Introduction to Geostatistics - Metamodeling with Gaussian processes

Fabrice Gamboa
Bertrand Iooss

28/02/2013

Main stakes of uncertainty management

- **Modeling phase**
  - Improve the model
  - Explore the best as possible different input combinations
  - Identify the predominant inputs and phenomena in order to priorize R&D

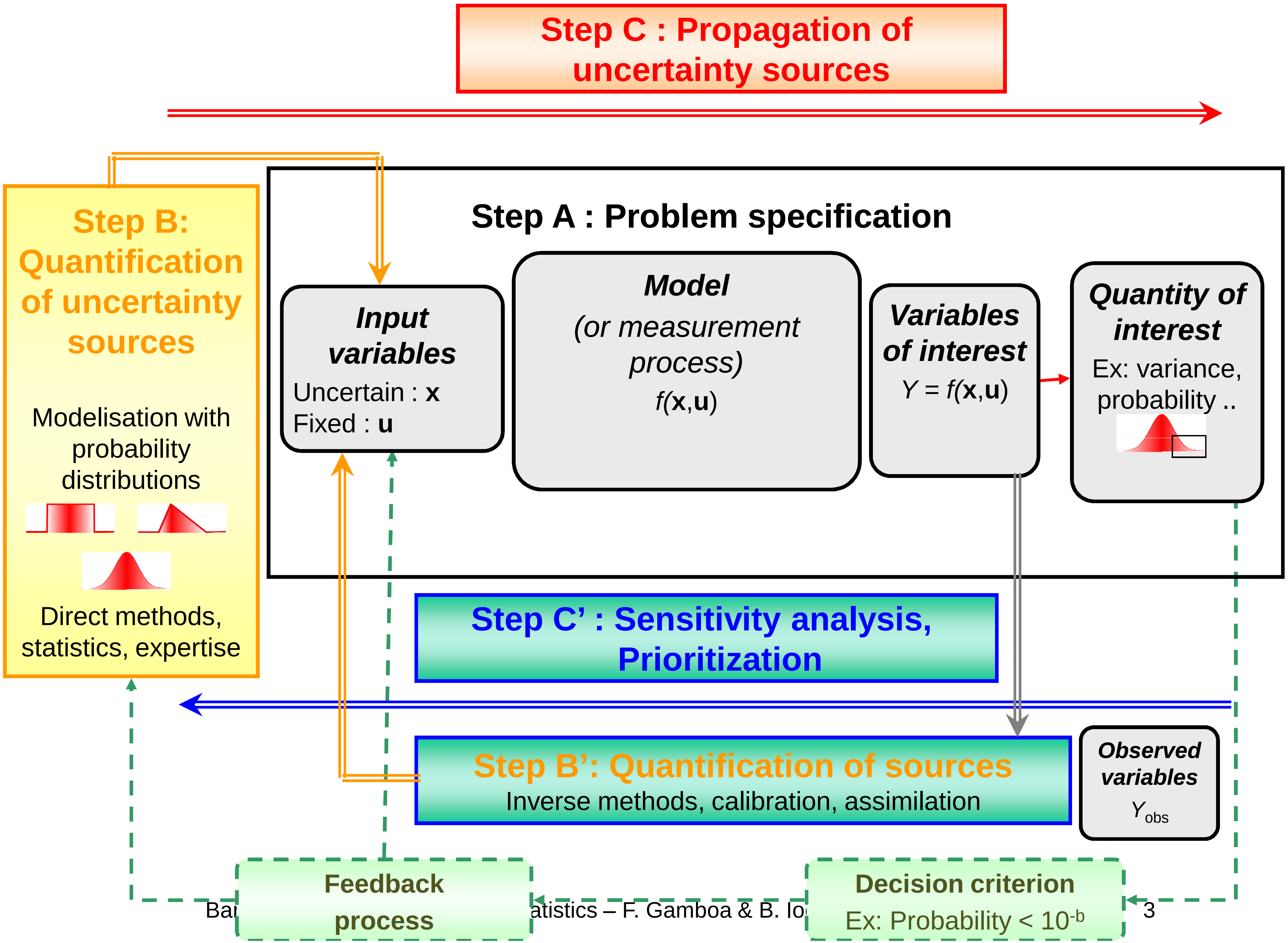
- **Validation phase**
  - Reduce prediction uncertainties
  - Calibrate the model parameters

- **Practical use of a model**
  - **Safety studies:** assess a risk of failure (rare events)
  - **Conception studies:** optimize system performances and robustness

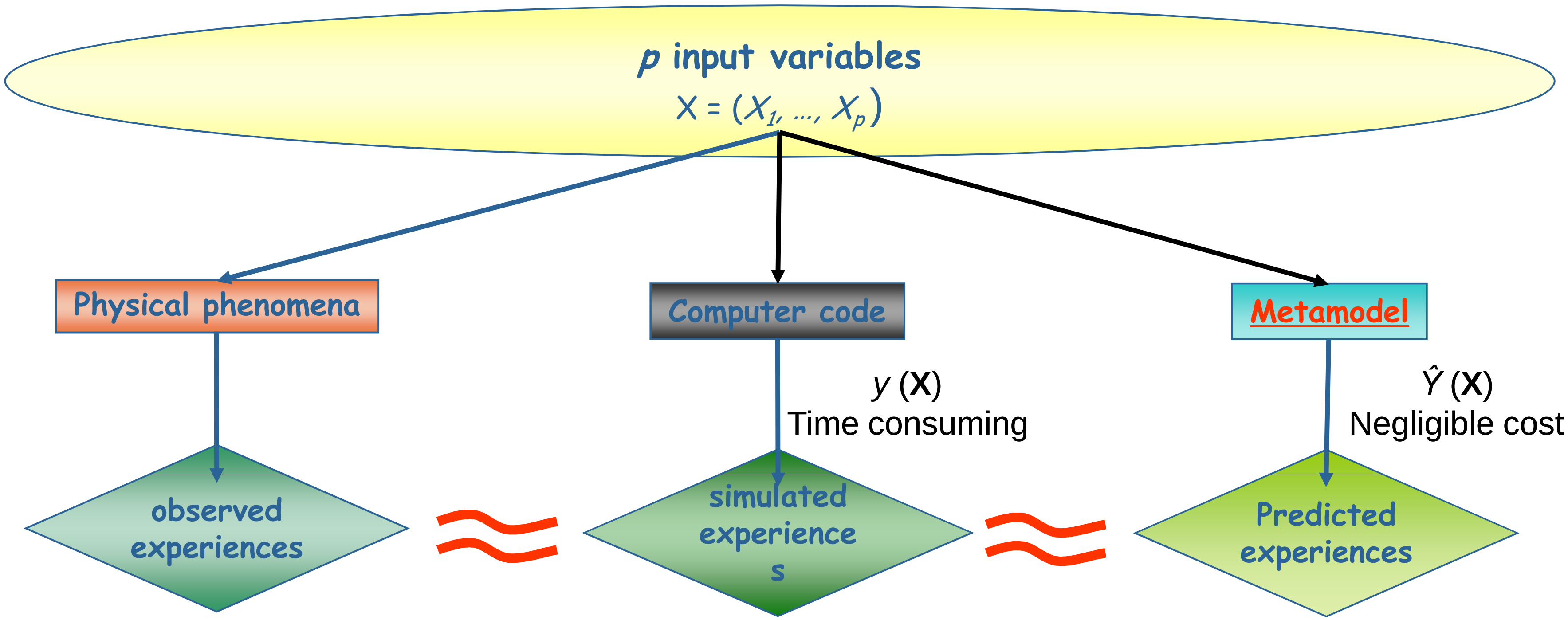
Advantages of a probabilistic approach

- to propagate the uncertainties, to perform global sensitivity analysis
- to design/optimize the system taking into account uncertainties
- to give rigorous safety margins, …

Uncertainty management - The generic methodology

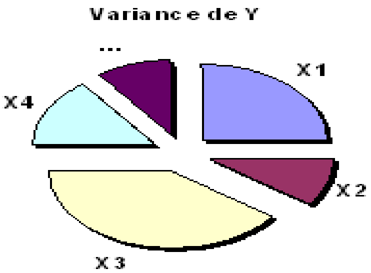


Uncertainties management for cpu time consuming models

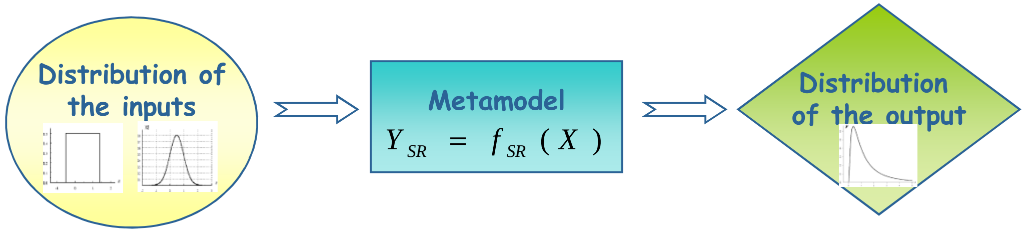


Use of the metamodel :

- C':Sensitivity analysis



- C:Uncertainty propagation (via Monte Carlo methods)



- B':Calibration

Identification of input parameters values

Adequation between observed and simulated experiences

Metamodel : definition

[Kleijnen 70's]

A metamodel is a mathematical function

- **which approximates the outputs of the model,**
- **with neglgible cpu cost,**
- **which allows to make new output predictions with a good accuracy**

- Synonyms:
  - Response surface
  - Simplified model
  - Emulator
  - Proxy model
  - Surrogate model

Metamodeling steps

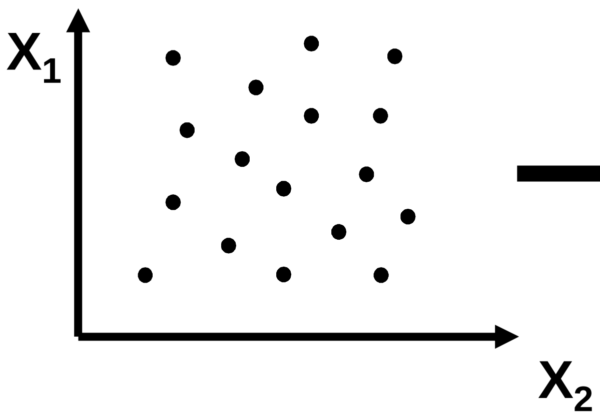
Design of experiments:
Points to perform simulations

→

Simulation :
Performing simulations

→

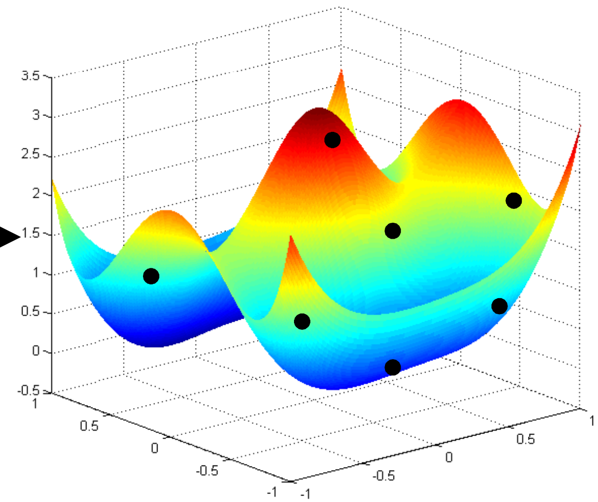
Metamodeling :
Approximation of the computer code



→

Computer code

→

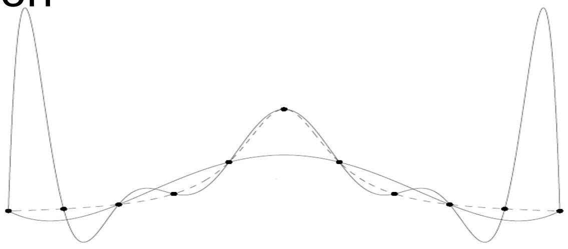


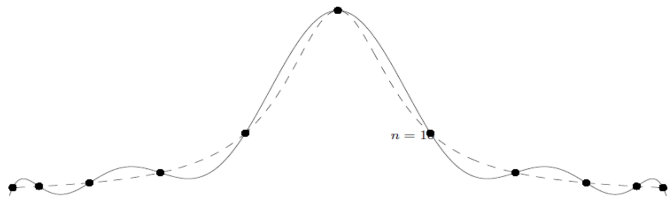
Differents types of metamodels

[Simpson et al. 2001]
[Storlie & Helton 2008]

- Linear regression

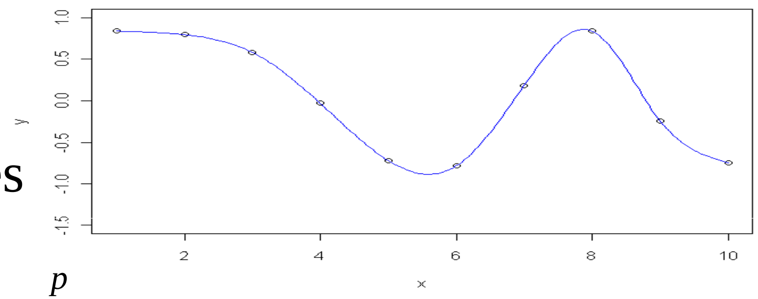
- Polynomials





- Splines

$$\hat{G}(\mathbf{x}) = \sum_{k=1}^{K} \hat{\beta}_k B_k(\mathbf{x}) \text{ with } K \text{ the number of nodes}$$



- Additive models, GAM $\hat{G}(\mathbf{x}) = \sum_{i=1}^{p} s_i(x_i) + \sum_{i<j}^{p} s_{ij}(x_i, x_j) + \cdots$

- Regression trees $\hat{G}(\mathbf{x}) = \sum_{k=1}^{K} \hat{\beta}_k I_k(\mathbf{x})$

- Neural networks
- Chaos polynomials
- Support Vector Machines
- Kriging – Gaussian process

Kriging metamodel

Kriging *[Matheron 63]* for computer codes relies on the idea to interpolate the code outputs in dimension *p* *[Sacks et al. 89]* as a spatial cartography

Kriging (or Gaussian process) is interesting because:

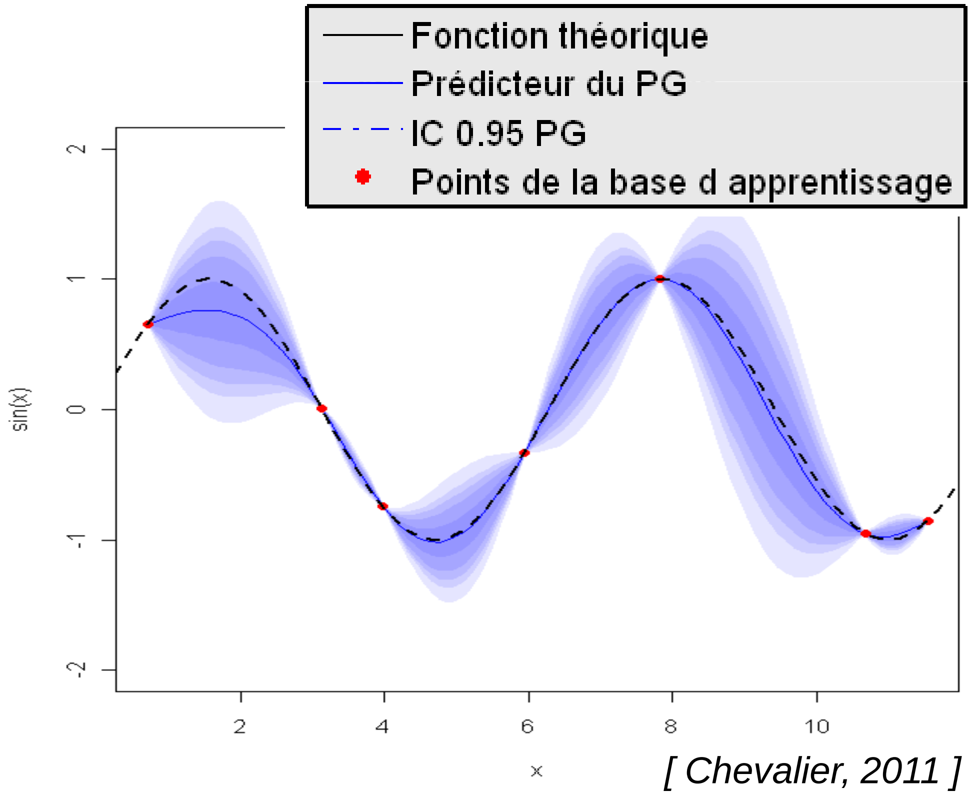
- **it interpolates the outputs,**
- **it gives predictor associated with confidence bands**

<u>Example in 1D :</u>

Theoretical function (p =1) :

$$Y = f(X) = \sin(X)$$

Simulation of N = 7 computation points



[Chevalier, 2011]

Introduction to Geostatistics

Introduction to Geostatistics

Objectives : treatment of **numerical data** with **spatial support** (or temporal) with **uncertainty quantification**

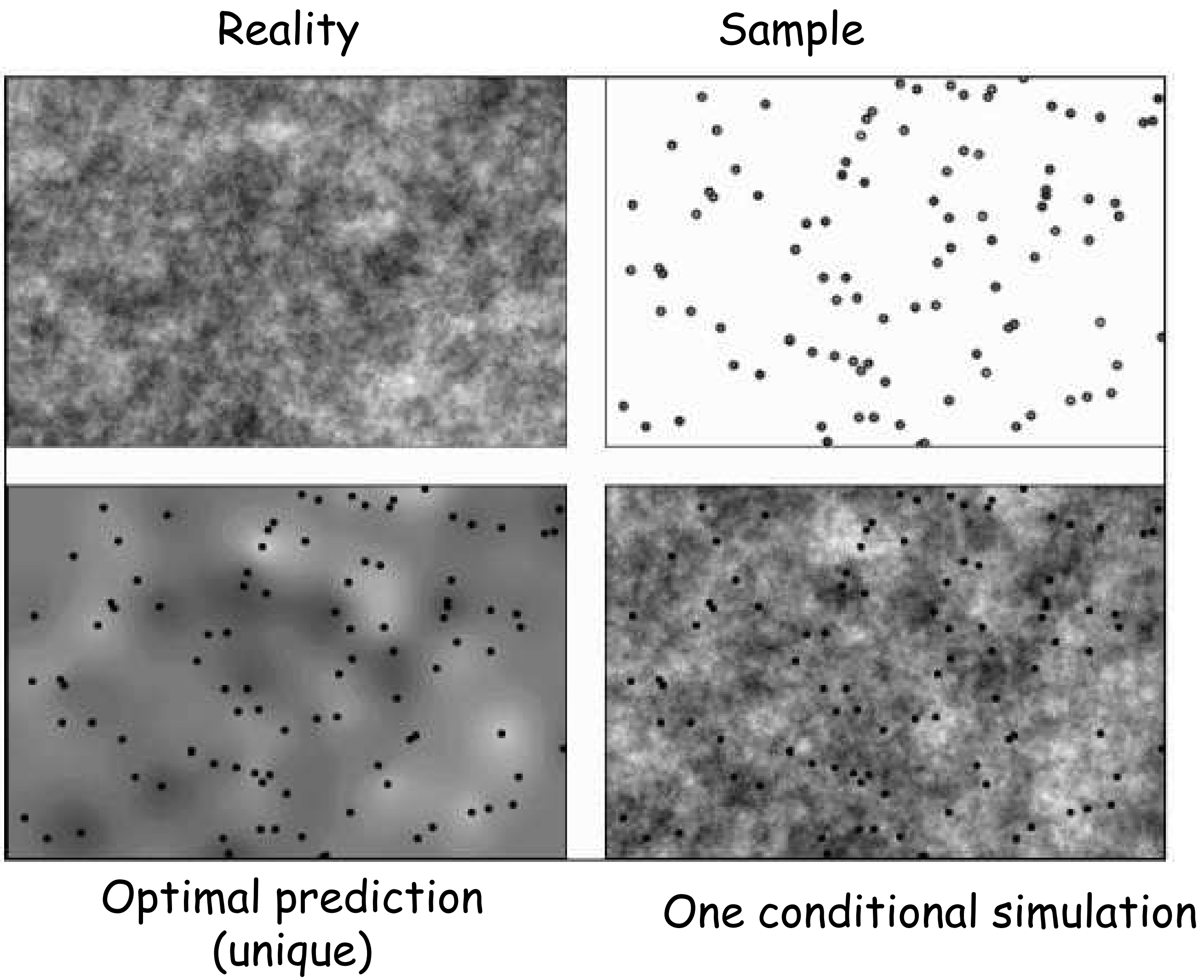
Principal aspects:

