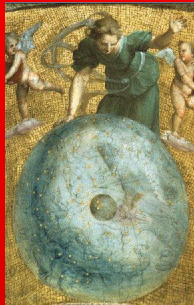


Uncertainties Propagation with URANIE



F. Gaudier, J.M. Martinez
G. Arnaud, A. Bruneton

CEA DANS/DM2S/STMF/LGLS

fabrice.gaudier@cea.fr

Workshop $P(ND)^2 - 2$
CEA DAM - TGCC
2014/10/16



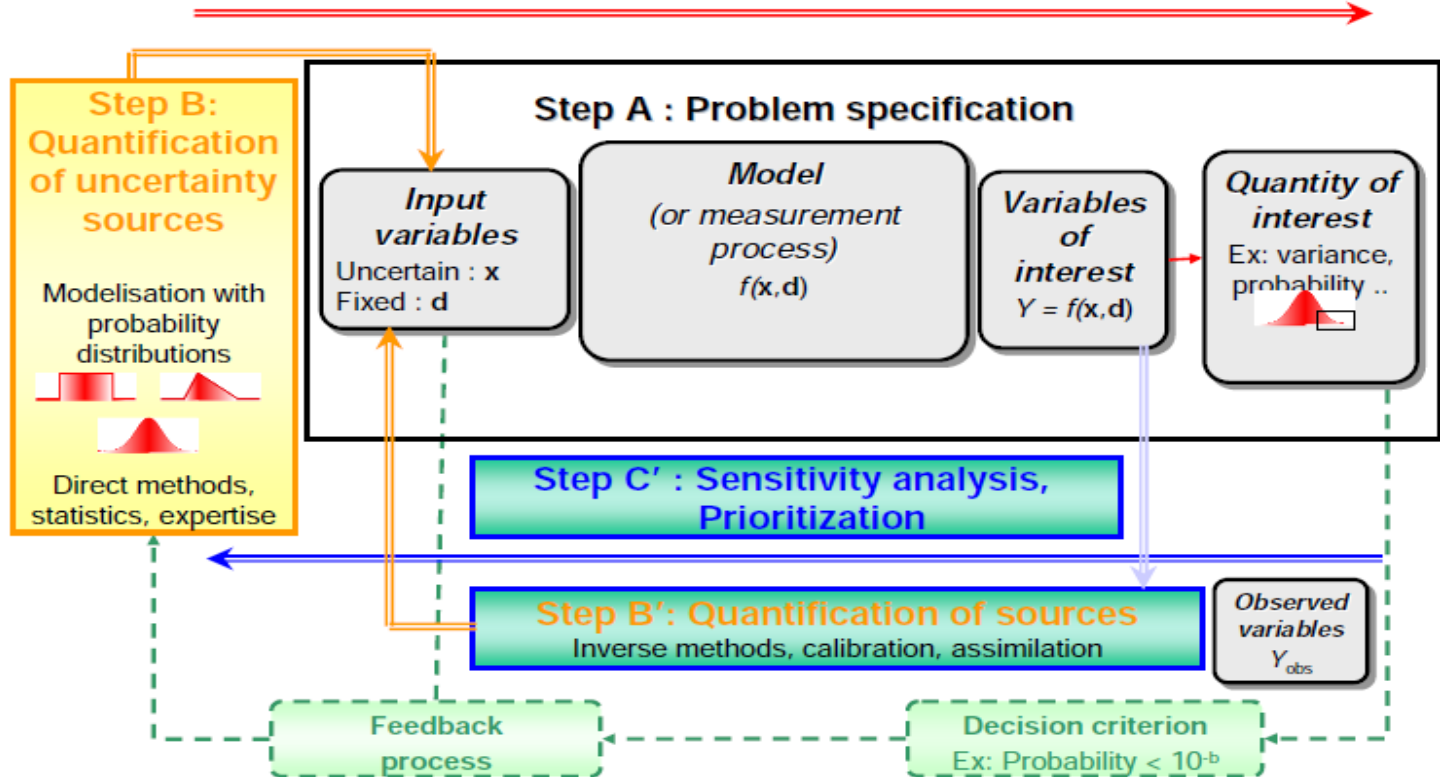
- Uncertainties Propagation
 - Input modelisation
 - Distribution of the computation code
 - Output analysis
- the Uranie platform
- Example of Uncertainties Propagation





[De Rocquigny et al., 2008]

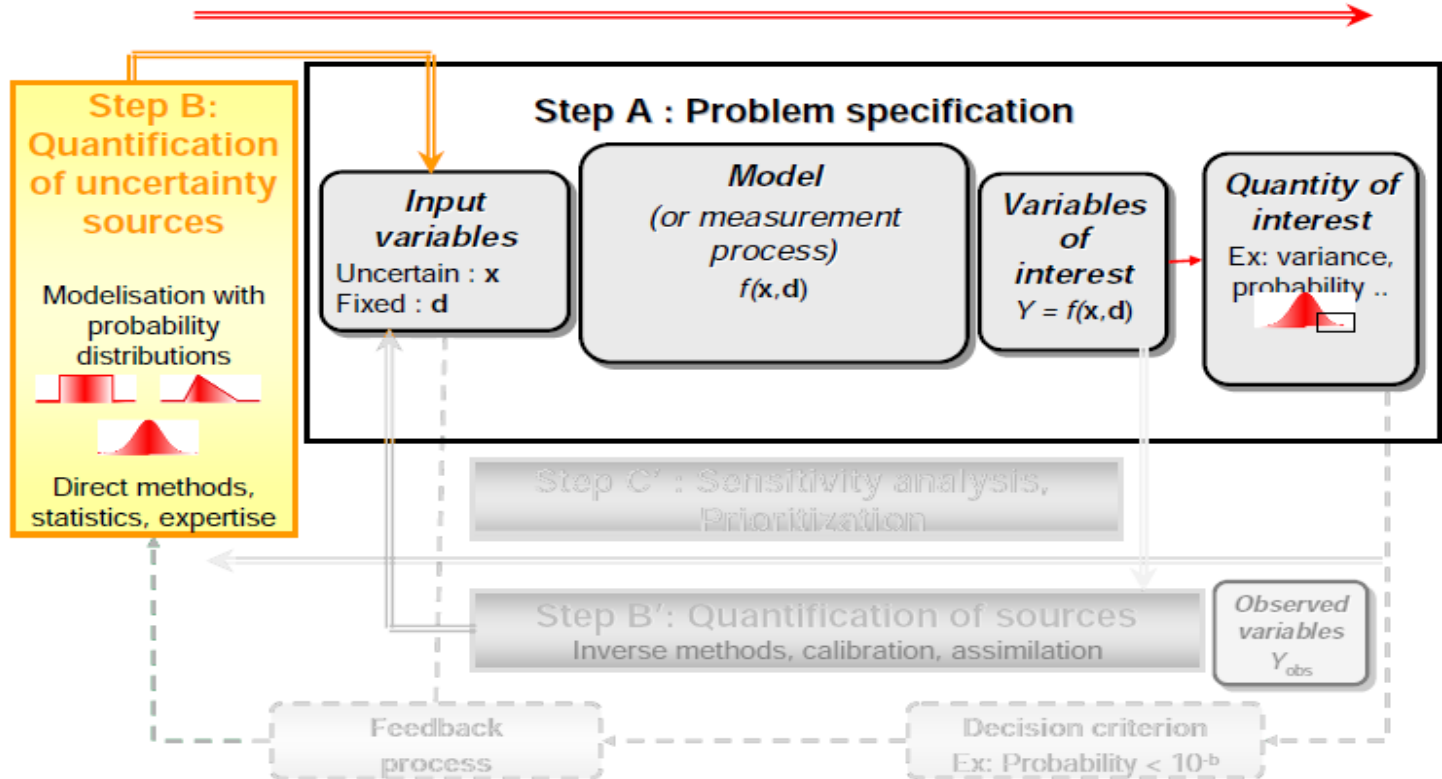
Step C : Propagation of uncertainty sources





