AQUIGROW - sustainable AQUIfer recharge to enhance resilience of GROundWater services under increased drought risk WP4 - Metamodeling as a tool to support decisions

Claire Lauvernet, Guerlain Lambert, Céline Helbert

AQUIGROW - sustainable AQUIfer recharge to enhance resilience of GROundWater services under increased drought risk

WP4 - Metamodeling as a tool to support decisions

Claire Lauvernet[1], Guerlain Lambert[1,2], Céline Helbert[2]

[1]RIVERLY, INRAE Lyon-Villeurbanne, FR
[2]Univ. Lyon, UMR CNRS 5208, Ecole Centrale de Lyon, FR

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

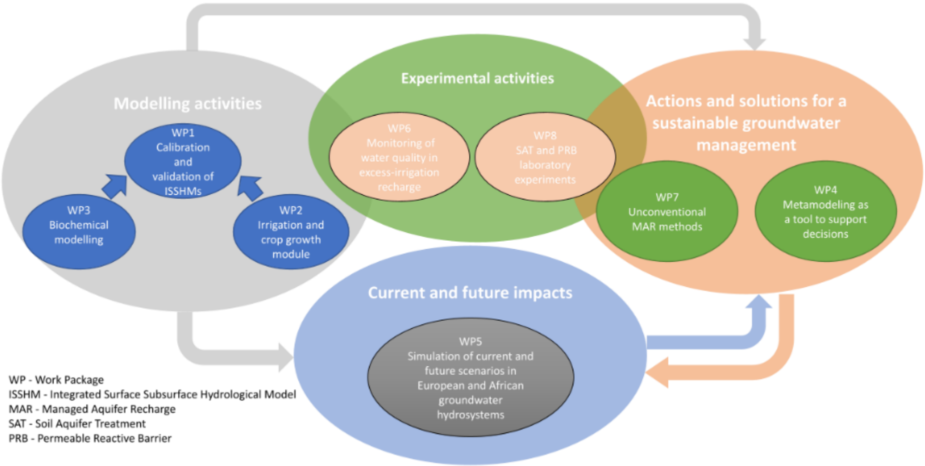
INRAE

The team

- **Claire Lauvernet**, Non-point source pollution team, unit RIVERLY, department AQUA
- **Céline Helbert**, Ecole Centrale de Lyon, teacher-researcher in statistics
- **Guerlain Lambert**, PhD 2023 - 2026
- a future post-doc (2025)
- a future internship (master 2)

WP4 Objectives

Metamodeling as a tool to support decisions



WP4 Objectives

Metamodeling as a tool to support decisions

- Physics-based models of water and pollutant transfer are computationally intensive and require a large number of parameters (soil characteristics, vegetation, chemical properties, etc.) and dynamic inputs (climate, agricultural practices, etc.).

$$S_w S_s \frac{\partial \psi}{\partial t} + \phi \frac{\partial S_w}{\partial t} = \nabla[K_s K_r(\nabla \psi + \eta_z)] + q_{ss}$$

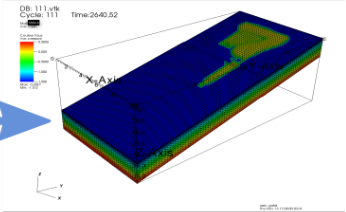
$$\frac{\partial Q}{\partial t} + c_k \frac{\partial Q}{\partial s} = D_h \frac{\partial^2 Q}{\partial s^2} + c_k q_s(h, \psi)$$

$$\frac{\partial \Theta_c}{\partial t} = \vec{\nabla}.(-\vec{U} c + D \vec{\nabla} c) + q_{tss}$$

$$\frac{\partial Q_m}{\partial t} + c_t \frac{\partial Q_m}{\partial s} = D_c \frac{\partial^2 Q_m}{\partial s^2} + c_t q_{ts}$$

CATHY-Pesticides

Camporese et al., 2010
Gatel et al., 2019



WP4 Objectives

Metamodeling as a tool to support decisions

- To be used by non-modelers as a tool to support decisions, they have:
  - to be simplified (to render them dependent on fewer inputs, ideally the most influential ones)
  - to address uncertainty with regards to the outputs.

⇒ Metamodeling (or surrogate modeling) offers a mean to achieve these two objectives

- still rarely used in coupled surface/subsurface hydrology and solute transport (processes are highly nonlinear and interrelated)
- a big methodolgical challenge in applied mathematics!
- attention paid to be generic and applicable to any ISSHM.

