University of Stuttgart
Germany

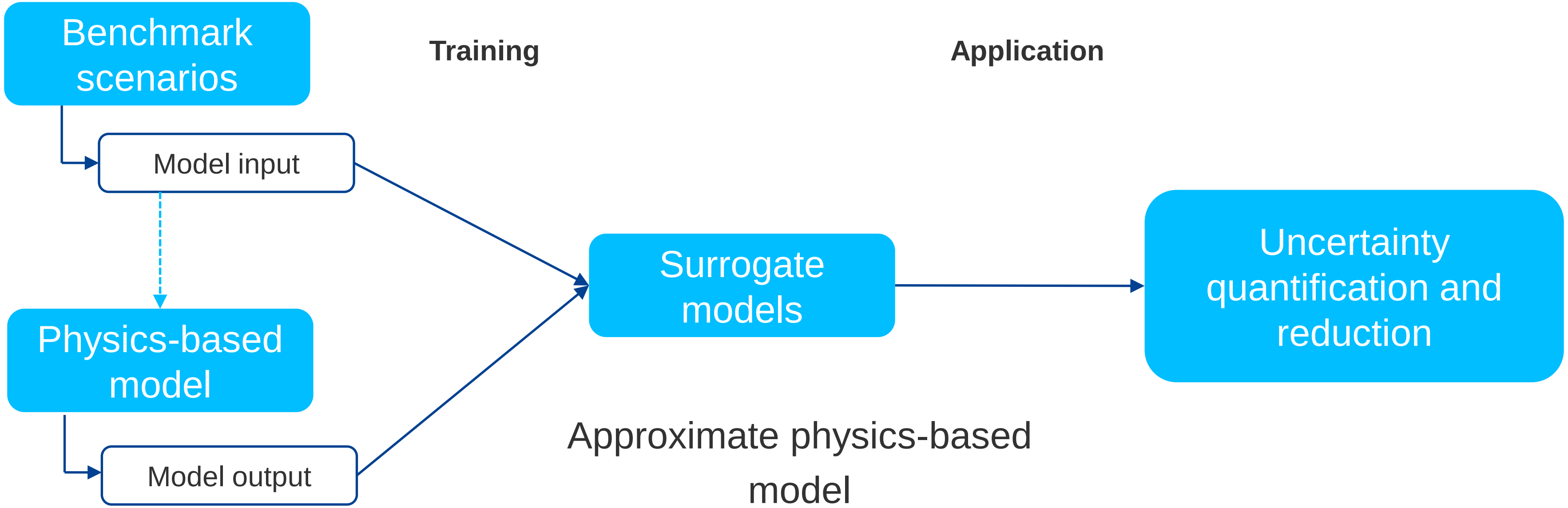
Input dimension reduction for surrogate model generation

Maria Fernanda Morales Oreamuno, M.Sc.

LS3

Recap

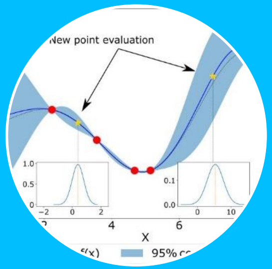
Surrogate modelling in the context of Smart Monitoring



