OPUR : Observatoire d'hydrologie urbaine en Île de France
Thème de recherche R1: Qualité microbiologique des eaux pluviales
Action de recherche :
- Action R1.1 : Sources et flux de pathogènes dans les rejets pluviaux
- Action R1.2 : Modélisation du bassin de la Villette

OPTIMISATION ET INCERTITUDE DE LA PREDICTION DES CONTAMINATIONS MICROBIOLOGIQUES DES EAUX DE SURFACE LORS DES EVENEMENTS PLUVIEUX

Rapport final
Thèse de doctorat de Naloufi Manel
Date (Mars, 2025)

- *Thèse réalisée au Laboratoire Eau Environnement et Systèmes Urbains (Leesu), sous la direction de Françoise Lucas et à la Direction de la Propreté et de l'Eau de la Ville de Paris - Service Technique de l'Eau et de l'Assainissement en collaboration avec le Laboratoire Image, Signaux et Systèmes Intelligents (LiSSi). La thèse est commune à plusieurs programmes de recherche : MeSeine Innovation, Piren-Seine et ForBath.*

UNIVERSITÉ PARIS EST

École Doctorale : Sciences, Ingénierie et Environnement

THÈSE DE DOCTORAT

Sciences et Techniques de l'Environnement

Optimisation et incertitude de la prédiction des contaminations microbiologiques des eaux de surface lors des événements pluvieux

---

présentée par

Manel Naloufi

Thèse réalisée au Laboratoire Eau Environnement et Systèmes Urbains (Leesu) Dirigée par Françoise Lucas et à la Direction de la Propreté et de l'Eau de la Ville de Paris - Service Technique de l'Eau et de l'Assainissement en collaboration avec le Laboratoire Image, Signaux et Systèmes Intelligents (LiSSi).

Soutenue le 12 mars 2025 à l'Université Paris-Est Créteil

Nicolas Flipo	Centre de Géosciences MINES Paris	Rapporteur
Emma Rochelle-Newall	Institut de recherche pour le développement	Rapporteur
Gilles Varrault	Université Paris Est-Créteil	Président du jury
Sébastien Wurtzer	Eau de Paris	Examinateur
Claire Beyeler	Métropôle du Grand Paris	Examinatrice
Sabine Herbin	Agence Nationale de Sécurité Sanitaire	Examinatrice
Françoise Lucas	Université Paris Est-Créteil	Directrice
Thiago Abreu	Université Paris Est-Créteil	Co-encadrant
Paul Kennouche	Ville de Paris	Co-encadrant
Marion Delarbre	Ville de Paris	Co-encadrante

Remerciements

C'est avec un grand plaisir et une immense gratitude que je réserve ces quelques lignes à tous ceux qui m'ont aidé à la réalisation de ce travail.

Avant toute chose, je tiens à exprimer ma plus profonde gratitude à Françoise Lucas, grâce à qui tout ce projet a été possible, pour sa patience, ses conseils précieux et ses encouragements tout au long de cette thèse ainsi que le soutien personnel qui m'a permis de réaliser ma thèse sereinement. Ces quatre années enrichissantes m'ont fait grandir. Je remercie également chaleureusement Thiago Abreu, Sami Souihi, Paul Kennouch et Marion Delarbre qui m'ont pris sous leurs tutelles et qui ont veillé, par leurs précieux conseils, à l'aboutissement de ce projet. Je leur suis très reconnaissante de tout le temps qu'ils m'ont consacré. Je remercie également Miguel Gillon-Ritz pour son accueil chaleureux à la Ville de Paris et ses conseils avisés.

Je souhaite commencer par remercier les membres de mon jury de thèse composé de Nicolas Flipo, Emma Rochelle-Newall , Claire Beyeler, Sébastien Wurtzer, Sabine Herbin et Gilles Varrault de m'avoir fait l'honneur d'évaluer ce travail de thèse. Je les remercie également pour le temps qu'ils ont consacré à la lecture de ce manuscrit. Je tiens également à remercier l'Université Paris-Est, et plus particulièrement l'école doctorale Sciences, Ingénierie et Environnement et le Leesu pour avoir rendu possible cette thèse.

Mes remerciements s'étendent à la Ville de Paris (Direction de la propreté et de l'eau – Service technique de l'eau et de l'assainissement), qui a contribué au financement et au fonctionnement du projet, ainsi qu'à l'Association Nationale de la Recherche et de la Technologie. Les analyses ont également été rendues possibles grâce aux programmes de recherche MeSeine Innovation, OPUR, Piren-Seine et ForBath.

Je remercie également Aurélie Janne et Christophe Debare, du Syndicat Marne Vive pour ses conseils pertinents et pour avoir facilité l'accès aux données de suivi estival de la Marne. De même je suis reconnaissante envers la Ville de Paris, Eau de Paris, les conseils départementaux de la Seine Saint Denis et du Val de Marne pour leur contribution à l'ensemble des données utilisées. Je remercie le Service des canaux de la ville de Paris pour l'accès aux sites de surveillance du Bassin de la Villette. Ainsi qu'au Club de voile VGA de Saint-Maur-des-Fossés, qui a permis l'installation des préleveurs et des capteurs. Je remercie le Syndicat Intercommunale d'Assainissement de Marne-la-Vallée de nous avoir fournis des échantillons de la station de traitement des eaux usées de St-Thibault-des-Vignes. Je remercie tous les

intervenants et gestionnaires du bassin de rétention de Sucy-en-Brie et de l'ouvrage cadre du centre urbain de Noisy-le-Grand de nous avoir permit l'installation des préleveurs automatiques pour nos analyses.

Mes remerciements vont également à Jean-Marie Mouchel (Université Sorbonne) pour la mise à disposition de la sonde d'oxygène dissous et Mohamed Aymen Labiod (LiSSi) pour l'aide apportée à l'installation et à la configuration des passerelles. J'exprime mes remerciements à Lamine Amour (ESME) pour m'avoir aidée avec les modèles d'apprentissage automatique. Je remercie également Laurent Moulin, Marion Goulet et Sébastien Wurtzer (Eau De Paris) pour l'analyse des données virologiques de nos campagnes de prélèvement.

Évidemment, cette thèse n'aurait été possible sans l'appui des membres du Leesu, qui m'ont mise dans de bonnes conditions pour accomplir ce travail et m'impliquer dans la vie du laboratoire. Toutes ces analyses n'auraient pas été possibles sans les mesures effectuées sur le terrain. Ainsi, un grand merci à ceux qui ont participé aux campagnes pour leur bonne humeur qu'ils ont pu apportée sur le terrain. Un grand merci à Mohamed Saad pour son aide sur les différentes campagnes de terrain réalisées.

Je tiens tout particulièrement à remercier Claire Thérial pour tous les moments que nous avons passés ensemble, pour ton aide précieuse et tes conseils qui m'ont permis de mener à bien mes analyses. Merci pour ton soutien quotidien, que ce soit sur le terrain ou en laboratoire, notamment lors des filtrations, des centrifugations, des comptages en microplaque et des qPCR, qui n'ont pas toujours été de tout repos. Un grand merci au reste de la cellule technique, notamment Émilie Caupos, Lila Boudahmane et Chandirane Partibane. Je remercie également Angélique Goffin pour la formation avec le spectrophotomètre et de m'avoir aidé pour le traitement des données de fluorescence 3D. Merci à Natalia Angelotti, Brigitte Vincon-leite et Phillipe Dubois pour leurs aide sur les capteurs et de m'avoir fournis les données des capteurs HYDROLAB Series 5. Enfin, je souhaite exprimer ma gratitude à tous mes collègues du LiSSi pour m'avoir accueillie au sein du laboratoire comme si j'étais l'une des vôtres. À ma stagiaire, Salimata bathily, merci pour l'aide précieuse que tu m'as apportée.

À tous les doctorants du Leesu et du LiSSi, merci pour votre gentillesse et votre bonne humeur, sans vous cette thèse n'aurait pas été la même. À tous mes amis, qui m'ont conseillée et encouragée durant toutes mes années d'études, un grand merci. Toute ma reconnaissance à Lamyae, Tharsana et Abhishek qui m'ont aidé au cours de ce projet.

Et enfin à toute ma famille, merci de m'avoir soutenue dans cette aventure, toute ma

reconnaissance ainsi que mon éternelle gratitude. À mes parents, qui m'ont toujours soutenue par leur amour, leur tendresse et leurs conseils précieux. À mes sœurs adorées, qui ont toujours été présentes à mes côtés. Et à mes neveux et nièces, dont les sourires et l'énergie m'ont toujours apporté une joie immense. Je remercie également ma belle-famille pour leur bienveillance. À mon mari, qui a toujours cru en moi, même lorsque je n'y croyais plus. Merci pour votre patience et pour votre soutien indéfectible.

Communications scientifiques

Publications dans des journaux scientifiques internationaux :

— **Naloufi, M.**, Lucas, F. S., Souihi, S., Servais, P., Janne, A., & Wanderley Matos De Abreu, T. (2021). Evaluating the performance of machine learning approaches to predict the microbial quality of surface waters and to optimize the sampling effort. *Water*, **13**(18), 2457. [MDPI].

— **Naloufi, M.**, Abreu, T., Souihi, S., Therial, C., Rodrigues, N. A. de P., Le Goff, A. G., Saad, M., Vinçon-Leite, B., Dubois, P., Delarbre, M., et al. (2024). Long-Term Stability of Low-Cost IoT System for Monitoring Water Quality in Urban Rivers. *Water*, **16**(12), 1708. [Multidisciplinary Digital Publishing Institute].

— Angelotti de Ponte Rodrigues, N., Carmigniani, R., Guillot-Le Goff, A., Lucas, F. S., Therial, C., **Naloufi, M.**, Janne, A., Piccioni, F., Saad, M., Dubois, P., et al. (2024). Fluorescence spectroscopy for tracking microbiological contamination in urban waterbodies. *Frontiers in Water*, **6**, 1358483. [Frontiers Media SA].

Publications en préparation :

— **Naloufi, M.**, Therial, C., Saad, M., Bathily, S., Souihi, S., Wanderley Matos De Abreu, T., Gillon-Ritz, M., Delarbre, M., Kennouche, P., & Lucas, F. S. Incorporating uncertainty and methodological variability into the measurement of surface water contamination indicators.

— **Naloufi, M.**, Therial, C., Delarbre, M., Kennouche, P., Janne, A., & Lucas, F. S. Determining bacteriological dynamics in the Marne and Seine rivers.

Publications dans des rapports scientifiques :

— Lucas, F. S., **Naloufi, M.**, Therial, C., Saad, M., Bathily, S., Delarbre, M., & Gillon-Ritz, M. Variabilité méthodologique pour la mesure des indicateurs de contamination de l'eau de surface. Rapport d'activités 2022 du programme de recherche MeSeine Innovation.

— Lucas, F. S., **Naloufi, M.**, Therial, C., Saad, M., Goulet, M., Wurtzer, S., & Moulin, L. Variabilité des contaminations microbiologiques des eaux de surface lors des

évènements pluvieux. Rapport d'activités 2023 du programme de recherche PIREN-Seine.

— Lucas, F. S., Therial, C., **Naloufi, M.**, Saad, M., Wurtzer, S., Moulin, L., & Goulet, M. Sources et flux de pathogènes dans les rejets pluviaux. Rapport d'activités 2023 du programme de recherche OPUR, Action 1.1.

— Bouleau, G., Lucas, F., Mouchel, J. M., Azimi, S., Barles, S., Delarbre, M., Euzen, A., Goffin, A., Guérin, S., Haghe, J. P., Guillot-Le Goff, A., Jauzein, V., Kennouche, P., Lestel, L., Moulin, L., Moutiez, J., **Naloufi, M.**, Rocher, V., Rouillé-Kielo, G., Varrault, G., Vinçon-Leite, B., & Wurtzer, S. La baignade en Seine et en Marne. Groupe de travail « Amélioration de la connaissance » piloté par la Ville de Paris, le syndicat Marne-Vive, HAROPA PORT. Piren-Seine, Paris, France, 2024.

Publications dans des congrès internationaux :

— Balachandran, T., Abreu, T., **Naloufi, M.**, Souihi, S., Lucas, F. S., & Janne, A. (2022). IoT and transfer learning based urban river quality prediction. Dans *GLOBECOM 2022—2022 IEEE Global Communications Conference*, 257–262. IEEE.

— Vinçon-Leite, B., de Ponte Rodrigues, N. A., Guillot-Le Goff, A., Carmigniani, R. A., Da Silva, R. L., Dubois, P., Saad, M., Lucas, F., **Naloufi, M.**, Therial, C. (2023). Modélisation hydrodynamique 3D pour l'évaluation de la qualité de l'eau en milieu urbain – application au Bassin de La Villette (Paris, France). *Novatech 2023*.

— de Ponte Rodrigues, N. A., Araújo, L. C., Guillot-Le Goff, A., Saad, M., Dubois, P., Lucas, F., Therial, C., **Naloufi, M.**, Carmigniani, R., Piccioni, F., et al. (2023). Fluorescence spectroscopy of dissolved organic matter for water quality monitoring in urban waterbodies. *Novatech 2023*.

— Dahane, A., Benameur, R., **Naloufi, M.**, Souihi, S., Abreu, T., Lucas, F. S., & Mellouk, A. (2024). IoT Urban River Water Quality System Using Federated Learning via Knowledge Distillation. Dans *ICC 2024—IEEE International Conference on Communications*, 1515–1520. IEEE.

— Dahane, A., Benameur, R., Souihi, S., **Naloufi, M.**, Belhadj Benziane, I., Lucas, F., Mellouk, A. (2025). FCL-IWQMS : Federated Continual Learning and IoT-Based Water Quality Monitoring System for Adaptive Real-Time Insights. *TincNET Re-*

search Team, Univ. Paris-Est Créteil, France.

Communications dans des congrès nationaux :

— **Naloufi, M.**, Lucas, F. S., Souihi, S., Servais, P., Janne, A., & Wanderley Matos De Abreu, T. (18 novembre 2021). Influence de la qualité et variabilité de la mesure sur la prédiction des concentrations en indicateurs de contamination fécale. *Colloque de l'Association Francophone d'Écologie Microbienne*.

— **Naloufi, M.**, Therial, C., Saad, M., Bathily, S., Delarbre, M., Gillon-Ritz, M., & Lucas, F. S. (18–19 octobre 2022). Variabilité méthodologique de la mesure des indicateurs de contamination de l'eau de surface. *Journées doctorales en Hydrologie urbaine*.

— **Naloufi, M.**, Therial, C., Saad, M., Delarbre, M., Kennouche, P., Moulin, L., Wurtzer, S., Goulet, M., & Lucas, F. S. (4–6 octobre 2023). Profilage de la variabilité de la qualité microbiologique des rejets et des eaux de surface lors des évènements pluvieux. *18e Congrès National de la Société Française de Microbiologie*.

— Lucas, F. S., **Naloufi, M.**, Therial, C., Saad, M., Goulet, M., Wurtzer, S., & Moulin, L. (4–6 octobre 2023). Impact du temps de pluie sur la qualité microbiologiqu des eaux urbaines. *18e Congrès National de la Société Française de Microbiologie*.

Table des matières

Liste des figures xix
Liste des tableaux xxii
Liste des abréviations xxiii

Introduction génerale **1**

1 Étude de la qualité microbiologique des eaux de surfaces : état de l'art **6**

1. Introduction 6
2. Historique de la baignade en ville en Ile-de-France 7
3. Sources de contamination 8
4. Risque sanitaire 10
5. Notion d'indicateur de contamination fécale 12
6. Evaluation de la qualité microbiologique de l'eau de baignade 14
7. Variabilité spatiale et temporelle de la qualité microbiologique 16
8. Décroissance des indicateurs de contamination fécale 18
9. Incertitudes sur la mesure des indicateurs bactériens 21
   - 9.1. Sources d'incertitudes 21
   - 9.2. Incertitude liée à la variabilité spatio-temporelle 22
   - 9.3. Incertitude liée à la méthodologie de prélèvement 23
   - 9.4. Incertitude liée à l'analyse 25
   - 9.5. Estimation de l'incertitude pour les mesures ponctuelles 26
10. Prédiction de la qualité microbiologique 27
    - 10.1. Modèles statistiques 29
    - 10.2. Modèles déterministes 30
    - 10.3. Modèles basés sur l'apprentissage 30
      - 10.3.1. Apprentissage automatique 30
      - 10.3.2. Apprentissage par transfert 34
      - 10.3.3. Apprentissage fédéré 36
      - 10.3.4. Réseau de neurones 36

11. Optimisation de la collecte de données 37
11.1. Apprentissage actif 39
11.2. Collecte automatisée de données 40

2 Optimisation de la collecte de données pour la modélisation de la qualité microbiologique des eaux de surface 45

1. Introduction 45
2. Evaluating the Performance of Machine Learning Approaches to Predict the Microbial Quality of Surface Waters and to Optimize the Sampling Effort . . . 47
2.1. Introduction 48
2.2. Materials and methods 51
2.2.1. Study site and water quality data collection 51
2.2.2. Data preparation 52
2.2.3. Machine-learning models 52
2.2.3.1. KNN 52
2.2.3.2. SVM 53
2.2.3.3. DT 53
2.2.3.4. Bagging 53
2.2.3.5. RF 53
2.2.3.6. Adaboost 53
2.2.4. Models evaluation 54
2.2.5. Identification of the weakness parts of the dataset 55
2.3. Results and discussion 56
2.3.1. The dataset used in this study 56
2.3.2. ML-based *E. coli* prediction comparison 58
2.3.3. Limits of ML-based *E. coli* estimation 59
2.3.4. Identification of the weaknesses in the dataset 60
2.4. Automated data collection 63
2.5. Conclusion 66
2.6. Appendix 67

3. Évaluation de la performance des approches d'apprentissage automatique et d'apprentissage par transfert pour prédire la qualité microbienne des eaux de surface en Seine et en Marne 69
3.1. Introduction 70
3.2. Matériel et méhodes 73
3.2.1. Site d'étude et collection de données sur la qualité de l'eau . 73
3.2.1.1. La Marne 73
3.2.1.2. La Seine 74
3.2.2. Préparation des données 74
3.2.3. Les modèles d'apprentissage automatique 75
3.2.4. Apprentissage par transfert 76
3.2.5. Evaluation des modèles 76
3.2.6. Identification des points faibles du jeu de données 77
3.3. Résultats 78
3.3.1. Jeux de données 78
3.3.1.1. La Marne 78
3.3.1.2. La Seine 78
3.3.2. Prédiction de la concentration en *E. coli* par les modèles par apprentissage automatique 79
3.3.2.1. Effet du nombre de paramètres sur les performances des modèles en Marne 79
3.3.2.2. Comparaison des performances de prédictions de *E. coli* avec les données de la Seine 80
3.3.3. Apprentissage par transfert 81
3.3.3.1. Evaluation de l'apprentissage par transfert pour prédire les concentrations en Marne 81
3.3.3.2. Evaluation de l'apprentissage par transfert pour prédire les concentrations en Seine 81
3.3.4. Limites de l'estimation de la concentration en *E. coli* basée sur le modèle RF de la Seine 82
3.4. Discussion 86
3.4.1. Comparaison des méthodes d'apprentissage automatique . . 86

3.4.2. Incertitude sur la prédiction des modèles RF 88
3.5. Conclusion 90
3.6. Annexe 91
4. Long-Term Stability of Low-Cost IoT System for Monitoring Water Quality in Urban Rivers 94
4.1. Introduction 95
4.2. Materials and Methods 97
4.2.1. Prototype Design 98
4.2.1.1. Low-Cost Sensors 98
4.2.1.2. Reference Sensors 100
4.2.2. Specifications and Price 100
4.2.3. Cleaning and Calibration 101
4.2.4. LoRa Gateway 102
4.2.5. Sensor Validation 102
4.2.5.1. Accuracy 102
4.2.5.2. Temperature Effect 104
4.2.5.3. Temporal Stability in the Laboratory 104
4.2.5.4. Temporal Stability in the Field 105
4.3. Results and Discussion 106
4.3.1. Accuracy of the Sensors 107
4.3.2. Reproducibility of the Sensors 108
4.3.3. Sensitivity to the Environment 109
4.3.4. Temporal Stability in the Laboratory 111
4.3.4.1. Short-Term Stability 111
4.3.4.2. Detection and Removal of Outlier for Long-Term Series . 111
4.3.4.3. Long-Term Stability 113
4.3.5. *In Situ* Validation 115
4.3.5.1. Light Interference with the Turbidity Sensor 115
4.3.5.2. Temporal Stability in the Field 116
4.3.6. LoRa Gateway Performance 120
4.4. Conclusions 122
4.5. Appendix 124

5. Conclusion 131

3 Incertitudes et variabilité des dynamiques bactériologiques dans la surveillance des eaux de surface 134

1. Introduction 134
2. Optimisation de la classification des échantillons en intégrant par la logique floue l'incertitude de la mesure des indicateurs de contamination fécale 137
   - 2.1. Introduction 138
   - 2.2. Matériel et méthodes 142
     - 2.2.1. Site d'échantillonnage 142
     - 2.2.2. Équipements d'échantillonnage ponctuel depuis la berge . . . 143
     - 2.2.3. Protocole de nettoyage des équipements d'échantillonnage ponctuel 143
     - 2.2.4. Protocole de nettoyage du préleveur automatique 144
     - 2.2.5. Protocole de transport et stockage 145
     - 2.2.6. Dénombrement des BIF 146
     - 2.2.7. Incertitude analytique et temps d'incubation 146
     - 2.2.8. Extraction et quantification de l'ADN 147
     - 2.2.9. Amplification des marqueurs spécifiques et des pathogènes . 147
     - 2.2.10. Analyse statistique 148
     - 2.2.11. Analyse et estimation de l'incertitude 148
     - 2.2.12. Prise de décision sous incertitude 149
   - 2.3. Résultats et discussion 151
     - 2.3.1. Variabilité liée à l'équipement pour le prélèvement ponctuel . 151
     - 2.3.2. Répétabilité dans le temps 152
     - 2.3.3. Protocole de nettoyage 153
       - 2.3.3.1. Equipements pour les prélèvements ponctuels 153
       - 2.3.3.2. Prélèveur automatique 154
     - 2.3.4. Protocole de transport et stockage 157
     - 2.3.5. Impact du temps d'incubation sur la lecture 159
     - 2.3.6. Synthèse globale des incertitudes 160
     - 2.3.7. Incertitudes retenues pour *E. coli* 162

2.3.8. Intégration de l'incertitude dans la prise de décision 164
2.4. Conclusion 168
2.5. Annexe 170
3. Dynamique temporelle de la qualité bactériologique en Seine et en Marne . . . 173
3.1. Introduction 174
3.2. Matériels et Méthodes 179
3.2.1. Sites d'étude en rivière 179
3.2.2. Bases de données 180
3.2.3. Expérimentation *in situ* 181
3.2.4. Quantification des BIF 182
3.2.5. Modélisation de la dynamique temporelle après une pluie . . 182
3.2.5.1. Modélisation exponentielle inverse 183
3.2.5.2. Estimation des taux de mortalité et de disparition 183
3.2.6. Indicateurs de résilience et de résistance 184
3.2.7. Traitements statistiques 185
3.3. Résultats 186
3.3.1. Détermination du taux de mortalité 186
3.3.2. Dynamique de disparition d'*E. coli* en rivière 187
3.3.2.1. Caractéristiques des événements sélectionnées 188
3.3.2.2. Estimation du taux de disparition 191
3.3.2.3. Impact des conditions environmentales sur les taux de disparition 195
3.3.3. Résilience et résistance 197
3.3.3.1. Estimation de la résilience et de la résistance avec les mesures réglementaires 197
3.3.3.2. Intervalle de temps minimum entre chaque mesure 198
3.3.3.3. Impact des conditions environnementales sur la résilience et la résistance des sites en Seine et en Marne 199
3.4. Discussion 201
3.4.1. Taux de disparition et temps de retour 201
3.4.2. Comparaison des bases de données 203
3.4.3. Pics de pollution 204

3.4.4. Temps et niveau de récupération après la pollution 205
3.5. Conclusion 207
3.6. Annexe 209
4. Conclusion 210

Conclusion génerale **212**

Références bibliographiques **218**

Liste des figures

1.1 Normes, directives et guide pour le prélèvement d'eau de baignade. 22
1.2 Le processus de fonctionnement de l'apprentissage par transfert pour un modèle donné (Peng et al., 2022). 35
1.3 Processus d'apprentissage actif (Russo et al., 2020). 39
1.4 Cadre basé sur l'IdO pour la surveillance de la qualité de l'eau (Rahu et al., 2024). 43

2.1 Marne River water quality monitoring stations. The light grey stars indicate the SMV sampling stations and the dark grey stars indicate the location of the rain gauges used. 51
2.2 Guideline to provide and select an adapted model for water quality prediction and for the identification of a set of data to optimize the sampling strategies. . . 56
2.3 Distribution of the data for each variable. The median is indicated as a solid black line inside each boxplot, outliers are indicated as black dots. On the ordinates are the values taken by each variable with the units specified in parenthesis. . . 57
2.4 Evaluation of the prediction performances of the 6 machine-learning models during the 10 trials. On the abscissa the model is indicated and on the ordinate the value of the statistical metrics are displayed (dimensionless) : (A) RMSE ; (B) MAE ; (C) RPD. 58
2.5 Relationship between the *E. coli* concentration predicted by the RF-based model and the measured concentration. The white circles indicate the values. The red line indicates theoretical values corresponding to an accurate prediction of the model compared to the measured values for the ten testing trials. Blue lines indicate the 50% confidence interval. 60
2.6 Vizualisation of the values that need enrichment in the dataset for the temperature, conductivity, 24 h cumulative rainfall of the previous day and the river flow. The abscissa displays the value range of each parameter. Predicted values giving a reasonable estimation are visualized with solid black bars, white spaces represent the values that need further enrichment in the dataset. 62
S1 Correlation analysis between water quality parameters and *E. coli* concentration estimated by RF for (A) reasonable estimation ; (B) inaccurate estimation of *E. coli* (n=10). 68
2.1 Stations de surveillance de la qualité de l'eau de la rivière en Marne (étoiles vertes) et en Seine (étoiles bleues). 73
2.2 Schéma récapitulatif de la stratégie utilisée pour l'entrainement et le test de l'ensemble des modèles sur les données de la Marne et de la Seine. 75

2.3 Description des données de la Seine pour les paramètres physico-chimiques, pluviométriques et microbiologiques (température en °C, conductivité en µS/cm, turbidité en NTU, nombre de jours secs après la dernière pluie, pluviométrie du jour cumulée sur 24 h en mm, pluviométrie de la veille cumulée sur 24 h en mm, débit au pont d'Austerlitz en m^3/s et le logarithme népérien de la concentration en *E. coli* (NPP/100 mL). 79
2.4 Évaluation des performances de prédiction des 6 modèles d'apprentissage automatique au cours des 10 essais avec 8 paramètres issus de la base de données de la Marne. Métriques statistiques : (A) RMSE ; (B) MAE ; (C) RPD. 80
2.5 Évaluation des performances de prédiction des 6 modèles d'apprentissage automatique au cours des 10 essais avec les données de la Seine. Métriques statistiques : (A) RMSE ; (B) MAE ; (C) RPD. 80
2.6 Évaluation des performances de prédiction des 6 modèles d'apprentissage automatique au cours des 10 essais avec les données de la Marne avec 8 paramètres, après entrainement avec les données de la Seine. Métriques statistiques : (A) RMSE ; (B) MAE ; (C) RPD. 81
2.7 Évaluation des performances de prédiction des 6 modèles d'apprentissage automatique au cours des 10 essais en Seine avec 8 paramètres après entrainement avec les données de la Marne. Métriques statistiques : (A) RMSE ; (B) MAE ; (C) RPD. 82
2.8 Nombre d'observations identifiées comme des estimations raisonnables ou inexactes selon les valeurs MAPE obtenues avec le modèle RF au cours des dix essais sur les données de la Seine. 83
2.9 Relation entre la concentration en *E. coli* (NPP/100 mL) prédite par le modèle RF et la concentration mesurée en Seine. Les cercles noirs indiquent les valeurs. La ligne rouge indique les valeurs théoriques correspondant à une prédiction exacte du modèle par rapport aux valeurs mesurées pour les dix essais et les courbes bleues indiquent l'intervalle d'incertitude de 50% autour de la valeur exacte de prédiction. 84
2.10 Valeurs des paramètres donnant une estimation raisonnable des concentrations en *E. coli* (NPP/100 mL) dans la plage de valeurs des prédicteurs pour l'ensemble de données mesurées en Seine pour (A) la température ; (B) la conductivité. . . 85
2.1 Synoptic view of the low-cost system for water quality monitoring in real time. 98
2.2 Installation sites (A,B) and parameters measured by each type of sensor (source : Google Maps). 106
2.3 Hardware components involved in the field experiment : 1 : temperature, 2 : turbidity, 3 : conductivity, 4 : pH-1, 5 : pH-2, 6 : dissolved oxygen, A : LoRa HAT gateway, B : LoRa Arduino Pro gateway. 107
2.4 Comparison of two unit sensors placed simultaneously in the same solution. In blue unit 1, and in red unit 2. (**A**) Temperature, (**B**) turbidity. 108

2.5 Temperature effect on the turbidity and dissolved oxygen. (**A**–**C**) Fixed temperature analysis, (**B**) turbidity measurement at different temperatures, (**D**) Dissolved oxygen before compensation in blue, after compensation in black and the PME sensor in red. 110
2.6 Long-term turbidity analysis. The blue dots correspond to measurements taken by the sensors and the red dots to measurements taken by the laboratory turbidimeter. (**A**) Raw data, and (**B**) after removal of outlier using ARIMA with a median filter (width of 5). 112
2.7 Long-term stability of sensors reading standard solutions. (**A**) Temperature, (**B**) pH, (**C**) conductivity measurement cleaned with an ARIMA model and median filter (width of 11), and (**D**) dissolved oxygen measurement cleaned with an ARIMA model and median filter (width of 5). 114
2.8 Effect of the ambient light on the reading of the turbidity sensor. (**A**) Before shading, and (**B**) after shading. 116
2.9 Temperature analysis at Site A at Bassin de La Villette. Values from OTT sensors are displayed in red, Arduino sensors in blue. Black dots indicate that the sensor has been calibrated, green dots that it has been cleaned, and gray dots that the sensor has been replaced. Replacements were carried out by alternating the two units of the same sensor every week. (**A**) From early September 2022 to early January 2023, and (**B**) from May to June 2023. 117
2.10 Conductivity measurement at site A at Bassin de la Villette. Values of the OTT sensors are displayed in red, and Arduino sensors in blue. Black dots indicate that the sensor has been calibrated, green dots that it has been cleaned, and gray dots that the unit has been replaced. Replacements were carried out by alternating the two units of the same sensor every week. (**A**) From early September 2022 to early January 2023, (**B**) from May to June 2023, and (**C**,**D**) data from (**A**,**B**) averaged over 4 h and cleaned by ARIMA. 117
2.11 Dissolved oxygen measurements at site B at Bassin de la Villette. Values of the PME sensor are displayed in red, Arduino sensors in blue, and Arduino temperature sensors in black. 119
2.12 Framework for testing the reliability of sensors. 120
2.13 LoRa gateway performance. Arduino LoRa gateway in pink, LoRa HAT gateway in blue light. (**A**) Received signal strength indicator (RSSI), (**B**) Signal-to-Noise Ratio (SNR). 121
S1 Analysis of the sensors calibration for 2 units (The results of the first unit in blue and the second unit in red). (**A**) Temperature. (**B**) pH. (**C**) Conductivity. (**D**) Turbidity. (**E**) Dissolved oxygen. 124

S2 Temperature effect on the pH and conductivity. (**A**–**C**) Reference, fixed temperature analysis. (**B**) pH measurement at different temperature. (**D**) Conductivity at different temperature without compensation (raw data) in blue and with compensation by using 2 compensation coefficients (coef of 0.0185 in black and 0.0265 in red). 125
S3 Effect of battery temperature in the box. In blue, the temperature in the waterproof box, and in red, the solution at laboratory temperature. 125
S4 Effect of long-term use of Dissolved Oxygen sensor Membrane Cap. (**A**) After 6 months of use. (**B**) New membrane cap. 126
S5 Comparison of two unit sensors placed simultaneously in the same solution : in blue, unit 1, and in red, unit 2. (**A**) pH-1, (**B**) pH-2, and (**C**) conductivity. 126
S6 pH analysis at site A at Bassin de la Villette. OTT sensors in red, Arduino sensors in blue. The black dot indicates the date of calibration, green only if the sensor was cleaned, and grey when the analysis process has changed, alternating between sensor units each week : (**A**) from early September 2022 to early January 2023 for the pH-1 meter, (**B**) in June 2023 for the pH-2 meter, (**C**) From May to June 2023 for the pH-1 meter, and (**D**) data from (**C**) averaged over 4 h and cleaned by ARIMA. 127
S7 Turbidity analysis at site A in Bassin de la Villette. OTT sensors in red, Arduino sensors in blue. The black dot indicates the date of calibration, green dots if the sensor has been cleaned, and grey dots when the analysis process has changed, alternating between sensor units each week and pink dots for external light protection : (**A**) from early September 2022 to early January 2023, and (**B**) from May to June 2023. 127
S8 Temperature (**A**) and conductivity (**B**) analysis at site B in Bassin de la Villette. OTT sensors in red, Arduino sensors in blue. The black dot indicates the date of calibration, green dots if the sensor has been cleaned, and grey dots when the analysis process has changed, alternating between sensor units each week. . . . 128
S9 Site 1 (Campus of Vitry) (**A**) and site 2 (residential area at Vitry) (**B**) for the 2 LoRa gateways tests. 128
S10 More detailed synopsis of the framework for testing the reliability of the sensors. 129
S11 Time gap between two measures. Arduino LoRa gateway in pink, LoRa HAT gateway in light blue. 129
S12 Performance analysis of the LoRa gateways in the two sites (site 1 (Campus of Vitry) in pink, site 2 (residential area at Vitry) in light blue). (**A**) Received signal strength indicator (RSSI), (**B**) Signal-to-Noise Ratio (SNR). 130

3.1 Comparaison des équipements de prélèvement ponctuel depuis la berge au niveau des 2 sites du lac de Créteil. (A) concentration en *E. coli* en NPP/100 ml, (B) Pourcentage moyen d'incertitude lié aux équipements et à la variabilité temporelle. La taille des cercles représente l'écart type du pourcentage d'incertitude. 152

3.2 Pourcentage d'incertitude pour l'estimation d'*E. coli* (A) et des différents marqueurs (B) par rapport à l'échantillon référence lors des différentes étapes du protocole de nettoyage du bécher et du tuyau de la pompe pour les équipements manuels (M) au niveau du site 1 du lac de Créteil et avec le préleveur automatique (A) à La Villette et à Saint-Maur-des-Fossés. 153
3.3 Blancs de terrain après prélèvement d'eau de surface au Bassin de la Villette (BV) et en Marne à Saint-Maur-des-Fossés (M) et d'eau résiduaire à l'ouvrage cadre du Centre Urbain (CU) et au bassin de rétention de Sucy-en-Brie (SB). Le symbole représente une comparaison des blancs par rapport à l'échantillon du site prélevé avant le blanc. 155
3.4 Blancs de terrain après décontamination et rinçage du préleveur au bassin de rétention de Sucy-en-Brie. Le symbole représente une comparaison des blancs par rapport à l'échantillon du site prélevé avant le blanc. 156
3.5 Pourcentage d'incertitude pour l'estimation d'*E. coli* par rapport à l'échantillon référence en fonction du temps (6 h ou 24 h) et de la température de stockage les 6 premières heures à 5°C (froid) ou à température ambiante (ambiant). . . . 157
3.6 Schéma récapitulatif de l'analyse de l'incertitude pour l'ensemble des indicateurs fécaux analysés ; en noir (équipements manuel), en vert (préleveur automatique) ; en bleu (tous les équipements). 160
3.7 Schéma récapitulatif de l'analyse de l'incertitude pour *E. coli*, en bleu les incertitudes sur les équipements automatiques pris en compte pour les mesures obtenues avec le système ColiMinder et en vert les incertitudes sur les équipements manuels pris en compte pour les mesures ponctuelles en culture * : (Bergeron et al., 2011; Noble et al., 2010). 163
3.8 Fonction d'appartenance avec la logique floue. En vert : qualité bonne, en Bleu : qualité moyenne, en rouge qualité mauvaise et les traits noirs représente les seuils réglementaires. 164
3.9 Comparaison de la méthode du (A) centre de gravité (COG) et de (B) moyenne des maxima (MOM) pour les mesures réglementaires au niveau des 3 sites avec une classification utilisant l'intervalle de 24 h de midi à midi. La couleur représente la classe d'appartenance et l'axe des abscisses représente la méthode de défuzzification. Le trait noir épais représente le seuil réglementaire et les traits fins représentent l'incertitude associée au seuil pour les analyses ponctuelles. . 165
3.10 Comparaison de la classification des données du système ColiMinder avec la logique floue pour les mesures réglementaires d'*E.coli* (Log_{10} NPP/100 mL) au niveau des 3 sites (2 : pont de Tolbiac rive droite, 3 : pont de Tolbiac rive gauche, 11 : pont de l'Alma) en utilisant différents intervalles de temps (A) midi à midi (B) 8 h à 8 h (C) minuit à midi (D) 20 h à 8 h (E) 4 h à 8 h (F) 8 h à midi. La couleur représente la classe d'appartenance. Les traits noirs épais représentent les seuils réglementaires de 1800 NPP/100 mL et de 100 NPP/100 mL et les traits fins représentent l'incertitude associée à la mesure manuelle. 167

S1 Pourcentage d'incertitude pour l'estimation des différents marqueurs bactériens par rapport à l'échantillon référence en fonction du temps (6 h ou 24 h) et de la température de stockage à 5°C (froid) ou à température ambiante (ambiant). . . 171
3.1 Facteurs pouvant avoir un effet sur la dynamique des concentrations en BIF dans les bases de données de suivi *in situ* (bleu), et sur la mortalité des BIF mesurées lors des l'expérimentations *in situ* (rouge) et en laboratoire (vert). 179
3.2 Schéma des sites étudiés. 180
3.3 Courbe modélisant la décroissance (ligne noire) avec les valeurs obtenues par expérimentation en sac à dialyse pour le dénombrement des *E. coli* avec l'eau du rejet. Les valeurs mesurées lors de l'expérimentation sont représentées par des carrés rouges. 186
3.4 Courbe modèle moyenne (ligne rouge) ajustée aux concentrations relatives en *E. coli* pour chaque site. (A) site SMV1, (B) site SMV10, (C) site SMV14, (D) pont de l'Alma, (E) pont de Tolbiac RD et (F) pont de Tolbiac RG. Les valeurs mesurées relatives exprimées en pourcentage de la concentration initiale sont représentées en bleu, les intervalles de confiance à 95% des courbes sont représentés en gris. 192
3.5 Comparaison des taux de disparition (jr^{-1}) entre les 6 sites. 192
3.6 Comparaison des taux de disparition (jr^{-1}) selon l'interval de temps entre 2 mesures (de 2 h à 24 h) pour *E. coli* pour les ponts de l'Alma et de Tolbiac. Les barres horizontales et les lettres (a, b, c) regroupent les intervalles sans différence significative. 194
3.7 Courbes modèles moyennes (ligne rouge) de la variation temporelle des concentrations relatives en *E. coli* pour les pluies $\geq$10 mm, selon le site. (A) Alma 2 h, (B) Alma 24 h, (C) Tolbiac 2 h et (D) Tolbiac 24 h. Les concentrations en *E. coli* relatives exprimées en pourcentage de la concentration au pic de pollution sont en bleu, les intervalles de confiance à 95% sont représentés en gris. 196
3.8 Courbes modèles moyennes (ligne rouge) de la variation temporelle des concentrations relatives en *E. coli* pour les pluies <10 mm pour chaque site.(A) Alma 2 h, (B) Alma 24 h, (C) Tolbiac 2 h et (D) Tolbiac 24 h.Les concentrations en *E. coli* relatives exprimées en pourcentage de la concentration au pic de pollution sont en bleu, les intervalles de confiance à 95% sont représentés en gris. 197
4.1 Cadre général pour la gestion de la qualité des rivières et la prise de décision en matière de baignade. * : (ML : machine learning ; BD : base de donnée ; TL : transfert learning, FL : federate learning ; AL : active learning), ¤ : Couplage, rapport coût/bénéfice, en violet : les avantages des méthodes, en gras : les futures perspectives. 217

Liste des tableaux

1.1 Seuils de qualité microbiologique pour le classement des sites de baignade selon la directive 2006/7/EC. Basé sur l'évaluation du percentile 95 (*) et 90 (**). . . 15

1.2 Valeurs limites de qualité microbiologique des eaux intérieures d'un site de baignade classé, pour la gestion active en cours de saison, proposées par l'Agence française de sécurité sanitaire de l'environnement et du travail (Duboudin et al., 2007). 15

1.3 Valeurs du taux de décroissance dans différents types d'eau, le terme utilisé dans la littérature et les facteurs pris en compte pour la mesure du taux. ^{a} : sans remise en suspension, ^{b} : avec remise en suspension, ^{c} : (Noble et al., 2004), ^{d} : (Servais et al., 2007a), ^{e} : (Chigbu et al., 2005), ^{f} : (Jozić et al., 2014), ^{g} : (Nakhle et al., 2021), ^{h} : (Dick et al., 2010), ^{i} : (Blaustein et al., 2013). 20

1.4 Méthodes utilisées pour l'estimation de l'incertitude à partir des données disponibles (Harmel et al., 2016). 27

2.1 Correlation coefficients (average r_s ± SD) for the relationship between the values of *E. coli* predicted by the RF model (reasonable and inaccurate) and the environmental variables. Significant coefficients are indicated with a * (coefficient significance test p <0.05). Significant differences between the correlation coefficients of the two datasets are indicated as t-test p-values < 0.01. MAPE values were used to identify reasonable (less than 50%) and inaccurate (over 50%) estimations of *E. coli* concentrations obtained with the RF model during the ten testing trials. 61

S1 Descriptive statistics of the parameters. 67

S2 Average and standard deviation of the statistic metrics (RMSE, MAE, RDP) obtained with each model during the ten testing trials. 67

2.1 Comparaison des coefficients de corrélation (moyenne ± écart-type) obtenus entre les variables prédictives et les concentrations en *E. coli* (NPP/100 mL) raisonnablement prédites par le modèle RF entrainé et testé sur les données de la Seine et celles pour lesquelles la prédiction est inexacte. Les p-valeurs des tests statistiques comparant les valeurs de corrélation entre les prédictions raisonnables et inexactes sont données. 84

S1 Moyenne et écart-type des mesures statistiques (RMSE, MAE, RPD) obtenues pour chaque modèle au cours des dix essais avec 11 paramétres et 8 paramètres avec les données de la Marne. 91

S2 Moyenne et écart-type des mesures statistiques (RMSE, MAE, RPD) obtenues avec chaque modèle au cours des dix essais avec les données de la Seine. . . . 92

S3 Moyenne et écart-type des mesures statistiques (RMSE, MAE, RPD) obtenues avec chaque modèle au cours des dix essais avec les données de la Marne avec 8 paramètres, après entrainement avec les données de la Seine. 92
S4 Moyenne et écart-type des mesures statistiques (RMSE, MAE, RPD) obtenues avec chaque modèle au cours des dix essais avec les données de la Seine avec 8 paramètres après entrainement avec les données de la Marne. 92
S5 Comparaison entre les jeux de prédictions raisonnables en Marne et en Seine des coefficients de corrélation (moyenne et écart-types) entre les variables prédictives et les valeurs de concentrations en *E. coli* prédites. 93
2.1 Characteristics and specifications of the Arduino sensors (Farouk et al., 2022; DS18B20, 2023; pH V2, 2023a,b; Conductivity Meter V2, 2023; Hakim et al., 2019; Arduino, 2023; DO, 2023; Villeneuve et al., 2006). 99
2.2 Characteristics and specifications of the Hydrolab multiprobes (OTT) (Hydrolab DS5X, 2024). 101
2.3 Descriptive analysis of sensors calibration for 2 units. 104
2.4 Short-term analysis of sensors (repeatability). 111
2.5 Stability analysis of sensors : standard deviation (* without missing values during the sensor regeneration, ** values cleaned with an ARIMA method using a median filter). 113

3.1 Pourcentage de vrais positifs pour les 5 méthodes de défuzzification par rapport aux 2 méthodes de référence (méthode 1 et méthode 2), (NC) non classé avec la méthode par la Ville de Paris pendant les JOP 2024. 165
S1 Liste des amorces et des sondes utilisées lors de la qPCR pour la recherche de marqueurs fécaux animaux et humains, des bactéries totales et des *Campylobacter*. La séquence et la concentration finale pour chaque amorce sens (F) et antisens (R), et pour la sonde TaqMan (P) sont présentées dans le tableau. . . . 170
S2 Incertitude sur la collecte des échantillons pour les *E. coli*, les Entérocoques intestinaux, HF183 et les bactéries totales (BacQuant). 171
S3 Incertitude dans la collecte des échantillons des marqueurs BacCan (Chien), CGOF1 (Oie), Gull2 (Mouette et Goéland) et le pathogène *C. jejuni*. 172
3.1 Valeurs des paramètres physico-chimiques de l'eau de la Marne et de l'eau contenue dans les sacs à dialyse mesurés lors du 2ème et 3ème jours de l'expérimentation *in situ*. 187
3.2 Nombre de jours de pluie et pluviométrie (mm) cumulée durant la période estivale (1er juin - 30 septembre) chaque année à Paris et sur l'aval de la Marne, en fonction des pluviomètres les plus proches de chaque station de prélèvement. 187
3.3 Concentrations en *E. coli* en NPP/100 mL durant la période estivale par année et par site en Marne selon le suivi réglementaire. 189
3.4 Concentrations en *E. coli* en NPP/100 mL durant la période estivale par année et par site en Seine selon le suivi réglementaire. 189

3.5 Concentrations en *E. coli* en NPP/100 mL par site pour les événements sélectionnés (temps sec avant la pluie (avant), concentration initiale au pic de pollution (pic) et temps sec après la pluie (après). 190
3.6 Concentrations en *E. coli* en NPP/100 mL estimées par l'analyseur ColiMinder durant toute la période estivale par année et par site 190
3.7 Constante de cinétique (K_1, jr^{-1}) obtenues par le modèle linéaire exponentiel (p-valeur et R^2), le taux de disparition (K_2, jr^{-1}). Moyenne ± écart type ou [Min : Max]. p-valeur significative (S), non significative (NS) au seuil 0,05. NA non applicable. 191
3.8 Constante de cinétique (K_1, jr^{-1}) dérivées du modèle linéaire exponentiel (p-valeur et R^2) et estimation du taux de disparition (K_2, jr^{-1}). Moyenne ± écart type ou [Min : Max]. p-valeur significative (S), non significative (NS) au seuil 0,05. 193
3.9 Relation entre les facteurs environnementaux (site, concentration en*E. coli* initiale au pic de contamination en NPP/100 mL, pluviométrie cumulée de l'événement en mm et taux de disparition d'*E. coli*) en jr^{-1}. Les p-valeurs (p) sont indiquées pour chaque paramètre, les interactions significatives et le modèle global. NA : les paramètres non retenus, lm : modèle linéaire, lmm : modèle linéaire mixte. 195
3.10 Valeurs moyennes du temps de retour (T_{90}, jr) et des amplitudes de variation de la pollution (AV_{avant}, %) et de récupération ($AV_{aprés}$, %) lors des campagnes réglementaires en Seine et en Marne. 198
3.11 Valeurs moyennes du temps de retour (T_{90}) et des amplitudes de variation de la pollution (AV_{avant}) et de récupération ($AV_{aprés}$) en fonction de l'intervalle de temps entre chaque mesure lors des campagnes ColiMinder en Seine. 199
3.12 Relation entre les facteurs environnementaux (débit en m^3/s), site, concentration en *E. coli* initiale au pic de contamination en NPP/100 mL, pluviométrie et taux de disparition d'*E. coli* en jr^{-1}). Les p-valeurs sont indiquées pour chaque paramètre, les interactions significatives et le modèle global. NA : les paramètres non retenus, lm : modèle linéaire, glm : modèle linéaire généralisé, ^{a} : pluviométrie en catégories, ^{b} : pluviométrie cumulée de l'événement en mm. 200
S1 Valeurs moyennes de la constante de cinétique (K1, jr^{-1}) obtenues par le modèle linéaire exponentiel (p-valeur et R^2) et le taux de disparition K_2 en jr^{-1}. Moyenne ± écart type ou [Min : Max]. p-valeur significative (S), non significative (NS) au seuil 0,05. 209

Liste des abréviations

ACP : analyse en composante principale AdaBoost : Adaptive boosting
AEE : Agence européenne pour l'environnement
AELB : Agence de l'Eau Loire-Bretagne
AFNOR : Association Française de Normalisation
AIC : Critère d'information d'Akaike
AL : Active Learning
ARIMA : Autoregressive Integrated Moving Average
ARS : Agences Régionales de la Santé
$AV_{après}$: Amplitude de variation de la récupération
AV_{avant} : Amplitude de variation de la pollution
BD : Base de donnée
BIF : Bactéries indicatrices fécales
BS : Bisecteur
C_0 : Concentration initiale
COG : Centre de gravité
Cres : Concentration de la population résiduelle
DBSCAN : Density-Based Spatial Clustering of Applications with Noise
DT : Decision tree
EI : Entérocoques intestinaux
E. coli : *Escherichia coli*
FL : Federate Learning
glm : Modèles linéaires généralisés
glmm : Modèles linéaires généralisés mixtes
IdO : Internet des objets
IoT : Internet of Things
ISO : International Organization for Standardization
JOP : Jeux Olympiques et Paralympiques
KNN : K-nearest neighbors
K_1 : Constante de cinétique
K_2 : Taux de disparition / mortalité
lm : Modèles linéaires simples
lmm : Modèles linéaires mixtes
LoRa : Long Range
LoRaWAN : Long-range Wide Area Network
LOM : Maximum le plus à gauche
MAE : Erreur absolue moyenne
MAPE : Pourcentage d'erreur absolue moyen
MAV : Marne Aval
MES : Matière en suspension

MIQE : Minimum Information for Publication of Quantitative Real-Time PCR Experiments
ML : Machine Learning
MOM : Moyenne des maxima
MPN : Most Probable Number
NC : Non Classifié
NF EN : Norme Française Européenne
NPP : Nombre le Plus Probable
NTU : Nephelometric Turbidity Unit
OMS : Organisation Mondiale de la Santé
O_2 : Oxygène
qPCR : PCR quantitative en temps réel
PCR : Polymerase Chain Reaction
RMSE : Erreur quadratique moyenne
RPD : Rapport performance/déviation
RD : Rive droite
RG : Rive gauche
RF : Random forest
ROM : Maximum le plus à droite
RSSI : Received Signal Strength Indicator
S : Période de latence
SIAAP : Syndicat d'Aménagement de l'Aglomération Parisienne
SMV : Syndicat Marne Vive
SNR : Signal-to-Noise Ratio
STEU : Stations de Traitement des Eaux Usées SVM : Support Vector Machines
T_{90} : Temps de retour
TL : Transfert Learning
TDS : Solides dissous totaux
TKN : total Kjeldahl Nitrogen
UV : Ultra-Violet USGS : United State Geological Survey

Introduction génerale

Depuis quelques années, l'attention des municipalités s'oriente vers les fleuves, les rivières, les canaux, les bras morts, et les plans d'eau, avec une volonté de reconquête de la baignade en ville. En effet, de nombreuses villes d'Europe favorisent l'ouverture de zones de baignade et organisent des compétitions de natation en eau libre dans leurs rivières (Kistemann et al., 2016; Mouchel et al., 2020). Dans le monde entier, les épisodes de canicule ont récemment intensifié le développement des activités récréatives aquatiques dans les mégapoles. Cette situation contribue à augmenter la fréquentation des zones de baignade en milieu urbain (Jang, 2016; Houtman, 2010).

Cette reconquête des espaces bleus s'accompagne d'une amélioration générale de la qualité des eaux de surface, grâce à des réglementations plus strictes et à des améliorations des infrastructures (Schreiber et al., 2015). Ainsi, en Ile-de-France (France), malgré l'interdiction historique de la baignade dans la Seine et la Marne, se développe une forte volonté politique et sociale de réhabiliter les rivières urbaines pour la baignade, avec également un engagement renouvelé en faveur de la restauration écologique des cours d'eau (Noury et al., 2018). En héritage des Jeux Olympiques et Paralympiques (JOP) de Paris en 2024, les municipalités de la région parisienne se sont fortement engagées à améliorer la qualité de l'eau de la Seine et de la Marne afin de permettre la baignade d'ici 2025 (Bouleau et al., 2024), avec pour objectif principal l'amélioration continue de la qualité de l'eau des rivières à des fins récréatives.

Cependant sur les territoires fortement urbanisés, ces différentes activités posent un risque sanitaire dû à l'exposition à des pollutions incluant les microorganismes pathogènes d'origine hydrique. Ces contaminations peuvent générer un risque sanitaire pour les nageurs, d'autant plus qu'il est à prévoir une intensification des usages récréatifs dans les cours d'eau urbains dans les prochaines années (Schijven and de Roda Husman, 2005; Islam et al., 2018). Différentes sources de contamination peuvent apporter un flux de pathogènes au niveau des sites de baignade (Guérineau et al., 2014). Lorsque le milieu reçoit des rejets d'origine animale ou humaine, les bactéries présentes peuvent rendre l'eau inappropriée pour différentes activités. Le groupe des entérocoques intestinaux qui appartient aux streptocoques fécaux, de même que les coliformes thermotolérants (dits fécaux), en particulier l'espèce *Escherichia coli*, sont des

microorganismes appartenant au microbiote du tube digestif des animaux à sang chaud et des humains (Paruch and Mæhlum, 2012; Boehm and Sassoubre, 2014). Ils sont excrétés dans les fèces et de ce fait ils servent d'indicateurs de la présence potentielle d'eaux usées (Hébert and Légaré, 2000). Les deux bactéries indicatrices fécales (BIF), sont relativement bien corrélées avec le risque de gastroentérite. Elles servent donc de proxy pour évaluer le risque sanitaire et donc la présence éventuelle de pathogènes (Payment and Locas, 2011). La directive européenne 2006/7/CE, concernant la gestion de la qualité des eaux de baignade et qui vise à améliorer la qualité de l'environnement et à protéger la santé humaine s'appuie donc sur ces deux BIF pour le suivi de la qualité microbiologique des eaux de baignade (Wade et al., 2003; Borja et al., 2020). Ce suivi implique un échantillonnage de terrain et des analyses de laboratoire dont la logistique peut être lourde et le coût élevé (Manjakkal et al., 2021). De plus, le rendu des résultats se fera au mieux dans les 24 h suivant le prélèvement.

Pour permettre une gestion quotidienne des ouvertures/fermetures des sites de baignade suite à des événements polluants temporaires, une surveillance de la qualité microbiologique en temps réel des eaux de surface est nécessaire. Afin de disposer d'outils de gestion des pollutions plus rapides et de mettre en place des systèmes d'alerte efficaces, l'Organisation Mondiale pour la Santé (OMS) préconise la modélisation dans le but de prédire les indicateurs de contamination dans l'eau (OMS, 2018). Il existe une variété de modèles qui sont proposés pour prédire la qualité de l'eau (Mälzer et al., 2016; Chen et al., 2020). Toutefois, la performance de ces différents modèles varie selon le jeu de données et le contexte (Mälzer et al., 2016; Chen et al., 2020).

La grande variabilité spatio-temporelle qui caractérise les concentrations en microorganismes d'origine fécale et la complexité des relations entre les caractéristiques du bassin versant d'apport et le comportement des différents indicateurs microbiens de contamination fécale et pathogènes rendent difficile la prédiction précise et fiable des niveaux de contamination microbiologiques des eaux de surface (Cha et al., 2016). Or la dynamique spatio-temporelle des pathogènes hydriques lors d'événements pluvieux qui vont dégrader fortement la qualité microbiologique des eaux de surface reste encore peu connue (Curriero et al., 2001).

Par conséquent, le sujet de thèse a pour but de caractériser la variabilité des niveaux de contaminations d'origine fécale dans les rejets pluviaux et leur impact sur la qualité microbiologique des eaux de surface en milieu urbain. Ceci permettra d'améliorer la compréhension des sources et flux de contaminations microbiologiques des eaux urbaines en vue de prédire la qualité microbiologique lors des événements polluants. Ce travail fournira un cadre conceptuel

et des outils seront proposés pour la surveillance de la qualité de l'eau dans les rivières urbaines et la gestion quotidienne des sites de baignade. Le manuscrit est divisé en 3 chapitres.

Le **premier chapitre** est un état de l'art de la connaissance scientifique sur la surveillance de la qualité des eaux de surface en milieu continental. Le **deuxieme chapitre** a pour objectif d'optimiser la prédiction des concentrations en BIF à l'aide de modèles d'apprentissage automatique. Il y a en effet encore peu d'études publiées qui explorent ces modèles pour prédire la qualité microbiologique dans les rivières urbaines. Le chapitre comprend un guide pour la sélection d'un modèle d'apprentissage automatique (machine learning, ML) permettant une estimation précise et immédiate (nowcast) des concentrations d'*E. coli* à partir de données historiques. Nous formulons l'hypothèse qu'une sélection des paramètres météorologiques et physico-chimiques les plus couramment suivis par les collectivités permet une modélisation fiable des concentrations en BIF dans les eaux de surface. Ainsi, nous avons étudié la capacité de prédiction des modèles sélectionnés afin d'évaluer leur valeur individuelle en tant qu'outil de prédiction. Les deux rivières Seine et Marne en région parisienne (France) ont été considérées comme un cas d'utilisation afin de prédire la concentration en *E. coli* qui est le critère le plus déclassant pour la gestion journalière (Mouchel et al., 2020). Afin d'améliorer la performance et la précision du modèle sélectionné, nous avons ensuite exploré plusieurs pistes pour augmenter la quantité et la qualité des jeux de données utilisés pour entraîner les modèles ML : i) l'apprentissage par transfert, ii) l'optimisation de la collecte des données réglementaires, iii) la mesure en continu des paramètres physico-chimiques servant de prédicteurs au modèle. L'approche de l'apprentissage par transfert se base sur l'hypothèse que les données réglementaires issues d'un autre bassin versant similaire permettent d'augmenter le jeu de données d'entraînement et de le diversifier. Pour ce faire, nous avons testé si les données de la Seine à Paris et celles de l'aval de Marne (qui appartiennent au bassin versant de la Seine) pouvaient être utilisées alternativement pour pré-entraîner les modèles de ces deux rivières et ainsi améliorer leurs performances respectives. Ensuite, nous avons émis l'hypothèse que les modèles ML sélectionnés pour prédire les concentrations en *E. coli* peuvent également être utilisés comme outils pour optimiser des stratégies d'échantillonnage réglementaire, en vue d'obtenir des données en quantité et qualité suffisante pour l'entraînement des modèles ML. Pour ce faire, nous proposons de mettre en place un système d'alerte sur les performances du modèle afin d'optimiser la collecte des données réglementaires en identifiant dans quelles conditions le modèle ne parvient pas à prédire. Enfin, un contrôle plus efficace de la qualité de l'eau devrait également reposer sur des méthodes ra-

pides, peu coûteuses, nécessitant un minimum d'échantillonnage et fournissant des résultats en temps réel en complément du suivi réglementaire. De ce fait, à terme, le système d'alerte devrait être relié à un réseau de capteurs à faible coût permettant un suivi en continu des différents paramètres physico-chimiques (Whelan et al., 2020; Yaroshenko et al., 2020). Une stratégie de surveillance continue en s'appuyant sur quelques paramètres sélectionnés avec des capteurs peu coûteux pourrait permettre le suivi d'indicateurs de la qualité de l'eau et aider les gestionnaires à détecter la contamination possible (Farouk et al., 2022; McGrane, 2016; Yaroshenko et al., 2020). Ainsi, ce chapitre aborde également la stratégie développée afin de vérifier la fiabilité et la stabilité des capteurs à faible coût et optimiser leur maintenance. Nous avons conçu comme cas d'usage un prototype à faible coût, en testant 6 sondes physico-chimiques utilisant la plateforme open-source Arduino, afin de surveiller la qualité des eaux de surface en utilisant la technologie IdO (internet des objets, Internet of Things ou IoT, en anglais). Ce prototype a été calibré, testé et une analyse de stabilité à long terme a été réalisée en laboratoire et sur le terrain au Bassin de la Villette. Afin de fournir une résolution spatiale et temporelle suffisante et de réduire le coût des surveillances. Combiner les capteurs *in situ* à l'apprentissage automatique pourrait contribuer à optimiser l'effort d'échantillonnage et serait ainsi utilisé comme outil de gestion quotidienne, qui vient en appui à la surveillance réglementaire selon la directive 2006/7/CE (Carvalho et al., 2019).

En complément, une surveillance optimale de la qualité microbiologique ne peut être atteinte que si l'incertitude au niveau de la mesure est identifiée et qu'un moyen pour la réduire est considéré lors de l'échantillonnage et de la mesure. Le **troisième chapitre** porte donc sur la définition de l'incertitude associée à la surveillance réglementaire des BIF et à celle des marqueurs de contamination fécale humaine et animale. Des approches expérimentales ont permis de mieux quantifier cette incertitude liée à l'échantillonnage et au stockage des échantillons avant l'analyse. Ceci permettra une amélioration des bases scientifiques des normes et des réglementations en vigueur, ainsi que des nouveaux outils de suivis des sources de contamination, pour la mise en œuvre d'un plan de gestion des eaux de surface via un guide d'échantillonnage précis. De plus, une gestion efficace de la qualité de l'eau exige une connaissance approfondie de la dynamique et du devenir des bactéries présentes dans les eaux de surface. Ces bactéries peuvent soit persister dans l'environnement, soit disparaître et leur survie dépendra de leur exposition à diverses influences environnementales (Devane et al., 2018).

Une gestion efficace de la qualité de l'eau exige une connaissance approfondie de la

dynamique spatiale et temporelle des BIF dans les habitats aquatiques, ainsi que des facteurs qui l'influencent. Si les taux de décroissance des BIF après un pic de pollution temporaire montrent une faible variabilité d'un événement polluant à un autre, cela pourrait permettre d'avoir une utilisation par les gestionnaires comme paramètre pour prédire le devenir des contaminations (Dick et al., 2010). Par ailleurs, l'analyse de la dynamique des BIF suite à une pollution de court terme peut aider à estimer la capacité d'un site de baignade à résister et à récupérer de cette perturbation. Ces informations sont cruciales pour l'implantation, la gestion et l'amélioration des futurs sites de baignade. Ce troisième chapitre présente donc une analyse de la dynamique temporelle d'*E. coli* lors des événements pluvieux, en exploitant les données d'échantillonnage réglementaire en Marne et en Seine avec 2 à 3 prélèvements par semaine, ainsi que les données du système de mesure automatisé ColiMinder en Seine avec une analyse toutes les 2 heures. Tout d'abord, les taux de mortalité *in situ* des *E. coli* dans la Marne ont été déterminés expérimentalement à l'aide de sacs à dialyse remplis d'eau d'un rejet de station d'épuration. Dans un deuxième temps, les taux de disparition des *E. coli* ont été estimés sur plusieurs futurs sites de baignade et des sites des JOP 2024. Les amplitudes des pics de pollution traduisent une partie de la résistance des sites de baignade et d'activité sportive aux perturbations temporaires générées par les événements pluvieux, et les taux de disparition témoignent d'un aspect du processus de récupération après le pic de pollution. Pour estimer les taux de mortalité et les taux de disparition, il s'agissait de modéliser et de quantifier la diminution des concentrations en *E. coli* au cours du temps. À partir des données extraites des deux bases de données, la résistance et la résilience des sites aux événements polluants (pluie) ont été estimées avec 3 métriques : le temps de retour, l'amplitude de la pollution et l'amplitude de la récupération du site à des niveaux en BIF typiques de temps sec. Ces métriques permettront d'aider à la gestion quotidienne et l'amélioration de la résistance et la résilience des sites des baignades face aux perturbations de court terme.

Chapitre 1 : Étude de la qualité microbiologique des eaux de surfaces : état de l'art

1. Introduction

Au début du XXIe siècle, tant en Europe qu'en Amérique du Nord, les municipalités se tournent progressivement vers leurs espaces bleus (rivières et plans d'eau), les considérant comme des composantes essentielles des projets urbains (Moutiez, 2021). Des efforts importants ont été consentis pour améliorer la qualité des eaux des rivières à des fins récréatives (Kistemann et al., 2016). Durant ces dernières décennies, la qualité des eaux de surface s'est généralement améliorée en Europe, grâce à l'application de la réglementation, à l'amélioration des stations de traitement des eaux usées (STEU) et des réseaux d'assainissement (Houtman, 2010). L'amélioration de la qualité de l'eau a de plus en plus mis l'accent sur la qualité microbiologique, celle-ci étant régulée par la directive 2006/7/CE pour les eaux de baignade (Schreiber et al., 2015).

De ce fait, de nombreuses villes comme Paris, Londres, ou Berlin promeuvent l'ouverture de baignades et l'organisation de compétitions de nage en eau libre dans leurs rivières (Rouillé-Kielo and Bouleau, 2021; Dominguez). Le développement de ces activités augmente le risque d'exposition des baigneurs aux agents pathogènes présents dans l'eau, ce qui peut entraîner des maladies gastro-intestinales, des infections oculaires ou des irritations cutanées (Soller et al., 2010; Mallin et al., 2000).

La France est le deuxième pays européen avec le plus de zones de baignade en eau douce contrôlées par l'Agence européenne pour l'environnement comptant 1286 sites en 2023. Selon l'Agence européenne pour l'environnement (AEE), le classement des eaux de baignade en 2023 en Europe montre que la France est classée en 19^{e} position en prenant en compte la proportion des eaux classées en qualité excellente (AEE, 2024). Sur les 1286 sites en eau douce suivis en France en 2023, 86,6% des zones ont été classées en qualité excellente ou bonne et en plus 3,4% en qualité suffisante, avec une légère dégradation de la qualité sanitaire des eaux depuis 2019 et une amélioration depuis 2023 (Gourmelon, 2023). Parmi ces sites, la région Île-de-France compte 6 baignades en plan d'eau et 3 baignades sur rivière ou canaux (Guide Îles de loisirs,

2024). À l'occasion des JOP 2024, un plan "Qualité de l'eau et baignade" a été lancé en 2016 par le ministère de la transition écologique français, afin de rendre la Seine et la Marne baignables à l'horizon 2024 (Préfecture de la région Île-de-France). Ce plan devrait se traduire par une ouverture de baignades à l'été 2025 sur la Seine et la Marne en région Parisienne (Noury et al., 2018).

2. Historique de la baignade en ville en Ile-de-France

La baignade en Seine et en Marne, aujourd'hui au cœur des débats de santé publique, a connu une évolution marquée par des réglementations variées. Initialement, les interdictions se concentraient davantage sur des questions de décence et de préservation du transport fluvial, comme l'illustre le fascicule de baignade du Piren Seine (Bouleau et al., 2024; Moutiez, 2021).

À partir du XVIIe siècle, l'accès aux rives parisiennes pour la baignade était strictement limité par des autorités soucieuses de garantir l'ordre public et la sûreté, et les dérogations n'étaient accordées que dans des établissements spécifiques, souvent situés sur des bateaux aménagés pour la toilette. Au XVIIe siècle, la baignade dans le fleuve en région parisienne connaît un tel succès que les premières installations apparaissent afin de protéger les baigneurs et de garantir leur sécurité (Moutiez, 2021). Au XIXe siècle, le bassin de la Villette est mis en eau et devient vite un lieu de loisirs aux portes de Paris (Moutiez, 2021). En dehors de la capitale, avec le développement de la banlieue et la croissance des transports ferroviaires, des plages et bains fixes s'installent en bord de Seine et de Marne, souvent associés à des guinguettes et divers services annexes sur la rive (Bouleau et al., 2024).

À partir du milieu du XIXème siècle, de nombreuses piscines municipales ont été installées le long des berges. Les Parisiens ont commencé à profiter des rives de la Seine et de la Marne pour se détendre et se baigner (Kistemann et al., 2016; Passerat et al., 2011). L'industrialisation, l'expansion concomitante de la population vivant dans les villes et l'augmentation de la densité de population au XIXe siècle ont changé cette situation (Houtman, 2010). La pratique de la baignade urbaine, autrefois répandue au début du XXe siècle, a graduellement décliné à mesure que les échanges par voie fluviale se sont accrus et que la qualité de l'eau s'est détériorée. À partir de l'ordonnance préfectorale du 17 avril 1923, la baignade dans la Seine à Paris a été interdite, bien que cette pratique ait perduré jusqu'aux années 1960, avant l'aménagement des voies automobiles le long des berges de la Seine (Guillot-Le Goff et al., 2023).

La baignade a été interdite par la suite en Marne dans le Val-de-Marne en 1970 par un

arrêté préfectoral (Qin et al., 2011). Cette interdiction, motivée par des niveaux de pollution alarmants, entraîne des fermetures massives et marque un tournant vers la prise en compte de la qualité microbiologique de l'eau (Bouleau et al., 2024).

Ces dernières décennies, un regain d'intérêt pour la réintroduction de la baignade urbaine s'est manifesté, reflétant l'engagement renouvelé de la zone métropolitaine en faveur de la restauration écologique des cours d'eau. L'évolution des concentrations en BIF, particulièrement visible dans les données de suivi historique à Ivry-sur-Seine, révèle qu'après des pics de pollution dans les années 1980, des efforts en matière d'assainissement ont permis une réduction significative des contaminants bactériens dans les années 1990 et 2000 (Bouleau et al., 2024).

Le désir politique et sociétal de reconquête des rivières urbaines pour la baignade est de plus en plus pressant en Ile-de-France, que ce soit pour la Seine ou pour la Marne. En prévision des Jeux Olympiques et Paralympiques de Paris en 2024, la municipalité s'est fortement engagée à inclure la Seine dans les épreuves de triathlon et de natation en eau libre (Moutiez, 2021). Cet engagement a donné un élan décisif à un "Plan d'action pour la qualité de l'eau et la baignade" lancé en 2016, visant à améliorer la qualité des eaux de la Seine et de la Marne pour permettre la baignade d'ici 2024, tout en préservant la biodiversité de ces cours d'eau (Guillot-Le Goff et al., 2023; Moutiez, 2021). En plus des événements sportifs, différentes communes prévoient l'ouverture de sites de baignade en héritage des Jeux Olympiques. L'objectif est d'améliorer la qualité de la rivière et d'accompagner les acteurs de bassin versant pour retrouver un jour une eau de baignade conforme.

