Sujet de recherche (2021)

Consolidation des aléas naturels extrêmes

1 Contexte

Ce sujet de stage s'inscrit dans le cadre des travaux de R&D d'EDF en traitement d'incertitude au sein des modèles numériques pour la sûreté et la conception, et plus largement dans celui de la construction du Groupement d'Intérêt Scientifique (GIS) LARTISSTE. Il concerne une problématique portée au sein d'EDF par le projet MADONE3, qui traite de la protection de sites contre certains types d'agressions externes naturelles (ex : crue, vent fort, froid intense) et la mitigation des risques associés.

Cette problématique porte plus spécifiquement sur la mise en oeuvre de plusieurs méthodes visant à produire des estimateurs d'indicateurs probabilistes, comme des probabilités ou quantiles de la variable $T = g(X, \theta)$, associés à la survenue d'un événement extrême $X \in \chi$. Plus précisément, on vise à déterminer les valeurs de (X, T) correspondant à un aléa de référence associé à une période de retour de 10 000 ans).

Dans ce formalisme :

— X est un ensemble de variables temporelles représentant les variables météorologiques (température de l'air, vitesse du vent, etc.), accessible via des données mesurées *in situ*. Toutefois, ces données sont de taille faible, et on gagnerait beaucoup à rechercher des sources publiques (notamment des données d'origine satellitaire) pour compléter ces jeux de données et produire des **modèles génératifs** tenant compte de l'évolution météorologique et climatique.

— T est une simulation de l'évolution temporelle température d'eau, qui est une variable importante pour l'étude de l'impact du changement climatique sur des biens ou services.

— g est un modèle de simulation phénoménologique de la température, qui représente la dynamique thermique d'un écoulement. Il est incarné par le code de dynamique des fluides TELEMAC 2D, disponible sur la plateforme `http://www.opentelemac.org`, couplé à un module de qualité de l'eau. Il est paramétré par θ, qui reflète des entrants connus (ex : température de l'eau en amont dans un cours d'eau) mais aussi la géométrie du lieu et qui est supposé connu dans les expérimentations (ex : θ peut typiquement représenter des contraintes liées à la géomorphologie) ; chaque simulation de g peut être lourde en temps de calcul (plusieurs heures).

La recherche de valeurs extrêmes de Y peut être traitée par des approches de statistique des extrêmes, à condition de disposer de suffisament de valeurs observées de Y. On peut également souhaiter utiliser des techniques de type Monte Carlo par réduction de variance, par régression quantile avec intégration de l'incertitude (ex : Conformalized Quantile Regression), des approches type Wilks, des approches tirant parti de la monotonie des phénomènes pour encadrer le résultat de façon déterministe. Voici quelques références utiles sur le sujet [3], [**?**],[2]

En définitive, l'objectif final est bien de proposer une démarche d'objectivisation d'un aléa de référence, fondé sur deux problèmes à régler :

1. la **production de modèles génératifs des variables d'entrée par apprentissage statistique**, par exemple fondé sur de la réduction de dimension / sur des approches par réseaux de neurones [1] ;
2. **l'utilisation de ces modèles pour mener des calculs de température, tout en contounrnant les difficultés liées au temps de calcul de TELEMAC 2D.**

2 Descriptif des cas d'étude

Le cas d'étude considéré porte sur l'étude d'un tronçon de la Garonne, pour lesquels la géométrie est connue. On considèrera plusieurs scénarios pour la valeur de l'eau amont, qui est l'une des dimensions de θ.

Les variables métérologiques X_T concernent entre autres la température de l'air, la vitesse du vent, l'humidité relative et la couverture nuageuse. Des données *in situ* sont disponibles, mais on cherche à compléter celles-ci par des données satellitaires et à produire un modèle génératif de ces séries temporelles. Proposer une stratégie de couplage entre ces données et produire des modèles constituent la première étape du travail proposé.

La génération de trajectoires de température en sortie de TELEMAC 2D,en aval du canal de la Garonne considéré, ainsi que le calcul de quantiles extrêmes de cette température constitue la seconde étape de ce travail. Pour des raisons pratiques liés à l'installation de l'instance de TELEMAC 2D sur ce cas Garonne, et au coût de calcul, il sera proposé aux étudiants de travailler sur des clusters de calcul de l'entreprise, sous réserve d'un accord de confidentialité spécifique.

Il est envisageable que deux groupes d'étudiants puissent travailler en interaction sur chacun des deux problèmes.

3 Support et encadrement

Un corpus de document internes décrivant les données et modèles historiques seront mis à la disposition des étudiants suivis par un ou deux enseignants. Des chercheurs d'EDF R&D viendront présenter le sujet et pourront guider l'évolution du travail.

4 Suivi industriel

Plusieurs ingénieurs-chercheurs d'EDF R&D suivront le travail : Nicolas Bousquet, Fabrice Zaoui et Cédric Goeury.

Références

[1] A. Koochali and S. Ahmed. If you like it, GAN it ! probabilistic multivariate times series forecast with GAN. *ArXiv :2005.01181v*.

[2] Y. Romano, E. Patterson, and E.J. Candès. Conformalized quantile regression. *NeurIPS*.

[3] R.Y. Rubinstein and D.P. Kroese. *Simulation and the Monte Carlo Method*. Wiley Series in Probability and Statistics. Wiley, 2011.

Sujet de recherche (2021)

Modélisation bayésienne robuste en statistique des extrêmes

1 Contexte

Ce sujet de stage s'inscrit dans le cadre des travaux de R&D d'EDF en traitement d'incertitude au sein des modèles numériques pour la sûreté et la conception, et plus largement dans celui de la construction du Groupement d'Intérêt Scientifique (GIS) LARTISSTE. Il concerne une problématique portée au sein d'EDF par le projet MADONE3, qui traite de la protection des sites de production d'EDF contre certains types d'agressions externes naturelles (ex : crue, vent fort, froid intense) et la mitigation des risques associés.

Cette problématique porte plus spécifiquement sur la construction et la mise en oeuvre d'une méthode de calcul bayésien *robuste* pour une loi d'extrême.

La construction de lois *a priori* sur les paramètres des distributions de valeurs extrêmes est une tâche difficile. Les experts n'ont en général aucune intuition sur la signification des paramètres et préfèrent faire des évaluations sur des quantités observables. Leurs jugements sont fondés sur l'expérience passée, où des événements intéressants se sont rarement produits, ce qui entraîne une incertitude importante dans les évaluations. La transformation de ces expertises en lois *a priori* peut être fortement affectée par l'arbitraire introduit par le statisticien, par exemple dans le choix de leurs formes fonctionnelles.

La robustesse bayésienne (ou analyse de sensibilité bayésienne : [2, 3]) est une

approche visant à lutter contre l'impossibilité pratique de spécifier exactement les distributions a priori, les modèles statistiques et les fonctions de perte, c'est-à-dire les trois ingrédients de l'approche bayésienne. En particulier, le choix de la distribution a priori est l'aspect le plus critique de cette approche. Son application aux lois issues de la théorie des valeurs extrêmes, dans des cadres appliquées où celle-ci est usuellement utilisée (par exemple pour modéliser le comportement de variables météorologiques extrêmes, telle la température, la pluviométrie, l'humidité relative, etc.), n'a à ce jour pas encore été menée.

Dans la pratique, l'approche bayésienne robuste se fonde sur la méthodologie suivante : on considère une classe de distribution *a priori*, et une gamme de valeur couverte par la quantité d'intérêt (ici, principalement les quantiles *a posteriori* et des périodes de retour) quand la loi *a priori* (prior) varie dans cette classe. Si l'intervalle est petit, alors tout prior dans la classe peut être choisi puisque le choix n'affecte pas l'estimation de la quantité d'intérêt. Si la fourchette est large, les experts doivent fournir des informations complémentaires afin de réduire la taille de la classe du prior. La procédure doit être répétée jusqu'à ce qu'une petite fourchette soit obtenue ou qu'aucun affinement ne soit possible. Dans ce dernier cas, l'analyse doit être effectuée en utilisant un seul prior (peut-être optimal au regard de certains critères) mais en indiquant la fourchette de la quantité d'intérêt et en reconnaissant comment l'estimation est affectée par ce choix particulier.

