Reconstruction des concentrations des matières en suspension par apprentissage profond et prédiction sous scénarios de changement climatique

A Deep Learning approach for reconstructing suspended sediment load and forecasting under various climate change scenarios

Taha Hamadene[1,2], Valérie Nicoulaud-Gouin[1], Hugo Lepage[1], Mitra Fouladirad[2]

(1) Institut de Radioprotection et de Sûreté Nucléaire, PSE-ENV, STAAR/LRTA, BP 3, 13115 Saint Paul Lez Durance, France
(2) M2P2 Aix Marseille Univ, CNRS, Centrale Med, M2P2, Marseille, France

RÉSUMÉ

Ces dernières années, de nombreuses études ont démontré la capacité des modèles d'apprentissage automatique (Machine Learning, ML) et d'apprentissage profond (Deep Learning, DL) à prédire les concentrations des matières en suspension (CMES). Cependant, ces approches nécessitent des mesures de CMES, limitant leur application aux rivières instrumentées. Dans le cadre de l'Observatoire des Sédiments du Rhône (OSR), nous proposons une nouvelle approche exploitant uniquement les données de débits et de précipitations collectées à différentes stations dans un bassin versant. Cette approche repose sur un modèle DL combinant une couche Convolution (CNN) et une couche Long Short-Term Memory (LSTM), capturant les caractéristiques spatiales et temporelles des variables explicatives. Nous appliquons ce modèle aux affluents du Rhône et comparons ses performances avec celles du modèle ML des forêts aléatoires (RandomForest, RF) et de l'approche empirique SiRCA (Simplified Rating Curve Approach) utilisée dans le cadre de l'OSR. Le modèle CNN-LSTM donne les meilleurs résultats, surpassant les autres modèles dans la plupart des affluents, avec des R^2 proches de 1. Le modèle RF surpasse également SiRCA, avec des R^2 allant de 0,8 à 0,94, contre des R^2 plus faibles de 0,2 à 0,7 pour SiRCA. La robustesse de la méthodologie nous permet d'utiliser le modèle pour évaluer l'impact du changement climatique sur la dynamique des sédiments en utilisant les projections futures.

ABSTRACT

In recent years, numerous studies have demonstrated the ability of machine learning (ML) and deep learning (DL) models to accurately predict suspended sediment load (SSL). As part of the Rhône Sediment Observatory (OSR), we propose a novel approach that utilizes streamflow and rainfall data collected from different stations for predicting SSL on each Rhône tributaries. This approach is based on a DL model that combines a Convolutional Layer (CNN) with a Long Short-Term Memory (LSTM) layer, enabling it to capture both spatial and temporal features of input variables. To highlight the advantages of using a complex model, we evaluate its performance against the random forest (RF) model and an empirical approach (Simplified Rating Curve Approach, SiRCA). The CNN-LSTM model shows the best results, outperforming other models in most tributaries, with R^2 values close to 1. The RF model also outperforms SiRCA, with R^2 values ranging from 0.8 to 0.94, compared to a lower range of R^2 from 0.2 to 0.7 for SiRCA. The robustness of our methodology allows us to use it to assess the impact of climate change on sediment dynamics by applying future streamflow and rainfall projections based on various climate scenarios from the Explore2 project.

MOTS CLÉS

Sédiment, apprentissage automatique, projection climatique, Rhône

KEYWORDS

Sediment, machine learning, climate change, Rhône River

1 INTRODUCTION

Etudier les matières en suspension (MES) transportées par les cours d'eau est cruciale car elles sont liées à la pollution du milieu aquatique, à la qualité de l'eau et à la dégradation des écosystèmes (Khabat Khosravi, 2022). Connaitre sa quantité joue donc un rôle important dans la gestion des rivières (Bibhuti Sahoo, 2023) et il n'est pas rare d'avoir recours à la modélisation pour connaitre ce paramètre car l'instrumentation des cours d'eau est parfois difficile à mettre en place.

Depuis plusieurs décennies, de nombreuses études ont utilisé des approches empiriques (Sediment Rating Curve, SRC) pour exploiter la relation entre le débit et la concentration de MES (CMES) (Crawford, 1991) (Mahrez Sadaoui, 2016) (Bárbara M. Jung, 2020). Le débit est en effet la variable la plus influente sur la CMES. Cependant, ces modèles empiriques manquent souvent de précision en raison de leur structure simpliste car une valeur de débit n'est associée qu'à une seule CMES, ne permettant ainsi pas de reconstruire les hystéresis et les systèmes complexes où l'origine de l'eau et de la matière est diverse.

