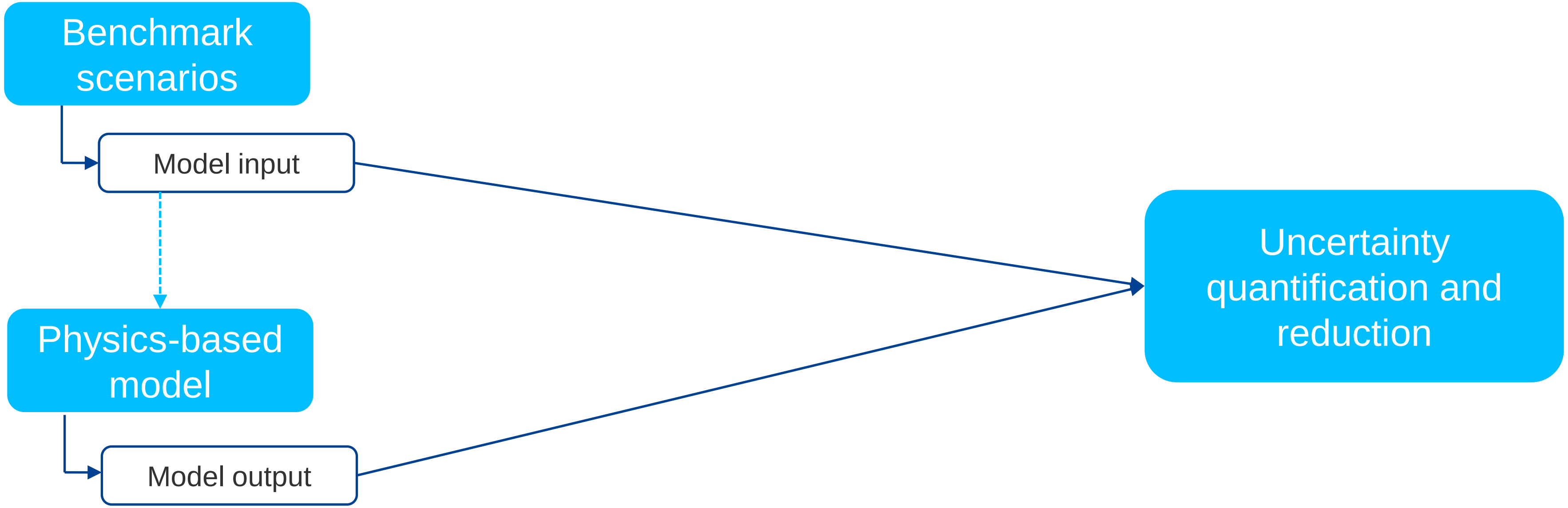
University of Stuttgart
Germany

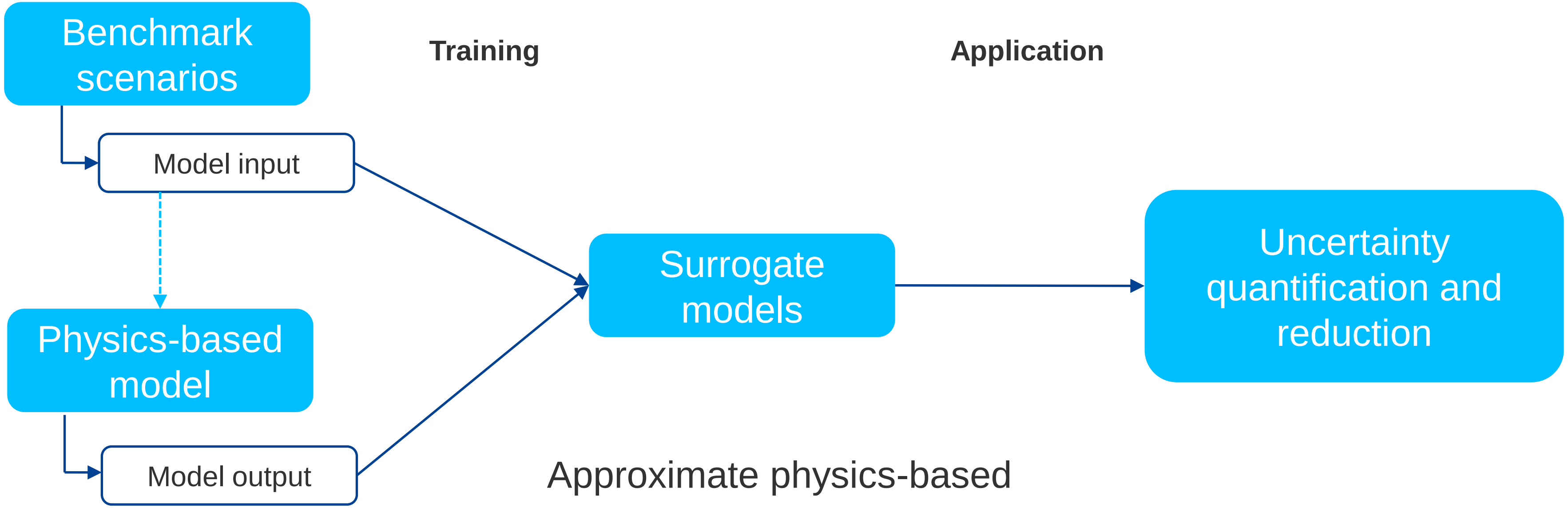
Surrogate model generation using Gaussian process regression and Bayesian active learning

Maria Fernanda Morales Oreamuno, M.Sc.