Approximate physics-based model

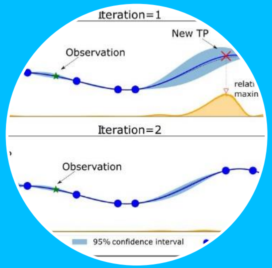
Run in a fraction of the time

Surrogate Modelling and Training Point Selection



Surrogate modelling

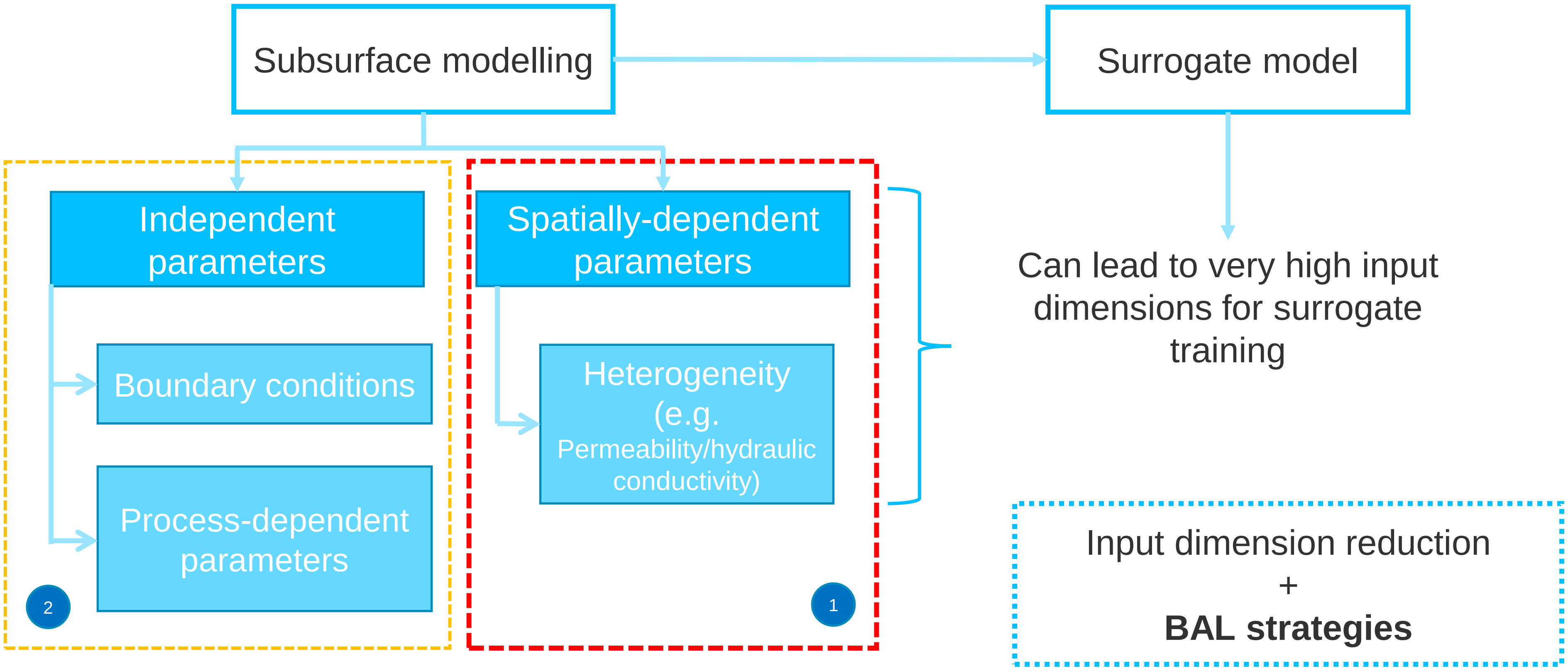
- Gaussian Process Emulator (GPE)
- (arbitrary) Polynomial Chaos Expansion (PCE)



Bayesian Active Learning

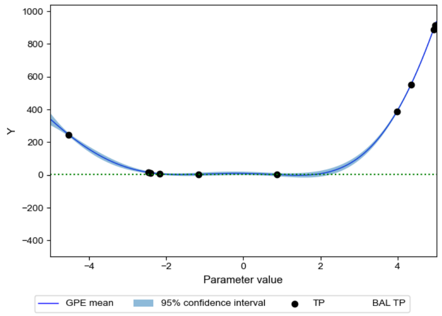
- Selection criteria
- Sampling criteria

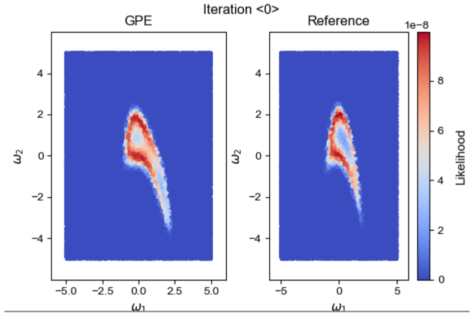
High input dimension problem



High input dimension problem

- Problems with high dimensions and surrogate models:
  - Visualization
  - Need more training points to cover parameter space
  - Computational power needed to train them increases
    - Gaussian Process Regression
    - Polynomial Chaos Expansion

