



RÉPUBLIQUE
FRANÇAISE

Liberté
Égalité
Fraternité

InfoGeol

Version 1.0



Géosciences pour une Terre durable

brgm

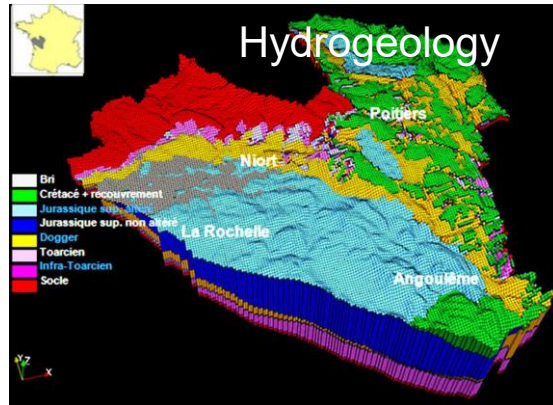
Uncertainty Quantification in Geosciences with the help of a surrogate model

Jérémy Rohmer j.rohmer@brgm.fr

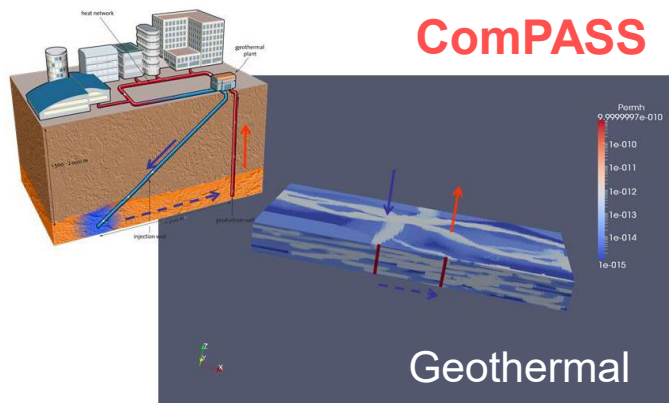
28 mai 2026

Key role of numerical models

Softwares developed w.r.t. BRGM activities



MARTHE

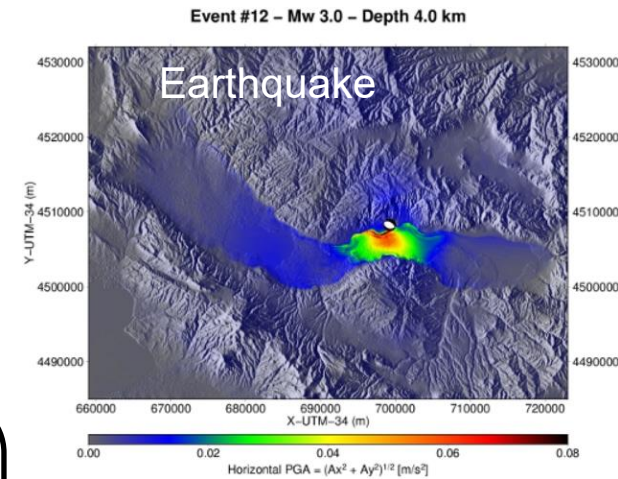


ComPASS

Computational Model

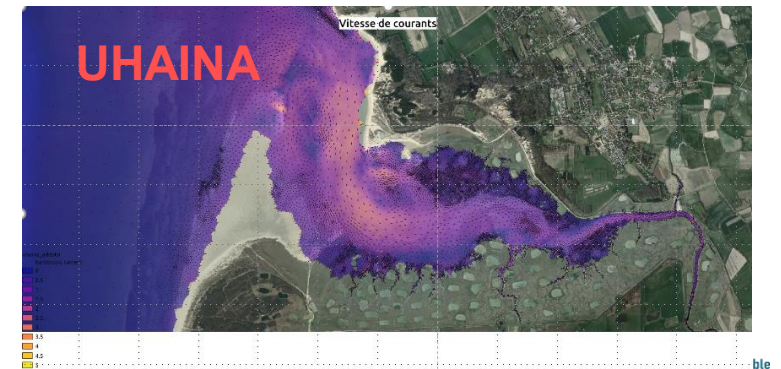
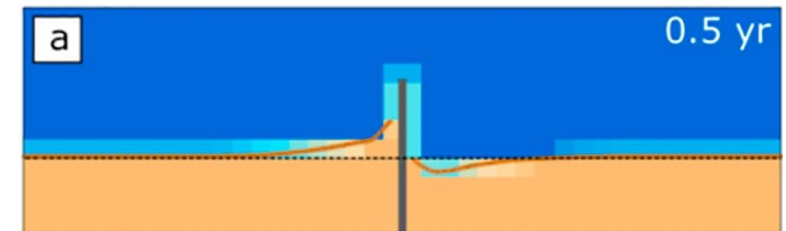
e.g. urban planning,
crisis management,
natural resource
exploitation, etc.

Support for
decision making

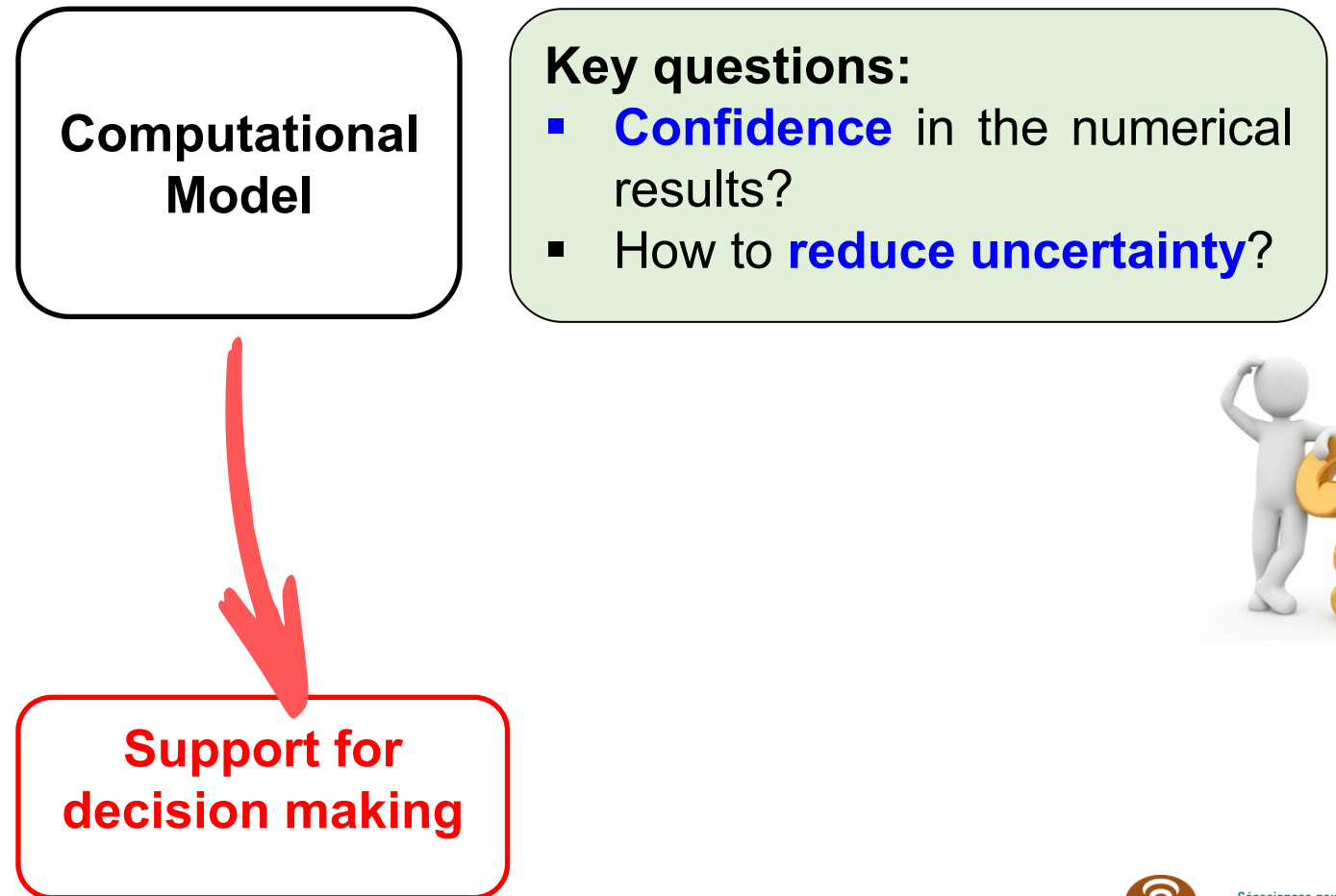


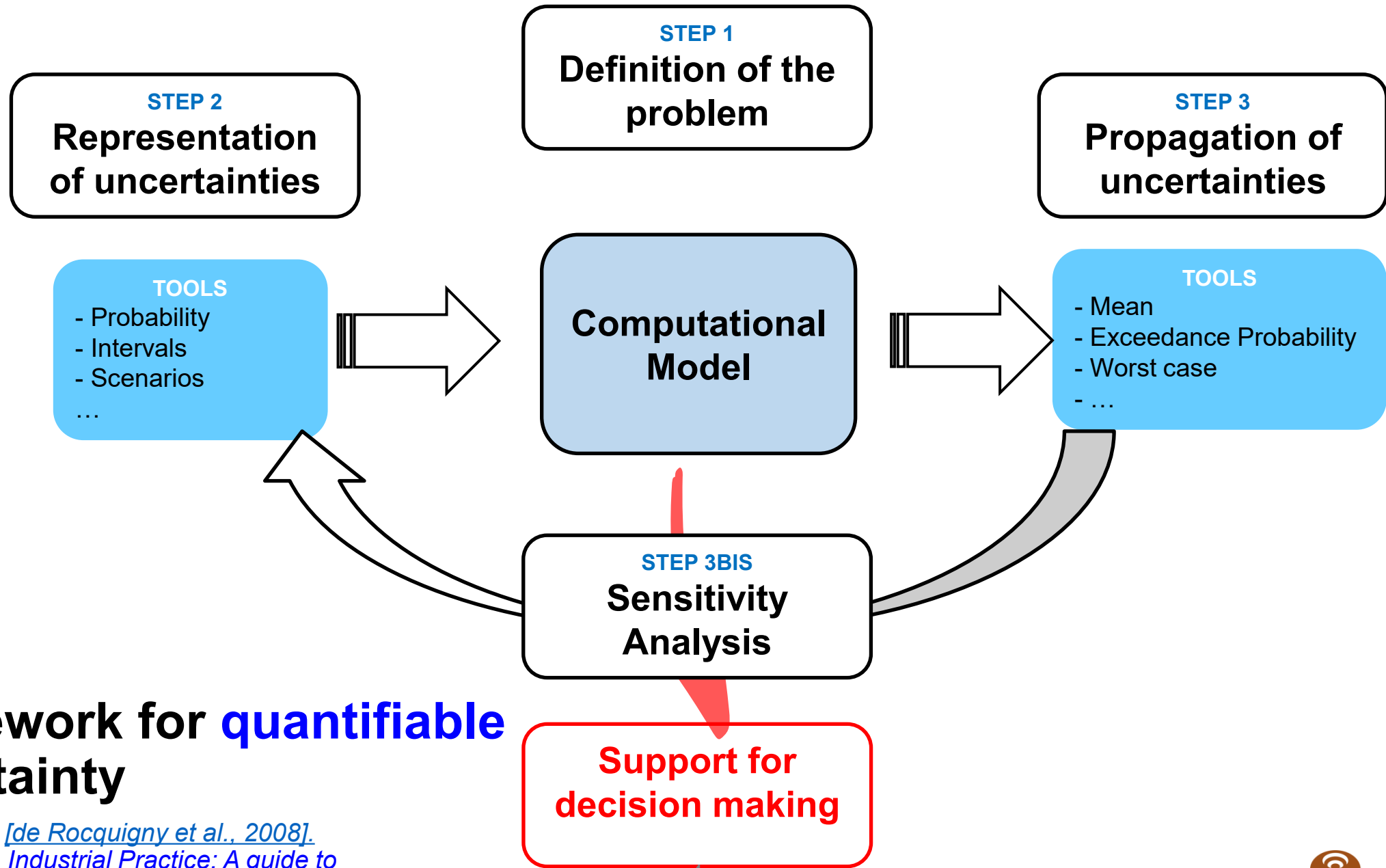
EFISPEC3D

LX-SHORE



Key role of numerical models





Framework for **quantifiable** uncertainty

Adapted from [de Rocquigny et al., 2008].
*Uncertainty in Industrial Practice; A guide to
Quantitative Uncertainty Management, Wiley*

STEP 2
**Representation
of uncertainties**

STEP 1
**Definition of the
problem**

STEP 3
**Propagation of
uncertainties**

TOOLS
- Probability
- Intervals
- Scenarios
...

