Bayesian calibration of computer models using the Openturns software

Guillaume Damblin[1,2]

[1]EDF R&D,*6 quai Watier 78401 Chatou*
[2]AgroParisTech,*16 rue Claude Bernard 75005 Paris*

Chatou, June 20th 2014

Summary

Introduction

Industrial processes deal with :

- complex physical systems
- high safety requirements

Computer models are increasingly used to complement prohibitively costly (or physically impossible) field experiments.

Goal : guarantee the safety of the installations.

Summary

Validation

Definition (from ASC, Advanced Simulation and Computing)

It is the process of confirming that the predictions of a code adequately represent measured physical phenomena.

Definition (from AIAA, American Institute of Aeronautics and Astronautics)

It is the process of determining the degree to which a model is an accurate representation of the real world from the perspective of the intented uses of the model.

Verification

Definition (from Advanced Simulation and Computing)

It is the process of confirming that a computer code correctly implements the algorithm that were intented.

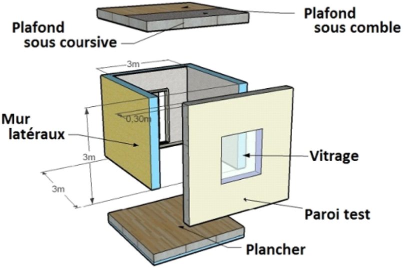
In theory, verification is a prerequisite to the validation.

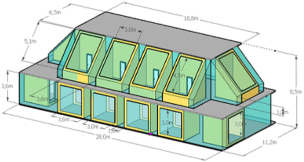
Notations

- The field is denoted by $R : x \in \mathcal{X} \subset \mathbb{R}^d \to \mathbb{R}$.
  - x : the physical variables (control variables).
- The computer code is denoted by $Y_\theta : x \in \mathcal{X} \times \mathcal{T} \subset \mathbb{R}^d \times \mathbb{R}^q \to \; Y_\theta(x) \in \mathbb{R}$:
  - x : aligned on R inputs.
  - θ : the simulator parameters, required to run Y.

An industrial case

We are interested in predicting the electric consumption of a building. Below is an experimental platform :



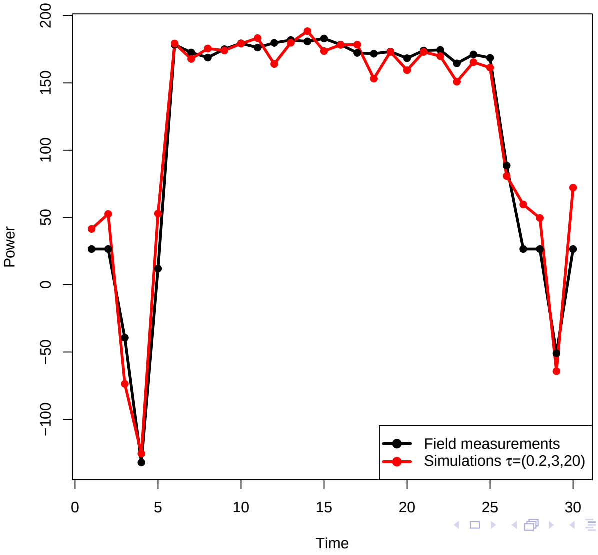


An industrial case

- $R(x(t))$ the electric power inside an experimental cell
  - t is a time step.
  - $x(t)$ includes temperature, both limits and weather conditions.

- $Y_\theta(x(t))$ a thermal computer model predicting the electric power inside the cell.
- θ includes:
  - albedo (calibration parameter).
  - convective factor (tuning parameter).
  - thermal bridge (tuning parameter).

Example : $\theta = (0.2, 3, 20)$



Calibration

Definition

It consists of searching for a set a values $\theta \in \mathcal{T}$ such that the computer model $Y(x, \theta)$ fits as closely as possible the field data.

Example : the method of least-square,

- $\bar{x} = (x_1, \cdots, x_n)$ the physical variables.
- $\bar{z} = (z_1, \cdots, z_n)$ the field measurements.
- $\theta = \underset{\tau}{\operatorname{argmin}} \sum_{i=1}^{n} \left(z_i - Y_\tau(x_i) \right)^2$

Bayesian statistics

$[\,]$ is a notation for a probability distribution.

