Projet Machine Learning:
Prévision de débits fluviaux

Julien ZHOU, Wendong LIANG

2022

Table des matières

Chapitre 1

Introduction

Dans ce travail, nous avons choisis d'étudier des séries temporelles concernant le débit de plusieurs fleuves en France, pour lesquels nous nous intéressons au problème de prévision.

On fait ci-dessous une petite revue rapide et non exhaustive de méthodes rencontrées lors de recherches rapides dans la littérature :

L'analyse des séries temporelles hydrologiques (concernant le cycle de l'eau) a un champ d'application dans de nombreux domaines, en géologie, en climatologie, dans l'étude des sols etc...

Des modèles physiques prenant en compte des équations de convection-diffusion et les différents processus hydrologiques sont utilisés dans [1], [2] et [3].

Des méthodes et grandeurs statistiques sont utilisées depuis assez longtemps, comme en témoigne [4], [5] et [6], on y retrouve des méthodes statistiques fondamentales, ainsi que quelques techniques avancées. Il y a notamment un intérêt pour la théorie des valeurs extrêmes dans [7], ce qui est compréhensible pour la prévision des crues. L'article [8] fait une revue assez récente (2018) des méthodes de machine learning utilisées pour la prédiction des crues. On retrouve des modèles autorégressifs dans [9] et [10] pour faire des prévisions. Les modèles de Deep Learning, avec des réseaux de neurones sont devenues assez populaires, on les retrouve dans [11], [12], [13], [14], [15], [16], [17] et [18]. On retrouve aussi parfois l'utilisation d'algorithmes de Support Vector Regression et de K-Nearest Neighbors. Les métriques utilisées sont classiques pour les statistiques, Root Mean Squared Error, Mean Absolute Error et R^2. On remarque l'utilisation du critère d'efficacité de Nash-Sutcliffe spécifique à l'hydrologie, qui se ramène cependant au R^2 d'un modèle de régression.

Dans notre travail, nous nous intéressons à des modèles statistiques, sans connaissance à priori des phénomènes physiques sous-jacents. Le problème que nous allons étudier est celui de la prévision à 1 semaine du débit d'eau journalier moyen à différentes stations sur la Seine, la Loire et la Garonne, en prenant en compte les débits mesurés actuels ainsi que des données météorologiques.

Dans la première partie de ce rapport, nous présentons les données qui ont servis de base à notre étude et la manière dont elles ont été retraitées. Puis dans une seconde partie, nous mettons en avant la méthodologie mise en place pour notre travail. La dernière partie est consacrée à la présentations des modèles mis en places ainsi qu'aux résultats obtenus.

Les codes utilisés sont disponibles sur Github : `https://github.com/JlnZhou/StatML_ProjectML` [19].

Chapitre 2

Données

2.1 Sources

Nous avons d'abord envisagé d'utiliser l'API "Hydrométrie" [20]. Cette dernière fournit des données provenant d'environ 3000 stations provenant des Directions Régionales de l'Environnement de l'Aménagement et du Logement (DREAL) et d'autres producteurs (collectivités, etc...). Elle fournit des données quasi temps réel de débits moyens journaliers ainsi que de débits mensuels. Cependant les requêtes possibles sont limités dans la profondeur d'historique disponible (1 mois selon la documentation).

Nous nous sommes donc tournés directement vers la plateforme dont sont issues les données de l'API, la banque HYDRO [21]. Il s'agit d'une banque de donnée qui dépend du ministère de l'Ecologie, du Développement Durable et de l'Energie. On peut y retrouver des mesures hydrologiques (hauteur, débit) à pas de temps variables de plusieurs milliers de stations de mesures implantées sur les cours d'eau français, avec les caractéristiques des stations (localisation altitude, etc...). Les mesures proviennent essentiellement de services de l'Etat, d'organismes de recherche ainsi que de compagnies d'aménagement.

A ce stade nous avons choisi d'étudier des données relatives à 4 fleuves : la Seine, la Loire, la Garonne et le Rhône. Nous avons extrait des données de débit à pas de temps constant (2h) sur 11 années, de 2010 à 2020, pour toutes les stations disponibles sur ces cours d'eau.

Il est à noter que la base a connu des problèmes fin 2021, elle a subi des attaques et a été indisponible pendant pas mal de temps. Cela a ralenti notre démarrage, mais ils ont été résolu courant Novembre / Décembre.

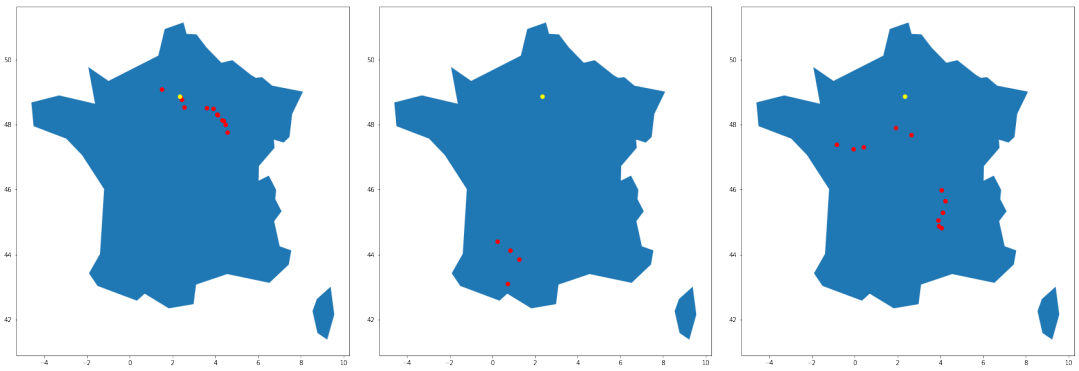
En ce qui concerne les données météorologiques, nous avons décidé de travailler avec la base de données SYNOP [22]. Les mesures proviennent du réseau de l'Organisation Météorologique Mondiale (OMM) et semble assez complète. On peut y trouver des mesures assez variées, de température, humidité, direction et force du vent, pression atmosphérique, hauteur de précipitations au pas de temps 3 heures.

2.2 Nettoyage des données

2.2.1 Débit

Après une première phase un peu exploratoire des possibilités de la base HYDRO, nous avons choisi d'extraire les données de débit à pas 2 heures pour toutes les stations de la Seine, la Loire, la Garonne et le Rhône. La base fournissant ces données séparément dans formats peu exploitables un premier travail a consisté à les rassembler dans un fichier facilement exploitables avec des outils d'analyse de données (des tableaux).

Nous avons étudié l'intégrité des données, la présence / répartition des données manquantes. Les stations avec une proportion de données manquantes trop importantes ont été supprimées, et une interpolation linéaire a été faite pour combler les quelques trous restant (faute de mieux). L'étude du Rhône a notamment été abandonnée, faute de données assez complètes.



(a) La Seine (b) La Garonne (c) La Loire

FIGURE 2.1 – Répartition des stations de mesures de débit

Ayant finalement choisi de travailler sur des données journalières, les données ont été moyennés par jour.

Le tableau A.1 en Annexe présente les 26 stations gardées est présent en Annexe , pour lesquelles nous avons 1 mesures par jour entre 2010 et 2020 inclus.

2.2.2 Météo

Le fichier obtenu contient des mesures météorologiques de plusieurs stations dans le monde, pas uniquement en France métropolitaine. Un premier tri a donc été fait avec un critère géographique.

D'autre part, plusieurs variables disponibles correspondent à la même grandeur physique mais parfois à des pas de temps différents. Un tri a été fait en fonction de leur intégrité, en comblant les "petits trous" par interpolation linéaire. Cela a permis d'obtenir des données au pas 3 heures (pas de temps d'origine de la base).

Un traitement spécifique a été fait pour le vent. En effet, on a accès à la vitesse du vent ainsi qu'à un angle correspondant à sa direction. Ces variables ont été converties en 3 variables contenant les coordonnées d'un vecteur dans le plan et sa norme.

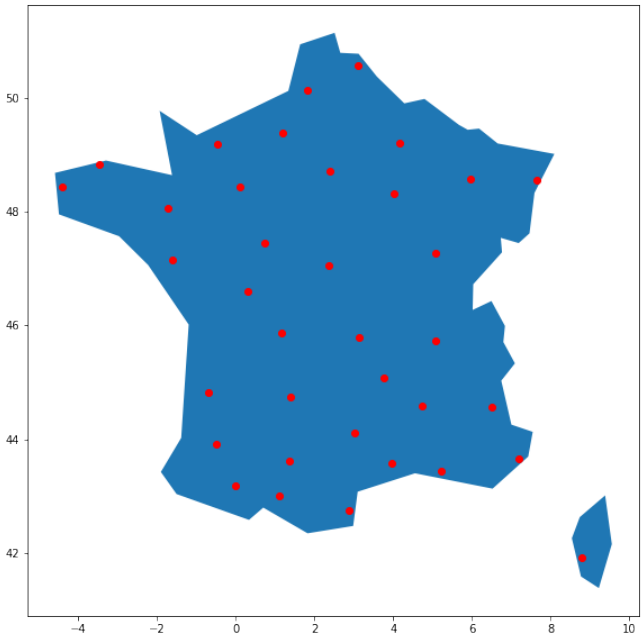


FIGURE 2.2 – Répartition des stations météo prises en compte

Au final, après les traitements, nous obtenons pour chaque station les données suivantes :
— la pression atmosphérique,
— le vent Nord,
— le vent Est,
— la vitesse du vent
— la température,
— l'humidité,
— les précipitation.
Les données ont été ici aussi agrégées par jour.

Le tableau A.2 en Annexe présente les 35 stations gardées est, pour lesquelles nous avons ces 7 grandeurs par jours entre 2010 et 2020 inclus.

Chapitre 3

Cadre

3.1 Problème étudié

Nous nous intéressons à un problème de régression des vecteurs des débits journaliers, à horizon une semaine (J+7). Les variables explicatives qui sont utilisées sont :

— la date,
— le vecteur des débits au jours J,
— les conditions météorologiques à J+7 (nous assimilons les données d'observation à des prévisions pour simplifier).

Afin de répondre à ce problème, nous testons plusieurs modèles (cf. prochaine partie).

