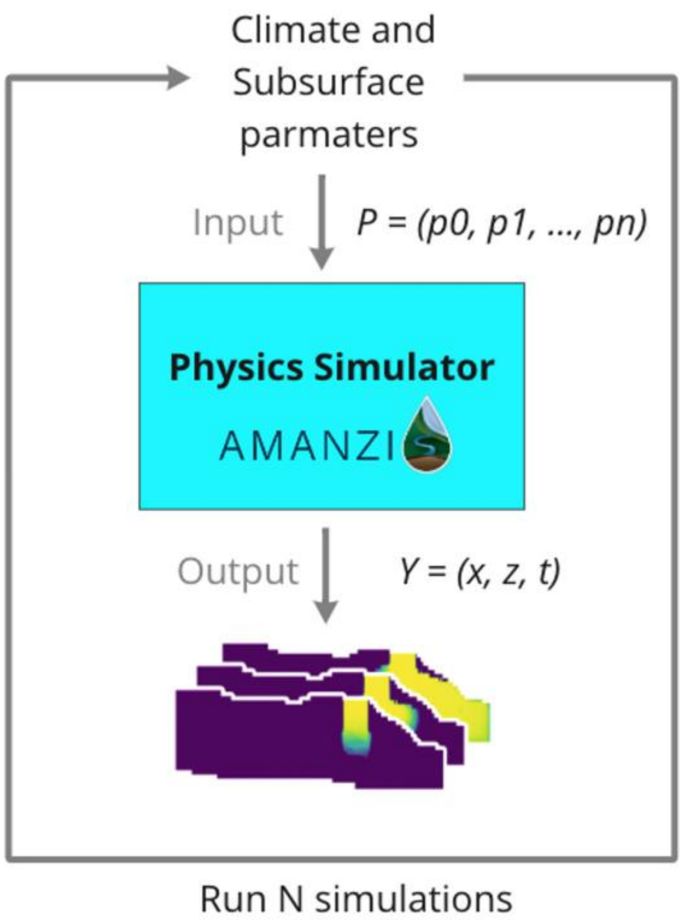
Emulator predictions

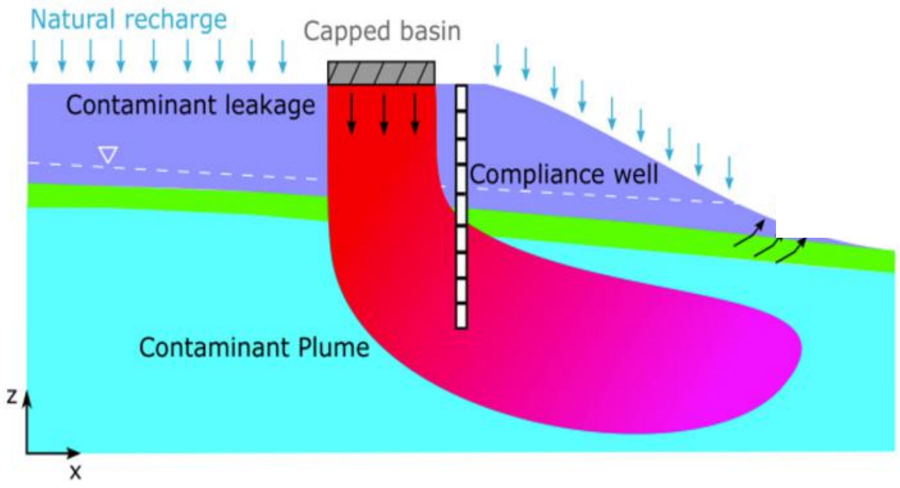
$\phi = f_{emulate}(p)$

Statistical representation of physical models



2D Flow and Transport simulator



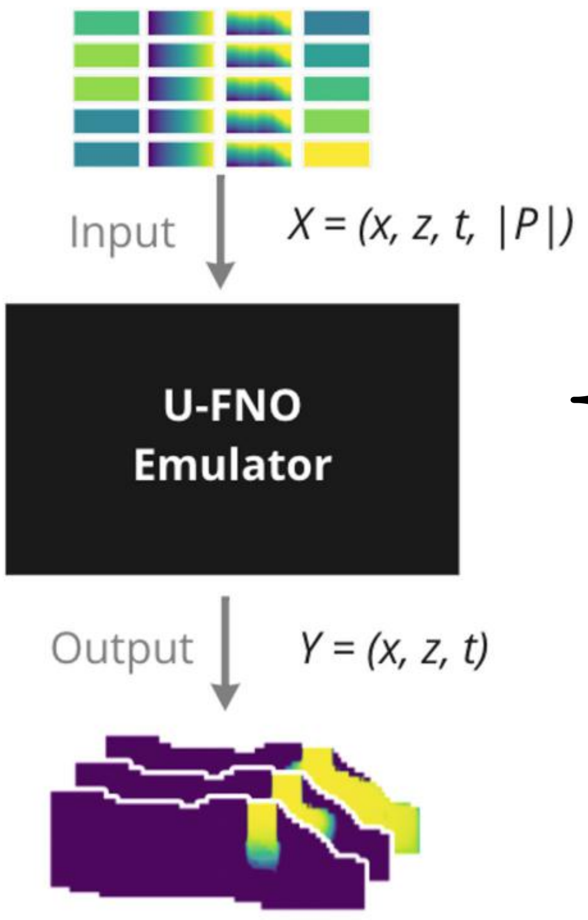


Parameter	Details
Permeability	upper layer
Porosity	upper layer
Alpha	inverse air entry suction
Sr	residual water content
m	m = 1-1/n, a measure of the pore-size distribution
Source concentration	**Initial contaminant concentration**
Discharge rate (Cap at 1988)	**Source/cap discharge rate in volumetric water**
Time-varying recharge	**Climate data (precip. & ET) (history, mid-century, late-century)**

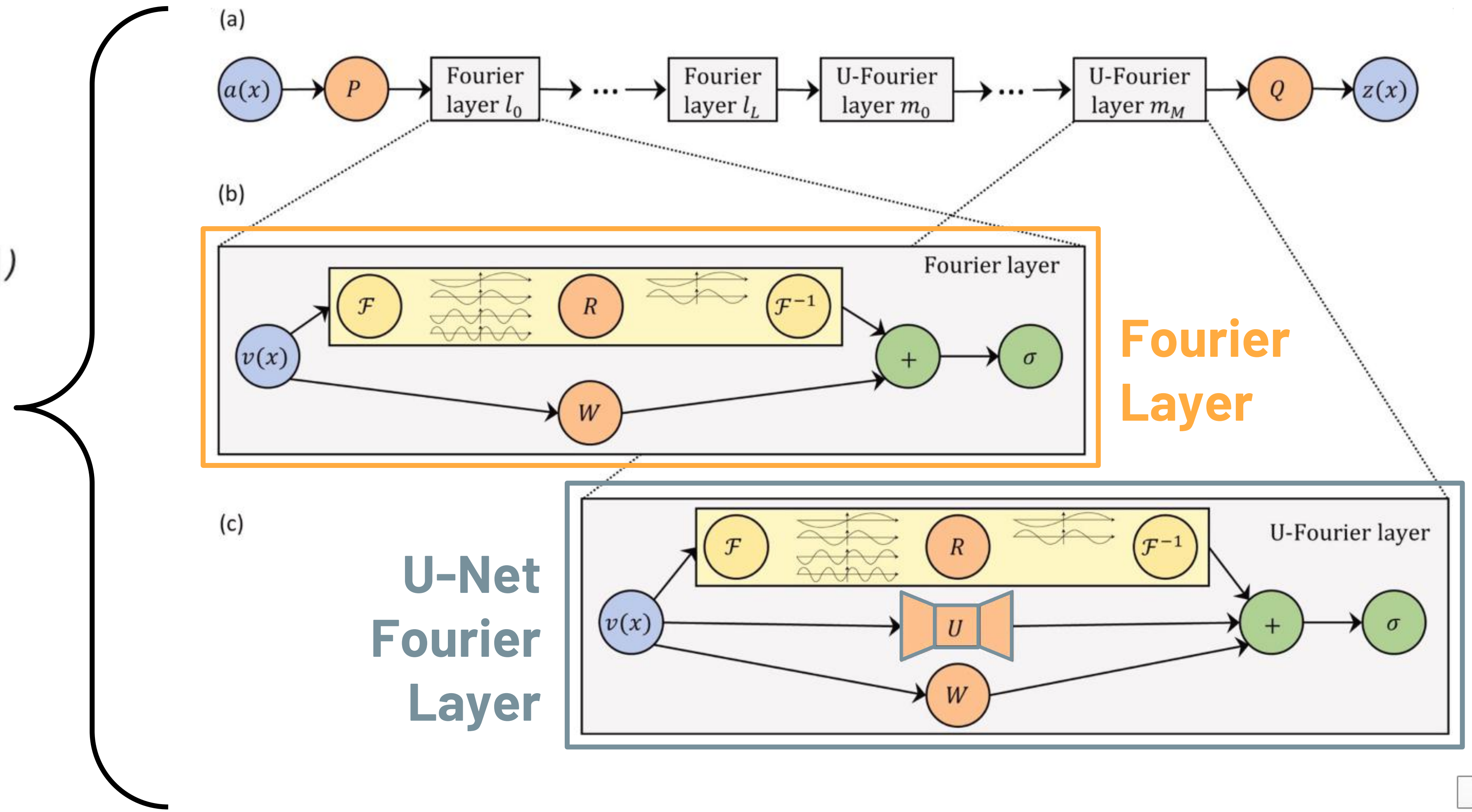
Flow and reactive transport model in 2D

Deep learning architecture: UFNO

Climate and Subsurface parmaters



An enhanced Fourier Neural Operator (Wen et al. 2022)



Data-driven & Physics-informed Loss

A combined loss function to minimizes multiple errors → UFNOB

$$\mathcal{L}(y,\hat{y}) = \mathcal{L}_{MRE}(y,\hat{y}) + \beta_1 \mathcal{L}_{der}(y,\hat{y}) + \beta_2 \mathcal{L}_{plume}(c',\hat{c'}) + \beta_3 \mathcal{L}_{BC}(\hat{y})$$