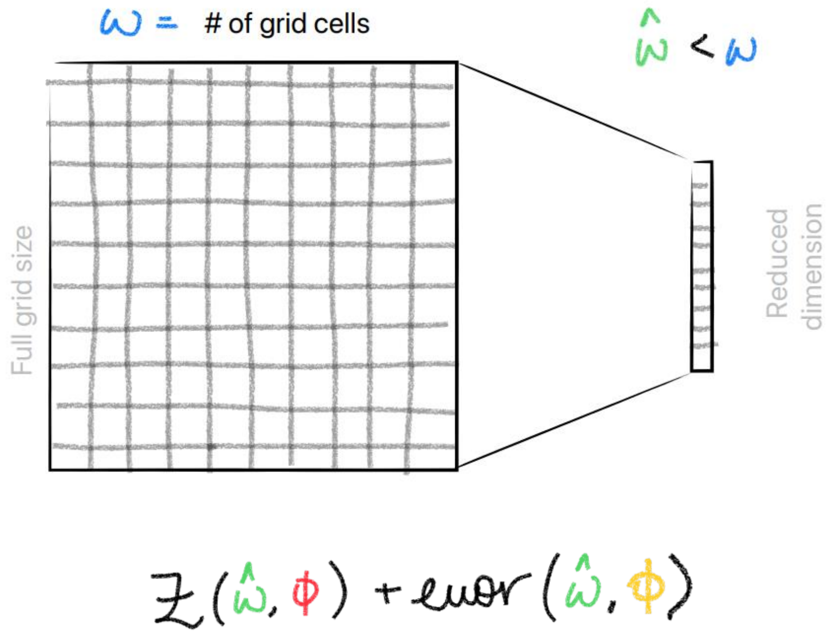




High input dimension problem

- Problems with high dimensions and surrogate models:
  - Visualization
  - Need more training points to cover parameter space
  - Computational power needed to train them increases
    - Gaussian Process Regression
    - Polynomial Chaos Expansion

- **Goal:** reduce the number of parameters sent to the surrogate (through truncations/transformations) while maintaining enough information on the QoI



Input dimension reduction case for spatially-dependent input parameters, through PCA approaches (e.g. for permeability / hydraulic conductivity fields)

Geostatistical Input Dimension Reduction

Spatially-dependent parameters

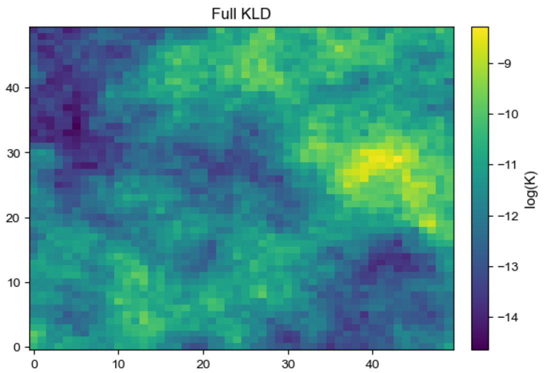
- Karhunen-Loeve decomposition: traditional approach
  - Global reduction
  - Only considers the input (no non-linear considerations)

What happens if the truncated value is still too large?

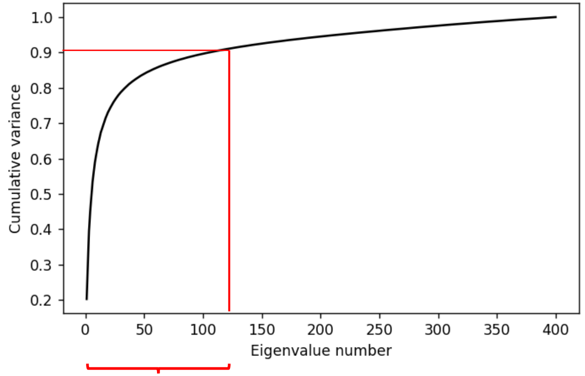
From covariance

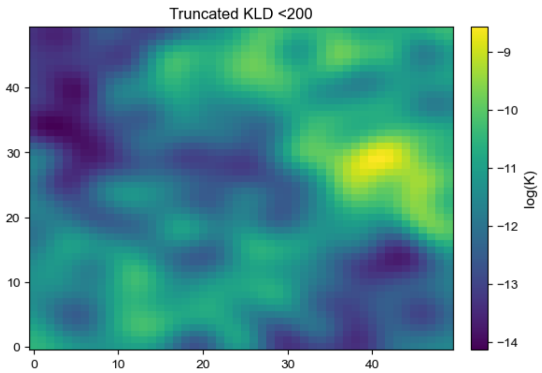
$N(0,1)$

$$Random\ field = Z^*(x) = \sum_{i=0}^{\hat{\omega}} \sqrt{\lambda_i} \cdot \vartheta_i(x) \cdot \xi_i$$



