OPUR : Observatoire des Polluants URbains

ACTION N° R2.4
MODÉLISATION DES FLUX DE MATIÈRES DANS LE RÉSEAU D'ASSAINISSEMENT DU SIAAP

Rapport d'activité

Auteurs : S. Laruelle

Ce travail est une action commune au programme OPUR et au programme MOCOPEE (Bloc 2.3 - Traitements des données et modélisation statistique des stations d'épuration des eaux usées et des processus connexes - Action n°2.3.2 – Construire des modèles statistiques fondés sur des séries de données environnementales, en vue de prédire la dynamique de la qualité des intrants des systèmes de traitement).

Ce travail est issu d'une collaboration entre le LAMA (UPEC-UGE-CNRS) et le SIAAP et réalisé avec le concours financier des partenaires opérationnels d'OPUR (AESN, SIAAP, Ville de Paris, CG93, CG94).

MODÉLISATION DES SÉRIES DE QUALITÉ DES EAUX USÉES À L'ENTRÉE DE SEINE-AVAL

Sophie LARUELLE (LAMA-UPEC)

DATE DE PUBLICATION DU RAPPORT : 1/12/2023

A l'aide de données de capteurs à 15 minutes à l'entrée de Seine-Aval, de données similaires à la station de Clichy, de données réseaux et de pluviométrie, le but de cette action est de construire un outil d'aide à la décision en ligne pour les exploitants de la station de Seine-Aval qui prédira à horizon 6 heures la qualité des eaux usées à venir pour optimiser leurs traitements.

1 Introduction

Ce rapport présente un résumé des travaux de validation des données et d'étude des corrélations entre les variables pour les données de capteurs à 15 minutes à l'entrée de Seine-Aval, celles de Clichy à des fréquences de 5 minutes ou 15 minutes selon les variables et celle de pluviométrie de janvier 2018 à décembre 2021.

Une première tâche a été d'analyser les données pour en tirer une série « propre » en vue de la modélisation. Des plages « raisonnables » pour les différentes données ont été établies à l'aide de statistiques et d'avis d'experts, et on a enlevé les fortes variations (2% de la distribution des accroissements). Puis l'analyse des autocorrélations des séries et des corrélations croisées avec décalage est étudiée pour orienter la modélisation vers des séries temporelles multivariées et estimer le pouvoir prédictif de la station de Clichy en amont du réseau sur les données de Seine-Aval.

2 Sélection des données pour la modélisation

Pour la station de Seine-Aval, on dispose des données suivantes à une fréquence de 15 minutes : conductivité (en µS/cm), matières en suspension (en mg/L), pH, température (en °C), débit total (en m3/s) et débits des 5 émissaire s(en m3/s) du 1er janvier 2018 au 31 décembre 2021.

Pour la station de Clichy, on dispose des données suivantes à une fréquence de 5 minutes (à l'exception de l'ammonium à 15 minutes) : conductivité (en µS/cm), turbidité (en NFU), pH, température (en °C), ultraviolets (unité inconnue) et ammonium (en mg/L) du 1er janvier 2018 au 31 octobre 2021.

Pour compléter ces mesures, nous disposons également de données horaires de pluviométrie collectées par la DSAR du SIAAP en mm par bassin versant.

3 Validation des données de capteurs de Seine-Aval et Clichy

3.1 Présentation des données brutes de Seine-Aval

La figure 1. ci-dessous présente les 4 séries de conductivité, MeS, température et pH de Seine-Aval du 1er janvier 2018 au 31 décembre 2021.