Ce projet vise donc à mettre en place une telle analyse.

2 Applications

On considère deux situations faisant intervenir les lois de maxima d'un phénomène naturel :

— **un jeu de données réelles de maxima journaliers annuels de débits de la Meuse** en une station de mesure située près de la ville de Liège (Belgique), téléchargeable à l'adresse

`https://www.lpsm.paris/pageperso/bousquet/coursM2-2021/examen/projets/flood-1/max-meuse.txt`

Les mesures sont données en m^3/s. On s'intéresse ici à l'estimation des niveaux de retour à 4 ans, puis aux niveaux correspondant à des probabilités de dépassement d'au plus 0.1 et 0.001. Une information *a priori* est donnée dans la table 1. Celle-ci est issue d'une expertise produite à partir de modèles de simulation qui tentent de prendre en compte la variation prédictive du débit dans des conditions de changement climatique au cours du 21ème siècle.

Percentile order	Discharge (m^3/s)
5%	1250 ($\pm$200)
50%	2000 ($\pm$100)
75%	2100 ($\pm$100)

TABLE 1 – Prior predictive information on daily maxima discharge per year, extrapolated by numerical analysis of physically-based climate models.

— **un jeu de données réelles de maxima journaliers annuels de la pluviométrie** à Penta-di-Casinca (Haute Corse), téléchargeable à l'adresse

`https://www.lpsm.paris/pageperso/bousquet/coursM2-2021/examen/projets/flood-1/pluviometry-corsica.csv`

Les mesures sont données en mm. On s'intéresse ici à l'estimation des niveaux de retour à 50 puis 100 ans. Une information historique *a priori*, issue de l'interrogation d'un expert de Météo-France, est donnée dans la table 2.

Percentile order	Pluviometry P (mm)
25%	75 ($\pm$20)
50%	100 ($\pm$20)
75%	150 ($\pm$20)

TABLE 2 – Prior predictive information on daily maxima pluviometry per year, extrapolated by an expert from daily maxima measured at a nearby station.

3 Formalisation

3.1 Principe général

On considère pour une grandeur X d'intérêt la loi des valeurs extrêmes généralisées (GEV) de fonction de répartition

$$F(x;\theta) = \exp\left\{ -\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]_+^{-1/\xi} \right\},$$

et de densité

$$f(x;\theta) = \frac{1}{\sigma}\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]_+^{-1/\xi-1} \exp\left\{ -\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]_+^{-1/\xi} \right\},$$

avec $\mu \in \mathbb{R}$, $\xi \in \mathbb{R}$ et $\sigma \in \mathbb{R}^+$. On note $\theta = (\mu, \sigma, \xi) \in \Theta$. Au regard des cas d'étude, on considère donc disposer pour chaque cas, outre un échantillon $x_1, \ldots, x_n$,

de spécifications *a priori* de la forme $F(x_{q_i}) = q_i, i = 1, \ldots, n$. En introduisant une distribution *a priori* de densité $\pi(\theta)$, on a alors :

$$F(x_{q_i}) = \int_\Theta F(x_{q_i}; \theta)\pi(\theta)d\theta = q_i, i = 1, \ldots, n.$$

Nous souhaitons évaluer comment les données expérimentales et les hypothèses sur la distribution priore pourraient affecter les quantiles et donc les niveaux de retour. De façon générale, les données observées $\mathbf{x} = (X_1, \ldots, X_m)$ conduisent à la vraisemblance $l_{\mathbf{x}}(\theta) = \prod_{j=1}^{m} f(X_j; \theta)$. L'intérêt réside dans l'obtention a posteriori d'une certaine quantité, disons $g(\theta)$, qui peut être estimée par

$$G(\pi) \quad = \quad \frac{\int_\Theta g(\theta) l_{\mathbf{x}}(\theta)\pi(d\theta)}{\int_\Theta l_{\mathbf{x}}(\theta)\pi(d\theta)}. \tag{1}$$

3.2 Approche robuste via une classe de moments contraints généralisés

Nous considérons une variable aléatoire X avec une distribution GEV. Nous supposons que l'expert est capable de spécifier uniquement des quantiles sur la quantité observable X (sans condition sur le paramètre θ), c'est-à-dire qu'il fournit uniquement des déclarations comme $F(x_{q_i}) = q_i, i = 1, \ldots, n$. En introduisant une distribution préalable $\pi(\theta)$, les instructions deviennent alors

$$F(x_{q_i}) = \int_\Theta F(x_{q_i}; \theta)\pi(\theta)d\theta = q_i, i = 1, \ldots, n.$$

Nous souhaitons évaluer comment les données expérimentales et les hypothèses sur la distribution a priori pourraient affecter ces quantiles (et même d'autres).

Une distribution a priori correspondant à ces quantiles pourrait être trouvée numériquement, au moins dans une approximation raisonnable, mais un tel choix serait sans doute arbitraire, fondé sur la commodité du statisticien plutôt que sur une évaluation efficace par l'expert. C'est pourquoi une approche bayésienne robuste est adoptée, en considérant toutes les distributions a priori compatibles avec les quantiles évalués et en étudiant ensuite l'influence d'un tel choix sur les quantités d'intérêt, à savoir les quantiles et les rendements.

La classe des priors admissibles est un cas particulier de la classe généralisée des moments contraints présentée dans [1], et donnée par

$$\Gamma = \{\pi : \int_\Theta H_i(\theta)\pi(\theta)d\theta \leq \alpha_i, \ ; i = 1, ..., n\}$$

où H_i sont des fonctions intégrées π et $\alpha_i, i = 1, \ldots, n$, sont des nombres réels fixes. Si nous prenons $H_i(\theta) = F(x_{q_i})$, $\alpha_i = q_i$ et, par souci de simplicité, l'égalité au lieu des bornes d'inégalité, alors la classe Γ est celle de tous les priors menant à ces quantiles, en supposant un modèle GEV.

L'approche bayésienne robuste s'intéresse à la mesure de l'effet d'une classe de priors sur la quantité d'intérêt (1). La mesure la plus courante est fournie par la gamme

$$\sup_{\pi\in\Gamma} G(\pi) - \inf_{\pi\in\Gamma} G(\pi).$$

La robustesse est atteinte lorsque cette gamme est *petite* (selon un jugement subjectif d'un décideur).

Nous nous concentrerons ici uniquement sur la manière de calculer $\sup_{\pi\in\Gamma} G(\pi)$, soit l'équivalent de $\inf_{\pi\in\Gamma} G(\pi)$.

Le théorème 3 de [1] montre que

$$\sup_{\pi\in\Gamma} G(\pi) = \sup_{(\theta,\mathbf{p})\in T} \frac{\sum_{j=1}^{n+1} g(\theta_j) l_{\mathbf{x}}(\theta_j) p_j}{\sum_{j=1}^{n+1} l_{\mathbf{x}}(\theta_j) p_j},$$

où $\theta = (\theta_1, \ldots, \theta_{n+1})'$, $\mathbf{p} = (p_1, \ldots, p_{n+1})'$ et l'ensemble $T \subset \Theta^{n+1} \times [0,1]^{n+1}$ est défini par les conditions suivantes :

- $\sum_{j=1}^{n+1} F(x_{q_i}; \theta_j) p_j = q_i, \ ; i = 1, \ldots, n$
- $\sum_{j=1}^{n+1} p_j = 1.$

Par conséquent, $\sup_{\pi\in\Gamma} G(\pi)$ est recherché dans le sous-ensemble des distributions extrêmes $\sum_{j=1}^{n+1} p_j \delta_{\theta_j}$, avec $\delta_\cdot$ la mesure de Dirac, satisfaisant les conditions ci-dessus.