- Taking into account the spatial structure of data,
- Dimension 1, 2, 3, …,
- Irregular sampling,
- Integrating external information

2 types of methods:

- **Estimation** (prediction, …) at a given point
- **Simulations** reproducing the variability of the phenomenon

Example : porosity of a geological medium

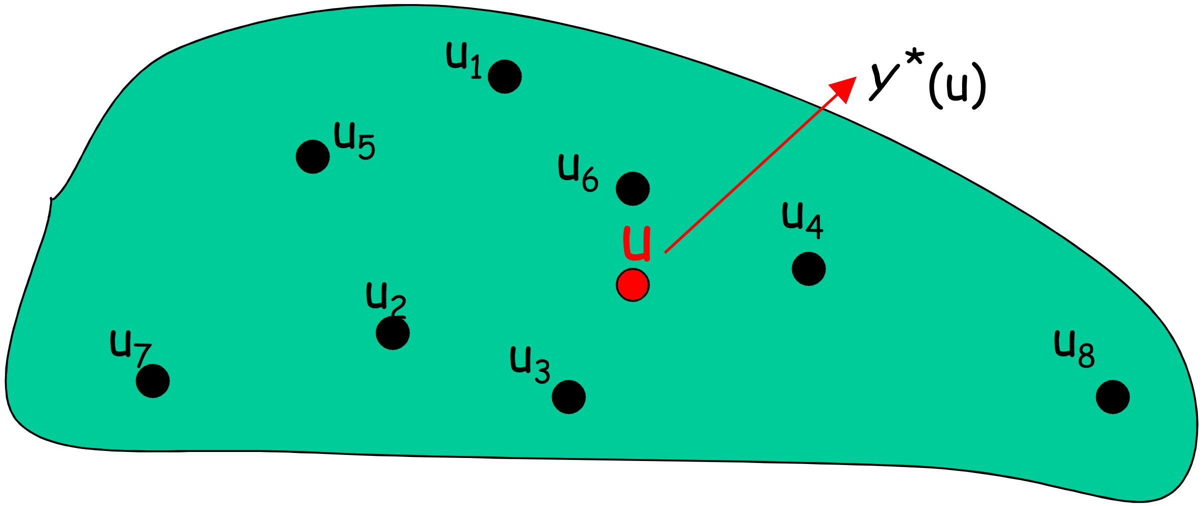


[Chilès]

Spatial statistics: kriging interpolation

Linear combination of N data: $$Y^*(u) = \sum_{i=1}^{N} \lambda_i Y(u_i)$$

Kriging can take into account the data configuration, the distance between data and target, the spatial correlations and potential external information



Probabilistic framework

Estimation without bias: $E[\,Y^*(u) - Y(u)\,] = 0$
the mean of the errors is zero

Estimation $Y^*(u)$ is optimal: $\mathrm{Var}[\,Y^*(u) - Y(u)\,]$ is minimal
the dispersion of the errors is reduced

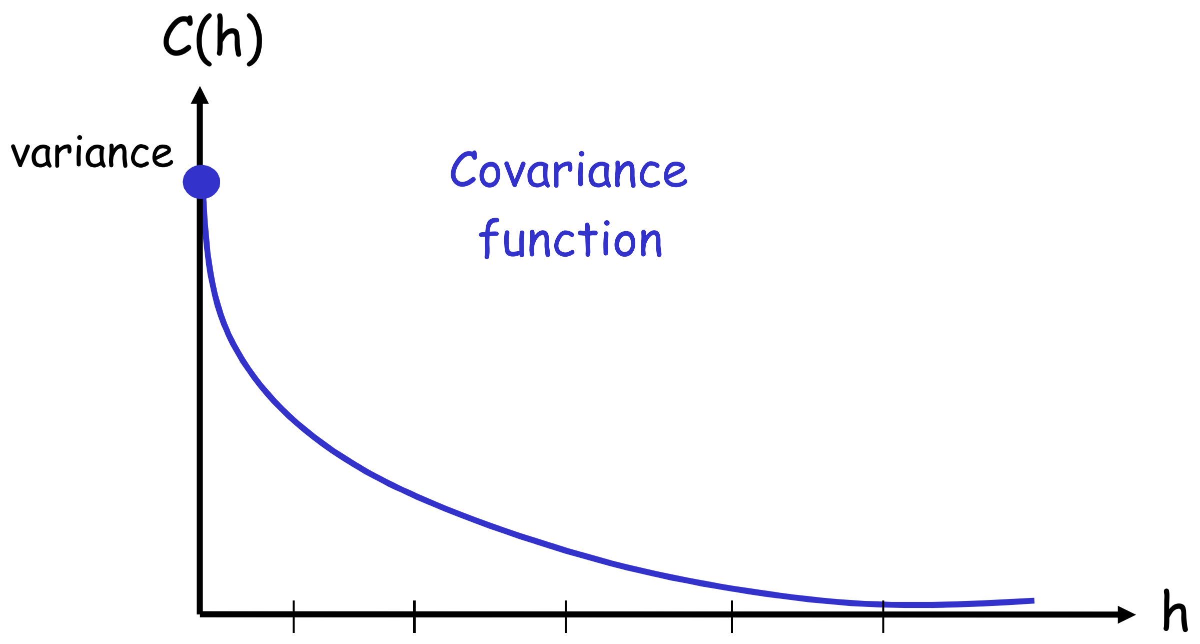
Stochastic model for Y(x)

The random field $Y(x)$, with $Y \in \Re$ and $x \in \Re^p$, is characterized by its mean and its covariance

$Y(x)$ is stationary of second order:

1. $E[Y(x)] = m$ does not depend on x
2. **Covariance:** $\mathrm{Cov}[Y(x), Y(x+h)] = \mathrm{E}[Y(x+h)Y(x)] - E[Y(x+h)]E[Y(x)] = C(h)$

does not depend on x



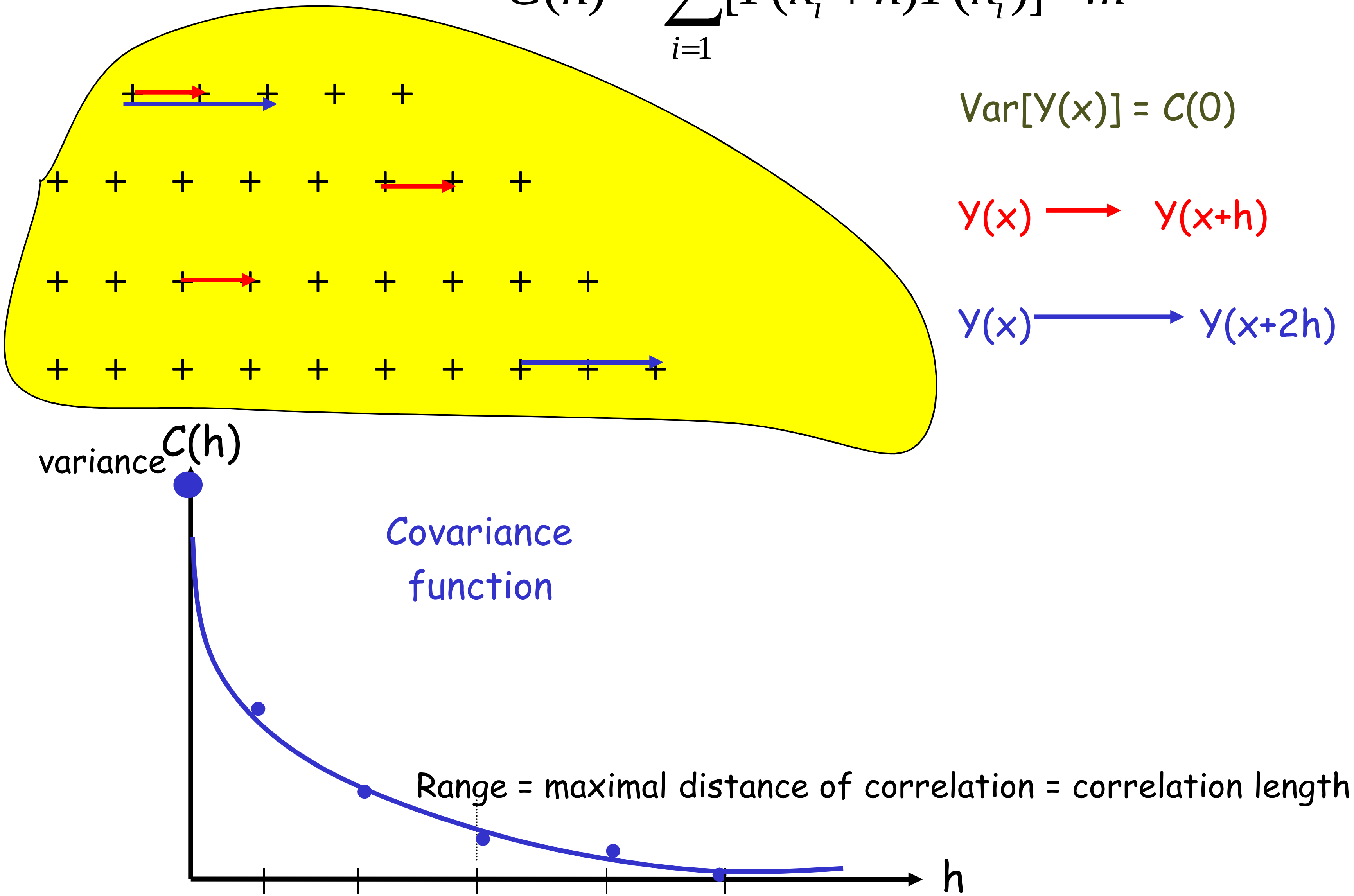
In practice

$$C(h) = \sum_{i=1}^{N(h)} \left[Y(x_i + h) Y(x_i) \right] - m^2$$

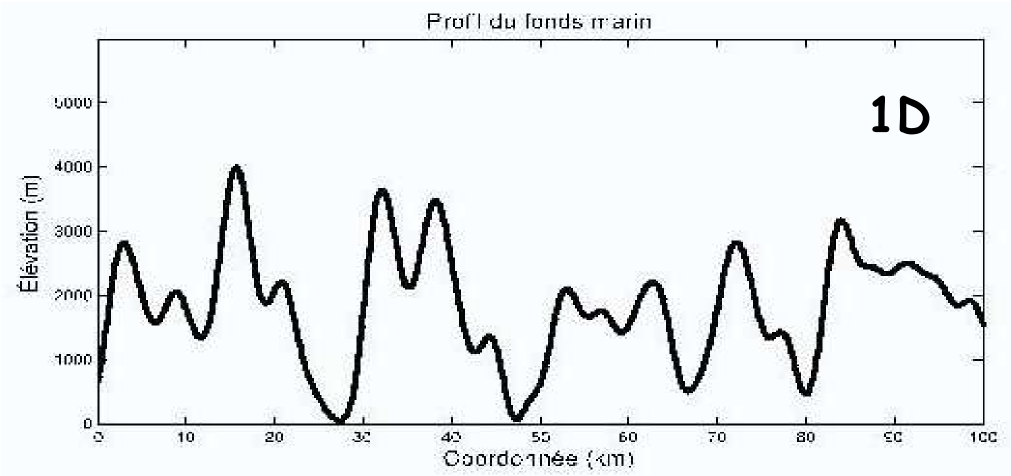
Var[Y(x)] = C(0)

Y(x) → Y(x+h)

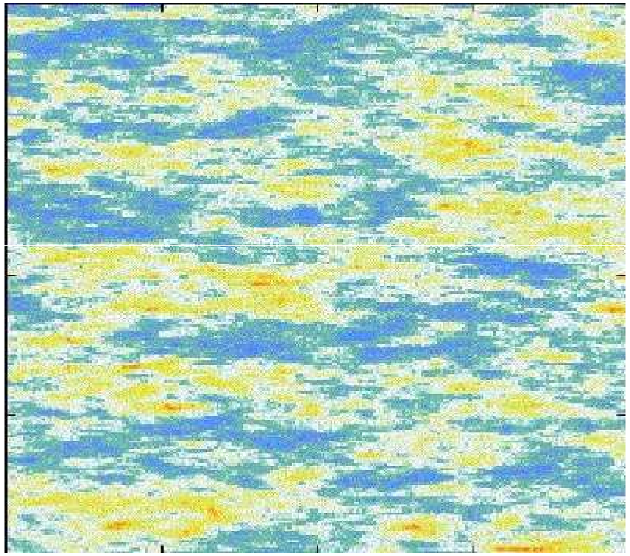
Y(x) → Y(x+2h)



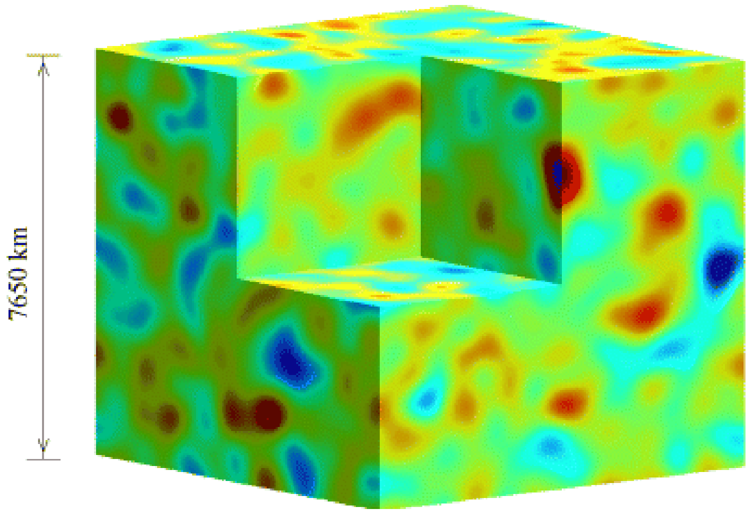
Examples of stochastic processes (Gaussian)



[from: Marcotte]



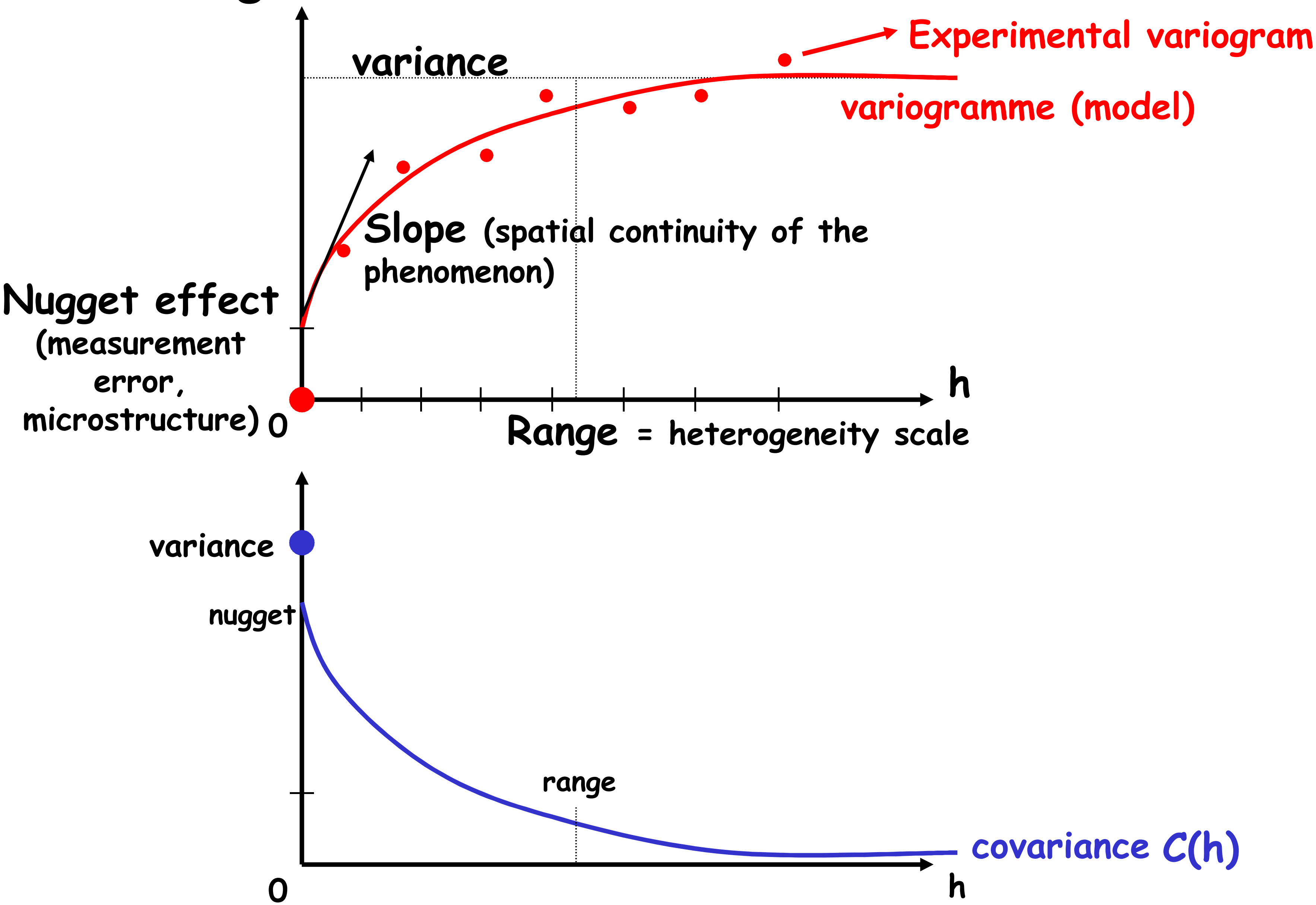
2D



3D

[from: Baig, 2003]

The variogram



Simple kriging (known mean)

$$Y^*(u) = \sum_{i=1}^{N} \lambda_i(u)\,[Y(u_i) - m] + m$$

(m = known constant)

Min { E [$Y^*(u)$ – $Y(u)$]^2 }

λ_i ➡ multiple linear regression by least squares

Best Linear Unbiased Predictor (BLUP)

Kriging weights $\lambda_i(u)$ for $Y(u_i)$ are obtained by:

$$\left\{ \sum_{j=1}^{N} \lambda_j(u)\, C(u_i - u_j) = C(u_i - u) \qquad \forall i = 1 \ldots N \right.$$

System of N linear equations with N unknowns which have an unique solution (for non singular covariance matrix)