Random field will all $\xi_{i,}\ i = 2500$



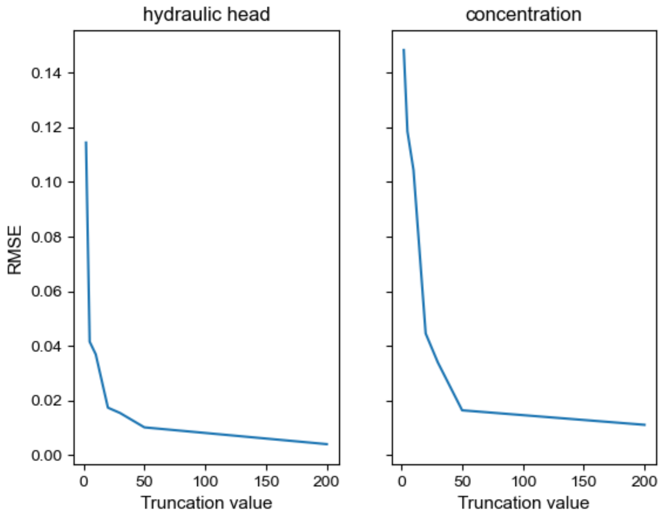


Random field will truncated $\xi_{i,}\ i = 200$

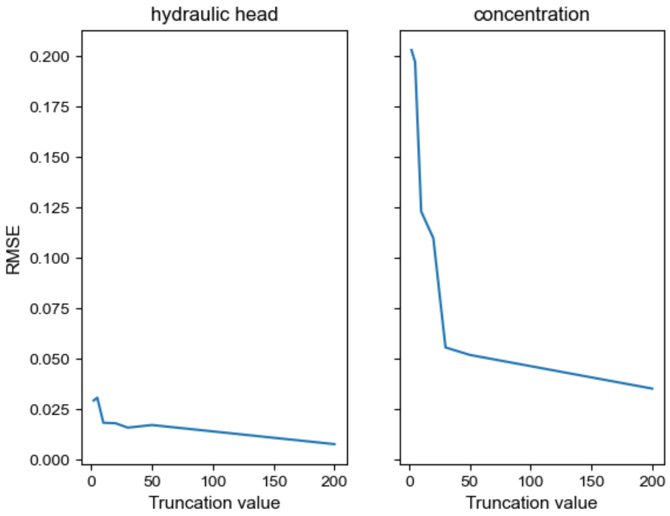
Geostatistical Input Dimension Reduction

Spatially-dependent parameters

- We did some tests with contaminant transport models (for relatively low heterogeneity)



50m x 50m grid (200 = 90% variance)



100 x 100 m grid (200 = 90% variance)

Question: how does the input dimension reduction affect the surrogate model performance?

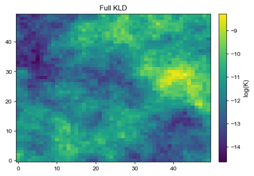
Geostatistical Input Dimension Reduction

Tests with Gaussian Process Regression

- Surrogates can only be trained using a reduced number of input parameters.
- The questions that remain are:
  - What outputs do we train with?
  - Is my surrogate providing sufficiently-accurate results (depending on my goals)
  - Is there an error associated with the input dimension reduction?

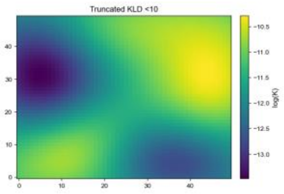
Scenario 1

Train surrogate with outputs obtained from full-KLD random field



Scenario 2

Train surrogate with outputs from smoothened-out field



Scenario 3

Train 2 surrogates:

- 1st with outputs from smoothened-out field
- 2nd with errors in relation with the truncation

