Méthodes de Monte-Carlo multi-niveaux pour la quantification d'incertitudes et l'assimilation de données - Application à la modélisation fluviale

Matthias De Lozzo[1,2], Paul Mycek[1], Sophie Ricci[1,2],
Mélanie Rochoux[1,2], Pamphile Roy[1] et Nicole Goutal[3]

[1] *CERFACS, 42 avenue Gaspard Coriolis, 31057 Toulouse Cedex 1, France*
(delozzo,mycek,ricci,rochoux,roy@cerfacs.fr)
[2] *CECI, CNRS - CERFACS 42 avenue Gaspard Coriolis, 31057 Toulouse Cedex 1, France*
[3] *Laboratoire d'Hydraulique Saint-Venant, EDF R&D, 6 Quai Watier, 78401 Chatou, France*
(nicole.goutal@edf.fr)

Résumé. Les simulateurs physico-numériques sont des outils usuels en modélisation hydraulique pour estimer le niveau et le débit d'un fleuve. La complexité de la physique qu'ils implémentent et les incertitudes portées par leurs entrées conduisent à mener des études de quantification d'incertitudes sur leurs sorties en estimant des moyennes, variances, distributions ainsi que des indices de sensibilité. Pour une précision donnée ε, l'estimation par des approches classiques de type Monte-Carlo a néanmoins le défaut de requérir $\mathcal{O}(\varepsilon^{-2})$ évaluations du simulateur dont le coût est souvent élevé. Pour pallier cette contrainte, des méthodes récentes d'échantillonnage de type Monte-Carlo multi-niveaux ont été développées et permettent de réduire ce coût calculatoire en faisant appel à des versions dégradées et moins coûteuses du simulateur initial. Initialement conçues pour l'estimation d'espérances, ces approches ont récemment été étendues à l'estimation d'autres moments statistiques. Nous avons appliqué avec succès les outils existants dans la littérature à un modèle simplifié d'hydraulique 1D avant d'étendre leur utilisation à un modèle fluvial de la Garonne décrivant les équations complètes de Saint-Venant 1D. Les perspectives de cette étude sont une analyse de sensibilité et le calcul de matrices de covariance d'erreurs intervenant dans l'algorithme d'assimilation de données du filtre de Kalman d'ensemble.

Mots-clés. Monte-Carlo multi-niveaux, multifidélité, quantification d'incertitudes, analyse de sensibilité, assimilation de données, modélisation fluviale.

Abstract. Physico-numerical simulators are usually considered in hydraulic modelling for the estimation of the water depth and flow rate. The physical complexity and the uncertainties associated to the inputs lead to the setting up of studies in uncertainty quantification concerning the model output, by means of various estimations: means, variances, distributions and sensitivity indices. However, for a given accuracy ε, this estimation by classical Monte-Carlo methods has the drawback of requiring $\mathcal{O}(\varepsilon^{-2})$ evaluations of the simulator whose computational cost is often high. To address this problem, new methods of Monte-Carlo sampling based on multi-level computer codes have been developed and can reduce significantly the CPU-time, using

calls to coarser and cheaper versions of the initial simulator. Initially designed for the estimation of means, these approaches have recently been extended to the estimation of statistical moments. We have successfully applied these tools to a 1D hydraulic model before extending the use of multi-level Monte-Carlo techniques to the Garonne river in a context of sensitivity analysis and data assimilation based on ensemble Kalman filtering.

Keywords. Multi-level Monte-Carlo, multifidelity, uncertainty quantification, sensitivity analysis, data assimilation, fluvial modelling.