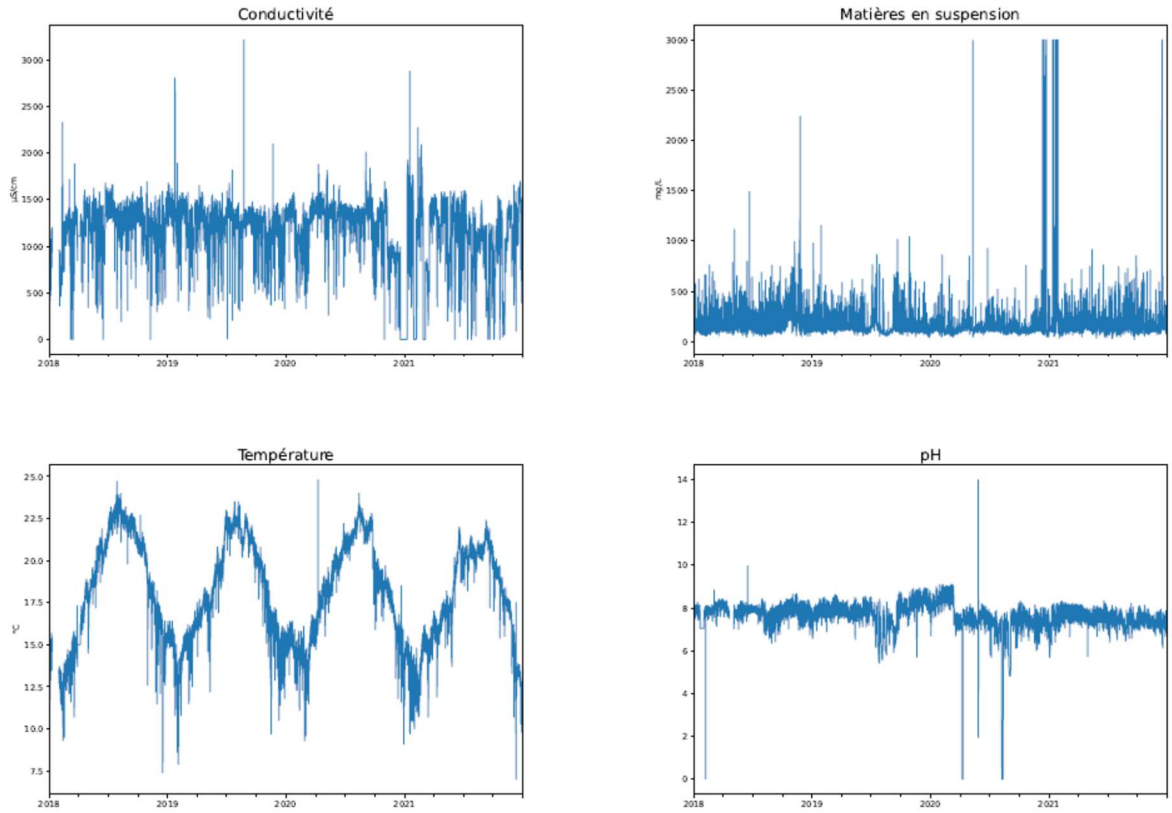

Figure 1. Séries temporelles des données de capteurs.

La figure 2. ci-dessous présente les 6 séries de débits des 5 émissaires et total de Seine-Aval du 1er janvier 2018 au 31 décembre 2021.