Ainsi, le contexte réglementaire s'est transformé, passant d'une interdiction motivée par des enjeux de pudeur à une véritable politique sanitaire. Ce retour progressif des baignades surveillées marque une reconquête symbolique et politique de la Seine et de la Marne, nourrie par des aspirations modernes à restaurer des écosystèmes et des lieux de loisirs naturels. Ainsi pour l'établissement d'une nouvelle baignade, la réglementation exige la mise en place d'un profil de baignade qui liste toutes les sources de contamination impactant le futur site en vue d'en faciliter la gestion (Commission européenne, 2006).

3. Sources de contamination

Cependant, les activités sportives et récréatives dans les eaux de surface d'un territoire fortement urbanisé posent des risques sanitaires (Davies-Colley et al., 2018). En effet, la qualité microbiologique des eaux de surface urbaines est fortement dégradée par des rejets d'eaux usées

insuffisamment traitées, comme cela a été précédemment montré pour la Seine (Moulin et al., 2010; Passerat et al., 2011; Lucas et al., 2014; Prevost et al., 2015). Différentes sources de contamination peuvent apporter des flux élevés de pathogènes d'origine fécale au niveau des sites de baignade (Lucas et al., 2019; Mouchel et al., 2020). Les contaminations fécales peuvent être d'origine humaine ou animale (animaux sauvages et domestiques) et provenir de sources ponctuelles telles que les effluents de STEU et de bateaux, les rejets de déversoirs d'orage et d'ouvrages cadres (Passerat et al., 2011; Guérineau et al., 2014; O'Mullan et al., 2017). À cela s'ajoutent des sources diffuses liées au ruissellement sur les surfaces urbaines et agricoles, à la re-suspension des sédiments et aux déjections directes des animaux (Droppo et al., 2011; Ahmed et al., 2019b).

Une estimation des flux moyens estivaux de bactéries fécales dans le bassin versant de l'agglomération parisienne, en amont du pont d'Iéna, détaillée dans le fascicule de baignade du Piren Seine, a révélé les principales sources de contamination bactérienne (Bouleau et al., 2024). Les STEU constituaient environ 38% des apports, suivies par les déversoirs d'orage (19%), les bateaux-logements non raccordés aux réseaux d'assainissement (6%), ainsi que les flux provenant des affluents amont de la Seine-et-Marne (3%), des rivières urbaines en amont (4%) et enfin le ruissellement urbain qui représente moins de 1%. L'analyse indiquait que de faibles rejets non traités suffisaient à compromettre localement la qualité de l'eau, rendant essentielle une vigilance accrue sur ces petites sources de contamination. L'ajout d'installations de désinfection en sortie des STEU Marne Aval et Seine Amont en 2023 a permis de réduire les apports bactériens d'environ 25%, aboutissant à une réduction globale proche de 50% des apports urbains (Bouleau et al., 2024).

Le caractère diffus de nombreuses sources rend difficile la quantification de l'influence relative de chaque source dans un bassin versant donné (Meays et al., 2004). Les rejets de temps de pluie sont souvent décrits comme étant à l'origine de fortes dégradations de la qualité des eaux de surface (Islam et al., 2017). Les événements météorologiques, tels que de fortes pluies, peuvent influencer les risques de contamination en perturbant le sol et en entraînant des débordements (Delamare et al., 2024). De plus, les facteurs météorologiques affectent les concentrations en microorganismes d'origine fécale dans les eaux de surface, notamment la température de l'eau et les caractéristiques des événements pluvieux (intensité, durée, période sèche précédant la pluie). De même, la qualité des eaux usées rejetées dans les rivières, peut varier d'une STEU à l'autre et même au sein d'une STEU en fonction du jour et de la saison

(Kadoya et al., 2019). De plus, l'accumulation et le lessivage de ces microorganismes dans un bassin versant sont influencés par l'usage des sols (Cha et al., 2016; Passerat et al., 2011; Dueker et al., 2017; Droppo et al., 2009; Garcia-Armisen and Servais, 2009).

L'ensemble de ces sources, multiples et variées, rend difficile une estimation précise des flux de contamination. Ceci pose problème car l'exposition à ces contaminants présente un risque pour la santé humaine.

4. Risque sanitaire

Une exposition à l'eau contaminée, pouvant contenir divers types de micro-organismes pathogènes, présente donc un risque accru de contracter des maladies infectieuses (DeNizio and Hewitt, 2019; Mouchel et al., 2020). Les personnes pratiquant des activités en eau douce peuvent présenter des niveaux de vulnérabilité différents en fonction de leur âge, de leur état de santé et de leur connaissance des risques associés à cette activité. Comparés aux individus jeunes et en bonne santé qui ont un système immunitaire plus performant, les personnes âgées, les enfants, les personnes immunodéprimées ou celles mal informées des risques encourus peuvent être plus exposées aux dangers sanitaires (Delamare et al., 2024).

En fonction de divers facteurs tels que la localisation géographique, l'environnement (type de sol, eaux stagnantes, boue, présence d'animaux sauvages ou de bétail) et les conditions météorologiques avant et pendant l'exposition (inondations, fortes pluies), les risques pour la santé liés à l'activité de baignade sont variés. Ils incluent des risques physiques, tels que les noyades, chutes, déshydratation, coups de soleil, qui sont les plus fréquents et graves, mais non liés à la qualité de l'eau (Martinez and Hooper, 2014; Pakasi, 2018). De plus, des micro-organismes tels que les bactéries, les virus et les parasites sont présents dans les milieux aquatiques (eaux côtières, rivières, lacs...), en quantité et diversité variables. Certains de ces micro-organismes peuvent être pathogènes pour l'Homme (Gourmelon, 2023). La présence de germes pathogènes dans l'eau peut entraîner des pathologies affectant, l'appareil digestif, les yeux, les oreilles ou la peau (OMS, 2018). Les pathogènes détectés dans les eaux des sites de récréation européens côtiers et continentaux sont principalement les virus entériques (Bouleau et al., 2024). Aux Pays-Bas, les épidémies associées aux baignades entre 1991 et 2007, étaient à 48% des infections de la peau et à 31% des gastro-entérites (Schets et al., 2011). Selon l'étude de Craun et al. (2005) portant sur la période 1971-2000, les épidémies associées aux eaux de récréation aux Etat-Unis étaient le plus souvent causées par les shigelles (21% des cas),

Naegleria fowleri (17%), *Pseudomonas aeruginosa* (14%), *E. coli* O157 (9%), les norovirus (6%), les leptospires (5%) et les *Giardia* (4%). Il faut toutefois noter que les agents éthiologiques principaux vont varier en fonction du pays, du climat. Par exemple *Vibrio cholerae*, l'agent du cholera, est fréquent dans les eaux de surface de pays Européens (Farrell et al., 2021), mais pas en France. De même, le virus de l'hépatite A et E n'a pas été détecté dans les eaux de la Seine (Prevost et al., 2015). Les bactéries gastro-intestinales transmises par des matières fécales dans l'environnement, telles que les genres *Campylobacter*, *Shigella*, la souche pathogénique d'*E. coli* O157, les salmonelles, sont des sources de gastro-entérite aiguë pouvant être associées aux activités récréatives dans les eaux de surface (Delamare et al., 2024). D'autres maladies comme la leptospirose, causée par la bactérie *Leptospira*, peuvent se transmettre par contact de la peau abîmée ou coupée, des muqueuses ou la conjonctive via l'exposition à de l'eau contaminée (via l'urine d'animaux infectés). Les manifestations cliniques sont comparables aux symptômes pseudo-grippaux, avec une fièvre simple dans la majorité des cas (Delamare et al., 2024). Les *Campylobacter* qui sont très présents dans les rejets de temps de pluie peuvent causer jusqu'à 5% des cas de maladies liée aux activités récréatives en Nouvelle-Zélande (Kistemann et al., 2016). Par contre, les salmonelles et les leptospires sont moins souvent rapportées comme présentes dans les eaux de récréation (Kistemann et al., 2016). De plus, la présence de cyanobactéries et de leurs toxines dans les eaux de baignade peut provoquer des éruptions cutanées, des démangeaisons, des gastro-entérites et atteintes neurologiques, par contact cutané ou ingestion de toxines. Le développement des cyanobactéries est favorisé par l'eutrophisation des eaux, les températures élevées et une faible agitation du milieu (Stewart et al., 2006). Comme mentionné dans le guide de recommandations sanitaires liés aux activités nautiques en eau douce, d'autres infections bactériennes peuvent survenir à la suite d'une exposition à l'eau douce (Agence Régionale de Santé Bretagne, 2017). Les personnes exposées à une forte concentration de *Pseudomonas aeruginosa* sont susceptibles de développer des infections cutanées, des otites, des conjonctivites ou des infections urinaires. Par exposition à de l'eau contaminée dans les environnements d'eau douce, ces infections peuvent entraîner divers problèmes de santé (Agence Régionale de Santé Bretagne, 2017; Delamare et al., 2024). Des pathogènes opportunistes autochtones du milieu aquatique comme les légionelles, les mycobactéries, et les *Aeromonas* peuvent provoquer des infections respiratoires (de Roda Husman and Schets, 2010).

Des gastro-entérites aïgues liées aux eaux récréatives sont souvent dues à des virus entériques (Mouchel et al., 2020). Leur transmission se fait par voie oro-fécale, soit par conta-

mination de contact, soit par la consommation d'eau contaminée, et leurs doses infectieuses sont très faibles, ce qui génère un risque important de gastro-entérite virale chez les nageurs (Bouleau et al., 2024). Les adénovirus et les norovirus sont très fréquents dans les eaux de surface et peuvent atteindre des concentrations relativement élevées même en dehors des périodes épidémiques (Prevost et al., 2015; Korajkic et al., 2018). Les épidémies de norovirus peuvent avoir un impact significatif sur le système de santé local et entraîner des épidémies secondaires avec transmission entre les malades et leurs proches (Delamare et al., 2024). Ces virus peuvent provoquer divers symptômes tels que des troubles intestinaux et respiratoires, des hépatites et des conjonctivites (Mouchel et al., 2020). Les norovirus sont la principale cause d'infection gastro-intestinale non bactérienne dans le monde. Les symptômes apparaissent après une période d'incubation moyenne de 24 à 48 h et durent généralement entre 12 et 72 h. Cependant, les formes sévères sont plus rares chez les patients adultes en bonne santé, comparés aux enfants (Delamare et al., 2024). D'autres virus entériques pouvant être impliqués dans les gastro-entérites humaines ont été identifiés au niveau de la Seine et de la Marne (aichivirus, rotavirus, entérovirus) (Prevost et al., 2015). La composition complexe des eaux et la sensibilité différente des espèces de virus rendent difficile la prévision du comportement des virus entériques (Kadoya et al., 2019). Les virus sont souvent très persistants dans l'environnement aquatique, et les variations de température, surtout les températures basses, favorisent leur survie (Ibrahim et al., 2019; Dean and Mitchell, 2022).

Parmi la large gamme de maladies infectieuses, il y a également les infections parasitaires qui peuvent être contractées dans les eaux de surface. Les contaminations par des parasites d'origine animale ou humaine (comme *Giardia* et *Cryptosporidium*) ou environnementale comme (*Naegleria floweri*) sont également les agents de maladies d'origine hydrique (Pakasi, 2018; Delamare et al., 2024). Les infections par les cryptosporidies sont toutefois intermittentes, essentiellement en lien avec un bassin versant agricole (Kistemann et al., 2016). L'amibe *Naegleria floweri* prolifère plutôt dans les eaux chaudes, et aucun cas n'a été rapporté en France dans les eaux de surface non polluées thermiquement (De Jonckheere, 2011).

5. Notion d'indicateur de contamination fécale

Pour s'assurer que le risque lié aux eaux récréatives est réduit au minimum pour le public, de nombreux gouvernements et autorités ont mis en place des mesures de qualité de l'eau (Avila et al., 2018; Visser et al., 2022). Il est difficilement faisable de mesurer l'ensemble

des pathogènes en routine, surtout qu'ils sont souvent en concentration faible dans les eaux de surface et que leur quantification demande des techniques de biologie moléculaire. La stratégie adoptée est de mesurer des microorganismes non-pathogènes, autochtones du tube digestif, qui sont soit présents en plus grand nombre que les pathogènes, soit présents en même temps que les pathogènes, et dont la mesure est peu coûteuse et facile à mettre en oeuvre. Ainsi, les paramètres recommandés par la directive 2006/7/EC pour évaluer la qualité de l'eau de baignade sont les BIF (*E. coli* et les entérocoques intestinaux). La qualité microbiologique des eaux récréatives est généralement évaluée par la présence de bactéries indicatrices de contamination fécale (Avila et al., 2018). *E. coli* et les entérocoques intestinaux sont des éléments du microbiote intestinal des mammifères et des oiseaux, de certains reptiles et des humains (Gordon, 2013; Byappanahalli et al., 2012; Staley et al., 2014; Silva et al., 2012). *Escherichia coli* est un bacille à Gram-négatif, appartenant au groupe des coliformes fécaux, classés dans le phylum des gamma-Protéobactéries et la famille des *Enterobacteriaceae*. Son habitat primaire est le bas intestin des animaux à sang chaud, incluant les humains (Ishii and Sadowsky, 2008). En général, on dénombre plus de 1 million de d'*E. coli* par g sec de fèces humaines (Ishii and Sadowsky, 2008). Les *Enterococcus* sont des coques Gram-positives, catalase-négatives, non sporulantes et anaérobies facultatives (Fisher and Phillips, 2009). Elles habitent généralement le tractus intestinal des humains, mais peuvent aussi être isolées de diverses sources environnementales et animales. Capables de résister à des conditions extrêmes, elles survivent à des températures allant de 5 à 65°C, à des pH entre 4,5 et 10,0, ainsi qu'à des concentrations élevées de NaCl, ce qui leur permet de coloniser divers milieux (Fisher and Phillips, 2009). Typiquement, elles représentent moins de 0,1% de la flore intestinale humaine (Schloissnig et al., 2013). Parmi les plus de 50 espèces du genre *Enterococcus* identifiées, *E. faecium* et *E. faecalis* sont les plus fréquentes dans le tractus gastro-intestinal humain et animal, avec des concentrations de l'ordre de 10000 à 1 million de cellules par g de fèces humaines (Boehm and Sassoubre, 2014).

Les niveaux de ces deux BIF sont indicatifs de la pollution fécale (Commission européenne, 2006). En effet, des études épidémiologiques ont montré la capacité des concentrations en BIF à prédire les risques de gastroentérites dans les eaux de surface et ont ainsi permis d'établir des seuils réglementaires (Prüss, 1998; Pond, 2005; Shuval, 2003). Prüss (1998) a passé en revue 37 études épidémiologiques sur les effets sur la santé de l'exposition aux eaux récréatives et a constaté pour la majorité des études une association positive, statistiquement significative, entre le nombre de BIF présentes et le risque de contracter une gastroentérite pour les nageurs.

Une méta-analyse réalisée par Wade et al. (2003) de plus de 900 études a révélé qu'au niveau des eaux douces *E. coli* était un prédicteur de maladie gastro-intestinale plus cohérent que les entérocoques et d'autres indicateurs bactériens. Ils ont constaté qu'une augmentation du nombre d'*E. coli* était associée à une augmentation non significative moyenne du risque relatif. Ces valeurs sont à mettre en regard des risques calculés par les études rétrospectives qui ont servi à fixer les seuils de qualité des eaux de baignade dans les réglementations de tous les pays. En Europe, les taux d'incidence de gastro-entérites considérés acceptables sont fixés à 3% (eau continentale de qualité "excellente") et 5% (eau continentale de qualité "bonne") pour le classement des sites de baignade dans la directive européenne 2006/7/EC (Fleisher et al., 1996). L'étude relative à la prévention des maladies gastro-intestinales par les agences de protection de l'environnement dans les eaux récréatives recommande que des études futures se concentrent sur la capacité de nouvelles méthodes microbiennes, plus rapides et plus spécifiques, à prédire les effets sur la santé et à estimer les risques d'exposition aux eaux chez les personnes sensibles (Wade et al., 2003).

6. Evaluation de la qualité microbiologique de l'eau de baignade

Actuellement, la qualité de l'eau de surface est principalement évaluée à l'aide d'échantillons d'eau collectés pour une analyse microbiologique et chimique en laboratoire et/ou à l'aide de capteurs spécifiques à haute précision placés à des endroits fixes. La surveillance réglementaire des eaux de baignade en Europe est basée sur la culture des *E. coli* et des entérocoques intestinaux (EI) (Commission européenne, 2006). L'abondance de ces bactéries indique le niveau de contamination fécale et donc la présence éventuelle de pathogènes pouvant être à l'origine de maladies gastro-intestinales (Commission européenne, 2006; OMS, 2018). Selon la directive 2006/7/CE, une eau de baignade continentale est jugée comme étant de qualité suffisante si la valeur du percentile 90 sur 16 mesures pendant 4 années est en dessous de 900 NPP/100 mL pour *E. coli*, et en dessous de 330 NPP/100 mL pour les EI (Tableau 1.1). En France, ces seuils spécifiques aux eaux continentales permettent aux Agences Régionales de la Santé (ARS) de classer chaque année les sites de baignade d'eau douce.

En cours de saison de baignade, la qualité microbiologique est évaluée en fonction des seuils définis pour les BIF, comme indiqué dans le tableau 1.2 et selon l'instruction n°

TABLE 1.1 – Seuils de qualité microbiologique pour le classement des sites de baignade selon la directive 2006/7/EC. Basé sur l'évaluation du percentile 95 (*) et 90 (**).

Paramètre	**Excellente**	**Bonne**	**Suffisante**
Escherichia coli (NPP/100 mL)	500 (*)	1 000 (*)	900 (**)
Entérocoques intestinaux (NPP/100 mL)	200 (*)	400 (*)	330 (**)

DGS/EA4/2022/168 du 17 juin 2022 relative aux modalités de recensement, gestion et classement des eaux de baignade. Les prélèvements dont les résultats sont classés comme "bon" ou "moyen" sont considérés conformes, tandis que les résultats qualifiés de "mauvais" sont jugés non conformes. Ces seuils ont été établis en lien avec les risques sanitaires observés, notamment un risque accru de gastro-entérite pour les concentrations comprises entre 900 et 1800 NPP/100 mL, avec un risque supérieur à 5% (Duboudin et al., 2007). Ces seuils servent pour la gestion active des baignades classées : prélèvements supplémentaires, ouverture et fermeture des zones de baignade (Bouleau et al., 2024).

TABLE 1.2 – Valeurs limites de qualité microbiologique des eaux intérieures d'un site de baignade classé, pour la gestion active en cours de saison, proposées par l'Agence française de sécurité sanitaire de l'environnement et du travail (Duboudin et al., 2007).

Paramètre	**Bonne**	**Moyenne**	**Mauvaise**
Escherichia coli (NPP/100 mL)	< 100	$< 1\,800$	$\geq 1\,800$
Entérocoques intestinaux (NPP/100 mL)	< 100	< 660	≥ 660

La gestion quotidienne des sites de baignade implique un suivi précis de la qualité microbiologique impactée par les pollutions de courte durée. Une pollution à court terme, définie à l'article D.1332-15 du code de la santé publique comme étant une contamination microbiologique affectant la qualité de l'eau de la baignade pendant moins de 72 h et dont les causes sont aisément identifiables, peut être déterminée par un dépassement de l'une des valeurs seuils proposées par l'agence française de sécurité sanitaire de l'environnement et du travail (AFSSET) pour les BIF (Duboudin et al., 2007). En cas de pollution de court terme, souvent provoquée par des précipitations importantes difficiles à prévoir, qui génèrent des rejets urbains, du ruissellement sur des surfaces contaminées et des rejets accidentels non maîtrisés, devrait donc entraîner des fermetures préventives (Penna et al., 2021; Bouleau et al., 2024).

Une gestion effective des fermetures doit à la fois permettre de préserver la santé publique et aussi l'économie locale liée aux activités de baignade (Penna et al., 2021). Depuis l'adoption de la directive sur les eaux de baignade en 2006, la proportion de sites classés comme "excellent" a augmenté, puis s'est stabilisée ces dernières années. En 2023, cette proportion représentait

85% de l'ensemble des eaux de baignade dans l'UE (22000 sites côtiers et continentaux) mais ce taux était seulement de 79% pour les sites continentaux (European Commission, 2023). Cela souligne la nécessité de mettre en place des systèmes d'alerte précoce fiables pour les eaux de baignade. L'OMS préconise d'utiliser la modélisation en complément du suivi réglementaire pour estimer ou prédire les contaminations et ainsi aider à la gestion quotidienne des sites de baignade (OMS, 2018). La modélisation pourrait être utilisée comme système d'alerte précoce en cas de pollution à court terme, en offrant une aide précieuse dans la gestion des fermetures temporaires des sites de baignade (OMS, 2018). Bien que l'ouverture reste soumise à une confirmation par une mesure réglementaire, la modélisation permettrait d'optimiser les efforts d'échantillonnage et de suivi (Seis et al., 2018). Cette approche contribuirait non seulement à raccourcir les périodes de fermeture mais aussi à anticiper les risques de contamination fécale et ainsi améliorer la gestion de la qualité des eaux de surface (Bouleau et al., 2024).

7. Variabilité spatiale et temporelle de la qualité microbiologique

Concernant le suivi de la qualité microbiologique, des questions subsistent quant à la stratégie d'échantillonnage nécessaire pour obtenir des mesures représentatives (Harmel et al., 2016; McCarthy et al., 2008). Il est essentiel de prendre en compte l'hétérogénéité spatiale et temporelle des sources de contamination, qui peut affecter significativement la précision des évaluations de la qualité de l'eau (Harmel et al., 2016). La variabilité spatiale des concentrations en BIF dans les milieux aquatiques peut s'observer à différentes échelles : d'une petite distance au sein d'un même site d'échantillonnage, à une grande distance le long d'un bassin versant ou à l'échelle de la région (Murphy et al., 2023). Cependant, les sites de baignade d'eau douce sont encore peu étudiés de ce point de vue comparés aux plages côtières. Cela est en particulier vrai pour les baignades en rivières. Plusieurs études (Quilliam et al., 2011; Weller et al., 2020) ont montré que le niveau en FIB variait significativement au sein d'un même site, par exemple d'une berge à l'autre pour une rivière (Quilliam et al., 2011). La distribution horizontale des BIF en rivière est influencée par les conditions d'écoulement du cours d'eau, qui modifient le degré de mélange et, par conséquent, la cohérence des concentrations bactériennes. Les débits plus faibles en bordure, par exemple, favorisent un dépôt bactérien plus élevé (Harmel et al., 2016; Salam et al., 2021). De plus, une incertitude spatiale peut être observée verticalement, en

raison de la remise en suspension des sédiments et de l'influence des UV qui pénètrent dans la colonne d'eau (McCarthy et al., 2008; Quilliam et al., 2011; Harmel et al., 2016). Ainsi, il a été montré sur une plage côtière que les concentrations en EI étaient 10 fois plus élevées dans les échantillons collectés à hauteur des genoux comparés à ceux prélevés à hauteur de la taille (Enns et al., 2012). À une échelle plus large, les sites échantillonnés au sein d'un même étang cessent de montrer une autocorrélation dans les mesures de BIF s'ils sont espacés de plus de 100 m. De ce fait, un seul échantillon par étang ou lac ne peut pas refléter toute la variabilité spatiale qui y est présente (Murphy et al., 2023).

A cette variabilité spatiale se superpose une incertitude temporelle des indicateurs fécaux à différentes échelles de temps : entre les heures du jour, entre les jours, entre les saisons, inter-annuelle. Ainsi, il a été montré que sur les plages de Chicago (USA), la profondeur de l'eau et l'heure du jour influençaient significativement la variation des concentrations en BIF, expliquant respectivement 7 et 13% de la variabilité (Whitman and Nevers, 2008). Une variation au cours d'une journée de baignade, avec des fluctuations de concentration peut être observée sur des intervalles de quelques minutes à quelques heures seulement (Wyer et al., 2018; Wymer et al., 2007; Boehm et al., 2002). Des études montrent ainsi des variations journalières significatives, avec des écarts pouvant atteindre 1 à 2 log_{10} dans une seule journée d'échantillonnage, et ce, même en l'absence de conditions météorologiques défavorables (Wyer et al., 2018). En effet, les prélèvements effectués le matin sont généralement plus représentatifs, car la qualité de l'eau tend à être meilleure l'après-midi, probablement en raison des rejets d'eaux usées du matin et des habitudes de vie (Jozić et al., 2024; Jovanovic et al., 2019). Toutefois, il est montré que le jour de l'échantillonnage a plus d'importance dans l'explication de la variation des concentrations en BIF des plages lacustres que l'heure du jour (Whitman and Nevers, 2008). A plus large échelle temporelle (inter-annuelle), l'analyse des variations des concentrations en BIF peut permettre de mieux comprendre les facteurs et les mesures de gestion qui influencent la qualité microbiologique de l'eau des plages d'eau douce (Weiskerger and Whitman, 2018). De ce fait, les variations de facteurs physiques, chimiques, biologiques de la rivière et le climat local et régional influencent directement les stratégies de surveillance (fréquence et positionnement des échantillonnages), l'interprétation des résultats de qualité de l'eau et leur modélisation (Quilliam et al., 2011). Différents paramètres tels que les sources de contamination proches, les événements en amont comme les rejets et les conditions météorologiques, ajoutent une couche de complexité temporelle et spatiale (Devane et al., 2020). Les facteurs environnementaux tels que

la température, les rayons UV, la sédimentation et le niveau d'étiage influencent la concentration et la dispersion des BIF en modifiant leur persistance et leur répartition dans l'eau (Ishii and Sadowsky, 2008). La remise en suspension des sédiments, qui peuvent représenter un réservoir important de BIF, ajoute une source importante de contamination, complexifiant ainsi les prévisions de qualité de l'eau (Piorkowski et al., 2014). Cependant, malgré la variabilité spatiale et temporelle intra-journalière et inter-saisonnière, les stratégies d'échantillonnage actuelles persistent à considérer qu'un seul prélèvement par semaine à un point situé à 2 m de la berge est représentatif de la qualité de l'eau pour l'ensemble du site et pour toute la journée de baignade (Wyer et al., 2018; Boehm, 2007). Or, cette approche ne reflète pas les fluctuations réelles de la journée, ni la variabilité spatiale du site de baignade, ni le délai de 24 h pour obtenir les résultats en réduit l'utilité opérationnelle, la qualité de l'eau pouvant changer considérablement entre le prélèvement et la disponibilité des données. Cette complexité a des implications cruciales pour les stratégies de suivi, car elle exige une prise en compte des dynamiques locales et globales pour une évaluation fiable de la qualité des eaux de baignade. La fréquence d'échantillonnage et la taille de l'échantillon déterminent la représentativité de la variabilité de la qualité de l'eau et peuvent ainsi avoir un effet crucial sur le calcul des percentiles 90 et 95, influençant ainsi la classification des sites de baignade (López et al., 2012). Ce point est particulièrement critique pour les sites dont la qualité est proche du niveau "suffisant". Ces connaissances sont essentielles pour adapter les modèles de prévision, qui peuvent servir de complément aux méthodes de surveillance directe, en particulier dans les collectivités de taille modeste où les ressources sont limitées.

8. Décroissance des indicateurs de contamination fécale

Le tube digestif des humains et animaux homéothermes offre aux BIF et aux pathogènes entériques des conditions favorables à leur croissance. Les proportions de BIF par gramme de fèces varient en fonction des espèces hôtes (Dean and Mitchell, 2022). Après le rejet d'eaux usées dans le milieu, la concentration des BIF et des agents pathogènes peut être modifiée par la dilution, le débit de l'eau et la capacité de persister de chaque espèce microbienne dans l'environnement (Devane et al., 2020). La proximité du site de baignade avec une source de pollution augmente la densité des BIF, qui diminue avec l'éloignement en raison de la dilution et de la dispersion des contaminants dans la masse d'eau (Jozić et al., 2024; Carneiro et al., 2018). Une fois excrétées et déversées dans l'environnement, les bactéries du tube digestif sont

exposées à divers facteurs, tels que la disponibilité en nutriments et en sources de carbone organique, les fluctuations de température, la salinité et la prédation, dont l'influence dépend des caractéristiques physiques et chimiques propres à chaque milieu (Sampson et al., 2006; Schloissnig et al., 2013; Solo-Gabriele et al., 2000; Nakhle et al., 2021). Dans cet habitat secondaire plus ou moins hostile, en théorie, les BIF vont soit mourir, soit entrer en dormance dans un état viable non cultivable. Cependant, il a été montré que dans certains compartiments du milieu aquatique tels que les sédiments, les litières ou les biofilms, des souches de ces BIF peuvent survivre, voire s'acclimater et croître (Korajkic et al., 2019). Par exemple, des températures élevées et la présence de matière organique peuvent favoriser la survie d'*E. coli* hors de son hôte, tandis que l'exposition à la lumière et la prédation contribuent à réduire leur présence. Étant donné la complexité des systèmes aquatiques, il est difficile de prédire l'influence de chaque facteur sur la survie et la croissance des bactéries dans des contextes variés (Solo-Gabriele et al., 2000; Sampson et al., 2006). La prédation a notamment démontré un rôle important dans la survie d'*E. coli* au sein des systèmes naturels (Solo-Gabriele et al., 2000).

Une gestion efficace de la qualité de l'eau exige une connaissance approfondie de la dynamique de décroissance des BIF dans les habitats aquatiques, ainsi que des facteurs qui l'influencent. Si la décroissance d'*E. coli* est relativement constante d'un événement polluant à un autre au sein d'un même site ou bien d'un site à un autre, cela pourrait permettre d'avoir une future utilisation de ce paramètre par les gestionnaires pour prédire le devenir des contaminations (Dick et al., 2010).

La plupart des études sur la décroissance des bactéries ont été menées dans des conditions contrôlées en laboratoire ou *in situ* pour déterminer le taux de mortalité (Dick et al., 2010; Korajkic et al., 2014; Tijdens et al., 2008). La majorité de ces études *in situ* utilise des microcosmes fermés (bouteilles) ou semi-ouverts composés de sacs à dialyse immergés dans l'eau de surface. Ces systèmes permettent de simuler la décroissance bactérienne suite à un rejet accidentel d'eaux usées tout en manipulant des facteurs tels que les UV ou la prédation. Par contre, ces expériences en laboratoire ou sur le terrain ne permettent pas de prendre en compte la dynamique liée aux flux d'apports amont, à la dilution et à la dispersion, ni à l'effet de la sédimentation et de la resuspension des sédiments (Maraccini et al., 2016; Ahmed et al., 2015; Korajkic et al., 2014; Nakhle et al., 2021).

L'étude de Jin et al. (2004) illustre la dynamique de décroissance après un événement

TABLE 1.3 – Valeurs du taux de décroissance dans différents types d'eau, le terme utilisé dans la littérature et les facteurs pris en compte pour la mesure du taux. [a] : sans remise en suspension, [b] : avec remise en suspension, [c] : (Noble et al., 2004), [d] : (Servais et al., 2007a), [e] : (Chigbu et al., 2005), [f] : (Jozić et al., 2014), [g] : (Nakhle et al., 2021), [h] : (Dick et al., 2010), [i] : (Blaustein et al., 2013).

Milieux	**Terme utilisé**	**Taux (Jr^{-1})**	**Facteurs pris en compte**
Mésocosme (Rivière)	Inactivation	3.12 et 1.12 (*E. coli*)[c]	Forte et faible irradiation solaire
Mésocosme (Rivière)	Inactivation	6.48 et 5.76 (EI) [c]	Forte et faible irradiation solaire
Laboratoire (Seine)	Mortalité	0.72 (*E. coli*) [d]	Prédation et stress physiologique
Laboratoire (Seine)	Mortalité	1.08 (Coliformes fécaux) [d]	Mortalité et perte de culturabilité (lumiére)
Rivière	Disparition	de 0.21 à 0.74 (Coliformes fécaux) [e]	Prédation, nutriment, sédimentation et lumière
Laboratoire	Inactivation	de 0.01 à 7.91 (*E. coli*) [f]	Irradiation solaire et obscurité
Mésocosme (Bassin versant) [a]	Décroissance apparente	1.43 ± 0.15 (*E. coli*) [g]	Sédimentation et lumière
Mésocosme (Bassin versant) [b]	Décroissance apparente	0.50 ± 0.15 (*E. coli*) [g]	Sédimentation et lumière
Laboratoire (Rivière et eau usée)	Décroissance	0.28 [h]	Réduction de la prédation
Laboratoire (rivière)	Inactivation	0.72 ± 0.07 (*E. coli*) [i]	Obscurité et température 20°C
Laboratoire (Eau usée)	Inactivation	0.67 ± 0.11 (*E. coli*) [i]	Obscurité et température 20°C

pluvieux et montre une diminution relativement rapide de ces indicateurs en deux à trois jours (Tableau 1.3). Cette décroissance rapide est attribuée à plusieurs facteurs, dont l'effet de dilution, la mortalité des microorganismes due à des conditions environnementales telles que les toxines algales, le pH, la prédation, la température, la salinité et la lumière solaire, ainsi qu'à la sédimentation des particules auxquelles les microorganismes peuvent être associés (Pendergrass et al., 2015; Gronewold et al., 2011). De plus, ce taux peut varier en fonction des saisons. Ainsi, des taux plus élevés étaient observés au centre-nord du golfe du Mexique en hiver avec des taux estimés en novembre/décembre de 0,64 ± 0,06 jr^{-1}, en janvier de 0,45 ± 0,03 jr^{-1} et en février/mars de 0,35 ± 0,03 jr^{-1}, probablement en raison des faibles températures de l'eau et de la baisse d'intensité du rayonnement solaire (Chigbu et al., 2005).

De même, la concentration initiale en BIF peut avoir un impact sur le taux d'inactivation (Gronewold et al., 2011). L'étude de Nakhle et al. (2021) a identifié un taux de décroissance plus

élevé avec une remise en suspension des sédiments. La sédimentation expliquait en moyenne 92% de la réduction de la concentration d'*E. coli*, tandis que le rayonnement solaire représentait environ 2% (Tableau 1.3).

9. Incertitudes sur la mesure des indicateurs bactériens

La variabilité spatiale et temporelle ajoute une part d'incertitude sur l'échantillonnage et donc sur les concentrations en BIF rapportées. Celle-ci est dépendante de l'effort d'échantillonnage (fréquence des dates d'échantillonnage et choix des points d'échantillonnage). À cela s'ajoute une part d'incertitude qui est liée à la méthodologie et aux équipements employés pour effectuer l'échantillonnage et l'analyse des échantillons, ainsi qu'une part d'incertitude liée à l'expertise des personnels du laboratoire. L'incertitude caractérise la dispersion des valeurs qui pourraient être raisonnablement attribuées à un ensemble de facteurs. L'incertitude associée à une mesure est un paramètre important à prendre en compte car elle nous renseigne sur la fiabilité des résultats et donc ensuite sur la confiance dans la prise de décision (Cazals et al., 2020).

9.1. Sources d'incertitudes

Les stratégies afin de diminuer l'incertitude sur la mesure et l'échantillonnage comportent la formation et la certification des personnels, l'accréditation du laboratoire, l'utilisation de protocoles normés, l'utilisation de standards et de contrôles, l'inter-comparaison entre laboratoires. Notons que les méthodes et normes de prélèvements et d'analyses peuvent différer d'un pays à l'autre (Europe, Asie, Pacifique) (Cazals et al., 2020). En Europe et plus particulièrement en France, les normes de prélèvements et de mesures sont celles figurant dans la figure 1.1. Cependant, ces normes laissent une marge d'interprétation qui peut être source d'incertitude.

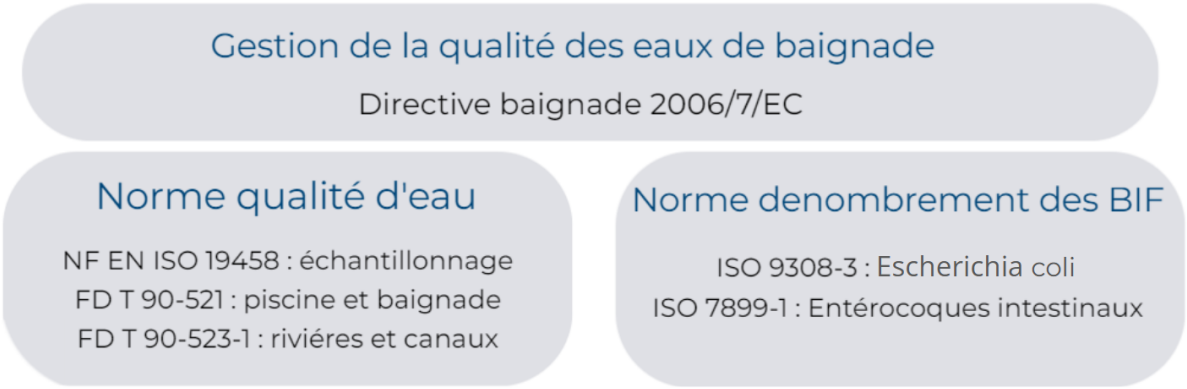

FIGURE 1.1 – Normes, directives et guide pour le prélèvement d'eau de baignade.

Globalement les différentes incertitudes vont se cumuler. La variabilité spatiale et temporelle constitue une première source majeure d'incertitude qui influence les ensembles de données. S'ajoute à cela une incertitude liée au processus de la collecte d'échantillons à l'analyse en laboratoire (Harmel et al., 2016). Une incertitude globale de ±33% (15-67%) a été calculée lors des prélèvements de temps de pluie à l'aide de préleveurs automatiques, en prenant en compte l'échantillonnage, le stockage et l'analyse (McCarthy et al., 2008). Cependant, en raison des variations d'incertitude selon la méthode, des recherches supplémentaires sont nécessaires pour chaque nouveau système de surveillance (McCarthy et al., 2008).

9.2. Incertitude liée à la variabilité spatio-temporelle

Lors du prélèvement d'un échantillon à un instant précis, de l'incertitude est introduite par le moment et le lieu de la collecte (Harmel et al., 2016). Une différence significative dans les concentrations en *E. coli* prélevées à des intervalles de deux heures dans un bassin hydrographique urbain de Houston (Texas) a été constatée (Desai and Rifai, 2013). En revanche, aucune corrélation significative entre les concentrations d'*E. coli* et l'heure de l'échantillonnage n'a été constatée dans des rivières échantillonnées par Pendergrass et al. (2015) et Sejkora et al. (2011). Une incertitude de ±(23 ± 16%) a été identifiée avec un échantillonnage répété espacé d'une minute (Pendergrass et al., 2015). À cela s'ajoute une incertitude spatiale verticale : les échantillons prélevés dans le haut et le bas de la colonne d'eau d'un rejet pluvial présentaient une incertitude moyenne de 1 ± 27% (McCarthy et al., 2008). Cette incertitude spatiale peut également être horizontale, des différences significatives dans les concentrations d'*E. coli* ayant été observées au sein du transect d'un système fluvial britannique, avec une incertitude moyenne de 62 ± 30% (Quilliam et al., 2011).

Cependant, l'impact des sources diffuses de pollution liées au ruissellement par temps de pluie ou à la défécation directe des animaux dans les rivières ou encore la variabilité des rejets ponctuels liée au caractère aléatoire des orages d'été vont impacter la variabilité spatiale et temporelle des concentrations en BIF dans les eaux de surface. Ainsi, les variations saisonnières de la distribution et l'activité des animaux sauvages ou les changements de pâturage pour le bétail peuvent constituer une source importante d'incertitude (Guérineau et al., 2014). Ces facteurs, qui varient considérablement dans le temps et l'espace, peuvent générer des niveaux d'incertitude élevés, comme le souligne Pendergrass et al. (2015), notamment en ce qui concerne l'impact potentiel des colonies d'oiseaux pouvant entraîner une incertitude dépassant 1000%.

9.3. Incertitude liée à la méthodologie de prélèvement

Lors du prélèvement, quelle que soit la méthode, l'incertitude est introduite par le moment et le lieu de la collecte de l'échantillon (incertitude temporelle et spatiale) mais également par le volume prélevé et l'équipement utilisé qui peut potentiellement générer des contaminations croisées entre les sites (Harmel et al., 2016; Hathaway et al., 2014). Les trois sources principales d'incertitude sur la concentration en *E. coli* dans l'échantillon sont l'échantillonnage, le stockage et l'analyse (McCarthy et al., 2008). Nous devons garder à l'esprit que la première incertitude dans les résultats peut être liée au prélèvement dans le cours d'eau. Plusieurs méthodes de prélèvements peuvent être appliquées sur le terrain (prélèvement automatique, prélèvement manuel avec une perche équipée d'un flacon ou une pompe) à une profondeur d'eau et à une distance de la berge qui peuvent dépendre des caractéristiques du site mais aussi de l'équipement (longueur de la perche, puissance de la pompe et longueur du tuyau par exemple). Les prélèvements bactériologiques y sont sensibles à ces caractéristiques et les normes NF EN ISO 19458 et FD T90-523-1, ainsi que la directive 2006/7/EC spécifient des profondeurs et distances minimales requises tout en laissant une marge de manœuvre.

Selon la norme FD T90-523-1 sur la qualité de l'eau dans l'environnement, plusieurs équipements peuvent être employés pour le prélèvement ponctuel : bécher associé à une perche télescopique, pompe dont le tuyau est associé à une perche télescopique, seau. Il existe aussi des bouteilles de prélèvements adaptées pour échantillonner à une profondeur donnée (exemple la bouteille de Niskin). Des mesures sont nécessaires pour limiter l'incertitude et les risques de contamination lors des prélèvements d'eau. Selon les normes NF EN ISO 19458 et FD T

90-52, l'utilisation de flacons stériles et de lingettes désinfectantes pour les béchers permet de minimiser les contaminations externes, mais les prélèvements effectués à l'aide de pompes et de tuyaux sont plus complexes à nettoyer et désinfecter que ce soit pour les prélèvements ponctuels ou intégrés (pompe associée à une perche ou préleveur automatique). En effet, ces dispositifs ont souvent un volume mort et de ce fait peuvent héberger une contamination résiduelle ou favoriser la formation de biofilms à l'intérieur des tuyaux, ce qui augmente le risque de contamination croisée entre échantillons (Solo-Gabriele et al., 2000). La norme FD T90-523-1 prévoit effectivement un rinçage du système avec de l'eau de rivière avant de réaliser un prélèvement, afin de minimiser les risques de contamination. Cependant, ce rinçage peut ne pas être suffisant pour garantir des résultats représentatifs dans certaines conditions. En effet, plusieurs facteurs peuvent influencer l'efficacité du rinçage, tels que la longueur du tuyau, son inclinaison et l'exposition au rayonnement solaire (Hathaway et al., 2014).

Ainsi, l'étude de Hathaway et al. (2014) a révélé une contamination <1% dans la tubulure d'un préleveur automatique après 7 jours à sec. Cependant, l'étude de Galfi et al. (2014) montre une influence de la longueur du tube ainsi qu'un potentiel de contamination croisée avec des échantillons successifs de concentrations variables. De plus, Hathaway et al. (2014) a montré un impact de l'inclinaison du tuyau sur la stagnation du volume mort d'échantillon prélevé. En inclinant le tuyau de prélèvement pour permettre son drainage complet entre les échantillons, l'incertitude diminue de 5.5 ± 0.05% à 1.7 ± 0.02%. En raison de la contamination potentielle, un lavage et un rinçage avec de l'eau déionisée et autoclavée les tuyaux d'échantillonnage entre les prélèvements est recommandé (Hathaway et al., 2010). Cette contamination résiduelle étant généralement négligeable pour les eaux de surface, un rinçage est parfois suffisant (Hathaway et al., 2014).

Une fois l'échantillon collecté, le délai entre la collecte et l'analyse d'un échantillon doit être aussi court que possible afin de limiter les changements dans les populations microbiennes (Salam et al., 2021). La température de stockage entre le prélèvement et l'analyse au laboratoire joue aussi un rôle important. D'une part, dans le cas des échantillonnages par préleveur automatique, le temps entre le premier prélèvement et le dernier peut représenter jusqu'à 24 h or il n'est pas toujours possible d'avoir une embase réfrigérée pour des raisons d'alimentation en électricité ou de place. D'autre part, quelque soit le type de prélèvement (manuel ou automatique) le temps de transit des échantillons entre le lieu de prélèvement et le laboratoire peut être long et dans des conditions de température pas toujours contrôlables. À température ambiante,

Salam et al. (2021) ont constaté dans 80% des cas étudiés, aucune différence significative entre les concentrations en *E. coli* des échantillons traités dans les 8 h et dans les 24 h, avec une incertitude moyenne de ±12% et une tendance à décroître au cours de la période de stockage. En effet, McCarthy et al. (2008) ont constaté que lors d'un stockage non réfrigéré dans un préleveur automatique, une augmentation des concentrations d'*E. coli* est observée entre 4 et 8 h, puis une diminution à 24 h à une température entre 10 et 15°C. Ainsi, après 24 h, l'incertitude moyenne due au stockage était de ±25% (McCarthy et al., 2008). Selon la température ambiante, les résultats peuvent fluctuer, et la saison du prélèvement est donc à prendre en compte, ainsi que la latitude du site de prélèvement. Ainsi, pour une durée de stockage de 6 h, l'incertitude moyenne est légèrement plus élevée lorsque les échantillons étaient conservés à 25°C (+8%) par rapport à 15°C (+6%) (Harmel et al., 2016). La réfrigération à des températures inférieures à 6°C permet une meilleure stabilité des échantillons. En effet, une étude de la Texas Commission on Environmental Quality n'a fait état que d'une faible incertitude de -4% des concentrations d'*E. coli* après 24 h par rapport à une durée de stockage de 8 h à une température inférieure à 4°C (Millican and Hauck, 2008). Un changement a été rapporté dans les échantillons d'eau stockés à une température inférieure à 10°C, avec une incertitude moyenne allant de 1% après 6 h à 20% après 24 h (Agency, 2006). Ainsi la norme FD T90-523-1 préconise un stockage maximal de 24 h à 5 ± 3 °C.

9.4. Incertitude liée à l'analyse

Il existe une incertitude pour la mesure de la concentration en bactéries liée à l'analyse en laboratoire de l'échantillon prélevé (Harmel et al., 2016). En comparaison avec l'incertitude d'échantillonnage qui est peu décrite et sous-estimée, l'incertitude analytique est quant à elle contrôlée et bien rapportée dans la littérature (Guigues et al., 2020). L'incertitude caractérise la dispersion des valeurs qui pourraient être raisonnablement attribuées à la méthode, à l'effet d'homogénéisation, de dilution choisie et aux facteurs humains (Harmel et al., 2016).

Il peut ainsi y avoir également une source d'incertitude supplémentaire liée à la distribution non homogène des micro-organismes dans l'échantillon. Il faut noter que les bactéries peuvent être associées à des particules, qui créent des amas de bactéries et sédimentent dans les flacons d'échantillonnage (Fries et al., 2006). Certaines espèces ont également tendance à s'adsorber sur les parois en raison de la nature de leur paroi cellulaire. Une moyenne de 38 ± 4%

de BIF est associée aux particules dans l'estuaire de la rivière Neuse (Fries et al., 2006). Ainsi, lors de la dilution en série des échantillons, l'incertitude augmente (Dufour, 2021). De même, si la concentration est faible, l'incertitude peut augmenter. En effet, les fluctuations aléatoires dans l'échantillon peuvent être plus prononcées, rendant les estimations moins précises (Harmel et al., 2016). Tout cela fait qu'à chaque étape, il y a une accumulation de l'incertitude. Ces données collectées présentent donc une incertitude de départ, difficilement quantifiable, mais qui ajoute une variabilité dans la mesure de la concentration en indicateurs bactériens qu'il faut avoir en tête, ce qui peut affecter la répétabilité ou la précision des mesures (Harmel et al., 2016).

En ce qui concerne la méthode d'analyse, Hamilton et al. (2005) a identifié une amélioration des mesures de concentration d'*E. coli* observée avec les milieux de culture spécifiques aux enzymes (quantitray Colilert ou en microplaque MUG/EC), en comparaison avec l'utilisation des milieux de culture conventionnels. L'incertitude associée aux estimations NPP découle du fait que cette méthode repose sur des estimations statistiques plutôt que sur une mesure directe, ce qui introduit une variabilité dans les résultats (McBride et al., 2003). Les normes sont conçues pour réduire l'incertitude dans les méthodes d'analyse microbiologique, en assurant des procédures uniformes et fiables. Pour la détection d'*E. coli* et des entérocoques intestinaux dans les eaux de surface, les normes ISO 9308-3, ISO 9308-2 et ISO 7899-1 visent à garantir des résultats précis et reproductibles. Par ailleurs, pour les méthodes de quantification par PCR (polymerase chain reaction), le guide MIQE (Minimum Information for Publication of Quantitative Real-Time PCR Experiments) offre des recommandations sur les bonnes pratiques en PCR quantitative, afin d'assurer la rigueur et la reproductibilité des résultats obtenus par cette méthode (Dooms et al., 2014).

9.5. Estimation de l'incertitude pour les mesures ponctuelles

L'incertitude représente la dispersion des données quantitatives qui peut être estimée par différents paramètres statistiques. Elle représente un doute sur les résultats de la mesure (Harmel et al., 2016). Afin d'estimer le pourcentage d'incertitude au niveau de la mesure de la concentration en BIF, une mesure du pourcentage d'erreur relative d'échantillonnage est réalisée (Esbensen and Wagner, 2014; Harmel et al., 2016).

Pour comprendre la part relative des sources de variabilité, il a été suggéré de calculer l'incertitude totale en cumulant l'incertitude liée à chaque source (McCarthy et al., 2008;

Topping, 2012). La méthode de calcul de l'incertitude globale consiste à cumuler l'ensemble des incertitudes. L'incertitude de la concentration en *E. coli* dans l'échantillon (xi/xi), peut être exprimée comme suit (McCarthy et al., 2008) :

$$\left(\frac{\Delta xi}{xi}\right)^2_{total} = \left(\frac{\Delta xi}{xi}\right)^2_{échantillonnage} + \left(\frac{\Delta xi}{xi}\right)^2_{stockage} + \left(\frac{\Delta xi}{xi}\right)^2_{analyse} \quad (1.1)$$

Pour calculer l'incertitude à chaque étape et en fonction du type de données disponibles, les formules présentées dans le tableau 1.4 peuvent être utilisées.

TABLE 1.4 – Méthodes utilisées pour l'estimation de l'incertitude à partir des données disponibles (Harmel et al., 2016).

Uncertainty estimation method	Comments	Equation
1. Used uncertainty estimate as directly reported	– Rarely were these estimates available	–
2. Used methods of Taylor and Kuyatt (1994) and McCarthy et al. (2008) (Eq. (1)) or Harmel and Smith (2007) to estimate uncertainty	– For random uncertainty – Used if necessary summary statistics (e.g., mean, standard deviation, number of samples collected) were reported	$\pm\%\ unc. = \frac{\Delta x_i}{x_i} \approx \frac{2u(x_i)}{x_i}$ (1)[a]
3.1. Used Eq. (1) to estimate uncertainty	– For random uncertainty – Used for raw data sets, after determination of mean and standard deviation	Eq. (1)
3.2. Used Eq. (2) to estimate uncertainty	– For systematic uncertainty – Used for paired values (a_i, b_i) with a_i assumed to be the "true" value	$\pm\%\ unc. = \frac{(a_i - b_i)}{a_i}$ (2)[b]
3.3. Used Eq. (3) to estimate uncertainty	– For random uncertainty – Used for paired values with no "true" value	$\pm\%\ unc. = \frac{\lvert a_i - b_i \rvert}{avg(a_i, b_i)}$ (3)
3.4. Used best professional judgment to assign an uncertainty estimate based on data for another constituent such as total suspended solids	Used when no data relevant to *E. coli* were available. – Used only as a contingency for knowledge gaps present for critical elements of *E. coli* monitoring; accounting for these uncertainty sources was necessary for a comprehensive uncertainty analysis	–

[a] Where x_i is the sample mean of a given data series, $\Delta x_i/x_i$ is the relative uncertainty of a quantity x_i, and $u(x_i)$ is the standard deviation of the mean.
[b] Where a_i and b_i are paired values.

La mesure des indicateurs de contamination pose également des problèmes pour la gestion, car un seul échantillon ne peut capturer la dynamique spatiale et temporelle des concentrations bactériennes sur tout le site. De plus, le moment de l'échantillonnage peut à lui seul influencer l'incertitude dans l'estimation des concentrations bactériennes, rendant ainsi les données peu fiables pour une gestion en temps réel et pour la modélisation en vue de prédire les concentrations en BIF (Wyer et al., 2018).

10. Prédiction de la qualité microbiologique

La concentration en bactéries fécales dépend beaucoup des conditions météorologiques, car en cas de fortes pluies, l'eau de surface est polluée par un apport de contaminant provenant des ruissellements, générant ainsi une grande variabilité dans les données. L'échantillonnage ponctuel réglementaire ne permet pas un suivi fin des variations de concentrations en BIF, car les méthodes sont coûteuses, chronophages et laborieuses (Chen et al., 2020). Toutefois, il est

important de surveiller et de prévoir la qualité d'eau de manière précise au moment opportun, pour gérer les baignades au quotidien et en temps réel. L'OMS recommande l'utilisation de la modélisation pour aider à la gestion des baignades (OMS, 2018).

Disposer de données fiables même en l'absence de mesures directes par la modélisation est particulièrement pertinent pour des alternatives aux plans d'échantillonnage intensifs, qui peuvent être coûteux pour les petites collectivités. En parallèle, ces connaissances peuvent servir à informer le public en temps réel et soutenir la mise en place de systèmes d'alerte sanitaire, comme requis par la directive européenne sur la qualité des eaux de baignade (van der Meulen et al., 2024). Les gestionnaires utilisent le nowcasting pour décider des avis de qualité de l'eau et des options de traitement. Le nowcasting est une technique de prévision à très court terme (quelques heures). L'objectif est de fournir des estimations des conditions actuelles ou proches en temps réel, à l'inverse des prévisions météorologiques classiques, le forecasting porte sur des conditions futures à l'échelle de quelques jours. Appliqué aux domaines environnementaux, le nowcasting utilise des modèles ou des techniques mathématiques pour évaluer rapidement les menaces sur la qualité de l'eau, par exemple en détectant des concentrations de contaminants ou de BIF comme *E. coli* dans un délai quasi-instantané (Francy et al., 2020).

A cet effet, plusieurs méthodes de modélisation ont été créées et mises en œuvre pour surveiller et prédire la qualité de l'eau (Chen et al., 2020). Sur la base des données collectées sur la qualité de l'eau, un modèle de prévision peut établir une relation de correspondance entre les données de surveillance multiples et les changements des paramètres de qualité de l'eau (Liu et al., 2019). Ces dernières années, l'établissement de modèles fiables de prévision de la qualité de l'eau est devenu l'un des points chauds de la recherche dans le domaine de la science environnementale de l'eau (Liu et al., 2019).

Dans la littérature scientifique, il existe une variété de modèles (statistique, déterministe, apprentissage automatique) qui sont proposés pour prédire la qualité de l'eau (Mälzer et al., 2016; Visser et al., 2022; Chen et al., 2020). Pour trouver l'outil de modélisation idéal, il faut examiner différents modèles prédictifs. Les modèles statistiques et l'apprentissage automatique sont deux approches utilisées pour analyser et interpréter des données dans le but de faire des prédictions. Les modèles statistiques sont souvent plus faciles à interpréter car ils sont basés sur des principes statistiques classiques et ont généralement des paramètres explicites avec des interprétations directes. Par contre, les modèles d'apprentissage automatique, peuvent être plus difficiles à interpréter en raison de leur nature non linéaire et de la présence de

nombreux paramètres qui font d'eux des "boites noires" (Mälzer et al., 2016). Visser et al. (2022) ont classé onze modèles en fonction de leurs performances prédictives et de leur niveau de transparence. L'étude a montré qu'il existe un compromis entre les performances prédictives et les niveaux de transparence des modèles. Les modèles d'apprentissage automatique ont les meilleures performances en matière de prédiction mais présentent des structures de modèle non transparentes comme les approches Random Forest et Boosting. Des régressions linéaires simples et multiples, un modèle hydrodynamique et un modèle de réseau de neurones ont été utilisés pour prédire la concentration en *E. coli* afin d'identifier les pollutions de courte durée dans la rivière Ruhr en Allemagne (Mälzer et al., 2016). Toutefois, la performance de ces différents modèles variait selon le jeu de données et le contexte (Mälzer et al., 2016; Chen et al., 2020). Ainsi, le long de la rivière Ruhr (Allemagne), la performance des différents modèles variait d'un site à l'autre (Mälzer et al., 2016). En comparaison à des modèles statistiques et déterministes, les performances des modèles de machine learning ont démontré leur capacité à prédire de manière fiable la concentration en *E. coli* (Mälzer et al., 2016). Parmi les différents outils de modélisation, les outils d'apprentissage automatique se sont avérés capables de prédire la qualité des eaux de surface des rivières avec une grande précision dans différentes situations (Ghahramani, 2015; Mälzer et al., 2016; Qiu et al., 2017).

10.1. Modèles statistiques

Les modèles statistiques, notamment les régressions linéaires simples et multiples, sont couramment utilisés pour prédire la qualité des eaux de baignade en se basant sur des corrélations avec des paramètres physico-chimiques et météorologiques tels que le pH, la turbidité, la conductivité, l'oxygène dissous et les nutriments (ammonium, nitrate et nitrite) (Mälzer et al., 2016). Ces approches sont souvent employées pour leur transparence et leur facilité d'interprétation, même si, dans certains cas, les régressions multiples peuvent présenter une précision de prédiction limitée, car elles ne peuvent pas prendre en compte des interactions complexes entre plusieurs facteurs et tendent à se limiter aux relations linéaires ou linéarisables (Nevers and Whitman, 2005).

Différents modèles statistiques ont été appliqués en corrélant les concentrations bactériennes avec divers paramètres de qualité de l'eau afin de prévoir les niveaux de contamination dans les zones de baignade à la suite du calcul de corrélations linéaires entre les bactéries et

plusieurs paramètres physico-chimiques (Nevers and Whitman, 2005; Mälzer et al., 2016). Ces modèles sont souvent utilisés par les collectivités pour développer des systèmes d'alerte précoce utilisant les paramètres physico-chimique et hydrométéorologiques les plus étroitement liés aux occurrences bactériennes (Mälzer et al., 2016; Seis et al., 2018).

10.2. Modèles déterministes

Les modèles hydrodynamiques sont essentiels pour simuler les processus dynamiques de l'eau, en intégrant divers paramètres comme le débit et les interactions biologiques et chimiques. Ils reposent sur la résolution numérique des équations de conservation de la masse et de la quantité de mouvement, notamment les équations de Navier-Stokes moyennées de Reynolds, qui décrivent le mouvement des fluides (Liu, 2018). En fonction du besoin, la modélisation peut être réalisée en 1D, 2D ou 3D. Le modèle 1D convient aux cours d'eau linéaires, le 2D est adapté aux estuaires ou rivières larges, et le 3D est utilisé pour des environnements complexes où les effets verticaux sont critiques. Ces modèles permettent aussi de simuler le transport des BIF, considérées comme traceurs passifs ou en intégrant des paramètres de mortalité, sédimentation et prédation (Liu, 2018). En région parisienne (France), plusieurs outils de modélisation déterministes sont en cours de développement ou validés sur la Seine et la Marne. Ainsi, le modèle PROSE, appliqué à la Seine en 2D, inclut des modules hydrauliques, de transport et biogéochimiques pour suivre la dynamique des BIF, tout en tenant compte des paramètres de dégradation pour mieux représenter leur comportement dans les écosystèmes fluviaux (Hasanyar, 2023). Un modèle hydrodynamique Telemac 3D pour la prédiction à court terme a été développé au bassin de La Villette à Paris, s'appuyant sur la mesure des BIF en amont du site de baignade et la simulation de leur transport. Le modèle permet l'estimation du temps de transfert des bactéries ainsi que de leur distribution spatiale (Guillot-Le Goff et al., 2023). Un modèle Telemac 2D a été développé dans le cadre de la démarche d'ouverture de sites de baignade dans la partie aval de la Marne et pour la Seine, afin de procurer un outil de gestion pour les collectivités (Van et al., 2022)

10.3. Modèles basés sur l'apprentissage

10.3.1. Apprentissage automatique

De nos jours, l'apprentissage automatique, aussi appelé machine learning (ML), est de plus en plus utilisé, dans une grande diversité d'applications. C'est une discipline donnant aux

algorithmes la capacité d'apprendre sans qu'ils ne soient explicitement programmés (Géron, 2019). Un système d'apprentissage automatique peut s'adapter à de nouvelles données et, bien sûr, à de gros volumes de données. Le machine learning est la science moderne qui va permettre de découvrir des patterns (motifs et structures) dans des données historiques et d'effectuer des prédictions en se basant sur des statistiques, des reconnaissances de pattern ou sur les analyses prédictives (Zhu et al., 2022).

Les outils de modélisation prédictive issus de l'apprentissage automatique, ont gagné en popularité dans de nombreux domaines de recherche, y compris celui de la modélisation hydrologique. Cette popularité peut s'expliquer par leurs qualités de prédiction relativement performantes (Visser et al., 2022). Des modèles prédictifs sont recommandés pour parvenir à une gestion active du site de baignade (OMS, 2018; Wuijts et al., 2022a). Plusieurs études antérieures ont utilisé des modèles d'apprentissage automatique pour prédire la qualité des eaux de surface à l'aide des paramètres physico-chimiques et hydrométéorologiques comme variables prédictives (Di et al., 2019; Avila et al., 2018; Mälzer et al., 2016; Cyterski et al., 2022). La modélisation prédictive des concentrations en BIF, comme *E. coli*, peut constituer un complément à la surveillance réglementaire de la qualité microbiologique des eaux de surface (Nevers and Whitman, 2005).

Les principales approches d'apprentissage sont l'apprentissage non supervisé et l'apprentissage supervisé. Quand il s'agit d'algorithmes non supervisés, nous parlons souvent d'algorithmes de regroupement, car les données à disposition ne seront pas étiquetées ou labellisées (un label étant une catégorie ou classe d'appartenance), les catégories (cluster ou classes) ne sont donc pas connues. Dans ce cas, l'algorithme va déterminer par lui-même des points similaires entre les caractéristiques pour pouvoir créer des groupes homogènes (Raul, 2017). Cela pourrait être par exemple un regroupement en trois catégories "bonne", "suffisante" ou "mauvaise" afin de caractériser la qualité de l'eau. Ces modèles comprennent l'analyse en composante principale (ACP), l'algorithme des k-moyennes (classification k-means), la classification hiérarchique et la classification probabiliste. Les algorithmes supervisés, quant à eux, utilisent des données labellisées car la catégorie (cluster ou classe) est déjà connue. De ce fait, cela permet de travailler alors avec des classes données et des exemples connus pour comprendre les patterns cachés. Les algorithmes apprendront soit de la classification soit de la régression en tant qu'algorithmes supervisés (Raul, 2017). Les modèles supervisés comprennent le modèle linéaire, les arbres de décision ou DT, les supports de vecteurs (machines à vecteur de support ou SVM), les réseaux

de neurones et les méthodes ensemblistes.

Par exemple, les arbres de régression ont été utilisés pour prédire en temps réel la concentration en *E. coli* et donc la qualité microbiologique des sites de baignade dans le Sud de la Nouvelle-Zélande (Avila et al., 2018). Cette prédiction est basée sur les valeurs passées de précipitation, le débit et la concentration en *E. coli* (Avila et al., 2018). D'autres modèles classiques d'apprentissage automatique, tels que la méthode des k-voisins les plus proches, les réseaux de neurones ou la machine à vecteur de support, ont également été utilisés pour la gestion et la prédiction de la qualité de l'eau (Chen et al., 2020; Qiu et al., 2017).

Il existe également l'apprentissage semi-supervisé qui combine ces deux approches, des algorithmes d'apprentissage non supervisé sont alors utilisés pour générer automatiquement des étiquettes, qui peuvent être introduites dans les algorithmes d'apprentissage supervisé. Enfin, l'apprentissage par renforcement permet un apprentissage d'une succession de tâches, combiné avec un feedback continu sous forme de récompense pour affiner la stratégie employée et ainsi améliorer la performance du modèle. Ces deux dernières approches sont encore peu utilisées pour la prédiction de la qualité de l'eau.

Avant d'appliquer l'apprentissage automatique, il est essentiel de procéder à l'acquisition et au nettoyage des données et éventuellement leur labellisation (Zhu et al., 2022). La première étape clé dans les applications d'apprentissage automatique est donc d'assurer un nettoyage de qualité des données. Dans un premier temps, il est important d'explorer la qualité des données en vérifiant leur exactitude, leur complétude, leur conformité, leur cohérence, leur fiabilité et leur pertinence. Ensuite, les données sont nettoyées. Il existe deux approches : soit en retirant les observations avec des valeurs manquantes, soit en remplaçant ces données manquantes par des moyennes, des médianes ou en utilisant des approches probabilistes. Le nettoyage inclut également la suppression des valeurs extrêmes (outliers), qui peuvent résulter d'erreurs ou d'événements exceptionnels (Gong et al., 2023). Enfin, les données sont prétraitées en les formatant, en réduisant leur taille par agrégation, en les normalisant ou en créant de nouvelles variables dérivées des données brutes (discrétisation, indices, rapports) (Zhu et al., 2022; Gong et al., 2023).

Une fois les données préparées, le jeu de données est divisé en deux parties : une pour l'entraînement du modèle (apprentissage) et l'autre pour le test et la validation. Les hyperparamètres du modèle sont ajustés et affinés pendant la phase d'entraînement et la performance du modèle est ensuite évaluée à l'aide de paramètres statistiques, afin de mesurer son efficacité et

sa capacité à généraliser sur de nouvelles données (Jovanovic et al., 2019). Pour les applications d'apprentissage automatique, la précision de la prédiction est généralement liée à deux aspects, à savoir la qualité de l'ensemble de données d'apprentissage et la sélection du modèle. Parmi ces processus, le choix de l'algorithme est crucial (Zhu et al., 2022). Après entraînement et validation du modèle, il faut sélectionner l'algorithme approprié (Zhu et al., 2022).

Une classe de modèles d'apprentissage automatique, les méthodes ensemblistes, améliore la stabilité et la précision des algorithmes d'apprentissage et représente donc un intérêt certain pour la gestion des eaux de baignade. Ainsi, les forêts d'arbres de décision (Random Forest ou RF) et le bootstrap aggregating (aussi appelé bagging) ont été utilisés pour le suivi de la qualité de l'eau de la rivière Talar au nord de l'Iran (Bui et al., 2020). Ce méta-algorithme de boosting utilise de manière répétée des sous-modèles développés séquentiellement sur un échantillon d'entraînement, les poids de chaque observation étant ajustés au fur et à mesure de leur développement. Ainsi, les régresseurs suivants se concentrent davantage sur les observations mal ajustées ou mal prédites (Hastie, 2009). Les modèles ensemblistes ont souvent des performances supérieures aux autres algorithmes pour prédire les concentrations en BIF dans les rivières (Weller et al., 2020). Ainsi, les modèles RF sont capables de mieux refléter la complexité et l'hétérogénéité des systèmes d'eau douce, car ils peuvent mieux prendre en charge les paramètres colinéaires, les données manquantes et les interactions entre paramètres (Weller et al., 2020).

La précision de prédiction des modèles d'apprentissage automatique dépend également des paramètres utilisés pour construire les modèles (Zhu et al., 2022). De nombreuses variables peuvent être utilisées pour prédire les concentrations d'indicateurs fécaux bactériens (Cyterski et al., 2022). Les variables redondantes réduiront la précision du modèle tout en augmentant sa complexité (Zhu et al., 2022). Cyterski et al. (2022) et Nevers and Whitman (2005) ont identifié que la pluviométrie était le paramètre explicatif le plus influent pour prédire les concentrations en indicateurs fécaux bactériens. En effet, les analyses indiquaient la présence d'eaux usées dans la rivière après de fortes pluies. L'oxygène dissous reflète l'état de l'écosystème aquatique et sa capacité à soutenir les organismes aquatiques (Zhu et al., 2022). Cyterski et al. (2022) ont aussi identifié la turbidité comme un paramètre important ayant une influence sur les modèles prédictifs des indicateurs microbiens, car sa fluctuation peut être un témoin d'apport de rejets urbains, de ruissellement et de remise en suspension des sédiments qui mobilisent des réservoirs et sources de BIF. D'autres paramètres de qualité de l'eau, comme la température, le pH ou la concentration

en nutriments (phosphore et azote) peuvent être utilisés comme prédicteurs dans les modèles de la prédiction de la qualité de l'eau de surface, car ils sont présents également dans les rejets d'eaux usées et les rejets de temps de pluie (Zhu et al., 2022). Cependant, en fonction des données à disposition, certains paramètres vont exercer une plus grande influence que d'autres sur les modèles prédictifs (Cyterski et al., 2022). Une amélioration de la performance des modèles avec plus de paramètres lors de l'entraînement a également été constatée (Chen et al., 2020). Toutefois, la sélection préalable de paramètres pertinents constitue également une stratégie d'amélioration de la performance des modèles. En effet, les modèles peuvent avoir une faible performance lorsque les données d'entrainement sont en quantité insuffisante ou de mauvaise qualité. Une des stratégies pour y pallier est de réduire le besoin en données du modèle en sélectionnant les prédicteurs les plus pertinents (Nafsin and Li, 2023; Wu et al., 2024). Une autre stratégie est d'augmenter la quantité et la diversité des données par une approche d'apprentissage par transfert à partir de bases de données provenant d'autres rivières de caractéristiques similaires (Wu et al., 2024).

10.3.2. Apprentissage par transfert

La plupart des données environnementales proviennent d'une minorité de sites bien surveillés (Willard et al., 2021). Dans des systèmes complexes et dynamiques comme une rivière, un corpus de données encore relativement modeste ne permet pas de rendre compte de toute la variabilité possible des paramètres mesurés. Les modèles de machine learning ne parviennent donc pas toujours à effectuer des prédictions fiables dans toutes les situations (Pachepsky et al., 2018; Chen et al., 2020). Il existe des solutions pour contourner ce problème à l'aide d'outils d'apprentissage automatique comme l'apprentissage par transfert. Le transfert des connaissances des sites surveillés vers les sites non surveillés constitue un défi, et les méthodes avancées d'apprentissage par transfert sont encore peu utilisées pour prédire la qualité microbiologique de l'eau (Willard et al., 2021; Wu et al., 2024).