1 Contexte

Les codes de calcul approchant un phénomène physique comportent souvent des entrées incertaines dont l'impact sur la quantité d'intérêt peut être important. On cherche alors à quantifier la part de chaque entrée ou groupe d'entrées dans la variabilité de cette sortie ; c'est le cadre de l'analyse de sensibilité [8,9]. L'incertitude sur la sortie peut être réduite en combinant les simulations *in silico* à des observations *in situ*, tout en tenant compte des modèles d'incertitudes des entrées et des observations. C'est le cadre de l'assimilation de données [1,10] où la combinaison entre simulations et observations se fait au sens d'un critère d'optimalité minimisant la variance. Lorsque les incertitudes sur les observations sont supposées gaussiennes, les algorithmes d'assimilation de données reposent sur l'estimation d'espérances, de variances et de covariances. Dans le cadre de méthodes d'analyse de sensibilité et d'assimilation de données, l'estimation des statistiques se fait classiquement au moyen d'une approche de type Monte-Carlo requérant un grand nombre de simulations. Cependant, ce nombre est en pratique restreint par les limites de temps de calcul et le coût computationnel du simulateur.

Pour pallier cette limitation, une première solution consiste à substituer au code de calcul un métamodèle purement mathématique [3], calibré par apprentissage statistique de quelques évaluations et dont le faible coût d'exécution permet d'obtenir un grand nombre d'évaluations. Néanmoins, cette approche a le défaut d'ajouter à l'erreur de Monte-Carlo l'erreur du métamodèle. Aussi, en présence de versions du code de calcul ayant des précisions et coûts d'évaluation moindres, une solution plus adaptée consiste à construire des estimateurs par Monte-Carlo faisant d'autant plus appel à une version du code que celle-ci est peu onéreuse. On parle alors de méthodes de Monte-Carlo multi-niveaux [2,4,5,7,11]. Ces techniques sont en plein essor depuis une dizaine d'années, essentiellement dans le cas où les codes de calcul résolvent numériquement des systèmes d'équations aux dérivées partielles ; le niveau du code de calcul correspond alors à la finesse de discrétisation de ces équations.

2 Principe de l'estimation Monte-Carlo multi-niveaux

On représente la quantité d'intérêt en sortie du code de calcul par une variable aléatoire P définie sur un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$ et dont on souhaite approcher l'espérance $\mathbb{E}[P]$. À partir d'un N-échantillon $\left(P^{(i)}\right)_{1\leq i\leq N}$ où $P^{(i)} := P\left(\omega^{(i)}\right)$, l'estimateur de Monte-Carlo s'écrit

naturellement $N^{-1}\sum_{i=1}^{N} P^{(i)}$, sa variance est égale à $N^{-1}\mathbb{V}[P]$ et son erreur quadratique moyenne est de l'ordre de N^{-1}. Ainsi, une précision de ε sur l'estimateur requiert un échantillon de taille importante de l'ordre de ε^{-2}. Face à cette limitation, on considère L versions différentes du code de calcul engendrant une séquence de variables aléatoires $P_0, P_1, \ldots, P_L$ qui approchent P avec une précision et un coût croissant selon $\ell \in \{0, 1, \ldots, L\}$. On obtient alors l'identité suivante :

$$\mathbb{E}[P_L] = \sum_{\ell=0}^{L} \mathbb{E}[P_\ell] - \mathbb{E}[P_{\ell-1}] = \sum_{\ell=0}^{L} \mathbb{E}[Y_\ell] \text{ où } Y_\ell = P_\ell - P_{\ell-1}$$

où $P_{-1} := 0$. L'estimateur sans biais de $\mathbb{E}[P_L]$, que l'on nomme "estimateur de Monte-Carlo multi-niveaux" (MLMC pour *multilevel Monte-Carlo*), s'écrit dans ce cas :

$$E_L^{\mathsf{ML}}[P] = \sum_{\ell=0}^{L} E_{N_\ell}[P_\ell] - E_{N_\ell}[P_{\ell-1}] = \sum_{\ell=0}^{L} E_{N_\ell}[Y_\ell]$$

où $E_{N_\ell}[P_k] = N_\ell^{-1}\sum_{i=1}^{N_\ell} P_k^{(\ell,i)}$, $P_k^{(\ell,i)} := P_k\left(\omega^{(\ell,i)}\right)$ et où les réalisations de ω sont indépendantes et identiquement distribuées. Il est facile de montrer que $\mathbb{E}\left[\left(E_L^{\mathsf{ML}}[P] - \mathbb{E}[P]\right)^2\right]$, l'erreur quadratique moyenne de $E_L^{\mathsf{ML}}[P]$, se décompose en la somme :

