

Rapport de Projet

Analyse de Sensibilité et Métamodélisation : Modèle Borehole

G. Aloui

Polytech Nice Sophia - Université Côte d'Azur

27 Février 2026

Table des matières

1	Introduction et Fonctions du Modèle	2
2	Mise en place	2
3	Propagation d'incertitudes par Monte Carlo	2
3.1	Évaluation empirique	2
3.2	Inégalité de Jensen	3
3.3	Calcul du quantile à 95% et intervalle de confiance (Bootstrap)	3
3.4	Probabilité de dépassement de seuil par Monte Carlo	4
4	Analyse de Sensibilité	5
4.1	Indices basés sur la régression linéaire	5
4.2	Calcul et interprétation des indices de Sobol'	6
5	Ajustement et utilisation d'un métamodèle de krigeage	8
5.1	Ajustement, validation (Q^2) et comparaison	8
5.2	Calcul des indices de Sobol' à l'aide du métamodèle	8

1 Introduction et Fonctions du Modèle

Ce document présente l'étude du modèle de forage borehole, évaluant le débit d'eau en fonction de 13 variables d'entrée incertaines.

Le débit d'eau y (en m^3/an) traversant le forage est modélisé par l'équation analytique suivante :

$$y = \frac{2\pi T_u (H_u - H_l)}{\ln\left(\frac{r}{r_w}\right) \left[1 + \frac{2LT_u}{\ln\left(\frac{r}{r_w}\right)r_w^2 K_w} + \frac{T_u}{T_l} \right]} \quad (1)$$

Avec les variables physiques définies ainsi :

- r_w : Rayon du forage (m)
- r : Rayon d'influence (m)
- T_u : Transmissivité de l'aquifère supérieur (m^2/an)
- H_u : Hauteur potentiométrique de l'aquifère supérieur (m)
- T_l : Transmissivité de l'aquifère inférieur (m^2/an)
- H_l : Hauteur potentiométrique de l'aquifère inférieur (m)
- L : Longueur du forage (m)
- K_w : Conductivité hydraulique du forage (m/an)

Remarque importante : Il est à noter que dans le script R fourni, les variables relatives à l'aquifère moyen (T_{um} , T_{lm} , H_{um} , H_{lm}) sont écrasées par la modélisation de la résistance hydraulique du forage. Elles sont donc rendues **inactives** et n'ont aucun impact sur le résultat final. Tout le reste de l'analyse sera basé sur ce constat.

2 Mise en place

Pour cette première étape d'appropriation, nous évaluons la fonction aux bornes de son domaine de définition ainsi qu'à son point central. La variable aléatoire du rayon d'influence r suit une loi lognormale. L'espérance mathématique se calcule via la formule :

$$\mathbb{E}(r) = \exp\left(\mu + \frac{\sigma^2}{2}\right) \quad (2)$$

Avec $\mu = 7.71$ et $\sigma = 1.0056$, nous obtenons environ 3698.24 m.

3 Propagation d'incertitudes par Monte Carlo

3.1 Évaluation empirique

On effectue 1000 échantillons de Monte Carlo pour calculer la moyenne et la variance de la sortie. Nous visualisons ensuite les résultats dans un histogramme.

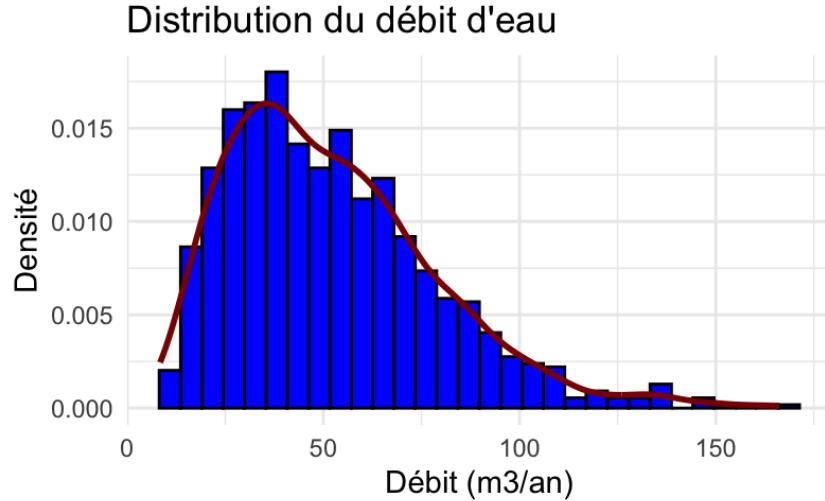


FIGURE 1 – Distribution du débit d'eau par Monte Carlo ($N = 1000$)