Dans la suite nous adopterons les notations génériques suivantes :

— n sera un nombre d'observations (d'entraînement / de test, selon les situations),
— X désignera un vecteur de données explicatives,
— Y désignera un vecteur à prédire,
— $\hat{Y}$ désignera une prédiction (sortie d'un modèle).

3.2 Métriques

Afin de vérifier l'adéquation de nos modèles, nous utilisons plusieurs critères usuels dans le cadre de la régression :

— le Root Mean Squared Error RMSE,
— le Mean Absolute Error MAE,
— le coefficient de détermination R^2.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \hat{Y}_i|$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \frac{1}{n}\sum_{j=1}^{n} Y_j)^2}$$

Le RMSE et le MAE caractérisent l'écart moyen à la valeur à prédire. Cependant, le RMSE est plus sensible à la présence de forts écarts à cause de la présence du carré. Ces deux critères sont à minimiser et le R^2 est à avoir aussi proche de 1 que possible.

Ces critères sont calculés pour chaque coordonnées de Y. Dans l'absolu, il faudrait les prendre en compte séparément pour chaque station, mais cela rend le problème difficile à aborder. C'est pourquoi nous décidons de regarder la moyenne (sur les stations) de ces critères, calculés sur les séries standardisées. Cela permet d'avoir 3 scores (RMSE, MAE, R2) pour chaque modèle.

3.2.1 Remarque sur le R^2

Malheureusement, une confusion s'est glissée dans nos implémentations.
En effet, dans le cas particulier de la régression linéaire, on a :

$$\begin{aligned} R^2 &= 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \frac{1}{n}\sum_{j=1}^{n} Y_j)^2} \\ &= \frac{\sum_{i=1}^{n}(\hat{Y}_i - \frac{1}{n}\sum_{j=1}^{n} Y_j)^2}{\sum_{i=1}^{n}(Y_i - \frac{1}{n}\sum_{j=1}^{n} Y_j)^2} \end{aligned}$$

mais c'est faux dans le cas général.

Or, dans nos implémentations, le R^2 a été défini comme $\frac{\sum_{i=1}^{n}(\hat{Y}_i - \frac{1}{n}\sum_{j=1}^{n} Y_j)^2}{\sum_{i=1}^{n}(Y_i - \frac{1}{n}\sum_{j=1}^{n} Y_j)^2}$. Ce dernier est assimilable à un rapport de de la variance de la série prédite, sur la variance de la série réelle.

Dans la suite, à la place du R^2, nous écrirons donc RV (rapport de variances) qui correspond à $\frac{\sum_{i=1}^{n}(\hat{Y}_i - \frac{1}{n}\sum_{j=1}^{n} Y_j)^2}{\sum_{i=1}^{n}(Y_i - \frac{1}{n}\sum_{j=1}^{n} Y_j)^2}$. Ce dernier doit aussi converger vers 1 aussi, mais on peut avoir $RV = 1$ avec un modèle mauvais alors que ce n'est pas possible pour le R^2 où un écart quadratique apparaît explicitement.

3.3 Approche de validation croisée

Pour avoir des estimations qui ne soient pas trop optimistes des risques de chaque modèle, nous avons choisis d'utiliser une procédure de validation croisée (notamment pour les choix d'hyperparamètres).

Dans le cas de séries temporelles, les données ne sont pas indépendantes et l'on ne peut pas faire de validation croisée au sens strict du terme. Plusieurs stratégies permettent tout de même de faire des estimations raisonnables.

Nous avons choisi de mettre l'année 2020 de coté, comme ensemble de test.

Puis les différents critères sont estimés avec les données de 2010 à 2019, de la manière suivante : Chaque année de 2011 à 2019 est séquentiellement utilisée comme ensemble de validation, et les modèles sont entraînés sur les années précédentes.

3.4 Quelques observations préliminaires

Afin de comprendre la structure intrinsèque des séries des débits, nous avons effectué quelques analyses préliminaires, notamment d'autocorrélation. On voit notamment des pics caractéristiques d'une saisonnalité dans les séries. Leur positions et leur nombre nous incite à croire que leur origine est liée à l'alternance des saisons. On peut en effet constater la disparition de ce phénomène après avoir retranché une composante annuelle calculée naïvement, en faisant les moyennes par date.

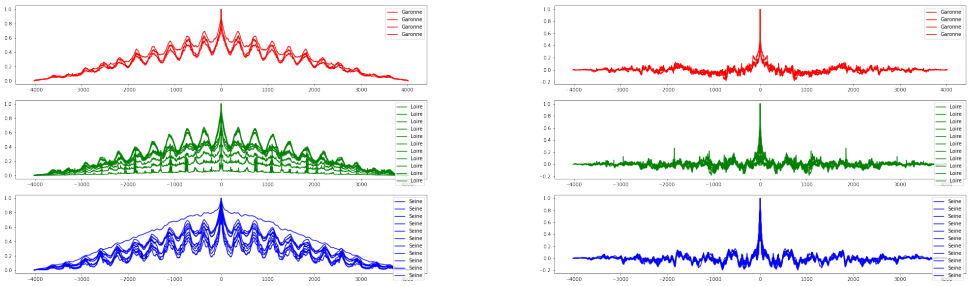


FIGURE 3.1 – Autocorrélationx des débits d'eau avant et après désaisonnalisation

On peut observer avec les autocorrélations croisées (par rapport aux stations les plus en amont pour chaque fleuve), l'influence des débits mesurés dans les stations en amont sur les débits en aval dans le futur, d'où la nécessité de considérer toutes les stations en même temps dans les modèles.

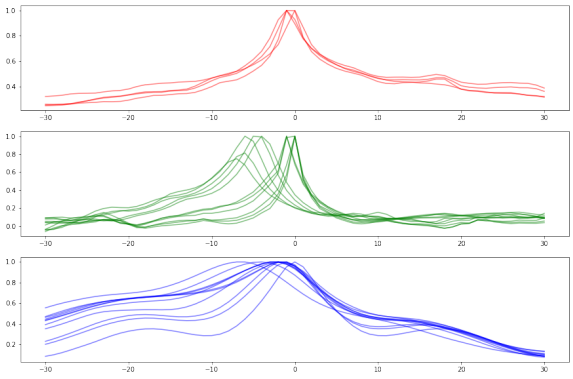


FIGURE 3.2 – Autocorrélations croisées par fleuve, avec la station amont

Chapitre 4

Modélisations

Dans cette partie, nous présentions les modèles utilisés et les résultats obtenus. D'une manière générale, la philosophie de l'étude a été d'améliorer incrémentalement différents modèles.

Les échelles de variations pour les débits pouvant être assez différents en fonction des stations, les données ont été standardisées pour l'entraînement des modèles, et pour le calcul des critères. Les hyperparamètres ont été choisis, lorsque c'est possible, par validation croisée. Des visualisation des résultats sont en Annexe B.

4.1 Deux Modèles naïfs

Nous présentons tous d'abord deux approches naïves qui servent de base de comparaisons.

4.1.1 Modèle Moyenne

Une première approche naïve consiste à estimer la moyennes des débits mesurés sur l'ensemble d'entraînement, et de l'utiliser pour la prédiction à une semaine. Ceci donne les résultats suivants :

	RMSE	MAE	RV
VC	1.10	0.76	0
Test	0.83	0.62	0

4.1.2 Modèle J+7

Une seconde approche consiste à utiliser le débit présent pour le débit à J+7.

	RMSE	MAE	RV
VC	0.91	0.44	0.99
Test	0.70	0.31	0.99

On constate que cette méthode est un peu meilleure car elle suit les variations des débits (quoique en décalés).

4.2 GAM

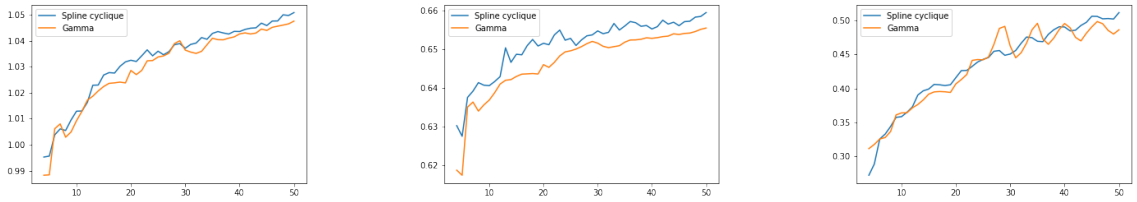
Une première tentative de modélisation utilise un modèle additif généralisé (GAM), avec les dates.

On suppose $Y = f(X) + \epsilon$ avec ϵ une variable aléatoire centrée gaussienne.

Dans le cas général, f est une fonction somme d'une partie linéaire et de fonctions lisses approximées par exemple avec des splines. Dans notre cas, f est une base de splines cubiques cycliques avec X la date comme variable explicative. Cela permet d'éviter des problèmes de sauts en fin d'année. D'autre part, un avantage de de cette modélisation est la simplicité pour estimer ses paramètres. Une fois

un nombre de noeuds dans l'année choisi (par validation croisée), l'estimation des paramètres se fait simplement par avec des outils classiques d'optimisation quadratique.

Un raffinement de cette méthode consiste à ajouter une fonction de "lien". On suppose alors que Y sachant X ne suit plus une loi normale mais une autre loi exponentielle. Nous avons utilisé une loi gamma (dont la fonction de lien est $x \to 1/x$ mais d'autres sont possibles (voir [23]).



(a) RMSE (b) MAE (c) RV

FIGURE 4.1 – Choix du nombre de noeuds dans le modèle GAM

Des splines cycliques avec et sans fonction de lien gamma sont testées pour différents nombre de noeuds (hyperparamètre). Pour éviter un problème de surapprentissage, nous avons choisis de garder 10 noeuds dans le modèle.

Nous avons les résultat pour des splines cycliques, avec une fonction de lien gamma et 10 noeuds :

	RMSE	MAE	RV
VC	1.01	0.63	0.36
Test	0.72	0.48	0.37

Les performances se situent entre les 2 modèles naïfs. Les prédictions suivent globalement les réalisation mais assez grossièrement. Dans la suite, nous utiliserons la sortie de ce modèle comme tendance saisonnière, et les modèles ayant pour tâche de capturer les variations plus fines seront entraînés sur les résidus.