Kriging variance (estimation error): $\sigma_K^2(u) = C(0) - \sum_{i=1}^{N} \lambda_i(u)\, C(u_i - u)$

does not depend on the Y values

=> Visualisation of regions with imprecise estimations

=> Put new observation points in these regions

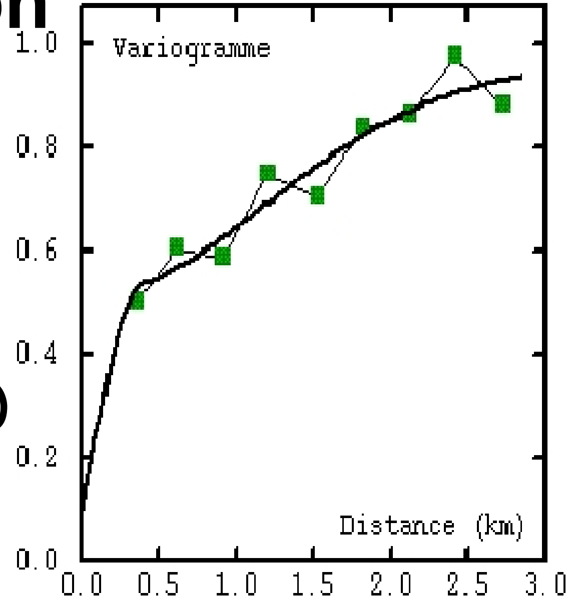
Example : cartography of air pollution

73 measures of benzene concentration (Rouen, France)

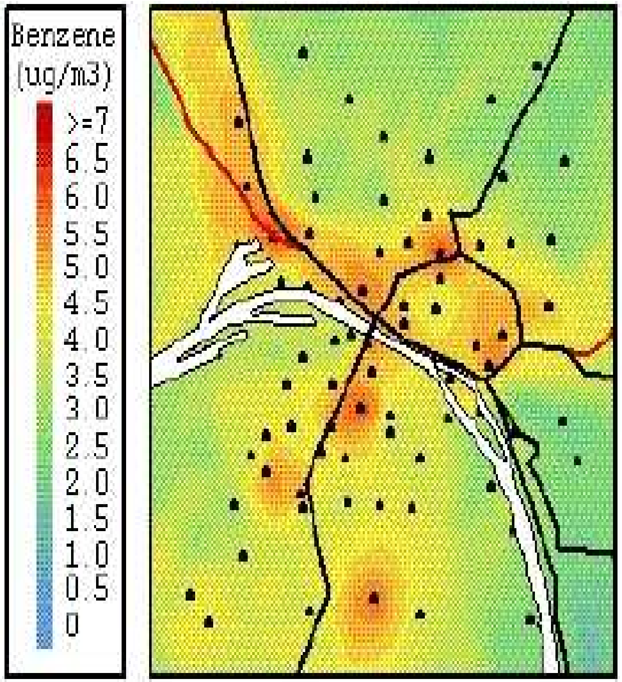
[from: Bobbia, Mietlicki & Roth, 2000]

$\gamma(h) = C(0) - C(h)$

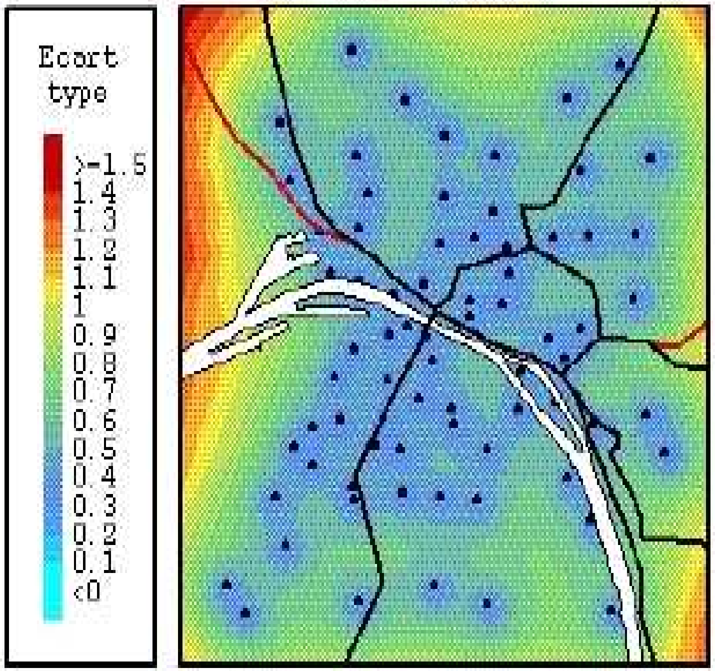
Variogram



Kriging mean



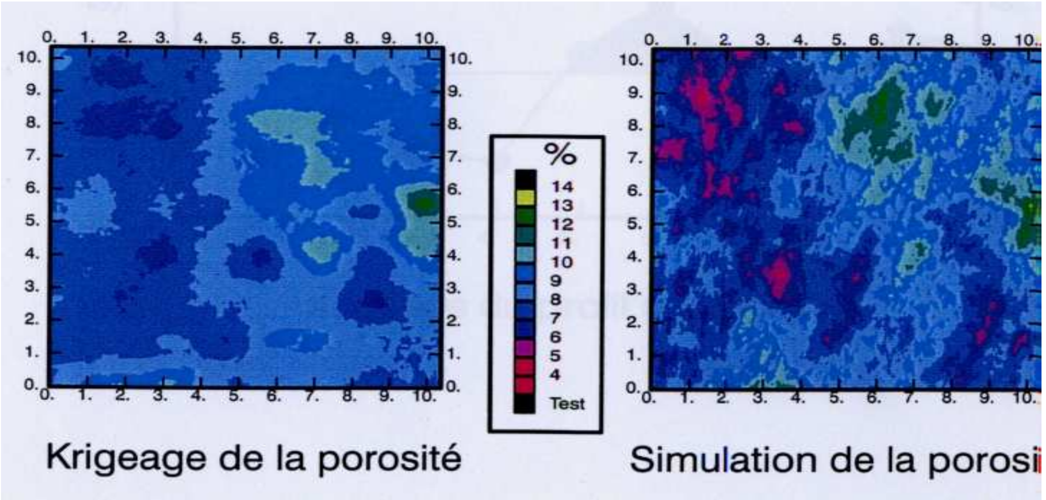
Kriging standard deviation



Simulations

- **Kriging give the optimal estimation** (unbiased, minimal error variance) of the variable at any point, from experimental data

- **A simulation represents a possible realization** of the real phenomenon
  It reproduces its true **variability** (distribution, variogram), with respecy to experimental data **(conditional simulation)**



Main goal of simulation : quantify the uncertainty via sampling (as Monte Carlo)

Numerous methods of random fields simulation (LU decomposition, turning bands method, spectral method, Karhunen-Loève, etc.)

Conditional simulations

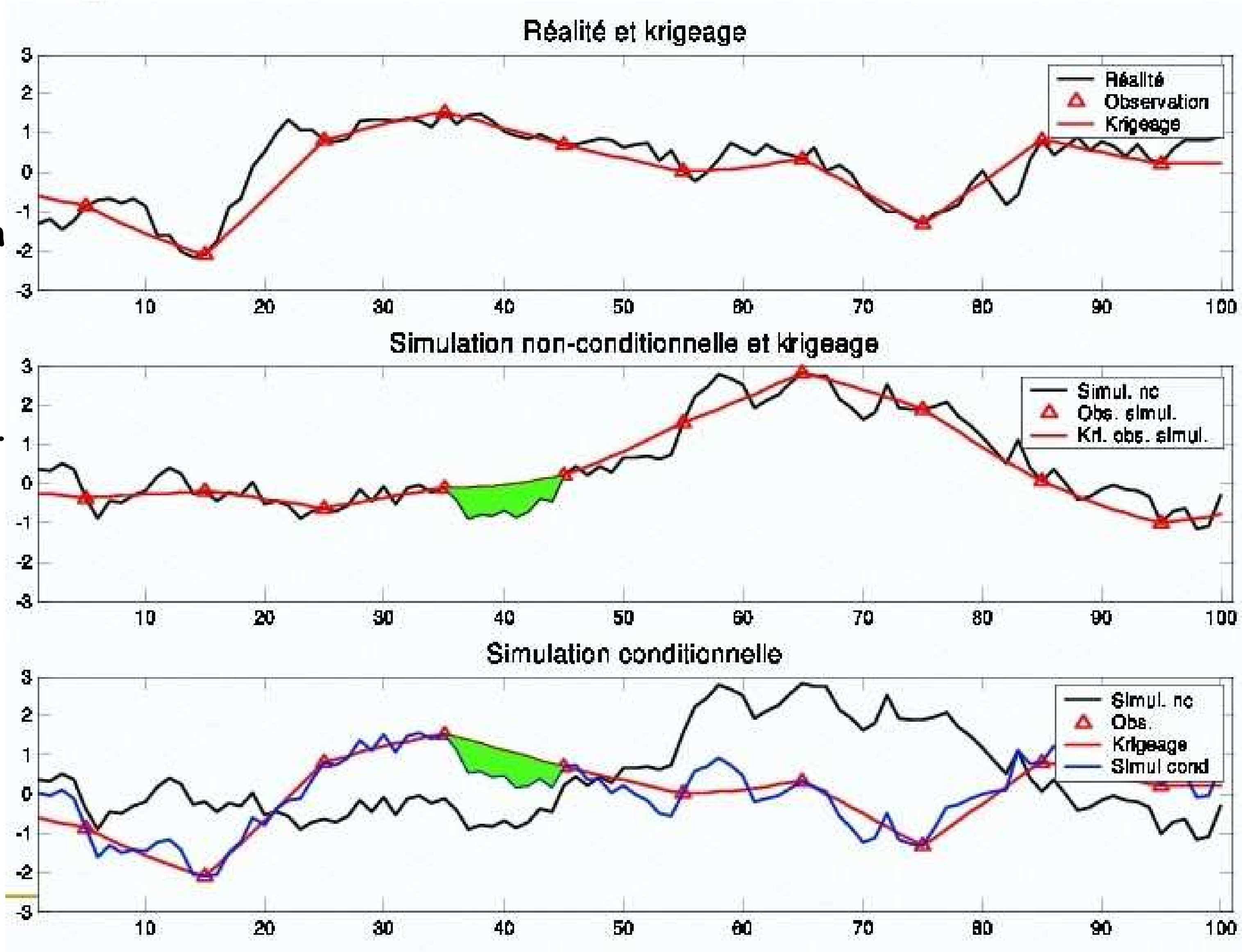
Kriging
of observation

+

Non cond. Simul.
-
Kriging of
non cond. simul

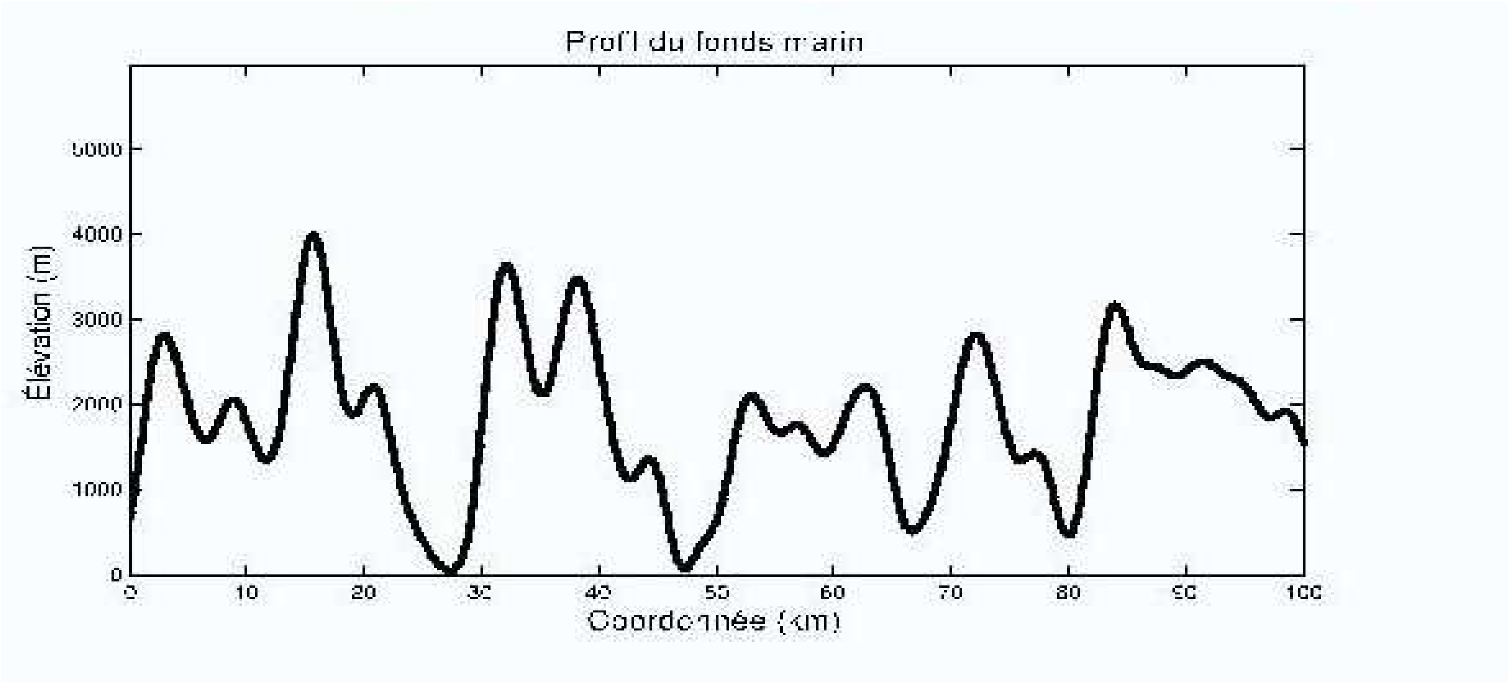
=

Conditional
simulation



[Marcotte, Cours EPM]

Example : profile of ocean bottom (1/5)



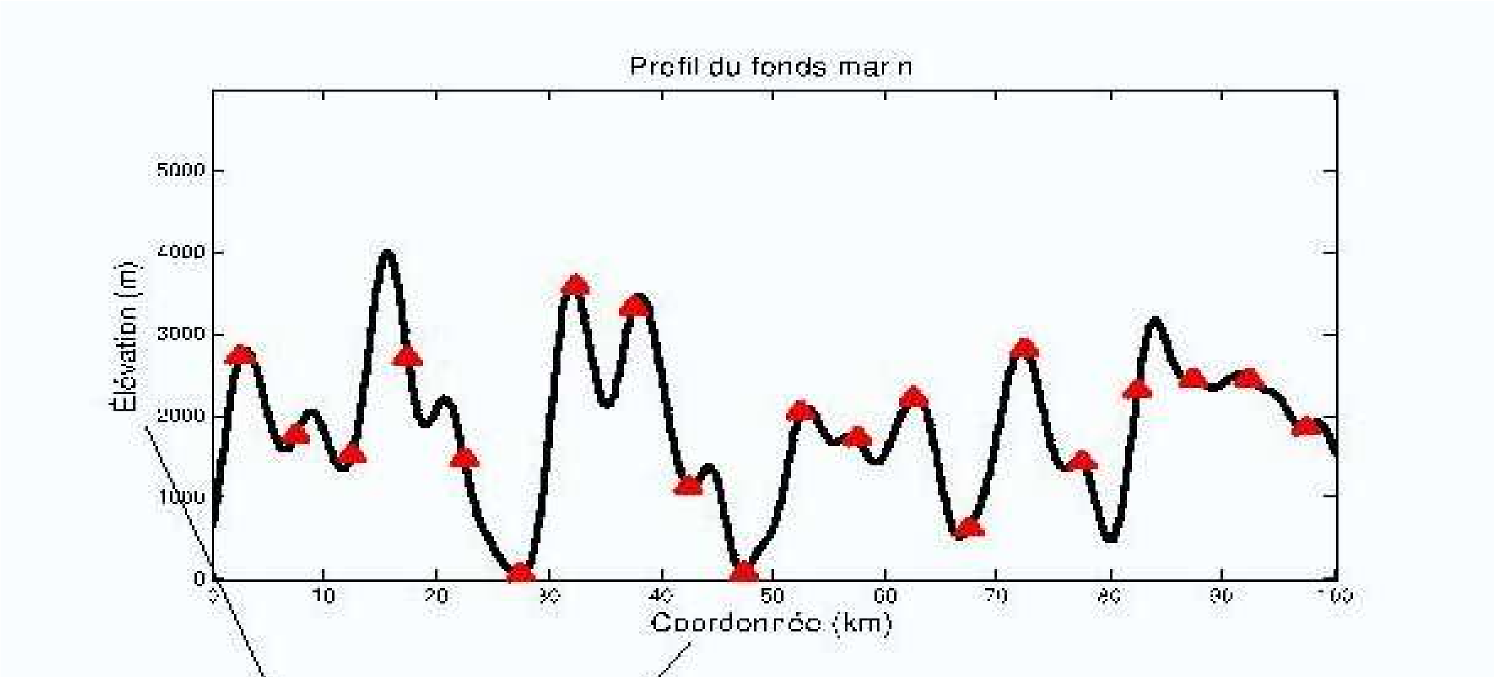
You have to put a cable on the ocean bottom

Question: what is the length of the cable?