**Computational
Model**

TOOLS
- Mean
- Exceedance Probability
- Worst case
- ...

Key question:
What **objective**?

STEP 3BIS
**Sensitivity
Analysis**

**Support for
decision making**

Understand
Accredit
Select
Comply

STEP 2
**Representation
of uncertainties**

TOOLS
- Probability
- Intervals
- ...

STEP 1
**Definition of the
problem**

**Impact
Assessment
Model**

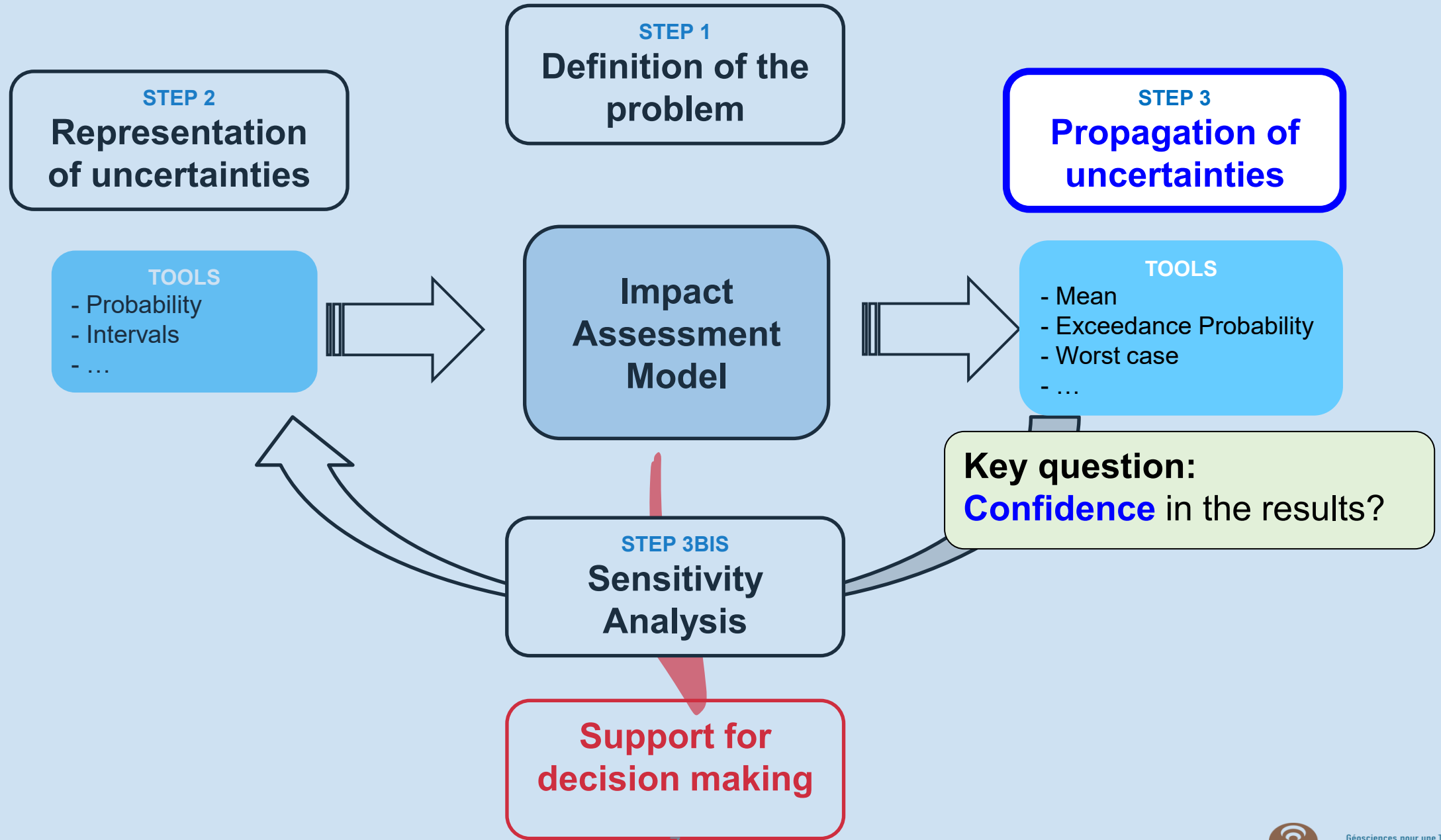
STEP 3
**Propagation of
uncertainties**

TOOLS
- Mean
- Exceedance Probability
- Worst case
- ...

Key question:
What do **I know**?
How to **model** this
knowledge?

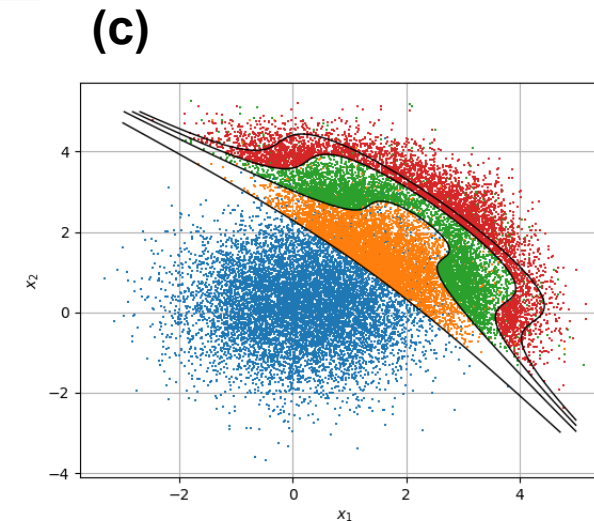
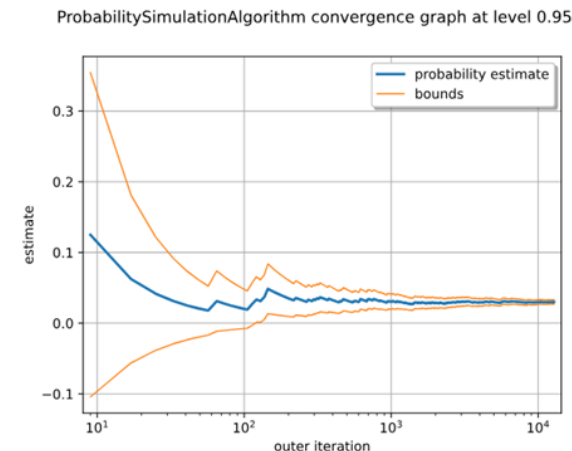
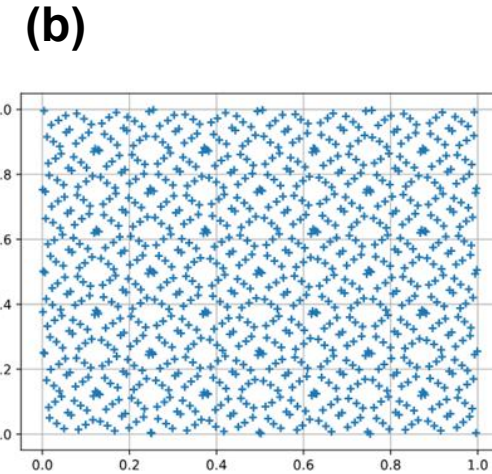
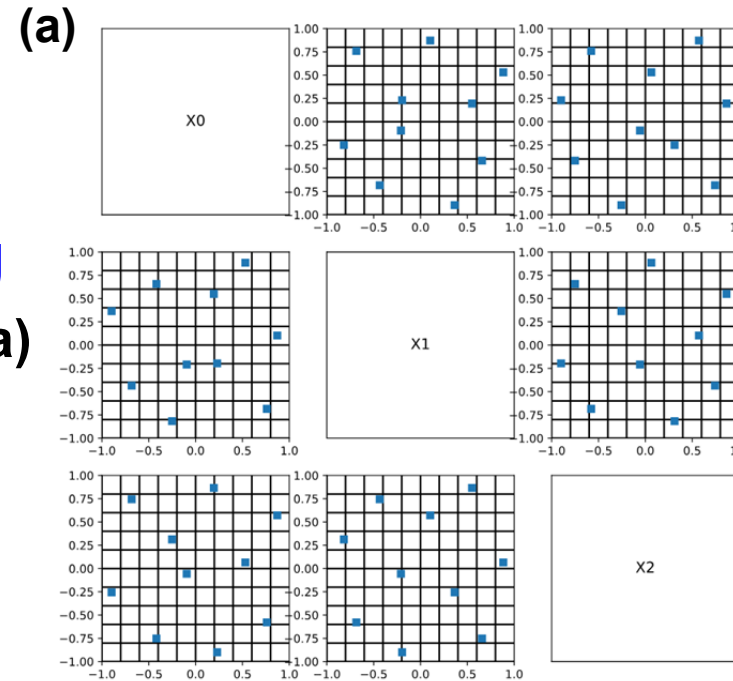
STEP 3BIS
**Sensitivity
Analysis**

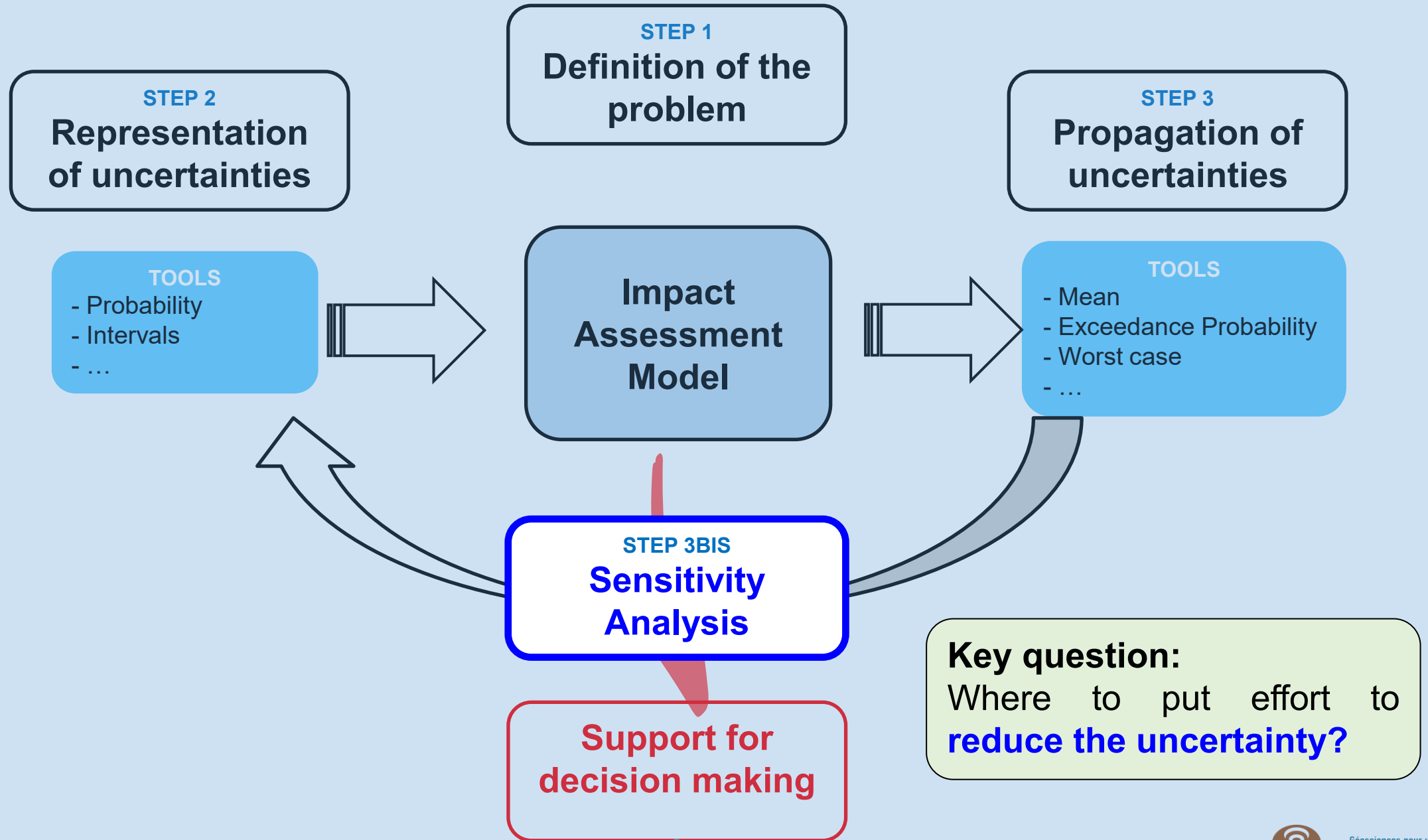
**Support for
decision making**



Depending on the objective, different approaches...