3.2 Inégalité de Jensen

Afin d'étudier la linéarité du modèle, nous avons comparé la moyenne empirique obtenue par l'échantillon de Monte Carlo avec la réponse du modèle calculée directement sur les valeurs moyennes des variables d'entrée.

Les résultats sont synthétisés dans le tableau ci-dessous :

Méthode d'évaluation	Débit estimé (m ³ /an)
Réponse à la moyenne des entrées : $f(\mathbb{E}[X])$	50.33
Moyenne empirique (Monte Carlo) : $\mathbb{E}[f(X)]$	53.64

TABLE 1 – Comparaison des moyennes illustrant l'inégalité de Jensen

On remarque très clairement que la moyenne de la réponse (53.64) est supérieure à la réponse de la moyenne (50.33). Ce phénomène est confirmé mathématiquement par l'inégalité de Jensen. Pour une fonction bornée non-linéaire f , nous avons :

$$\mathbb{E}[f(X)] \neq f(\mathbb{E}[X]) \quad (3)$$

Puisqu'on n'a pas d'égalité entre ces deux quantités, cela démontre formellement la **non-linéarité** du modèle borehole.

3.3 Calcul du quantile à 95% et intervalle de confiance (Bootstrap)

À partir de notre échantillon de Monte Carlo initial (de taille $N = 1000$), nous avons estimé le quantile d'ordre 95% empirique, ainsi que son intervalle de confiance via la méthode de rééchantillonnage Bootstrap ($B = 10\,000$ itérations) :

- Quantile d'ordre 95% estimé : $100.75\text{ m}^3/\text{an}$
- Intervalle de confiance (95%) : $[95.88 ; 106.61]\text{ m}^3/\text{an}$

Interprétation et détails de la méthode :

1. **Signification du quantile à 95% :** Dans le contexte de notre modèle, le quantile à 95% représente une valeur seuil de débit d'eau. Concrètement, cela signifie que selon notre modèle et les incertitudes de nos variables d'entrée, il y a 95% de chances que le débit du forage soit inférieur ou égal à cette valeur (et inversement, seulement 5% de chances qu'il la dépasse). C'est un indicateur crucial pour l'évaluation des risques extrêmes.
2. **Principe du rééchantillonnage Bootstrap :** L'estimation de ce quantile (calculée sur notre unique échantillon Monte Carlo de taille $N = 1000$) est soumise à une incertitude statistique. Pour quantifier cette incertitude sans avoir à relancer le modèle physique `borehole()` des milliers de fois (ce qui serait très coûteux en temps de calcul dans un cas réel), nous utilisons la méthode du **Bootstrap non-paramétrique**.
3. **La démarche algorithmique étape par étape :**
 - **Tirage avec remise :** À partir de notre échantillon initial de 1000 débits, nous "piochons" aléatoirement 1000 valeurs *avec remise*. Certaines valeurs de l'échantillon initial seront sélectionnées plusieurs fois, d'autres ignorées. Cela crée un "nouvel" échantillon virtuel.
 - **Calcul itératif :** Sur ce nouvel échantillon, nous recalculons le quantile à 95%.
 - **Répétition :** Nous répétons cette opération un grand nombre de fois ($B = 10\,000$ fois) pour obtenir non plus une seule estimation du quantile, mais une véritable *distribution statistique* de ce quantile.
 - **Extraction de l'intervalle :** Enfin, nous trions ces 10 000 valeurs de quantiles simulés et nous prenons les valeurs situées à 2.5% (borne inférieure) et 97.5% (borne supérieure). L'écart entre ces deux bornes nous donne notre intervalle de confiance à 95%, garantissant la robustesse de notre estimation initiale.