Review


Benchmark scenarios
Model input
Physics-based model
Model output
Uncertainty quantification and reduction

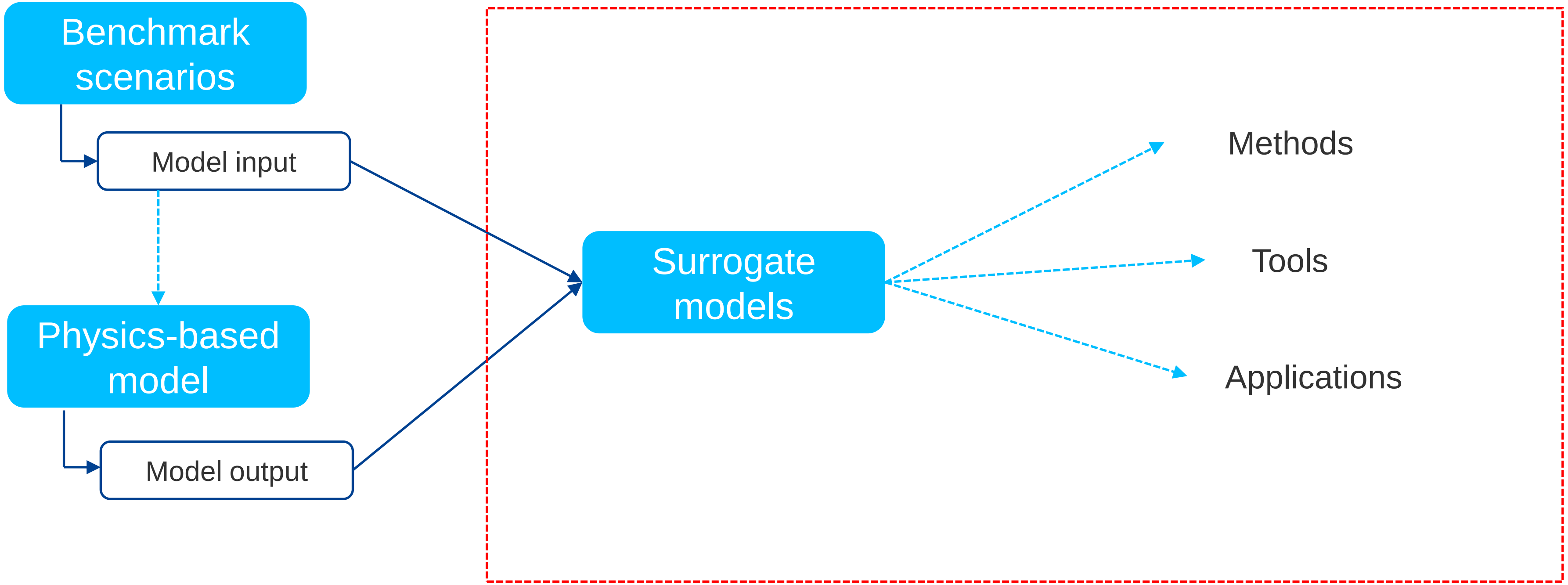
Review



Approximate physics-based model

Run in a fraction of the time

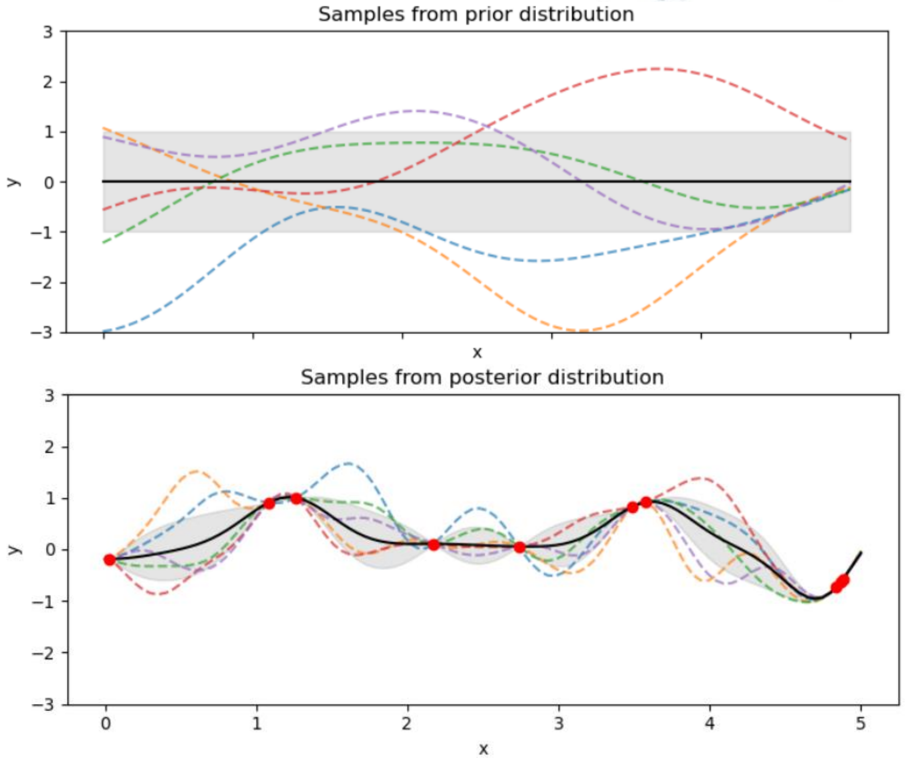
Review


Benchmark scenarios
Model input
Physics-based model
Model output
Surrogate models
Methods
Tools
Applications

Gaussian process regression

Methodology

- **Assumption**: a given function f(**x**) resembles a realization of a Gaussian stochastic process, described by a[1]
  - Mean (μ): assumed 0
  - Covariance (Σ): a given kernel
- Trained through input-output pairs, generated by the simulator