- ❑ Monte-Carlo random sampling
- ❑ 'Accelerated' Monte Carlo sampling
 - ❑ Latin Hypercube Sampling (&variants) (a)
 - ❑ Low discrepancy random sequences (b)
 - ❑ Importance sampling
 - ❑ Multi-level sampling (c)
 - ❑ ...
- ❑ Reliability methods
 - ❑ FORM/SORM
 - ❑ Monotonic Reliability Method
- ❑ Interval computation and optimisation





*“For every dollar that is spent trying to quantify uncertainty, we should spend 10 dollars collecting and analyzing data that would **reduce uncertainty**”.*

[Gail Atkinson] (2004 World Conference on Earthquake Engineering)



Two possible approaches

❑ Local

- One parameter is varied 'one at a time' OAT
- Quantify a deviation from a reference solution

❑ Global

- All parameters are varied all to together
- Account for probabilistic information
- Account for interactions

❑ Objective:

- Explore the influence of inputs X_i ($i=1..d$) on the output y ;
- Over the whole range of variation of X_i ;
- With account of interactions
- Model-free

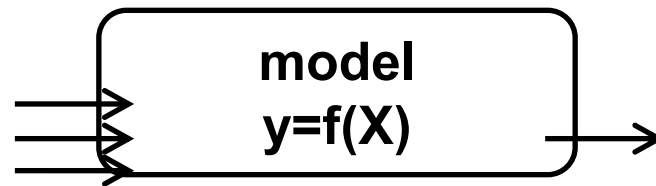
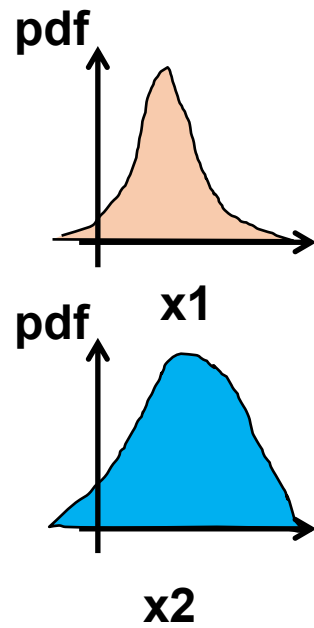
❑ Principle:

- Estimate the variance reduction of Y if I knew the true value x_i^* of X_i
- The input leading to the largest reduction has the largest influence

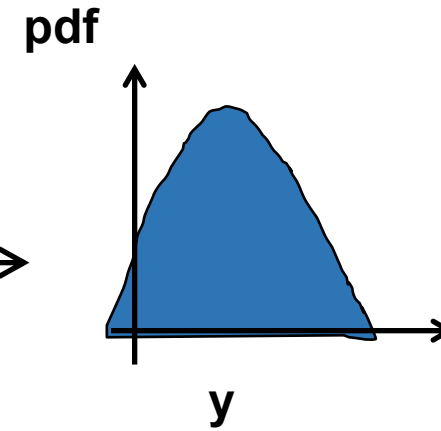
❑ Assumptions in the following

- $\text{Variance}(Y) \sim$ measure of uncertainty
- X_i are independent

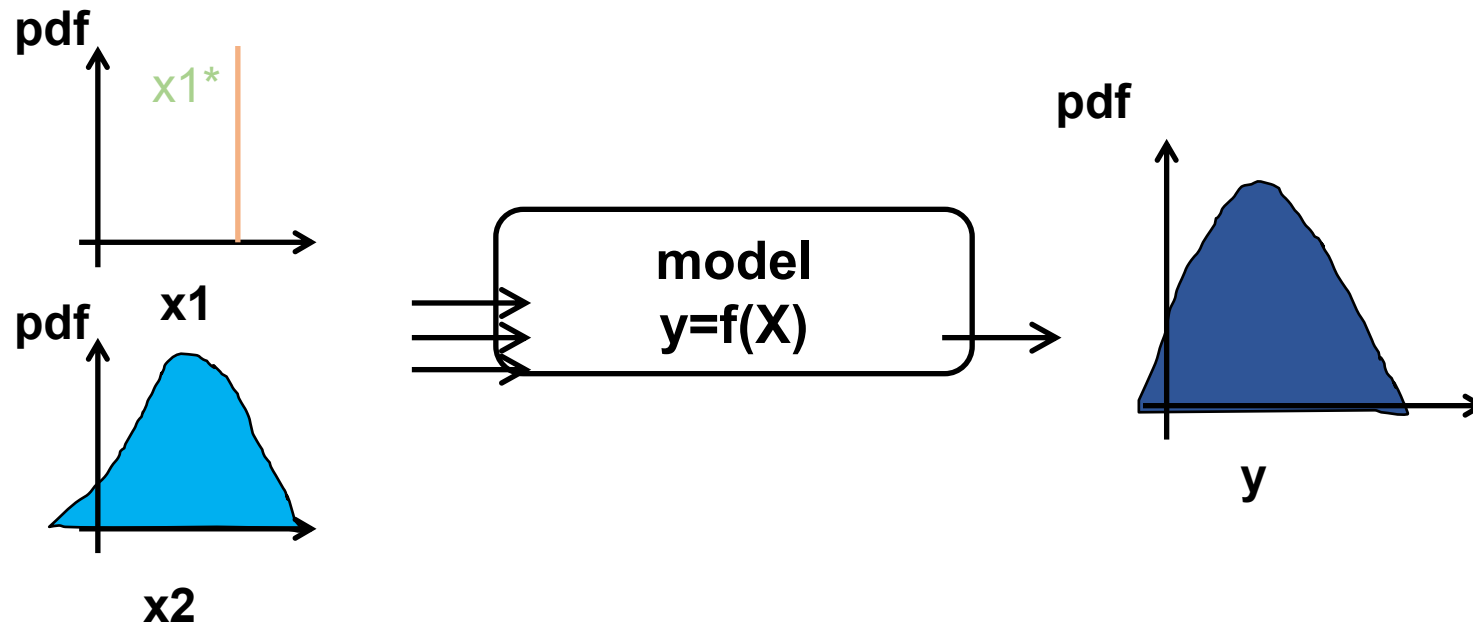
Inputs



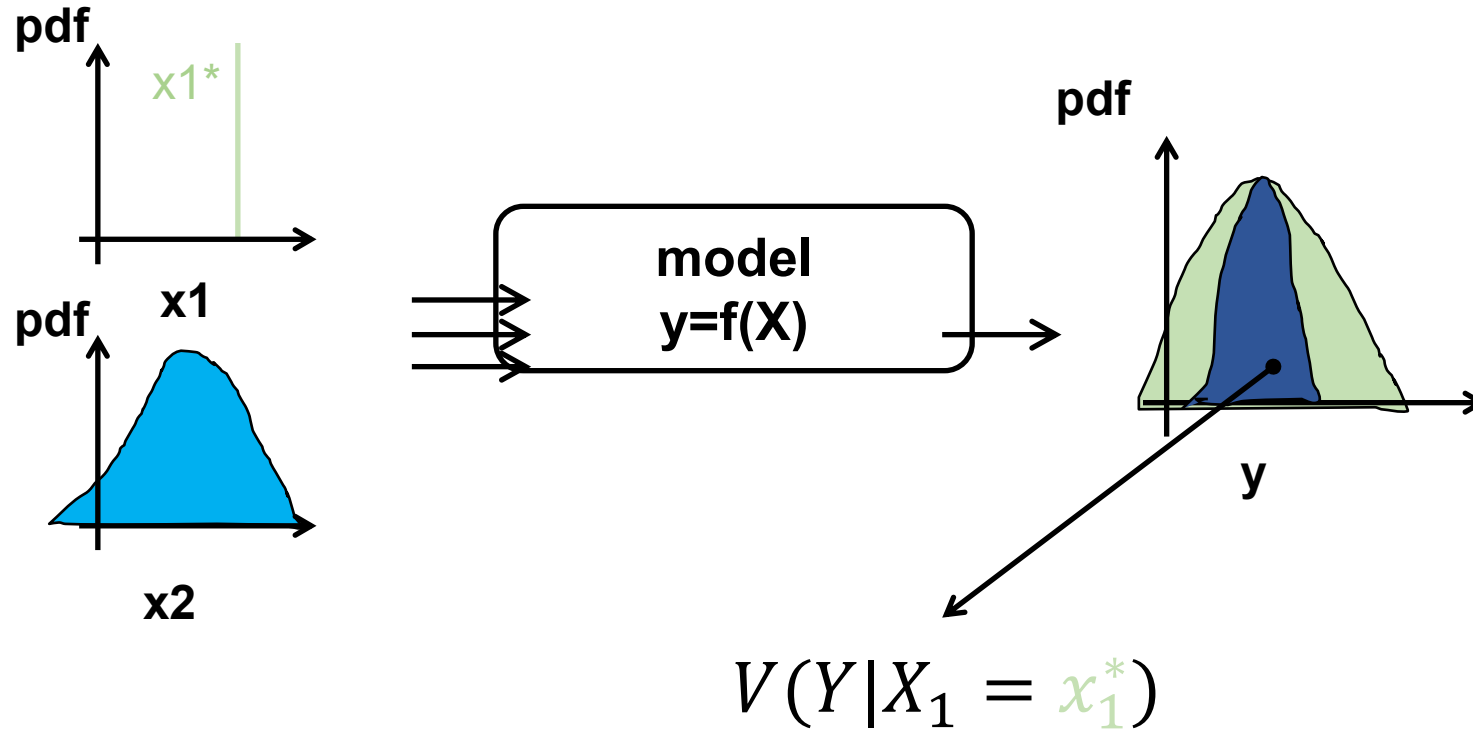
Output



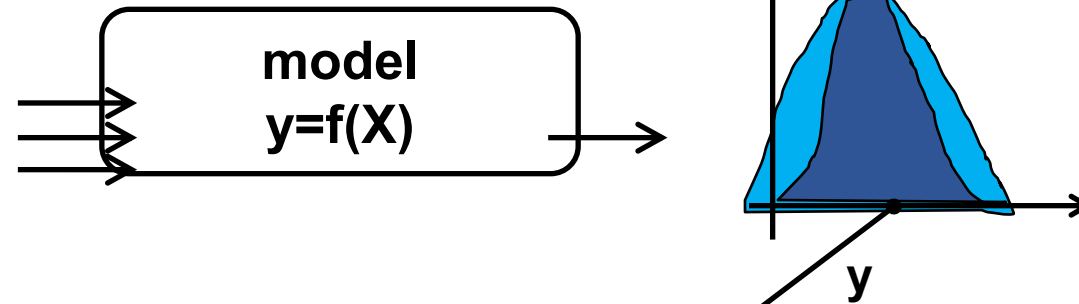
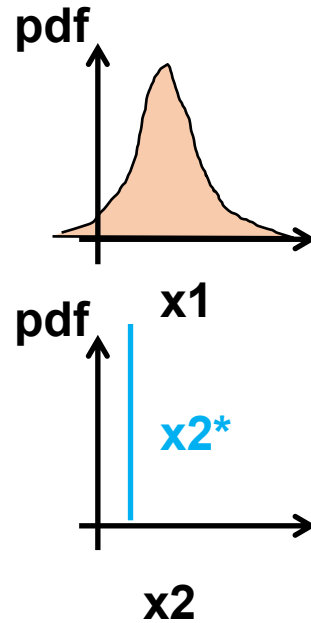
If the true value x_1^* was known the variance $V(Y)$ would be reduced by $V(1)$