4.3 GAM+VAR

Maintenant que la composante de variations annuelles est gérés, on considère que la série des résidus est stationnaire. En simplifiant et en enlevant l'influence de variables exogènes (la météo particulièrement), on peut essayer de mettre en oeuvre. Nous regardons tout particulièrement le modèles Vector Autoregression (VAR).

Dans notre cas, on a $p = 27$ variables qui sont les stations de débit d'eau. Dans le modèle VAR(k) (k hyperparamètres à déterminer), on a l'équation :

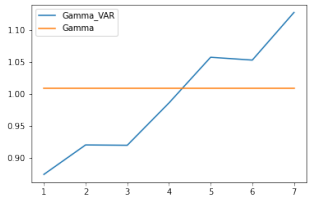
$$Y_t = \nu + \sum_{i=1}^{p} A_i Y_{t-i} + \epsilon_t$$

où :

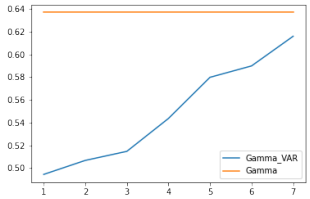
- Y_t est le vecteur des débits au temps t,
- ν est un vecteur moyen,
- les A_i sont des matrices à déterminer, caractérisant les autocorrélation et corrélations croisées des séries,
- ϵ_t est un terme d'erreur, que l'on considérera iid.

On constate sur la figure 4.2 que pour de faibles valeurs de k, le modèle GAM + VAR a de meilleures performances que le GAM simple. Augmenter le nombre d'observations passées ont tendance à augmenter les critères de RMSE et MAE (ce qui sera globalement le cas aussi par la suite).

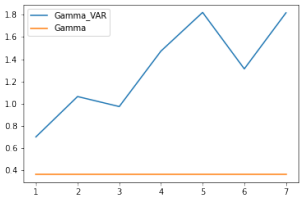
Pour le modèle GAM+VAR final, nous choisissons de garder un lag de 2 et on a les résultats suivant :



(a) RMSE



(b) MAE



(c) RV

FIGURE 4.2 – Comparaison des modèles GAM gamma et GAM+VAR en fonction de k

	RMSE	MAE	RV
VC	0.92	0.51	1.06
Test	0.57	0.33	0.76

4.4 GAM+LASSO+AR

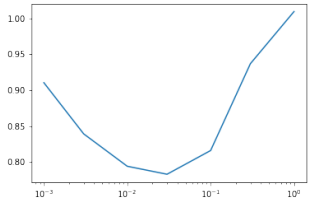
Nous n'avons pas encore utilisé les données météo à disposition. Un problème s'impose, nous avons 7 grandeurs mesurées pour chaque station en France (une cinquantaine). Il semble qu'il faille faire de la réduction de dimensions. Plusieurs options sont possibles, l'une d'entre elle est de faire une régression LASSO.

La régression LASSO permet de sélectionner les variables importantes dans la régression linéaire. Il s'agit de résoudre un problème de minimisation du type :

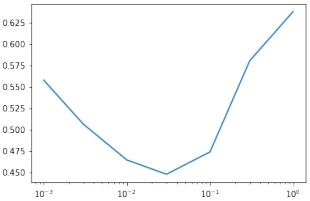
$$\begin{aligned}\beta_\lambda &\in \arg\min_{\beta\in\mathbb{R}^p} \sum_{i=1}^{n}(Y_i - \beta_1 x_i^1 - \beta_2 x_i^2 - \ldots - \beta_p x_i^p)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \\ &\in \arg\min_{\beta\in\mathbb{R}^p} ||Y - X\beta||_2^2 + \lambda||\beta||_1\end{aligned}$$

pour un λ à calibrer.

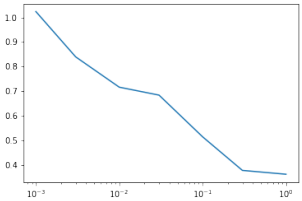
Ici aussi, on enlève d'abord la partie périodique calculée par GAM, ensuite on fait une prédiction à partir des données du jour J, de plus ou moins d'historique, et de données météorologique du jour J+7.



(a) RMSE



(b) MAE

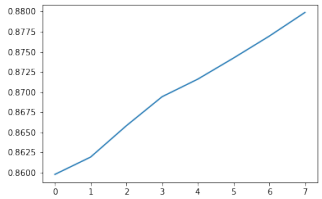


(c) RV

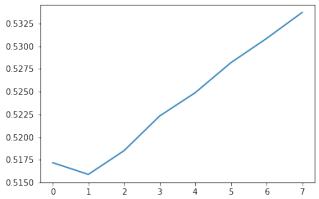
FIGURE 4.3 – Choix du λ pour GAM+LASSO+AR

Un choix raisonnable d'hyperparamètres est de prendre $\lambda = 0.03$ et un lag de 1, ce qui donne les performances :

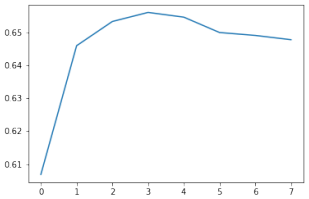
	RMSE	MAE	RV
VC	0.78	0.45	0.69
Test	0.55	0.31	0.55



(a) RMSE



(b) MAE



(c) RV

FIGURE 4.4 – Choix du lag pour GAM+LASSO+AR

Choisir $\lambda = 0.03$ permet de grandement réduire le nombre de variables dans la régression. On observe dans la table B.1 que pour la régression sur chaque station de débit, il ne reste qu'une vingtaine de variables prises en compte, par rapport aux 246 variables présentes dans les données. Il s'agit principalement de variables concernant l'historique des stations (variables sous forme de "[code]_[lag]"), ainsi que des variables météorologiques.

4.5 GAM+PLS+AR

Le LASSO fait une sélection de variable. Une autre méthode de réduction de dimension consiste à faire une analyse en composantes principales (ACP) pour trouver des variables explicatives les plus orthogonales possibles. Cependant faire une ACP sur les variables explicatives sans prendre en compte leur influence sur la cible, puis faire une régression sur les composantes principales (RCP) peut mener à des prédictions de piètre qualité.

Nous avons ici mis en oeuvre une approche de régression des moindres carrés partiels (Partial Least Square Regression) qui fait une ACP mais prend en compte l'importance des composantes choisies sur la variables à prédire. C'est un modèle particulièrement indiqué lorsque les variables explicatives sont correlées entre elles, ce qui risque d'être le cas pour notre problème.

On reprend le modèle de régression linéaires :

$$Y = X\theta + \epsilon$$

Avec :
— Y de dimension $n \times m$
— X de dimension $n \times p$

On cherche une matrice de poids W de taille $p \times l$ (l le nombre de composantes gardées) et on pose $T = XW$. Cette matrice T de taille contient alors les l variables explicatives (contre p pour X) sur lesquelles nous effectuons la régression.

Une régression de Y sur T donne :

$$Y = TQ + E = XWQ + E$$

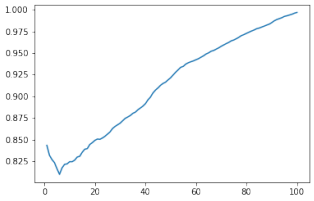
avec $\theta = WQ$.

On ne développe pas la méthode utilisée pour effectuer la régression. C'est une méthode itérative, et on renvoie à [24] pour plus de détails.

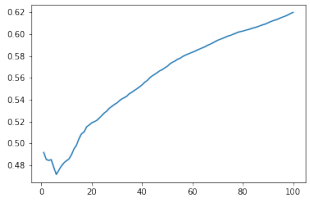
Les résultats de validation croisée nous encouragent à choisir de l'ordre de 10 composantes, et un lag de 0 (on prend en compte l'observation présente, mais aucune du passé) :

	RMSE	MAE	RV
VC	0.81	0.48	0.77
Test	0.56	0.33	0.59

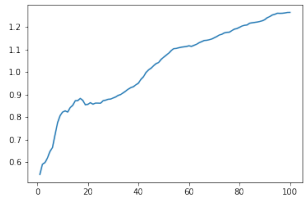
Les performances sont légèrement moins bonnes que pour le LASSO même si elles sont assez comparables.



(a) RMSE

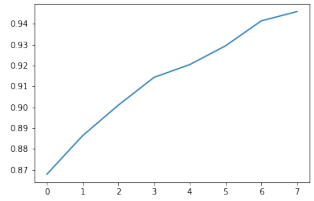


(b) MAE

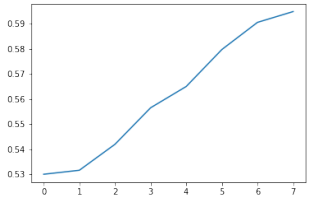


(c) RV

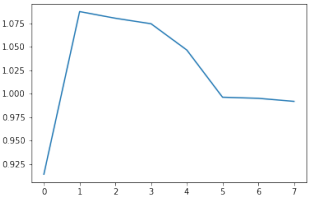
FIGURE 4.5 – Choix du nombre de composantes pour GAM+PLS+AR



(a) RMSE



(b) MAE



(c) R2

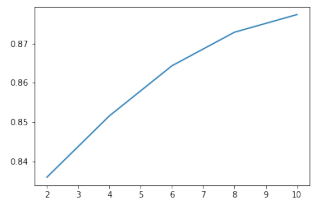
FIGURE 4.6 – Choix du lag pour GAM+PLS+AR

4.6 GAM+RF+AR

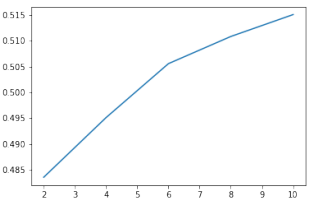
Jusque là, nous avons utilisé des modélisation linéaires, mais certaines interactions peuvent ne pas l'être. Nous mettons eu oeuvre ici une méthode à base d'arbres de régression. Un des avantages des arbres est notamment leur sélection de variable automatique.