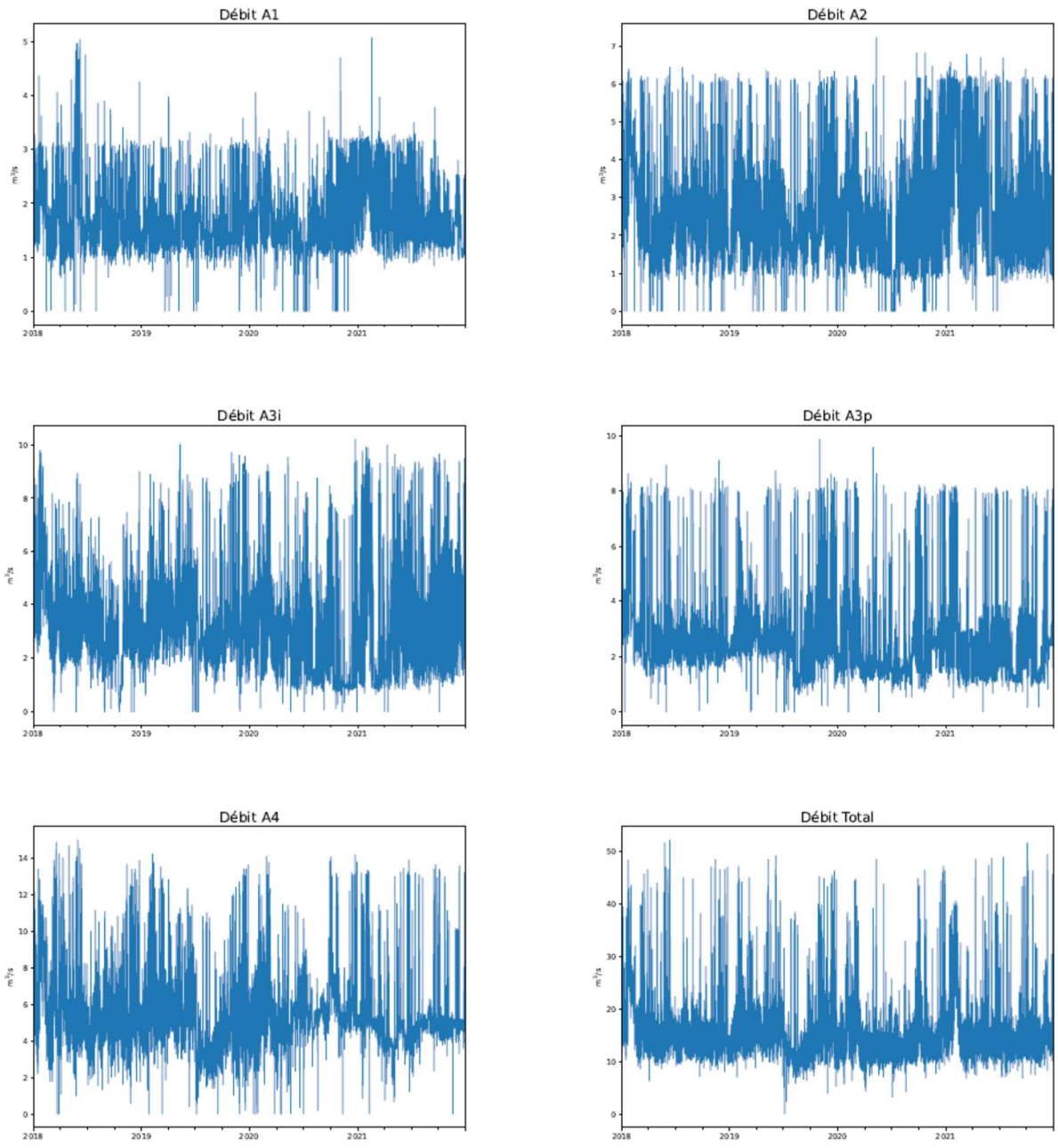

Figure 2. Séries temporelles des données de capteurs de débits.

3.2 Validation des données brutes de Seine-Aval

Au regard de cette table et d'avis d'expert, nous avons choisi les plages suivantes pour les données :
— conductivité : entre 200 et 1 600 ;
— matières en suspension : entre 120 et 570 ;
— pH : entre 6 et 9 ;