Les quantités d'intérêt sont des quantiles à des probabilités données (et des temps de retour conséquents), mais elles ne peuvent être obtenues à partir d'une fonction $G(\pi)$ pour un choix adéquat de la fonction $g(\theta)$. Par conséquent, nous calculerons les limites supérieures et inférieures de $F(x|\mathbf{x})$, $x > 0$, et nous obtiendrons les limites des quantiles par une fonction inverse. Le processus est exigeant en termes de calculs (et conduit à des solutions approximatives), car de nombreux problèmes d'optimisation doivent être résolus pour obtenir les limites supérieures et inférieures de $F(x|\mathbf{x})$ sur une grille suffisamment fine, puis une fonction inverse doit être calculée numériquement pour obtenir les limites des quantiles.

Par conséquent, nous calculons d'abord $\sup$ et $\inf$ de

$$\frac{\sum_{j=1}^{n+1} F(x; \theta_j) l_{\mathbf{x}}(\theta_j) p_j}{\sum_{j=1}^{n+1} l_{\mathbf{x}}(\theta_j) p_j}$$

sur une grille de valeurs x et tracer les deux courbes $\inf_{\pi\in\Gamma} F(x|\mathbf{x})$ et $\sup_{\pi\in\Gamma} F(x|\mathbf{x})$. Supposons que l'intérêt se situe dans la plage a posteriori du quantile x_α d'ordre α. Nous traçons une ligne horizontale en correspondance de α et nous l'intersectons avec les courbes ci-dessus : les points d'intersection donnent les limites inférieure et supérieure de x_α.

On cherche donc à produire des algorithmes d'optimisation permettant de réaliser ces calculs. Il est à prévoir que lorsqu'un seul quantile prédictif *a priori* est pris en compte, le calcul ne soit pas trop compliqué. Lorsque plusieurs quantiles sont pris en compte, une solution pourrait être fondée sur l'interprétation du critère à optimiser comme une distribution de probabilité multivariée *a posteriori*, dont il faut chercher le mode joint. Dans ce cas, des approches par simulation, via (par exemple) un algorithme de type MCMC (en grande dimension) pourrait se révéler fructueux.

Si le temps le permet, on cherchera à réaliser le même type d'analyse avec une loi GPD (Pareto généralisée), et on pourra chercher à évaluer comment l'approche pourrait se généraliser au cas des données extrêmes multivariées.

4 Quelques indications pour guider ce travail

Ce travail de recherche vise à produire des estimateurs des quantités d'intérêt proposées plus haut, pour les deux cas d'étude, en mettant en oeuvre la démarche robuste proposée ci-dessus. Il sera intéressant de comparer les résultats avec une démarche classique où une loi *a priori* relativement arbitraire est choisie.

Il est donc conseillé de commencer par formaliser le problème, notamment en rappelant les principau résultats et en détaillant les formules plus haut, puis de choisir un cas simple de loi *a priori* auquel comparer les résultats. Il semble aussi important de se munir d'une procédure de simulation de données, afin de reproduire l'expérimentation et de tester la "robustesse" générale de l'approche (elle-même dite robuste).

On attend de ce travail, outre une formalisation et une mise en oeuvre, une rédaction de code Python ou R bien documentée.

5 Suivi industriel

Deux ingénieurs-chercheurs d'EDF R&D du département PRISME suivront ce travail : Nicolas Bousquet et Merlin Keller.

Références

[1] B. Betrò, M. Męczarski, and F. Ruggeri. Robust bayesian analysis under generalized moments conditions. *Journal of Statistical Planning and Inference*, 41 :257–266, 1994.

[2] D. Ríos Insua, F. Ruggeri, and B. Vidakovic. Some results on posterior regret γ–minimax estimation. *Statistics and Decision*, 13 :315–331, 1995.

[3] D. Ríos Insua and F. (eds) Ruggeri. *Robust Bayesian Analysis*. Lecture Notes in Statistics, Springer :New York, 2000.

Proposition de sujet de travaux étudiant (2021)

Détermination de scénarii pénalisants (pire cas ou superquantile) dans le cadre de modèles hydrauliques d'inondation

1 Contexte

A ce jour, la démarche de sûreté des installations industrielles et civiles face au risque d'inondation repose essentiellement sur la définition de scénarii déterministes d'inondation. Par exemple, selon la méthodologie nationale, les PPRI (Plan de Prévention du Risque Inondation) doivent prendre en compte le plus fort événement connu à condition que celui-ci soit au minimum un événement de type centennal : c'est-à-dire ayant 1 chance sur 100 de se reproduire chaque année (on parle aussi de « période de retour de 100 ans »). Par la suite, lorsque le scénario de référence de l'étude a été identifié, son impact sur la zone d'intérêt est évalué par modélisation numérique, à travers des codes de calcul permettant de simuler des écoulements à surface libre grâce à la discrétisation d'équations aux dérivées partielles (équations de Saint-Venant ou Navier-Stokes).

Les études inondation dépendent de plusieurs paramètres, tels que la géométrie de la rivière, le débit et les courbes de tarage. Certains de ces paramètres, tels que le coefficient de frottement, sont incertains et mal connus. Par ailleurs, ces études doivent permettre de simuler, entre autres, le fonctionnement des ouvrages hydrauliques et d'autres phénomènes spécifiques des crues étudiées, tel que le phénomène de rupture de digues (et formation de brèches) qui sont « attendus » pour ce type de scénario. Le résultat de ces études est une caractérisation déterministe d'un écoulement (cartographie des hauteurs d'eau et des vitesses) pour une configuration donnée. Il en résulte que :

— la construction du modèle sur une zone d'étude nécessite la définition de maillages

exigeants d'un point de vue coût de calcul (plusieurs milliers de nœuds de calcul) ;
— les incertitudes sont intégrées de manière simplifiée à l'étude numérique (majoration forfaitaire, scénario « enveloppe » définit à dire d'expert, etc...).

Dans ce contexte, EDF R&D travaille depuis des années à la mise en place de méthodologies permettant de caractériser rigoureusement (i) les incertitudes liées aux paramètres physiques utilisées pour les études inondations (ex. frottements, bathymétrie, débits, brèches) et (ii) le (ou les) scénario (scénarii) le plus pénalisant pour une zone d'intérêt.

Plus précisément, le stage porte sur la mise en œuvre de plusieurs méthodes visant à identifier le scénario le plus impactant pour une quantité d'intérêt (ex. hauteur d'eau maximale ou hauteur d'eau de fréquence d'occurrence faible et inférieure à 0.001) dans le cadre d'une étude inondation classique. Une attention particulière sera portée sur l'évaluation de l'impact de la formation de brèches fluviales et des paramètres de frottement modélisant la résistance de l'écoulement sur le sol sur les résultats de modélisation. D'un point de vue technique, l'idée de base est d'utiliser des approches d'inversion déterministe adaptées aux expériences numériques, i.e. l'analyse des résultats de simulations coûteuses en temps de calcul à partir des techniques dite de "métamodélisation".

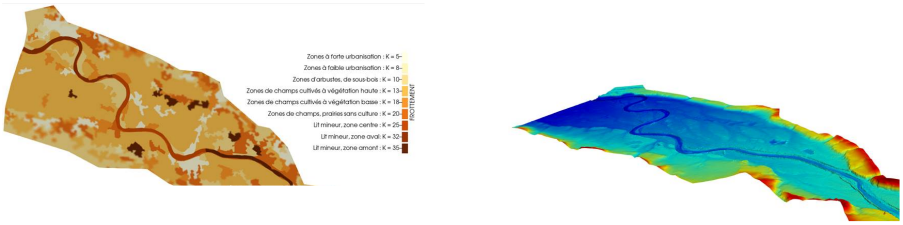
2 Quelques pistes issues de la bibliographie

L'objectif de ce travail est de la mise en oeuvre de différentes méthodes pour la détermination de scénarios "pire cas" ou de superquantile sur un jeu de données issus d'un modèle hydraulique. A l'heure actuelle, un travail basé sur la méthodologie décrite dans [3] est en cours d'investigation. En complément de celui-ci un travail d'exploration d'autres approches est souhaité. Certains travaux ont déjà été appliqué avec succès au BRGM pour l'analyse du risque de submersion marine à l'aide de simulateur numérique couteux en temps de calcul (ex. en référence [5] et [6]), et plus récemment par l'IRSN dans le cadre de la "qualification"de l'aléa tsunamis sur le littoral français [1], ou en contexte fluvial pour l'analyse des incertitudes de modélisation des brèches sur les niveaux d'eau en zone inondable. L'ambition d'inverser systématiquement et exhaustivement exige d'être économe en nombre d'appels au simulateur. Pour répondre à cette exigence, des techniques d'apprentissage actif ("active learning"), i.e. d'optimisation des scénarios de simulation à réaliser, ont été développés comme la famille d'algorithmes "Stepwise Uncertainty Reduction SUR" [2], qui ont déjà été mis en place sur un cas d'inondation fluviale ([4]).