[Marcotte, Cours EPM]

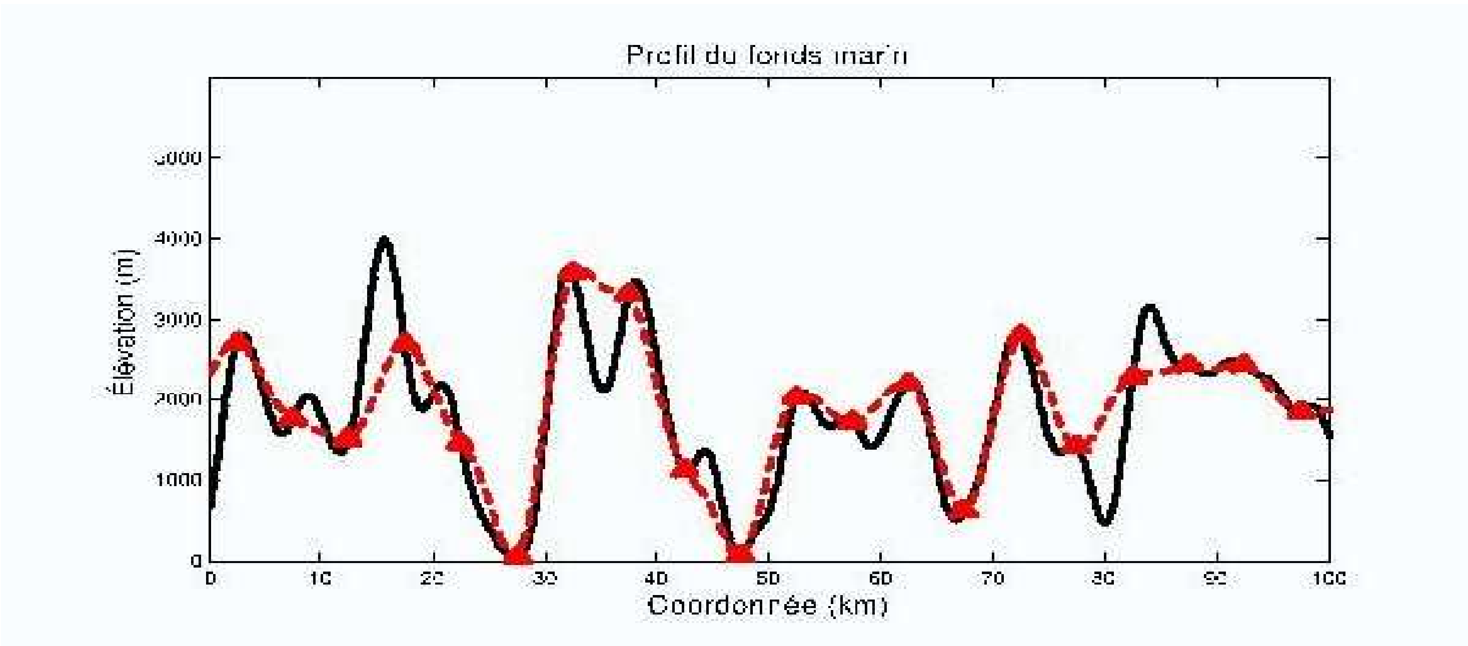
Example : profile of ocean bottom (2/5)

The exact depth is uniquely known at the observation points (survey)



Example : profile of ocean bottom (3/5)

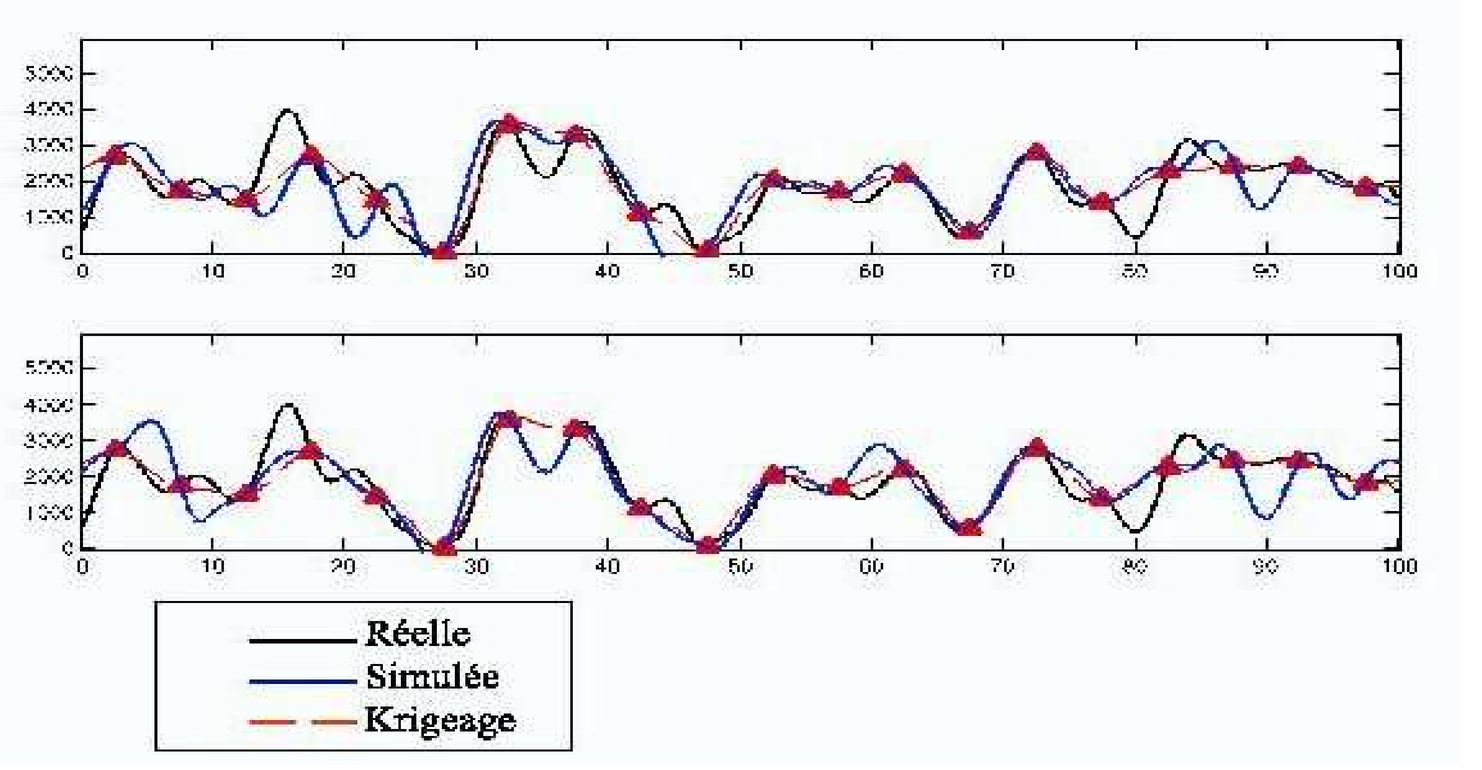
Kriging of the ocean depth



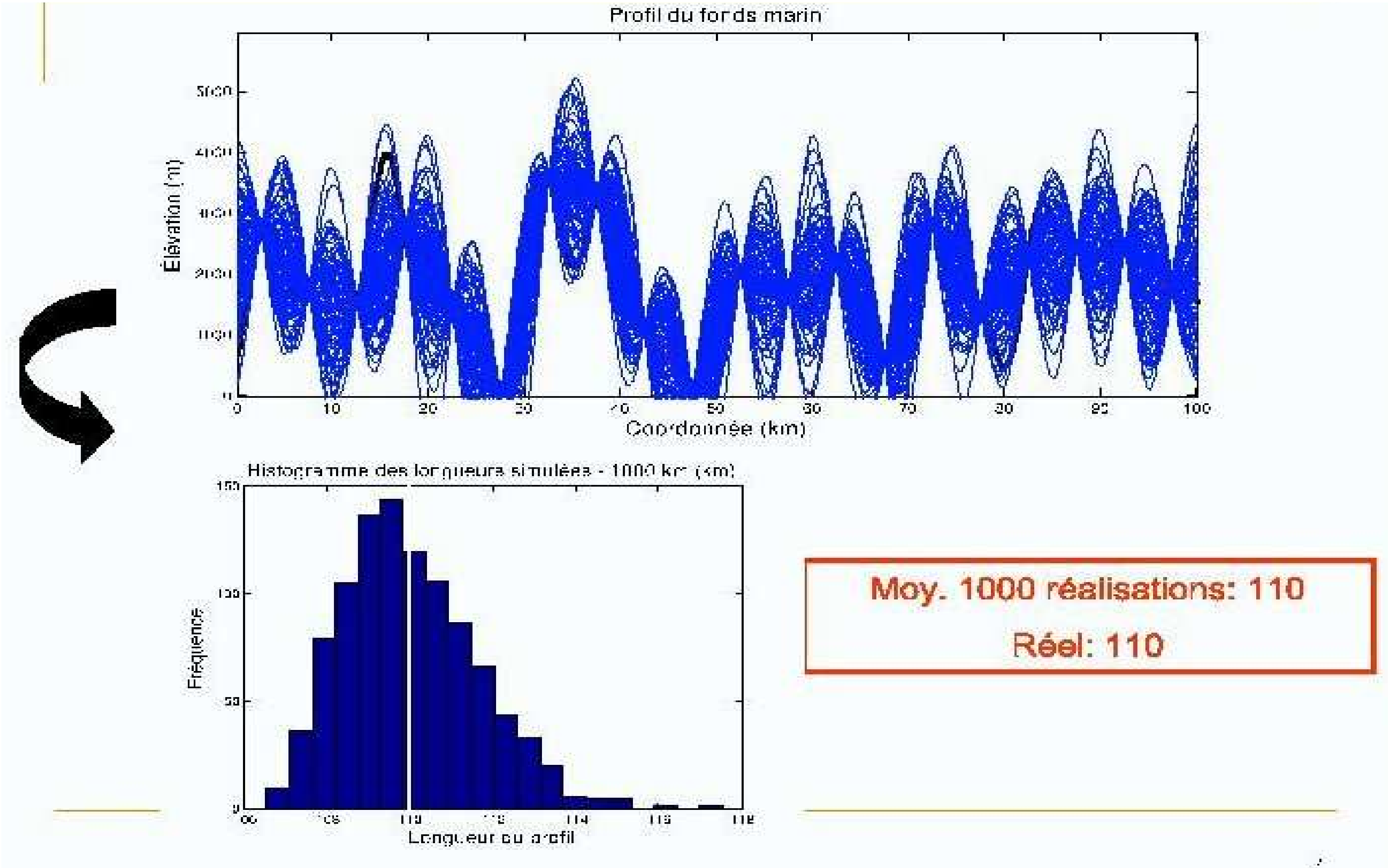
The true length is 110 km while kriging gives 104.6 km
=> some cable is missing

Example : profile of ocean bottom (4/5)

Another approach: the conditional simulations



Example : profile of ocean bottom (5/5)



The 95%-confidence interval from conditional simulations is [108.8,113.5]

Same problem for probability of failure estimation (non linear transfer fct)

Gaussian process metamodel

Gaussian process metamodel (1/2)

- Idea: Computer code results are interpolated with the kriging technique
- Necessary hypothesis: Gaussian process

Stochastic process Z with :

- $E[Z(x)] = 0$
- $Cov(Z(x), Z(u)) = \sigma^2 R(x, u)$

  where σ^2 is the variance

  and R the correlation function
- $Z \sim N(0, \sigma^2 R)$

- Definition:

  $Y(x) = \beta F(x) + Z(x)$

  $\beta F(x)$: Regression — $Z(x)$: stochastic part

- Parametric choices:
  - F : polynomial of degree 1 $\beta F(\mathrm{x}) = \beta_0 + \sum_{i=1}^{p} \beta_{\mathrm{i}} x_i$
  - R : stationary => covariance function

Example: Gaussian covariance $\mathrm{R}(\mathrm{x}, \mathrm{u}) = \mathrm{R}(\mathrm{x} - \mathrm{u}) = \exp\left(-\sum_{i=1}^{p} \theta_i \left| x_i - u_i \right|^2 \right)$

Anisotropy: θ_is are not equal (correlation length of each input variable)

Gaussian process metamodel (2/2)

- Joint distribution :
  - Gaussian process (Gp) model : $Y(x) = \beta F(x) + Z(x),\ x \in \Re^p$
  - Learning sample (LS) of N simulations : (X_{LS}, Y_{LS})

$$X_{LS} = \left(x^{(1)},\dots,x^{(N)}\right),\ F_{LS} = F(X_{LS}),\ R_{LS} = \left(R\left(x^{(i)},x^{(k)}\right)\right)_{i,k}$$
$$Y_{LS} \sim N(\beta F_{LS}, \sigma^2 R_{LS})$$

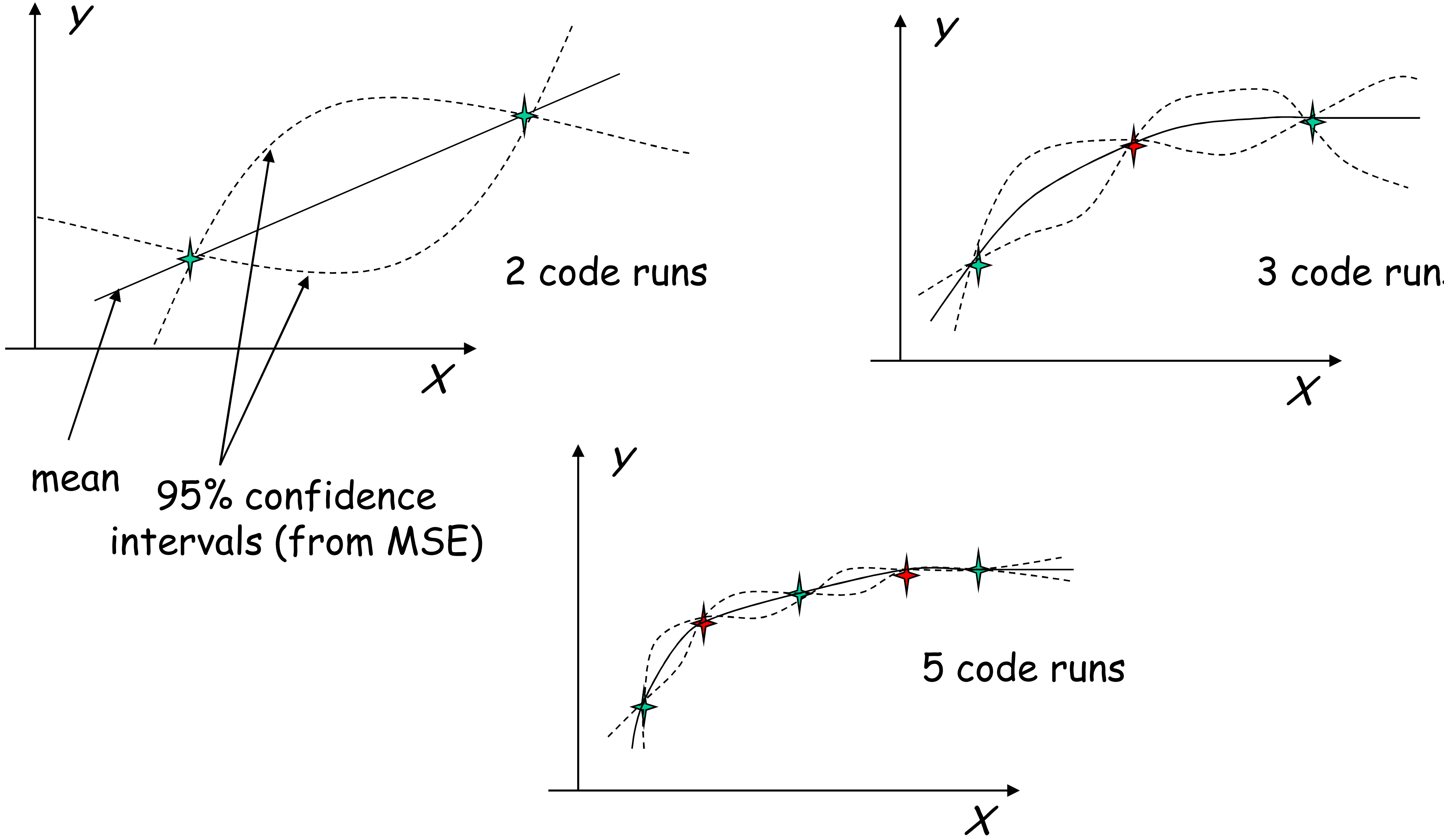
  - Conditional Gp metamodel :

$$\Rightarrow \quad Y(x)_{|X_{LS},Y_{LS}} \sim Gp$$

$$\text{Mean}: \hat{Y}(x) = \mathrm{E}\left[Y(x)_{|X_{LS},Y_{LS}}\right] = \beta F(x) + r(x) R_{LS}^{-1}\left[Y_{LS} - \beta F_{LS}\right]$$
$$\text{with } r(x) = \left[R\left(x^{(1)},x\right),\dots,R\left(x^{(N)},x\right)\right]$$

$$\text{Covariance}: \mathrm{Cov}\left(Y(u)_{|X_{LS},Y_{LS}}, Y(v)_{|X_{LS},Y_{LS}}\right) = \sigma^2\left(R(u,v) + {}^t r(u) R_{LS}^{-1} r(v)\right)$$
$$\Rightarrow \text{Variance} \Rightarrow \text{Mean Square Error (MSE)}$$

Illustration



Conclusion: given a sufficient number of points, we obtain an accurate metamodel

Hyperparameters estimation

- **Maximum likelihood method**

  - Likelihood maximisation on the learning basis (X_s,Y_s):

$$\left(\beta^*,\theta^*,\sigma^*\right)=\underset{(\beta,\theta,\sigma)}{\text{Argmax}}\ \ln L\left(Y_{LS},\beta,\theta,\sigma\right)$$

  - with

$$\ln L\left(Y_{LS},\beta,\theta,\sigma\right)=-\frac{N}{2}\ln\left(2\pi\sigma^2\right)-\frac{1}{2}\ln\left(\det R_{LS}\right)-\frac{1}{2}\sigma^2\ {}^t\left[Y_{LS}-\beta F_{LS}\right]R_{LS}^{-1}\left[Y_{LS}-\beta F_{LS}\right]$$

  - Joint estimation of β and σ :

$$\begin{cases}\beta^*=\left[{}^tF_{LS}R_{LS}^{-1}F_{LS}\right]^{-1}{}^tF_{LS}R_{LS}^{-1}Y_{LS}\\ \sigma^{2^*}=\dfrac{1}{N}\,{}^t\left[Y_{LS}-\beta^*F_{LS}\right]R_{LS}^{-1}\left[Y_{LS}-\beta^*F_{LS}\right]\end{cases}$$

  - Estimation of correlation parameters θ :

$$\left(\theta^*\right)=\underset{\theta}{\text{Argmin}}\ \psi\left(\theta\right)\quad \text{with}\quad \psi\left(\theta\right)=\left|R_{LS}\right|^{-1/N}\sigma^{2^*}$$

Estimation and validation

$$R(\mathbf{u},\mathbf{v}) = R(\mathbf{u}-\mathbf{v}) = \exp\left(-\sum_{i=1}^{p}\theta_i |u_i - v_i|^2\right)$$

- **Hyperparameters $(\theta_i)_{i=1\ldots p}$ estimated by likelihood maximization**

  Simplex method, **stochastic algorithms**

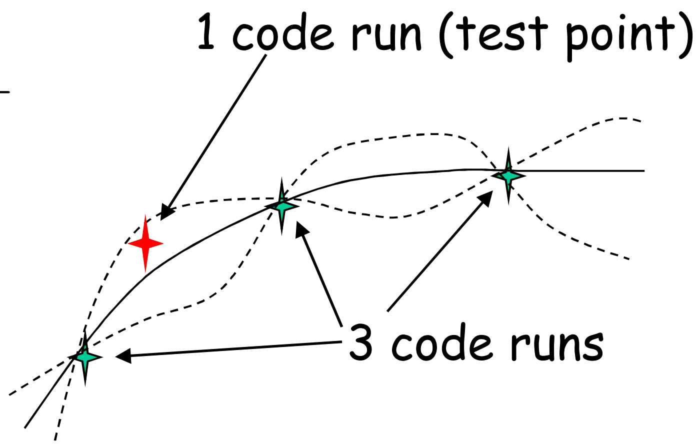
  Problems in high dimensional context ($p > 10$), can be solved by sequential algorithms *[Marrel et al. 2008]*

- **Predictor validation:**

  Predictivity coefficient

$$Q_2(Y,\hat{Y}) = 1 - \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^{n}\left(\overline{Y} - Y_i\right)^2}$$

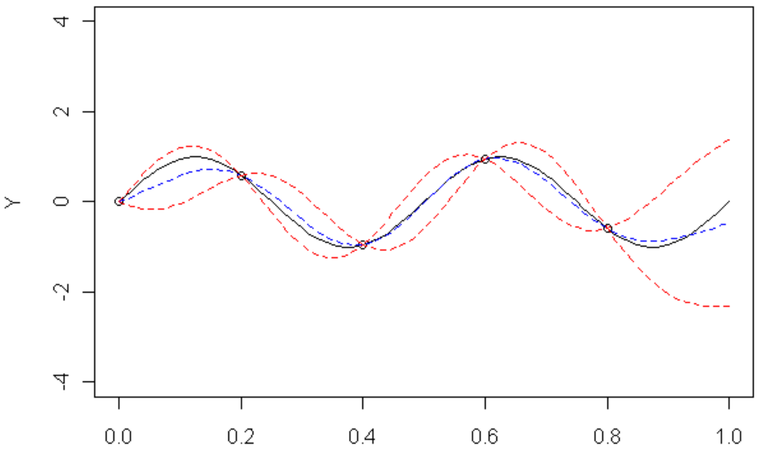
- Test sample
- or leave-one-out
- or *k*-fold cross validation