Geostatistical Input Dimension Reduction

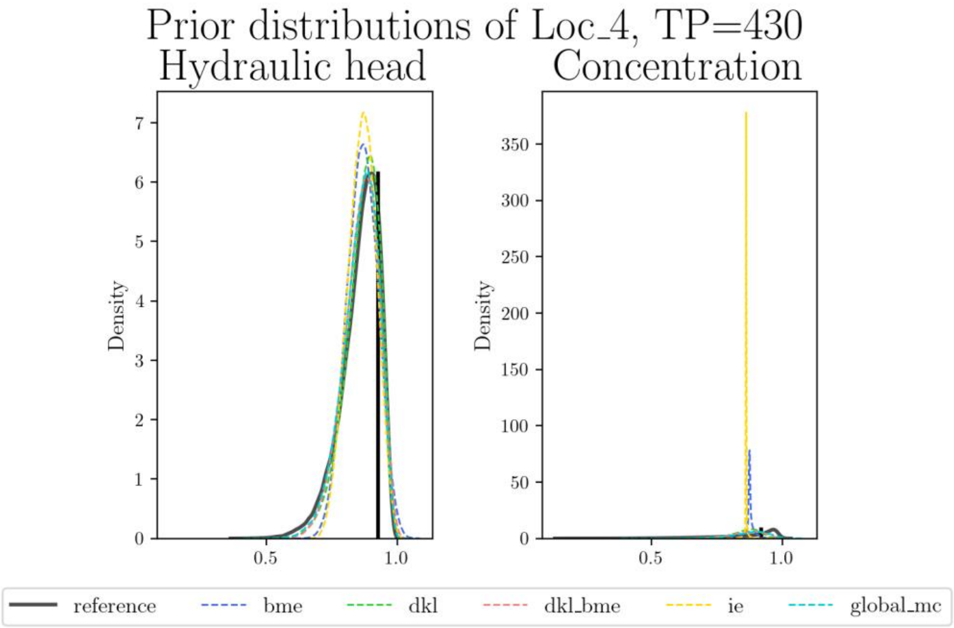
Tests with Gaussian Process Regression

We do tests for:

- Different truncation values (number of parameters)
- Different BAL strategies

We consider different evaluation criteria

- **Output distributions (PDF, CDFs)**
- Bayesian criteria
- Validation criteria (RMSE, normalized errors)
- Posterior distributions (if observations were available)



Prior-based output distributions,

Scenario 1, KLD coefficients = 10

Geostatistical Input Dimension Reduction

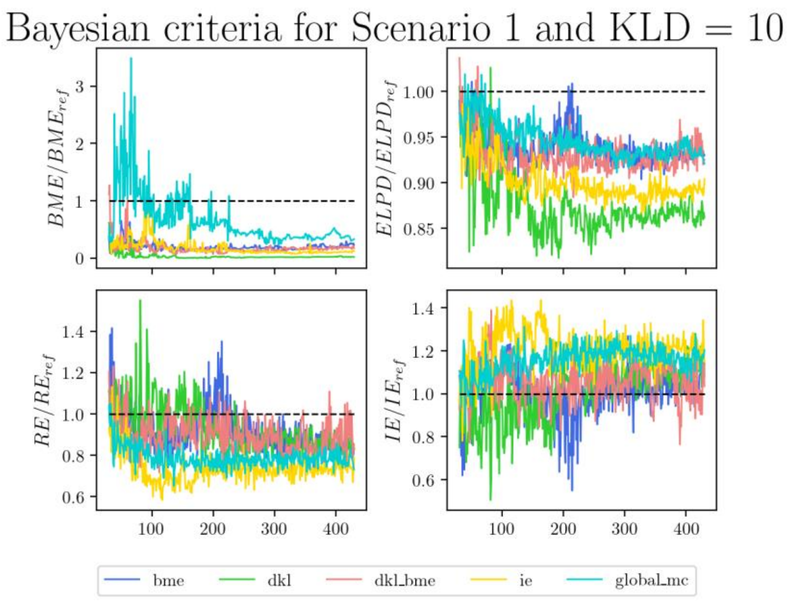
Tests with Gaussian Process Regression

We do tests for:

- Different truncation values (number of parameters)
- Different BAL strategies

We consider different evaluation criteria

- Output distributions (PDF, CDFs)
- **Bayesian criteria**
- Validation criteria (RMSE, normalized errors)
- Posterior distributions (if observations were available)



Bayesian criteria compared to a reference solution, Scenario 1, KLD coefficients = 10