Nous étudions le modèle Random Forest (RF) (voir [25]), qui consiste à agréger les prédiction d'un grand nombre de prédicteurs "faibles". La méthode consiste à combiner les prédiction d'un grand nombre d'arbres faiblement corrélés car entraînés sur des ensembles tirés aléatoirement, et prenant en compte des variables explicatives elles aussi tirées aléatoirement. (Nous utilisons l'implémentation présente dans le package sklearn.)

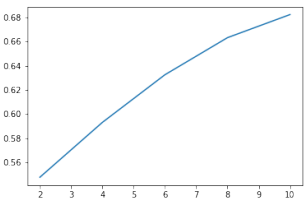
Nous nous limitons dans un premier temps à un nombre de 100 arbres, à cause du coût calculatoire prohibitif d'entraînement. Les hyperparamètres restant, auxquels nous nous intéressons sont la proportion de l'ensemble d'entraînement utilisé pour chaque arbre, et ici aussi la profondeur d'historique sur les débits pris en compte. Pour le nombre de variables prises en compte, nous gardons par défaut la racine carrée du nombre total de variables.



(a) RMSE



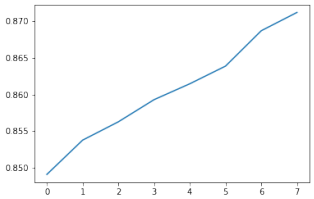
(b) MAE



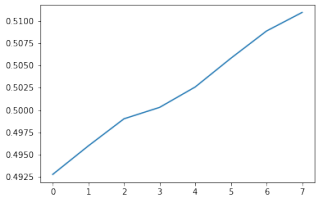
(c) RV

FIGURE 4.7 – Choix de la proportion de l'ensemble d'entrainement pour GAM+RF+AR

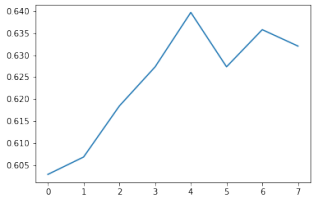
On obtient les meilleurs performances pour l'utilisation de 20% (ce qui est logique pour rendre les arbres faiblement corrélés) de l'ensemble d'entraînement, et un lag de 0.



(a) RMSE



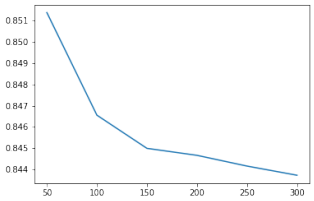
(b) MAE



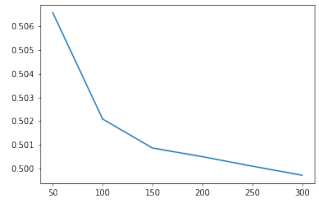
(c) RV

FIGURE 4.8 – Choix du lag pour GAM+RF+AR

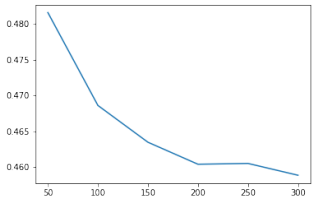
Pour améliorer les performances du modèles, on essaie maintenant d'augmenter les nombre d'arbres, et jouer sur la profondeur des arbres.



(a) RMSE

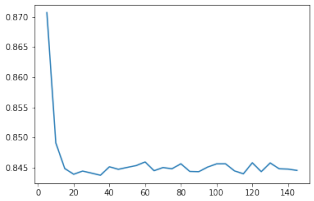


(b) MAE

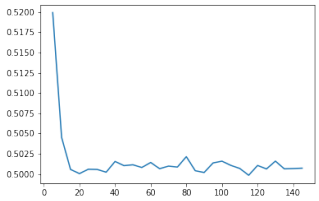


(c) RV

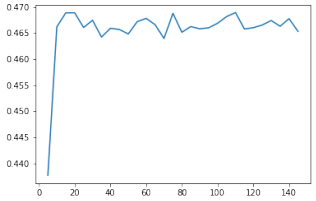
FIGURE 4.9 – Choix du nombre d'arbres pour GAM+RF+AR



(a) RMSE



(b) MAE



(c) RV

FIGURE 4.10 – Choix de la profondeur des arbres pour GAM+RF+AR

On voit que les performances ne s'améliorent pas énormément par rapport à ce que l'on avait auparavant avec le nombre d'arbre, et que la profondeur des arbres n'influe pas beaucoup.

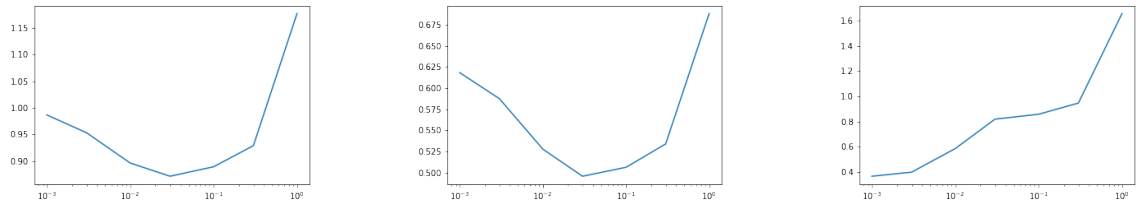
Finalement, pour les hyperparamètres :
— proportion d'observations par arbre : 20%,
— lag : 0,
— nombre d'arbres : 100
les performances :

	RMSE	MAE	RV
VC	0.82	0.47	0.53
Test	0.59	0.33	0.60

4.7 GAM+GBoost+AR

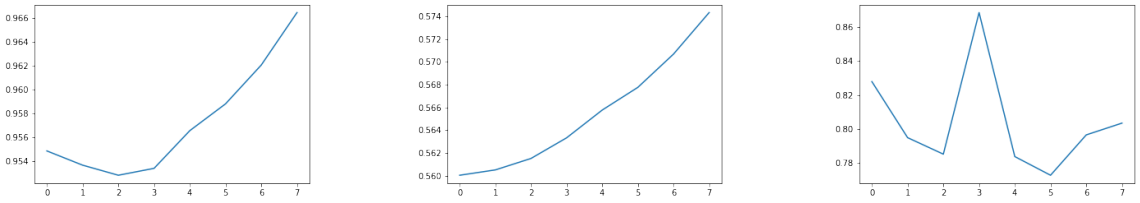
Un autre méta-modèle d'apprentissage à base d'arbres de régression consiste à en entraîner de manière incrémentale, de sorte que chaque arbre apprenne les erreurs du précédent, et les réduise.

Nous étudions les performances du modèle Gradient Boosting (GB), avec ici aussi une centaine d'arbre (implémentation de sklearn). Nous cherchons à optimiser le "learning rate" (lr) de l'algorithme, et la profondeur d'historique de débits pris en compte.



(a) RMSE (b) MAE (c) RV

FIGURE 4.11 – Choix du lr pour GAM+GBoost+AR



(a) RMSE (b) MAE (c) RV

FIGURE 4.12 – Choix du lag pour GAM+GBoost+AR

Le choix du learning rate de l'ordre de 0.03 semble pertinent. Pour la profondeur d'historique à prendre en compte il y a un compromis entre le RMSE et le MAE. Nous choisissons de prendre un lag de 0, en privilégiant donc la robustesse (caractérisée par le MAE).

Ceci donne les performances :

	RMSE	MAE	RV
VC	0.87	0.49	0.83
Test	0.57	0.30	0.57

4.8 Bilan des modélisation

	RMSE	MAE	RV
Moyenne	1.10	0.76	0
J+7	0.91	0.44	0.99
GAM	1.01	0.63	0.36
GAM+VAR	0.92	0.51	1.06
GAM+LASSO+AR	0.78	0.45	0.69
GAM+PLS+AR	0.81	0.48	0.77
GAM+RF+AR	0.82	0.47	0.53
GAM+GBoost+AR	0.87	0.49	0.83

FIGURE 4.13 – Tableau des scores de validation croisée

On compare les scores de validation croisée de l'ensemble des modèles testés, avec les hyperparamètres choisis. Le modèle GAM+LASSO+AR semble le plus performant (de peu), et c'est confirmé par les scores sur l'ensemble de test. Lorsqu'on regarde les courbes des débits prédits par rapport aux

	RMSE	MAE	RV
Moyenne	0.83	0.62	0
J+7	0.70	0.31	0.99
GAM	0.72	0.48	0.37
GAM+VAR	0.57	0.33	0.76
GAM+LASSO+AR	0.55	0.31	0.55
GAM+PLS+AR	0.56	0.33	0.59
GAM+RF+AR	0.59	0.33	0.60
GAM+GBoost+AR	0.57	0.30	0.57

FIGURE 4.14 – Tableau des scores de test

courbes réelles, le résultat n'est toutefois pas encore pleinement satisfaisant. On voit notamment un échec pour prédire certaines variations subites. Cela pourrait être dû au modèle, ou alors au manque de certaines variables explicatives (par exemples relatives à l'activité humaine, la présence de barrage ?).

Cependant, en ce qui concerne les modèles non-linéaires testés, nous nous sommes limités sur certains paramètres à cause de la puissance de calculs disponibles (le nombre d'estimateurs agrégés), et ils pourraient s'avérer meilleurs avec un réglage plus fin des hyperparamètres.

D'autre part, les critères ont été agrégés sur les stations et les fleuves. Certains modèles peuvent être bons pour certains fleuves / certaines stations et pas pour d'autres. On pourrait faire une moyenne pondérée des scores en fonction d'une critère d'importance (par fleuve / station), et/ou se donner la possibilité d'utiliser des modèles différents pour prédire les débits à différentes stations.

Chapitre 5

Conclusion

Dans ce projet, nous avons entraîné des modèles de régression dans le but de prédire l'évolution des débits d'eau de certains fleuves en France. Nous avons adopté une approche incrémentale, en estimant une tendance générale (avec un GAM prenant en compte la date) avant de lui ajouter un modèle plus fin.

Notre modèle le plus performant en ce qui concerne les critères RMSE et MAE combine un GAM avec une régression LASSO sur les résidus à partir de variables autorégressives et météorologiques.