3.4 Probabilité de dépassement de seuil par Monte Carlo

L'objectif est d'estimer la probabilité $p = \mathbb{P}(y > 250 \text{ m}^3/\text{an})$. L'estimateur empirique de cette probabilité est :

$$\hat{p} = \frac{1}{N} \sum_{i=1}^N \mathbf{1}_{\{y_i > 250\}} \quad (4)$$

Nous fixons une condition d'arrêt basée sur l'erreur relative $\leq 10\%$.

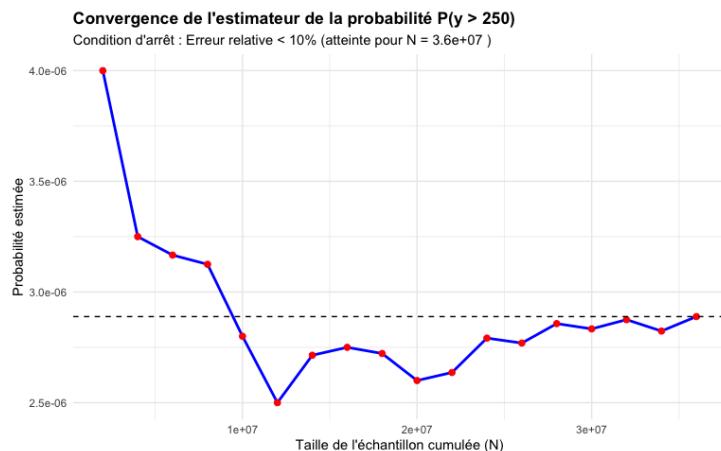


FIGURE 2 – Convergence de l'estimateur de la probabilité $P(y > 250)$

Résultats de l'algorithme à convergence :

- Taille d'échantillon totale nécessaire (N) : 3.6×10^7 simulations
- Probabilité estimée $\mathbb{P}(y > 250)$: 2.888889×10^{-6}
- Erreur relative finale atteinte : 9.81%

4 Analyse de Sensibilité

4.1 Indices basés sur la régression linéaire

Nous étudions d'abord la linéarité des relations avec un échantillon de taille $N = 500$ via des nuages de points (scatterplots).