Geostatistical Input Dimension Reduction

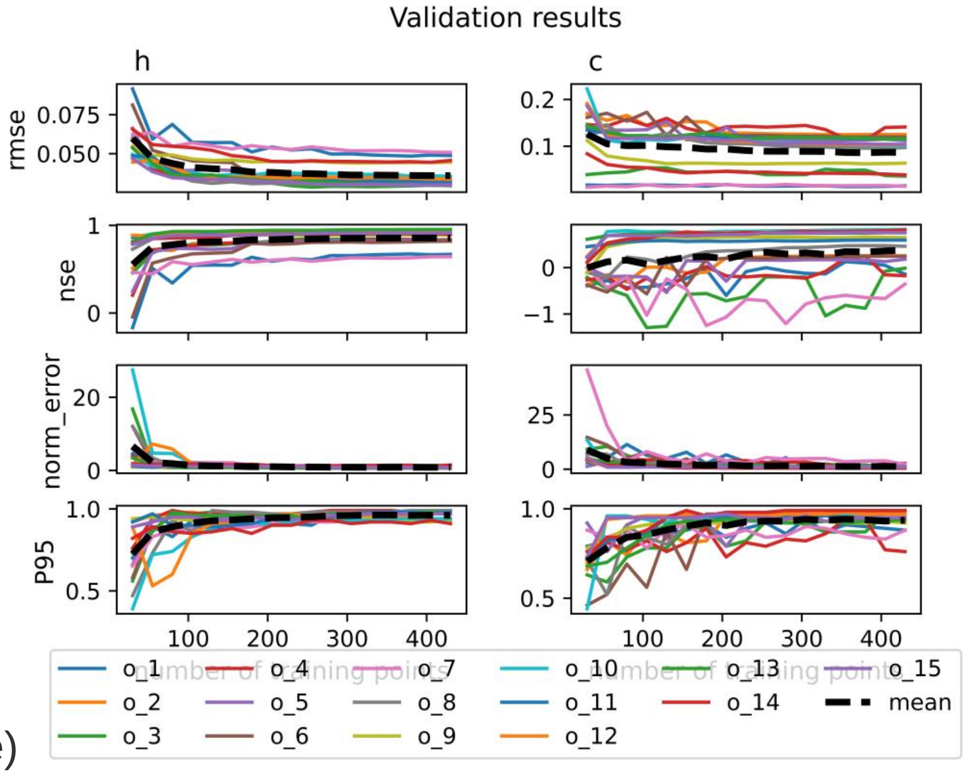
Tests with Gaussian Process Regression

We do tests for:

- Different truncation values (number of parameters)
- Different BAL strategies

We consider different evaluation criteria

- Output distributions (PDF, CDFs)
- Bayesian criteria
- **Validation criteria (RMSE, normalized errors)**
- Posterior distributions (if observations were available)



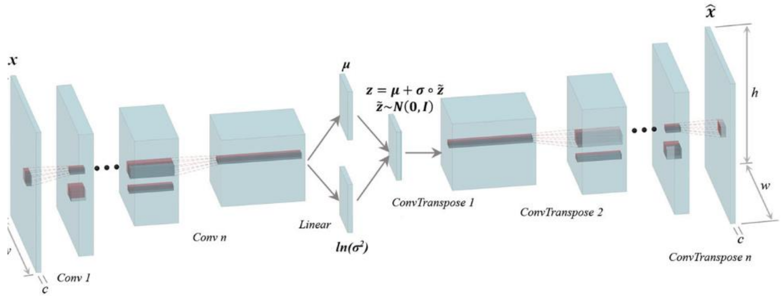
Validation criteria

Scenario 1, KLD coefficients = 10

Outlook

Input Dimension Reduction

- Test how aPCE works with input dimension reduction techniques
- Test other input-dimension reduction techniques for spatially-dependent parameters
  - Variational Auto-Encoders



- Test input dimension reductions that work for dependent **and** independent parameters
  - Methods that also contemplate non-linear relationships between inputs and outputs

Outlook

TransPyREnd

- Look into number of (independent) input parameters required by the model
  - Related to each radioactive nuclide being considered
  - Stratigraphy (homogeneous cases)
- Test surrogate model approaches to TransPyREnd (or 1D model from Qian)
  - PCE and GPR
- Implement input-dimension reduction techniques for domain-specific test cases (TransPyREnd)
- Reliability engineering-related surrogates

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 (LS^3)
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.