3 Applications

Pour la réalisation des travaux, un jeu de données d'entrée et des sorties associées sera fourni. Celui-ci est issu de 10000 simulations 2D réalisées avec le module TELEMAC-2D du systeme hydro-informatique TELEMAC-MASCARET. La zone d'étude est une portion de la Loire. Les différents calculs sont issus d'un échantillonnage de Monte

Carlo pour lequel les variables incertaines considérées dans l'étude sont tous les coefficients de frottement des différentes zones (de l'ordre de 10^2 (cf. Figure 1a)) ainsi que les différents paramètres de chacune des brèches considérées (de l'ordre de 10 (trois paramètres par brèche multipliés par 12 brèches possibles dans la zone d'étude)). La quantité d'intérêt analysée au cours de cette étude est la hauteur d'eau.



(a) zones de frottement (b) fond

FIGURE 1 – Zoom du domaine d'étude

4 Quelques indications pour guider ce travail

Le programme de travail suivant est proposé :

— Définition du cadre théorique au regard de l'objectif visé (h-max ou quantile extrême)
— Benchmark d'algorithmes d'inversion à partir d'un jeu de données inondations (plan d'expérience). Objectifs :
  — Identification des algorithmes les plus précis au regard de la variable d'intérêt (h-max ou quantile) ;
  — Identification des algorithmes les plus performants (qualité de prédiction, nombre de scénarii nécessaire pour la construction de l'émulateur, etc. . .)
— Propositions d'outils adaptés au cadre industriel de l'étude.

5 Stage potentiel

Ce sujet pourra éventuellement faire l'objet d'un stage en 2022, sous réserve d'un accord hiérarchique prévu pour fin 2021.

6 Contacts EDF

— julien.pelamatti@edf.fr
— cedric.goeury@edf.fr

Références

[1] V. Bacchi, E. Antoshchenkova, H. Jomard, L. Bardet, C.-M. Duluc, O. Scotti, and H. Hebert. Development of a methodological framework for the assessment of seismic induced tsunami hazard through uncertainty quantification : application to the azores-gibraltar fracture zone. *Natural Hazards and Earth System Sciences Discussions*, 2018 :1–40, 2018.

[2] J. Bect, D. Ginsbourger, L. Li, V. Picheny, and E. Vazquez. Sequential design of computer experiments for the estimation of a probability of failure. *Statistics and Computing*, 22(3) :773–793, 2012.

[3] A. Marrel, B. Iooss, and V Chabridon. The ICSCREAM methodology : Identification of penalizing configurations in computer experiments using screening and metamodel – Applications in thermal-hydraulics. working paper or preprint, August 2021.

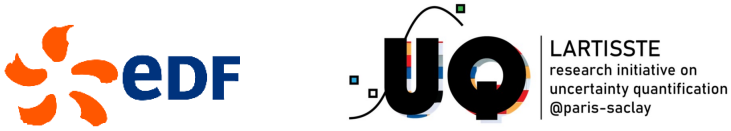
[4] Y. Richet and V. Bacchi. Inversion algorithm for civil flood defense optimization : Application to two-dimensional numerical model of the garonne river in france. *Frontiers in Environmental Science*, 7 :160, 2019.

[5] J. Rohmer and D. Idier. A meta-modelling strategy to identify the critical offshore conditions for coastal flooding. *Natural Hazards and Earth System Sciences*, 12(9) :2943–2955, 2012.

[6] J. Rohmer, D. Idier, F. Paris, R. Pedreros, and J. Louisor. Casting light on forcing and breaching scenarios that lead to marine inundation : Combining numerical simulations with a random-forest classification approach. *Environmental Modelling and Software*, 104 :64–80, June 2018.

Proposition de sujet de travaux étudiant (2021)

Prise en compte des incertitudes des données géométriques pour des études d'inondation



Cédric Goeury - *cedric.goeury@edf.fr*
Mathieu Couplet - *mathieu.couplet@edf.fr*

1 Contexte

La construction d'un modèle numérique en hydraulique fluviale commence par un recueil de données sur la topographie, la bathymétrie, l'occupation des sols, le fonctionnement hydraulique des aménagements et leurs caractéristiques hydrologiques. Puis on construit des éléments géométriques représentatifs du relief, qui servent ensuite de support pour la construction d'un maillage. En effet, afin de représenter les éléments déterminants pour l'écoulement des crues dans le domaine d'étude, le maillage s'appuie sur des lignes de contrainte qui définissent des lignes de rupture de pente, soit naturelles (coteaux abrupts, berges,...) soit artificielles (routes en remblais, digues,...). Le maillage est plus raffiné dans les zones d'intérêt du modèle (lit mineur, levées et déversoirs, ...). Il reste alors à interpoler les données altimétriques sur ce maillage pour constituer un modèle numérique de terrain (MNT). Cependant, les données à notre disposition ne sont pas exhaustives et peuvent être entachées d'incertitude de mesure. En effet, la précision altimétrique d'un MNT dépend de la méthode d'acquisition des données (technologie LiDAR, Radar, photographies aériennes, etc.) et des traitements effectués selon les caractéristiques des zones traitées (littorales, inondables, forestières, urbaines, rurales, etc.). Par exemple, pour les zones inondables et littorales, la précision altimétrique du MNT fourni par l'Institut National de l'Information Géographique et Forestière (IGN) est comprise entre 0,2 m et 0,5 m (RGE ALTI, 2018). Par ailleurs, certains phénomènes physiques tels que la migration des barres en sédimentologie engendrent des modifications significatives de la bathymétrie des rivières.

Ce projet vise à fournir une modélisation des incertitudes de relief à partir des

informations disponibles et d'exploiter le modèle obtenu. Dans la suite, seul le paradigme des probabilités est retenu pour la modélisation des incertitudes et le cadre de la géostatistique est adopté. Plus précisément, les techniques du krigeage et de manipulation de processus gaussiens (PG) sont considérés dans l'objectif de pouvoir générer aléatoirement des réalisations d'un PG préalablement construit à partir des données, et ainsi de mettre en œuvre certaines méthodes d'analyse de sensibilité ou de propagation d'incertitudes dans le cadre d'étude d'inondation.

1.1 Données altimétriques : topographie et bathymétrie

Les données altimétriques (topographie et bathymétrie) concernent un même objet physique : le relief du sol. Considérons une projection cartographique associée à des coordonnées planes notées x et y (longitude et latitude usuelles par exemple). Ce relief peut être représenté mathématiquement par un champ bidimensionnel

$$\begin{array}{rccc} h \ : & \mathcal{D} \subset \mathbb{R}^2 & \longrightarrow & \mathbb{R} \\ & (x,\, y) & \longmapsto & h(x,\, y) \end{array}$$

qui associe à un point donné d'une étendue $\mathcal{D}$ de la surface du sol, repéré par deux coordonnées $(x,\, y) \in \mathbb{R}^2$, sa hauteur $h(x,\, y) \in \mathbb{R}$. Dans l'étude des crues, il faut considérer une étendue $\mathcal{D}$, appelée lit majeur, bien plus large que la zone $\mathcal{B}$ où les eaux s'écoulent en temps normal, appelée lit mineur : $\mathcal{B} \subset \mathcal{D}$ (inclusion stricte). Ici, on appelle *bathymétrie* la restriction du champ $h(.)$ au lit mineur, notée $h_{\mathcal{B}}$:

$$\begin{array}{rccc} h_{\mathcal{B}} \ : & \mathcal{B} & \longrightarrow & \mathbb{R} \\ & (x,\, y) & \longmapsto & h_{\mathcal{B}}(x,\, y) = h(x,\, y) \end{array} .$$

En outre, on appelle *topographie* la restriction, notée $h_{\mathcal{T}}$, du champ $h(.)$ qui permet de compléter la bathymétrie pour retrouver l'ensemble du relief considéré :

$$\begin{array}{rccc} h_{\mathcal{T}} \ : & \mathcal{T} & \longrightarrow & \mathbb{R} \\ & (x,\, y) & \longmapsto & h_{\mathcal{T}}(x,\, y) = h(x,\, y) \end{array} \quad \text{où} \quad \mathcal{T} = \mathcal{D} \setminus \mathcal{B}.$$

Dans la suite, on notera $x \in \mathbb{R}^2$ plutôt que (x, y) les coordonnées pour simplifier.