GPE training using a square-exponential covariance kernel

Source: https://scikit-learn.org/stable/modules/gaussian_process.html

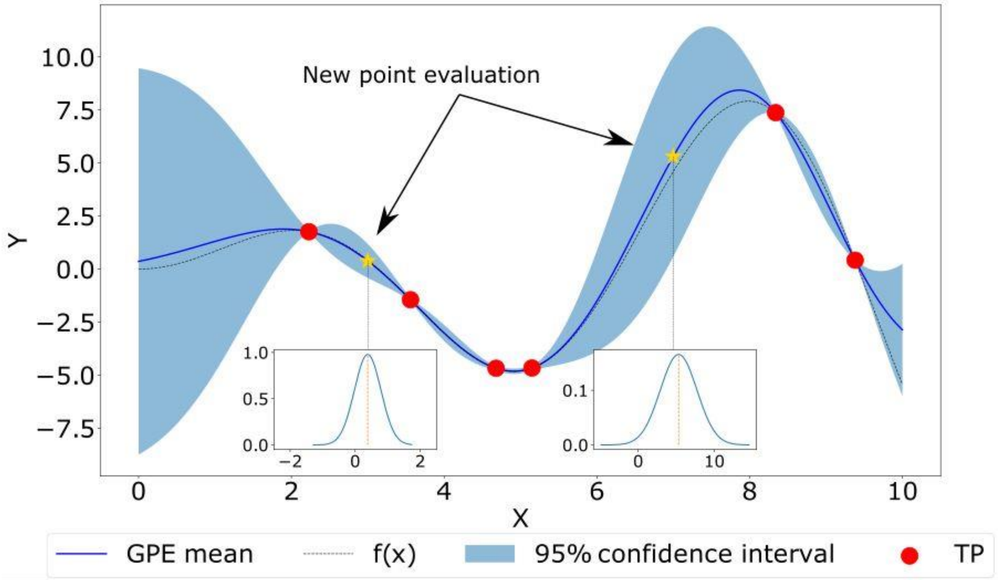
[1]Zhang, J., Li, W., Zeng, L., & Wu, L. (2016). An adaptive Gaussian process-based method for efficient Bayesian experimental design in groundwater contaminant source identification problems. *Water Resources Research*, *52*(8), 5971-5984.

Gaussian process regression

Methodology

- Trained through input-output pairs, generated by the simulator
- Predictions for all (future) parameter combinations are described by:
  - Mean
  - Variance

- Surrogate prediction error



1D input – 1D output example using Gaussian process regression

Williams, C. K., & Rasmussen, C. E. (2006). *Gaussian processes for machine learning* (Vol. 2, No. 3, p. 4). Cambridge, MA: MIT press.

Crevillen-Garcia, D., Wilkinson, R. D., Shah, A. A., & Power, H. (2017). Gaussian process modelling for uncertainty quantification in convectively-enhanced dissolution processes in porous media. *Advances in water resources*, *99*, 1-14.

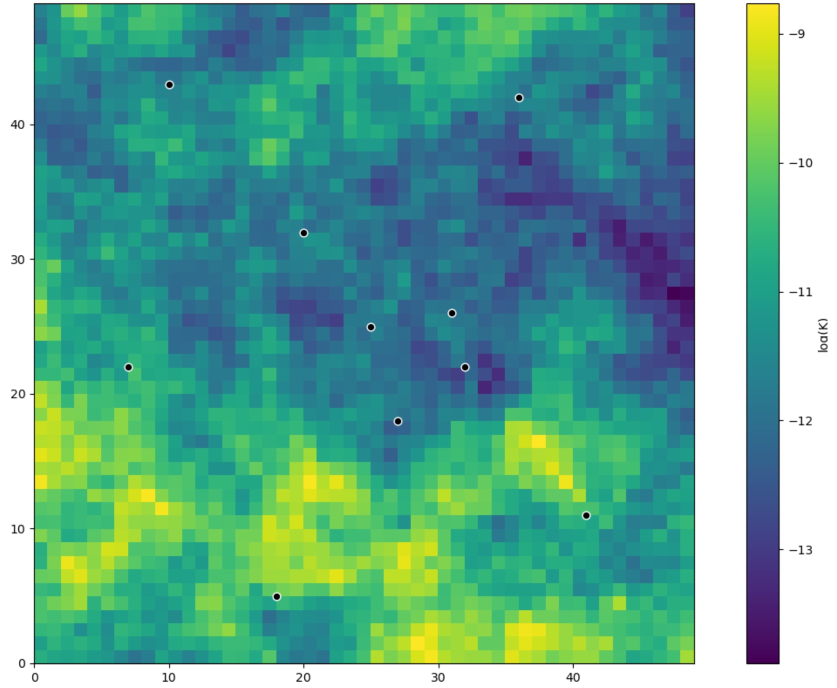
Gaussian process regression

Challenges

- High input dimensions:
  - Heterogeneity
  - Parameters for different processes

  Needs large number of training points
- High output dimension
  - Train GPE for each cell and/or each time step