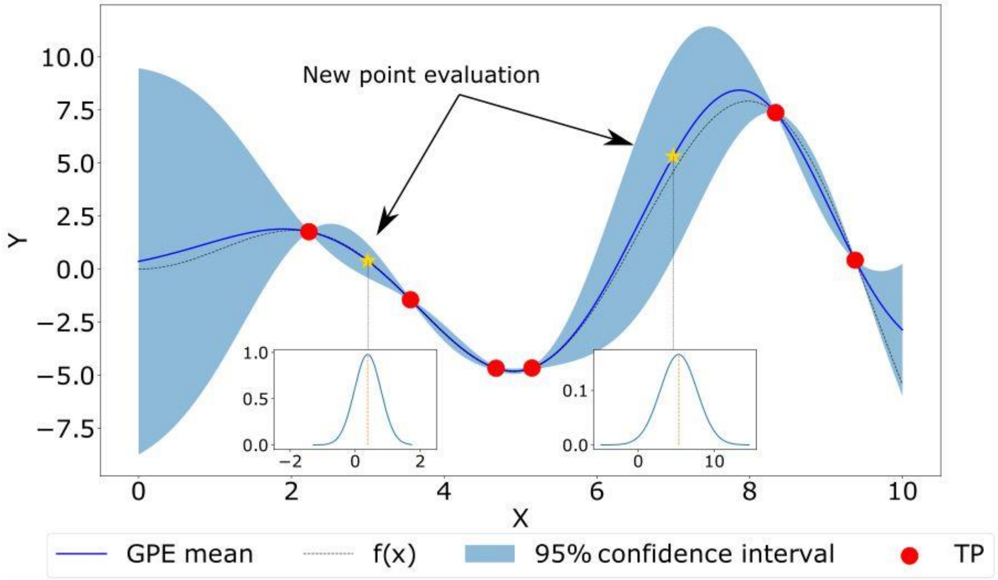
Surrogate modelling: Gaussian process regression

Recap

- Approximates the full-complexity model (simulator)
- Trained through input-output pairs (TP), generated by the simulator
- Predictions for any (future) parameter combinations are described by:
  - Mean
  - Variance

+ Reduces computation time

- Induces approximation error



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

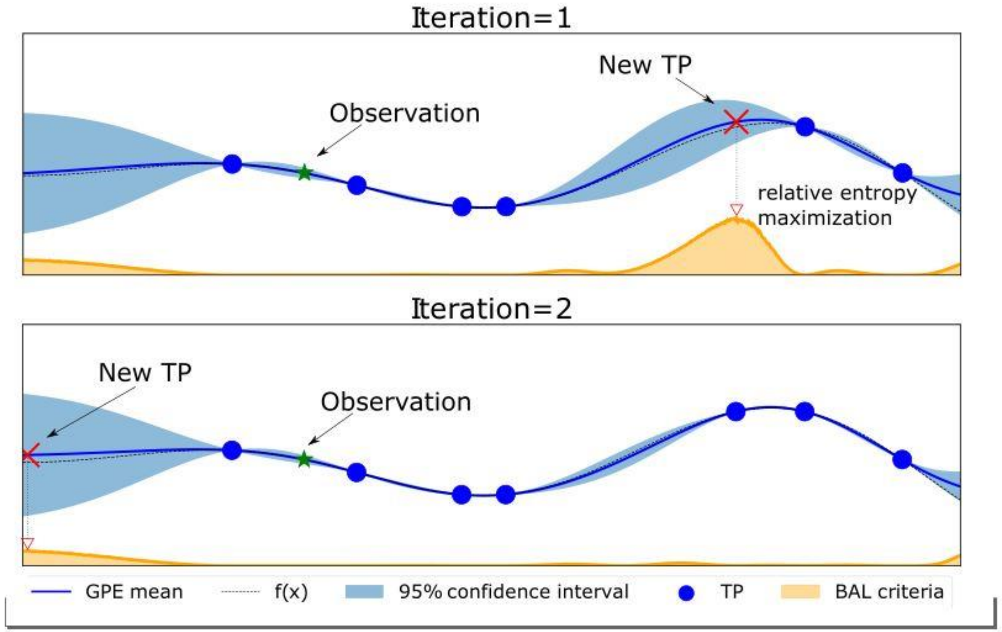
Williams, C. K., & Rasmussen, C. E. (2006). *Gaussian processes for machine learning* (Vol. 2, No. 3, p. 4). Cambridge, MA: MIT press.

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.

Surrogate modelling: Bayesian Active Learning

Recap

- Methodology to select training points using field observations and Bayesian criteria
- Goals:
  - **Improve** the surrogate in a region, where it is more likely that the true parameters are
  - **Reduce** the number of training points needed



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.

