



HAL
open science

Inverse uncertainty quantification applied to thermal-hydraulic simulations

G. Damblin, Pierre Gaillard

► **To cite this version:**

G. Damblin, Pierre Gaillard. Inverse uncertainty quantification applied to thermal-hydraulic simulations. MIA PARIS-Rochebrune 2018 (Statistics seminar), Nov 2018, Rochebrune, France. cea-02338934

HAL Id: cea-02338934

<https://cea.hal.science/cea-02338934>

Submitted on 12 Dec 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

DE LA RECHERCHE À L'INDUSTRIE



Inverse uncertainty quantification applied to thermal-hydraulic simulations

*Guillaume Damblin*¹ *Pierre Gaillard*²

¹ CEA DEN/DANS/DM2S/STMF/LGLS
² AREVA NP

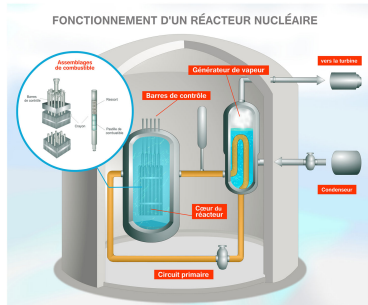
Statistics seminar in Rochebrune

26/03/2018

- 1 Motivations for inverse UQ in TH simulations
- 2 The CIRCE method
- 3 The Bayesian counterpart of CIRCE
- 4 Non linear generalization
- 5 Conclusions

Motivations for inverse UQ in TH simulations

- Best Estimate thermal-hydraulic system code developed by CEA,
- Based on six balance equations: mass, momentum and energy conservation
 - require building **correlations** (also called closure laws or physical models)
- Nuclear simulations with several levels of complexity:
 - Separate/Combined Effects Test (SET, CET)
 - at reduced scale, few physical phenomena
 - Integral Effect Test (IET)
 - many phenomena together.
- Simulating accidental transients for safety analysis
- Great effort devoted to V&V implementation
 - Verification : Are the equations solved right ?
 - Validation : Are the right equations solved ?



CONCEPTION OF CORRELATIONS

- ex: heat transfers (convection, condensation, etc):

$$C_{nom}(\mathbf{x}, \theta)$$

where \mathbf{x} is a vector of physical variables and θ is a fitting parameter.

- parameter uncertainty affecting θ (neglected by physicists)

V&V IMPLEMENTATION

- Verification stage : numerical uncertainties (ex: mesh convergence)
- Validation stage : where CATHARE 2 predictions are confronted to experimental data from SET.
 - correlation uncertainty assessed from differences between both of them
 ⇒ **inverse UQ process**

The CIRCE method

MAIN ASSUMPTIONS

- Model uncertainty is multiplicative:

$$C_{\Lambda}(\mathbf{x}) = \Lambda \times C_{nom}(\mathbf{x})$$

- Λ follows a probability distribution
- Λ is log-Gaussian, calculated by the **CIRCE method** (De Crécy and Bazin, 2001).

CIRCE = *Calcul des Incertitudes Relatives aux Corrélations Élémentaires*.

STATISTICAL MODELING

- $z_i \in \mathbb{R}$ the QoI experimentally measured at $\mathbf{x}_i \in \mathbb{R}^m$
- $Y(\cdot)$ the CATHARE 2 code (used as a black-box function)
- For $i \in \llbracket 1; n \rrbracket$, we assume that

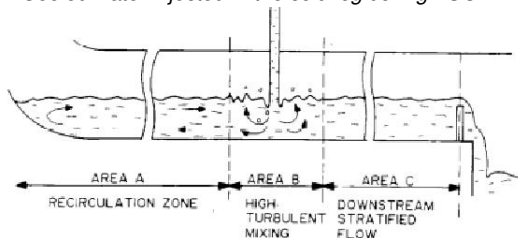
$$\begin{aligned} z_i &= Y(C_{\lambda_{1,i}}(\mathbf{x}_i), \dots, C_{\lambda_{p,i}}(\mathbf{x}_i)) + \epsilon_i \\ &= Y_{(\lambda_{1,i}, \dots, \lambda_{p,i})}(\mathbf{x}_i) + \epsilon_i \end{aligned}$$