Lors de la modélisation du relief, il convient de distinguer la topographie et la bathymétrie du fait de leurs spécificités. Les premières spécificités concernent la nature des données à disposition. L'IGN a employé différentes techniques de mesures du relief, notamment la télédétection par laser (*light detection and ranging*, ou LiDAR) aéroportée pour les zones inondables et littorales, et les grands massifs forestiers ; d'autres moyens moins précis sont utilisés sur le reste du territoire (RGE ALTI, 2018). Nous n'avons pas accès aux données brutes correspondantes, mais à un MNT « maillé » qui résulte notamment d'une interpolation des mesures par LiDAR sur une grille cartésienne de résolution homogène de 1 m (c'est-à-dire 1000×1000 points par km^2). Ce MNT est appelé RGE [1] ALTI® (version 2.0). Les techniques employées pour le constituer ne permettent pas de mesurer le relief sous les eaux. Il n'a en effet pas vocation

1. Référentiel Géographique à Grande Échelle

à décrire le sol habituellement immergé (fond des cours d'eau ou plan d'eau), même s'il couvre en partie l'estran et parfois la bande littorale (bathymétrie) par l'incorporation de données du référentiel géométrique Litto3D® (autour des ports en particulier). Aussi, le MNT modélise la surface des plans d'eau en extrapolant la position du sol sur les berges ou éventuellement la position mesurée de la surface de l'eau. Le RGE ALTI® n'est pas toujours précisément conforme à la réalité : outre les zones où les données initiales sont peu denses ou absentes et les zones de surplombs, des inexactitudes sont par exemple possibles dans les zones inondables scannées par LiDAR en cas de de végétation dense de faible hauteur [2]. Dans ces zones inondables les mieux couvertes par LiDAR (densité souhaitée de 1 pts/m²), les erreurs quadratiques moyennes estimées sont respectivement de 0,2 m en altitude et 0,6 m dans le plan (RGE ALTI, 2018, Annexe A). Notons qu'un programme de couverture nationale en données LiDAR Haute Densité piloté par l'IGN est en cours (avec un objectif de 10 pts/m² en moyenne sur l'ensemble du territoire national). À terme (2025), la diffusion en *open data* des nuages de points et des résultats des traitements dans une infrastructure numérique nationale (Géoplateforme), mais aussi l'accompagnement des acteurs des politiques publiques dans la manipulation des nuages de points, sont prévus [3].

La bathymétrie fait donc l'objet de mesures spécifiques, adaptés aux conditions aquatiques, typiquement des sondeurs acoustiques mono-faisceau ou, plus rarement, multi-faisceaux [4]. Souvent, la « résolution » de la bathymétrie est bien inférieure à celle de la topographie (c'est-à-dire un nombre de mesures par unité de surface beaucoup plus faible pour la bathymétrie) et, de plus, les points où le relief est mesuré sont distribués de manière beaucoup plus hétérogène pour la bathymétrie (le long de trajets suivi par un sondeur déplacé par bateau le long d'une section transverse ou d'amont en aval, typiquement) que la topographie.

Enfin, si le relief résulte notamment de l'érosion et du transport de matières par l'action mécanique de l'eau, dont le parcours dépend du relief, « l'interaction physique » entre relief et écoulement de l'eau est particulièrement forte dans le lit mineur $\mathcal{B}$.

Ce travail a pour vocation d'étudier uniquement les données de topographie.

2 Problèmes induits par un nombre importants de points

La démarche proposée pour la prise en compte des incertitudes de relief requiert l'échantillonnage aléatoire du processus $\mathcal{H}(.)$ conditionné par les observations, que nous appellerons processus conditionnel et que nous noterons $\mathcal{H}(.)|h(X)$, sur le maillage du modèle de simulation numérique hydraulique. Les points de ce maillage sont notés $x^{(I+j)}$ pour $1 \leq j \leq J$. Il faut noter que le vecteur $H \in \mathbb{R}^J$ formé par les valeurs du processus conditionnel aux J points du maillage, autrement dit formé par $\mathcal{H}(x^{(I+j)})|h(X)$ pour $1 \leq j \leq J$, est un vecteur aléatoire qui suit une loi gaussienne multivariée $\mathcal{N}(\mu, \Sigma)$ définie par son espérance μ et sa matrice de covariance Σ. Celles-

2. Voir `http://wikhydro.developpement-durable.gouv.fr/index.php/Utilisation_des_donn%C3%A9es_LIDAR_pour_la_directive_inondation`.

3. `https://www.ign.fr/institut/nos-activites/lidar-hd-une-couverture-nationale-dici-2025`

4. Une technologie LiDAR existe également pour la bathymétrie.

ci peuvent être théoriquement obtenues par les équations du krigeage, qui constitue en fait un préalable à toutes ces méthodes.Le krigeage a été exploité dès la fin des années 1980 pour modéliser la bathymétrie dans un environnement marin (Herzfeld, 1989). Cette approche géostatistique est liée à celle de la simulation gaussienne séquentielle (*sequential Gaussian simulation*, SGS), exploitée par Legleiter *et al.*, 2011 (Legleiter *et al.*, 2011) pour étudier l'impact de l'incertitude de la bathymétrie sur la simulation hydraulique d'une rivière. Nous n'avons pas identifié d'autres approches exploitée dans la littérature pour la modélisation de l'incertitude du relief.

Les méthodes de simulation conditionnelles sont toutes limitées par le nombre total de points à considérer dans le domaine spatial $\mathcal{D}$. Dans le cadre de la simulation hydraulique numérique, le nombre de données est de l'ordre de quelques millions. Il faut noter que leur mise en œuvre requiert toujours des calculs d'espérance et de covariance conditionnelles qui impliquent des formules de krigeage, donc la constitution et l'inversion de matrices de covariance de grande taille (terme de covariance dans les équations de prédiction de l'espérance et de la covariance conditionnée du Krigeage).

Cela a trois conséquences pratiques potentielles :

1. l'impossibilité de stocker en mémoire les matrices de covariance nécessaires,
2. un problème de conditionnement des matrices à inverser, donc de robustesse numérique (Ababou *et al.*, 1994),
3. un problème de coût et de temps CPU d'inversion de ces matrices.

Une pratique courante qui permet d'éviter ces écueils consiste à découper le domaine spatial d'étude en plusieurs zones qui se recouvrent et où le krigeage sera appliqué de manière indépendante (à part l'estimation de la dérive éventuellement), ou en associant à chaque point un voisinage de points à considérer pour conditionner le processus gaussien. Un intérêt de cette approche est de pouvoir adapter le découpage de manière à ce que l'hypothèse de fonctions de covariance (localement) stationnaires soit adaptée, alors que l'hypothèse d'une fonction de covariance (globale) stationnaire ne le serait pas. Cette approche pragmatique peut être exploitée dans notre contexte. En effet, un découpage naturel peut être considéré en fonction de l'occupation des sols. Un tel découpage est accessible via la base de données géographiques CORINE LAND COVER (`https://data.europa.eu/euodp/fr/data/dataset/DAT-214-en`)

3 Applications

Les travaux réalisés pourront être mis en place sur une portion de la Garonne comprise entre Tonneins, en aval de la confluence avec le Lot, et La Réole (limite de l'influence hydrodynamique de la marée), soit environ 50 km de rivière (cf. Figure ci-dessous).

