Evaluation de quantiles de codes de calcul

B. Iooss, N. Devictor (CEA Cadarache)
A. De Crecy, P. Bazin (CEA Grenoble)

Projet CEMRACS 2006

27/07/2006

Contexte : impact des incertitudes dans un processus

Modélisation des incertitudes (variables physiques, paramètres de modèles, …)

Processus (codes, expériences…)

- **Incertitudes sur les réponses**
- **Variables les plus influentes (indices de sensibilité)**
- **Probabilité Y > Y_{target}**

Évaluation des critères

Données

✓Modèle probabiliste des incertitudes sur les variables physiques, les paramètres…;

Exemples :

- Propriétés matériaux, géométrie, conditions environnementales
- Paramètres de convergence

✓Modèle de comportement, de dégradation ou de défaillance ;

✓Définition d'un seuil d'acceptation

Contexte CEA / Direction de l'Énergie Nucléaire

Recherche de méthodologies génériques sur les problématiques liées à la **gestion des incertitudes dans des processus.**

➡ **Développement d'outils non intrusifs, indépendants de la physique.**

Quelques exemples de domaines d'application à la DEN :

- Réacteurs : durée de vie des cuves (calculs mécaniques) ;
- Sûreté : simulation d'accident grave dans le cœur (thermohydraulique) ;
- Combustibles : modules thermomécaniques de calcul d'irradiation ;
- Maîtrise des Risques : calculs d'impact environnemental des radionucléides.

Quelques acteurs du CEA/DEN sur les incertitudes :

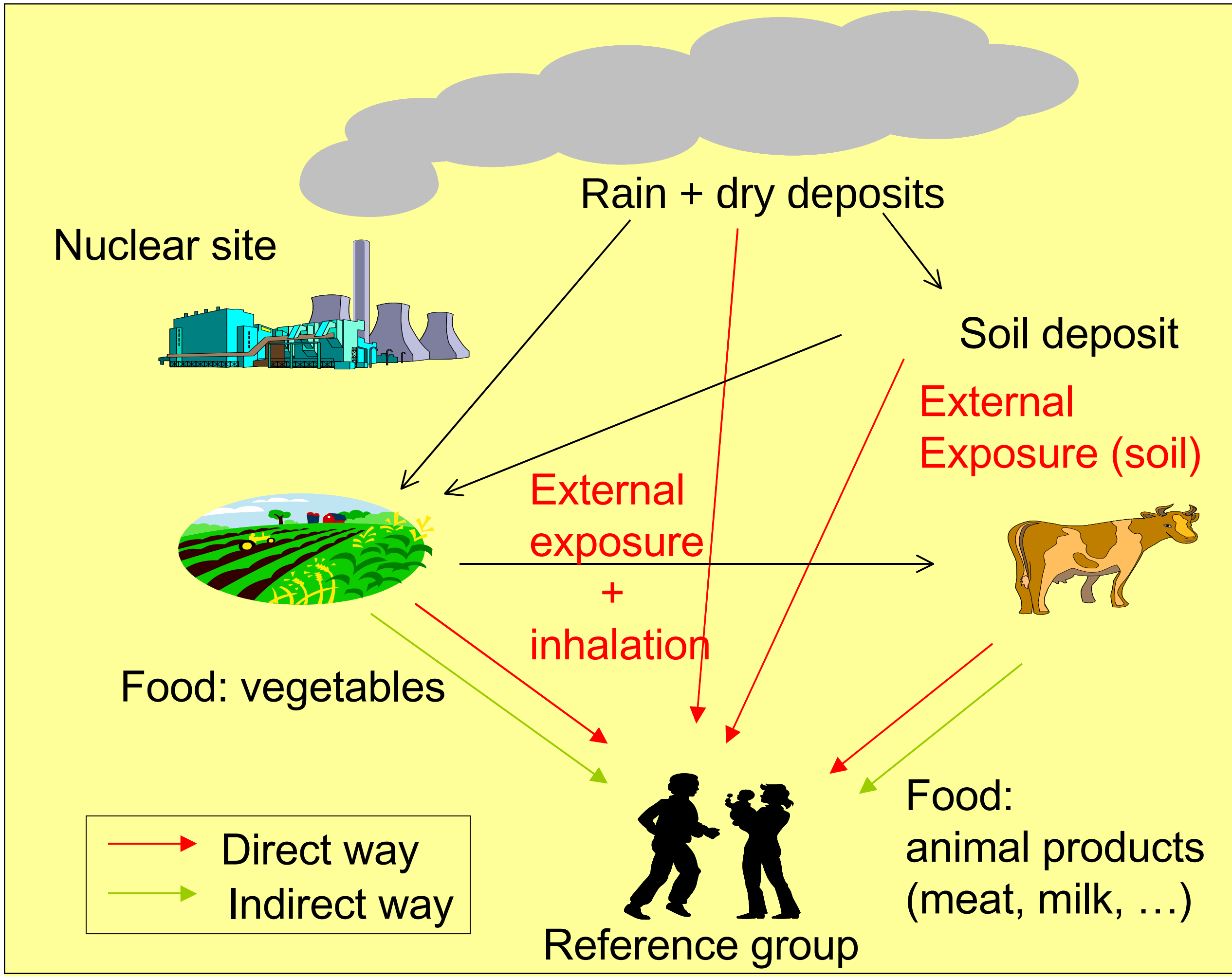
- Le Laboratoire de Conduite et Fiabilité des Réacteurs (Cadarache) travaille sur la thématique de la maîtrise des risques, équipe « incertitudes ».
- A Saclay, le LETR développe des méthodes et outils en modélisation, optimisation, thématique « apprentissage ».
- A Grenoble, le LDAS travaille sur le développement et les applications de modèles et codes thermohydrauliques.