If the true value x_1^* was known the variance $V(Y)$ would be reduced by $V(1)$

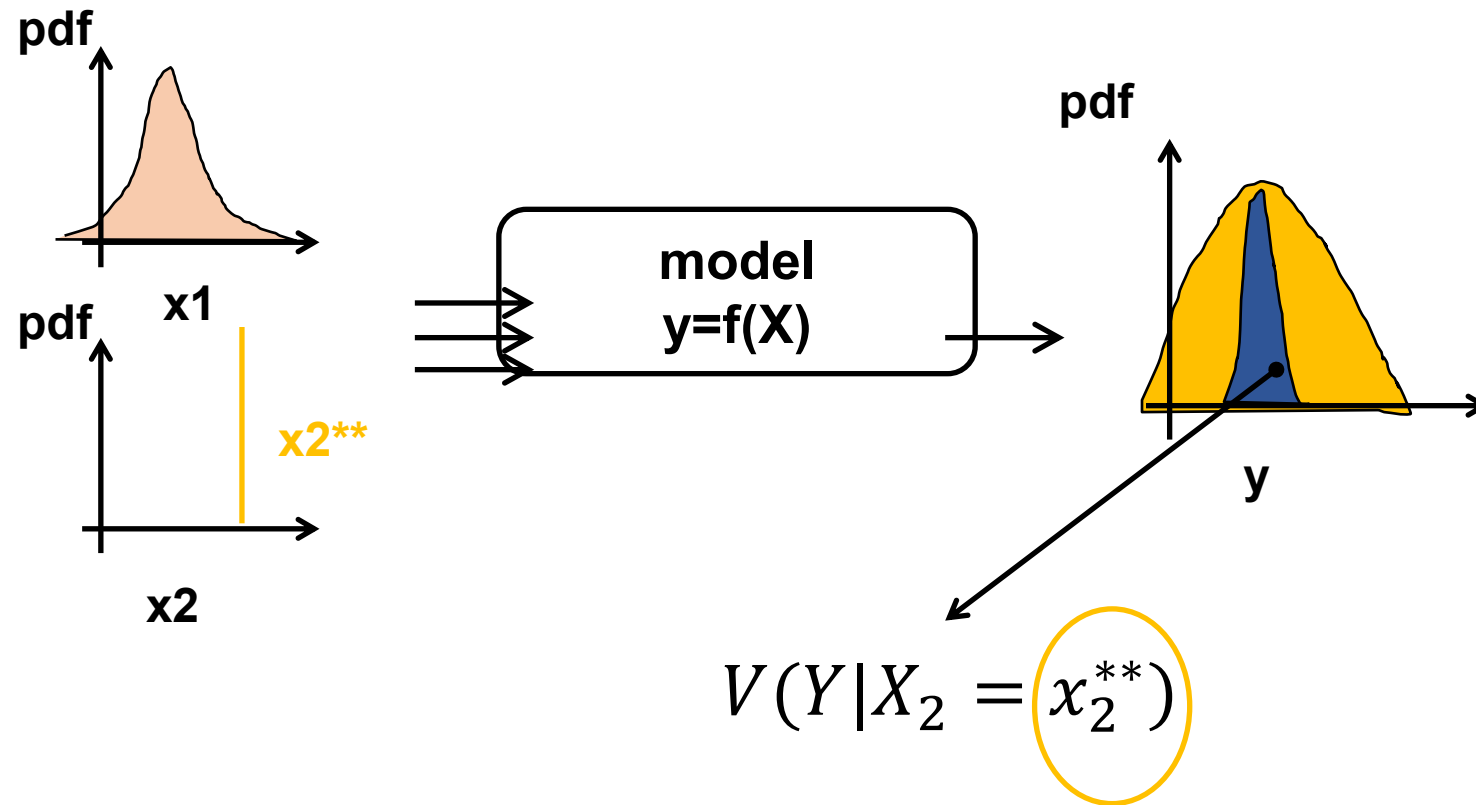


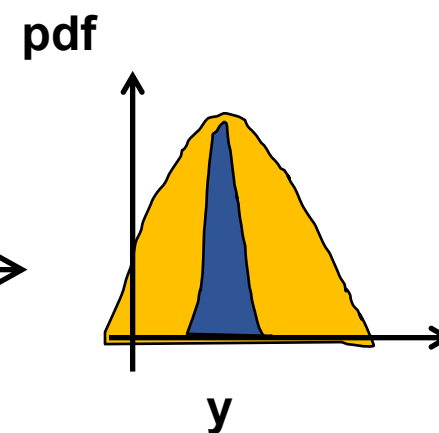
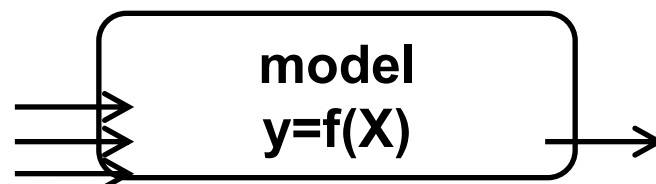
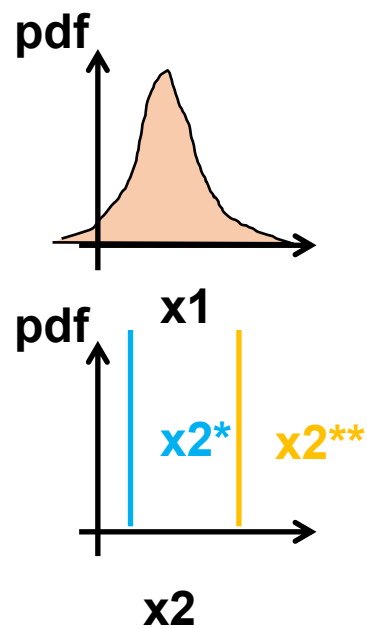
If the true value x_2^* was known the variance $V(Y)$ would be reduced by $V(2)$



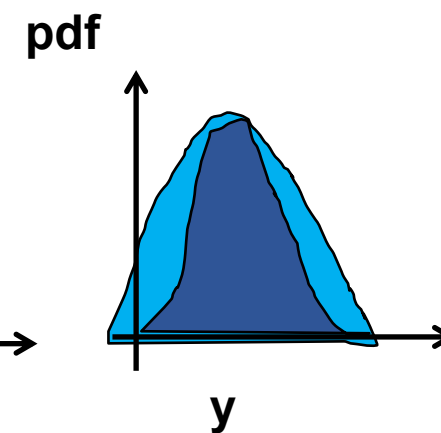
$$V(Y|X_2 = x_2^{**})$$

If the true value x_2^{**} was known the variance $V(Y)$ would be reduced by $V(2)'$





?



Since we do not know the true value of x_2

➔ $E(V(Y|X_2))$

Definition of sensitivity indices

[Saltelli et al., 2008]

Law of Total Variance

$$V(Y) = E\left(V(Y|X_i = x_i^*)\right) + V\left(E(Y|X_i = x_i^*)\right)$$

Sobol' index of first order (main effect):

$$S_i = \frac{V(E(Y|X_i = x_i^*))}{V(Y)}$$

Functional ANOVA

$$V(Y) = \sum_i V_i + \sum_{i < j} V_{ij} + \sum_{i < j < l} V_{ijl} + \dots + V_{ij\dots d} \quad 2^d - 1 \text{ terms !!}$$

Total effect

$$S_{Ti} = 1 - \frac{V(E(Y|X_{-i}))}{V(Y)} \quad \text{Main effects + interactions}$$

with $X_{-i} = (X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_d)$

$$0 \leq S_i \leq S_{Ti} \leq 1$$

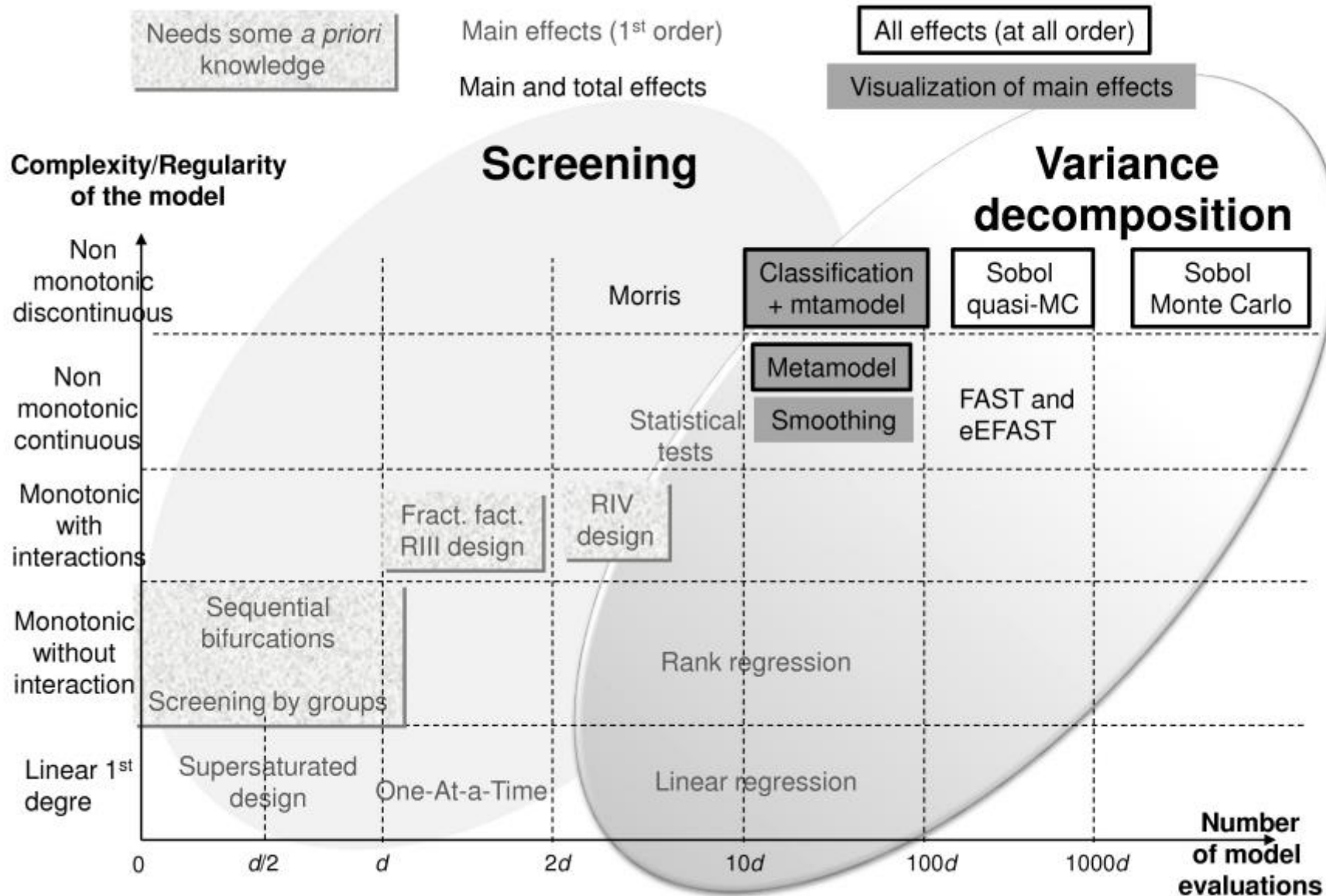
□ $S_{Ti} - S_i$: interaction betw. X_i and all other inputs