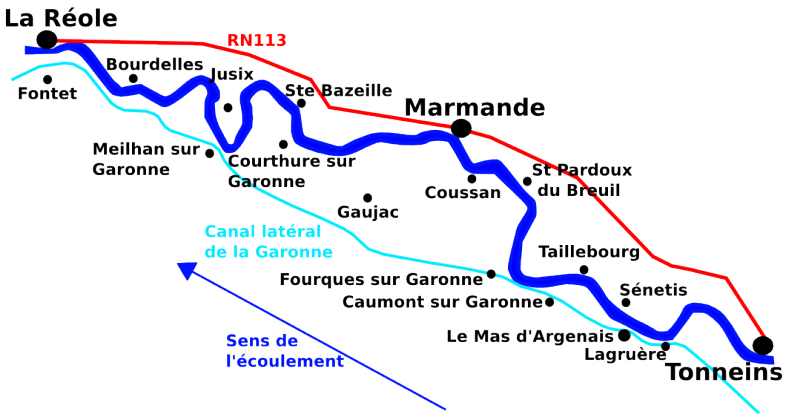


FIGURE 1 – Domaine d'étude

Cette zone ne comporte pas d'installations particulières mais présente l'intérêt d'être fortement aménagée (endiguement, déversoirs,...) pour protéger les riverains des crues de la Garonne.

4 Quelques indications pour guider ce travail

Le programme de travail suivant est proposé pour traiter les incertitudes de relief.

1. Identification des données IGN contenues dans l'emprise du modèle numérique de la Garonne et modélisation des incertitudes associées. Plusieurs modélisations de la structure de dépendance spatiale des erreurs qui entachent ces données pourront être construites étant donné le manque de connaissance sur ces dépendances.
2. Modélisation de l'incertitude de topographie en exploitant le découpage du domaine défini par la base de donnée d'exploitation des sols CORINE LAND COVER.
3. Exploitation des résultats issus de l'étape précédente pour effectuer une propagation d'incertitudes et une analyse de la sensibilité de la hauteur d'eau (via la génération aléatoire de reliefs par simulation séquentielle gaussienne). Une analyse de sensibilité, par estimation d'indices de Sobol, qui tienne compte du relief est attendue (Iooss et Ribatet, 2009; Saint-Geours, 2012).
4. Pour répondre aux problèmes indiqués à la section 2, outre l'exploitation d'un découpage préalable du domaine, l'utilisation de méthodes de krigeage adaptées à un grand nombre de données pourra être étudiée, notamment celles pro-

posées par Cressie et Johannesson (2008); Rullière *et al.* (2018), ou l'utilisation de $\mathcal{H}$-matrix comme dans le logiciel OpenTURNS (voir `openturns.org` et Gorshechnikova (2019)).

L'environnement informatique Python sera privilégié pour tous ces traitements.

Références

ABABOU, R., BAGTZOGLOU, A. C. et WOOD, E. F. (1994). On the condition number of covariance matrices in kriging, estimation, and simulation of random fields. *Mathematical Geology*, 26(1):99–133.

CRESSIE, N. et JOHANNESSON, G. (2008). Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 70(1):209–226.

GORSHECHNIKOVA, A. (2019). *Likelihood approximation and prediction for large spatial and spatio-temporal datasets using H-matrix approach*. Thèse de doctorat, Università degli Studi di Padova.

HERZFELD, U. C. (1989). Variography of submarine morphology : Problems of deregularization, and cartographical implications. *Mathematical Geology*, 21(7):693–713.

IOOSS, B. et RIBATET, M. (2009). Global sensitivity analysis of computer models with functional inputs. *Reliability Engineering and System Safety*, 94:1194—-1204.

LEGLEITER, C. J., KYRIAKIDIS, P. C., MCDONALD, R. R. et NELSON, J. M. (2011). Effects of uncertain topographic input data on two-dimensional flow modeling in a gravel-bed river. *Water Resources Research*, 47(3).

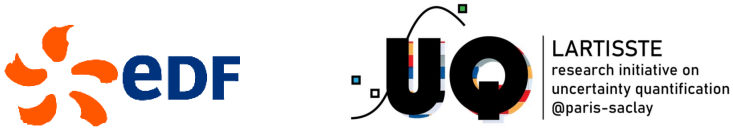
RGE ALTI (2018). RGE ALTI® Version 2.0 – Descriptif de contenu. Institut National de l'Information Géographique et Forestière. `https://geoservices.ign.fr/ressources_documentaires/Espace_documentaire/MODELES_3D/RGE_ALTI/DC_RGEALTI_2-0.pdf`.

RULLIÈRE, D., DURRANDE, N., BACHOC, F. et CHEVALIER, C. (2018). Nested kriging predictions for datasets with a large number of observations. *Statistics and Computing*, 28(4):849–867.

SAINT-GEOURS, N. (2012). *Analyse de sensibilité de modèles spatialisés - Application à l'analyse coût-bénéfice de projets de prévention des inondations*. Thèse de doctorat, Université Montpellier 2.

Utilisation des Generative Adversarial Networks pour l'analyse de sensibilité en gestion d'actifs industriels

7 octobre 2021



1 Contexte

L'analyse et l'optimisation du cycle de vie d'actifs industriels passe par la modélisation du système technico-économique qui est, pour les systèmes les plus complexes, simulé à l'aide d'algorithme de Monte-Carlo pour prendre en compte la stochasticité de la vie d'un actif physique, notamment les dates de défaillances. Un tel code de calcul peut s'avérer en temps de calcul, or certaines analyses réalisées par l'ingénierie peuvent nécessiter de nombreux appels au code de calcul. Une approche classique pour réduire les temps de calcul est de récourir à des métamodèles pour approximer la sortie du code de calcul. Si l'analyse porte sur une statistique de la sortie stochastique (moyenne, quantile...) on pourra utiliser des métamodèles connus pour être efficaces à approximer une sortie scalaire, même bruitée (processus gaussiens...). Cependant il existe des analyses nécessitant l'accès à l'ensemble de la distribution de la sortie du simulateur et pour circonvenir au problème de temps de calcul il est nécessaire d'utiliser un émulateur stochastique. Nous nous intéresserons ici à l'utilisation des Generative Adversarial Network [1].

2 Applications

EDF a développé plusieurs simulateurs pour la gestion d'actifs pour les différentes filières de production d'électricité (hydraulique, éolien, nucléaire...) afin d'apporter un

éclairage robuste aux décideurs dans les choix des stratégies d'investissements. Dans le cadre de la gestion d'actifs industriels, l'émulation de simulateurs stochastiques peut être utile pour plusieurs types d'analyses telles que l'optimisation des dates de maintenance ou l'analyse de sensibilité. Le stage se concentrera sur une application pour l'analyse de sensibilité d'un modèle pour la gestion d'une ferme éolienne offshore.

3 Quelques indications pour guider ce travail

Sur la base de travaux préliminaires réalisés par la R&D et après une étude bibliographique approfondie il conviendra d'implémenter et d'optimiser un GAN sur ce cas d'application à l'aide des librairies classiques de Machine Learning. L'étape suivante consistera à évaluer des indicateurs de sensibilité à l'aide de cet émulateur (par exemple des indices de Sobol adaptés au cas stochastique [2]) et comparer avec des implémantations utilisant d'autres métamodèles.

4 Recherche d'un candidat

Le candidat devra avoir les compétences en Mathématiques Appliquées et plus particulièrement en Machine Learning. Le candidat devra avoir un goût réel pour la programmation et maîtriser des langages bas niveau (C, C++...) ou haut niveau (Python, Java...). Une bonne connaissance de librairies de Machine Learning est également attendue.