Plus récemment, les modèles d'apprentissage automatiques ont été largement adoptés dans divers domaines, y compris l'hydrologie. Par conséquent, de nombreux chercheurs ont appliqué diverses méthodes d'apprentissage automatique (Deepak Gupta, 2021) démontrant que les modèles de ML surpassent les approches empiriques traditionnelles.

Les réseaux de neurones récurrents (RNN), dont les modèles Long Short-Term Memory (LSTM) et Gated Recurrent Unit (GRU) ont également de nombreuses applications dans la prédiction des CMES De plus, les réseaux de neurones convolutionnels (CNN), sont de plus en plus appliqués aux séries temporelles car ils ont la capacité d'extraire des caractéristiques implicites dans les données. En combinant les CNN avec des réseaux RNN, on exploite les forces des deux modèles : L'extraction des caractéristiques implicites et la prise en compte des dépendances temporelles. Néanmoins, l'application de ces modèles pour prédire la CMES sur le long terme reste peu étudiée car l'utilisation de l'historique de la CMES comme données d'entrée ne permet pas de faire des prévisions lointaines dû à l'accumulation des erreurs.

Pour pallier ce problème et pouvoir prédire sur du long terme, nous avons développé une méthode basée uniquement sur des observations de débit et de précipitation recueillies à partir de plusieurs stations distantes tout au long de la rivière pour modéliser la CMES. A la différence des autres études, nous développons un modèle CNN-LSTM qui n'utilise que les valeurs retardées des variables explicatives, c'est-à-dire la prise en compte du temps de transit de l'amont vers l'aval établie par une analyse de corrélation, et qui ne prend pas en compte l'historique de la CMES.

2 METHODOLOGIE ET DONNEES

2.1 Variables environnementales

Ce travail est réalisé dans le cadre de l'Observatoire des Sédiments du Rhône et bénéficie ainsi d'une importante base de données de mesure du débit et de la CMES les principaux affluents du Rhône (Tableau 1) (Lepage et al., 2022). Une partie des données de débit et les données de précipitation sont également issues dans la banque de données HydroPortail, édité par le Service Central d'Hydrométéorologie et d'Appui à la Prévision des Inondations (SCHAPI) en France.

Affluents	Années	N observation	Variables Explicatives
Drôme	2019-2023	30361	4 Débits
Durance	2014-2023	39920	4 Débits, 3 Précipitations
Ain	2012-2015	3064	3 Débits
Saône	2014-2023	47361	4 Débits, 3 Précipitations
Bourbre	2011-2013	17679	4 Débits
Isère	2019-2023	29761	4 Débits, 3 Précipitations

Tableau 1 : Description des données

2.2 Modélisation

Dans cette étude, nous utilisons trois modèles différents :

- Le modèle empirique SiRCA (Simplified Rating Curve Approach) consiste à déterminer un point de rupture dans la relation entre la CMES et le débit, permettant d'obtenir deux équations de régression différentes. C'est la méthode actuellement utilisée dans l'OSR.
- Le modèle d'apprentissage automatique RF est un ensemble d'arbres de décision, où chaque arbre est construit sur une version bootstrap de l'ensemble de données originales. La prédiction finale du modèle est simplement la valeur moyenne des prédictions de chaque arbre,
- Le modèle d'apprentissage profond CNN-LSTM utilisé se compose d'une couche de convolution qui applique des filtres aux données d'entrée pour extraire des caractéristiques importantes et d'une couche LSTM qui permet de traiter les tâches sur de longues séquences en maintenant les dépendances temporelles sur de plus longues périodes.

Le coefficient de détermination (R^2) et la racine l'erreur quadratique moyenne (RMSE) ont été utilisées comme indicateurs de performance de nos modèles.

L'analyse de sensibilité, c'est-à-dire l'impact des variables explicatives (débit et pluviométrie des différentes stations), a été réalisée avec 3 méthodes différentes afin d'avoir une évaluation diversifiée de l'influence de ces variables sur le comportement de la CMES :

- La diminution moyenne d'impureté (Mean Decrease in Impurity, MDI) :la méthode quantifie l'importance des variables en mesurant dans quelle mesure chaque variable contribue à réduire l'impureté des nœuds des arbres de décision qui composent la forêt.
- Les valeurs de Shapley : Pour chaque variable, la méthode considère toutes combinaisons possibles des variables et mesure l'impact (et donc la différence) de l'ajout de la variable en question sur les prédictions du modèle. Les différences sont ensuite moyennées pour obtenir une valeur de Shapley.
- L'importance des caractéristiques par permutation (PFI, Permutation Feature Improtance) : l'importance d'une caractéristique est mesurée en calculant l'augmentation de l'erreur de prédiction du modèle après permutation des valeurs de la caractéristique.