where

- $\lambda_{j,i} \sim \Lambda_j = \mathcal{LN}(m_j, \sigma_j^2)$, $j \in \llbracket 1; p \rrbracket$
- $\alpha_{j,i} = \log(\lambda_{j,i}) \sim A_j = \mathcal{N}(m_j, \sigma_j^2)$,
- $\epsilon_i \sim \mathcal{N}(0, \sigma_{\epsilon_i}^2)$.

The CIRCE method consists in estimating m_j and σ_j^2 for $j \in \llbracket 1; p \rrbracket$

Cooled water injected in the cold leg during LOCA



- One correlation per area,
- We focus on area B and C ,
- Condensation higher in the area B than area C ,
- Q_i = condensation flow rate measurement to the edge of area C (kg/s),

■ The CATHARE 2 code can predict Q by using two correlations of condensation $C_{\lambda_1}(\mathbf{x}_i)$ (area B) and $C_{\lambda_2}(\mathbf{x}_i)$ (area C)

■ $Q_i = Y_{(\lambda_{1,i}, \lambda_{2,i})}(\mathbf{x}_i) + \epsilon_i$
 where \mathbf{x}_i includes injection pressure, injection temperature, water height in the cold leg, etc.

OTHER ASSUMPTIONS UNDERLYING CIRCE:

- The factors are not correlated each other: $Cov(\Lambda_j, \Lambda_k) = 0$; $1 \leq j \neq k \leq p$,
- The experimental variances $\sigma_{\epsilon_i}^2$ are assumed known.

CIRCE IMPLEMENTATION:

1. Linearization at $\alpha_{nom} = \log(\lambda_{nom})$, typically at the nominal model $\mathbf{0}_p = \log \mathbf{1}_p$
 - $z_i - Y_i^{nom} = h_i^T (\alpha_i - \alpha_{nom}) + \epsilon_i$ with $\alpha_i := \log \lambda_i$
 - Identifiability: $\text{rank}(H) = p$ where $H = [h_1, \dots, h_n]^T \in \mathcal{M}_{np}$.
2. Computation of MLE estimates $(\hat{m}_j, \hat{\sigma}_j^2)$ using an EM algorithm:
 - Both E and M steps are explicit,
 - ECME to speed up the convergence (Celeux et al., 2010).
3. Post treatment:
 - Statistical analysis of residuals, LOO cross validation,
 - Check the linearity assumption on

$$IF_{0.95}(A_j) = [\hat{m}_j - 1.96\hat{\sigma}_j, \hat{m}_j + 1.96\hat{\sigma}_j], \quad j \in \llbracket 1; p \rrbracket,$$

- Deduce the 95%-interval of Λ_j :
 $IF_{0.95}(\Lambda_j) = [\exp(\hat{m}_j - 1.96\hat{\sigma}_j), \exp(\hat{m}_j + 1.96\hat{\sigma}_j)], \quad j \in \llbracket 1; p \rrbracket$

The Bayesian counterpart of CIRCE

Notations:

- $\mathbf{z} = [z_1, \dots, z_n]^T \in \mathbb{R}^n$ the matrix of field measurements
- $\alpha = [\alpha_1, \dots, \alpha_n]^T \in \mathcal{M}_{np}$ the matrix of missing model log-samples:
- $m = (m_1, \dots, m_p)^T \in \mathbb{R}^p$ and $\sigma^2 = (\sigma_1^2, \dots, \sigma_p^2)^T \in \mathbb{R}^p$.