- d'un terme de biais associé au niveau le plus fin L : $\mathbb{E}\left[P_L - P\right]^2$,
- d'un terme de variance prenant en compte la variance de chaque estimateur Y_ℓ pondérée par la taille de l'échantillon associé : $\sum_{\ell=0}^{L} \frac{1}{N_\ell}\mathbb{V}[Y_\ell]$.

Pour une précision $\varepsilon > 0$ fixée, l'objectif est alors de chercher un nombre de niveaux L, une suite de complexités de codes $\{M_\ell\}_{0\le\ell\le L}$ et une suite de tailles d'échantillons $\{N_\ell\}_{0\le\ell\le L}$ équilibrant ces deux termes afin que la somme soit de l'ordre de ε^2. Ce problème revient à optimiser le coût total et la précision de l'estimateur MLMC. En pratique, une première approche consiste à prendre pour tout $\ell \in \{0, 1, \ldots, L\}$, $M_\ell := 2^\ell$, de sorte que la complexité algorithmique varie d'un rapport de 2 entre deux niveaux de codes. Dans le cadre de différences finies, ceci revient à diviser par 2 le pas d'espace en passant du niveau ℓ au niveau $\ell - 1$.

Théorème 1 ([2], adaptation de [4]) *On note respectivement C_ℓ et V_ℓ le coût et la variance de Y_ℓ. Soit $\{M_\ell\}_{\ell=0}^{\infty}$ une séquence d'entiers positifs, à croissance exponentielle et satisfaisant $M_\ell/M_{\ell-1} \ge a$ pour un $a > 1$. Soient α, β, γ des constantes positives telles que :*

$$(i)\quad |\mathbb{E}[P_\ell - P]| \lesssim M_\ell^{-\alpha} \qquad (ii)\quad V_\ell \lesssim M_\ell^{-\beta} \qquad (iii)\quad C_\ell \lesssim M_\ell^{\gamma}.$$

Alors pour toute précision $\varepsilon > 0$, il existe des valeurs de L et de $\{N_\ell\}_{0\le\ell\le L}$ pour lesquelles l'estimateur $E_L^{ML}[P]$ a une erreur quadratique moyenne :

$$\mathbb{E}\left[\left(E_L^{ML}[P] - \mathbb{E}[P]\right)^2\right] < \varepsilon^2.$$

Le coût total $\text{Coût}\left(E_L^{ML}[P]\right)$ *vérifie quant à lui :*

$$\text{Coût}\left(E_L^{ML}[P]\right) \lesssim \varepsilon^{-\frac{\gamma}{\alpha}} + \begin{cases} \varepsilon^{-2}, & \beta > \gamma, \\ \varepsilon^{-2}\left|\log(\varepsilon)\right|^2, & \beta = \gamma, \\ \varepsilon^{-2-\frac{\gamma-\beta}{\alpha}}, & \beta < \gamma. \end{cases}$$

À un terme logarithmal près, le coût total se résume à $\text{Coût}\left(E_L^{\mathsf{ML}}[P]\right) \lesssim \varepsilon^{-2-\frac{\gamma}{\alpha}+\frac{\min(2\alpha,\beta,\gamma)}{\alpha}}$, entraînant ainsi une réduction de $\varepsilon^{\frac{\min(2\alpha,\beta,\gamma)}{\alpha}}$ par rapport à une méthode de Monte-Carlo classique appliquée uniquement au niveau le plus fin L.