Definition : metamodel/surrogate model/emulator

- a statistical model is built between inputs and outputs
- a "model of the model" (2nd level of abstraction from reality)

→ MM is dependant of much less input parameters than the model

→ MM is very fast to run on new scenario

Definition : metamodel/surrogate model/emulator

- a statistical model is built between inputs and outputs
- a “model of the model” (2nd level of abstraction from reality)

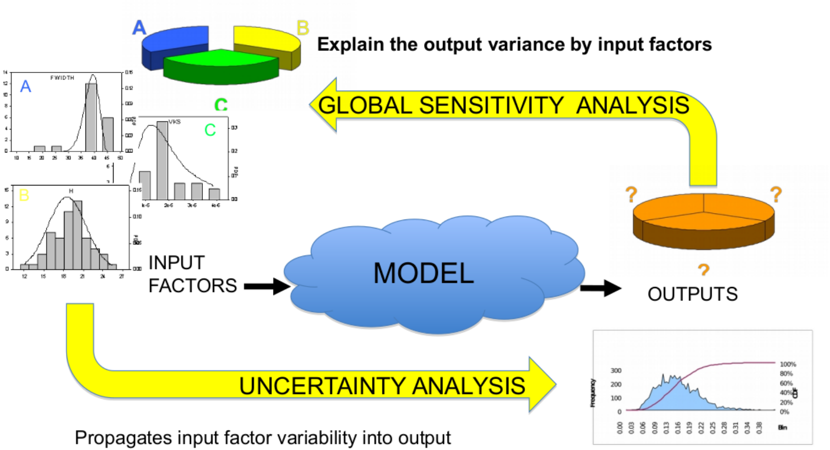
→ MM is dependant of much less input parameters than the model

→ MM is very fast to run on new scenario

Building a metamodel of your model is necessary when

- simulations are time-consuming
- the model is too difficult to be used (too many input parameters)
- you want to analyse the uncertainty and sensitivity of your model (to sort/fix inputs)

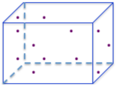
Uncertainty and sensitivity analysis of a model (Sobol indices)



Classical building of a metamodel: in practice

$\mathcal{G}$ the code function, $\mathcal{M}$ the metamodel, $K \ll N$

$$Y = \mathcal{G}(X_1, \ldots, X_N) \approx \mathcal{M}(X_i, \ldots, X_K)$$

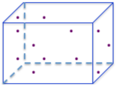


1. DoE = sampling in the input space

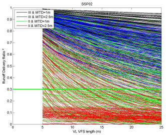
Classical building of a metamodel: in practice

$\mathcal{G}$ the code function, $\mathcal{M}$ the metamodel, $K \ll N$

$$Y = \mathcal{G}(X_1, \ldots, X_N) \approx \mathcal{M}(X_i, \ldots, X_K)$$



1. DoE = sampling in the input space



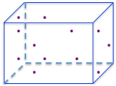
2. Evaluate the model on the sampling

Classical building of a metamodel: in practice

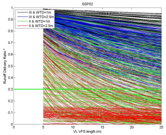
$\mathcal{G}$ the code function, $\mathcal{M}$ the metamodel, $K \ll N$

$$Y = \mathcal{G}(X_1, \ldots, X_N) \approx \mathcal{M}(X_i, \ldots, X_K)$$

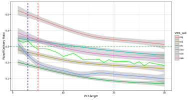
1. DoE = sampling in the input space



2. Evaluate the model on the sampling



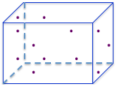
3. Approximate the model by $\mathcal{M}$ on this DOE:
   $Y = \mathcal{M}(X_{i1}, X_{i2}, \ldots, X_{iJ}) + \eta$



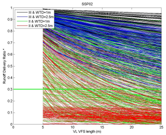
Classical building of a metamodel: in practice

$\mathcal{G}$ the code function, $\mathcal{M}$ the metamodel, $K \ll N$

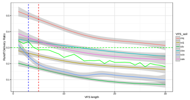
$$Y = \mathcal{G}(X_1, \dots, X_N) \approx \mathcal{M}(X_i, \dots, X_K)$$



1. DoE = sampling in the input space

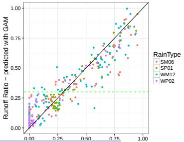


2. Evaluate the model on the sampling



3. Approximate the model by $\mathcal{M}$ on this DOE:
   $Y = \mathcal{M}(X_{i1}, X_{i2}, \dots, X_{iJ}) + \eta$

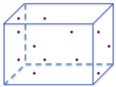
4. Validate $\mathcal{M}$ on a independent DoE and calculate sensitivity indices