  High computational time

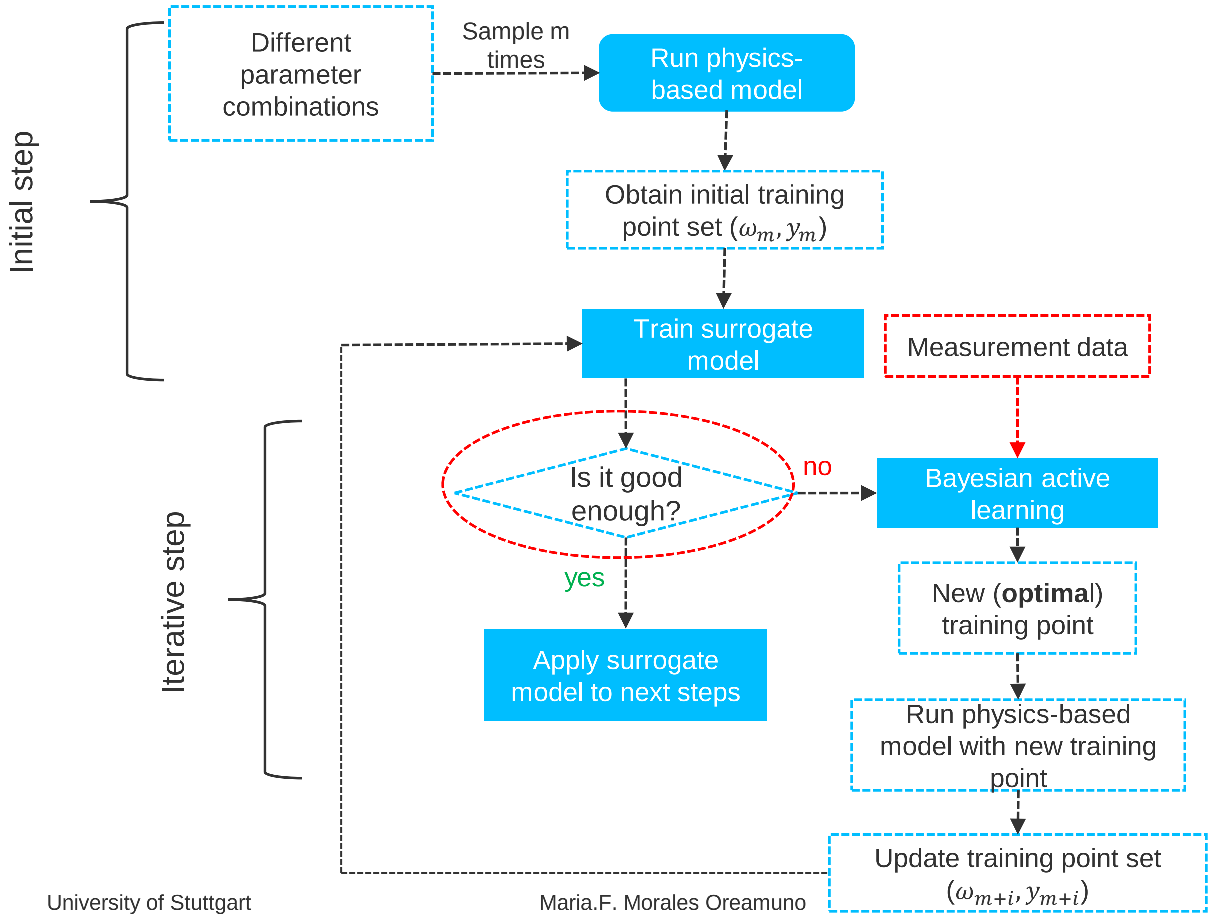


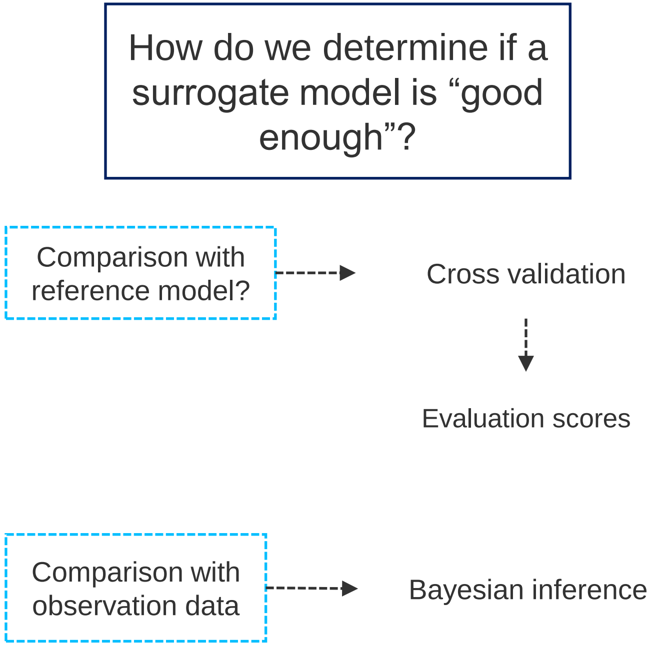
Configuration of 2D groundwater model, with a 50 m x 50 m grid. Allows for different input-output scenarios

Williams, C. K., & Rasmussen, C. E. (2006). *Gaussian processes for machine learning* (Vol. 2, No. 3, p. 4). Cambridge, MA: MIT press.