Mean Relative Error

Derivatives

Contaminant boundary

No flow boundary

$$\mathcal{L}_{MRE}(y,\hat{y}) = \frac{||y-\hat{y}||_2}{||y||_2}$$

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

MCL: maximum contaminant level

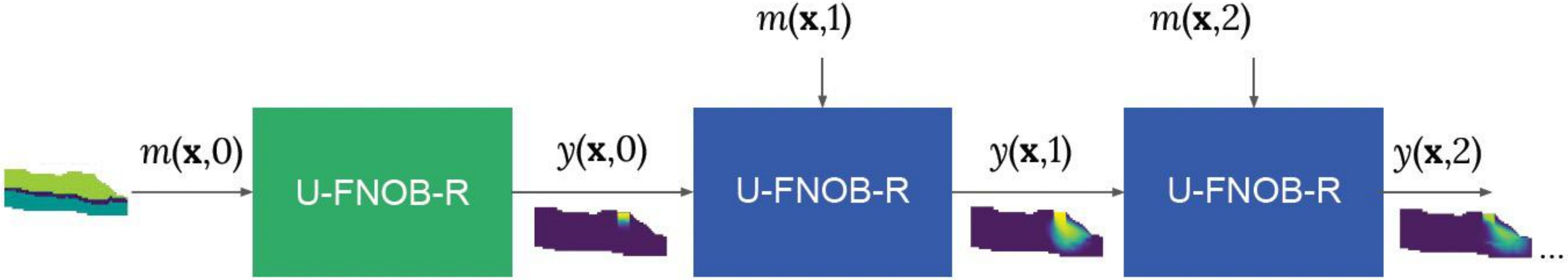
$$\mathcal{L}_{plume}(c',\hat{c'}) = \frac{||\partial c'/\partial x - \partial \hat{c'}/\partial x||_2}{||\partial c'/\partial x||_2} + \frac{||\partial c'/\partial z - \partial \hat{c'}/\partial z||_2}{||\partial c'/\partial z||_2}, ;\text{ where } c' = \begin{cases} 0, & c < MCL \\ 1, & c \geq MCL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

$$\mathcal{L}_{BC}(\hat{y}) = ||\hat{q}_x|_{\partial D}||_2 + ||\hat{q}_z|_{\partial D}||_2 + ||\partial \hat{h}|_{\partial D}||_2$$