Nos modèles parviennent à suivre les évolution générales des débits, mais échouent à capter les évolution trop rapides. Plusieurs raisons sont possibles. Il se peut que nous n'ayons pas toutes les variables explicatives nécessaires, nous n'avons par exemple pas pris en compte les activités humaines pouvant avoir une influence (barrage ? ville ?...). Il se peut aussi que la qualité de nos données laisse à désirer. Une autre explication peut provenir de nos modèles qui ne sont peut-être pas assez élaborés. Des méthodes plus ambitieuses (mais nécessitant davantage de puissance de calcul, comme du Deep Learning) pourrait aboutir à de meilleurs prédictions.

D'autre part, nous avons utilisé des critères agrégés (sur toutes les station), mais il se peut que certains modèles performants pour certaines station ne le soient pas pour d'autres et inversement. Un étude plus fine, choisissant par exemple des modèles différents pour chaque station (puis allant jusque l'agrégation en ligne d'experts) est susceptible de fournir de meilleurs résultats.

Annexe A

Tableaux des stations hydrologiques et météo étudiées

Code station	Altitude	Long	Lat	Ordre (Amont/Aval)	Cours eau
O0200020	357.0	0.70	43.09	0	Garonne
O2620010	90.0	1.24	43.85	1	Garonne
O6140010	46.0	0.83	44.12	2	Garonne
O9000010	0.0	0.22	44.41	3	Garonne
K0030020	881.34	4.04	44.81	0	Loire
K0100020	0.0	3.92	44.88	1	Loire
K0260020	589.5	3.90	45.06	2	Loire
K0550010	442.0	4.11	45.29	3	Loire
K0690010	339.11	4.22	45.64	4	Loire
K0910050	267.51	4.0	45.99	5	Loire
K4180010	120.97	2.63	47.68	6	Loire
K4350020	89.82	1.92	47.89	7	Loire
K6830020	34.51	0.40	47.31	8	Loire
L8000020	24.12	-0.07	47.26	9	Loire
M5300010	9.58	-0.86	47.39	10	Loire
H0100010	248.25	4.57	47.76	0	Seine
H0100020	179.35	4.48	47.99	1	Seine
H0400010	148.0	4.37	48.11	2	Seine
H0400020	0.0	4.31	48.14	3	Seine
H0800011	102.0	4.10	48.30	4	Seine
H0800012	100.0	4.07	48.31	5	Seine
H0810010	78.0	3.88	48.50	6	Seine
H1700010	60.0	3.59	48.52	7	Seine
H3930020	34.72	2.55	48.53	8	Seine
H4340020	29.46	2.41	48.78	9	Seine
H8100021	9.16	1.48	49.09	10	Seine

Table A.1 – Tableau des stations hydrologiques étudiées

ID	Nom	Latitude	Longitude	Altitude
7005	ABBEVILLE	50.14	1.83	69
7015	LILLE-LESQUIN	50.57	3.10	47
7027	CAEN-CARPIQUET	49.18	-0.46	67
7037	ROUEN-BOOS	49.38	1.18	151
7110	BREST-GUIPAVAS	48.44	-4.41	94
7117	PLOUMANAC'H	48.83	-3.47	55
7130	RENNES-ST JACQUES	48.07	-1.73	36
7139	ALENCON	48.45	0.11	143
7149	ORLY	48.72	2.38	89
7168	TROYES-BARBEREY	48.32	4.02	112
7181	NANCY-OCHEY	48.58	5.96	336
7190	STRASBOURG-ENTZHEIM	48.55	7.64	150
7222	NANTES-BOUGUENAIS	47.15	-1.61	26
7240	TOURS	47.44	0.73	108
7255	BOURGES	47.06	2.36	161
7280	DIJON-LONGVIC	47.27	5.09	219
7335	POITIERS-BIARD	46.59	0.31	123
7434	LIMOGES-BELLEGARDE	45.86	1.18	402
7460	CLERMONT-FD	45.79	3.15	331
7471	LE PUY-LOUDES	45.07	3.76	833
7481	LYON-ST EXUPERY	45.73	5.08	235
7510	BORDEAUX-MERIGNAC	44.83	-0.69	47
7535	GOURDON	44.75	1.40	260
7558	MILLAU	44.12	3.02	712
7577	MONTELIMAR	44.58	4.73	73
7591	EMBRUN	44.57	6.50	871
7607	MONT-DE-MARSAN	43.91	-0.50	59
7621	TARBES-OSSUN	43.19	0.00	360
7627	ST GIRONS	43.01	1.11	414
7630	TOULOUSE-BLAGNAC	43.62	1.38	151
7643	MONTPELLIER	43.58	3.96	2
7650	MARIGNANE	43.44	5.22	9
7690	NICE	43.65	7.21	2
7747	PERPIGNAN	42.74	2.87	42
7761	AJACCIO	41.92	8.79	5