Exemple 1 : calcul d'impact radiologique

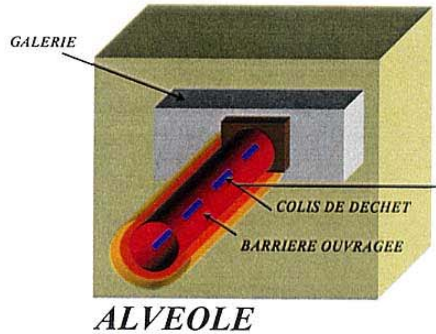
- Beaucoup de paramètres
- mal connus

→ Analyse de sensibilité des paramètres du code

→ Orientation des axes de R&D

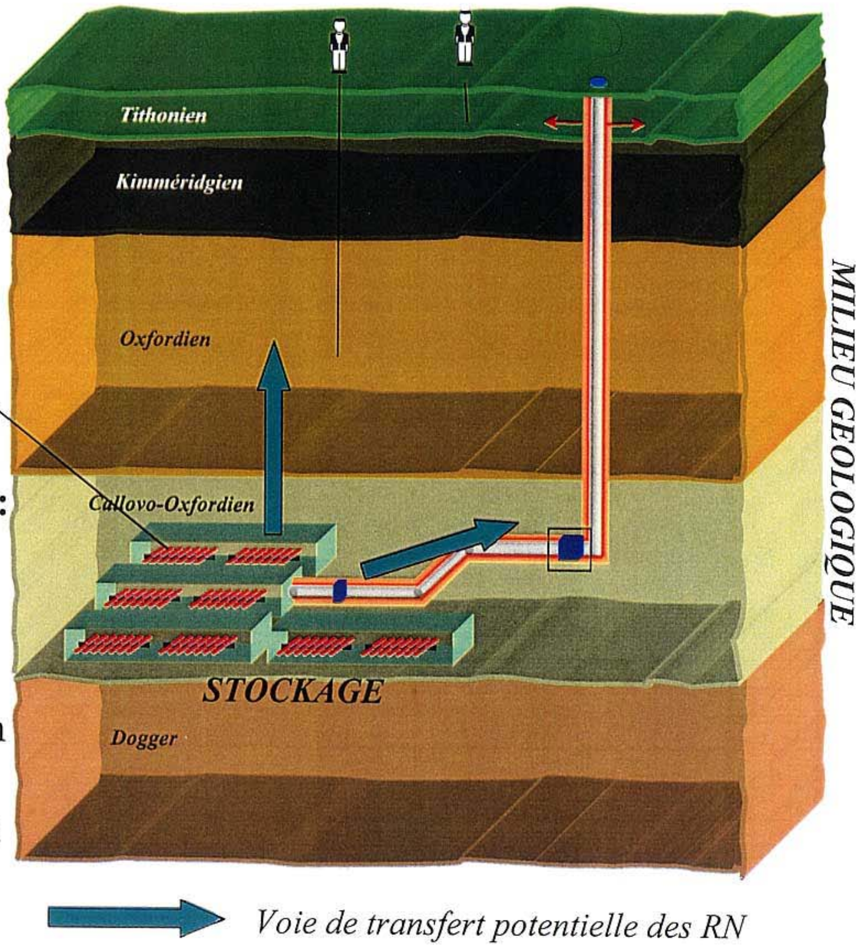


Exemple 2 : stockage profond (1/2)



Etude du transfert de la radioactivité (RN) depuis les colis de déchet jusqu'à l'homme :

