Statistical approaches in nuclear safety problems: recent advances around sensitivity analysis and metamodeling

FROM RESEARCH TO INDUSTRY

Amandine MARREL (CEA DES/IRESNE/DER)

In collaboration with my peers from CEA (G. Sarazin, R. El Amri [at IFPEN now]) and EDF R&D (B. Iooss, V. Chabridon)

Workshop "Statistical methods for safety and decommissioning“

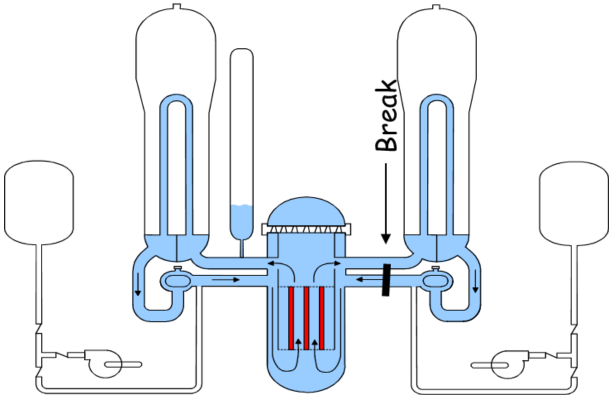
University of Avignon

IRESNE | DER | SESI | LEMS

French Alternative Energies and Atomic Energy Commission

Numerical Simulation – Experimental results – Database

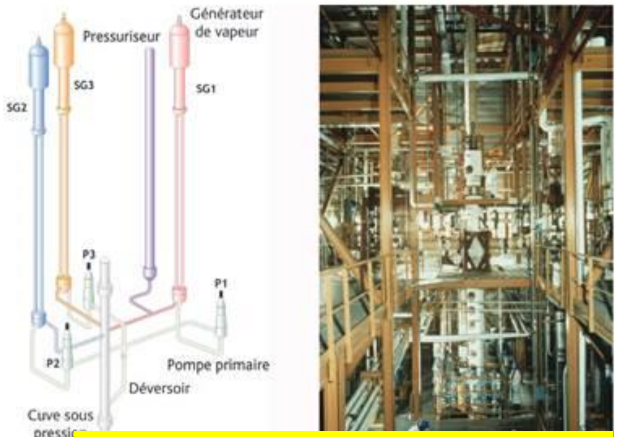
Assessment of a accidental scenario on PWR: Break Loss Of Coolant Accident (B-LOCA)

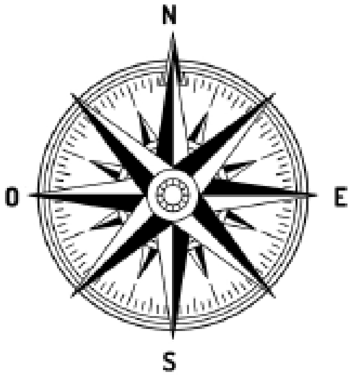


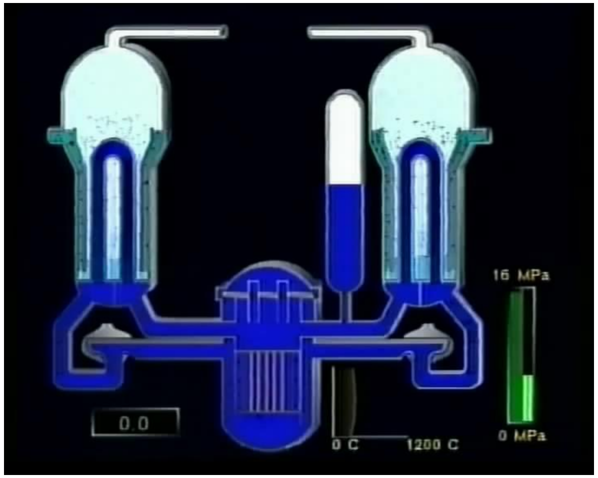
3/ Numerical simulation

CATHARE simulator

Thermal-hydraulic simulation *of multiphase flow dynamics*
developed by the CEA
with EDF, FRAMATOME and IRSN







1/ Experimental results

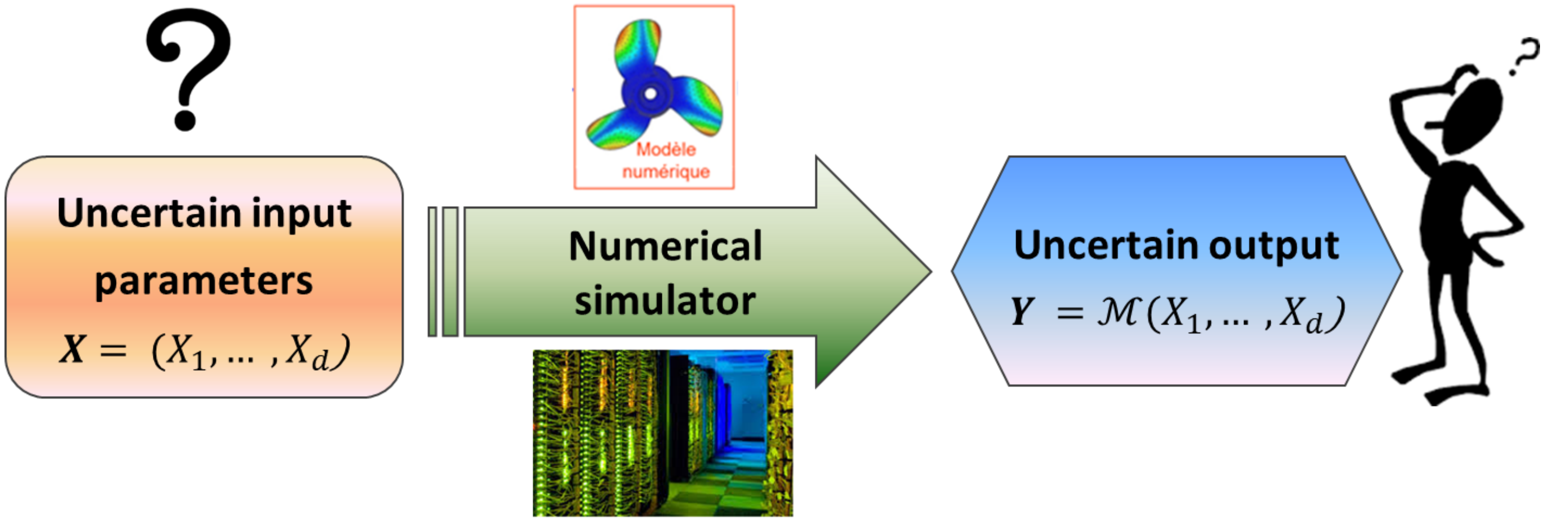
BETHSY experimental facility: 3-loop reduced scale model (1/100 in vol., real size in height) of a 900 MWe Framatome pressurized water reactor (PWR)

2/ Data

Thermal-hydraulics variables, physical properties and coefficients, ...

Risk assessment in nuclear accident analysis

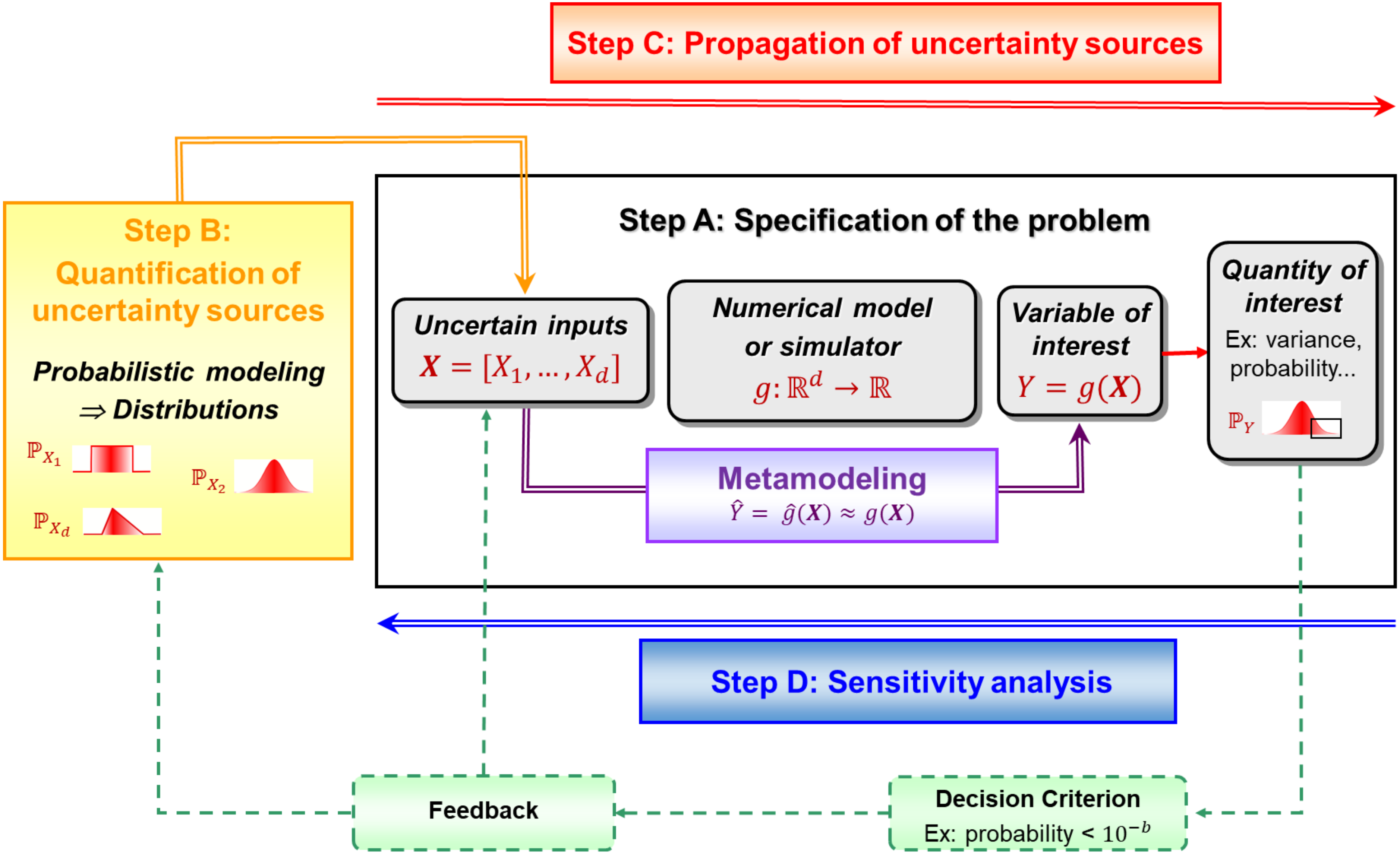
- **Safety studies:** compute a failure risk (margins, rare events) and prioritize the risk indicators, with validated computer/numerical models
- **Numerical simulators:** fundamental tools to understand, model & predict physical phenomena.
- **Large number of input parameters**, characterizing the studied phenomenon or related to its physical and numerical modelling.
- **Uncertainty on some input parameters** → impacts the **uncertainty on the output, the evaluation of safety margins**
- **BEPU (Best Estimate Plus Uncertainties): realistic models & uncertain inputs** → Better assessment of the real margins