L'apprentissage par transfert (Transfer Learning, TL) est un sous-ensemble de l'apprentissage automatique. Comme son nom l'indique, il regroupe l'ensemble des méthodes permettant de transférer des connaissances acquises à partir de la résolution d'un problème, pour en traiter un autre. Il est basé sur la création de modèles d'apprentissage sur des données et ces modèles peuvent être réutilisés sur des jeux de données plus petits (Willard et al., 2021).

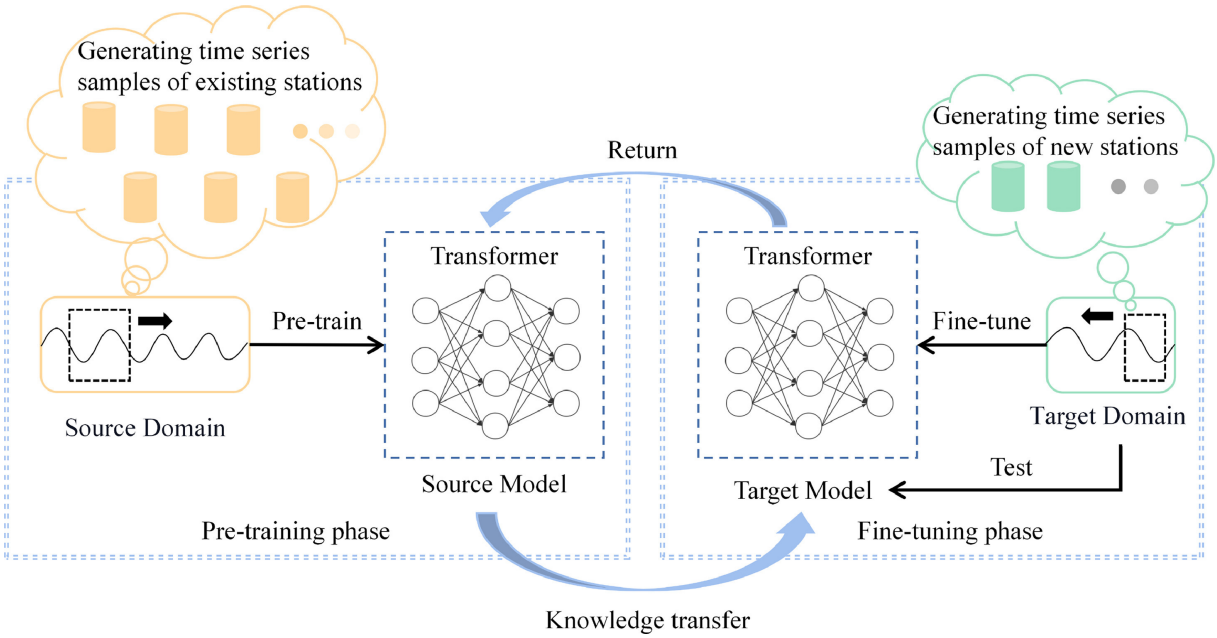

FIGURE 1.2 – Le processus de fonctionnement de l'apprentissage par transfert pour un modèle donné (Peng et al., 2022).

Le modèle proposé par Peng et al. (2022) est composé de deux parties principales (Figure 1.2) :

- Une méthode de prédiction est réalisée à partir d'un jeu de données composé de plusieurs stations de mesures afin d'obtenir le modèle source (noté transformer).

- Un transfert des connaissances (modèle de prédiction) est réalisé vers le nouveau jeu de données cible.

De plus, l'apprentissage par transfert est avantageux dans le sens où la création d'un modèle utilise beaucoup de ressources (Uddin et al., 2019; Shahid Iqbal et al., 2018). En utilisant des modèles pré-entraînés selon le contexte, nous pouvons pallier ce manque de données et réduire les ressources utilisées (Dipanjan, 2018). Les informations obtenues sur la dynamique d'un lac pourraient par exemple être transférées à d'autres lacs similaires (Willard et al., 2021). Dans le cas des écosystèmes ayant des caractéristiques physiques et une dynamique de la qualité de l'eau similaires ou même très proches, cela pourrait permettre le transfert stratégique de modèles spécifiques à un site bien surveillé afin de faire des prédictions dans des systèmes moins surveillés (Willard et al., 2021; Naloufi et al., 2021).

Des modèles de prédiction de la qualité de la rivière Haihe (Chine) ont été testés (Peng et al., 2022). L'étude a permis d'identifier de meilleurs résultats en utilisant l'approche d'apprentissage par transfert à partir des données de 10 stations sur la rivière Huaihe (Chine) pour le pH et l'oxygène dissous. Par contre, pour l'azote ammoniacal, les modèles pré-entrainés par

apprentissage par transfert étaient moins performants que ceux entrainés sans apprentissage par transfert. Ce phénomène s'appelle l'apprentissage négatif. La perte de connaissance suite à un apprentissage par transfert peut être due à une similarité limitée entre les sites des deux bases de données utilisées. Le transfert négatif est encore peu considéré et de ce fait la connaissance sur ce type de résultat est encore manquante (Wu et al., 2024).

10.3.3. Apprentissage fédéré

L'apprentissage fédéré (Federate Learning, FL) est une autre application, analogue à l'apprentissage par transfert. Il s'agit d'entraîner plusieurs modèles de différentes entités localement et de créer un modèle global basé sur les mises à jour des modèles locaux (Lo et al., 2021). L'apprentissage fédéré permet à plusieurs appareils de former en collaboration un modèle partagé tout en conservant les données locales à chaque appareil (Vellingiri et al., 2023). Ainsi, l'apprentissage fédéré va utiliser les paramètres des différents modèles locaux pour créer un modèle centralisé qu'il distribuera à chacune des entités, sans diffuser les données sources (Lo et al., 2021). Avec cette approche, une précision de prédiction de 87% pour l'évaluation de la qualité d'une rivière au sud de l'Inde a été obtenue (Vellingiri et al., 2023). Dahane et al. (2024) ont également utilisé l'apprentissage fédéré pour rendre les données plus privées dans le contexte de l'analyse de la qualité de l'eau pour la baignade en rivière.

L'approche d'apprentissage fédéré peut également être utilisée dans d'autres systèmes innovants de gestion de la qualité de l'eau. Park et al. (2021) présentent un réseau sophistiqué d'apprentissage fédéré intégrant des capteurs. Ce réseau exploite des données de qualité de l'eau en temps réel, géographiquement distribuées, afin d'améliorer la précision des prédictions et de proposer une approche proactive.

10.3.4. Réseau de neurones

La dynamique des BIF dans les habitats aquatiques est causée par un certain nombre de facteurs environnementaux qui peuvent avoir une influence sur la distribution et le devenir des microorganismes (Devane et al., 2018). De ce fait, il serait intéressant de disposer d'un modèle prenant en compte les conditions des jours précédents. Nous pourrions ainsi utiliser des réseaux de neurones qui prennent en compte les séries temporelles (Liu et al., 2019). Les réseaux de neurones utilisent une cascade de couches multiples d'unités de traitement non linéaires pour l'extraction et la transformation des caractéristiques, ce qui fait qu'ils sont adaptés

à l'analyse et à l'extraction de connaissances utiles à partir de grandes quantités de données et de données collectées à partir de différentes sources (Shinde and Shah, 2018). Parmi les modèles de réseaux de neurones, il y a les Long Short Term Memory (LSTM) présentés par Hochreiter and Schmidhuber (1997). Ce sont des modèles dits récurrents qui utilisent des données qui doivent être sous la forme de séries temporelles. Les LSTM ont la capacité d'apprendre les dépendances à long terme. Avec cette approche, l'évolution de la pollution pourrait être prise en compte lors de la modélisation (Shinde and Shah, 2018). Les chercheurs ont vérifié que les LSTM peuvent traiter des séries temporelles de données sur la qualité de l'eau qui sont fluctuantes et non saisonnières (Zhu et al., 2022).

Les résultats de l'étude de Liu et al. (2019) révèlent le potentiel de l'application des LSTM et de l'apprentissage profond pour prédire la qualité de l'eau. Les valeurs prédites par leur modèle et les valeurs réelles étaient en accord et révélaient avec précision la tendance future de la qualité de l'eau (Liu et al., 2019). Les modèles de réseau de neurones vont se baser sur les données pour créer les modèles avec une extraction des caractéristiques de ces données et apprendre à partir de ces données (Shinde and Shah, 2018). Par exemple, la précision d'un modèle LSTM pour la prédiction des concentrations en oxygène dissous mesurées par une station de surveillance automatique était meilleure que celle du SVR (support vector regression) qui, à long terme, devenait inexact avec un ensemble de données d'apprentissage relativement petit (Liu et al., 2019). Au niveau de l'étude de Mälzer et al. (2016), les réseaux de neurones ont donné de bons résultats pour prédire les concentrations en *E. coli* pour la plupart des sites le long de la rivière Ruhr (Allemagne) à l'exception d'une station où la régression linéaire et la régression multiple donnaient de meilleurs résultats. Ce résultat montre que les réseaux de neurones sont suffisamment versatiles pour s'adapter à des sites aux caractéristiques variées.

11. Optimisation de la collecte de données

Cependant, en raison de la petite taille de la plupart des jeux de données disponibles pour le suivi de la qualité des eaux de surface, la performance des modèles peut être faible (Chen et al., 2020; Ghahramani, 2015). En effet, l'entraînement et le test des modèles prédictifs des concentrations en indicateurs de contamination fécale nécessitent des données de haute précision qui sont difficiles et coûteuses à collecter (Jovanovic et al., 2019). Les données physico-chimiques et hydrométéorologiques sont souvent utilisées comme données d'entrée dans les modèles de prédiction de la concentration des BIF, car la mesure en temps réel de ces paramètres fournit

des données de haute qualité à faible coût (Nnane et al., 2011; Bui et al., 2020; Banda and Kumarasamy, 2020). Toutefois, il reste nécessaire d'entraîner et de valider le modèle avec des données de concentrations en BIF, données qui sont acquises au mieux une fois par jour et plus généralement une fois par semaine. La détermination de la taille d'échantillonnage minimale et de la stratégie d'échantillonnage appropriée requises pour la construction, l'entraînement et le test des modèles est donc cruciale (OMS, 2018).

Plusieurs stratégies visant à améliorer l'ensemble des données d'entrée des modèles d'apprentissage automatique existent, cependant leur utilité pour optimiser l'acquisition de données pour la prédiction de la qualité de l'eau doit encore être évaluée. Premièrement, les observations les plus pertinentes pendant le processus d'apprentissage des modèles pourraient être identifiées afin de maximiser le gain d'information ou bien par l'augmentation des données (Qian et al., 2020). Ce processus peut être effectué de différentes manières, notamment pour les mesures avec des données manquantes, en utilisant des moyennes ou des médianes, ou en utilisant une combinaison de méthodes d'apprentissage automatique et de complétion matricielle pour compléter les données manquantes (Zhu et al., 2022). Deuxièmement, en comblant les lacunes du jeu d'entraînement avec des données supplémentaires via des méthodes comme l'apprentissage actif (Bouneffouf, 2016) ou l'apprentissage par transfert (Wu et al., 2024). Troisièmement, en déployant un réseau de capteurs à faible coût sur le site de baignade, cela permettrait de fournir suffisamment de données d'entrée pour alimenter les modèles d'apprentissage automatique (KnowFLow, 2021).

Afin de fournir une résolution spatiale et temporelle suffisante et de réduire le coût du suivi, une surveillance avec des capteurs *in situ* combinés à l'apprentissage automatique pourrait aider à optimiser l'effort d'échantillonnage (Carvalho et al., 2019; Whelan et al., 2020). Par exemple, dans le cas de la surveillance d'un site de baignade, la mise en place d'un réseau de capteurs et d'appareils de mesure enzymatique en temps quasi réel fournirait suffisamment de données à la fois pour les BIF (mesure enzymatique) et pour les prédicteurs (capteurs physico-chimiques) pour modéliser la qualité microbiologique de l'eau et améliorerait également la quantité et la qualité des données (Pule et al., 2017).

11.1. Apprentissage actif

La collecte de données sur la qualité de l'eau peut être coûteuse en termes de temps, d'argent et de ressources. En effet, le coût important en main-d'œuvre humaine et en matériel pour collecter les données représente un frein (Jia et al., 2021). La faiblesse dans le jeu d'entraînement pourrait être déterminée afin d'identifier les classes de données minoritaires dans le jeu de données pour chaque prédicteur. À partir de cette information, trois stratégies sont possibles pour réduire le déséquilibre dans le jeu de données : i) soit d'utiliser un algorithme qui génère des données synthétiques pour les classes minoritaires, ii) soit d'utiliser un transfert à partir d'un jeu de données d'un site similaire pour amender les classes minoritaires, iii) soit d'optimiser l'échantillonnage sur le terrain pour renforcer ces classes minoritaires (Krishnan et al., 2024; Wu et al., 2024). Ainsi, les données supplémentaires nécessaires pour améliorer les performances du modèle pourraient être ciblées précisément (Bouneffouf, 2016). L'apprentissage actif est une méthode qui offre une certaine flexibilité pour identifier les instances qui doivent être ajoutées au jeu d'entraînement. L'objectif est d'améliorer l'efficacité de l'apprentissage en utilisant de manière sélective les données les plus informatives pour l'entraînement du modèle (Cacciarelli et al., 2022).

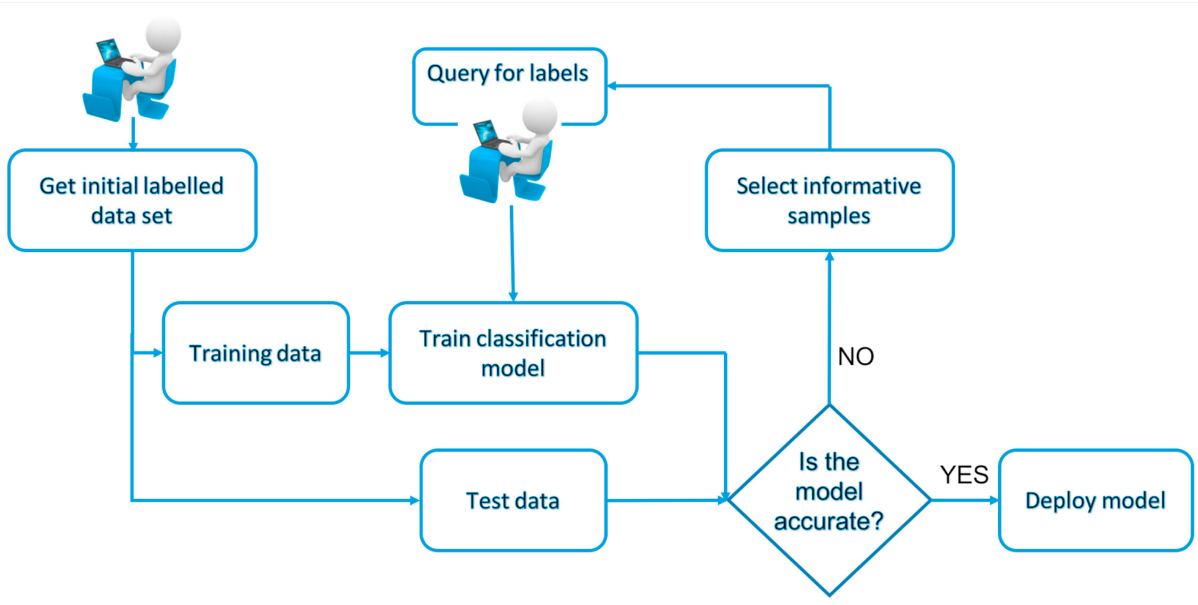

FIGURE 1.3 – Processus d'apprentissage actif (Russo et al., 2020).

Le but est de sélectionner activement les données de manière à apprendre une bonne hypothèse avec moins d'entraînement. La stratégie populaire consiste à utiliser l'échantillonnage d'incertitude pour identifier le point où la prédiction est incertaine dans le modèle (Bouneffouf,

2016). En effet, les efforts d'échantillonnage peuvent être considérablement réduits en utilisant l'approche d'échantillonnage d'incertitude (Russo et al., 2020). Une fois les données ajoutées, le jeu de données est actualisé et utilisé pour réentraîner le modèle (Figure 1.3). Cette opération peut être répétée, en fonction des propriétés des données, jusqu'à ce que le modèle atteigne des performances satisfaisantes (Russo et al., 2020).

En optimisant les ressources de collecte de données grâce à l'identification des informations les plus pertinentes, il serait possible d'améliorer la précision du modèle. Cela permettrait de réduire la quantité de données à collecter tout en augmentant la fiabilité des prédictions (Cacciarelli et al., 2022). Dans le domaine de la surveillance environnementale, les applications de détection d'anomalies permettront de développer les outils de gestion de la qualité (Russo et al., 2020). Chen et al. (2020) ont montré que de meilleures performances pour la prédiction de la qualité de l'eau pouvaient être obtenues après avoir augmenté le nombre de paramètres d'entrée pour la modélisation et les données d'entraînement pour un ensemble de modèles d'apprentissage automatique testés. Avec une augmentation des données d'entraînement de 1% à 10%, une amélioration des performances de prédiction des modèles a été constatée jusqu'à 22,76%. Cependant, cette amélioration était limitée lorsqu'une augmentation aléatoire des données jusqu'à 100% était utilisée (Chen et al., 2020). Cela illustre bien l'importance d'augmenter les données de manière efficace en sélectionnant les données à ajouter. Identifier la faiblesse de performance du modèle pourrait de ce fait être envisagé pour mettre en place un système d'alerte qui pointe sur les paramètres et sur les données nécessaires pour renforcer la prédiction du modèle (Qian et al., 2020; Jiang et al., 2020).

11.2. Collecte automatisée de données

Du contexte d'apprentissage actif résulte le problème de la collecte des données car les décisions d'échantillonnage doivent être prises immédiatement après l'observation du déséquilibre dans le jeu de données qui génère une performance médiocre du modèle. Une méthode de suivi de la qualité de l'eau consisterait à déployer des systèmes de mesures à haut débit de la concentration en *E. coli*. Les systèmes automatisés ou semi-automatisés permettant l'estimation des BIF en temps quasi réel sont relativement chers. Ils sont basés sur de la mesure enzymatique (systèmes automatisés ColiMinder (Vienna Water Monitoring, VWM) et BACTcontrol (Bionef) par exemple (Cazals et al., 2020)), sur de la culture bactérienne avec une détection entre 2 et 12 h

(par exemple le système ALERT (Fluidion) (Angelescu et al., 2019)) ou la mesure de la matière organique urbaine par spectre d'émission-excitation (sondes Proteus-Instruments, Fluocopee (SIAAP) ou BacTrack (NKE) (Bouleau et al., 2024)) qui peuvent être utilisées pour la gestion quotidienne de la qualité de l'eau en rivière. De plus, pour l'acquisition des données des prédicteurs physico-chimiques, des capteurs à haute résolution peuvent être utilisés et positionnés à des emplacements stratégiques en amont et au niveau du site de baignade. Cependant, un de ces dispositifs peut coûter plusieurs dizaines de milliers d'euros, ce qui rend leur acquisition difficile pour les petites administrations (Tatari et al., 2016). En outre, même des villes plus riches comme Paris limitent la quantité d'équipements à utiliser et leur couverture géographique en raison des coûts élevés de maintenance de ces dispositifs et des analyses de laboratoire associées. Ces problèmes peuvent réduire considérablement le nombre de données disponibles pour la modélisation. En raison des changements dynamiques complexes des systèmes fluviaux au fil du temps, le moyen le plus efficace de gérer les rivières est de surveiller la qualité de l'eau en temps réel ou de faire des prédictions basées sur ces données en temps réel (Zhu et al., 2022; Jia et al., 2021).

Une façon d'améliorer la quantité de données pour l'entraînement consiste à déployer sur le site de baignade et en amont, un grand nombre de capteurs à faible coût qui viennent compléter les capteurs physico-chimiques à haute résolution en offrant une couverture et un maillage spatial accrus. La faiblesse de la qualité de la donnée acquise par ces capteurs bas-coût peut être corrigée par la haute résolution spatiale et temporelle en tirant parti d'un déploiement d'un réseau dense de capteurs (Wang et al., 2019a). Chaque capteur individuel peut présenter une marge d'erreur légèrement supérieure à celle des équipements coûteux de haute précision, mais la multitude de capteurs permet de construire un réseau dense qui, en moyenne, est capable de fournir suffisamment d'informations pour les modèles d'apprentissage automatique (KnowFLow, 2021). Cette réduction de la qualité peut également être atténuée par l'association avec des dispositifs de haute précision, qui aideront à la calibration des capteurs à bas coût, afin de fournir des résultats précis (Abegaz et al., 2018). Différentes sondes pour différents paramètres peuvent être associées et utilisées pour former des systèmes de capteurs multiparamétriques (KnowFLow, 2021; Wang et al., 2019a). De nombreuses initiatives ont vu le jour pour mettre en place un réseau de capteurs à faible coût (Hong et al., 2021; Trevathan et al., 2021; Wong et al., 2021; de Camargo et al., 2023). Cheniti et al. (2023) ont testé leur système de surveillance de la qualité de l'eau basé sur des capteurs Arduino à court terme pendant 24 h dans l'eau du robinet.

D'autres études Gowri et al. (2023); Sekhar et al. (2023); Bogdan et al. (2023), ont également développé des systèmes de surveillance de la qualité de l'eau pour la baignade, avec une mesure des paramètres physico-chimiques.

Pour l'ensemble des capteurs se pose ensuite le problème de leur étalonnage, leur entretien et de la stratégie de déploiement. Il n'existe pas de lignes directrices sur les meilleures pratiques pour l'étalonnage et la validation des réseaux de capteurs à faible coût. De nombreuses initiatives ont vu le jour pour mettre en place un réseau de capteurs en temps réel à faible coût, mais peu d'entre elles se concentrent sur la fiabilité et la viabilité d'une utilisation à long terme (Hong et al., 2021). Ce manque de validation rend les résultats obtenus moins fiables (de Camargo et al., 2023). Un prototype de surveillance de la qualité de l'eau basé sur la technologie Arduino a été développé par Hong et al. (2021) composé de 4 sondes (pH, température, turbidité et solides dissous totaux (TDS)) dans un petit ruisseau artificiel de l'Université Brunei Darussalam pendant une courte durée de 20 jours. Il a été constaté que le système fonctionnait de manière fiable, mais il était dépendant de l'intervention humaine. Wong et al. (2021) ont développé un système de surveillance de la qualité de l'eau qui mesure la turbidité et les niveaux d'eau. Cependant, des erreurs causées par le dépôt de débris et l'encrassement biologique sur les capteurs ont été identifiées. Trevathan et al. (2021) ont souligné également la nécessité d'un entretien régulier et d'un mécanisme de nettoyage des capteurs. Comme le montrent ces études, pour mettre en œuvre les systèmes de surveillance de la qualité de l'eau, des tests sont nécessaires avant installation pour déterminer la validité des données afin de permettre un bon suivi de la qualité.

Enfin, un tel système pourrait bénéficier d'une observation en temps réel pour faciliter l'utilisation par le grand public et les administrations. Cela pose le défi de déterminer quand et où nous devrions déployer des instruments de mesure (par exemple, des capteurs *in situ*) pour collecter des données de manière efficace (Jia et al., 2021). Pour améliorer la collecte de données, il ne suffit pas d'avoir plus de capteurs mais une étude doit être menée concernant leur déploiement pour la construction d'un réseau adapté (Senouci and Mellouk, 2016). Actuellement, certains travaux, comme Ciaponi et al. (2018); Ramesh et al. (2017), ont abordé cette question dans le contexte de la surveillance de la qualité de l'eau et ont proposé différentes méthodologies pour le placement des capteurs. La décision de placement dépendra de la technologie de transmission et de la durée de vie de la batterie des appareils. Il est donc conseillé d'utiliser des réseaux étendus de faible puissance (LPWAN). Les technologies de communication comme LoRaWAN ou Sigfox, couramment utilisées dans le contexte de l'IdO, sont capables de diffuser les données

en temps réel avec une transmission à faible coût énergétique et une longue portée (Jiang et al., 2020). La perte de données due à la distance de communication limitée entre le capteur et la passerelle est également un problème crucial (Huan et al., 2020). En outre, les réseaux de capteurs à faible coût publiés sont généralement testés sur un type limité de qualité de l'eau, ce qui fait que la gamme de performances et les limites de détection des capteurs sont rarement vérifiées. Dans une étude récente, de Camargo et al. (2023) ont souligné que des tests supplémentaires sont nécessaires pour déterminer la validité des données et l'opérabilité des systèmes recommandés afin de mettre en œuvre sur le terrain une surveillance continue et fiable de la qualité de l'eau. Occasionnellement, l'utilisation de la technologie 5G peut également être intéressante si la quantité de données est importante (Rahimi et al., 2018). Bien que les passerelles LoRa soient économes en énergie pour des transmissions limitées, leur capacité de transmission des données est restreinte. À l'inverse, la 5G permet des échanges de données plus importants, mais elle consomme davantage d'énergie.

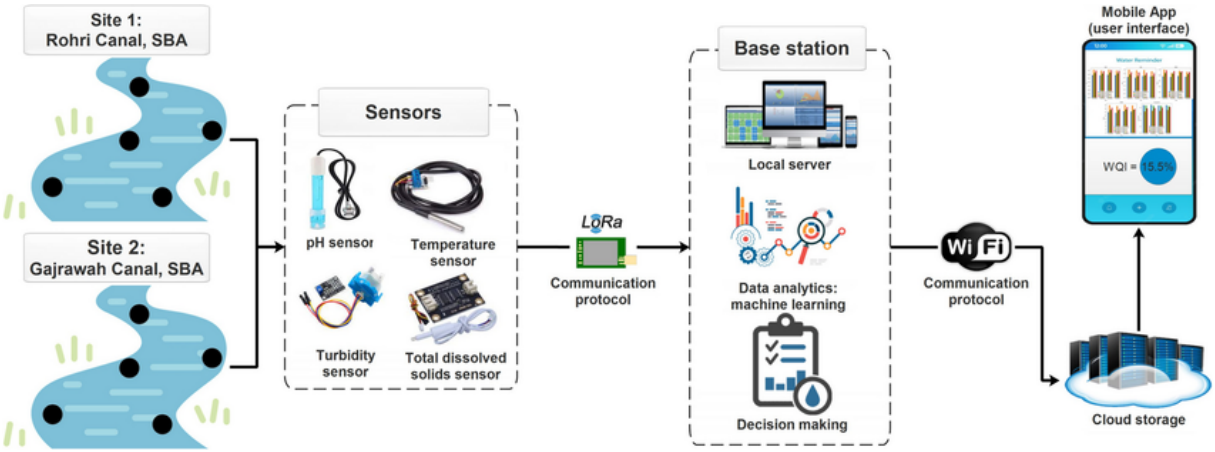

FIGURE 1.4 – Cadre basé sur l'IdO pour la surveillance de la qualité de l'eau (Rahu et al., 2024).

Les technologies d'IdO peuvent répondre aux besoins de surveillance de la qualité de l'eau en temps réel et à grande échelle (Figure 1.4). Les méthodes de mesure en temps réel utilisent les informations contextuelles spatiales et temporelles pour identifier des échantillons représentatifs dans un cadre de renforcement de la base de données (Jia et al., 2021). Du fait d'une détection en temps réel, ces outils peuvent également être utilisés pour suivre spatio-temporellement la migration des contaminants qui sont difficiles à détecter à l'aide de méthodes conventionnelles avec des mesures ponctuelles, en fonction de leurs limites de détection (Zhu et al., 2022). Un système de surveillance de la qualité de l'eau basé sur la technologie IdO mesurant la turbidité a été réalisé par Huan et al. (2020). Cependant, à mesure que la distance

de communication entre le capteur et la passerelle augmentait, le système subissait des pertes de paquets de données.

Couplés avec les modèles d'apprentissage automatique ces systèmes de mesure permettraient de construire un dispositif de suivi en temps réel de la qualité de l'eau et ainsi minimiser les risques sanitaires (Salam, 2020). En complément, utiliser ces données pour faire la prédiction en temps réel serait un atout considérable pour une gestion continue de la qualité de l'eau (Zhu et al., 2022). La mise en place de réseaux de capteurs pour suivre la qualité des eaux de baignade urbaines peut également aider à répondre à d'autres besoins en matière de qualité de l'eau, tels que les ressources en eau potable, la connaissance de l'état écologique, permettant une gestion intégrée des masses d'eau et de leurs usages multiples (Wuijts et al., 2022b). Les paramètres utilisés pour le suivi de la qualité microbiologique de l'eau, tels que la concentration en BIF ou en pathogènes, ne peuvent pas être mesurés directement par des capteurs *in situ*, car ces derniers ne sont pas optiquement actifs ou ne disposent pas de données hyperspectrales à haute résolution spatiale. Cependant, ces paramètres peuvent être estimés indirectement à l'aide d'autres données plus facilement mesurables (comme leur enzymes, la fluorescence des substances protéiniques, la température de l'eau, les nutriments, la turbidité, la conductivité, l'intensité des événements pluvieux et d'autres paramétres physico-chimiques) (Zhu et al., 2022; Cha et al., 2016; Passerat et al., 2011; Dueker et al., 2017; Bouleau et al., 2024). L'apprentissage automatique basé sur l'expérience permettait une optimisation sophistiquée des prédictions (Zhu et al., 2022). Le système d'alerte sur la donnée pourrait être alimenté par un réseau de capteurs permettant un suivi en temps réel des différents paramètres de l'eau (Luccio et al., 2020). Cela permettrait de fournir des recommandations au gestionnaire en indiquant de manière efficace quand l'échantillonnage manuel est nécessaire. En combinant la modélisation avec une collecte de données via des capteurs de surveillance en temps réel, cela offre une possibilité prometteuse d'une utilisation opérationnelle de capteurs pour la surveillance de la qualité de l'eau et la prise de décision (Sagan et al., 2020).

Chapitre 2 : Optimisation de la collecte de données pour la modélisation de la qualité microbiologique des eaux de surface

1. Introduction

Les épisodes de canicule et l'essor des activités récréatives aquatiques dans les grandes villes ont renforcé l'intérêt pour l'ouverture ou la réouverture de zones de baignade dans des rivières urbaines dans les mégapoles (e.g. Paris, Berlin, Londres...). Cependant, cette tendance expose les baigneurs à des risques sanitaires liés aux pathogènes présents dans les eaux de surface, notamment en raison des rejets d'eaux usées, du ruissellement pluvial et des excréments d'animaux (Soller et al., 2010; Passerat et al., 2011; Ahmed et al., 2019b). Afin de minimiser ces risques, il est essentiel de surveiller la qualité microbiologique de l'eau, principalement à travers la mesure d'indicateurs fécaux tels que *E. coli* et les entérocoques intestinaux (OMS, 2018; Commission européenne, 2006). En Europe, cette surveillance repose sur des analyses en laboratoire, mais leur fréquence et leur coût limitent l'efficacité de la gestion en temps réel des sites de baignade (Mälzer et al., 2016; Jovanovic et al., 2019).

Dans ce contexte, les techniques de modélisation, et en particulier les outils d'apprentissage automatique, apparaissent comme une solution prometteuse pour prédire la qualité de l'eau et anticiper les épisodes de pollution de courte durée (Ghahramani, 2015; Avila et al., 2018; Chen et al., 2020). Ces modèles permettent d'analyser de grandes quantités de données issues de paramètres physico-chimiques, météorologiques et microbiologiques. Toutefois, leur performance est fortement influencée par la taille et la précision des bases de données disponibles. Les jeux de données limités, souvent dus à des échantillonnages peu fréquents et/ou à des données coûteuses à acquérir, réduisent la capacité des modèles à fournir des prédictions fiables (Banda and Kumarasamy, 2020; Ghahramani, 2015). Dans ce chapitre, le premier et le deuxième article explorent les approches d'apprentissage automatique pour prédire la qualité microbiologique des eaux de surface en Seine et en Marne, en optimisant l'effort d'échantillonnage. Ces études

mettent en lumière différentes méthodes de machine learning, afin d'identifier celles qui offrent les meilleures performances dans la prédiction des contaminations microbiennes des eaux de surface, en mettant l'accent sur l'optimisation de l'effort d'échantillonnage et la précision des prévisions. Ce type d'approche est de plus en plus pertinent face à la complexité croissante des sources de pollution, tant naturelles qu'anthropiques. L'apprentissage actif permet d'identifier les données les plus informatives à ajouter au jeu d'entraînement, en se concentrant sur les zones d'incertitude du modèle (Bouneffouf, 2016). Cette approche permet d'apprendre plus efficacement avec moins de données, en maximisant la pertinence de chaque échantillon collecté.

Pour optimiser la collecte de données et améliorer les performances des modèles, plusieurs stratégies peuvent être mises en place. En parallèle de apprentissage actif, le déploiement de réseaux de capteurs à faible coût peut également renforcer la densité des données disponibles, même si chaque capteur présente une marge d'erreur plus élevée que les équipements de laboratoire (KnowFLow, 2021; Wang et al., 2019a). Le troisième article de ce chapitre explore la stabilité à long terme des systèmes basés sur l'IdO pour la surveillance continue des paramètres physico-chimiques de l'eau. Il met en avant les avantages d'un réseau de capteurs à faible coût pour assurer une couverture spatiale et temporelle accrue. Toutefois, il souligne également les défis liés à la fiabilité des données, notamment en raison de la dérive des capteurs et des besoins de maintenance régulière.

L'intégration de l'apprentissage actif et de réseaux de capteurs dans un cadre de modélisation permettra non seulement d'optimiser la collecte de données, mais aussi d'améliorer la précision des modèles de prédiction microbiologique. Ces outils constituent un atout précieux pour les gestionnaires des sites de baignade urbains, en leur offrant des moyens plus efficaces de surveiller et d'anticiper la qualité de l'eau, notamment dans des rivières comme la Seine et la Marne, où des projets ambitieux de réouverture de zones de baignade sont en cours (Noury et al., 2018).

En combinant ces différentes perspectives, ce chapitre vise à illustrer les avancées technologiques récentes dans la surveillance de la qualité de l'eau. Cela offre une vision globale des solutions pratiques et accessibles pour surmonter les défis complexes de la gestion des ressources hydriques dans les environnements urbains.

2. Evaluating the Performance of Machine Learning Approaches to Predict the Microbial Quality of Surface Waters and to Optimize the Sampling Effort

Published in : Water (2021)

https://doi.org/10.3390/w13182457

Manel Naloufi[1,2,*], Françoise S. Lucas[2], Sami Souihi[3], Pierre Servais[4], Aurélie Janne[5] and Thiago Wanderley Matos De Abreu [3,*]

[1] Direction de la Propreté et de l'Eau - Service Technique de l'Eau et de l'Assainissement, 27 rue du Commandeur 75014 Paris, France ; manel.naloufi@paris.fr,

[2] Leesu, Université Paris-Est Créteil, École des Ponts ParisTech, 61 avenue du Général de Gaulle, 94010 Créteil Cedex, France ; lucas@u-pec.fr

[3] Image, Signal and Intelligent Systems (LiSSi) Laboratory, University of Paris-Est Créteil Val de Marne, 122 rue Paul Armangot, 94400 Vitry sur Seine, France ; thiago.wanderley-matos-de-abreu@u-pec.fr ; sami.souihi@u-pec.fr

[4] Ecology of Aquatic Systems, Université Libre de Bruxelles, Brussels, Belgium ; Pierre.Servais@ulb.be

[5] Syndicat Marne Vive, Maison de la Nature, 77 quai de la Pie, 94100 Saint-Maur-des-Fossés, France ; aurelie.janne@marne-vive.com

Correspondence : manel.naloufi@paris.fr ; thiago.wanderley-matos-de-abreu@u-pec.fr

Abstract : Exposure to contaminated water during aquatic recreational activities can lead to gastrointestinal diseases. In order to decrease the exposure risk, the fecal indicator bacteria *Escherichia coli* is routinely monitored, which is time-consuming, labor-intensive and costly. To assist the stakeholders in the daily management of bathing sites, models have been developed to predict the microbiological quality. However model performances are highly dependant on the quality of the input data which are usually scarce. In our study, we proposed a conceptual framework for optimizing the selection of the most adapted model, and to enrich the training dataset.

This frameword was successfully applied to the prediction of *Escherichia coli* concentrations in the Marne River (Paris Area, France). We compared the performance of six machine-learning (ML) based models : K-nearest neighbors, Decision Tree, Support Vector Machines, Bagging, Random Forest and Adaptive boosting. Based on several statistical metrics, the Random Forest model presented the best accuracy compared to the other models. However, 53.2±3.5% of the predicted *E. coli* densities were inaccurately estimated according to the mean absolute percentage error (MAPE). Four parameters (temperature, conductivity, 24 h cumulative rainfall of the previous day the sampling and the river flow) were identified as key variables to be monitored for optimization of the ML model. The set of values to be optimized will feed an alert system for monitoring the microbiological quality of the water through combined strategy of *in situ* manual sampling and the deployment of a network of sensors. Based on these results we propose a guideline for ML model selection and sampling optimization.

Keywords : Water quality prediction ; Machine learning ; *Escherichia coli* concentration ; Optimized sampling ; River monitoring

2.1. Introduction

Worldwide the heat wave episodes have recently intensified the development of aquatic recreational activities in megapoles, increasing the interactions between citizens and freshwater in urban context (Jang, 2016). Indeed, many cities such as Paris, London or Berlin promote the opening of bathing areas and organize open water swimming competitions in their rivers. However, the development of these activities increases the risk of exposure of bathers to waterborne pathogens, which could result in gastrointestinal diseases, eye infections or skin irritations (Davies-Colley et al., 2018; Soller et al., 2010; Mallin et al., 2000).

In highly urbanized areas, the microbiological quality of surface waters is strongly degraded by different diffuse and point sources of contamination that may bring high pathogen flow into the rivers (Passerat et al., 2011; Dueker et al., 2017; Droppo et al., 2009; Garcia-Armisen and Servais, 2009). Fecal contaminations due to sewer discharges, animal feces and rain runoff are among the main factors impacting the quality of surface waters (Ahmed et al., 2019b). As the climate change is expected to modify precipitation patterns, with higher frequency of extreme events, these new conditions should negatively impact the water quality (Whitehead et al., 2009). Currently, the water quality is mainly assessed using a collection of water samples for biological

and chemical analysis in the laboratory and/or highly accurate sensors at fixed position. The regulatory monitoring of the bathing waters is based on the enumeration of culturable fecal indicator bacteria, *Escherichia coli* and intestinal enterococci (e.g. European Bathing directive 2006/7/EC). Such surveys are costly, time-consuming and labor-intensive, as a consequence weekly or montly sampling strategies are routinely implemented with additional event-based sampling (WHO, 2018; Weiskerger and Phanikumar, 2020).

For the daily management of urban bathing sites, models could also be used instead of collecting additional samples to check the microbial quality of the water after each short-term pollution event (WHO, 2018). However building, training and validation of predictive models require high accuracy data that are difficult and expensive to collect (Jovanovic et al., 2019). Environmental stressors such as physico-chemical, hydrological and meteorological variables are often used as input data in models to predict the concentration of fecal indicator bacteria since real-time measurement of these parameters provides cost effective and high quality data (Nnane et al., 2011; Bui et al., 2020; Banda and Kumarasamy, 2020). Among the different predictive models, machine-learning tools have been proved to predict surface water quality in rivers with high accuracy in different situations, including traditional machine-learning models or ensemblist methods (Ghahramani, 2015; Mälzer et al., 2016; Qiu et al., 2017). However due to the small size of most stakeholder datasets, the performance of the model can be low (Chen et al., 2020; Ghahramani, 2015). The determination of the minimum sampling size and the appropriate sampling strategy required for building, training and validation of models is thus crucial (WHO, 2018).

Several strategies to improve the input dataset of machine learning models exist, however their usefulness for rationalizing the data acquisition for water quality prediction still needs to be evaluated. First, the most relevant observations during the learning process of the models could be identified in order to maximize the information gain (Qian et al., 2020). Second, the weakness in the training dataset could be determined in order to identify which and how much additional data are needed to improve the model performance. For instance, active learning is a method that gives flexibility to identify which instances need to be added to the training set (Zhu et al., 2017). Another popular strategy is to use uncertainty sampling, to identify the point where the prediction is uncertain in the model (Bouneffouf, 2016). Third, another way to enhance the amount of training data is to deploy on site a large number of low cost sensors. Each individual sensor may present a slightly greater error margin than the costly high precision

equipment, however the multitude of sensors allows to build a dense network which in average is capable of providing enough information for the machine learning models (KnowFLow, 2021). However, enrichment of training datasets with high quality data of extreme events is particularly important in the context of climate change with the expected rise of temperature and increase in the frequency and intensity of storm events (Weiskerger and Phanikumar, 2020). Therefore, the objective of this study is to explore these three strategies to improve the input datasets for training and testing machine learning models, particularly study the relevance of the active learning strategy. The ultimate goal is to provide a conceptual framework and an operating mode to assist the stakeholders in the daily management of the bathing sites. The framework thus includes i) a guideline for selecting from a toolbox of six machine-learning models, the one most adapted to their bathing site context and ii) a strategy to improve the training and testing of their model via the sub-optimization of the sampling strategies. The Marne River (Paris Area, France) was considered as a use case. Indeed, several municipalities wish to open bathing sites on the border of the Marne river by 2022. Environmental stressor dataset used to predict *E. coli* concentrations were acquired from the Syndicat Marne Vive.

Using this database, we tested the following strategy :

- 1) We propose to compare the performance of six machine-learning models, including three traditional models and three ensemblist models, to predict the concentrations of the fecal indicator bacteria *Escherichia coli*. To train and test the models, meteorological data and river flow data should be aggregated with physico-chemical data.

-2) For the chosen model, we propose to set up an alert system on the performance of the model in order to optimize the data collection. This alert should consist in identifying under which conditions the model fails to make the prediction and thus alerting the managers to carry out on site analysis in order to enrich the database.

- 3) The usefulness of a network of low cost sensors for sampling optimization as a complementary strategy to improve the dataset is discussed.

2.2. Materials and methods

2.2.1. Study site and water quality data collection

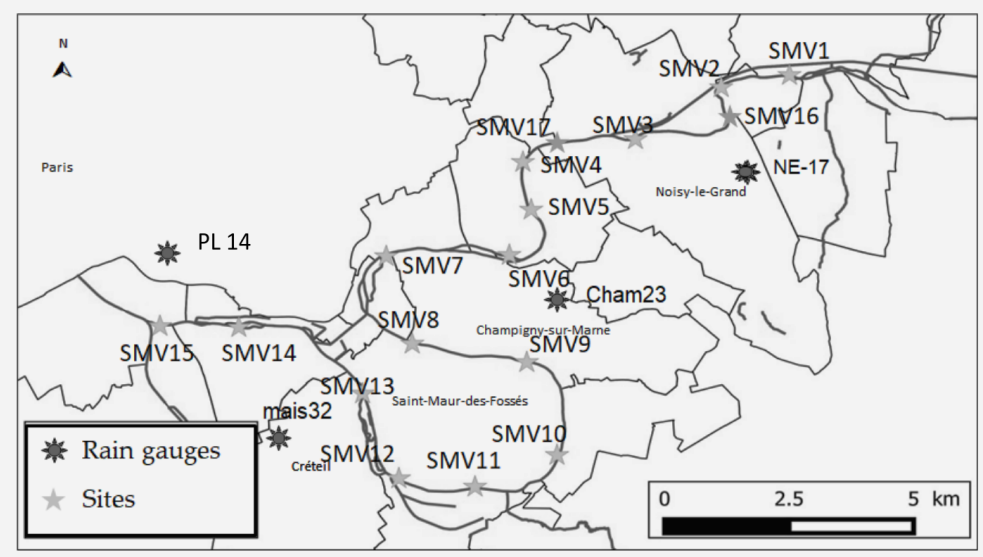

FIGURE 2.1 – Marne River water quality monitoring stations. The light grey stars indicate the SMV sampling stations and the dark grey stars indicate the location of the rain gauges used.

From mid-June to mid-September for 5 years (2015, 2017-2020), samplings were carried out weekly or bi-weekly by the Syndicat Marne Vive (SMV) on 18 stations (SMV0 to SMV17) in the lower Marne River (France) (Figure 2.1). For each sampling site the following parameters were measured : *E. coli* concentrations (Most Probable Number or MPN/100 mL), temperature (°C), turbidity (FTU), conductivity (µS/cm), Total Suspended Solids or TSS (mg/L), NH_4^+ (mg of N/L), Total Kjeldahl Nitrogen or TKN (mg of N/L), number of dry days after the last rainfall, 24 h cumulative rainfall of the day (mm), 24 h cumulative rainfall of the previous day (mm) and the river flow (m^3/s) measured at Gournay-sur-Marne (Paris area, France). The sampling protocole for surface water was carried out according to the French standard FD T 90-523-1 (2008) for physicochemical parameters and according to the 2006/7/EC directive for *E. coli* concentrations. Microbiological and physico-chemical measurements were respectively carried out by Aquamesures and Eurofins (2015) and the Val de Marne Departmental Environmental Health Laboratory (2017-2020) following the French standard methods NF EN ISO 9308-3, NF EN ISO 7027-1, NF EN 27888, NF EN 872, NF T 90-015-2, NF EN 25663.

Rainfall data were obtained from the network of rain gauges of the Departmental Councils of Val-de-Marne (station CHAM23, MAIS32), Seine-Saint-Denis (station NE-17) and the City

of Paris (station PL14). For each sampling point, the meteorological data of the nearest measuring station were used. For the year 2020, rainfall data of the stations SMV5 to SMV13 were not yet available. Flow data measured at the Gournay-sur-Marne station were retrieved from the Banque Hydro (`http://www.hydro.eaufrance.fr/`).

2.2.2. Data preparation

A total of 1696 measures were obtained after data cleaning which consisted of removing the entries with missing and aberrant values. The ID of the station (ordered from upstream to downstream) and the ten measured physico-chemical and hydro-meteorological parameters were used as inputs for our modeling. The output of the models was the concentration of *E. coli*. Then, the dataset was divided randomly in two parts, the training set (90%, 1526 observations) and the test set (10%, 170 observations).

In order to keep all the input parameters with the same degree of influence on the final outcomes, we performed a Z-score standardization for each feature of the datasets (inputs and output) (Chen et al., 2020). The training dataset was used for the standardization in order to block access to the values of the test set during the training of the models.

2.2.3. Machine-learning models

In order to evaluate the performance of the estimation of *E. coli* concentration by the machine-learning models, three traditional machine-learning models (KNN (K-nearest neighbors (Cover and Hart, 1967)), DT (Decision tree (Swain and Hauska, 1977)) and SVM (Support vector machines, (Vapnik, 1995 - 1995)) and three ensemblist learning models (bagging (Breiman, 1996), RF (Random forest (Breiman, 2001)) and AdaBoost (adaptive boosting (Freund and Schapire, 1996)), that combines several base models, were selected and used in this study. All the models were carried out in python 3.7.10 with the Scikit-learn packages (Pedregosa et al., 2011). The GridSearchCV technique was applied to select the hyperparameter that gives the most optimal model by 5-fold cross-validation, over a parameter grid. A 10-fold cross-validation was used to train and to estimate the performance of each model, by randomly generating 10 different sub-sets of the training and test datasets.

2.2.3.1. KNN

The k nearest neighbor method consists in considering the k nearest samples in the training dataset as an input to predict each new observation (Hastie, 2009). For each test datum

the closeness to all the training data is calculated with an Euclidean distance. This allows finding the k observations closest in input space to assign the test datum to a class label, and the output value of each class label is used to estimate the value to predict. The value of k thus varied from 1 to 30 with a step of 2, depending on the dataset.

2.2.3.2. SVM

The support-vector machine goal is to find the optimal hyper-plan from which the distance to all the data point is minimum, it can be applied to classification and regression problems. It consists in transforming the training data representation space into a higher dimensional space, infinite in some cases, and in constructing a hyperplane or set of hyperplanes in a high dimensional space (Hastie, 2009). The idea is to find a solution to flatten the projections of the training points in space without moving too far away from the training points.

2.2.3.3. DT

Tree-based models are used to estimate a quantitative variable or classify observations by reapeatedly separating data into mutually exclusive groupes. The tree-based method slices the variable space and recursively partitions each variable into subsets based on the values of the input variable and then fits a model in each of them (Hastie, 2009).

2.2.3.4. Bagging

Bagging, also known as bootstrap aggregation, uses portions of the data and combines them by generating random subsets of the data through sampling, with repositioning (Barboza et al., 2017). The prediction is obtained by averaging the outcomes of all models. The goal is to reduce the overfitting of predictions in the model.

2.2.3.5. RF

Random forests combine multiple DT at training time. Each tree uses a sample obtained by bootstrap. Given a training set with N measures, the bootstrap aggregation randomly selects N samples with replacement of the training set (Chen et al., 2020). Then a subset of features is randomly selected, in order to construct a collection of decision trees with controlled variance, and fits trees to these samples. The results of the predictions from each tree are averaged (Hastie, 2009).

2.2.3.6. Adaboost

Adaboost repeatedly uses a regression tree developed sequentially on a training sample with weights for each observation adjusted as they are developed (Shrestha and Solomatine, 2006). It starts with fitting a regression to the original dataset and then adjusts the weights of

the variables based on the error of the prediction. Thus, subsequent regressors focus more on poorly fitted or poorly predicted observations (Hastie, 2009). Finally, the results from each weak machine-learning model are combined using the weighted median.

2.2.4. Models evaluation

In order to select the model that performs the best in predicting *E. coli* concentration, the testing phase was carried out with 10 random trials for each model with the 10 test datasets. The prediction performances of each model was evaluated by four statistical metrics. They included Root-mean-square error (RMSE) (Qiu et al., 2017), mean absolute error (MAE) (Bui et al., 2020), the ratio of performance to deviation (RPD) (Wang et al., 2017), and Mean absolute percentage error (MAPE) (Lewis, 1982; Yan et al., 2020). These metrics are calculated as follows :

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - y_i')^2}(1)$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - y_i'|(2)$$

$$RPD = \frac{SD}{RMSE}(3)$$

$$SD = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \bar{y})^2}{N}}(4)$$

In these formulas, (y_i) is the measured value, (y_i') is the predicted value, (N) is the total number of samples, and (SD) is the standard deviation of the tested dataset ($\bar{y}$ is the mean of the measured values). The smaller the RMSE or the MAE, the more stable is the predictive capacity of the model. RPD values < 1.4 indicate that the model is not reliable. For RPD values between 1.4 and 2, the model is moderately accurate and when the value is higher than 2 the model presents a high level of predictive ability (Wang et al., 2017). Mean absolute percentage error (MAPE), which measures the goodness of fit, was also applied.

$$MAPE = \frac{|y_i - y_i'|}{y_i} * 100(5)$$

The lower the MAPE value, the more accurate is the prediction (Lu and Ma, 2020). Values < 50% can be evaluated as "reasonable" even good if < 20%. MAPE values greater than 50%, are indicative of an "inaccurate" prediction. A MAPE value of 50% indicates an overestimation or an underestimation of 50% with regard to the measured value.

2.2.5. Identification of the weakness parts of the dataset

The MAPE values calculated during the 10 trials were used to separate the predicted values in two datasets : the reasonable (MAPE < 50%) and inaccurate estimations of the *E. coli* densities (MAPE $\leq$ 50%), generated by the best model on the Marne River dataset. In order to determine the physico-chemical and hydro-meteorological parameters that potentially influenced the predictive capacity of the best model, a spearman-correlation analysis was performed between the physico-chemical or hydro-meteorological parameters and the predicted values of *E. coli* (V3.5.1, (R-Core-Team, 2018)). All Spearman coefficients (r_s) were tested for their significance based on 5% error. Then the correlation coefficients obtained with the reasonably and inacurately predicted concentrations were compared using a t-test in order to identify the set of hydro-meteorological and physico-chemical parameters that are influent in the model (significant r_s) and that need improvement (t-test, p-value<0.01), or parameters that are less influent (non significant r_s) but could be worth checking after improvement (t-test p-value <0.01). Next, we identified for each parameter that could be improved (t-test p-value <0.01), which data were weakly represented in the dataset. For each parameter, the 10 test sets have been merged together. The set of values contributing to the reasonable dataset were identified and the set of values that gave at least one inaccurate prediction were removed and inspected to identify which additional data are needed to improve the model. This allowed us to identify the set of values that give at least a reasonable or good prediction for our dataset. The guideline for selecting the best model for *E. coli* concentration prediction among the six machine-learning models, and the strategy to identify a set of parameters and values range needed to optimize the sampling strategies are displayed in the Figure 2.2. The python and R script of the framework are avaible on GitHub (`https://github.com/naloufi-manel/ML-performance-microbial-quality.git`).

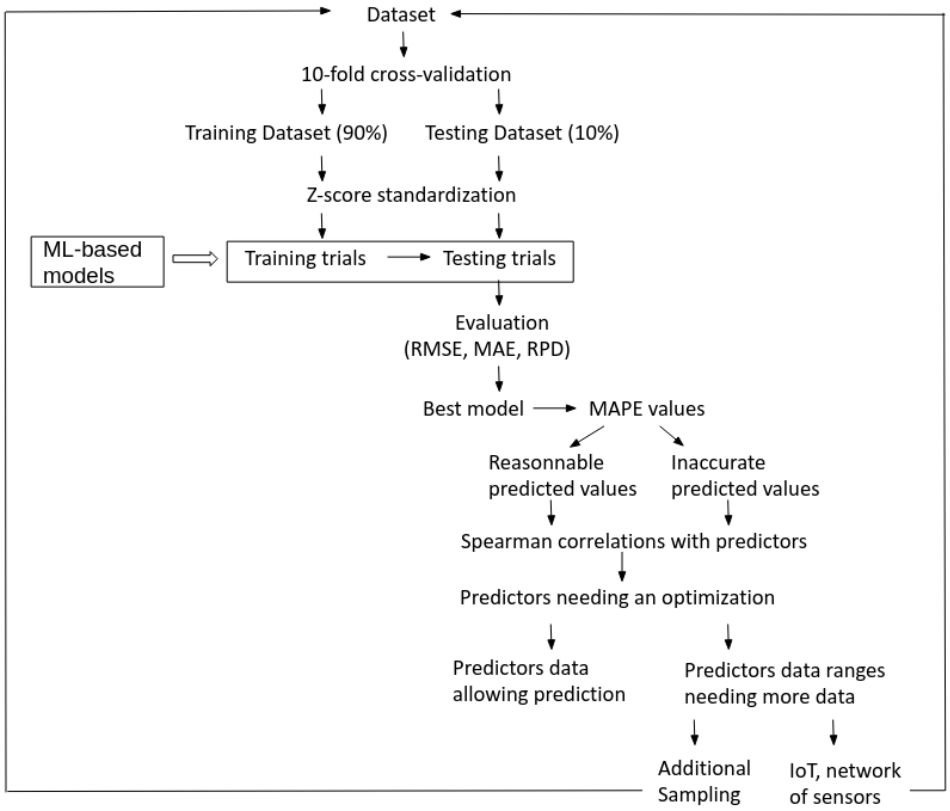

FIGURE 2.2 – Guideline to provide and select an adapted model for water quality prediction and for the identification of a set of data to optimize the sampling strategies.

2.3. Results and discussion

2.3.1. The dataset used in this study

The Marne River dataset was characterized by a high heterogeneity concerning the number of observations per station (13 to 47 entries). The summary sample statistics of the dataset are reported in Table S1. The temperature and the conductivity displayed a fair representativeness (Figure 2.3). However, most parameters presented a skewed distribution and the presence of upper and lower outliers (Figure 2.3). Indeed, for each parameter (except the temperature and the conductivity) a range of values were rarely measured and therefore not well represented in the dataset. This indicates that our dataset is not yet representative of all possible measurements.

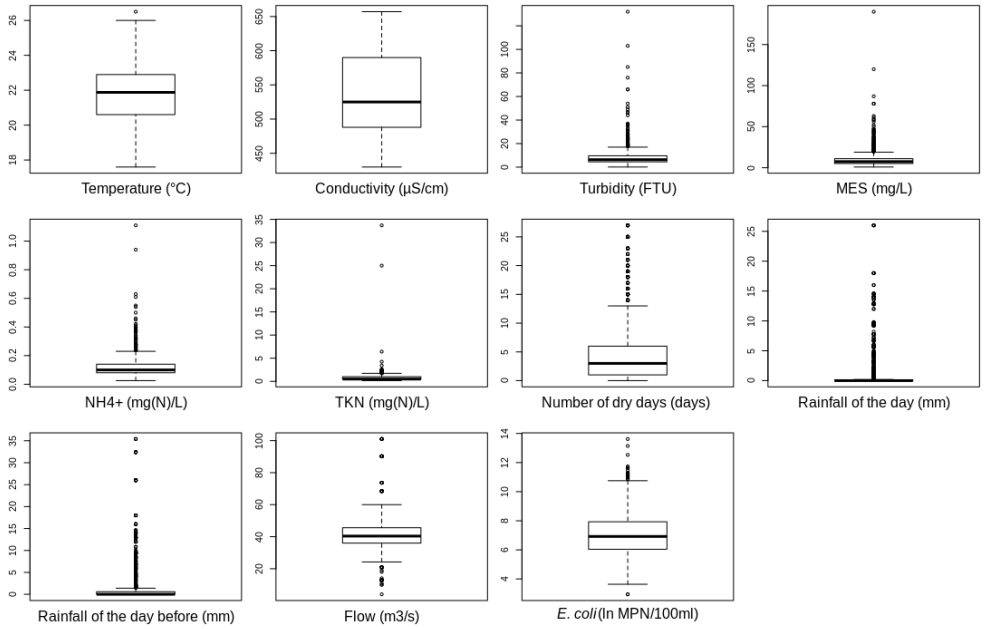

FIGURE 2.3 – Distribution of the data for each variable. The median is indicated as a solid black line inside each boxplot, outliers are indicated as black dots. On the ordinates are the values taken by each variable with the units specified in parenthesis.

The concentration of *E. coli* (4337.61 ± 25983.50 MPN/100 mL) measured during the 5 summers in the Marne river ranged from 19 to 820670 MPN/100 mL. Three pollution events producing very high concentrations of *E. coli* could be identified. For instance, the maximum *E. coli* value (820670 MPN/100 mL), corresponds to high values of turbidity, TSS and TKN levels (respectively 28 FTU, 33 mg/L and 2.6 mg of N/L) compared to the majority of the measurements. Extreme pollution events are often under-represented in the datasets due to their low frequency. For instance rainfalls >10 mm which lead to peaks of pollution occur less than 20 days per years in Quebec region (Sylvestre et al., 2020). However, removing extreme values from the dataset can lead to a decrease in the predictive capacity of the model during events with high pollution. Chen et al. (2020) have shown that a better performance could be achieved after increasing the training data for each of the learning models. Considering the biased distribution of most parameters in the Marne River dataset, it may be necessary to add additional measurements to increase the size of the database and improve the training of the ML models. This would provide a better representation of the set of possible values. However, the high cost of field sampling and laboratory analyses for monitoring microbiological quality (about 100 € according to the Syndicat Marne Vive) requires an optimisation of the collection in order to identify the necessary measures to efficiently complete the datasets.

2.3.2. ML-based *E. coli* prediction comparison

Various machine-learning models have been used previously to predict water quality and their predictive performance was compared to other models by assessing their ability for prediction (e.g. Mälzer et al. (2016); Avila et al. (2018); Bui et al. (2020)). In this study, we compared the performance of six machine-learning based algorithms (KNN, DT, SVM, Bagging, RF and AdaBoost) to predict *E. coli* concentration in an urban river, to identify the best suited model. We performed a trial-and-error procedure, using the RMSE, MAE and RPD metrics to evaluate the performance of each model. Average values of these statistics metrics for each random trial are available in Table S2. The RF model exhibited the highest prediction power among all the models with the weakest error (average value 0.37 ± 0.20 for RMSE and 0.09 ± 0.02 for MAE) followed by KNN and Bagging (respectively 0.41±0.28 and 0.38±0.19 for RMSE and 0.09±0.03 and 0.14±0.06 for MAE) (Figure 2.4).

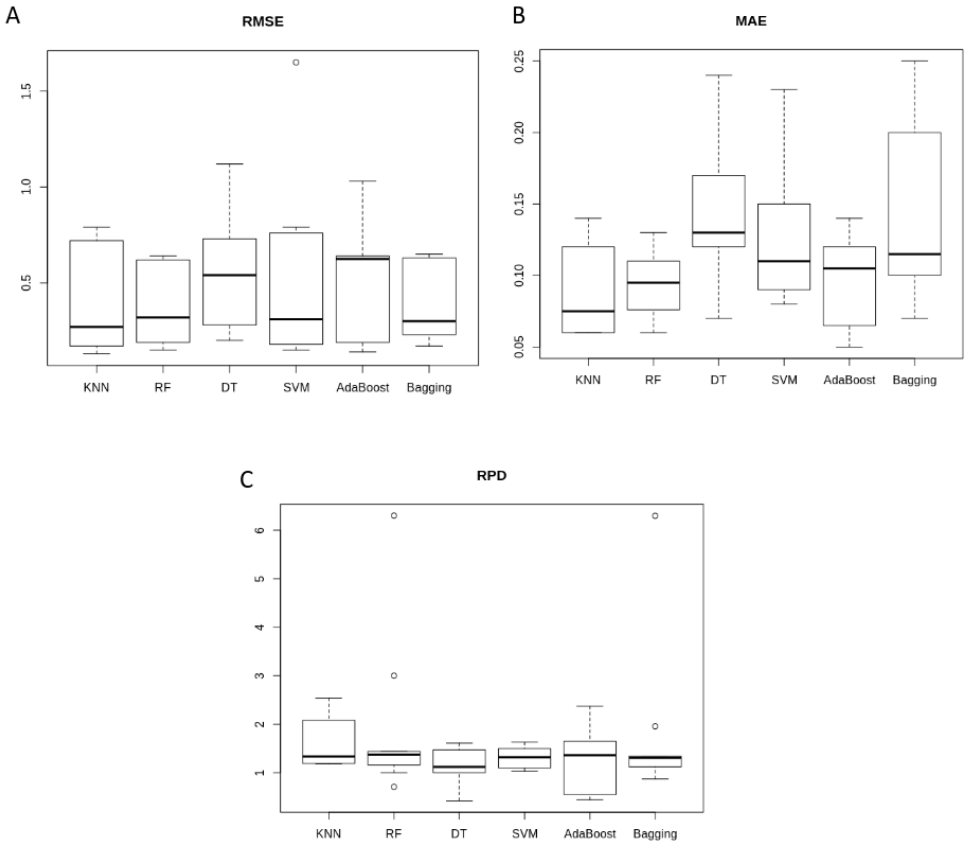

FIGURE 2.4 – Evaluation of the prediction performances of the 6 machine-learning models during the 10 trials. On the abscissa the model is indicated and on the ordinate the value of the statistical metrics are displayed (dimensionless) : (A) RMSE ; (B) MAE ; (C) RPD.

An analysis of the accuracy and reliability of the model was also performed using the RPD metrics (Figure 2.4C). Three models (KNN, Bagging and RF) were estimated as moderately

accurate and presented acceptable results. The 3 other models were not reliable, with an RPD < 1.4 (Figure 2.4). For the RF model, the RDP value was close to 2 (1.91 ± 1.65), indicating that the model had a high predictive capacity. In conclusion, the RF model gave better *E. coli* concentration estimation compared to other machine-learning models. This result is in agreement with Bui et al. (2020) but disagree with the results of Chen et al. (2020). Both studies compared the performances of DT models with RF models in their ability to predict water quality. We also checked if the performance of the models will increase by compacting the sampling sites together, however without the station ID the performance of all models slightly decreased (data not shown).

Our results confirm that Ensemblist learning models have a better performance compared to traditional models (e.g. KNN and SVM). This conclusion is in agreement with some previous studies (e.g. Ahmed et al. (2019a); Bui et al. (2020)). However we must bear in mind that the performance of a model depends on external uncertainty conditions (Chen et al., 2020). Thus, for each specific dataset, several algorithms should be tested in order to find the models with the best fitting to *E. coli* concentrations. Indeed Mälzer et al. (2016) found that the performance of models could differ from one site to another along the Ruhr River in Germany. For this reason we proposed this set of six machine-learning models as a basic toolbox to be used.

2.3.3. Limits of ML-based *E. coli* estimation

Identifying observations with uncertain predictions is an approach to determine the set of data requiring optimization and thus find a way to optimize the collection and to efficiently complete our training set, allowing for a better prediction in the future by re-running the model with the newly collected measurements. Indeed, recent studies have shown that increasing the quality and quantity of the dataset by adding complementary measures allows to effectivly increase the training set and to improve prediction accuracy (Pachepsky et al., 2018; Chen et al., 2020).

To further analyze the performance, the MAPE indice, which measures the goodness of fit and examines the performance of models based on their tendency to estimate the *E. coli* concentration, was calculated for all testing trials. For 46.7 ± 3.5% of the predicted values generated by the RF model, the percentage of the absolute error was less than 50%, which indicates that the estimates were reasonable or even good. The remaining 53.2 ± 3.5% of the predicted values were associated with MAPE values equal or exceeding 50%, corresponding to

inaccurate estimates. These results indicate that the RF-based model did not properly predict *E. coli* values in all contexts and that our dataset is not sufficient to efficiently train the RF-based model. Figure 2.5 indicates uncertainty in the prediction for some of the *E. coli* measurements.

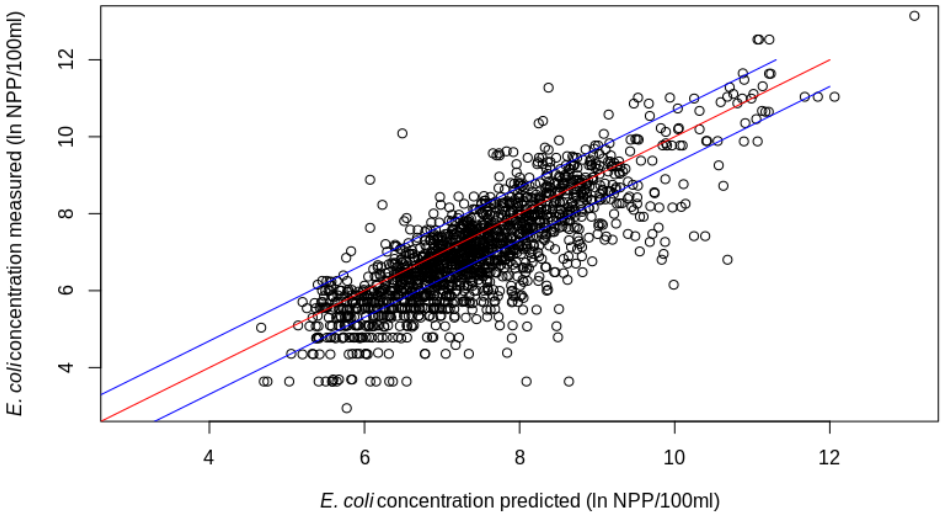

FIGURE 2.5 – Relationship between the *E. coli* concentration predicted by the RF-based model and the measured concentration. The white circles indicate the values. The red line indicates theoretical values corresponding to an accurate prediction of the model compared to the measured values for the ten testing trials. Blue lines indicate the 50% confidence interval.

2.3.4. Identification of the weaknesses in the dataset

Different methods can be used to improve the input datasets. Some studies focuses on finding the best combination of input variables to improve the algorithm's predictions (e.g. Bui et al. (2020); Hameed et al. (2017)). However, weak features also represent a powerful source of information, that can be used in combination with the features that are adequate for learning the target concept (Muslea et al., 2006). Thus, in our study, we propose to use the second strategy. For this purpose, the prediction limits and biases of the RF-based model were further examined in order to identify among the physicochemical and hydro-meterological variables the weaknesses in the training and testing datasets.

We hypothetized that the variability induced by the low representativeness of some parameters can affect the predictive capacity of the model. In order to identify the key parameters allowing a reasonable estimation of *E. coli* concentrations and those leading to an inaccurate estimation, the predicted values were separated in two datasets (inaccurate and reasonable

TABLE 2.1 – Correlation coefficients (average r_s ± SD) for the relationship between the values of *E. coli* predicted by the RF model (reasonable and inaccurate) and the environmental variables. Significant coefficients are indicated with a * (coefficient significance test p <0.05). Significant differences between the correlation coefficients of the two datasets are indicated as t-test p-values < 0.01. MAPE values were used to identify reasonable (less than 50%) and inaccurate (over 50%) estimations of *E. coli* concentrations obtained with the RF model during the ten testing trials.

Parameter	**Reasonable predictions (r_s)**	**Inaccurate predictions (r_s)**	**p-value**
Water temperature	-0.17 ± 0.05	$-0.28^* \pm 0.07$	0.001
Conductivity	-0.05 ± 0.11	-0.18 ± 0.09	0.009
Turbidity	$0.42^* \pm 0.07$	$0.39^* \pm 0.08$	0.43
TSS	$0.43^* \pm 0.09$	$0.40^* \pm 0.04$	0.42
NH_4^+	$0.54^* \pm 0.06$	$0.48^* \pm 0.07$	0.05
TKN	-0.03 ± 0.08	0.001 ± 0.06	0.26
Number of dry days	-0.10 ± 0.09	-0.01 ± 0.09	0.02
24 h cumulative rainfall (day)	0.09 ± 0.10	-0.02 ± 0.11	0.02
24 h cumulative rainfall (previous day)	0.17 ± 0.08	0.03 ± 0.10	0.002
River flow	$0.54^* \pm 0.09$	$0.39^* \pm 0.09$	0.001

estimations) based on the MAPE indice 50% threshold. Then an analysis of the relationship between the differents physico-chemical and hydro-meteorological variables and the predicted values was carried out on the inaccurate and reasonable datasets. We assumed that a significant difference in the coefficient correlation between the two datasets would point out the variables that had an impact on the outcome of the model but needed optimization. To compare the correlation coefficients obtained with the reasonable and inacurrate datasets, a t-test was used (n = 10). The p-values obtained are displayed in (Table 2.1).

Turbidity, TSS, NH_4^+, were important predictors (significant r_s above 0.40), and no significant differences in the two datasets arised (t-test, $p \geq 0.05$, Table 2.1). We classified these parameters has having an impact of the RF-model output, with no urgent need for additional data. The river flow also contributed to the model output (significant $r_s > 0.40$), but there was a significant difference between the two datasets (t-test, $p < 0.01$, Table 2.1). It was thus considered as an important parameter that needs additional data. Finally the water temperature, the conductivity, the 24 h cumulative rainfall of the previous day (Table 2.1, t-test, $p < 0.01$), as well as for the number of dry days after the last rain and 24 h cumulative rainfall of the day (Table 2.1, t-test, $p < 0.05$) showed weak correlations with *E. coli* values, but a difference between the two datasets. Since it is not certain if these weak correlations are an artifact due to the skewed distribution of these parameters or if these parameters are just bad predictors, we

decided to further explore the parameters with a highly significant difference in the correlation obtained with the reasonable and inaccurate estimates. Thus for the river flow, temperature, the conductivity and the 24 h cumulative rainfall of the previous day (t-test, $p < 0.01$), it was considered that additional data were needed to provide the dataset with enough information to reduce the uncertainty in the model's estimates. The reasons for this uncertainty may be that the measurements have not yet been tested and it is not yet known whether the model will be able to reasonably estimate the *E. coli* concentration, or that the distribution of the data is skewed and the correlation of some environmental variables with the *E. coli* concentration is not yet obvious.