STATISTICAL MODEL

- $z_i = h_i^T \alpha_i + \epsilon_i$ for $i \in \llbracket 1; n \rrbracket$,
 - $z_i \in \mathbb{R}^q$; $h_i \in \mathbb{R}^p$; $\alpha_i \sim \mathcal{N}(m, \sigma^2) \in \mathbb{R}^p$;

POSTERIOR DISTRIBUTION

- Bayes formula gives $[m, \sigma^2 | \mathbf{z}, \alpha] \propto [\mathbf{z}, \alpha | \sigma^2, m][m, \sigma^2]$
 - Likelihood: $\mathbf{z} | \alpha, \sigma^2, m \sim \otimes_{i=1}^n \mathcal{N}(h_i^T \alpha_i, R_i)$,
 - Prior: $[m, \sigma^2] = [m | \sigma^2][\sigma^2]$
 - Conjugate Gaussian-inverse-gamma,
 - Gaussian for $m | \sigma^2$ along with a folded non-standardized-t for σ (Gelman, 2006).

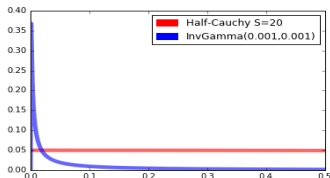
INVERSE-GAMMA (ϵ, ϵ) FOR σ_j^2

- Leads to an improper posterior as $\epsilon \rightarrow 0$.
 - Spiegelhalter et al. (2004) took $\epsilon = 0.001$,
 - Inference is sensible to ϵ (mainly when low values of σ provide large likelihood values),
 - Such diffuse priors cannot fix troubles with improper posteriors (Kass and Wasserman, 1996).

FOLDED NON-STANDARDIZED STUDENT DISTRIBUTION FOR σ_j (via the augmented model)

- $z_i = h_i^T \times (C\tilde{\alpha}_i) + \epsilon_i$ for $i \in \llbracket 1; n \rrbracket$, with $\alpha_i = C\tilde{\alpha}_i$
 - Priors: $C \sim \mathcal{N}(m_C, 1)$ and $\sigma_{\tilde{\alpha}}^2 \sim \mathcal{IG}(0.5 \times \nu, S)$
 - Thus, $\sigma = |C|\tilde{\sigma}$ is a folded noncentral-t
 - half-t if $m_C = 0$,
 - half Cauchy if $m_C = 0$ and $\nu = 1$ (which tends to be uniform on \mathbb{R}^+ as $S \rightarrow +\infty$)

$$[\sigma] \propto \frac{1}{\sigma^2 + S}$$



SUBSTITUTION (OR DATA-AUG.) SAMPLING (Gelfand and Smith, 1990)

By following the hierarchical structure $[m, \sigma^2, \alpha | \mathbf{z}] = [m, \sigma^2 | \alpha, \mathbf{z}][\alpha | \mathbf{z}]$

- Start with a first sample (m_0, σ_0^2)
- In a loop $k \geq 1$, sample until convergence :
 1. $\alpha_k \sim \alpha | \mathbf{z}, m_{k-1}, \sigma_{k-1}^2$ (Gaussian),
 2. $m_k, \sigma_k^2 \sim m, \sigma^2 | \alpha_k, \mathbf{z}$ (Gaussian-inverse-gamma).

GIBBS SAMPLING

Based on the full conditional posterior distributions

- Start with a first sample (m_0, σ_0^2)
- In a loop for $k \geq 1$, sample until convergence :
 1. $\alpha_k \sim \alpha | \mathbf{z}, m_{k-1}, \sigma_{k-1}^2$ (Gaussian),
 1. $m_k \sim m | \sigma_k^2, \alpha_k, \mathbf{z}$ (Gaussian),
 2. $\sigma_k^2 \sim \sigma^2 | b_k, \alpha_k, \mathbf{z}$ (inverse-gamma),

SUBSTITUTION (OR DATA-AUG.) SAMPLING (Gelfand and Smith, 1990)

Full posterior $[\tilde{m}, \tilde{\sigma}^2, \tilde{\alpha}, C | \mathbf{z}] = [\tilde{m}, \tilde{\sigma}^2 | \tilde{\alpha}, \mathbf{z}, C][\tilde{\alpha}, C | \mathbf{z}]$