5 Contact EDF

: Jérôme Lonchampt (jerome.lonchampt@edf.fr)

— Durée du stage : 6 mois.
— Date de début du stage : Printemps 2022.
— Durée hebdomadaire de travail : 35 heures.
— Lieu du stage : le site EDF-Lab Chatou de la R&D (6 quai Watier), 78400 CHATOU (accès RER A Rueil Malmaison) ;

Références

[1] Ian J. Goodfellow et al. Generative Adversarial Networks. 2014. NIPS
[2] Gildas Mazo. The Sobol method in sensitivity analysis for stochastic computer models. 2019. ⟨hal-02113448v1

Proposition de sujet de travaux étudiant (2021)

Techniques pour l'analyse de sensibilité et la propagation d'incertitudes pour systèmes multidisciplinaires

ONERA
THE FRENCH AEROSPACE LAB

LARTISSTE
research initiative on
uncertainty quantification
@paris-saclay

1 Contexte

Ce projet étudiant s'inscrit dans le cadre des travaux de l'ONERA en traitement d'incertitudes au sein des modèles numériques pour la sûreté et la conception, et plus largement dans celui de la construction du Groupement d'Intérêt Scientifique (GIS) LARTISSTE.

La prise en compte des incertitudes dans des études de conception de véhicules aérospatiaux soulèvent de nombreuses difficultés méthodologiques. En effet, la conception des systèmes aérospatiaux est couramment effectuée au moyen de processus multidisciplinaires qui couplent un certain nombre de codes de calculs modélisant différentes disciplines (propulsion, aérodynamique, structure, trajectoire, *etc.*). Un exemple classique de couplages interdisciplinaires pour véhicule aérospatial porte sur le couplage aérodynamique - structure (aérostructure - Figure 1). En effet, pour une aile d'avion, afin de simuler la structure de l'aile il est nécessaire d'obtenir les efforts aérodynamiques qui sont une sortie de la discipline aérodynamique mais pour calculer les efforts aérodynamiques il est nécessaire d'avoir la géométrie déformée qui est une sortie de la discipline structure.

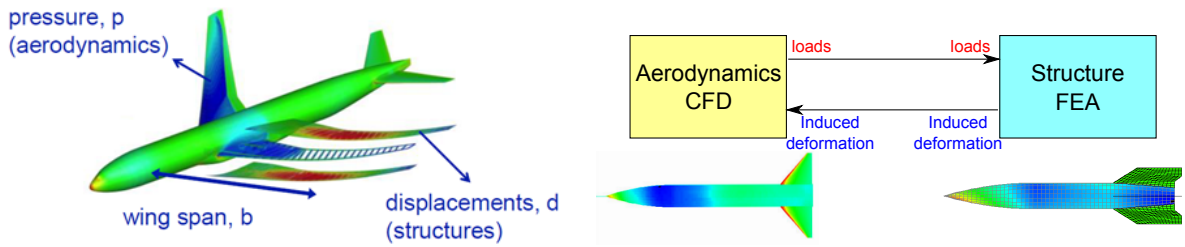


FIGURE 1 – Couplage aérostructure

La gestion de ces couplages interdisciplinaires fait appel à des méthodes itératives résolvant le système couplé. Celles-ci permettent de garantir la viabilité du système multi-physique à concevoir. Ainsi, les approches par point fixe (algorithmes Gauss-Seidel, Jacobi) sont souvent utilisées pour résoudre le couplage interdisciplinaire et calculer la valeur des couplages convergés entre les disciplines.

Ainsi, les méthodologies de propagation d'incertitudes pour systèmes multidisciplinaires sont souvent très gourmandes en temps de calculs car elles superposent la gestion des couplages interdisciplinaires et la propagation d'incertitudes (Balesdent et al., 2016; Brevault et al., 2020). Ainsi, un axe de recherche actif consiste à développer des approches de quantification d'incertitudes efficaces pour ce type de systèmes tout en maîtrisant le coût de calculs associé.

2 Descriptif de la problématique

Une stratégie pour limiter les temps de calculs consiste à faire appel à des méthodes d'analyse de sensibilité pour un système couplé. Celles-ci permettent ainsi de classer l'influence des paramètres incertains (qui peuvent être très nombreux) sur les performances du système multi-physique. Ces paramètres incertains proviennent de méconnaissances de modélisation impliqués dans les différentes disciplines (incertitudes portant sur la masse des structure, sur les performances propulsives, *etc*.) Cette analyse permet d'identifier les incertitudes les plus critiques qui pourraient nécessiter une meilleure caractérisation. Par ailleurs, elle permet de figer les incertitudes peu influentes et ainsi diminuer la dimension du problème étudié. Les analyses de sensibilité reposent sur différents indices de sensibilité (*e.g.*, Sobol', Shapley, HSIC) (Sobol, 2001; Da Veiga, 2015; Kucherenko and Song, 2016) et nécessitent en général le calcul d'intégrales multidimensionnelles et utilisent des approches de propagation d'incertitudes.

La difficulté ici repose sur l'aspect multidisciplinaire et les couplages entre différentes disciplines. En effet, l'analyse de sensibilité doit être traitée au regard des performances globales du système couplé étant donné qu'une variable peut-être très influente sur une discipline prise isolément mais n'avoir que peu d'influence sur le système couplé (et réciproquement). Par conséquent, les méthodes développées prendront en compte cette problématique. Malheureusement, la plupart des approches existantes ne tiennent pas compte de la spécificité des processus multidisciplinaires et de la présence de couplage entre différents codes de calculs dans l'analyse de sensibilité.

3 Descriptif de la démarche du projet étudiant

Pour cela, les étudiants prendront en main la problématique de la propagation d'incertitudes pour un système couplé. Dans un premier temps, une approche "couplée" de référence sera mise en oeuvre combinant une technique de point fixe pour la résolution des couplages interdisciplinaires et une analyse de sensibilité de type Sobol'. Les étudiants se familiariseront avec ces techniques sur des cas analytiques de référence (par exemple le cas classique Sellar et ses dérivés (Sellar et al., 1996)).

Une fois cette approche de référence mise en oeuvre, on visera à développer une approche dite "hybride - découplée" afin de supprimer la résolution systématique des couplages par point fixe afin de définir la sensibilité des performances globales du système à partir des sensibilités de chacune des disciplines prises séparément. Ainsi, on tâchera de limiter les appels aux méthodes permettant de résoudre le système multidisciplinaire couplé (qui engendrent des coûts de calculs très importants) et on maîtrisera l'approximation associée à la non résolution systématique du système couplé.

On pourra pour cela explorer différents types de satisfaction des couplages interdisciplinaires en présence d'incertitudes (*e.g.*, convergence en loi, convergence pour chaque réalisation des incertitudes) et associer des techniques de propagation d'incertitudes pour système multidisciplinaire (Balesdent et al., 2016; Brevault et al., 2020). Par ailleurs, on pourra analyser divers indices de sensibilité (*e.g.*, Sobol', Shapley, HSIC) en fonction des besoins associés aux propagation d'incertitudes.

Pour ce faire, projet étudiant se déroulera de la manière suivante :

- État de l'art sur les techniques de propagation d'incertitudes et d'analyse de sensibilité sur systèmes multidisciplinaires,
- Mise en place d'une stratégie "couplée" de référence d'analyse de sensibilité pour systèmes multidisciplinaires,
- Développement d'une stratégie "hybride - découplée" d'analyse de sensibilité pour systèmes multidisciplinaires,
- Implémentation des approches élaborées sur des cas tests analytiques et physiques représentatifs de véhicules aérospatiaux.

4 Compétences souhaitées

Outre présenter des compétences fortes en probabilité, les étudiant(e)s devront maîtriser Python. Les développements mis en oeuvre s'appuyeront en particulier sur les librairies Python telles que OpenTURNS (développé par EDF, Airbus, ONERA, Phimeca, IMACS) (Baudin et al., 2015) et OpenMDAO (développé par la NASA) (Gray et al., 2019).

5 Suivi industriel

Mathieu Balesent et Loïc Brevault (ingénieurs-chercheurs à l'ONERA) suivront ce projet étudiant.

Références

Balesdent, M., Brevault, L., Price, N. B., Defoort, S., Le Riche, R., Kim, N.-H., Haftka, R. T., and Bérend, N. (2016). Advanced space vehicle design taking into account multidisciplinary couplings and mixed epistemic/aleatory uncertainties. In *Space Engineering*, pages 1–48. Springer.