- *A priori* distribution $[\theta]$
- Likelihood $[\bar{z}|\bar{x},\theta]$
- *A posteriori* distribution $[\theta|\bar{z},\bar{x}] \propto [\bar{z}|\theta,\bar{x}][\theta]$

Summary

Summary

Unbiaised calibration

$$z_i = R(x_i) + \epsilon_i = Y_\theta(x_i) + \epsilon_i$$

where $\epsilon_i \sim \mathcal{N}(0, \lambda^{-1})$.

Unbiaised calibration

$$z_i = R(x_i) + \epsilon_i = Y_\theta(x_i) + \epsilon_i$$

where $\epsilon_i \sim \mathcal{N}(0, \lambda^{-1})$.

Hence, the likelihood is gaussian

$$[\bar{z}|\bar{x}, \theta, \lambda] \sim \mathcal{N}\left(Y_\theta(\bar{x}), \lambda^{-1}\mathbf{I}\right)$$

Unbiaised calibration using Openturns

RandomWalkMetropolisHastings(prior,conditionnal,model,z,θ_0,proposal)

- a **prior** distribution $[\theta]$ *(Distribution)*
- likelihood model :
  - a **conditional** distribution $[\bar{z}|\bar{x},\theta]$ *(Distribution)*
  - **model** parameters *(NumericalMathFunction)* $\Longrightarrow$ Wrapper
- **z** =field data (*NumericalSample*)
- $\theta_{\mathbf{0}}$ =initialization (*NumericalPoint*)
- **proposal** =proposal distribution *(Distribution)*

A thermal industrial case

Requirements :

- **prior:** uniform $[\theta] = [\theta_1][\theta_2][\theta_3]$ *(Distribution)*
- **conditional:** gaussian distibution *(Distribution)*
- **model:** $M : (\theta, \lambda) \longrightarrow (Y_\theta(x(t)), \lambda^{-1})_{t\in[0,T]}$ *(NMF)*
- $\mathbf{z}$ =field data (*NumericalSample*)
- $\theta_{\mathbf{0}}$ =initialization (*NumericalPoint*)
- **proposal:** gaussian distribution *(Distribution)*

Material

a Dymola software $\theta \longrightarrow \Big(z(t), Y_\theta\big(x(t)\big)\Big)_{t\in[0,T]}$.

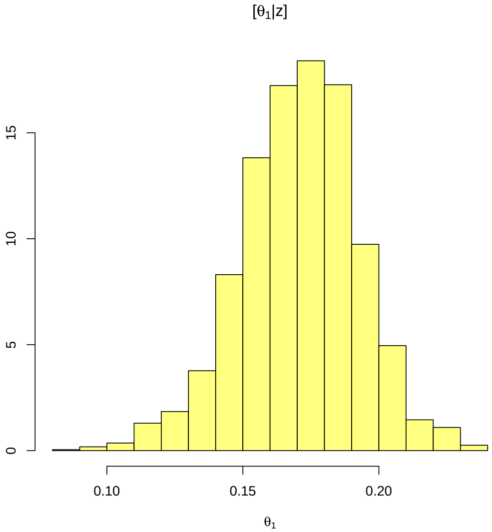
Wrapper implementation

It consists in making the NumericalMathFunction (**model**)

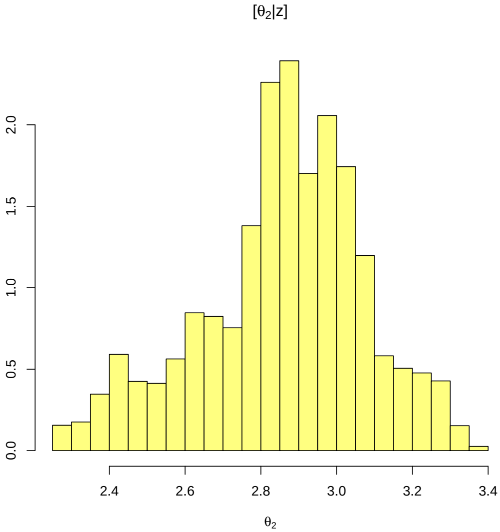
$$M : (\theta, \lambda) \longrightarrow (Y_\theta(x(t)), \lambda^{-1})_{t\in[0,T]}$$

$$\Big(Y_\theta(x(t)), \lambda^{-1}\Big)_{t\in[0,T]} = \Big(\big(Y_{x(t_1)}(\theta), \lambda^{-1}\big), \cdots, \big(Y_{x(t_n)}(\theta), \lambda^{-1}\big)\Big)$$