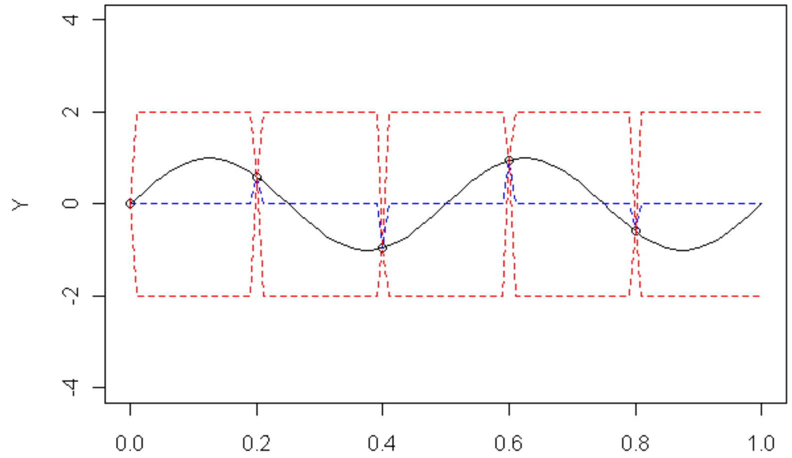
- **MSE validation:** Percentage of predicted values inside confidence bounds

Effects of the hyperparameters $\boldsymbol{\theta}$ and $\boldsymbol{\sigma}$

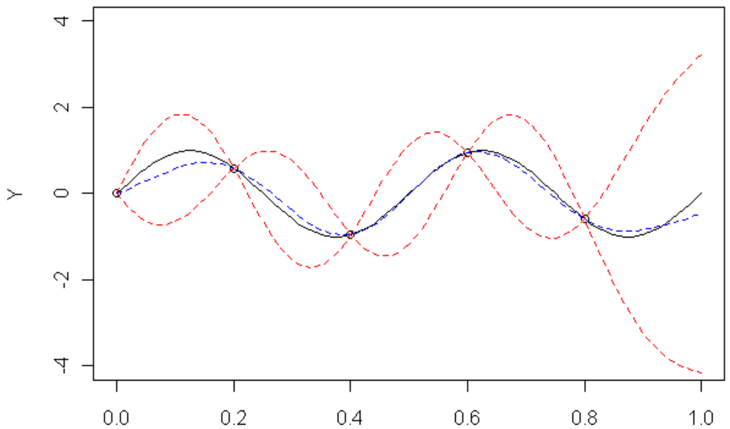
$$f(x) = \sin(4\pi x)$$



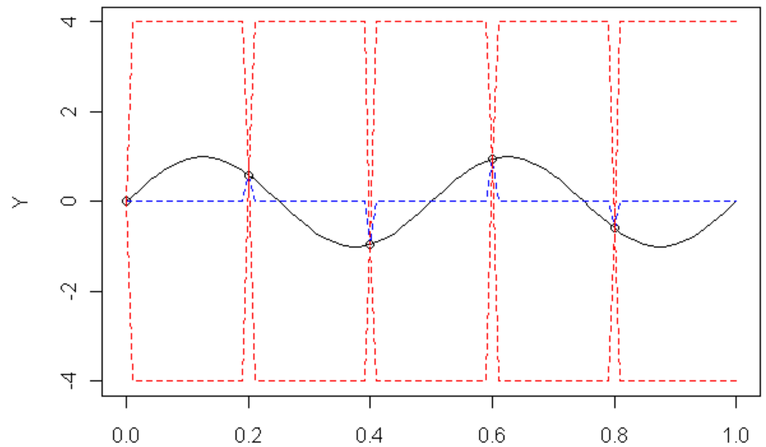
$$\sigma^2 = 1; \theta = 0.2$$



$$\sigma^2 = 1; \theta = 10^{-4}$$



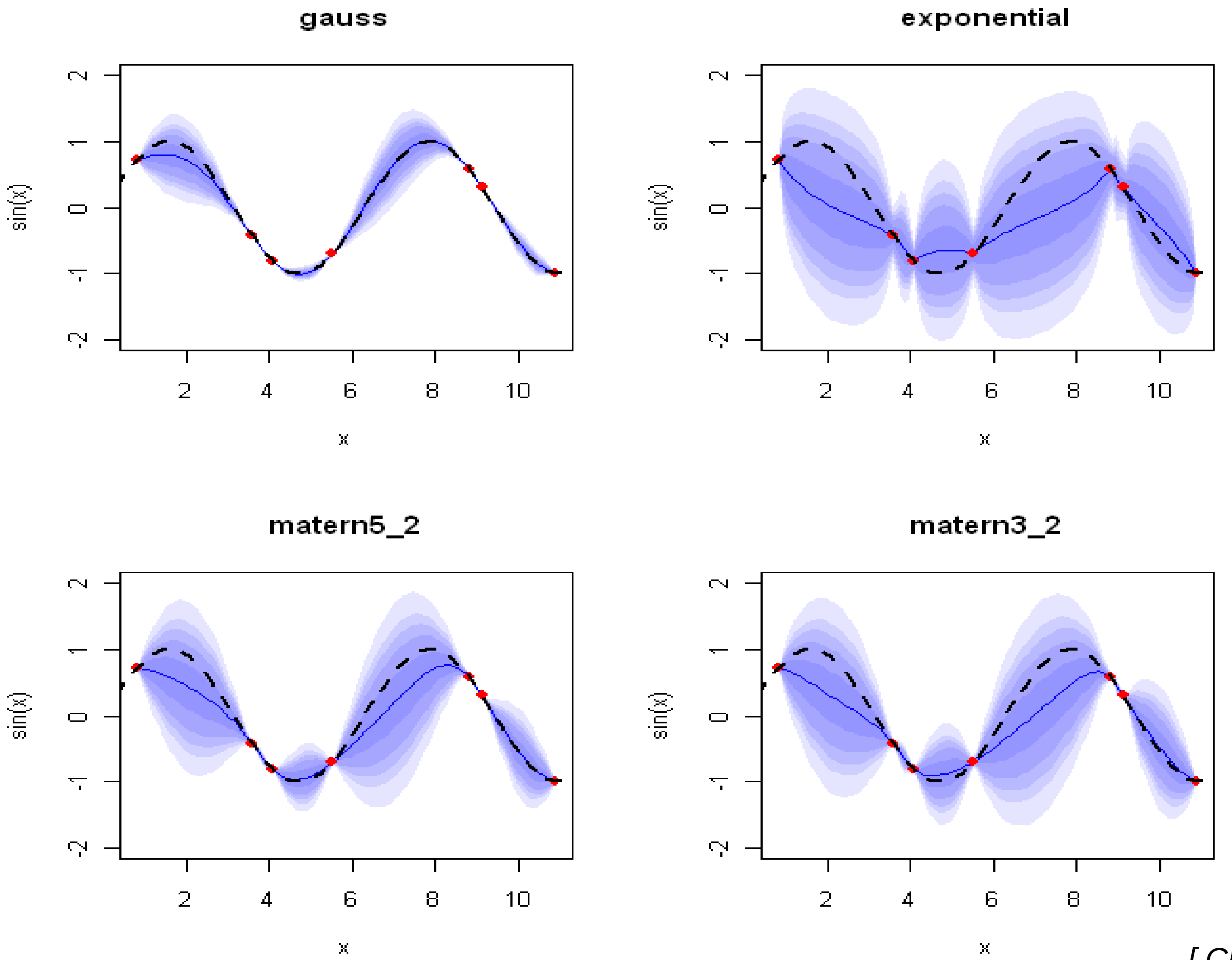
$$\sigma^2 = 4; \theta = 0.2$$



$$\sigma^2 = 4; \theta = 10^{-4}$$

[Le Gratiet, 2011]

Effects of the covariance structure



[Chevalier, 2011]

Adaptive designs using Gaussian process metamodel

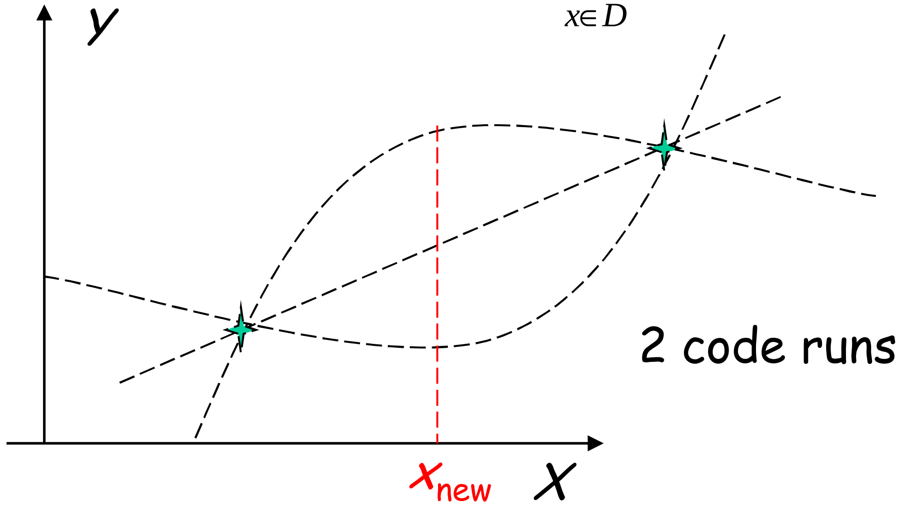
The best way to build Gp: model-based adaptive designs

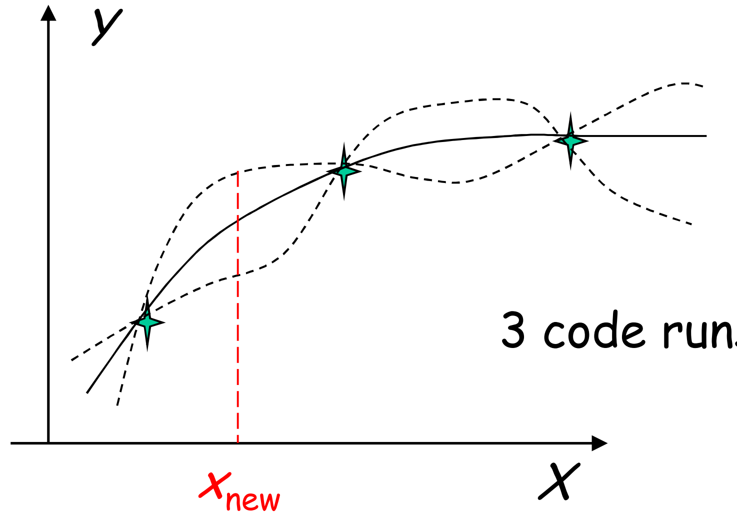
Example: criterion of the Gaussian process MSE (Mean Square Error)

$$MSE(x) = \sigma^2 + {}^t r(x) R_{LS}^{-1} r(x) + u(x) ({}^t(\beta F_{LS}) R_{LS}^{-1} \beta F_{LS})^t u(x)$$

$$u(x) = \beta F(x) - {}^t k(x) R_{LS}^{-1} \beta F_{LS}$$

$$x_{\text{new}} = \arg\max_{x \in D} MSE(x)$$





Remark: other criteria are possible (e.g. focusing to active variables)

Conclusion: Model-based adaptive designs are the most efficient ones,
but are not always applicable
In practice, we need to initiate the process with a space-filling design

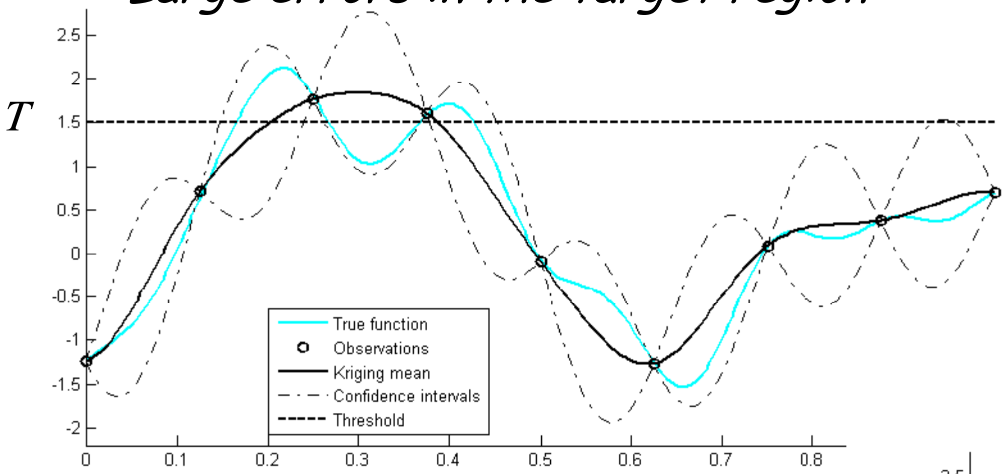
Estimation of rare events probability using kriging

[Bect et al. 2012]

Industrial problems: safety analysis with computer code (nuclear, transport, …)

<u>Problem:</u> find P_f = Prob [$f(X) > T$] with X = random inputs ; T = treshold

Reasonable variance everywhere
Large errors in the target region



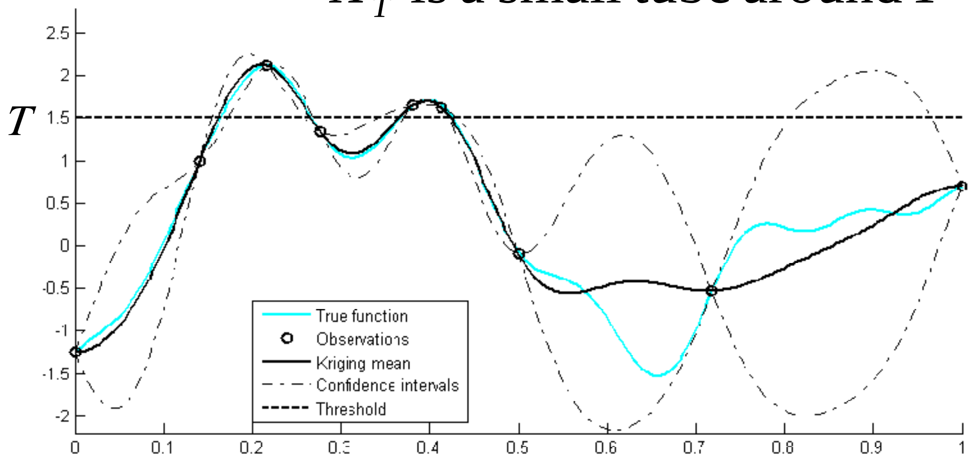
[from: Picheny et al. 2010]

☺ **New adaptive design**

$$X^* = \arg\min_X (IMSE_T)$$

$$IMSE_T = \int MSE(x) 1_{X_T}(x) dx$$

X_T is a small tube around T



Large variance in non-target region
Good accuracy in target region

Optimisation of a model output using kriging

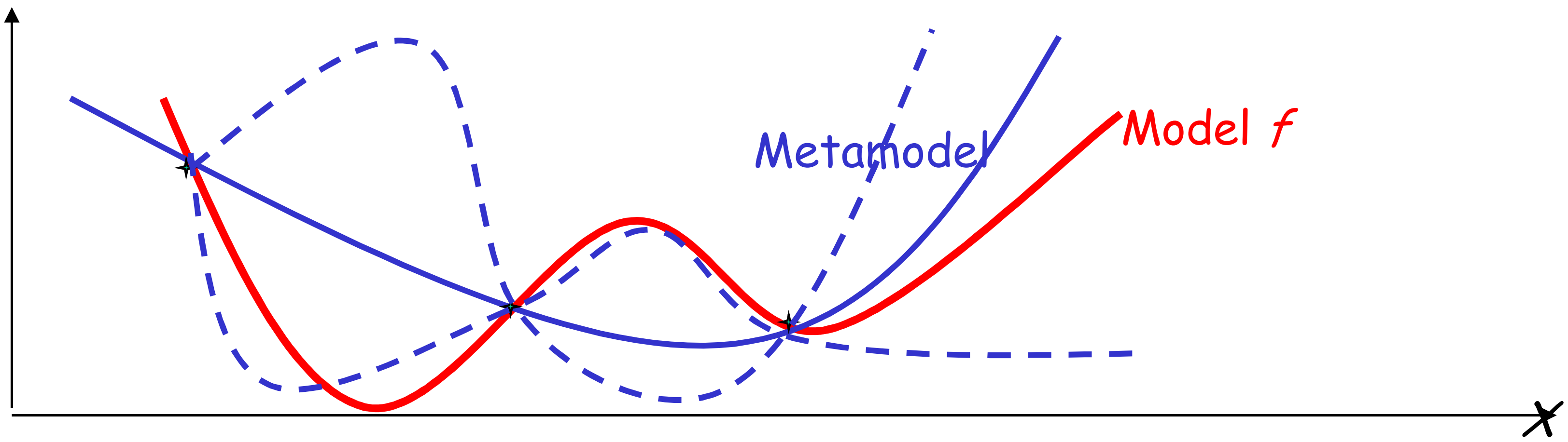
Industrial problems: conception with costly computer code (automobile, nuclear, aeronautics, …)

<u>Problem:</u> find the values of X which minimize the model output

$$X^* = \arg\min_{X \in D} f(X)$$

☠ **If *f* is costly, a natural solution would be to optimize a metamodel of *f* :** dangerous idea because the metamodel tends to smooth the true model

☺ Gp metamodel allows to take into account the metamodel error, and to define the expected improvement EI*(X)* for each $X \in D$

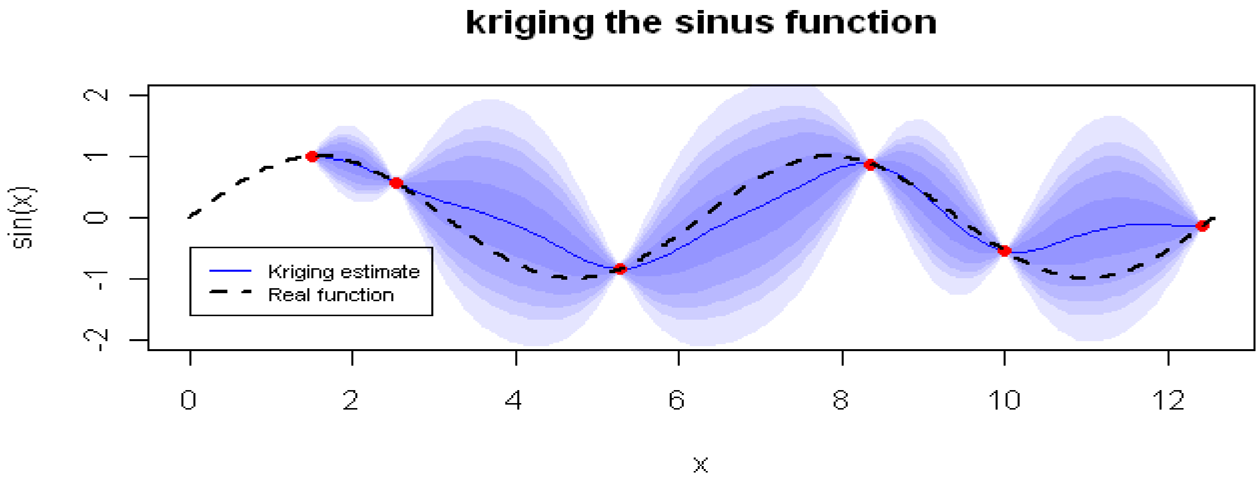


EI(x) = E[max(0 , observed minimum – f(x))]