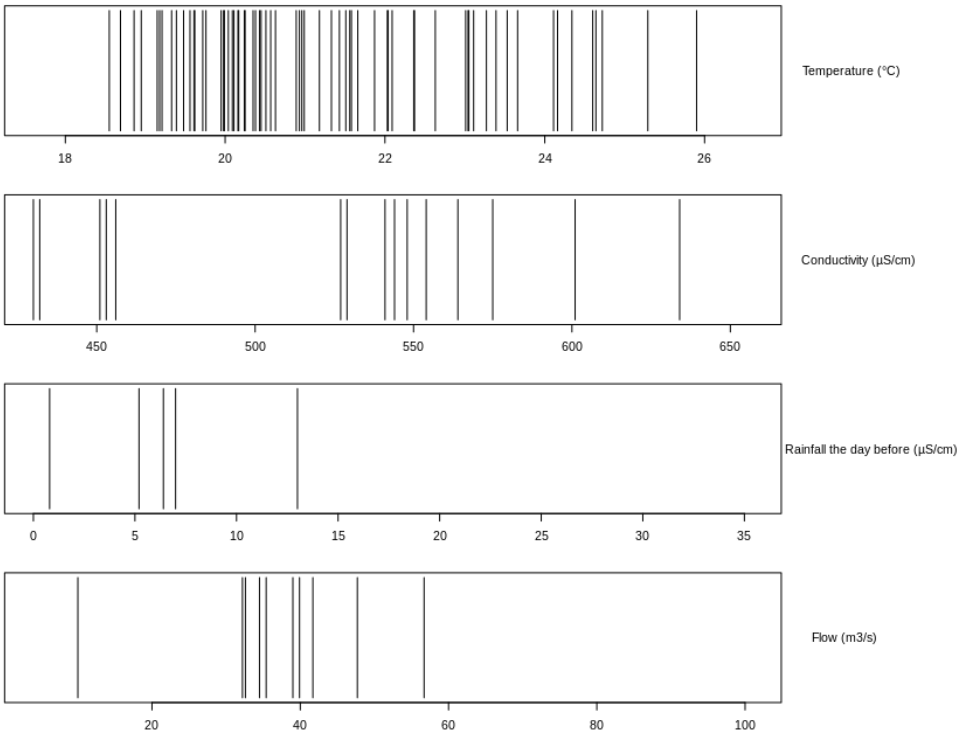

FIGURE 2.6 – Vizualisation of the values that need enrichment in the dataset for the temperature, conductivity, 24 h cumulative rainfall of the previous day and the river flow. The abscissa displays the value range of each parameter. Predicted values giving a reasonable estimation are visualized with solid black bars, white spaces represent the values that need further enrichment in the dataset.

The next step was to identify the value ranges of the four parameters that needed extra measurements to efficiently complete the training and testing datasets. A deeper understanding of the behavior of these parameters in the model should help optimizing the sampling process while minimizing additional cost and efforts of sample collection and analysis. The temperature is the parameter for which the reasonable predicted values of *E. coli* densities covered pretty well the whole range of values [17.6-26.5 °C] (Figure 2.6). For the conductivity [data range 430-657

µS/cm] and the flow [data range 4-101 (m^3/s)] the distribution of the reasonable estimates was not regularly disseminated along the data range, and the 24 h cumulative rainfall of the day there was only 4 reasonable values in the [data range 0-35.4 mm] (Figure 2.6). The Figure 2.6, is a valuable tool to identify which data are missing in the data range, and thus help to determine where the sampling efforts should be carried out.

In our study, the RF-based model produced a versatile modeling in prediction. Based on this observation, we were able to identify a set of parameters and values needed to complete the dataset. An alert system based on the analysis of the reasonable and inacurate estimates would be a valuable tool for stakeholders to optimize their sampling and measurement efforts. However manual sampling and laboratory analysis maybe too costly and labor intensive to realistically implement the training dataset. A network of sensors allowing continuous monitoring of physico-chemical parameters and the monitoring of rainfalls as well as dry weather, could help in optimizing the sampling. Such approach may help developping models able to adapt under environmental perturbations such as accidental pollutions or heavy rainfalls (> 30 mm), which are usually under-represented in the datasets due to their scarcity, and/or the fact that weekly/monthly routine survey often miss such events.

2.4. Automated data collection

From the results, it is clear that the machine-learning models are capable of delivering interesting results, as long as one can provide enough good-quality data. Thus, the use of data sensors in addition to manual collection should be investigated as means of feeding these models. Concerning the water quality parameters that we have investigated in this work, there are a myriad of sensors that could perform their collection with acceptable data quality. Those sensors may vary in price, accuracy, usage, lifespan among other characteristics, as they were extensively studied in Abegaz et al. (2018) and Kruse (2018). Therefore, to incorporate sensors as a permanent brick in the data collection system, further studies must be conducted to determine their optimal and sub-optimal numbers to be deployed on a given site, the expected accuracy and the available budget for their acquisition. In this direction, Abegaz et al. (2018) have thoroughly discussed the nature of different sensors (piezoelectric, optical, etc.) and how they fit for different use cases, while Kruse (2018) provide interesting inputs concerning their usages for different use cases.

One strategy for monitoring the bathing water quality and decide when to open or close a bathing site is to use online measurement systems that detect Beta-D-Galactosidase or Beta-D-Glucuronidase activity. For instance, the ColiMinder automated measurement system (Vienna Water Monitoring, VWM GmbH) (Cazals et al., 2020), ALERT system (Fluidion) (Angelescu et al., 2019), Colifast ALARM™ (Tryland et al., 2015), TECTA-B16 (Endetec, Veolia) (Bramburger et al., 2015), have been demonstrated to be useful to monitor *E. coli* in rivers, but the price of these devices may be economically prohibitive for numerous cities, since one unit may cost up to tens of thousand of euros. Alternatively the use of sensing technologies to measure proxies or surrogate parameters procures high frequency, precise and accurate data. Based on electrodes, fluorescence, colorimetry, wet analytical chemistry, or flow cytometry techniques, these devices are deployed at fixed strategic locations (Rode et al., 2016). However these sensors are often costly ($\sim$10K euros unit price), for instance multiparameter sensors such as Proteus Multi-parameter Water Quality Sensor based on tryptophane-like fluorescent detection or sensors plateforms measuring physico-chemical proxies (such as YSI, Sea-Bird or NKE instrument) are often used to monitor water quality.

One interesting way of integrating a network of sensors to data collection is to build an Internet of Things (IoT) network, mixing high-quality (expensive) and medium-quality (cheaper) devices capable of delivering real-time analysis. In comparison, cheaper sensors can be used to deliver good enough approximations of the correct data. For instance, the KnowFlow platform KnowFLow (2021), based on Arduino computers and IoT long-range communication can be a significant addition to the network. A recent review of low-cost sensors is provided by Wang et al. (2019a).

Concerning the deployment of these heterogeneous sensors, some approaches exist to maximize the quantity and quality of gathered data. The collection system may rely on i) deterministic deployment, where sensors position is calculated before the collection begins, based on the environmental and economic conditions (Nguyen et al., 2019) ; ii) random deployment, in the case where areas are hard to achieve and to position sensors (Priyadarshi et al., 2020) ; iii) hybrid deployment, a mix of aforementioned approaches, which is used indicated to very large networks, covering vastly heterogeneous areas (Senouci and Mellouk, 2016). Some studies have investigated this topic, with a further analysis on the advantages of IoT networks to enhance data collection (Ciaponi et al., 2018; Ramesh et al., 2017). For instance, in Ciaponi et al. (2018) authors proposed a methodology to derive the optimal placement of sensors in an aquatic envi-

ronment, based on a "divide-and-conquer" approach, which could reduce the complexity of this task for large scenarios.

The deployment of sensors will heavily depend on the battery lifespan of devices, as much as on their communication range. Therefore, IoT-based measurement networks should be based on Low Power Wide Area Networks (LPWAN) technologies, as LoRaWAN, Sigfox or NB-IoT. Such technologies allow communications range up to kilometers and ensure very low energy consumption, when compared to 4G, Wi-Fi or Bluetooth networks (Mekki et al., 2019). Users can also consider the utilization of new 5G cellular technology, which is adapted for large-scale sensor networks and IoT communications (Rahimi et al., 2018).

One remaining challenge to enhance the use of IoT networks for water quality assessment is the real-data collection and visualization mechanisms. For example, Grafana allows users to analyze sensor metrics through dashboards, messaging and alerts in real time (Betke and Kunkel, 2017). The Elastic stack application allows a deeper analysis of data logs and provides so-called intelligent dashboards, capable of adapting screens to environmental, economic or user contexts (e.g., what a researcher sees is not what a common user would see) (Protopsaltis et al., 2020). In Ramesh et al. (2017), authors developed an IoT-based system within a town, capable of sensing the environmental parameters and effectively delivering real-time information on water quality. This clearly shows that the automation of the collection process is possible and viable for the estimation of water quality in urban sites Chen and Han (2018).

Although the use IoT network composed of heterogeneous sensors is an interesting solution to enhance surveillance systems, the use of low-quality devices must be taken with caution due to their less accurate results. Therefore, the calibration of sensors remains an important issue to be investigated. As discussed in Abegaz et al. (2018), the errors, margins and durability of devices vary a lot. Therefore, an automated data collection must take into account a mechanism to estimate which sensors are no longer in optimal operation conditions, which is more likely to happen to low-quality models. One simple solution consists in compare their output to nearby high-quality devices and analyze when important deviations occur. More complex solutions would consist in estimating their lifespan based on already collected data to perform changes preemptively.

2.5. Conclusion

In this paper, we proposed a framework and statistical indicators to select among a toolbox of six supervised learning algorithms (KNN, SVM, DT, RF, Bagging and AdaBoost) the most suitable model for the prediction of fecal indicator bacteria in an urban river. This framework was sucessfully applied to the Marne River (Greater Paris, France). Netherless, with regard to the actual dataset, *E. coli* concentration could not be predicted in all contexts (53.2 ± 3.5% of inaccurate predicted values). This result illustrates well the fact that predicting the microbial quality of surface waters in urban rivers remains complex. Refining the models to be able to adapt to environmental changes represents a future challenge in the context of the global change which may increase the frequency of extreme rainfalls and floods Sylvestre et al. (2020). Further amelioration and testing of predictive models is needed to reproduce and predict the temporal and spatial dynamic of fecal indicators in changing and complex aquatic environments. Due to the fact that our dataset was not representative of all the possible values in the data range, some values have not yet been trained or tested by the RF-based model. For these values it is not clear yet whether our model is able to estimate the *E. coli* concentration in a reasonable way at the moment. To address this problem, we proposed a strategy and tools to help improving the quality and quantity of the training data. The distribution of the accurate values along the data range of each parameters seems an appropriate approach to identify which additional data are needed for which parameter, in order to achieve a good predictive efficiency.

Acquiring additional data is usually costly because it's a manual process that requires human action. As a consequence our proposed approach aims to optimize the sampling process. It requires to focus on the following points :

i) How and where to use of microbiological high-quality monitoring systems to feed itself ; ii) How to install low cost physico-chemical sensors on an IoT network for the prediction of microbiological quality and iii) When to perform sampling by human operators when the model fails to correctly estimate the *E. coli* concentration and the microbiological quality of surface water ?

Overall the proposed framework will help rationalize and optimize the sampling effort, thus saving time and cost of microbiological analyses.

Data Availability Statement : Datasets are deposited in the CapGeo database of a

working group directed by the City of Paris to study the water quality of the Seine and the Marne river. This dataset is not yet openly accessible.

Acknowledgments : We thank the Departmental Councils of Val-de-Marne and of Seine-Saint-Denis (France) and the city of Paris for their contribution to the dataset. We are grateful to Lamine Amour for his constructive preliminary work. We are grateful to Miguel Gillon-Ritz for his sound advices and for the access to CapGeo database (City of Paris, Direction de la Propreté et de l'Eau - Service Technique de l'Eau et de l'Assainissement). We thank Miguel Gillon-Ritz and Marion Delarbre for their kind welcome and supervision at the Service Technique de l'Eau et de l'Assainissement (City of Paris).

2.6. Appendix

TABLE S1 – Descriptive statistics of the parameters.

Parameter	Mean	Standard deviation	Minimum	Maximum
Water temperature	21.77	1.59	17.60	26.50
Conductivity	537.36	56.09	430.00	657.00
Turbidity	7.91	7.33	0.12	132.00
TSS	9.76	9.13	0.90	190.00
NH_4^+	0.12	0.07	0.03	1.11
TKN	0.72	1.08	0.15	33.70
Number of dry days	4.80	5.72	0.00	27.00
24 h cumulative rainfall (day)	0.97	2.90	0.00	26.00
24 h cumulative rainfall (previous day)	1.86	4.69	0.00	35.40
River flow	41.68	10.59	4.00	101.00

TABLE S2 – Average and standard deviation of the statistic metrics (RMSE, MAE, RDP) obtained with each model during the ten testing trials.

Metric	KNN	RF	DT	SVM	AdaBoost	Bagging
RMSE	0.41 ± 0.28	0.37 ± 0.20	0.54 ± 0.29	0.53 ± 0.48	0.53 ± 0.28	0.38 ± 0.19
MAE	0.09 ± 0.03	0.09 ± 0.02	0.14 ± 0.05	0.13 ± 0.05	0.10 ± 0.03	0.14 ± 0.06
RDP	1.60 ± 0.49	1.91 ± 1.65	1.12 ± 0.36	1.32 ± 0.22	1.28 ± 0.62	1.77 ± 1.62

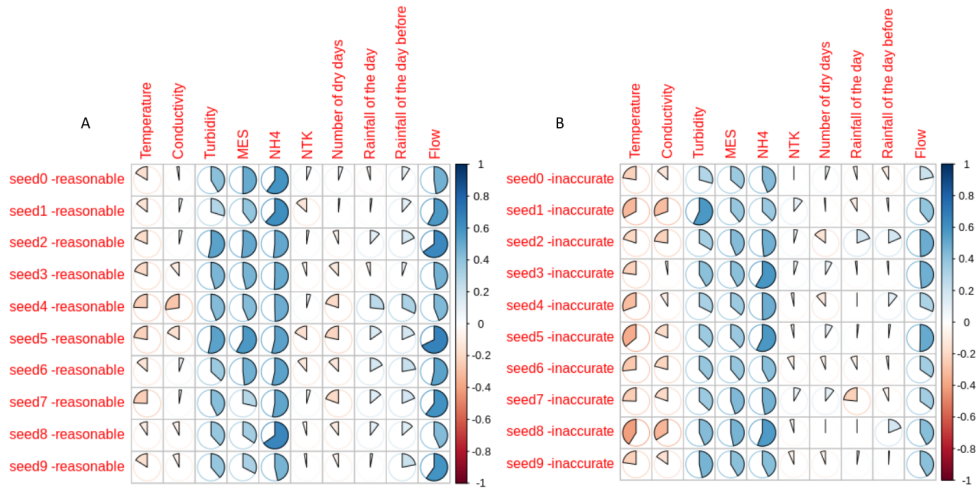

FIGURE S1 – Correlation analysis between water quality parameters and *E. coli* concentration estimated by RF for (A) reasonable estimation ; (B) inaccurate estimation of *E. coli* (n=10).

3. Évaluation de la performance des approches d'apprentissage automatique et d'apprentissage par transfert pour prédire la qualité microbienne des eaux de surface en Seine et en Marne

Manel Naloufi[1,2,*], Françoise S. Lucas[2], Sami Souihi[3], Pierre Servais[4], Aurélie Janne[5] and Thiago Wanderley Matos De Abreu [3,*]

[1] Direction de la Propreté et de l'Eau - Service Technique de l'Eau et de l'Assainissement, 27 rue du Commandeur 75014 Paris, France ; manel.naloufi@paris.fr,

[2] Leesu, Université Paris-Est Créteil, École des Ponts ParisTech, 61 avenue du Général de Gaulle, 94010 Créteil Cedex, France ; lucas@u-pec.fr

[3] Image, Signal and Intelligent Systems (LiSSi) Laboratory, University of Paris-Est Créteil Val de Marne, 122 rue Paul Armangot, 94400 Vitry sur Seine, France ; thiago.wanderley-matos-de-abreu@u-pec.fr ; sami.souihi@u-pec.fr

[4] Ecology of Aquatic Systems, Université Libre de Bruxelles, Brussels, Belgium ; Pierre.Servais@ulb.be

[5] Syndicat Marne Vive, Maison de la Nature, 77 quai de la Pie, 94100 Saint-Maur-des-Fossés, France ; aurelie.janne@marne-vive.com

Résumé : Pour améliorer la gestion quotidienne des sites de baignade, la surveillance d'*E. coli* est essentielle. Cependant, ce suivi est souvent limité temporellement et spatialement en raison de contraintes méthodologiques, logistiques et financières. La modélisation constitue un outil précieux pour gérer les pollutions à court terme et pourrait également optimiser la collecte de données. Néanmoins, la performance des modèles varie selon les sites, rendant crucial le choix du modèle le plus approprié. Dans notre étude, nous avons comparé les performances de six algorithmes d'apprentissage automatique : K-nearest neighbors (KNN), Decision tree (DT), Support vector machines (SVM), bagging, Random forest (RF) et adaptive boosting (AdaBoost) pour prédire les concentrations d'*E. coli* dans la Marne et la Seine en région parisienne, France. Nous avons proposé un cadre conceptuel pour sélectionner le modèle le plus adapté et rationaliser l'effort d'échantillonnage afin d'optimiser le jeu de données d'entraînement. Selon plusieurs

mesures statistiques, le modèle RF a démontré la meilleure précision. Il apparaît que l'origine du jeu de données d'entraînement, ainsi que la distribution et le nombre de paramètres explicatifs, influencent significativement la performance du modèle. Si certains paramètres explicatifs sont bien représentés dans le jeu de données, d'autres, comme la température et la conductivité, nécessitent une optimisation pour la Seine. Ainsi, le modèle prédictif pourrait alimenter un système combinant échantillonnage manuel in situ et déploiement de capteurs, optimisant ainsi le suivi de la qualité microbiologique de l'eau.

Mots clés : Prédiction de la qualité de l'eau, apprentissage automatique, concentration en *E. coli*, optimisation de l'échantillonnage, surveillance des rivières

3.1. Introduction

Ces dernières décennies, les vagues de chaleur mondiales ont intensifié les activités aquatiques dans les mégapoles, augmentant les interactions entre les citoyens et l'eau douce urbaine (Jang, 2016). Cependant, nager dans des eaux de surface urbaines, rivière ou lac, présente des risques sanitaires liés à la contamination des eaux par des agents pathogènes issus de rejets ponctuels ou diffus (Soller et al., 2010). La surveillance actuelle repose sur des analyses biologiques et chimiques en laboratoire, mais des modèles prédictifs pourraient compléter et aider à rationaliser l'échantillonnage réglementaire, afin de faciliter la gestion quotidienne des zones de baignade (OMS, 2018).

Plusieurs modèles prédictifs, allant des régressions linéaires aux réseaux de neurones, peuvent être utilisés pour estimer les concentrations en *Escherichia coli* (van der Meulen et al., 2024). Par exemple, un modèle basé sur les arbres de régression a été utilisé pour prédire en temps réel les concentrations en *E. coli* dans les rivières du sud de la Nouvelle-Zélande, en se basant sur des données météorologiques et hydrologiques (Avila et al., 2018). Cependant, la performance des modèles peut varier en fonction du site et/ou du contexte météorologique et hydrologique (Mälzer et al., 2016). Les modèles d'apprentissage automatique, notamment les méthodes ensemblistes comme les forêts d'arbre de décision (Random Forest) et le bootstrap aggregating (aussi appelé bagging), ont montré une grande précision dans la prédiction de la qualité des eaux de surface (Bui et al., 2020). Toutefois, l'efficacité des modèles basés sur l'apprentissage automatique est souvent limitée par la taille des ensembles de données disponibles (Chen et al., 2020). Or, la collecte de données de haute qualité reste coûteuse

et complexe, ce qui tend à limiter la fréquence des prélèvements au minimum exigé par la réglementation et à limiter la période à la saison de baignade de juin à septembre (van der Meulen et al., 2024). Par ailleurs, les données de concentration en bactéries indicatrices fécales (BIF) sont généralement acquises de manière épisodique et ne couvrent pas toujours les différentes conditions environnementales qui peuvent affecter le site de baignage telles que les périodes d'étiage et de hautes eaux (Jovanovic et al., 2019). En effet, il est important d'utiliser des jeux de données qui couvrent différentes conditions environnementales qui se produisent sur un site, telles que les débits élevés et les débits faibles (Herrig et al., 2019).

Les connaissances sont augmentées en explorant d'autres bases de données pour identifier des similitudes entre différents sites ou contextes, en utilisant l'apprentissage par transfert (Dipanjan, 2018). Pour pallier la faible taille et diversité des ensembles de données, il existe différentes stratégies tout en minimisant l'effort et le coût de l'échantillonnage (Wu et al., 2024). D'une part, il est possible d'orienter les campagnes de mesure et d'optimiser la collecte de données en ciblant des périodes ou des conditions environnementales stratégiques, afin d'améliorer la quantité et la représentativité des données disponibles. D'autre part, il est possible d'augmenter la taille et la qualité de la base de données existante sans faire de prélèvements supplémentaires à l'aide de techniques d'apprentissage automatique, en générant des données synthétiques, ou bien en réduisant les besoins en données du modèle, ou bien entransférant les connaissances d'une autre base de données disponible (Wu et al., 2024). Pour cette dernière approche, en effet, le transfert de connaissance d'une base de données à une autre peut servir soit à compléter une série temporelle avec des données manquantes, soit à pré-entrainer un modèle. Le transfert de connaissance consiste à tirer parti de jeux de données riches et diversifiés pour améliorer les performances des prédictions (Noam, 2016; Segev et al., 2015). Par exemple, un modèle initialement entraîné sur des données issues de plusieurs bassins versants peut être adapté à d'autres sites. Les connaissances sont alors augmentées en explorant d'autres bases de données pour identifier des similitudes entre différents sites ou contextes, en utilisant l'apprentissage par transfert (Dipanjan, 2018).

L'objectif de notre étude est de développer un cadre conceptuel et pratique pour aider les gestionnaires des sites de baignade à prédire les concentrations en *E. coli*, notamment dans les rivières Seine et Marne en région parisienne (Île-de-France). Ce cadre inclut la sélection des modèles les plus performants et l'optimisation des stratégies d'échantillonnage. Avec la demande croissante pour des sites de baignade en Île-de-France, en particulier en vue des Jeux

Olympiques et Paralympiques de 2024, il est crucial d'améliorer les méthodes de gestion de la qualité de l'eau (Noury et al., 2018). Nous avons comparé les performances de six modèles d'apprentissage automatique, incluant des modèles traditionnels et des méthodes ensemblistes, pour la prédiction des concentrations en *E. coli*. L'entrainement a été réalisé avec des variables prédictives physico-chimiques et des variables hydrométérologiques. Nous faisons l'hypothèse que les concentrations en BIF peuvent être prédites avec des données acquises en routine par les collectivités. Le modèle avec la meilleure capacité de prédiction et la meilleure précision a été sélectionné. Afin de tenter d'améliorer la performance des modèles, nous avons testé trois stratégies. Tout d'abord, nous avons manipulé le nombre de variables prédictives pour évaluer si la sélection d'un nombre limité de prédicteurs peut permettre une bonne prédiction. Ainsi, nous avons comparé les performances de six modèles d'apprentissage automatique en utilisant 11 et 8 paramètres de la base de données de la Marne. Nous avons également testé l'approche d'apprentissage par transfert entre deux bases de données de rivières proches localement, pour vérifier si cet enrichissement de l'entraînement apporte une amélioration de la prédiction. Pour cela, nous avons utilisé alternativement les bases de données de la Marne et la Seine pour entraîner et tester les modèles. Enfin, nous proposons une approche pour optimiser la collecte des données, en identifiant les conditions physico-chimiques qui génèrent des incertitudes dans les prédictions des concentrations en *E. coli* dans la Seine. Ce travail vient compléter l'article de Naloufi et al. (2021) présenté au niveau de la section 2.

3.2. Matériel et méhodes

3.2.1. Site d'étude et collection de données sur la qualité de l'eau

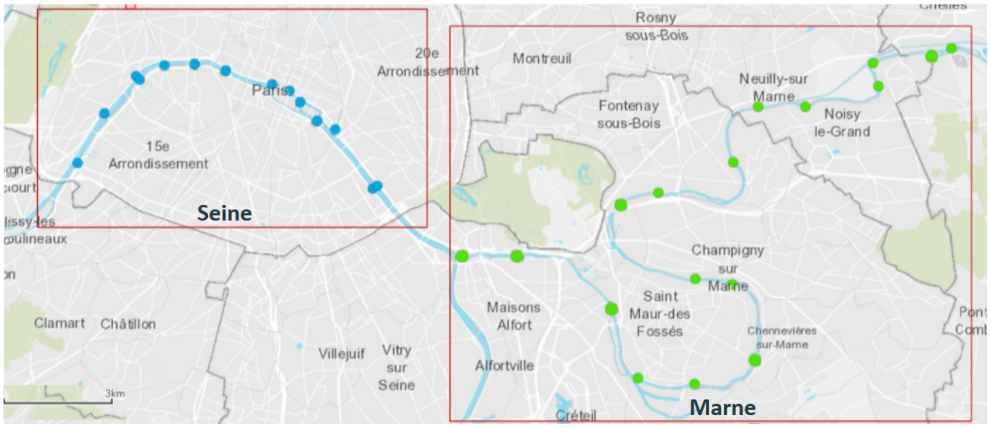

FIGURE 2.1 – Stations de surveillance de la qualité de l'eau de la rivière en Marne (étoiles vertes) et en Seine (étoiles bleues).

Plusieurs sites de suivi de la qualité des eaux de surface de la Marne (Syndicat Marne Vive) et de la Seine (Ville de Paris) ont été retenus, leurs données étant regroupées dans la base de données CapGeo de la Ville de Paris (Figure 2.1). Cette base de données est issue de l'activité du groupe de travail "Amélioration de la connaissance de la qualité microbiologique de la Seine et de la Marne" qui est piloté par la Ville de Paris et elle n'est pas en open source pour l'instant.

À partir de la base de données CapGeo, les stations de suivi pour lesquelles à la fois les données microbiologiques, physico-chimiques et météorologiques étaient disponibles sur 5 ou 6 ans ont été sélectionnées. Pour l'ensemble de ces stations, le protocole d'échantillonnage des eaux de surface a été réalisé selon la norme française FD T 90-523-1 (2008) pour les paramètres physico-chimiques et selon la directive 2006/7/CE pour les concentrations en *E. coli*. Huit paramètres physico-chimiques et microbiologiques sont communs pour les deux jeux de données (Seine et Marne). Pour chaque point de prélèvement, les données météorologiques de la station de mesure la plus proche ont été utilisées.

3.2.1.1. La Marne

De mi-juin à mi-septembre pendant 5 ans (2015, 2017-2020), des prélèvements ont été effectués de manière hebdomadaire ou bi-hebdomadaire sur 18 stations situées dans la section aval de la rivière Marne (France). Pour chaque site d'échantillonnage, les paramètres suivants ont été mesurés : concentrations en *E. coli* (NPP/100 mL), température (°C), turbidité (FNU), conductivité (µS/cm), MES (mg/L), NH_4^+ (mg de N/L), NTK (mg de N /L), nombre de jours secs après

la dernière pluie (jours), pluviométrie du jour cumulée sur 24 h (mm), pluviométrie de la veille cumulée sur 24 h (mm) et le débit (m^3/s),mesuré à Gournay-sur-Marne (station hydrométrique F664 0001 04, (`https://www.hydro.eaufrance.fr`). Les mesures ont été réalisées selon les méthodes normalisées françaises. Les pluviomètres (station CHAM23, MAIS32) du conseil départemental du Val-de-Marne, (station NE-17) du conseil départemental des Seine-Saint-Denis et (station PL14) de la Ville de Paris ont fourni les données de pluviométrie.

3.2.1.2. La Seine

De début-juin à fin-septembre pendant 6 ans (2015-2020), des prélèvements ont été effectués de manière hebdomadaire ou bi-hebdomadaire sur 14 stations de la rivière Seine (France). Pour chaque site d'échantillonnage, les paramètres suivants ont été mesurés : concentrations en *E. coli* (NPP/100 mL), température (°C), turbidité (FNU), conductivité (µS/cm), nombre de jours secs après la dernière pluie (jours), pluviométrie du jour cumulée sur 24 h (mm), pluviométrie de la veille cumulée sur 24 h (mm) et le débit (m^3/s) mesuré à Austerlitz (Station hydrométrique F700 0001 03, (`https://www.hydro.eaufrance.fr`). Les mesures microbiologiques et physico-chimiques ont été réalisées par Eau de Paris selon les méthodes normalisées françaises NF EN ISO 9308-3, NF EN ISO 7027-1, NF EN 27888. Les données pluviométriques ont été obtenues à partir du réseau de pluviomètres de la Ville de Paris (stations PL1, PL4, PL5).

3.2.2. Préparation des données

Après le nettoyage des données qui a consisté à supprimer les entrées avec des valeurs manquantes, un total de 1696 mesures a été obtenu pour le jeu de données de la Marne et un total de 985 mesures pour la Seine. Les modélisations ont été réalisées séparément sur les deux jeux de données (Marne et Seine). L'ID de la station (ordonnée d'amont en aval) et les paramètres physico-chimiques et météorologiques en Seine (8 paramètres) et en Marne (11 paramètres), ont été utilisés comme entrées des modèles. La sortie des modèles était la concentration en *E. coli* prédite. Chaque jeu de données (Marne et Seine) a été divisé aléatoirement en deux parties, le jeu de données d'entraînement (90%, 1526 observations pour la Marne et 886 observations pour la Seine) et le jeu de données test (10%, 170 observations pour la Marne et 99 observations pour la Seine).

Afin que tous les paramètres d'entrée aient le même degré d'influence sur les résultats finaux, nous avons effectué une normalisation Z-score pour chaque caractéristique de l'ensemble des données (entrées et sorties) (Chen et al., 2020). Le jeu d'entraînement a été utilisé pour la

standardisation afin de bloquer l'accès aux valeurs des données du jeu test pendant l'entraînement des modèles.

3.2.3. Les modèles d'apprentissage automatique

Afin d'évaluer la performance de l'estimation de la concentration en *E. coli* par les modèles d'apprentissage automatique, trois modèles traditionnels d'apprentissage automatique (KNN, DT et SVM) et trois modèles d'apprentissage ensemblistes (Bagging, RF et AdaBoost), qui combinent plusieurs modèles de base, ont été sélectionnés et utilisés dans cette étude.

Tous les modèles ont été réalisés avec les packages Scikit-learn comme décrit dans notre étude précédente (Naloufi et al., 2021). La technique GridSearchCV a été appliquée pour sélectionner l'hyperparamètre qui donne le modèle le plus optimal par validation croisée 5 fois, sur une grille de paramètres. Dix répliques ont été générées en divisant aléatoirement les ensembles de données pour générer 10 jeux d'entraînement et de test différents. Les modèles ont été formés avec les 10 jeux d'entrainement, puis chaque modèle a été testé avec un des 10 jeux test.

Après l'entraînement des 6 modèles, en vue de sélectionner le modèle le plus performant pour prédire la concentration en *E. coli*, une phase de test a été réalisée avec les 10 jeux aléatoires de données test (Figure 2.2).

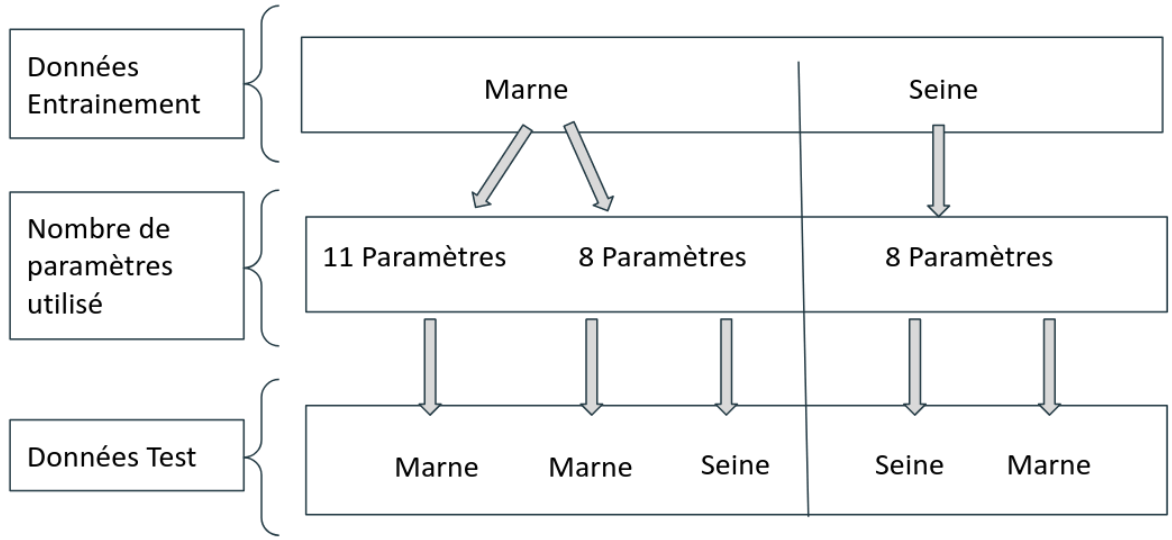

FIGURE 2.2 – Schéma récapitulatif de la stratégie utilisée pour l'entrainement et le test de l'ensemble des modèles sur les données de la Marne et de la Seine.

Pour la base de données de la Marne, l'entrainement à été réalisé sur 10 jeux aléatoires

avec les 11 paramètres, puis avec uniquement les 8 paramètres physico-chimiques et hydro-météorologiques en commun avec la base de données de la Seine (ID de la station, température, conductivité, turbidité, nombre de jours de temps secs après la dernière pluie, pluviométrie du jour cumulée sur 24 h, pluviométrie de la veille cumulée sur 24 h et le débit de la rivière).

3.2.4. Apprentissage par transfert

Afin d'évaluer l'approche d'apprentissage par transfert, dix tests aléatoires ont été réalisés pour évaluer les performances des modèles entraînés. D'une part, des modèles ont été entraînés sur les données de la Marne avant d'être testés sur celles de la Seine, et d'autre part, l'inverse a été effectué, avec des modèles entraînés sur la Seine puis appliqués à la Marne. Seuls les huit paramètres communs entre les jeux de données de la Seine et de la Marne ont été utilisés afin de garantir la comparabilité et de maximiser la transférabilité des modèles.

3.2.5. Evaluation des modèles

Les tests et les calculs de métriques d'erreur et de performance des modèles ont été réalisés alternativement sur chacun des deux jeux de données (Marne et Seine) comme explicité dans la figure 2.2 présentant la stratégie employée.

Les performances de prédiction de chaque modèle pendant les 10 essais aléatoires ont été évaluées par quatre mesures statistiques. Il s'agit de l'RMSE, de l'MAE, du RPD et du MAPE (Naloufi et al., 2021). Ces métriques sont calculées comme suit :

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(yi - y'i)^2}\,(1)$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|yi - y'i|\,(2)$$

$$RPD = \frac{SD}{RMSE}\,(3)$$

Dans ces formules, (y_i) est la valeur mesurée, (y_i') est la valeur prédite, (N) est le nombre total d'échantillons, et (SD) est l'écart-type de l'ensemble des données testées. Plus le RMSE ou le MAE sont petits, plus la capacité de prédiction du modèle est stable. Des valeurs de RPD <1.4 indiquent que le modèle n'est pas fiable. Pour des valeurs de RPD comprises entre 1,4 et 2,0, le modèle est modérément précis et lorsque la valeur est supérieure à 2,0, le modèle présente

un niveau élevé de capacité prédictive (Wang et al., 2017; Naloufi et al., 2021). Le pourcentage d'erreur absolue moyenne (MAPE), qui mesure la qualité de l'ajustement, a également été appliqué.

$$MAPE = \frac{|yi - y'i|}{yi} * 100 \, (4)$$

Plus la valeur MAPE est faible, plus la prédiction est précise (Lu and Ma, 2020). Les valeurs <50% peuvent être évaluées comme "raisonnables", voire bonnes si elles sont <20%. Les valeurs MAPE supérieures à 50%, indiquent une prédiction "inexacte". Une valeur MAPE de 50% indique une surestimation ou une sous-estimation de 50% par rapport à la valeur mesurée (Naloufi et al., 2021).

3.2.6. Identification des points faibles du jeu de données

Les valeurs MAPE calculées au cours des 10 essais aléatoires ont été utilisées pour séparer en deux les valeurs prédites : les estimations raisonnables de la concentration en *E. coli* et les estimations inexactes. L'analyse a été effectuée sur la base des prédictions générées par le meilleur modèle et cela sur les deux jeux de données (Marne et Seine).

Afin de déterminer les paramètres physico-chimiques et météorologiques qui ont potentiellement influencé la capacité prédictive du meilleur modèle, une analyse de corrélation a été réalisée pour les deux jeux de données (raisonnable et inexact). Sachant que les concentrations prédites en *E. coli* n'avaient pas une distribution normale, un test du coefficent de corrélation de Spearman a été réalisé entre chaque paramètres physico-chimiques et la concentration en *E. coli* prédite par les modèles ou les valeurs mesurées (R Project V3.5.1, (R-Core-Team, 2018)). Pour tous les tests statistiques, le niveau de signification était basé sur 5% et 1%.

Les résultats ont été utilisés pour identifier l'ensemble de paramètres présentant une différence de corrélation entre les valeurs prédites raisonnables et inexactes. Par la suite, après l'analyse de corrélation, les 10 jeux de test aléatoires ont été fusionnés. Pour chaque variable physico-chimique et météorologique, l'ensemble des valeurs permettant une estimation raisonnable a été identifié et l'ensemble des valeurs donnant une prédiction inexacte a été retiré. Le résultat a été inspecté afin d'identifier les données supplémentaires nécessaires pour améliorer le modèle. Cela nous a permis d'identifier l'ensemble des valeurs qui donnent au moins une prédiction raisonnable ou bonne pour la Seine.

La stratégie utilisée pour sélectionner le meilleur modèle pour la prédiction de la concen-

tration en *E. coli* et pour identifier sur l'ensemble des paramètres les plages de valeurs nécessaires pour optimiser les stratégies d'échantillonnage pour la Marne a été présentée dans notre étude précédente (Naloufi et al., 2021). Les scripts python et R utilisés sont disponibles sur GitHub (`https://github.com/naloufi-manel/ML_qualite_microbiologique_eau.git`).

3.3. Résultats

3.3.1. Jeux de données

Deux bases de données ont été analysées, celle du suivi estival de la Marne et celle de la Seine.

3.3.1.1. La Marne

Le jeu de données de la rivière Marne est caractérisé par une grande hétérogénéité concernant le nombre d'observations par station (24 à 167 entrées pour les 18 stations). L'analyse descriptive des données est détaillée dans l'étude de Naloufi et al. (2021) au niveau de la section 2.

3.3.1.2. La Seine

Le jeu de données de la rivière Seine est caractérisé par une grande hétérogénéité concernant le nombre d'observations par station (15 à 200 entrées pour les 14 stations). La concentration en *E. coli* mesurée au cours des 6 étés dans la rivière varie entre 30 et 35000 NPP/100 mL avec une valeur moyenne de 3434 ± 6987 NPP/100 mL. La distribution comprend plusieurs valeurs extrêmes (Figure 2.3). Concernant les variables physico-chimique, hormis la température, l'ensemble des paramètres présentent une distribution asymétrique avec la présence de nombreuses valeurs extrêmes (Figure 2.3).

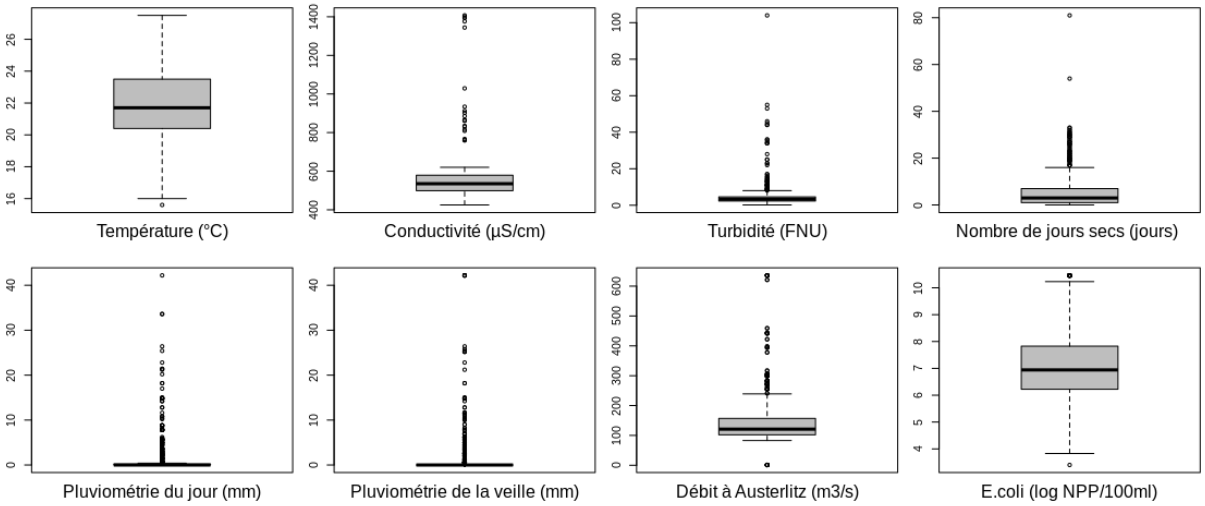

FIGURE 2.3 – Description des données de la Seine pour les paramètres physico-chimiques, pluviométriques et microbiologiques (température en °C, conductivité en µS/cm, turbidité en NTU, nombre de jours secs après la dernière pluie, pluviométrie du jour cumulée sur 24 h en mm, pluviométrie de la veille cumulée sur 24 h en mm, débit au pont d'Austerlitz en m^3/s et le logarithme népérien de la concentration en *E. coli* (NPP/100 mL).

3.3.2. Prédiction de la concentration en *E. coli* par les modèles par apprentissage automatique

Dans cette étude, nous avons comparé la performance de six algorithmes basés sur l'apprentissage automatique (KNN, DT, SVM, Bagging, RF et AdaBoost) pour prédire la concentration en *E. coli* dans deux rivières urbaines (Marne et Seine). Afin d'identifier le modèle le mieux adapté, nous avons analysé l'erreur test de chaque modèle entrainé sur 10 jeux de données aléatoires, en utilisant les métriques RMSE, MAE et RPD pour évaluer la performance de chaque modèle.

3.3.2.1. Effet du nombre de paramètres sur les performances des modèles en Marne

Les six modèles ont été entrainés et testés avec les données de la Marne (avec 11 paramètres et 8 paramètres). Les valeurs moyennes des métriques calculées pour chaque essai aléatoire sont présentées au niveau du tableau S1. Lorsque les modèles sont entrainés avec les 11 paramètres, on observe que pour le modèle RF, la valeur RPD était proche de 2 (1,91 ± 1,65), ce qui indique que le modèle avait une capacité de prédiction élevée. Une description plus détaillée des résultats est présentée au niveau de la section 2. Par la suite, les modèles ont été entrainés avec 8 paramètres, le modèle AdaBoost présenté le pouvoir de prédiction le plus élevé, avec l'erreur la plus faible (valeur moyenne de 0.52 ± 0.29 pour la RMSE et 0,11 ± 0,02 pour la MAE) (Figure 2.4), suivi par les modèles RF et SVM (respectivement 0,58 ± 0,33 et

0,59 ± 0,35 pour la RMSE et 0,15 ± 0,04 et 0,14 ± 0,03 pour la MAE). En effet, en retirant 3 paramètres du modèle, aucun des modèles testés ne pouvait être considéré comme fiable (RPD <1,4, Tableau S1).

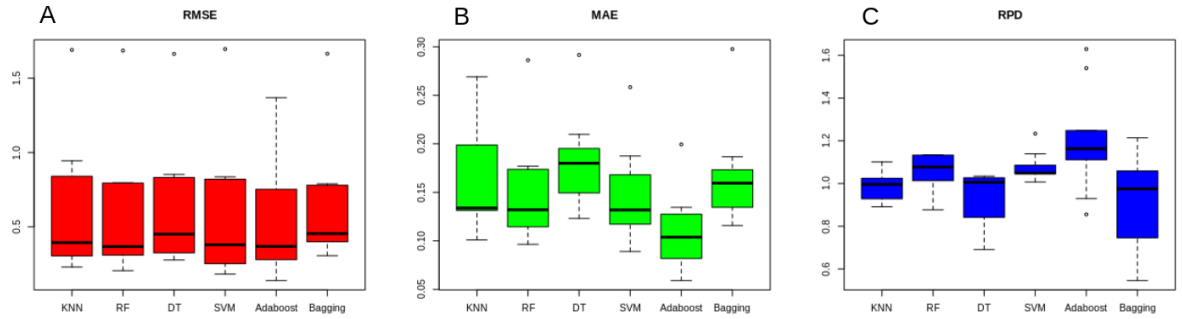

FIGURE 2.4 – Évaluation des performances de prédiction des 6 modèles d'apprentissage automatique au cours des 10 essais avec 8 paramètres issus de la base de données de la Marne. Métriques statistiques : (A) RMSE ; (B) MAE ; (C) RPD.

3.3.2.2. Comparaison des performances de prédictions de *E. coli* avec les données de la Seine

Aprés entrainement et test des 6 modèles sur les données de la Seine (Tableau S2), les valeurs RMSE et RPD indiquaient que les modèles RF et Bagging étaient les plus performants et pouvaient être considérés comme modérément précis et fiables (présentant des résultats acceptables). L'analyse de l'indice MAE a montré que le modèle Bagging présentait l'erreur la plus faible suivi par les modèles SVM et RF (Figure 2.5). Le modèle KNN a été estimé également comme modérément précis avec des résultats acceptables (Tableau S2). Ainsi, le modèle RF semble donner la meilleure estimation de la concentration en *E. coli* avec un rapport de performance le plus élevé pour les données de la Seine.

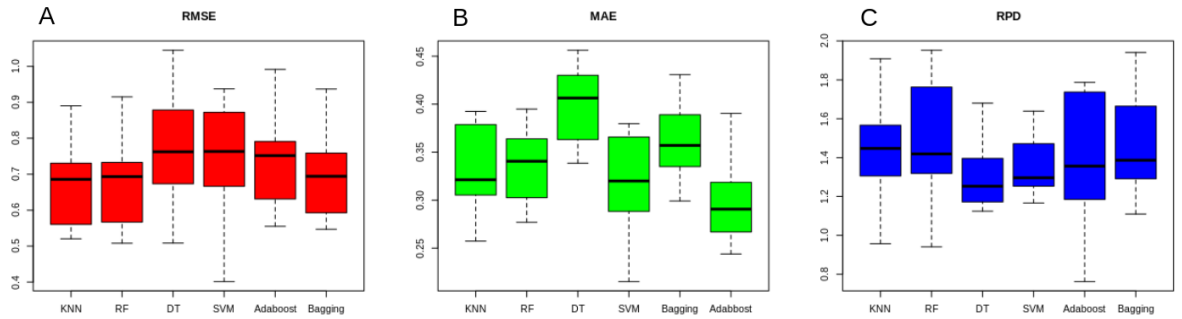

FIGURE 2.5 – Évaluation des performances de prédiction des 6 modèles d'apprentissage automatique au cours des 10 essais avec les données de la Seine. Métriques statistiques : (A) RMSE ; (B) MAE ; (C) RPD.

3.3.3. Apprentissage par transfert

Étant donné que la base de données de la Seine présente moins de données que celle de la Marne, nous avons testé l'approche d'apprentissage par transfert entre les 2 bases de données de bassins versants de la même région géographique. Les différents modèles ont tout d'abord été entrainés en utilisant le jeu d'entrainement de la Seine, puis testés sur le jeu de données test de la Marne. Nous avons également évalué un entrainement des modèles sur les données de la Marne suivi par un test sur le jeu de données de la Seine. Pour cela nous avons utilisé pour la Marne la base de données avec 8 paramètres.

3.3.3.1. Evaluation de l'apprentissage par transfert pour prédire les concentrations en Marne

Aprés entrainement des modèles sur les données de la Seine, des tests ont été effectués avec les jeux de données de la Marne (Figure 2.6). Les valeurs moyennes des métriques calculées sont présentées dans le tableau S3. Les modèles présentaient des performances moyennes similaires (RPD entre 0,98 et 1,00).

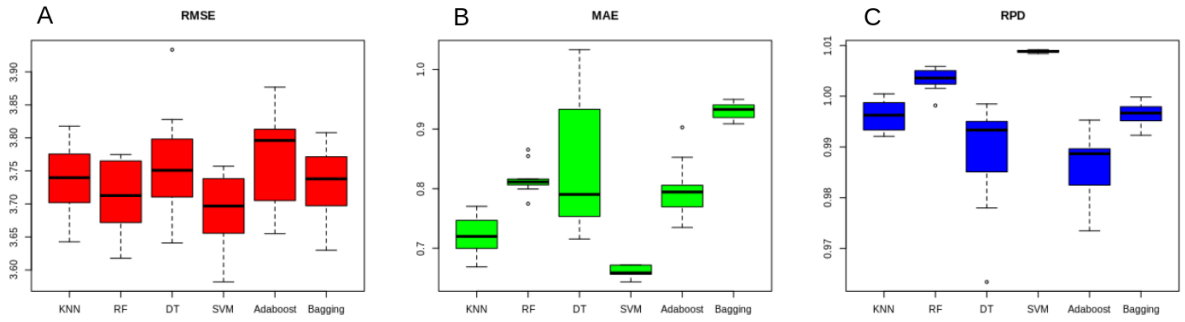

FIGURE 2.6 – Évaluation des performances de prédiction des 6 modèles d'apprentissage automatique au cours des 10 essais avec les données de la Marne avec 8 paramètres, après entrainement avec les données de la Seine. Métriques statistiques : (A) RMSE ; (B) MAE ; (C) RPD.

3.3.3.2. Evaluation de l'apprentissage par transfert pour prédire les concentrations en Seine

Sur les modèles entraînés par le jeu de données Marne, des tests ont été effectués avec les jeux de données de la Seine. Les valeurs moyennes des métriques calculées pour chaque essai aléatoire sont présentées dans le tableau S4. Les modèles SVM, Adaboost et RF présentaient de meilleures performances de prédiction par rapport aux autres modèles d'apprentissage (Figure 2.7). Toutefois, aucun de ces modèles ne semblait fiable (RPD < 1.4, Tableau S4).

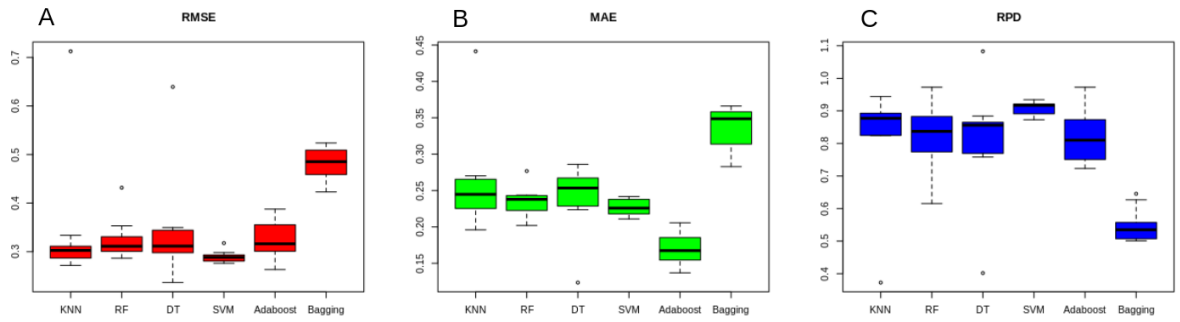

FIGURE 2.7 – Évaluation des performances de prédiction des 6 modèles d'apprentissage automatique au cours des 10 essais en Seine avec 8 paramètres après entrainement avec les données de la Marne. Métriques statistiques : (A) RMSE ; (B) MAE ; (C) RPD.

Ces résultats montrent que le jeu de données d'entraînement détermine la capacité de prédiction au sein de chaque bassin versant. Par exemple, les modèles entraînés avec les données de la Seine avaient une meilleure capacité prédictive pour la Seine et de même les modèles entraînés avec les données de la Marne avaient une meilleure capacité prédictive pour la Marne (Tableau S1 et S2). L'ensemble de ces résultats a été utilisé afin de sélectionner le meilleur modèle pour prédire la concentration en *E. coli* en Seine et en Marne. C'est le modèle RF, que se soit en Seine ou en Marne, qui a été le plus performant (entrainé sans apprentissage par transfert). Ce dernier a été sélectionné pour une analyse plus détaillée des performances de de prédiction. Suite à notre étude précédente sur les limites de l'estimation de la concentration en *E. coli* en Marne (Naloufi et al., 2021), cette étude se focalise sur les limites des modèles d'estimation en Seine.

3.3.4. Limites de l'estimation de la concentration en *E. coli* basée sur le modèle RF de la Seine

L'identification des observations avec des prédictions incertaines est une approche permettant de déterminer l'ensemble des données nécessitant une optimisation et donc de trouver un moyen d'optimiser la collecte et de compléter efficacement le jeu d'entraînement, permettant une meilleure prédiction dans le futur en réexécutant le modèle avec les mesures complémentaires nouvellement collectées de manière ciblée.

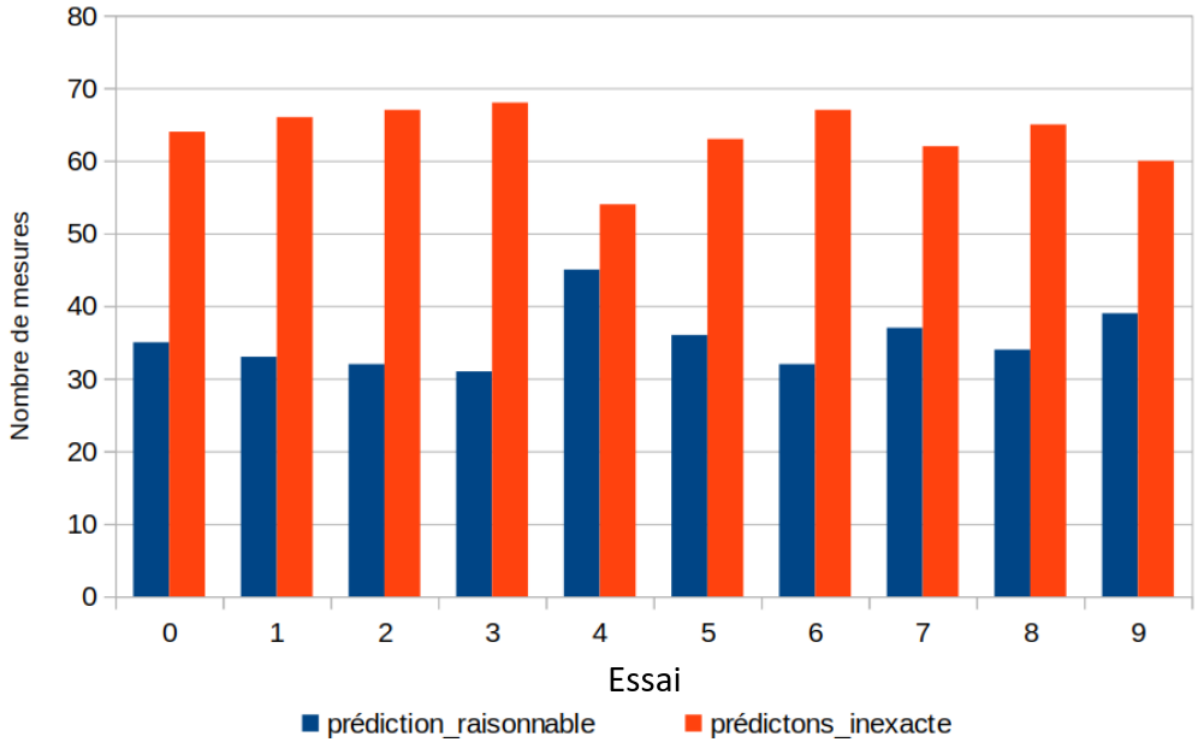

FIGURE 2.8 – Nombre d'observations identifiées comme des estimations raisonnables ou inexactes selon les valeurs MAPE obtenues avec le modèle RF au cours des dix essais sur les données de la Seine.

L'indice MAPE, qui mesure la qualité de l'ajustement et examine la performance des modèles en fonction de leur tendance à estimer la concentration en *E. coli*, a été calculé pour tous les essais effectués avec les deux modèles RF respectifs pour les données de la Marne ou de la Seine . En se basant sur la valeur MAPE calculée pour chaque observation, une distinction a été effectuée entre les estimations raisonnables (<50%) et les estimations inexactes (>=50%), permettant de séparer les données prédites en deux selon ces catégories, respectivement pour la Marne et pour la Seine.

Pour les données de la Seine, sur la base des 10 essais, 35.75 ± 3.11% des concentrations prédites en *E. coli* générées par le modèle RF correspondent à des estimations raisonnables. Par contre 64,25 ± 3,11% des valeurs ont été identifiées comme des estimations inexactes, le plus souvent surestimées (Figure 2.8).

En effet, la figure 2.9 indique une incertitude de la prédiction pour certaines des mesures de *E. coli* plus élevées que pour le modèle RF de la Marne (Figure 2.5). Ainsi, les limites de prédiction du modèle RF ont été examinées plus en détail afin d'identifier les faiblesses dans le jeu de données de la Seine.

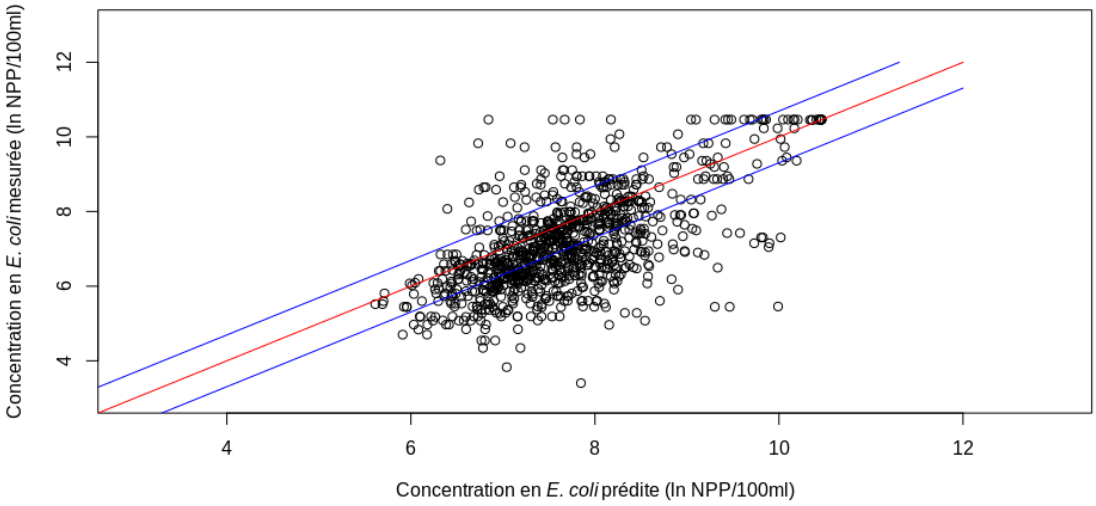

FIGURE 2.9 – Relation entre la concentration en *E. coli* (NPP/100 mL) prédite par le modèle RF et la concentration mesurée en Seine. Les cercles noirs indiquent les valeurs. La ligne rouge indique les valeurs théoriques correspondant à une prédiction exacte du modèle par rapport aux valeurs mesurées pour les dix essais et les courbes bleues indiquent l'intervalle d'incertitude de 50% autour de la valeur exacte de prédiction.

Ensuite les paramètres permettant une estimation raisonnable des concentrations d'*E. coli* et ceux conduisant à une estimation inexacte ont été identifiés pour ce jeu de données de la Seine. La relation entre les différentes variables explicatives et les concentrations prédites d'*E. coli* a été explorée sur les données test de la Seine.

TABLE 2.1 – Comparaison des coefficients de corrélation (moyenne ± écart-type) obtenus entre les variables prédictives et les concentrations en *E. coli* (NPP/100 mL) raisonnablement prédites par le modèle RF entrainé et testé sur les données de la Seine et celles pour lesquelles la prédiction est inexacte. Les p-valeurs des tests statistiques comparant les valeurs de corrélation entre les prédictions raisonnables et inexactes sont données.

Paramètres	**Prédictions raisonnables**	**Prédictions inexactes**	**p-valeur**
Température	-0,51 ± 0,06	-0,37 ± 0,06	0,002
Conductivité	-0,45 ± 0,10	-0,26 ± 0,08	0,002
Turbidité	0,39 ± 0,15	0,28 ± 0,09	0.173
Nombre de jours secs	-0,45 ± 0,16	-0,32 ± 0,09	0,139
Pluviométrie du jour	0,44 ± 0,15	0,27 ± 0,09	0,01
Pluviométrie de la veille	0,56 ± 0,08	0,45 ± 0,08	0,10
Débit à Austerlitz	0,26 ± 0,10	0,33 ± 0,08	0,18

Pour le jeu de données de la Seine, la pluviométrie de la veille cumulée sur 24 h semble un prédicteur important avec un coefficient de corrélation élevé et aucune différence significative

n'a été observée entre les prédictions raisonnables et les prédictions inexactes (Test de Wilcoxon, n=10, p>0,05, Tableau 2.1). La température et la conductivité semblent également être de bons prédicteurs, mais il y avait une différence significative entre les deux ensembles de données (Test-t, n=10, p<0,01, Tableau 2.1). Ces deux paramètres ont été classés comme ayant un impact sur la prédiction mais nécessitant des données supplémentaires. L'ensemble des paramètres restants ont approximativement les mêmes niveaux de corrélation à l'exception du débit au Pont d'Austerlitz qui présente le niveau de corrélation le plus faible (r=0,26 et 0,33, Tableau 2.1). Aucune différence significative n'étant observée entre les données raisonnablement prédites et celles inexactes pour le débit de la Seine, ce prédicteur reste néanmoins intéressant.

Par la suite, une exploration des deux paramètres présentant une différence de corrélation hautement significative entre les valeurs raisonnables et celles inexactes a permis d'identifier l'ensemble des mesures permettant une estimation raisonnable. En ce qui concerne la température, les valeurs bien prédites sont disséminées le long de la plage de données [15,60-27,50 °C], mais nécessite principalement une optimisation dans l'intervalle de valeur [19-24°C] (Figure 2.10). Pour la conductivité, les valeurs bien prédites couvrent le début de la plage de valeurs (Figure 2.10). Cependant, un grand intervalle de valeur [606-1393 µS/cm] nécessite des échantillonnages supplémentaires.

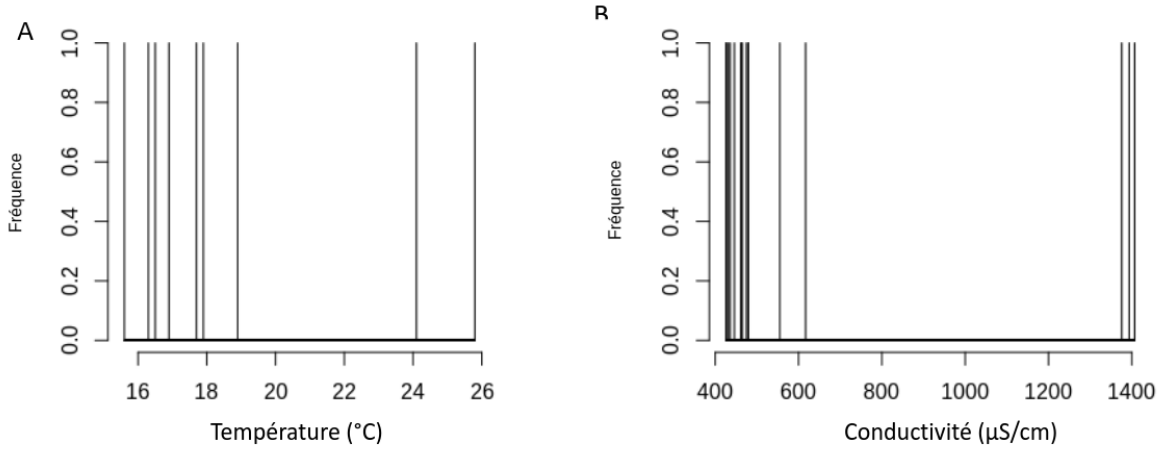

FIGURE 2.10 – Valeurs des paramètres donnant une estimation raisonnable des concentrations en *E. coli* (NPP/100 mL) dans la plage de valeurs des prédicteurs pour l'ensemble de données mesurées en Seine pour (A) la température ; (B) la conductivité.

3.4. Discussion

E. coli est un des deux indicateurs biologiques utilisés pour évaluer la qualité de l'eau de baignade (directive européenne sur les baignades 2006/7/CE) car il peut caractériser les risques de contamination fécale de l'eau testée et sa présence est relativement bien corrélée avec le risque de contracter une gastro-entérite (van Asperen et al., 1998). Ainsi dans le cadre de l'ouverture de site de baignades en ville, disposer d'outils pour prédire la qualité microbiologique est nécessaire pour la surveillance quotidienne de la qualité de l'eau des rivières urbaines (OMS, 2018). Dans ce travail de thèse, différents modèles d'apprentissage automatique ont été utilisés pour prédire la concentration en *E. coli* dans deux rivières urbaines et leurs performances prédictives ont été comparées et évaluées. Nous avons également proposé une méthode d'analyse détaillée des jeux de données afin d'évaluer les faiblesses à renforcer pour améliorer la prédiction des modèles.

3.4.1. Comparaison des méthodes d'apprentissage automatique

Plusieurs études antérieures ont utilisé des modèles d'apprentissage automatique pour prédire la qualité des eaux de surface à l'aide des paramètres physico-chimiques (Di et al., 2019; Avila et al., 2018; van der Meulen et al., 2024). Leur performance prédictive a été comparée à des modèles statistiques et/ou déterministes en évaluant leur capacité de prédiction, montrant ainsi tout leur intérêt et notamment leur grande capacité à prédire de manière fiable dans différentes configurations de sites (par exemple Mälzer et al. (2016); Avila et al. (2018); Bui et al. (2020)). Les modèles d'apprentissage automatique ont la capacité d'identifier des motifs ou structures et des relations non-linéaires parmi les variables utilisées, ce qui explique le fait qu'ils ont souvent des performances meilleures que les modèles de régression linéaire (Nafsin and Li, 2023). Dans le cadre de notre étude, six algorithmes d'apprentissage automatique (KNN, DT, SVM, Bagging, RF et AdaBoost) ont été utilisés pour prédire la concentration d'*E. coli* dans la Marne et la Seine. Les performances de ces modèles ont été comparées. De meilleures performances de prédiction et une plus grande fiabilité ont été observées lorsque les modèles ont été entraînés sur un jeu de données provenant de la même rivière. Ainsi, au niveau de la Seine, le modèle basé sur la méthode RF et entraîné avec les données de la Seine a donné une meilleure estimation de la concentration en *E. coli* que lorsqu'il est entrainé avec le jeu de données d'une autre rivière. La même conclusion a été observée sur les données de la Marne. Ceci confirme les résultats de Mälzer et al. (2016) qui ont constaté que la performance d'un même

modèle d'apprentissage automatique pouvait différer d'un site à l'autre le long d'une rivière. En effet, pour chaque site, des interactions complexes entre des facteurs physico-chimiques, des facteurs hydrométéorologiques et des caractéristiques géospatiales telles que l'usage des sols vont déterminer la dynamique des concentrations en BIF (Nafsin and Li, 2023).

Globalement nos résultats montrent que les modèles d'apprentissage ensemblistes ont une meilleure performance par rapport aux modèles traditionnels (Naloufi et al., 2021). Ceci est en accord avec la littérature, avec plusieurs études qui montrent que les méthodes d'apprentissage d'ensembles tels que le bagging ou le boosting ont souvent une bonne capacité à prédire tout en étant précis (Ahmed et al., 2019a; Bui et al., 2020; Qiu et al., 2017). Pour nos deux bases de données, le modèle RF a donné une meilleure estimation de la concentration en *E. coli* par rapport aux autres modèles d'apprentissage automatique. Ce résultat est en accord avec notre précédente étude Naloufi et al. (2021) et avec plusieurs autres études (e.g. Bui et al. (2020); Choi and Seo (2018); Sokolova et al. (2022); Iyer (2024); Weller et al. (2020)) qui ont identifié les modèles RF et Bagging comme ayant les meilleures performances pour prédire les concentrations en BIF dans les eaux de surface des rivières. Si les modèles RF ont souvent les meilleures performance, ils ont aussi tendance au sur-apprentissage (Sokolova et al., 2022). De plus, les modèles ensemblistes sont l'objet d'un compromis entre l'interprétabilité et la précision, car ils se présentent plus comme des "boites noires" comparés aux modèles basés sur les arbres de décision (Weller et al., 2020). Il reste néanmoins difficile de comparer nos résultats avec ceux de la littérature car le nombre et le type de paramètres diffèrent d'une étude à l'autre, de même que les conditions climatiques, l'usage des sols et l'urbanisation.

Nous avons également pu constater avec le jeu de données de la Marne qu'avoir des paramètres supplémentaires en entrée du modèle permettait une amélioration de la performance de la prédiction. Ceci est en concordance avec l'étude de Chen et al. (2020) qui a également constaté une diminution de la performance des modèles en enlevant un ou deux paramètres lors de l'entraînement. La sélection d'une combinaison optimale de variables et l'optimisation des paramètres clef du modèle font d'ailleurs partie de stratégies pour augmenter la performance d'un modèle d'apprentissage automatique (Nafsin and Li, 2023).

