**University of Stuttgart**
Institute for Modelling Hydraulic and Environmental Systems
Department of Stochastic Simulation and Safety Research for Hydrosystems

YUAN Tian

# Training a Gaussian Process emulator using MCMC-based Bayesian Active Learning: application to a groundwater transport problem

Master's Thesis

# Training a Gaussian Process Emulator using MCMC-based Bayesian Active Learning: application to a groundwater transport problem


Submitted by

YUAN Tian

Matriculation Number: 3506792

Examiners: apl. Prof. Dr.-Ing. Sergey Oladyshkin

Supervisors: M.Sc. Maria Fernanda Morales Oreamuno

Institute for Modeling Hydraulic and Environmental Systems
Department of Stochastic Simulation and Safety Research for Hydrosystems


Stuttgart, 01 December 2023

# Author declaration

I declare that I have developed and written the enclosed thesis completely by myself and that I have not used sources or means without declaration in the text. Any thoughts from others or literal quotations are clearly marked.

The thesis was not used in the same or in a similar version to achieve an academic grading or is being published elsewhere.

The enclosed electronic version is identical to the printed versions.

| | |
|---|---|
| Date | Signature |

# Abstract

This study focuses on employing Gaussian Process Emulators (GPEs) as surrogate models to replicate complex systems with restricted training data while also aiming to enhance the efficiency of the GPEs' training process. aims to improve the efficiency of the GPEs' training process and test two stages to enhance the efficacy of building the GPEs surrogate model. For the construction processes, three adaptive Bayesian Active Learning (BAL) selection criteria are employed: Bayesian model evidence (BME), relative entropy (RE), and information entropy (IE). In the first stage, we use two distinct methods to generate the exploration parameter set: random sampling and Markov Chain Monte Carlo (MCMC). In the subsequent stage, we utilize two approaches to estimate the BAL selection criteria: prior-based and posterior-based methods. Specifically, for posterior-based estimation, we use Monte Carlo (MC) and rejection sampling along with MCMC, comparing these outcomes with the results obtained by prior-based estimation. The performance of these strategies is demonstrated within the context of a 2D groundwater transport model. Our study emphasizes the convergence capability and the capacity to quantify post-calibration uncertainty, showcasing the effectiveness of these BAL strategies in refining surrogate models with limited training data. Our findings indicate that the posterior-based estimation of BAL selection criteria exhibits faster convergence and reduced fluctuations, signifying its superiority over the prior-based estimation, especially concerning the posterior-based estimation of RE-based selection criteria using MCMC. Overall, this study underscores the superiority of posterior-based estimation methods over prior-based methods, highlighting the evident advantages of MCMC over other methods in both stages.

***Keywords***— Machine learning, Information Theory, Bayesian model selection, Bayesian inference