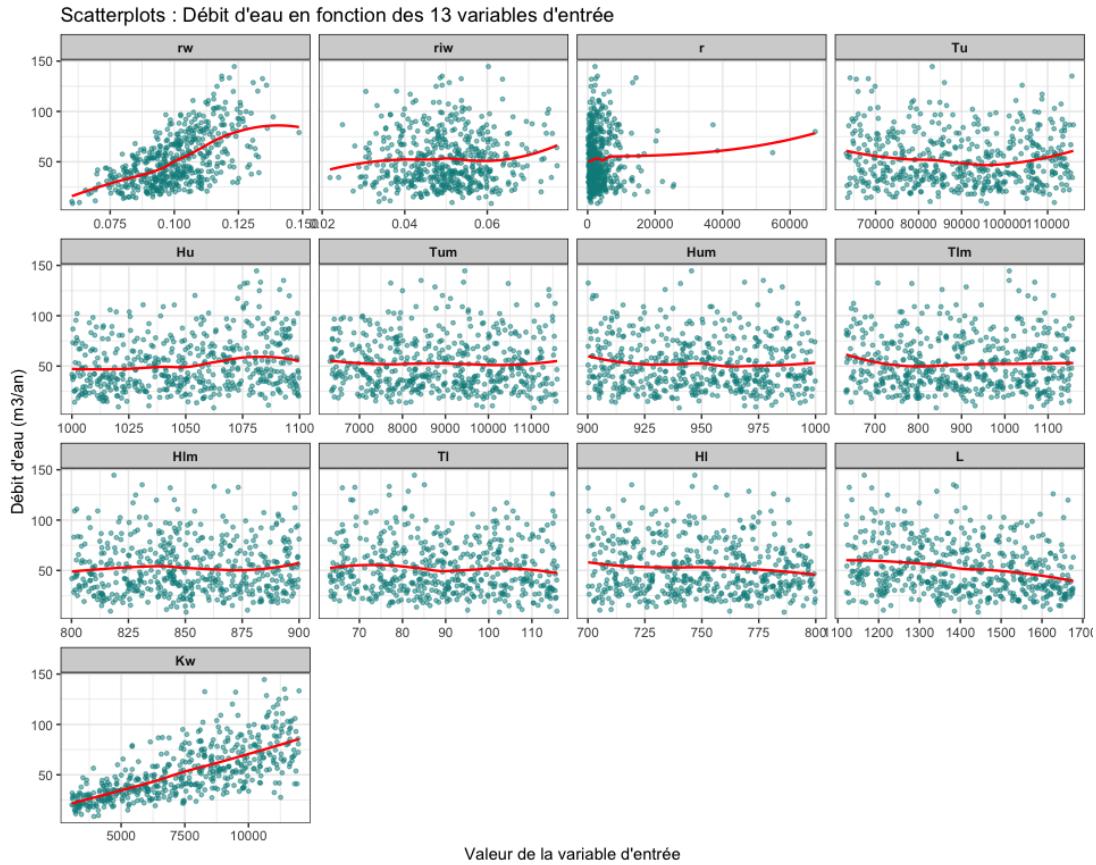


FIGURE 3 – Scatterplots : Débit d'eau en fonction des 13 variables d'entrée

Interprétation visuelle des Scatterplots :

L'analyse des nuages de points et de leurs courbes de tendance (en rouge) nous permet de tirer plusieurs conclusions préliminaires sur le comportement du modèle :

1. **Variables très influentes et corrélations positives :** Les variables **rw** (rayon du forage) et **Kw** (conductivité hydraulique du forage) se détachent très nettement. On observe une forte tendance croissante : plus ces paramètres augmentent, plus le débit d'eau augmente. De plus, la courbure marquée pour **rw** indique que cette relation est **non-linéaire**.

2. **Variables faiblement influentes** : Certaines variables comme L (longueur du forage) montrent une très légère tendance (légèrement décroissante pour L), mais la dispersion des points reste très large, ce qui suggère une influence de premier ordre modérée.
3. **Confirmation visuelle des variables inactives** : C'est l'observation la plus importante. Les courbes de tendance pour les variables de l'aquifère moyen (T_{um} , Hum , T_{lm} , H_{lm}) sont **parfaitement plates**. Cela confirme visuellement notre analyse analytique précédente : ces variables sont écrasées dans le script R de la fonction et n'ont strictement aucun impact sur le débit de sortie.
4. **Limites de cette représentation** : Bien que ces graphiques soient utiles pour repérer les effets marginaux principaux (effets de premier ordre), ils ne permettent pas de visualiser les **interactions** entre les variables (par exemple, l'effet combiné de rw et L).

Conclusion de l'analyse visuelle : La forte non-linéarité observée (notamment sur rw) justifie le calcul du R^2 de la régression linéaire globale. Si ce dernier est faible, les indices SRC basés sur cette régression seront insuffisants, ce qui nous amènera naturellement à utiliser la méthode de décomposition de la variance (Indices de Sobol).