En conclusion, pour nos deux bases de données (Seine et Marne), le modèle RF a donné une meilleure estimation de la concentration en *E. coli* par rapport aux autres modèles d'apprentissage automatique. Nos résultats confirment que les modèles d'apprentissage ensemblistes ont une meilleure performance par rapport aux modèles traditionnels (par exemple DT et SVM)

(Naloufi et al., 2021). De plus, les modèles RF ont produit une modélisation polyvalente en matière de prédiction. Ces données sont précieuses car il y a encore peu d'études utilisant l'apprentissage automatique pour prédire les concentrations en *E. coli* dans les rivières urbaines, ce qui limite notre compréhension de leur capacité à prédire les pics de contaminations en BIF (van der Meulen et al., 2024).

3.4.2. Incertitude sur la prédiction des modèles RF

En plus de la structure du modèle, nous nous sommes intéressés à la précision de la prédiction en fonction de la qualité et de la taille des jeux de données test. En dépit de la performance du modèle RF, une grande incertitude dans la prédiction peut être observée que ce soit pour la Marne ou pour la Seine. Il a été établi que la qualité des eaux de surface dépend de multiples conditions. Cependant, comme nous l'avons observé dans cette étude, une grande variabilité de la distribution des paramètres physico-chimiques et hydrologiques et des concentrations en *E. coli* en raison des faibles quantités de données d'apprentissage peuvent conduire à une faible précision des modèles d'apprentissage automatique (Bui et al., 2020; Naloufi et al., 2021; Nafsin and Li, 2023). Pour améliorer la capacité de prédiction des modèles d'apprentissage automatique, ce n'est pas seulement la taille du jeu de données qu'il faut augmenter mais aussi sa diversité. Dans le cas de notre modélisation des concentrations en *E. coli* par la méthode RF, une part de l'incertitude sur la prédiction pourrait avoir pour origine le fait que toute la gamme des mesures des différents paramètres explicatifs n'a pas encore été testée et que l'on ne sait pas encore si le modèle RF serait capable d'estimer raisonnablement la concentration en *E. coli*, avec toute l'étendue des valeurs que peuvent prendre les variables explicatives. Une explication serait que la distribution des données était asymétrique et que la corrélation de certains paramètres physico-chimiques et/ou hydrométéorologiques avec la concentration en *E. coli* était faible dans ces jeux de données (Naloufi et al., 2021). Cela était probablement dû à la complexité spatiale des processus dans chacun des bassins versants et aux différentes sources de pollution qui entraînaient des relations non linéaires entre les paramètres de l'eau et la concentration en *E. coli* (Bui et al., 2020). En effet, le tableau S5 montre bien des niveaux de corrélation différentes entre la Marne et la Seine pour les différents paramètres de l'eau avec les concentrations en *E. coli* prédites raisonnablement. De plus, pour plusieurs variables, certaines classes de données étaient minoritaires (peu de données collectées sur les événements polluants extrêmes par exemple) comparées à des classes majoritaires (conditions de

temps sec, faibles pluies par exemple). Ce déséquilibre peut présenter un défi pour les algorithmes d'apprentissage automatique qui tendent à être biaisés vers les classes de données majoritaires et ainsi présentent de faibles capacités prédictives sur les classes de données minoritaires (Li et al., 2021). Pour contourner le problème des données de qualité et quantité insuffisantes pour entrainer les modèles, trois stratégies peuvent être employées : i) la génération artificielle de données, ii) l'apprentissage par transfert et iii) la réduction des besoins du modèle en sélectionnant les variables explicatives utilisées (Wu et al., 2024). Toutefois ces approches présentent des limitations comme l'a montré le transfert de connaissance entre la Seine et la Marne. Un transfert de connaissance pauvre peut être dû à une similarité limitée entre les deux rivières malgré leur appartenance à la même hydroécorégion, leur taille et leur débit diffèrent, la Marne étant un affluent de la Seine (Elbaz-Poulichet et al., 2006). La solution qui consiste à générer des données synthétiques doit maintenir les caractéristiques réalistes d'un jeu de données réel, et cette approche ne peut pas simuler de nouvelles conditions *in situ* (Wu et al., 2024). Acquérir des données réelles par des échantillonnage plus fréquents peut être une solution pour augmenter la base de données. Toutefois cette approche est coûteuse et laborieuse. Nous avons donc exploré une autre stratégie qui consiste à utiliser les résultats du modèle pour identifier les classes minoritaires dans le jeu de données existant et ainsi rationaliser l'effort d'échantillonnage pour renforcer ces classes minoritaires et limiter le coût et l'effort de collecte. Optimiser le processus d'échantillonnage permettrait d'obtenir une meilleure représentation de l'ensemble des valeurs possibles sur le site évalué. L'approche que nous avons développée a permis d'identifier les paramètres sur lesquels le focus devrait être porté, ainsi que les classes de données minoritaires à acquérir ou renforcer dans le jeu de données. Ainsi, pour le cas de la Marne et celui de la Seine, la température de l'eau et la conductivité ont été identifiées comme étant les paramètres nécessitant des mesures supplémentaires. Par contre, pour la pluviométrie du jour cumulée sur 24 h et le débit de la rivière, il a été considéré que des données supplémentaires étaient nécessaires uniquement pour le jeu de données de la Marne. En effet, les jeux de données acquis pour le suivi des sites de baignade sont souvent limités à la saison estivale, et à des mesures au mieux journalière mais le plus souvent hebdomadaires. De ce fait, l'étendue des mesures de température fluctue peu, et pour les autres paramètres certains événements générant des variations intenses peuvent ne pas avoir été échantillonnés (accidents sur le réseau, pluies extrêmes, crues). Des capteurs en temps réel peuvent aider à augmenter le jeu de données. Toutefois certaines classes de données minoritaires peuvent être difficiles à renforcer avec des mesures sur le terrain. Par

exemple dans le cas de la pluviométrie, autant il est aisé de se procurer des données temporelles avec un pas de temps fin (5 min), il est plus compliqué d'acquérir des données avec une résolution spatiale fine car le maillage des pluviomètres est relativement large sur le territoire Francilien. De plus, l'intensité des précipitations n'est pas régulière en Ile-de-France, les pluies < 5 mm étant plus couramment observées que les pluies >10 mm ou encore les pluies > 20 mm qui sont plus exceptionnelles (Lucas et al., 2020). La stratégie de l'augmentation des données par génération de données synthétiques peut alors constituer une solution pour les données difficiles ou impossibles à acquérir (Wu et al., 2024).

3.5. Conclusion

Dans cette étude, nous avons discuté d'un modèle basé sur l'apprentissage automatique pour la prédiction afin d'évaluer la qualité de l'eau dans des deux rivières franciliennes. D'après les résultats, les modèles basés sur la méthode Random Forest ont donné la meilleure précision dans la prédiction de la concentration en *E. coli* (RMSE de 0,37 ± 0,20 en Marne et 0,67 ± 0,09 en Seine). Néanmoins, selon le pourcentage d'erreur absolu moyen (MAPE) permettant de distinguer entre les estimations raisonnables et inexactes, la concentration en *E. coli* ne peut être prédite dans tous les contextes (la valeur de MAPE est supérieur à 50%, avec 53,20 ± 3,50% et 63,25 ± 3,11% de prédictions inexactes respectivement pour la Marne et la Seine). Étant donné que notre jeu de données n'est pas représentatif de toutes les valeurs possibles dans la gamme de données, il est raisonnable de penser que les modèles RF n'ont pas été encore entrainés ou testés avec toute l'étendue des valeurs que peuvent prendre les paramètres prédictifs clefs. Pour ces valeurs, il n'est donc pas encore clair si notre modèle est capable d'estimer la concentration en *E. coli* de manière raisonnable.

Pour augmenter le jeu de données deux stratégies ont été explorées. Le transfert de connaissance, ne s'est pas révélé concluant, conduisant à une performance moindre des modèles (phénomène appelé "transfert négatif"). Cependant nous avons utilisé une approche simple se limitant à utiliser les paramètres communs entre les deux jeux de données. Il existe des approches plus sophistiquées d'apprentissage par transfer qu'il pourrait être intéressant d'explorer (Wu et al., 2024). En effet, il est nécessaire de conduire des recherches sur le transfert négatif et la façon de l'éviter (Wang et al., 2019b). Notre deuxième stratégie a été d'utiliser les résultats du modèle RF pour identifier les paramètres clefs à optimiser en premiers lieu. Cette approche

semble appropriée afin d'augmenter de manière ciblée et rationelle les bases de données. De plus, pour ces paramètres, l'analyse de la distribution des valeurs donnant une prédiction raisonnable le long de la plage de données permettrait d'identifier quelles données minoritaires mal prédites nécessitent d'être renforcées dans la base de données, afin d'obtenir une meilleure efficacité prédictive.

Afin d'améliorer les modèles prédictifs des concentrations en bactéries indicatrices de contamination fécale (*E. coli*, entérocoques intestinaux), l'apprentissage actif permet d'identifier les observations les plus pertinentes, et le déploiement de capteurs à faible coût peut aider à densifier la collecte de données physico-chimiques (qui servent de variables explicatives dans les modèles) en temps réel tout en réduisant les coûts (Bouneffouf, 2016; KnowFLow, 2021). Ces capteurs, bien qu'individuellement moins précis, peuvent ainsi collectivement fournir des informations fiables pour optimiser les bases de données pour les modèles de prédiction.

Acknowledgments : Nous remercions les Conseils départementaux du Val-de-Marne et de la Seine-Saint-Denis, pour leur contribution au jeu de données. Nous sommes reconnaissants envers Miguel Gillon-Ritz pour ses conseils avisés et pour l'accès à la base de données Cap-Geo (Ville de Paris, Direction de la Propreté et de l'Eau - Service Technique de l'Eau et de l'Assainissement).

3.6. Annexe

TABLE S1 – Moyenne et écart-type des mesures statistiques (RMSE, MAE, RPD) obtenues pour chaque modèle au cours des dix essais avec 11 paramétres et 8 paramètres avec les données de la Marne.

Modèle	**KNN**	**RF**	**DT**	**SVM**	**AdaBoost**	**Bagging**
11 paramètres						
RMSE	0,41 ± 0,28	0,37 ± 0,20	0,54 ± 0,29	0,53 ± 0,48	0,53 ± 0,28	0.38 ± 0,19
MAE	0,09 ± 0,03	0,09 ± 0,02	0,14 ± 0,05	0,13 ± 0,05	0,10 ± 0,03	0,14 ± 0,06
RPD	1,60 ± 0,49	1,91 ± 1,65	1,12 ± 0,36	1,32 ± 0,22	1,28 ± 0,62	1,77 ± 1,62
8 paramétres						
RMSE	0,62 ± 0,35	0,58 ± 0,33	0,63 ± 0,32	0,59 ± 0,35	0,52 ± 0,29	0,63 ± 0,28
MAE	0,16 ± 0,04	0,15 ± 0,04	0,18 ± 0,03	0,14 ± 0,03	0,11 ± 0,02	0,16 ± 0,03
RPD	0,98 ± 0,04	1,08 ± 0,06	0,94 ± 0,09	1,07 ± 0,04	1,19 ± 0,16	0,92 ± 0,16

TABLE S2 – Moyenne et écart-type des mesures statistiques (RMSE, MAE, RPD) obtenues avec chaque modèle au cours des dix essais avec les données de la Seine.

Modèle	**KNN**	**RF**	**DT**	**SVM**	**AdaBoost**	**Bagging**
RMSE	0,69 ± 0,09	0,67 ± 0,09	0,77 ± 0,12	0,75 ± 0,12	0,73 ± 0,10	0,68 ± 0,09
MAE	0,34 ± 0,03	0,33 ± 0,02	0,40 ± 0,03	0,32 ± 0,04	0,35 ± 0,02	0,30 ± 0,03
RPD	1,43 ± 0,21	1,47 ± 0,23	1,30 ± 0,13	1,33 ± 0,11	1,37 ± 0,25	1,44 ± 0,18

TABLE S3 – Moyenne et écart-type des mesures statistiques (RMSE, MAE, RPD) obtenues avec chaque modèle au cours des dix essais avec les données de la Marne avec 8 paramètres, après entrainement avec les données de la Seine.

Modèle	**KNN**	**RF**	**DT**	**SVM**	**AdaBoost**	**Bagging**
RMSE	3,73 ± 0,04	3,70 ± 0,04	3,76 ± 0,06	3,68 ± 0,04	3,77 ± 0,06	3,73 ± 0,04
MAE	0,72 ± 0,03	0,81 ± 0,01	0,83 ± 0,09	0,66 ± 0,01	0,79 ± 0,03	0,93 ± 0,01
RPD	0,99 ± 0,02	1,00 ± 0,01	0,98 ± 0,01	1,00 ± 0,01	0,98 ± 0,01	0,99 ± 0,01

TABLE S4 – Moyenne et écart-type des mesures statistiques (RMSE, MAE, RPD) obtenues avec chaque modèle au cours des dix essais avec les données de la Seine avec 8 paramètres après entrainement avec les données de la Marne.

Modèle	**KNN**	**RF**	**DT**	**SVM**	**AdaBoost**	**Bagging**
RMSE	0,33 ± 0,07	0,32 ± 0,02	0,34 ± 0,06	0,28 ± 0,01	0,32 ± 0,03	0,48 ± 0,02
MAE	0,25 ± 0,04	0,23 ± 0,01	0,24 ± 0,02	0,22 ± 0,01	0,17 ± 0,01	0,33 ± 0,02
RPD	0,83 ± 0,09	0,81 ± 0,07	0,81 ± 0,10	0,90 ± 0,01	0,82 ± 0,06	0,54 ± 0,03

TABLE S5 – Comparaison entre les jeux de prédictions raisonnables en Marne et en Seine des coefficients de corrélation (moyenne et écart-types) entre les variables prédictives et les valeurs de concentrations en *E. coli* prédites.

paramètres	**Prédictions raisonnables pour la Marne**	**Prédictions raisonnables pour la Seine**
Température	-0,17 ± 0,05	-0,51 ± 0,06
Conductivité	-0,05 ± 0,11	-0,45 ± 0,10
Turbidité	0,42 ± 0,07	0,39 ± 0,15
MES	0,43 ± 0,09	NA
NH_4^+	0,54 ± 0,06	NA
NTK	-0,03 ± 0,08	NA
Nombre de jours secs	-0,10 ± 0,09	-0,45 ± 0,16
Pluviométrie du jour	0,09 ± 0,10	0,44 ± 0,15
Pluviométrie de la veille	0,17 ± 0,08	0,56 ± 0,08
Débit	0,54 ± 0,09	0,26 ± 0,10

4. Long-Term Stability of Low-Cost IoT System for Monitoring Water Quality in Urban Rivers

Published in : Water (2024)

https://doi.org/10.3390/w16121708

Manel Naloufi [1,2,*], Thiago Abreu [3,*], Sami Souihi [3], Claire Thérial [2], Natália Angelotti de Ponte Rodrigues [2], Arthur Guillot Le Goff [2], Mohamed Saad [2], Brigitte Vinçon-Leite [2], Philippe Dubois [2], Marion Delarbre [1], Paul Kennouche [1] and Françoise S. Lucas [2,*]

[1] Direction de la Propreté et de l'Eau—Service Technique de l'Eau et de l'Assainissement, 27 Rue du Commandeur, 75014 Paris, France ; marion.delarbre@paris.fr (M.D.) ; paul.kennouche@paris.fr (P.K.)

[2] Laboratoire Eau Environnement et Systèmes Urbains (Leesu), Université Paris-Est Créteil, École des Ponts ParisTech, university. 61 Avenue du Général de Gaulle, 94010 Créteil , France ; claire.therial@u-pec.fr (C.T.) ; natalia.angelotti-de-ponte-rodrigues@enpc.fr (N.A.d.P.R.) ; arthur.guillot-legoff@enpc.fr (A.G.L.G.) ; mohamed.saad@enpc.fr (M.S.) ; b.vincon-leite@enpc.fr (B.V.-L.) ; philippe.dubois@enpc.fr (P.D.)

[3] Image, Signal and Intelligent Systems (LiSSi) Laboratory, Université Paris-Est Créteil, 122 Rue Paul Armangot, 94400 Vitry sur Seine, France ; sami.souihi@u-pec.fr

Correspondence : manel.naloufi@gmail.com (M.N.) ; thiago.wanderley-matos-de-abreu@u-pec.fr (T.A.) ; lucas@u-pec.fr (F.S.L.)

Abstract : Monitoring water quality in urban rivers is crucial for water resource management since point and non-point source pollution remain a major challenge. However, traditional water quality monitoring methods are costly and limited in frequency and spatial coverage. To optimize the monitoring, techniques such as modeling have been proposed. These methods rely on networks of low-cost multiprobes integrated with IoT networks to offer continuous real-time monitoring, with sufficient spatial coverage. But challenges persist in terms of data quality. Here, we propose a framework to verify the reliability and stability of low-cost sensors, focusing on the

implementation of multiparameter probes embedding six sensors. Various tests have been developed to validate these sensors. First of all, a calibration check was carried out, indicating good accuracy. We then analyzed the influence of temperature. This revealed that for the conductivity and the oxygen sensors, a temperature compensation was required, and correction coefficients were identified. Temporal stability was verified in the laboratory and in the field (from 3 h to 3 months), which helped identify the frequency of maintenance procedures. To compensate for the sensor drift, weekly calibration and cleaning were required. This paper also explores the feasibility of LoRa technology for real-time data retrieval. However, with the LoRa gateways tested, the communication distance with the sensing device did not exceed 200 m. Based on these results, we propose a validation method to verify and to assure the performance of the low-cost sensors for water quality monitoring.

Keywords : Arduino sensor ; stability ; water quality ; chemical parameters ; urban rivers

4.1. Introduction

Monitoring the water quality of urban rivers is one of the most important issues in water resources management (Bunsen et al., 2021). However water quality degradation is still problematic, due to leaky sewers, rain runoff on contaminated surfaces, and untreated wastewater discharge in surface waters during rain events (Whelan et al., 2022). The spatial and temporal monitoring of water quality in rivers is crucial to optimize the management of freshwater resources since it provides important information to guide stakeholders (Sutadian et al., 2016; Carvalho et al., 2019; Whelan et al., 2020). However for most regulatory parameters, expensive and time-consuming field collection and laboratory analysis are necessary. For instance, for the management of bathing sites, the regulatory monitoring of the bathing waters is based on the enumeration of cultivable fecal indicator bacteria following the European Bathing directive 2006/7/EC (WHO, 2018; Mouchel et al., 2020). Such a monitoring approach is restrictive both in terms of frequency and spatial coverage, resulting in poor comprehension of the actual water quality in a particular area at a particular time (Yaroshenko et al., 2020; Sutadian et al., 2016).

More effective water quality control should rely on methods that are rapid and low cost with minimum sampling required, and, ultimately, it should provide real-time results (Farouk et al., 2022; McGrane, 2016; Yaroshenko et al., 2020). In addition, in situ sensing devices combined with machine learning could help stakeholders to detect in real time the possible

contamination and to optimize the sampling effort (Carvalho et al., 2019; Whelan et al., 2020). Cost-effective strategies should rely on few selected parameters with available low-cost sensors that will serve as indicators of water quality. As pointed out by Zhu et al. (2023), there is no consensus definition of ‘low-cost’ sensors. The cheapest sensors available on the market are usually considered "low cost", and price ranges can depend on the parameter (Zhu et al., 2023). Several physico-chemical parameters can easily be measured in situ, with sensors. For instance, (Kannel et al., 2007) showed the usefulness of monitoring temperature, pH, dissolved oxygen concentrations, conductivity and turbidity to assess the spatial and temporal changes of water pollution and to classify rivers according to their water quality.

A high number of low-cost sensors could be deployed in networks at large spatial scale (Internet of Things, IoT). Each individual sensing device may present a slightly greater error margin than the precision obtained with high-cost equipment. However, the multitude of sensors should compensate by increasing the amount of information both temporally and spatially (Wang et al., 2019a). The continuous development of IoT solutions based on non-proprietary methods during the last decade allows a viable real-time measurement of the water quality for a large spectrum of applications such as monitoring drinking water resources and bathing sites (Bogdan et al., 2023; Wuijts et al., 2022b,a). Many initiatives have arisen, and the interest of the research community has tremendously increased over time (de Camargo et al., 2023). Real-time water quality monitoring through IoT application is expected to help reduce costs associated with logistics and increasing the number of sites monitored. However, the energy autonomy of the monitoring devices deployed on the field needs to be considered. Usually, the sensor is powered by batteries or solar cells. Data are then transmitted either using SMS or long-range (LoRA) technology. In order to be energy-efficient, the long-range (LoRA) technology offers an interesting solution, making it suitable for devices deployed over long periods of time (de Camargo et al., 2023; Huan et al., 2020).

Many challenges remain and need to be covered, such as the reliability, the stability and the repeatability of the measurement, the similarity of performance between sensor units and their interoperability in order to implement in the field reliable continuous monitoring of the water quality (de Camargo et al., 2023). Therefore, the general objective of this paper is to propose a framework to verify the reliability and the stability of the readings and to identify the necessary maintenance of low-cost sensors in order to optimize the quality of the acquired data to assist the stakeholders in the daily management of river water.

Indeed, few studies have focused on the long-term reliability and viability of the sensors, and were restricted to a maximum of 20–30 days despite the fact that river monitoring requires longer periods (Hong et al., 2021; Gowri et al., 2023; Sekhar et al., 2023; Bogdan et al., 2023; de Camargo et al., 2023; Cheniti et al., 2023; Hacker, 2023). As a consequence, our first objective was to analyze the stability over a longer period of 3 months.

Moreover, previous papers highlighted the need for maintenance and cleaning routines to avoid the deposition of debris and biofouling of the sensors that would impair the measurement (Trevathan et al., 2021; Wong et al., 2021). However, no best-practice guideline for the calibration and validation of low-cost sensor networks exists. As the consequence, our second objective was to propose a framework for validation of low-cost sensors.

An additional crucial issue is to consider the data loss due to the limited communication distance between the sensors and the LoRa gateway (Huan et al., 2020). As a consequence, our third objective was to test two LoRa gateways in order to determine the maximum distance between the devices and the gateway without data loss.

In order to address these three objectives, we designed a low-cost multiparameter prototype that can monitor surface water quality using IoT technology. Several sensors such as temperature, pH, conductivity, turbidity and dissolved oxygen were embedded in this device. After the calibration of each sensor, their precision and stability were analyzed in laboratory using reference solutions. The low-cost sensing device was validated for long-term monitoring in the field by comparing it with highly accurate monitoring platforms. In order to validate the possibility of using the prototype in networks, two units were compared, and the performance of the LoRA gateways was assessed.

4.2. Materials and Methods

To monitor water quality, we implemented a LoRa-based wireless system network which includes a LoRa gateway and a network of low-cost sensing devices with real-time data recovery (Figure 2.1). Arduino technology was selected to design this multiparameter sensing device. To execute instructions, process the data, and perform data transmission, two boards can be used : the Arduino UNO R3 based on the Microchip ATmega328P, or the Arduino Mega 2560 microcontroller board based on ATmega2560. The latter was chosen for its compatibility with a high number of monitoring devices (Abotaleb, 2023). Indeed, the Arduino mega board has 8

times more memory space than the UNO R3 board (Abotaleb, 2023; RANDIKA et al., 2022).

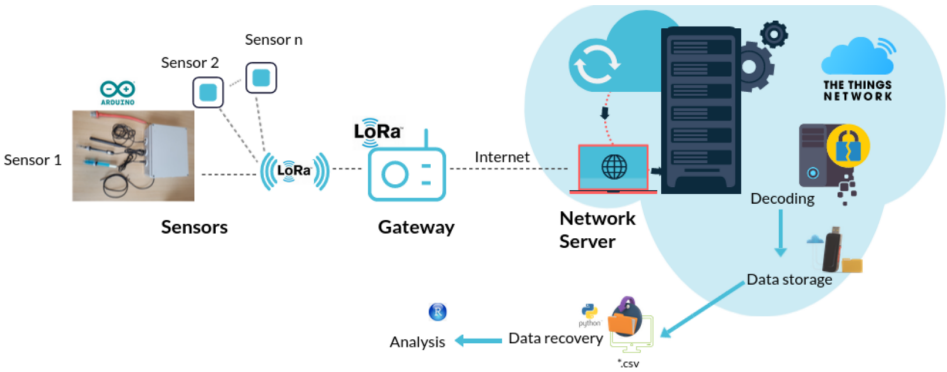

FIGURE 2.1 – Synoptic view of the low-cost system for water quality monitoring in real time.

4.2.1. Prototype Design

Each monitoring device (called "unit") included an external battery (20,000 mAh), 6 analog or digital sensors from DFRobot (Shanghai, China) (temperature, 2 pH, conductivity, turbidity, and dissolved oxygen), a micro SD module/card for data storage, a 16 Bit ADC module V1.0 to increase the precision of the conductivity, turbidity and dissolved oxygen sensors, and a LoRa Shield to connect to a LoRa network (DFROBOT, 2023; Gravity, 2023). Zhu et al. (2023) and de Camargo et al. (2023) compared a list of low-cost water quality sensors with their specifications and a summary of their performance characteristics. These studies were used to select the sensors for our device in order to have a range of reliable low-cost and medium-cost sensors (Zhu et al., 2023). No true low-cost sensors exist for monitoring nutrient concentrations, such as nitrogen and phosphorus. The cheapest from Vernier costs around EUR 300 (Zhu et al., 2023). For pH, two different types of sensors were mounted in the sensing devices in order to compare their performance, which are later named “pH-1” and “pH-2”. All parts of the system were contained in a waterproof box. Analog isolators were used to avoid any signal interference among the sensors, except for the pH sensor. The code allowing the measurement of all parameters at regular intervals was uploaded to the Arduino board and is available on GitHub (`https://github.com/naloufi-manel/low_cost_sensor.git` (accessed on 20 March 2024) with the Python (version 3.8.1) and R scripts (version 4.1.1).

4.2.1.1. Low-Cost Sensors

The pH sensor, which measures the hydrogen ion activity in solution, comprises a pH glass electrode and a silver/silver chloride reference electrode (Bogdan et al., 2023). The pH-1 sensor

(SEN0161-V2, DFRobot) was cheaper than the industrial pH-2 sensor (SEN0169-V2, DFRobot, Shanghai, China).

The specific conductivity reflects the number of electrolytes dissolved in the water (Conductivity Meter V2, 2023). We selected the DFR0300 (DFRobot) sensor since it is the cheapest sensor compatible with Arduino (Zhu et al., 2023). However, its detection range may be more adapted for coastal environments than rivers (Table 2.1). For the conductivity measurements, Equation (3.2) is commonly used to correct the measurements by comparing with a reference measurement at 25 °C :

$$\mathrm{EC}_{25} = \frac{EC_T}{1 + a(T - 25)} \tag{2.1}$$

where EC_T is the conductivity at temperature T (°C), EC_{25} is the conductivity at 25 °C, and (°C^{-1}) is a temperature compensation factor corresponding to the percentage increase per degree (Hem, 1985).

TABLE 2.1 – Characteristics and specifications of the Arduino sensors (Farouk et al., 2022; DS18B20, 2023; pH V2, 2023a,b; Conductivity Meter V2, 2023; Hakim et al., 2019; Arduino, 2023; DO, 2023; Villeneuve et al., 2006).

Parameters	**Temperature (°C)**	**pH-1**	**pH-2**	**Conductivity (mS·cm^{-1})**	**Turbidity (NTU)**	**Dissolved Oxygen (%)**
Sensor	DS18B20	SEN0161-V2	SEN0169-V2	DFR0300	SEN0189	SEN0237-A
Detection range	−10 to 85	0 to 14		0 to 20	0 to 1000	0 to 100
Resolution	0.010	0.010		0.001	1.000	0.050
Measurement Accuracy	±0.5	±0.5	±0.1	±1.0	±3.6	±2.0
Price (EUR)	8	39	65	70	9	148

For turbidity measurement, the selected sensor (SEN0189, DFRobot) measures the light transmittance and scattering rate which changes with the amount of total suspended solids (Arduino, 2023). The sensor uses an infrared LED as a light source and an infrared phototransistor to detect the amount of light not blocked by the water. A change in voltage is obtained and converted into unit measuring turbidity NTU (Nephelometric Turbidity unit) using Equation (3.1) in a

range from 1 to 1000 NTU (Hakim et al., 2019; Arduino, 2023). The upper part of the sensor is covered with a heat-shrink sheath to make it waterproof, and the sensor is shielded from external light using an opaque plastic cover (Trevathan et al., 2020) :

$$\text{Turbidity} = \frac{3.9994 - voltage}{0.0008} \tag{2.2}$$

For measuring dissolved oxygen (SEN0237-A, DFRobot), we select a galvanic sensor with a filling solution and a membrane cap. Its response time stands within a few seconds. Since dissolved oxygen concentration is directly influenced by temperature, we include a temperature compensation in our code (DO, 2023; Villeneuve et al., 2006). Equation (3.3) is usually used to take into account the temperature effect (DO, 2023) :

$$\text{DO} = \frac{volt + b * T - b * T_{cal}}{volt_s + b * T - b * T_{cal}} * 100 \tag{2.3}$$

where DO is the dissolved oxygen (in saturation (%)), volt is the voltage measured at a temperature T, $volt_s$ is the voltage corresponding to the saturated dissolved oxygen measured at a temperature T_{cal}, and b (°C^{-1}) is a temperature compensation factor (DO, 2023).

4.2.1.2. Reference Sensors

To validate the low-cost sensors (noted Arduino sensor), we compared their readings with 2 high-end HYDROLAB Series 5 multiparameters (OTT, Aix-En-Provence, France), which embedded 4 sensors (noted OTT). For dissolved oxygen, we also compared the low-cost sensor with a MINIDOT LOGGER sensor (PME, California, United States), which recorded data on an internal SD card (PME, 2023). The PME sensor measures dissolved oxygen concentration in water using a fluorescence method (PME, 2023).

4.2.2. Specifications and Price

Tables 2.1 and 2.2 show the specifications, operating range, accuracy, and the price of each sensor. The price of the monitoring devices includes the 6 sensors prices added to EUR 159 for the total price of the other components (battery, microSD card and reader, ADC module, box, Arduino card, and isolators) and EUR 93 for the LoRa connection. The total price of each monitoring device was between EUR 285 and EUR 400. For the Hydrolab multiprobes from OTT, the price reached EUR 3050 and the PME sensor cost EUR 1775.

TABLE 2.2 – Characteristics and specifications of the Hydrolab multiprobes (OTT) (Hydrolab DS5X, 2024).

Parameters	**Temperature (°C)**	**pH**	**Conductivity (mS·cm^{-1})**	**Turbidity (NTU)**
Detection range	−5 to 50	0 to 14	0 to 100	0 to 3000
Resolution	0.01	0.01	0.0001	0.10
Measurement Accuracy	±0.100	±0.200	±0.001	±1.000
Price (EUR)		480	380	1540

4.2.3. Cleaning and Calibration

Standard solutions at different concentrations were used to calibrate each sensor except for the temperature sensor. The standard solutions were checked using an Eutech multiparameter probe for pH and conductivity, the CellOx® 325 sensor for the dissolved oxygen and the 2100P turbidimeter (HACH) for turbidity. For the pH, we used standard buffer solutions (pH 4, 7 and 10) from VWR. To remove contamination, which leads to a reduction in slope and unstable readings, every month, the pH sensor must be immersed in 0.1 mol $\cdot$ L^{-1} of HCL solution for a few hours then rinsed with deionized water. For conductivity, the standard solutions were prepared from a 1 M stock solution of potassium chloride. Standard solutions were diluted in deionized water to reach 0.36 mS $\cdot$ cm^{-1}, 0.72 mS $\cdot$ cm^{-1} and 1.41 mS $\cdot$ cm^{-1}. For the turbidity sensor, we used a Formazin stock solution at 4000 NTU (prepared from dissolved hydrazine sulfate and dissolved hexamethylenetetramine). The stock solution was diluted to 0, 20 and 200 NTU in deionized water. Finally, for the dissolved oxygen sensor, a sodium sulfite solution was used for the zero point (VWR), and tap water maintained at saturation with a bubbler served as a 100% standard solution. The oxygen sensor needed to be prepared before use by adding a filling solution into the membrane cap, which consisted of a 0.5 mol $\cdot$ L^{-1} NaOH solution. The filling solution needed to be changed every month. Then, the sensor was calibrated at a fixed temperature (between 20 and 25 °C) in the 100% saturated water.

Each sensor was carefully washed with deionized water and wiped before calibration. The calibration took place at a fixed temperature and under agitation at 700 rpm using a magnetic stirrer. The sensor was kept in the standard solution for a few minutes to stabilize, after which the calibration point could be set. Each calibration point was measured 10 times, and the fitting regression curve (y = cx + d) was determined. For each parameter, the coefficients (c and d)

were used to correct the measured values after data recovery. Calibration needed to be performed once a week.

4.2.4. LoRa Gateway

The LoRa Shield v1.4 from Dragino with SX1276 LoRa Chip fully compatible with Arduino models was associated with the Arduino Mega 2560, which operates at a frequency of 868 MHz (European Union) and contains an external antenna (Dragino LoRa Shield, 2023). The LoRa modules were configured at a bandwidth of 125 kHz, transmit power of 14 dBm, and spread factor of 12. We tested 2 different models of the LoRa Gateway to compare their performance in terms of range coverage. The first gateway is a Raspberry gateway made of LoRa hat for RPi (Raspberry Pi) with a SX1276 LoRa Chip associated to a RPi 3 and implemented with a single-channel gateway program (LoRa, 2023). The second gateway is the Arduino pro Gateway LoRa connectivity. It allows up to 8 LoRa Channels in the 868 Mhz frequency (Semtech solution) and includes a microchip SX1301 with two SX1257 and an on-board UFL antenna. According to the manufacturer, LoRa gateways allow connecting devices within several kilometers (Arduino Pro Gateway Documentation, 2023). For the two gateways, we estimated the spatial coverage of the gateways by measuring the distance between the end node and the gateways using a signal levels analysis. The transmission distance was tested regarding the quality of the signal by analyzing the Received Signal Strength Indicator (RSSI), the Signal-to-Noise Ratio (SNR) measured by the gateway and the time interval between the reception of 2 successive data. RSSI measures the distance between a transmitter and a receiver and SNR quantifies the strength of the signal regarding the amplitude of the ambient noise (Tsanousa et al., 2021; Audéoud et al., 2020). These indicators are commonly used for the estimation of the maximum distance (Guidara et al., 2021). The tests were performed in dense and residential urban zones (Greater Paris area), with the gateway placed at a fixed position and the end device at different positions (Figure S9).

4.2.5. Sensor Validation

The reliability and the long-term stability of the tested low-cost sensors were checked in laboratory and in the field. The field tests were conducted in Bassin de La Villette (Paris, France), where OTT sensors were already deployed (Guillot-Le Goff et al., 2023).

4.2.5.1. Accuracy

The accuracy of each sensor after a calibration was tested for 2 sensing device units in order to evaluate the linearity and the repeatability of each sensor (norm ISO 21748 : 2017 and NF EN 17075 2018) (Venelinov, 2016). The tests were performed in the laboratory at ambient temperature (20.97 $\pm$ 0.12 °C) under agitation at 700 rpm. To validate the temperature sensor, the reading was performed in a water bath with a range of temperature from 5 °C to 30 °C, incremented by 5 °C every 8 min, followed by stabilization for 15 min at the same temperature. For the other sensors, between 2 and 7 standard solutions at varying concentrations were used. For each sensors, readings were repeated 10 times for each standard solution (Table 2.3). Repeatability was estimated by calculating the standard deviation of the sensor's measurements during the repeated trials. Trueness and linearity were evaluated by comparing the readings with the value of the standard solutions (true value). A linear regression was generated by plotting the low-cost sensor measurements against the known concentration of the standard solutions. Reproducibility of the sensing devices was evaluated by inter-comparison of the performance of two sensing devices. For each sensor, 2 units were tested in parallel for a week with the same standard solutions. Each parameter except for oxygen (due to high-cost of the sensor) was measured every 15 min. The temperature was maintained at around 20 °C, the pH sensors were placed in a pH 4.22 solution, the conductivity sensors were placed in a 1.42 $mS \cdot cm^{-1}$ solution and finally, the turbidity sensors were placed in a 10 NTU solution. Reproducibility was estimated by calculating the standard deviation between 2 units.

TABLE 2.3 – Descriptive analysis of sensors calibration for 2 units.

Parameters	**Temperature** (°C)	**pH-1**	**pH-2**	**Conductivity** ($\mathbf{mS \cdot cm^{-1}}$)	**Turbidity** (NTU)	**Dissolved Oxygen** (%)
Number of measures (n)	368	30	30	44	77	20
Standard solutions	Temperature from 5 to 30°C	4, 7 and 10		4 standards from 0.22 to 1.42	7 standards from 0 to 800	0 and 100
Linearity (units 1–2)	0.999	0.999	0.999	0.998–0.993	0.998	0.999
Slope of the curve						
Unit 1	0.999	0.938	0.959	1.060	0.947	1.038
Unit 2	0.999	0.950	0.984	1.083	0.916	
Repeatability						
Unit 1	0.01	0.02	0.01	0.02	3.66	1.74
Unit 2	0.01	0.02	0.01	0.02	3.69	
Reproducibility	0.03	0.02	0.01	0.02	3.54	

4.2.5.2. Temperature Effect

In order to analyze the effect of temperature on the measurement by all the sensors and to identify the correct parameters for compensation, each sensor measured every 15 min under agitation at 700 rpm a standard solution previously cooled at 10 °C, allowing the solutions to reach an ambient temperature for 3 h (from 10 to 19 °C). The standard solutions were the following : pH 10.2; conductivity 0.72 $mS \cdot cm^{-1}$; turbidity 20 NTU; and dissolved oxygen 100% O_2 saturated water via a bubbler. For dissolved oxygen, the age of the membrane cap was also taken in consideration by using a 6-month-old membrane and a new membrane. For the new membrane, the temperature variation analyzed was between 14 and 25 °C. In order to distinguish variations due to temperature fluctuations from sensor errors, the results were compared to readings of the same standard solutions at a fixed temperature of 20.97 $\pm$ 0.12 °C for 3 h.

4.2.5.3. Temporal Stability in the Laboratory

Testing a probe's stability in the laboratory, where environmental conditions are tightly regulated, provides reliable test conditions (de Camargo et al., 2023). The controlled conditions of the laboratory enable the probe's readings to be compared with known standards to verify that the measurements are accurate and consistent. The short-term and long-term stability of the sensors was tested in the laboratory at a steady ambient temperature of 19 $\pm$ 2 °C. To evaluate the short-term stability, we collected 3 replicates of 1 L water samples from Créteil Lake and from the lower Marne River (Paris area, France) in April 2022. The samples were placed under agitation, and measurements were taken continuously with the sensors every 10 s for 3 to 6 h. This short-term analysis was carried out under continuous supervision in order to immediately detect any problem or rapid variations. The pH-2 was not tested, as it was bought later.

The long-term stability was analyzed by placing each sensor in a standard solution (pH : 7 ; conductivity : 0.72 mS $\cdot$ cm^{-1} ; turbidity : 20 NTU ; and dissolved oxygen : 100% O_2 saturated water via a bubbler). Measurements were taken every 3 to 5 min for approximately 3 months. For dissolved oxygen, the PME sensor was used as a reference.

4.2.5.4. Temporal Stability in the Field

To test the long-term stability of the sensors in the field, we installed the low-cost monitoring devices at two sites 1 km apart from each other (A and B) at Bassin de la Villette (Paris area, France), as shown in Figure 2.2. Site B is in front of the bathing site of Paris Plage, and site A is upstream of site B, enabling contamination to be anticipated at the bathing site. Every year, analyses are regularly carried out by the City of Paris during the summer period (June to September) to monitor the microbiological quality in the proximity of site B. In 2022, the results indicated a good microbiological quality, with an average concentration of *Escherichia coli* and intestinal enterococci of 101 $\pm$ 78 MPN/100 mL and 44 $\pm$ 50 MPN/100 mL, respectively. As for the physico-chemical parameters measured, the temperature was 21.27 $\pm$ 2.83 °C, the conductivity was 0.65 $\pm$ 0.03 mS $\cdot$ cm^{-1}, and the turbidity was 7.32 $\pm$ 2.61 NTU. These 2 selected sites are part of a research project where high-precision OTT multiparameter probes have been deployed continuously since 2020. This long deployment was regularly verified and maintained in order to provide reliable data. OTT multiparameters were used as a reference. Measurements were performed in situ at site A from early September 2022 to early January 2023 and then from May 2023 to June 2023, and at site B from early September 2022 to the end of November 2022 . For site B, the OTT probe only measures temperature and conductivity. Occasional loss of data occurred due to unit malfunction or installation problems on site.

The installed low-cost devices were changed every week in order to clean the sensors and to check their calibration. In the beginning, cleaning and calibration were carried out directly in the field on the same unit. However, because of the length and complexity of the process starting from ”15 October 2022”, the method was modified by alternating between two units each week. The sensors of device N°2 were cleaned, calibrated and stabilized for a few hours in the laboratory before replacing the device N°1 in the field, and vice versa. The measurement interval was also optimized during these stability tests.

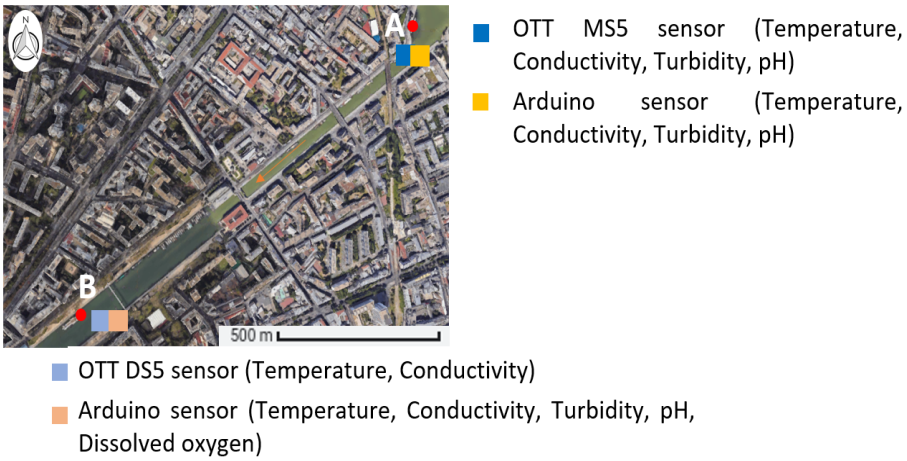

FIGURE 2.2 – Installation sites (A,B) and parameters measured by each type of sensor (source : Google Maps).

4.3. Results and Discussion

Figure 2.3 shows a low-cost sensing device once it is completely assembled.

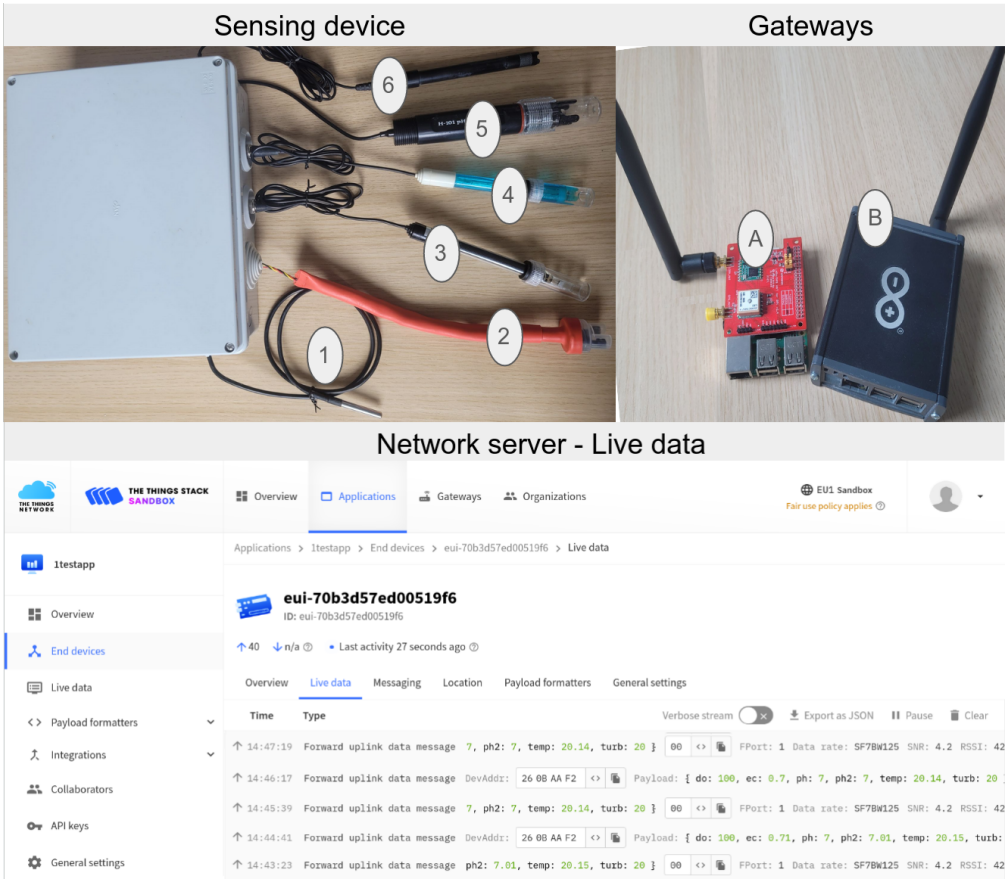

FIGURE 2.3 – Hardware components involved in the field experiment : 1 : temperature, 2 : turbidity, 3 : conductivity, 4 : pH-1, 5 : pH-2, 6 : dissolved oxygen, A : LoRa HAT gateway, B : LoRa Arduino Pro gateway.

4.3.1. Accuracy of the Sensors

After calibration, the accuracy of each sensor was evaluated with the linearity and repeatability (Table 2.3). Correlation between measured values and the expected values of the standard solutions showed good linearity with a significant rh > 0.99 ($p < 0.01$) for all sensors (Figure S1). The slopes were between 0.92 and 1.08, which showed a good precision of the measure compared to the true value (Table 2.3). Each sensor from both devices showed high repeatability with low standard deviation values between the repeated measures, with values ranging from 0.01 to 0.02 for temperature, pH and conductivity sensors (Table 2.3). Turbidity and dissolved oxygen sensors showed less accuracy with higher standard deviation between repeated measures. The reproducibility between units was satisfactory for all sensors except the turbidity since measurements of the 2 sensing devices were in good agreement as demonstrated by the low standard deviation values. A recent review of Zhu et al. (2023) compiled performance indicators of several low-cost sensors, including the SEN0169, DFR0300, and SEN0189 sensors selected in our study (Moyón Rivera and Ordóñez Berrones, 2019; Saputra et al., 2017; Rozaq

et al., 2020; Hakim et al., 2019; Trevathan et al., 2020). Zhu et al. (2023) noticed that the information was heterogeneous and somewhat difficult to compare for trueness and linearity, and most of the time repeatability and reproducibility were not estimated.

4.3.2. Reproducibility of the Sensors

In order to verify if there is a difference in the accuracy of different units of the same type of sensor, a one-week experiment was carried out with two units of each sensor placed in the same standard solutions (Figures 2.4 and S5). The temperature measurements of the two units matched almost perfectly. The average difference between the 2 units was only 0.07 °C, with a significant correlation of Spearman (Figure 2.4A, r = 0.98, $p < 0.01$, $n = 434$). There was fairly good reproducibility between the 2 conductivity sensors, with a mean deviation of 0.04 mS $\cdot$ cm^{-1}, a low coefficient of variation of 2.83% for unit 1 and 2.14% for unit 2 and a low but significant correlation (Figure 2.4B, r = 0.30, $p < 0.01$, $n = 434$).

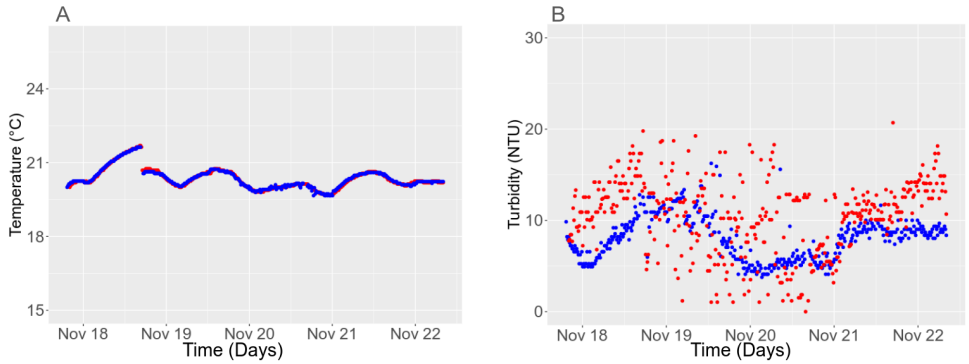

FIGURE 2.4 – Comparison of two unit sensors placed simultaneously in the same solution. In blue unit 1, and in red unit 2. (**A**) Temperature, (**B**) turbidity.

The pH-1 sensor needed a few hours to stabilize its reading and the pH-2 meter took one day (Figure S5A,B). After stabilization, the mean deviation between the 2 units was low for both sensors (pH-1 : 0.09 and pH-2 : 0.02), with a low coefficient of variation for pH-1 of 0.47% for units 1 and 1.71% for unit 2 and for pH-2 0.23% for unit 1 and 0.47% for unit 2.

For the turbidity sensor, the two units differed by an average of 3.91 NTU (monitoring of a 10 NTU solution, Figure 2.4B). The correlation was significant but weak (r = 0.32, $p < 0.01$, $n = 434$). Figure 2.4B shows that the 2 units displayed the same trend over time but with a greater dispersion for the 2nd unit (coefficient of variation 29.81% for units 1 and 38.06% for units 2). The reproducibility appears rather poor for the sensor. This may be due to the difference in

performance of the infrared LED and phototransistor inside the sensors (Zhu et al., 2023).

4.3.3. Sensitivity to the Environment

Low-cost sensors are usually sensitive to the environmental conditions and need retrofit actions such as compensation equations, waterproof enclosure, or coating (Zhu et al., 2023). For instance, water temperature is known to influence the measure of some parameters and the sensitivity to sensor current (Hayashi, 2004; Jeroschewski and Zur Linden, 1997). We analyzed the effect of temperature by comparing between 3 h series of measurements under increasing temperature conditions with measurement at fixed ambient temperature. Under fixed conditions of temperature, for all of the sensors, the fluctuation over time of the measurement was low, showing a good stability of the measure. Compensation for the temperature effect was not necessary for the 2 pH meters and the turbidity sensor. The coefficients of variation of the stable temperature series for pH-1 and pH-2 meters were 0.45%, 0.22% (respectively), and 12.76% for the turbidity sensor. Under fluctuating temperature condition, the coefficients of variation were higher (0.52% and 0.47% for pH meters and 14.38% for the turbidity sensor). This slight variation as confirmed by Figures S2A,B and 2.5A,B was rather due to random variations observed over time.

In the case of conductivity and dissolved oxygen sensors, there was a noticeable deviation in the measurement under varying temperature (3.09% and 5.88%) compared with the fixed-temperature measurements (1.60% and 0.63%) (Figures S2C,D and 2.5C,D). This indicates that a compensation for temperature effect was required. Compensation coefficients were determined by fitting a model linear curve to the data. Several values of the 'a' coefficient (Equation (3.2)) are commonly cited in the literature. For example, Hayashi (2004) reported an average a value of 0.0187 $°C^{-1}$(minimum–maximum : 0.0175–0.0198 $°C^{-1}$), which is in accordance with the 0.019 $°C^{-1}$ value recommended by Clesceri (1998). Based on the EC-temperature relation, we identified a compensation factor of 0.0265 $°C^{-1}$, which is comparable to the 0.025 $°C^{-1}$ value reported by Keller and Frank (1966). After compensation of the measured values using the coefficient 0.0265 $°C^{-1}$, the coefficient of variation displayed a lower value (1.19%), close to the coefficient of variation obtained at a fixed temperature. The 0.0265 coefficient provided a better fit (Figure S2D) compared to the 0.0185 factor recommended by the manufacturer (Conductivity Meter V2, 2023).

For the dissolved oxygen sensor, there was an effect of temperature on the readings

(Figure 2.5C,D). This result is not surprising since the saturation of oxygen in water is dependent on the temperature and due to the change in permeability of the sensor membrane (Hitchman, 1978; Villeneuve et al., 2006). By fitting Equation (3.3) to the increasing temperature series, a factor 'b' of 14.48 °C^{-1} was determined and used for the temperature compensation of the sensor signal. After compensation, the coefficient of variation decreased from 5.88% to 1.78%, which is closer to the coefficient of variation of 0.63% obtained for the reference analysis at a fixed temperature (Figure 2.5D). Moreover, the cap membrane should be replaced at least every 6 months since the coefficient of variation with a new membrane was 1.78%, whereas it was 7% with a membrane used for 6 months (Figure S4).

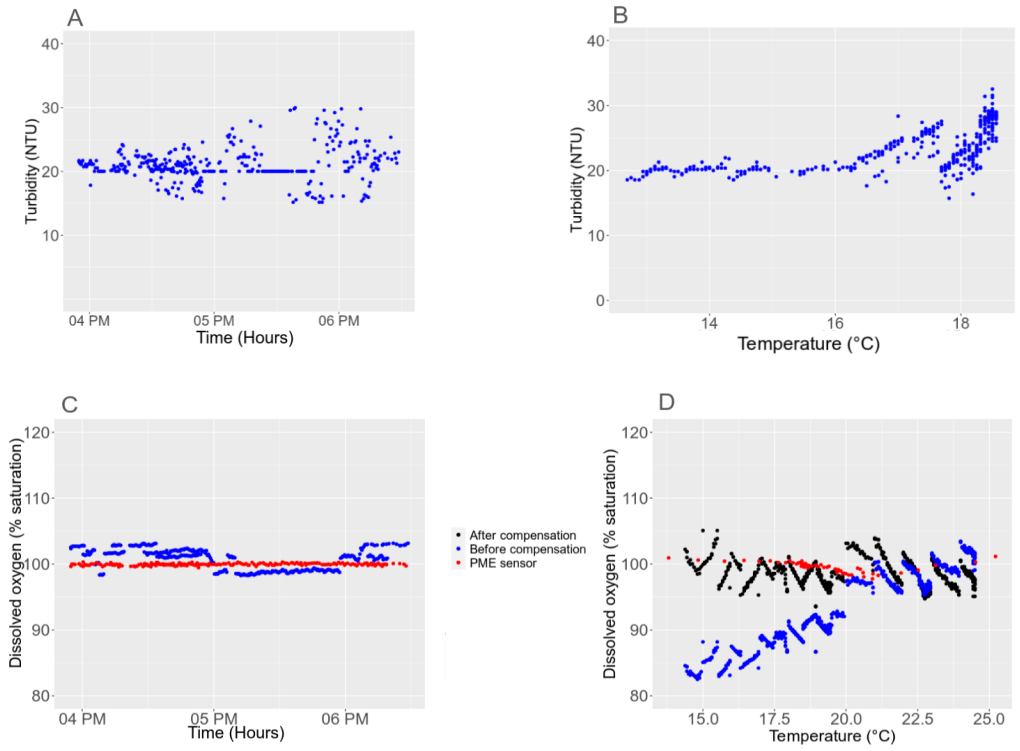

FIGURE 2.5 – Temperature effect on the turbidity and dissolved oxygen. (**A**–**C**) Fixed temperature analysis, (**B**) turbidity measurement at different temperatures, (**D**) Dissolved oxygen before compensation in blue, after compensation in black and the PME sensor in red.

Other external factors may affect the sensing device, causing irreversible damage and reducing its lifespan. We tested if the battery was overheating the sensors housing and whether this might affect vulnerable components on the Arduino board. The manufacturer specify that Arduino boards should be operated between −25 °C and +70 °C (Arduino Boards, 2023). During 3 months of monitoring temperature in the laboratory, the temperature in the box (19.97 ± 1.77 °C) and the water temperature (19.15 ± 1.91 °C) remained steady. This indicates

that at ambient temperature, the battery did not overheat the waterproof box.

4.3.4. Temporal Stability in the Laboratory

Following calibration, a short- and long-term stability analysis was carried out with all sensors. This checking was rarely performed for the low-cost water quality sensors (Zhu et al., 2023).

4.3.4.1. Short-Term Stability

Different surface water samples were monitored for 3 to 6 h at room temperature. The readings showed a relatively satisfactory temporal stability with average standard deviation values not significantly different from those obtained during the calibration, except for temperature (*t* test, n = 3, $p > 0.05$) (Table 2.4). Different studies checked the stability of DFRobot sensors using standard solutions but only for a few minutes to several hours (Saputra et al., 2017; Trevathan et al., 2021; Alimorong et al., 2020; Saha et al., 2018) (and Atlas Scientific (Méndez-Barroso et al., 2020)). Generally speaking, in situ water measurement with low-cost sensors appears promising, with relatively satisfactory temporal stability for all parameters (temperature (Méndez-Barroso et al., 2020; Alimorong et al., 2020; Saha et al., 2018), pH (Méndez-Barroso et al., 2020; Saha et al., 2018), turbidity (Trevathan et al., 2020; Alimorong et al., 2020), conductivity (Méndez-Barroso et al., 2020; Alimorong et al., 2020; Saha et al., 2018; Saputra et al., 2017), and dissolved oxygen (Méndez-Barroso et al., 2020)).

TABLE 2.4 – Short-term analysis of sensors (repeatability).

Parameters	**Average**	**Min-Max**
Temperature (°C)	0.78	0.38–1.01
pH-1	0.04	0.02–0.08
pH-2	0.03	0.02–0.06
Conductivity ($mS \cdot cm^{-1}$)	0.04	0.02–0.08
Turbidity (NTU)	4.68	3.89–5.46
Dissolved Oxygen (%) Arduino Sensor	2.33	1.55–3.71
Dissolved Oxygen (%) PME Sensor	0.31	0.16–0.59

4.3.4.2. Detection and Removal of Outlier for Long-Term Series

The long-term stability was checked by placing each sensor in a standard solution for

3 months. The turbidity sensor showed a wide dispersion, which required rectification (Figure 2.6A). Indeed,the measurement of the 20 NTU standard solution gave values ranging from 0 to 1000 NTU. Filtering noise is a common pre-processing step of real-time datasets, and numerous noise-reduction methods have been used to detect and remove outliers (Xu et al., 2015; Le Deunf et al., 2020). A set of filtering methods was tested to identify the most optimal one : interquartile range, density-based methods K-means, DBSCAN (Density-Based Spatial Clustering of Applications with Noise) clustering, combining DBSCAN with Local Outlier Factor, Mean-shift and the ARIMA (Autoregressive Integrated Moving Average) model with the median filter approach (Xu et al., 2015; Wang and Wang, 2019; Sedaghat et al., 2013; Yang et al., 2021; Bianco et al., 2001).

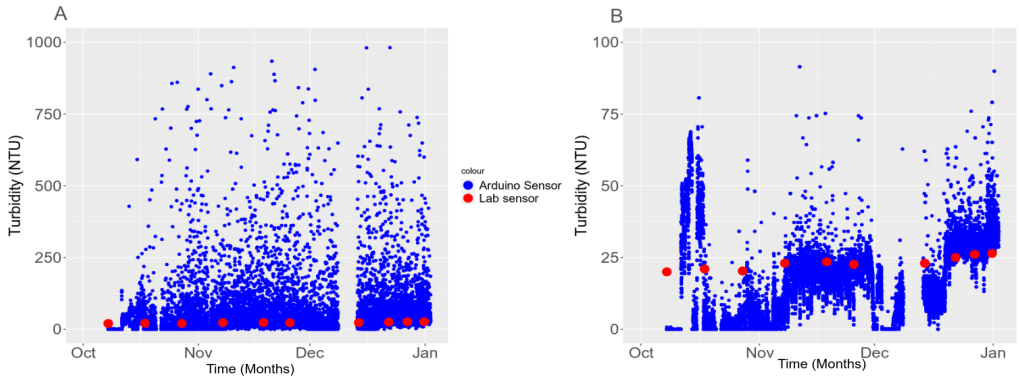

FIGURE 2.6 – Long-term turbidity analysis. The blue dots correspond to measurements taken by the sensors and the red dots to measurements taken by the laboratory turbidimeter. (**A**) Raw data, and (**B**) after removal of outlier using ARIMA with a median filter (width of 5).

The first five methods removed mainly extreme outliers, corresponding to 2.69% to 9.84% of the data. The ARIMA model, which can be used for data cleaning of non-stationary time series (Wang and Wang, 2019; Bianco et al., 2001), seemed the best for cleansing this turbidity dataset (Figure 2.6B). With a moving window of 3, 5 and 8 points, 18.84%, 26.77% and 37.12% of the data were identified as outliers, respectively. Considering the data density and trend, the optimal window seemed to be 5, but this parameter is data-dependent (Le Deunf et al., 2020). The conductivity and the dissolved oxygen datasets did not require extensive cleaning since less than 0.01% and 0.001% (respectively) of the data were removed using the ARIMA approach. Time series methods are robust, and efficient data cleaning tools can process a dynamic dataset within a solid theoretical framework and detect outliers with different properties (Xu et al., 2015; Liu et al., 2004). Since the cleaning process should be based on a minimum modification of the original data Liu et al. (2004), for each dataset of the different

sensors, different parameters were tested and retained.

4.3.4.3. Long-Term Stability

After cleaning of the datasets, the long-term stability was estimated using the standard deviation. Given the manufacturer's precision values for each sensor, the calculated standard deviation values could be considered reasonable (Tables 2.1 and 2.5). Long-term measurements remained quite stable for most of the sensors, with the exception of turbidity and dissolved oxygen, which showed greater variability (Figures 2.6B and 2.7). It could be noted that the temperature sensor correctly measured two air-conditioning incidents in the laboratory in early November and early December (Figure 2.7A).

Figure 2.7B shows that both pH meters were fairly stable (standard deviations of 0.04). However, after 3 months, the pH-1 meter drifted by 1.0 pH unit (Figure 2.7B). This was due to the fouling of the electrode which was removed by soaking the sensor in a 0.1 M solution of HCl for at least 8 h to a maximum of 24 h (pH V2, 2023a). After regeneration at the end of December, the pH-1 meter was back to a stable reading (Figure 2.7B). The pH-2 meter is more suitable for long-term online detection due to its ring PTFE membrane that confers resistance to clogging (pH V2, 2023b).

TABLE 2.5 – Stability analysis of sensors : standard deviation (* without missing values during the sensor regeneration, ** values cleaned with an ARIMA method using a median filter).

Parameters	**Value**
Temperature (°C)	1.91
pH-1 *	0.04
pH-2	0.04
Conductivity ($mS \cdot cm^{-1}$)	0.03
Turbidity (NTU)	64.85
Turbidity ** (NTU)	13.23
Dissolved Oxygen (%) Arduino Sensor	12.42
Dissolved Oxygen (%) PME Sensor	0.73

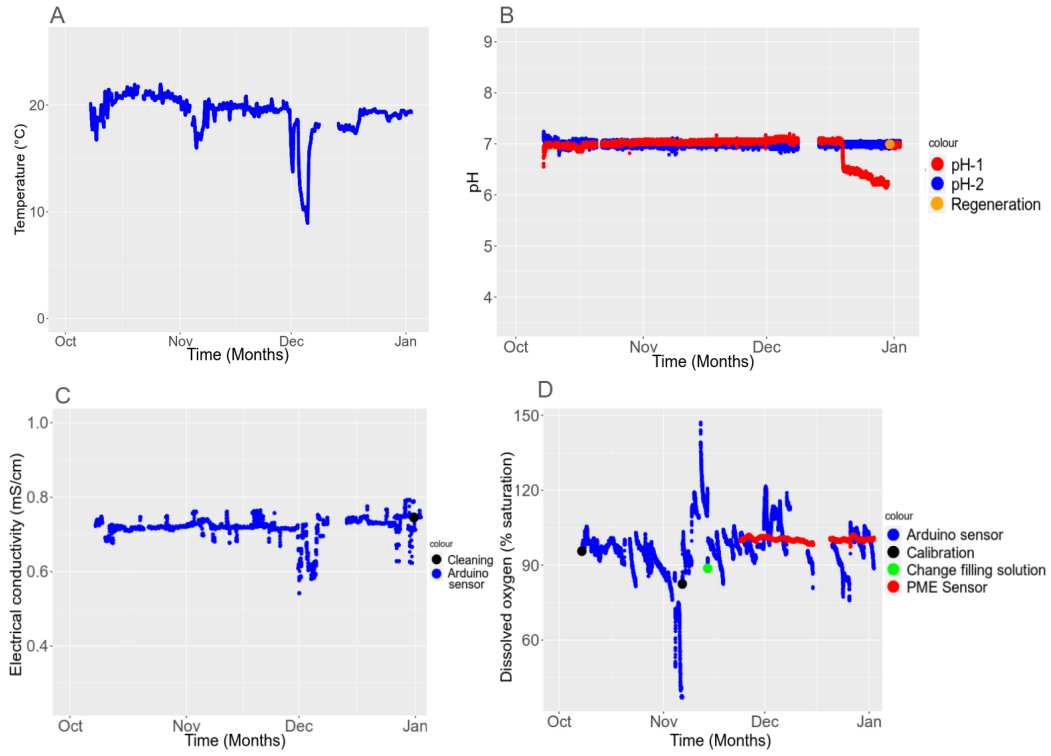

FIGURE 2.7 – Long-term stability of sensors reading standard solutions. (**A**) Temperature, (**B**) pH, (**C**) conductivity measurement cleaned with an ARIMA model and median filter (width of 11), and (**D**) dissolved oxygen measurement cleaned with an ARIMA model and median filter (width of 5).

For the conductivity sensor, only minor variations could be observed (Figure 2.7C). A sharp decrease in the reading happened in early December due to a sudden drop in the laboratory temperature to 11 °C (Figure 2.7C). The compensation equation was not sufficient to make up for this sudden temperature variation. It may have originated from a desynchronization between the water temperature variation and the optical components heat change (Shi et al., 2022). Special care should be taken with rapid temperature variations, as the reading will not be totally reliable. Finally, at the end of the 3-month period, the sensor needed to be cleaned to restore a stable monitoring.

For the oxygen sensor, after one month of stable measurements, the percentage of oxygen decreased from 95.00 ± 4.37% to 36.87% (Figure 2.7D). A new calibration only helped to stabilize the reading for a few days until the measures raised to 154.93%. A change in filling solution is in fact necessary every month. Finally, for the turbidity sensor, the long-term standard deviation remains quite high 13.23 NTU, though the data cleaning tremendously improved the situation (Figure 2.6B). This high variability indicates a certain instability of the sensor. In fact, the study of Trevathan et al. (2020) also reported low reliability and accuracy for the same sensor with values below 100 NTU. The difference in performance of the infrared LED and

phototransistor of this equipment probably affects the detection limit, making the sensors more adapted for monitoring high-turbidity waters (Zhu et al., 2023; Trevathan et al., 2021).

Overall, laboratory experiments showed that the measurements were relatively stable over the short and long term. Readings were concordant between two units of the same sensor, with the exception of turbidity, which fluctuated considerably and was not reliable. In terms of sensor maintenance, the pH-1 and the dissolved oxygen sensor needs to be maintained monthly. In addition to the oxygen sensor, the membrane should be changed twice a year. Finally, for the conductivity sensor, care must be taken when dealing with sharp temperature variations fully. The longevity of the sensors was not checked ; however, the manufacturer datasheets usually indicate a lifespan > 6 months (Zhu et al., 2023).

4.3.5. *In Situ* Validation

The accuracy and stability of the low-cost sensors was estimated by comparing with high-end probes at two sites in Bassin de la Villette (Paris), which were already equipped with OTT multiparameter probes (Hydrolab Sensor, 2024). Field monitoring also raised concerns about the interferences of environmental parameters (such as sunlight and temperature variation) with the reading signal of the low-cost sensors, especially with the turbidity sensor (Trevathan et al., 2020).

4.3.5.1. Light Interference with the Turbidity Sensor

The ambient infrared radiation interfered with the detection of the sensor infrared LED by the infrared photo transistor. This resulted in a daily oscillation of the turbidity readings, with a peak in the late afternoon and evening (Figure 2.8A). Trevathan et al. (2020) also identified a degree of ambient infrared interference during the daytime using the same sensor. To avoid light interference from external light, the sensor should be shaded by a cover, an opaque box or a tubing, with the bottom open to allow water to circulate freely. Half of the bottom of the sensor with the infrared LED and the infrared phototransistor is not in the opaque box. This allows water to circulate between the two ends, without affecting the results obtained. We partly solved this light interference problem by shading the sensor using an opaque shell held with a weight above the sensor submerged in the water (Figure 2.8B). However, some variations were still present (Figure 2.8B), probably due to the inherent instability of this sensor and due to the indirect refracted light penetrating the water (Trevathan et al., 2020).

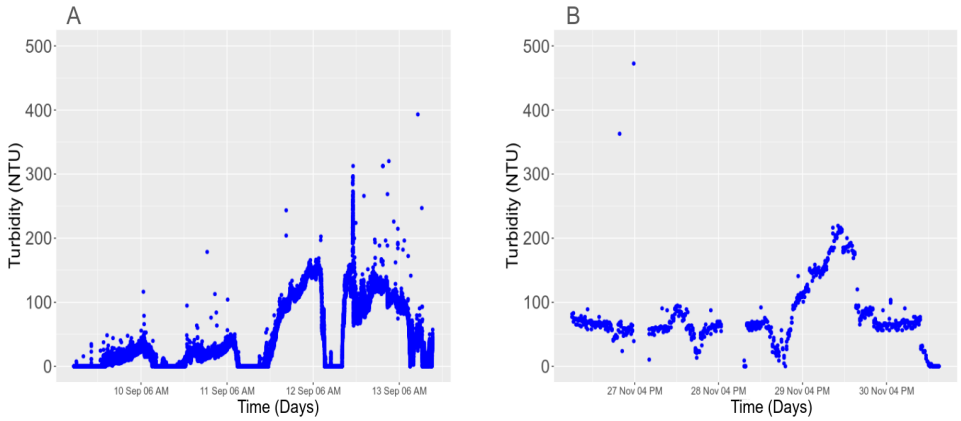

FIGURE 2.8 – Effect of the ambient light on the reading of the turbidity sensor. (**A**) Before shading, and (**B**) after shading.

4.3.5.2. Temporal Stability in the Field

Although calibration with standard solutions is crucial to improve the accuracy of sensors, it is not sufficient. It is also essential to compare the results obtained from low-cost sensors with those of reference devices, such as high-resolution sensors, to ensure their validity (de Camargo et al., 2023).