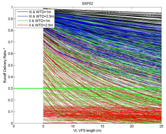


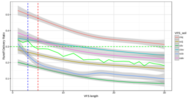
Classical building of a metamodel: in practice

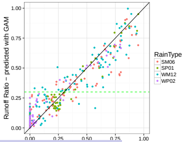
$\mathcal{G}$ the code function, $\mathcal{M}$ the metamodel, $K \ll N$

$$Y = \mathcal{G}(X_1, \ldots, X_N) \approx \mathcal{M}(X_i, \ldots, X_K)$$







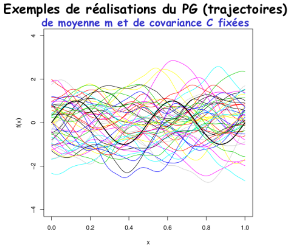


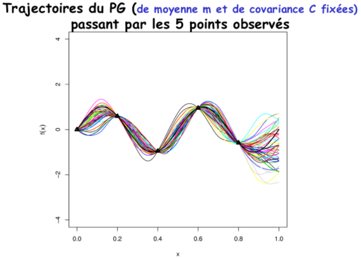
Several methods:

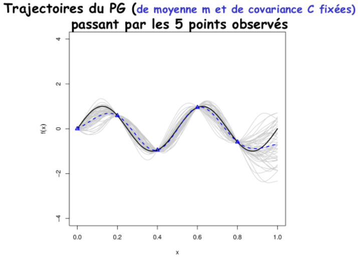
- Parametric regression (linear, polynomial)
- Nonparametric regression : **GAM**, Spline,...
- Gaussian Process regression : **Kriging**
- Learning methods : NN, SVM, Random Forests,...

Metamodels/Surrogate : Gaussian Process regression (kriging)

- The deterministic output of the model f is the realization of a GP $(Y(\mathbf{x}))_{\mathbf{x}} \sim GP(m, k(.,.))$
- The relation between points is expressed by a covariance structure between the obs
- Prediction is the mean of the conditionned GP realisations : $\hat{f}(x) = \mathbb{E}(Y(\mathbf{x})|Y(\mathbb{X}) = \mathbf{y}\}$
- Interpolation, non parametric approach, all is in the prior







Metamodeling vs model : an example with BUVARD

VFS optimal length

=

Modeling toolkit $\mathcal{G}$

- Rainfall hyetograph
- Runoff dynamic hydrograph
- Season
- Curve Number
- Slope, Area
- VFS Water table depth
- Soil type (K_{sat}, θ_s, VG par., ...)
- Sediments characteristics
- Roughness, grass height
- ...

>70 parameters

$\approx$

Metamodel $\mathcal{M}$

- Curve Number
- Slope
- Contributive area
- VFS Water table depth
- VFS Soil type
- Rainfall typical event

6 parameters

WP tasks

- Task 4.1 Setup of CATHY with reactive solute on one study site (months 1-6)
- Task 4.2 Development of robust methods for the surrogate of ISSHMs (months 7-30)
- Task 4.3 Combining data and CATHY metamodel (months 18-30)
- Task 4.4 Metamodeling in coupled groundwater-vadose zone models (months 18-30)

Task 4.1 Setup of CATHY with reactive solute on the Italian site (M1-M6)

- version of CATHY coupled with reactive solutes (Weill et al. 2011, Gatel et al., 2019) implemented on the Morcille catchment
- calibration of and evaluation on one of the two Italian AQUIGROW sites
- taking into account the *correlation or dependance between some of the inputs* (for example, VG/hydrodynamical properties)
- first case test = an 1D with analytical solution
- a small sample to explore the links between inputs and outputs in CATHY

$\rightsquigarrow$ collab. with Claudio Paniconi (INRS-ETE) for the coupling water-solutes + calibration, sensitivity and uncertainty analysis of CATHY on the Italian site

Task 4.2 Development of robust methods for the surrogate of ISSHMs (M7-M30)

- fitting of several metamodels/surrogates of CATHY : test of GP, Random Forest, Polynomial Chaos Expansion and Deep GP
- comparison of the methods → better understanding of their behavior with spatial and dynamic processes, and how each of them deal with these surface/subsurface and water/pesticides interactions along with functional outputs
- first focus on GP
- methodological challenge with a dynamic distributed model such as CATHY !
- importance also given to computational cost and carbon footprint

⇝ methodological development with Guerlain

On-going work from Guerlain Lambert, PhD

Sequential approach to metamodeling and sensitivity analysis of expensive computational codes (in the presence of correlated and spatio-temporal random variables)

- DoE with dependent inputs[1]
- Sensitivity analysis / Sobol indices => need for a rich sampling => time-consuming
- metamodeling the model with Gaussian Processes => costly to build too!
- make a first GP with a low-size initial DoE (sampling)
- augment the DoE with points that are deemed most impactful according to an acquisition function.
- goal = to improve the efficiency and accuracy of Sobol indices estimation by optimising the DoE used in fitting GP metamodels.

