Statistical analysis of simulation experiments:

Challenges for industrial applications

Bertrand Iooss

Commissariat à l'Energie Atomique

Direction de l'Energie Nucléaire

France

Main industrial stakes for Nuclear Energy

- NPP life time
- Fuel performance
- Safety, control of radiological release

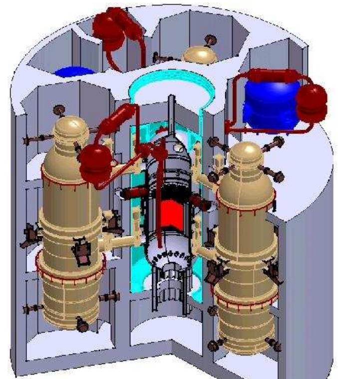
- Sustainability (better use of resources, recycling, management of waste)
- Non proliferation
- Dismantling

Present NPP

GEN-III

Continuity / present NPP

GEN-IV



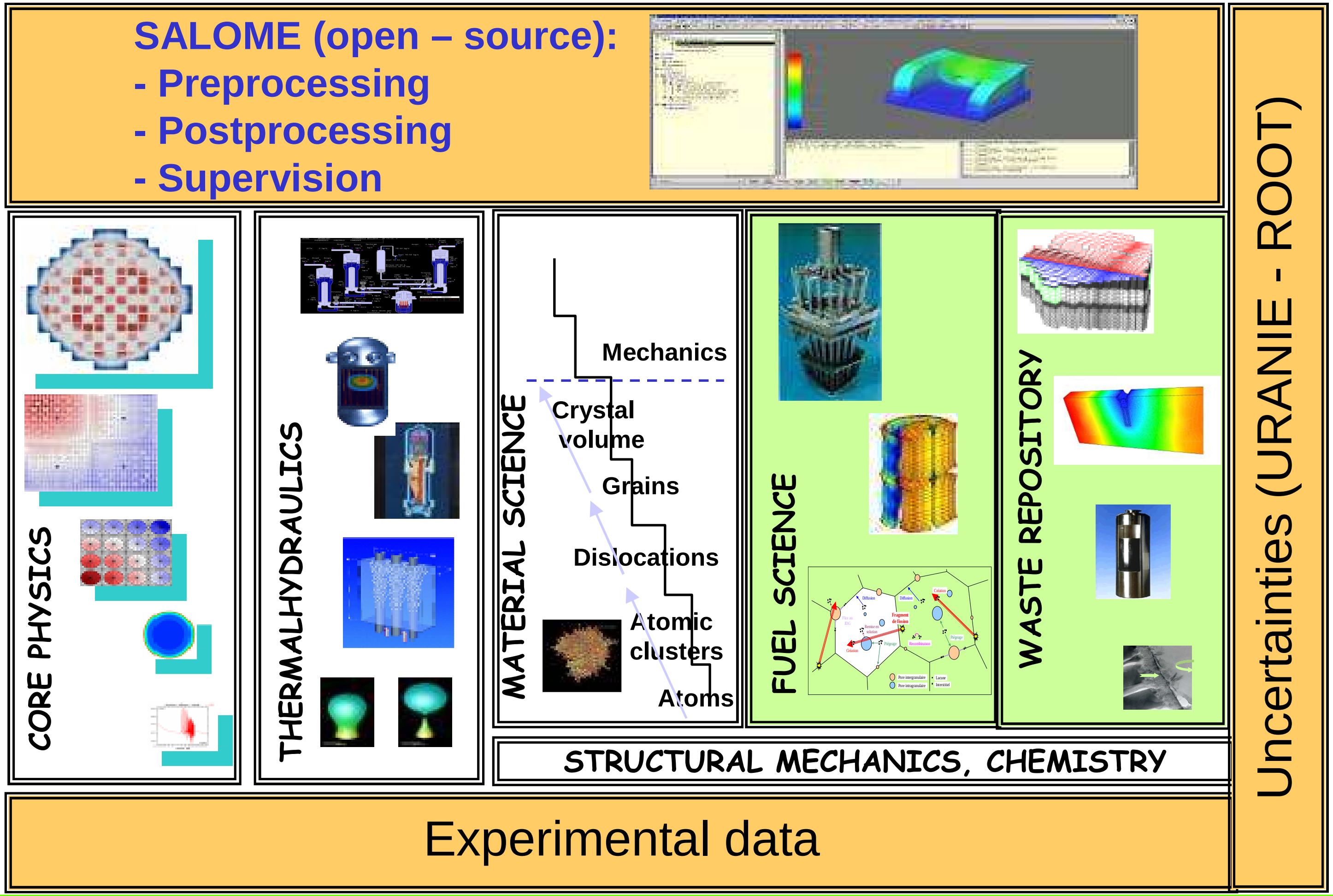
Technological rupture

Quantification of Uncertainties (operating conditions, unavailable data, physics ignorance, external events, …) is necessary to address these stakes

The DEN Simulation Platform for Nuclear Applications

Multiscale, Multiphysics, Multipurpose

- Partnership
- Portability
- Generic solutions



Outline

Challenge for industrial applications

1. Main difficulties for industrial dissemination : transfer to users

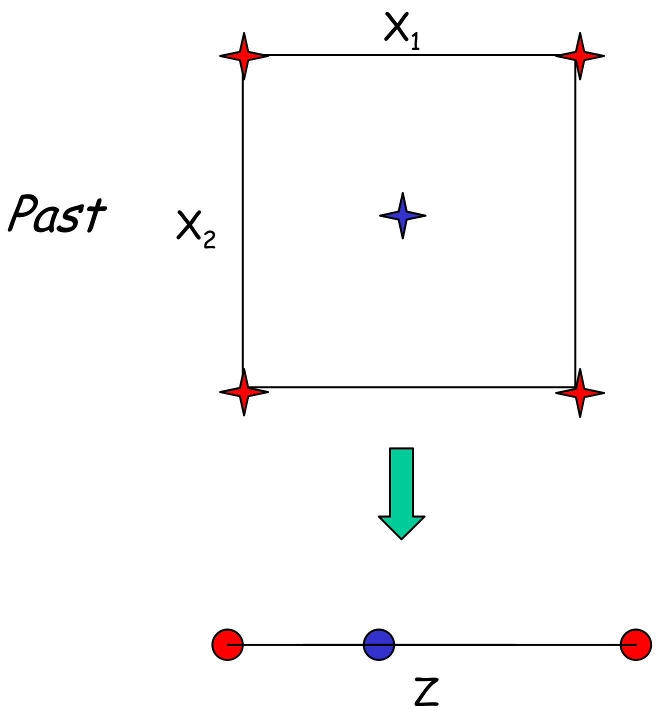
2. Example of a successful industrial dissemination

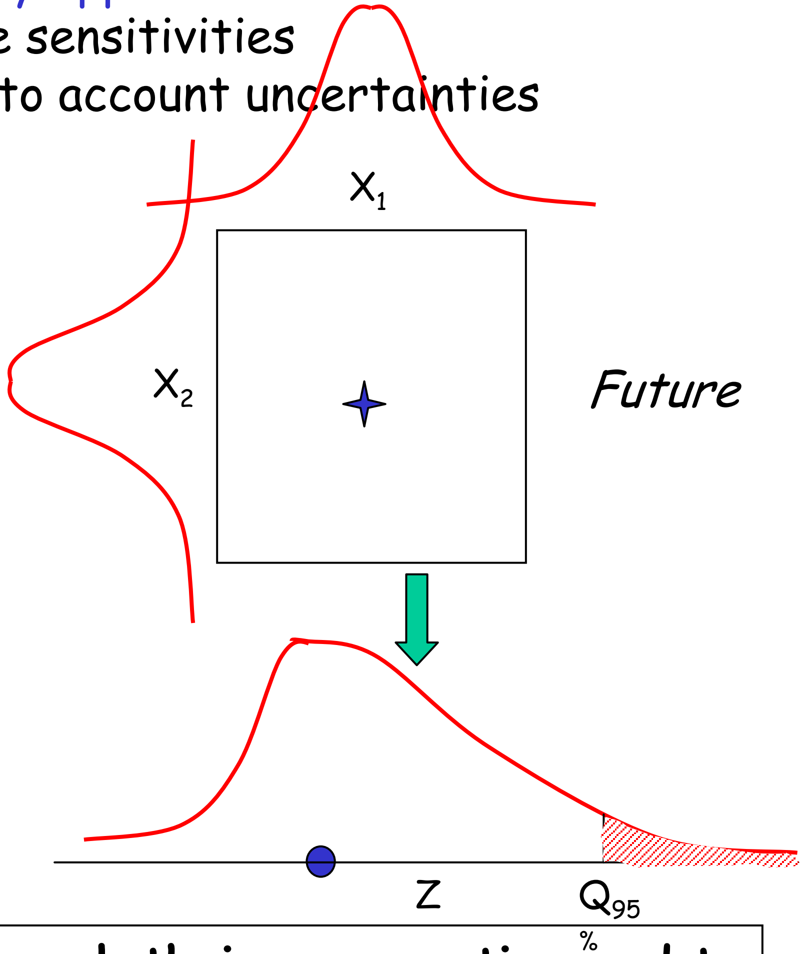
3. Some scientific challenges

1.0 On uncertainties in simulation experiments

We are all convinced that all **computer code modellers and users** should adopt a carefully thought out exploratory approach in order to :