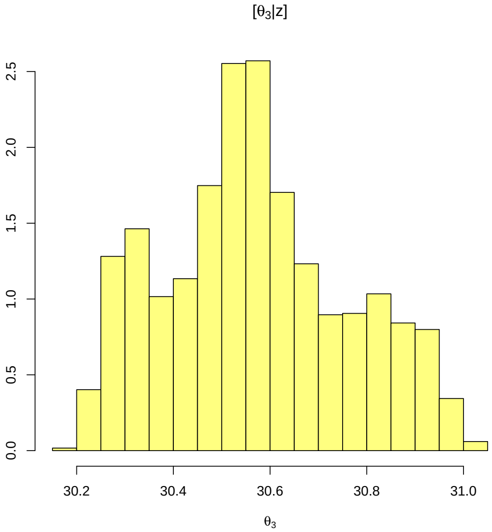
Albedo a posteriori distribution



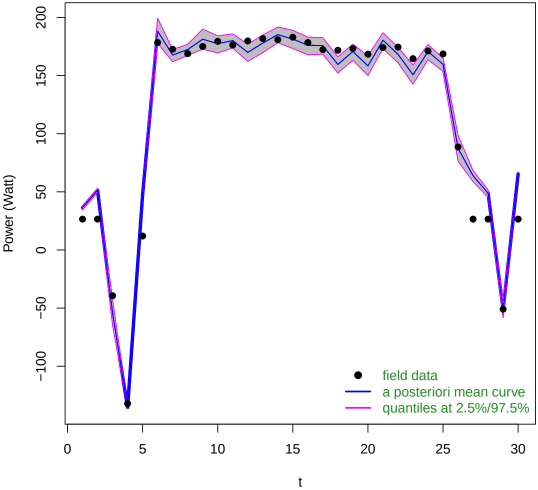
Thermal bridge a posteriori distribution



Convective factor a posteriori distribution



Prediction

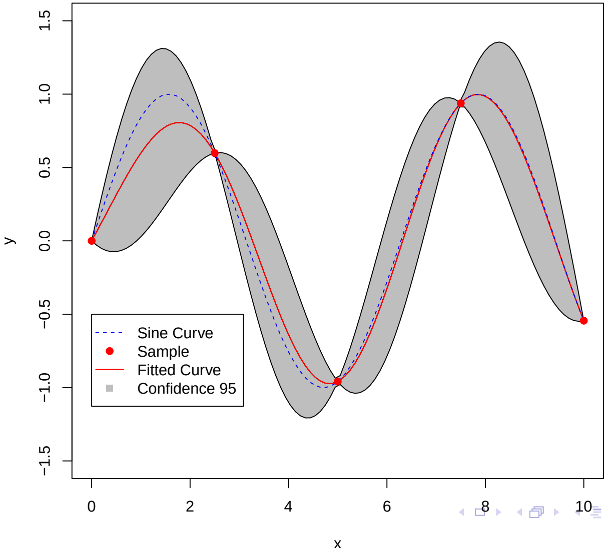

Power (Watt)
t
field data
a posteriori mean curve
quantiles at 2.5%/97.5%
200
150
100
50
0
−50
−100
0
5
10
15
20
25
30

Limited model runs

Notations:

- $D = \{(x_j, \theta_j)\}_{j=1,\cdots,m}$ a design of experiments.
- Only $y = Y(D)$ can be performed.
- A Gaussian process $\hat{Y} \sim \mathcal{N}(\mu_\phi^Y, \Sigma_\phi^Y | D)$ replaces Y in $[\bar{z}|\bar{x}, \theta]$.

Gaussian process



Summary

Model with discrepancy

$$z_i = R(x_i) + \epsilon_i = \ Y_\theta(x_i) + e(x_i) + \epsilon_i$$

where $\epsilon_i \sim \mathcal{N}(0, \lambda^{-1})$ be a white noise. Hence,

$$\bar{z} \sim \mathcal{N}(Y_\theta(\bar{x}) + e(\bar{x}), \lambda^{-1}\mathbf{I})$$

Critical issue : $e(x)$ **is unknown !**

Model reformulation

Assuming $e(x) \approx e(x + \delta x)$,

$$e(x) \sim \mathcal{PG}(0, \Sigma^{e}_{\phi^e}(x))$$

Then,

$$\bar{z} \sim \mathcal{N}(Y_\theta(\bar{x}), \Sigma^{e}_{\phi^e}(\bar{x}) + \lambda^{-1}\mathbf{I_n})$$

- $[\bar{z}|\bar{x}, \theta, \lambda, \phi_e]$ is Gaussian.
- Identifiability holds.