Gaussian process regression

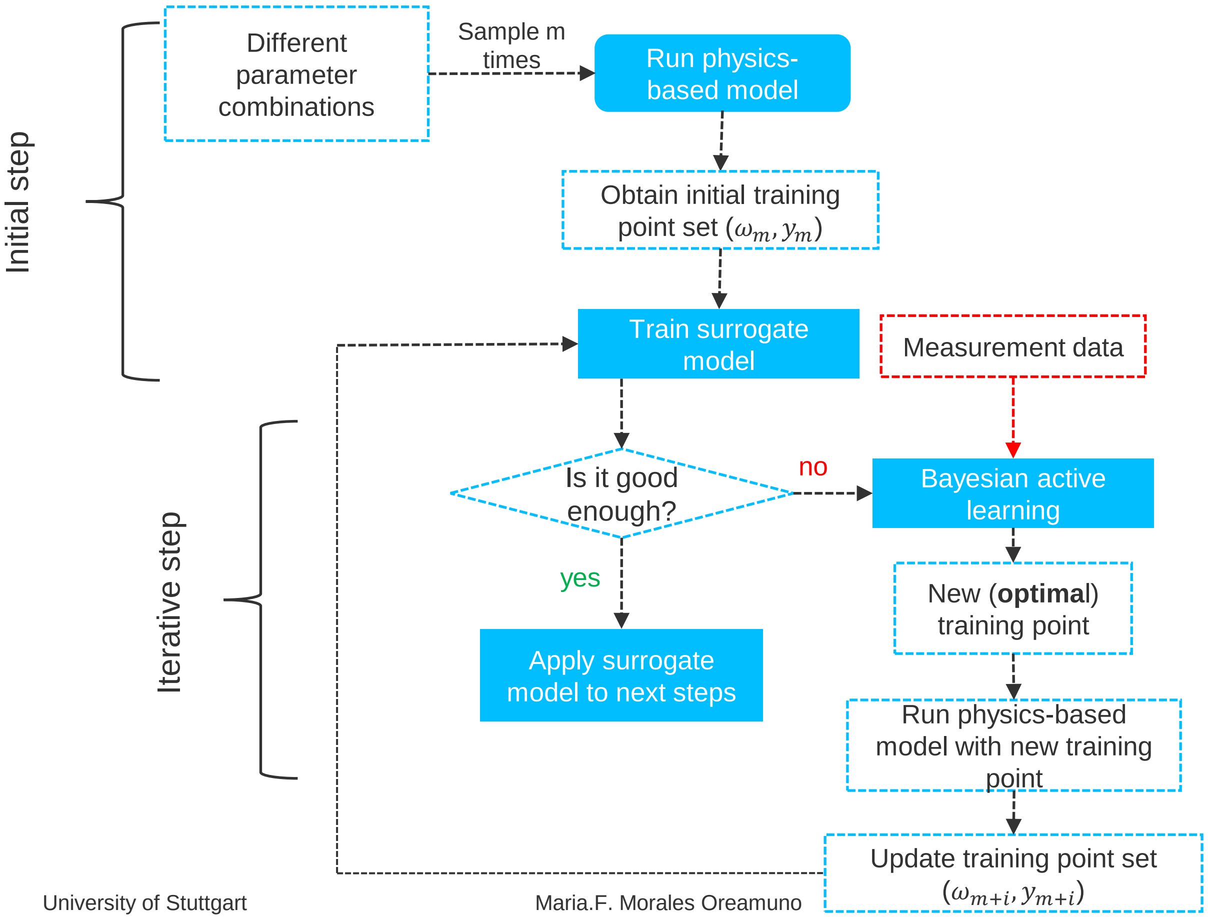
Work flow

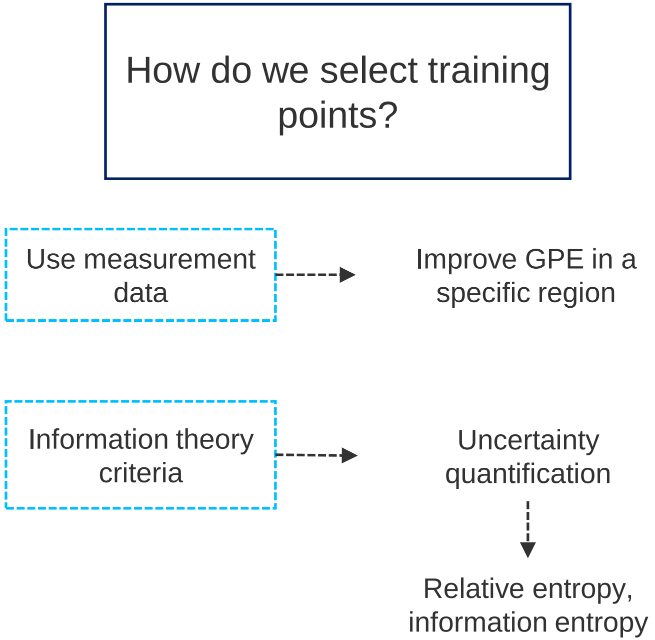




Gaussian process regression

Work flow



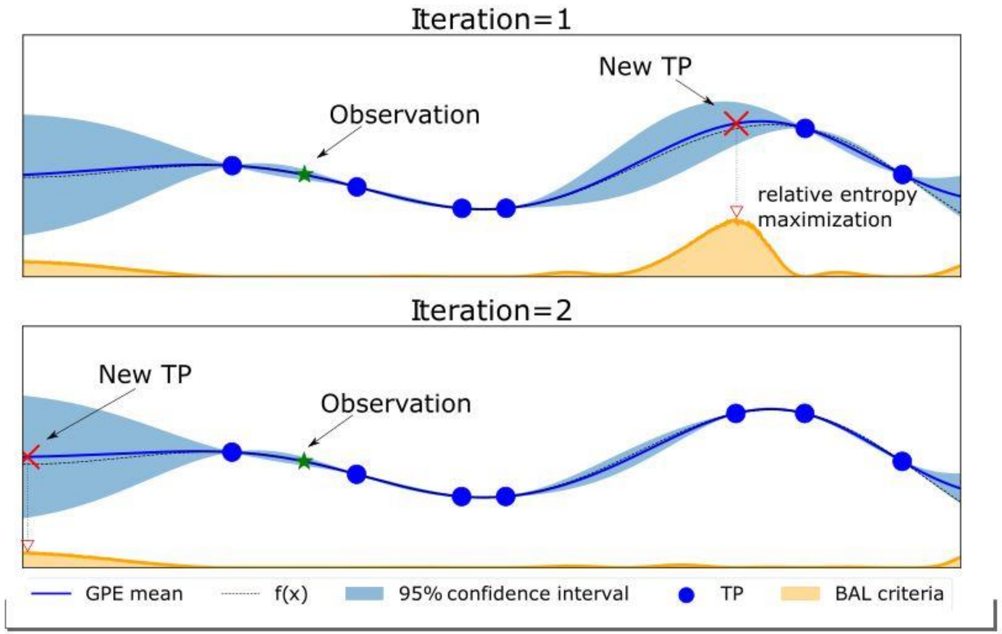


Bayesian Active Learning

- For each parameter set, one can get an output distribution
- We can sample from each output distribution
- **Bayesian inference** → obtain information from measurements

+

- **Information theory:** uncertainty associated with predictions



Information theory scores as training point selection criteria using Bayesian active learning

Oladyshkin, S., Mohammadi, F., Kroeker, I., & Nowak, W. (2020). Bayesian^3 active learning for the gaussian process emulator using information theory. *Entropy*, *22*(8), 890.

Zhao, H., & Kowalski, J. (2022). Bayesian active learning for parameter calibration of landslide run-out models. *Landslides*, 1-13.