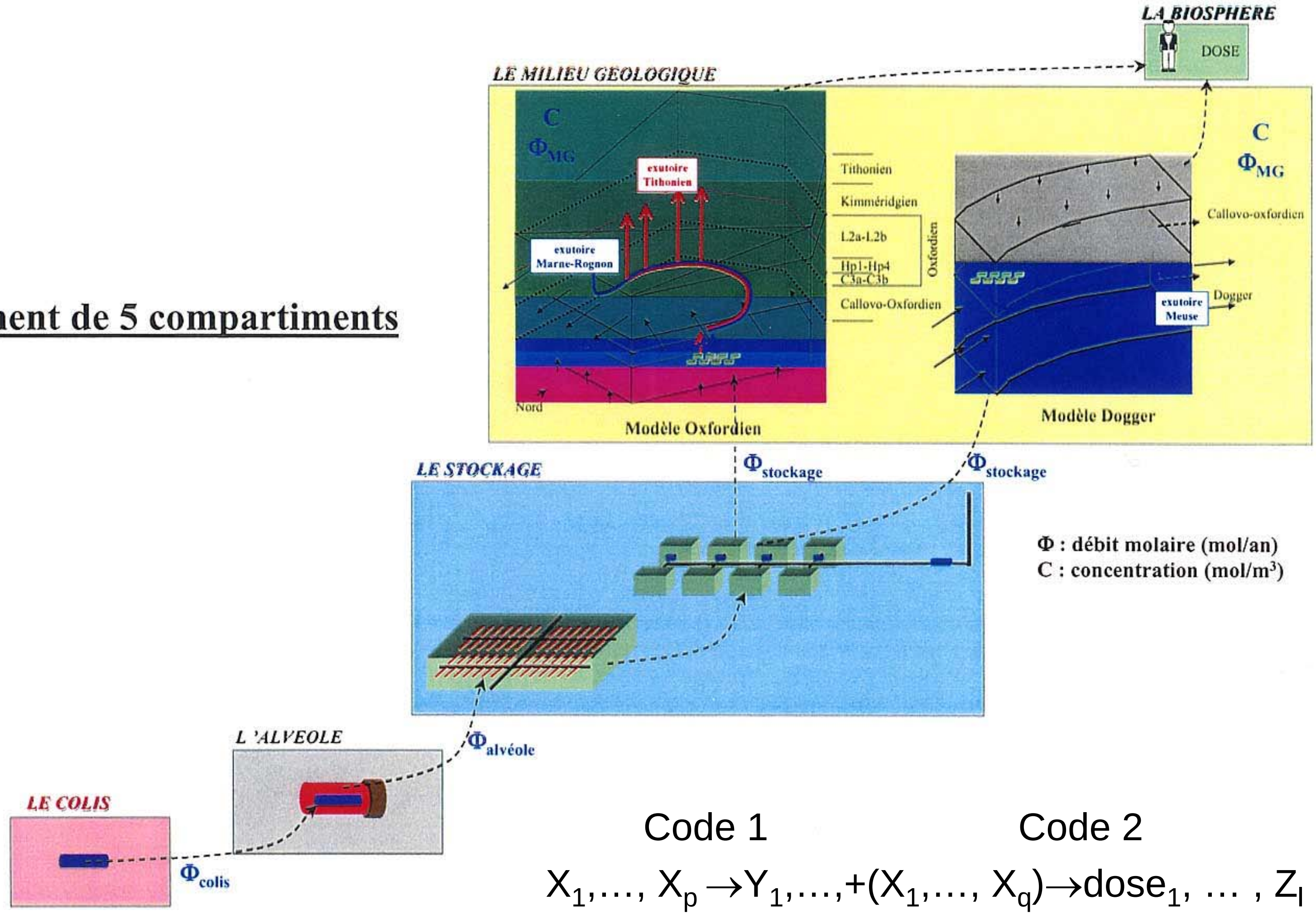
- transfert (par l'eau) en milieu poreux saturé à travers ≠ matériaux de caractéristiques spécifiques
- transport par convection/diffusion/dispersion
- prise en compte de la décroissance/filiation
- prise en compte des phénomènes de sorption et de précipitation chimique



- *Extrait d'une présentation de G. Pépin (ANDRA DS/CS)*

Stockage profond (2/2)

Enchaînement de 5 compartiments



Code 1 Code 2

$$X_1,\ldots, X_p \rightarrow Y_1,\ldots,+(X_1,\ldots, X_q)\rightarrow dose_1, \ldots , Z_l$$

$$\rightarrow Prob(dose_j > dose_{seuil})$$

- *Extrait d'une présentation de G. Pépin (ANDRA DS/CS)*

Propagation des incertitudes (PI) : Méthodes

- Question importante : Sous quelle forme est attendu le résultat? (dépendante de l'utilisation future)

- Distribution de probabilité
  - Simulation + ajustement sur l'échantillon généré + tests statistiques (en général asymptotiques)