□ In practice

- $S_{Ti} = 0 \rightarrow f(X)$ does not depend on X_i
- $S_i = 1 \rightarrow f(X)$ only depends on X_i
- $S_i = S_i^{tot}$ no interaction
- $\sum_{i=1}^n S_i = 1$ additive function

□ Estimation:

- ❖ Monte-Carlo 'Pick-freeze' methods [[Sobol 2001](#); [Saltelli et al. 2010](#)] ← requires specific experimental random design
- ❖ Given Samples [rank-based: [Gamboa et al. 2020](#); [da Veiga et al. \(2025\)](#)] ← only uses N random samples



A few challenges

- ❑ Correlated inputs → Shapley effects [[Iooss & Prieur 2019](#)]
- ❑ Beyond variance → HSIC [[da Veiga 2015](#)]
- ❑ Complex outputs → [[Gamboa et al., 2021](#)]
- ❑ Computational burden → **use of surrogate models**

**Huge computation time
cost of a single run**

Versus

**Large number of model
runs needed**

**Key ingredient:
High-Performance Computing**

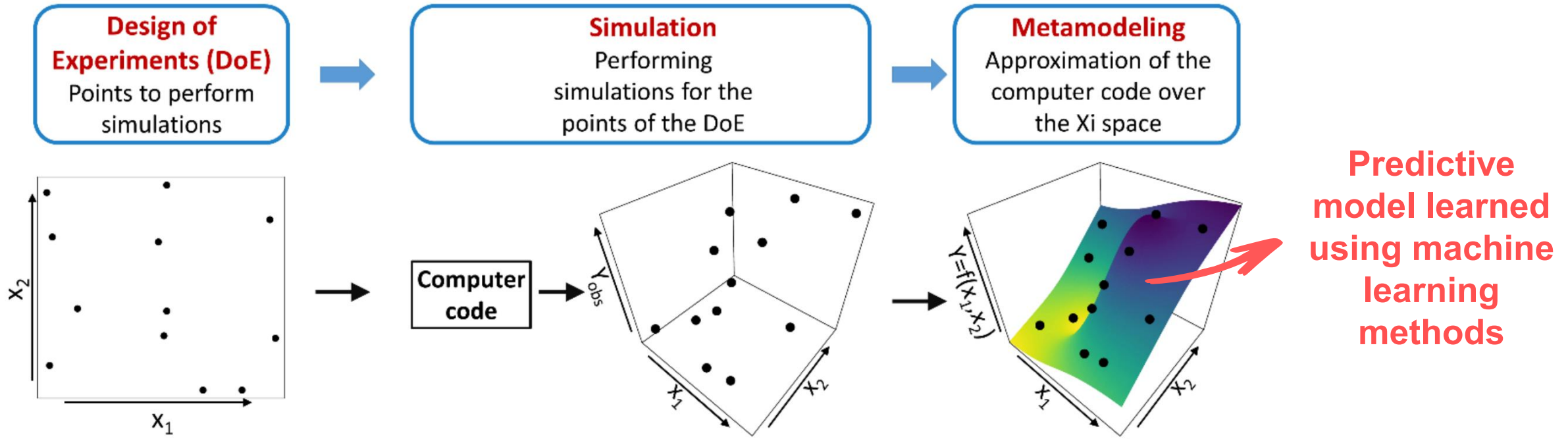


**Huge computation time
cost of a single run**

Versus

**Large number of model
runs needed**

**Complementary ingredient:
Surrogate modelling
(aka metamodeling)**



Multiple methods

- Polynomial regression
 - Smoothing techniques
 - Kriging, aka Gaussian Process
 - Polynomial Chaos expansion
 - Support Vector Machine
 - Neural network
 - **All techniques of data-science / machine learning**
 - Etc.
-
- **Validity?**
 - Comparison to the results of test set not used for training
 - In practice, cross-validation (LOO or K-fold)

A flexible surrogate model: Gaussian process (Gp) regression

[Williams and Rasmussen, 2006]

▪ **Objective:** predict Y_* for new inputs x_*
without running the computer code

▪ Assumptions

- distribution of Y_* depends on n simulation results $Y_n = y$
- distribution of $Y_* | \{Y_n = y\}$ is Gaussian

With parameters (simple kriging equations)

▪ Mean $m = k^{*t} \cdot K^{-1} y$

▪ Variance $\sigma^2 = k(x_*, x_*) - k^{*t} \cdot K^{-1} \cdot k^*$

With

Covariance matrix $K = (k(x_i, x_j))_{1 \leq i, j \leq n}$

Covariance vector $k^* = (k(x_*, x_1), k(x_*, x_2), \dots, k(x_*, x_n))$

A flexible surrogate model: Gaussian process (Gp) regression

[Williams and Rasmussen, 2006]

▪ **Objective:** predict Y_* for new inputs x_*
without running the computer code

▪ Assumptions

- distribution of Y_* depends on n simulation results $Y_n = y$
- distribution of $Y_* | \{Y_n = y\}$ is Gaussian

Uncertainty
quantification

With parameters (simple kriging equations)

▪ Mean $m = k^{*t} \cdot K^{-1} y$

▪ Variance $\sigma^2 = k(x_*, x_*) - k^{*t} \cdot K^{-1} \cdot k^*$

Flexibility via kernels

Function modelling the notion of **neighbouring** i.e. $Y(x_{(1)})$ and $Y(x_{(2)})$ should be close if $x_{(1)}$ and $x_{(2)}$ are

With

Covariance matrix $K = (k(x_i, x_j))_{1 \leq i, j \leq n}$

Covariance vector $k^* = (k(x_*, x_1), k(x_*, x_2), \dots, k(x_*, x_n))$

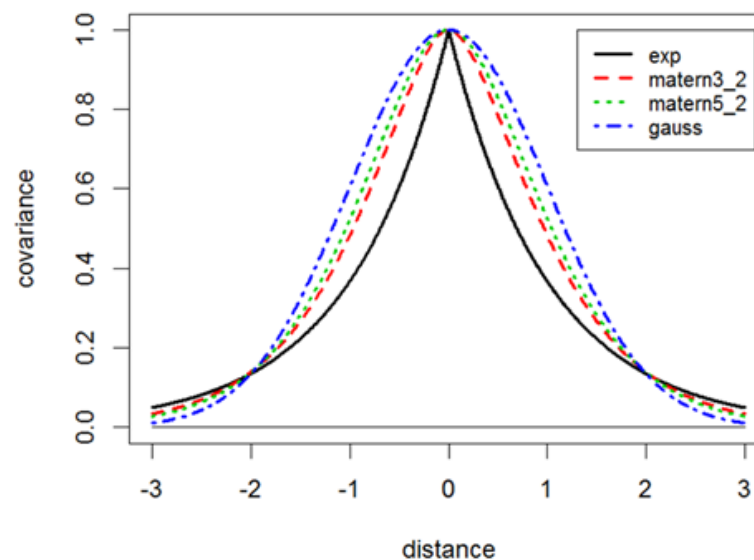
■ Possible choices

- Exponentiel $k(x_i, x_j) = \sigma^2 \exp\left(-\frac{|x_i - x_j|}{\theta}\right)$
- Gaussian (squared exponential) $k(x_i, x_j) = \sigma^2 \exp\left(-\frac{(x_i - x_j)^2}{2\theta^2}\right)$
- Matérn 3/2 $k(x_i, x_j) = \sigma^2 \left(1 + \sqrt{3} \frac{|x_i - x_j|}{\theta}\right) \exp\left(-\sqrt{3} \frac{|x_i - x_j|}{\theta}\right)$
- ...

■ Key parameters

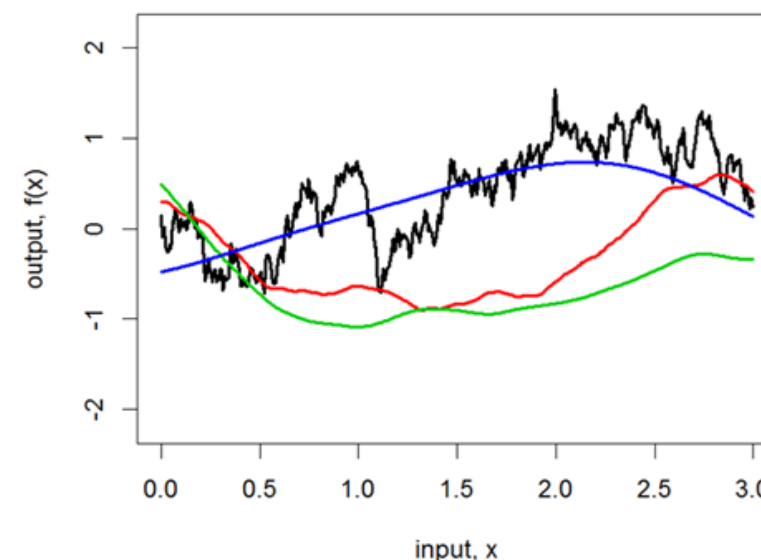
- σ^2 amplitude
- θ correlation length

Covariance function



$\sigma^2=1$; lengthscale $\theta=1$

Example of 1 random sample



■ Possible choices

- Exponentiel $k(x_i, x_j) = \sigma^2 \exp\left(-\frac{|x_i - x_j|}{\theta}\right)$
- Gaussian (squared exponential) $k(x_i, x_j) = \sigma^2 \exp\left(-\frac{(x_i - x_j)^2}{2\theta^2}\right)$
- Matérn 3/2 $k(x_i, x_j) = \sigma^2 \left(1 + \sqrt{3} \frac{|x_i - x_j|}{\theta}\right) \exp\left(-\sqrt{3} \frac{|x_i - x_j|}{\theta}\right)$
- ...