3 RESULTATS ET DISCUSSION

Le modèle empirique (SiRCA) offre déjà de bonnes performances pour la Durance, la Saône, et l'Ain (avec des R^2 de 0,68, 0,59, 0,70 respectivement), performances qui sont significativement améliorées par le RF et par le CNN-LSTM (0,94, 0,95, et 0,92 (RF), et 0.98, 0.97, et 0.97 (CNN-LSTM)). Pour la Drôme, le modèle empirique est beaucoup moins performant que les techniques de ML (passant de 0,44 à 0,90 (RF) et 0.94 (CNN-LSTM)). Enfin, le modèle SiRCA ne s'avère pas performant pour l'Isère et le Bourbre (0,26 et 0.10 respectivement) qui sont bien modélisés avec les modèles ML une (0.97 (CNN-LSTM), 0.92 (RF) pour l'Isère, le Bourbre respectivement). Nous constatons que de manière générale, le réseau neuronal est plus proche de la mesure que le RF sur la quasi-totalité des affluents (Voir Figure 1 et Tableau 2). Par ailleurs, en utilisant le modèle SiRCA pour estimer les valeurs manquantes des CMES dans la base de données OSR, utilisées pour calculer les flux de chaque affluent et, par conséquent, le bilan annuel à l'exutoire (Beaucaire) (A. Gruat, 2022)), on constate que les bilans ne se rebouclent pas toujours correctement. En effet, les apports des affluents ne correspondent pas systématiquement au flux mesuré à Beaucaire, qui est parfois sous-estimé ou sur-estimé. Le modèle CNN-LSTM, quant à lui arrive à mieux estimer le bilan annuel (avec un gain en précision de 4 MT/ans pour les années hydrologiques 2012-2013 et 2019-2020).

Affluents	Modèle	R^2	RMSE (mg/L)
Durance	SiRCA	0.68	432
	RF	0.94	168
	CNN-LSTM	0.98	114
Drôme	SiRCA	0.44	330
	RF	0.9	139
	CNN-LSTM	0.94	113
Isère	SiRCA	0.26	305
	RF	0.92	303
	CNN-LSTM	0.97	237
Saône	SiRCA	0.59	13

	RF CNN-LSTM	0.95 0.97	5 4
Bourbre	SiRCA RF CNN-LSTM	0.1 0.92 0.9	49 14 17
Ain	SiRCA RF CNN-LSTM	0.7 0.92 0.97	9 5 3

Tableau 2 : Comparaison des performances des modèles

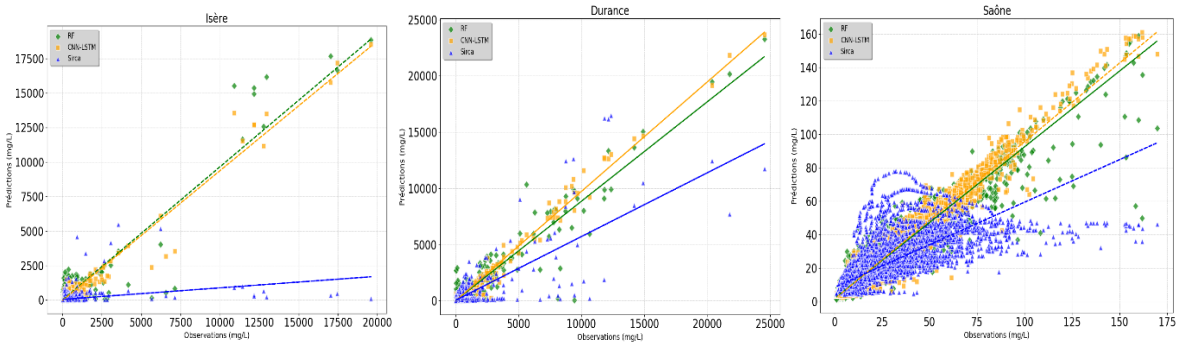

Figure 1 : Comparaisons des modèles : Prédictions vs Observations (RNN en jaune, SiRCA en bleu et RF en vert)

Il est également intéressant de regarder le comportement des crues et des différentes hystérèses qu'elles dessinent. Le modèle CNN-LSTM arrive à reproduire parfaitement le comportement des crues et des différentes hystérèses, aussi bien dans les boucles horaires qu'anti-horaire (voir Figure 2).