Table A.2 – Tableau des stations météo étudiées

Annexe B

Visualisations des prédictions

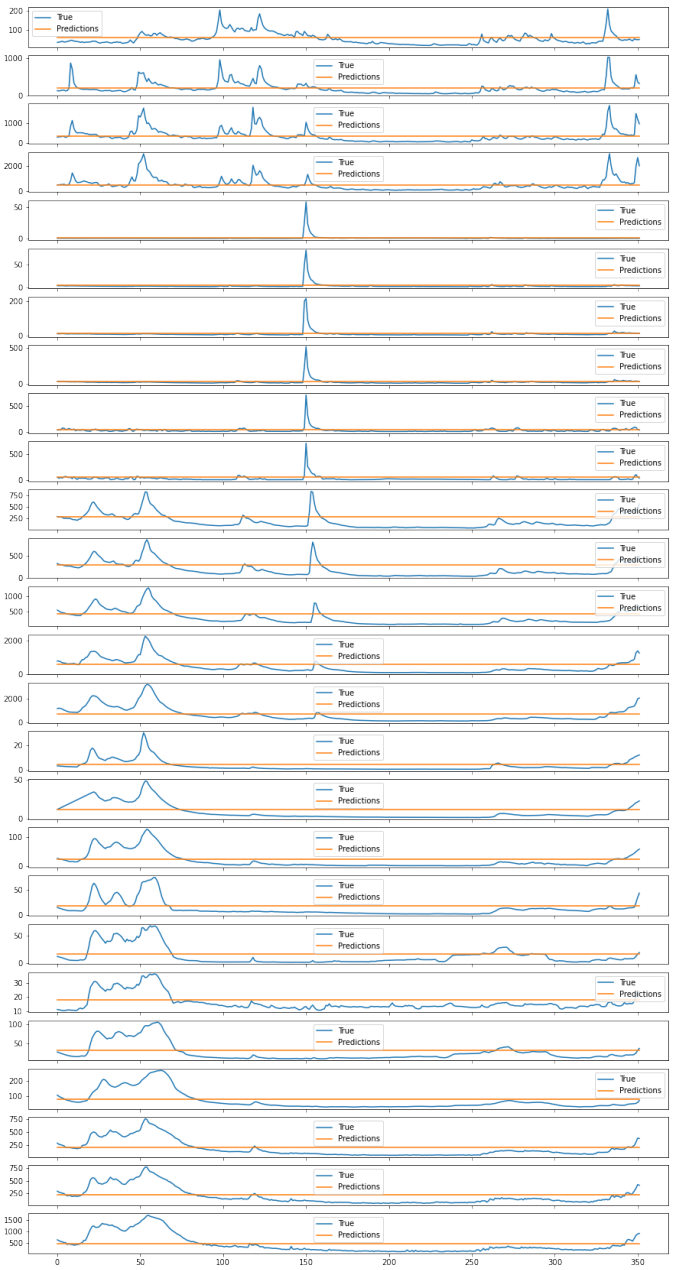


FIGURE B.1 – Prédiction du modèle naif, Moyenne

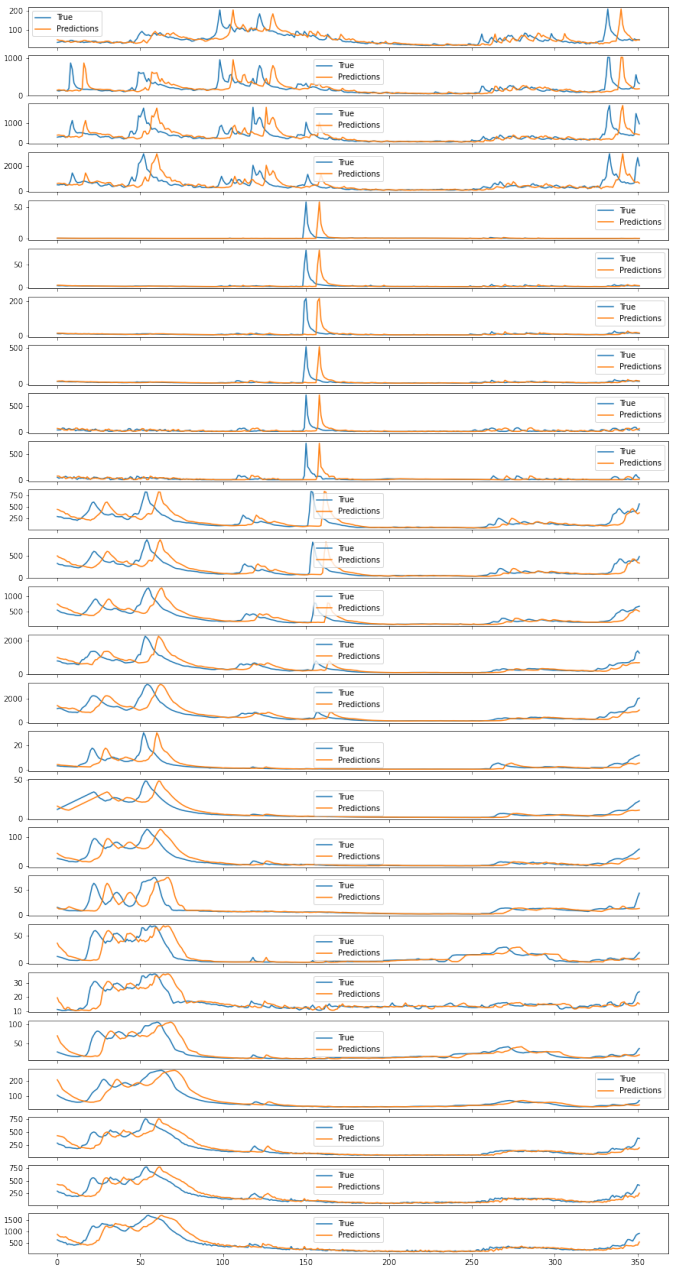


FIGURE B.2 – Prédiction du modèle naif, J+7

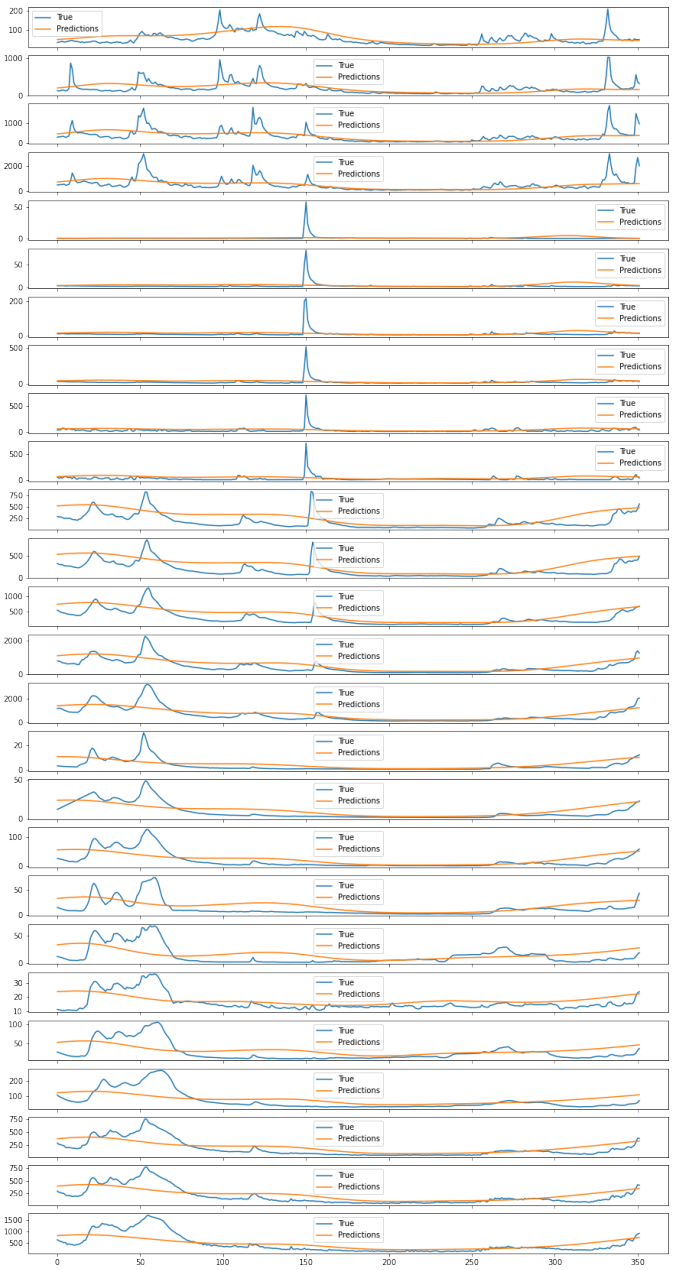


FIGURE B.3 – Prédiction du modèle GAM

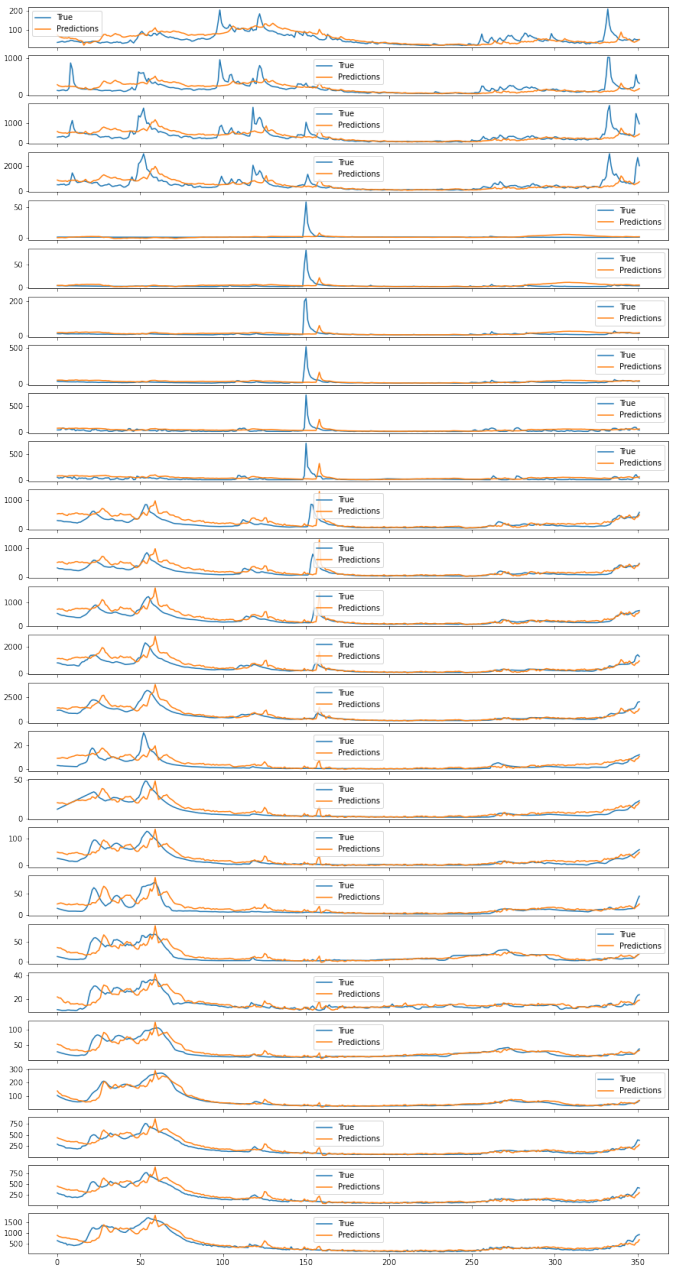


FIGURE B.4 – Prédiction du modèle GAM+VAR

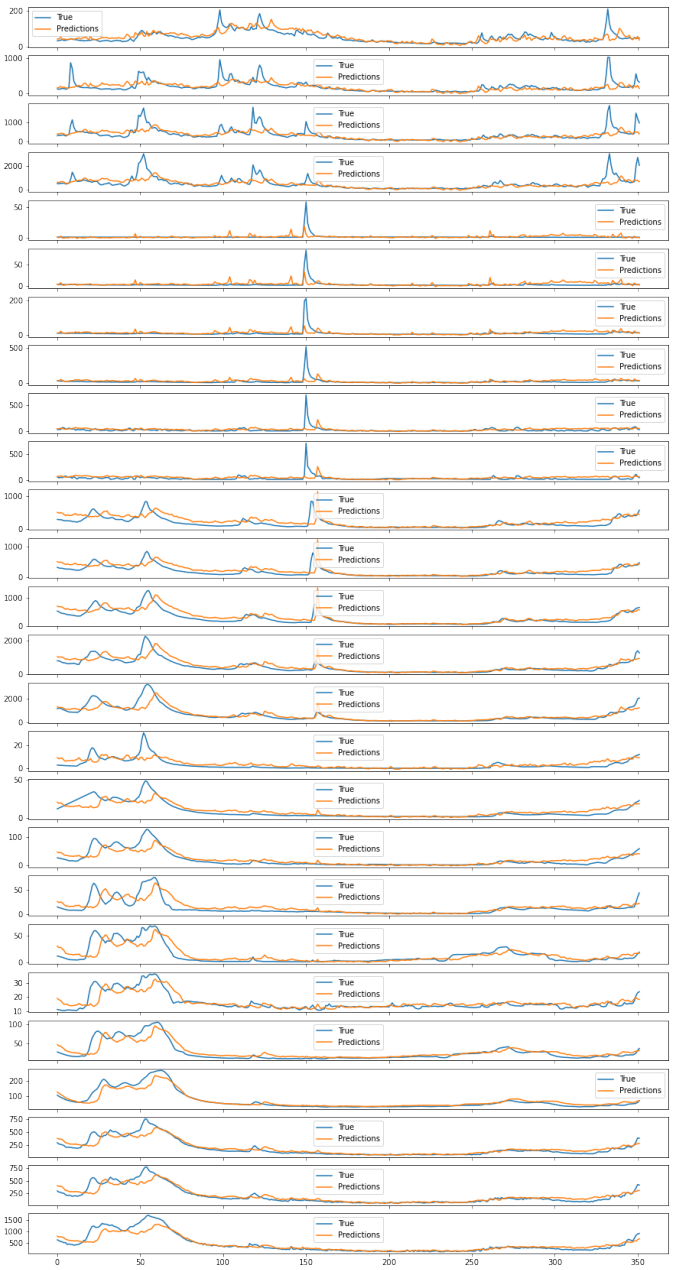


FIGURE B.5 – Prédiction du modèle GAM+LASSO+AR

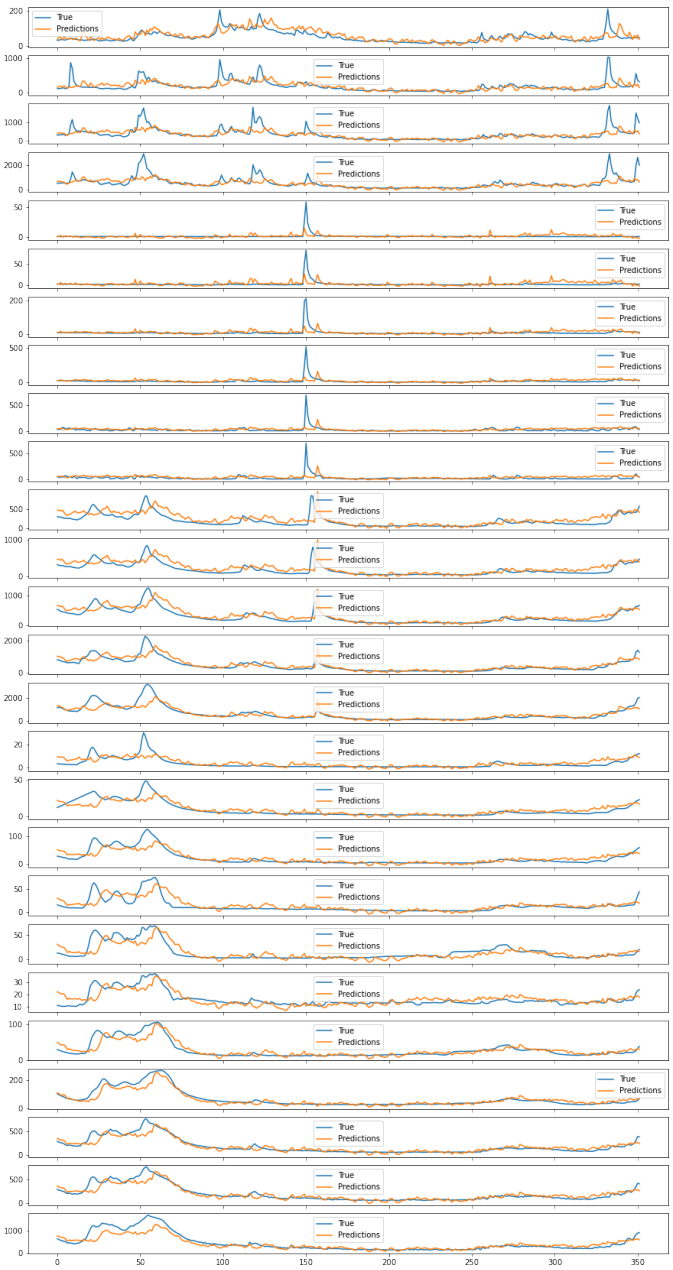


FIGURE B.6 – Prédiction du modèle GAM+PLS+AR

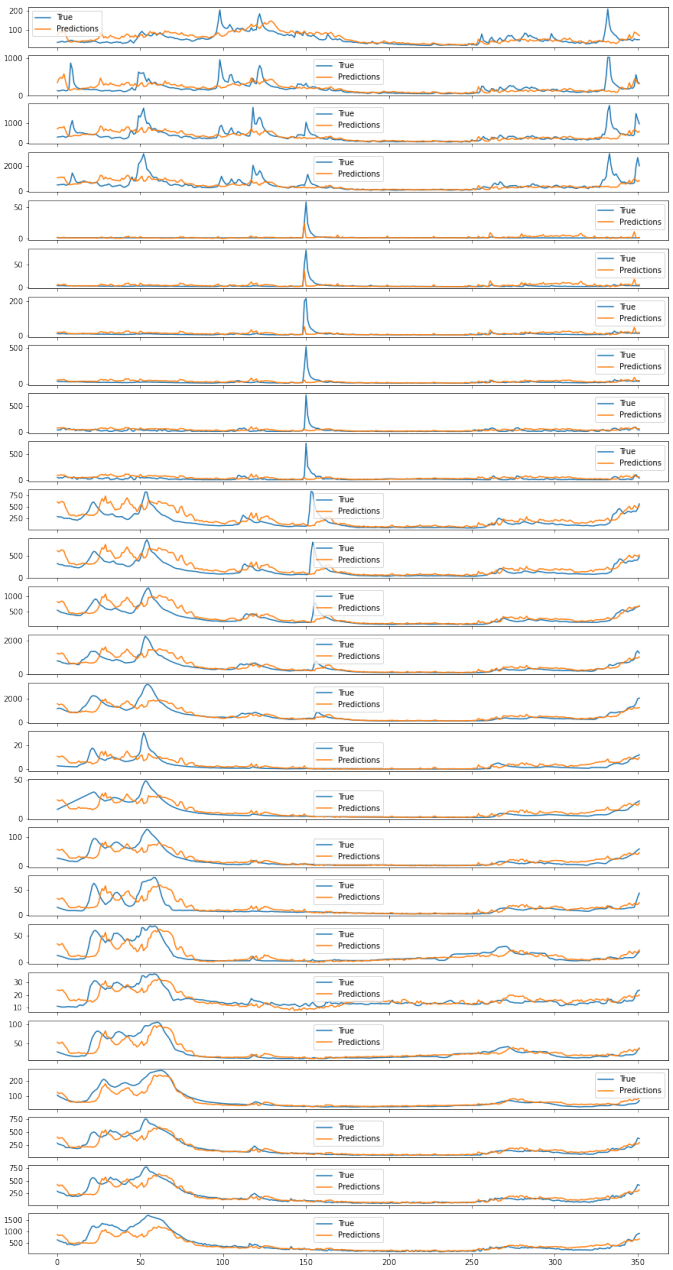


FIGURE B.7 – Prédiction du modèle GAM+RF+AR

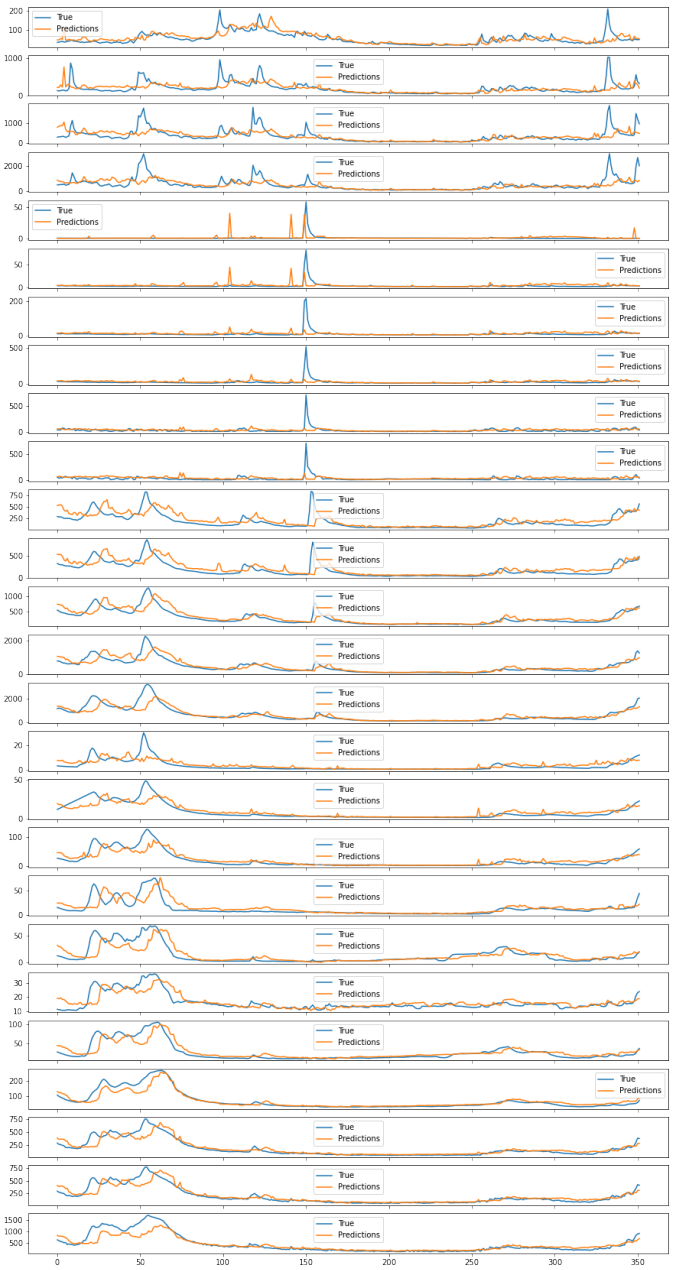


FIGURE B.8 – Prédiction du modèle GAM+GBoost+AR

O0200020 top features		O2620010 top features		O6140010 top features	
Weight	Feature	Weight	Feature	Weight	Feature
0.303	O0200020_0	0.227	O0200020_0	0.162	O0200020_0
0.132	O0200020_1	0.079	O2620010_1	0.138	O9000010_0
0.072	Precipitations_07627	0.057	Vitesse_vent_07627	0.088	O9000010_1
0.055	Precipitations_07621	0.047	O9000010_1	0.048	M5300010_0
0.046	H1700010_1	0.043	Humidite_07621	0.047	Vitesse_vent_07627
0.043	Humidite_07621	0.038	Precipitations_07627	0.04	Humidite_07621
0.042	O9000010_1	... 12 more positive ...		... 14 more positive ...	
... 13 more positive ...		... 4 more negative ...		... 2 more negative ...	
... 4 more negative ...		-0.065	Pression_07761	-0.094	Pression_07761
O9000010 top features		K0030020 top features		K0100020 top features	
Weight	Feature	Weight	Feature	Weight	Feature
0.187	O9000010_0	0.221	Precipitations_07690	0.192	Precipitations_07690
0.134	M5300010_0	0.092	Precipitations_07460	0.088	Precipitations_07471
0.094	O0200020_0	0.067	Vitesse_vent_07643	0.077	Precipitations_07460
0.077	O9000010_1	0.059	Precipitations_07471	0.063	Vitesse_vent_07643
0.052	K0690010_0	0.029	Vent_Est_07761	0.059	K0550010_0
... 11 more positive ...		... 12 more positive ...		0.047	Precipitations_07577
... 1 more negative ...		... 5 more negative ...		... 15 more positive ...	
-0.05	Pression_07761	-0.031	Precipitations_07535	... 4 more negative ...	
-0.069	Pression_07690	-0.098	Vent_Nord_07643	-0.075	Vent_Nord_07643
K0260020 top features		K0550010 top features		K0690010 top features	
Weight	Feature	Weight	Feature	Weight	Feature
0.172	Precipitations_07690	0.125	Precipitations_07690	0.222	K0690010_0
0.124	Precipitations_07471	0.11	K0690010_0	0.082	K0550010_0
0.065	Humidite_07591	0.102	K0550010_0	0.08	K0690010_1
0.062	Humidite_07690	0.081	K0690010_1	0.079	Humidite_07591
0.062	K0550010_0	0.08	Humidite_07591	0.065	Humidite_07690
0.051	Vitesse_vent_07591	0.071	Humidite_07690	... 15 more positive ...	
0.051	K0690010_0	0.068	Precipitations_07471	... 2 more negative ...	
... 16 more positive ...		... 15 more positive ...		-0.054	K0030020_1
... 10 more negative ...		... 8 more negative ...		-0.061	Pression_07761
K0910050 top features		K4180010 top features		K4350020 top features	
Weight	Feature	Weight	Feature	Weight	Feature
0.176	K0690010_0	0.18	K0690010_0	0.172	K0690010_0
0.094	K0550010_0	0.157	L8000020_0	0.156	L8000020_0
0.073	K0690010_1	0.061	O9000010_0	0.07	K0910050_0
0.064	Humidite_07591	0.052	K4180010_0	0.066	H4340020_0
0.059	M5300010_1	0.046	H0400020_0	0.061	K4180010_0
... 16 more positive ...		0.045	K0910050_0	0.049	O9000010_0
... 2 more negative ...		... 14 more positive ...		0.048	K0550010_0
-0.047	K0030020_1	... 2 more negative ...		... 13 more positive ...	
-0.058	Pression_07761	-0.041	K0030020_1	... 2 more negative ...	
K6830020 top features		L8000020 top features		M5300010 top features	