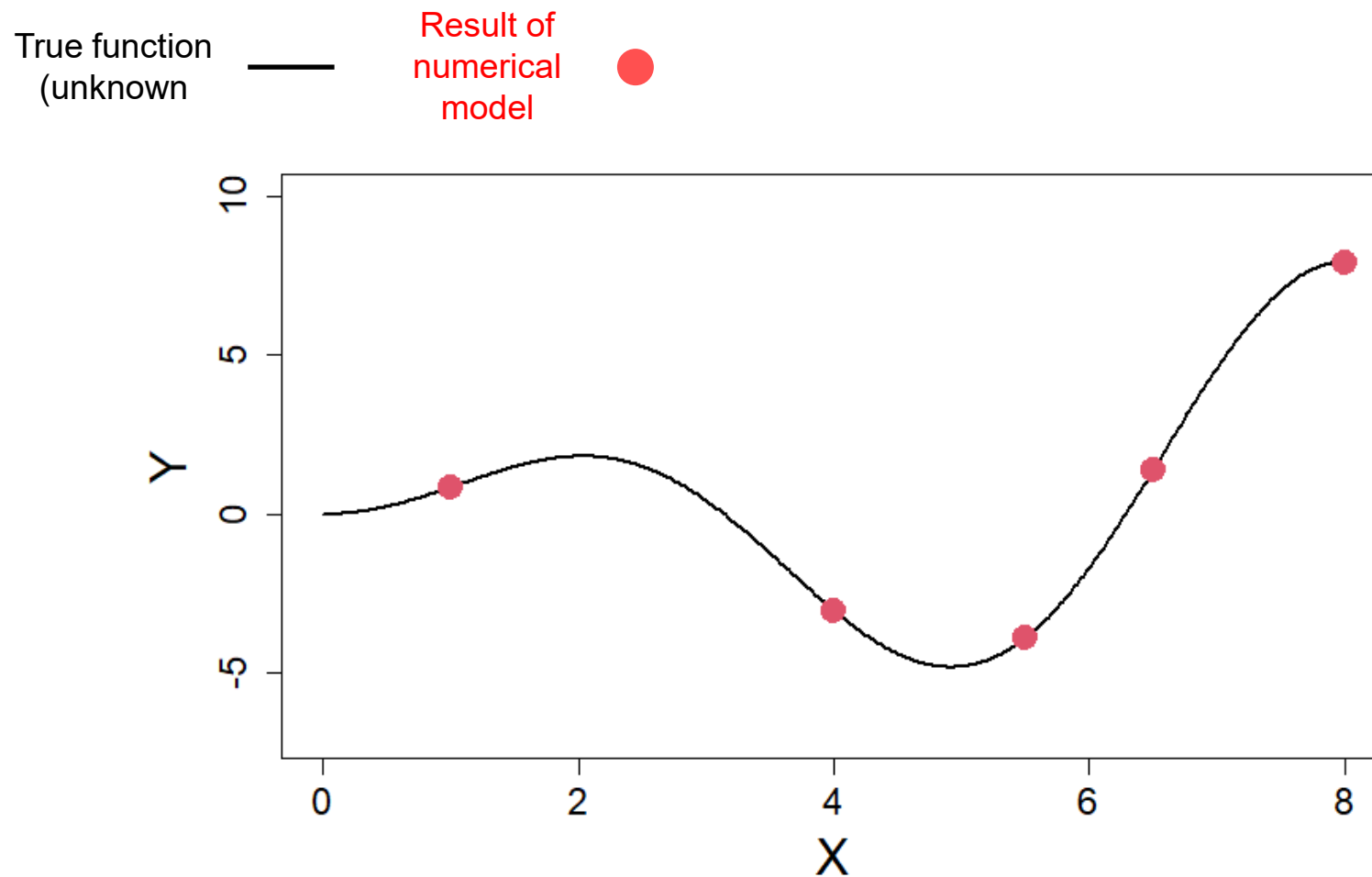
■ Key parameters

- σ^2 amplitude
- θ correlation length

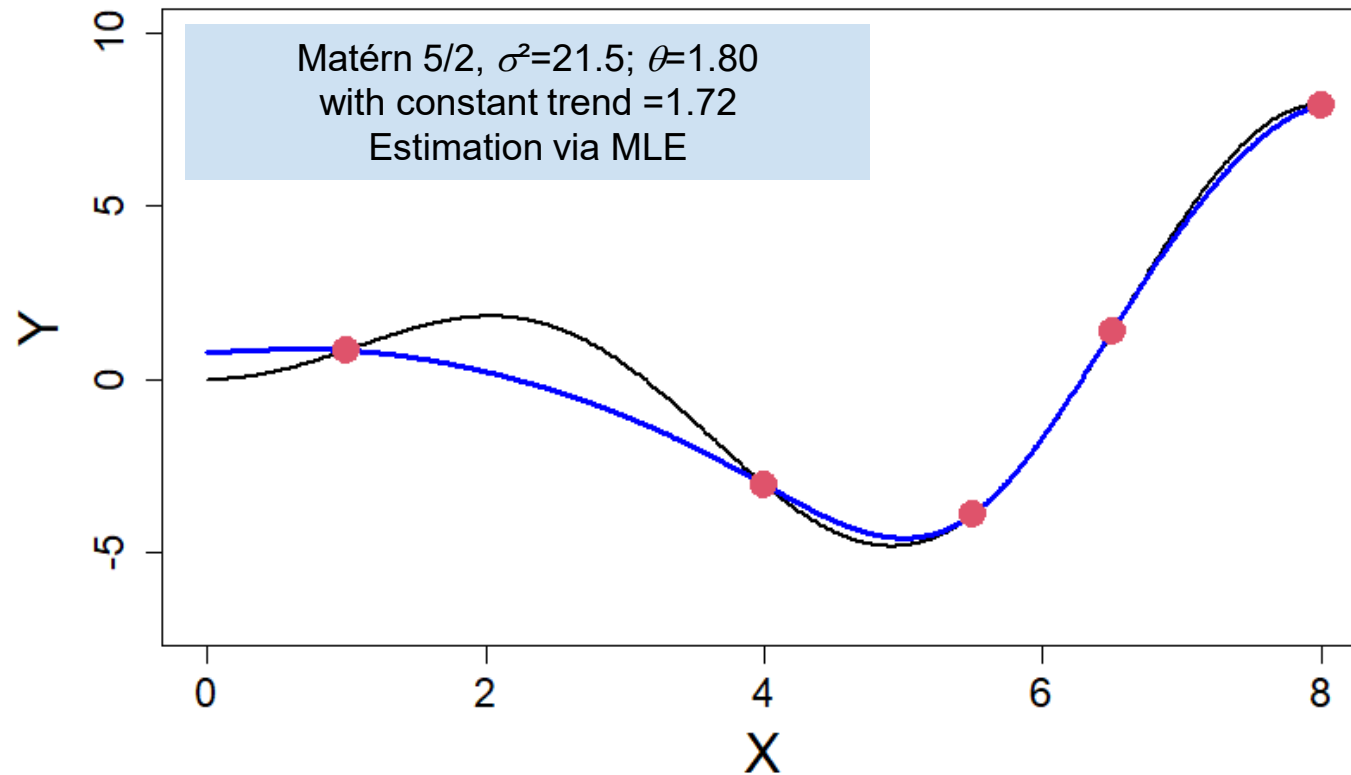
■ Definition for multiple inputs: through tensorisation of 1d- kernels

■ Estimation via maximimisation of log-likelihood (other choices possible)

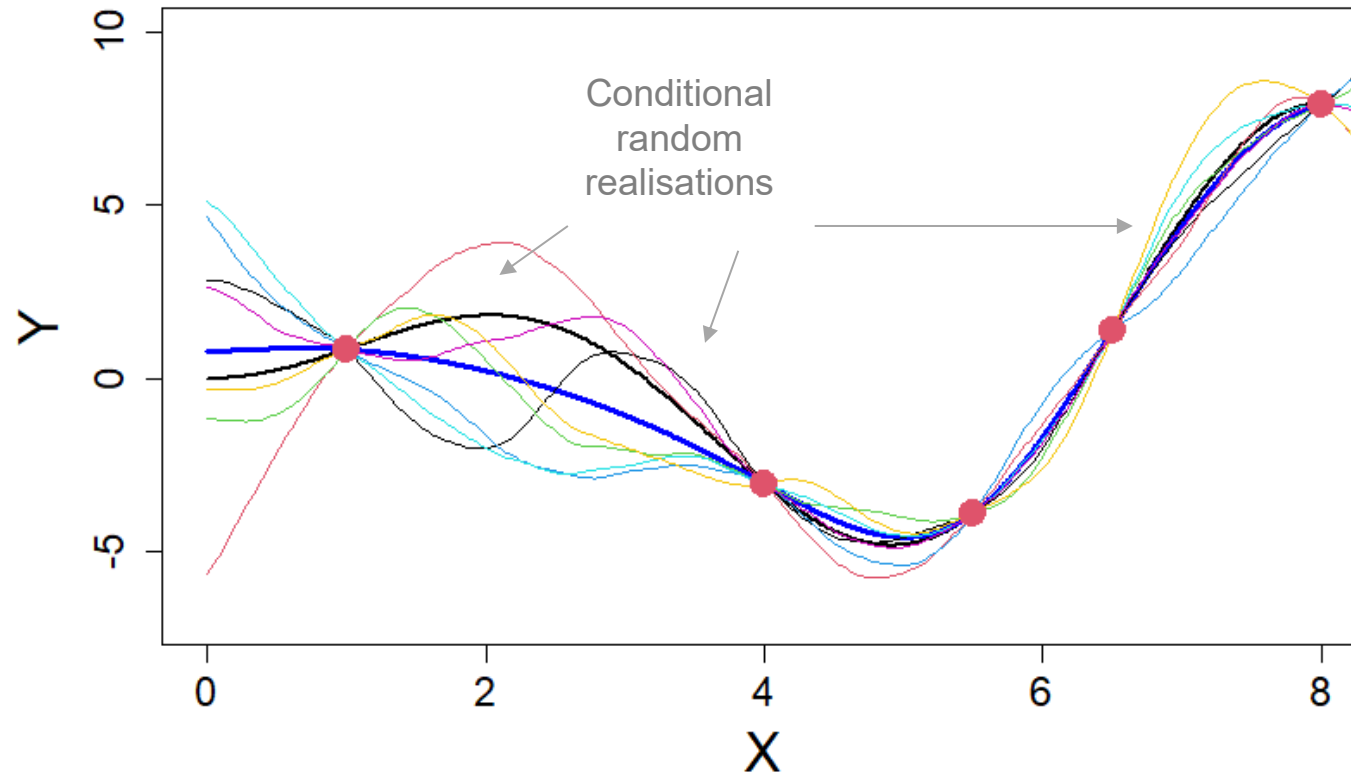
$$2\log(L) = -\mathbf{y}^T \mathbf{K}^{-1} \mathbf{y} - \log|\mathbf{K}| - n\log(2\pi)$$



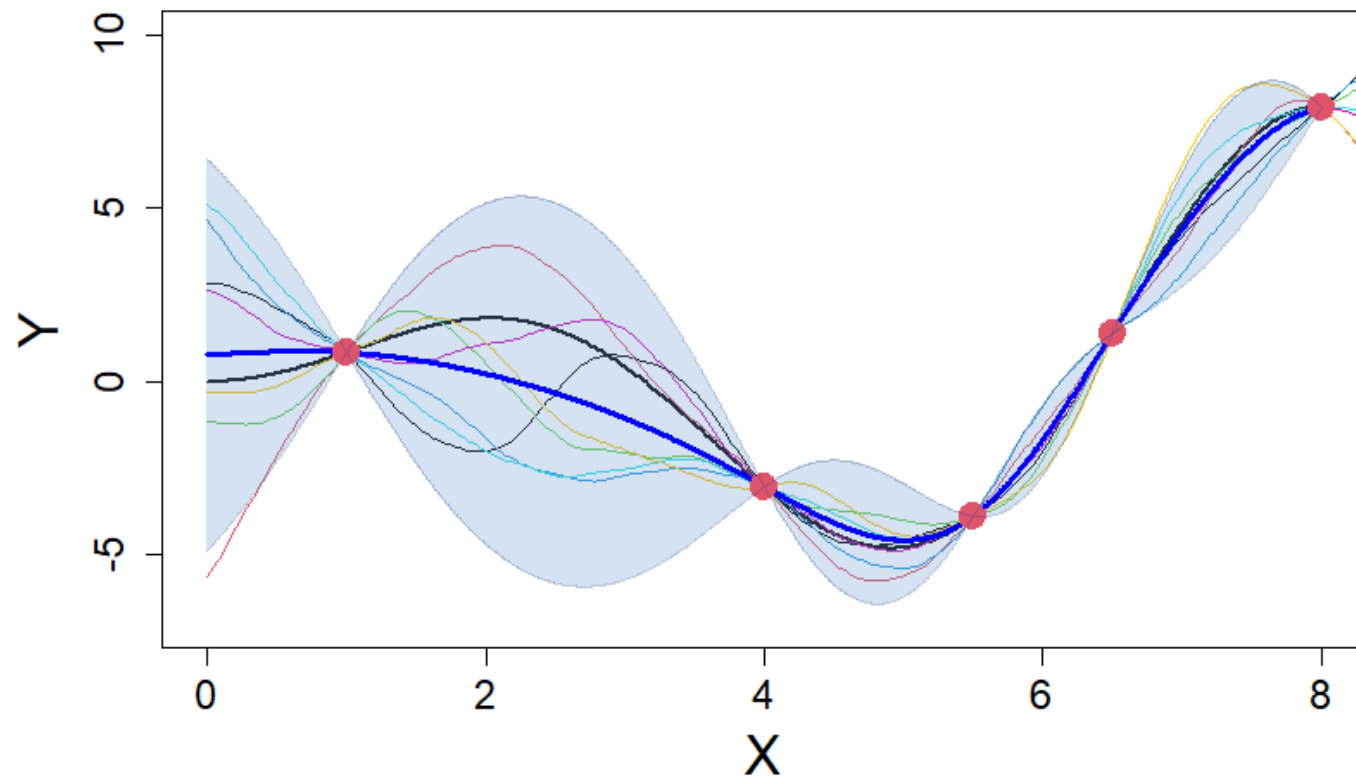
True function (unknown) — Result of numerical model ● Surrogate (Gp mean) —