- propagate their uncertainties, compute sensitivities
- design/optimize their system taking into account uncertainties
- give rigorous safety margins, …

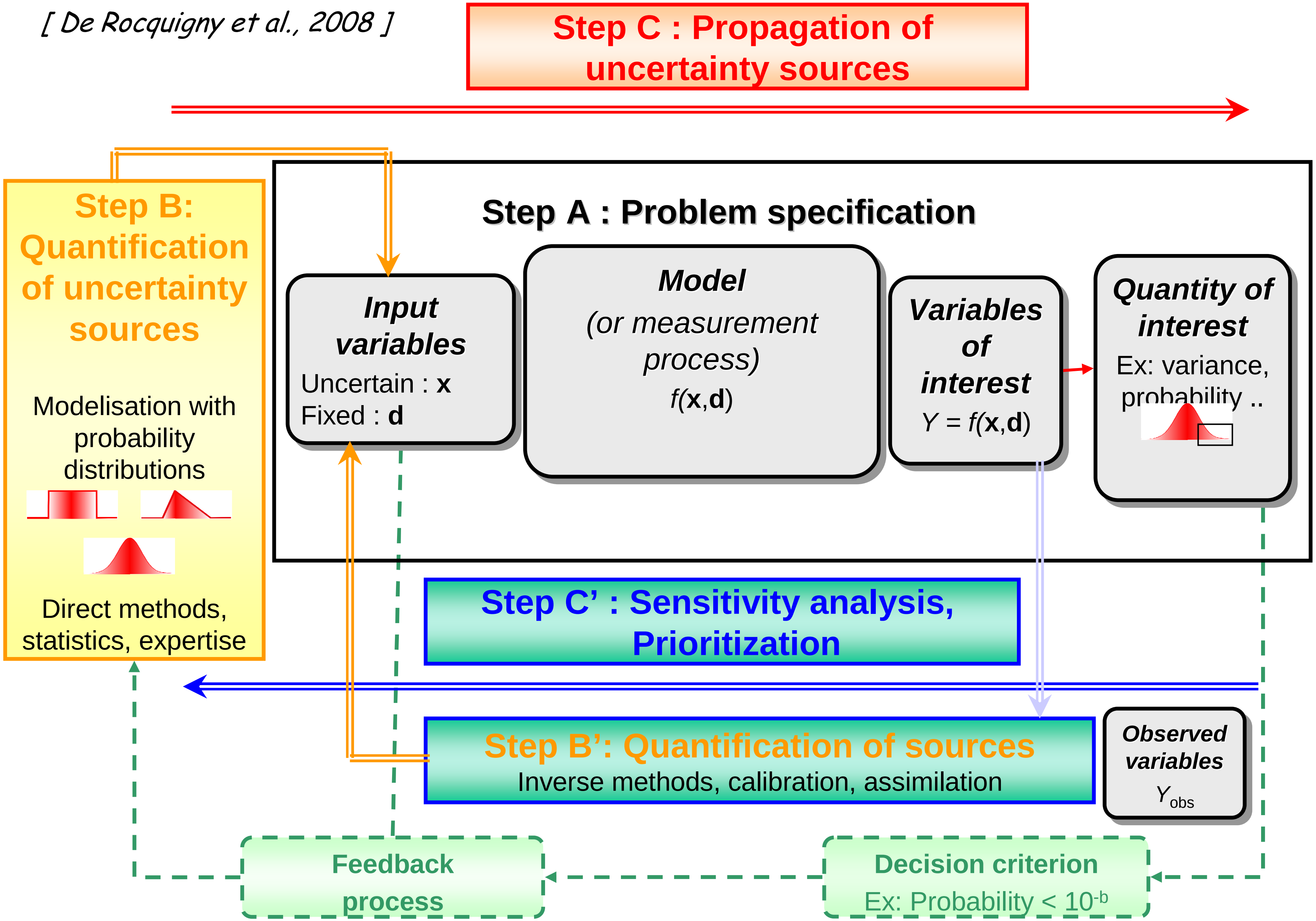




Once said that, how to convince them to break their conservatism and to pursue somme efforts to learn new methods ?

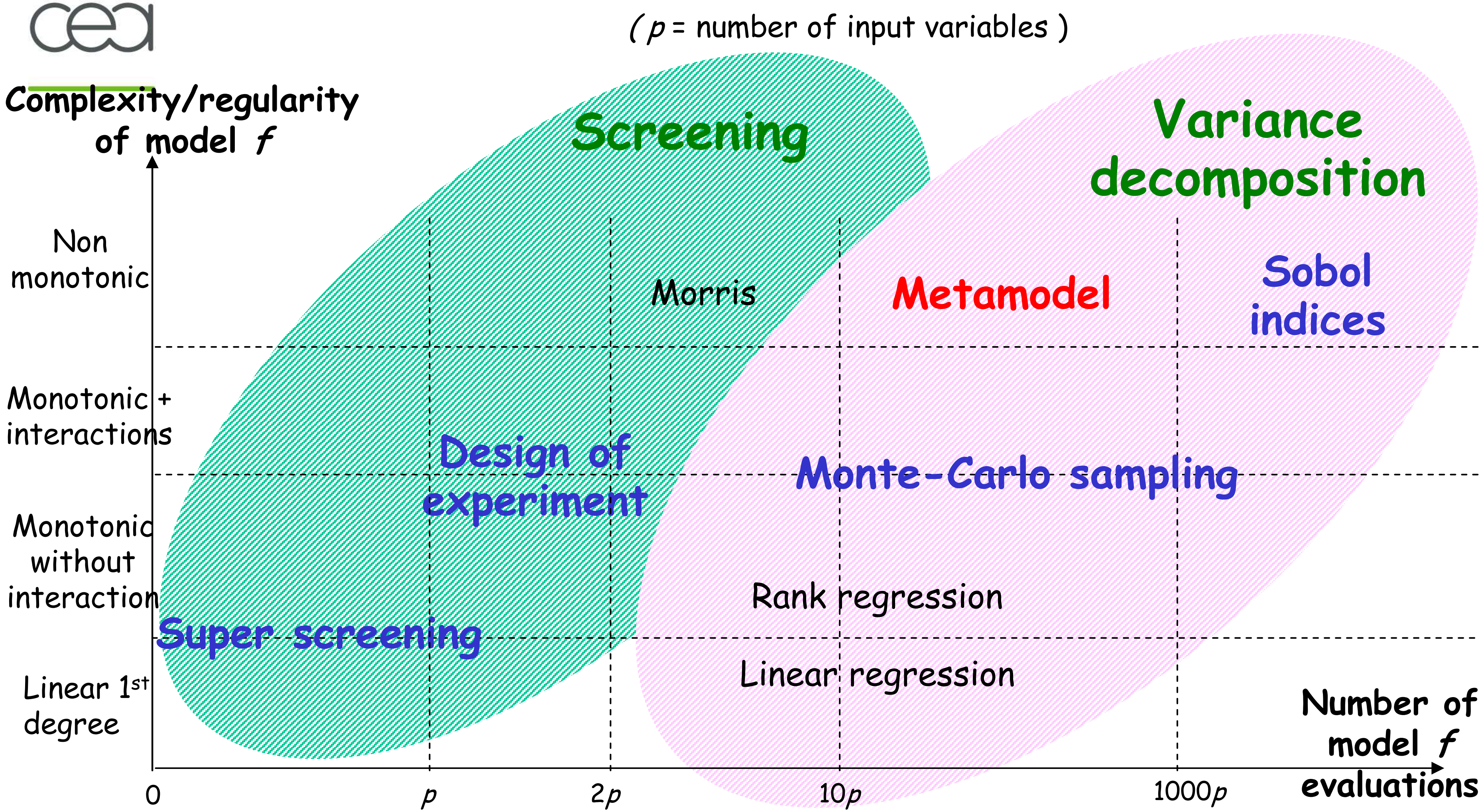
1.1 Bring clear (even schematic) methodology