Weight	Feature	Weight	Feature	Weight	Feature
0.183	L8000020_0	0.209	M5300010_0	0.419	M5300010_0
0.133	H4340020_0	0.123	L8000020_0	0.155	O9000010_0
0.094	K0690010_0	0.12	O9000010_0	0.103	H0100010_0
0.093	O9000010_0	0.094	H4340020_0	0.081	K0690010_0
0.089	K0910050_0	0.083	K0690010_0	0.033	K0550010_0
0.061	H0100010_0	0.071	H0100010_0	0.024	H4340020_0
0.055	M5300010_0	... 10 more positive ...		... 6 more positive ...	
... 9 more positive ...		... 1 more negative ...		... 2 more negative ...	
... 2 more negative ...		-0.035	Pression_07761	-0.029	Pression_07761
H0100010 top features		H0100020 top features		H0400010 top features	
Weight	Feature	Weight	Feature	Weight	Feature
0.161	H5920014_0	0.228	H0100010_0	0.256	H0100010_0
0.132	H0100010_0	0.195	H5920014_0	0.234	H5920014_0
0.115	H0800012_0	0.126	H0800012_0	0.129	H0800012_0
0.055	Humidite_07591	0.058	H0800011_0	0.059	H0800011_0
0.035	Vitesse_vent_07627	0.03	Humidite_07591	0.03	Vitesse_vent_07627
0.025	Vitesse_vent_07650	0.023	Vitesse_vent_07627	0.03	Humidite_07591
... 8 more positive ...		0.021	Vitesse_vent_07650	0.016	Vitesse_vent_07650
... 9 more negative ...		... 4 more positive ...		... 4 more positive ...	
-0.027	Vent_Est_07280	... 4 more negative ...		... 4 more negative ...	
H0400020 top features		H0800011 top features		H0800012 top features	
Weight	Feature	Weight	Feature	Weight	Feature
0.238	H4340020_0	0.533	H0800011_0	0.685	H0800012_0
0.211	H0100010_0	0.155	H0100010_0	0.078	H0100010_0
0.197	H0400020_0	0.147	H4340020_0	0.059	H5920014_0
0.087	H0800011_0	0.023	Humidite_07591	0.014	K0690010_0
0.048	H0800012_0	0.015	Vitesse_vent_07627	... 6 more positive ...	
0.039	Humidite_07591	... 6 more positive ...		... 4 more negative ...	
... 5 more positive ...		... 2 more negative ...		-0.015	Pression_07190
... 2 more negative ...		-0.017	K0030020_1	-0.016	Vent_Est_07471
-0.028	K0030020_1	-0.027	Vent_Est_07130	-0.02	Vent_Est_07255
H0810010 top features		H1700010 top features		H3930020 top features	
Weight	Feature	Weight	Feature	Weight	Feature
0.42	H0800011_0	0.336	H0800011_0	0.278	H5920014_0
0.194	H4340020_0	0.193	H5920014_0	0.228	H4340020_0
0.17	H0800012_0	0.189	H1700010_0	0.136	H0800011_0
0.142	H0100010_0	0.142	H0100010_0	0.098	H0800012_0
0.01	Humidite_07591	0.063	H0800012_0	0.034	H0100010_0
... 4 more positive ...		0.059	H4340020_0	0.022	Humidite_07591
-0.005	Pression_07690	0.005	O0200020_0	0.015	Humidite_07149
-0.021	Vent_Est_07130	... 2 more positive ...		... 4 more positive ...	
				... 4 more negative ...	
H4340020 top features		H5920014 top features		H8100021 top features	
Weight	Feature	Weight	Feature	Weight	Feature
0.293	H5920014_0	0.587	H5920014_0	0.487	H5920014_0
0.243	H4340020_0	0.094	H0800012_0	0.156	H8100021_0
0.108	H0800011_0	0.074	H0100010_0	0.081	H0100010_0
0.095	H0800012_0	0.04	H0800011_0	0.058	H0800012_0
0.025	Humidite_07591	0.023	Humidite_07591	0.017	Humidite_07591
0.024	H0100010_0	0.011	Humidite_07149	... 7 more positive ...	
0.02	Humidite_07149	0.011	Vitesse_vent_07627	... 1 more negative ...	
... 5 more positive ...		... 4 more positive ...		-0.015	Pression_07690
... 4 more negative ...		... 4 more negative ...		-0.022	Vent_Est_07255

TABLE B.1 – Importances des features de GAMLASSOAR

Bibliographie

[1] Jean Cunge. Modèles mathématiques en hydraulique et en hydrologie. *Techniques de l'Ingénieur*, 1995.

[2] Frédéric Jordan. *Modèle de prévision et de gestion des crues : optimisation des opérations des aménagements hydroélectriques à accumulation pour la réduction des débits de crue.* PhD thesis, EPFL, 2007.

[3] Florine Garcia. *Amélioration d'une modélisation hydrologique régionalisée pour estimer les statistiques d'étiage.* Theses, Université Pierre et Marie Curie - Paris VI, December 2016.