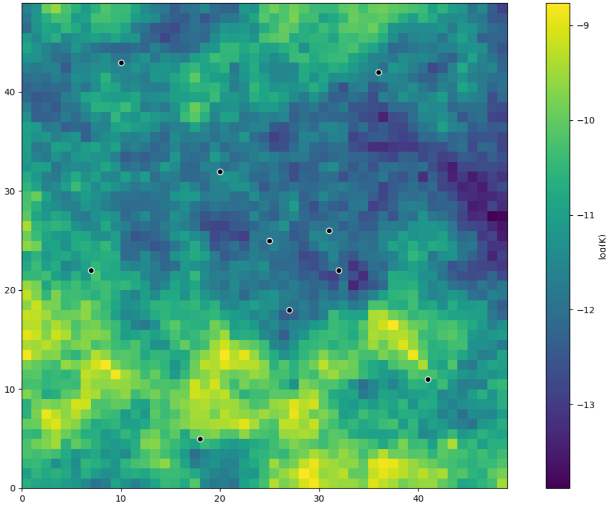
Tools

- Main coding language:
  - Python
- Gaussian process libraries:
  - Scikit Learn
  - GPyTorch
- Project collaboration:
  - GitHub
- Literature review
  - Zotero

Applications and Outlook

- **Develop** methodologies to train the Gaussian process emulator(s) using Bayesian active learning
  - Convergence criteria
  - BAL selection criteria (information theory)
- Consider/reduce potentially high input and output dimensions

Step 1: test methods using a simple, fast model, to be able to compare GPE results with a reference model



Configuration of 2D groundwater model, with a 50 m x 50 m grid. Allows for different input-output scenarios

Step 2:

- Apply methodologies to geophysical models
- Implement geophysical inversion methods, model calibration

University of Stuttgart
Germany

Thank you for your attention!

Maria Fernanda Morales Oreamuno

maria.morales@iws.uni-stuttgart.de

University of Stuttgart
Stochastic Simulation and Safety Research for Hydrosystems (LS³)
iws-ls3.uni-stuttgart.de

References

Crevillen-Garcia, D., Wilkinson, R. D., Shah, A. A., & Power, H. (2017). Gaussian process modelling for uncertainty quantification in convectively-enhanced dissolution processes in porous media. *Advances in water resources*, *99*, 1-14.

Gardner, J. R., Pleiss, G., Bindel, D., Weinberger, K. Q., and Wilson, A. G. (2018). Gpytorch: Blackbox matrix-matrix gaussian process inference with GPU acceleration. In Advances in Neural Information Processing Systems. https://gpytorch.ai/

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Van-derplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12:2825–2830. https://scikit-learn.org/stable/modules/gaussian_process.html

References

Oladyshkin, S., Mohammadi, F., Kroeker, I., & Nowak, W. (2020). Bayesian^3 active learning for the gaussian process emulator using information theory. *Entropy*, *22*(8), 890.

Williams, C. K., & Rasmussen, C. E. (2006). *Gaussian processes for machine learning* (Vol. 2, No. 3, p. 4). Cambridge, MA: MIT press.

Zhao, H., & Kowalski, J. (2022). Bayesian active learning for parameter calibration of landslide run-out models. *Landslides*, 1-13.

Overview

Surrogate modelling

- Main goal: overcome computational time constraints for expensive models

Also referred to as…

- Meta-modelling
- Response surface
- Model emulator

Different methods available…

- Neural networks
- Gaussian process regression
- aPCE

Gaussian process regression:

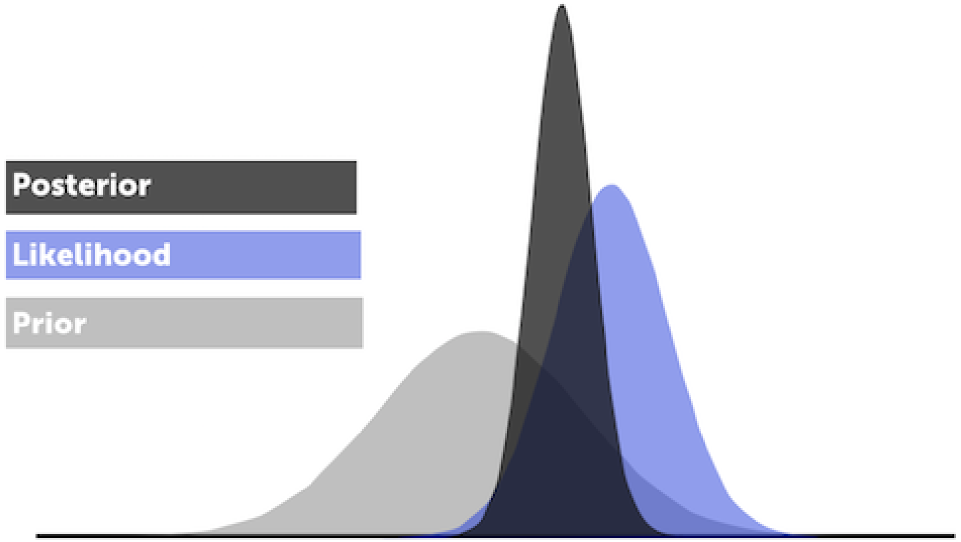
- Works for a smaller number of training points
- Provides uncertainty quantification

Data-driven models

Bayesian inference

- Bayes' theorem: update a prior state of knowledge to a posterior based on observation data