3 Implémentation de l'estimation MLMC de l'espérance

Dans le cadre de l'estimation de l'espérance avec une précision donnée, [5] a développé un algorithme itératif afin d'échantillonner les différents niveaux de codes de façon séquentielle. Celui-ci est disponible[1] en Matlab, C++, Python et R.

L'algorithme est initialisé en fixant un nombre minimum ℓ_{min}, un nombre maximum ℓ_{max} et un nombre courant $L := \ell_{\mathsf{min}}$ de niveaux et en donnant de faibles tailles aux échantillons de départ : $N_0^{[1]}, N_1^{[1]}, \ldots, N_L^{[1]}$ pour chacun des niveaux $0, 1, \ldots, L$. On définit également pour chaque niveau ℓ le coût calculatoire C_ℓ d'une évaluation.

Par la suite, à chaque itération it, on simule pour chaque niveau ℓ le $N_\ell^{[it]}$-échantillon $\left(P_\ell^{(\ell),it,i} - P_{\ell-1}^{(\ell),it,i}\right)_{1\leq i\leq N_\ell^{[it]}}$ et on calcule sa moyenne empirique μ_ℓ et sa variance empirique σ_ℓ^2. En utilisant ces quantités et les relations (i) et (ii) du Théorème 1, on estime alors par regression linéaire les paramètres α et β. Pour chaque niveau ℓ, le nombre optimal de simulations est calculé via la formule $N_\ell = (1-\theta)^{-1}\varepsilon^{-2}\sigma_\ell C_\ell^{-1/2}\sum_{k=0}^{L}\sqrt{\sigma_k^2 C_k}$, pour un $\theta \in]0,1[$, et le cas échéant le nombre de simulations supplémentaires $N_\ell^{[it+1]} := N_\ell - N_\ell^{[it]}$. Si ce dernier est significatif, on repète les précédentes étapes en itérant $it := it + 1$. Sinon, dans le cas où $L < \ell_{\mathsf{max}}$, on ajoute un nouveau niveau de finesse $L := L + 1$ avant d'initialiser sa variance σ_L^2 et son coût C_L, d'actualiser les nombres optimaux et supplémentaires de simulations et de répéter les étapes précédentes en itérant $it := it + 1$.

Lorsque l'algorithme a convergé ou que $L := \ell_{\mathsf{max}} + 1$, on calcule l'estimateur MLMC de l'espérance à partir des différents échantillons $\left(P_\ell^{(\ell),it,i} - P_{\ell-1}^{(\ell),it,i}\right)_{\substack{1\leq i\leq N_\ell^{[it]} \\ 1\leq it\leq it_{\mathsf{final}}}}$.

4 Extension de la méthode MLMC à la variance

De manière analogue à l'espérance de P, l'estimateur MLMC de sa variance s'écrit :

$$V_L^{\mathsf{ML}}[P] := \sum_{\ell=0}^{L} V_{N_\ell}\left[P_\ell\right] - V_{N_\ell}\left[P_{\ell-1}\right]$$

[1] `http://people.maths.ox.ac.uk/gilesm/mlmc/` et `https://bitbucket.org/pefarrell/pymlmc`

où $P_{-1} := 0$ et $V_{N_\ell}[P_k] := \frac{1}{N_\ell - 1} \sum_{i=1}^{N_\ell} \left(P_k^{(\ell,i)} - E_{N_\ell}[P_k] \right)^2$. On observe que $V_L^{\mathsf{ML}}[P]$ est un estimateur sans biais de $\mathbb{V}[P_L]$.