[De Rocquigny et al., 2008]

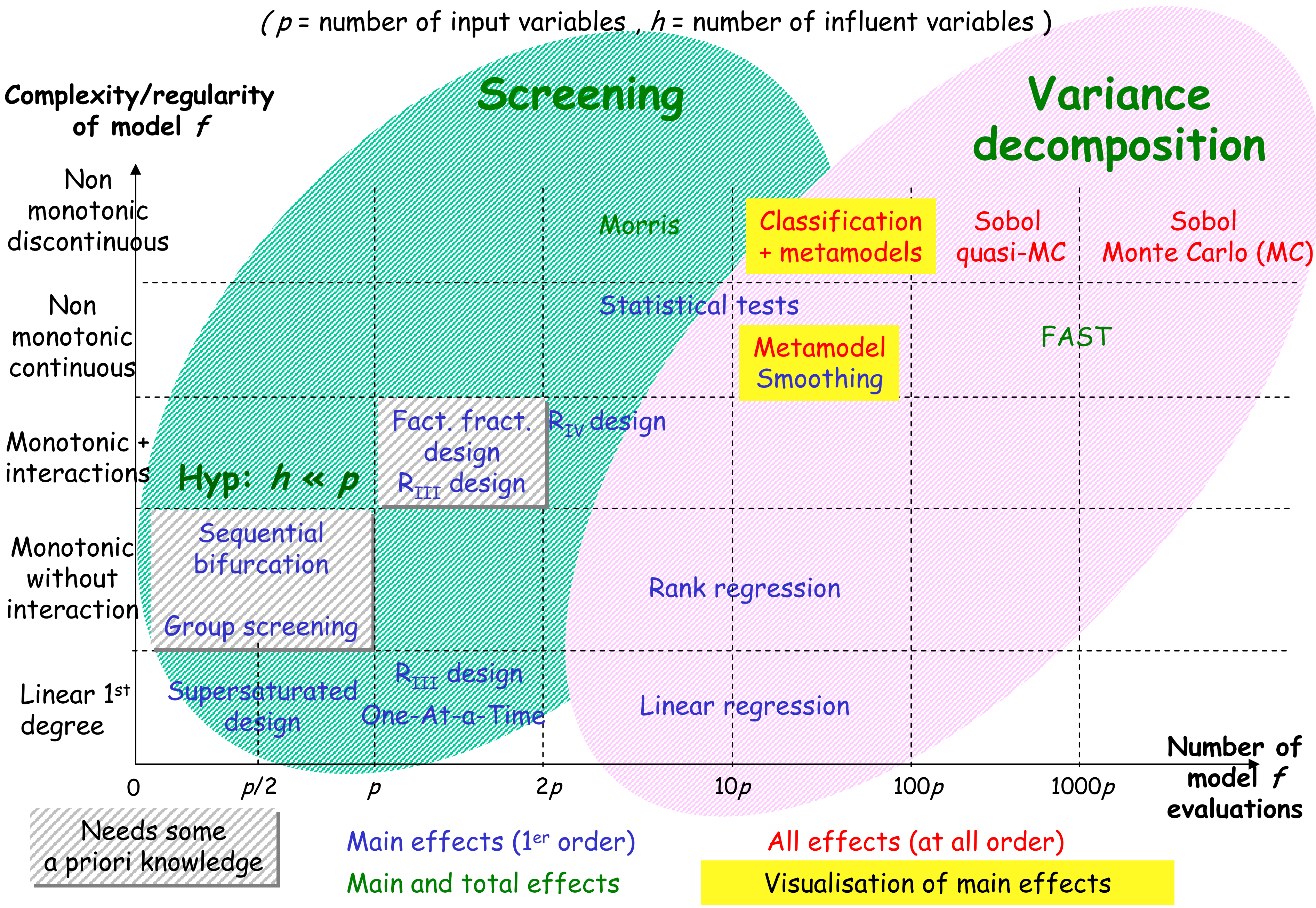


1.2 Give clear classification of most useful methods

Example : Classification of sensitivity analysis methods



1.3 Reserve details for (very) advanced practicioners



1.4 Offer strong services

• **Educating engineers**

Train our industrial colleagues and future engineers on past and modern statistical methods and the use of statistical softwares

⟹ Statistical education

• **Propose user-frienfly, all inclusive** (« clef en main ») **and integrated softwares**

⟹ Tempt users with easy-to-use softs (avoid additional difficulties)

• **Individualized steady for beginners** in order to avoid:

- mistakes
- strong disapointments

Outline

Challenge for industrial applications

1. Main difficulties for industrial dissemination : transfer to users

2. Example of a successful industrial dissemination

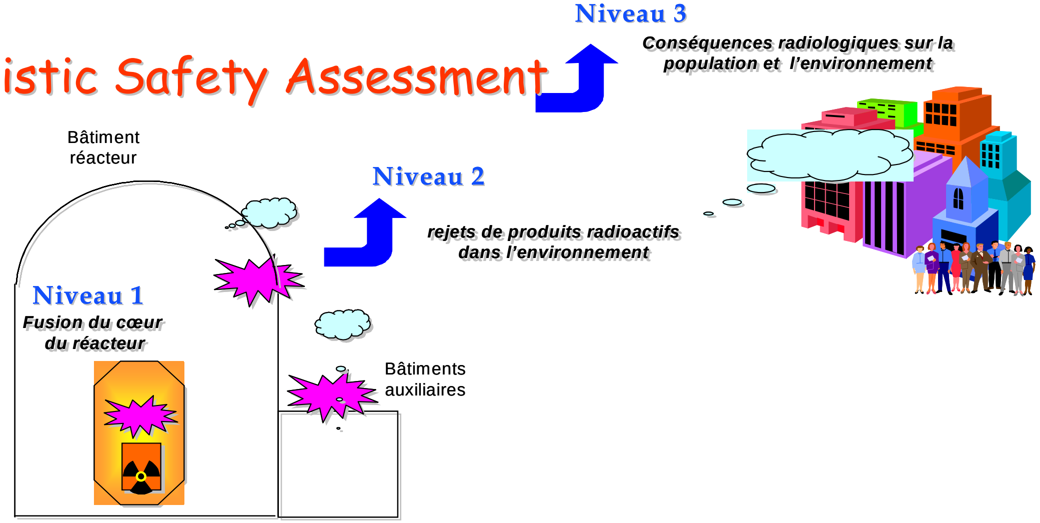
   The strength of the graphical tools

3. Some scientific challenges

The industrial stakes for nuclear PWR severe accident study

Important issue for the safety core of the industrial operator

Probabilistic Safety Assessment



Main questions after a severe accident (PSA level 2):

- Is there some radionuclide release outside the containment (and what is the time of this release) ?
- What is the efficiency of mitigation actions?
- Influence of the large parameter uncertainties on the failure probability?