Model assumptions

Let us consider the following hypothesis :

1. $e(x) \sim \mathcal{PG}(0, \Sigma^e_{\phi_e}(x))$
2. Y is cheap
3. $\epsilon_i = 0$ (no measure error)

with the Gaussian kernel :

$$\Sigma^e_{\gamma,\beta}(x, x') = \frac{1}{\gamma} \exp \sum_{k=1}^{d} -\beta_k |x_k - x'_k|^2$$

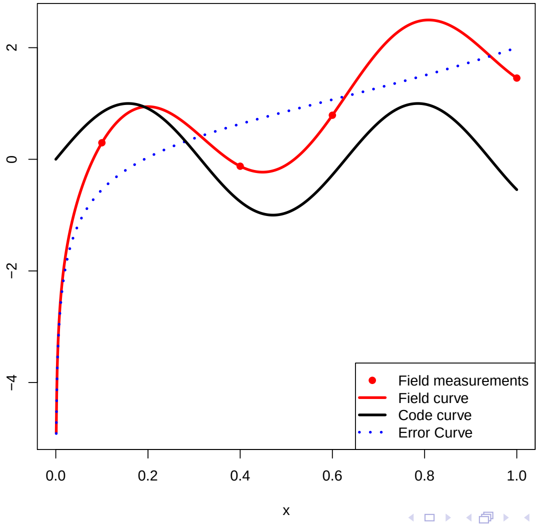
Pedagogic Example

- $\theta = 10$.
- $Y_\theta(x) = \sin(x\theta)$
- $e(x) = \exp(x)\log(x) + 2$.

Simulated field data :

- $x_i \in [0, 1]$.
- $z_i = R(x_i) = \sin(10\,x_i) + \exp(x_i)\log(x_i) + 2$.

Illustration $\bar{x} = \{0.1, 0.4, 0.6, 1\}$



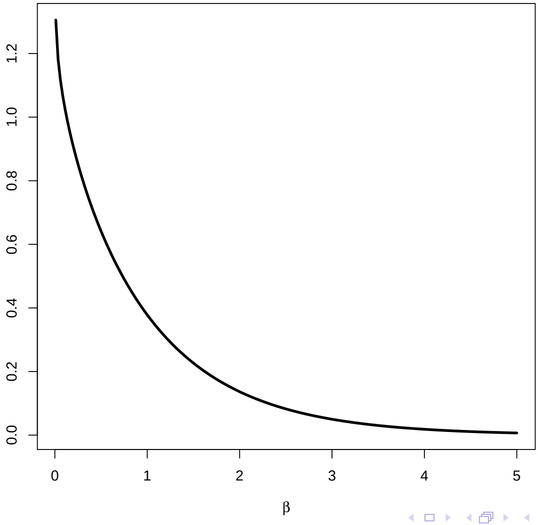
Bayesian inference

A priori hypothesis :

- $[\gamma] \sim \mathcal{G}(2,2)$
- $[\theta] \sim \mathcal{U}[5,15]$.
- $[\beta] \propto (1 - \exp -\beta)^{-0.06} \exp -\beta \ \mathbf{1}_{\beta>0}$

This β prior gives a slight tendency.

β prior

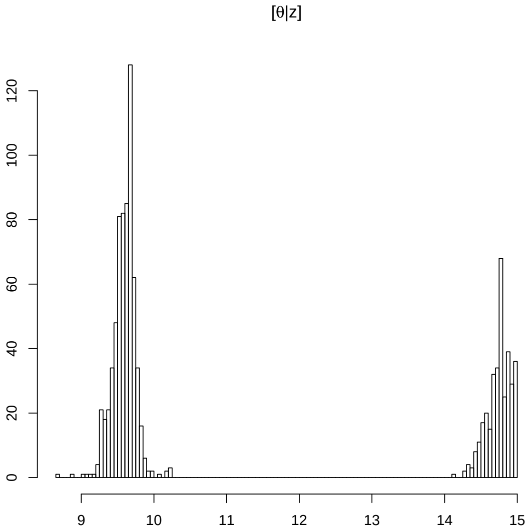


Gibbs algorithm

Iterate :

1. MH to simulate $[\theta_k|z, \beta_{k-1}, \gamma_{k-1}]$.
2. MH to simulate $[\beta_k|z, \gamma_{k-1}, \theta_k]$.
3. MH to simulate $[\gamma_k|z, \beta_k, \theta_k]$.