U-FNOB and U-FNOB-Recurrent

a) **Architecture 1:** U-FNOB

b) **Architecture 2:** U-FNOB-R **(Recurrent)**

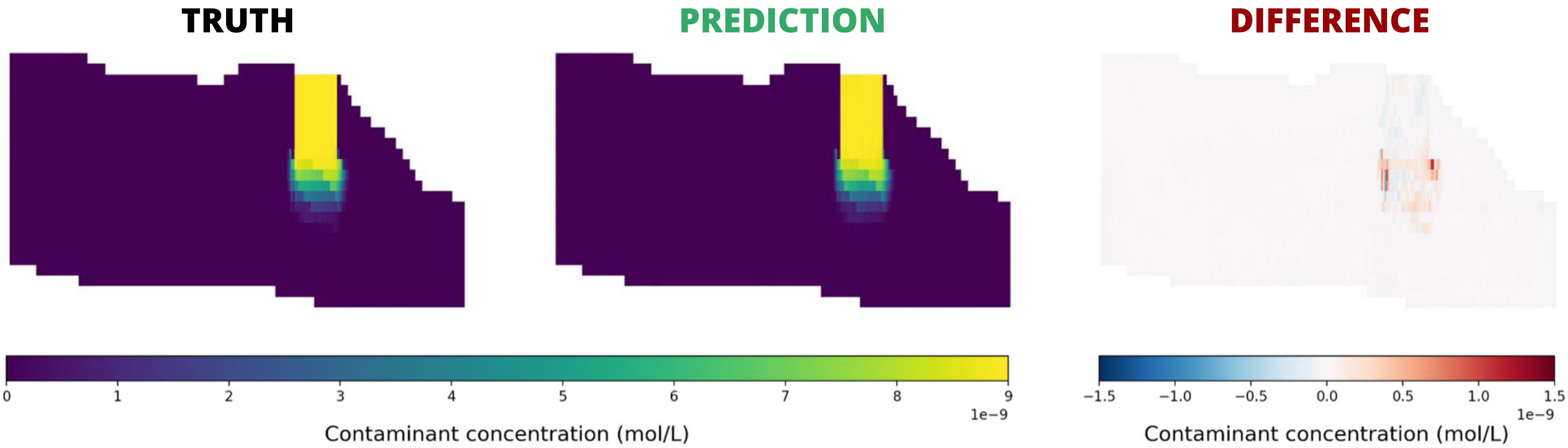


Learn physics for Δt

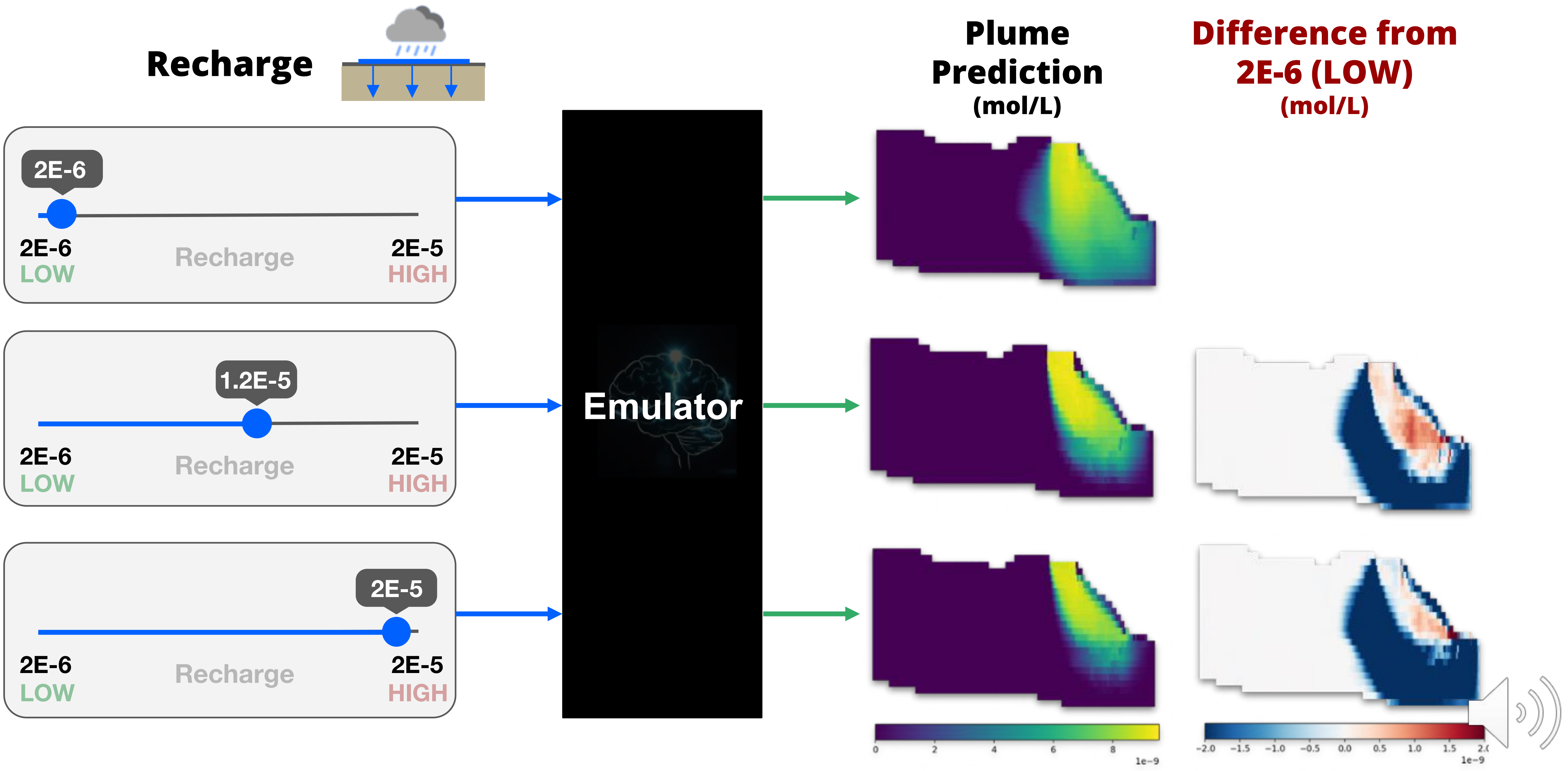
Emulator-based Plume Prediction

1955

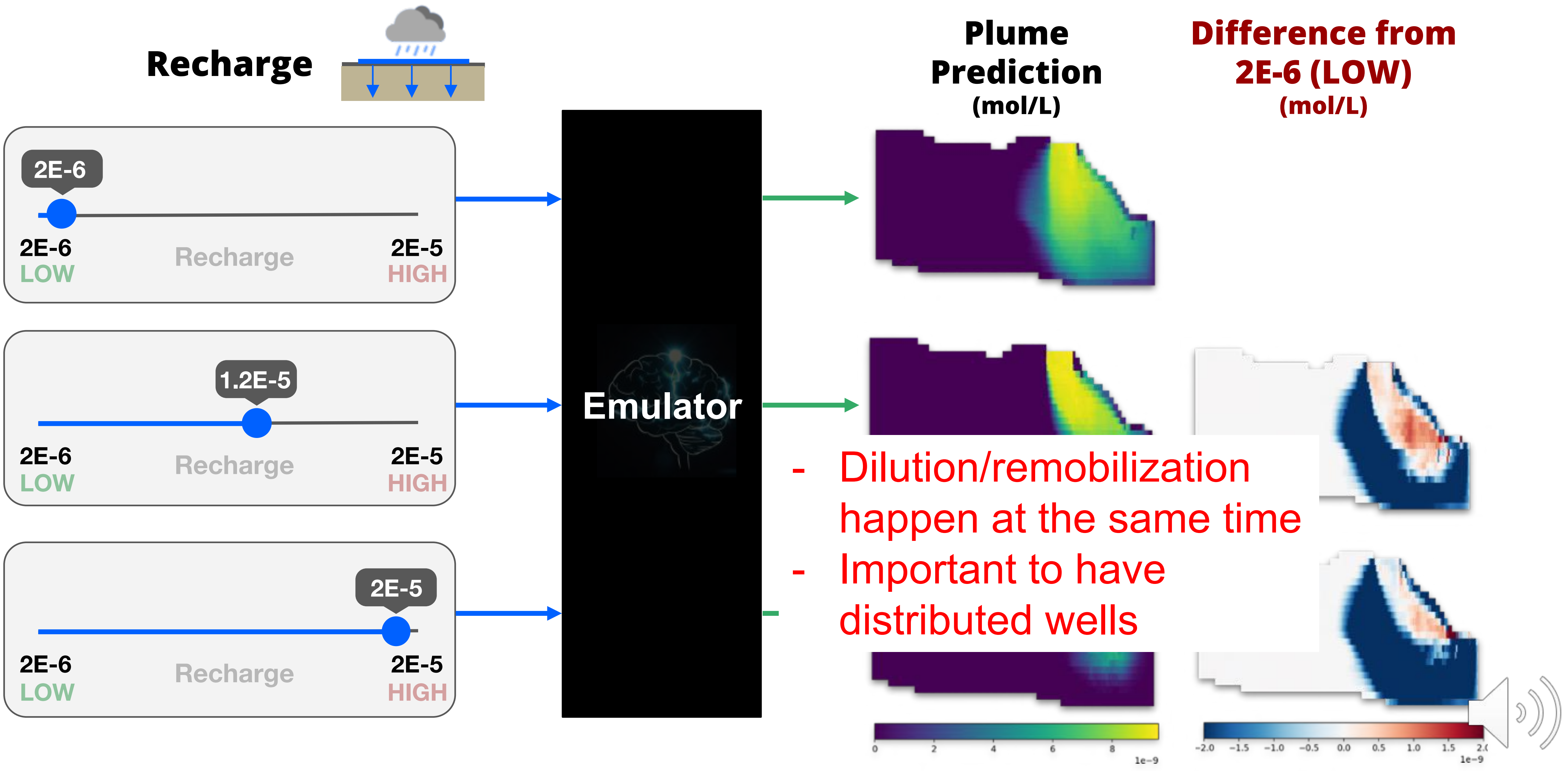
Contaminant Concentration



Off-Line Climate Change Assessment



Off-Line Climate Change Assessment



- Dilution/remobilization happen at the same time
- Important to have distributed wells

Comparison of Different Strategies

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

- FNOB is better than FNOB-Recurrent

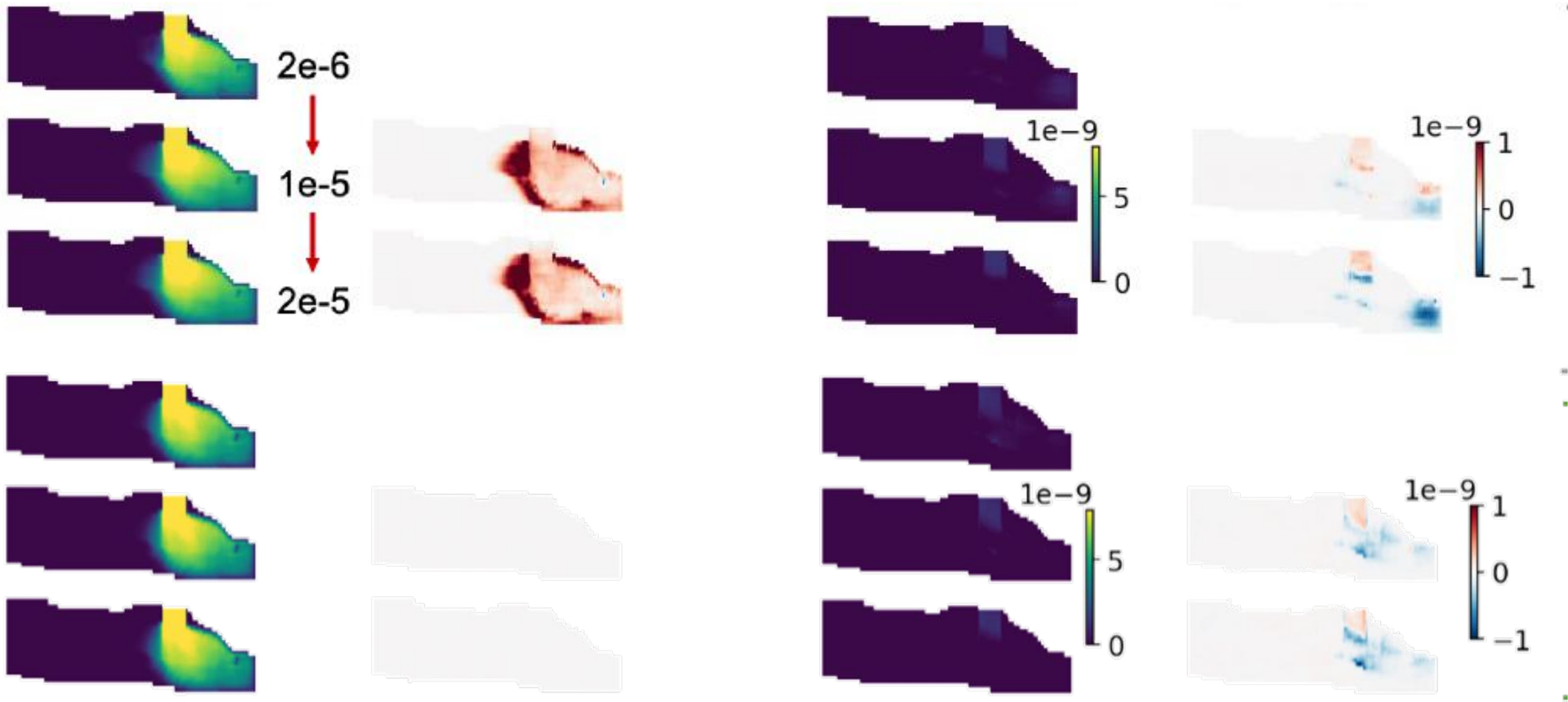
Full Time Series vs Recurrent

Changing the mid-century precipitation (kg-water $m^{-2}s^{-1}$) between 2020 and 2060

1985 | 2025

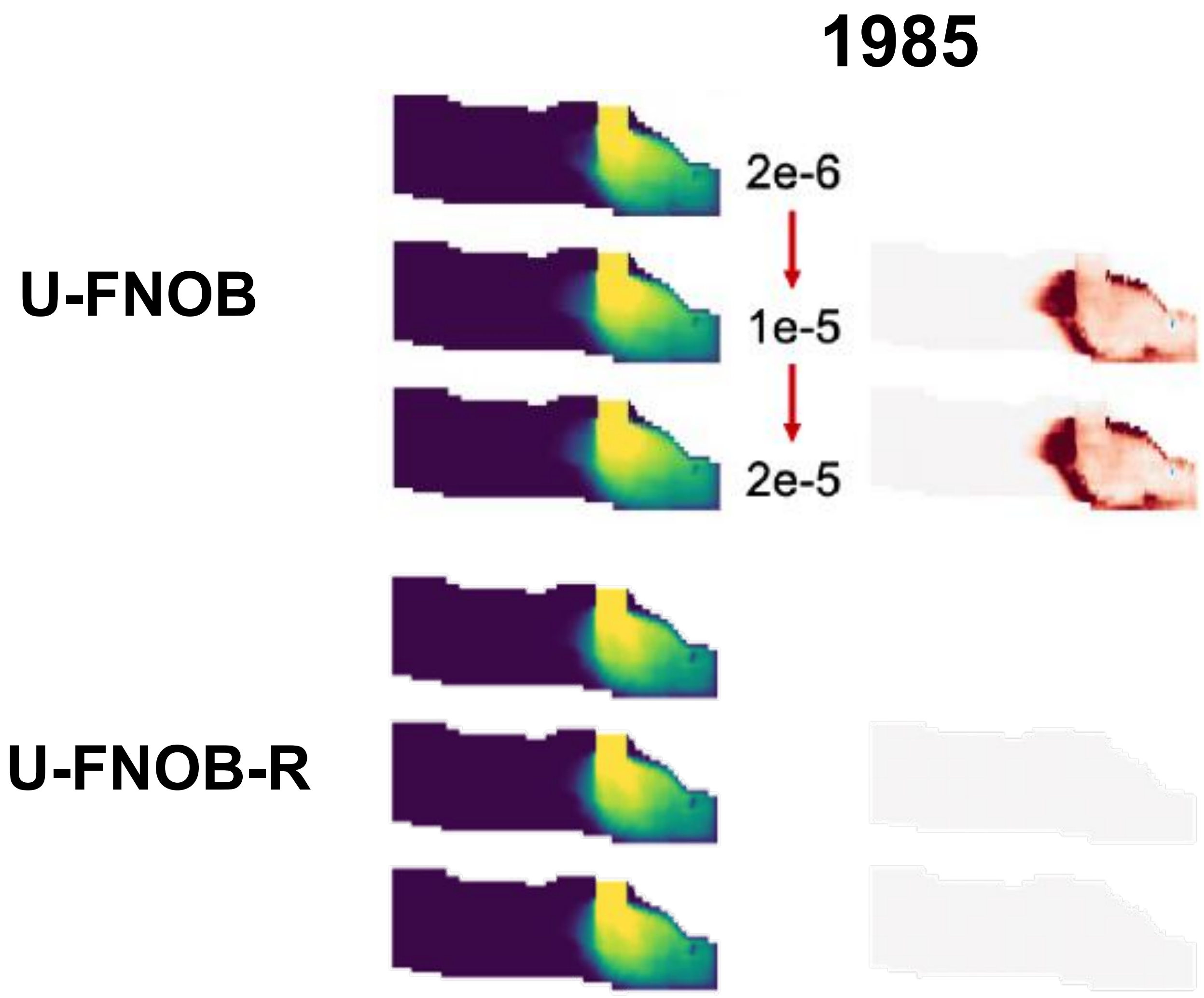
U-FNOB

U-FNOB-R



Full Time Series vs Recurrent

Changing the mid-century precipitation (kg-water $m^{-2}s^{-1}$) between 2020 and 2060

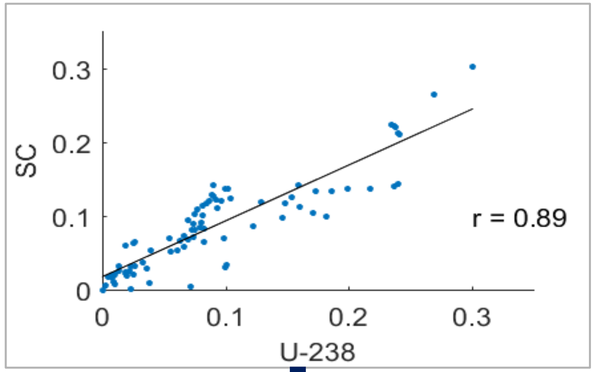


- **FNOB changes plumes in the past**
- **Regression with full time-series does not know past/future**
- **Recurrent one is more realistic**

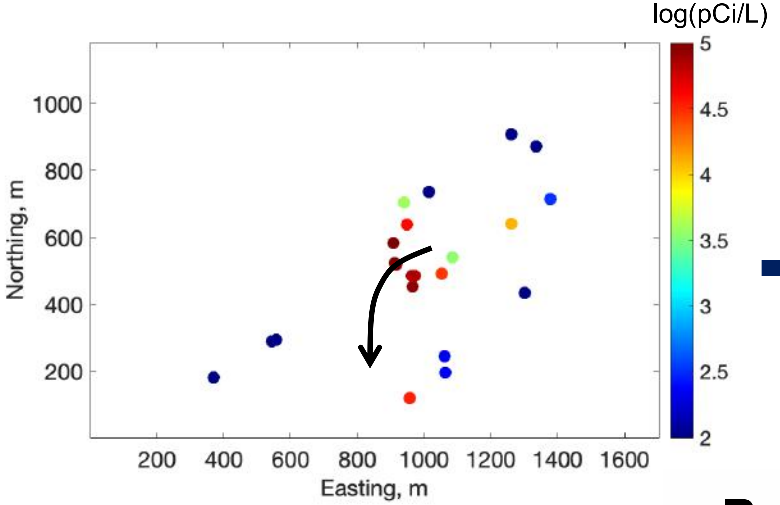
→ Different strategies for time-dependent parameters?