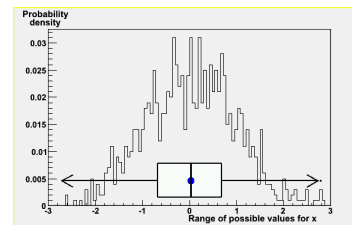
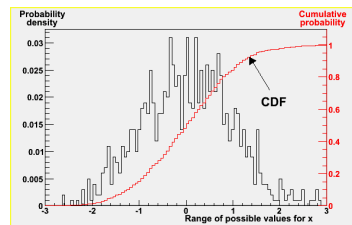
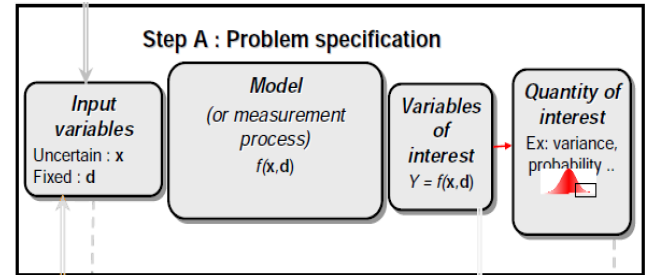
[De Rocquigny et al., 2008]

Step C : Propagation of uncertainty sources



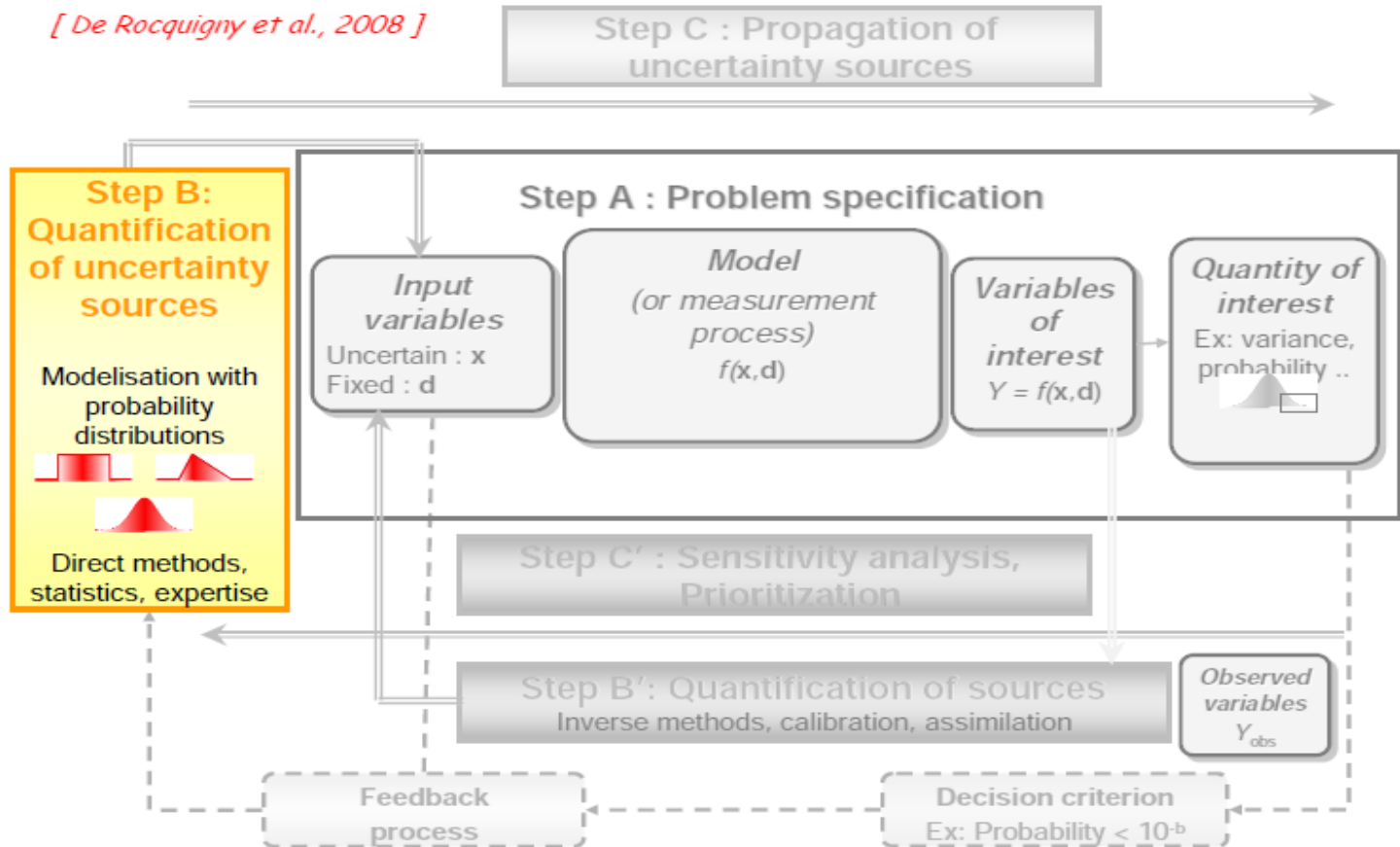


- Computer code f
- Two types of input parameters (x, d)
 - Fixed parameters d
 - Uncertain parameters x
- Outputs of interest $y = f(x, d)$
- Quantity of interest on the Uncertainties Propagation
 - *Location* : Mean μ , Min, Max, Mode, Median, Quantile q_α
 - *Dispersion* : Standard-deviation σ , Variance σ^2 , Range ($Max - Min$), Coefficient of Variation ($\delta = \sigma/\mu$)
 - Probability Density Function (PDF), Cumulative Density Function (CDF)





[De Rocquigny et al., 2008]





- Expert judgment
- With "large" dataset : Fitting the parameters of the PDF
 - Parametric methods
 - Non-parametric methods
 - Statistical Tests
- With small dataset :
 - Bayesian methods
 - Bootstrap methods (resampling)

Step B: Quantification of uncertainty sources

Modelisation with
probability
distributions



Direct methods,
statistics, expertise

Continuous

Bounded

Uniform
Beta
Triangular
Trapezium
Uniform by parts
LogUniform
LogTriangular
...

positive

Exponential
LogNormal
Weibull
Gamma
Khi-two
Pareto
...

Unbounded

Normal
Cauchy
Gumbel
...

Discrete

Binomial
Multinomial
Poisson
...



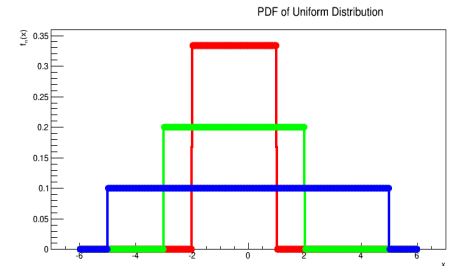


1. Uniform Distribution

- The values in the interval $[a, b]$ are equally probable
- 2 parameters a ("Minimum") and b ("Maximum")

$$f(x) = \frac{1}{b-a} \mathbb{I}_{[a,b]}(x)$$

- Mean : $\mu = \frac{b-a}{2}$ (Median)
- Mode : any value in $[a, b]$
- Variance : $\sigma^2 = \frac{(b-a)^2}{12}$

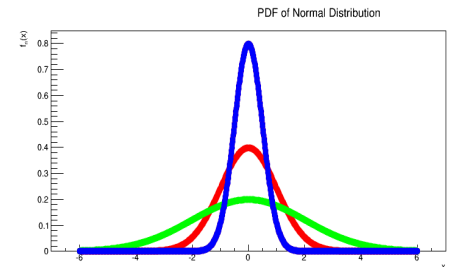


2. Normal Distribution

- 2 parameters μ ("Mean") and σ ("Standard-Deviation")

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- Mean : μ (Mode, Median)
- Variance : σ^2



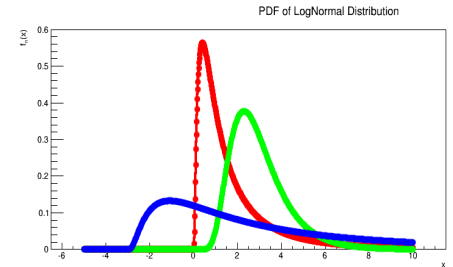


3. LogNormal Distribution

- A **positive** random variable x is said to follow a *LogNormal* law when $\ln x \sim \mathcal{N}$
- 3 parameters x_0 (lower bound) and (μ, σ) when $\ln(X) \sim \mathcal{N}(\mu, \sigma)$

$$f(x) = \frac{1}{(x - x_0)\sigma\sqrt{2\pi}} \exp\left\{-\frac{(\ln(x-x_0)-\mu)^2}{2\sigma^2}\right\} \quad \forall x > x_0$$