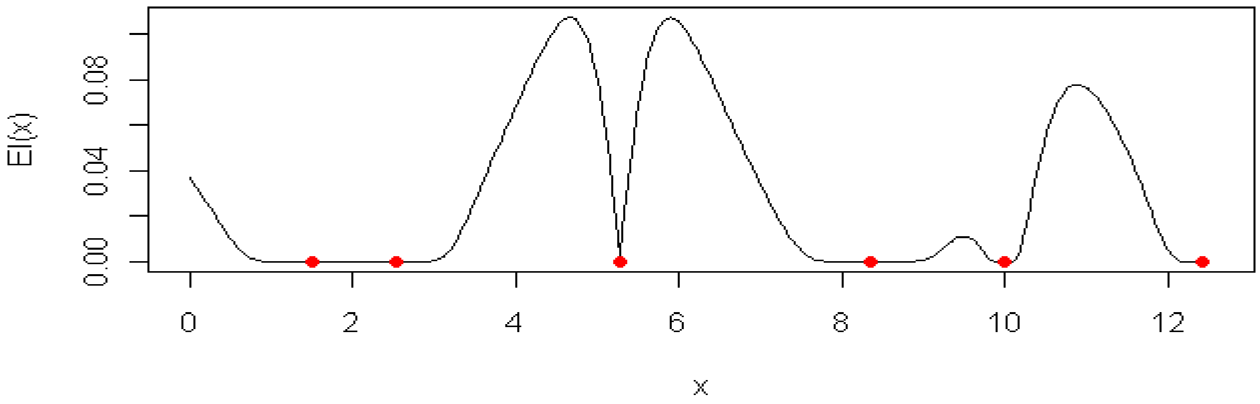
Adaptive design for optimization: EGO algorithm

[Jones et al. 1998]

EGO: step 0

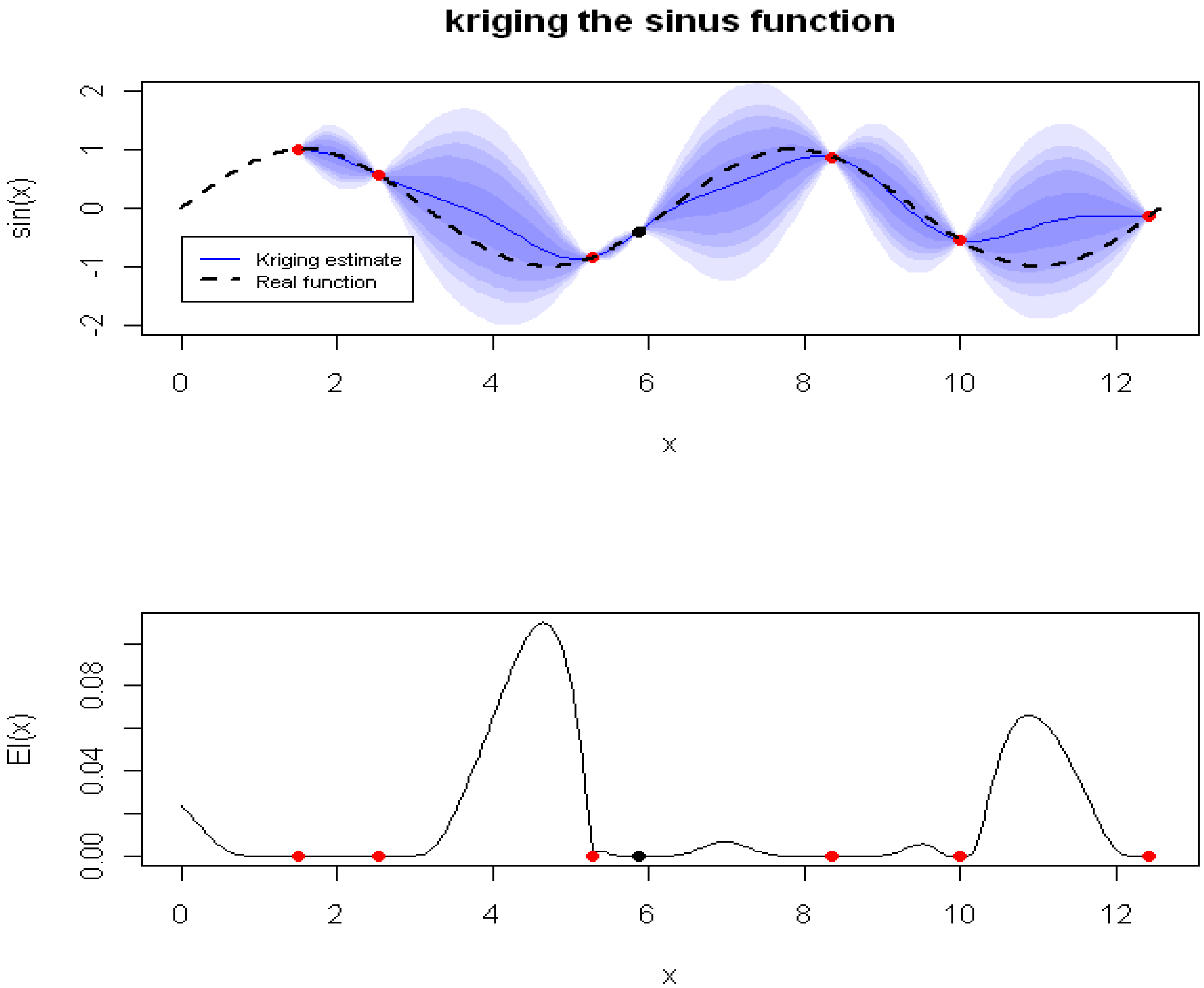


[Chevalier, 2011]



Adaptive design for optimization: EGO algorithm

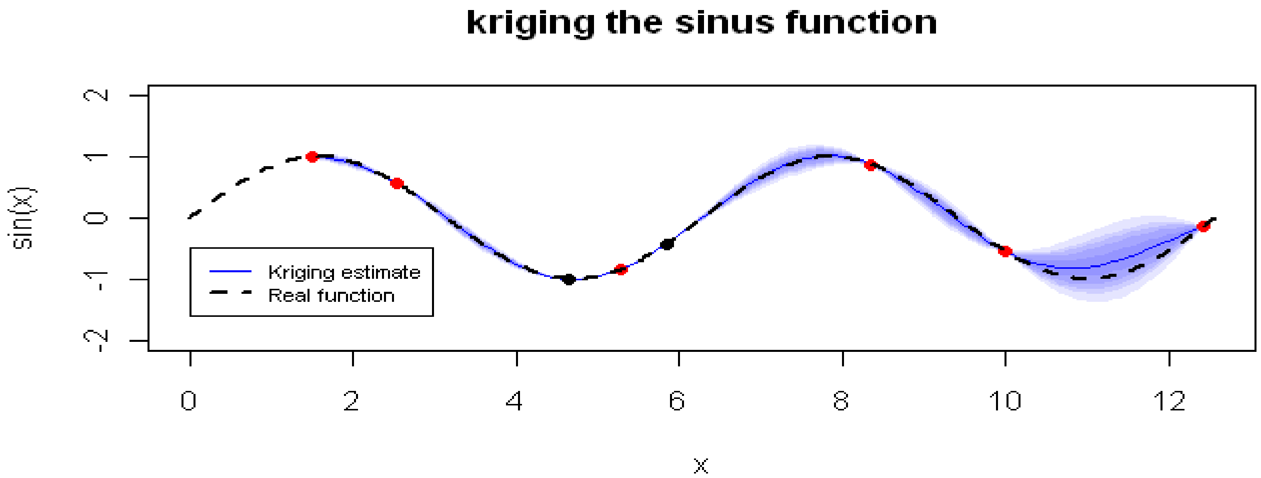
EGO: step 1



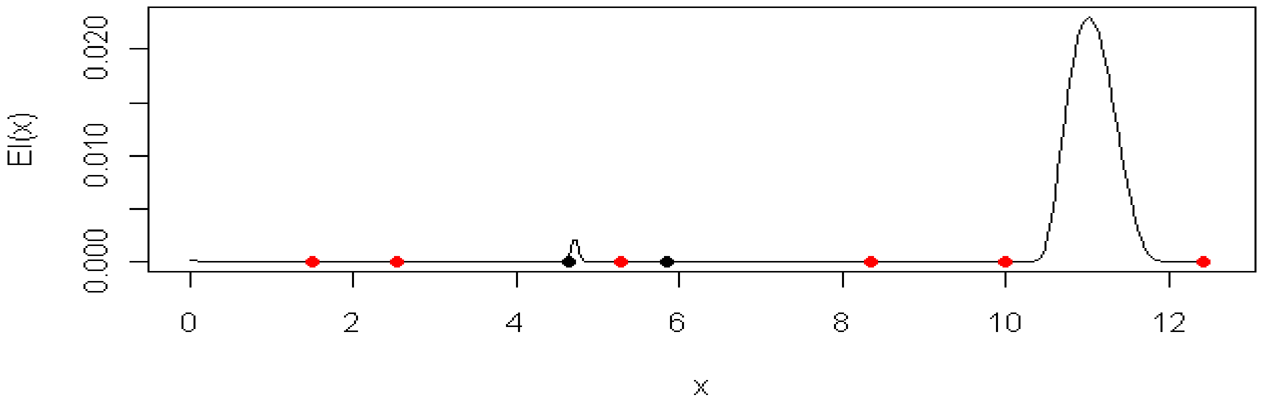
[Chevalier, 2011]

Adaptive design for optimization: EGO algorithm

EGO: step 2

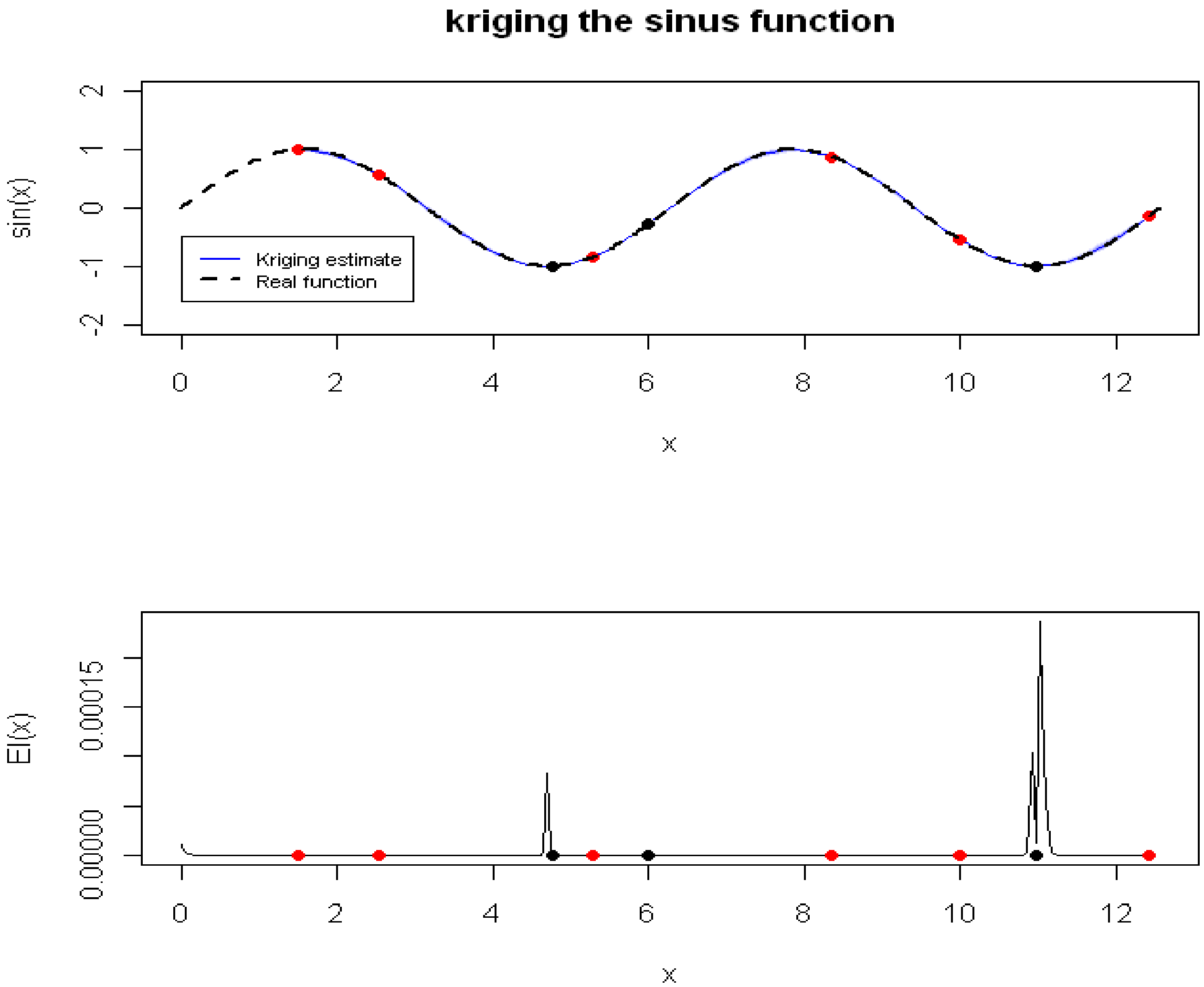


[Chevalier, 2011]



Adaptive design for optimization: EGO algorithm

EGO: step 3



[Chevalier, 2011]

Conclusions on the Gaussian process metamodel

- Gp model construction is possible even in high dimensional case

- Main advantage of Gp: probabilistic metamodel which gives confidence bands in addition to a predictor

- Fitting quality is dependent of the initial design
  Gp model is well adapted to sequential and adaptative designs

- Caveats : it can require a large amount of effort during the fitting process and cases with more than 1000 points begin to be difficult (matrix inversion)

- Designs for specific objectives (optimization, quantile, probability, etc.)

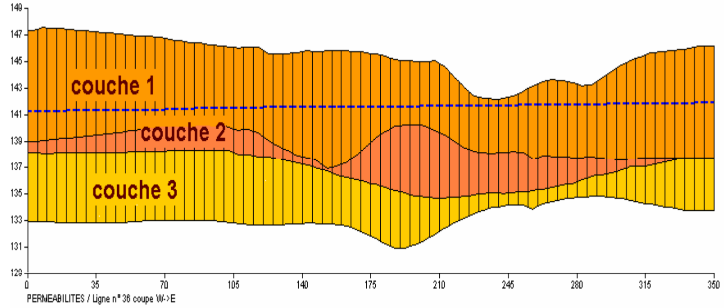
Application

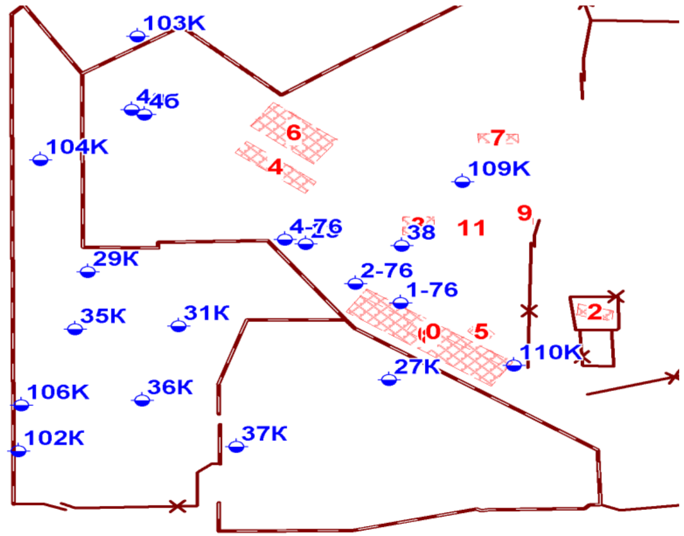
Motivating example: hydrogeological modeling

Collaboration Kurchatov Institute/CEA

[Volkova et al. 2008]

- Site (2 ha) near Moscow
- From 1943 to 1974 : radioactive waste repositories
- 1990 : site recognition with 20 piezometers
- Upper aquifer contamination in ^{90}Sr





Questions:

- impact assessment of the contamination on the environment

- degree of rehabilitation of the site

Uncertainty management in hydrogeological modeling

- Computation of the spatio-temporal evolution of ^{90}Sr concentration in an ancient radwaste disposal site between 2002 and 2010

- Goal: Estimate the contamination impact on the environment
  Identify the influent inputs on predicted outputs

Step A

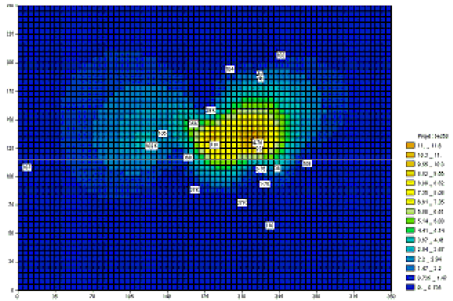
Numerical modeling:
Hydrogeological transport (Darcy's law) scenario of ^{90}Sr with the MARTHE software