Physics-informed Spatiotemporal Interpolation

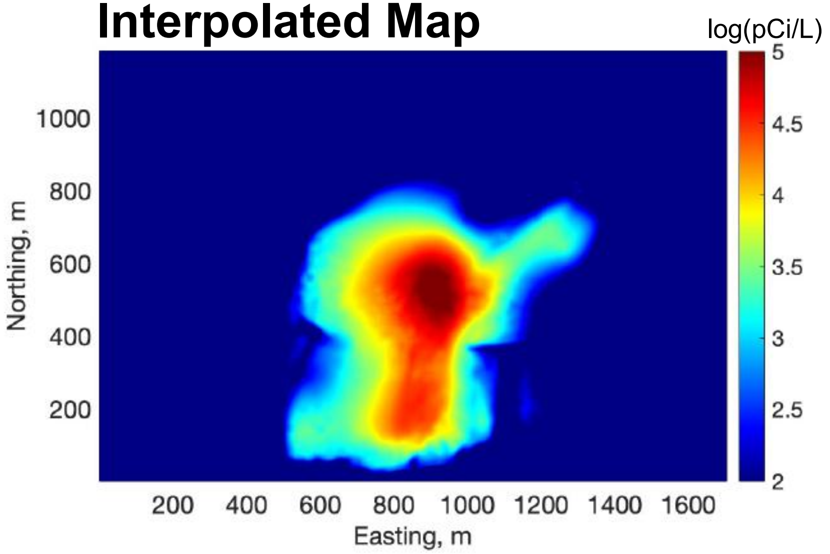
In situ Sensor Data



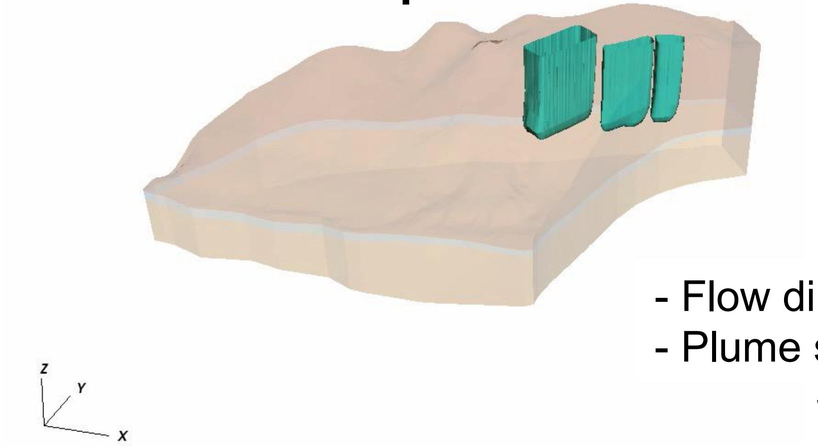
Concentrations at Wells (2015)



Interpolated Map



Reactive Transport Model



- Flow direction
- Plume source

Wainwright et al., 2019

- Ensemble simulations

$$\{y_1, y_2, \dots y_T\}$$

Advances in GP for Large Datasets

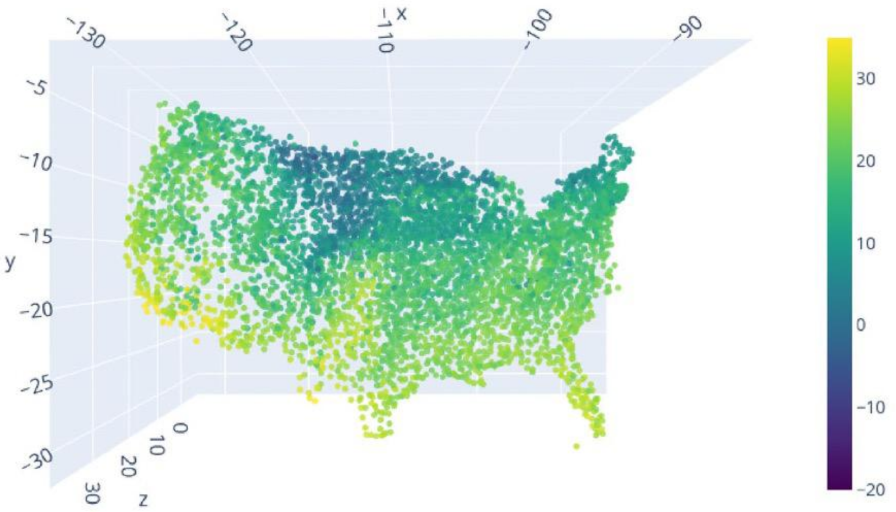
scientific reports

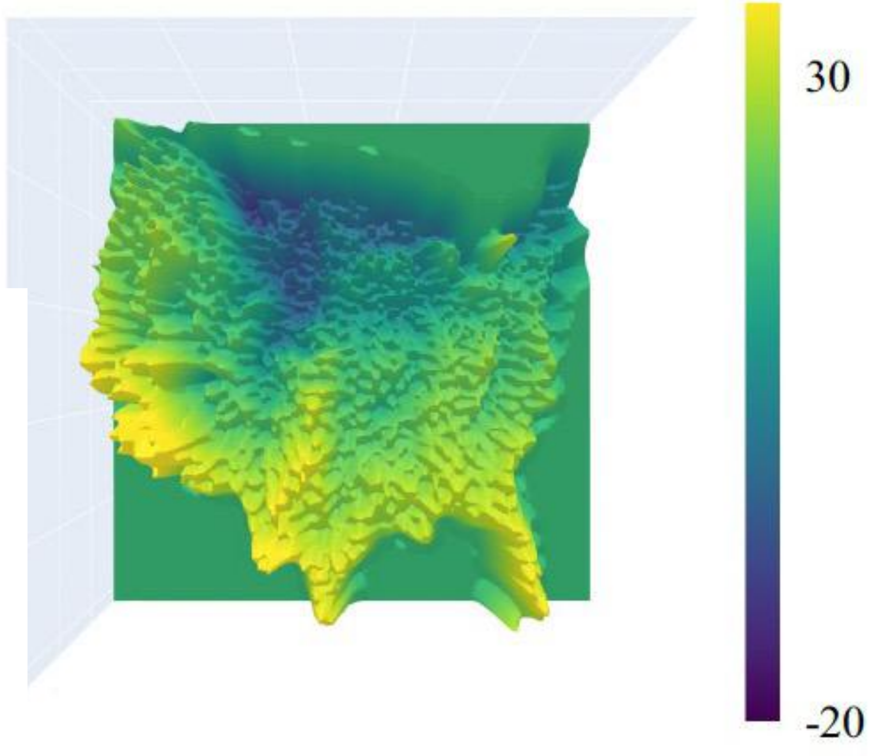
Check for updates

OPEN

Exact Gaussian processes for massive datasets via non-stationary sparsity-discovering kernels

Marcus M. Noack[1]✉, Harinarayan Krishnan[1], Mark D. Risser[2] & Kristofer G. Reyes[3]





(c)

Bayesian Hierarchical Model for Physics-informed Interpolation

- Estimate the spatiotemporal distribution of contaminant concentrations y (ppm) conditioned on groundwater sampling data (z_G) and in situ sensor (z_S)

$$p(\boldsymbol{y}|\boldsymbol{z}_G, \boldsymbol{z}_s)$$

- Ensemble simulations of plume and concentrations: $\boldsymbol{\phi} = \boldsymbol{f}(\mathbf{p})$
- Address the bias and errors of simulations: $\boldsymbol{y} = \boldsymbol{g}(\boldsymbol{\phi}) + \boldsymbol{\varepsilon}$
  - Probably not good for long-term prediction/extrapolation
  - Good for improving the current interpolation

- Gaussian Process Model: $\boldsymbol{y} \sim N(\boldsymbol{g}(\boldsymbol{\phi}), \boldsymbol{C})$, $\boldsymbol{C} = \text{spatially correlated covariance}$

Bayesian Hierarchical Model for Physics-informed Interpolation

- **Posterior distribution**

$$p(\boldsymbol{y}|\boldsymbol{z}_G, \boldsymbol{z}_S) \propto \int p(\boldsymbol{z}_S|\boldsymbol{y})p(\boldsymbol{z}_G|\boldsymbol{y})p(\boldsymbol{y}|\boldsymbol{\phi}(\mathbf{p}), \boldsymbol{\theta})d\boldsymbol{\theta}\, d\mathbf{p}$$

- Data models:
  - $p(\boldsymbol{z}_S|\boldsymbol{y})$: the correlation between sensor data and concentrations

    $p(\boldsymbol{z}_S|\boldsymbol{y}) = N(h(\boldsymbol{y}), \tau^2), \tau^2$ is the measurement error
  - $\boldsymbol{z}_G$: Concentrations plus i.i.d errors

- Prior model:
  - $p(\boldsymbol{\theta})$ → GP parameters
  - $p(\mathbf{p})$ → model parameters

Bayesian Hierarchical Model for Physics-informed Interpolation

- **Ensemble simulations:** $\{\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \dots., \boldsymbol{\phi}_N\}$
- **Jeffrey's prior for $\boldsymbol{\theta}$**