- Mean : $\mu_X = \exp\left(\mu + \frac{\sigma^2}{2}\right)$
- Median : $\exp(\mu)$
- Mode : $\exp(\mu - \sigma^2)$
- Variance : $\mu^2 \times (\exp\sigma^2 - 1.)$





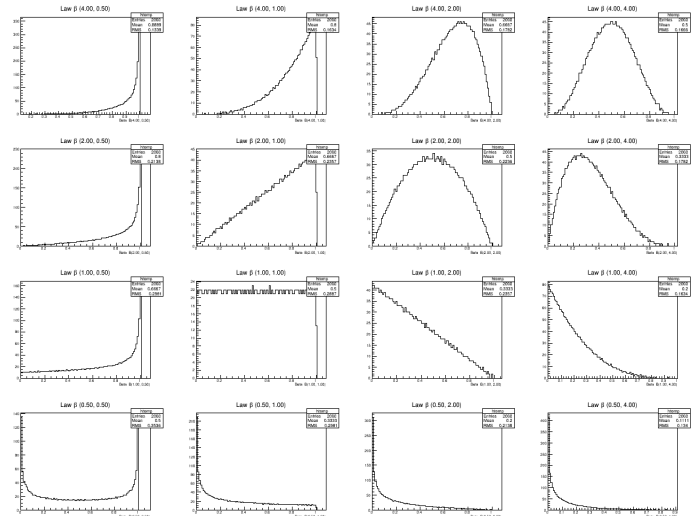
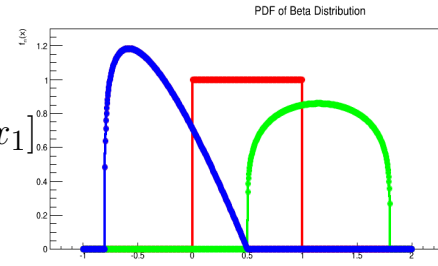
4. Beta Distribution

- 4 parameters α, β (shapes) & $x_0 < x_1$ (bounds)

$$f(x) = \frac{u^{\alpha-1} * (1-u)^{\beta-1}}{B(\alpha, \beta)} \quad \forall x \in [x_0, x_1]$$

$$\text{with } u = \frac{x-x_0}{x_1-x_0}$$

- Mean : $x_0 + (x_1 - x_0) \frac{\alpha}{\alpha+\beta}$
- Mode : depends on (α, β)
- Variance : $(x_1 - x_0)^2 \frac{\alpha\beta}{\alpha+\beta+1}$





- Let (x_1, x_2, \dots, x_n) an *i.i.d* sample of a PDF $f(x, \theta)$ where $\theta \in \Theta$ is a vector of parameters for this family. The true value of the parameter θ^* is unknown
- Build an estimator $\hat{\theta}$ which would be as close to the true value θ^* as possible

1. Maximum Likelihood (MLE)

The method of maximum likelihood selects the set of values of the model parameters that maximizes the *likelihood* function. This function measures the "agreement" of the selected model with the observed data.

$$\hat{\theta}_{MLE} = \arg \max_{\theta \in \Theta} \frac{1}{n} \ln \left(\prod_{i=1}^n f(x_i | \theta) \right) \quad \dots \text{ if any maximum exists}$$

2. Moments Method (MM)

- One starts with deriving equations that relate the population moments to the parameters θ
- The moments are estimated from the given sample
- The equations are then solved for the parameters θ , using the sample moments in place of the (unknown) population moments

$$g_k(\theta_1, \theta_2, \dots, \theta_k) = \mathbb{E}[X]^k = \mu_k \quad \text{and} \quad \widehat{\mu}_k = \frac{1}{n} \sum_{i=1}^n x_i^k = g_k(\widehat{\theta}_1, \widehat{\theta}_2, \dots, \widehat{\theta}_k)$$





- The **histograms** are classical density estimation
- The followings steps are needed to build the histogram:
 - Arrange the sample in increasing order;
 - Subdivide the range of the sample into several equal intervals, and count the number of observations in each intervals;
 - plot the number of observations in each interval versus the random variable
- but the form depends on the number of bins

1. **Sturges**

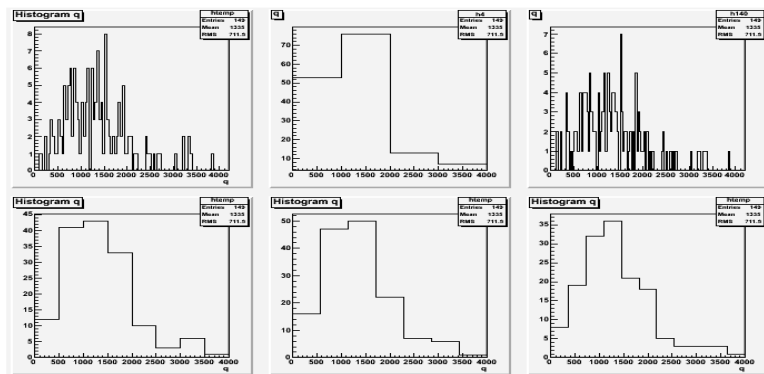
2. **Scott**

3. **Freedman & Diaconis**

$$N_{bin} = \log_2(n) + 1$$

$$N_{bin} = (x_{max} - x_{min}) * \sqrt[3]{n} / 3.5 \hat{\sigma}_x$$

$$N_{bin} = (x_{max} - x_{min}) * \sqrt[3]{n} / 2 * (Q_x^{0.75} - Q_x^{0.25})$$





- A function $K : \mathbb{R} \rightarrow \mathbb{R}$ is said a **Kernel** if

$$\int K(u) \, du = 1.$$

- Often, but not necessarily,

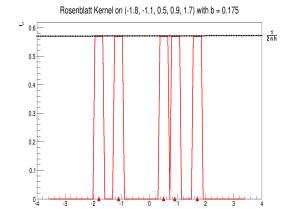
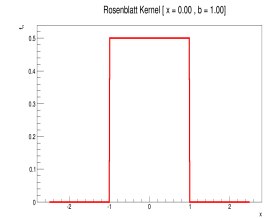
- K is symmetric around the origin: $K(-u) = K(u) \quad \forall u$
- K is positive: $K(u) > 0 \quad \forall u$

- $\forall h > 0$,

$$\hat{f}_{n,h}(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h} K\left(\frac{X_i - x}{h}\right)$$

is a **kernel estimator** of the density f ($\int \hat{f}_{n,h}(x) \, dx = 1$)

- Kernel approach is a histogram which, for estimating the density of $f(x)$, has been shifted so that x , say, lies at the center of a mesh interval. And For evaluating the density at another point, say y , the mesh is shifted again, so that y is at the center of a mesh interval.
- The parameter h is a *smoothing* parameter called **bandwidth**.
More greater h is, more the estimation $\hat{f}_{n,h}$ is smooth.