True function (unknown) — Result of numerical model ● Surrogate (Gp mean) —

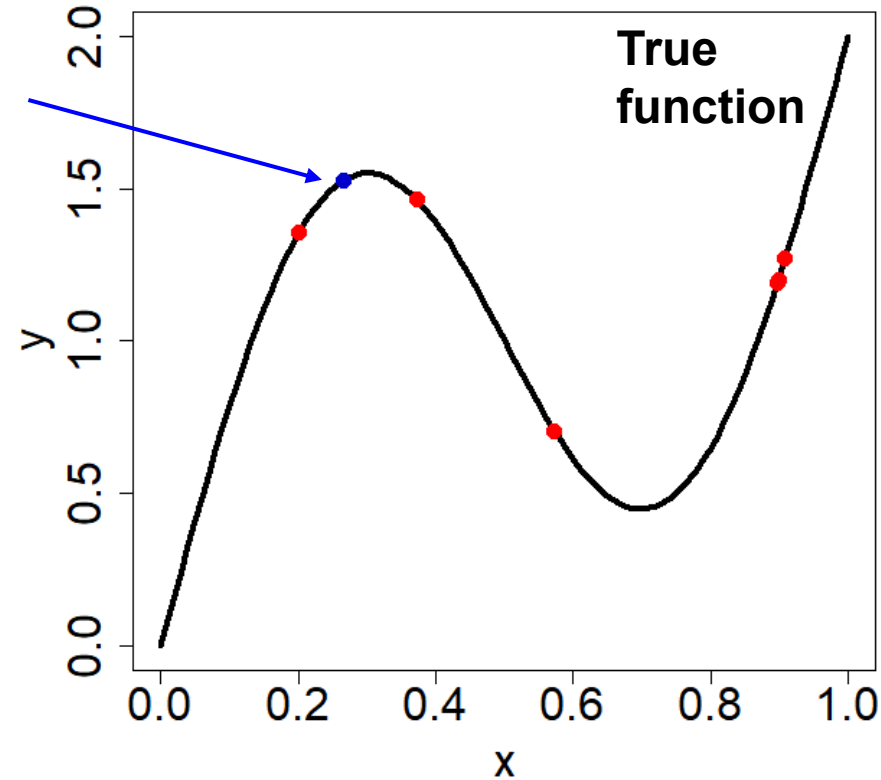


True function (unknown) — Result of numerical model ● Surrogate (Gp mean) — 95% uncertainty band ($\propto \sigma$ of Gp) 

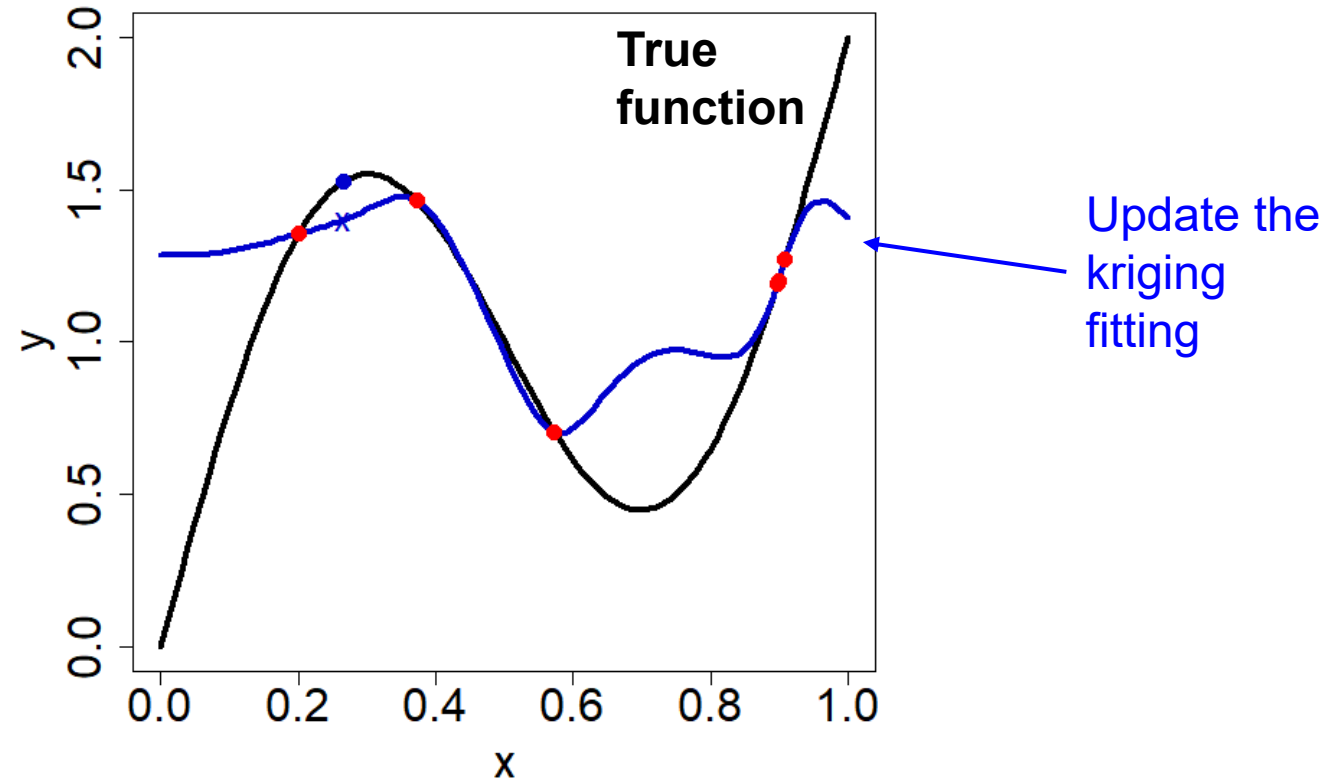


Principle of cross-validation Leave-One-Out

Extract an observation from the training dataset

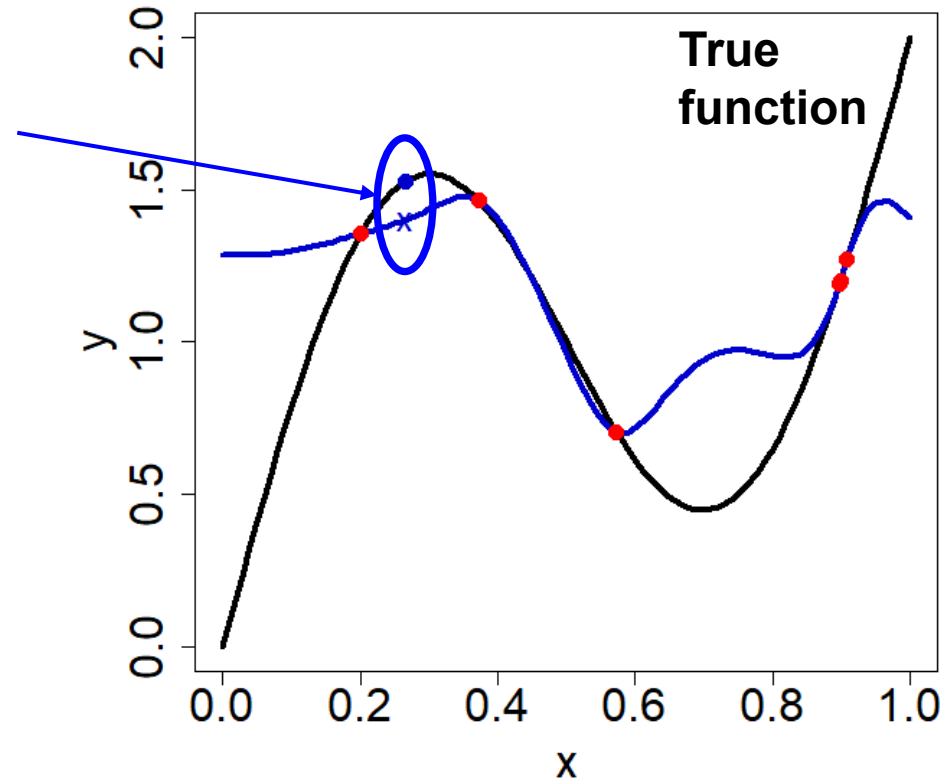


Principle of cross-validation Leave-One-Out

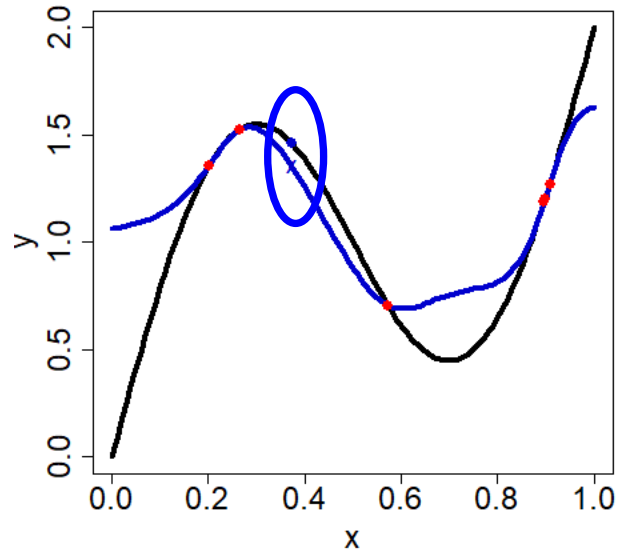


Principle of cross-validation Leave-One-Out

Estimate the error on the prediction



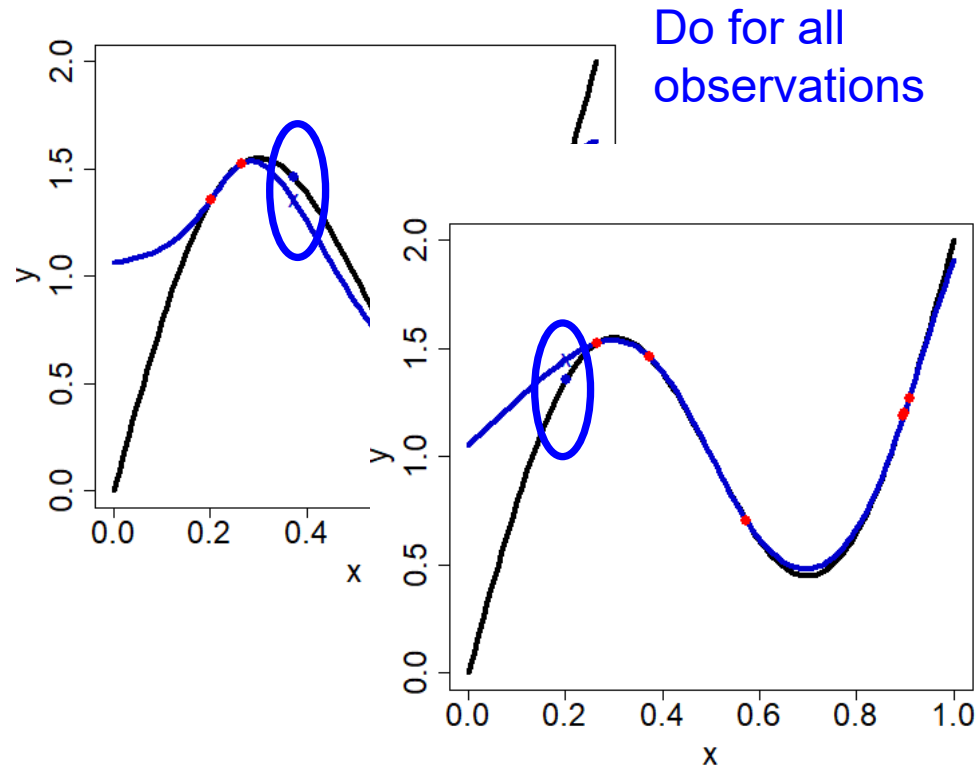
Principle of cross-validation Leave-One-Out