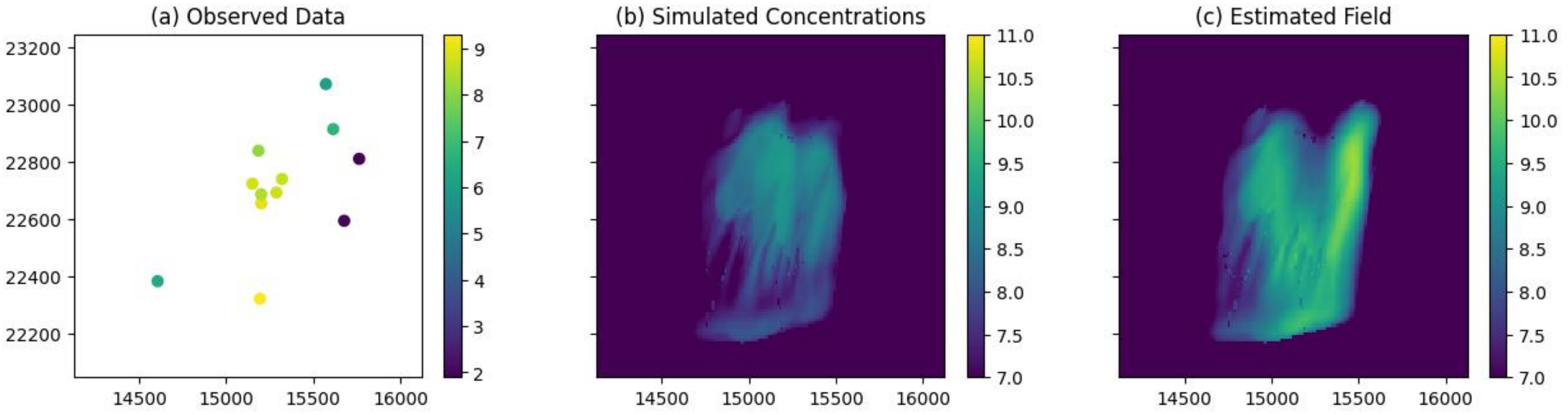
Algorithm 1: Sampling-Resampling Scheme

1. For k from 1 to N:
    1.1. Read the k-th ensemble simulation $\boldsymbol{\phi}_k$.
    1.2. Sample the hyperparameters $\boldsymbol{\theta}$
    1.3. Apply fvGP and estimate $\boldsymbol{y}_k$ and likelihood

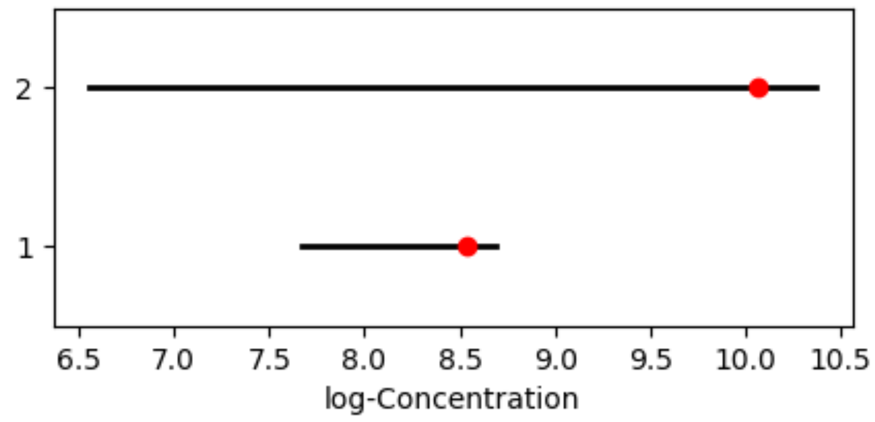
$$L_k = p(\boldsymbol{z_G}|\boldsymbol{y})p(\boldsymbol{z_S}|\boldsymbol{y})p(\boldsymbol{y}|\boldsymbol{\phi}, \boldsymbol{\theta})$$

2. Resample $\boldsymbol{y}_k$ based on the likelihood L_k for the posterior distribution

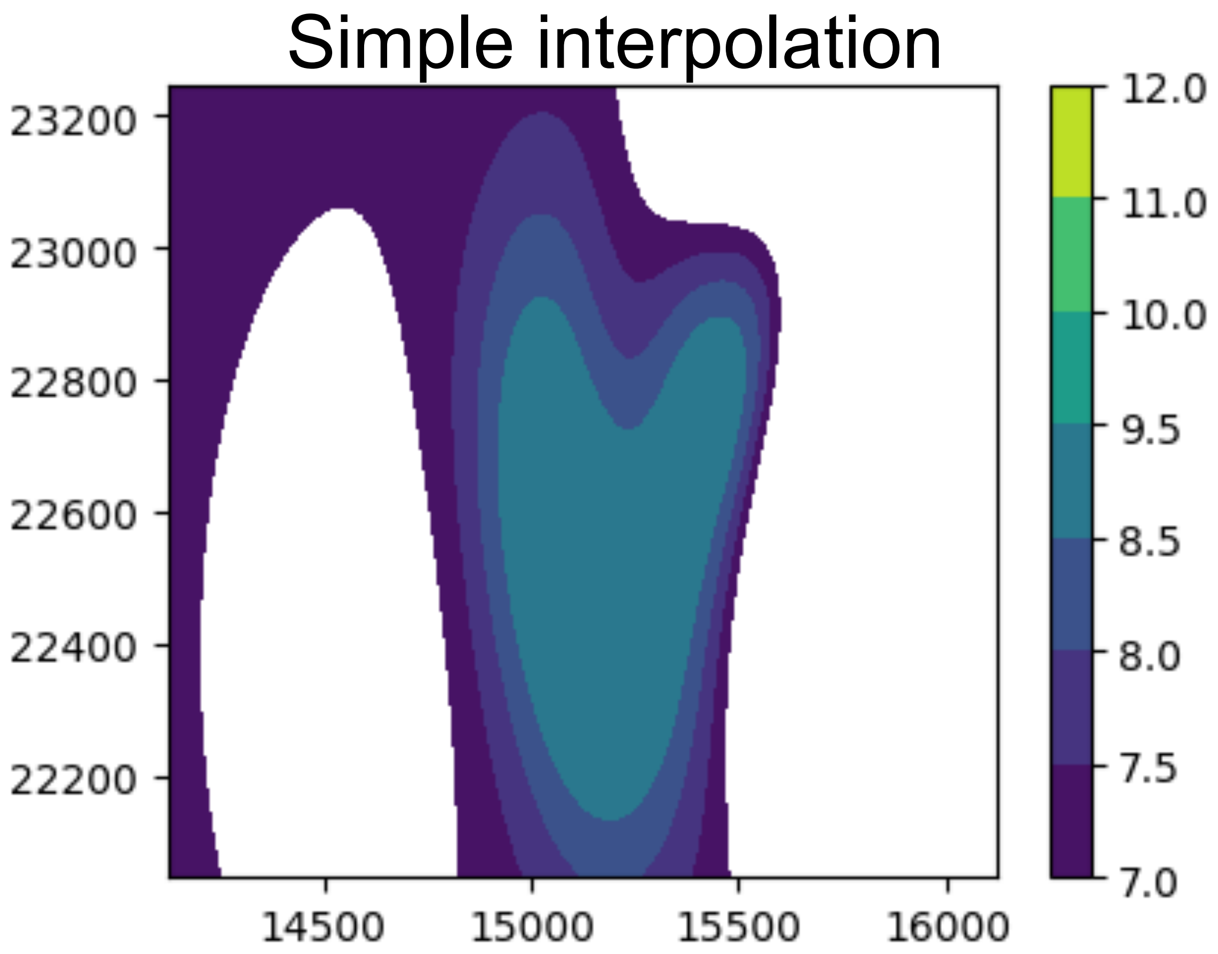
Physics-informed Spatiotemporal Interpolation

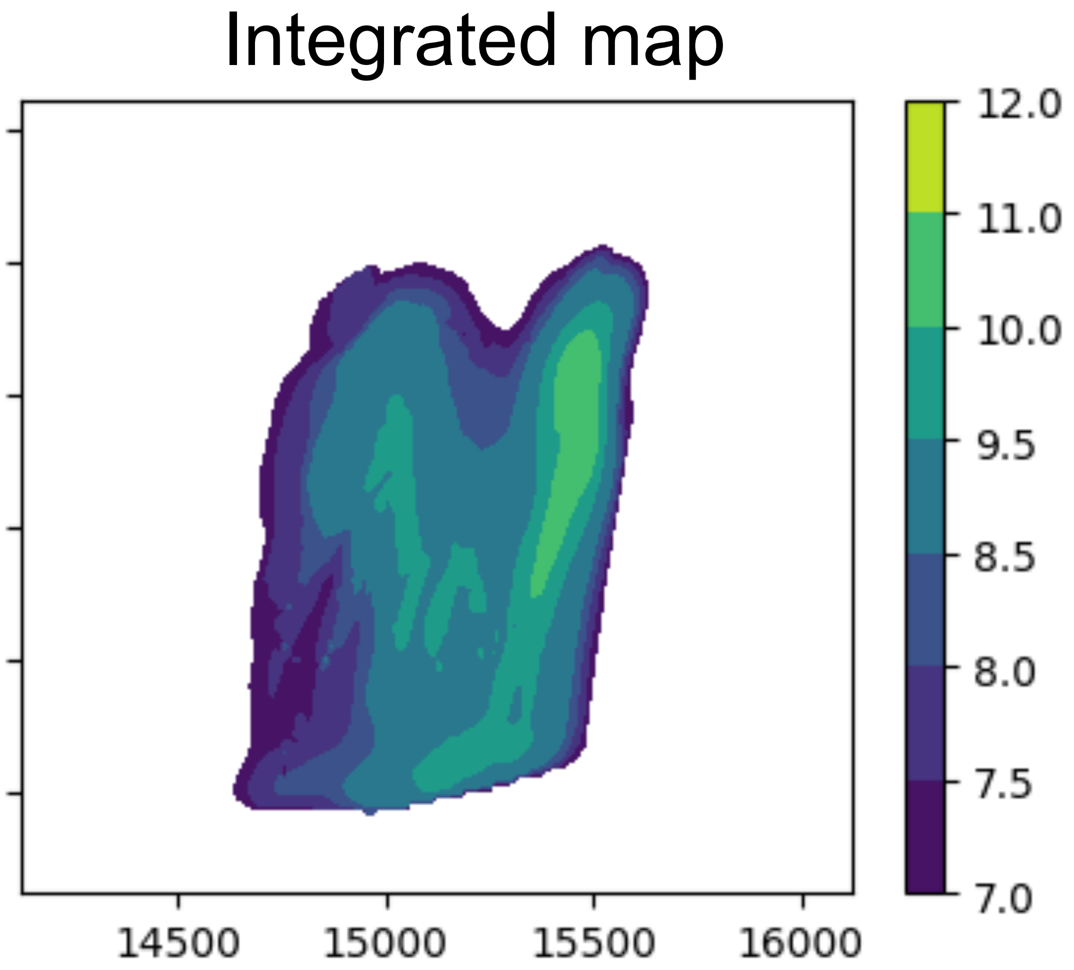


Performance Confirmation
- Confidence intervals
- Points not included in the estimation