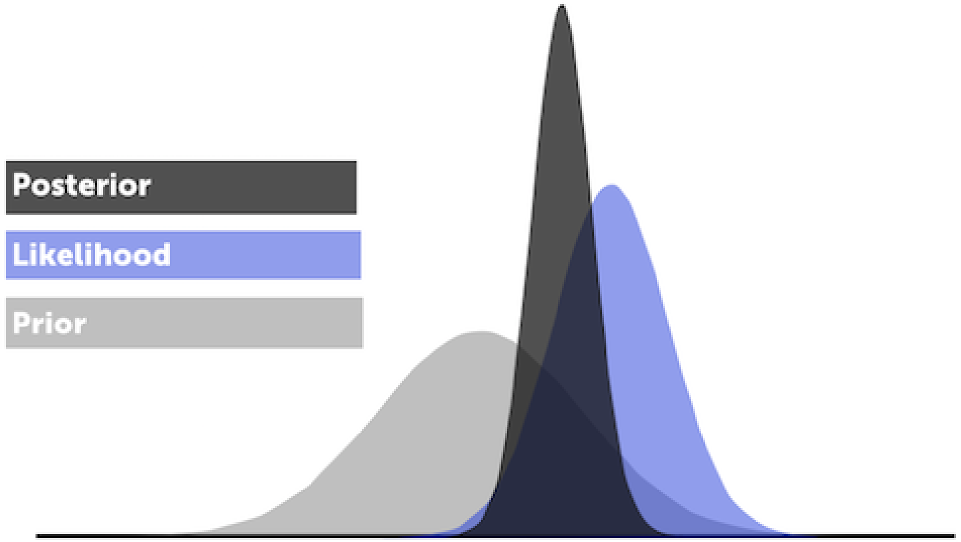
Zhao, H., & Kowalski, J. (2022). Bayesian active learning for parameter calibration of landslide run-out models. *Landslides*, 1-13.

Bayesian inference

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

posterior likelihood prior

$$p(\omega|y_o) = \frac{p(y_o|\omega)p(\omega)}{p(y_o)}$$

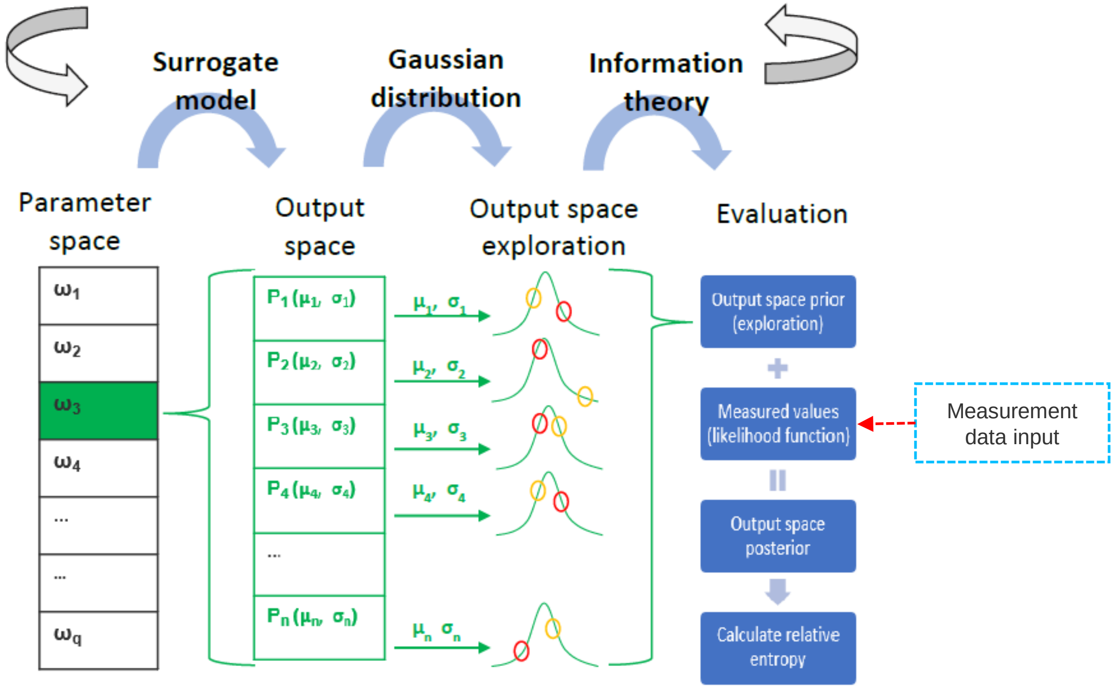


Bayesian inference: updating a prior to a posterior using observation data, through a likelihood function

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

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)