- Premiers moments statistiques
  - Statistiques sur l'échantillon (convergence, bootstrap)
  - Composition de la variance (dérivées locales « faibles »)

$$s\left[f\left(x_1,\cdots,x_n\right)\right]=\left[\sum_i\left(\frac{\partial f}{\partial x_i}\right)^2 s^2\left(x_i\right)+2\sum_{i=1}^{n}\sum_{j=i+1}^{n}\left(\frac{\partial f}{\partial x_i}\right)\left(\frac{\partial f}{\partial x_j}\right)\operatorname{cov}\left(x_i;x_j\right)\right]^{1/2}$$

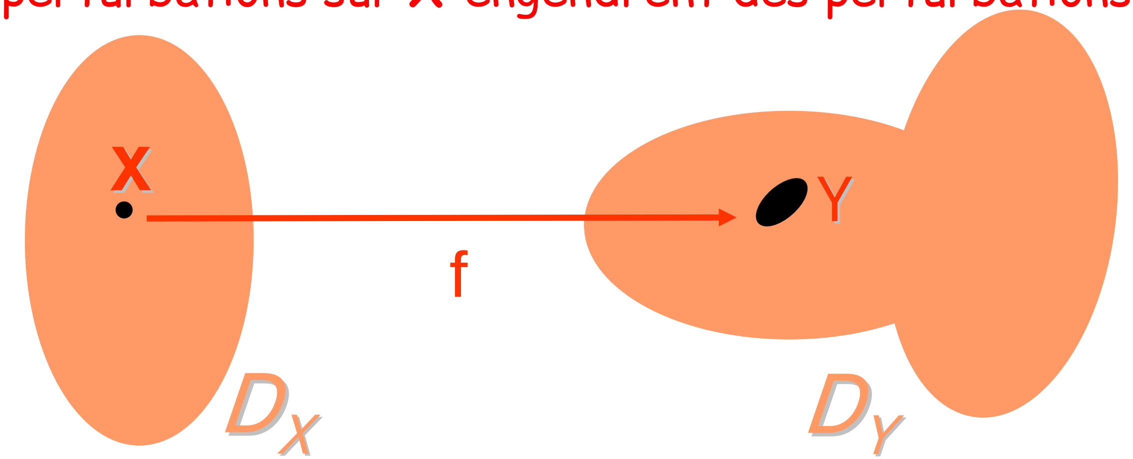
- Intervalle de confiance
  - A partir de la distribution de probabilité $P\{P(m\leq Y\leq M)\geq\alpha\}\geq\beta$
  - Formule de Wilks $1-\alpha^N-N(1-\alpha)\alpha^{N-1}\geq\beta$

« Analyse de sensibilité » ?

Soit le modèle

$$f : \mathfrak{R}^p \rightarrow \mathfrak{R}$$
$$X \rightarrow Y = f(X)$$

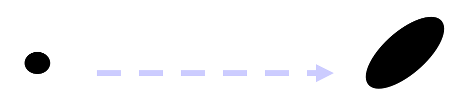
L'analyse de sensibilité (AS) étudie comment des perturbations sur **X** engendrent des perturbations sur Y



<u>BUTS :</u> développement, analyse, utilisation de modèles numériques

paramètres les + influents ⇨ **R&D prioritaire sur ces paramètres**

- AS locale
  Variation de la réponse par rapport à la variation d'une entrée (approche souvent déterministe, analyse différentielle)



- AS globale
  Quel est la contribution de l'incertitude de X_i (ou d'un ensemble de X_i) dans l'incertitude de Y ?

Méthodologie d'analyse de sensibilité globale

[Saltelli et al. (2000), McKay, Helton et al., Kleijnen, …]

Modèle déterministe : $f : \mathfrak{R}^p \rightarrow \mathfrak{R}$

$X \rightarrow Y = f(X)$

X est un vecteur de paramètres contrôlables

- Analyse d'incertitudes sur la réponse
- **Analyse de sensibilité entre la réponse et les entrées**

? Relation linéaire (R^2)

Oui → Coefficient de régression standard

$$SRC_i = \frac{\beta_i^2 \sigma(X_i)}{\sigma(Y)}$$

Non → **? Relation monotone** (R^2*)

- **Oui** → Coefficient de régression basés sur les rangs
- **Non** → Indices de Sobol

$$S_i = \frac{Var_{X_i}[E(Y/X_i)]}{Var(Y)}$$

nécessitent des milliers de calculs

Code trop coûteux ⟹ **Métamodèle**

Indices de sensibilité

Métamodèle (surface de réponse)

[Box & Draper (1987), Sacks et al. (1989), Kleijnen, …]

- **Fonction mathématique ou numérique qui peut remplacer le code de calcul dans le domaine de variation des paramètres influents.**

- *Exemples : polynômes, GLM, splines, réseaux de neurones, krigeage, …*

- **Qualités attendues d'une surface de réponse :**
  - représentative du code étudié, i.e. une **bonne approximation**,
  - ayant de **bonnes capacités de prédiction**,
  - **temps de calcul négligeable (<< 1 sec.).**

- **Utilités d'un métamodèle :**
  - Prédicteur rapide et facile à implémenter ;
  - Propagation d'incertitudes ;
  - Analyse de sensibilité globale quantitative ;
  - Optimisation/calibration de paramètres (très coûteux).

Construction et validation d'une surface de réponse

- La construction dépend :
  - de l'échantillon **D** de points ($\mathbf{x}(i)$, $z(i)$), i variant de 1 à N,
  - de la famille F de fonctions $f(\mathbf{x},\mathbf{c})$, avec **c** vecteur de paramètres,
  - de la minimisation d'une fonction risque.

$$R_E(f) = \frac{1}{N}\sum_{i=1}^{N} L\left(z(i), f(x(i), c)\right)$$

- Compromis « biais/variance » : modèle simple/modèle complexe.

- Définition de « bonne approximation » : utilisation de la SR (calibration, modélisation, prédiction), compatible avec l'erreur du modèle, contraintes (conservatisme, bonne représentation dans un sous-domaine, …).

- **Outils :**
  - Comparaison des résultats à une référence.
  - **Critères statistiques** : analyse de la régression, analyses des résidus (R^2, MSE, biais, normalité, …) et des résidus relatifs.
  - Base de construction (apprentissage) / base de test (généralisation).
  - **Validation croisée (qq centaines de données).**
  - Méthode de rééchantillonnage, de bootstrap (< 100 données).

Le problème : estimation de quantiles

- **Déterminer Y_s tel que $P(Y < Y_s) \geq \alpha$**

  Où Y=f(X) est la réponse d'un code de calcul dont les entrées X sont aléatoires.

  On aimerait aussi avoir un intervalle de confiance sur ce quantile

- Pour assurer le conservatisme, on introduit parfois un niveau de confiance β sur la probabilité :

  $P[\, P(Y < Y_s) \geq \alpha \,] \geq \beta$

- **Solution simple** : simulations Monte-Carlo.

  Si α=95%, cela nécessite plusieurs centaines de calculs.

  Cela n'est souvent pas possible en pratique.

- **Objectif du projet : estimation précise d'un quantile à 95% avec moins de 200 calculs.**

- **Outil classique dans le nucléaire : la formule de Wilks.**

Formule de Wilks

De manière générale, Wilks fournit pour un N-échantillon suivant une loi de probabilité (inconnue) que l'on a ordonné : $Y_{(1)},\ldots,Y_{(p1)},\ldots Y_{(p2)},\ldots,Y_{(N)}$ la probabilité P de l'intervalle $[Y_{(p1)},Y_{(p2)}]$ (resp. que la valeur $Y_{(N-r+1)}$ où r représente l'ordre décroissant) de représenter un α-fractile bilatéral (resp. un α-fractile unilatéral). Il en fournit la fonction de répartition (G(P)).

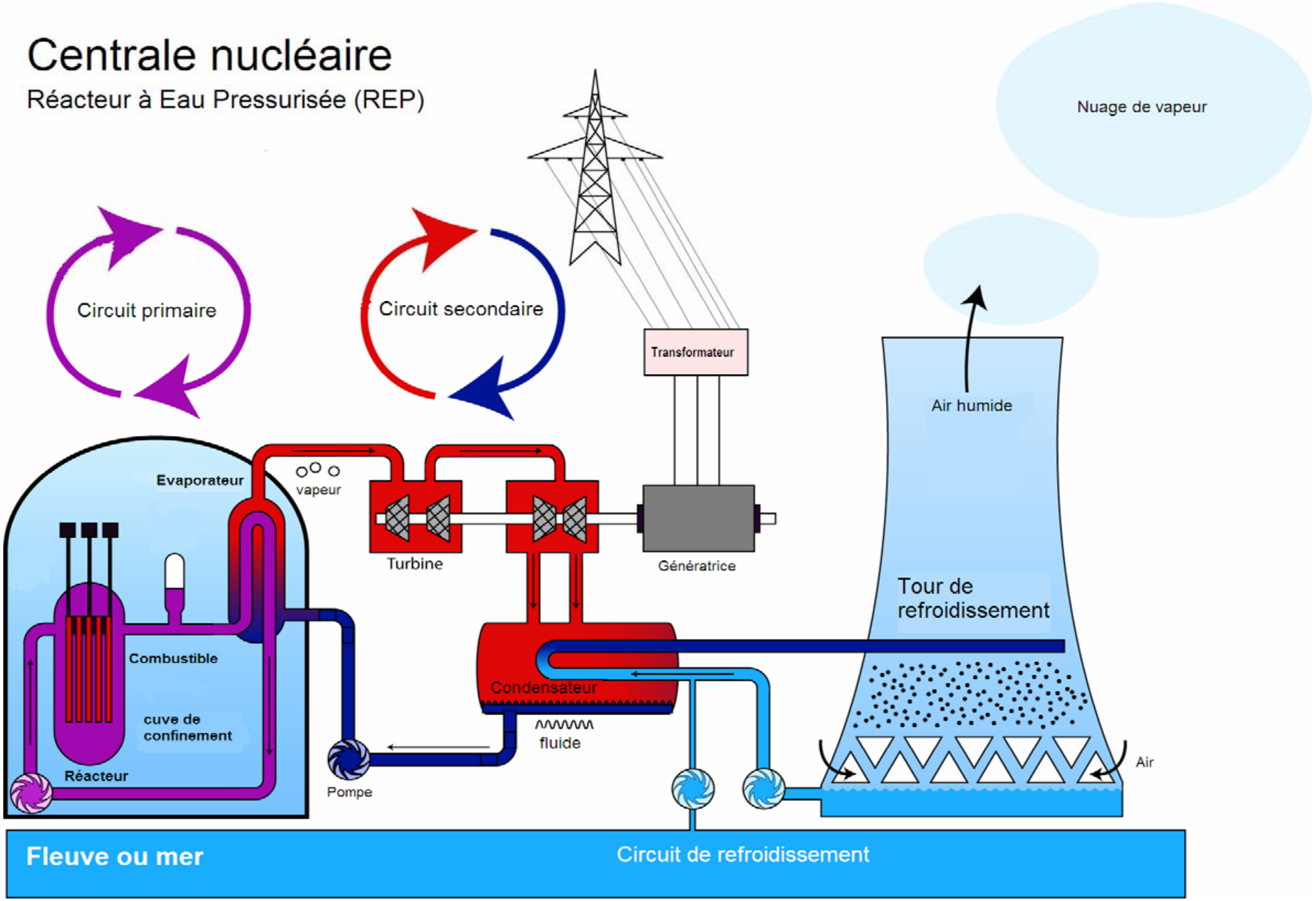
En général, on cherche à déterminer un α-fractile unilatéral (avec α = 95 %) c.'à.d. tel que $\mathrm{Prob}\{Y < Y_\alpha\} = \alpha$ avec un degré de confiance égal à β ; le nombre de calculs **N** et l'ordre **r** sont liés par la relation :

$$1-G(\alpha)=1-\sum_{N-r+1}^{N} C_N^i \alpha^i (1-\alpha)^{(N-i)} \geq \beta \quad \text{qui devient pour } r=1 : 1-\alpha^N \geq \beta$$

Pour $\alpha = \beta$ = 95% les couples **r** (ordre) et **N** remplissent ces conditions :

- r = 1, N =59
- r = 2, N = 93
- r = 3, N = 124
- …
- r = 39, N = 1000

Application à la sûreté des réacteurs



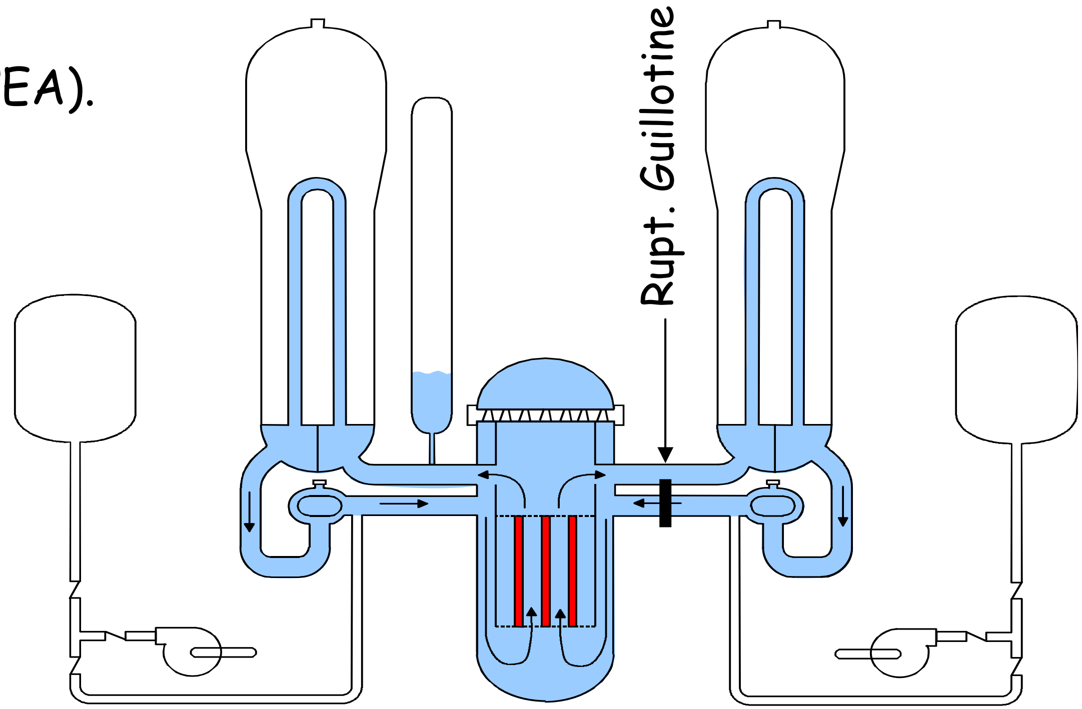
Estimation de quantiles : le projet BEMUSE

- Piloté par le CEA Grenoble/LDAS (A. De Crecy, P. Bazin), participation de : IRSN, GRS, KAERI, JNES, …

- Benchmark OCDE/NEA : **analyses d'incertitudes et de sensibilité sur les calculs thermohydrauliques de sûreté des réacteurs nucléaires.**

  Scénario : Accident de Perte de Refroidissement Primaire – GB

- Logiciel CATHARE (CEA).



Objectifs de l'analyse d'incertitude

Vérifier que le calcul de ce scénario accidentel, compte tenu des incertitudes sur les résultats, respecte un certain nombre de critères.

Par exemple, température maximale de gaine reste inférieure à 1204 C :

$$T_{gaine\_max} + \text{incertitude} < 1204\ C$$

Objectif important de BEMUSE : obtention du quantile à 95% du premier pic de température de gaine (PCT)

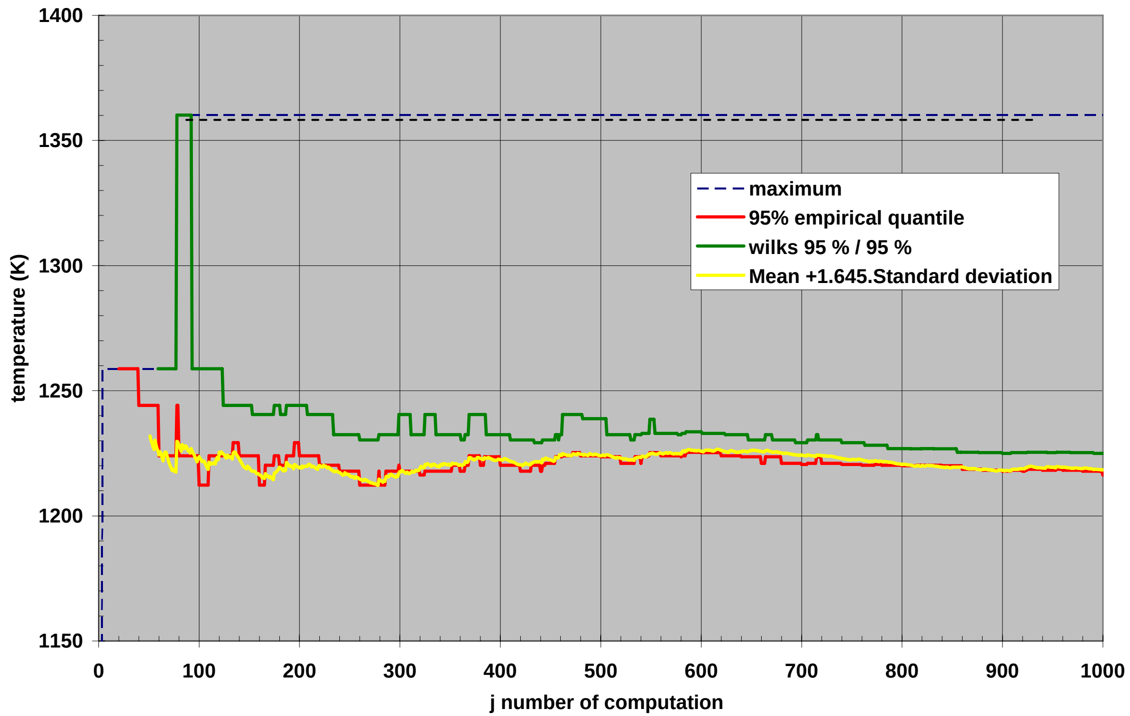
Étude CEA : 52 paramètres incertains de CATHARE sont identifiés avec leur loi de probabilité.

Pour estimer la variabilité de différents estimateurs, 1000 calculs sont réalisés ; en pratique on souhaiterait en réaliser moins de 200.

Estimation du quantile :

- ✓ Formule de Wilks
- ✓ Étude préliminaire sur l'utilisation des métamodèles

Bilan de l'étude



On rééchantillonne dans la base des 1000 points :

Wilks à l'ordre 1 (N=59) : très conservatif, forte dispersion du quantile, risqué (parfois < q_{emp}).