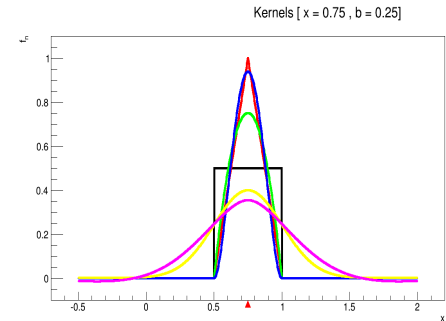


- Optimal bandwidth with the Silverman Rule (1996)

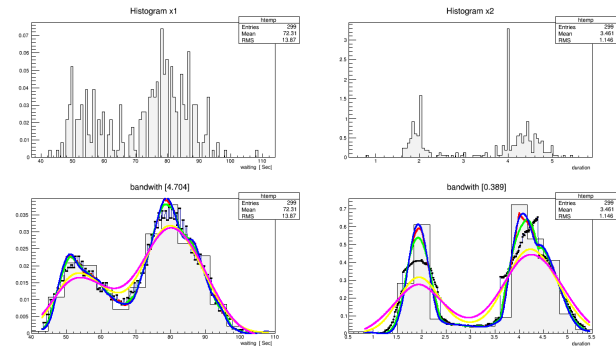
$$h_n = 1.364 \times \alpha_K \times \text{MIN}\left\{\hat{\sigma}, \frac{\text{IQR}}{1.349}\right\} \times n^{-1/5}$$

with

- $\hat{\sigma}$ is the sample standard deviation
- IQR is the "InterQuartile Range"
($\text{IQR} = q_{0.75} - q_{0.25}$)
- α_K is a constant that only depends on the used kernel



Kernel	$K(x)$	α_K
Rectangular	$1/2, x < 1$	1.3510
Triangular	$1 - x , x < 1$	1.8882
Epanechnikov	$\frac{3}{4}(1 - x^2), x < 1$	1.7188
Biweight	$\frac{15}{16}(1 - x^2)^2, x < 1$	2.0362
Gaussian	$\frac{\exp^{-x^2/2}}{\sqrt{2\pi}}$	0.7764



Geyser database for Gaussian
Kernel (*left*) waiting $b = 4.70$,
(*right*) duration $b = 0.39$



- **QQPlot** (Graphical methods)

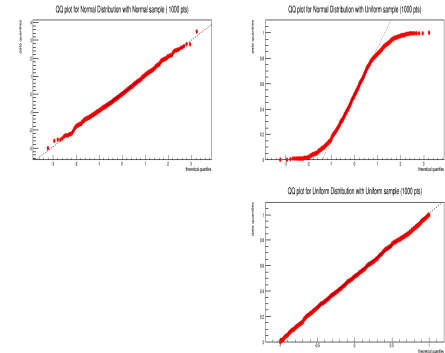
- a **QQ-plot** ("Q" stands for *Quantile*) is a probability plot to compare two probability distributions by plotting their quantiles against each other

- A point (x, y) on the plot corresponds to one of the quantiles of the second distribution (y-coordinate) plotted against the same quantile of the first distribution (x-coordinate).

- If the two distributions being compared are similar, the points in the QQ-plot will approximately lie on the line $y = x$

- If the distributions are linearly related, the points in the QQ-plot will approximately lie on a line, but not necessarily on the line $y = x$.

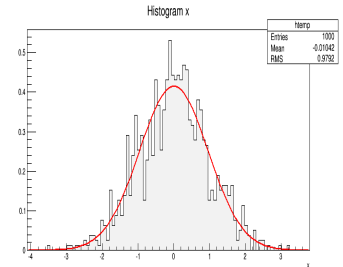
- Select one axe for the theoretical distribution for Goodness-of-Fit test





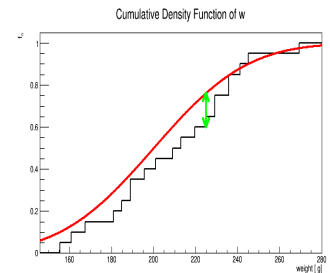
- Statistical Tests
 - **Chi-Squared** : The basic idea is to partition the range of the sample into k cells, and compare the observed frequency O_i with the expected frequency E_i in each cell i

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$$



which follows a χ^2 distribution with $(k - 1 - t)$ degrees of freedom, where t is the number of parameters of the distribution to estimate

- **Tests based on EDF Statistics** ("*Empirical Distribution Function*")
 - ★ Measures the discrepancy between the empirical and the theoretical CDFs (based on the differences between $F_n(x)$ and $F(x)$)
 - ★ Two classes : the **supremum** and the **quadratic**



$$D = \sup_x |F_n(x) - F(x)|$$

$$Q = n \int_{-\infty}^{+\infty} (F_n(x) - F(x))^2 \psi(x) dx \quad \text{where } \psi \text{ is a weight function}$$

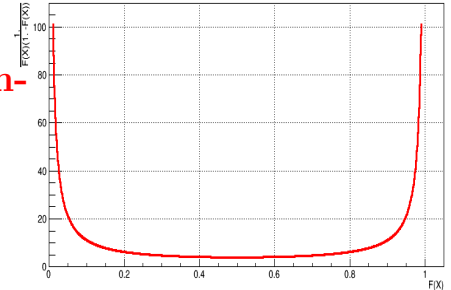


- For $\psi(x) = 1$ we obtain the **Cramer-von Mises** Tests, denoted as W^2 :

$$W^2 = n \int_{-\infty}^{+\infty} (F_n(x) - F(x))^2 dx$$

- For $\psi(x) = \frac{1}{F(x)(1.0-F(x))}$ we obtain the **Anderson-Darling** test, denoted A^2 :

$$A^2 = n \int_{-\infty}^{+\infty} \frac{(F_n(x) - F(x))^2}{F(x)(1.0 - F(x))} dx$$

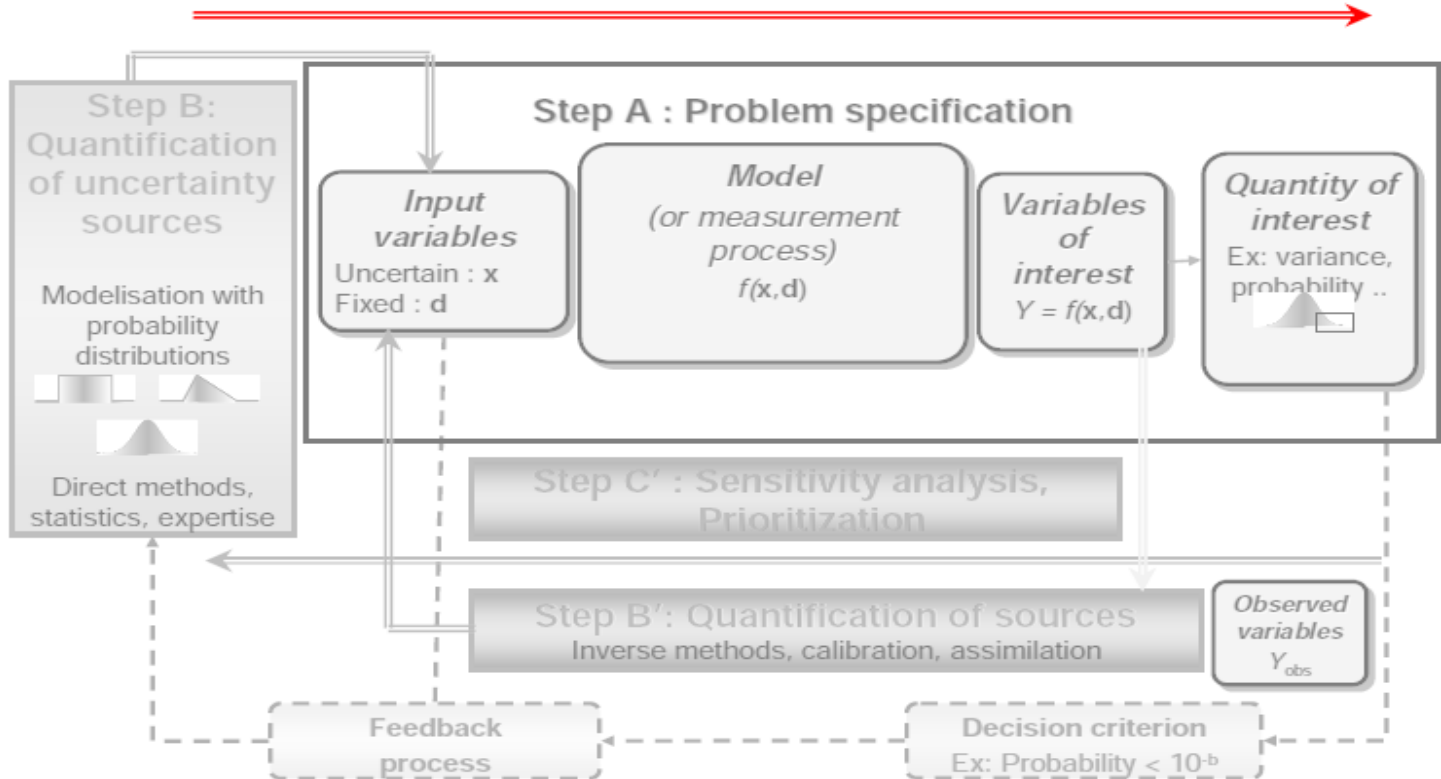


- The χ^2 statistic is the lower powerful for continuous PDF
- EDF statistics are usually much more powerful than the χ^2 statistic (where data must be grouped, then loss of information)
- the **Kolmogorov-Smirnov** D statistic is the most well-known of the EDF statistics, but it is often much less powerful than the quadratic statistics W^2 and A^2
- A^2 and W^2 give often similarly values, but A^2 is on the whole more powerful when the distribution F departs from the true distribution in the tails (weight function)