— température : entre 10 et 25.

Les débits ne seront pas concernés par la validation. Toutes les valeurs en dehors de ces bornes sont remplacées par une interpolation linéaire entre les autres dates. On obtient alors les résultats suivants

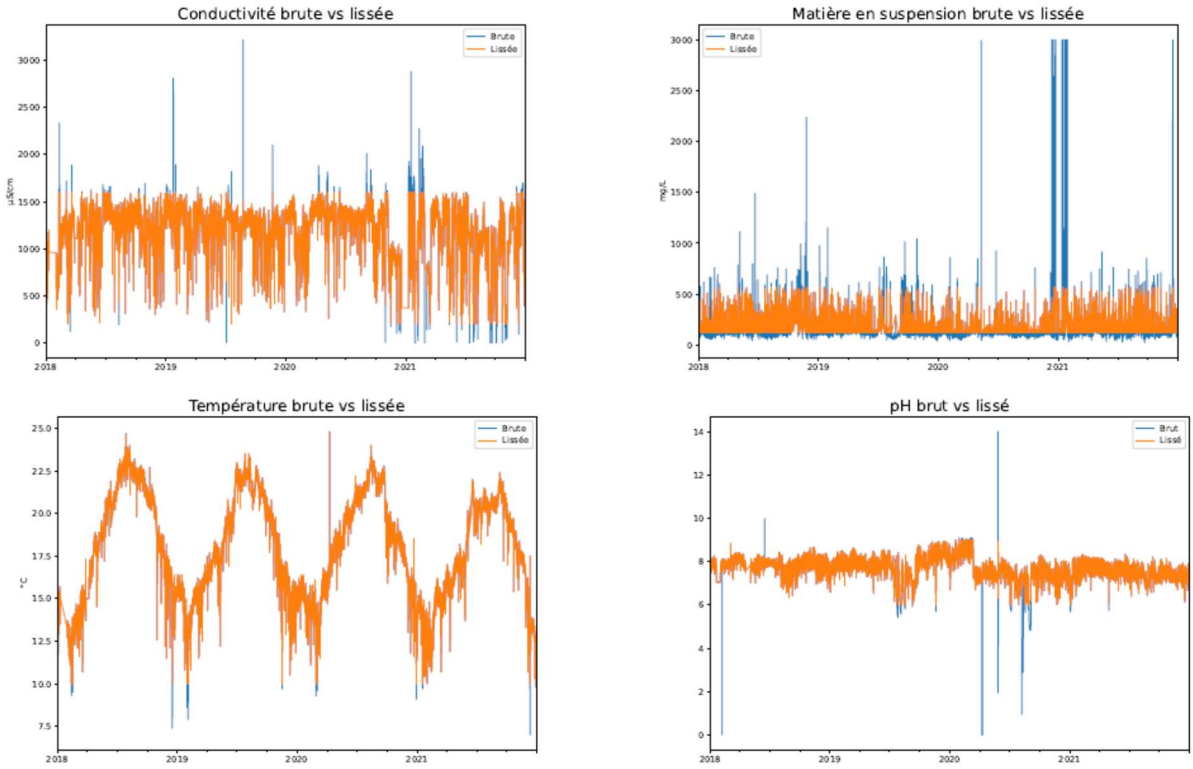

Figure 3. Séries temporelles des données de Seine-Aval lissées.