To provide reliable data, the frequency of data acquisition should be selected to compromise between noise minimization and time resolution. During the first week of monitoring at La Villette, the time interval of 10 sec was too short and produced noised time series (Figure S6). Later, a setup of three measurements with 10-second intervals every 20 min helped in optimizing the data quality for the remaining monitoring period (Figure S6). The mean standard deviation between the three measurements was low for the temperature sensor (0.010 ± 0.004 °C), the 2 pH meters (0.028 ± 0.012 for pH-1 and 0.010 ± 0.007 for pH-2) and for the conductivity sensor ($0.004 \pm 0.001 \text{mS} \cdot \text{cm}^{-1}$). However, the difference between the repeat measurements of the turbidity sensor was high (42.4 ± 43.9 NTU), indicating low repeatability.

As already observed in the laboratory, the two units of the temperature sensor were highly reliable and accurate. The readings of the Arduino sensor were similar to the readings of the OTT sensor at both sites (A and B) (Figures 2.9 and S8A). Similarly, Méndez-Barroso et al. (2020) obtained very good performance results of the DS18b20 temperature sensor.

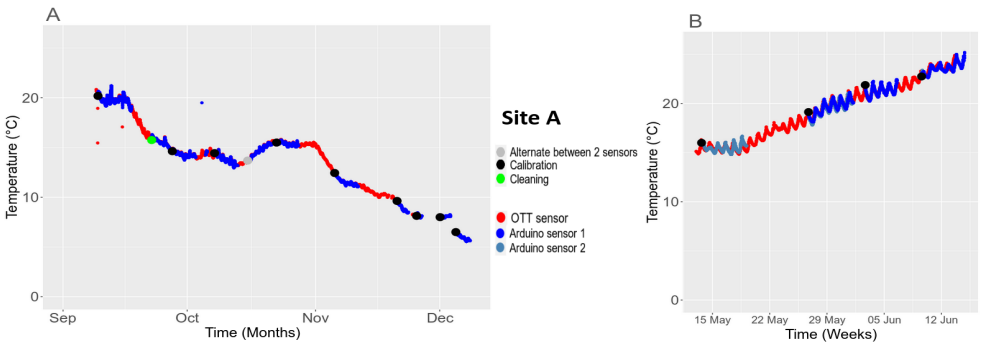

FIGURE 2.9 – Temperature analysis at Site A at Bassin de La Villette. Values from OTT sensors are displayed in red, Arduino sensors in blue. Black dots indicate that the sensor has been calibrated, green dots that it has been cleaned, and gray dots that the sensor has been replaced. Replacements were carried out by alternating the two units of the same sensor every week. (**A**) From early September 2022 to early January 2023, and (**B**) from May to June 2023.

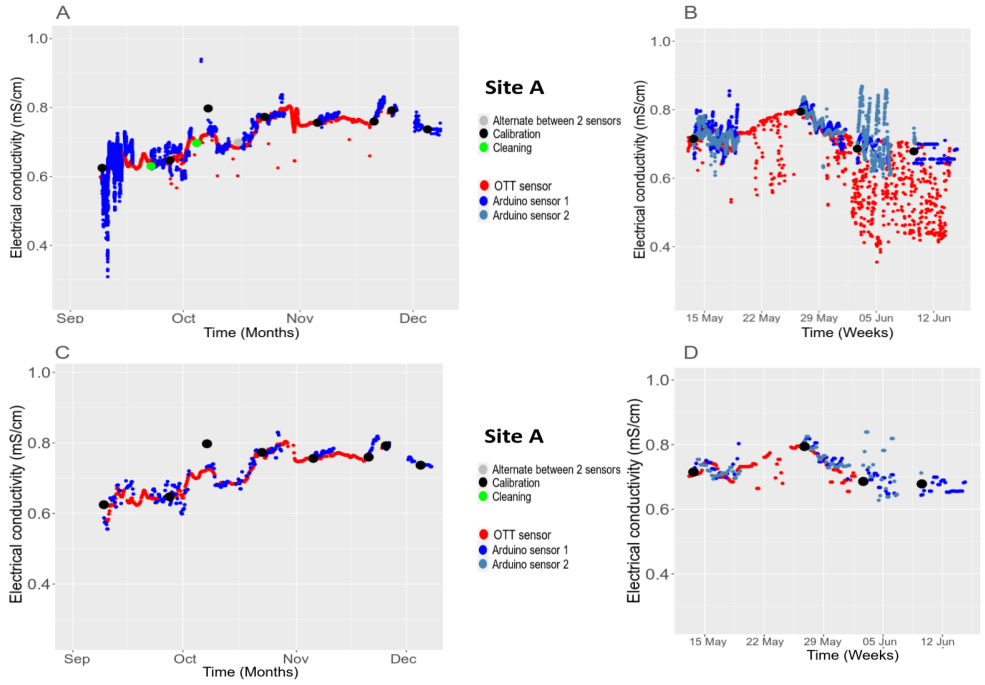

FIGURE 2.10 – Conductivity measurement at site A at Bassin de la Villette. Values of the OTT sensors are displayed in red, and Arduino sensors in blue. Black dots indicate that the sensor has been calibrated, green dots that it has been cleaned, and gray dots that the unit has been replaced. Replacements were carried out by alternating the two units of the same sensor every week. (**A**) From early September 2022 to early January 2023, (**B**) from May to June 2023, and (**C**,**D**) data from (**A**,**B**) averaged over 4 h and cleaned by ARIMA.

Field campaigns also confirmed that the pH-1 sensor, although reliable enough to enable monitoring, was less accurate and stable than the pH-2 meter (Figure S6). The standard deviation for the pH-1 meter was 0.14 (unit 1) and 0.33 (unit 2), whereas for the pH-2 meter, the deviation was slightly lower, at 0.10 and 0.22 for each unit, respectively. The OTT sensor was the most reliable, with a standard deviation of 0.09. Indeed, Demetillo et al. (2019) also identified an average error of 0.18 for Atlas scientific sensors (which are slightly more costly than the pH-1 sensor) during a two-week test. This indicates the need to find the right balance between the cost

and the accuracy of the sensor, which will depend on its intended use.

The Arduino conductivity sensors displayed a similar trend compared with the OTT sensors at both sites (Figures 2.10 and S8B), although in May and June, few measurement errors could be observed due to soiling. During the spring and summer, regular maintenance is required due to biofouling as is visible for both the Arduino and the OTT sensors (Figure 2.10B). Data post-treatment (averaging over 4 h and removal of the outliers with ARIMA model) helped in providing time series of sufficient quality. Overall, the data obtained from the Arduino sensors agreed well with the OTT sensors, indicating that the low-cost sensors were effective in providing usable data. However, for setting an IoT of low-cost sensors, it should be kept in mind that the reproducibility of the two units of Arduino conductivity sensors was sometimes low (standard deviation of 0.17 $mS \cdot cm^{-1}$ and 0.02 $mS \cdot cm^{-1}$, respectively). It should not be forgotten that this sensor has low accuracy (factory certificate) since it is more suitable for monitoring water quality in mariculture (Conductivity Meter V2, 2023). Some other sensors are more accurate and more suitable for freshwater water ; however, they are three times more expensive. For instance, the SEN0451 sensor from DFRobot displays an accuracy of 0.1 $mS \cdot cm^{-1}$ (Conductivity Meter, 2024; de Camargo et al., 2023; Zhu et al., 2023).

Concerning the turbidity sensor, the readings were highly noised due to the instability of the sensor and light interference (Figure S7). Hacker (2023) tested for a month the same turbidity sensor and also identified an instability in the measurement. As noted by Hong et al. (2021), the cable being too short, the sensor floats at the surface, leading to light interference. Fouling, as indicated by the gradual increase in NTU values (Figure S7B), triggered the requirement for regular maintenance.

Finally, the dissolved oxygen was measured over a few days, both by the Arduino sensor and the PME sensor at site B (Figure 2.11). The two sensors displayed similar trends, although the variation deviation was slightly greater for the Arduino sensor compared to the PME sensor (respectively 3.56% and 1.80%). Huan et al. (2020) designed a low-cost dissolved oxygen sensor, which displayed an average error of 2.47%. Using this sensor, they also observed daily oscillations like we did, with peaks in the afternoon when the temperature increased. To demonstrate that low-cost sensors operate properly on site and to help in establishing their accuracy and reliability, long-term exposure in the field is a recommended procedure (de Camargo et al., 2023). The low-cost temperature sensor was highly reliable, while the pH, conductivity and dissolved oxygen sensors gave relatively satisfactory results. Over time, small measurement

errors tended to appear. This phenomenon was more pronounced for the Arduino sensors than for the OTT sensors. Similarly, other sensors from DFRobot or Atlas Scientific showed good stability and effectiveness with small measurement errors (Demetillo et al., 2019; Huan et al., 2020; de Camargo et al., 2023). Considering the cost of the sensors tested in our study and their relatively low margin of error, their utilization for continuous measurement in the field was validated, given regular maintenance to ensure the reliability of the results. The results we obtained indicate that weekly cleaning and calibration of the Arduino sensors are necessary for some parameters. The labor cost associated with the weekly maintenance is hard to quantify since it depends on a variety of factors, such as the installation time, the number of sensors, and also the sites.

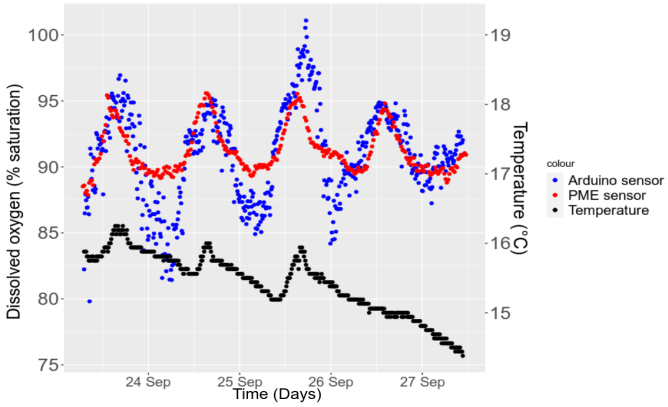

FIGURE 2.11 – Dissolved oxygen measurements at site B at Bassin de la Villette. Values of the PME sensor are displayed in red, Arduino sensors in blue, and Arduino temperature sensors in black.

After each calibration, we recommend letting the sensors stabilize for a few hours in the standard solution before installation in the field. Finally, the turbidity sensor does not seem suitable for the continuous monitoring of fresh waters. In environmental conditions, the turbidity sensor quickly becomes soiled by biofilm, and the slightest particle or element that passes through, such as a leaf, may cause a variation in the readings. Trevathan et al. (2020) also identified a fast negative impact of fouling (less then 48 h) on signal transmission. Zhu et al. (2023) showed that even with other brands (TSD-10 and TSW-10 from Amphenol), the reproducibility appears rather poor for these low-cost turbidity sensors since they are all built on the same principle. The turbidity sensor should potentially be more suitable for detecting particular events with significantly high turbidity levels, such as wastewater (Trevathan et al., 2020; Hakim et al., 2019). The error rate decreased with increasing turbidity (Zhu et al., 2023). This rate was higher for turbidity levels above 100 NTU (Hakim et al., 2019; Gusri and Harmadi,

2021).

Based on this sensing device performance, we propose a framework to verify the reliability and stability and to identify necessary maintenance measurement intervals for each of the sensors (Figure 2.12). This framework can be generalized to all types of sensors other than those presented in this study so that they can be verified before installation and data processing. A more detailed synopsis of the framework is presented in Figure S10.

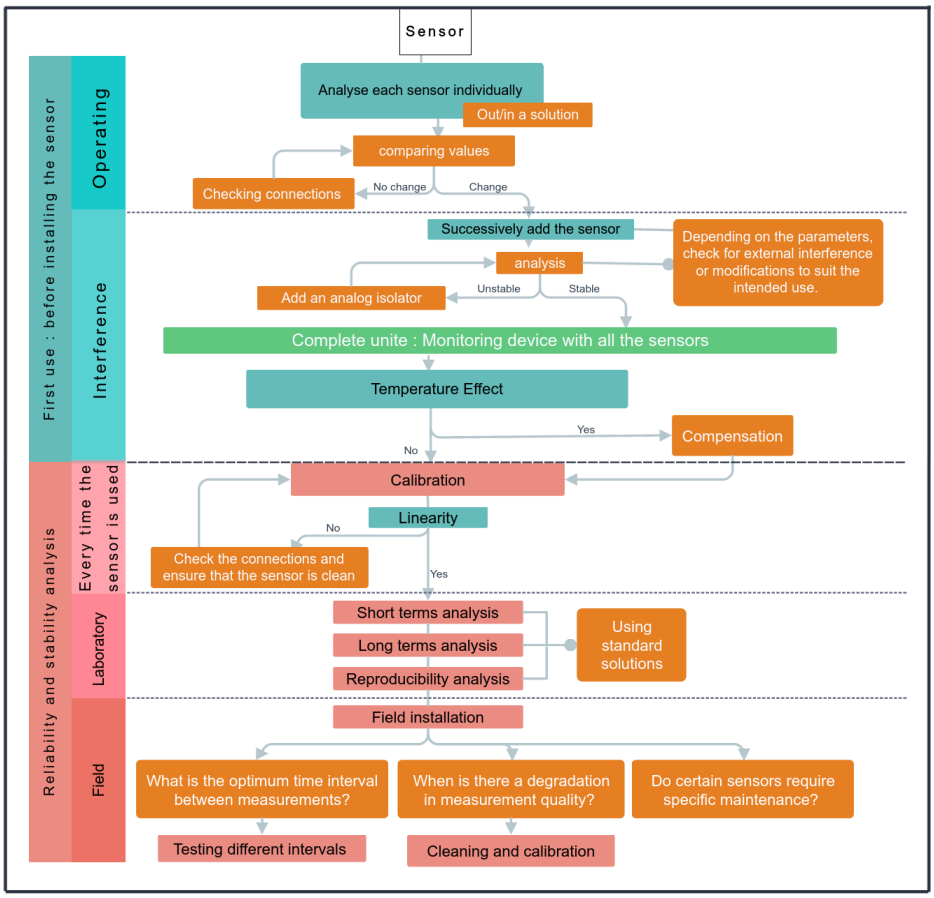

FIGURE 2.12 – Framework for testing the reliability of sensors.

4.3.6. LoRa Gateway Performance

Long-range wide-area networks (LoRaWANs) were recently introduced as a promising low-power technology for several IoT applications, including networks to monitor water quality (Jiang et al., 2020; Wang et al., 2019a). We analyzed the performance of two different LoRa gateways (a LoRa Arduino Pro gateway and a LoRa HAT gateway) in their ability to retrieve data from the end node device and to send them to the server without data corruption and loss. Both gateways were first tested in a dense urban area (Campus of Vitry, France). The maximum distance at which the node managed to send data was 200 m for the LoRa Arduino Pro gateway

and 170 m for the LoRa HAT gateway, which is far below the potential distance announced by the manufacturer for the Arduino gateway (Figure 2.13).

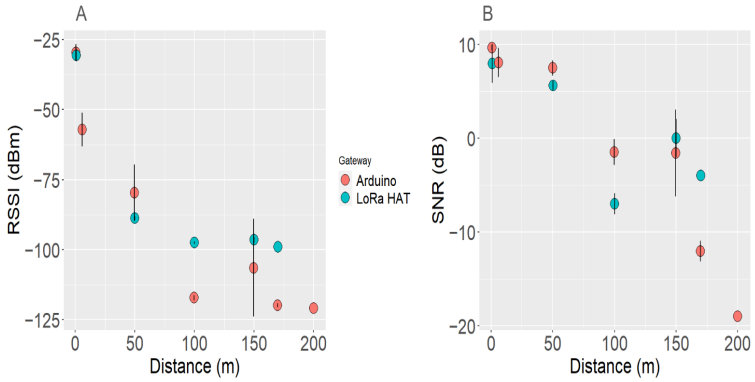

FIGURE 2.13 – LoRa gateway performance. Arduino LoRa gateway in pink, LoRa HAT gateway in blue light. (**A**) Received signal strength indicator (RSSI), (**B**) Signal-to-Noise Ratio (SNR).

Sendra et al. (2023) similarly identified a maximum distance of 150 m with the same LoRa HAT gateway we used. Interference and path loss can occur due to structural obstacles, such as glass, metallic surfaces or walls, and due to the interference of other electronic components (Sendra et al., 2023; Guidara et al., 2021; Zourmand et al., 2019). As a consequence, the signal propagation is obstructed, resulting in deterioration of the SNR and reduction in the RSSI levels with the increasing distance. After 100 m, we observed a rapid decline in the signal quality (RSSI levels) of both gateways, in the zone with the most obstacles. In the zone with fewer obstacles, the quality remained relatively unchanged between 100 and 200 m. Under 100 m, the LoRa HAT gateway exhibited better performance than the Arduino gateway, while it was the opposite between 100 and 200 m (Figure 2.13). Using a gateway combining the sx1278 (433 MHz) and ESP8266 modules, Zourmand et al. (2019) also found that the quality signal decreased above 120 m from the gateway as indicated by the negative SNR (below the noise floor).

We also assessed the performance of the LoRa gateways with the time interval between the reception of two successive data (Figure S11). Up to 100 m, the interval between two measurements was short 5.25 ± 5.20 min for both gateways, though the LoRa HAT gateway displayed a better signal quality. Above 150 m, the time interval increased beyond 15 min for the LoRa HAT gateway, and over 20 min above 200 m for the LoRa Arduino gateway. However, even with a longer time reception, the quality and quantity of the data were still integral without any loss or degradation of the data collected. Beyond this distance limit, no data were received by the LoRa gateways.

The effect of the environment on the signal quality was tested with the LoRa Arduino gateway positioned in two different sites at a distance of 50 and 100 m. The first site was densely built, while the second site (residential area at Vitry, France) presented fewer buildings, and therefore fewer obstacles. Figure S12A,B show that for site 2, the signal quality was slightly better with higher RSSI at 50 m. However, there was no significant difference between the two sites (Wilcoxon test, $p = 0.25$, n = 72). This result is not surprising since coverage is usually much lower in urban areas than in open land such as rural areas, reaching up to several kilometers for the latter (Petrariu et al., 2019).

4.4. Conclusions

Our study demonstrated the suitability of the Arduino sensors (except the turbidity sensor) for monitoring water quality. In particular, the low-cost temperature sensor performed very well, as well as the two pH sensors, showing good repeatability and stability in the laboratory and in the field. However the pH-1 meter requires monthly maintenance, including regeneration of the sensor to remove any residue on the electrode. The low-cost conductivity sensor gave more variable results with lower accuracy. Similarly, the dissolved oxygen sensor was satisfying in terms of data acquisition and in terms of required maintenance. The filling solution should be changed every month and the membrane every 6 months (depending on frequency of use). The turbidity sensor is not recommended since it is too unstable and sensitive to external light. For a reliable low-cost sensing device, a balance has to be struck between cost and sensor reliability, depending on the sensor's intended use.

A framework was then proposed to help characterizing and validating the sensing devices. This flexible framework makes it possible to integrate various sensors, to add or replace sensors as required, and to create a variety of devices to meet different measurement objectives and different water matrices. Finally, with a view to having a network of monitoring system, we tested two LoRa communication modules (LoRa HAT gateway and the LoRa Arduino pro gateway). Both performed well, with maximum communication distances, respectively, of 170 m and 200 m.

This low-cost monitoring device will be used in networks for the continuous acquisition of water quality data in a river. The provision of a dense multiprobes network integrated into an IoT system would enable real-time monitoring with greater precision due to the multitude

of sensors. Coupling with a real-time anomaly detection system, like a nonlinear cooperative control algorithm based on game theory (Casado-Vara et al., 2018), would help in improving the continuous monitoring of surface water and reducing maintenance costs. Further studies are required to verify this hypothesis. The data collected with the devices will also feed machine learning models to predict the water quality and set up an alert system for urban bathing sites. It will also help with rationalizing the sampling strategy during the bathing season to measure bacterial indicators of fecal pollution. These combined approaches will improve sensor performance, reduce cost, and accelerate decision-making processes.

Data Availability Statement : This dataset is not yet openly accessible.

Acknowledgments : We thank the Service des canaux of the city of Paris (France) for the access to monitoring sites at Bassin de la Villette. We are grateful to Jean-Marie Mouchel for providing us with the dissolved oxygen sensor (PME, MINIDOT LOGGER) and to Mohamed Aymen Labiod for helping with the installation and configuration of the gateways.

4.5. Appendix

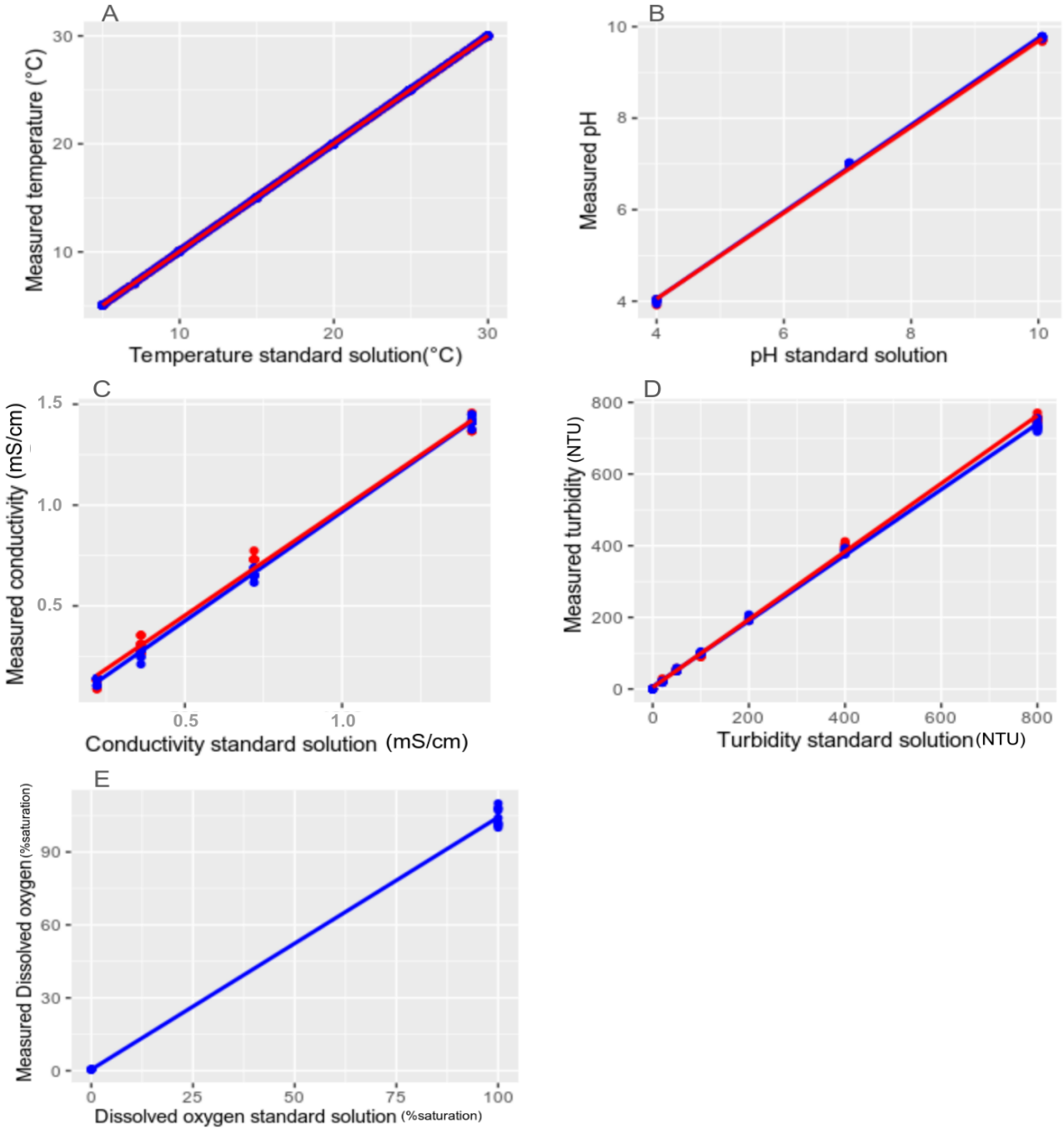

FIGURE S1 – Analysis of the sensors calibration for 2 units (The results of the first unit in blue and the second unit in red). (**A**) Temperature. (**B**) pH. (**C**) Conductivity. (**D**) Turbidity. (**E**) Dissolved oxygen.

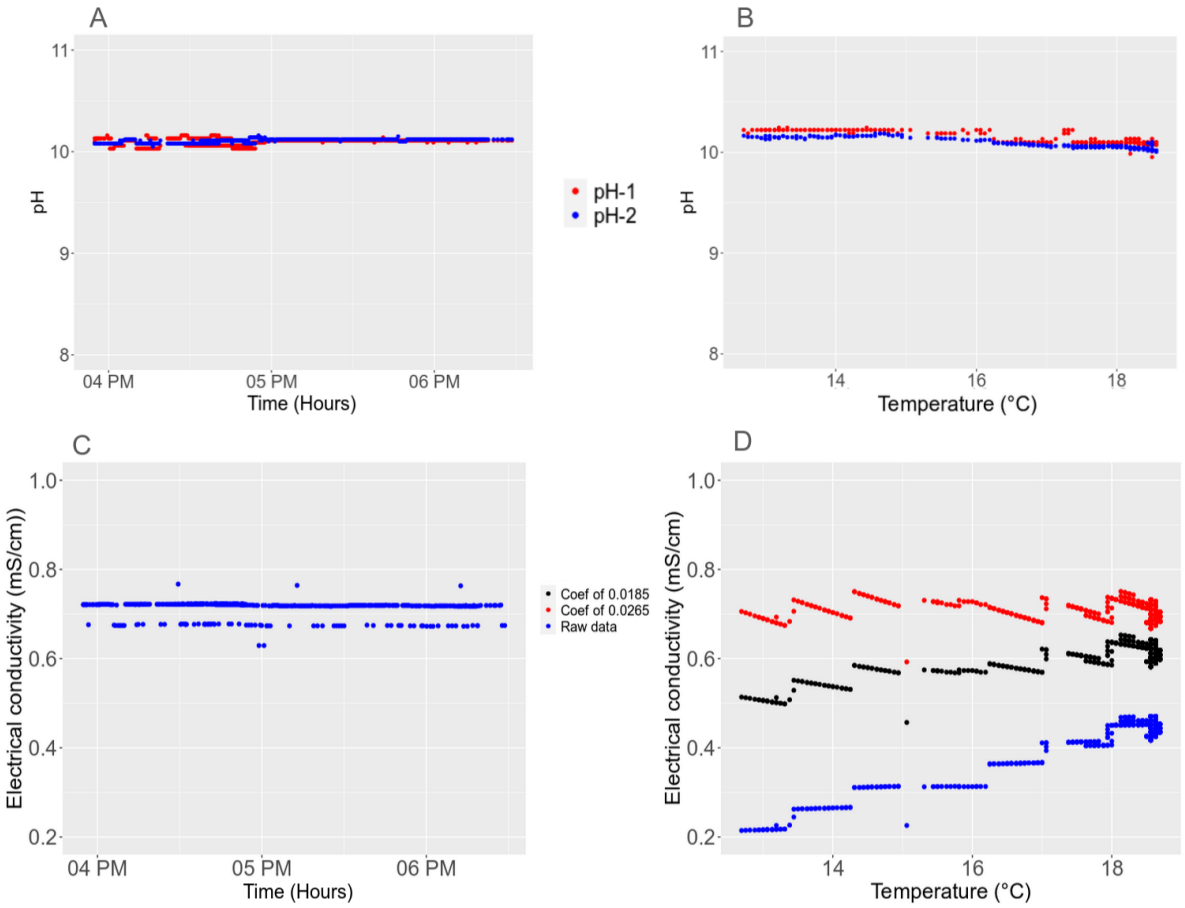

FIGURE S2 – Temperature effect on the pH and conductivity. (**A**–**C**) Reference, fixed temperature analysis. (**B**) pH measurement at different temperature. (**D**) Conductivity at different temperature without compensation (raw data) in blue and with compensation by using 2 compensation coefficients (coef of 0.0185 in black and 0.0265 in red).

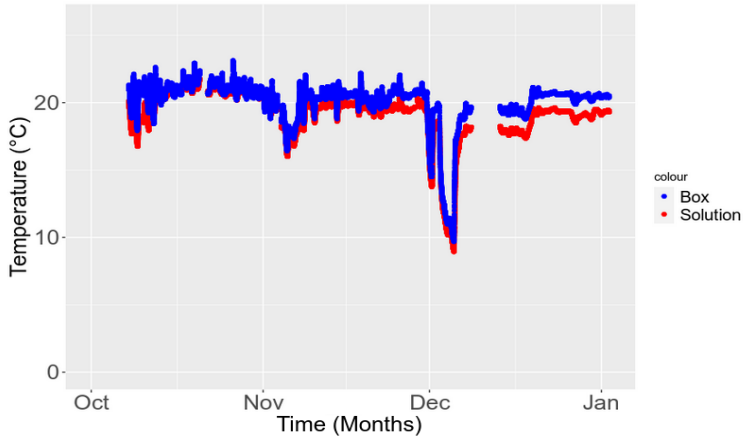

FIGURE S3 – Effect of battery temperature in the box. In blue, the temperature in the waterproof box, and in red, the solution at laboratory temperature.

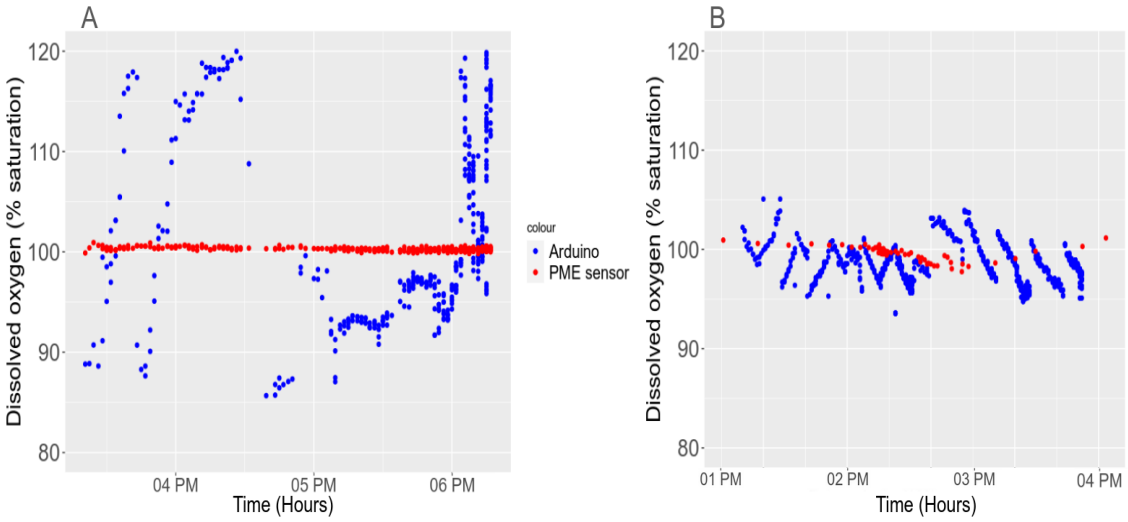

FIGURE S4 – Effect of long-term use of Dissolved Oxygen sensor Membrane Cap. (**A**) After 6 months of use. (**B**) New membrane cap.

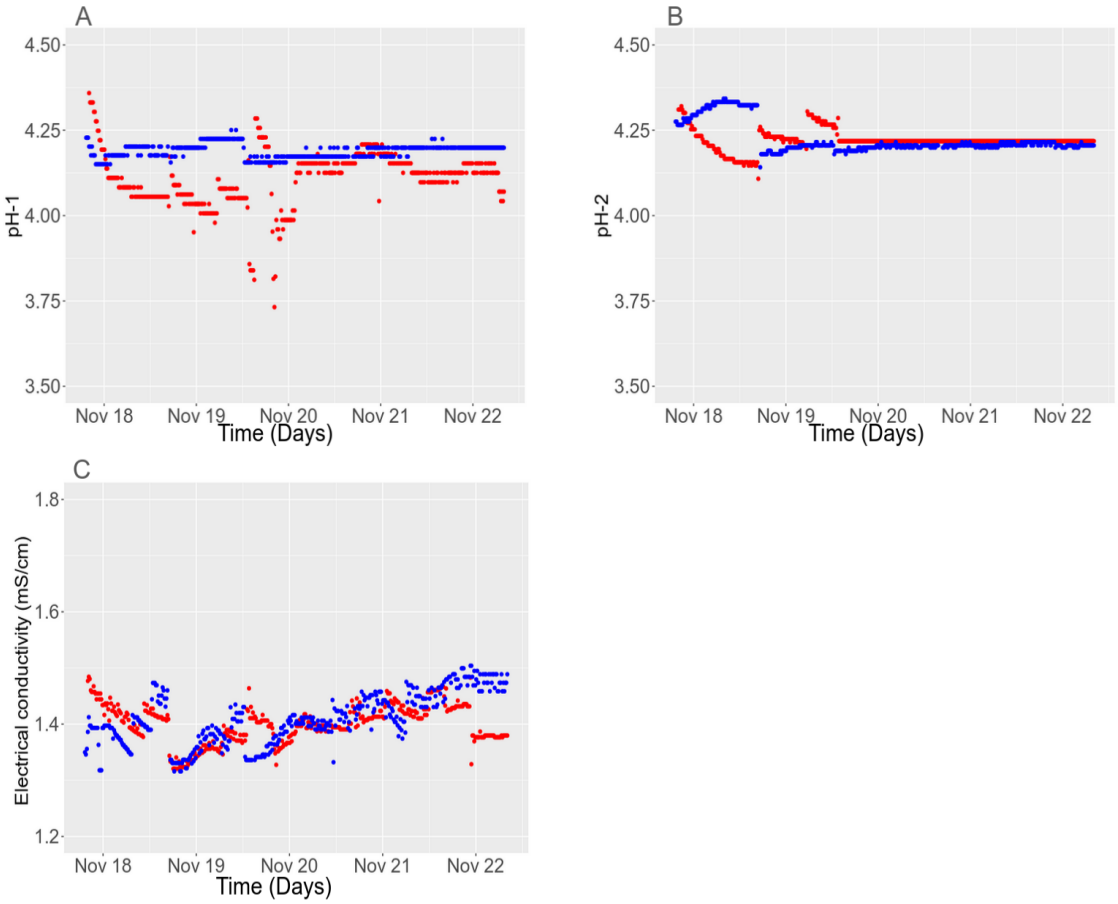

FIGURE S5 – Comparison of two unit sensors placed simultaneously in the same solution : in blue, unit 1, and in red, unit 2. (**A**) pH-1, (**B**) pH-2, and (**C**) conductivity.

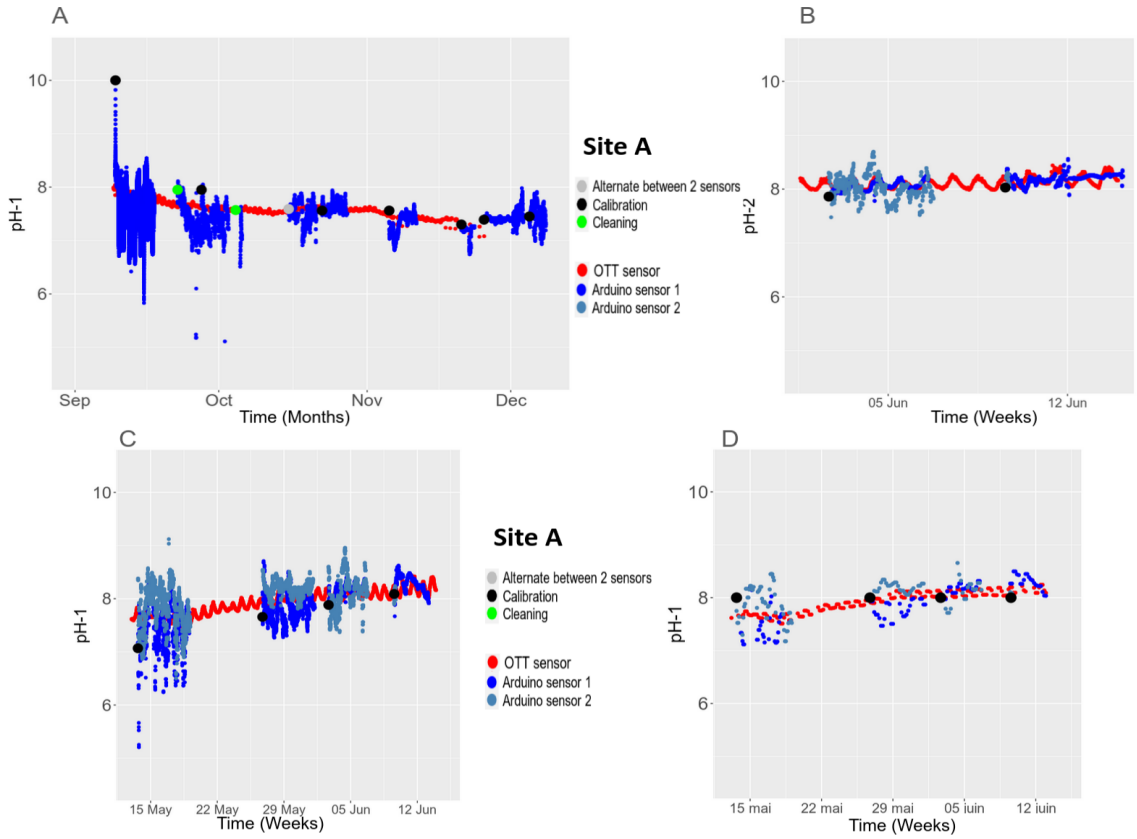

FIGURE S6 – pH analysis at site A at Bassin de la Villette. OTT sensors in red, Arduino sensors in blue. The black dot indicates the date of calibration, green only if the sensor was cleaned, and grey when the analysis process has changed, alternating between sensor units each week : (**A**) from early September 2022 to early January 2023 for the pH-1 meter, (**B**) in June 2023 for the pH-2 meter, (**C**) From May to June 2023 for the pH-1 meter, and (**D**) data from (**C**) averaged over 4 h and cleaned by ARIMA.

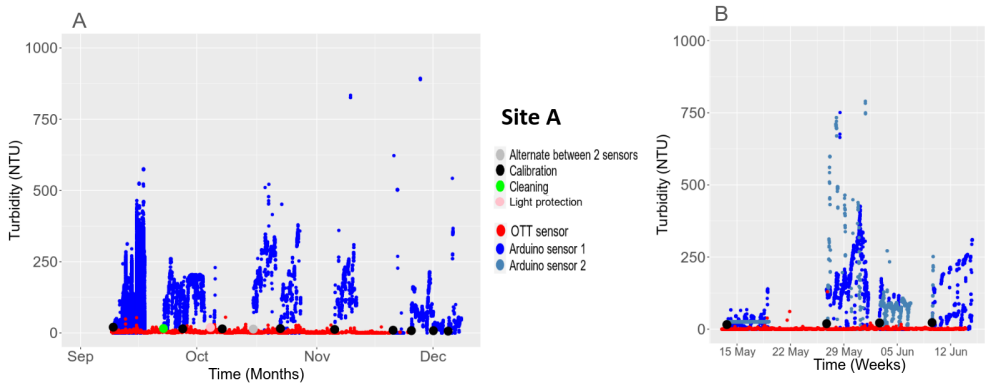

FIGURE S7 – Turbidity analysis at site A in Bassin de la Villette. OTT sensors in red, Arduino sensors in blue. The black dot indicates the date of calibration, green dots if the sensor has been cleaned, and grey dots when the analysis process has changed, alternating between sensor units each week and pink dots for external light protection : (**A**) from early September 2022 to early January 2023, and (**B**) from May to June 2023.

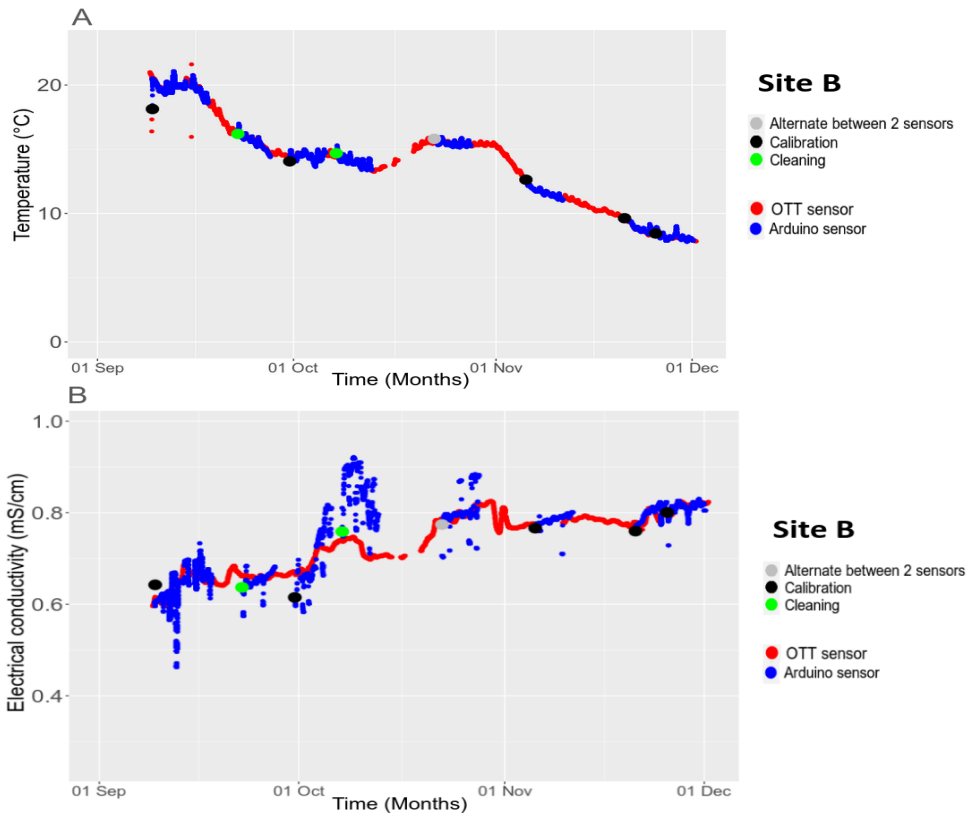

FIGURE S8 – Temperature (**A**) and conductivity (**B**) analysis at site B in Bassin de la Villette. OTT sensors in red, Arduino sensors in blue. The black dot indicates the date of calibration, green dots if the sensor has been cleaned, and grey dots when the analysis process has changed, alternating between sensor units each week.

FIGURE S9 – Site 1 (Campus of Vitry) (**A**) and site 2 (residential area at Vitry) (**B**) for the 2 LoRa gateways tests.

Analyse each sensor (first use) :

- Put the sensors in the solution then out :
    - If measurements do not change : check the connections between the Arduino board and the sensor
    - If measurements change : the sensor is operational
- Check for interference : check that the measurement remains stable over time
    - If measurements are unstable : add an analog isolator for the unstable sensor
    - If measurements are stable: the monitoring device is complete and operational
- Check temperature effect : check if the measurement changes with increasing temperature
    - Measurement changes : identify the right compensation factor
    - Measurement does not change : the sensor is operational

Reliability and stability analysis :

- Calibration : analysis of the linearity
    - If the correlation is < 0.99 : Check the connections and ensure that the sensor is clean
    - If the correlation is > 0.99 : good linearity
- Laboratory analysis : using standard solutions in a controlled environment
    - Short term analysis : during few hours
    - Long term analysis : during few months
    - Reproducibility of the sensors : comparison between 2 units in the same solutions
- Field analysis : identify the optimal process for collecting reliable data
    - Optimize measurement interval : increase the interval until a suitable interval is found for continuous, energy-optimized monitoring
    - Optimize the maintenance process : identify when there is a degradation in measurement quality in order to identify the most effective cleaning and calibration method
    - Check if certain sensors require specific maintenance : e.g. pH-1 requires regeneration once a month

FIGURE S10 – More detailed synopsis of the framework for testing the reliability of the sensors.

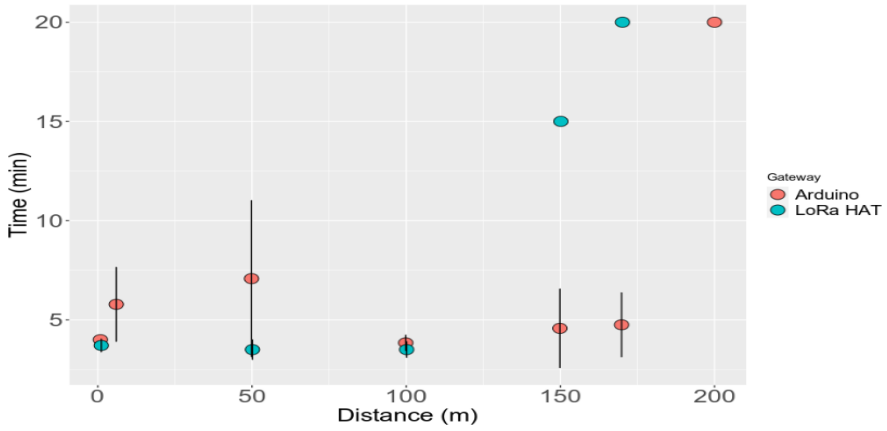

FIGURE S11 – Time gap between two measures. Arduino LoRa gateway in pink, LoRa HAT gateway in light blue.

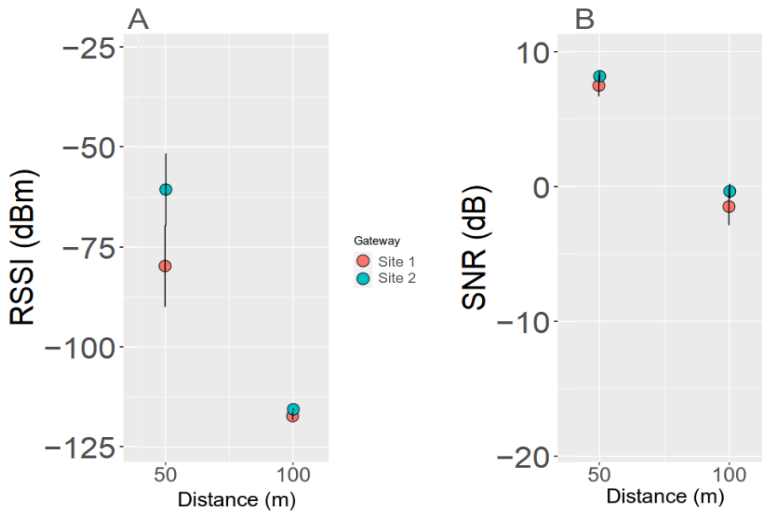

FIGURE S12 – Performance analysis of the LoRa gateways in the two sites (site 1 (Campus of Vitry) in pink, site 2 (residential area at Vitry) in light blue). (**A**) Received signal strength indicator (RSSI), (**B**) Signal-to-Noise Ratio (SNR).

5. Conclusion

Permettre la baignade en rivière est un enjeu à plusieurs niveaux : i) améliorer la qualité de l'eau, ii) favoriser le cadre de vie des habitants et iii) diminuer, *in fine*, les risques pour les futurs baigneurs, incluant dangers physiques, tels que les courants forts, les obstacles submergés ou les risques liés aux bateaux et aux infrastructures hydrauliques, nécessitant une vigilance accrue et le risque microbiologique. Un suivi de la qualité bactériologique est requis par la directive 2006/7/EC pour autoriser la baignade et assurer une surveillance continue. Dans le cadre de notre approche cela nécessite de trouver le bon équilibre entre :

- L'utilisation de systèmes de surveillance microbiologique de haute qualité comme ColiMinder, dont le coût est élevé, à des positions stratégiques au niveau de la rivière pour la zone de baignade.

- L'installation des capteurs physico-chimiques à faible coût sur un réseau Internet des Objets pour la prédiction de la qualité microbiologique permettant un suivi en continue des différents paramètres.

- En combinant des modèles d'apprentissage automatique et des capteurs comme outils de prédiction de la qualité microbiologique et d'alerte, il est possible d'optimiser, tant sur le plan temporel que spatial, l'effort d'échantillonnage effectué par des opérateurs humains.

Ces interventions sont particulièrement nécessaires lorsque le modèle ne parvient pas à estimer correctement la concentration en *E. coli*, contribuant ainsi à enrichir la base de données et à améliorer les performances des modèles de prédiction.

Les progrès de la technique de surveillance de la qualité de l'eau soulignent l'importance d'optimiser l'échantillonnage, car ceci impacte fortement le coût du suivi de routine (Jiang et al., 2020). S'ajoute à cela la mise en place d'un réseau de capteurs dont les données collectées serviront à alimenter des modèles de machine learning pour prédire la qualité de l'eau et optimiser les stratégies d'échantillonnage dans les zones de baignade urbaines. Un cadre flexible a été proposé au niveau de notre étude pour intégrer divers capteurs et créer des dispositifs adaptés à différents besoins. Notre étude a montré que les capteurs Arduino, à l'exception du capteur de turbidité, sont adaptés pour surveiller la qualité de l'eau. Le capteur de température à faible coût et les capteurs de pH se sont révélés fiables, bien que le capteur pH-1 nécessite un entretien mensuel. Le capteur de conductivité a offert des résultats plus variables, et le capteur d'oxygène dissous, bien qu'efficace, requiert un entretien régulier. Le capteur de turbidité, en revanche, est

trop instable et sensible à la lumière ambiante. Nous avons testé un système d'ombrage qui c'est révélé insuffisant. Il serait intéressant d'explorer l'utilisation d'un tube opaque pour protéger le capteur de la lumière. Corriger l'interférence crée par la lumière ambiante sur le signal pourrait constituer une autre piste d'amélioration comme suggéré par Trevathan et al. (2021). Concernant l'envoi des données, la communication via LoRa nécessite encore des améliorations, notamment en testant d'autres modules, avant son déploiement pour la surveillance en temps réel de la qualité de l'eau. En effet, la portée actuelle de communication avec les passerelles testées ne dépasse pas 200 mètres.

Avoir une meilleure représentation des données permettrait d'augmenter la performance et la fiabilité des modèles pour le développement d'un système de surveillance en temps réel et d'alerte précoce (Jiang et al., 2020). Dans le cadre des méthodes d'apprentissage automatique, les algorithmes reposent sur l'hypothèse que le jeu de données utilisées pour l'entraînement et pour le test présentent les mêmes caractéristiques (Noam, 2016). La méthode d'apprentissage par transfert repose sur cette hypothèse. Cependant, nos résultats montrent que transposer directement à un site un modèle entrainé sur un jeu de données d'un autre site ne donne pas toujours des résultats fiables. Une autre possibilité qui peut être explorée afin d'améliorer les performances des modèles de prédictions est de créer un méta-modèle regroupant par exemple 2 modèles ayant des bonnes performances de prédictions. En effet, nous avons constaté que les modèles peuvent être complémentaires pour certains paramètres. En sélectionnant deux modèles présentant un bonne performance de prédiction chacun, par exemple, Random forest et KNN, nous avons observé que pour 32,9% des mesures un des deux modèles permettait une estimation raisonable de la concentration en *E. coli*. Toutefois, l'apprentissage par transfert peut être plus optimisé en combinant un méta-modèle avec l'utilisation de larges bases de données publiques. Explorer d'autres bases de données avec un plus grand jeu de données permettrait d'augmenter les connaissances en identifiant les similarités avec notre cas d'étude. Ainsi grâce à l'apprentissage par transfert, les connaissances acquises à partir d'un jeu de données d'entraînement (source) peuvent être transférées sur un autre jeu de données (cible) (Noam, 2016). Cette stratégie peut offrir un avantage car l'entraînement de plusieurs modèles peut être gourmand en données et coûteux en temps de calcul. Ainsi, une collaboration a été mise en place avec le Syndicat Marne Vive pour explorer l'approche d'apprentissage par transfert à l'aide de base de données internationales pour prédire la qualité microbiologique de la Marne en utilisant la base de données de la nouvelle zélande, 3 fois plus grande (Balachandran et al.,

2022). Les résultats ont montré qu'en combinant les meilleurs modèles obtenus sur les deux jeux de données pour prédire la qualité de l'eau en Marne, on obtient un modèle plus précis, avec un RPD passant de 1,25 pour le meilleur modèle de machine learning initial (entraîné et testé sur la Marne) à 1,47 pour le métamodèle final.

Enfin, comme nous avons pu l'observer à travers les modèles de prédiction appliqués sur la Marne et la Seine, il existe un défaut de transférabilité d'un modèle entraîné sur une rivière vers une autre. Cette limitation pourrait être liée à des différences spécifiques entre les deux environnements. Ainsi, une meilleure compréhension des sources d'incertitude au niveau du modèle, ainsi qu'une étude approfondie de la dynamique des contaminations microbiologiques, permettrait potentiellement d'améliorer la précision des prédictions et de mieux saisir la variabilité des résultats entre les différentes rivières.

Chapitre 3 : Incertitudes et variabilité des dynamiques bactériologiques dans la surveillance des eaux de surface

1. Introduction

L'accès à une eau douce de qualité est fondamental pour la santé humaine et l'environnement. Environ 60 à 80% des besoins mondiaux en eau douce sont satisfaits par les eaux de surface (Bunsen et al., 2021), faisant de la gestion durable des ressources en eau une priorité de l'agenda 2030 des Nations Unies (Bunsen et al., 2021). Divers services écosystémiques et besoins sociaux dépendent de la qualité et de la quantité d'eau douce disponible : le soutien de la faune aquatique et de la biodiversité, l'irrigation, les activités récréatives, ainsi que les usages industriels (Giri, 2021).

En région parisienne, les activités récréatives en lien avec les eaux douces constituent un enjeu relativement ancien qui réémerge récemment. En effet, depuis le milieu du XIXe siècle, de nombreuses piscines municipales ont vu le jour le long des rives de la Seine et de la Marne. Les Parisiens ont ainsi commencé à profiter de ces espaces pour se détendre et se baigner (Pardailhé-Galabrun, 1983; Kistemann et al., 2016). Cependant, au début du XXe siècle, en raison de la mauvaise qualité des eaux, la baignade en Marne a été interdite dans le Val de Marne en 1970 par un arrêté préfectoral (Schaffner et al., 2009; Qin et al., 2011). Au cours des cinquante dernières années, l'amélioration de la qualité de l'eau a progressivement mis l'accent sur la qualité microbiologique, régulée par la directive 2006/7/CE pour les eaux de baignade (Schreiber et al., 2015). Ce désir politique et sociétal de reconquête des rivières urbaines pour la baignade est de plus en plus pressant en Île-de-France, tant pour la Seine que pour la Marne, avec l'ouverture de plusieurs sites de baignade prévue pour l'été 2025 en héritage des Jeux Olympiques et Paralympiques 2024.

Dans cette région, les projets de réhabilitation des rivières et de création de zones de baignade symbolisent cette volonté de reconquête. Ces initiatives s'inscrivent dans une démarche

de préservation des écosystèmes aquatiques et visent à restaurer la qualité des eaux, notamment en réduisant la contamination microbiologique d'origine anthropique, qui constitue une source majeure de risques pour la santé publique. L'ouverture de sites de baignade en Marne et en Seine illustre cette dynamique (Schaffner et al., 2009; Qin et al., 2011).

Bien que la qualité de l'eau se soit globalement améliorée en Europe depuis le XXe siècle, malgré la croissance démographique, cette amélioration est principalement attribuée à la gestion des sources de pollution ponctuelles, favorisée par les législations européennes (91/271/CEE, 2000/60/CE, 2006/07/CE). Ces législations ont conduit à une meilleure gestion des réseaux d'assainissement, à la modernisation des stations d'épuration et à une réduction des émissions polluantes (Mouchel et al., 2020; Whelan et al., 2022). Cependant, les sources de pollution diffuses demeurent problématiques et encore peu étudiées comparé aux sources ponctuelles, plus faciles à identifier (Garcia-Armisen and Servais, 2007; APE États-Unis, 2022). Or, les schémas de pluie et de ruissellement peuvent conduire à des déversements d'eaux usées non traitées dans les eaux de surface (Whelan et al., 2022) et la pollution chimique et microbienne limite l'utilisation de l'eau en raison des risques sanitaires, impactant ainsi l'état écologique des plans d'eau et des rivières.

L'intégration de l'incertitude dans les modèles de gestion de la qualité de l'eau est cruciale pour prendre des décisions éclairées sur l'ouverture des zones de baignade. Les évaluations de la qualité de l'eau, dans le cadre de programmes de gestion des ressources en eau, comparent souvent les concentrations mesurées d'indicateurs de contamination fécale à des normes de qualité établies sur la base de risques épidémiologiques (Benham et al., 2006; Gronewold and Wolpert, 2008). Toutefois, la variabilité méthodologique associée à la quantification des BIF peut avoir un impact significatif sur les actions de gestion (Griffin et al., 2001; McBride et al., 2003; Gronewold and Wolpert, 2008). Une meilleure compréhension des sources de variabilité, y compris celles introduites par les méthodes de mesure et les conditions environnementales, est nécessaire pour générer des décisions de gestion robustes, telles que l'ouverture ou la fermeture des sites de baignade.

La dynamique de dégradation de *E. coli* après un événement pluvial ou une pollution accidentelle est un autre facteur essentiel à considérer. Les modèles utilisés pour évaluer la qualité de l'eau doivent intégrer des paramètres reflétant le taux effectif de perte des BIF au fil du temps, en tenant compte de divers facteurs environnementaux (Auer and Niehaus, 1993; Ferguson et al., 2003). La modélisation de la décroissance bactérienne, souvent basée

sur des modèles de décroissance de premier ordre, est couramment appliquée dans les études sur la décroissance des BIF (également désigné comme un taux de « disparition » ou de « mortalité ») (Sinton et al., 1999; Noble et al., 2004). Ce taux varie selon différentes conditions environnementales, telles que l'irradiation solaire et la température de l'eau ; nous désignerons donc ce taux comme un « taux de disparition ». Cependant, la variabilité du taux de disparition des BIF en réponse à d'autres facteurs, y compris la concentration initiale, n'est pas encore bien comprise (Gronewold et al., 2011). La plupart des études sur la décroissance des bactéries ont été menées dans des conditions contrôlées en laboratoire ou *in situ* (Korajkic et al., 2014; Dick et al., 2010; Tijdens et al., 2008). Il est donc nécessaire d'incorporer aussi une approche qui prenne en compte l'effet de la variabilité des différents paramètres environnementaux sur le taux de disparition pour améliorer la précision des prévisions concernant la qualité de l'eau. Si les taux de décroissance pour *E. coli* sont relativement stables d'un pic de pollution à l'autre pour un même site, cela pourrait permettre d'avoir un outil utilisable par les gestionnaires dans le futur (Dick et al., 2010).

Ainsi, la prise en compte de l'incertitude et de la dynamique bactériologique dans l'évaluation de la qualité de l'eau est cruciale pour une évaluation précise et fiable des risques sanitaires. En intégrant ces éléments, il sera possible d'établir des prévisions plus précises concernant la qualité de l'eau et d'assurer la protection de la santé publique tout en favorisant la reconquête des rivières pour des usages récréatifs (Dick et al., 2010).

2. Optimisation de la classification des échantillons en intégrant par la logique floue l'incertitude de la mesure des indicateurs de contamination fécale

Manel Naloufi[1,2], Claire Thérial [2], Mohamed Saad [2], Sami Souihi[3], Thiago Wanderley Matos De Abreu [3], Miguel Gillon-Ritz [1], Marion Delarbre [1], Paul Kennouche [1], Françoise S. Lucas[2]

[1] Direction de la Propreté et de l'Eau - Service Technique de l'Eau et de l'Assainissement, 27 rue du Commandeur 75014 Paris, France ; manel.naloufi@paris.fr,

[2] Leesu, Université Paris-Est Créteil, École des Ponts ParisTech, 61 avenue du Général de Gaulle, 94010 Créteil Cedex, France ; lucas@u-pec.fr

[3] Image, Signal and Intelligent Systems (LiSSi) Laboratory, University of Paris-Est Créteil Val de Marne, 122 rue Paul Armangot, 94400 Vitry sur Seine, France ; thiago.wanderley-matos-de-abreu@u-pec.fr ; sami.souihi@u-pec.fr

Résumé : La gestion quotidienne des sites de baignade implique un suivi de la qualité microbiologique. Or, une incertitude de mesure peut exister au niveau des différentes étapes du processus, du prélèvement jusqu'à l'analyse de l'échantillon en laboratoire. En effet, la réglementation et les normes laissent une marge de liberté qui peut induire des pratiques différentes d'échantillonnage et d'analyse.

Dans notre étude, nous avons analysé la variabilité liée à la méthodologie de prélèvement ponctuel et automatique pour la mesure des bactéries indicatrices fécales réglementaires, de 3 indicateurs de sources animales, d'un indicateur de source humaine et de 2 pathogènes du genre *Campylobacter*. Aucune différence significative de concentration en BIF dans les eaux de surface n'a été constatée en comparant différents modes de prélèvement ponctuel depuis la berge (seau, bécher, pompe), quelque soit le site de prélèvement. Les résultats des équipements de prélèvement montraient que les rinçages avec l'eau du site préconisés par la réglementation étaient suffisants pour éviter les contaminations croisées pour les équipements de prélèvement ponctuel des eaux de surface même lorsque le site précédent était 10 fois plus contaminé. Pour le préleveur automatique entre deux prélèvements d'eau de surface, un nettoyage du système à l'eau stérile suffisait. Par contre, pour les eaux résiduaires, une désinfection à l'eau de Javel

suivie de 3 rinçages à l'eau stérile pouvait s'avérer nécessaire.

Les résultats indiquaient qu'il est recommandé de limiter le temps de stockage de l'échantillon, en privilégiant un transport réfrigéré. Une fois l'échantillon ensemencé sur le milieu de culture, le temps minimal d'incubation des BIF (24 h, 48 h, 72 h) présentait une variation inverse à la concentration de l'échantillon. L'identification et l'estimation de ces sources de variabilité permettront ainsi la mise en place d'un guide d'échantillonnage pour une surveillance optimale des sites de baignade.

L'intégration de la logique floue dans l'évaluation de la qualité de l'eau, notamment en ce qui concerne les concentrations en *Escherichia coli*, s'est révélée être une approche efficace pour la gestion des zones de baignade. En combinant des méthodes de défuzzification appropriées avec des dispositifs de surveillance en temps réel tels que le ColiMinder, il est possible de classer rapidement et de manière fiable les sites de baignade, tout en tenant compte des incertitudes inhérentes aux mesures. Les résultats obtenus montraient une forte concordance avec les méthodes couramment utilisées par les gestionnaires, offrant ainsi une évaluation plus nuancée des données et une prise de décision accélérée.

Mots clés : Bactéries indicatrices fécales ; incertitude ; échantillonnage, prélèvement, logique floue

2.1. Introduction

Différentes sources de contamination ponctuelles et diffuses peuvent apporter un flux de pathogènes au niveau des sites de baignade et ainsi générer un risque sanitaire lié au contact ou à l'ingestion des eaux contaminées (Guérineau et al., 2014). En zone urbaine, les rivières sont particulièrement sujettes à des dégradations de la qualité microbiologique lors des événements pluvieux qui génèrent du ruissellement sur des surfaces contaminées par des déjections animales et des rejets urbains de temps de pluie pouvant contenir des eaux usées non traitées. Ainsi, les principales sources de pathogènes d'origine hydrique sont les fèces humaines et animales, provenant d'individus porteurs sains ou malades (Passerat et al., 2011). Les sources animales typiquement associées aux contaminations fécales des eaux de surface en zone urbaine incluent les chiens, les chats et les oiseaux aquatiques, dont la présence peut contribuer à augmenter fortement les quantités en indicateurs bactériens de contamination fécale (Simpson et al., 2002;

Wright et al., 2009).

Dans le cadre de la directive européenne 2006/7/CE qui porte sur la gestion des eaux de baignade, la qualité microbiologique des eaux de surface est actuellement estimée à l'aide de deux groupes de bactéries, dites bactéries indicatrices fécales (BIF), les *Escherichia coli* et les entérocoques intestinaux (EI), dont l'analyse permet d'évaluer la conformité des eaux aux normes de qualité microbiologique. Il est possible de compléter le diagnostic de l'origine des sources de contamination en utilisant des bactéries ou des virus spécifiques des sources humaines ou animales (Devane et al., 2007).

La gestion quotidienne des zones de baignade nécessite donc un suivi régulier de la qualité microbiologique de l'eau pour limiter les risques sanitaires (OMS, 2018). En cours de saison, l'ouverture ou la fermeture d'un site de baignade est basée sur la confrontation des mesures de concentration en BIF à des valeurs seuils suivant l'instruction N° DGS/EA4/2022/168 du 17 juin 2022 relative aux modalités de recensement, gestion et classement des eaux de baignade. Toutefois, cette prise de décision peut être rendue délicate lorsque les concentrations mesurées sont proches des seuils, sachant qu'il existe plusieurs sources d'incertitude sur le prélèvement, le stockage et la mesure des BIF dans les échantillons d'eau de surface. Une surveillance optimale de la qualité ne peut être atteinte que si l'incertitude sur les niveaux de BIF mesurés est identifiée et qu'un moyen pour la réduire est considéré lors de l'échantillonnage et de la mesure. De ce fait, les laboratoires habilités pour le suivi de la qualité des eaux de baignade sont accrédités et les méthodes d'échantillonnage et d'analyse sont normées. Malgré tout, il subsiste une incertitude non négligeable qui peut rendre l'interprétation des résultats délicate. Pour la gestion quotidienne des sites de baignade, une prise de décision éclairée est nécessaire, souvent face à une incertitude significative (Brandão et al., 2022). L'hétérogénéité et l'incertitude liées aux échantillons compliquent cette tâche lorsque la qualité de l'eau est proche d'un seuil, car des valeurs peuvent chevaucher les valeurs limites réglementaires, créant des doutes quant à leur conformité (Rabinovici et al., 2004).

En effet, les textes réglementaires et normatifs fournissent un corpus de recommandations pour le prélèvement d'eau de baignade et permettent ainsi d'avoir un référentiel commun pour l'analyse de la qualité microbiologique de l'eau. Par exemple en France, dans le cadre Européen il existe un corpus réglementaire traduit en droit français et des guides produits par les Agences de l'eau qui donnent des directives sur le prélèvement et l'analyse, tels que la directive européenne 2006/7/EC, le guide de prélèvement de l'Agence de l'Eau Loire-Bretagne (AELB, 2006), ainsi

que l'ensemble de normes NF EN ISO 19458, FD T 90-521 et FD T 90-523-1. Ce référentiel contribue à diminuer l'incertitude sur l'échantillonnage, mais même avec ces instructions il existe une certaine liberté d'interprétation et d'adaptation entrainant potentiellement une incertitude. Ainsi, différents équipements peuvent être utilisés lors du prélèvement ponctuel : flacon ou pompe reliés à une perche télescopique depuis la berge ou encore seau lesté lancé depuis un pont ou une berge (guide FD T 90-523-1). Avec la perche, le prélèvement peut s'effectuer directement dans le flacon stérile ou par un flacon de prélèvement intermédiaire (que nous appellerons bécher) pour transvaser dans le flacon stérile (guide FD T 90-523-1). De plus, des contaminations croisées des équipements peuvent avoir lieu entre deux sites consécutifs et des procédures de désinfection avec des lingettes et/ou de rinçage avec l'eau du site sont prescrites. Plusieurs pratiques sont aussi constatées sur le terrain quant aux conditions de transport et de stockage des échantillons avant l'analyse. Ainsi, entre les différents textes francais, la température et le temps de transport recommandés sont de 5 ± 3°C avec un temps de retour au laboratoire le plus rapide et une analyse au plus tard dans les 24 heures. À celà s'ajoute des contraintes de terrain et d'éloignement des sites qui demandent parfois une adaptation de la part des personnels. L'ensemencement et la lecture des milieux comportent également une certaine incertitude liée aux erreurs de manipulation, à l'homogénéisation de l'échantillon, à l'équipement et aux réactifs. Dans le cadre du suivi réglementaire, l'incertitude est minimisée par le fait que des laboratoires accrédités réalisent le suivi. Par contre, pour les échantillonnages avec les préleveurs automatiques, il n'existe pas de protocole normé et l'incertitude peut atteindre des valeurs de l'ordre de 15 à 67% (McCarthy et al., 2008). Lorsque les indicateurs spécifiques de sources humaines ou animales (bactériens ou viraux) sont suivis, la mesure en laboratoire ne fait pas l'objet de normes (hormis le marqueur humain HF183 aux USA qui fait l'objet d'une norme US-EPA 1696.1) et en dehors du guide MIQE pour l'analyse en PCR quantitative en temps réel (qPCR), il n'y a pas ou peu d'effort de normalisation ou de tests interlaboratoires (EPA, 2019; Layton et al., 2013; Lane, 2019; Ahmed et al., 2020).

Mieux définir l'incertitude associée à la surveillance des BIF et des marqueurs bactériens ou viraux spécifiques de sources de contamination permettra une amélioration des bases scientifiques des normes et des réglementations en vigueur. Intégrer cette notion d'incertitude dans la prise de décision lors de la gestion quotidienne des sites de baignade est essentiel pour garantir la sécurité des usagers et prévenir les risques sanitaires. Les décisions liées à la qualité de l'eau ne se limitent pas à des critères quantitatifs simples mais intègrent également des éléments

plus subjectifs qui peuvent rajouter de l'ambiguïté. Dans cette optique, Gharibi et al. (2012) soulignent l'importance d'une approche fondée sur la logique floue qui permet de modéliser l'incertitude dans l'évaluation de la qualité de l'eau comme la qualité microbiologique. La logique floue (Zadeh, 1965) est devenue une approche courante particulièrement adaptée pour des indices environnementaux. Cette approche se base sur une méthodologie robuste, intégrant les incertitudes associées aux données et aux décisions (Ross, 2005). Elle permet d'incorporer les réflexions et l'expertise humaine dans les indices, ce qui facilite la gestion d'informations non linéaires, incertaines, ambiguës et subjectives (Gharibi et al., 2012). Un aspect particulièrement délicat, principalement pour les gestionnaires, est que lorsque les valeurs mesurées approchent la valeur seuil, la prise de décision devient complexe. Dans ces situations, la logique floue offre un cadre permettant d'évaluer les nuances et de rationaliser le processus décisionnel en tenant compte de cette incertitude supplémentaire. Quelle que soit la source d'information, elle sera associée à un certain degré d'incertitude (Ross, 2005).