[De Rocquigny et al., 2008]

Step C : Propagation of uncertainty sources





- Generate Design of Experiments ("*DoE*")
 - Monte-Carlo Sampling ("*SRS*"), Latin HyperCube Sampling ("*LHS*")
 - quasi Monte-Carlo Sampling ("*qMC*")
 - Low Discrepancy Sequences ("*Space-Filling Design*")
 - Take into account correlations between variables
- Evaluate the code for each points of the DoE (sequential on a PC, or parallel on MultiCore PC/Cluster)
 - Substitute the values on the current point into the input files of the code
 - Launch the code with the new input files
 - Catch the output values of the variables of interests
 - Using "*Surrogate Model*" (linear, polynomial, Artificial Neural Network, Kriging) to reduce the computational times of the code evaluation
- Analyze the *Quantity of interest* by statistics
 - Univariate attribute
 - Data Modelisation with **PDF** or **Kernel** (*as Step B*)
 - Goodness-of-Fit Techniques (*as Step B*)



The effect of the "location" parameter is to translate the graph relative to the standard distribution

- **Mean** μ :

$$\mu = \frac{1}{nS} \sum_{i=1}^{nS} x_i$$

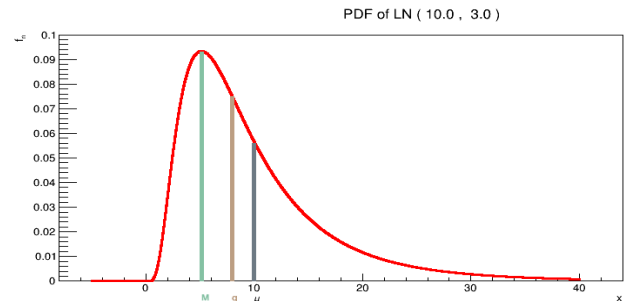
- **Mode** M : Value where the probability is the greatest value
- **Median** $q_{0.5}$: it is the 0.5-quantile

$$q_{0.5} \quad \text{as} \quad IP[X \leq q_{0.5}] = 0.5 = IP[X \geq q_{0.5}]$$

- **α -Quantile** q_α with $\alpha \in [0, 1]$:

$$q_\alpha \quad \text{as} \quad IP[X \leq q_\alpha] = \alpha$$

- **Quartiles** $q_{0.25}, q_{0.50}, q_{0.75}$
- **Extremes values** min, max





The effect of a "dispersion" parameter is to stretch|shrink the standard distribution

- **Variance** $Var[X]$: measure of spread in the data about the mean $Var[X] = \mathbb{E}[(X - \mathbb{E}[X])^2]$, and can be estimated by :

$$Var[X] = \frac{1}{nS - 1} \sum_{i=1}^{nS} (x_i - \mu)^2$$

- **Standard Deviation** σ : to have an information in the same unit as the variable

$$\sigma = \sqrt{Var[X]}$$

- **Coefficient of Variation** δ : σ does not indicate **the degree** (%) of dispersion around the mean value μ , a **nondimensional** term can be introduced :

$$\delta = \frac{\sigma}{\mu}$$

- **Range** R :

$$R = Max - Min$$

- **InterQuartile Range** IQR :

$$IQR = q_{0.75} - q_{0.25}$$





A "shape" parameter is any parameter of a PDF that is neither a location parameter nor a scale parameter. Such a parameter must affect the shape of a distribution rather than simply shifting it (location parameter) or stretching/shrinking it (dispersion parameter).

- **Moment order p :** $\mu_p := \mathbb{E}[(X - \mathbb{E}[X])^p]$

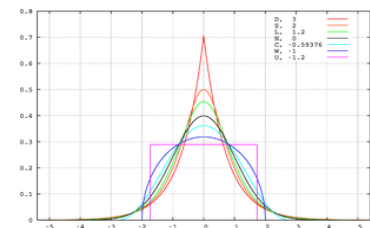
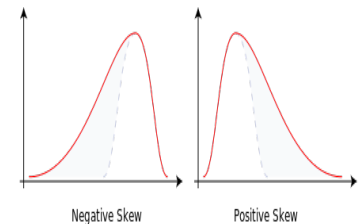
$$\mu_p = \frac{1}{nS} \sum_{i=1}^{nS} (x_i - \mu)^p$$

- **Skewness :** γ_1 is a measure of the asymmetry of the PDF about its mean. The skewness value can be positive or negative, or even undefined.

$$\gamma_1 := \mathbb{E} \left[\left(\frac{X - \mu}{\sigma} \right)^3 \right] = \frac{\mu_3}{\sigma^3} = \frac{\mathbb{E}[X^3] - 3\mu\sigma^2 - \mu^3}{\sigma^3}$$

- **Kurtosis :** γ_2 is a measure of the "peakedness/flatness" of the PDF

$$\gamma_2 = \frac{\mu_4}{\sigma^4} - 3.0$$

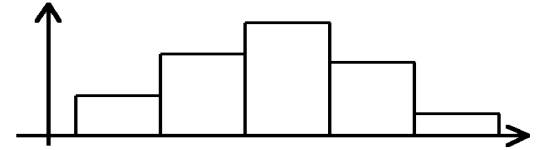


from Wikipedia



- **Histogram**

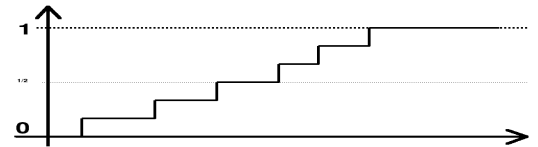
$$H(x) = \frac{1}{nS} \sum_{i=1}^{nS} \frac{\mathbb{I}_{[t_i, t_{i+1}]}(x)}{(t_{i+1} - t_i)} \quad \text{when } x \in [t_i, t_{i+1}]$$



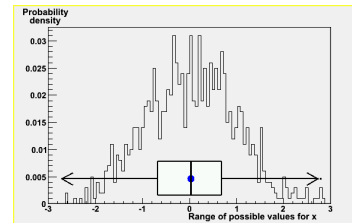
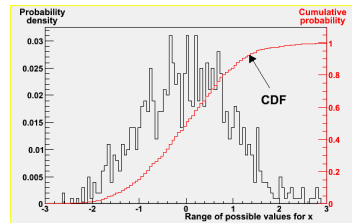
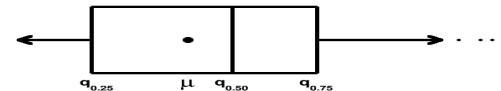
where $[a, b] = \bigcup_i [t_i, t_{i+1}]$

- **Empirical Cumulative Density Function (eCDF)**

$$F_n(x) = \frac{1}{nS} \sum_{i=1}^{nS} \mathbb{I}(X_i \leq x)$$



- **Boxplot (Tukey)**








- Uncertainties Methodology
 - Input modelisation
 - Distribution of Computation
 - Output analysis
- **the Uranie platform**
- Example of Uncertainties Propagation





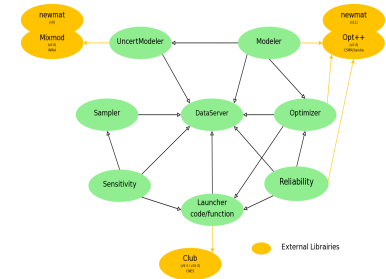
-  ROOT (CERN),
  MIXMOD (Gaussian Mixtures - INRIA),
  OPT++ (Optimization - Sandia), NLOpt (Optimization - MIT)

- Data access :

- Flat file with header ("Salomé Table")
- TTree (internal ROOT)
- SQL Data base (MySQL, PostgreSQL, ...)