- Start with a first sample $(\tilde{m}_0, \tilde{\sigma}_0^2, C_0)$
- In a loop for $k \geq 1$, sample until convergence :
 1. $\tilde{\alpha}_k \sim \tilde{\alpha} | \mathbf{z}, \tilde{m}_{k-1}, \tilde{\sigma}_{k-1}^2, C_{k-1}$ (Gaussian),
 2. $C_k \sim C | \mathbf{z}, \tilde{\alpha}_k$ (Gaussian),
 3. $\tilde{m}_k, \tilde{\sigma}_k^2 \sim \tilde{m}, \tilde{\sigma}^2 | \tilde{\alpha}_k, \mathbf{z}, C_k$ (Gaussian-inverse-gamma).

GIBBS SAMPLING

Based on the full conditional posterior distributions:

- Start with a first sample $(\tilde{m}_0, \tilde{\sigma}_0^2, C_0)$
- In a loop for $k \geq 1$, sample until convergence :
 1. $\tilde{\alpha}_k \sim \tilde{\alpha} | \mathbf{z}, \tilde{m}_{k-1}, \tilde{\sigma}_{k-1}^2, C_{k-1}$ (Gaussian),
 2. $C_k \sim C | \mathbf{z}, \tilde{\alpha}_k$ (Gaussian),
 3. $\tilde{\sigma}_k^2 \sim \tilde{\sigma}^2 | \tilde{m}_k, \tilde{\alpha}_k$ (inverse-gamma),
 4. $\tilde{m}_k \sim \tilde{m} | \tilde{\sigma}_k^2, \tilde{\alpha}_k$ (Gaussian),

The Sobol indice for Model A_j and experiment i quantifies the fraction of the output variance that is due to A_j .

$$S_j(\mathbf{x}_i) = \frac{\mathbb{V}[z_i] - \mathbb{E}[\mathbb{V}[z_i|A_j]]}{\mathbb{V}[z_i]} = \frac{\mathbb{V}[\mathbb{E}[z_i|A_j]]}{\mathbb{V}[z_i]} = \frac{\sigma_j^2 \times h_i(j)^2}{h_i^T \text{diag}(\sigma^2) h_i + \sigma_{\epsilon_i}^2}$$

Based on the marginal likelihood $[\mathbf{z}|m, \sigma^2]$ after integrating with respect to the missing samples, we can prove that the Fisher information matrix is written as

$$I_n(m, \sigma^2) = \begin{pmatrix} I_n(m) & 0 \\ 0 & I_n(\sigma^2) \end{pmatrix}$$

where

$$I_n(m)_{j,k} = \sum_{i=1}^n \frac{h_i(j)h_i(k)}{h_i^T \text{diag}(\sigma^2) h_i + \sigma_{\epsilon_i}^2} \quad 1 \leq j, k \leq p$$

Therefore, we can get

$$I_n(m_j) = \frac{n\bar{S}_j}{\sigma_j^2}$$

and

$$I_n(\sigma^2)_{j,k} = \sum_{i=1}^n \frac{0.5 \times h_i^2(j)h_i^2(k)}{h_i^T \text{diag}(\sigma^2) h_i + \sigma_{\epsilon_i}^2} \quad 1 \leq j, k \leq p$$

and

$$I_n(\sigma_j^2) = \frac{nS_j^2}{2\sigma_j^4}$$

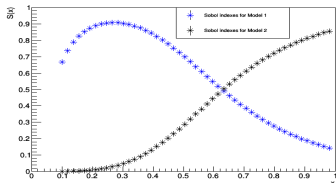
- The smaller the Sobol index S_Λ , the less accurate the estimates (Celeux et al., 2010):
 - Bayesian counterpart?
 - studying the role of the prior in terms of size of credible regions.

- Well-posedness principles in inverse UQ:
 - in the Hadamard sense (condition number as low as possible)
 - in the Sobol sense: $S_\Lambda > S_\epsilon$ (i.e. the input contribution to the randomness of \mathbf{Z} is larger than that of the noise)
 - in the entropy sense, in the Fisher sense (Bousquet and Blazère, 2016).

