Making surrogates robust against model misspecification: A residual-aware combination of Gaussian processes and U-Net architectures

Waqas Ahmed, Ahsan Qasam Khan, Wolfgang Nowak
University of Stuttgart, Institute for Modelling Hydraulic and Environmental Systems, Stochastic Simulation and Safety Research for Hydrosystems, Germany
(waqas.ahmed@iws.uni-stuttgart.de)

Abstract

Motivation

In large-scale groundwater models high quality data, and full representation of model complexity is rare. There can be structural errors due to:

Sparse and noisy data	Model simplification	Resolution mismatch	Computation schemes

Challenge

Develop a surrogate model that is robust to data-quality gaps and model misspecification.

Goal

To develop a surrogate framework for making high-fidelity forecasts while being robust against low-fidelity inputs. This is relevant when estimating, e.g., groundwater maps, in study regions where only simplified (low-fidelity) data and inaccurate (misspecified) models are available.

Theoretical Background

Groundwater system (High-fidelity version):

$$z = f(x;\boldsymbol{\theta}) + \epsilon_r$$

Simplified model with structural error as an additive term:

$$z = \underbrace{g(\bar{x};\bar{\theta})}_{\text{Misspecified model}} + \underbrace{\boldsymbol{\delta}_r(\bar{x})}_{\text{Structural Error Term}} + \underbrace{\epsilon_r}_{\text{Classical (i.i.d) Error Term}}$$

Learning Error Terms

Defining the fidelity: difference between numerical model output with high-fidelity and low-fidelity inputs: $\boldsymbol{\delta}_r(\bar{x}) = f(x;\boldsymbol{\theta}) - g(\bar{x};\bar{\theta})$

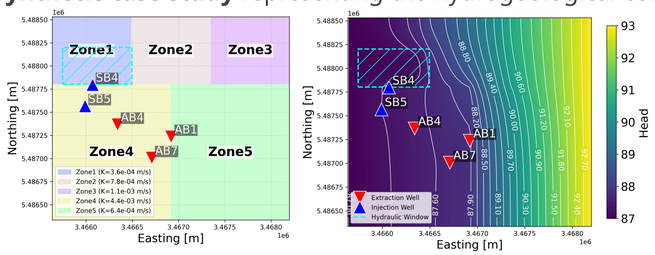
Training a U-Net to represent the structural error bias as a function of low-fidelity inputs $\bar{x}$: $\boldsymbol{\delta}_r(\bar{x}) \sim \mathrm{CNN}(\bar{x};\phi)$

Representing the remaining residuals and $\epsilon_r = z - g(\bar{x};\bar{\theta}) - \boldsymbol{\delta}_r(\bar{x})$ (the part not explainable by $\bar{x}$) as Gaussian Process over the spatial coordinates: $\epsilon_r(\cdot,\cdot) \sim GP\left(\mu(\cdot,\beta); \Sigma_{\sigma^2,l}(\cdot,\cdot)\right)$

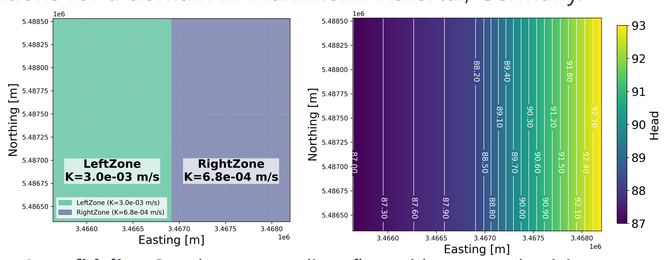
Universität Stuttgart

Synthetic Data

Synthetic case study representing the hydrogeological conditions for a domain in Mannheim-Käfertal, Germany.



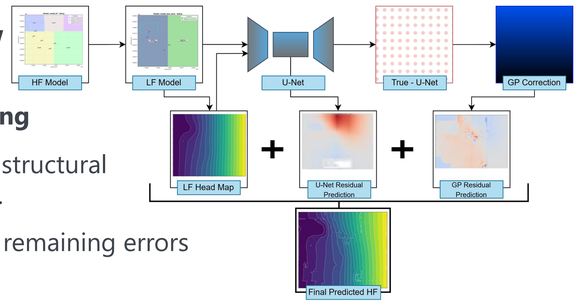
High-fidelity: Steady-state gradient flow with five conductivity zones, a hydraulic window, three injection wells, and two pumping wells.



Low-fidelity: Steady-state gradient flow with two conductivity zones, no hydraulic window, no injection and pumping wells.

Methods

Overall workflow

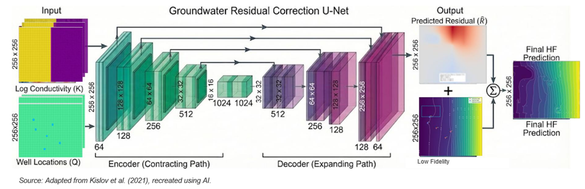


Two-Step Learning

Step 1: Learning structural errors with UNET.

Step 2: Learning remaining errors with GPR.

Architecture



Source: Adapted from Kislov et al. (2021), recreated using AI.

Fully convolution neural network with skip connections, known as **UNET.**

Results

High Fidelity

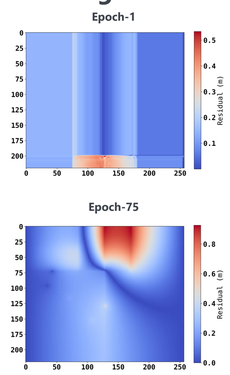
Scenario-I: Sub-surface simplified.

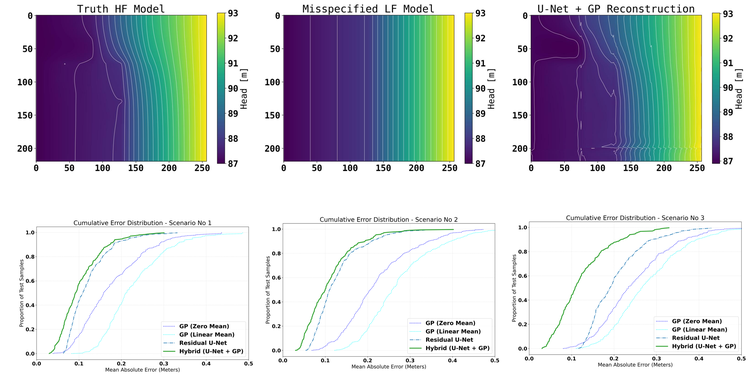
Scenario-II: Subsurface simplified, with no hydraulic window.

Scenario-III: Sub-surface Simplified, no pumping, no injection, and no hydraulic window.

Low Fidelity

Learning residuals





The combination of **UNET and GPR** performed better in all scenarios.

Conclusions

- With increasing LF, the U-Net loses the physical context and alone struggles to reconstruct the high fidelity truth.
- A two-step strategy where the U-Net identifies large-scale spatial patterns, while the GP refines the remaining systematic errors to reach a high-fidelity output is an effective approach.

Future Work

- **Loss functions:** Spatially correlated errors, Tikhonov Regularization.
- **Spatio-temporal:** Combine learning of spatio-temporal processes.
- **Large scale:** Test the approach for a large-scale hydrogeological domain.
- **Architecture:** Learn errors by diffusion models.

References

LS³ · ENWAT International Doctoral Program Environment Water · DAAD Deutscher Akademischer Austauschdienst German Academic Exchange Service · HEC · EGU Outstanding Student & PhD candidate Presentation contest