Dans l'évaluation de la conformité avec une limite de spécification supérieure, des scénarios typiques émergent lorsque les résultats de mesure et leurs incertitudes sont pris en compte. Lorsque la valeur mesurée, ajoutée à son incertitude, est clairement au-dessus ou en dessous de cette limite, la décision est évidente. Cependant, des erreurs peuvent survenir en raison du chevauchement partiel des bandes d'incertitude autour des limites de spécification (Brandão et al., 2022). De plus, un article souligne que «le problème de la prise de décision sous incertitude est que la majorité des informations que nous avons sur les résultats possibles est généralement vague, ambiguë et autrement floue» (Ross, 2005). Intégrer ces perspectives dans le cadre de l'évaluation de la qualité microbiologique des eaux de baignade vise à établir un cadre de décision robuste qui tienne compte de la complexité et de l'incertitude inhérentes aux contextes de baignade. Deux études offrent un éclairage précieux sur l'intégration des méthodes de décision dans des contextes environnementaux complexes, renforçant ainsi la pertinence de cette approche (Ross, 2005; Zhou and Chen, 2023).

L'objectif de notre étude est donc de proposer une nouvelle approche de prise de décision sous incertitude utilisant la logique floue pour aider à la gestion quotidienne des eaux de baignade. Dans un premier temps, les sources de variabilité seront identifiées et l'incertitude associée à l'échantillonnage dans l'analyse des BIF, ainsi que dans celle de six marqueurs de contamination fécale d'origine animale et humaine, sera quantifiée. Pour ce faire, nous avons procédé à une analyse statistique des modalités de prélèvement, des protocoles de nettoyage,

ainsi que de toutes les étapes de transport et de stockage des échantillons, en incluant la recherche et le dénombrement des BIF. Cette démarche vise à établir une incertitude globale qui facilitera dans un deuxième temps la mise en place d'une prise de décision s'appuyant sur une approche de logique floue. Pour tester l'efficacité de cette nouvelle approche de classement des échantillons selon la réglementation française pour la gestion quotidienne, nous utiliserons les données provenant du dispositif de suivi en continu de la qualité microbiologique, ColiMinder, installé sur plusieurs sites dans la Seine et la Marne en région parisienne (France). Placé en amont d'un site de baignade, ce type d'équipement permet d'accélérer le processus décisionnel d'ouverture et de fermeture du site le jour même, et de rationaliser l'effort d'échantillonnage supplémentaire avec les analyses réglementaires qui rendront un résultat au plus tôt 18 h ou 24 h plus tard (Angelotti et al., 2022). Cet équipement installé sur berge estime la concentration en BIF à partir de mesures enzymatiques sur un volume d'eau prélevé et filtré (Cazals et al., 2020).

2.2. Matériel et méthodes

2.2.1. Site d'échantillonnage

Au cours de cette étude, des prélèvements d'eau de surface ont été effectués de mai à octobre sur 4 sites en Ile-de-France (France) entre 2022 et 2023. Ces sites représentent un gradient de concentrations en BIF allant d'un site de baignade de bonne qualité microbiologique à des eaux usées non-traitées. Deux sites étaient situés au lac de Créteil (Val-de-Marne, France), où les concentrations en BIF étaient 10 fois plus élevées sur le site 2 que sur le site 1. Ces deux sites lacustres ont fait l'objet de prélèvements ponctuels depuis la berge avec différents équipements. Des prélèvements moyens sur 24 h à l'aide d'un préleveur automatique Bühler 2000 réfrigéré (Hach) ont été réalisés au pont de Crimée sur l'eau du canal de l'Ourcq au niveau du second bassin de la Villette (Paris, France) et dans la Marne à Saint-Maur-des-Fossés (Val-de-Marne). De plus, les eaux brutes en entrée de la station de traitement des eaux usées de Saint-Thibault-des-Vignes (Seine-Saint-Denis) et les eaux pluviales en amont du bassin de rétention de Sucy-en-Brie (Val-de-Marne) et au rejet de l'ouvrage cadre du centre urbain de Noisy-le-Grand (Seine-Saint-Denis) ont également été prélevées pendant 24 h à l'aide d'un préleveur automatique. Différentes techniques et protocoles ont été testés afin d'étudier la variabilité liée à la méthodologie de prélèvement, de transport, de stockage des échantillons et de mesure des BIF, des marqueurs de contamination fécale animale et humaine et des pathogènes du genre

Campylobacter. Pour les protocoles d'échantillonnage, de lavage, de transport et de stockage des échantillons, un total de 5 prélèvements a été effectué avec chaque équipement au niveau de chaque site.

De plus, nous avons exploité les résultats entre 2020 et 2023 des prélèvements hebdomadaires ou bi- hebdomadaires effectués par la Ville de Paris entre 7 h et 12 h au niveau de 3 sites dans la Seine (Pont de l'Alma et Pont de Tolbiac en rive gauche et en rive droite ; Paris, France). Les mesures ont été effectuées selon la méthode de référence NF EN ISO 9308-3 pour *E. coli*. Nous avons également utilisé des mesures estivales effectuées entre 2020 et 2023 par le système de suivi automatisé ColiMinder (Vienna Water Monitoring, VWM) au niveau de 2 sites en Seine (en rive gauche à Pont de l'Alma et Pont de Tolbiac à Paris) mis en place par la Ville de Paris.

2.2.2. Équipements d'échantillonnage ponctuel depuis la berge

Les eaux de surface du lac de Créteil ont été prélevées dans les 30 premiers cm à 1-2 m de la berge selon la norme FD T90-523-1. Selon cette même norme, trois équipements peuvent être employés pour le prélèvement ponctuel : un bécher associé à une perche (ou canne) télescopique, une pompe dont le tuyau est associé à une perche télescopique et un seau lancé depuis un pont ou une berge. Ainsi, au niveau des 2 sites du lac de Créteil, ces 3 équipements ont été testés depuis la berge. Nous avons choisi un milieu lentique pour nous affranchir d'une trop grande hétérogénéité spatio-temporelle entre chaque prélèvement comme dans le cas d'une rivière. L'analyse de l'incertitude a été réalisée en utilisant l'équation 3.1.

2.2.3. Protocole de nettoyage des équipements d'échantillonnage ponctuel

Il est généralement recommandé d'avoir les mains propres, de nettoyer le matériel de prélèvement, d'utiliser un flaconnage stérile et d'effectuer le prélèvement de manière aseptique, mais sans plus de précision (norme de prélèvement FD T90-521, Directive 2006/7/CE, Arrêté du 19 octobre 2017 sur les méthodes d'analyse pour le contrôle sanitaire des eaux). Le guide des directions régionales et départementales des affaires sanitaires et sociales de la région Rhône-Alpes (2006) précise de flamber la canne télescopique sur la partie en contact avec l'eau ou de désinfecter avec un produit adapté. Le guide FD T90-523-1 préconise pour la pompe de laisser couler l'eau le temps nécessaire pour rincer le tuyau avant le prélèvement et pour le bécher intermédiaire de bien le rincer avec l'eau du site. Sur la base de ces recommandations, nous

avons testé plusieurs protocoles de désinfection et/ou rinçage du bécher de prélèvement et du tuyau de la pompe, en simulant un risque de contamination croisée entre deux sites avec un écart de contamination d'1 Log_{10} en concentration de BIF.

Ainsi au lac de Créteil, la contamination croisée a été simulée en prélevant d'abord sur un site peu contaminé (site 1), puis sur un site 10 fois plus contaminé (site 2) et à nouveau au site 1, en utilisant les mêmes équipements ayant subi ou non un protocole de nettoyage. Trois protocoles de nettoyage ont été testés pour le tuyau de la pompe et pour le bécher : i) rinçage 3 fois à l'eau du lac avant prélèvement, ii) désinfection à l'éthanol et séchage à l'air ou iii) désinfection à l'éthanol puis rinçage 3 fois avec l'eau du site de prélèvement. L'analyse de l'incertitude a été réalisée en utilisant l'équation 3.2.

2.2.4. Protocole de nettoyage du préleveur automatique

Les prélèveurs automatiques sont utiles pour effectuer des échantillonnages sur un intervalle de temps (par exemple un échantillon moyen sur 24 h) ou pour échantillonner un événement pluvieux au pas de temps ou en fonction d'un débit, ou d'un seuil. Le prélèvement peut être initié à un temps donné ou sur des paramètres hydrologiques (débit, hauteur d'eau...). Par ailleurs, le pourcentage d'incertitude ne semble pas dépendre de la concentration en BIF de chaque site comme il a pu être démontré sur les rejets pluviaux (McCarthy et al., 2008). Toutefois, une contamination du système de prélèvement peut survenir après l'échantillonnage en contaminant les échantillons suivants (Hathaway et al., 2014).

Le système de prélèvement de ces échantillonneurs équipés d'une pompe péristaltique, comporte un tuyau d'aspiration, un tuyau d'écrasement, un bol de prélèvement, un logement pouvant accueillir une rosette de 24 flacons d'un litre. Ces éléments peuvent générer une contamination, il est donc nécessaire de bien les nettoyer (Wilson et al., 2024). La réglementation française est assez succincte quant à leur protocole de nettoyage pour l'analyse microbiologique. Le guide FD T90-523-1 et celui de l'Agence de l'eau Loire-Bretagne ne donnent pas de recommandation sur le nettoyage mais plus sur l'installation du tuyau de prélèvement et les conditions de stockage des flacons de prélèvement. L'Institut d'études géologiques des États-Unis (USGS) propose un guide plus détaillé dans lequel il est recommandé, entre chaque prélèvement, de démonter le système de prélèvement pour autoclaver les flacons et les tuyaux, et, si nécessaire, d'utiliser une solution de Javel domestique diluée à 5%, suivie de plusieurs rinçages à l'eau désionisée (Wilson et al., 2024). Des blancs de terrain du système de prélèvement servent de

contrôle qualité.

Le but de notre expérience est de déterminer s'il existe une contamination résiduelle après un prélèvement et si le protocole de rinçage à l'eau du site programmé sur le préleveur suffit, ou s'il est nécessaire d'effectuer une désinfection à l'eau de Javel puis un rinçage à l'eau stérile. Pour ce faire, pour les eaux de surface du canal de l'Ourcq, nous avons comparé les niveaux de BIF mesurés dans des prélèvements ponctuels pris au préleveur automatique (Bühler 2000 réfrigéré Hach) avant et après désinfection suivis de rinçages à l'eau du robinet stérile. Le préleveur automatique est équipé de deux électrodes de remplissage conductrices qui nécessitent une conductivité minimale de 50 μS/cm pour détecter correctement le niveau de liquide et gérer les prélèvements (Lange, 2012). Or, l'eau distillée, en raison de sa conductivité très faible (généralement inférieure à 1 μS/cm), ne permet pas le bon fonctionnement de ces électrodes. Pour éviter ces problèmes, nous avons utilisé de l'eau du robinet autoclavée pendant 20 minutes à 120°C qui présente une conductivité suffisante. Un prélèvement manuel au bécher directement à côté du tuyau du préleveur représentait la référence avec laquelle les prélèvements à l'échantillonneur Hach ont été comparés (avant et après stérilisation et rinçage). En effet, il a été démontré que lors de prélèvements ponctuels par méthode manuelle au bécher et par préleveur automatique, les concentrations en BIF ne différaient pas significativement (Ferguson, 1994; Galfi et al., 2014). Le système de pompage (tuyau et bol) du préleveur automatique a été nettoyé avec l'eau de Javel à 0,5% (degré chlorhydrique) suivi de 3 rinçages à l'eau du robinet autoclavée 20 min à 120°C. Ces tests ont été effectués 1 à 63 jours après un prélèvement automatique pour vérifier si une contamination résiduelle persiste entre deux prélèvements plus ou moins éloignés dans le temps (par exemple entre deux événements pluvieux). L'analyse de l'incertitude a été réalisée en utilisant l'équation 3.2. De plus, des blancs de terrain ont également été réalisés en prélevant de l'eau du robinet stérile soit après désinfection suivie de trois rinçages à l'eau du robinet autoclavée, soit directement après un prélèvement d'eau sur site, ou encore 3 à 10 jours après le dernier prélèvement.

2.2.5. Protocole de transport et stockage

Selon la directive 2006/7/CE, l'analyse en laboratoire doit être effectuée le plus rapidement possible après prélèvement. Cependant, le transport peut parfois prendre du temps suivant l'éloignement du site d'échantillonnage et la circulation routière. Selon la norme FD T90-523-1, l'échantillon doit être conservé à une température de 5 ± 3 °C et l'analyse doit être effectuée

au mieux dans les 8 h et au plus tard dans les 24 h. Cependant, le système de réfrigération des échantillons peut dysfonctionner ou ne pas être installé en raison des limitations de l'infrastructure de la zone d'installation. Ainsi, l'effet de la réfrigération pendant le transport et le stockage des échantillons a été testé au lac de Créteil sur le site 2 avec un temps de transport de 0,5 h et 6 h à 5°C (glacière) ou à température ambiante (19,2 ± 2,4°C), à l'ombre, avec un ensemencement immédiatement dès le retour au laboratoire ainsi qu'après 24 h de stockage des échantillons au réfrigérateur à 5°C. L'analyse de l'incertitude a été réalisée en utilisant l'équation 3.2.

2.2.6. Dénombrement des BIF

Afin d'estimer la concentration (exprimée en nombre le plus probable NPP/100 mL), d'*E. coli* et des EI, les échantillons ont été ensemencés sur les microplaques MUG/EC et MUD/SF (BioRad) selon la méthode de référence NF EN ISO 9308-3 pour *E. coli* et NF EN ISO 7899-1 pour EI. Les microplaques ont été incubées à 44°C pendant 24 à 48 h selon les normes précitées et le nombre de puits positifs a été dénombré sous lampe UV. Le calcul du NPP/100 mL dans un intervalle de confiance de 95% a été réalisé à l'aide d'une feuille de calcul Excel® publiée par Jarvis et al. (2010).

2.2.7. Incertitude analytique et temps d'incubation

Nous avons procédé à une analyse de l'incertitude analytique par ensemencement d'un même échantillon sur 5 microplaques différentes représentant 5 réplicats. Pour chaque réplicat, des dilutions ont été préparées à nouveau, en homogénéisant entre chaque dilution et avant ensemencement. L'analyse de l'incertitude a été réalisée en utilisant l'équation 3.1.

Lors du dénombrement des BIF, les normes NF EN ISO 9308-3 et NF EN ISO 7899-1 précisent que la lecture doit être réalisée au minimum 36 h et au maximum 72 h après l'ensemencement. Cependant, dans la pratique, une lecture est souvent faite dès 24 h. La variabilité des concentrations liées au temps d'incubation a été évaluée sur des échantillons représentant un gradient de contamination : les eaux usées en entrée et sortie de la station de traitement des eaux usées de Saint-Thibault-des Vignes (échantillon moyen sur 24 h), les eaux de surface du canal de l'Ourcq (échantillon moyen sur 24 h) et les eaux de surface du Lac de Créteil (échantillons ponctuels au bécher ou à la pompe). Les microplaques MUG/EC et MUD/SF ont été lues après 24, 48 et 72 h.

2.2.8. Extraction et quantification de l'ADN

La variabilité liée au prélèvement, transport et stockage a également été évaluée pour des indicateurs bactériens de contamination fécale humaine ou animale (oies bernaches, chiens, mouettes et goélands), deux espèces du genre *Campylobacter* et pour les bactéries totales. Dans ce cas, uniquement 2 prélèvements sur 5 ont été utilisés. La filtration des différents échantillons a été réalisée sur des cartouches Sterivex^{TM} de 0.22 μm de porosité (Milipore) qui ont été stockées à -20° C avant extraction de l'ADN. Chaque Sterivex^{TM} a été ouvert stérilement et le filtre a été découpé en morceaux d'environ 1-2 mm à l'aide d'un scalpel stérilisé, selon Roguet (2015). Les fragments ont été insérés dans un tube Lysing Matrix E du kit FAST DNA SPIN KIT for soil (MP Biomedical) pour en extraire l'ADN total selon les instructions du fabricant modifiées par Roguet (2015). La concentration et la pureté de l'ADN extrait ont été mesurées à 230, 260 et 280 nm avec un spectrophotomètre (WPA, BioWave DNA). Puis l'ADN a été stocké à -20°C en attendant son amplification.

2.2.9. Amplification des marqueurs spécifiques et des pathogènes

Six espèces bactériennes ont été utilisées comme marqueurs de contamination fécale animale et humaine. Pour les marqueurs bactériens suivants, le gène de l'ARNr 16S a été amplifié et quantifié : *Catellicoccus marimammalium* (*Bacillota*) pour les mouettes et goélands (Gull2, Ryu et al. (2012)) ; et trois espèces du groupe des *Bacteroidales* pour les chiens (BacCan,Kildare et al. (2007)), les oies bernaches (CGOF1, Fremaux et al. (2010)), et les humains (HF183, Green et al. (2014)) . De plus, les pathogènes du genre *Campylobacter* ont été quantifiés en amplifiant un fragment du gène hipO codant pour l'hippurate hydrolase pour *C. jejuni* et un fragment du gène codant pour la peptidase T pour *C. lari* (He et al., 2010; Vondrakova et al., 2014). En plus de ces marqueurs, une estimation de la charge bactérienne totale a été réalisée par une quantification du nombre de copies du gène de l'ARNr 16S (BactQuant, Liu et al. (2012)).

La PCR quantitative en temps réel (qPCR) a été réalisée sur les ADN extraits en utilisant le thermocycleur CFX96 (BioRad). Un contrôle positif interne (β-actine) a été ajouté pour évaluer la présence d'inhibiteurs résiduels selon Wurtzer et al. (2014). Un contrôle négatif (eau stérile) a été inclus également lors des amplifications. Pour chaque cycle, des courbes standard en triple ont été générées (de 10^1 à 10^7 copies/µL) en utilisant un plasmide linéarisé ou des gBlocks contenant la séquence cible. Un coefficient de corrélation supérieur à 0,995 a été observé pour

chaque courbe standard de dosage. Les réactions contenaient 1X de iTaq^{TM} Universal probes Supermix (Bio Rad), les amorces sens et antisens de la β-actine et de la cible bactérienne, une sonde à hydrolyse spécifique de chaque cible, 10^4 copies/µl d'ADN de β-actine et 1 µL d'ADN matrice pour un volume total de 20 µL. Le tableau S1 montre les séquences des amorces et des sondes et leur concentration finale. Les réactions ont été soumises à une dénaturation initiale à 95°C pendant 10 min puis 40 cycles de dénaturation à 95°C pendant 20 sec et hybridation et élongation à 60°C pendant 1 min.

2.2.10. Analyse statistique

Afin d'évaluer l'effet significatif de chaque protocole et technique, une analyse statistique a été effectuée avec le logiciel R pour les BIF par des tests appariés : de Friedman, de Wilcoxon ou de Student (R Core Team, 2021). La normalité des données a été vérifiée avec un test de Shapiro-Wilk. Dans le cas des tests de Wilcoxon ou test t répétés pour tous les échantillons 2 à 2, la correction de Bonferroni a été appliquée. Pour tous les tests statistiques, le niveau de signification était basé sur 5%.

2.2.11. Analyse et estimation de l'incertitude

L'incertitude représente le manque partiel de connaissance, réduite par une amélioration de la collecte de données. Afin d'estimer le pourcentage d'incertitude au niveau de la mesure de la concentration en BIF et de la quantification des indicateurs de sources et des 2 pathogènes du genre Campylobacter. Une mesure du pourcentage d'erreur relative d'échantillonnage a été réalisée (Harmel et al., 2016; Esbensen and Wagner, 2014). Pour cela, 3 équations différentes peuvent être utilisées en fonction des données à disposition.

Pour plusieurs échantillons :

$$\pm\ \%\mathrm{Inc} = \frac{2 * ecartype(X_i)}{moyenne(X_i)} * 100 \tag{3.1}$$

Si X_{rf} est la valeur de référence :

$$\%\mathrm{Inc} = \frac{X_2 - X_{rf}}{X_{rf}} * 100 \tag{3.2}$$

Si la valeur de référence est inconnue :

$$\pm\ \%\mathrm{Inc} = \frac{|X_2 - X_1|}{moyenne(X_1, X_2)} * 100 \tag{3.3}$$

Au niveau de ces formules, %Inc représente le pourcentage d'incertitude, (X_i, X_1 et X_2) sont des valeurs de concentration d'un échantillon et X_{rf} est supposée être la vraie valeur (échantillon de référence).

2.2.12. Prise de décision sous incertitude

Afin de proposer une aide à l'utilisation de la qualification des échantillons en cours de saison en tenant compte de l'incertitude liée au prélèvement et à l'analyse, nous avons implémenté un outil d'aide à la décision basé sur la logique floue. Les seuils utilisés sont de 100 et 1800 NPP/100 mL servant au classement des échantillons ponctuels en cours de saison sur un site de baignade classé (Instruction DGS/EA4/2022/168 du 17 juin 2022 relative aux modalités de recensement, gestion et classement des eaux de baignade). Pour ce faire, nous avons utilisé les résultats des prélèvements hebdomadaires ou bi-hebdomadaires au niveau de 3 sites en Seine (Pont de l'Alma et Pont de Tolbiac en rive gauche et en rive droite). Au niveau du pont de Tolbiac, un système de mesure en continu ColiMinder avait également été installé sur la rive gauche et a servi à classer les deux sites rive-gauche et rive-droite. En effet, il n'y avait pas de différence significative entre les 2 sites avec les mesures réglementaires (Test de Wilcoxon, p>0.57, n=547). L'intégration de l'incertitude dans la classification de la qualité de l'eau a été réalisée en appliquant une approche basée sur la logique floue (Ross, 2005). L'incertitude globale, de l'échantillonnage à la lecture des milieux de culture, a été calculée comme la racine carrée de la somme quadratique de l'incertitude sur le nettoyage, le stockage et la mesure, conformément à la méthode proposée par Brandão et al. (2022). Cette approche permet d'englober l'ensemble des incertitudes associées à la mesure de la concentration d'*E. coli* obtenue.

La logique floue a ensuite été utilisée pour intégrer ces incertitudes dans les valeurs seuils utilisées en cours de saison, selon l'instruction n°DGS/EA4/2020/111 du 2 juillet 2020 : Bonne (<100 NPP/100 mL), Moyenne (<1800 NPP/100 mL) et Mauvaise (>1800 NPP/100 mL). Nous nous sommes focalisés sur le paramètre le plus déclassant pour ces sites en Seine, à savoir la concentration en *E. coli* (NPP/100 mL). Des ensembles flous ont été définis pour refléter les

niveaux de qualité avec des sous-ensembles se chevauchant pour représenter l'incertitude dans les valeurs seuils. Les fonctions d'appartenance ont été définies à l'aide de fonctions sigmoïdes, permettant d'incorporer un pourcentage d'incertitude associé à chaque sous-ensemble. Chaque mesure s'est vue attribuer un degré d'appartenance à plusieurs ensembles flous simultanément. Ensuite, l'inférence floue a été effectuée en utilisant la méthode de défuzzification pour prendre la décision (Ross, 2005). Une méthode dite de défuzzification a permis de convertir les sorties floues résultant du moteur d'inférence floue en une valeur numérique non floue. Il existe plusieurs méthodes de défuzzification, telles que le centre de gravité (COG), le bisecteur (BS), la moyenne des maxima (MOM), le maximum le plus à gauche (LOM) et le maximum le plus à droite (ROM) (Akkurt et al., 2004; Jantzen, 1999). Parmi ces méthodes, le COG, souvent appelé méthode du centroïde, est la plus couramment utilisée. Cette méthode calcule le barycentre des valeurs d'appartenance et fournit ainsi une estimation précise de la qualité de l'eau (Ross, 2005). La méthode du bisecteur (BS) divise l'aire sous la courbe d'appartenance en deux parties égales pour estimer le résultat. Quant à la méthode de la moyenne des maxima (MOM), elle prend la moyenne des points où la fonction d'appartenance atteint son maximum, tandis que les méthodes du maximum le plus à gauche (LOM) et du maximum le plus à droite (ROM) sélectionnent respectivement le premier et le dernier maximum rencontré. Ces méthodes permettent chacune une approche différente de la défuzzification et ont été toutes testées dans cette étude pour évaluer la qualité de l'eau, en fonction des critères flous définis.

Afin d'appliquer les méthodes de défuzzification, nous avons utilisé les données collectées par le système ColiMinder avec la valeur d'incertitude associée. La qualité de l'eau a été évaluée pour chaque jour en testant la méthode de logique floue sur plusieurs intervalles de temps permettant de classer la qualité de l'eau pour le matin (8 h) ou l'après-midi (12 h). Nous avons testé des intervalles de 24 heures (de midi la veille à midi, ou de 8 h la veille à 8 h), des intervalles de 12 heures (de minuit à midi, ou de 20 h la veille à 8 h) permettant une analyse temporelle plus fine des variations de la qualité de l'eau dans les différentes stations de prélèvement. De plus, un intervalle de 4 heures (de 4 h à 8 h et de 8 h à 12 h) comparable à celui utilisé par la Ville de Paris durant les jeux olympiques a également été testé (Desir, 2024). Un total de 6 intervalles de temps a été testé.

Les résultats de classification obtenus avec les 5 méthodes de défuzzification mises en œuvre sur les données du ColiMinder ont été comparés en utilisant deux approches : i) la classe identifiée après défuzzification a été comparée à la classification des données du

suivi réglementaire par rapport aux valeurs seuils de gestion en cours de saison (Instruction n°DGS/EA4/2020/111 du 2 juillet 2020), ii) la classe de défuzzification est comparée au résultat de la méthode de classification utilisée par la Ville de Paris pendant les Jeux Olympiques 2024. Cette dernière calcule une moyenne glissante de 4 heures, qui est comparée à celle des 24 heures précédentes, pour évaluer la dégradation, l'amélioration ou la stabilité de la qualité de l'eau, et réalise une classification (Desir, 2024). Ceci a permis de valider l'efficacité de la logique floue sous incertitude par rapport à la méthode réglementaire dont les résultats ne sont généralement disponibles qu'après 36 h d'incubation des microplaques.

2.3. Résultats et discussion

2.3.1. Variabilité liée à l'équipement pour le prélèvement ponctuel

Le type d'équipement utilisé (bécher, pompe, seau) pour le prélèvement ponctuel depuis la berge ne semble pas avoir d'impact sur les concentrations en BIF de l'eau de surface. En effet, aucune différence significative n'a été observée entre les résultats obtenus avec les trois systèmes, quel que soit le site de prélèvement (2 sites du lac de Créteil) et que ce soit pour les *E. coli* ou pour les EI (Test de Friedman apparié, n=30, $p>0,05$, Figure 3.1A). Ce résultat conforte le fait que le guide FD T 90-523-1 propose ces trois équipements comme options possibles pour le prélèvement. Toutefois, il est recommandé de n'utiliser le seau qu'en dernier recours du fait de la difficulté à maintenir propre cet équipement. La suite de l'étude se concentrera donc sur la pompe et le bécher comme méthode de prélèvement. Nos résultats montrent globalement que les incertitudes sur la mesure des indicateurs étaient similaires entre les équipements, confortant ainsi la polyvalence du protocole (AELB, 2006). Combinés à une perche, ces équipements permettent d'effectuer un prélèvement à 2 m de la berge et dans les 30 premiers centimètres (AFNOR). La variabilité estimée pour les deux équipements était la plus faible pour les BIF, le marqueur Humain HF183 et le marqueur oie CGOF1 (Figure 3.1B).

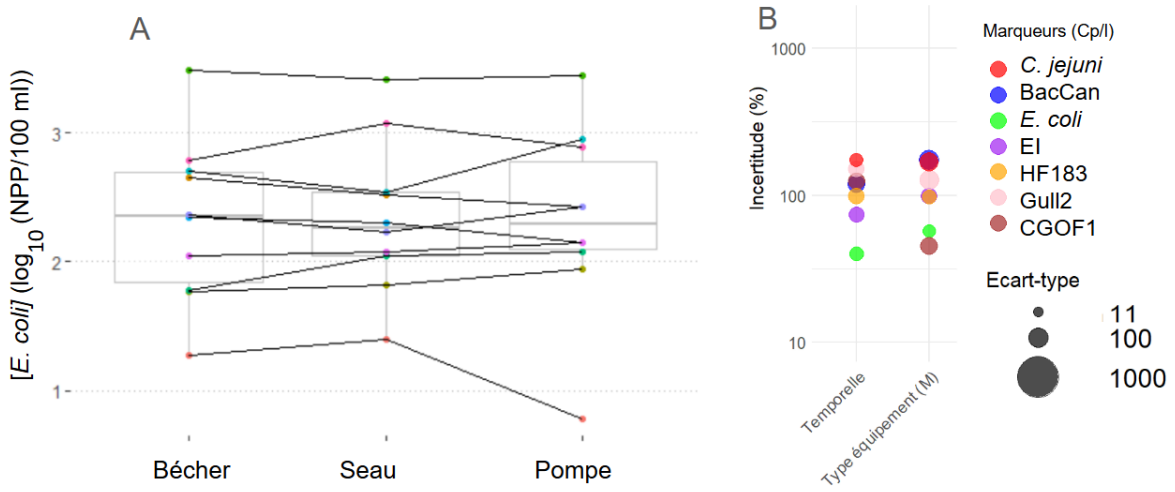

FIGURE 3.1 – Comparaison des équipements de prélèvement ponctuel depuis la berge au niveau des 2 sites du lac de Créteil. (A) concentration en *E. coli* en NPP/100 ml, (B) Pourcentage moyen d'incertitude lié aux équipements et à la variabilité temporelle. La taille des cercles représente l'écart type du pourcentage d'incertitude.

2.3.2. Répétabilité dans le temps

La mise en oeuvre d'échantillonnages répétés sur un intervalle de temps court permet d'évaluer une partie de l'incertitude liée à l'échantillonnage que l'on nomme la répétabilité. Si l'intervalle de temps est plus long, il est possible alors d'évaluer l'incertitude temporelle de la qualité de l'eau. Dans la littérature, il est rapporté une incertitude moyenne de répétabilité pour *E. coli* de ±(23 ± 16)%, pour un intervalle de 1 minute d'échantillonnage ponctuel avec des flacons plongés dans l'eau de rivière (Pendergrass et al., 2015; Harmel et al., 2016). Dans notre étude, l'incertitude liée à la variabilité temporelle (toutes les 10 minutes) a été estimée avec l'équation 3.3 pour le bécher de prélèvement et la pompe utilisés au site 2 du Lac de Créteil. Pour les deux équipements, l'incertitude était relativement similaire quel que soit l'indicateur microbien (Figure 3.1B). L'incertitude moyenne des deux équipements était respectivement pour *E. coli* de ±(40 ± 40)% et pour les EI de ±(74 ± 59)%. Cette disparité entre BIF a également été remarquée par une étude antérieure Jin et al. (2004), qui a montré que l'incertitude temporelle était plus élevée pour les entérocoques intestinaux que pour les coliformes fécaux, à la fois en surface et en profondeur dans la colonne d'eau du lac d'eau saumâtre Pontchartrain (USA). Du fait que les EI sont plus résistants. Ils peuvent ainsi mieux survivre dans divers environnements (Alm et al., 2003).

2.3.3. Protocole de nettoyage

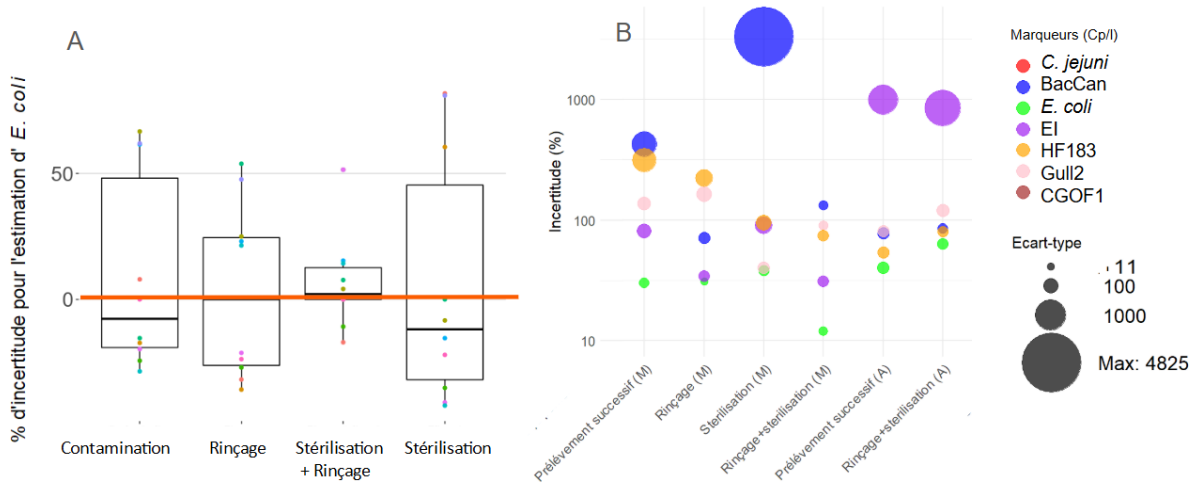

FIGURE 3.2 – Pourcentage d'incertitude pour l'estimation d'*E. coli* (A) et des différents marqueurs (B) par rapport à l'échantillon référence lors des différentes étapes du protocole de nettoyage du bécher et du tuyau de la pompe pour les équipements manuels (M) au niveau du site 1 du lac de Créteil et avec le préleveur automatique (A) à La Villette et à Saint-Maur-des-Fossés.

2.3.3.1. Equipements pour les prélèvements ponctuels

Au cours des campagnes de prélèvement, une contamination croisée peut avoir lieu d'un site à l'autre et les protocoles de nettoyage des équipements sont relativement peu détaillés dans les textes réglementaires, normes et guides. La stratégie mise en oeuvre a donc été d'estimer l'incertitude liée au nettoyage des différents équipements disponibles pour le prélèvement manuel depuis la berge. Pour cela nous avons utilisé le site 1 du lac de Créteil qui présente une concentration moyenne en *E. coli* de 164 ± 102 NPP/100 mL et le site 2 avec une concentration moyenne en *E. coli* de 853 ± 1070 NPP/100 mL. L'analyse du pourcentage d'incertitude a montré qu'après un prélèvement par un site plus contaminé (site 2), l'incertitude du prélèvement sur le site 1 moins contaminé était en moyenne pour *E. coli* de 30 ± 24% et pour les EI de 81 ± 90%. Un rinçage de l'équipement 3 fois avec l'eau du site ou une désinfection à l'éthanol sans rinçage suffisait pour diminuer l'incertitude (Figure 3.2A). La combinaison de la désinfection avec le rinçage triple à l'eau du site réduisait l'écart type de l'incertitude pour *E. coli* (31 ± 11%) (Figure 3.2A), pour les EI (34 ± 34%) et pour le marqueur humain HF183 (de 313 ± 503% à 222 ± 189%) (Figure 3.2B). Cependant, aucune différence significative n'a été constatée pour les 2 BIF entre le protocole sans rinçage et le protocole avec désinfection et rinçage (Test de Wilcoxon, p=0.677 pour *E. coli*, p=0.288 pour les EI, n=20). En ce qui concerne les autres indicateurs bactériens, les protocoles de désinfection et de lavage ne modifiaient pas l'incertitude

moyenne (Figure 3.2B). Cependant, pour le marqueur canin BacCan une très grande incertitude était observée après la stérilisation (3314 ± 4825%).

Globalement les résultats indiquent, qu'entre des sites avec des concentrations d'environ 1 Log_{10} d'écart en BIF, un rinçage 3 fois avec l'eau du site est suffisant quelque soit l'équipement utilisé. Ce résultat est conforme au guide de prélèvement de l'Agence de l'eau Loire-Bretagne (AELB, 2006). Il est également recommandé de nettoyer le bécher et la perche avec une lingette désinfectante dans les normes et guides, mais ce protocole ne peut pas s'appliquer à l'intérieur du tuyau de la pompe. Ainsi procéder à une étape de désinfection de l'intérieur et extérieur du tuyau à l'éthanol et rinçage peut être une alternative afin de réduire l'incertitude liée à la contamination du tuyau de prélèvement de la pompe. Il faudrait toutefois vérifier qu'il en va de même avec des sites qui ont un écart de qualité microbiologique plus élevé.

2.3.3.2. Prélèveur automatique

L'échantillonnage ponctuel réglementaire est généralement effectué à des dates fixes écartées d'une semaine à un mois, au mieux il peut être réalisé une fois par jour, mais pendant la semaine de travail (Burnet et al., 2021). De ce fait, des événements polluants de court terme peuvent ne pas être échantillonnés (Burnet et al., 2021). Les échantillonneurs automatiques peuvent alors être utilisés pour échantillonner sur 24 h ou à l'événement un échantillon composite ou des échantillons discrets multiples (Wilson et al., 2024). Le protocole de nettoyage des lignes de prélèvement des préleveurs automatiques recommandé par l'USGS (Wilson et al., 2024) est relativement lourd puisqu'il nécessite un démontage et nettoyage en laboratoire. Nous avons donc testé si une désinfection à l'eau de Javel et des rinçages directement sur le terrain étaient suffisants, à l'aide d'un préleveur automatique installé au canal de l'Ourcq à la Villette, et en Marne à Saint-Maur-des-Fossés. Les concentrations en BIF sur les deux sites différaient légèrement puis la concentration moyenne en *E. coli* était respectivement de 266 ± 142 NPP/100 mL pour l'Ourcq et de 616 ± 500 NPP/100 mL pour la Marne. Quel que soit le temps écoulé depuis la dernière utilisation du préleveur automatique, les résultats ont montré que l'utilisation sans nettoyage supplémentaire autre que la purge automatique donne des niveaux de BIF qui ne différaient pas significativement des prélèvements effectués après une stérilisation et trois rinçages du préleveur à l'eau du robinet stérile, ni des valeurs de référence par échantillonnage manuel, malgré une faible augmentation de l'incertitude (de 40 ± 46% avant stérilisation et rinçage à 63 ± 32% après stérilisation et rinçage pour *E. coli* par exemple) (Test de Wilcoxon apparié, n=16, p>0,05, Figure 3.2B). De même, dans notre étude, l'étape de stérilisation et

rinçage s'accompagnait d'une augmentation de l'incertitude pour les autres marqueurs (Tableau S2 et S3). Cependant, une incertitude plus faible entre ± 7 à 9% pour les concentrations en BIF a été mesurée dans les eaux d'un rejet pluvial (McCarthy et al., 2008). Nos résultats étaient probablement liés à des résidus d'eau de Javel dans les tuyaux et le bol de prélèvement qui généraient pour certains essais une sous-estimation des concentrations en BIF. Ces résultats indiquent que pour les eaux de surface avec des concentrations faibles en BIF, un rinçage de la ligne de prélèvement à l'eau du robinet autoclavée serait suffisant même après plusieurs semaines sans utilisation du préleveur.

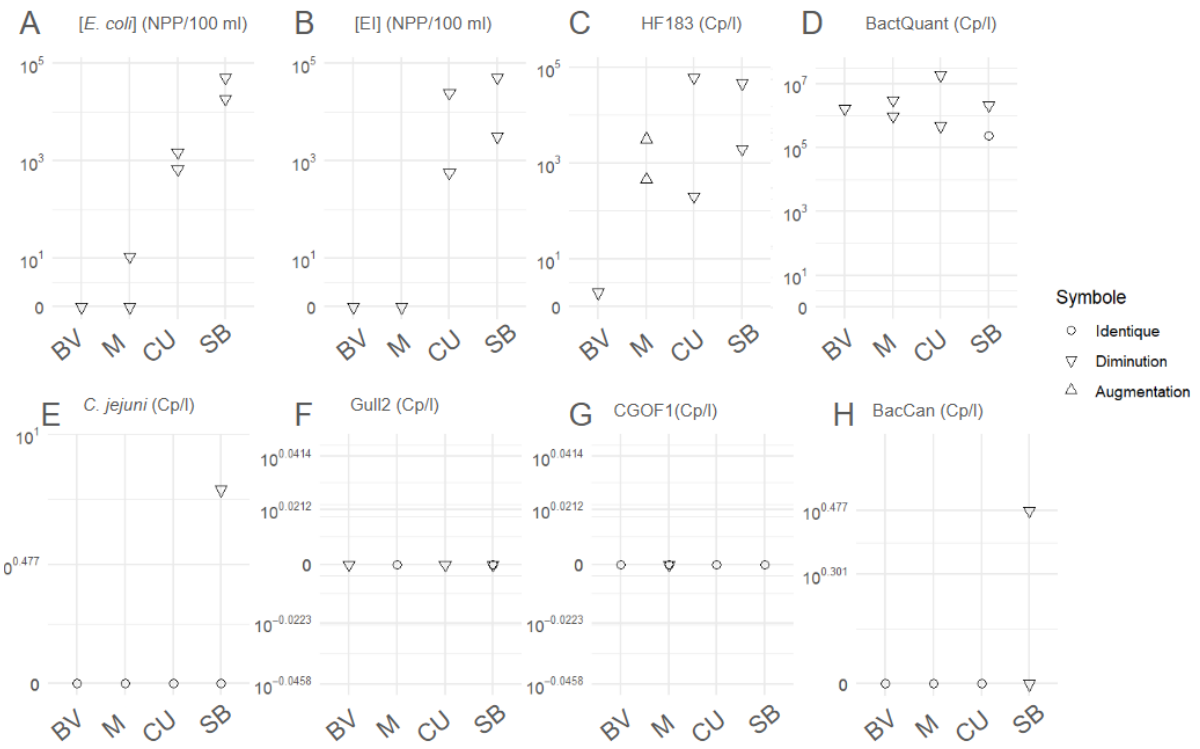

FIGURE 3.3 – Blancs de terrain après prélèvement d'eau de surface au Bassin de la Villette (BV) et en Marne à Saint-Maur-des-Fossés (M) et d'eau résiduaire à l'ouvrage cadre du Centre Urbain (CU) et au bassin de rétention de Sucy-en-Brie (SB). Le symbole représente une comparaison des blancs par rapport à l'échantillon du site prélevé avant le blanc.

Par contre, comme le montre la figure 3.3, pour les eaux résiduaires, une étape de décontamination de la ligne de prélèvement était nécessaire, surtout pour l'analyse des BIF. En effet, les blancs de terrain (eau du robinet autoclavée) après un prélèvement d'eau de rejet pluvial montraient une concentration résiduelle élevée de 35000 ± 16000 NPP/100 mL pour les eaux pluviales en entrée du bassin de rétention de Sucy-en-Brie et de 1090 ± 410 NPP/100 mL au rejet de l'ouvrage cadre du Centre Urbain de Noisy-le-Grand. Après décontamination à la Javel et rinçage à l'eau du robinet stérile, une diminution supplémentaire d'environ 0.5 à 1 Log_{10}

(Figure 3.4A) de la contamination résiduelle en BIF était observée dans les blancs terrains. Bien qu'il restait encore 2 à 3 Log_{10} de BIF dans les blancs, l'impact de cette contamination résiduelle pouvait être considéré négligeable sur des échantillons d'eau résiduaires qui présentaient des concentrations en *E. coli* entre 2.5 et 4.7 Log_{10} /100 mL. Le prélèvement d'un échantillon très contaminé en BIF (6.3-6.4 Log_{10} NPP/100 mL) avait entraîné une contamination croisée 10 fois supérieure sur 1 à 3 prélèvements successifs d'un échantillon de concentration plus faible (4,6-4,9 Log_{10} NPP/100 mL), et ceci malgré la purge automatique de la ligne de prélèvement (Galfi et al., 2014). Ces contaminations résiduelles du système de prélèvement peuvent entraîner un biais lorsque l'étude vise à analyser la dynamique temporelle des concentrations en BIF pendant un événement pluvieux. Dans notre étude, cette contamination croisée s'atténuait de moitié lorsque 7 jours s'étaient écoulés entre les prélèvements (500 ± 200 NPP/100 mL) au rejet de l'ouvrage cadre du Centre Urbain de Noisy-le-Grand. Par contre, pour des eaux de ruissellement d'un parking en entrée et sortie de filtre de roseaux plantés, une contamination <1% dans le tuyau du préleveur après 7 jours secs a été constatée dans une étude précédente (Hathaway et al., 2014).

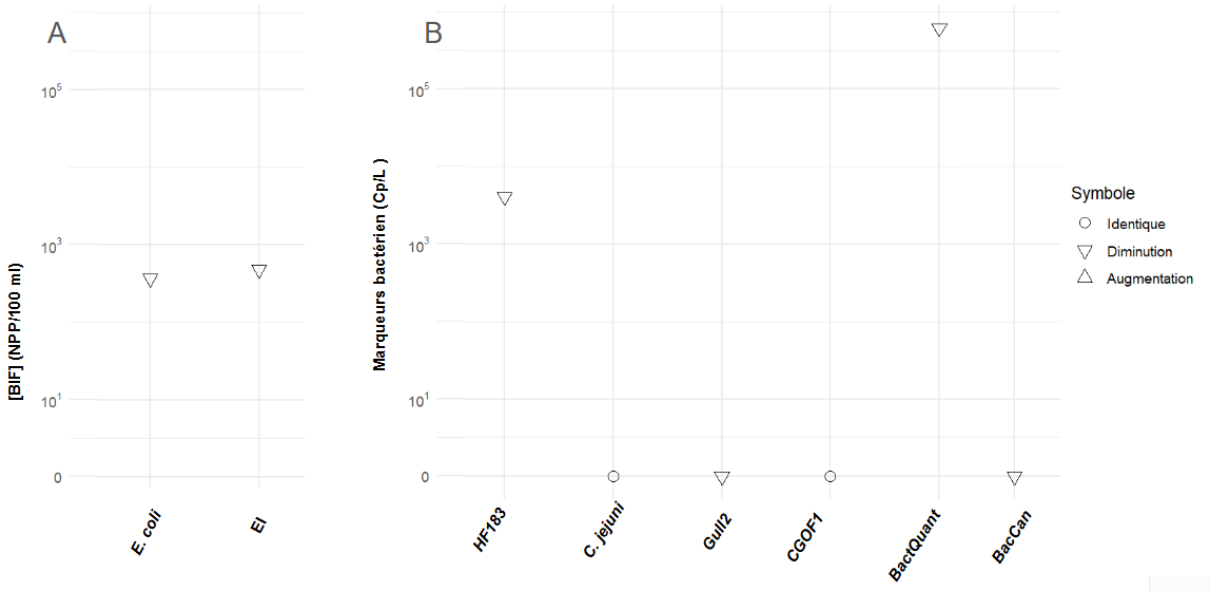

FIGURE 3.4 – Blancs de terrain après décontamination et rinçage du préleveur au bassin de rétention de Sucy-en-Brie. Le symbole représente une comparaison des blancs par rapport à l'échantillon du site prélevé avant le blanc.

En ce qui concerne les autres marqueurs bactériens, une contamination résiduelle de la ligne de prélèvement a été observée pour les blancs lorsque les eaux de surface étaient prélevées préalablement, principalement pour le marqueur humain HF183 et les bactéries totales

BactQuant (Figure 3.3). Les résultats étaient aléatoires, parfois la désinfection était efficace (blancs terrains négatifs) comme pour les marqueurs canin et aviaires. La désinfection permettait également une diminution de l'incertitude sur la mesure des marqueurs spécifiques pour les échantillonnages de rejets pluviaux (Figure 3.4B).

Il ne faut pas oublier de prendre en compte l'influence de la longueur et de l'inclinaison du tuyau de prélèvement qui peut se contaminer du fait d'un volume mort d'eau qui reste à l'intérieur malgré la purge automatique (Galfi et al., 2014; Hathaway et al., 2014). En effet, une incertitude moyenne plus faible (1.7%) avait été mesurée lorsque le tuyau était incliné, alors qu'elle était de 5.5% avec un tuyau droit (Hathaway et al., 2014). La longueur du tuyau (1,5 vs 5 m) ne semblait pas influencer la contamination croisée lors d'un passage d'un échantillon très contaminé (6.3-6.4 Log_{10} NPP/100 mL) à un échantillon moins contaminé (4,6-4,9 Log_{10} NPP/100 mL) (Galfi et al., 2014). Il est donc important de privilégier une installation du préleveur automatique proche du bord et en hauteur, comme il est recommandé dans les guides FD t90-523-1 et de l'agence de l'eau Loire-Bretagne. La désinfection terrain suivie de 3 rinçages à l'eau stérile couplée à une installation adéquate, conduit donc à diminuer l'incertitude liée à la contamination résiduelle de la ligne de prélèvement.

2.3.4. Protocole de transport et stockage

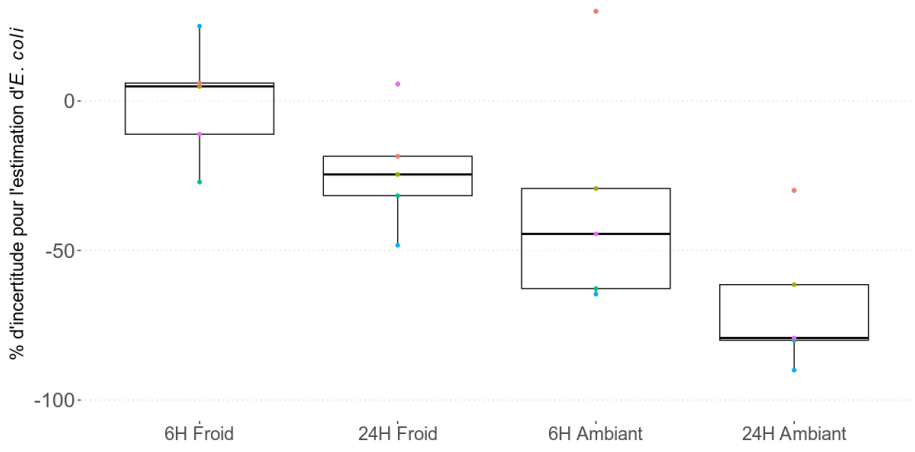

FIGURE 3.5 – Pourcentage d'incertitude pour l'estimation d'*E. coli* par rapport à l'échantillon référence en fonction du temps (6 h ou 24 h) et de la température de stockage les 6 premières heures à 5°C (froid) ou à température ambiante (ambiant).

Pour évaluer l'incertitude liée aux conditions de température durant le transport et le stockage des échantillons, des prélèvements manuels ont été effectués avec la pompe et la

perche au niveau du site 2 du lac de Créteil, avec une concentration moyenne en *E. coli* de 353 ± 233 NPP/100 mL et en EI de 568 ± 882 NPP/100 mL (n=5). Pour les deux groupes de BIF (respectivement *E. coli* et EI), aucune différence significative n'a été observée entre les échantillons transportés à température ambiante ou à 5°C, que les analyses aient été réalisées après 6 h ou 24 h (Test t apparié, $p>0.06$, $p>0.18$, n=36, Figure 3.5). Toutefois, à 24 h le pourcentage de perte des BIF était légèrement plus élevé avec un stockage les 6 premières heures à température ambiante (67 ± 21% pour *E. coli* et 57 ± 22% pour EI) qu'avec un stockage tout le temps à 5°C (55 ± 26% pour *E. coli* et 31 ± 13% pour EI). Les résultats sont pas tout à fait en concordance avec une étude antérieure qui a montré que jusqu'à 24 h le stockage des échantillons d'eau pluviale à température ambiante ne constitue pas un facteur significatif de variation de la concentration en *E. coli* (McCarthy et al., 2008). Toutefois, à 5°C, une faible décroissance au cours du temps a été rapportée par Harmel et al. (2016). Pour une période de stockage de 24 h à 10-15°C d'eau douce et marine, l'existence d'une réduction de la concentration en *E. coli* de 28% en moyenne a été démontrée (Crane and Moore, 1986). De plus, cette variabilité est dépendante du niveau de contamination de l'échantillon (Ferguson, 1994). En effet, pour les échantillons de rivière relativement contaminés (coliformes fécaux 2,7 à 3,9 Log_{10} NPP/100 mL), la concentration ne différait pas entre un stockage froid pendant 9 h ou 18 h. Par contre, pour une rivière 10 fois moins contaminée, une variation de 0,1 Log_{10} de la concentration en coliformes fécaux était visible entre les deux durées (Ferguson, 1994). Toutefois, aucune analyse statistique n'avait été menée dans cet article.

En ce qui concerne le marqueur d'ADN total (BacQuant), la variabilité était la plus faible avec 6 h de stockage à 5°C (Tableau S2) et pour le marqueur humain HF183, le profil de variabilité était relativement similaire à celui des BIF (Figure S1). Par contre, une plus grande variabilité a été observée avec les marqueurs aviaires (CGOF1 et gull2) pour le stockage à température ambiante, l'incertitude étant la plus élevée au-delà de 6 h qui se traduit par une décroissance de ces marqueurs (Figure S1). Pour le marqueur canin, il est plus délicat de conclure car il n'a été détecté que lors d'une seule campagne. Une décroissance après 6 h à 5°C a été observée, alors qu'après 24 h à 5°C, la concentration était élevée (Tableau S3).

Les résultats de notre étude montrent donc que le transport à température ambiante peut entraîner une forte variabilité des concentrations en BIF, ainsi que les autres marqueurs fécaux bactériens si la mesure n'est pas effectuée avant 24 h. L'ensemble de ces résultats indique qu'il est recommandé de limiter le temps de stockage à moins de 6 h, en privilégiant un transport à 5°C.

Ceci apporte une précision et un éclairage sur les textes réglementaires, normes et guides français qui recommandent pour le transport de placer les échantillons dans une enceinte réfrigérée à 5 ± 3°C au maximum pendant 24 h, à l'abri du rayonnement solaire (NF EN ISO 19458). Le guide FD T90-523-1 précise que pour les échantillons de rivière, la température doit être de 4 ± 2°C pendant <8 h au mieux et dans les 24 h au plus. L'analyse doit avoir lieu le jour même de préférence et au plus tard dans les 24 h s'il existe une impossibilité géographique (Arrêté du 19 octobre 2017).

2.3.5. Impact du temps d'incubation sur la lecture

Pour la mesure des BIF, les Normes NF EN ISO 9308-3 et NF EN ISO 7899-1 spécifient un temps d'incubation de 36 à 72 h et une lecture à 36 h. Cependant, une lecture dès 24 h est souvent pratiquée pour la surveillance des eaux de baignade. Afin de savoir quel est l'impact du temps d'incubation sur l'incertitude de la mesure des BIF, nous avons comparé les lectures à 24, 48 et 72 h sur une large gamme d'échantillons provenant de 4 sites. La gamme de concentrations moyennes en *E. coli* s'étendait de $1{,}56.10^2$ NPP/100 mL (site 1 du Lac de Créteil) à $7{,}77.10^6$ NPP/100 mL en entrée de la station de traitement des eaux usées de St-Thibault-des-Vignes. L'ensemble des résultats indiquait que plus la concentration au niveau d'un site était faible, plus le temps d'incubation nécessaire avant une stabilisation de la lecture était long. Ainsi, par exemple, à St-Thibault-des-Vignes la lecture était stable dès 24 h d'incubation (Test de Wilcoxon et test t apparié, $p>0{,}05$, n=12) alors que pour le site 1 du Lac de Créteil une différence significative était observée entre les 3 temps d'incubation (Test de Friedman, n=177, p=0,02). En effet, les tests microbiologiques sont souvent moins précis pour des échantillons complexes. Par exemple, les résultats faux positifs pour la mesure d'*E. coli* atteignent 4% dans l'eau de pluie mais jusqu'à 40% dans les eaux usées (McLain et al., 2011). Cependant, il faut être vigilant que d'autres micro-organismes non ciblés ne se développent pas dans ces microplaques au-delà de 48 h d'incubation à 44°C (Ndione, 2022). Sachant que la lecture à 24 h peut générer une sous-estimation pour certains échantillons, il s'agit donc d'un compromis entre rapidité et exactitude du résultat. Or, pour la gestion d'un site de baignade, connaitre le résultat le plus tôt possible est crucial. La méthode NPP ColiLert (IDEXX) peut offrir une alternative intéressante à la méthode par microplaque puisqu'elle permet une lecture en 18 h pour *E. coli* (ISO 9308-2 :2012) et en 24 h pour les entérocoques intestinaux (certification NF Validation du test Enterolert®-E).

2.3.6. Synthèse globale des incertitudes

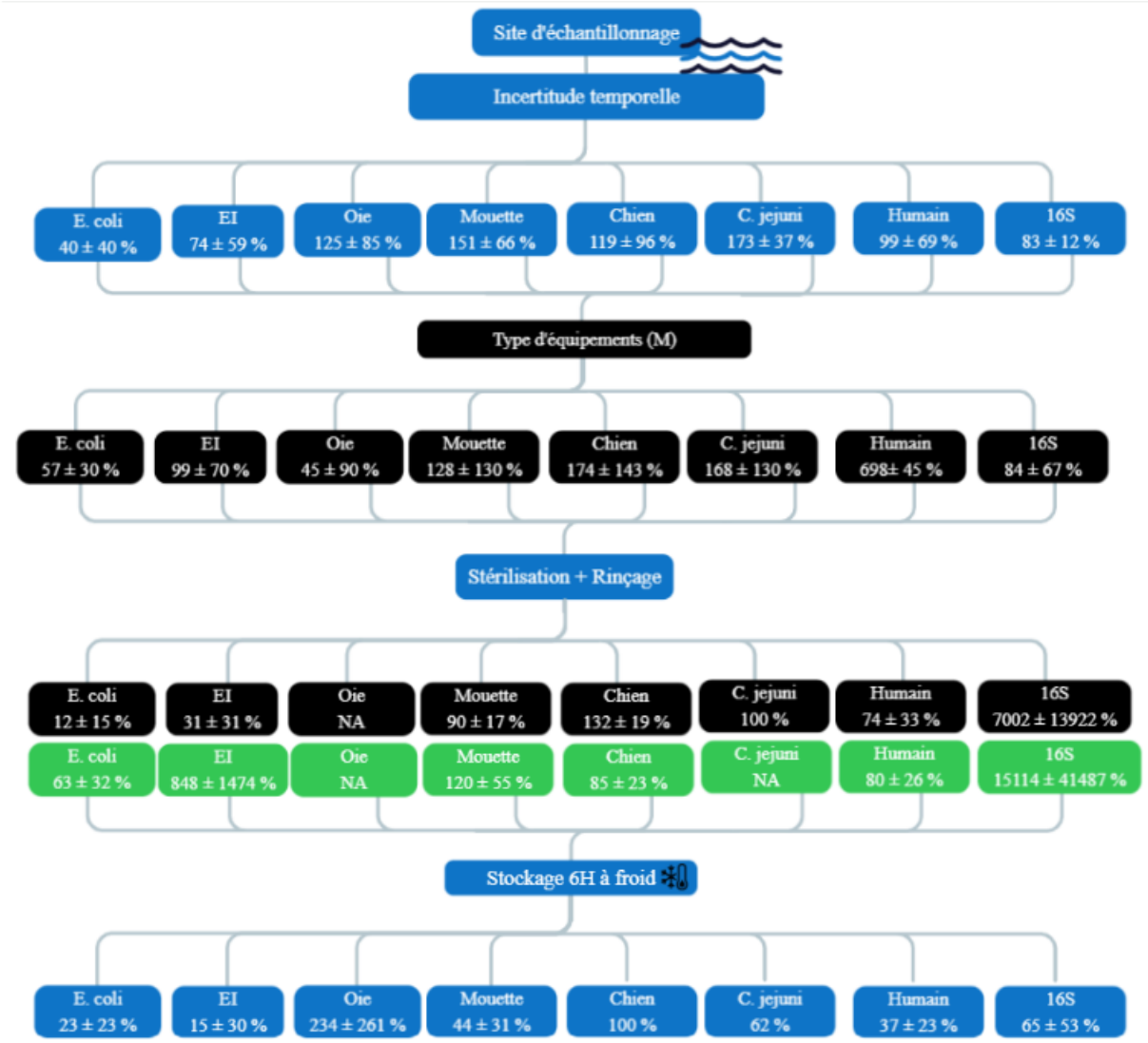

FIGURE 3.6 – Schéma récapitulatif de l'analyse de l'incertitude pour l'ensemble des indicateurs fécaux analysés ; en noir (équipements manuel), en vert (préleveur automatique) ; en bleu (tous les équipements).

La figure 3.6 récapitule l'ensemble des pourcentages d'incertitude liés au prélèvement, au transport et au stockage des échantillons. Afin de limiter l'incertitude globale liée au prélèvement, nos recommandations sont les suivantes pour le prélèvement manuel depuis la berge :

- privilégier le bécher ou la pompe associés à une perche télescopique depuis la berge ou un bateau.

- un rinçage à l'eau du site est généralement suffisant pour des écarts de concentration en BIF d'environ 1 Log_{10} pour les eaux de surface. Toutefois, pour des écarts de concentration plus élevés, une désinfection préalable du bécher ou du tuyau de prélèvement permettra de réduire

l'incertitude liée aux contaminations croisées.

- réaliser des blancs de terrain avec de l'eau stérile.

Pour les préleveurs automatiques, les recommandations sont les suivantes :

- limiter la longueur du tuyau de prélèvement et favoriser une inclinaison.

- la purge automatique n'est pas suffisante pour éviter les contaminations croisées, un rinçage à l'eau du robinet stérile de la ligne de prélèvement est toutefois suffisant pour des eaux de surface peu contaminées. Pour les sites très contaminés ou pour les eaux résiduaires il est nécessaire de procéder à une désinfection à la Javel à 5% comme recommandé par l'USGS (Wilson et al., 2024), suivie d'*a minima* trois rinçages à l'eau du robinet stérile.

- limiter le stockage dans le préleveur à 24 h maximum, et privilégier une embase réfrigérée ou l'ajout de pains de glace dans le logement des flacons. En cas d'impossibilité, les échantillons pourront être à température ambiante tant que le préleveur est à l'ombre à l'extérieur ou à l'intérieur du réseau.

Une fois l'échantillon collecté, leur transport et leur stockage doivent, idéalement, être limités à moins de 6 heures avec un stockage à 5 ± 3°C. Au-delà de 24 heures ou à température ambiante, une variabilité accrue des résultats a été observée, notamment pour les marqueurs animaux et humains.

Il est également crucial de connaitre l'incertitude associée aux méthodes analytiques, car celles-ci peuvent avoir un impact sur l'incertitude finale à prendre en compte. Les erreurs liées à la mesure de la qualité de l'eau peuvent être attribuées à plusieurs facteurs, notamment la méthode de mesure, les dilutions en série, et la distribution hétérogène des microorganismes dans le volume prélevé, les réactifs et équipements, les erreurs humaines (Harmel et al., 2016). Un temps d'incubation de 36 ou 48 heures est recommandé pour stabiliser les résultats de lecture des microplaques. Pour les BIF, nous avons estimé une incertitude analytique de 31% [IC95% 26 : 35] pour *E. coli* et 45% [IC95% 35 : 56] pour les EI par les méthodes miniaturisées en microplaques. Chaque étape de dilution augmente l'incertitude, surtout avec des faibles concentrations (Harmel et al., 2016). Pour les quantitray Colilert (IDEXX), l'incertitude rapportée dans la littérature est de 22 ± 15% (McCarthy et al., 2008). De plus, Tiwari et al. (2016) ont montré que la méthode Colilert-18 et la méthode en microplaque ISO 9308-3, avaient un taux de concordance supérieur à 90%, la concentration estimée par les deux méthodes n'étant pas significativement différente. Pour les estimations par PCR quantitative (qPCR), l'incertitude peut être estimée en s'appuyant sur une étude comparative qui rapporte une incertitude de 67% pour *E. coli* et 27% pour EI dans

un premier essai inter-laboratoire, contre 25% pour *E. coli* et de 21% pour EI dans un second essai (Noble et al., 2010). De même, une incertitude sur la quantification par qPCR d'*E. coli* a estimé des valeurs inférieures à 25% (Bergeron et al., 2011). Concernant l'incertitude liée à la mesure des BIF par le système online ColiMinder, une analyse a été menée en laboratoire par Eau de Paris avec des échantillons de la Seine, ce qui a permis d'estimer une incertitude de 14% (n=10) avec un nettoyage automatique entre les échantillons prélevés successivement (Loiodice, 2024). Au niveau de la littérature, une incertitude analytique sur *E. coli* avec le système ColiMinder a été estimée à 6% en laboratoire à 22% en terrain (Cazals, 2019).

Enfin pour la détection par qPCR des marqueurs spécifiques de sources humaines et animales, une répétabilité de la mesure (coefficient de variation) a été estimée à <5% pour le marqueur aviaire Gull2, <6% pour le marqueur humain HF183 et <3% pour le marqueur canin BacCan lors d'essais intra-laboratoire (Ebentier et al., 2013). Dans cette même étude inter-laboratoires, la répétabilité entre laboratoire était estimée à <20% pour le marqueur Gull2, <10% pour le marqueur HF183 et <6% pour le marqueur BacCan. En effet, les espèces et souches bactériennes peuvent réagir différemment, selon leur métabolisme et les conditions de stress au moment de l'analyse (Sutton, 2011). Ces éléments sont essentiels à considérer pour interpréter les résultats.

2.3.7. Incertitudes retenues pour *E. coli*

Pour aider à la prise de décision lors de la gestion active des sites de baignade, nous avons pris comme un cas d'utilisation la concentration en *E. coli* car ce paramètre est le plus déclassant pour la gestion journalière des rivières franciliennes telle que la Seine ou la Marne (Mouchel et al., 2020). Ainsi pour les mesures avec le système ColiMinder en Seine (Ville de Paris), le calcul de l'incertitude globale a regroupé l'ensemble des incertitudes rapportées dans la figure 3.6. Ainsi l'incertitude analytique de 14% estimée par Eau de Paris (Loiodice, 2024) qui comprend l'étape de nettoyage automatique entre les échantillons, l'incertitude sur les mesures qui sont réalisées rapidement sans stockage. Cette incertitude ayant été évaluée en laboratoire, une incertitude temporelle de 40% a également été incluse (Figure 3.7). L'incertitude globale qui a été retenue était donc de 42%. Pour les analyses réglementaires d'*E. coli*, le guide technique FD T 90-521 est le référentiel suivi par les équipes de prélèvement de la Ville de Paris. Une incertitude de 12% a été retenue pour la stérilisation et le rincage des équipements, une incertitude de 23% liée au stockage à froid et analyse le jour même et une incertitude

analytique de 31% (Figure 3.6). Au total cela représentait une incertitude globale de 40% (Figure 3.7). Des coefficients de variations allant de 0.9 à 7.2% pour le système ColiMinder, et de 12.4 à 34.0% pour la mesure avec les quantitray Colilert (IDEXX) sur des mesures repliquées 6 fois ont été rapportés dans la litérature (Burnet et al., 2019). Des coefficients de variation allant de 31 à 105% ont été estimé dans une étude antérieure pour la méthode de mesure en microplaque (NPP) (Servais et al., 2005). En effet, les méthodes basées sur le NPP peuvent présenter des incertitudes qui dépassent 30% du fait de la probabilité de distribution du nombre le plus probable (Gronewold and Wolpert, 2008). Considérant ces valeurs de la litérature, nos estimations pouvaient donc être considérées raisonnables.

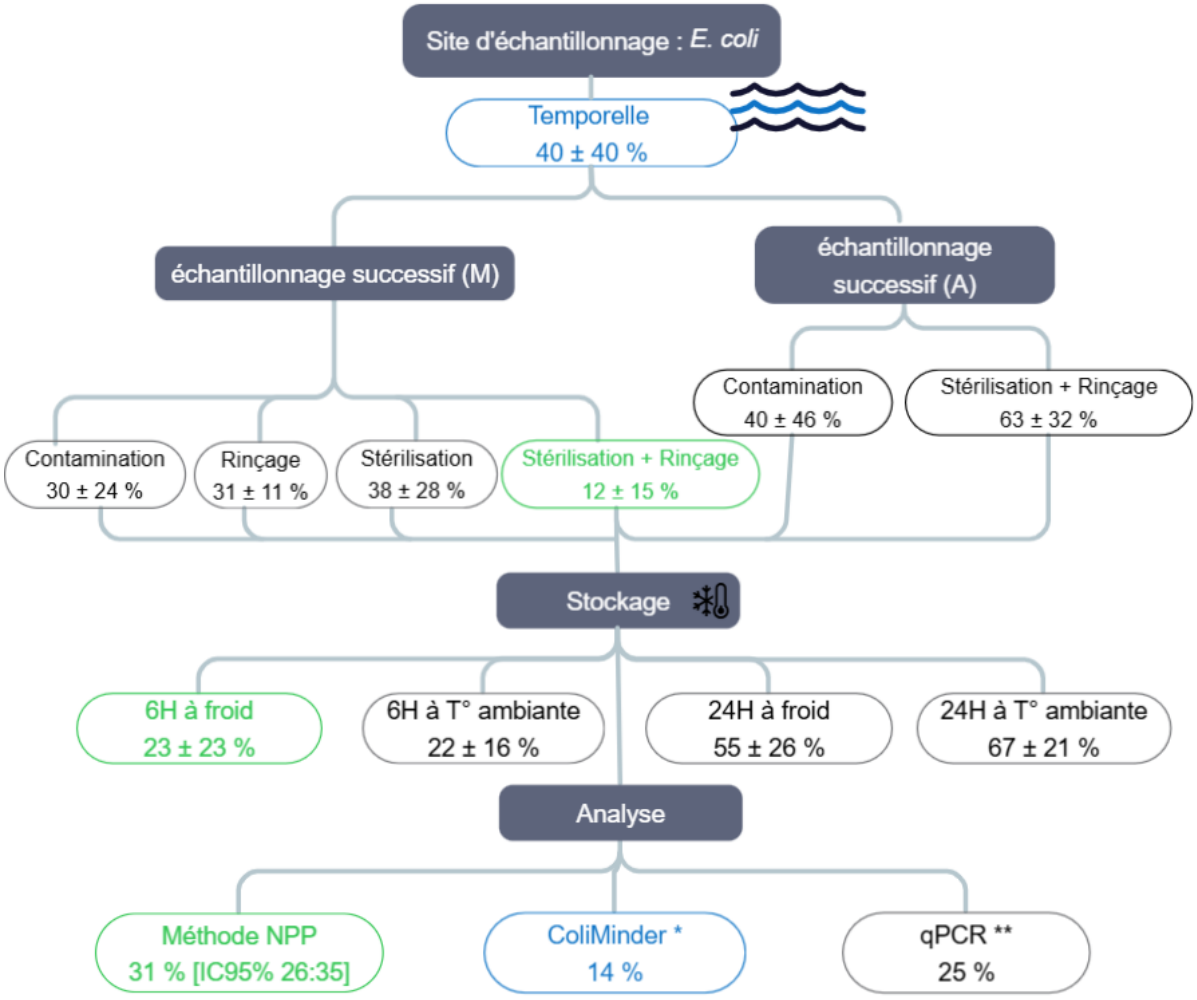

FIGURE 3.7 – Schéma récapitulatif de l'analyse de l'incertitude pour *E. coli*, en bleu les incertitudes sur les équipements automatiques pris en compte pour les mesures obtenues avec le système ColiMinder et en vert les incertitudes sur les équipements manuels pris en compte pour les mesures ponctuelles en culture * : (Bergeron et al., 2011; Noble et al., 2010).