- **How to deal with uncertainties in numerical simulation?**
  - → **Probabilistic** framework and **statistical methods**
  - → **Monte Carlo-**based approaches and data analysis ⇒ **Data Sciences**
  - → Essential use of **machine learning**

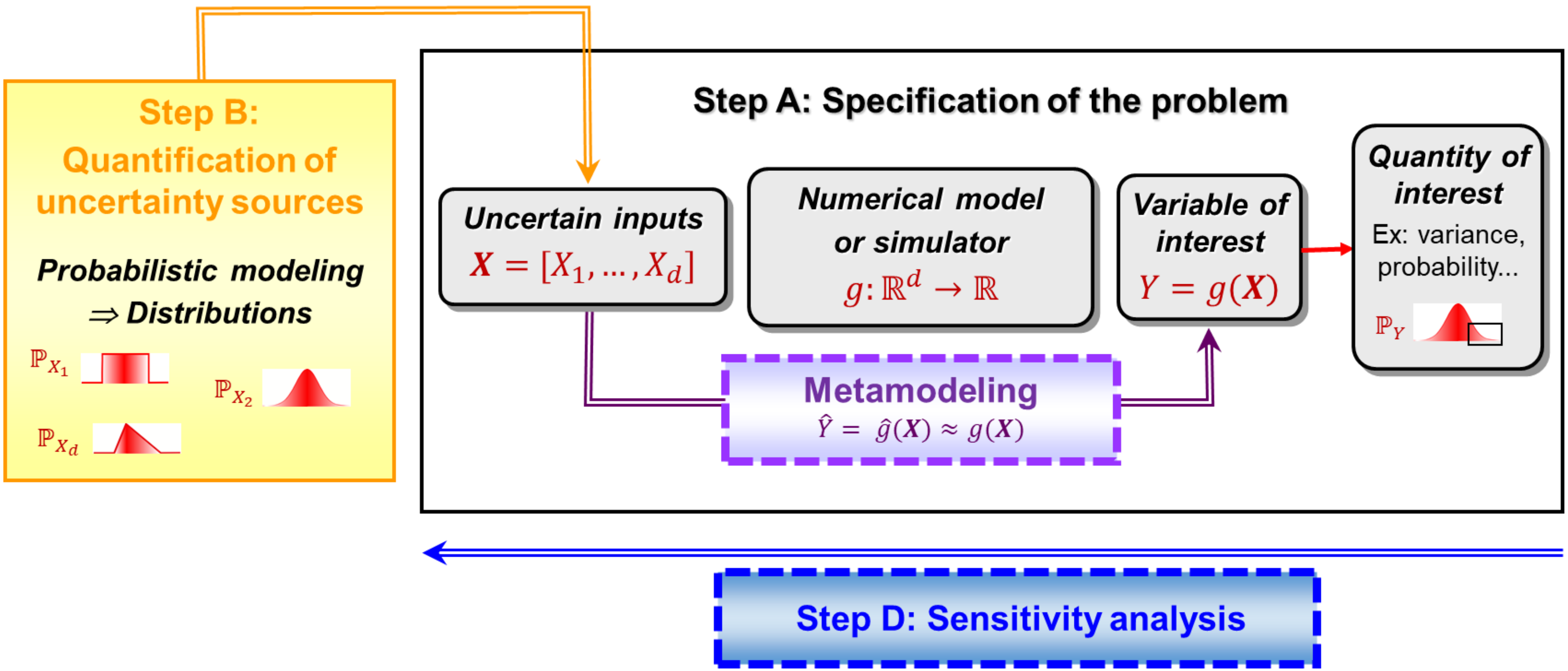
- ***Data-driven* methods in support of physical modeling, analysis and forecasting**
  - → To **propagate the uncertainties** of the inputs
  - → Assess their impact on the simulator predictions
  - → Estimate **probabilities of failure, quantiles, safety margins**
  - → Identify the most influential uncertain inputs: **sensitivity analysis**
  - → **Calibrate** modeling parameters & input uncertainty w.r.t. experimental results
  - → **Validate** the numerical simulator accuracy w.r.t. experimental results
  - → Identify optimal configurations

General uncertainty quantification methodology



Extracted and modified from *De Rocquigny et al. (2008)*

General uncertainty quantification methodology



Recent advances in Sensitivity Analysis

$\Rightarrow$ Focus on HSIC measures

Sensitivity Analysis (SA) in Uncertainty treatment

⇒ Quantify how the variability of the input parameters influences the output
→ Aim of **Sensitivity Analysis (SA)**

- **Quantitative SA and ranking purpose:**
  - Quantify the impact of each uncertain input and interaction → Ranking

  → Identify the variables to be fixed or further characterized in order to obtain the largest reduction of the output uncertainty

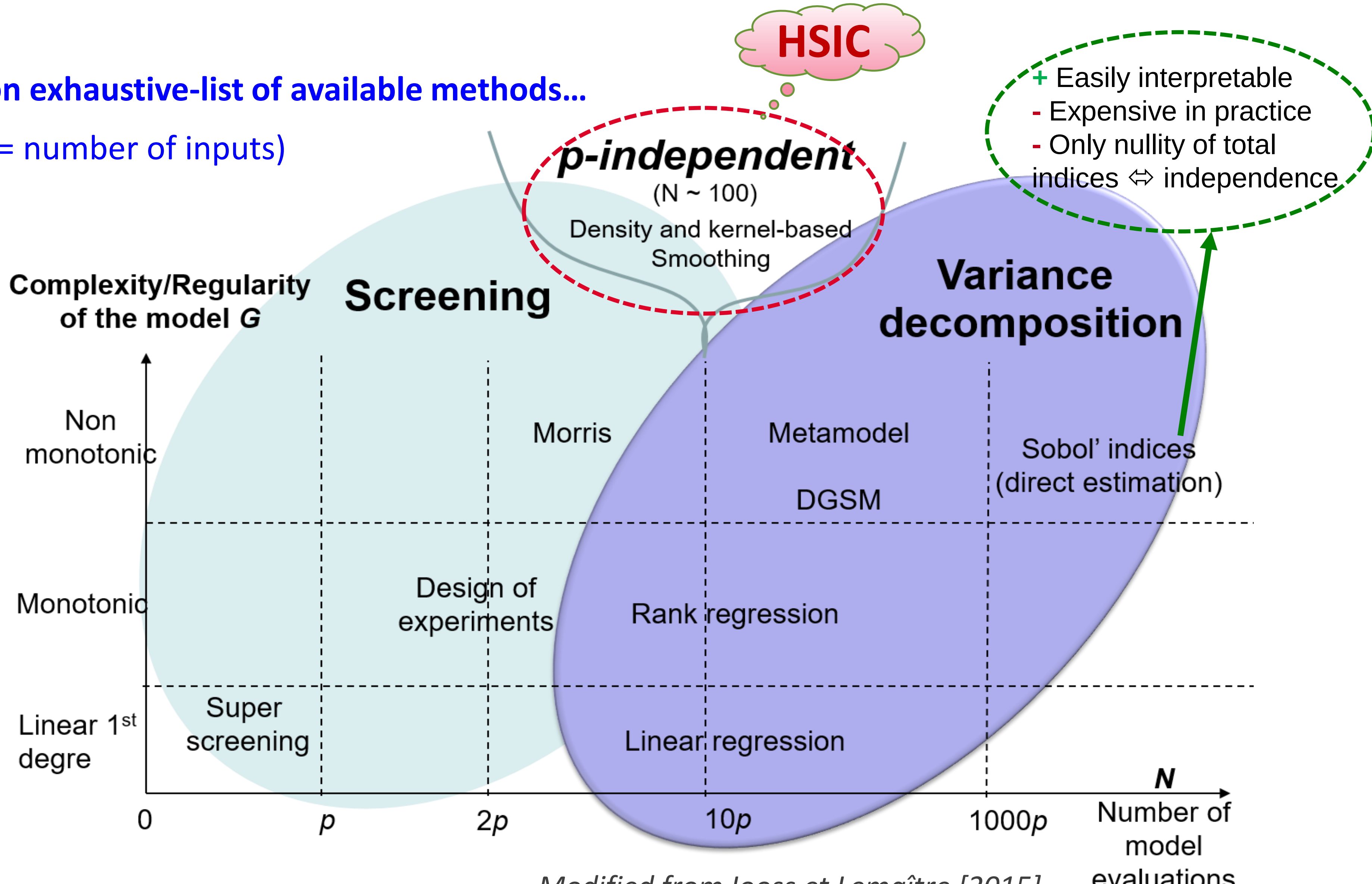
- **Screening purpose. Separate the inputs into two groups: influential and non-influential**
  - Non-influential variables fixed without consequences on the output uncertainty
  - In support of model reduction
  - **To build a simplified model, a metamodel ⇒ ICSCREAM metholodogy**

Global SA within a probabilistic framework

→ Valuable information to understand G and underlying phenomenon

Non exhaustive-list of available methods…

(p = number of inputs)



Modified from Iooss et Lemaître [2015]

Black-box model

$$\boldsymbol{Y} = \mathcal{M}(X_1, \dots, X_d)$$

- $X_1, \dots, X_d$ are *d* independent inputs, evolving in domain $\mathcal{X}_1, \dots, \mathcal{X}_d$
- $\boldsymbol{Y}$ evolves in domain $\mathcal{Y}$
- $\boldsymbol{P_X}$ denotes the probability distribution of X
- $\boldsymbol{P_{X,Y}}$: the joint probability measure and $\boldsymbol{P_Y} \otimes \boldsymbol{P_X}$ the product of marginal distributions

Only a *n*-sample of simulations is available

$\mathcal{M}$ unknown, only Monte-Carlo sample $\left(\boldsymbol{X}^{(j)}, \boldsymbol{Y}^{(j)}\right)_{1 \le j \le n}$ where $Y^{(j)} = \mathcal{M}\left(X^{(j)}\right)$

HSIC review: a kernel-based GSA method

▶ **How to evaluate the sensitivity in a probabilistic way? ⇔ <u>Independence</u>**

→ **By comparing $P_{X_i Y}$ with $P_{X_i} \otimes P_Y$**

$$S_i = d\left(P_{X_i Y}, P_{X_i} \otimes P_Y\right)$$

where *d* a dissimilarity measure between two probablity distributions

d can be based on **Maximum Mean Discrepancy**:

$$\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) = \sup_{f \in \mathcal{H}} \left[\mathbb{E}_{\mathbb{P}} f(Y) - \mathbb{E}_{\mathbb{Q}} f(Y)\right]$$

With $\mathcal{H}$= **<u>unit ball</u> in a (characteristic) <u>RKHS</u> (Reproducing Kernel Hilbert Space)**

Sriperumbudur et al. [2008]

$$\Rightarrow \mathrm{S_i} = MMD^2\left(P_{X_i Y}, P_{X_i} \otimes P_Y\right) = HSIC(X_i, Y)$$

Hilbert-Schmidt Independence Criterion

▶ MMD² applied between $P_{X_i\,Y}$ and $P_{X_i} \otimes P_Y \Rightarrow HSIC(X_i, Y)_{\mathcal{H}_{X_i}, \mathcal{H}_Y}$

$\mathcal{H}_{X_i}$ and $\mathcal{H}_Y$ **RKHS** associated to X_i and Y, resp :

Kernel $k_{X_i}: \mathcal{X}_i \times \mathcal{X}_i \to \mathbb{R}$ with feature space $\mathcal{H}_{X_i}$ and feature map φ_{X_i} (not unique)

Kernel $k_Y: \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ with feature space $\mathcal{H}_Y$ and feature map φ_Y

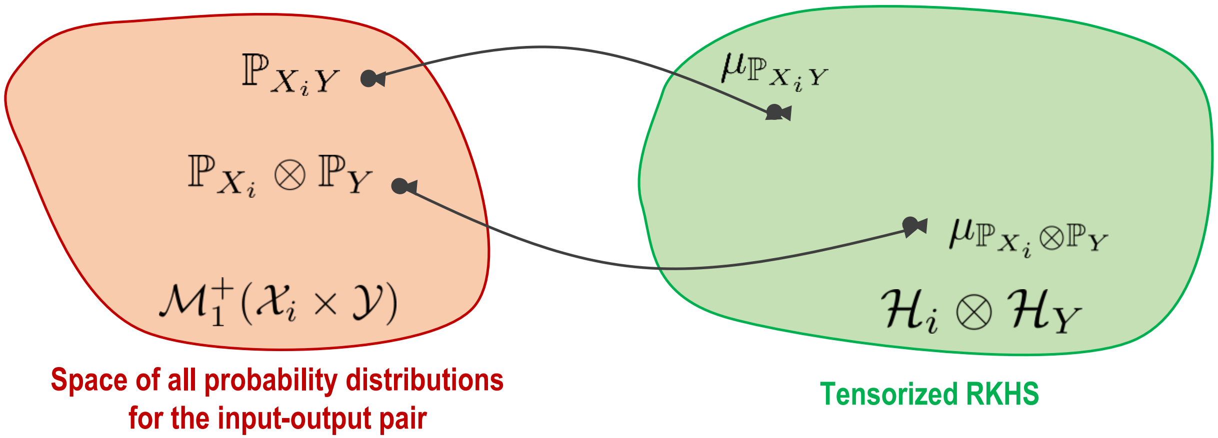
$K_{X_i}(x, x') = \langle \varphi_{X_i}(x), \varphi_{X_i}(x') \rangle_{\mathcal{H}_{X_i}}$ and $K_Y(y, y') = \langle \varphi_Y(y), \varphi_Y(y') \rangle_{\mathcal{H}_Y}$

kernel defines the inner product in the RKHS

Kernel embedding of a distribution $\mathbb{P}_Z$ into RKHS with kernel K_Z :

$$\mu_{\mathbb{P}_Z}(z) = \mathbb{E}_{Z \sim \mathbb{P}_Z}[K_Z(Z, z)] = \langle \mu_{\mathbb{P}_Z}, K_Z(.\,, z) \rangle_{\mathcal{H}_Z}$$

Muandet et al. [2017]



Picture extracted from G. Sarazin's (CEA) slides

HSIC review: a kernel-based GSA method

▶ MMD² applied between $P_{X_i Y}$ and $P_{X_i} \otimes P_Y \Rightarrow HSIC(X_i, Y)_{\mathcal{H}_{X_i}, \mathcal{H}_Y}$

$\mathcal{H}_{X_i}$ and $\mathcal{H}_Y$ **RKHS** associated to X_i and Y, resp :

Kernel $k_{X_i}: \mathcal{X}_i \times \mathcal{X}_i \to \mathbb{R}$ with feature space $\mathcal{H}_{X_i}$ and feature map φ_{X_i}

Kernel $k_Y: \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ with feature space $\mathcal{H}_Y$ and feature map φ_Y

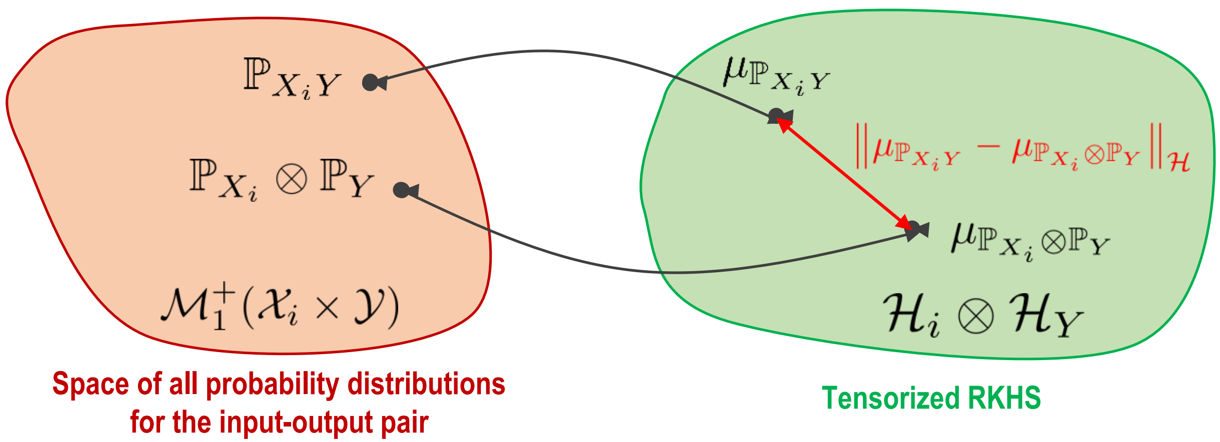
$K_{X_i}(x, x') = \langle \varphi_{X_i}(x), \varphi_{X_i}(x') \rangle_{\mathcal{H}_{X_i}}$ and $K_Y(y, y') = \langle \varphi_Y(y), \varphi_Y(y') \rangle_{\mathcal{H}_Y}$

kernel defines the inner product in the RKHS

HSIC = distance in the RKHS between the images of the two distributions of interest

$$\Rightarrow HSIC(X_i, Y)_{\mathcal{H}_{X_i}, \mathcal{H}_Y} = MMD^2_{\mathcal{H}_{X_i}, \mathcal{H}_Y}(P_{X_i Y}, P_{X_i} \otimes P_Y) = \left\| \mu_{\mathbb{P}_{X_i Y}} - \mu_{\mathbb{P}_{X_i} \otimes \mathbb{P}_Y} \right\|^2_{\mathcal{H}_{X_i}, \mathcal{H}_Y}$$

Gretton et al. [2005]



Extracted from G. Sarazin's (CEA) slides

HSIC review: a kernel-based GSA method

▶ MMD² applied between $P_{X_i\,Y}$ and $P_{X_i}\otimes P_Y \Rightarrow HSIC(X_i, Y)_{\mathcal{H}_{X_i},\mathcal{H}_Y}$

$\mathcal{H}_{X_i}$ and $\mathcal{H}_Y$ **RKHS** associated to X_i and Y, resp :

Kernel $k_{X_i}: \mathcal{X}_i \times \mathcal{X}_i \to \mathbb{R}$ with feature space $\mathcal{H}_{X_i}$ and feature map φ_{X_i}

Kernel $k_Y: \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ with feature space $\mathcal{H}_Y$ and feature map φ_Y

$K_{X_i}(x, x') = \langle \varphi_{X_i}(x), \varphi_{X_i}(x') \rangle_{\mathcal{H}_{X_i}}$ and $K_Y(y, y') = \langle \varphi_Y(y), \varphi_Y(y') \rangle_{\mathcal{H}_Y}$

kernel defines the inner product in the RKHS

HSIC = distance in the RKHS between the images of the two distributions of interest

$$\Rightarrow HSIC(X_i, Y)_{\mathcal{H}_{X_i},\mathcal{H}_Y} = MMD^2_{\mathcal{H}_{X_i},\mathcal{H}_Y}\left(P_{X_i\,Y}, P_{X_i}\otimes P_Y\right) = \left\| C_{X,Y} \right\|^2_{HS}$$

With $C_{X,Y}$ **the covariance operator** between features maps:

$$C_{X,Y} = \mathbb{E}_{X,Y}\left[\varphi_X(X)\otimes\varphi_Y(Y)\right] - \mathbb{E}_X\left[\varphi_X(X)\right]\otimes\mathbb{E}_Y\left[\varphi_Y(Y)\right]$$

*HSIC **"summarizes" the cross-cov** between **feature maps***

*⇒ **Large panel of input-output dependency** can be captured.*

Gretton et al. [2005]

HSIC review: a kernel-based GSA method

► **Characteristic kernels and RKHS** ⇨ *Injective canonical feature map*

⇨ **Equivalence to independence:** $HSIC(X, Y) = 0 \Leftrightarrow X \perp Y$

Ex: Gaussian Kernel

$$k(x_i, x_i') = exp\left(-\frac{(x_i - x_i')^2}{2\lambda^2}\right)$$

► **Estimation:** **Kernel Trick** ⇒ *Feature map linked to kernel function*

Very simple M-C estimator from a *n*-sample of simulations $\left(X_i^{(j)}, Y^{(j)}\right)_{1\leq j\leq n}$

$$\widehat{\text{HSIC}}(X_i, Y) = \frac{1}{n-1} Tr(K_i HLH)$$

where $H = I_n - \frac{1}{n}$, $K_i = \left(k_i\left(X_i^{(j)}, X_i^{(j')}\right)\right)_{1\leq j,j'\leq n}$ and $L = \left(k\left(Y^{(j)}, Y^{(j')}\right)\right)_{1\leq j,j'\leq n}$

► **Statistical Properties:**

- Asymptotically unbiased, variance of order *O*(1/n)
- If $X \perp Y$, $n\widehat{\text{HSIC}}(X, Y)$ converges asymptotically to a Gamma distribution

▶ Normalization for sensitivity analysis:

Da Veiga [2015]

$$R^2_{HSIC} = \frac{HSIC(X,Y)}{\sqrt{HSIC(X,X)HSIC(Y,Y)}}$$

⇨ $R^2_{HSIC} \in [0,1]$ for easier interpretation

$$\textbf{Influence}(X_{[1]}) > \textbf{Influence}(X_{[2]}) > \cdots > \textbf{Influence}(X_{[d]})$$

Where order $[\cdot]$ is such that $\hat{R}^2_{H,X_{[1]}} > \hat{R}^2_{H,X_{[2]}} > \cdots > \hat{R}^2_{H,X_{[d]}}$

⇒ Use for ranking of inputs

▶ Independence tests: $HSIC(X,Y) = 0 \Leftrightarrow X \perp Y$ (*with characteristic kernels!*)

- Null hypothesis: $\mathcal{H}_0 : X \perp Y$ against $\mathcal{H}_1 : X \not\perp Y$
- Test statistics: $n\widehat{\text{HSIC}}(X,Y)$
- Decision rule: $\mathcal{H}_0$ **rejected iff** $n\widehat{\text{HSIC}}(X,Y) > q_{1-\alpha}$

where $q_{1-\alpha}$ is the $(1-\alpha)$quantile of $n\widehat{\text{HSIC}}(X,Y)$ under $\mathcal{H}_0$

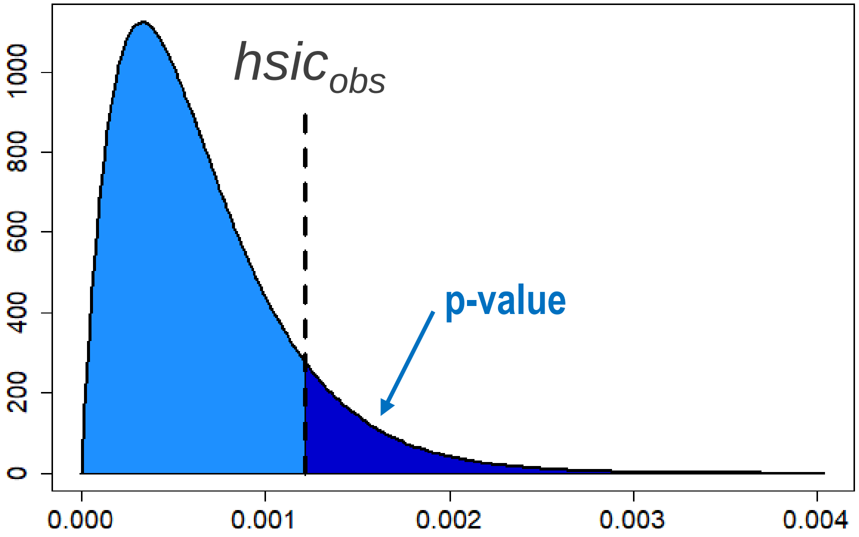
⇒ Use for screening of inputs

HSIC-based independence tests for screening

How to have the distribution $n\widehat{\text{HSIC}}(X_i, Y)$ under $\mathcal{H}_0$ to compute *p-value*?

- **If *n* large: asymptotic test based** on approximation with Gamma distribution *(Gretton et al. (2008])*
- **If *n* small: Permutation-based** approximation *(De Lozzo & Marrel (2016a], Meynaoui [2019], El Amri & Marrel [2021a])*

Gamma distribution



P-value = $Pr\ [\ \widehat{\text{HSIC}}(X_i, Y) > hsic_{obs}\ ]$

<u>Interpretation of *p-value* for a level α ($\alpha = 5\%$ or 10%) for screening:</u>

- **<u>pval < α</u>** ⇒ H_0 (Independence) rejected ⇒ X_i **is significantly influential**

HSIC as indices of Sensitivity Analysis

- Focus the SA analysis on the difference between $P_{X,Y}$ with $P_X \otimes P_Y$
- Power of RKHS → HSIC=one of the most successful non-parametric dependence measure
- Capture a **large spectrum of relationships**
- Able to deal with **many types of variables** and **purposes**:
  - **Goal-oriented SA for safety studies** *(Marrel & Chabridon [2021], Iooss & Marrel[2019]) :* To measure the input influence in a restricted output domain: $Y \in \mathcal{C}$

    ⇒ Numerous applications for safety and risk assessment ($\mathcal{C}$: critical safety domain, e.g. $\mathcal{C} = \{Y | Y > \text{critical value}\}$)
  - **SA of multivariate or functional output (or inputs)** ⇒ **definition of specific kernels** Atmospheric dispersion model with spatio-temporal output *(De Lozzo & Marrel [2016b]),* Dynamic compartmental epidemiologic model on COVID-19 *(El Amri & Marrel [2021b])*

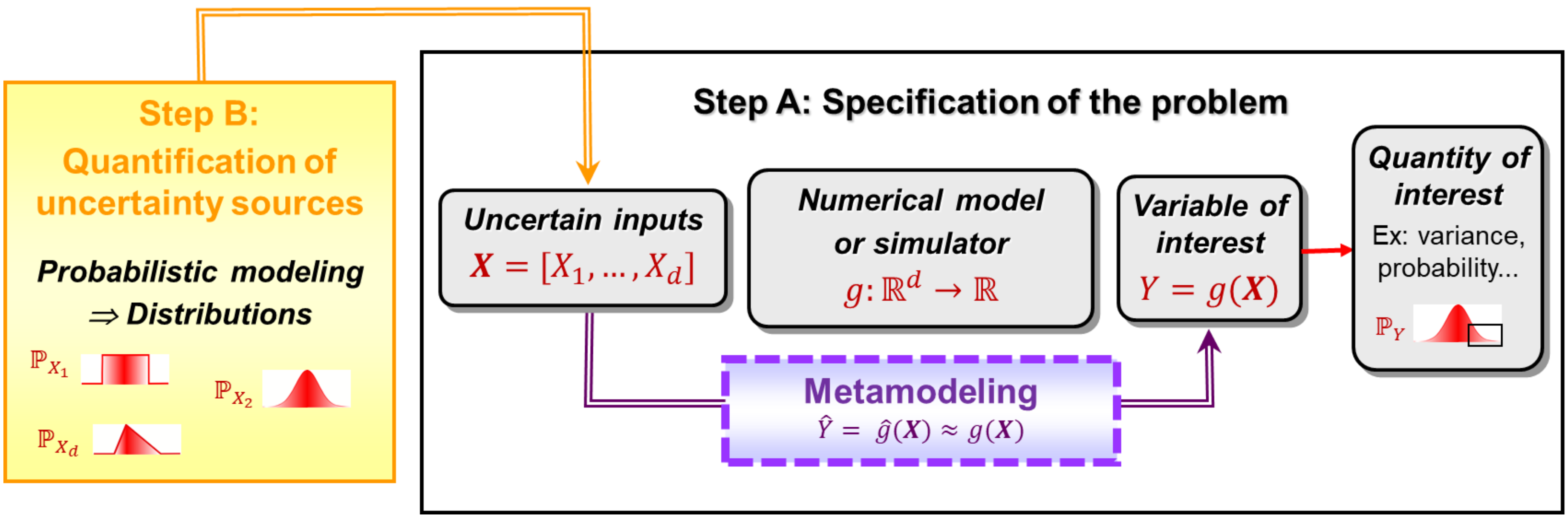
Efficiency demonstrated in numerous industrial applications, especially with small sample size *n* and large dimension *d*

- **Use in support of metamodeling in large dimension** → ICSCREAM Methodology

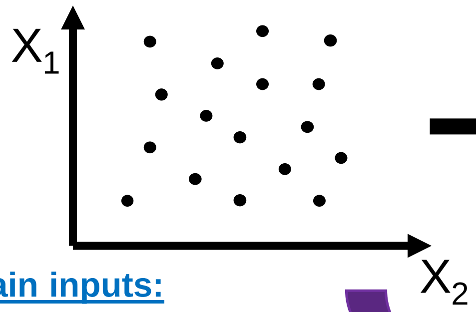
Recent advances in Metamodeling

⇒ Focus on Gaussian Process (GP) Regression

General uncertainty quantification methodology



Design of numerical experiments → **Numerical simulations** → **Analysis of outputs**



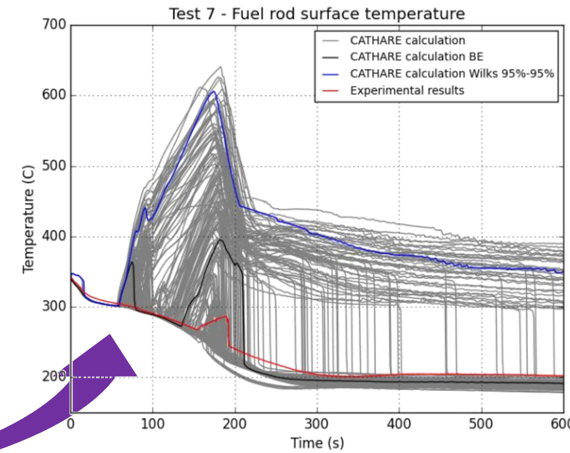
Incertain inputs:

- Environmental variables
- Physical parameters
- Modeling and numerical parameters

Simulator

$Y = G(X_1, \ldots, X_d)$

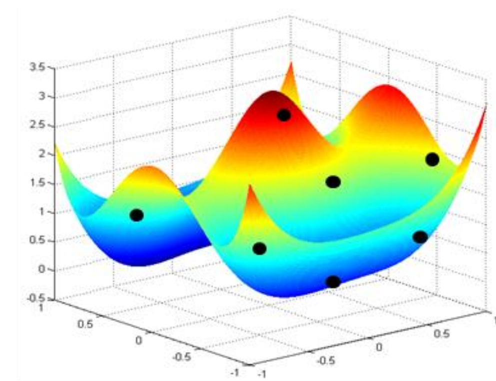
In case of costly *G*:
Model reduction or Approximation with Machine learning (ML) **metamodel** $Z_{app} \approx g(\boldsymbol{X})$



Thermal-hydraulics transients for accidental B-LOCA scenario

- ✓ With good approximation and prediction capabilities ⇒ to be controled
- ✓ With a negligible cpu cost for prediction
- ✓ Built from a Monte Carlo sample of n experiments ($n \sim 10\ d$)

<u>*Ex :*</u> *Polynomials, splines, neural networks, regression trees…*

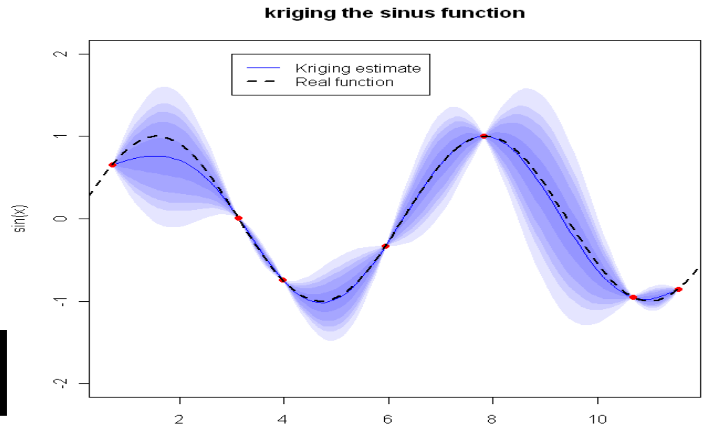


Choice: Gaussian process (GP) metamodel

see *Rasmussen & Williams [2005]*

Part of *Supervised Machine Learning*

Advantage: gives a prediction with an associated error bound (Gaussian distribution at each point)



→ How to build the GP in large dimension?

→ How to build the GP for chaotic code?