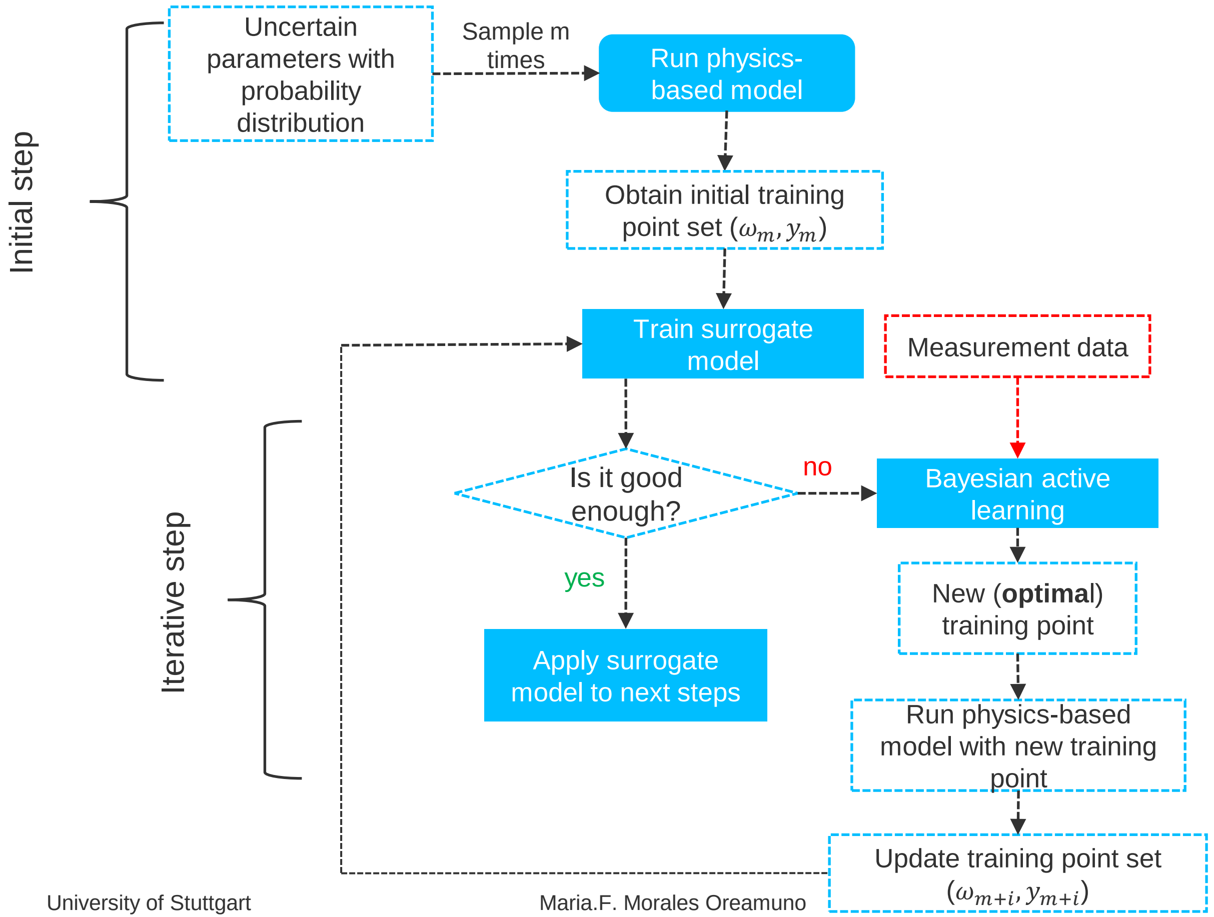
$$\underbrace{p(\omega|y_o)}_{\text{posterior}} = \frac{\overbrace{p(y_o|\omega)}^{\text{likelihood}}\overbrace{p(\omega)}^{\text{prior}}}{p(y_o)}$$



Bayesian inference: updating a prior to a posterior using observation data, through a likelihood function

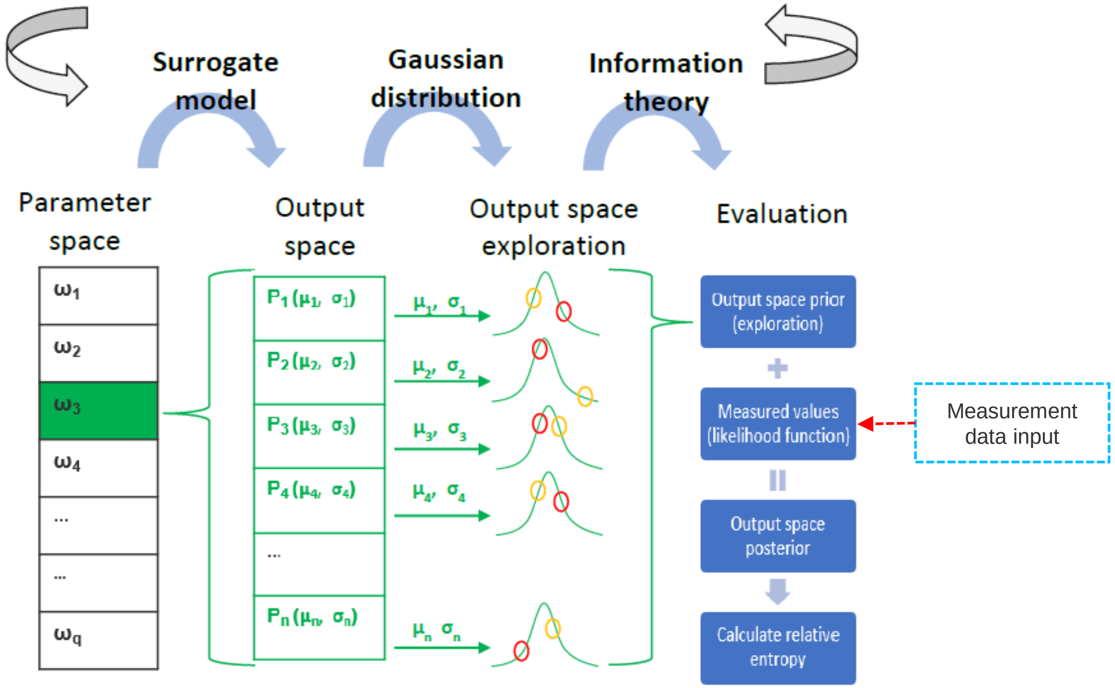
Source: Oladyshkin (2022), IWS lecture, University of Stuttgart