EDF asked CEA to develop a software (LEONAR) based on its physical model expertise and on the CEA uncertainty software (URANIE)

Main objectives of these uncertainty studies

Complex and coupled numerical models which require sensitivity analysis to:

- Identify the main sources of uncertainties
- Understand the relations between uncertain inputs and interest outputs

Scenario :

Core degradation, corium transfer and interaction (inside in-vessel and ex-vessel)

23 Interest output variables :

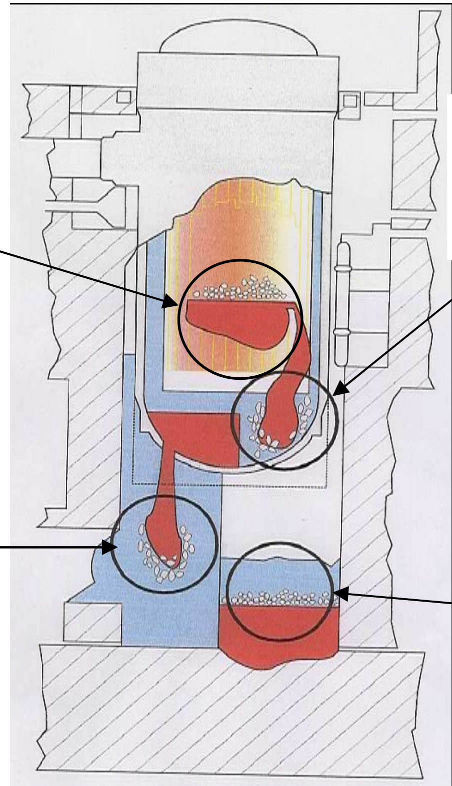
Corium masses

Time of vessel failure

Time of reactor pit failure

32 input random variables (uniform laws) :

Water management, physical properties, …



Additional problem: provide tools for engineers knowing nothing little about stats

3 use levels – 1) Precise and punctual needs

Example: relative study about the failure probability

Goal: **understand** the influence of water injection in the vessel and/or in the reactor pit during the accident

Random var. **X** : 22 uniform - 1 binary (for the water presence/absence)

N =500 random simulations (LHS design)

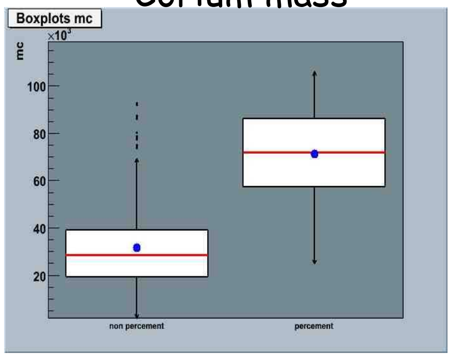
Failure probability is estimated by Monte-Carlo method

P(failure without water injection)/ P(failure with water injection) = 2.5

This kind of result could help the definition of management strategies

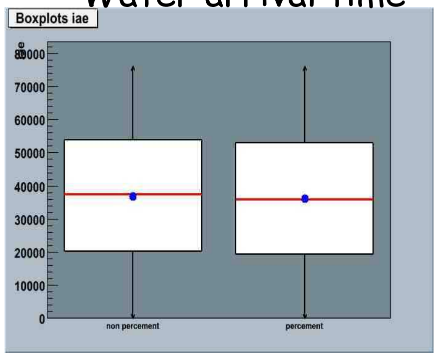
Sensitivity analysis via boxplots

Corium mass



No failure Failure

Water arrival time



No failure Failure

3 use levels – 2) Detailed analyses for confirmed users

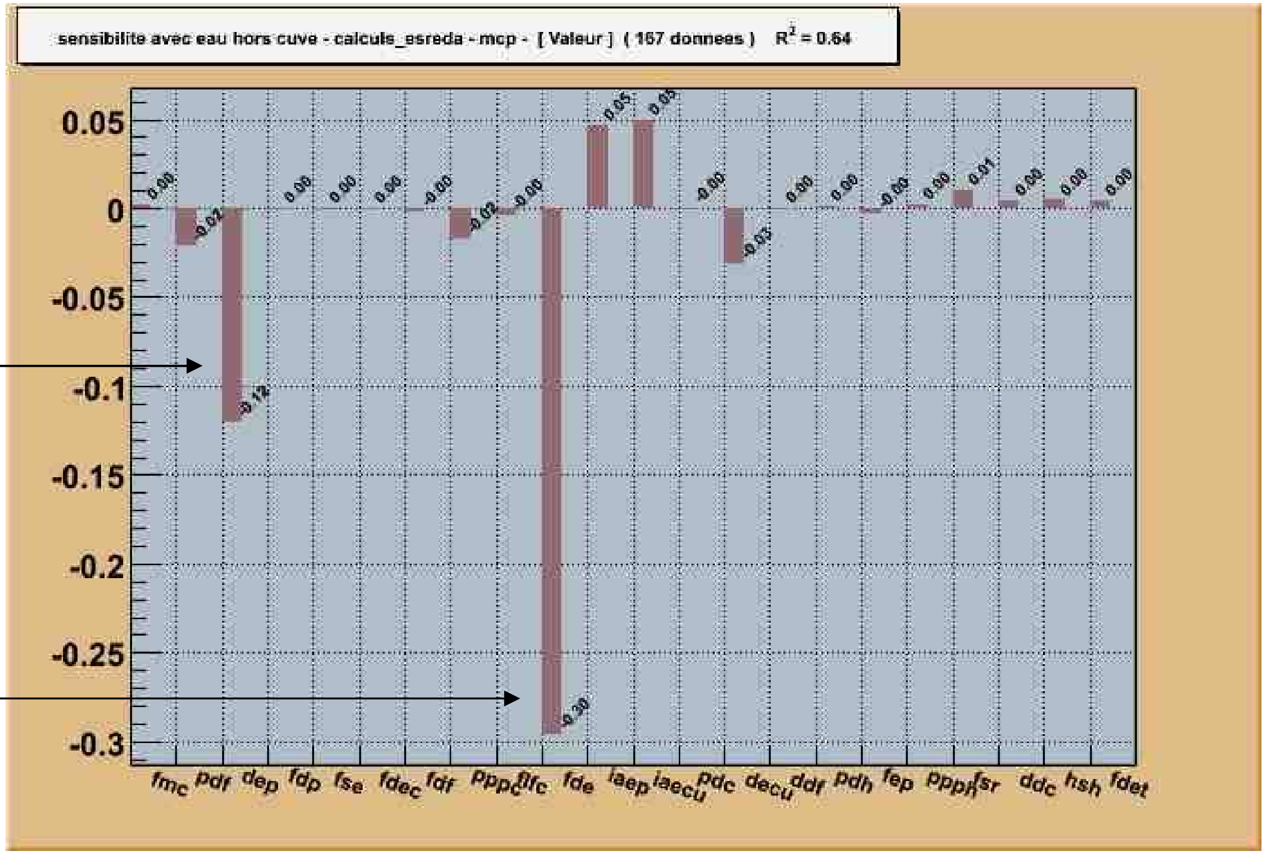
Example: sensitivity analysis related to an output variable (corium mass)

Goal: **understand** which input variables strongly affect the corium mass

$$SRC_i = \mathrm{sgn}(\beta_i)\,\frac{\beta_i^2\sigma^2(X_i)}{\sigma^2(Z)}$$

Water flow in the pit

Rubble fraction in the water

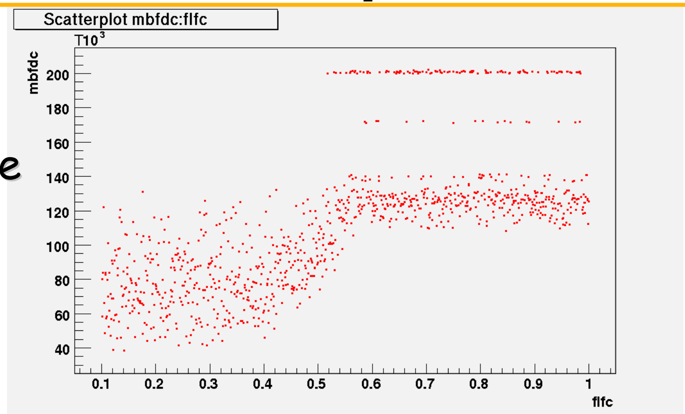


3 use levels – 3) Help to physical model developers

Example: detection of anomalies

500 MC simulations

Corium mass in the vessel bottom



Limit factor of the critical flux

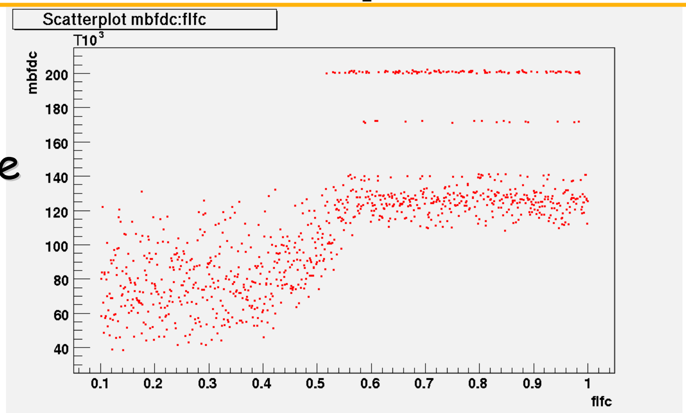
It has allowed to detect that important phenomena related to the corium transfer were not included in the physical model