- In real inverse UQ problems, the Sobol indices are unknown
 - the matrix H can provide a local sensitivity measure

Synthetic example: $x_i \in [0.1, 1]$, $\alpha = (\alpha_1, \alpha_2)$, $n = 50$

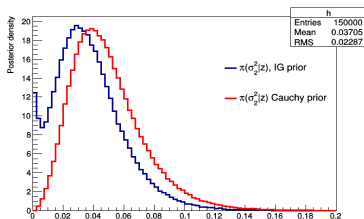
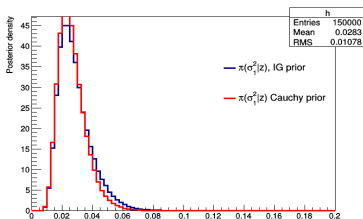
- $z_i = x_i \alpha_{1,i} + \mathbf{1.6} \times x_i^3 \alpha_{2,i} + \epsilon_i$
 - $\alpha_{1,i} \sim \mathcal{N}(2, 0.02^2)$ and $\alpha_{2,i} \sim \mathcal{N}(2, 0.05^2)$
 - $\epsilon_i \sim \mathcal{N}(0, 0.01^2)$



- Sobol indices $S(x_i)$ against $x \in [0.1, 1]$
- In averaging over $x \in [0.1, 1]$, Model 1 gets higher Sobol indices than Model 2 :
 - $\bar{S}_1 = 0.57$, $\bar{S}_1 = 0.39$ and $\bar{S}_\epsilon = 0.04$

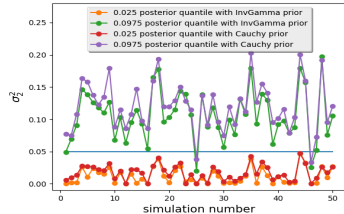
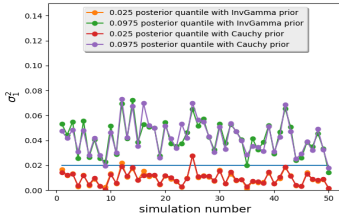
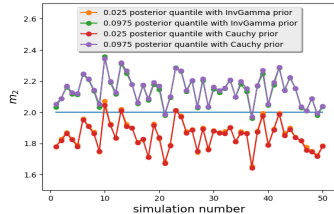
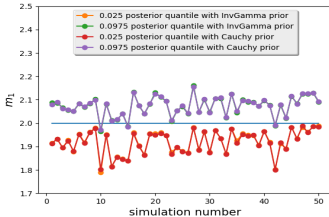
■ Comparison of marginal posterior distributions $[\sigma_j^2 | \mathbf{z}]$ according to the prior distribution in attempting to make a default Bayesian estimation:

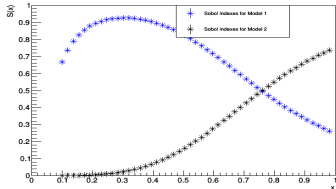
- $\mathcal{IG}(\epsilon, \epsilon)$ with $\epsilon = 10^{-3}$ for σ_j^2 vs half-Cauchy with $S = 20$ for σ_j ($j = 1, 2$)



■ Comparison being done over 50 simulated data set:

- credible intervals at 95% are calculated in two cases: $\mathcal{IG}(0.001, 0.001)$ vs half-Cauchy with $S = 20$



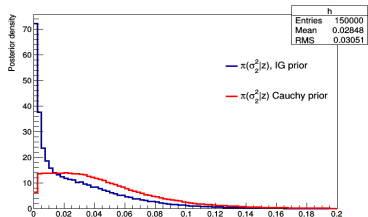
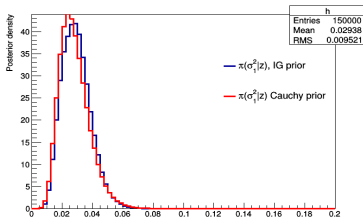


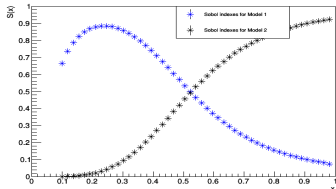
Synthetic example: $x_i \in [0.1, 1]$,
 $\alpha = (\alpha_1, \alpha_2)$, $n = 50$

- $z_i = x_i \alpha_{1,i} + 1.1 \times x_i^3 \alpha_{2,i} + \epsilon_i$
- $\bar{S}_1 = 0.67$, $\bar{S}_1 = 0.28$ and $\bar{S}_\epsilon = 0.04$

■ Comparison of marginal posterior distributions $[\sigma_j^2 | \mathbf{z}]$

- $\mathcal{IG}(\epsilon, \epsilon)$ with $\epsilon = 10^{-3}$ for σ_i^2 vs half-Cauchy with $S = 20$ for σ_j ($j = 1, 2$)



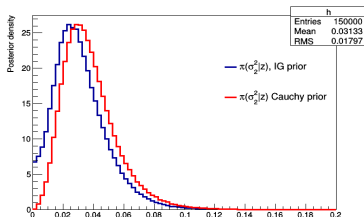
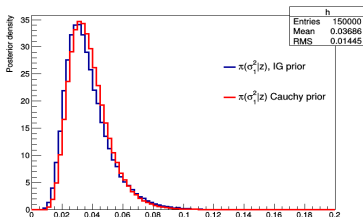


Synthetic example: $x_i \in [0.1, 1]$,
 $\alpha = (\alpha_1, \alpha_2)$, $n = 50$

- $z_i = x_i \alpha_{1,i} + \mathbf{2.3} \times x_i^3 \alpha_{2,i} + \epsilon_i$
- $\bar{S}_1 = 0.47$, $\bar{S}_1 = 0.49$ and $\bar{S}_\epsilon = 0.04$

■ Comparison of marginal posterior distributions $[\sigma_j^2 | \mathbf{z}]$

- $\mathcal{IG}(\epsilon, \epsilon)$ with $\epsilon = 10^{-3}$ for σ_j^2 vs half-Cauchy with $S = 20$ for σ_j ($i = 1, 2$)



STEPS FOR BAYESIAN CIRCE:

1. Make linear approximation at $\alpha_{nom} = \log \lambda_{nom}$ (begin at the nominal $\alpha_{nom} = \mathbf{0}_p$)
2. Sample the joint posterior distribution $m, \sigma^2 | \mathbf{z}$
 - if the MAP for m is close to α_0 , then go to the next step
 - else, go back to Step 1 (Iterative Bayesian Circe)
3. Calculate the marginal distribution of A :

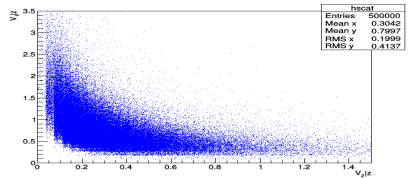
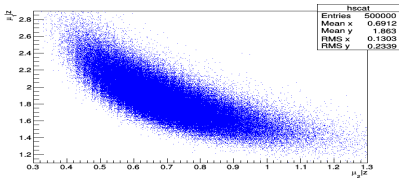
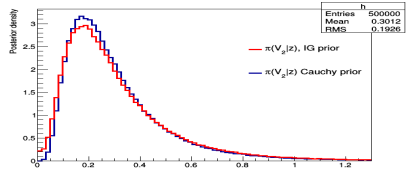
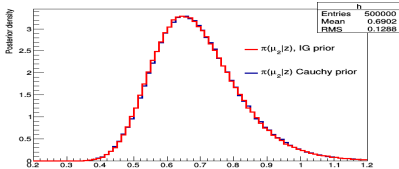
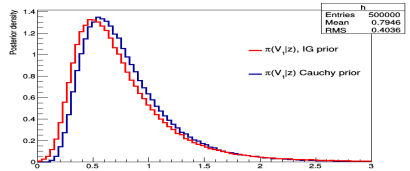
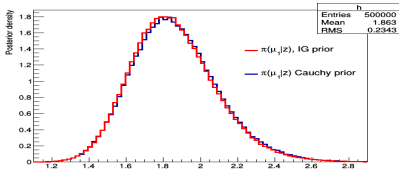
$$[A] = \int [A|m, \sigma^2][m, \sigma^2 | \mathbf{z}] dm d\sigma^2$$

4. Check the validity of the linear assumption on $\text{IF}_{0.95}(A_j)$ for $1 \leq j \leq p$,
5. Deduce $\text{IF}_{0.95}(\Lambda_j)$ for $1 \leq j \leq p$.

APPLICATION TO THE CONDENSATION MODELS

- 50 tests with two output values: condensation flow rate and temperature
 - ⇒ $n = 100$ physical measurements. Two models are considered:
 - Model Λ_1 (flow rate on the free surface in area B)
 - Model Λ_2 (flow rate due to the turbulent mixing in area C)

Results (Gibbs implemented with the ROOT library)



Non linear generalization

Instead of linearizing the computer code, we aim to tackle the exact situation where $Y(\cdot)$ is non-linear with respect to α :

- $\tilde{\alpha}_k \sim \tilde{\alpha} | \mathbf{z}, \tilde{m}_{k-1}, \tilde{\sigma}_{k-1}^2, C_{k-1}$ is no longer Gaussian \implies MH sampling \implies **MH within Gibbs algorithm**
- $Y(\cdot)$ is thermal-hydraulic system code, moderately time-consuming (several minutes per simulation)
 - \implies several weeks for a converged Gibbs sampler, along with possible failed simulations.
- Emulator is needed such as **Gaussian process (GP)**, neural networks. GP interpolates the learning simulations, which is expected for deterministic ones:
 - $\implies \tilde{\alpha}_k \sim \tilde{\alpha} | \mathbf{z}, \tilde{m}_{k-1}, \tilde{\sigma}_{k-1}^2, C_{k-1}$ is based on both mean and variance of the GP emulator.
- How to control the gap between the GP-based posterior distribution and the actual one?
 - see Barbillon (2017) in the context of mixed models from a classic point of view (SAEM algorithm instead of MH-within Gibbs).

Some questions/future works

- How to specify priors for scale parameters in hierarchical models when an objective Bayesian estimation is expected ?
 - Should we specify a prior for the scale S of the half-Cauchy prior?
 - How to measure how strong the estimation is data-dominated?
 - Studying the frequentist properties of credible intervals obtained from various priors proposed in the literature including the half-Cauchy.

- Statistical modeling to carry out in future works:
 - Estimating the experimental variances $\sigma_{\epsilon_i}^2$ when they are unknown, promoting the multidimensional version, taking into account a model Λ_k that is already known,
 - Assuming a functional multiplier coefficient $\Lambda(x)$ as a log-Gaussian process (functional Bayesian CIRCE).

- Convergence diagnostics to implement for future users in CEA (I hope so !).

- Barbillon, Pierre ; Barthélémy, C. S. A. (2017). Parameter estimation of complex mixed models based on meta-model approach. *Statistics and Computing*, 27.
- Bousquet, N. and Blazère, M. (2016). Predicted sensitivity for establishing well-posedness conditions in stochastic inversion problems. *International Conference on Sensitivity Analysis of Model Output*.
- Celeux, G., Grimaud, A., Lefèbvre, Y., and de Rocquigny, E. (2010). Identifying intrinsic variability in multivariate systems through linearized inverse methods. *Inverse Problems in Science and Engineering*, 18.
- De Crécy, A. and Bazin, P. (2001). Determination of the uncertainties of the constitutive relationship of the CATHARE 2 code. *M&C 2001*.
- Gelfand, A. and Smith, A. (1990). Sampling based approaches to calculating marginal densities. *Journal Amer. Stat. Assoc.*, 85 :398–409.
- Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1 :1–19.
- Kass, R. and Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91(435) :1343–1370.
- Spiegelhalter, D. J., Abrams, K. R., and Myles, J. P. (2004). *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*, section 5.7.3. Chichester : Wiley.