[1] Lambert, G., C. Helbert, and C. Lauvernet. Quantization-Based Latin Hypercube Sampling for Dependent Inputs With an Application to Sensitivity Analysis of Environmental Models. doi:10.1002/asmb.2899.

On-going work from Guerlain Lambert, PhD

- Started the sequential approach for enriching the sampling to compute the metamodel and get finer sensitivity indices **on a toy model**
- the approach relies on the use of Derivative-based Global Sensitivity Measures (DGSM[1]) or its variants to effectively handle complex inputs

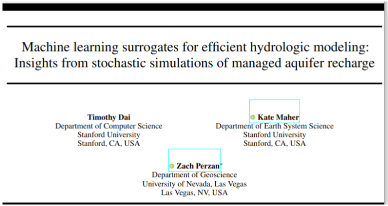
Next steps

- include groups of dependent random vectors and functional inputs for spatio-temporal models
- Application to water and pesticide transfer models in an agricultural context : from BUVARD (VFS scale) **to ISSHM CATHY, coupled with reactive solutes (AQUIGROW catchment)**

Task 4.4 Metamodeling in coupled groundwater-vadose zone models on the Vistula Spit site (M18-M30)

- test on MODFLOW-2005 (saturated zone flow) + HYDRUS-1D (unsaturated zone flow) models from Task 1.2.
- to obtain optimal and less costly metamodels predicting time variable groundwater recharge
- a large number of simulations performed with the vadose zone model HYDRUS-1D (varying weather data, soil types, plant cover and depth to water table)
- not too risky since the method will be well tested before on CATHY
- collab. with Adam Szymkiewicz/Anna Gumuła-Kawęcka: may be done earlier than expected! probable visit of Anna in Lyon in Feb 2025

A very recent paper on the topic (subm. to JoH

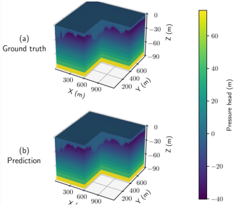


Machine learning surrogates for efficient hydrologic modeling: Insights from stochastic simulations of managed aquifer recharge

Timothy Dai
Department of Computer Science
Stanford University
Stanford, CA, USA

Kate Maher
Department of Earth System Science
Stanford University
Stanford, CA, USA

Zach Perzan*
Department of Geoscience
University of Nevada, Las Vegas
Las Vegas, NV, USA



- A case study on a propsective MAR, 800mx400m with an unconfined aquifer 150 m depth
- simulations of variably saturated groundwater flow shows that the ML surrogate models can achieve under 10% mean absolute percentage error
- with ParFlow (Richards)-CLM (water and energy fluxes at the land surface), surrogate built with 7 surrogate architectures (CNN4d, ViT4d, U-FNO4d, CNN3d, ViT3d, U-FNO3d, PredRNN++)
- high limitations with spatial and temporal variations

⇝ in AQUIGROW: focus on dependance of spatial and temporal inputs (and outputs?)

thanks!

To define

- version of CATHY with/without vegetation (depending on the catchment study)
- need for better understanding the processes of the project catchments
- need for water quality data on the italian site (push later the first deliverable?)
- Outputs of the MM should be relevant for the later operationnal use → to discuss with stakeholders ?
- water content at several profiles (water table level?), and contaminants concentration in the soil and in the groundwater
- it will be short to fond a student for the first 6 months !
- need to find the best postdoc for next year
- 3 years = short!

Task 4.3 Combining data and CATHY metamodel (M18-M30)

- combination of the available data on the Italian site (e.g., contaminants in the groundwater) as external data (with their own uncertainty) with CATHY's outputs in the metamodeling building.
- innovative method to improve the metamodel with observed data to make it closer to one specific catchment
- a bonus/risky task ?

⇝ Deliverable 4.3 - Python package to perform a metamodel coupling models' outputs and available data

- Hyp. = the deterministic output of the model is the realization of a GP
- The GP Z is conditioned by points from the model simulations (still a GP)