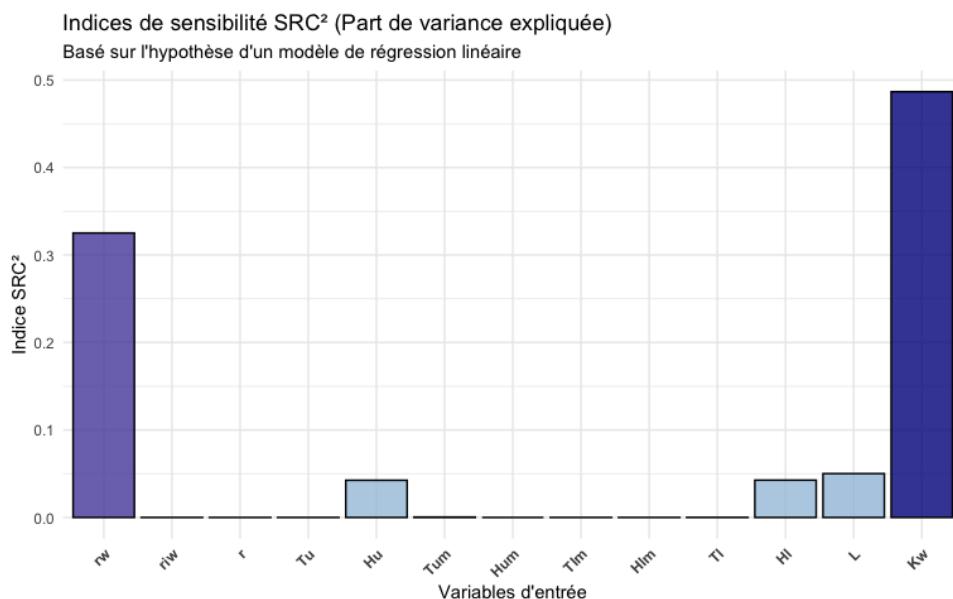


FIGURE 4 – Indices de sensibilité SRC² (Part de variance expliquée)

Bien que ce graphique identifie K_w et r_w comme ayant un fort effet proportionnel, l'indice SRC^2 est aveugle aux **interactions** entre les variables et aux effets fortement non-linéaires. C'est pourquoi nous devons passer aux indices de Sobol'.

4.2 Calcul et interprétation des indices de Sobol'

Nous utilisons l'estimateur de **Martinez** (`sobolmartinez()`) avec $N = 10\,000$ pour capturer l'ensemble des effets (linéaires, non-linéaires et interactions).

Indices de Sobol' du 1er Ordre et Totaux (Est Échantillon N = 10000 | Intervalles de confiance à 95%

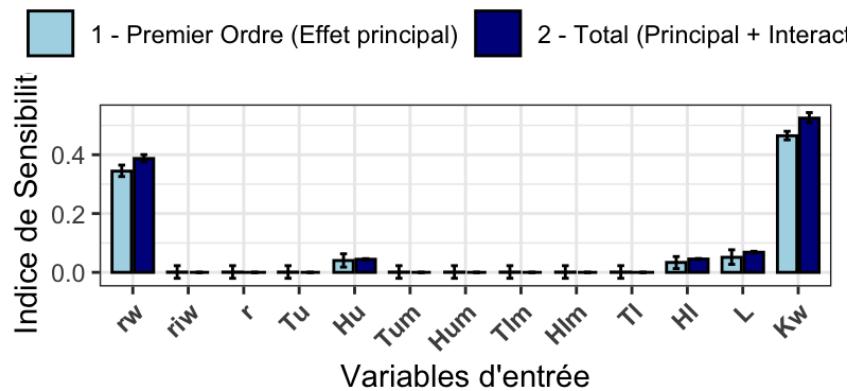


FIGURE 5 – Indices de Sobol' du 1er Ordre et Totaux (Estimateur de Martinez)

Synthèse de l'interprétation quantitative :

Afin d'aller plus loin, nous appliquons les outils de diagnostic quantitatifs :

Variable	S_i (1er ordre)	S_Ti (Total)	Interaction Pure	Statut
Kw	0.465	0.525	0.060	Influente
rw	0.344	0.387	0.043	Influente
L	0.051	0.068	0.017	Influente
Hu	0.040	0.045	0.004	Influente
Hl	0.034	0.045	0.012	Influente
riw	0.001	0.000	-0.001	Négligeable
r	0.001	0.000	-0.001	Négligeable
Tu	0.001	0.000	-0.001	Négligeable
Tum	0.001	0.000	-0.001	Négligeable
Hum	0.001	0.000	-0.001	Négligeable
Tlm	0.001	0.000	-0.001	Négligeable
Hlm	0.001	0.000	-0.001	Négligeable
Tl	0.001	0.000	-0.001	Négligeable

TABLE 2 – Diagnostic quantitatif des indices de Sobol'

La somme des indices de premier ordre s'élève à **0.942**, prouvant que le modèle est majoritairement additif. Toutefois, **5.8 % de la variance** est pilotée par des interactions (surtout entre K_w et r_w). Le modèle de forage peut donc être mathématiquement réduit à 5 variables actives influentes sans perte d'information majeure.

5 Ajustement et utilisation d'un métamodèle de krigage

5.1 Ajustement, validation (Q^2) et comparaison

Nous ajustons un métamodèle basé sur les processus gaussiens (Krigage) sur un échantillon d'apprentissage de taille $N = 500$, avec un noyau de covariance **Matérn 5/2**. La validation est réalisée par Leave-One-Out (LOO).

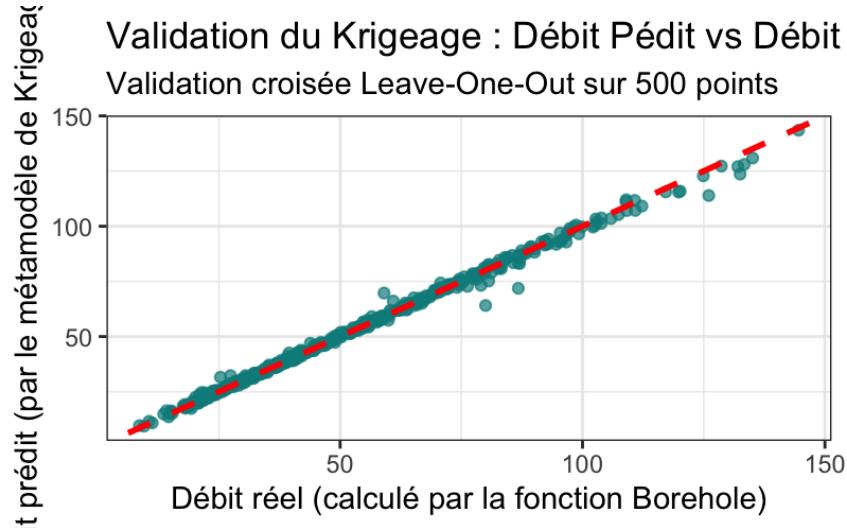


FIGURE 6 – Validation du Krigage : Débit Prédit vs Débit Réel

Améliore-t-il le modèle de régression linéaire ? Oui, magistralement. La régression linéaire classique obtenait un R^2 de **0.9419** (ce qui correspond exactement à la part additive trouvée par Sobol), ignorant ainsi les interactions. Le métamodèle de krigage atteint un coefficient de prédictivité Q^2 de **0.9952**. Il a implicitement "appris" les interactions entre les paramètres.

5.2 Calcul des indices de Sobol' à l'aide du métamodèle

Puisque le Krigage a un Q^2 exceptionnel, nous l'utilisons pour remplacer le modèle physique lourd et recalculer les indices de Sobol' sur 10 000 nouveaux points via la fonction `predict.km()`.

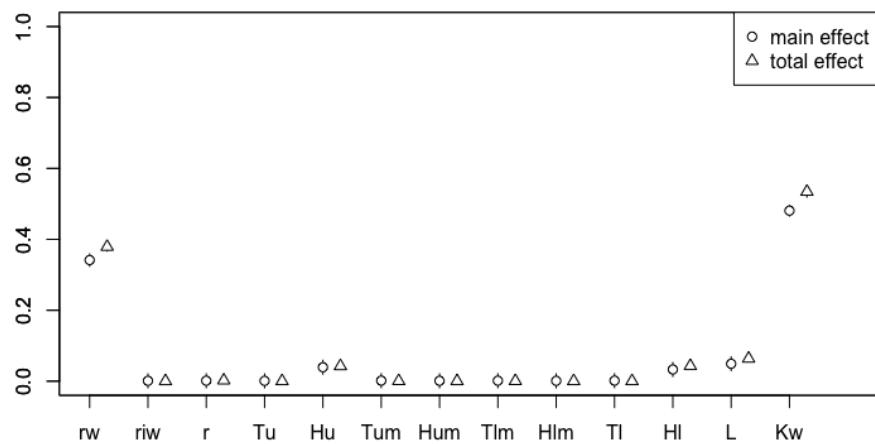


FIGURE 7 – Indices de Sobol' estimés sur le métamodèle de Krigeage

Conclusion finale : Le métamodèle reproduit virtuellement à l'identique les résultats du modèle physique d'origine (même hiérarchie, même détection des variables inactives), validant l'utilisation du Krigeage pour réaliser des analyses de sensibilité lourdes à un coût de calcul drastiquement réduit.