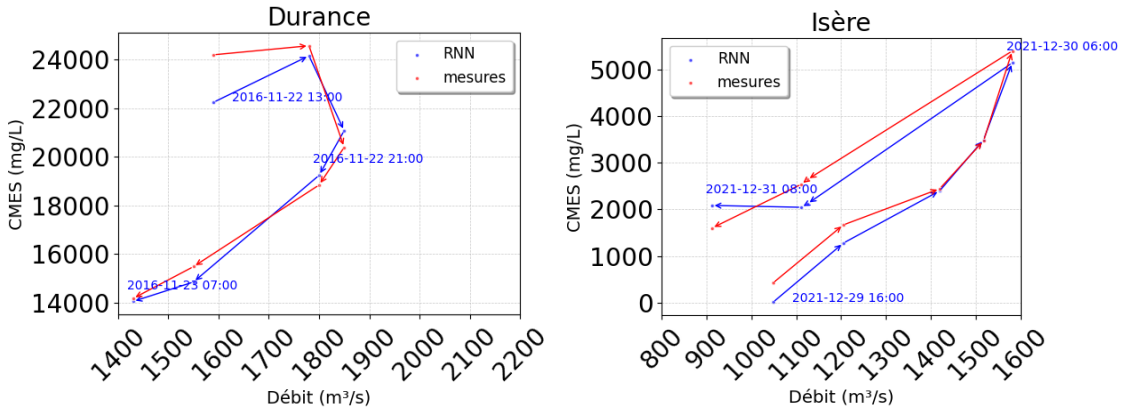

Figure 2 : Dynamique des crues : Reproduction des hystérèses par le CNN-LSTM

Les résultats de l'analyse de sensibilité des trois méthodes (MDI, PFI et Shapley Values) montrent qu'il n'y a pas une tendance générale. Par exemple sur la Drôme, le débit de la deuxième station (d'aval en amont) est le plus influent (contribution = 50%) alors que pour la Durance, le débit de la station située complètement en aval est la variable la plus importante (contribution > 60%) pour 2 des 3 méthodes (MDI et Shapley Values). Nous notons aussi que les précipitations ne contribuent que très peu dans tous les affluents (<1%).

Nos travaux démontrent ainsi l'utilité d'une approche qui ne se base pas sur l'historique des mesures de CMES comme c'est le cas dans la littérature. Les modèles de ML et DL permettent de mieux modéliser les CMES et les flux associés qu'une approche empirique. Les performances obtenues dans la modélisation suggèrent que notre méthodologie est pertinente. Par conséquent, elle peut être appliquée sur les projections futures des débits et des précipitations afin que l'on puisse étudier le comportement des CMES selon les différents scénarios climatiques élaborés par le projet Explore2 (Explore2, 2024).

BIBLIOGRAPHIE

A. Gruat, M. C. (2022). Rapport sur le fonctionnement du réseau d'observation des flux de matières en suspension et de contaminants particulaires pour l'année 2023. Rapport final, Observatoire des Sédiments du Rhône – 6ème programme d'action. Action E1,.

Bárbara M. Jung, E. H.-R. (2020). Estimating suspended sediment concentrations from river discharge data for reconstructing.

Bibhuti Sahoo, S. S. (2023). novel smoothing-based deep learning. *Water Resources*. doi:0.1007/s11269-023-03552-7

Crawford, C. (1991). Estimation of suspended-sediment rating curves and mean suspended-sediment loads. *J. Hydrol.*

Deepak Gupta, B. B. (2021). rtifcial intelligence for suspended sediment load prediction : a review.

Explore2. (2024, 11). *Explore2 - des futurs de l'eau*. Récupéré sur Institut national de recherche pour l'agriculture, l'alimentation et l'environnement (Inrae): https://professionnels.ofb.fr/fr/node/1244

HydroPortail. (s.d.). Récupéré sur https://hydro.eaufrance.fr

Khabat Khosravi, A. G. (2022). Suspended sediment load modeling using advanced hybrid rotation forest based elastic network. *Journal of Hydrology*. doi:: https://doi.org/10.1016/j.jhydrol.2022.127963

Lepage, H. G.-P.-P. (2022). Concentrations and fluxes of suspended particulate matter and associated contaminants in the Rhône River from Lake Geneva to the Mediterranean Sea. *Earth System Science Data*. doi:https://doi.org/10.5194/essd-14-2369-2022

Mahrez Sadaoui, W. L. (2016). Estimation of suspended-sediment rating curves and mean suspended-sediment loads. *Journal of Hydrology*.