[4] Mauro Naghettini (eds.). *Fundamentals of Statistical Hydrology.* Springer International Publishing, 2017.

[5] Rajib Maity. Statistical methods in hydrology and hydroclimatology. *Springer Transactions in Civil and Environmental Engineering*, 2018.

[6] Benjamin Renard. Probabilités et Statistiques appliquées à l'Hydrologie. Lecture, 2014.

[7] Hang Zeng, Jiaqi Huang, Zhengzui Li, Weihou Yu, and Hui Zhou. Nonstationary bayesian modeling of extreme flood risk and return period affected by climate variables for xiangjiang river basin, in south-central china. *Water*, 14(1), 2022.

[8] Amir Mosavi, Pinar Ozturk, and Kwok-wing Chau. Flood prediction using machine learning models : Literature review. *Water*, 10(11), 2018.

[9] DEEPESH MACHIWAL and MADAN K. JHA. Comparative evaluation of statistical tests for time series analysis : application to hydrological time series / evaluation comparative de tests statistiques pour l'analyse de séries temporelles : application à des séries temporelles hydrologiques. *Hydrological Sciences Journal*, 53(2) :353–366, 2008.

[10] Muhammad Ali Musarat, Wesam Salah Alaloul, Muhammad Babar Ali Rabbani, Mujahid Ali, Muhammad Altaf, Roman Fediuk, Nikolai Vatin, Sergey Klyuev, Hamna Bukhari, Alishba Sadiq, Waqas Rafiq, and Waqas Farooq. Kabul river flow prediction using automated arima forecasting : A machine learning approach. *Sustainability*, 13(19), 2021.

[11] Paulin Coulibaly, François Anctil, and Bernard Bobée. Prévision hydrologique par réseaux de neurones artificiels : état de l'art. *Canadian Journal of Civil Engineering*, 26(3) :293–304, 1999.

[12] Haytham Assem, Salem Gharbia, Gábor Makrai, Paul Johnston, Laurence Gill, and Francesco Pilla. *Urban Water Flow and Water Level Prediction Based on Deep Learning*, pages 317–329. Springer International Publishing, 01 2017.

[13] Mu Lin Ha Si, Liu Darong. Prediction of yangtze river streamflow based on deep learning neural network with el niño–southern oscillation. *Scientific Reports*, 2021.

[14] Zaher Mundher Yaseen et al. Hybrid data intelligent models and applications for water level prediction. *Handbook of Research on Predictive Modeling and Optimization Methods in Science and Engineering*, 2018.

[15] Mohamed Samir Toukourou. *Application de l'apprentissage artificiel à la prévision des crues éclair.* PhD thesis, École Nationale Supérieure des Mines de Paris, 2009.

[16] Thomas Darras. *Prévision de crues rapides par apprentissage statistique.* Theses, Université Montpellier, November 2015.

[17] Zahra Alizadeh, Jafar Yazdi, Joong Hoon Kim, and Abobakr Khalil Al-Shamiri. Assessment of machine learning techniques for monthly flow prediction. *Water*, 10(11), 2018.

[18] Amir Sahraei, Alejandro Chamorro, Philipp Kraft, and Lutz Breuer. Application of machine learning models to predict maximum event water fractions in streamflow. *Frontiers in Water*, 3 :52, 2021.

[19] Julien Zhou and Wendong Liang. Github repository for the project. https://github.com/JlnZhou/StatML_ProjectML.

[20] Api hydrométrie. https://hubeau.eaufrance.fr/page/api-hydrometrie.

[21] Banque hydro. http://www.hydro.eaufrance.fr.

[22] Observation météorologique historiques france (synop). https://public.opendatasoft.com/explore/dataset/donnees-synop-essentielles-omm/information/.

[23] Trevor J Hastie and Robert J Tibshirani. *Generalized additive models*. Routledge, 2017.

[24] Michel Tenenhaus. *La régression PLS*. Technip, 1995.

[25] Adele Cutler Leo Breiman. Random forests. https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm.