Gaussian process metamodel and wavelet decomposition for sensitivity analysis

Application to functional output of computer code

Amandine Marrel
IFP Rueil-Malmaison

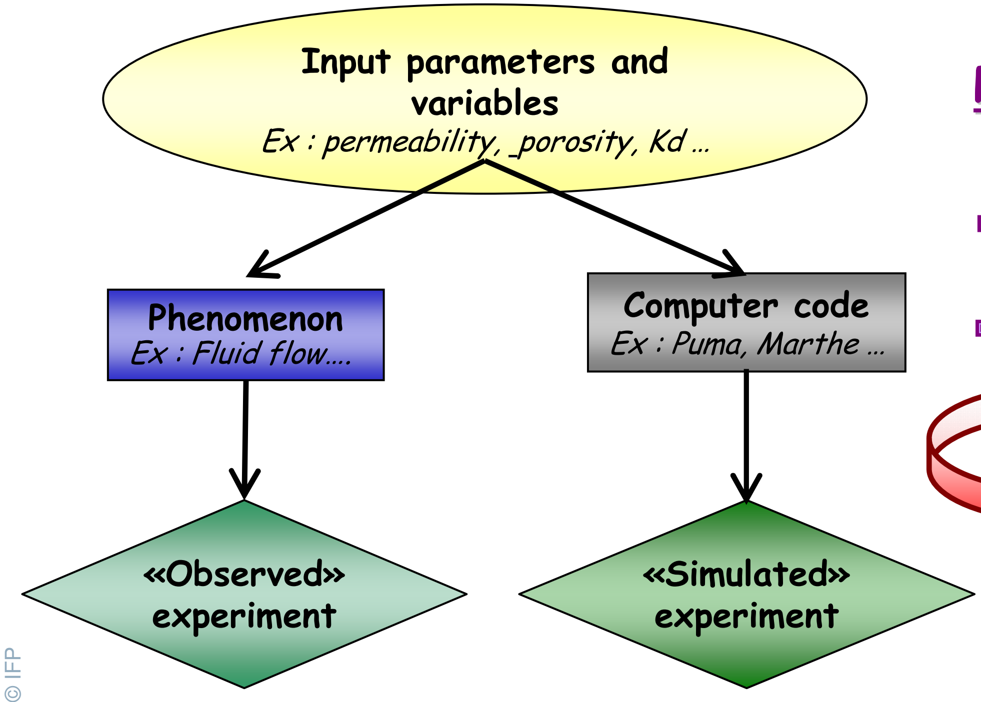
Framework and goals (1)

Modeling process :

- Real phenomena represented by **deterministic equations**
  Ex : fluid flow described by Darcy equations…

- **Input parameters and associated uncertainties** $X = [X_1, \ldots, X_d]$
  Ex : model parameters (estimation or literature), physical variables (in situ or laboratory)…

- Implementation : **computer code** $Y_{code}(X)$
  Ex : oil reservoir simulation, migration of pollutant in saturated porous media…



Purposes :

Identify the most influent parameters

⇒ Improvement of the knowledge of these parameters

⇒ Reduction of uncertainties on prediction and optimization

SENSITIVITY ANALYSIS

Framework and goals (2)

Extension to functional outputs of computer code :

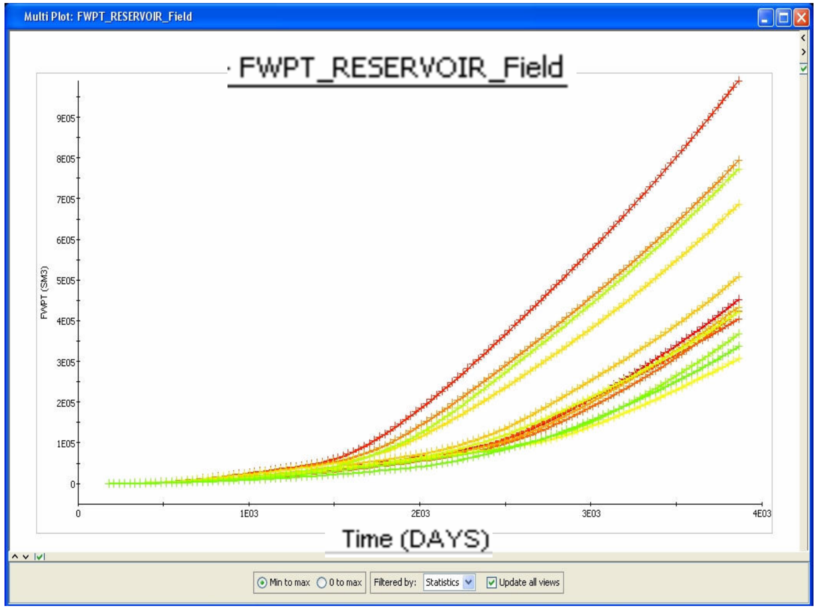
$$X = [X_1, \ldots, X_d] \Rightarrow Y_{code}(z, X),\ z \in D$$

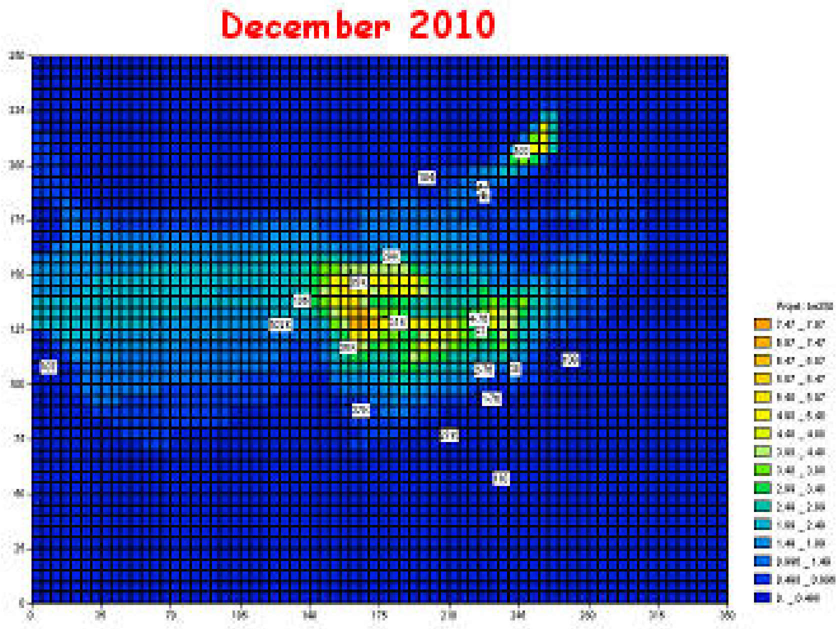
Temporal output

Ex : Oil production history...

Spatial Output

Ex : Concentration of component...





SENSITIVITY ANALYSIS OF A FUNCTIONAL OUTPUT

Content

Functional metamodel : wavelet decomposition + *Gp* modeling

- Gaussian process metamodel (Gp metamodel)
- Building methodology of Gp metamodel
- Extension to a functional output

Sensitivity analysis of a functional output

Application to a hydrogeological transfer code

- Data presentation
- Approximation by the functional Gp metamodel
- Implementation of spatial sensitivity indices
- Interpretation

Conclusion

Gp metamodel (1)

Problem :

- Computer code time expensive (~ 30mn / evaluation)
- Complex model
- High number of inputs (> 10)

→ **Sensitivity analysis with a direct use of computer code is difficult**

Solution : Replace computer code by a **metamodel**

- Approximation of computer code with a negligible computer time
- Ex : Polynomials, splines, neural networks, regression trees…

Choice : conditional Gaussian process (Gp)

Development of an efficient methodology for modeling functional complex computer code with Gp metamodel

Gp metamodel (2)

Definition :

Gaussian process defined on $R^d \times \Omega$

$Y(x,\omega) = F(x) + Z(x,\omega)$

Regression — stochastic part

Stochastic process Z with :

$E_\Omega[Z(x)] = 0$

$Cov_\Omega(Z(x), Z(u)) = \sigma^2 R(x, u)$

where σ^2 is the variance

and R the correlation function

$Z \sim N(0, \sigma^2 R)$

Parametric choices :

- F : polynomial of degree 1 $\mathbf{F(x)} = \beta_0 + \sum_{i=1}^{d} \beta_i x_i$
- R : stationary process with generalized exponential covariance

$$R(x, u) = R(x - u) = \exp\left(-\sum_{i=1}^{d} \theta_i |x_i - u_i|^{p_i}\right) + \varepsilon^2 1_{x=u}$$

Gp metamodel (3)

Joint and conditional distributions :

- Learning sample (LS) of *n* simulations : (X_{LS}, Y_{LS})

$$X_{LS} = \left[x^{(1)}, \dots, x^{(n)}\right],\; F_{LS} = F(X_{LS}),\; R_{LS} = \left(R\left(x^{(i)}, x^{(k)}\right)\right)_{i,k}$$

- Joint distribution of LS : $Y_{LS} \sim N\,(\beta F_{LS}, \sigma^2 R_{LS})$

- Conditional Gp metamodel :

$$\Rightarrow \quad Y(x,\omega)_{|X_{LS},Y_{LS}} \sim Gp$$

$$E_\Omega\left[Y(x,\omega)_{|X_{LS},Y_{LS}}\right] = \beta F(x) + r(x) R_{LS}^{-1}\left[Y_{LS} - \beta F_{LS}\right]$$

$$\textbf{with} \quad r(x) = \left[R\left(x^{(1)}, x\right), \dots, R\left(x^{(n)}, x\right)\right]$$

$$Cov_\Omega\left(Y(u,\omega)_{|X_{LS},Y_{LS}}, Y(v,\omega)_{|X_{LS},Y_{LS}}\right) = \sigma^2\left(R(u,v) + {}^t r(u) R_{LS}^{-1} r(v)\right)$$

$$\text{Predictor notation : } \hat{Y}(x) = E_\Omega\left[Y(x,\omega)_{|X_{LS},Y_{LS}}\right]$$

Gp metamodel (4)

Building methodology of Gp metamodel

- Parameter estimation by **maximum likelihood** : $(\beta, \theta, p, \sigma, \varepsilon)$
- **Sequential algorithm :**

  Stochastic algorithm followed by a descent step (Hookes & Jeeves)

  Sequential insertion of inputs in regression and covariance functions

- **Double selection of inputs**

  - Minimization of AICC criterion ⇨ regression

  - Maximization of predictivity coeff. Q_2 ⇨ covariance

- **Final validation by Q_2** : $Q_2(Y, \hat{Y}) = 1 - \dfrac{\sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{N} (\overline{Y} - Y_i)^2}$

All the details of the building algorithm in Marrel et al.[4]

Extension to a spatial output ?

Which functional metamodel ?

Use of full discretization of the function

- Building of a metamodel and the associated sensitivity analysis for each point of discretization

⇨ Possible but computer time expensive with metamodel like Gp

⇨ **Necessary reduction of data or isolation of main information**

Replace the function with few parameters of interest (final value, max, …)

- Reduced exploitation, strongly related to the initial problematic

Kriging / Cokriging (Santner et al., 2003, Fang et al., 2006)

- simultaneous treatment of a large dimension data

Spectral decomposition of covariance (Karhunen-Loeve)

Decomposition on a functional basis (Fourier, wavelets,…)

Extension of Gp metamodel to a spatial output

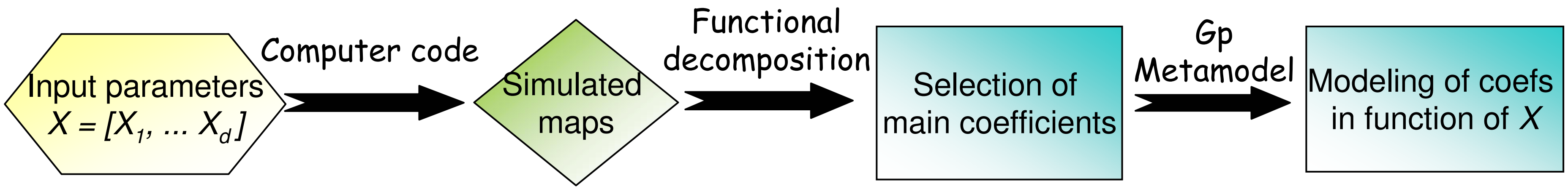
Wavelet decomposition + Gp modeling of coefficients

Step 1 : Spatial decomposition on a wavelet basis

⇨ Selection of the *k* main coefficients of the decomposition

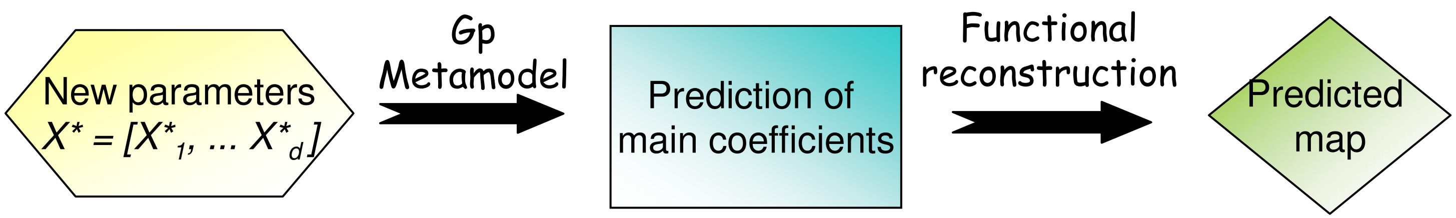
$$Y_k(\mathbf{X},z) = \mu(z) + \sum_{j=1}^{k} \alpha_j(\mathbf{X})\phi_j(z) \text{ avec } \alpha_j(\mathbf{X}) = \int_D (Y(\mathbf{X},z) - \mu(z))\phi_j(z)dz$$

Step 2 : Gp modeling of the *k* main coeffs in function of the inputs X



Step 3 : Prediction for any inputs

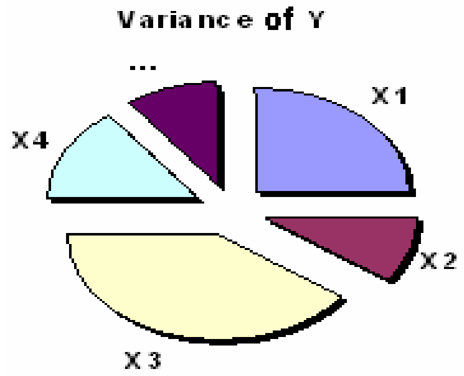
Sensitivity analysis, uncertainties propagation, …



Innovation Energy Environment

Sensitivity analysis : Sobol indices

<u>Global sensitivity analysis</u>



➡ Based on variance decomposition :
Part of the variance of each input in the output variance

➡ Input/output relation neither linear, nor monotonous :
Computation of Sobol indices

<u>Definitions for a spatial output</u> $Y(X_1, \ldots, X_d, z)$

- Main Effects :

$$a(X_i, z) = \oint Y(x_1, \ldots, x_{i-1}, X_i, x_{i+1}, \ldots, x_d, z) \prod_{j=1, j \neq i}^{d} dx_j = E\left[Y(X_1, \ldots, X_d, z) / X_i\right]$$

- Sobol indices :

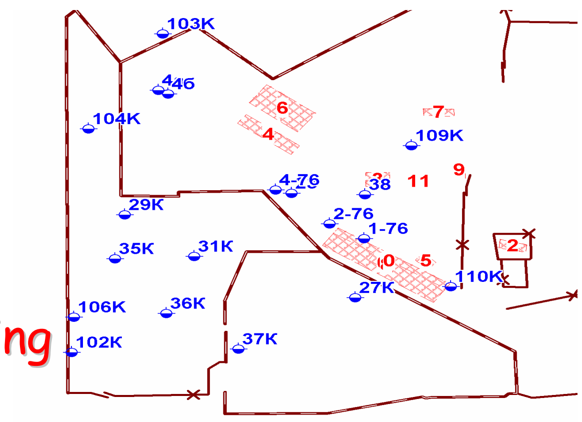
$$S_i(z) = \frac{Var_{X_i}[E(Y(X_1, \ldots, X_d, z) / X_i)]}{Var(Y(z))} = \frac{Var_{X_i}[a(X_i, z)]}{Var(Y(z))}$$

Details of sensitivity analysis with Gp in Marrel et al.[3]

ifp Innovation Energy Environment

Illustration on an hydrogeological transport code (1)

- **Kurchatov site (Moscow suburbs):**
  - Temporary storage of radioactive waste from 1943 to 1974
  - Study of the site in 1990 : 20 piezometers
  - Contamination of upper layer by ^{90}Sr



- **Purpose :**

  Estimation of the sort-term evolution of ^{90}Sr transport to help in decision rehabilitation making

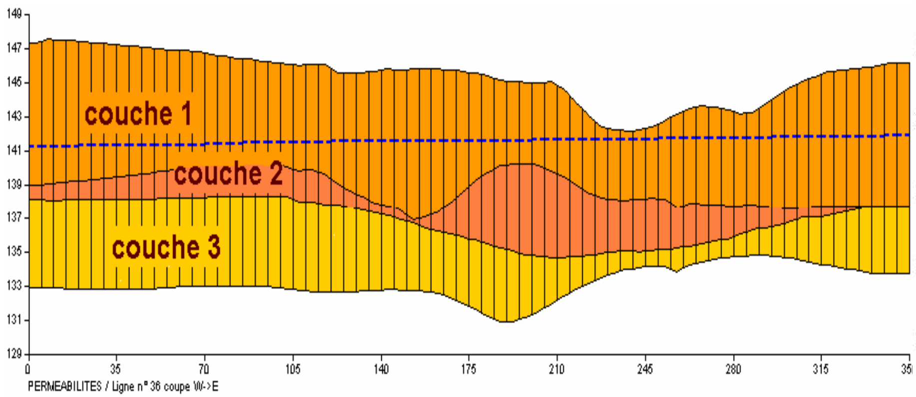
- **Modeling :**
  - Development of ^{90}Sr generalized transport scenario on the site from 2002 and 2010, with MARTHE software
  - Identification and partial characterization of transport parameters

→ **Determine the most influent parameters:**

Response surface building + Sensitivity analysis of parameters

Illustration on an hydrogeological transport code (2)

- Geological modeling & simplification

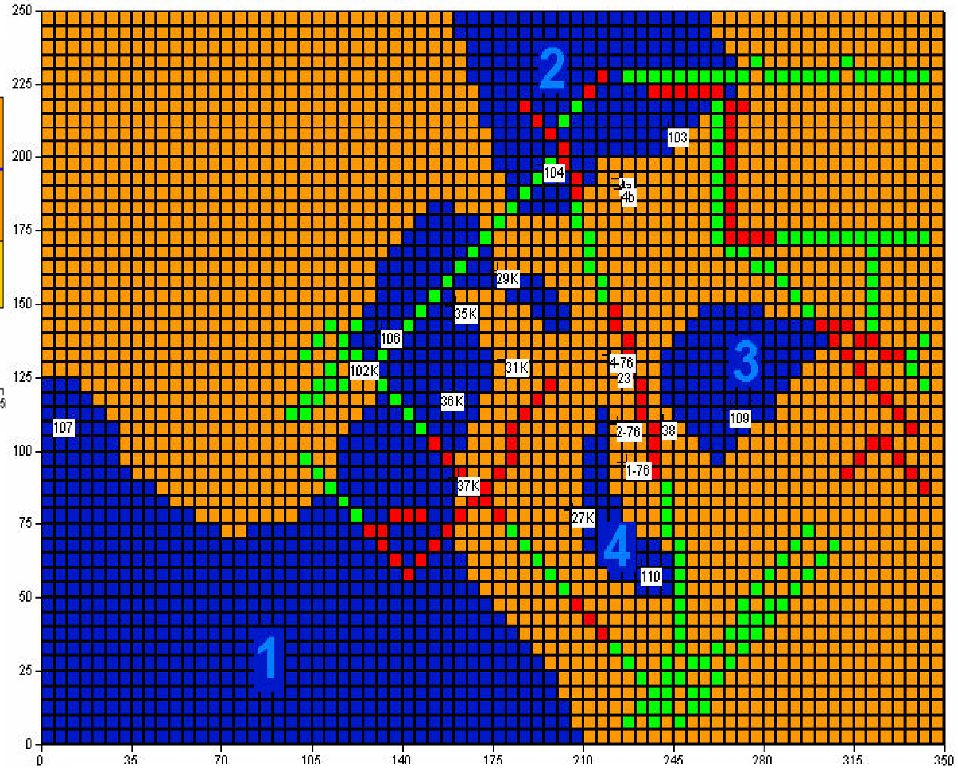


Identification of 20 parameters:

Permeability, porosity, sorption coefficient Kd, infiltration intensities…

&

Associated uncertainties



- areas without the layer 2
- areas without the layer 2
- moderated infiltrations around the pipes
- strong infiltrations around the pipes
- p4a piezometers

Illustration on an hydrogeological transport code (3)

- Mathematical and numerical modeling : Marthe data
  - 20 input parameters :permeability, Kd, porosity, infiltration intensities, …
  - Functional output (discretization):

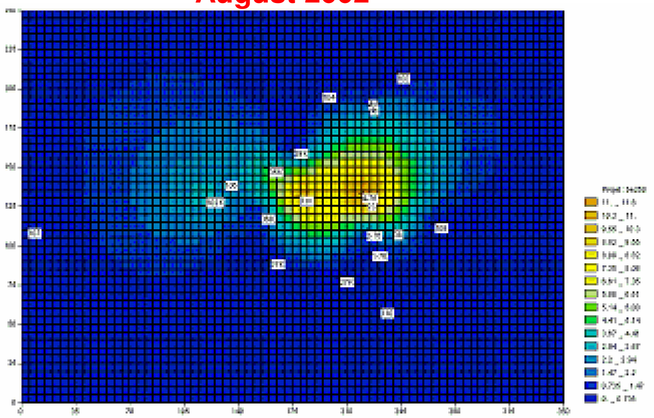
    Map of concentration in 2010 computed by MARTHE software

    Spatial discretization : 64x64 = 4096 sites

Simulation of the 20 inputs (LHS)

300 LHS simulations

Initial concentration map
August 2002



MARTHE Code

Final concentration map
December 2011

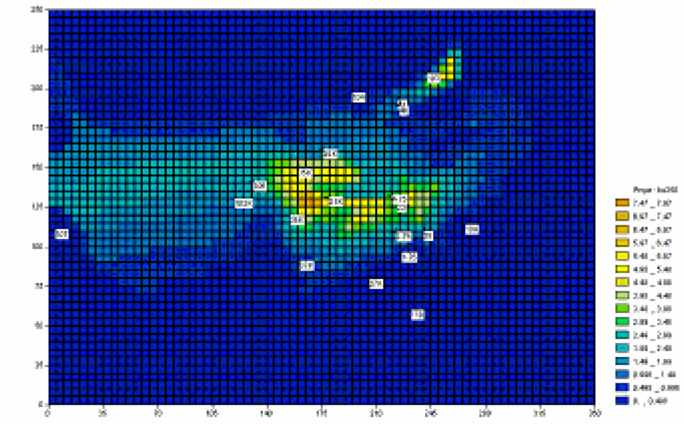


Illustration on an hydrogeological transport code (4)

Few examples of concentration maps (few outputs of the code)

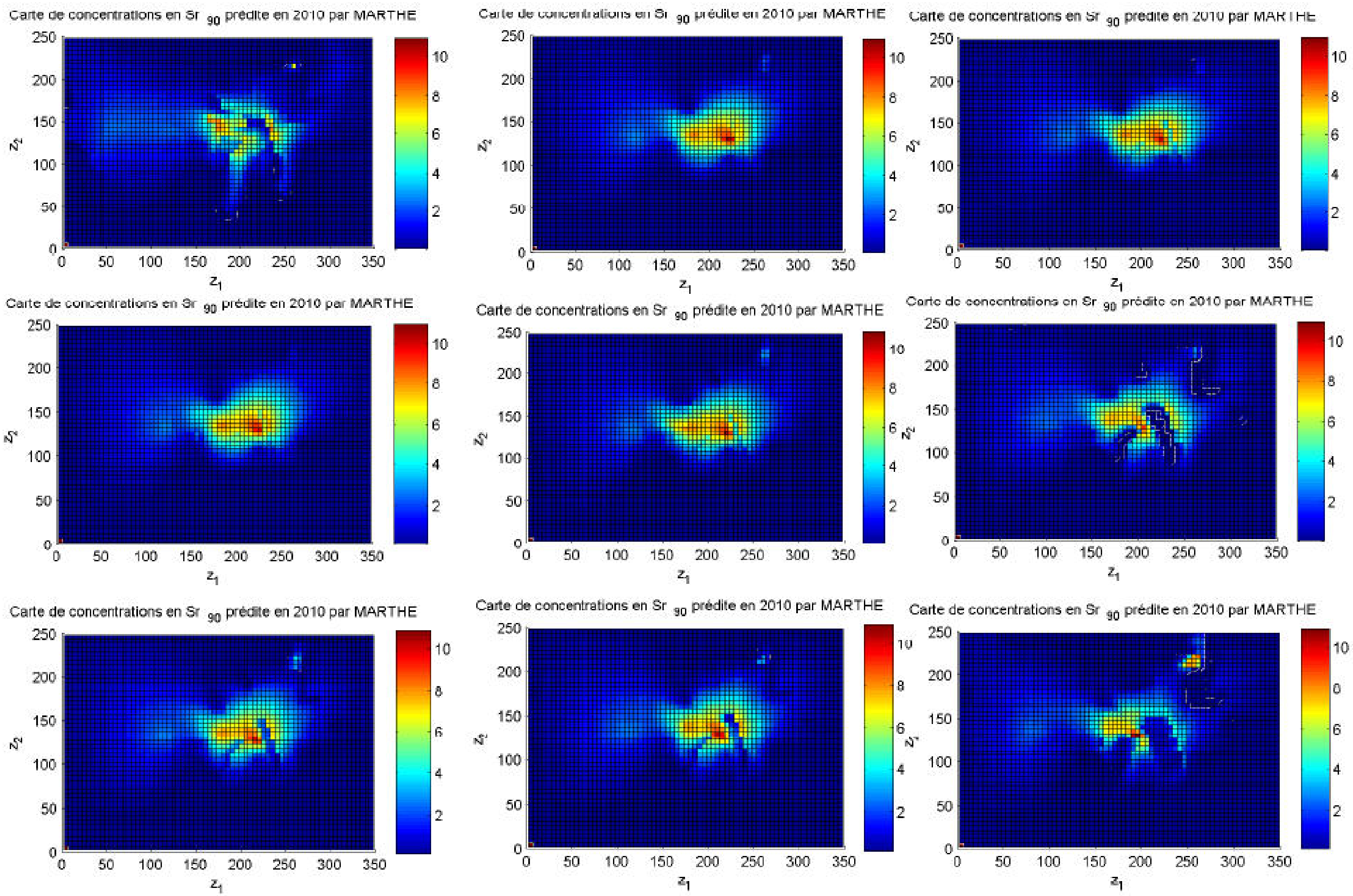


Illustration on an hydrogeological transport code (5)

- **Step 1 : Decomposition of each map on a wavelet basis**
  - Centering of the maps (empirical mean)
  - Decomposition on a wavelet basis (Daubechies wavelets)
  - Sort of and selection of the 100 main coefficients (L_2 norm)

- **Step 2 : Gp Modeling of coefficients in function of inputs X**
  - Modeling by Gp metamodel of the 100 main coeffs (accuracy controled by Q_2 estimated by cross validation)

- Step 3 : Prediction of a map for any set of inputs x^*

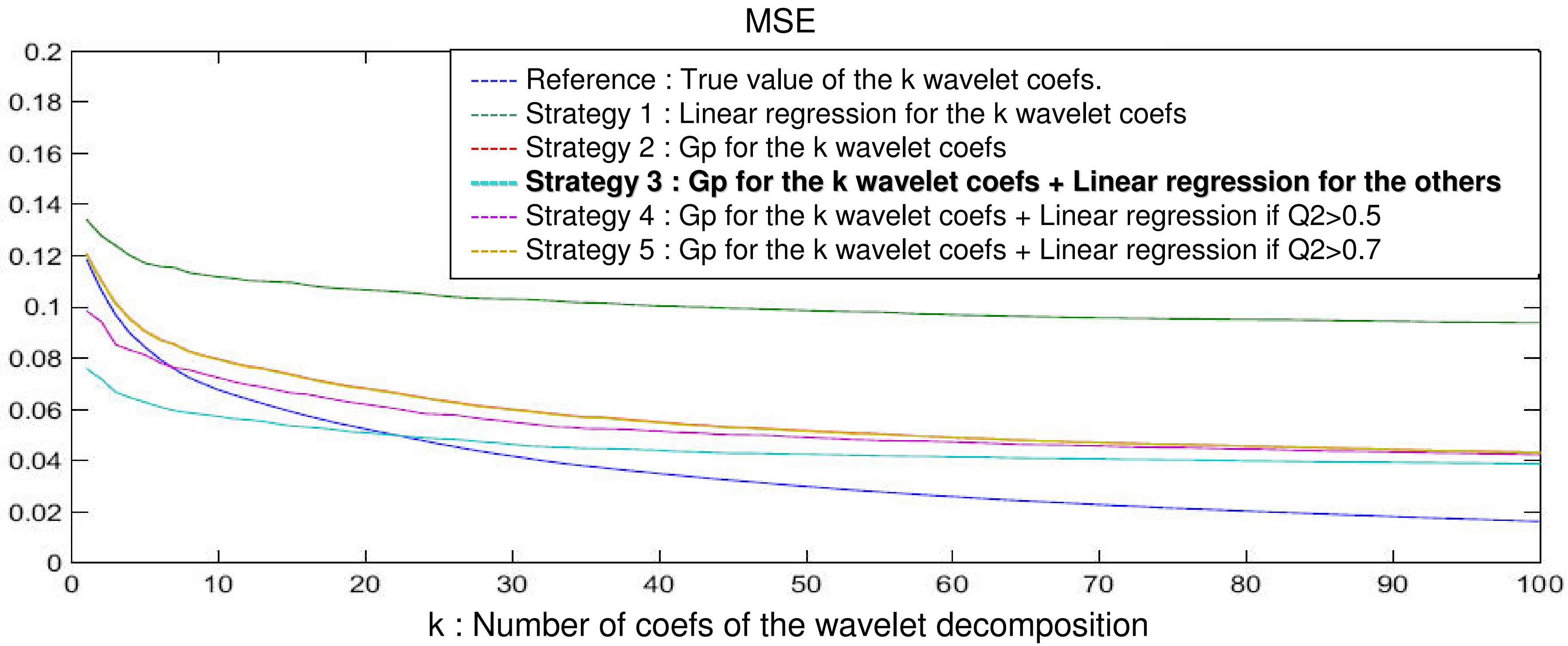
x^* => prediction of the coeffs => reconstitution of the concentration map

Sensitivity analysis :

Spatial maps of Sobol indices

Illustration on an hydrogeological transport code (6)

Different strategies for the modeling of the coefficients of the wavelet decomposition

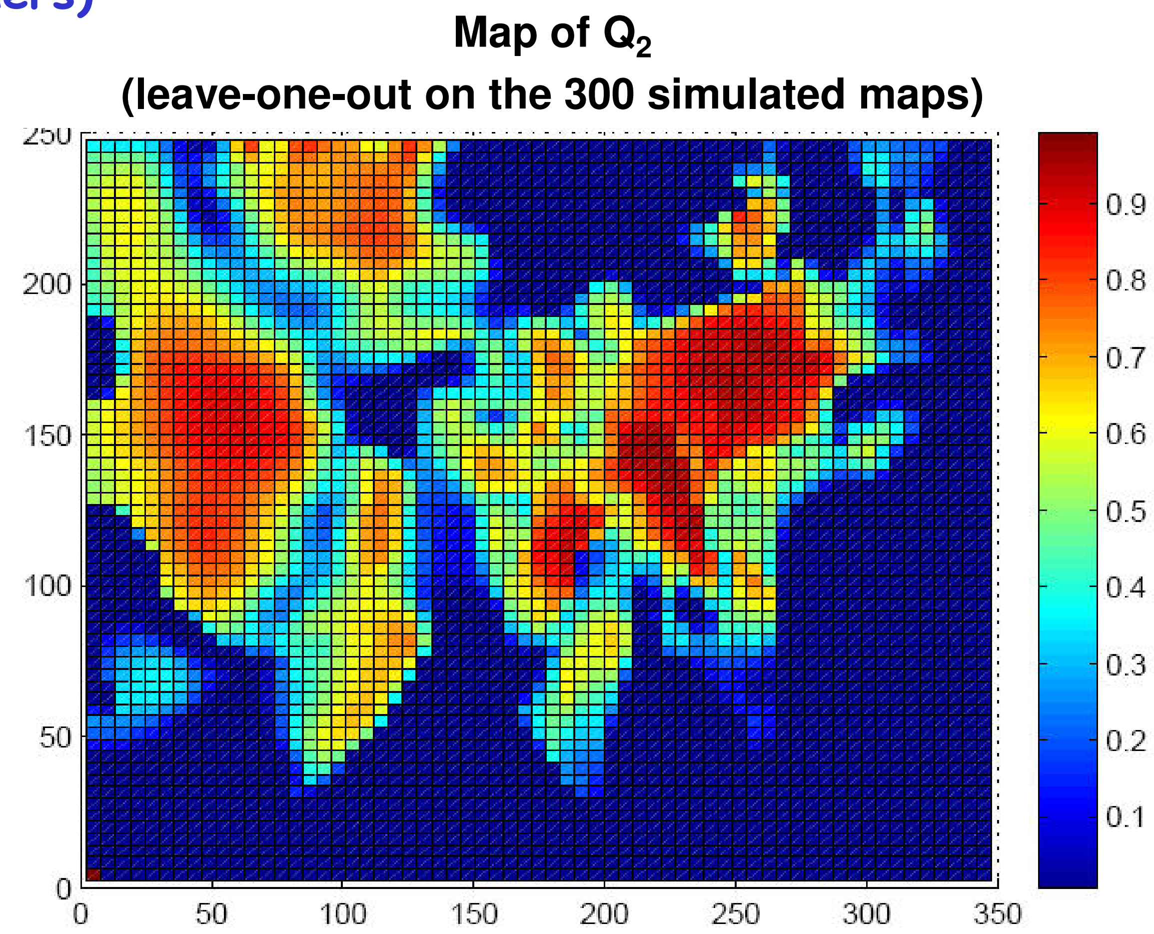


Remark : linear regression is combined with a selection procedure based on criterion like AIC ,BIC or Cp Mallows.