Baudin, M., Dutfoy, A., Iooss, B., and Popelin, A.-L. (2015). Open turns : An industrial software for uncertainty quantification in simulation. *arXiv preprint arXiv :1501.05242*.

Brevault, L., Balesdent, M., Morio, J., et al. (2020). *Aerospace System Analysis and Optimization in Uncertainty*. Springer.

Da Veiga, S. (2015). Global sensitivity analysis with dependence measures. *Journal of Statistical Computation and Simulation*, 85(7) :1283–1305.

Gray, J. S., Hwang, J. T., Martins, J. R., Moore, K. T., and Naylor, B. A. (2019). Openmdao : An open-source framework for multidisciplinary design, analysis, and optimization. *Structural and Multidisciplinary Optimization*, 59(4) :1075–1104.

Kucherenko, S. and Song, S. (2016). Derivative-based global sensitivity measures and their link with sobol'sensitivity indices. In *Monte Carlo and Quasi-Monte Carlo Methods*, pages 455–469. Springer.

Sellar, R., Batill, S., and Renaud, J. (1996). Response surface based, concurrent subspace optimization for multidisciplinary system design. In *34th aerospace sciences meeting and exhibit*, page 714.

Sobol, I. M. (2001). Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. *Mathematics and computers in simulation*, 55(1-3) :271–280.

Real Fluids Modeling approach using Deep Learning

Introduction

Accurate and robust two-phase flow models are required to predict phase change and mixing processes due to injections of various hydrogen or hydrogen-based fuels in sub-transcritical conditions. These Real Fluid Models (RFM) are based on a two-phase fully compressible four-equation model under mechanical and thermal equilibrium assumptions and closed by thermodynamic equilibrium computations.

The thermodynamic equilibrium computations (also known as “flash” computations) are used to determine the stable phases at equilibrium (liquid, aqueous, gas or a combination of these) and the distribution of the species in each stable phase. Unfortunately, these flash calculations involve complex non-linear physics and require long computation times, slowing down the overall RFM resolution. Therefore, a common strategy is to pre-tabulate the solution of these equations using the IFPEN thermodynamics library *Carnot*, and simply interpolate the resulting table during the resolution of the RFM model. This approach is however limited to a small dimension of the inputs (and hence of the number of chemical species considered) as the table becomes impractically large due to the curse of dimensionality.

The objective of the project is to extend the capacity of the RFM model to handle mixtures with a larger number of species. We will use deep learning algorithms to overcome the limitation of the current tabulation approach due to the growing dimensionality of the inputs with the number of species. Since exact numerical calculations are available on-demand using the *Carnot* library, we are interested in using active learning methods, iteratively adding specific new data to the training database selected in order to reduce the training error.

Methodology

I – Learning model with fixed database

First, we focus on the learning algorithm: Given a fluid characterized by its pressure, temperature and composition (mass fractions of the chemical species in the fluid), the objective is to classify the various phases present at equilibrium (or alternatively predict their volume fraction), and determine the molar fractions of each species in each phase.

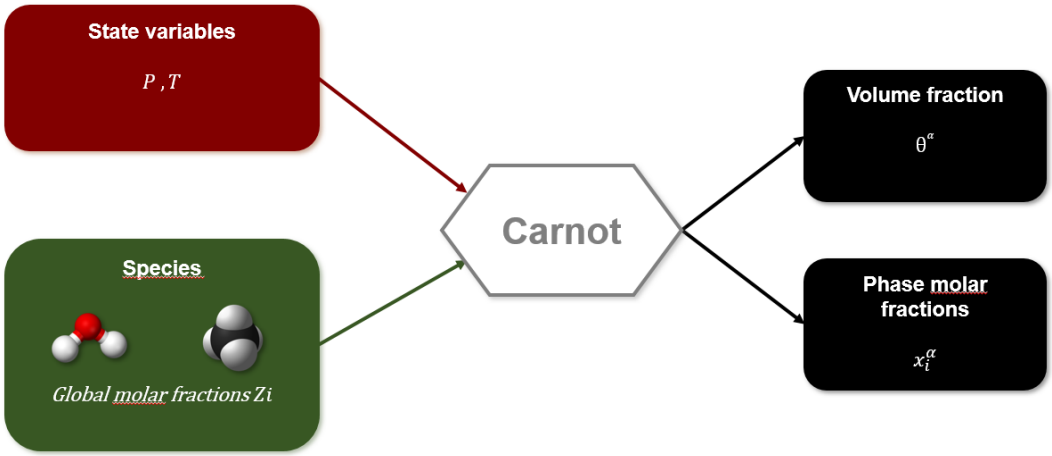


Figure 1: Schematic of the flash calculations performed by Carnot.

The students will use a standard Latin Hypercube Sampling (LHS) method to determine the input data that will be entered in *Carnot* and create the database.

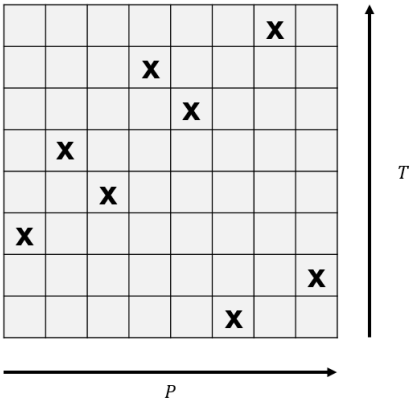


Figure 2: The LHS algorithm samples without replacement among a range of input variables. The objective is to "fill the space'' (illustration for two input variables P and T).

Once the database is created, the students will train a deep learning algorithm to classify the equilibrium phases and predict the molar fractions of chemical species. The results will be compared with a traditional tabular approach in term of accuracy and execution time. We are also interested in determining the best neural network architecture for a given problem size (number of species) as well as the database size influence on the results.

II – Active Learning

Once a benchmark learning algorithm has been developed, the students will explore active learning methods to improve the results. The idea is to characterize the region in input space where the model performs worse, use the Carnot library to over-sample new data in this specific region, and retrain the model with the additional data.

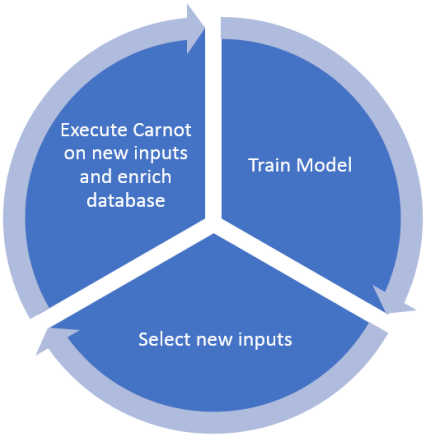


Figure 3: Illustration of the active learning process.

This approach is especially important as only a small region in the input space corresponds to large non-linearities in the response function. Therefore, we expect to be able to obtain a more accurate model with only a fraction of the data (and hence the training time).

III – Transfer Learning

With the current approach, when performing RFM on a new fluid with different chemical species, we need to re-train the model from scratch (including possibly an expensive neural architecture search). However, problems sharing a common set of chemical species usually share common a physical behavior and hence a common response function. Therefore, the students will explore transfer learning approaches where some weights of a learned network will be transferred to a network that will be trained on a problem with a different database corresponding to different species.

Test cases

The researchers at IFPEN will provide the students with a variety of test cases of increasing complexity. We will provide a simple test case consisting of a binary mixture (2 chemical species) and compare the results with a tabulation approach. We will then move to test cases with complex fluids (> 10 chemical species) where the physics is highly non-linear and the tabulation approach fails.

IFPEN contacts

The students will be able to interact with several researchers at IFPEN in the fields of thermodynamics and data science. We will be available for in-person meetings on a regular basis (at least once a month) and frequent video meetings.

- Chawki Habchi - chawki.habchi@ifpen.fr (Thermodynamics)
- Julien Bohbot - Julien.bohbot@ifpen.fr (Thermodynamics)
- Thibault Faney – thibault.faney@ifpen.fr (Data Science)