Théorème 2 ([2]) *Soient $1 \leq p \leq q \leq \infty$ tels que $\frac{1}{p} + \frac{1}{q} = 1$ et $P : \Omega \to B^{2q}$ un champ aléatoire à valeurs dans l'espace de Banach $B^{2q} \in \{L^{2q}(D), W^{1,2q}(D), \mathbb{R}\}$ où $D \subset \mathbb{R}$. Soit $\{M_\ell\}_{\ell=0}^{\infty}$ une séquence d'entiers positifs, à croissance exponentielle et satisfaisant $M_\ell / M_{\ell-1} \geq a$ pour un $a > 1$. Soient α, β, γ des constantes positives telles que :*

$$(i) \quad \|\mathbb{V}[P_\ell] - \mathbb{V}[P]\|_{B^2} \lesssim M_\ell^{-\alpha} \qquad (ii) \quad \|Y_\ell - \mathbb{E}[Y_\ell]\|_{L^{2p}(\Omega, B^{2p})}^2 \lesssim M_\ell^{-\beta} \qquad (iii) \quad C_\ell \lesssim M_\ell^{\gamma}$$

avec $P, P_\ell \in L^{2q}(\Omega, B^{2q})$ de normes uniformément bornées. Alors, pour toute précision $\varepsilon > 0$ telle que $\left\| V_L^{\mathsf{ML}}[X] - \mathbb{V}[X] \right\|_{L^2(\Omega, B^2)} < \varepsilon$, il existe un entier L et une séquence $\{N_\ell\}_{\ell=0}^{\infty}$ tels que le coût total de l'estimateur MLMC de la variance satisfait :

$$\text{Coût}\left(V_L^{\mathsf{ML}}[P]\right) \lesssim \varepsilon^{-\frac{\gamma}{\alpha}} + \begin{cases} \varepsilon^{-2}, & \beta > \gamma, \\ \varepsilon^{-2} \log(\varepsilon)^2, & \beta = \gamma, \\ \varepsilon^{-2 - \frac{\gamma - \beta}{2}}, & \beta < \gamma. \end{cases}$$

Les estimateurs MLMC de l'espérance et de la variance ont été appliqués avec succès à un modèle simple de courbes de remous : $\frac{\partial h(x)}{\partial x} = S_0 \left(1 - \left(\frac{h(x)}{h_n}\right)^{-1/3}\right) \left(1 - \left(\frac{h(x)}{h_c}\right)^{-0.3}\right)^{-1}$ sur $[0, x_{\mathsf{aval}}]$, avec $h(x_{\mathsf{aval}}) := h_{\mathsf{aval}}$ [m] la hauteur d'eau en aval, $h(x)$ [m] la hauteur d'eau au point x, $h_n = (Q^2 g^{-1} W^{-2})^{1/3}$ [m] la hauteur d'eau normale, $h_c = \left(Q^2 S_0^{-1} W^{-2} K_s^{-2}\right)^{0.3}$ [m] la hauteur critique, S_0 la pente du lit du canal, Q [$\text{m}^3.\text{s}^{-1}$] le débit en amont, K_s [$\text{m}^{1/3}.\text{s}^{-1}$] le coefficent de frottement, W [m] la largeur du canal rectangulaire et g [$\text{m}.\text{s}^{-2}$] l'accélération gravitationelle. Les paramètres Q et K_s sont considérés comme aléatoires et de lois de probabilité connues tandis que les autres sont fixés à des valeurs nominales. Selon la précision désirée, le coût calculatoire diminue d'un facteur variant de quelques unités à quelques dizaines d'unités en utilisant l'algorithme de la Section 3, celui-ci créant de nouveaux niveaux de codes et sollicitant d'autant moins un niveau que celui-ci est onéreux.