On calcule les accroissements des séries lissées à fréquence de 15 minutes, c'est-à-dire la différence de deux valeurs consécutives. On observe encore quelques pics sûrement dus à des valeurs aberrantes. Nous allons regarder de plus près leurs distributions pour les retirer.

	Condu (μS/cm)	pH	Temp (°C)	MeS (mg/L)
Moyenne	0,00	0,00	0,00	0,00
Ecart-type	24,91	0,044	0,087	21,043
Min	−1 097,00	−0,990	−8,400	−418,000
$q_{1\%}$	−65,00	−0,11	−0,20	−57,00
$q_{25\%}$	−5,00	−0,010	0,000	−2,113
$q_{50\%}$	0,00	0,00	0,00	0,00
$q_{75\%}$	5,00	0,010	0,000	2,000
$q_{99\%}$	59,00	0,14	0,20	66,00
Max	1 075,00	1,680	8,600	421,000

Au vu de la table statistique, nous décidons de retirer les accroissements supérieurs au quantile à 99% et inférieurs à celui à 1%. Nous remplaçons les données par leur interpolation linéaire. Les résultats obtenus sont présentés dans la figure suivante.

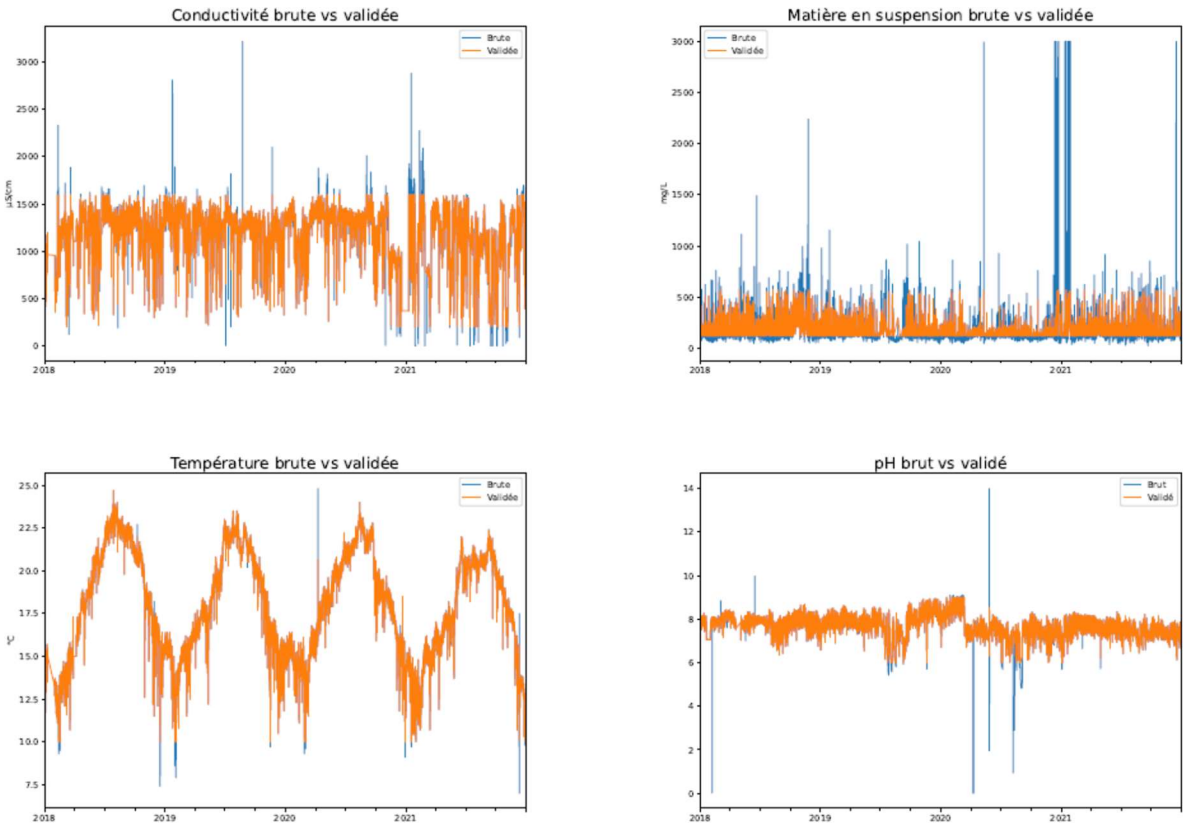

Figure 4. Séries temporelles des données de Seine-Aval validées.

On remarque dans la série du pH un décrochage en 2020, sûrement dû à un changement de sonde ou à un rééchantillonnage, qui pourra influencer l'analyse des corrélations entre les variables.

3.3 Validation des données brutes de Clichy

Les données de Clichy (à l'exception de l'ammonium) sont mesurées à une fréquence de 5 minutes. Pour avoir la même échelle de temps que pour les données de Seine-Aval, on procède tout d'abord à un rééchantillonnage à la fréquence de 15 minutes en prenant la moyenne des 3 valeurs pour la conductivité, la turbidité, la température, le pH et les ultraviolets. Les valeurs nulles ou négatives seront considérées comme manquantes.

Nous procèderons ensuite de la même façon que pour Seine-Aval pour la validation de ces données.

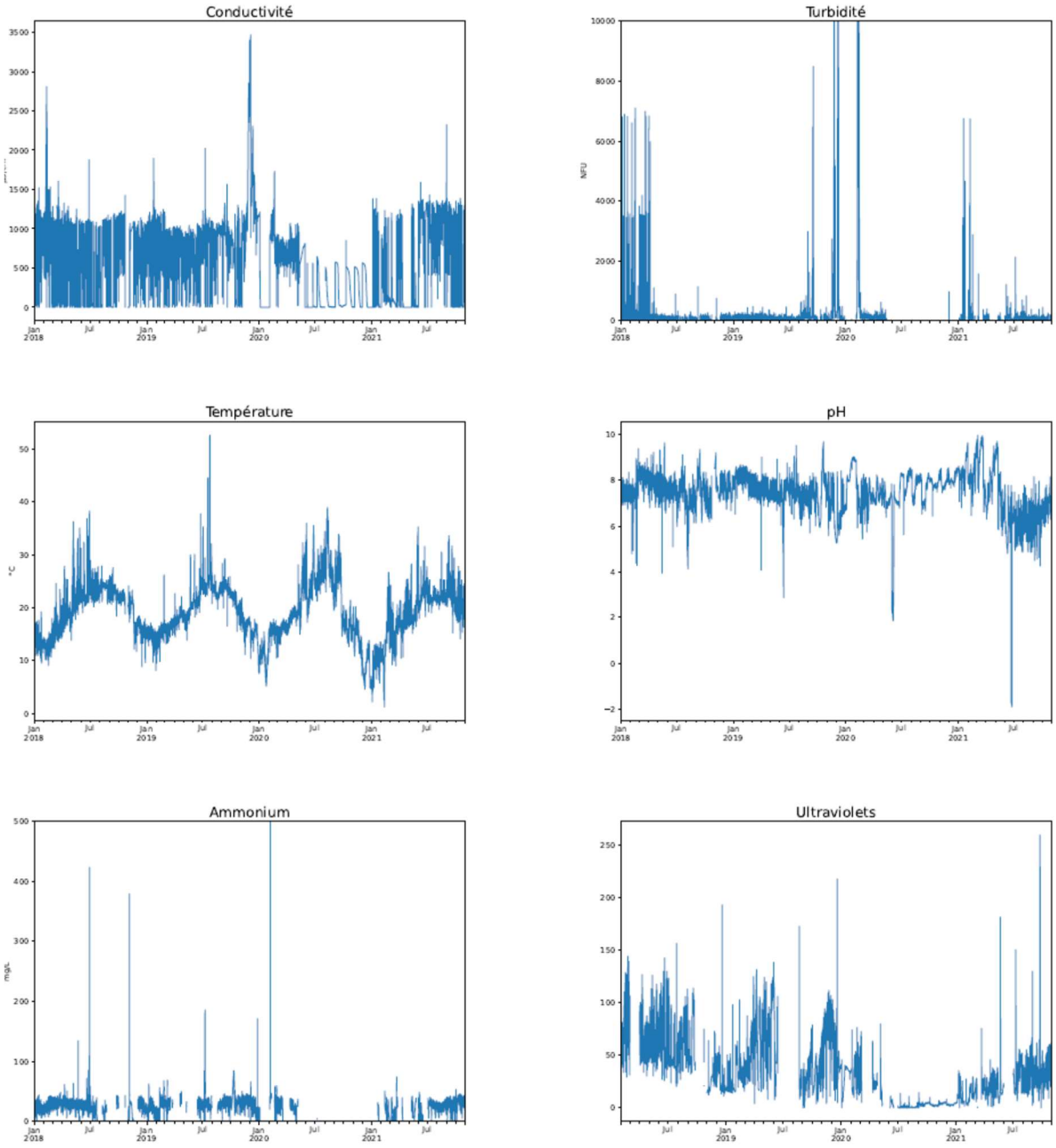

Figure 5. Séries temporelles des données de Clichy brutes.

On observe ici de très fortes valeurs, de nombreux pics et de nombreuses données manquantes qui vont rendre la validation un peu plus compliquée (surtout pour l'ammonium). Nous regardons les distributions de ces séries (histogrammes), puis une table statistique pour déterminer les plages de valeurs retenues, tout en maintenant une cohérence avec celles obtenues pour Seine-Aval.

3.4 Validation des données brutes de Clichy

Nous disposons de 134 400 données sur 2 ans et 10 mois (du 1er janvier 2018 au 31 octobre 2021) à une fréquence de 15 minutes.

Au regard de cette table et par cohérence avec les données de Seine-Aval, nous avons choisi les plages suivantes pour les données :
— conductivité : entre 200 et 1 600 ;
— turbidité : entre 2 et 600 ;
— pH : entre 6 et 9 ;
— température : entre 10 et 30 ;
— ammonium : entre 11 et 45 ;
— ultraviolets : entre 0 et 100.

Toutes les valeurs en dehors de ces bornes sont remplacées par une interpolation linéaire entre les autres dates. Les résultats obtenus sont présentés ci-dessous

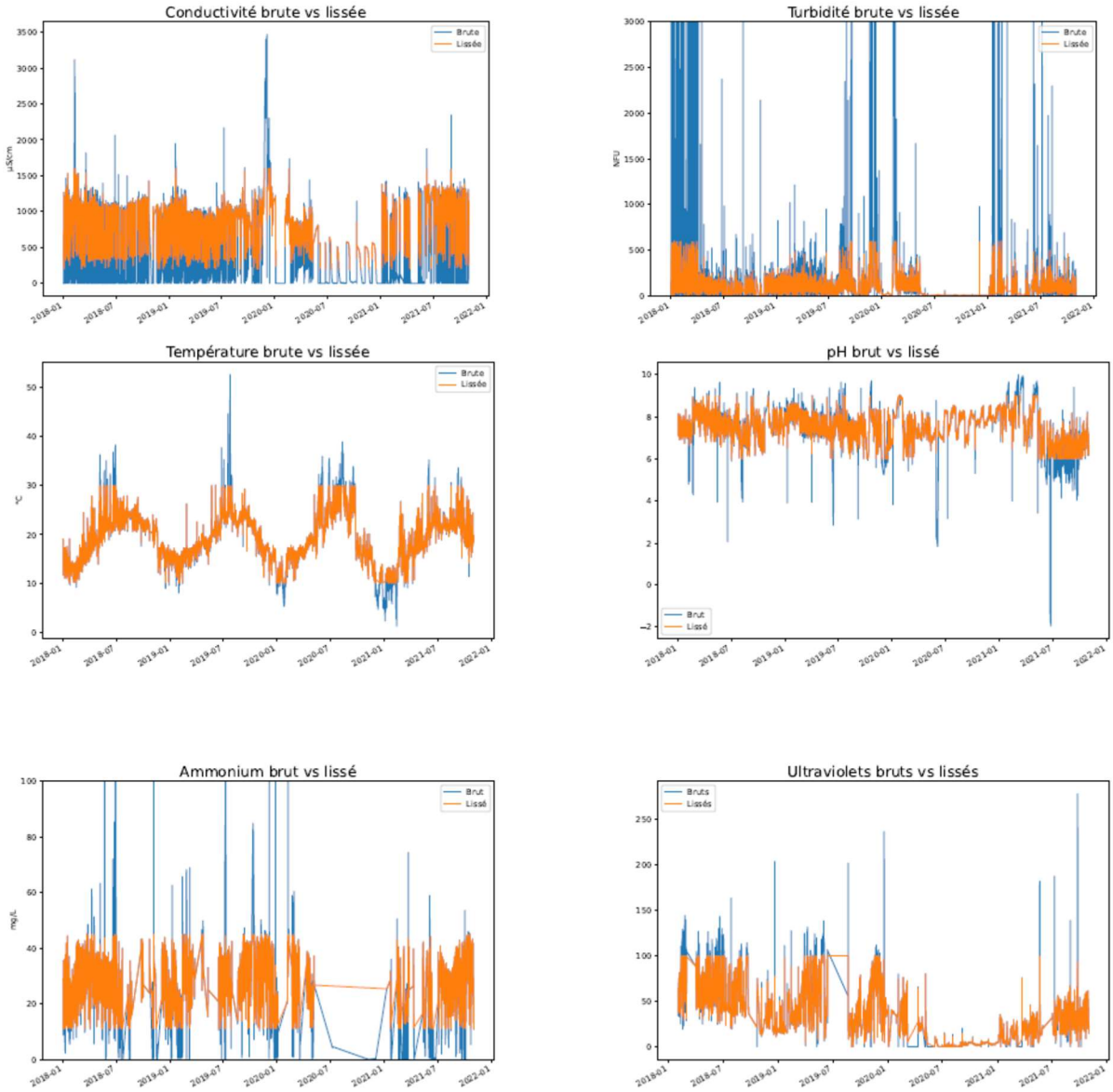

Figure 6. Séries temporelles des données de Clichy lissées.

Pour compléter la validation, nous allons regarder maintenant les accroissements pour supprimer les trop fortes variations.
On calcule les accroissements des séries lissées à fréquence de 15 minutes, c'est-à-dire la différence de deux valeurs consécutives.

	Condu (μS/cm)	pH	Temp (°C)	Turbi (NFU)	NH_4 (mg/L)	UV
Moyenne	0,01	0,0	0,0	0,0	0,0	0,0
Ecart-type	140,06	0,06	0,13	28,44	0,71	1,2
Min	−991,64	−1,7	−6,63	−552,58	−16,48	−55,55
$q_{1\%}$	−376,49	−0,16	−0,25	−63,58	−2,02	−2,59
$q_{25\%}$	−1,86	−0,01	−0,04	−1,36	−0,03	−0,08
$q_{50\%}$	0,13	0,0	−0,01	0,0	−0,0	0,0
$q_{75\%}$	2,56	0,01	0,03	1,52	0,03	0,06
$q_{99\%}$	374,34	0,17	0,35	63,88	2,0	2,59
Max	1144,59	1,58	5,49	555,76	17,83	55,46

Les données de Clichy présentent de plus fortes variations que celle de Seine-Aval, nous décidons alors de retirer les accroissements supérieurs au quantile à 99% et inférieurs à celui à 1%. Nous remplaçons les données par leur interpolation linéaire. Les résultats obtenus sont présentés dans la figure suivante.**

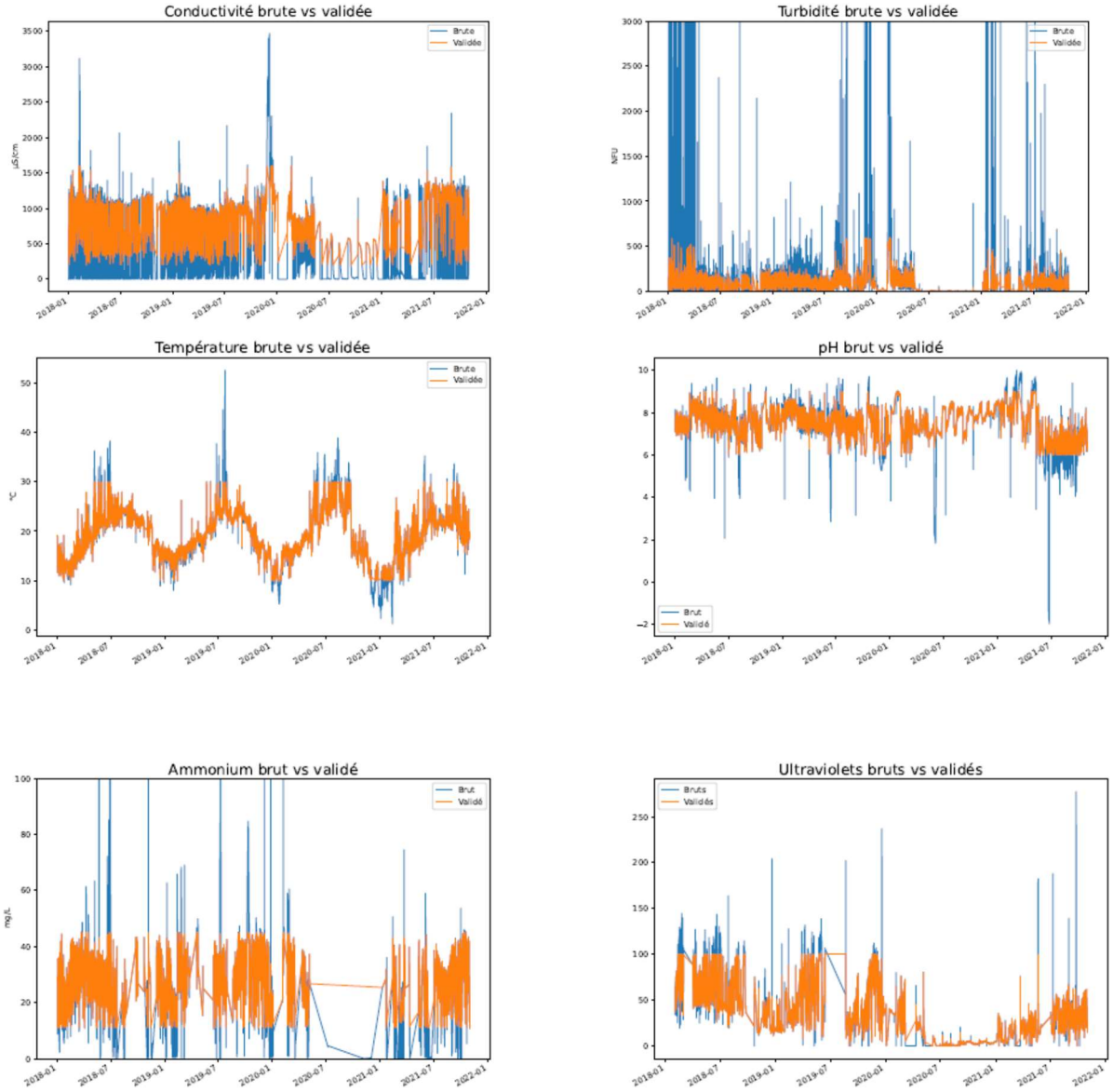

Figure 7. Séries temporelles des données de Clichy validées.

4 Autocorrélations et corrélations croisées entre les données

Dans ce chapitre, nous allons analyser les autocorrélations des séries, de leurs accroissements et les corrélations entre les séries avec décalage temporel pour voir si les données de la station de Clichy qui se trouve en amont du réseau peut nous aider à prédire la qualité des eaux arrivant à Seine-Aval.

Pour une série temporelle $(X_t)_t$ nous définissons la fonction d'autocovariance comme suit

$$\gamma(h) = Cov\big(X_t, X_{\{t+h\}}\big), \quad h \in Z,$$

et sa fonction d'autocorrélation

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)}, \quad h \in Z.$$

Pour deux séries temporelles $(X_t)_t$ et $(Y_t)_t$, nous définissons la fonction de corrélation croisée avec décalage temporel comme suit

$$\eta(h) = \frac{Cov\big(X_t, Y_{\{t+h\}}\big)}{\sigma_X \ \ \sigma_Y}, \quad h \in Z,$$

où σ_X et σ_Y sont les écart-type de chacune des séries. Pour h = 0, on a donc la corrélation entre les deux séries, pour h < 0 on mesure l'influence du passé de Y sur X, et pour h > 0, on mesure celle de X sur le futur de Y.

Pour observer les saisonnalités annuelles et inférieures, nous allons choisir de calculer les autocorrélations des séries sur une période d'une journée, soit pour des décalages h entre −96 et 96.

Cette analyse est à compléter d'ici le début de l'année prochaine pour comprendre le pouvoir prédictif de Clichy sur Seine-Aval et pour mieux comprendre la structure de corrélations entre les variables sur chaque station d'épuration (notamment regarder des corrélations à des périodes de plus longue portée comme l'année).

5 Conclusion et Perspectives

Cette étude a permis de mettre en évidence la nécessité de valider les données avant de faire de la modélisation et de prendre en considération les corrélations croisées entre les séries et donc d'orienter la modélisation vers des séries vectorielles. Cependant, il reste encore à approfondir la modélisation en analysant les saisonnalités, les tendances, l'utilisation des modèles de type ARMA-GARCH et leurs résidus. Ensuite, il faudra analyser la stabilité du modèle : combien de données à utiliser pour l'estimer ? Les coefficients sont-ils stables dans le temps ou faut-il réestimer le modèle régulièrement pour s'adapter aux nouvelles données ? Enfin une étude plus poussée de la qualité des prévisions est à faire, en considérant le plus de périodes possibles pour avoir des intervalles de confiance et leur distribution.

Il faudra aussi coupler toutes les données de Seine-Aval avec celles de Clichy et de pluviométrie pour améliorer la qualité des prévisions de la qualité des eaux usées à l'entrée de Seine-Aval. Une thèse a démarré en octobre 2023 (Mme Innocentia Sossou est la doctorante), elle permettra d'exploiter toutes ces données pour construire un outil d'aide à la décision en ligne pour les exploitants de la station d'épuration.