3 use levels – 3) Help to physical model developers

Example: detection of anomalies

500 MC simulations

Corium mass in the vessel bottom

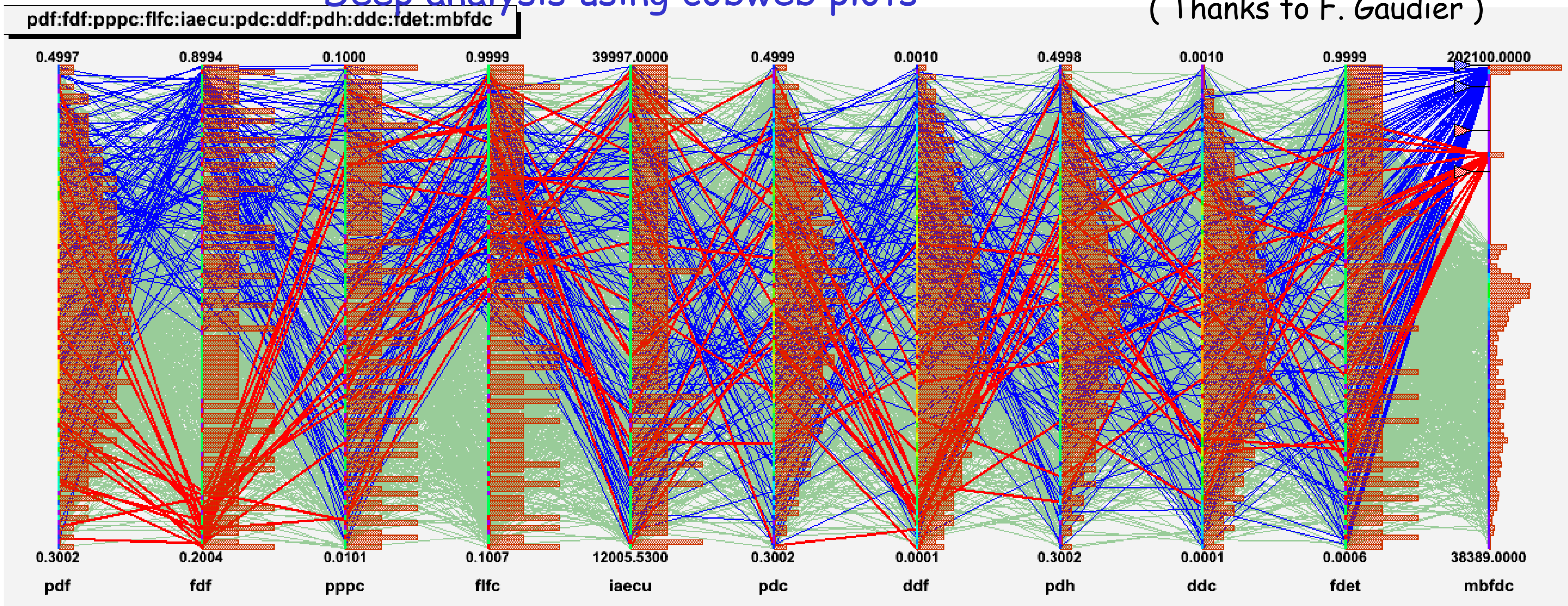


Limit factor of the critical flux

It has allowed to detect that important phenomena related to the corium transfer were not included in the physical model

Deep analysis using cobweb plots

(Thanks to F. Gaudier)



Conclusions of this small success story

Main objectives during the LEONAR development:

- To be used by engineers knowing nothing about uncertainties and very little about statistics

  ⇒ Easy-to-use software, **analyses with graphics**
  ⇒ Parcimony principle: LEONAR contains only the necessary basics (elementary functionalities of URANIE),
  ⇒ Abundant documentation and training.

- To be used by modelers interested by exploring their models

  ⇒ Dealing with large number of input and output variables,
  ⇒ Possibility for the users to perform an advanced level analysis via the ROOT environment (used by URANIE).

After a few years of use, we will integrate more advanced/complex methods (*e.g.* metamodels and related designs, Sobol indices, …)

Outline

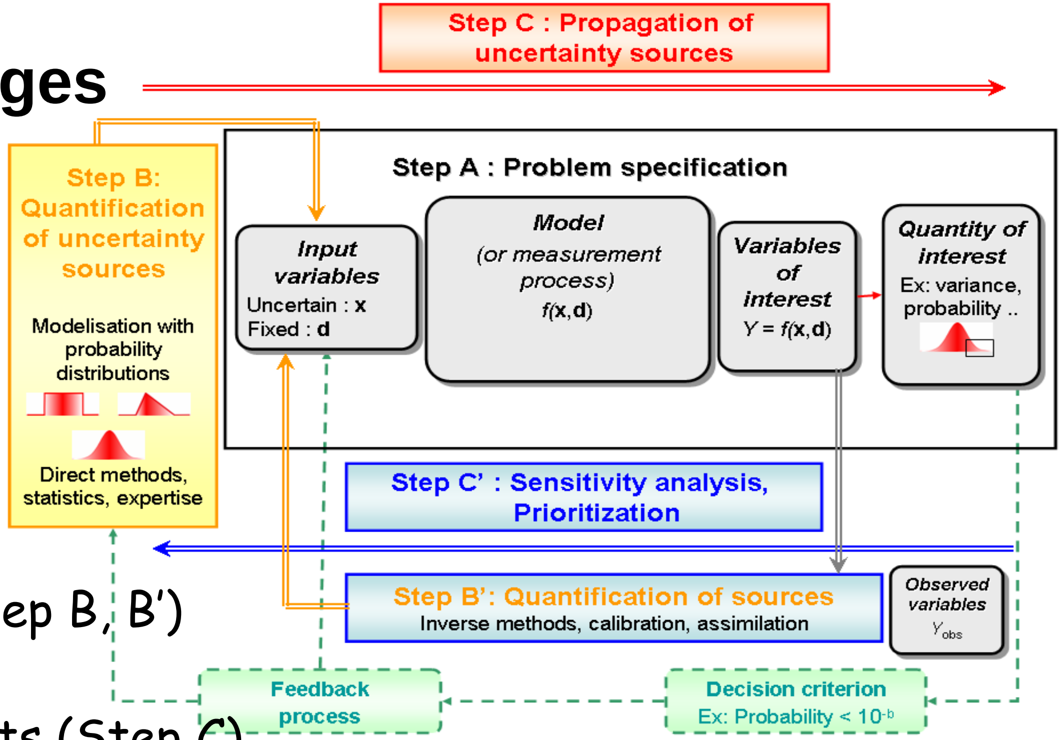
Challenge for industrial applications

1. Main difficulties for industrial dissemination : transfer to users

2. Example of a successful industrial dissemination

3. Some scientific challenges

A lot of scientific challenges



1. Uncertainty representation (Step B, B')
2. Design of simulation experiments (Step C)
3. Estimation of extremal events (Step C)
4. Treatment of functional input and output variables (Steps B, B', C, C')
5. Mixing variational/stochastic methods (Steps C, C')
6. Stochastic computer codes (Steps C, C')

3.1 Combine many representations of inputs

- **Qualitative (categorial) factors**

Present in a lot of situations

For ex., in CFD computer codes (numerical scheme choice)

Recent works have introduced Gp metamodel mixing continuous and categorical inputs *[Qian et al., 2008]*

Main challenge : build specific designs ?

- **Uncertainty modelling**

Integrate and aggregate expert information ?