A posteriori distribution $[\theta|z]$



Calibrated prediction

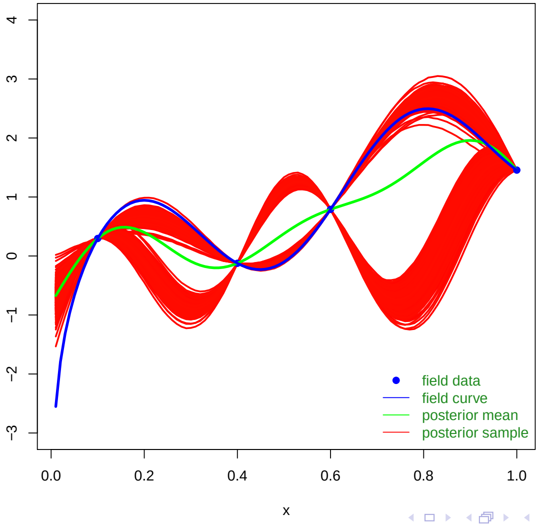
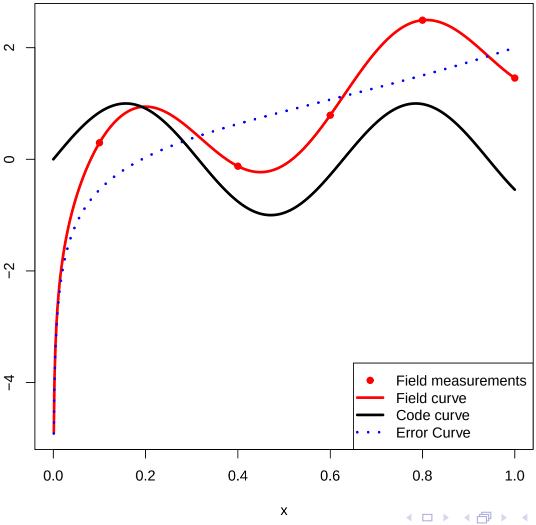
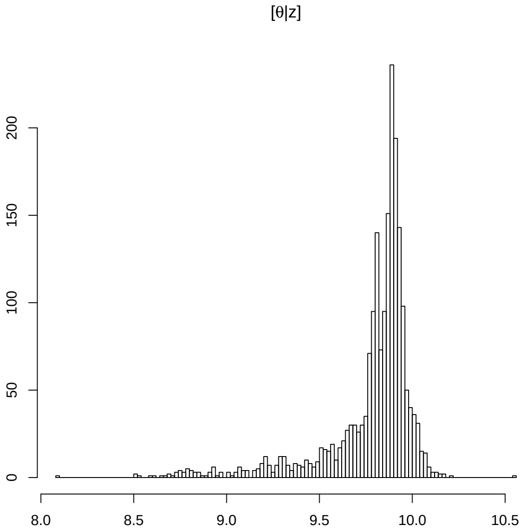


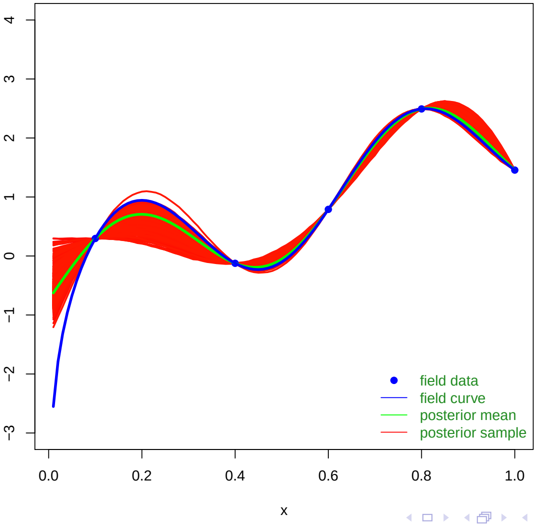
Illustration $\bar{x} = \{0.1, 0.4, 0.6, 0.8, 1\}$



A posteriori distribution $[\theta|z]$



Calibrated prediction



Summary

Validation Goals

1. the quality of Y_θ
   - interested in the meaning of θ
   - interested in the closeness of Y_θ to R (model error in question).
2. the prediction of R (a mean predicted value with a tolerance bound)
   - pure model prediction if $e = 0$.
   - bias-corrected prediction if $e \neq 0$.