***p* = 20 uncertain inputs X:**
Permeability, dispersivity, Kd, infiltration intensity, …

Output of interest:
Concentration values Y

Quantity of interest :
Distribution, variance

Carte de concentration initiale
Août 2002



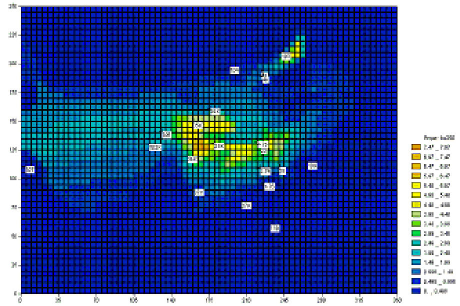
+

Jeu de valeurs
pour les 20 paramètres d'éntrée

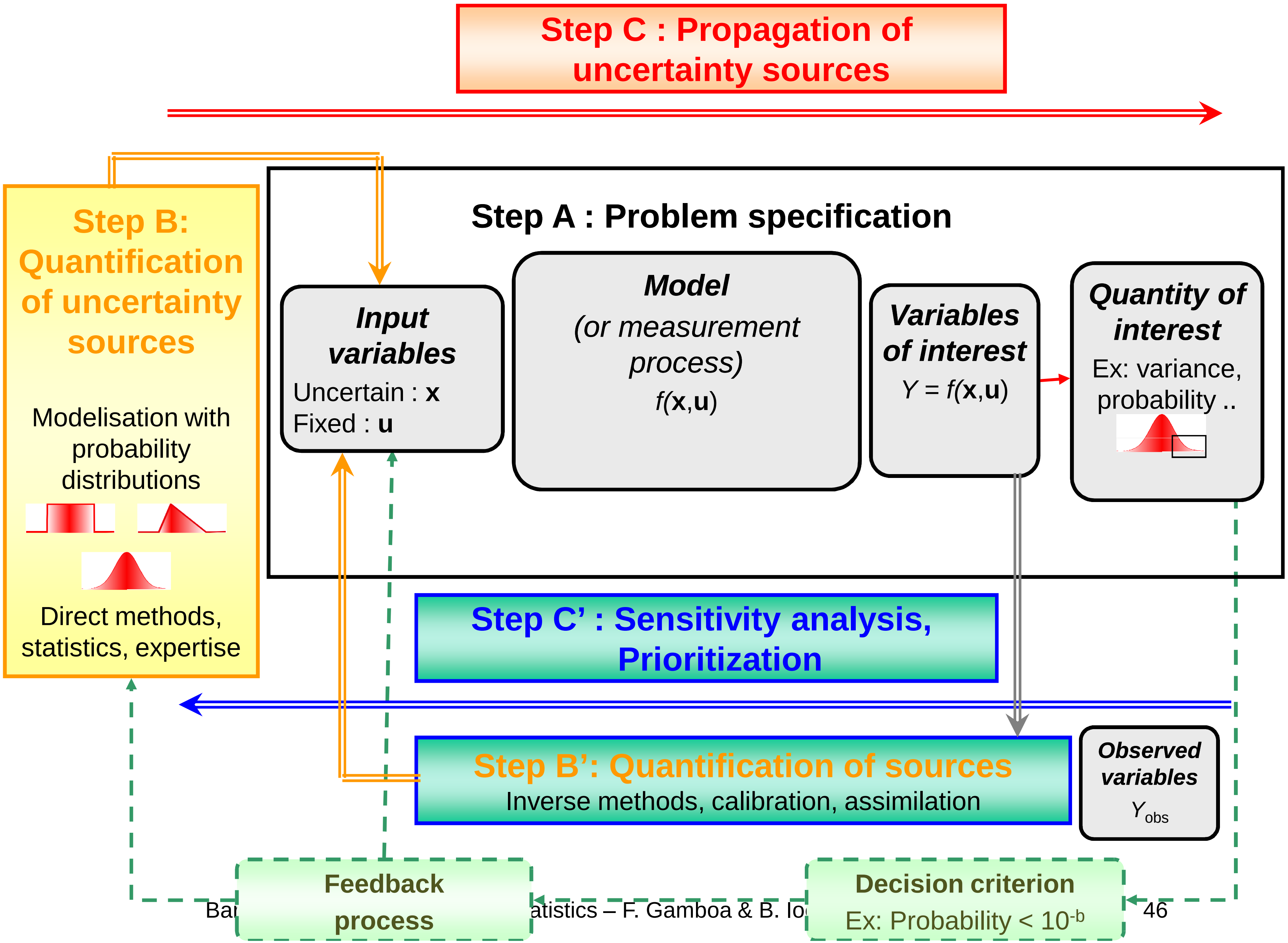
per1	perz3	dt1	kd3
per2	perz4	dt2	poros
per3	d1	dt3	i1
perz1	d2	kd1	i2
perz2	d3	kd2	i3

MARTHE →
30 mn / run

Carte de concentration finale
Décembre 2010



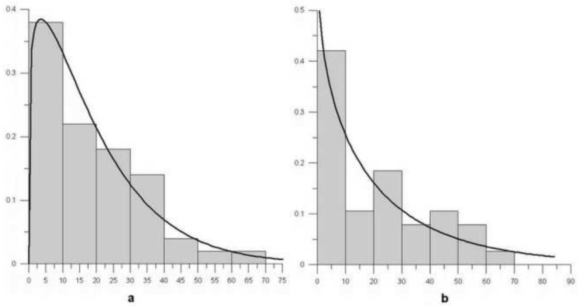
Uncertainty management - The generic methodology



Step B: Quantification of uncertainty sources

Statistical modeling of the uncertainty of each input

Different cases:



1. A lot of data
   - Fitting of probability distributions
   - Statistical hypothesis test

   MARTHE case : hydraulic conductivity (lognormal distribution)

2. Few data ($n < 10$)
   - Bayesian inference

3. No data
   - Bibliography
   - Expert judgment techniques

   MARTHE case : dispersivity, permeability, … (uniform distributions)

Step B: Quantification of uncertainty sources

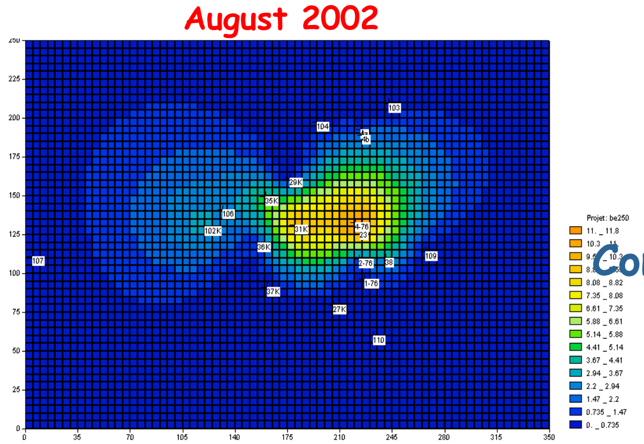
- **Input variables of the model**

→ nominal value, type of distribution and parameters

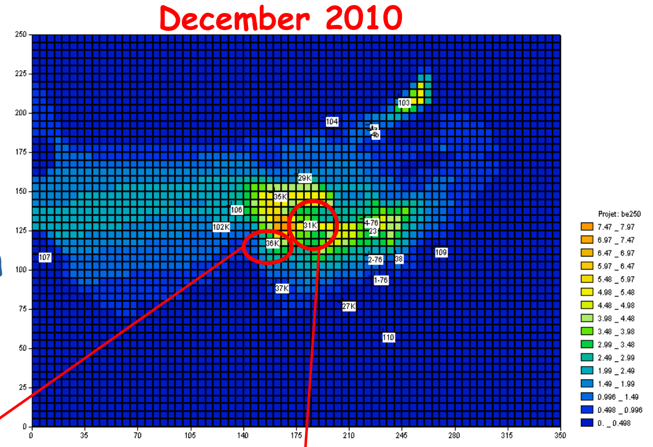
	Paramètres	Indicateur	Valeur du modèle	Type de distribution	Intervalle ou paramètres de distribution
1	Perméabilité couche 1	per1	8	Uniforme	1 - 15
2	Perméabilité couche 2	per2	15	Uniforme	5 - 20
3	Perméabilité couche 3	per3	8	Uniforme	1 - 15
4	Perméabilité zone 1	perz1	8	Uniforme	1 - 15
5	Perméabilité zone 2	perz2	8	Uniforme	1 - 15
6	Perméabilité zone 3	perz3	8	Uniforme	1 - 15
7	Perméabilité zone 4	perz4	8	Uniforme	1 - 15
8	Dispersivité longitudinale couche 1	d1	0,8	Uniforme	0,05 - 2
9	Dispersivité longitudinale couche 2	d2	0,8	Uniforme	0,05 - 2
10	Dispersivité longitudinale couche 3	d3	0,8	Uniforme	0,05 - 2
11	Dispersivité transversale couche 1	dt1	0,08	Uniforme	0,01*d1 - 0,1*d1
12	Dispersivité transversale couche 2	dt2	0,08	Uniforme	0,01*d2 - 0,1*d2
13	Dispersivité transversale couche 3	dt3	0,08	Uniforme	0,01*d3 - 0,1*d3
14	Coefficient de partage volumique c. 1	kd1	5,1	Weibull	1.1597, 19.9875
15	Coefficient de partage volumique c. 2	kd2	0,34	Weibull	0.891597, 24.4455
16	Coefficient de partage volumique c. 3	kd3	5,1	Weibull	1.27363, 22.4986
17	Porosité tous les couches	poros	0,3	Uniforme	0,3 - 0,37
18	Infiltration type 1	i1	0,0001	Uniforme	0 - 0,0001
19	Infiltration type 2	i2	0,004	Uniforme	i1 - 0,01
20	Infiltration type 3	i3	0,02	Uniforme	i2 - 0,1

Step C: Uncertainty propagation

Initial concentration

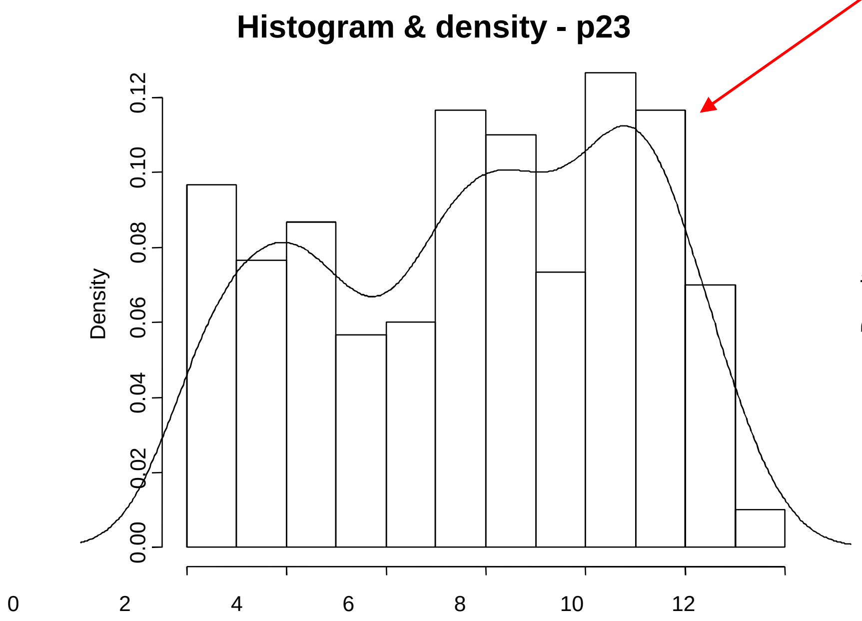


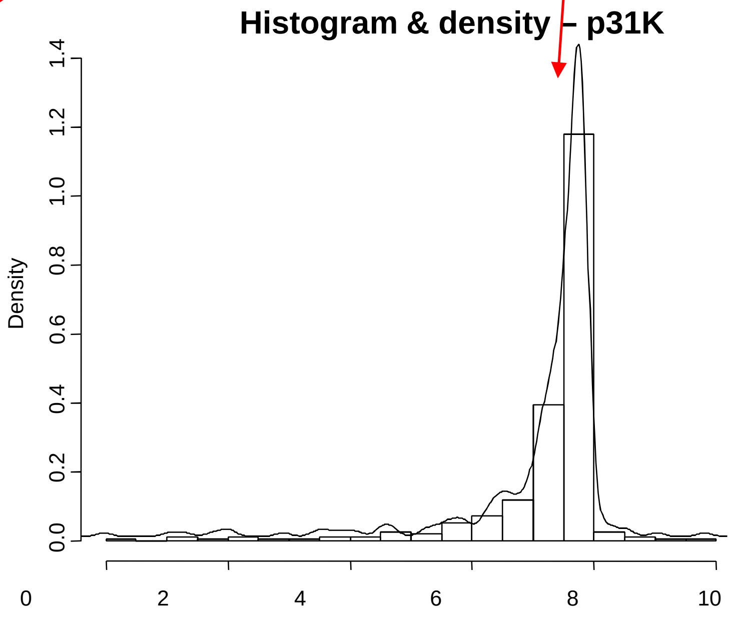
Concentration maps



N = 300 runs from Latin Hypercube Sample

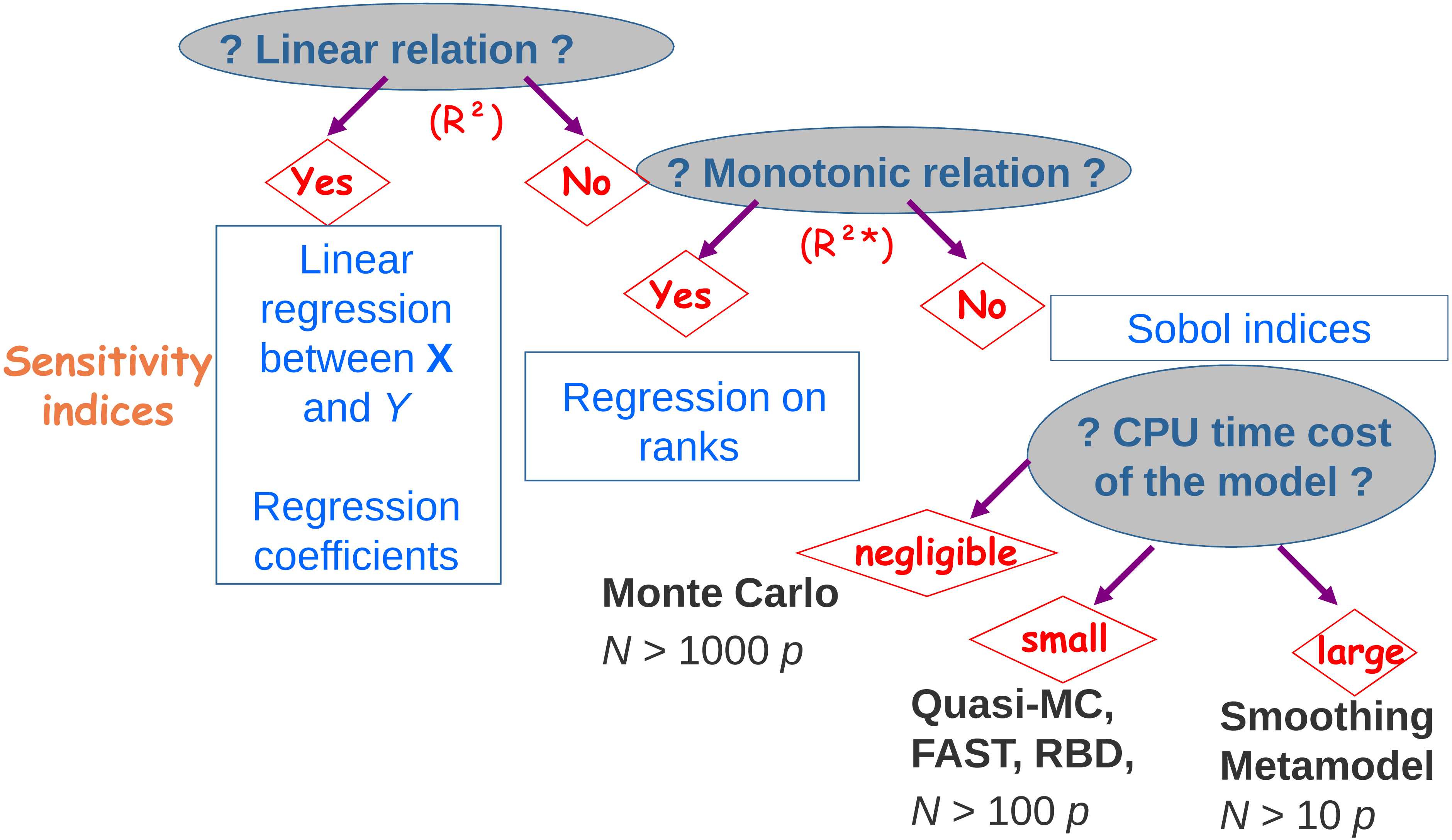
concentration distributions at the 20 piezometers (Bq/l)





Step C’: Sensitivity analysis - Sampling-based approaches

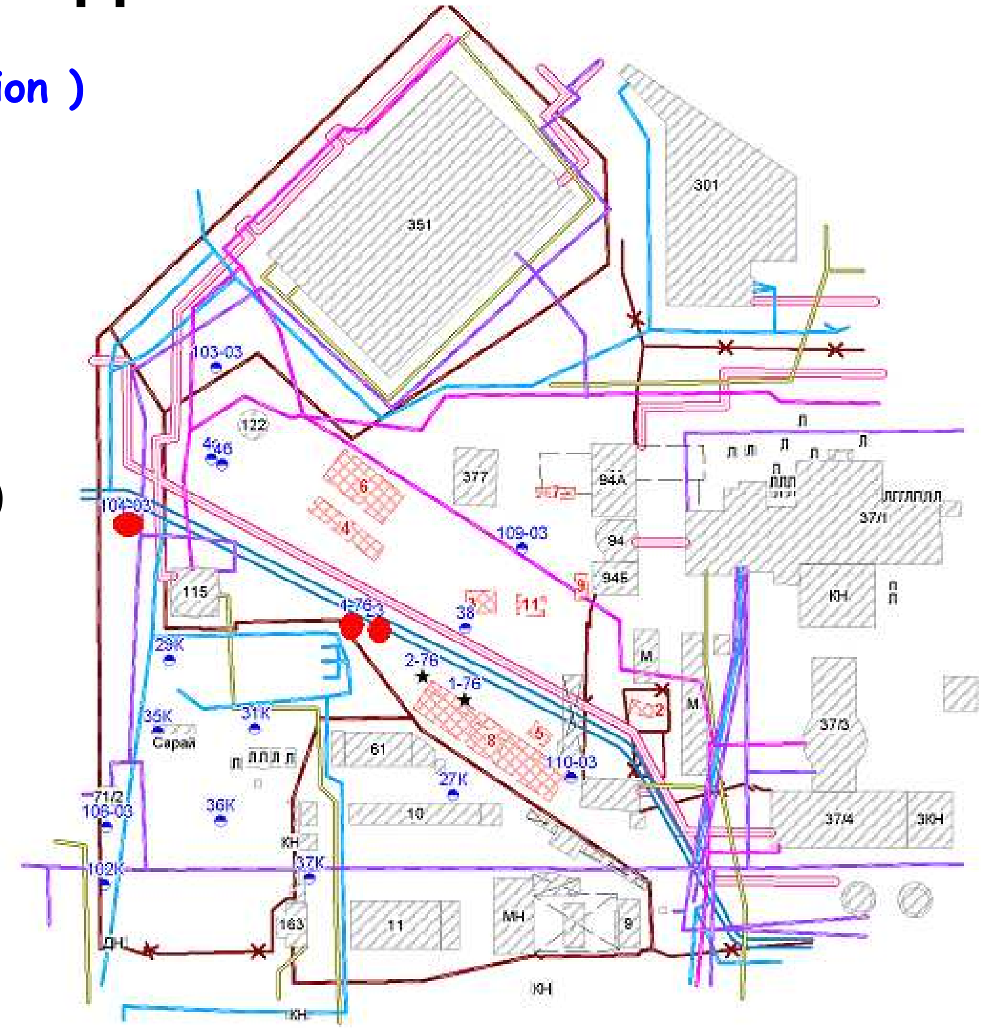
Sample ($\mathbf{X} \in \Re^p$, $Y(\mathbf{X}) \in \Re$) of size $N > p$



Step C’: Regression-based approach

- **Outputs with larger R^2** **(linear relation)**
- p23 (R^2 = 0,78) p104 (R^2=0,68)
- p4-76 (R^2 = 0,71)

- **Outputs with larger R^{2*}** **(monotonic relation)**
- p4-76 (R^{2*} = 0,95) p102K (R^{2*} = 0,90)
- p107 (R^{2*} = 0,92) p23 (R^{2*} = 0,90)
- p104 (R^{2*} = 0,91) p29K (R^{2*} = 0,83)



- **Most influent inputs**
- Distribution coefficient, layer 2
- Distribution coefficient, layer 1
- Infiltration intensity
- permeability, layer 2

PROBLEM: 14 outputs are non monotonic and we have only 300 simulation runs

Sensitivity analysis results for one scalar output « p104 »

Gaussian process (Gp) metamodel :

Predictivity coefficient: Q_2 = 93% - Linear regression : Q_2 = 68%

Sobol indices estimation + confidence intervals

(en %)	SRC_i^2 (linear regression)	$\mu_i = E_\Omega[\tilde{S}_i]$ Gaussian process	IC- 90% $(\tilde{S}_i)$ Gaussian process
per1	2	8	[5 ; 11]
kd1	52	69	[56 ; 83]
i3	13	13	[10 ; 17]