- Uncertainty/Sensitivity/Optimisation methods in URANIE

- Design Of Experiments (SRS, LHS, ROA, qMC, MCMC, Copulas)
- Clustering methods
- Surrogate models (Polynomial, Artificial Neural Networks, Kriging, GLM)
- Non Intrusive Spectral Projection : Generalized Polynomial Chaos
- Inverse Quantification of Uncertainty (CIRCE)
- Sensitivity Analysis: Local, Morris, Regressions (*Pearson, Spearmann*), Sobol, FAST & RBD
- Optimization, Multi-Criteria (**Vizir** library : Genetic Algorithms)
- Computing distribution (**HPC** : TGCC, CCRT)





- 115 000 lines & 235 classes
- Version of ROOT :
v5.34.13 (2013 Nov.)
v5.32 (2011 Dec.)
v5.34.19 (2014/07/09)

- Compilation with **cmake**
(Linux-Makefiles/Windows-Visual Project)

- CDash reporting
 - Unitary tests with **CppUnit**
 - Coverage with **gcov**
 - Memory check with **valgrind**

- **Exceptions** (Warning, Error, Deprecated)
- Open source since 2013/05

<http://sourceforge.net/projects/uranie>

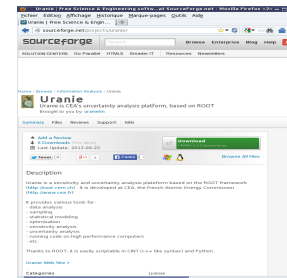
Uranie											
No update data as of Wednesday, September 25 2013 - 00:00 CEST											
Nightly											
Site	Build ID	Build Name	Files	Error	Warn	Error	Warn	Not Run	Fail	Pass	Build Time
ii220576	6001	CentOS-5.5-64bits	0	0	0	0	0	0	0	11	Sep 25, 2013 - 02:00 CEST
ii222287@ra0a08	6002	Fedora-18-64bits	0	0	0	0	0	0	0	10	Sep 25, 2013 - 02:40 CEST
ii212796	6000	Mandriva-2010-6-64bits	0	0	0	0	0	0	0	11	Sep 25, 2013 - 03:31 CEST
ii214467	6004	Windows	0	0	0	0	879	0	0	7	Sep 25, 2013 - 02:36 CEST
Coverage											
Site	Build ID	Build Name	Percentage	LOC Tested	LOC Untested	Date					
ii220576	6001	CentOS-5.5-64bits	55.18%	49939	39781	Sep 25, 2013 - 02:00 CEST					
ii222287@ra0a08	6002	Fedora-18-64bits	0%	0	0	Sep 25, 2013 - 02:40 CEST					
ii212796	6000	Mandriva-2010-6-64bits	55.23%	45204	35882	Sep 25, 2013 - 03:31 CEST					
Dynamic Analysis											
Site	Build ID	Build Name	Checker	Defect Count	Date						
ii220576	6001	CentOS-5.5-64bits	Valgrind	1236	Sep 25, 2013 - 02:00 CEST						
ii222287@ra0a08	6002	Fedora-18-64bits	Valgrind	1361	Sep 25, 2013 - 02:40 CEST						
ii212796	6000	Mandriva-2010-6-64bits	Valgrind	26	Sep 25, 2013 - 03:31 CEST						

CDash reports

```

--- <WARNING> URANIE::MessageLogger : ===== WARNING =====
--- <WARNING> URANIE::MessageLogger : Depreciated constructor since v0.3 to v0.5
--- <WARNING> URANIE::MessageLogger : Using the same constructor with a TDataServer object
--- <WARNING> URANIE::MessageLogger : ===== End Of Warning =====
  
```

Depreciated message



Uranie SourceForge site

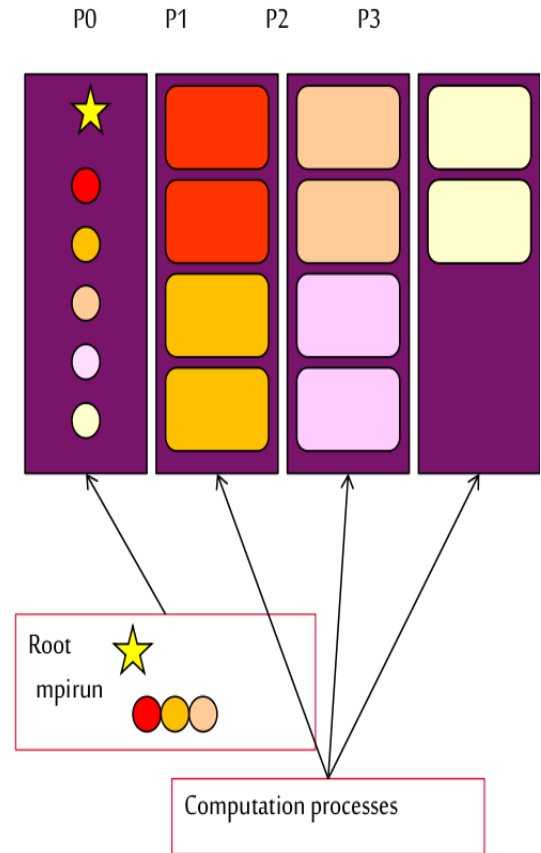


- Distribution of computations:
 - Sequential on PCs
 - Parallel on Multicore PCs
 - Parallel on Cluster (LSF, SGE, SLURM, LoadLeveler) with
`bsub` < `BsubFile`
- The mechanism for launching computations in URANIE is transparent for the user : the URANIE script is the same whether you run it on the local machine, a cluster or a supercomputer
- The sequence is the following:
 - The design of experiments is created (depending on the method and on the uncertainties on the input parameters)
 - URANIE analyses the machine through the environment variables and deduces the number of available processes
 - A pool of processes is managed in order to distribute computations as the processors become available





- Chosen strategy
 - One job in which the computations are hosted
- Aim : being able to run design of experiments
 - On serial codes
 - On MPI-based parallel codes
 - On coupled simulations (with SALOME platform or with MPI)
- Difficulty related to the fact that MPIRUN cannot call itself
- Chosen implementation
 - The master node manages the distribution of computations as processors become available
 - When a processor group becomes available, the master process is forked and runs MPIRUN
 - The end of the job execution is detected by analyzing the state of the child process





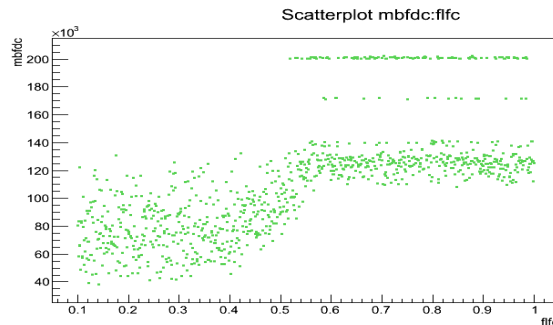
- Uncertainties Methodology
 - Input modelisation
 - Distribution of Computation
 - Output analysis
- the Uranie platform
- **Example of Uncertainties Propagation**





- Thermal hydraulic code (study in 2009)
- CPU time for single computation :
~ 5 minutes (approximation of the true code)
- Design of Experiments
 - $nX = 32$ input attributes with Uniform and Normal Distributions
 - $nY = 23$ output attributes
 - $nS = 1500$ points

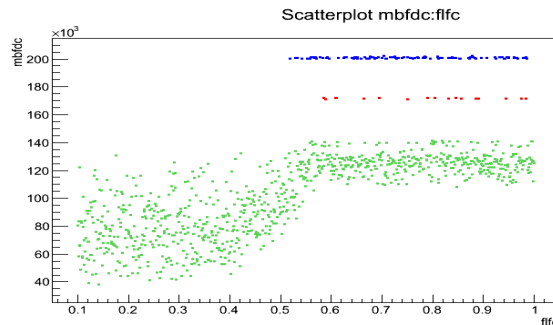
The Code's developer use to see one output y_i versus one input x_j :