Physics-informed Spatiotemporal Interpolation





Impact on Plume Extent: The integrate plume map capture the source locations, plume direction and dispersion

ML Pathway to Adaptation: Challenge

- **Conceptual model development is difficult**
  - Geological heterogeneity, unknowns
- **It is a sensitive topic**
  - Failed weather prediction → 🤷‍♀️
  - Failed contaminant transport prediction → legal actions…
  - QA/QC of codes
- **Regulations**
  - Processes/paperwork for sensor installment, monitoring modification

ML Pathway to Adaptation: Opportunities

- **Understand regulations**
  - Stepwise implementation: in situ sensor deployment
  - Reducing sampling frequencies is easier
  - Then reducing # variables and reducing # wells
- **Emphasize additional safety assurance**
  - Continuous monitoring → early warning, explaining anomalies
  - Guide monitoring strategies (e.g., climate change)
- **Autonomous/autonomous monitoring → AI-assisted monitoring**
  - Anomaly detection → instrument failure, system changes
  - Realistic plume visualization
  - Digital twin → simulate what can happen in the future

NuclearNewswire

POWER & OPERATIONS

Importance of environmental monitoring for consent-based siting of nuclear facilities

Sat, Nov 19, 2022, 6:04AM | Nuclear News | Haruko Wainwright and Carol Eddy-Dilek

a

Distributed Sources: Agriculture Runoff

- **Runoff water may contain:**
  - Soil
  - Nutrients: nitrogen, phosphorous, trace metals
  - Pesticides: herbicides, insecticides, fungicides
- **Impacts on water quality:**
  - Decreased water quality
  - Harmful algal blooms
  - Fish kills

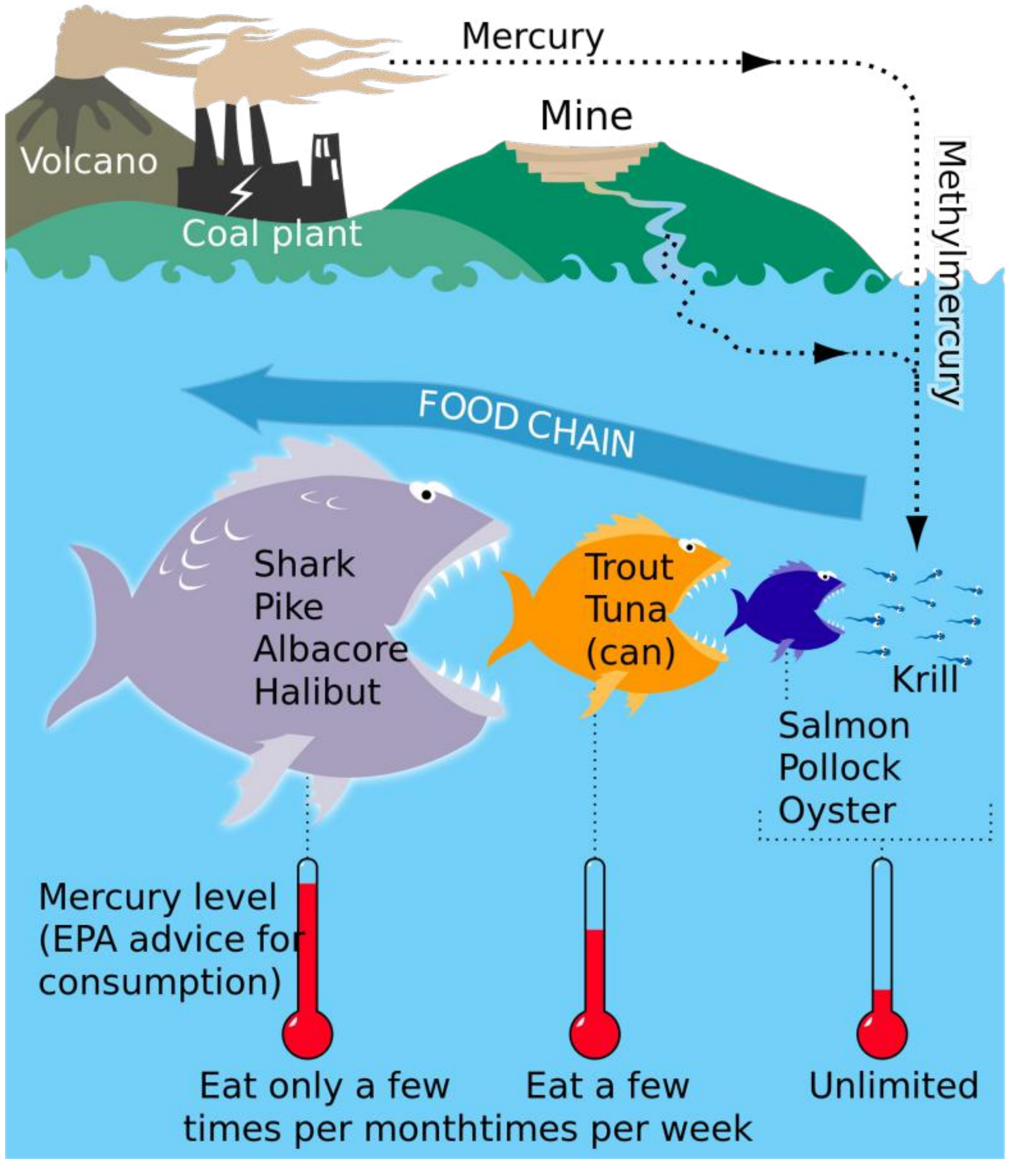
At least one pesticide was found in about:
- 94 percent of water samples and
- 90 percent of fish samples taken from streams across the Nation
- In nearly 60 percent of shallow wells sampled.

(https://www.usgs.gov/mission-areas/water-resources/science/agricultural-contaminants)

https://www.usgs.gov/media/images/farmers-application-fertilizer

Distributed Sources: Mercury

The biggest single source of mercury is the **burning of fossil fuels, especially coal**, which releases 160 tons of mercury a year into the air in the US alone. (Woods Hole Oceanographic Institution)



http://www.groundtruthtrekking.org/Graphics/MercuryFoodChain.html)

Other Distributed Sources/Contamination

Science

Current Issue | First release papers | Archive | About

PERSPECTIVE | TOXICOLOGY

Microplastics and human health

Knowledge gaps should be addressed to ascertain the health risks of microplastics

A. DICK VETHAAK AND JULIETTE LEGLER Authors Info & Affiliations

SCIENCE • 12 Feb 2021 • Vol 371, Issue 6530 • pp. 672-674 • DOI: 10.1126/science.abe5041

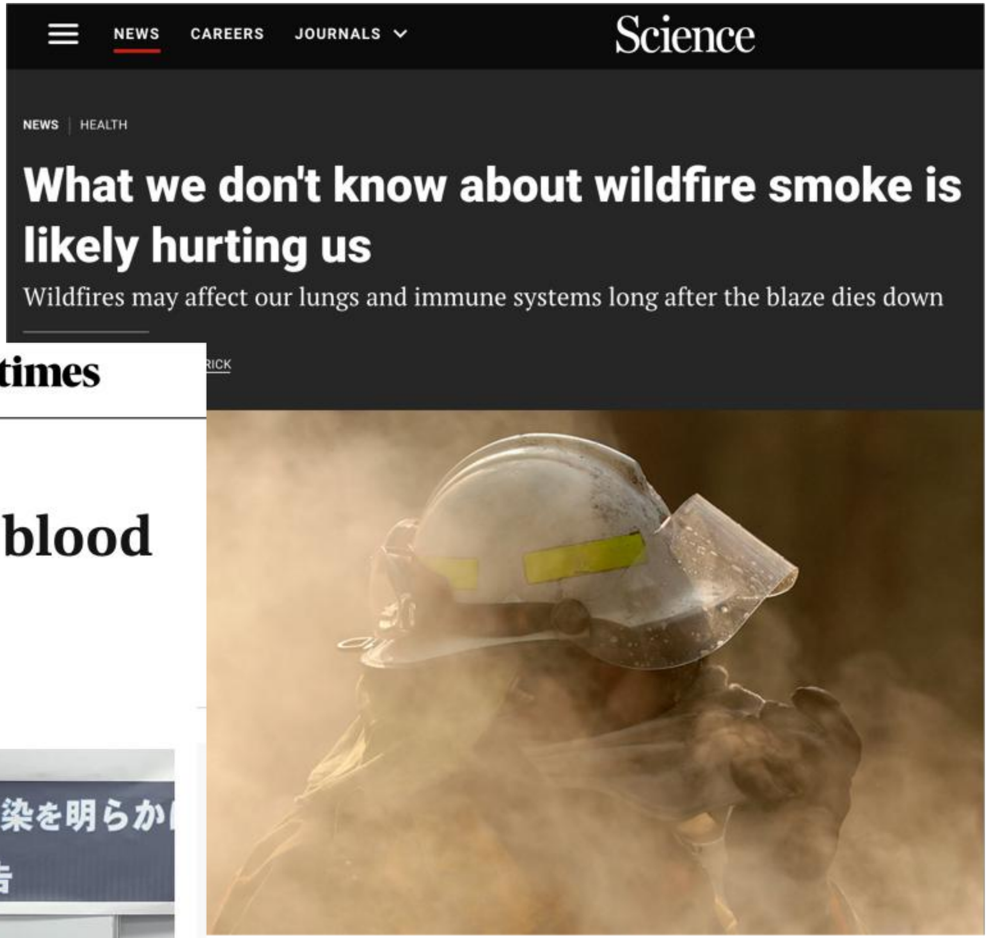
20,681 | 265

Humans are exposed to different types of fibers and particles, including micr… effects of microplastics are largely unknown. PHOTO: DICK VETHAAK

the japan times

JAPAN / SCIENCE & HEALTH

PFAS in western Tokyoites' blood more than double national average



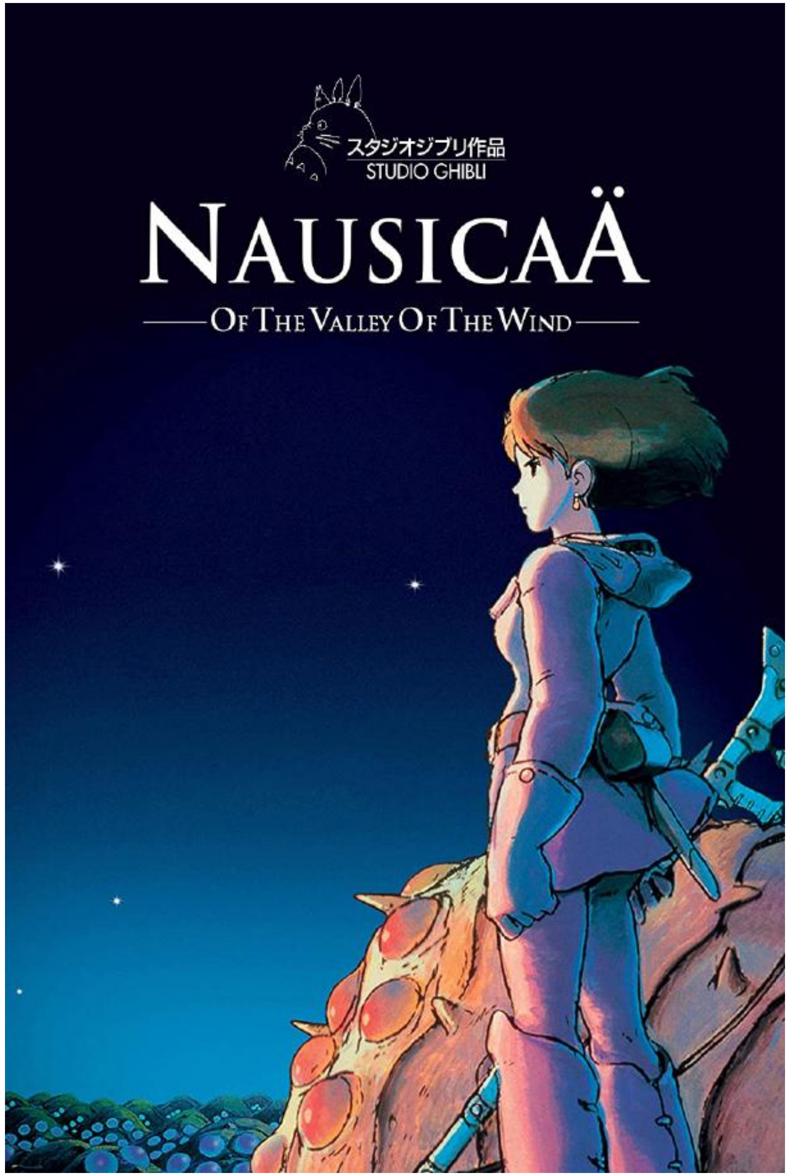
NEWS | CAREERS | JOURNALS | Science

NEWS | HEALTH

What we don't know about wildfire smoke is likely hurting us

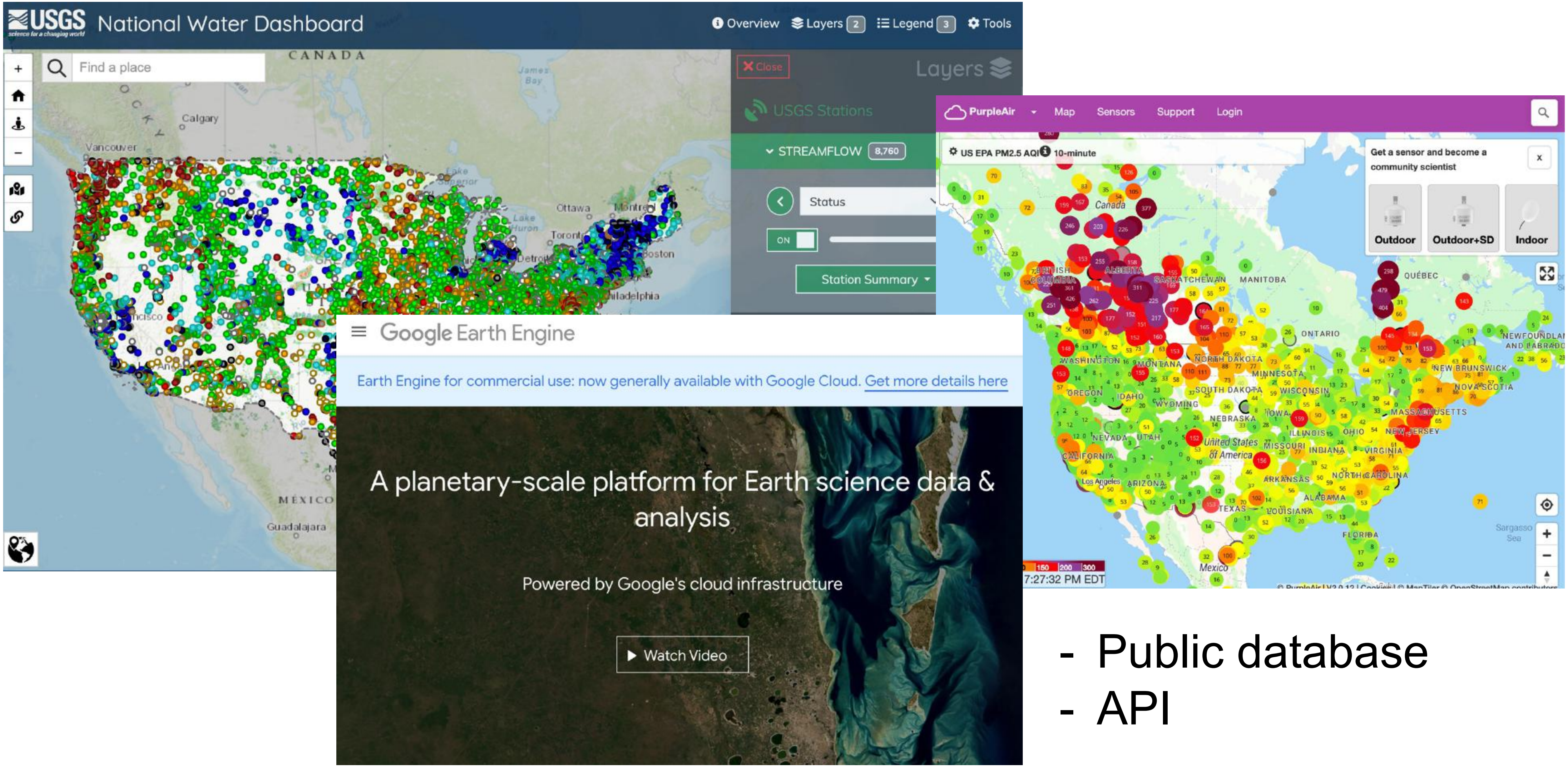
Wildfires may affect our lungs and immune systems long after the blaze dies down

Environmental Science is Critical



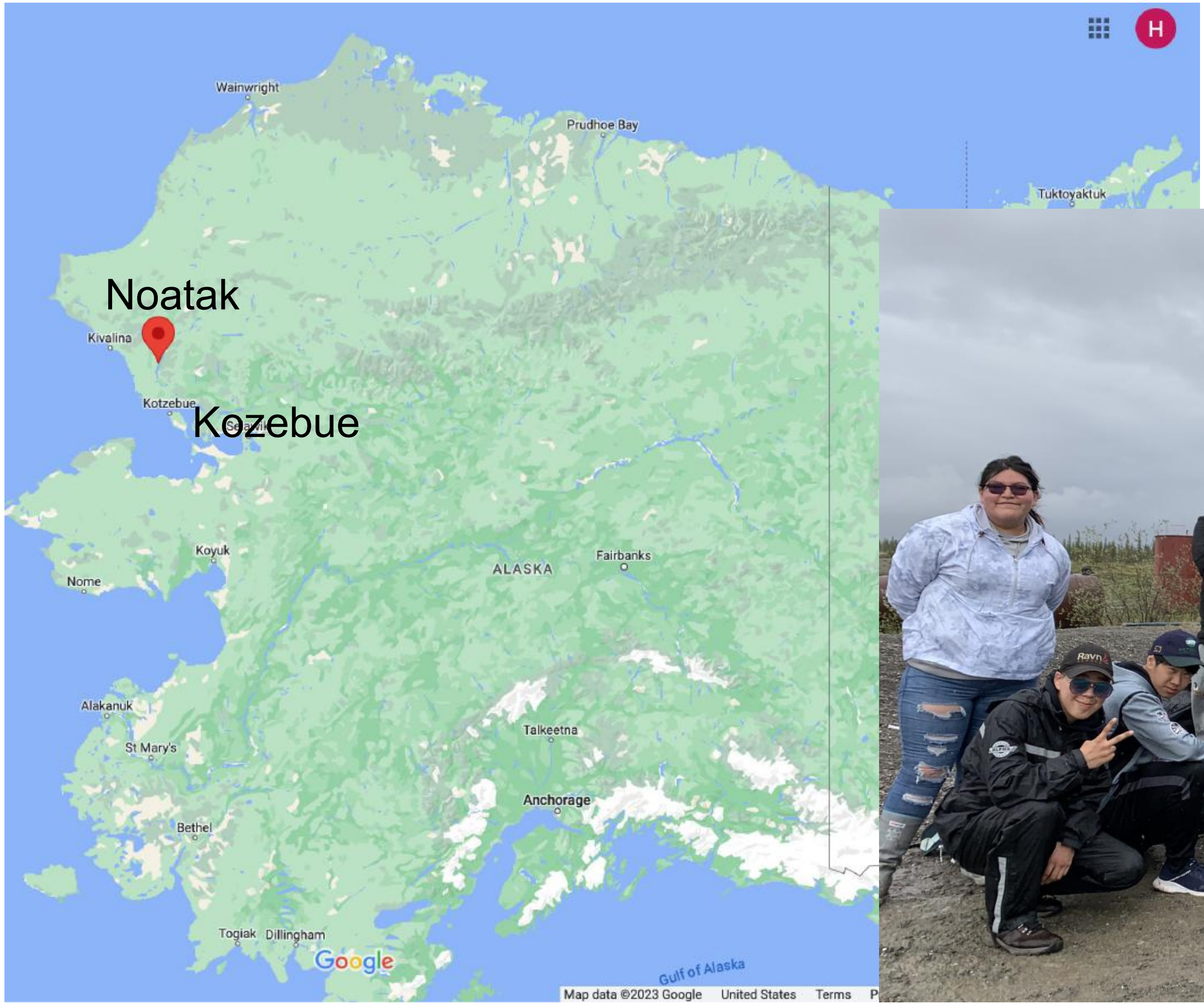
- Our environment is more polluted/contaminated than people think
- Often substances that people don’t worry too much end up spreading out widely and impacting our life
- Everyone needs to be more aware of pollution issues, and more vigilant to protect our health