→ How to build the GP for functional or other type of data?

→ Integration of physical constraints? *cf. Bachoc's talk*

Crucial use of metamodel (*machine learning*)

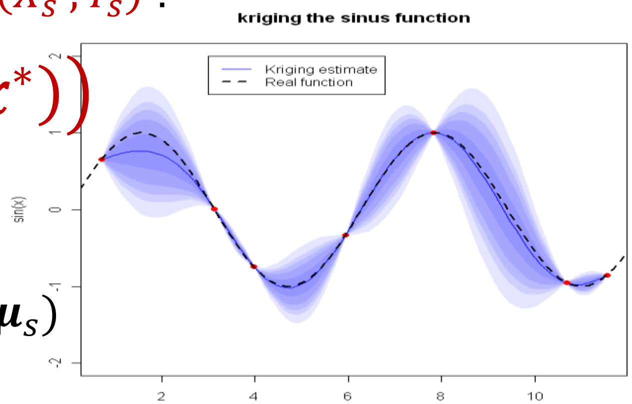
✓ Kernel-based method of supervised learning from (X_s, Y_s). Response is considered as a realization of a random GP field:

$$Y(\boldsymbol{x}) \sim GP(\mu(\boldsymbol{x}), k(\boldsymbol{x}', \boldsymbol{x}))$$

With $\mu(\boldsymbol{x})$ the mean and $k(\boldsymbol{x}', \boldsymbol{x})$ the covariance function.

⇒ Predictive GP is the GP conditioned by the observations (X_s, Y_s) :

$$Y(\boldsymbol{x}^*)_{|Y(X_s)=Y_s} \sim GP\left(\hat{\mu}(\boldsymbol{x}^*), \hat{s}(\boldsymbol{x}', \boldsymbol{x}^*)\right)$$



With

- $\hat{\mu}(\boldsymbol{x}^*) = \mathrm{E}[Y(\boldsymbol{x}^*)|Y(X_s) = \boldsymbol{Y}_s] = \mu(\boldsymbol{x}^*) + k^T_{X_s,\boldsymbol{x}^*} K^{-1}_{X_s,X_s} (\boldsymbol{Y}_s - \boldsymbol{\mu}_s)$
- $\hat{s}(\boldsymbol{x}', \boldsymbol{x}^*) = \mathrm{Cov}[Y(\boldsymbol{x}^*)|Y(X_s) = Y_s] = k^T_{X_s,\boldsymbol{x}^*} K^{-1}_{X_s,X_s} k_{X_s,\boldsymbol{x}^*}$

where $\boldsymbol{\mu}_s$ corresponds to μ evaluated at X_s, $k_{X_s,\boldsymbol{x}^*}$ the covariance between X_s and $\boldsymbol{x}^*$ and K_{X_s,X_s} the covariance matrix for X_s

⇒ Conditional mean $\hat{\mu}(\boldsymbol{x}^*)$ serves as the predictor at location $\boldsymbol{x}^*$

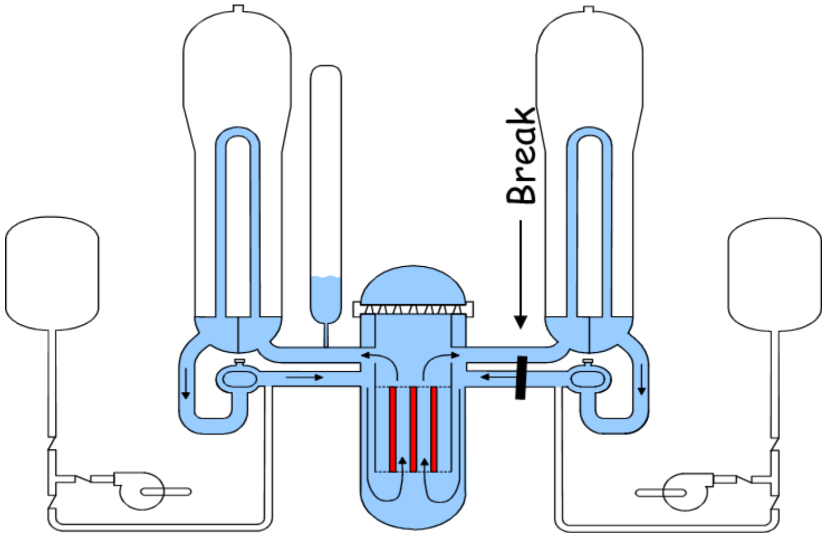
⇒ Prediction variance (*i.e.* mean squared error) given by conditional covariance $\hat{s}(\boldsymbol{x}^*, \boldsymbol{x}^*)$

Illustration: the ICSCREAM methodology for the IB-LOCA nuclear accident

Simulation of IB-LOCA nuclear accident

Accidental scenario on pressurized water reactor: IB-LOCA

LOss of primary **C**oolant **A**ccident due to a **I**ntermediate **B**reak in cold leg

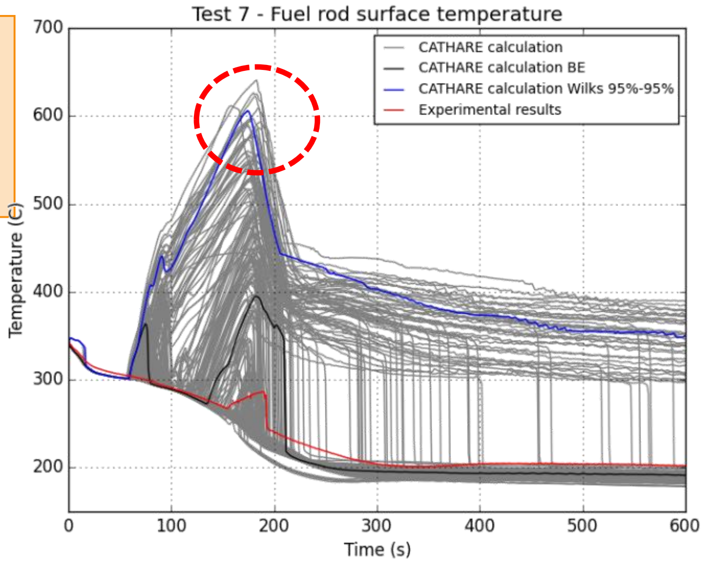


***d* (~ 100) input random variables:**

Critical flowrates, initial/boundary conditions, phys. eq. coef., …

Modelled with **CATHARE2 code**:

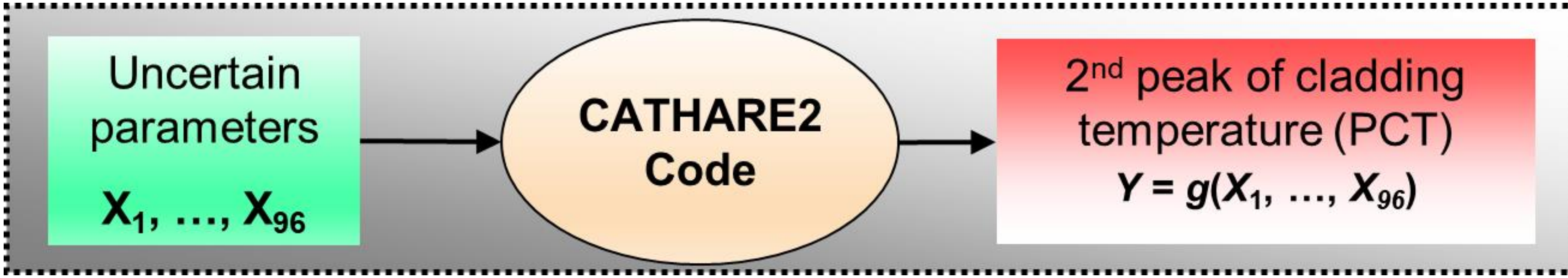
- Models complex thermal-hydraulic phenomena
- **Large CPU cost for one code run (> 1 hour)**



Variable of Interest:

2^{nd} peak of cladding temperature (PCT) = scalar output

Objective in IB-LOCA safety study



- In IB-LOCA modeling framework, uncertain input parameters are:
  - (Type 1) Initial conditions, physical model parameters ⇒ Probabilistic ($\mathcal{U}, \mathcal{LU}, \mathcal{N}, \mathcal{LN}$)
  - (Type 2) **Scenario parameters (min / max bounds)** ⇒ **No probabilistic**

Objective in support of safety studies

Identify the most **penalizing configurations** for **Type 2** inputs, under the uncertainties of **type 1** inputs.

Penalizing configurations ⇔ leading to high PCT values

IB-LOCA: *Intermediate Break LOss of Coolant Accident*

Objective in IB-LOCA safety study

Problems & constraints

- **Very large number of inputs (~100),** but ***effective*** dimension might be lower
- Each CATHARE simulation ~ 1 hour ⇒ around 1000 simulations available
- Phenomena involved are complex **with strong non-linearities**
- Black-box model: intrusive methods not possible

⇨ **Monte Carlo sampling + advanced statistical tools for data analysis**

- ✓ **Screening and sensitivity analysis**
- ✓ **Approximation with metamodel**
- ✓ **Uncertainty propagation**

⇒ **Adapted to VERY HIGH DIMENSIONAL test case** (~100 uncertain inputs)

⇒ **ICSCREAM* methodology in 4 Main Steps**

Identification of penalizing **C**onfigurations using **SCRE**ening **A**nd **M**etamodel*

STEP 1: Monte Carlo Sampling design

Uncertainty quantification of uncertain inputs + scenario inputs to be penalized X_{pen}

Uncertain inputs $\boldsymbol{X} = (X_1, \dots, X_{d'})$ with probability distributions + scenario inputs $\boldsymbol{X}_{pen}$

↓

Step 1: Learning sample of *n* simulations $(\boldsymbol{X}_S, \boldsymbol{Y}_S)$

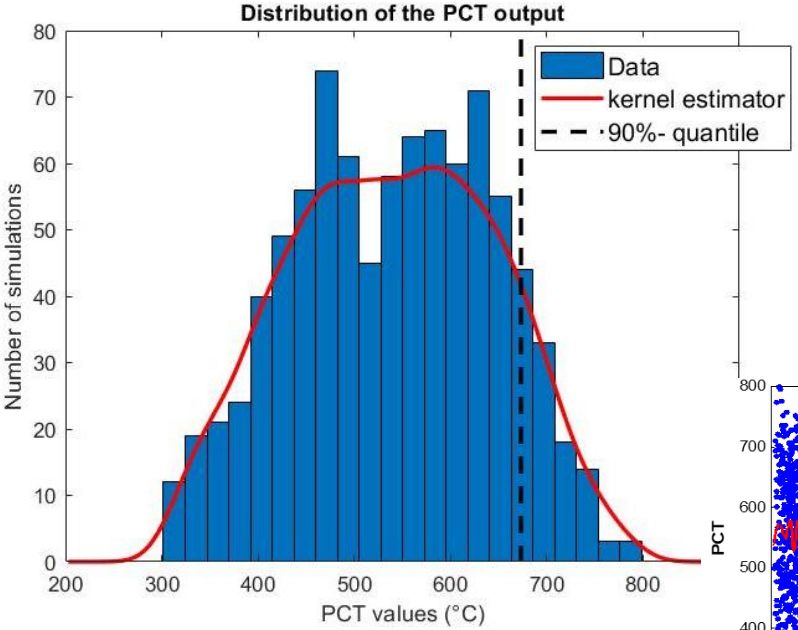
Monte-Carlo design of *n* experiments $\boldsymbol{X}_S = \{\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(n)}\}$ and associated CATHARE2 PCT outputs $\boldsymbol{Y}_S$

STEP 1: Monte Carlo Sampling design
Illustration on the IB-LOCA test case

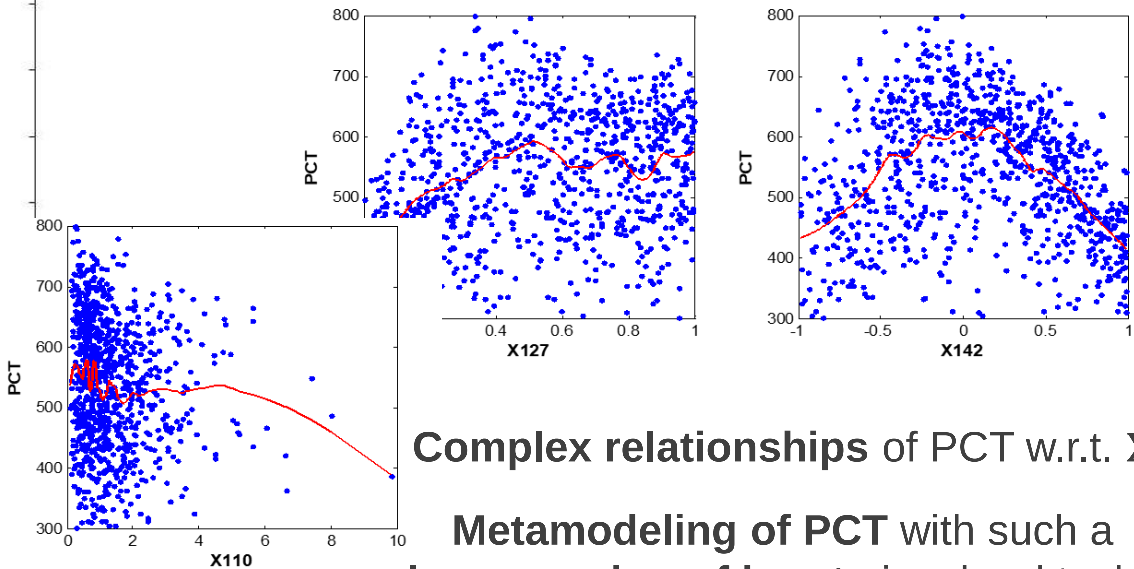
- ***d* = 96 uncertain variables** with probability distributions $\mathbb{P}_X$ (almost indep.)
- ***n* = 889 CATHARE2 simulations** : Monte-Carlo sample ($\mathbf{X} \sim \mathbb{P}_X$)

Empirical quantile 90%: $q_{0.9} \approx$ **673.18 °C**

Critical configurations are defined as: **PCT > $q_{0.9}$**



Scatter plots with 1-D local polynomials for trends



Complex relationships of PCT w.r.t. $\mathbf{X}$

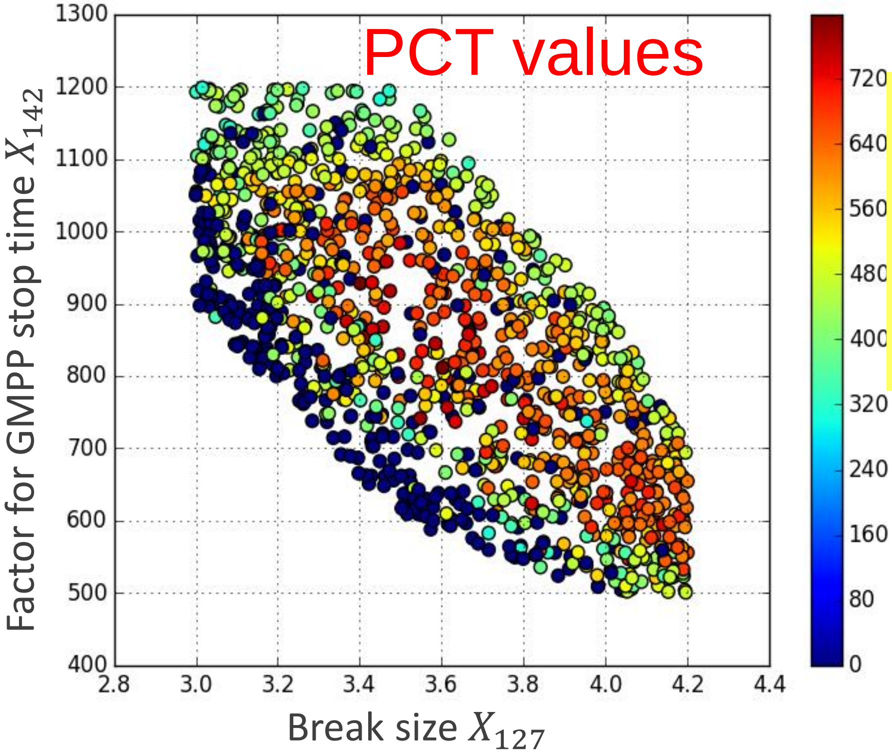
Metamodeling of PCT with such a **large number of inputs** is a hard task!!

STEP 1: Monte Carlo Sampling design

Illustration on the IB-LOCA test case

Among 96 inputs, 2 scenario inputs to be penalized (here dependent):

- X_{127} **(break size):** uniform distribution on [3, 4.2] inches
- X_{142} **(factor for GMPP stop time):** uniform random variable whose range of variation depends on the value of X_{127}



➢ **Objective:**

Precisely capture **critical configurations of (X_{127}, X_{142})** which lead to the highest probability of PCT exceeding $q_{0.9}$ (≈ 673.18 °C)

$$\boldsymbol{X}_{pen} = \{X_{127}, X_{142}\} \subset \boldsymbol{X}$$

GMPP : *group of primary pumps*

STEP 2: Screening & ranking of inputs

Uncertainty quantification of uncertain inputs + scenario inputs to be penalized X_{pen}

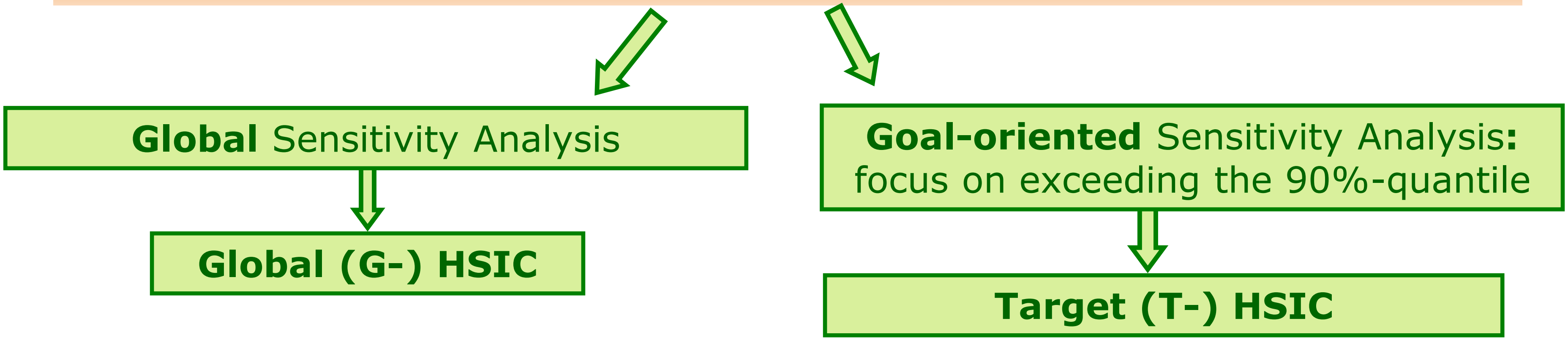
↓

Step 1: Learning sample of *n* simulations (X_S, Y_S)

↓

Step 2: Screening and ranking with HSIC-based independence tests from (X_S, Y_S)

Identify and **rank** the inputs of primary influence with HSIC-based tests

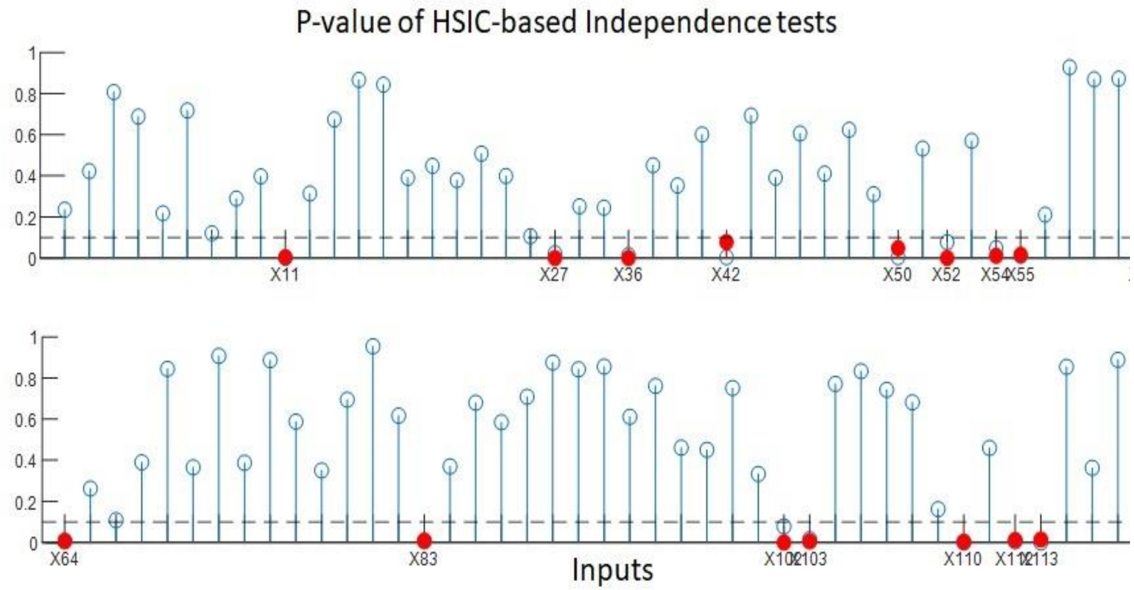


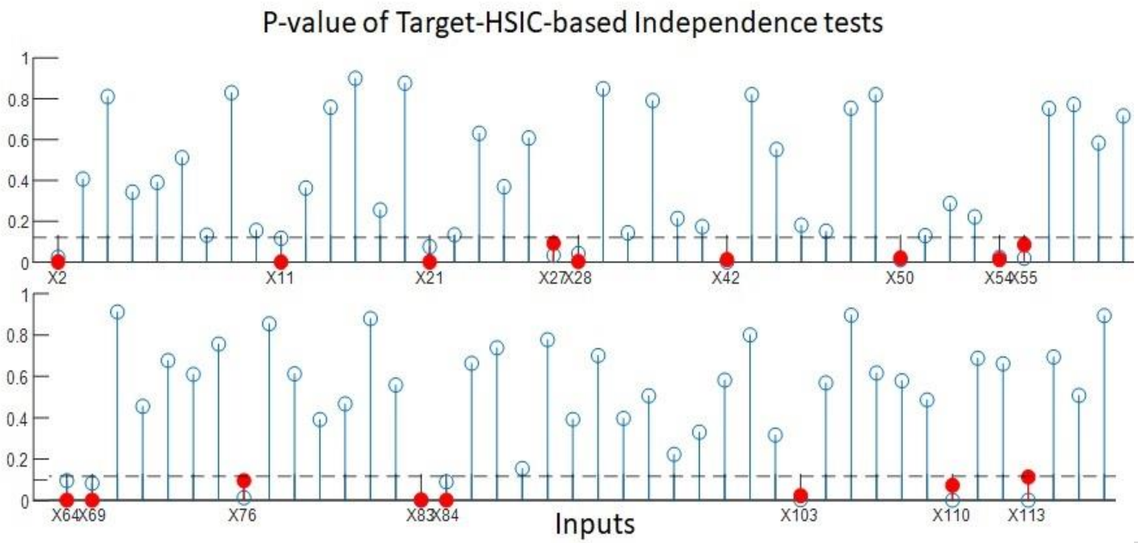
STEP 2: Screening & ranking of inputs

Illustration on the IB-LOCA test case

Global-HSIC tests

~ 18 influential inputs in GSA
Influence ++ : X142 *(GMPP time)*
Influence + : X127 *(break size)*
Influence : X113, X110, X11
Lower influence : X50, X42, X112, X83, X64, X125, X55 , X103, X36, X27, X54, X102, X52





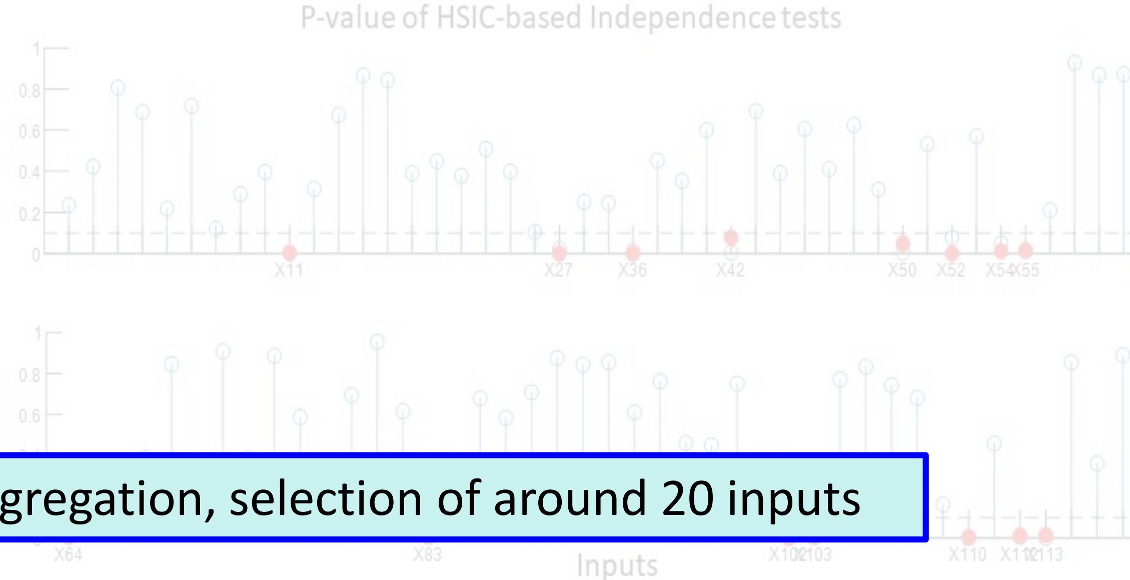
T-HSIC ⇨ Impact on exceeding the 90%-quantile $\hat{q}_{0.9}(Y)$

~ 19 influential inputs in TSA
Influence ++ : X142 *(GMPP time)*
Influence + : X113, X110, X127, X125, X83
Lower influence : X42, X103, X76, X50, X55, X54, X2, X27, X28, X21, X84, X64, X11

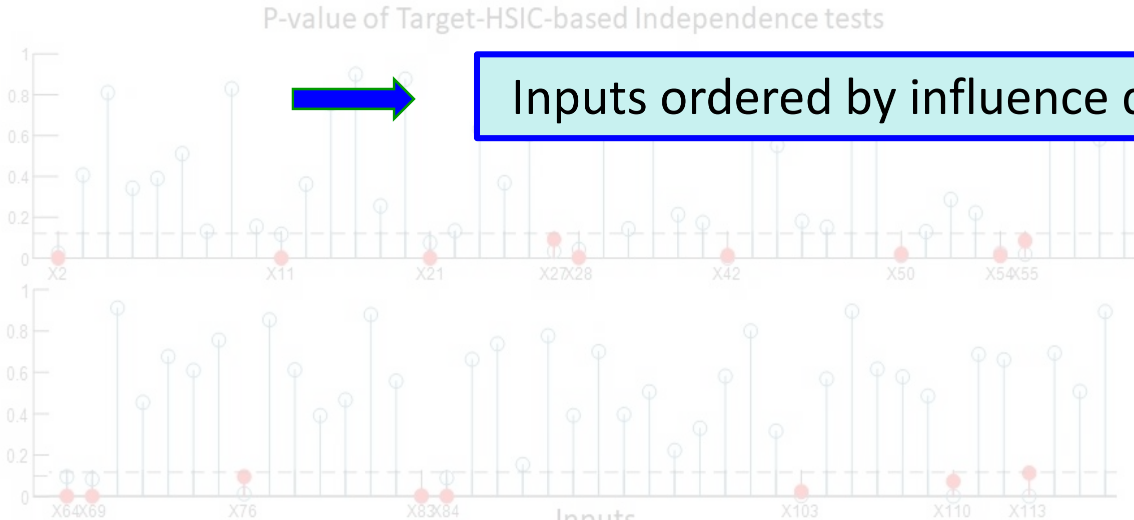
Illustration on the IB-LOCA test case

Global-HSIC tests

~ **18 influential inputs** in **GSA**
Influence ++ : X142
Influence + : X127
Influence : X113, X110, X11
Lower influence : X50, X42, X112, X83, X64, X125, X55 , X103, X36, X27, X54, X102, X52



From aggregation, selection of around 20 inputs



Inputs ordered by influence d°, using *P-values*

...ng the 90%-quantile $\hat{q}_{0.9}(Y)$

~ **19 influential inputs** in **TSA**
Influence ++ : X142
Influence + : X113, X110, X127, X125, X83
Lower influence : X42, X103, X76, X50, X55, X54, X2, X27, X28, X21, X84, X64, X11

Uncertainty quantification of uncertain inputs + scenario inputs to be penalized X_{pen}

↓

Step 1: Learning sample of *n* simulations (X_S, Y_S)

↓

Step 2: Screening and ranking with HSIC and T-HSIC independence tests from $(\boldsymbol{X_S}, \boldsymbol{Y_S})$

↓

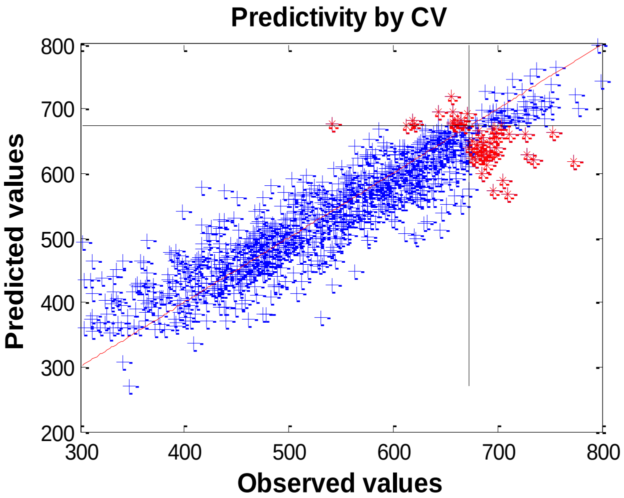
Step 3: Sequential Metamodeling → Gaussian process (GP) regression from $(\boldsymbol{X_S}, \boldsymbol{Y_S})$

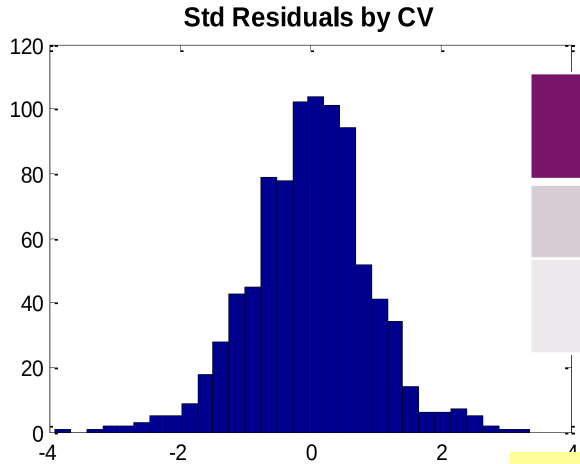
→ Challenge to be addressed here: how to build the GP in large dimension (d~100) ?

→ Use the information of screening and ranking from HSIC

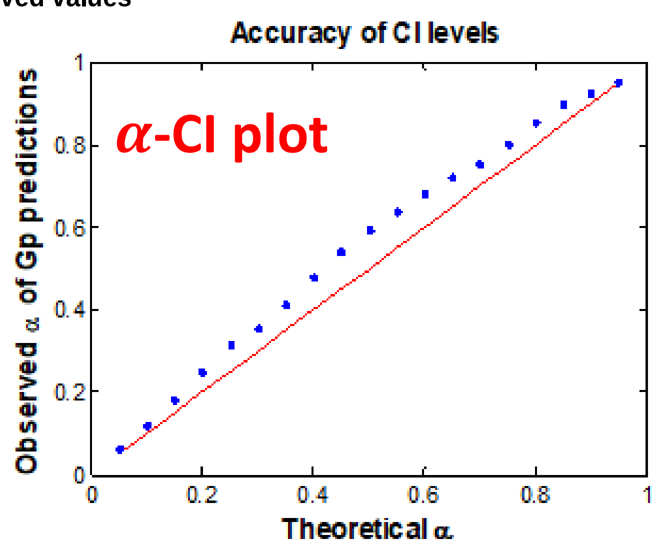
⇒ **Sequential estimation of GP hyperparameters**

Assessment of accuracy and predictivity of final GP metamodel built on *N = 889* simulations





Accuracy criterion	Sequential GP
Q²	0.82
PVA	0.15



⇒ *Q²* : 82 % of PCT variance explained by the GP built with the 20 selected 96 inputs

⇒ 18% of variance unexplained: inaccuracy of the GP + total effect of the 76 neglected inputs

⇒ **Low PVA and good *α*-CI plot:** accurate confidence intervals in prediction

STEP 4: Uncertainty propagation with the GP

Uncertainty quantification of uncertain inputs + scenario inputs to be penalized X_{pen}

↓

Step 1: Learning sample of *n* simulations $(\boldsymbol{X_S}, \boldsymbol{Y_S})$

↓

Step 2: Screening and ranking with HSIC and T-HSIC independence tests from $(\boldsymbol{X_S}, \boldsymbol{Y_S})$

↓

Step 3: Sequential Metamodeling with Gaussian process (GP) from $(\boldsymbol{X_S}, \boldsymbol{Y_S})$

↓

Step 4: Uncertainty propagation with GP metamodel

⇒ **Identify penalizing values of X_{pen}** under the uncertainty of the other inputs $\{\boldsymbol{X} \backslash \boldsymbol{X}_{pen}\}$

Step 4: Uncertainty propagation with GP metamodel to identify the penalizing values of X_{pen} under the uncertainty of the other inputs $\{X \backslash X_{pen}\}$

⇒ *Precisely capture critical* configurations of $\boldsymbol{X_{pen}} = \{X_{127}, X_{143}\}$ which lead to the highest probability of PCT > $\hat{q}_{0.9}(Y)$ (under randomness of the other variables)

Notations :

- $\boldsymbol{X_{exp}}$ are explanatory inputs of the GP
- $\widetilde{\boldsymbol{X}}_{exp} = \boldsymbol{X_{exp}} \setminus \boldsymbol{X_{pen}}$

$$\hat{P}(\mathbf{X_{pen}}) = P[Y_{Gp}(\mathbf{X_{exp}}) > \hat{q}_{0.9} \,|\mathbf{X_{pen}}]$$

$$= 1 - \int_{\mathscr{X}_{exp}} \Phi\left(\frac{\hat{q}_{0.9} - \hat{Y}_{Gp}(\tilde{\mathbf{x}}_{\mathbf{exp}}, \mathbf{X_{pen}})}{\sqrt{MSE[\hat{Y}_{Gp}(\tilde{\mathbf{x}}_{\mathbf{exp}}, \mathbf{X_{pen}})]}}\right) d\mathbb{P}_{\tilde{\mathbf{X}}_{\mathbf{exp}}}(\tilde{\mathbf{x}}_{\mathbf{exp}})$$

$\widetilde{\boldsymbol{X}}_{exp}$ and $\boldsymbol{X_{pen}}$ are independent (necessary condition)

$\boldsymbol{\Phi}$: CDF of standard Gaussian distribution

Variation domain of $\widetilde{\boldsymbol{X}}_{exp}$

Joint distribution of $\widetilde{\boldsymbol{X}}_{exp}$

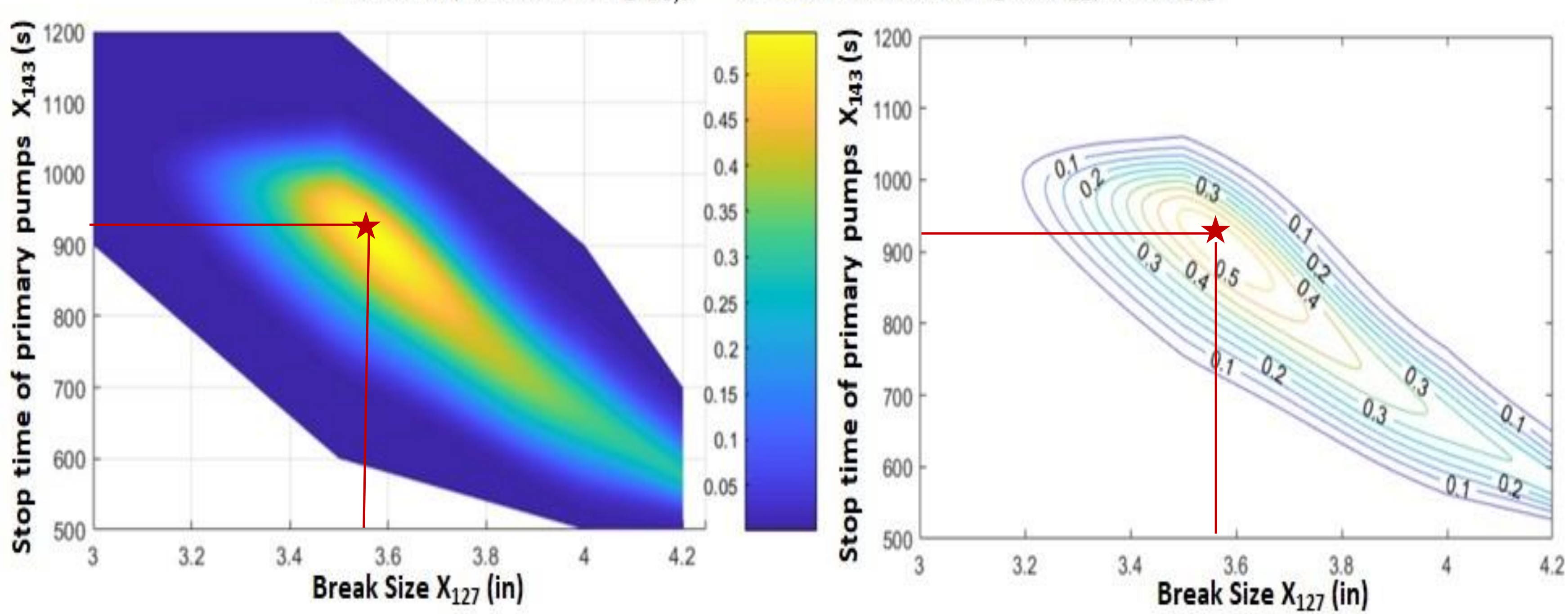
- In practice, for each value of $\boldsymbol{X_{pen}} = \{X_{127}, X_{143}\}$, $\hat{P}(\boldsymbol{X_{pen}})$ is estimated by intensive Monte-Carlo computation (here integral in dimension 18 in the use-case)

STEP 4: Uncertainty propagation with the GP

Illustration on the IB-LOCA test case

Computation of $\hat{P}(\boldsymbol{X_{pen}})$

Probability of exceeding $\hat{q}_{0,9} = 673.18°C$, according to X_{127} and X_{143}



- ► Strong interaction between the two scenario parameters
- ► Worst case: (3.57 inches, 907.8 seconds) ⇒ $\hat{P} \approx 0.55$
- ► Physical explanation: these two parameters drive the degradation of the water inventory
  - The smaller X_{127}, the longer the pump will have to run for the same inventory degradation
  - If $X_{127} < 3.3$ ⇒ the water inventory does not degrade too much (whatever GMPP)
  - If $X_{127} > 3.9$ ⇒ break tends to be prevailing and reduces the impact of stop time of GMPP

Some axes of research for uncertainty treatment

- **Model exploration : numerical Designs of Experiments (DoE)**
  - → Space-filling designs for large number of uncertain inputs
  - → How to tackle the curse of dimensionality?
  - → Extension to functional (temporal/spatial) inputs?
  - → Adaptive/sequential DoE (tractability in large dimension)
- **Sensitivity analysis techniques**
  - → Advanced and robust screening (dimension reduction) and ranking techniques
  - → Extension to functional (temporal/spatial) outputs?
  - → Extension to correlated inputs?
- **Metamodeling for large number of uncertain inputs**
  - → How to build accurate and reliable GP metamodel in very large dimension *d* ?
  - → Scalability with large sample size *n ?*
- **Validation/Calibration of model (real experiments vs. calculations)**
  - → Bayesian approaches
  - → Definition of relevant metrics for validation

Some axes of research for uncertainty treatment

Some are notably addressed within the ANR SAMOURAI project

→ Advanced and robust screening and ranking

Decomposition into main effects & interactions must be investigated

⇒ Assess the use of **HSIC with ANOVA-like kernels** (*Da Veiga [2021]*)

⇒ Build associated independence tests

⇒ Relevancy in support of metamodel building

anr agence nationale de la recherche — SAMOURAI

Simulation **A**nalytics and **M**eta-model-based solutions for **O**ptimization, **U**ncertainty and **R**eliability **A**naly**I**s

→ BuildGP in large dimension: improve reliability

⇒ More reliable estimation of hyperparameters

⇒ Bayesian approach and sparse GP

→ Adaptive/sequential DoE

Some References

- De Rocquigny et al. Uncertainty in industrial practice, Wiley, 2008.
- Da Veiga. Global sensitivity analysis with dependence measures, *Journal of Statistical Computation & Simulation*, 2015.
- De Lozzo and Marrel, New improvements in the use of dependence measures for sensitivity analysis and screening, *Journal of Statistical Computation and Simulation*, 86:3038-3058, 2016.
- Demay, Iooss, Le Gratiet, Marrel. Model selection for Gaussian Process regression: applications on the characterization of chemical segregation, *Quality and Reliability Engineering International*, 2021.
- El Amri and Marrel. HSIC-based independence tests with optimal sequential permutations: application for sensitivity analysis of numerical simulators, *Quality and Reliability Engineering International*, 2021.
- Ghanem et al. Handbook of uncertainty quantification. Springer, 2017.
- Gramacy. Gaussian Process Modeling, Design, and Optimization for the Applied Sciences. Chapman and Hall/CRC, 2021.
- Iooss, and Lemaître. A review on global sensitivity analysis methods. *In Uncertainty management in Simulation-Optimization of Complex Systems: Algorithms and Applications*, Springer, 2015.
- Gretton, Bousquet, Smola and Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. *In: Proceedings Algorithmic Learning Theory*, 2015.
- Marrel, Iooss and Chabridon, Statistical identification of penalizing configurations in high-dimensional thermalhydraulic numerical experiments: The ICSCREAM methodology, Nuclear Science & Engineering, 2021.
- Marrel and Chabridon. Statistical developments for target and conditional sensitivity analysis: application on safety studies for nuclear reactor, *Reliability Engineering and System Safety* (214), 2021.
- Mazgaj, Vacher and Carnaveli. Comparison of CATHARE results with the experimental results of cold leg intermediate break LOCA obtained during ROSA-2/LSTF test 7, *EPJN,* 2, 2016.
- Meynaoui. New developments around dependence measures for sensitivity analysis : application to severe accident studies for generation IV reactors. PhD INSA Toulouse, 2019.
- Muandet, Fukumizu, Sriperumbudur and Schölkopf. Kernel mean embedding of distributions: A review and beyond. Foundations and Trends in Machine Learning. 2017
- Rasmussen and Williams. Gaussian processes for machine learning. MIT Press, 2006.