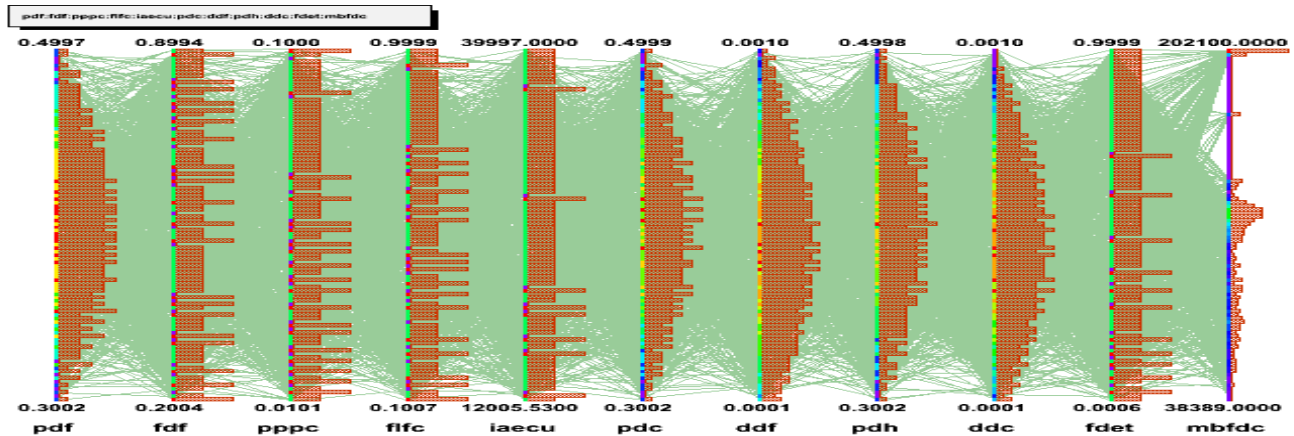
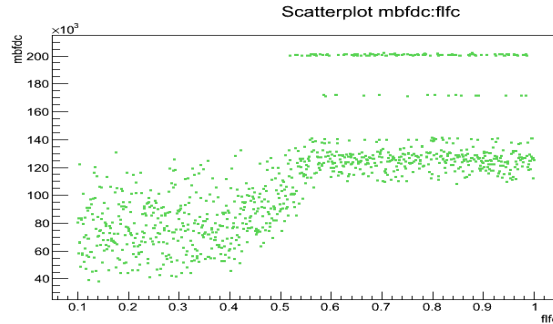




- Thermal hydraulic code (study in 2009)
- CPU time for single computation :
 - ~ 5 minutes (approximation of the true code)
- Design of Experiments
 - $nX = 32$ input attributes with Uniform and Normal Distributions
 - $nY = 23$ output attributes
 - $nS = 1500$ points

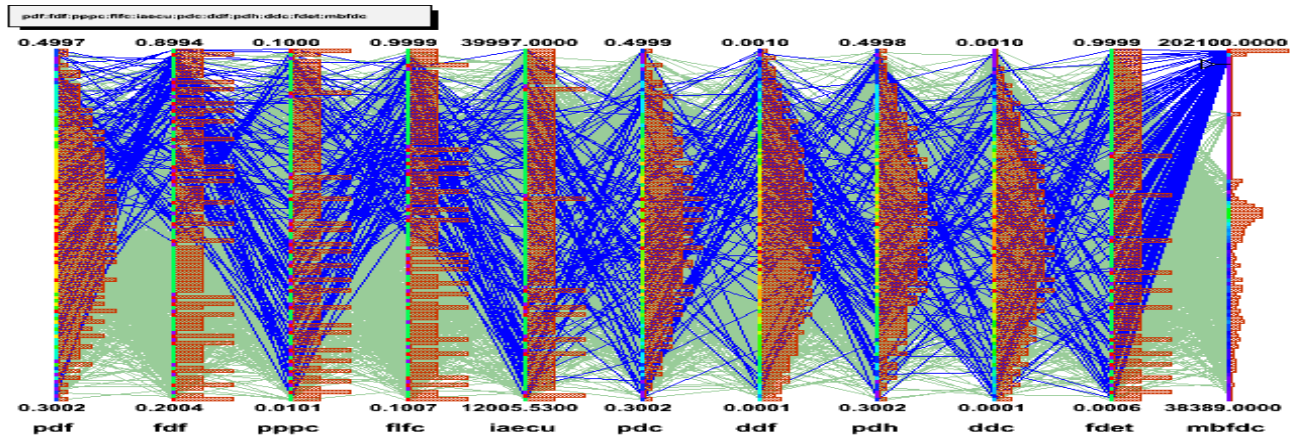
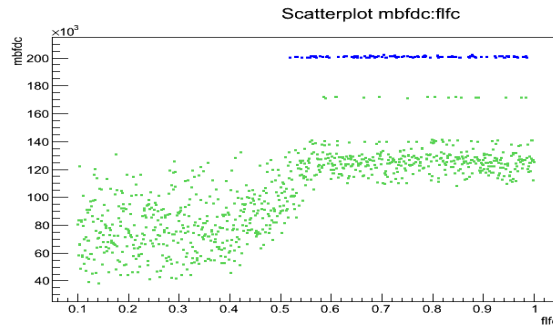
The Code's developer use to see one output y_i versus one input x_j :

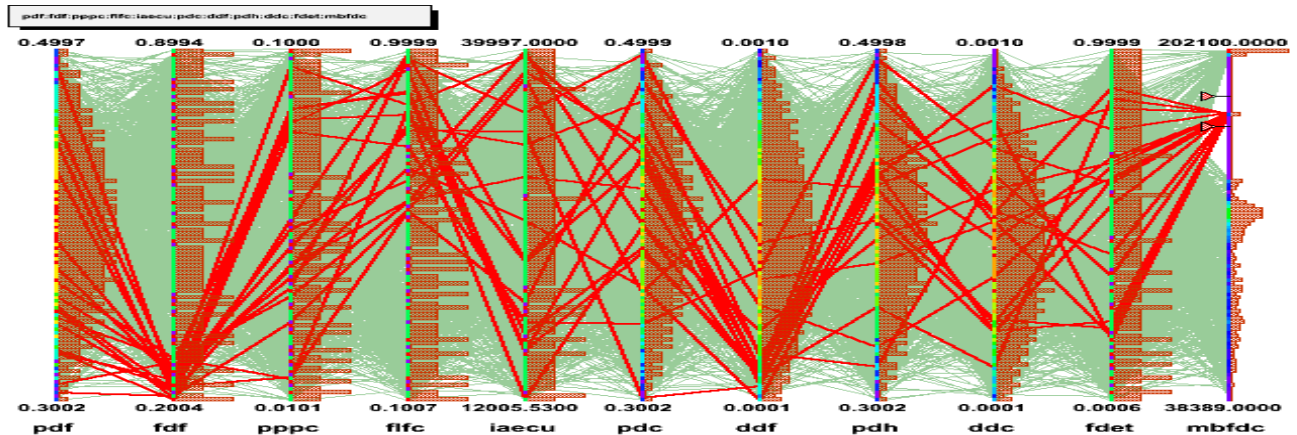
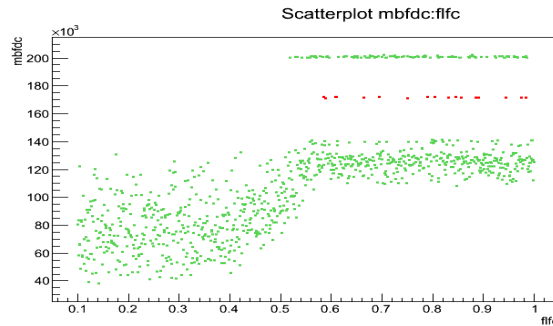