Extra probabilistic representation :

Single probability / Bayesian approach (BA) / Imprecise probability (IP)

Goal: Model the ignorance about the probability laws

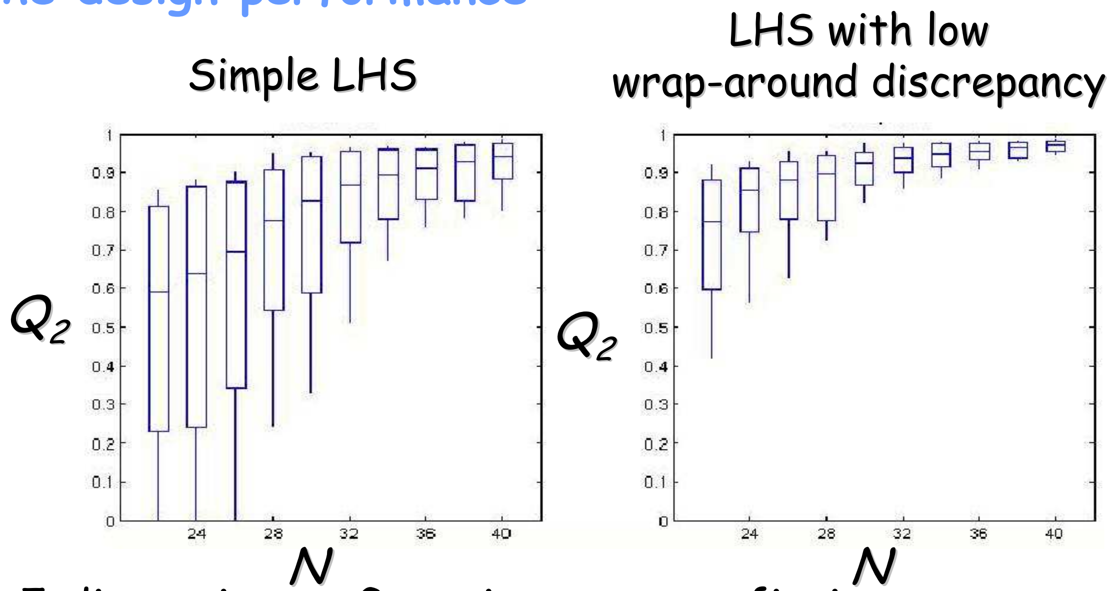
A lot of work have to be done to explain and use BA and IP for computer experiments

3.2 Best initial designs for metamodel fitting

Best initial design to use with deterministic computer experiments?

Ex: LHS, optimal LHS (correlation-based and/or distances-based), uniform, low-discrepancy, Strauss

- Comparisons about the design performance



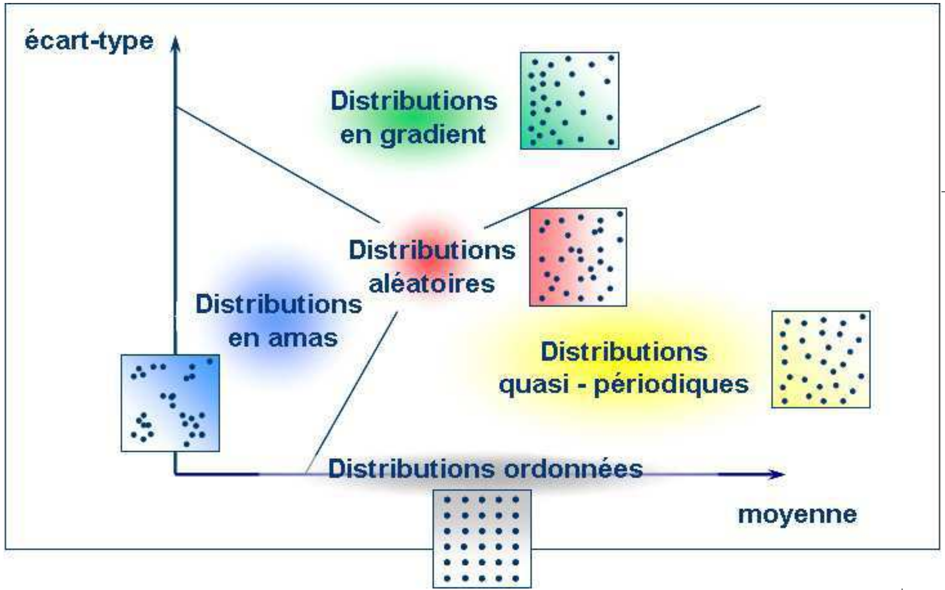
Sobol g-function in 5 dimensions - Gaussian process fitting *[Marrel, 2008]*

- Intensive numerical tests are needed
- **Theoretical arguments ?**

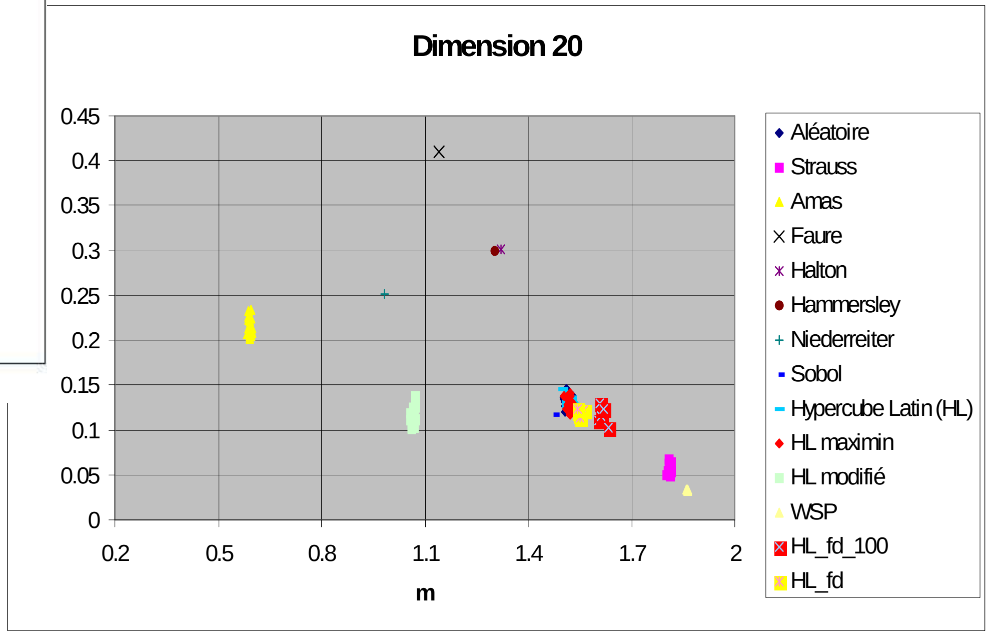
3.2 Best initial designs for metamodel fitting

Best initial design to use with deterministic computer experiments?

- Distinction of spatial patterns in high dimensional space



Criteria based on the Minimal Spanning Tree *[Franco et al., 2009]*

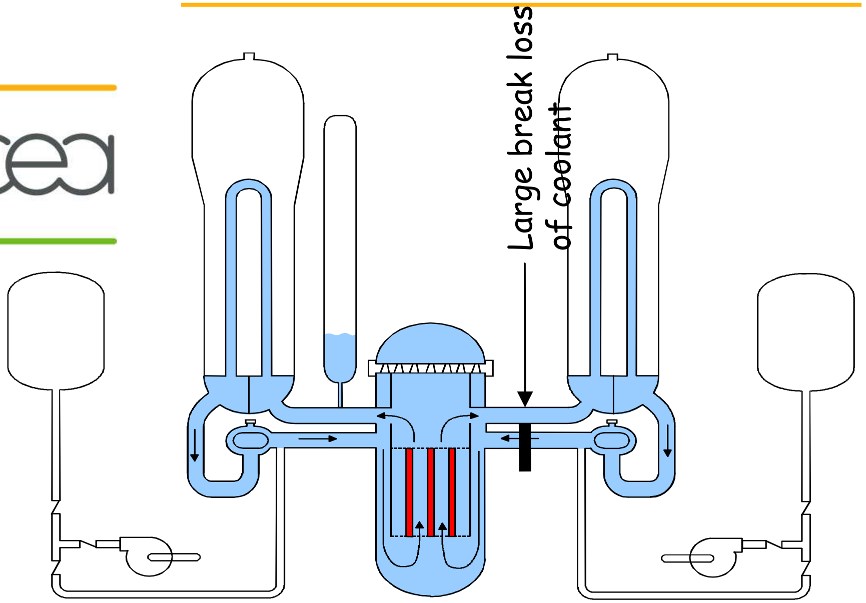


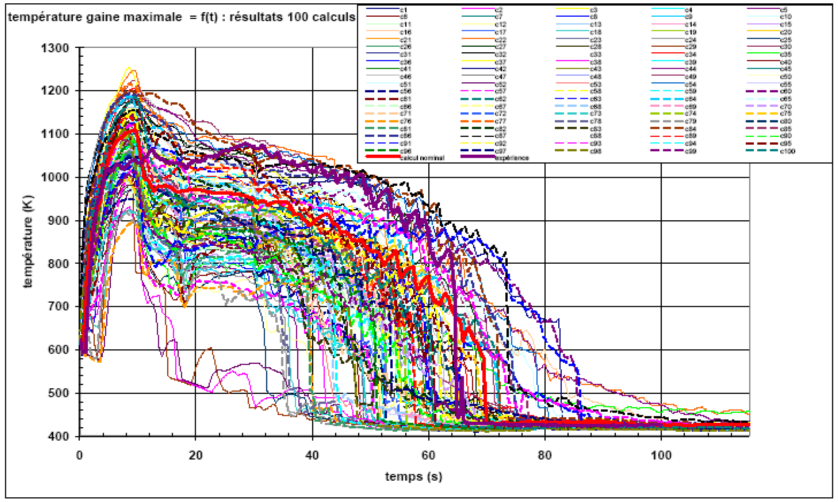
[Vasseur et al., 2009]