This can be done for several outputs …

Step C’: sensitivity analysis for 20 scalar outputs

Main influent inputs

➢ **Group 1 : kd1**

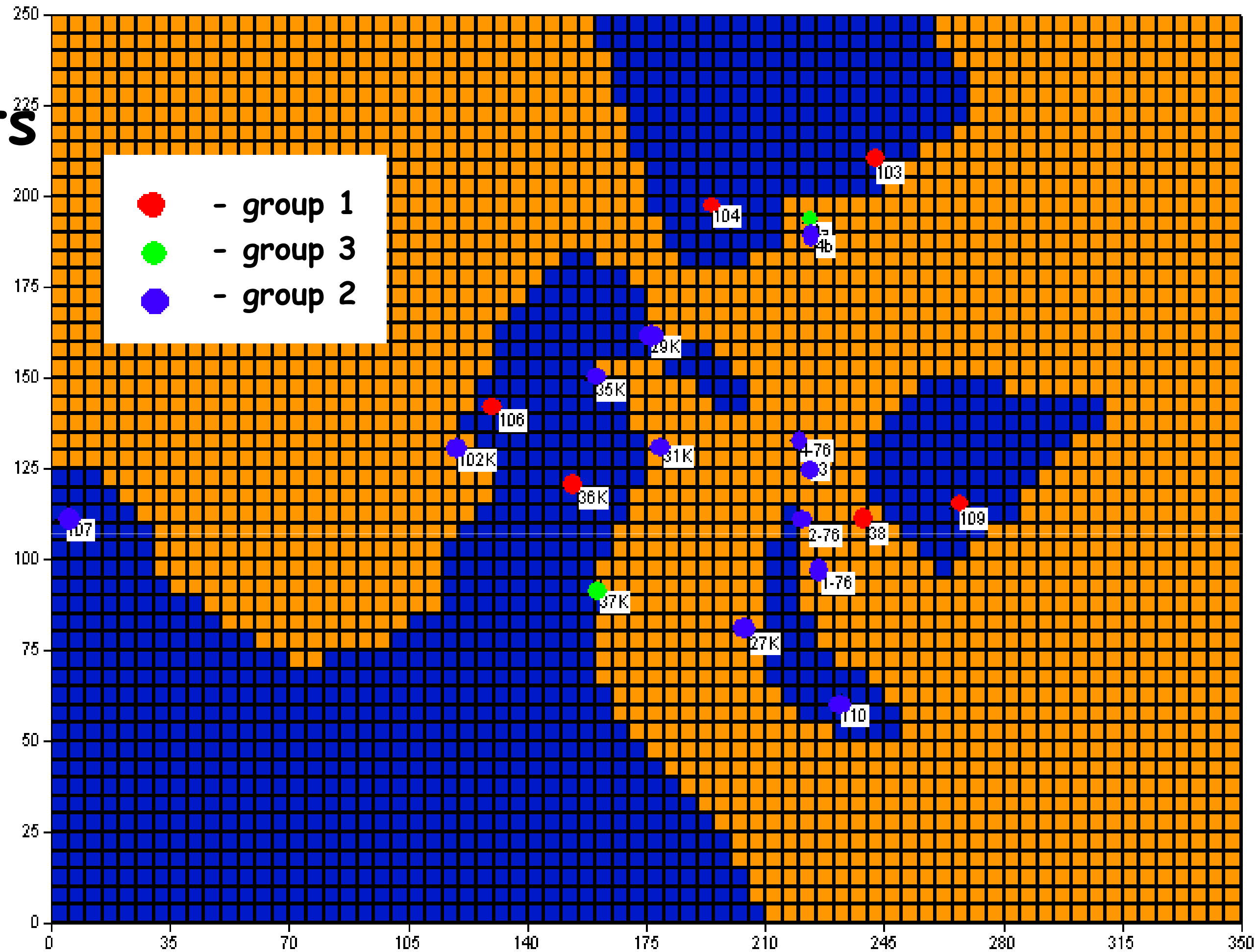
(distribution coef. of layer 1)

➢ **Group 2 : kd2**

(distribution coef. of layer 2)

➢ **Group 3 : i3**

(infiltration intensity)



Spatial location map

… but the results are difficult to synthesize, then to interpret

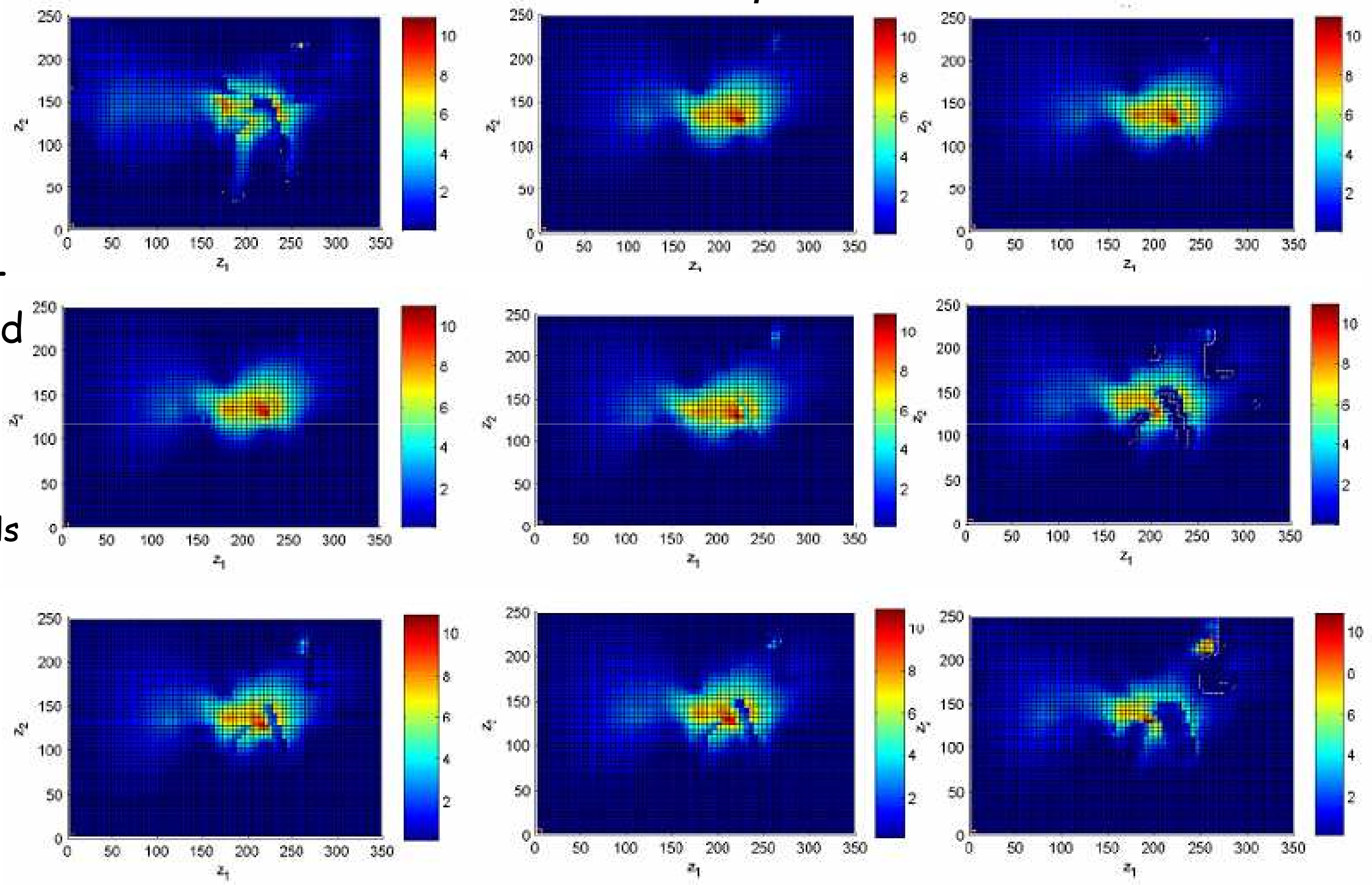
Considering functional (spatial) output

We have simulated N = 300 maps (Monte Carlo runs) with p = 20 random inputs

Concentration in ^{90}Sr predicted in 2010

9 output simulated maps

4096 pixels



Discretized spatial output can be considered as a functional 2D output

Sensitivity analysis when model ouputs are functions

Elementary cases (sensitivity analysis on each scalar output):

- Very small CPU time consuming model
- Linear or monotonic model

Difficult cases:

- Complex/Non linear model ⇒ need of Sobol indices
- CPU time expensive model ⇒ need of metamodel

Sensitivity analysis for spatial outputs: methodology

[Marrel et al. 2011]

- Computer code $f(.)$:

  Input: $\mathbf{X} = (X_1, \ldots, X_p)$ random vector

  Output for input $\mathbf{x}^*$: $y = f(\mathbf{x}^*, \mathbf{z})$, $\mathbf{z} \in D_{\mathbf{z}} \subset \mathbb{R}^2$

In practice, $D_{\mathbf{z}}$ is discretized in $n_{\mathbf{z}}$ points (here: 64 x 64 = 4096 points)

$(\mathbf{X}, Y(\mathbf{X},\mathbf{z}))$ = input-output sample of size N

- **Decomposition of $Y(\mathbf{z})$ on an orthogonal function basis** (fixed basis)

For example, a wavelet basis is well-suited if there are discontinuities

- **Modeling of the decomposition coefficients by a Gp metamodel**

  Selection procedure of the most important coefficients

- **Prediction: $\mathbf{x}^*$** => prediction of coeff. => spatial output map reconstruction

Functional metamodel → **Sensitivity analysis :** Spatial maps of sensitivity indices

Metamodel fitting: methodology for spatial output

- Step 1 : Wavelet decomposition of each map (300 maps)
- Step 2 : Kriging metamodeling of the main wavelet coefficients (the most variables) in fonction of X ; constant for the other coef.
- Step 3 : Prediction for a new input x*
  x* => prediction of the coefficients => spatial output map reconstruction
- Example of use: Sensitivity analysis (spatial map of sensitivity indices)

Sensitivity indices (Sobol indices, based on functional ANOVA):

$$S_i = \frac{Var_{X_i}[E(Y|X_i)]}{Var(Y)} \quad ; \quad S_{ij} \quad ; \quad \ldots \quad ; \quad S_{Ti} = S_i + \sum_j S_{ij} + \sum_{j,k} S_{ijk} + \ldots$$

1st order — 2nd order — Total indices

The estimation of Sobol indices require a large number of simulations

→ computation of the Sobol indices by the way of the metamodel

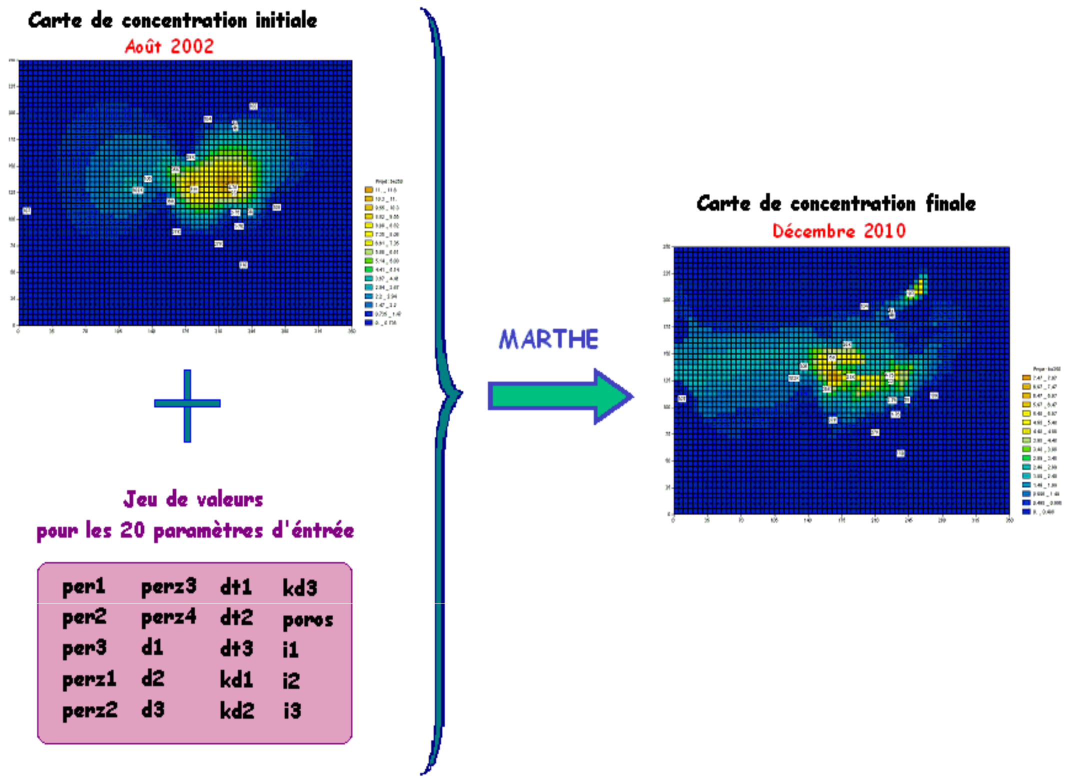
Application on our test case (hydrogeological pollution)

N = 300 simulations

p = 20 random input variables

K = 4096 pixels

k = 100 wavelet coefficients
modeled by Gaussian process



Mean predictivity (functional metamodel): $Q_2 = 72\%$

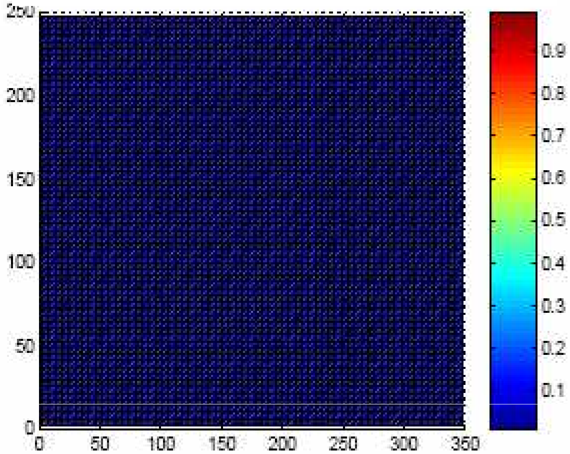
Estimation of first order and total Sobol' indices maps by Monte Carlo (22000 runs with the fonctional metamodel)

→ 20 maps of sensitivity indices

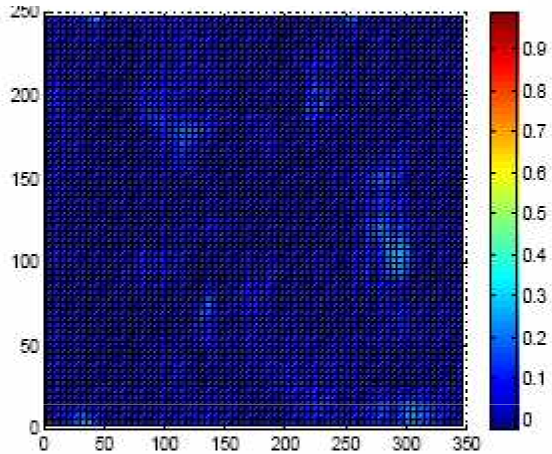
Spatial output: results of sensitivity analysis

Spatial maps of Sobol sensitivity indices of first order, for 6 inputs

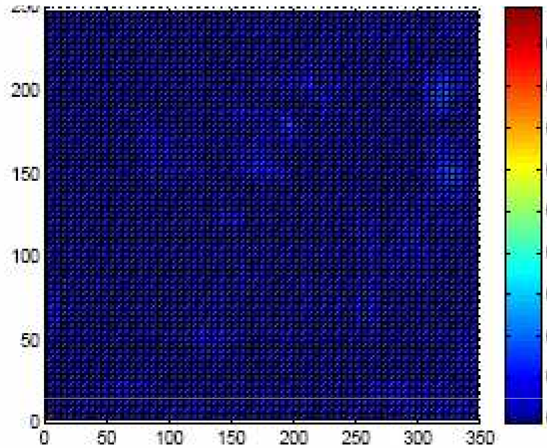
Permeability layer 1



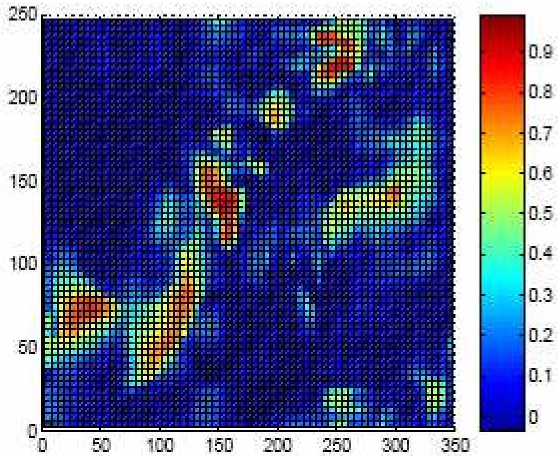
Permeability layer 2



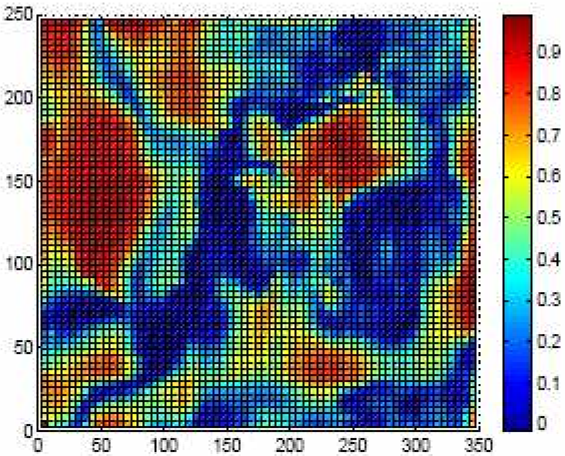
Permeability layer 3



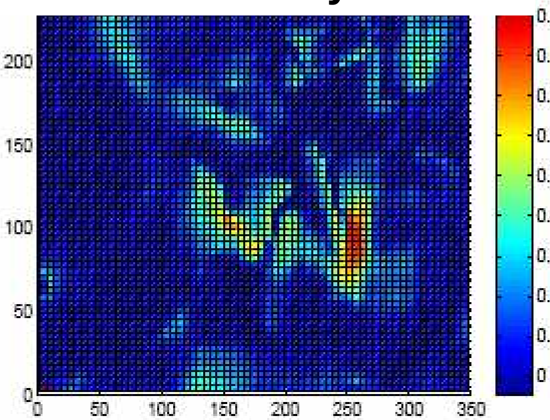
Kd layer 1



Kd layer 2



Strong infiltration intensity



Bibliography

- Chilès & Delfiner, Geostatistics, Wiley, 1999
- Fang, Li & Sudjianto, D*esign and modeling for computer experiments,* Chapman, 2006
- Hastie, Tibshirani & Friedman, *The elements of statistical learning theory,* Springer, 2002
- Kleijnen, *Design and analysis of simulation experiments,* Springer, 2008
- Marcotte, Cours de l'Ecole Polytechnique de Montreal
- Marrel et al., Global sensitivity analysis for models with spatially dependent outputs, *Environmetrics,* 2011
- T. Santner, B. Williams and W. Notz: The design and analysis of computer experiments. Springer, 2003.
- Volkova et al., *Stoch. Environ. Res. Risk Assess.,* 2008