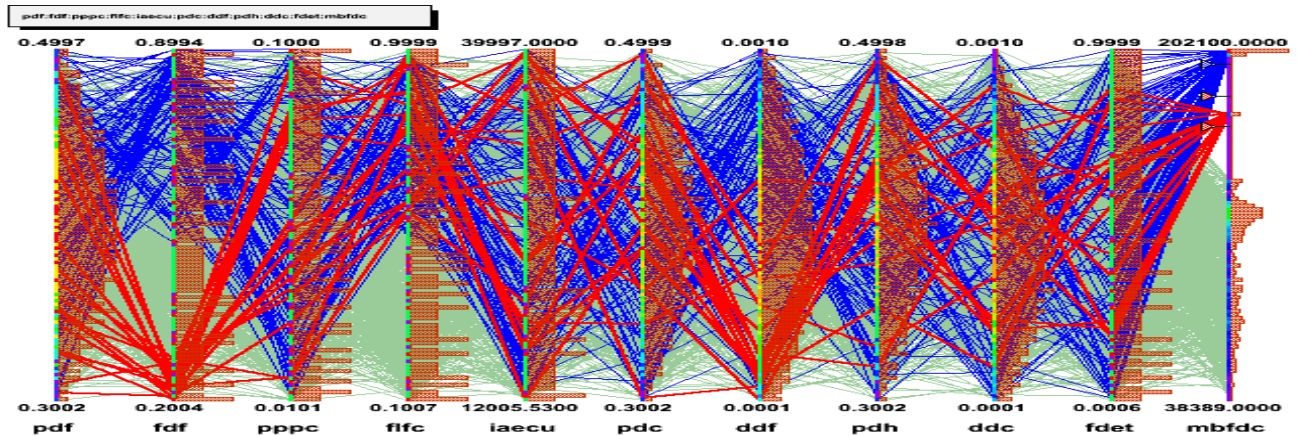
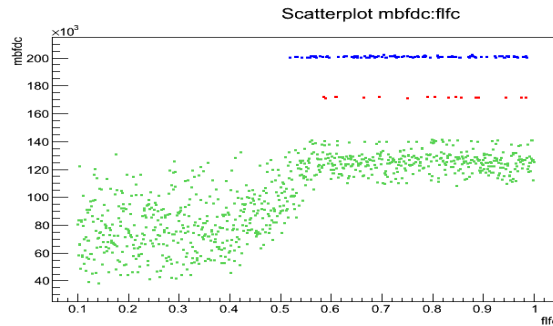


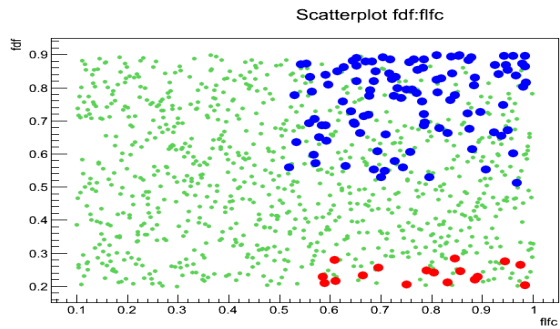
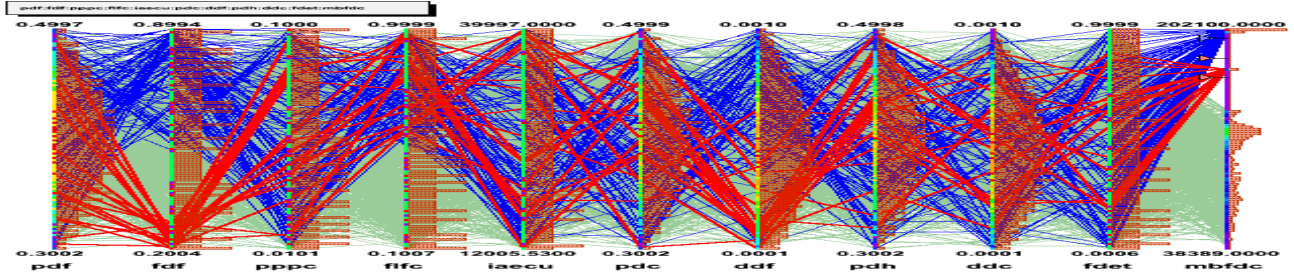
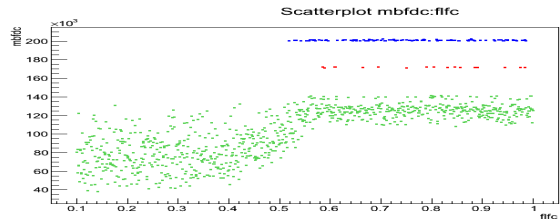
```
tds->Draw( "pdf:fdff:pppc:fffc:iaecu:pdc:ddf:pdh:ddc:fdet:mbfdc", "", "para");
```

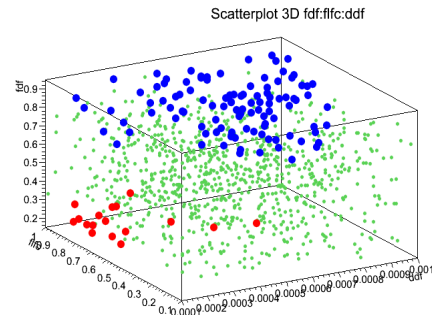
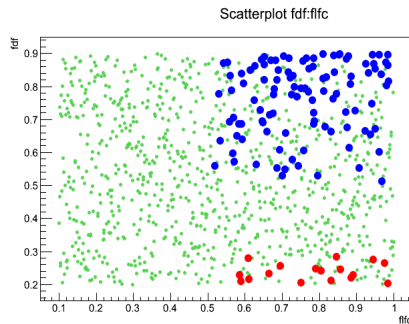
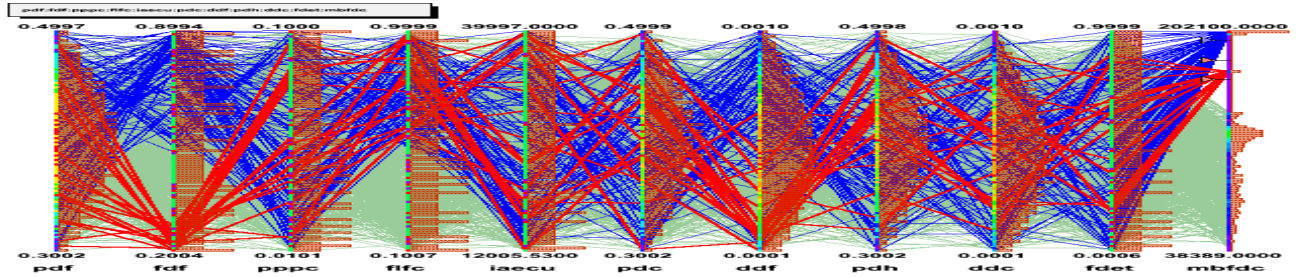
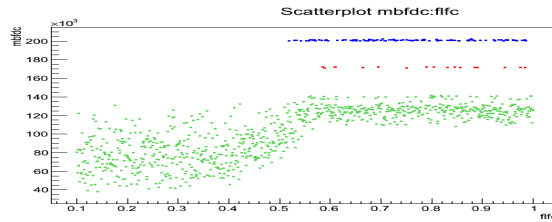














- The Uncertainties Propagation procedure
- A presentation of the Uranie Platform
- UQ example (*"powerful"* **CobWeb/Parallel Coordinates** graphic)





D'Agostini, R. and Stephens, M. (1986). *Goodness-of-Fit techniques*. Dekker, Statistics monographs, vol 68.

McKay, M. D. (1995). Evaluating Prediction Uncertainty. NUREG/CR-6311.



Saltelli, A., Chan, K., and Scoot, M., editors (2000). *Sensitivity Analysis*. John Wiley and Sons.



Saltelli, A., Tarantola, S., Campolongo, F. and Ratto, M. (2005). *Sensitivity Analysis in Practise*. John Wiley and Sons.



Fang, K.T., Li, R., and Sudjianto, A. (2006). *Design and Modeling for Computers Experiments*. Chapman.



Saltelli, and all (2008). *Global Sensitivity Analysis : the Primer*. John Wiley and Sons.



de Rocquigny E., Devictor N. and Tarantola S. (2008). *Uncertainty in Industrial Practice : A Guide to Quantitative Uncertainty Management*. John Wiley and Sons.





Thank you for your attention!

Questions?