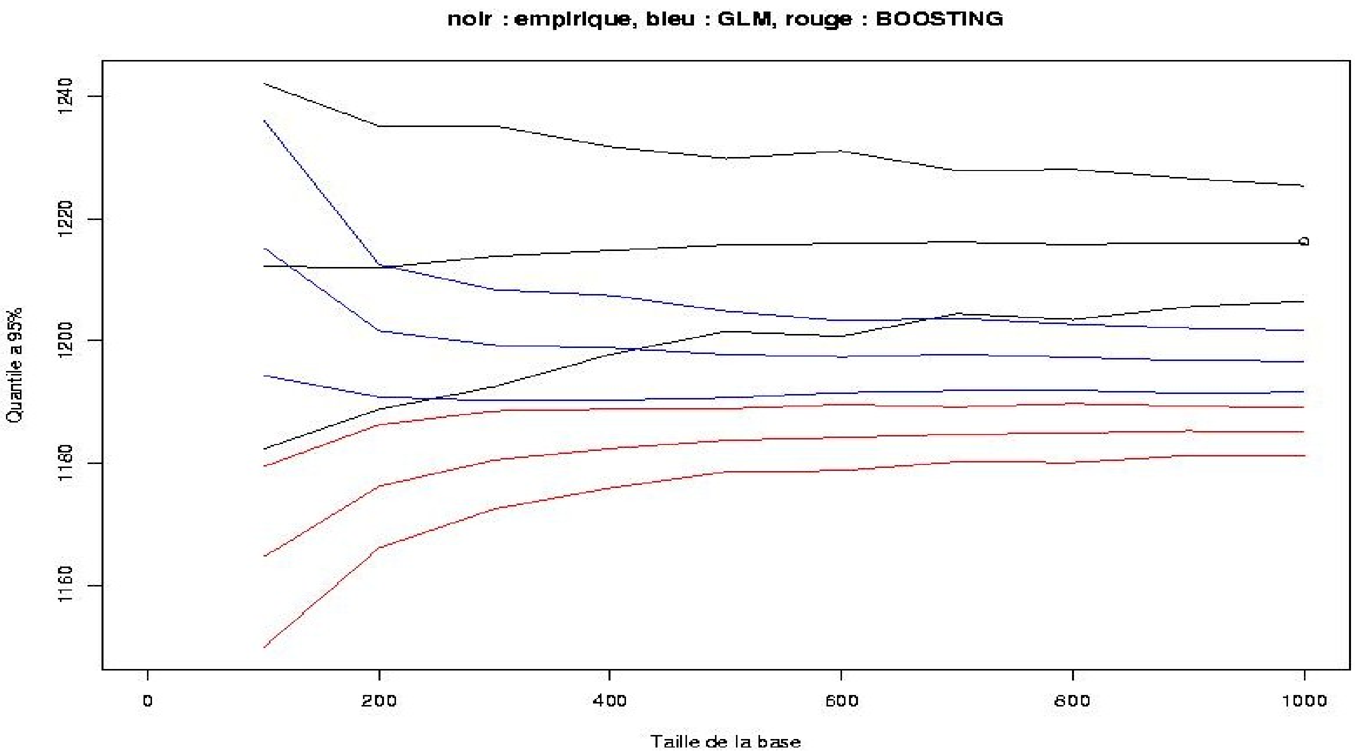
Conseil : estimation avec Wilks à l'ordre 3 (N=124) ou 4 (N=153) (la dispersion est + réduite), ou alors quantile emp. avec un IC bootstrap.

Estimation de quantiles à l'aide de SR

- Construction de SR (GLM=modèle linéaire et Boosting trees) en faisant varier le nombre de points (N=100 à 1000) ; R^2 sur base de test
- Estimation de quantiles à partir des SR ($n=10^4$)
- Bootstrap sur la base totale → SR → IC ($\pm 2\sigma$) sur quantiles

Conclusion : sous-estimation importante du quantile par SR :

les SR ajustées sur tous les points lissent les données et ne reproduisent pas la variabilité haute fréquence du phénomène.



N	GLM	Boosting
100	74%	63%
400	80%	84%
900	80%	87%

Valeurs du R^2 (test) en fct de N

Quelques pistes pour estimer les quantiles

- Travailler sur l'échantillonnage :
  - Tirage d'importance pour viser le quantile ;
  - Plans d'expérience séquentiels.

- Construction de SR spécifiques pour estimer les quantiles :
  - Utiliser la variance des résidus.

    Par exemple, Oakley (2004) utilise le modèle des processus gaussien et le MSE associé.
  - Régression sur les quantiles (Koenker, 2005) ; que fait-on du quantile conditionnel (déconditionnement) ?
  - Utilisation des résidus : addition de quantiles, modélisation ?
  - Autres ?

Nota Bene :

1. Avant l'application sur CATHARE, on souhaite étudier les diverses stratégies sur des fcts analytiques simples.

2. Pour l'application sur CATHARE, on peut utiliser les 1000 calculs pour étudier la variabilité des estimateurs.