Illustration on an hydrogeological transport code (7)

<u>Predictivity of functional metamodel</u>

(Gp modeling of the 100 selected coefs + lin. reg. & AIC selection for the others)



Innovation Energy Environment

Illustration on an hydrogeological transport code (8)

- **Sensitivity analysis :** Main maps of sensitivity indices

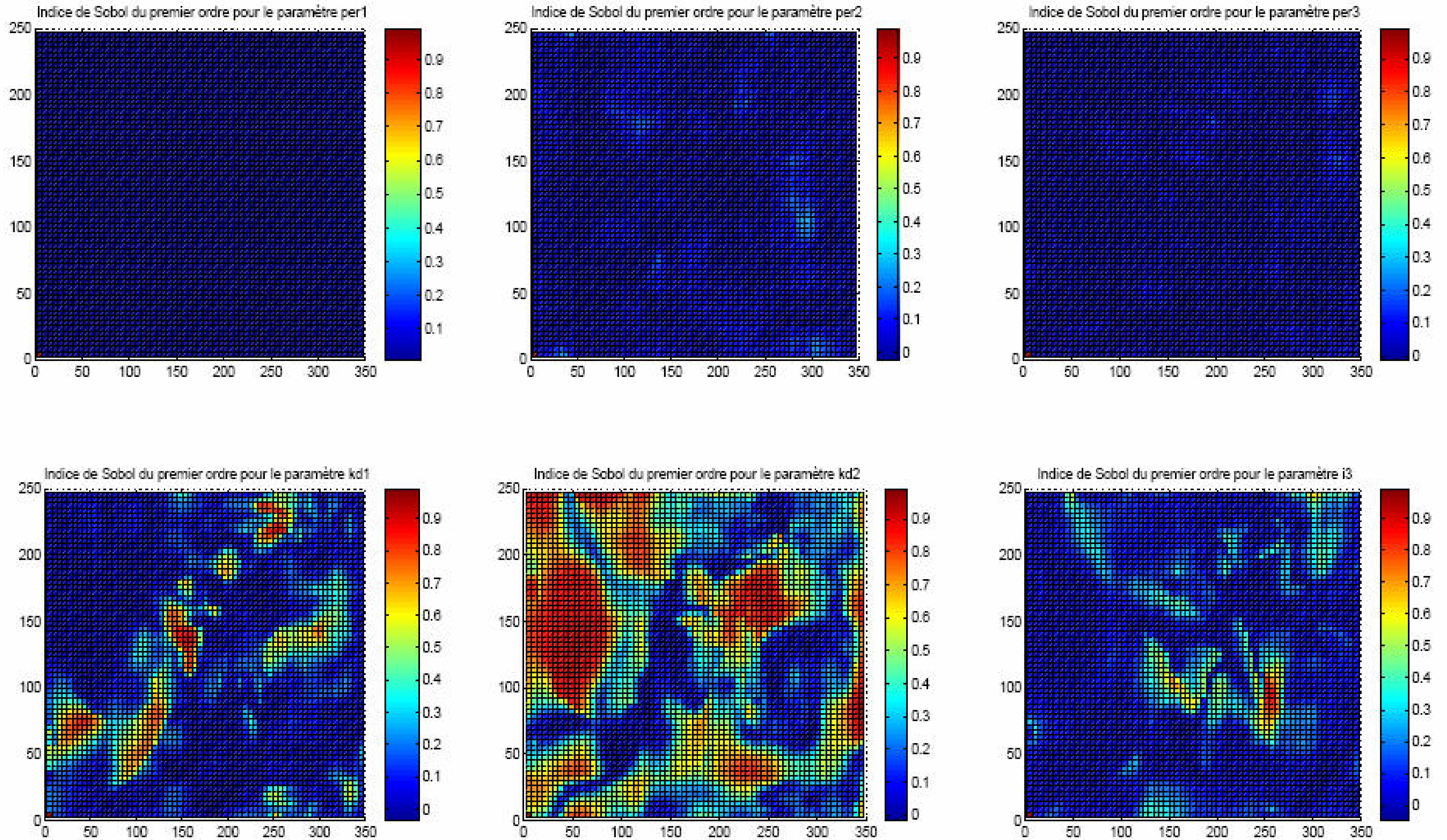
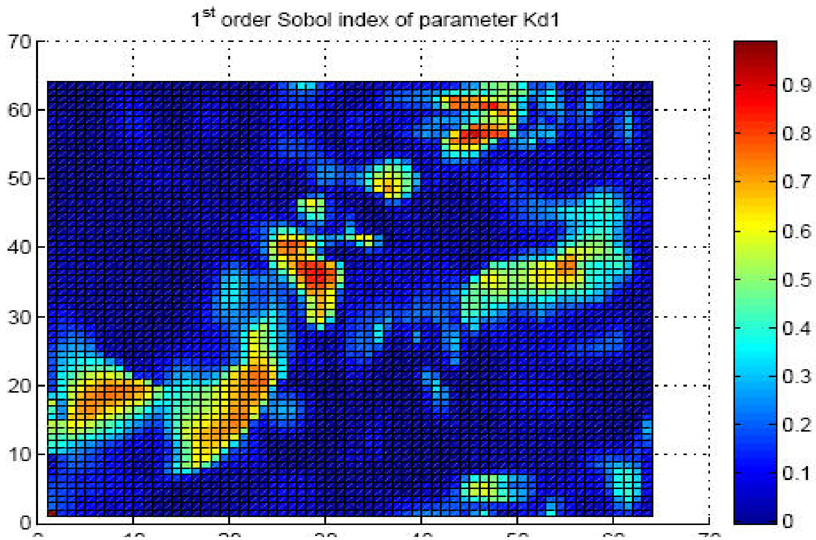
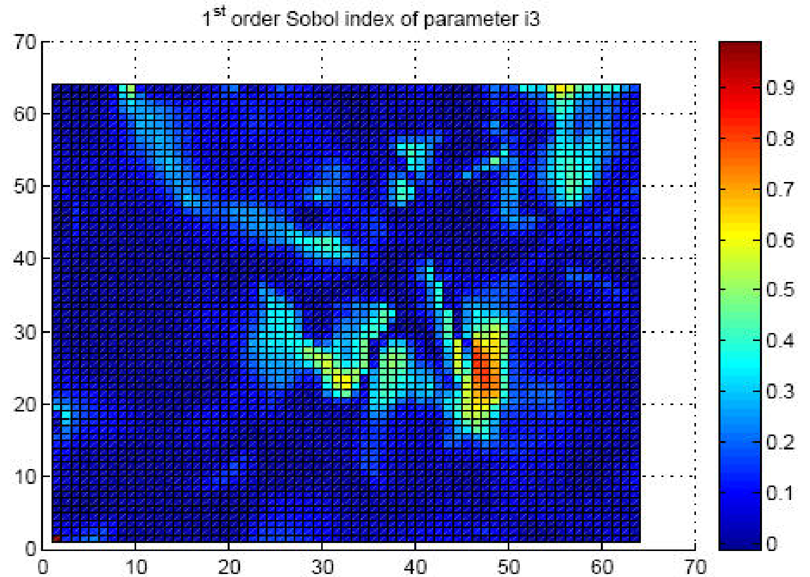
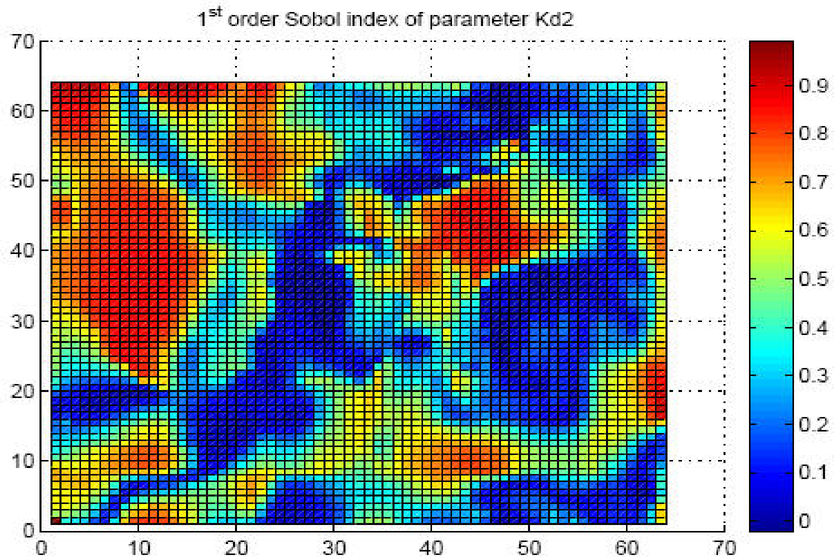


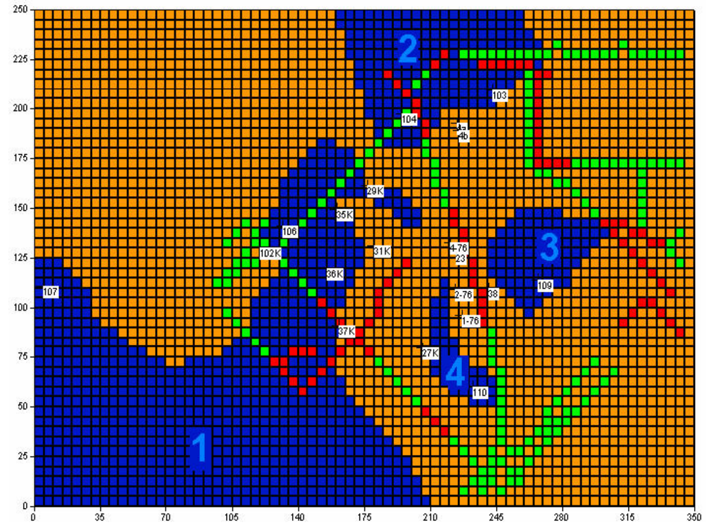
Illustration on an hydrogeological transport code (9)

- **Sensitivity analysis :** ⇨ Sensitivity maps for the 3 most influent inputs









Innovation
Energy
Environment

Conclusion

Functional metamodel : Gp + wavelet decomposition

- **An efficient strategy to reduce the large dimension of output**
  ⇨application to temporal or spatial output (environmental problems)

- **Maps of sensitivity indices:**
  ⇨influence of each input on the output of computer code.

- **Global but also local information:** for each spatial point, the most influent inputs are known.

- **Re-characterization strategy available :**
  ⇨ The most influent inputs (Kd, infiltration coeff., ..) are identified and their uncertainties must be reduced to decrease the uncertainty of the code prediction.

- **Confidence intervals for the code prediction of the output**
  ⇨ Help for the final decision.
  Ex : decontamination perimeter, optimisation of oil production

Bibliography

[1] Fang K-T, Li R., Sudjianto A.: Design and modelling for computer experiments, *Chapman & Hall/CRC* (2006).

[2] Higdon D., Gattiker J., Williams B., Rightley M.: Computer model calibration using high dimensional output, *Journal of the American Statistical Association*, **103**, 570-583 (2008).

[3] Marrel A, Iooss B, Laurent B, Roustant O: Calculations of Sobol indices for the Gaussian process metamodel, *Reliability Engineering and System Safety*, **94**, 742-751 (2009).

[4] Marrel A, Iooss B, Van Dorpe F, Volkova E: An efficient methodology for modeling complex computer codes with Gaussian processes, *Computational Statistics and Data Analysis*, **52**, 4731-4744 (2008).

[5] Marrel A, Mise en oeuvre et utilisation du métamodèle processus gaussien pour l'analyse de sensibilité de modèles numériques, *Thèse de l'INSA de Toulouse* (2008).

[6] Oakley JE, O'Hagan A: Probabilistic sensitivity analysis of complex models: A Bayesian approach, *Journal of the Royal Statistical Society, Series B*, **66**, 751-769 (2004).

[7] Sacks J, Welch WJ, Mitchell TJ, Wynn HP: Design and analysis of computer experiments, *Statistical Science*, **4**, 409-435 (1989).

[8] Saltelli A, Chan K, Scott EM, editors: *Sensitivity Analysis*, Wiley, 2000.

[9] Sobol I.M., Sensitivity estimates for non linear mathematical models, *Mathematical Modelling and Computational Experiements*, 1:407 414, 1993.

Innovation Energy Environment