- **Intensive numerical tests are needed**
- **Theoretical arguments ?**

Time is coming to mix the criteria (LH/orthogonality/maximin/low discrepancy/…)

3.3 Estimation of extremal events





100 evolutions of the maximal fuel cladding temperature in the nuclear reactor

Safety issue: demonstration with the model that *T* + **uncertainty < 1204°C**

Problems: Estimation of very low probabilities and estimation of high order quantile

Methods and ideas:

- FORM/SORM, Monte Carlo, variance reduction MC
- Mixing metamodel and MC *[Cannamela et al. 2008]*
- sequential Gaussian process *[Oakley 2004]*, maxima of Gaussian process
- extreme value theory, …

3.4 Functional data in computer experiments

- Classical model writes $Y = f(\mathbf{X})$, where Y is a scalar output variable and **X** is a vector of scalar input variables

- **However, model inputs/outputs can be curves, surfaces, volumes, …**

  → **Functional data**

  Model with functional data writes for example $Y(t) = f(X(t))$

  More generally : $Y(v) = f(X_1(u_1),\dots, X_p(u_p),v)$,

  v and u_i are real parameters (possibly multidimensional)

  Ex. for u and v : time t, spatial coordinates (x,y,z), temperature T, …

  Ex. for Y and X : Y(v) = pollutant concentration in function of time

  X(u) = groundwater porosity in function of space

  $X_i(u_i)$ and Y(v) become random functions

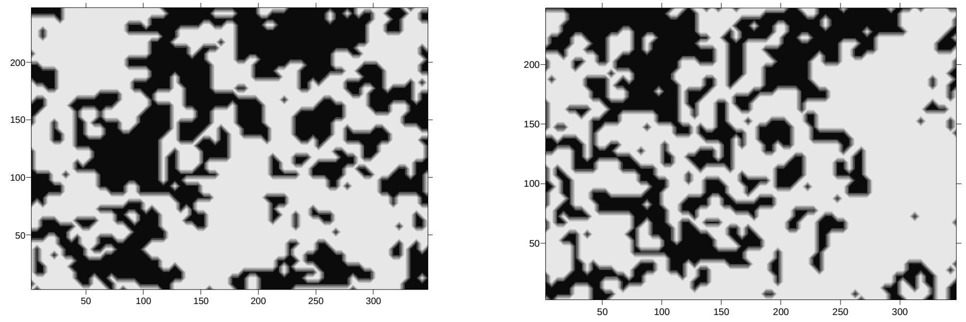
3.4 An example : hydrogeological model

Scenario of pollutant (^{90}Sr) transport (radwaste source) in porous media

- 16 scalar input variables (porosity, Kd, infiltration rates, …)
- Permeability spatial heterogeneity represented by a random field ε(x,y)

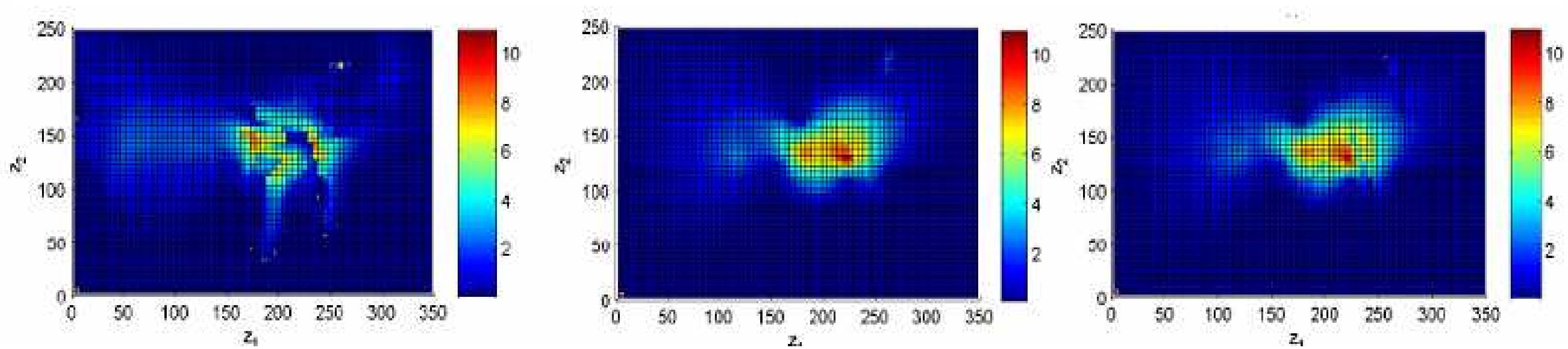
Realizations of this random field (geostatistical simulations)

Inputs



2 possible realisations of the permeability

Outputs



3 different simulations : output concentration maps

3.4 Open issues for functional inputs

Model complex uncertainty sources

Efficient designs for stochastic processes, random fields ?

- LHS for Gaussian random fields *[Pebesma & Heuvelink, JASA, 1999]*
- Functional distances to realize space filling designs *[Scheidt & Caers,07]*

Sensitivity analysis and metamodels

- Labelize the fonction with scalar parameters (scenario parameter) : each realization is associated with a number *[Lilburne & Tarantola, 2008]*
- Consider the function as an uncontrollable variable *[Zabalza, 2000]*
  ⟹ Stochastic computer code
- Functional decomposition
- …

3.4 Open issues for functional (and multidimensionnal) outputs

Sensitivity analysis and metamodels

Functional decomposition

[Campbell et al., 2006, Fang et al., 2006, Higdon et al., 2008, Bayarri et al., 2007, Marrel, 2008]

Sequential/adaptive designs

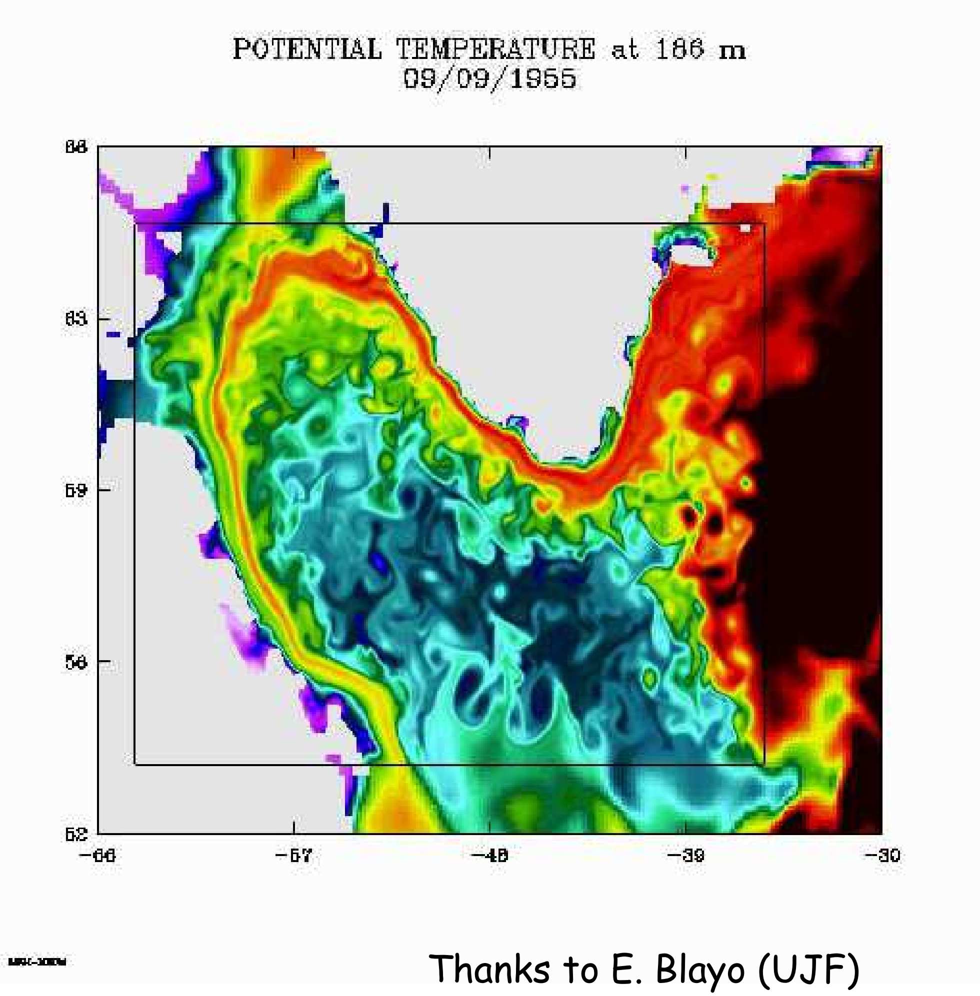
Multivariate analysis

Optimization algorithms, EGO

Multiobjective optimization

3.5 Mixing variational and stochastic methods

In meteorology/oceanography/hydrology/…, the state vector can contain billions of variables → adjoint methods are essential