Do for all
observations

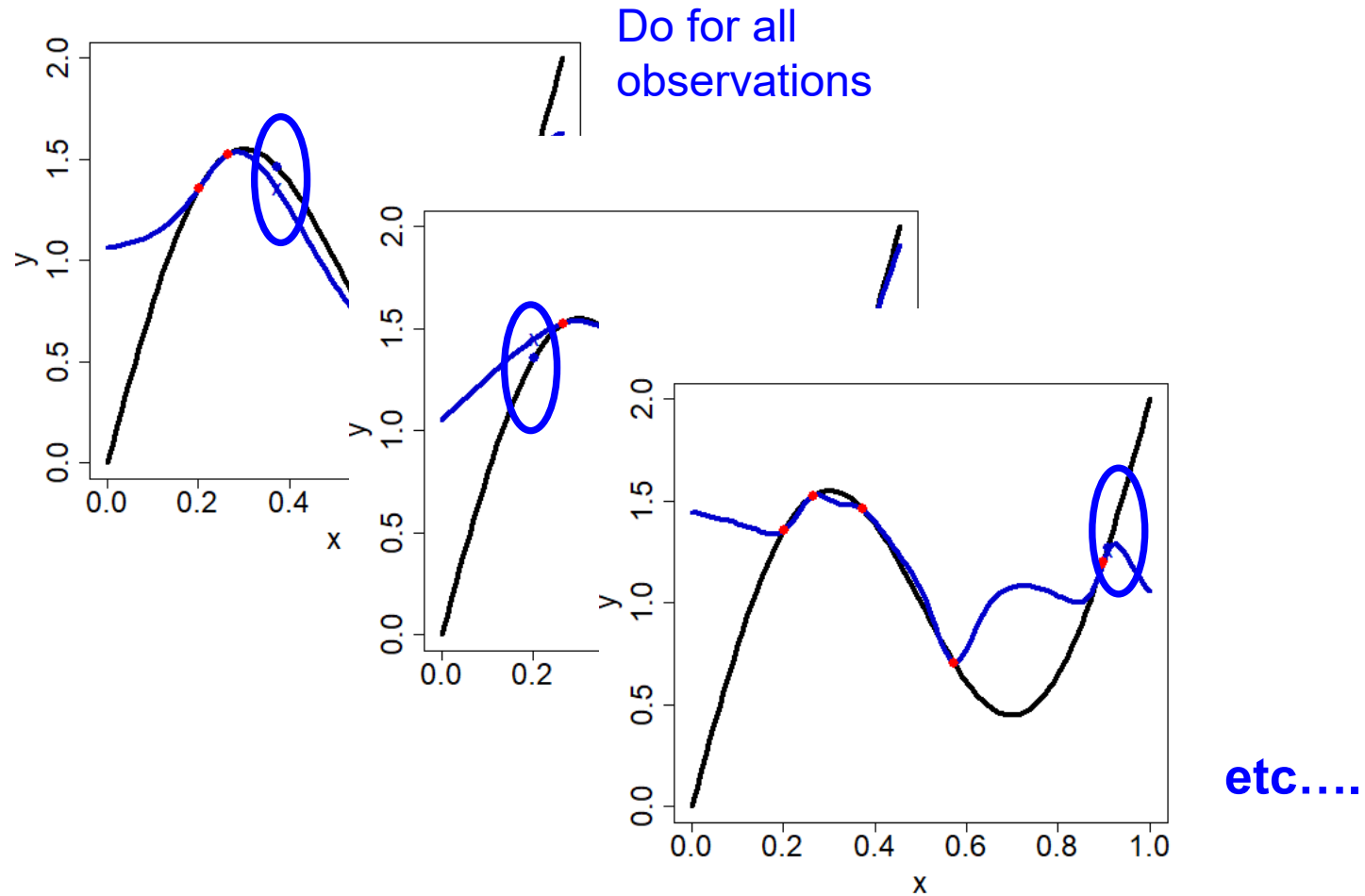
etc....

Principle of cross-validation Leave-One-Out



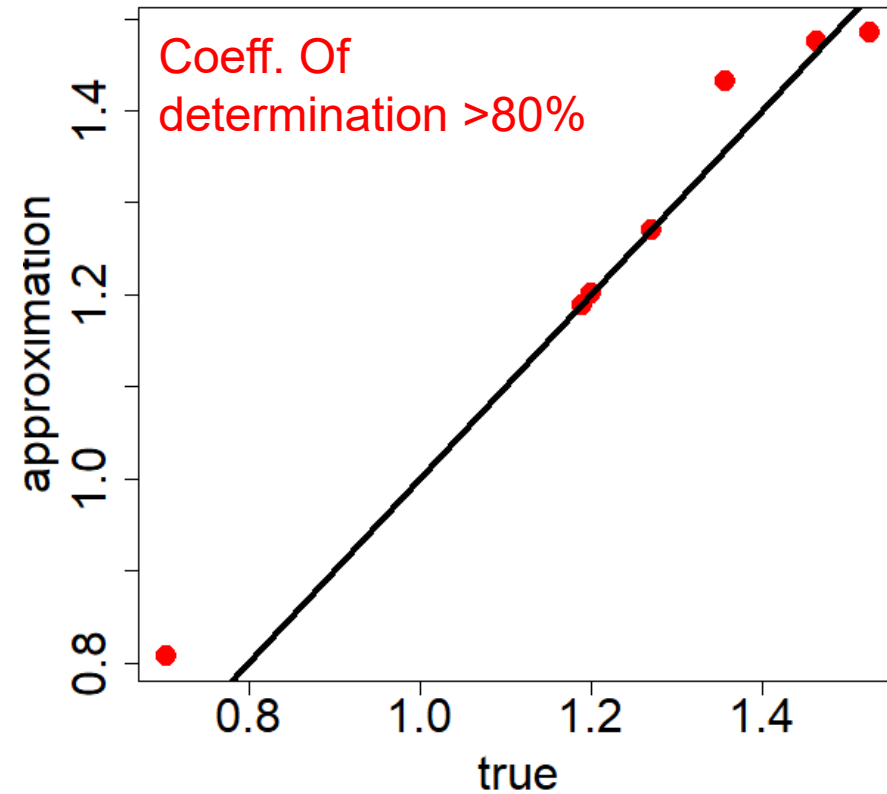
etc....

Principle of cross-validation Leave-One-Out



Principle of cross-validation Leave-One-Out

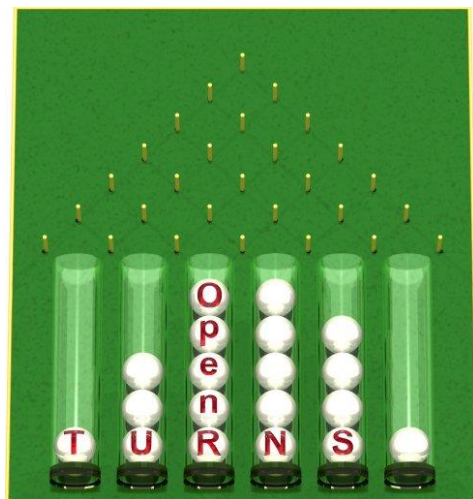
$$Q^2 = 1 - \frac{\sum_{i \in T} (y_i - \hat{y}_i)^2}{\sum_{i \in T} (y_i - \bar{y})^2}$$



A few references to go beyond the scalar-on-scalar case

- ❑ Adaptative design of experiments [[Bect et al. 2012](#)]
- ❑ Scenario-like, categorical inputs [[Roustant et al., 2020](#); [Rohmer et al. 2022](#)]
- ❑ Complex inputs: time series [[Betancourt et al., 2020](#)], spatial inputs [[Liu and Guillas 2016](#); [Rohmer 2015](#)]
- ❑ Complex output [[El Garroussi et al., 2022](#); [Chang et al., 2014](#); [Perrin et al. 2020](#); [Li et al., 2020](#); [Gu and Berger, 2016](#)]
- ❑ Mixing with low fidelity simulations [[Rohmer et al., 2023](#); [Thenon et al. 2016](#)]
- ❑ For calibration [[Higdon et al., 2008](#); [Salter et al. 2019](#)]

In practice



<https://openturns.github.io/www/>

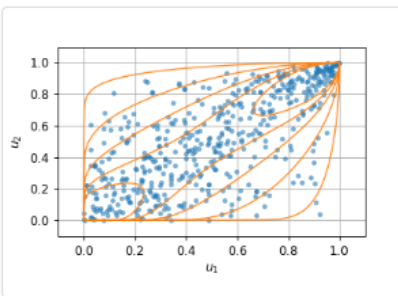
Module python, [github](#)

Reference:
[Baudin et al. \(2016\)](#)



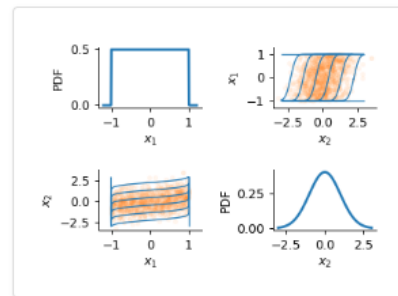
Tutorial today

Data analysis



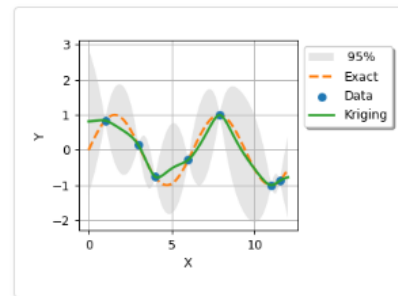
Manage data and samples
Sample analysis
Distribution fitting
Statistical hypothesis testing
Estimate dependency and copulas
Estimate stochastic processes
Graphics

Probabilistic modeling



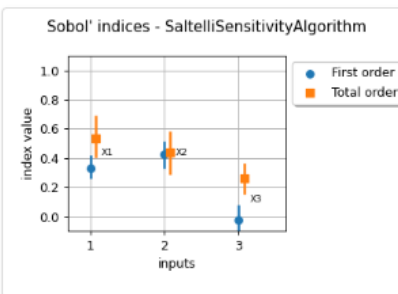
Distributions
Copulas
Stochastic processes

Surrogate modeling



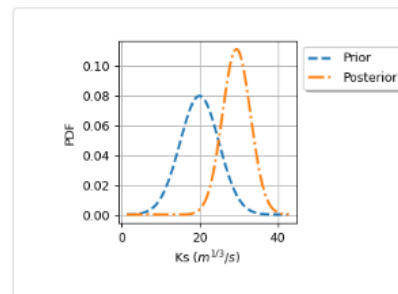
General purpose metamodels
Polynomial chaos metamodel
Gaussian process regression
Fields metamodels

Reliability sensitivity



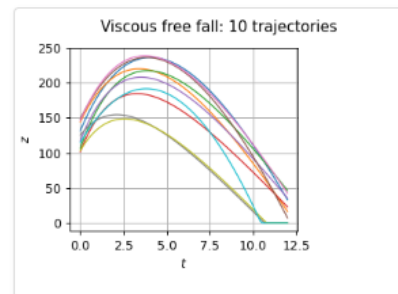
Central dispersion
Reliability
Reliability: processes
Sensitivity analysis
Design of experiments

Calibration



Least squares and gaussian calibration
Bayesian calibration

Functional modeling



Vectorial functions
Field functions
Link to an external code

Merci pour votre attention!
Des questions?

Contact: j.rohmer@brgm.fr