Open Data, Open Science



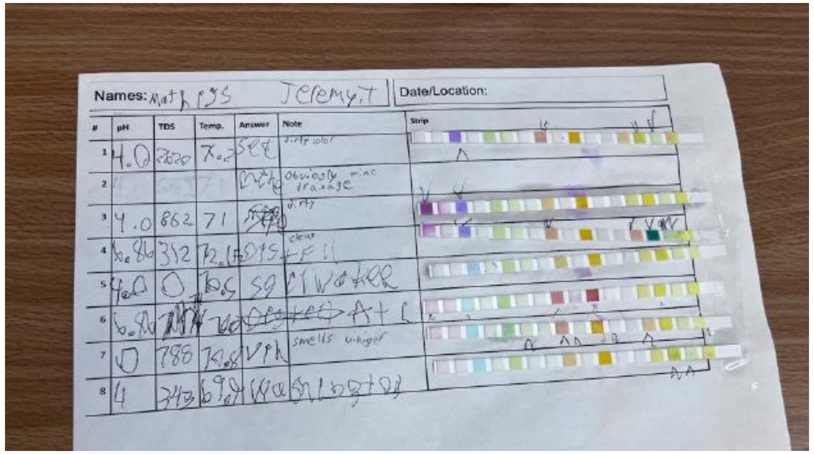
- Public database
- API

Toward Citizen Science: K-12 Education

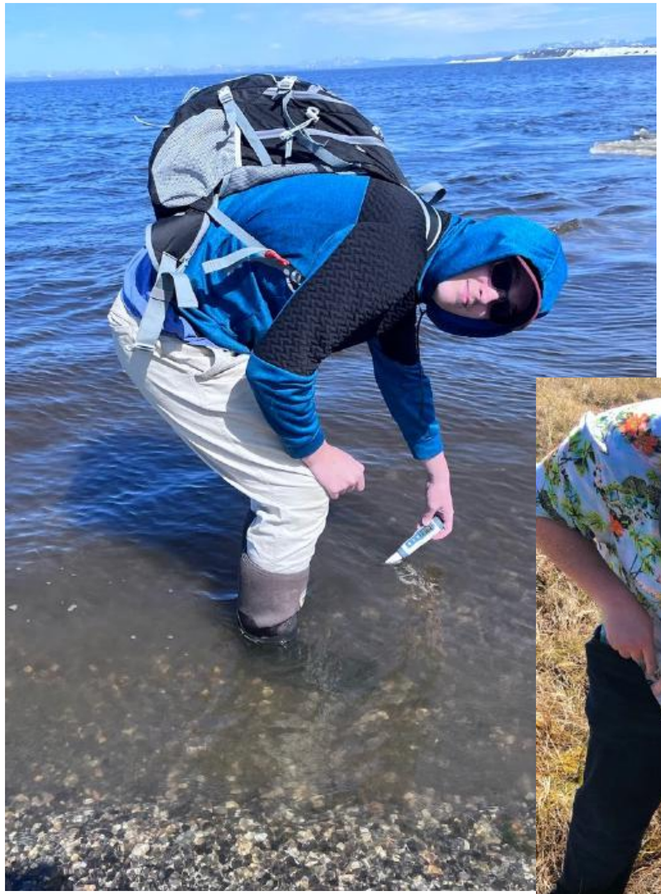


T3 Teaching Through Technologies

Hands on Activities with Sensors



Sensor Technologies for Teaching



GP Data Integration for Air Quality

GP4AQ

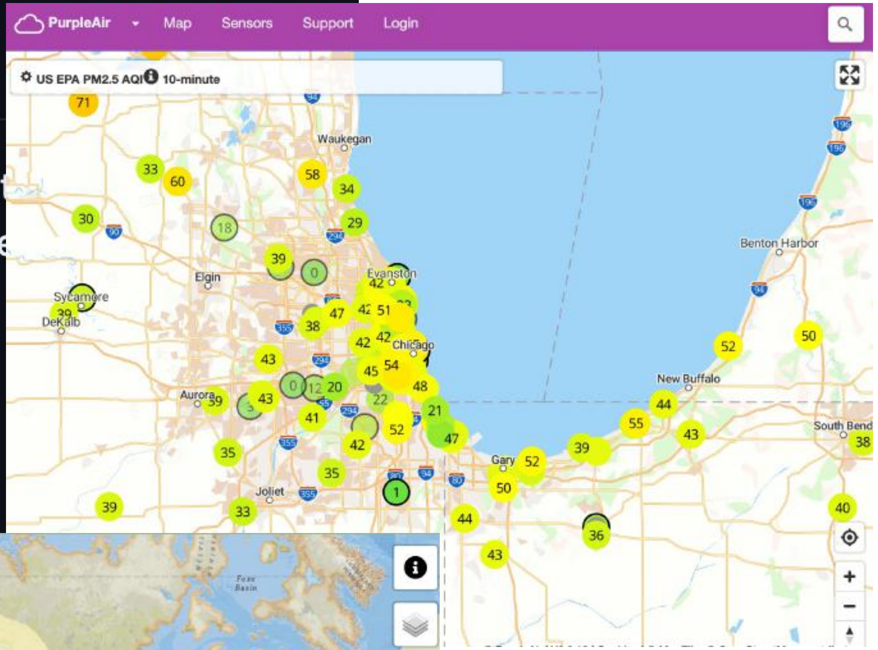
github.com/hmwainw/GP4AQ

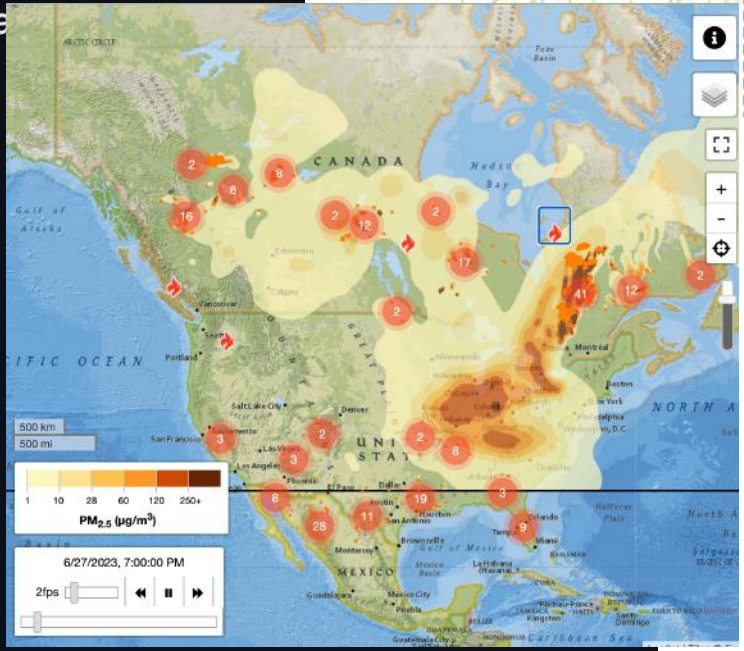
These Jupyter notebooks demonstrate the data integration of EPA and Purple air sensor dat
simulation results, using the Gaussian Process regression, for interpolating and mapping the
over space.

Note:

- You have to have purpleair_chicago.txt in the Google Drive root directory.
- The same functions appear in multiple notebooks. These notebooks a
purpose.

1. Define the grid and domain for spatial estimation
2. Download EPA air quality data
3. Download Purple air quality data
4. Download plume simulation maps
5. Interpolate EPA air quality data
6. Interpolate Purple air quality data
7. Integrate EPA and Purple air quality data
8. Integrate EPA, Purple air, and plume simulation data





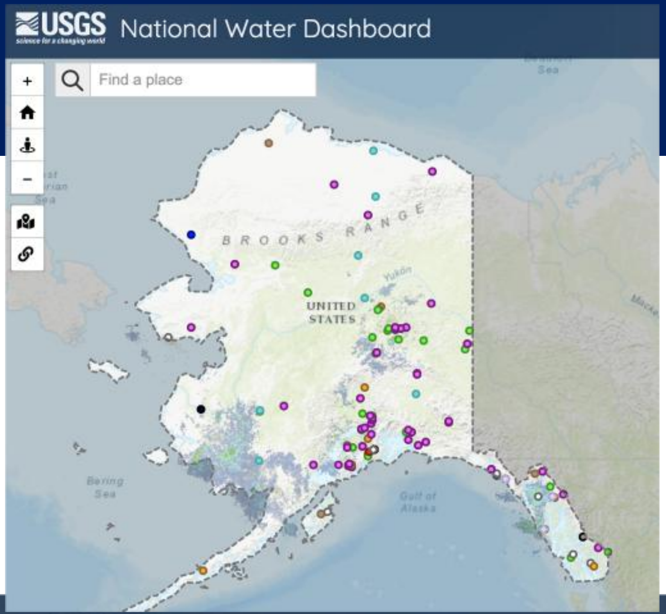
Next Step

- Environmental Monitoring Network for rural America

- USGS Database does not cover rural regions
  - Too remote
  - Data quality concern

- Can high schools be the base for environmental network in rural regions?

- Improve STEM education
  - More college/PhD from rural regions!



Challenges….. Tech/AI in Environment/Climate

- **Pollution monitoring is not exciting when nothing happens**

  → **Attributes more relevant to daily life?**

  - River temperature for fishing?
  - Soil moisture sensors for gardening?

- **Students good at math/science are not interested in the environment and climate**

→ **Environmental data in math/statistics education?**

  - Open data and problem sets?

Summary

- **Long-term monitoring of soil and groundwater contamination**
  - Sustainable remediation: long-term institutional control
  - Ensure the stability/safety of contaminated sites and detect anomalies
- **ALTEMIS: Multiscale multi-type data integration**
  - Integration of proxy information (e.g., spatial data, in situ sensors)
  - PyLenM: Framework from various data to ML and decision making
  - Model simulations to inform monitoring and management
- **Simulation Intelligence: Simulations x ML/AI**
  - U-FNO for emulating simulation results to understand climate change impact on residual contamination
  - Bayesian hierarchical models with GP for physics-informed spatial interpolation (physics-informed monitoring)
- **General contaminations: Democratizing environmental science**
  - Citizen science for water/air quality, tackling environmental justice issues
  - AI/ML for environmental science

Thank You!

Contact
Haruko Wainwright
HMWainw@MIT.EDU

Acknowledgment
DOE Office of Environmental Management
DOE Office of Science