5 Application à l'analyse de sensibilité et à l'assimilation de données pour un modèle fluvial

Dans la philosophie de [2], nous développons des estimateurs MLMC pour les indices de Sobol' en analyse de sensibilité et pour les matrices de covariances du filtre de Kalman d'ensemble en assimilation de données. Nous les accompagnons d'un algorithme d'échantillonnage itératif des différents niveaux analogue à celui de la Section 3. Ces développements sont appliqués au logiciel de simulation hydraulique MASCARET [6] résolvant les équations non conservatives de Saint-Venant monodimensionnelles :

$$\frac{\partial A(h)}{\partial t} + \frac{\partial Q}{\partial x} = 0, \quad \frac{\partial Q}{\partial t} + \frac{\partial Q / A(h)}{\partial x} + g A(h) \left(\frac{\partial h}{\partial x} - S_0 + S_f \right) = 0$$

avec $S_f = \frac{Q^2}{K_S^2 A^2(h) R(h)^{4/3}}$, où A [m^2] est la section du fleuve, Q [m^3.s^{-1}] le débit, h $[m]$ la hauteur d'eau, K_s [m$^{1/3}$.s^{-1}] le coefficent de frottement, R le rayon hydraulique, g [m.s^{-2}] l'accélération gravitationelle et S_0 et S_f les pentes respectivement de la rivière et de frottement. Le logiciel MASCARET renvoie en sortie l'état hydraulique (h, Q) sur un maillage du réseau $[x_{\text{amont}}, x_{\text{aval}}]$ et sur une fenêtre temporelle donnée et discrétisée. Ce travail s'inscrit dans le contexte de la prévision des crues et de la gestion de la ressource en eau dans le cadre d'une collaboration avec EDF et le SCHAPI (Service Central d'Hydrométéorologie d'Appui à la Prévision des Inondations). On s'intéresse à titre illustratif à un tronçon du fleuve Garonne s'étendant sur 50 km entre Tonneins et La Réole. Les incertitudes sur l'état hydraulique simulé sont supposées liées à des incertitudes sur la description du coefficient de frottement (discrétisé par un nombre limité de zones géographiques) et du débit d'apport à l'amont du réseau (variable en temps en régime instationnaire). On cherche à estimer la matrice de covariance entre les erreurs sur les variables incertaines en entrée de la simulation et la hauteur d'eau, ainsi que la matrice de covariance entre la hauteur d'eau en différents points du réseau.

Bibliographie

[1] M. Asch, M. Bocquet et M. Nodet (2016), *Data assimilation: methods, algorithms, and applications*, SIAM, Fundamentals of Algorithms, pp. xviii + 306.

[2] C. Bierig et A. Chernov (2015), *Convergence analysis of multilevel Monte Carlo variance estimators and application for random obstacle problems*, Numerische Mathematik, 130(4):579-613

[3] M. De Lozzo (2015), *Substitution de modèle et approche multifidélité en expérimentation numérique*, journal de la Société Française de Statistique, 156(3).

[4] M.B. Giles (2008), *Multi-level Monte Carlo path simulation*, Operations Research, 56(3):607-617.

[5] M.B. Giles (2015), *Multilevel Monte Carlo methods*, Acta Numerica, 24:259-328.

[6] N. Goutal et F. Maurel (2002), *A finite volume solver for 1D shallow-water equations applied to an actual river. International Journal for Numerical Methods in Fluids*, 38(1):1-19.

[7] H. Hoel, K.J.H. Law et R. Tempone (2016), *Multilevel ensemble Kalman filter*, SIAM, Journal of Numerical Analysis, 54(3):1813-1839.

[8] B. Iooss (2011), *Revue sur l'analyse de sensibilité globale de modèles numériques*, Journal de la Société Française de Statistique, 152(1).

[9] A. Janon, M. Nodet, C. Prieur et C. Prieur (2016), *Global sensitivity analysis for the boundary control of an open channel*, Mathematics of Control, Signals, and Systems, 28(1).

[10] F.X. Le Dimet et J. Blum (2002), *Assimilation de données pour les fluides géophysiques*, Matapli, 67:33-55.

[11] S. Mishra, Ch. Schwab et J. Šukys (2011), *Multi-level Monte Carlo finite volume methods for shallow water equations with uncertain topography in multi-dimensions*, SIAM, Journal on Scientific Computing, 34(6):761-784.