Thanks to E. Blayo (UJF)

Labrador sea

NEMO model (Nucleaus for European Modelling of the Ocean)

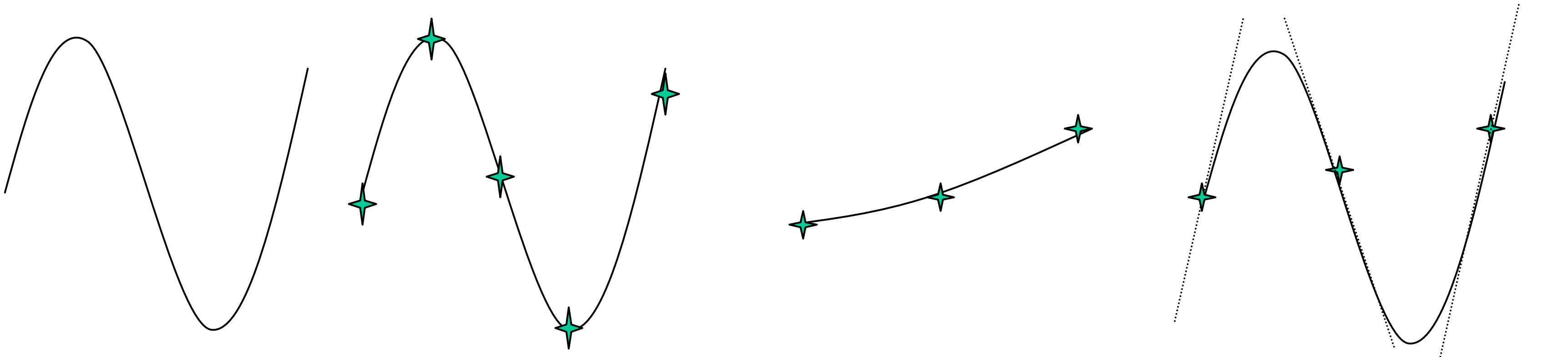
3.5 Mixing variational and stochastic methods

From our practice, metamodels can approximate computer codes until $d \sim 50$ with the heuristics $N \sim 10\ d$

What can we do when $d \gg 100$?

One solution could be to use extra-information

For example: derivatives (adjoint methods, automatic differentiation)
[Isukapalli 2000, Morris et al. 1993]



Model | Approximation with large N | Approximation with small N | Approximation with small N + derivatives

3.6 Stochastic computer codes

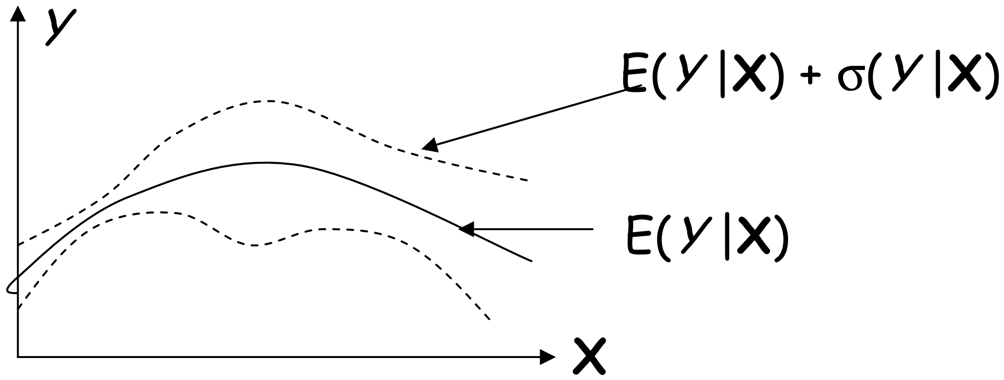
Stochastic model : $f : \mathfrak{R}^p \longrightarrow \mathfrak{R}$

$$\mathbf{X} \longrightarrow Y = h(\mathbf{X}) + \nu(\varepsilon, \mathbf{X} ; \varepsilon)$$

ν is random and caused by « uncontrollable » variables ε, interacting with $\mathbf{X}$

Example : code based on Stoch. Diff. Equations, evolution models (disease, population, …), lagrangian approaches (particle trajectories, …)

Some works have been done on the modelisation of the conditional mean and variance (joint metamodels)



[Vining & Myers 1990, Smyth 1989, Zabalza-Mezghani 2000, Iooss & Ribatet 2009, Boukouvalas 2009]

3.6 Stochastic computer codes

- Sensitivity analysis

$Y_m(\mathbf{X}) = E(Y|\mathbf{X})$ and $Y_d(\mathbf{X}) = \mathrm{Var}(Y|\mathbf{X})$, we have

$$\mathrm{Var}[Y(\mathbf{X},\varepsilon)] = \mathrm{Var}[Y_m(\mathbf{X})] + E[Y_d(\mathbf{X})]$$

Sensitivity indices for **X** : $S_{X_i} = \dfrac{\mathrm{Var}[E(Y_m|X_i)]}{\mathrm{Var}(Y)}$

Sensitivity indices for ε : $S_{T\varepsilon} = \dfrac{E(Y_d)}{\mathrm{Var}(Y)}$

?? Sensitivity indices for the interactions between ε et X_i ??

- Specific designs: No replications / Replications on **X** / Replications on ε if possible

- Metamodels of the conditional distribution p(Y | **X**)

Stochastic computer models return a distribution, rather a scalar, for each set of inputs *[Reich et al., 2009]*

Conclusion: The main challenge: Place of « DACE » teams

Problem: How to maintain team mixing mathematical research, application exercises and dissemination to the users in the computer experiments domain

This seems to be a classical problem in industrial statistics !

[see Steinberg (ed), The future of industrial statistics, Technometrics (50), 2008]

- Integration in support department ? ☠ No research and small term views
- Integration in engineering department ? ☠ Priority to big project
- Integration in computing engineering department ? ☠ Lack of methodology
- Integration in modelling department ? ☠ Limited views, physics dependencies

Solutions : create strong partnerships between industrial and academic teams

Consortium DICE seems to have chosen this key of success

Another powerful initiatives : national collaborations as MUCM (UK), SFCME (USA), **GdR MASCOT-NUM (France): www.gdr-mascotnum.fr**, …

THANK YOU FOR YOUR ATTENTION

References

- Bayarri et al., *The Annals of Statistics*, 2007
- Boukouvalas & Cornford, Technical Report, Aston University, 2009
- Cannamela et al., *Annals of Applied Statistics,* 2008
- De Rocquigny et al., *Uncertainty in industrial practice,* Wiley, 2008
- Fang et al., D*esign and modeling for computer experiments,* Chapman, 2006
- Franco et al., *Chem. Intell. Lab. Systems,* 2009
- Higdon et al., *Journal of American Statistical Association,* 2008
- Iooss & Ribatet, *Reliability Engineering & System Safety*, 2009
- Isukapalli et al., *Risk Analysis,* 2000
- Lilburne & Tarantola, *Intern. Journal of Geographical Information Science,* 2009
- Marrel, Thèse INSA Toulouse, 2008
- Morris et al., *Technometrics,* 1993
- Oakley, *JRSS C,* 2004
- Pebesma & Heuvelink, *Technometrics,* 1999
- Qian et al., *Technometrics*, 2008
- Reich et al., *Technometrics*, submitted
- Scheidt & Caers, *Petroleum Geostatistics,* 2007
- Smyth, *Journal Royal Statistical Society B,* 1989
- Vasseur et al., Journées de la SFdS, 2009
- Vining & Myers, *Journal of Quality Technology,* 1990
- Zabalza-Mezghani, Thèse de l'Université de Pau, 2000