Gaussian process regression



Bayesian active learning

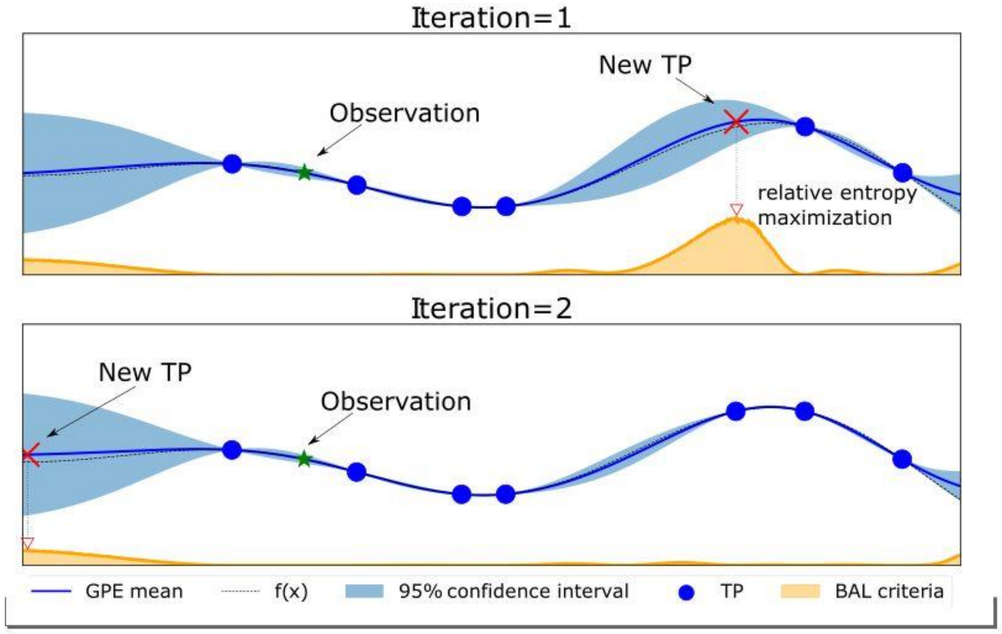
- (Bayesian) Active Learning allows to select training points located in regions of **high posterior likelihood**:
  - to improve the surrogate model prediction
  - reduce the number of total training points needed.
- For each iteration of the surrogate training, one selects the parameter set $\boldsymbol{\omega_i}$ which presents the highest gain in information as the next training point



Source: Acuna Espinoza (2021)

Bayesian Active Learning

- Main goals:
  - Reduce the number of training points
  - Uses observation data
  - Improve GPE in a specific region
- **Criteria**
  - Information theory scores
  - Chooses points with high uncertainty



Information theory scores as training point selection criteria using Bayesian active learning

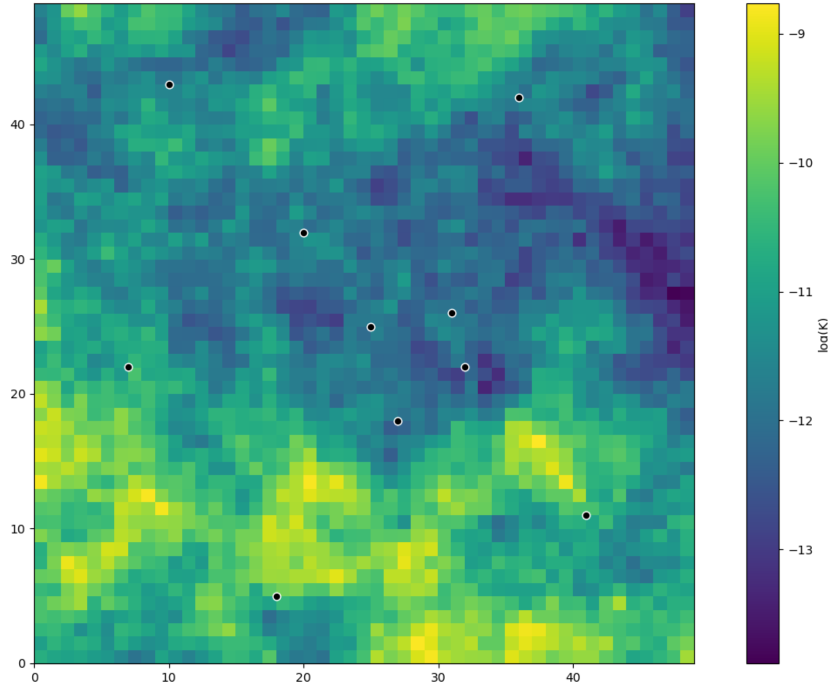
Oladyshkin, S., Mohammadi, F., Kroeker, I., & Nowak, W. (2020). Bayesian^3 active learning for the gaussian process emulator using information theory. *Entropy*, *22*(8), 890.

Zhao, H., & Kowalski, J. (2022). Bayesian active learning for parameter calibration of landslide run-out models. *Landslides*, 1-13.

Application

Simple test case

- 2D flow and transport groundwater model
  - Self-made finite-element solver in MATLAB
  - < 1 s run time
- Allows for different input parameters
  - Heterogeneity
  - Transport parameters
  - Boundary conditions
- High output dimension scenario
  - 1 GPE for each output cell
- **Future**: test with a 1D geophysical model for a more case-specific example



Configuration of 2D groundwater model, with a 50 m x 50 m grid. Allows for different input-output scenarios

Summary and outlook

- Develop methodologies to train Gaussian process emulators (GPE)
  - To reduce computational time and allow for uncertainty quantification and geophysical inversion
- Train GPEs using Bayesian active learning
  - Consider observations → give additional information about true processes
  - Reduce number of (expensive) runs of physics-based model
- Test/apply methodologies on simple (fast) test cases to compare GPE with reference model

Outlook

- Apply methodologies to (more expensive) geophysical models
- Use GPEs for Bayesian inversion and model calibration
- Apply GPEs for optimal experimental design and smart monitoring strategies