2.3.8. Intégration de l'incertitude dans la prise de décision

En France, pour la gestion quotidienne d'un site de baignade, la qualité microbiologique instantanée d'un prélèvement est qualifiée suivant des valeurs seuils de l'instruction n°DGS/EA4/2022/168 du 17 juin 2022 relative aux modalités de recensement, gestion et classement des eaux de baignades. Ces valeurs seuils sont basées sur un rapport de l'AFSSET de septembre 2007 (Duboudin et al., 2007) et s'appliquent sur un site de baignade classé par une Agence Régionale de la Santé (ARS). Elles sont utilsées en cours de saison pour aider les gestionnaires à décider de l'ouverture ou la fermeture du site de baignade. Pour les eaux douces, ces seuils sont de 100 et 1800 NPP/100 mL, afin de catégoriser l'échantillon en qualité bonne, moyenne ou mauvaise.

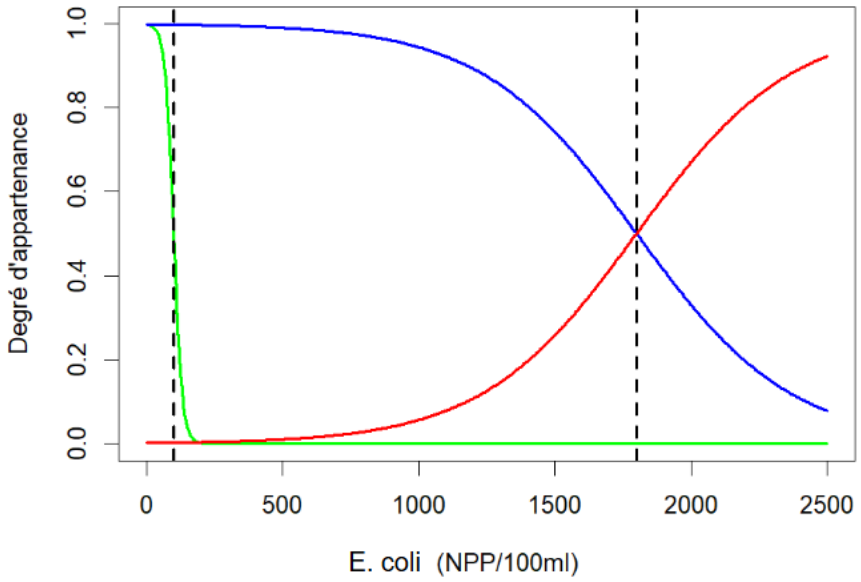

FIGURE 3.8 – Fonction d'appartenance avec la logique floue. En vert : qualité bonne, en Bleu : qualité moyenne, en rouge qualité mauvaise et les traits noirs représente les seuils réglementaires.

Lors de la décision, les gestionnaires vont classiquement comparer la valeur mesurée de l'échantillon ponctuel à la valeur seuil, sans tenir compte de son incertitude sur la détermination du NPP la plupart du temps (Sylvestre et al., 2020), pour décider si l'échantillon est conforme avec une ouverture de la baignade (méthode de référence 1). Dans le cas de la Ville de Paris, pour la prise de décision de tenue des épreuves de nage ou triathlon lors des Jeux Olympiques, l'historique des mesures du système ColiMinder a été pris en compte sur une fenêtre de 4 h précédant le matin (méthode de référence 2). En vue d'améliorer ce processus de prise en compte de l'historique des données des systèmes de mesure en temps réel ou quasi réel, nous proposons

d'utiliser la logique floue pour intégrer d'une part l'incertitude sur la mesure et d'autre part l'historique des données précédant le matin. Nous avons testé 6 intervalles de temps allant de 4 à 24 h. Nous avons comparé les résultats de classification obtenus, avec ceux obtenus en mettant en oeuvre les méthodes de référence 1 et 2 utilisées par les gestionnaires.

TABLE 3.1 – Pourcentage de vrais positifs pour les 5 méthodes de défuzzification par rapport aux 2 méthodes de référence (méthode 1 et méthode 2), (NC) non classé avec la méthode par la Ville de Paris pendant les JOP 2024.

Intervalle d'analyse	**Méthode de référence**	**COG**	**MOM**	**LOM**	**ROM**	**BS**	**NC**
[12 h : 12 h]	1	81	79	79	79	78	0
	2	30	31	31	31	27	51
[8 h : 8 h]	1	80	80	80	80	77	0
	2	31	31	31	31	28	53
[0 h : 12 h]	1	77	77	77	77	77	0
	2	30	32	32	32	27	51
[20 h : 8 h]	1	78	77	77	77	77	0
	2	31	30	30	30	29	53
[4 h : 8 h]	1	77	77	77	77	75	0
	2	29	29	29	29	28	53
[8 h : 12 h]	1	78	78	78	78	76	0
	2	32	32	32	32	32	52

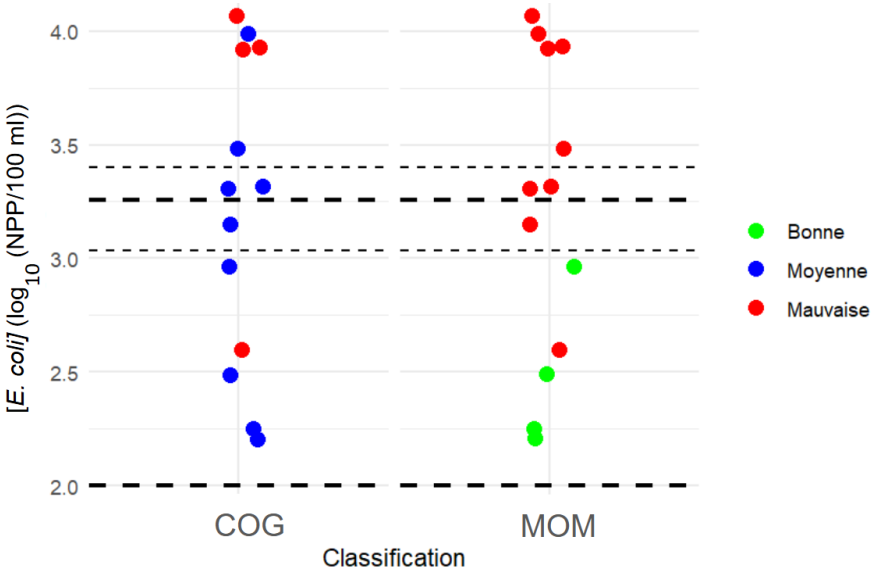

FIGURE 3.9 – Comparaison de la méthode du (A) centre de gravité (COG) et de (B) moyenne des maxima (MOM) pour les mesures réglementaires au niveau des 3 sites avec une classification utilisant l'intervalle de 24 h de midi à midi. La couleur représente la classe d'appartenance et l'axe des abscisses représente la méthode de défuzzification. Le trait noir épais représente le seuil réglementaire et les traits fins représentent l'incertitude associée au seuil pour les analyses ponctuelles.

Après fuzzification pour déterminer les probabilités d'appartenance aux 3 classes de qualité, pour l'étape de défuzzication qui permet de classer les valeurs mesurées, nous avons

testé 5 méthodes de calcul (COG, MOM, LOM, BS et ROM). Nous avons classé l'ensemble des mesures réglementaires du suivi estival sur la Seine de 2020 à 2023. Les 3 méthodes (MOM, LOM et ROM) présentaient les mêmes résultats de classification. Aucune différence significative n'est observée pour les 5 méthodes de défuzzification après comparaison des classifications avec la (méthode de référence 1) et la (méthode référence 2) (Test de Khi² d'ajustement, $p<0.001$, $n=154$). Toutefois, en ce qui concerne le pourcentage de vrais positifs, après comparaison aux méthodes de référence 1 ou 2, pour les différents intervalles de temps analysés, la méthode BS présentait des pourcentages de vrais positifs similaires ou légèrement plus faibles (Tableau S1). De plus, l'emploi d'une des méthodes des maxima (MOM, LOM et ROM) pour la défuzzification entraînait des classements très incohérents pour certains jours, avec des échantillons très contaminés classés comme étant de qualité "bonne" (Figure 3.9). Le centre de gravité qui donnait des résultats corrects a été ainsi retenu comme méthode de défuzzification. En effet, parmi les méthodes de défuzzification, le calcul du centre de gravité est l'un des plus utilisés (Mahabir et al., 2003).

La méthode de défuzzification (COG) retenue générait en moyenne 78 ± 2% de vrais positifs communs avec la méthode de référence 1 et 30 ± 1% avec la méthode de référence 2. Il faut noter que la méthode de référence 2 classait en moyenne 52 ± 1% des valeurs comme incertaines. La logique floue est particulièrement adaptée pour traiter des données aux connaissances très variables, vagues ou incertaines, permettant ainsi un flux d'information logique, fiable et transparent depuis la collecte des données jusqu'à leur utilisation dans des contextes environnementaux (Icaga, 2007). Pour les différents intervalles de temps analysés, les résultats étaient statistiquement similaires, la majorité (entre 87 et 97%) des mesures étaient classées de la même manière (Test de Khi² d'ajustement, $p<0.001$, $n=154$). Ces résultats indiquaient qu'au pont de l'Alma et Tolbiac, les 6 intervalles de temps donnaient la même classification. Il y avait majoritairement (72%) des mesures classées selon la méthode de référence 1 comme étant de qualité moyenne. Parmi celles-ci, une large proportion (entre 84 et 90%) de ces mesures était attribuée par la méthode de logique floue à la classe de qualité moyenne. De même, entre 52 et 57% des valeurs au-dessus de 1800 NPP/100 mL (classe mauvaise, selon la méthode de référence 1) étaient classées en qualité mauvaise par la méthode de logique floue. Le processus de logique floue utilise une approche simple basée sur des règles pour résoudre des problèmes de contrôle car capable d'intégrer différents types d'observations de qualité (Elmas, 2003; Icaga, 2007).

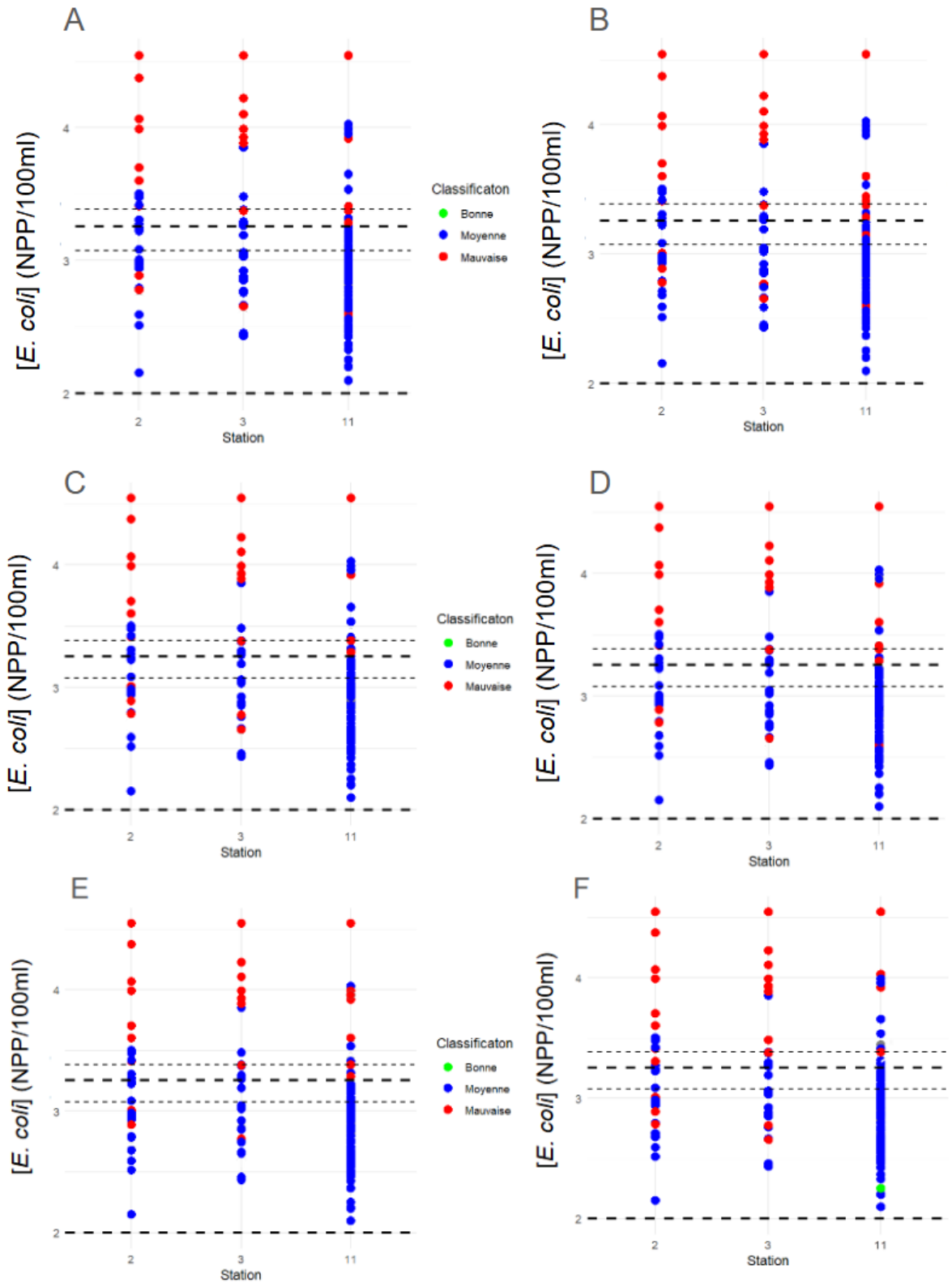

FIGURE 3.10 – Comparaison de la classification des données du système ColiMinder avec la logique floue pour les mesures réglementaires d'*E.coli* (Log_{10} NPP/100 mL) au niveau des 3 sites (2 : pont de Tolbiac rive droite, 3 : pont de Tolbiac rive gauche, 11 : pont de l'Alma) en utilisant différents intervalles de temps (A) midi à midi (B) 8 h à 8 h (C) minuit à midi (D) 20 h à 8 h (E) 4 h à 8 h (F) 8 h à midi. La couleur représente la classe d'appartenance. Les traits noirs épais représentent les seuils réglementaires de 1800 NPP/100 mL et de 100 NPP/100 mL et les traits fins représentent l'incertitude associée à la mesure manuelle.

La logique floue permet d'obtenir des informations plus précises en utilisant une forme continue. Cela est particulièrement pertinent dans le contexte de l'évaluation de la pollution de l'eau, où il est crucial de considérer la variabilité et la complexité des données (Icaga, 2007). En effet, en fonction de la variabilité spatio-temporelle, des sources de contamination en amont du site, des conditions environnementales et des caractéristiques du site, mais également de la position du ColiMinder par rapport au site, l'intervalle de temps à prendre en compte peut être variable (Quilliam et al., 2011; Rossi et al., 2020; Briciu-Burghina et al., 2019). L'avantage de cette approche pour classer les futurs sites de baignade est qu'elle permet de lisser sur plusieurs heures et de prendre en compte l'historique plutôt qu'une mesure ponctuelle, tout en associant l'incertitude sur la mesure. Cette méthode permet aussi une prise de décision objective pour des valeurs proches de la valeur seuil. Utiliser cette approche combinée avec des mesures en temps (quasi) réel acquises par des appareils de surveillance de haute qualité, comme ColiMinder, permet ainsi une classification rapide et fiable. Un système ColiMinder a été utilisé pour surveiller des eaux récréatives à quatre emplacements le long de rivières (Makris et al., 2023). Cette même étude a observé des relations spécifiques entre l'activité enzymatique et les niveaux de contamination par *E. coli*, indiquant que la surveillance en ligne pourrait constituer un complément aux méthodes de laboratoire traditionnelles, surtout en cas de contamination élevée ou lors de déversements combinés (Makris et al., 2023). Cela ouvrirait la voie à une évaluation encore plus complète et précise des problèmes de pollution, en renforçant la capacité de la logique floue à modéliser des systèmes complexes et à gérer les incertitudes inhérentes aux données environnementales.

2.4. Conclusion

Nos résultats présentent l'originalité de ne pas se limiter à l'analyse de l'incertitude de la mesure des BIF et d'évaluer d'autres marqueurs bactériens spécifiques de sources ou encore des pathogènes, contrairement à de nombreux articles scientifiques traitant uniquement de l'incertitude de la mesure des BIF. L'analyse de l'incertitude au niveau de la mesure de la concentration en BIF et d'autres marqueurs bactériens montre globalement que les équipements manuels étaient statistiquement similaires, confirmant la flexibilité des protocoles d'échantillonnage. Pour le protocole de nettoyage que ce soit pour le prélèvement automatique ou ponctuel, la désinfection ne semblait pas nécessaire pour les eaux de surface avec des concentrations allant

de 45 à 3800 NPP/100 mL, un simple rinçage à l'eau du site ou à l'eau stérile n'entrainant pas de contamination croisée entre sites avec des écarts de concentration de 1 Log_{10}. Par contre, en ce qui concerne les eaux résiduaires, une désinfection était nécessaire, mais l'eau de Javel à 0,5 % n'était pas toujours suffisante. Toutefois, il est possible de la réaliser sur le terrain sans avoir à démonter l'équipement et le ramener au laboratoire pour stériliser. Il serait nécessaire d'utiliser une solution de Javel plus concentrée puis de veiller à bien rincer l'équipement. En ce qui concerne le stockage et le transport des échantillons, une sous-estimation de la concentration en BIF a été observée pour les échantillons transportés et stockés à température ambiante lorsque la mesure était réalisée 24 h plus tard. Enfin, le temps d'incubation de l'ensemble des échantillons avant lecture des microplaques dépend de la concentration de l'échantillon. Ce temps pouvant être réduit à 24 h sur les eaux fortement contaminées. De plus, s'ajoute à cela une incertitude liée à la méthode d'analyse qui doit être prise en compte.

L'ensemble des résultats pourrait aider à l'écriture d'un guide pratique d'échantillonnage en complément des normes et réglementations sur le prélèvement. Un tel guide aurait pour but de permettre une harmonisation du suivi de la qualité des eaux de surface par les différents acteurs. Par ailleurs, il est aussi important de considérer l'intercalibration des méthodes de mesure entre les laboratoires d'analyses lorsque des résultats acquis par différents acteurs sont agrégés pour réaliser des études de séries temporelles longues où lorsque des études sont menées à l'échelle du bassin versant, ou à l'échelle régionale, ou nationale. En effet, il a été montré que le coefficient de variation entre des mesures d'*E. coli* réalisées par 49 laboratoires différents sur 2 aliquots d'un même échantillon pouvait atteindre 119 à 128% (Bremser et al., 2011).

L'intégration de la logique floue dans l'évaluation de la qualité de l'eau, notamment à travers la concentration en *E. coli*, s'est révélée être une approche efficace et objective pour la prise de décision en matière de gestion des baignades. En combinant des méthodes de défuzzification adaptées et des appareils de surveillance en temps réel comme le ColiMinder, il est possible de classer rapidement et avec fiabilité les sites de baignade, en tenant compte des incertitudes associées aux mesures. Les résultats montrent une forte concordance avec les méthodes couramment utilisées par les gestionnaires tout en permettant une évaluation plus nuancée des données et une prise de décision plus rapide. Il serait intéressant de tester cette approche avec les entérocoques intestinaux qui constituent un indicateur pertinent dans les eaux côtières (EPA, 2019). Cette approche offre une réponse proactive aux enjeux de pollution de l'eau, améliorant ainsi la sécurité et la qualité des activités récréatives en milieu aquatique urbain.

En outre, l'utilisation de la logique floue pour l'évaluation de la pollution de l'eau pourrait non seulement améliorer la précision des informations obtenues, mais également fournir une méthode robuste pour intégrer divers aspects de la qualité de l'eau, ce qui est essentiel pour des décisions éclairées en matière de gestion environnementale.

Remerciements : Les analyses ont été financées par les programmes de recherche Me-Seine InnEauvation (https ://inneauvation.fr/) ; OPUR (https ://www.leesu.fr/opur/) et Piren-Seine (https ://www.piren-seine.fr/). Nous remercions également le service des canaux de la Ville de Paris (Olivier Lalouette et Thierry Mareschal) et le Syndicat Intercommunal d'Assainissement de Marne-la-Vallée (Sophie Masnada et ses collègues de la SAUR). Nous remercions Laurent Moulin (Eau de Paris) pour le partage d'informations sur le système ColiMinder et pour ses conseils éclairés.

2.5. Annexe

TABLE S1 – Liste des amorces et des sondes utilisées lors de la qPCR pour la recherche de marqueurs fécaux animaux et humains, des bactéries totales et des *Campylobacter*. La séquence et la concentration finale pour chaque amorce sens (F) et antisens (R), et pour la sonde TaqMan (P) sont présentées dans le tableau.

Cible	**Séquence 5'—-3'**	**finale (µM)**
BacCan F	GGA GCG CAG ACG GGT TTT	0,2
BacCan R	CAA TCG GAG TTC TTC GTG ATA TCT A	0,2
BacCan P	FAM-TGG TGT AGC GGT GAA A-TAMRA-MGB (life tech)	0,1
Gull2 F	CTT GCA TCG ACC TAA AGT TTT GAG	0,1
gull2 R	GGT TCT CTG TAT TAT GCG GTA TTA GCA	0,2
gull2 P	FAM-ACA CCT GGG TAA CCT CAG A - BHQ1	0,2
CGOF1 F	GTA GGC CCT GTT TTA AGT CAG C	0,2
CGOF1 R	AGT TCC CGC TGC CTT GTC TA	0,2
CGOF1 P	FAM - CCG TGC GGT CCT GAC ACA CTT GGA - BHQ1	0,2
β-actine F	GCAAGAGGGAGGAGAAGGACAGAGT	0,05
β-actine R	CAAAGAGGGAGGAGAAAGGAAGT	0,05
β-actine P	HEX-CCCCCTCCTACTGCTCCACCCGAAAATG-BHQ1	0,05

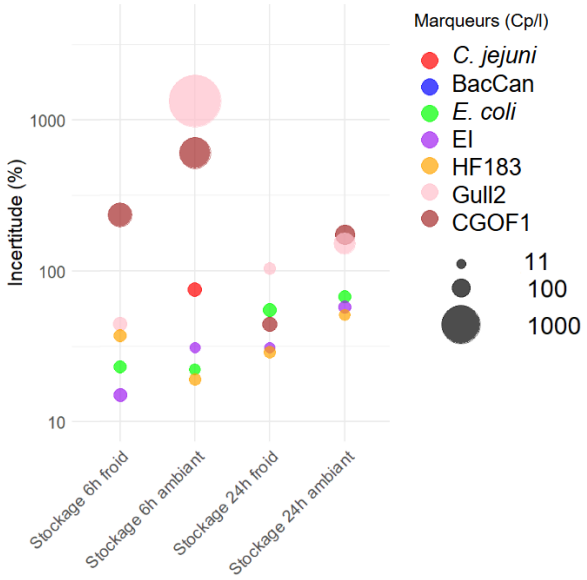

FIGURE S1 – Pourcentage d'incertitude pour l'estimation des différents marqueurs bactériens par rapport à l'échantillon référence en fonction du temps (6 h ou 24 h) et de la température de stockage à 5°C (froid) ou à température ambiante (ambiant).

TABLE S2 – Incertitude sur la collecte des échantillons pour les *E. coli*, les Entérocoques intestinaux, HF183 et les bactéries totales (BacQuant).

Parameter	**EC**	**EI**	**BacQuant**	**HF183**
Temporelle	40 ± 40	74 ± 59	83 ± 12	99 ± 69
Type équipement (M)	57 ± 30	99 ± 70	84 ± 67	98 ± 45
Prélèvement successif (M)	30 ± 24	81 ± 90	8322 ± 16566	313 ± 503
Rinçage (M)	31 ± 11	34 ± 34	6107 ± 12112	222 ± 189
Stérilisation (M)	38 ± 28	90 ± 187	1232802 ± 2040662	95 ± 118
Rinçage + Stérilisation (M)	12 ± 15	31 ± 31	7002 ± 13922	74 ± 33
Prélèvement successif (A)	40 ± 46	986 ± 917	40352 ± 65554	54 ± 38
Rinçage + Stérilisation (A)	63 ± 32	848 ± 1474	15114 ± 41487	80 ± 26
Stockage 6 h froid	23 ± 23	15 ± 30	27 ± 7	37 ± 23
Stockage 6 h ambiant	22 ± 16	31 ± 12	40 ± 28	19 ± 17
Stockage 24 h froid	55 ± 26	31 ± 13	45 ± 35	29 ± 19
Stockage 24 h ambiant	67 ± 21	57 ± 22	65 ± 53	51 ± 14

TABLE S3 – Incertitude dans la collecte des échantillons des marqueurs BacCan (Chien), CGOF1 (Oie), Gull2 (Mouette et Goéland) et le pathogène *C. jejuni*.

Parameter	**BacCan**	**CGOF1**	**Gull2**	***C. jejuni***
Temporelle	119 ± 96	125 ± 85	151 ± 66	173 ± 37
Type équipement (M)	174 ± 143	45 ± 90	128 ± 130	168 ± 118
Prélèvement successif (M)	428 ± 581	NA	137 ± 64	50
Rinçage (M)	71 ± 42	NA	164 ± 112	100
Stérilisation (M)	3314 ± 4825	NA	40 ± 51	NA
Rinçage + Stérilisation (M)	132 ± 19	NA	90 ± 17	100
Prélèvement successif (A)	78 ± 42	NA	81 ± 43	NA
Rinçage + Stérilisation (A)	85 ± 23	NA	120 ± 55	NA
Stockage 6 h froid	100	234 ± 261	44 ± 31	62
Stockage 6 h ambiant	NA	602 ± 582	1339 ± 2109	75 ± 35
Stockage 24 h froid	92	44 ± 44	104 ± 18	100
Stockage 24 h ambiant	NA	174 ± 135	151 ± 183	100

3. Dynamique temporelle de la qualité bactériologique en Seine et en Marne

Manel Naloufi[1,2], Claire Thérial [2], Aurélie Janne[3], Marion Delarbre [1], Paul Kennouche [1], Françoise S. Lucas[2]

[1] Direction de la Propreté et de l'Eau - Service Technique de l'Eau et de l'Assainissement, 27 rue du Commandeur 75014 Paris, France ; manel.naloufi@paris.fr,

[2] Leesu, Université Paris-Est Créteil, École des Ponts ParisTech, 61 avenue du Général de Gaulle, 94010 Créteil Cedex, France ; lucas@u-pec.fr

[3] Syndicat Marne Vive, Maison de la Nature, 77 quai de la Pie, 94100 Saint-Maur-des-Fossés, France ; aurelie.janne@marne-vive.com

Résumé : Dans le cadre de l'ouverture de sites de baignade en Marne et en Seine, il a été observé que, par temps de pluie, les seuils réglementaires pour le classement des baignades sont souvent dépassés dans ces deux rivières. Des modélisations exponentielles inverses ont permis d'estimer un taux de disparition et trois indicateurs de résilience. L'analyse du taux de disparition estimé à partir des mesures réglementaires pour *E. coli* a révélé des valeurs de 0,44 ± 0,35 jr^{-1} en Marne et de 0,47 ± 0,32 jr^{-1} en Seine. De plus, une analyse des mesures effectuées avec le dispositif ColiMinder a été réalisée en Seine avec des intervalles de mesure variant de 2 à 24 heures, et a révélé une sous-estimation de la résilience et de la résistance à mesure que l'intervalle augmente. Cette tendance est particulièrement marquée avec un intervalle de 24 heures, où une différence significative a été observée. Le taux de disparition ainsi que l'analyse de la résilience estimés à partir des mesures réglementaires étaient similaires pour les différentes stations étudiées sur la Seine et la Marne. Ce qui n'exclut pas la possibilité de généraliser à l'ensemble des sites en région parisienne. De plus, la simulation d'un rejet en Marne a permis d'estimer un taux de mortalité pour *E. coli* de 0,97 ± 0,48 jr^{-1}. Ces paramètres pourront alimenter des modèles hydrodynamiques pour la gestion des futurs sites de baignades en Marne.

Mots clés : baignades, riviére urbaine, sources, contaminations, fèces d'animaux, pluies, *E. coli*, taux de décroissance, taux de disparition

3.1. Introduction

Depuis quelques décennies, la reconquête des zones de baignade est devenue une priorité dans de nombreuses régions d'Europe, en réponse aux aspirations croissantes des citoyens pour des activités de loisir en plein air et une meilleure qualité de vie (Kistemann et al., 2016; Schreiber et al., 2015). Dans un contexte de prise de conscience environnementale, la qualité des eaux de baignade a fait l'objet de mesures réglementaires strictes, encadrées par la directive européenne 2006/7/CE, qui vise à protéger la santé publique en garantissant la qualité microbiologique des eaux destinées à la baignade.

Ainsi, en Île-de-France, une volonté marquée de réouvrir l'accès aux rivières urbaines est observée, notamment pour la Marne et la Seine, afin de permettre à nouveau la baignade (Bouleau et al., 2024). Toutefois, l'urbanisation de ce territoire génère des risques sanitaires forts en raison de contaminations d'origine domestique et industrielle, y compris les micro-organismes pathogènes d'origine hydrique, émanant des diverses sources fécales. Les principales sources de contamination fécale dans les rivières urbaines incluent les rejets des stations d'épuration, les rejets des réseaux d'assainissement pluvial, les dysfonctionnements des réseaux d'assainissement, les habitations mal raccordées, les rejets des embarcations, ainsi que les déjections animales (Passerat et al., 2011; Droppo et al., 2009; Guérineau et al., 2014). Il y a également les sources diffuses le long de la rivière, notamment celles issues du ruissellement lors des précipitations, lessivant les surfaces urbaines et transportant divers contaminants chimiques et microbiologiques. À cela s'ajoute les événements de remise en suspension des sédiments liés au débit et au transport fluvial (Kay et al., 2008; Devane et al., 2020; Wuijts et al., 2022b; Droppo et al., 2011; Garcia-Armisen and Servais, 2009; Fries et al., 2008). Bien que l'application de la réglementation en Europe ait permis une amélioration de la qualité des eaux de surface, les sources diffuses restent problématiques, et les événements pluvieux exacerbent les risques de déversement d'eaux usées non traitées (Whelan et al., 2022).

La directive européenne 2006/7/CE, concernant la gestion de la qualité des eaux de baignade, vise à protéger, préserver et améliorer la qualité de l'environnement en continu ainsi qu'à protéger la santé humaine. Du point de vue de cette directive, le suivi de la qualité microbiologique des eaux de baignade est actuellement quantifié à l'aide des bactéries indicatrices fécales (BIF). La concentration en *Escherichia coli*, espèce membre du microbiote intestinal des humains et animaux homéothermes, est couramment utilisée comme indicateur de la contami-

nation fécale des eaux. Sa présence est associée aux risques de gastro-entérites liés à des agents pathogènes d'origine hydrique, notamment dans les eaux de baignade (Lucas and Servais, 2016).

La distribution et le devenir des bactéries, présentes dans les fèces ou dans un rejet du réseau d'assainissement, dans les eaux de surface dépend généralement de la dilution du rejet par la rivière, de la dispersion des bactéries dans la colonne d'eau, de leur taux et vitesse de sédimentation et de leur taux de mortalité (Davies et al., 1995; Cho et al., 2010). Des modèles transport-dispersion sont utilisés pour mieux comprendre les mécanismes de transport et de propagation des BIF en prenant en compte les processus d'advection, dispersion, sorption, sédimentation, resuspension et de mortalité (Jalliffier-Verne et al., 2017). Les BIF peuvent soit persister dans l'environnement, soit disparaître rapidement et leur survie dépendra de l'espèce mais également de leur exposition à diverses influences environnementales (Devane et al., 2018; Korajkic et al., 2019). La capacité de survie des BIF est due probablement à des facteurs qui permettent, par exemple, à *E. coli* de survivre et/ou de croître à l'extérieur de l'hôte tels que la température et la matière organique (Ishii and Sadowsky, 2008). Des études ont montré que la disparition des BIF est causée par un certain nombre de facteurs environnementaux, dont la température, la matière organique, la lumière du soleil et le microbiote aquatique (prédation par les protozoaires et métazoaires, compétition bactérienne et lyse virale) (Davies et al., 1995; Korajkic et al., 2019). De plus, il a également été démontré que l'association avec les sédiments améliore la capacité de survie d'*E. coli* dans le milieu aquatique, dû à la présence de matière organique et de nutriments (Zimmer-Faust et al., 2017; Korajkic et al., 2019). Les sédiments servant potentiellement d'habitat secondaire (par rapport à l'intestin des espèces endothermes), peuvent alors représenter une source de BIF de par leur remise en suspension (Devane et al., 2018; Petersen and Hubbart, 2020). Toutefois, il existe peu d'études sur les facteurs impactant la survie des BIF dans les habitats secondaires (eau, sédiments, sols) et les connaissances souvent basées sur quelques expériences en laboratoire limitent la capacité à quantifier et prédire l'impact de ces processus sur les variations de concentrations en BIF dans la colonne, sous différentes conditions climatiques (Petersen and Hubbart, 2020).

La dynamique temporelle et spatiale des BIF pendant et après un événement polluant reste encore mal comprise, notamment durant les pollutions de court terme affectant la qualité de l'eau de la baignade pendant moins de 72 heures selon l'agence française de sécurité sanitaire de l'environnement (Duboudin et al., 2007). Par exemple, lors d'un incident sur le réseau ou lors d'un événement pluvieux, le pollutographe montre que la concentration en BIF s'élève après une

phase de latence, atteint un pic, puis décroît pour revenir proche de la ligne de base antérieure à l'événement de pollution (Stumpf et al., 2010; Tornevi et al., 2014; Oliver et al., 2015). Habituellement, le terme utilisé pour évaluer la diminution des concentrations en BIF au sein d'un événement polluant est le taux de décroissance (Gronewold et al., 2011; Nakhle et al., 2021; Passerat et al., 2011; Beaudeau et al., 2001; Korajkic et al., 2014). Ce taux exprime généralement la réduction des concentrations, influencée par les mécanismes de mortalité microbienne, liés à la prédation par des bactériovores, la lyse virale, ainsi que par l'exposition à la lumière solaire et à la température (Passerat et al., 2011; Nakhle et al., 2021; Servais et al., 2007a). D'autres termes, comme le taux d'inactivation (Gronewold et al., 2011; Carneiro et al., 2018; Blaustein et al., 2013; Noble et al., 2004), et le taux de survie (Ogorzaly et al., 2010; Carneiro et al., 2018) ont été également employés au niveau de la littérature pour des expériences réalisées en laboratoire ou *in situ* en rivière, généralement pour évaluer l'impact d'un ou de plusieurs paramètres. Lors de l'estimation des taux de décroissance par expérimentation *in situ*, les microcosmes le plus souvent utilisés sont des bouteilles ou des sacs à dialyse immergés dans l'eau de surface (Ahmed et al., 2015; Maraccini et al., 2016). Par rapport aux bouteilles, les sacs à dialyse offrent l'avantage de permettre l'échange d'eau et de nutriments entre l'intérieur du sac et la rivière, tout en retenant les cibles microbiennes (Mattioli et al., 2017; Maraccini et al., 2016). Ces expériences en microcosmes, si elles sont plus réalistes que les expériences en laboratoire, ne permettent pas néanmoins d'évaluer les apports ni la dilution des contaminants en amont, ni l'effet de la sédimentation.

Sur un site donné, la dynamique temporelle observée lors d'un événement polluant dans les bases de données de suivi de la qualité de l'eau va résulter des caractéristiques hydrologiques de la rivière, de phénomènes physiques de dilution, de dispersion, de sédimentation et de transport, combinés au taux de décroissance des bactéries. Des termes comme le taux de disparition (Servais et al., 2007a; Schultz-Fademrecht et al., 2008) ou taux de perte (Schultz-Fademrecht et al., 2008) ou taux de dissipation (Xiao et al., 2024) sont le plus souvent employés pour quantifier la perte observée de BIF au fil du temps sur un site donné et qui résulte de l'action combinée de divers processus présentés au niveau de la figure 3.1 (Servais et al., 2007a; Devane et al., 2007; Carneiro et al., 2018). Le taux de disparition peut être défini comme le taux auquel les bactéries sont éliminées et disparaissent au cours du temps (Gronewold et al., 2011; Servais et al., 2007a; Schultz-Fademrecht et al., 2008). Nous avons choisi pour la suite de notre étude, d'utiliser les termes de taux de mortalité pour la décroissance mesurée lors des expériences en

mésocosmes et de taux de disparition pour la décroissance observée dans les données de qualité de l'eau collectées en rivière.

Les taux de mortalité et de disparition sont généralement estimés en ajustant une équation exponentielle aux concentrations bactériennes mesurées au fil du temps dans les expérimentations ou lors d'un suivi temporel à un même site de la rivière. Cette équation est exprimée comme une cinétique de décroissance de premier ordre (Nakhle et al., 2021; Geeraerd et al., 2005; Gronewold et al., 2011). Ce modèle mathématique est privilégié, étant souvent appliqué dans les études de décroissance des BIF. Il constitue une base solide pour suivre la diminution de la concentration dans le temps. L'estimation des taux de décroissance des BIF est cruciale pour nourrir les modèles hydrodynamiques utilisés pour la prédiction des concentrations en BIF sur les zones de baignade. Par défaut, des valeurs issues de la littérature sont souvent utilisées dans les modèles déterministes, car les paramètres de décroissance sont rarement évalués expérimentalement. Or, ceci introduit une forte incertitude sur les concentrations prédites, car les taux de décroissance utilisés peuvent fortement affecter la prédiction (Eregno et al., 2018).

Une mesure complémentaire à la décroissance de la contamination est le calcul de la résilience et de la résistance du site face aux perturbations polluantes, ce qui permet d'évaluer la vulnérabilité d'un site ou d'un écosystème (Imani et al., 2021; Xiao et al., 2024). La résistance est la capacité d'un système à résister à une perturbation et la résilience est largement interprétée comme la capacité d'un système à absorber et supporter, puis à se rétablir rapidement après une perturbation (Mirauda et al., 2021). Il existe une diversité d'explications associées à la notion de résilience, souvent définie de manière vague. Cependant, pour établir une théorie utilisable dans différents domaines, il est crucial de partir de définitions précises et d'offrir une comparaison mathématique des diverses mesures de résilience (Krakovská et al., 2024). La résilience de la qualité de l'eau est définie comme la capacité des systèmes aquatiques à se rétablir après une détérioration de la qualité de l'eau due à un évènement polluant (Xiao et al., 2024). Pour une pollution fécale, cette capacité de résistance ou de résilience va dépendre de l'intensité de la pollution, du taux de décroissance des BIF mais également des caractéristiques hydromorphologiques du site et du bassin versant. L'estimation de la résilience offre une approche robuste pour la prise de décision en matière de gestion de la qualité de l'eau mais reste encore peu utilisée (Mirauda et al., 2021). Les indicateurs de résilience permettent d'évaluer l'impact des pollutions pluviales, offrant ainsi des outils précieux pour l'évaluation de la qualité de l'eau dans les zones étudiées (Noble et al., 2004). Plus le taux de décroissance ou de disparition est

élevé, le temps de retour est court et l'amplitude de variation est faible, plus la résilience du système aquatique est forte (Xiao et al., 2024; Krakovská et al., 2024).

Dans ce contexte, nous avons souhaité établir une approche qui englobe une estimation de ces différents aspects de la dynamique temporelle des concentrations en BIF. Cette approche combine de l'expérimentation *in situ* avec des sacs à dialyse pour déterminer le taux de mortalité (Figure 3.1), avec l'utilisation des données de suivi microbiologique réglementaire et des données de suivi en temps quasi réel par le système ColiMinder pour estimer le taux de disparition et le niveau de résistance et de résilience sur des sites contrastés (Figure 3.1). Les taux de décroissance ont été quantifiés par modélisation de la diminution des concentrations en *E. coli* à l'aide de courbes exponentielles inverses. La Marne et la Seine ont été investiguées pour savoir si les mêmes taux pouvaient être appliqués sur les deux fleuves franciliens. En effet, il existe encore peu de données disponibles pour ces deux rivières en région parisienne. Les modèles actuellement utilisés (Prose, Telemac) pour prédire les concentrations en *E. coli* dans ces deux rivières se basent généralement sur les données expérimentales de taux de décroissance en présence et absence de broutage par les protozoaires qui sont disponibles pour la Seine et la Marne (Servais et al., 2007b), et alternativement sur des taux de disparition qui ont été estimés à partir de données issues d'une unique campagne d'échantillonnage sur la Marne (Van et al., 2022). Notre étude vise à procurer des outils et des données pour la gestion des baignades qui devraient s'ouvrir à l'horizon 2025 en Seine et en Marne, en se focalisant sur l'estimation des taux de décroissance et de disparition et sur l'analyse de la résistance et de la résilience d'*E. coli*, qui reste le critère le plus déclassant pour la gestion journalière de ces deux rivières en Ile-de-France (Lucas et al., 2020; Mouchel et al., 2020).

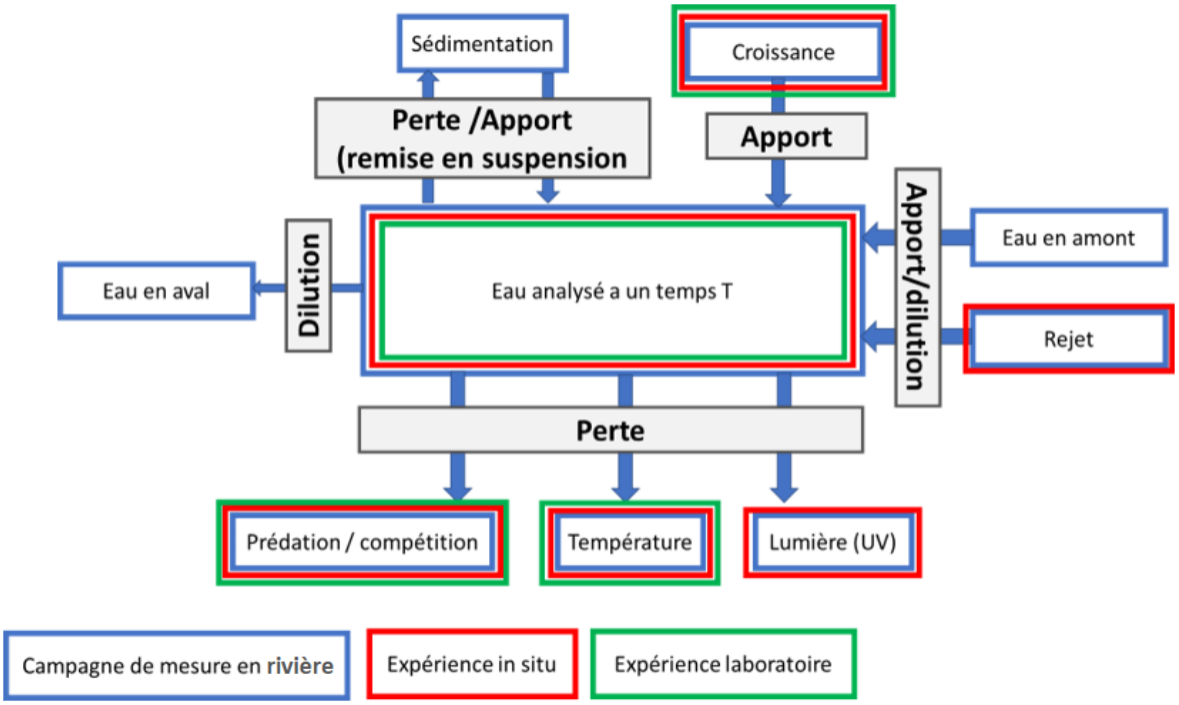

FIGURE 3.1 – Facteurs pouvant avoir un effet sur la dynamique des concentrations en BIF dans les bases de données de suivi *in situ* (bleu), et sur la mortalité des BIF mesurées lors des l'expérimentations *in situ* (rouge) et en laboratoire (vert).

3.2. Matériels et Méthodes

3.2.1. Sites d'étude en rivière

L'étude s'est focalisée sur 6 sites en région parisienne (France) : 3 sites en Marne (SMV1, SMV10 et SMV14) et 3 sites en Seine (pont de l'Alma, pont de Tolbiac rive droite, Tolbiac rive gauche) (Figure 3.2). Les sites en Marne ont été sélectionnés car ce sont des candidats pour l'ouverture de baignade en 2025, vu leur qualité microbiologique et leur facilité d'aménagement (Noury et al., 2018) et parce qu'ils représentent des situations contrastées. En ce qui concerne la Seine, les sites ont été sélectionnés du fait de la présence d'un système ColiMinder en amont et qu'il s'agissait de sites suivis par la Ville de Paris pour les épreuves des Jeux Olympiques et Paralympiques 2024.

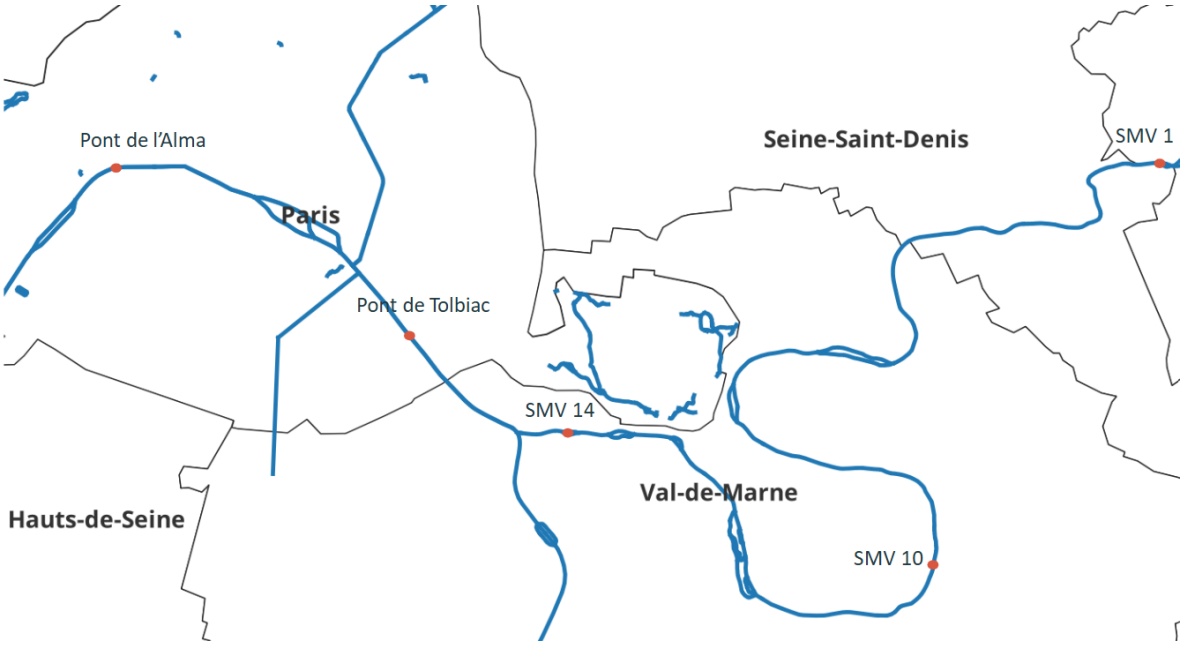

FIGURE 3.2 – Schéma des sites étudiés.

3.2.2. Bases de données

La base de données du Syndicat Marne Vive regroupe les données des mesures bactériologiques et physico-chimiques réalisées de début juin à mi-septembre les années 2015, 2017 à 2022, une à deux fois par semaine sur la Marne. Les données ont été mesurées par Eurofins en 2015 et par le laboratoire départemental des eaux du Val de Marne de 2017 à 2022. Les échantillons ont été prélevés selon la norme française FD T 90-523-1, les mesures microbiologiques ont été réalisées selon les méthodes normalisées françaises NF EN ISO 9308-3 pour *E. coli*. Les mesures physico-chimiques ont été réalisées selon les méthodes normalisées françaises, NF EN 27888 pour la conductivité électrique, NF EN 872 pour matières en suspension (MES). Pour la pluviométrie, les données ont été fournies par les réseaux de pluviomètres des Conseils départementaux du Val-de-Marne (stations CHAM23 et MAIS32), de la Seine-Saint-Denis (station NE-17) et de la Ville de Paris (station PL14). Pour le site SMV14, les prélèvements ont été réalisés en 2015 en rive gauche (Maisons-Alfort) puis en rive droite (Saint-Maurice) pour les années suivantes. Le débit à la station de Gournay-sur-Marne a été obtenu à partir de la Station hydrométrique - F664 0001 04 sur le site HydroPortail (https ://hydro.eaufrance.fr/stationhydro/).

La base de données de la Ville de Paris regroupait les mesures effectuées de début-juin à fin-septembre sur la période 2015-2023, de manière hebdomadaire ou bi-hebdomadaire sur la Seine (France). En 2015 et de 2018 à 2023 les échantillonnages au pont de Tolbiac ont été effectués en rive droite (RD), et de 2017 à 2023 en rive gauche (RG). Pour le pont de

l'Alma, les échantillonnages ont été réalisés entre 2017 et 2023. Les mesures microbiologiques et physico-chimiques ont été réalisées par le laboratoire d'Eau de Paris selon les méthodes normalisées françaises NF EN ISO 9308-3 pour *E. coli*, NF EN ISO 7027-1 pour la turbidité, et NF EN 27888 pour la conductivité électrique. Les données pluviométriques ont été obtenues à partir du réseau de pluviomètres de la Ville de Paris (stations PL1 et PL5). Le débit au pont d'Austerlitz a été obtenu à partir de la station hydrométrique F700 0001 03 sur le site HydroPortail (https ://hydro.eaufrance.fr/stationhydro/).

Nous avons en plus utilisé les données du système automatisé ColiMinder (Vienna Water Monitoring, VWM) au niveau de 2 sites en Seine en rive gauche (Pont de l'Alma entre 2020 et 2023 et Pont de Tolbiac RG entre 2021 et 2023 ; Paris, France). Les analyses ont été réalisées toutes les 2 heures. Les données pluviométriques ont été obtenues à partir du réseau de pluviomètres de la Ville de Paris (stations PL1 et PL5).

À partir de ces bases de données, une sélection des données a été réalisée sur deux critères : i) uniquement les concentrations en *E. coli* mesurées *a minima* deux fois par semaine entre juin et septembre pour les bases de données avec suivi réglementaire et ii) uniquement des événements pluvieux isolés qui génèrent une pollution, suivis de 3 jours de temps sec (données réglementaires et ColiMinder). Pour les données réglementaires, les événements pluvieux retenus devaient inclure une première analyse bactériologique le lendemain de la pluie ou le dernier jour de pluie (concentration initiale), suivie par au moins 3 jours de temps sec incluant *a minima* un prélèvement pendant cette période, ceci afin d'observer et de modéliser la diminution des concentrations en *E. coli*.

À partir de ces événements sélectionnés dans les deux bases de données, la concentration initiale correspond à la première mesure de concentration pour l'événement pluvial, tandis que la concentration en temps sec est celle observée après trois jours de temps sec. Le taux de disparition ainsi que les paramètres de résilience et de résistance ont ensuite été calculés.

3.2.3. Expérimentation *in situ*

La vitesse de mortalité d'*E. coli* a été étudiée à l'aide de sacs à dialyse. Cette expérience simule un rejet de station d'épuration dans la Marne, afin de calculer le taux de mortalité des *E. coli*. Cette expérience a été réalisée du 27 mai au 1er juin 2019 à quelques mètres en amont du site SMV14. Les autorisations ont été obtenues auprès de Voies Navigables de France pour l'accès au site et l'installation des systèmes expérimentaux dans la Marne. Un total de 18 L

d'eau de la Marne au droit du rejet de l'usine de traitement des eaux usées Marne Aval (Syndicat Intercommunal d'Aménagement de l'Agglomération Parisienne) a été collecté au seau, dont 10 L ont été autoclavés 20 min à 120°C afin d'être utilisés comme contrôle. Les sacs à dialyse Spectra/Por®1 (seuil de coupure de 6 à 8 kD) ont été remplis avec environ 180 ml d'eau du rejet autoclavé ou non. Les sacs ont ensuite été fermés à l'aide de pinces et attachés par du fil de pêche à des cagettes en plastique de 32 cm x 28 cm. Les cagettes lestées ont été placées à environ 10-20 cm sous la surface et arrimées à la berge à l'aide d'une corde. Deux traitements ont été réalisés en cinq exemplaires pour 4 pas de temps (n=30) : (1) eau autoclavée, (2) eau du rejet pure (non diluée). Les échantillons ont été collectés à quatre moments différents : après 24 h (T1), 48 h (T2) et 72 h (T3). Les sacs à dialyse ont été collectés le matin entre 10 et 11 h et placés dans des sacs plastiques scellés, partiellement remplis d'eau de la Marne provenant du site de prélèvements et placés dans une glacière pour le transport jusqu'au laboratoire.

3.2.4. Quantification des BIF

Pour les expériences *in situ*, les densités (NPP/100 mL) en *E. coli* dans l'eau du rejet autoclavée et pure à T0 et l'eau contenue dans chaque sac à dialyse (T1 à T3) ont été ensemencées sur les microplaques MUG/EC (BioRad) selon la méthode de référence NF EN ISO 9308-3 pour *E. coli*. Le calcul du NPP/100 mL dans un intervalle de confiance de 95% a été réalisé à l'aide d'une feuille de calcul Excel proposée par Jarvis et al. (2010). Elle fournit également un indicateur de rareté qui permet de détecter des incohérences dans les comptages obtenus pour l'ensemble des dilutions. La turbidité (Turbidimètre Hach),la conductivité et le pH (sonde multiparamétrique Eutech) ont été mesurés dans les échantillons d'eau à T2 et T3.

3.2.5. Modélisation de la dynamique temporelle après une pluie

L'analyse de la décroissance bactérienne s'effectue par une estimation de la vitesse de diminution des concentrations en *E. coli* à l'aide d'un modèle exponentiel inverse. À partir des données mesurées expérimentalement, des suivis de qualité microbiologique réglementaire aux 6 sites, ou des suivis en Seine avec le système ColiMinder, la constante de cinétique des courbes de décroissance des différents événements sélectionnés a été déterminée empiriquement avec un modèle linéaire exponentiel. En utilisant ce paramètre cinétique déterminé, les courbes de décroissance ont été modélisées pour générer des valeurs prédites par le modèle pour chaque événement sélectionné. Ces valeurs servaient ensuite à déterminer avec un modèle log-linéaire

les taux de mortalité à partir des expériences, et les taux de disparition à partir des données de suivi en Marne et en Seine et des données du système ColiMinder. L'ensemble des étapes de modélisation ont été effectuées sous R (R-Core-Team, 2018).

3.2.5.1. Modélisation exponentielle inverse

A partir des valeurs de la constante de cinétique, des tableaux contenant les valeurs prédites de concentration pour chaque événement pour les 6 sites (issues de la base de données réglementaires) et pour les 2 sites équipés du ColiMinder, ainsi que pour les expériences *in situ* ont été générés en modélisant la courbe de décroissance à l'aide d'une équation exponentielle inverse d'ordre 1 (Gronewold et al., 2011; Geeraerd et al., 2005).

$$C(t) = (C_0 - C_{res}).e^{-K_1 t}\left(\frac{e^{K_1 S}}{1 + e^{-K_1 t}\left(e^{K_1 S} - 1\right)}\right) + C_{res} \tag{3.4}$$

Cette équation utilise la constante de cinétique (K_1, en jr_{-1}), l'épaulement (S, en jr), la concentration initiale (C_0, en NPP/100 mL), la densité de la population (C_{res}, en NPP/100 mL) et le temps (t, en jr) comme données d'entrée (Geeraerd et al. (2005), équation 1). K_1 est une inconnue qui est estimée de manière empirique en utilisant les données mesurées des concentrations en BIF au cours du temps. Dans la suite de notre étude, l'outil de modélisation de la décroissance de Geeraerd et al. (2005) a été utilisé pour établir des modèles log linéaire en prenant en compte la population résiduelle (C_{res}) et l'épaulement S. La plupart des études sur la modélisation de la décroissance bactérienne n'incluent pas la période de latence (S) lors de la modélisation. Or cette dernière peut exister, dû à l'état de la population microbienne ou à un artéfact expérimental (Mattioli et al., 2017; Brooks and Field, 2016).

Pour chaque site, une courbe moyenne des courbes modèles a été réalisée. Pour cela, les données des courbes modèles et les valeurs mesurées ont été exprimées en pourcentage (de manière à avoir à t0 un pourcentage de 100% puis une décroissance au cours du temps). Les courbes moyennes ont été réalisées avec un intervalle de confiance de 95% (R-Core-Team, 2018). Par la suite, les données prédites pour chaque événement ont servi à estimer les valeurs de taux de disparition ou de taux de mortalité (K_2 - pente des modèles).

3.2.5.2. Estimation des taux de mortalité et de disparition

Le taux de mortalité/disparition (K_2) a été calculé en utilisant un modèle linéaire exponentiel. K_2 représente un taux de décroissance qui peut correspondre au taux de mortalité d'une même population de BIF dont les concentrations évoluent dans le temps, et qui est déterminé

dans les expériences (Korajkic et al., 2014; Carneiro et al., 2018; Passerat et al., 2011). K_2 peut aussi représenter un taux de disparition dans le cas de la modélisation d'une série temporelle d'un suivi de qualité réalisé sur un site de baignade. Le taux K_2 est exprimé en unités de temps, généralement en jr^{-1}.

3.2.6. Indicateurs de résilience et de résistance

La résilience et la résistance désignent la capacité des systèmes aquatiques à se transformer, s'adapter et se maintenir face à des perturbations, telles que des épisodes de pollution de courte durée (Xiao et al., 2024; Krakovská et al., 2024). En ce qui concerne la qualité de l'eau d'une rivière, la résilience ou la résistance peut être quantifiée en comparant les concentrations observées lors d'un événement de pollution à celles enregistrées dans des conditions normales (Xiao et al., 2024). Dans le cas de notre étude, les temps secs (minimum 3 jours après une pluie) ont été considérés comme des conditions normales, et les événements pluvieux comme des perturbations.

Au niveau de notre étude, nous avons utilisé 3 paramètres :

- Le temps de retour (T_{90}) est l'un des indicateurs les plus couramment utilisés pour l'analyse de la résilience (Nakhle et al., 2021; Ogorzaly et al., 2010; Carneiro et al., 2018; Xiao et al., 2024; Schultz-Fademrecht et al., 2008; Noble et al., 2004). Il représente le temps nécessaire pour atteindre une réduction de 90% des concentrations initiales en *E.coli*, indiquant le retour à un état proche de l'équilibre, soit les concentrations par temps sec.

- L'amplitude de récupération ($AV_{après}$) représente l'amplitude de changement entre la concentration initiale au pic de pollution et celle mesurée en période de temps sec après la perturbation. Elle est également exprimée en pourcentage de différence. Ce paramètre permet de mesurer l'ampleur de la récupération après une pollution (Krakovská et al., 2024).

- L'amplitude de variation de la pollution (AV_{avant}) quantifie l'écart des concentrations en *E. coli* entre le pic de pollution (concentration dite initiale pour la mesure de décroissance) et le temps sec précédent chaque événement pluvial analysé. Exprimée en pourcentage de différence, ce paramètre permet de mesurer la capacité d'un système à résister et l'ampleur des changements subis par la qualité de l'eau suite à une pollution (Krakovská et al., 2024; Mirauda et al., 2021).

3.2.7. Traitements statistiques

Toutes les analyses statistiques ont été réalisées à l'aide du logiciel R (V3.5.1, (R-Core-Team, 2018). Les tests de Kruskal-Wallis ou de Friedman (appariés) étaient suivis de tests post-hoc (tests de Wilcoxon multiples par paire). Les p-valeurs des comparaisons par paires ont été ajustées avec une correction de Bonferroni. Pour l'analyse de variance, le test post-hoc HSD de Tukey a été utilisé. Afin de déterminer les paramètres qui peuvent influencer les valeurs K_2 et les 3 paramètres de résilience ou de résistance obtenues pour chaque événement à partir des données réglementaires mesurées en Seine et en Marne et des données du système ColiMinder, des modèles linéaires ont été réalisés en utilisant des paramètres physico-chimiques, hydro-météorologiques et bactériologiques (turbidité, conductivité, température, pluviométrie cumulée sur l'événement avant prélèvement (48 h), débit, concentration initiale en *E. coli*), station de mesure). Les données de pluviométrie ont été testées sous forme quantitative (mm cumulés), mais ont également été discrétisées en deux catégories : pluie faible (<10 mm cumulés) (Islam et al., 2017) et pluie élevée (>10 mm cumulés) (Gebremichael et al., 2014).

Les variables colinéaires ont d'abord été éliminées après calcul du facteur d'inflation de la variance pour tester la multicolinéarité (librairies usdm, MASS et leaps). La température et la conductivité, en raison de leur colinéarité, n'ont pas été retenues. La turbidité a été testée pour chaque modèle, mais n'a pas été identifiée comme un paramètre significatif du modèle. Elle a été écartée lors du processus de validation du modèle linéaire final.

Sur les variables retenues, une vérification de la distribution des données quantitatives a été effectuée à l'aide des diagrammes quantiles-quantiles (librairies fitdistrplus et car). Nous avons appliqué différentes méthodes de modélisation, notamment les modèles linéaires simples (lm), les modèles linéaires généralisés (glm), ainsi que les modèles linéaires mixtes (lmm) et les modèles linéaires généralisés mixtes (glmm), afin de prendre en compte à la fois les effets fixes et aléatoires entre les sites. Une sélection du modèle le plus significatif a été réalisée en utilisant la méthode descendante et en se basant sur le critère d'information d'Akaike (AIC, librairies lme et car) suivie par une validation du modèle en vérifiant la distribution et l'indépendance des résidus (Zuur et al., 2009). Pour l'ensemble des analyses statistiques, le seuil de significativité a été fixé à 5%.

3.3. Résultats

3.3.1. Détermination du taux de mortalité

Pour l'expérience *in situ* proche du site SMV14 en Marne, les modèles linéaires exponentiels utilisés pour le calcul de la constante de cinétique sur chaque réplicat séparément (1,17 ± 0,58 jr^{-1}, n=5) étaient non significatifs (p > 0,06) avec un r² moyen de 0,98 ± 0,01. Cependant, en regroupant l'ensemble des réplicats, le modèle s'ajustait de manière significative aux données (n=5, p < 0,001). Le modèle de décroissance présenté en figure 3.3 a ensuite permis l'estimation du taux de mortalité (K_2) d'*E. coli*, avec une valeur moyenne de 0,97 ± 0,48 jr^{-1}, soit une décroissance après plus d'un jour.

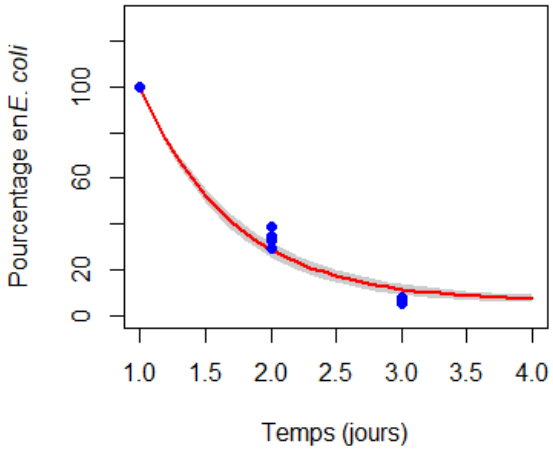

FIGURE 3.3 – Courbe modélisant la décroissance (ligne noire) avec les valeurs obtenues par expérimentation en sac à dialyse pour le dénombrement des *E. coli* avec l'eau du rejet. Les valeurs mesurées lors de l'expérimentation sont représentées par des carrés rouges.

Lors de l'analyse de la mortalité d'*E. coli* en Marne à proximité du site de baignade potentielle SMV14, les contrôles (eau de rejet autoclavée) étaient restés stériles, montrant ainsi que les sacs à dialyse étaient étanches. De plus, les mesures de pH et de conductivité étaient similaires entre l'eau de la Marne et l'intérieur des sacs à dialyse, montrant que le système était bien semi-ouvert (Tableau 3.1).

TABLE 3.1 – Valeurs des paramètres physico-chimiques de l'eau de la Marne et de l'eau contenue dans les sacs à dialyse mesurés lors du 2ème et 3ème jours de l'expérimentation *in situ*.

Date de prélèvement	**Eau / Traitement**	**pH**	**Conductivité (S/cm)**	**Turbidité (FTU)**	**Concentration initiale (NPP/100 mL)**
2ème jour	Marne	8,19	558	30,24	412
	Rejet	8,06	560	5,11	120000
	Rejet autoclavé	8,19	568	13,13	0
3ème jour	Marne	8,07	550	5,71	
	Rejet	8,01	556	6,96	
	Rejet autoclavé	8,15	552	7,75	

3.3.2. Dynamique de disparition d'*E. coli* en rivière

Une variabilité interannuelle a été observée sur la saison de baignade de juin à septembre en ce qui concerne la pluviométrie (Tableau 3.2). De même, il existait une hétérogénéité spatiale sur la Seine à Paris et l'aval de la Marne depuis Gournay-sur-Marne à la confluence avec la Seine. Globalement, les étés les plus pluvieux étaient 2017 (188 à 303 mm cumulés, et 42 à 66 jours de pluie), 2021 (288 à 382 mm soit 42 à 63 jours de pluie), et 2023 (204 à 299 mm de pluie, et 32 à 44 jours de pluie) (Tableau 3.2).

TABLE 3.2 – Nombre de jours de pluie et pluviométrie (mm) cumulée durant la période estivale (1er juin - 30 septembre) chaque année à Paris et sur l'aval de la Marne, en fonction des pluviomètres les plus proches de chaque station de prélèvement.

Année	**SMV 1**		**SMV 10**		**SMV 14**		**Alma**		**Tolbiac**	
	Nb de jours	Cumul	Nb de jours	Cumul	Nb de jours	Cumul	Nb de jours	Cumul	Nb de jours	Cumul
2015	40	186	20	111	20	140	29	160	37	177
2016	46	171	42	188	58	127	38	112	39	131
2017	66	303	42	193	40	188	57	226	58	389
2018	33	244	32	192	42	175	20	130	26	139
2019	31	210	29	168	28	168	22	82	23	119
2020	38	181	38	197	44	171	35	136	37	145
2021	54	382	51	363	63	303	42	288	43	310
2022	42	197	41	217	38	184	35	158	40	242
2023	42	299	35	225	37	204	36	250	34	240

Cette variabilité a permis de pouvoir sélectionner un ensemble d'événements pluvieux entre 2 et 42 mm de pluie cumulée, suivis au minimum par 3 jours de temps sec, associant une décroissance en *E. coli*. Pour la base de données réglementaires, ont été sélectionnés 33

événements en Marne (12 au site SMV1, 11 au site SMV10 et 10 au site SMV14) et 13 en Seine dont 5 au pont de l'Alma, 5 au pont de Tolbiac en RD et 3 en RG. Pour la base de données du dispositif ColiMinder, ont été retenus 13 événements pluvieux au pont de l'Alma et 21 au pont de Tolbiac. L'ensemble des pluies sélectionnées se caractérise d'une part par 54% de pluie <10 mm et 46% de pluie ≥10 mm pour les mesures réglementaires, et d'autre part par 62% de pluie <10 mm et 38% de pluie ≥10 mm pour les analyses avec le système ColiMinder.

3.3.2.1. Caractéristiques des événements sélectionnées

3.3.2.1.1. Caractéristiques des données réglementaires

Une variabilité interannuelle a été observée au niveau de chaque site (SMV1, SMV10, SMV14 et pont de Tolbiac RD et RG) (Test de Kruskal-Wallis, p<0,001, p<0,001, p<0,001, p<0,023, p<0,002, n=293, n= 274, n=286, n=208, n=187), à l'exception du pont de l'Alma qui ne présentait pas de différence significative entre les années (Test de Kruskal-Wallis, p=0,058, n=180). Ces données montraient que les années 2018 et 2021 se démarquaient par des concentrations *E. coli* élevées en Marne, et les concentrations de l'année 2020 étaient particulièrement plus basses (Tableau 3.3). Ainsi, en Seine, cette différence entre années était la plus marquée au pont de Tolbiac RG entre 2018 et 2020, avec un niveau de contamination le plus faible en 2020 (Test post-hoc, p=0,003, n=187, Tableau 3.4). En Marne, une plus grande variabilité interannuelle était observée entre les différentes années (Tableau 3.3). Les concentrations en 2021 étaient significativement plus élevées que les autres années pour les 3 sites (Test de Kruskal-Wallis suivi de tests post-hoc, p<0,002 pour tous les sites et les années, n=853) à l'exception du SMV1 pour lequel les concentrations ne différaient pas entre 2021 et 2023 (Test post-hoc, p=1,000, n=78). Ces concentrations élevées en 2021 correspondaient au fait que les cumuls de pluviométrie en 2021 étaient les plus élevés (Tableau 3.2). En 2018, des concentrations nettement élevées ont été constatées de Gournay-sur-Marne à la confluence avec la Seine, atteignant une moyenne de 8708 ± 4416 NPP/100 mL à SMV1 (Tableau 3.4) en raison d'un incident sur le réseau d'assainissement à l'amont de SMV1 (Tests post-hoc, p<0,001, n=293).

Pour l'ensemble des événements sélectionnés dans les bases de données réglementaires de la Seine et la Marne, la concentration en *E. coli* augmentait après la pluie et dépassait le seuil de qualité suffisante de 1800 NPP/100 mL (instruction N°DGS/EA4/2020/111 du 2 juillet 2020) dans 69% des pluies sélectionnées. Au niveau de la Marne, 63% des pluies sélectionnées ont dépassé le seuil de 1800 NPP/100 mL alors qu'au niveau de la Seine, 84% des pluies ont dépassé ce seuil.

TABLE 3.3 – Concentrations en *E. coli* en NPP/100 mL durant la période estivale par année et par site en Marne selon le suivi réglementaire.

Année	SMV 1	SMV 10	SMV 14
2015	950 ± 1759	1512 ± 2825	637 ± 833
2017	1119 ± 2080	2599 ± 2444	580 ± 1227
2018	8708 ± 4416	3927 ± 6046	2182 ± 3531
2019	1316 ± 2362	1222 ± 1533	421 ± 419
2020	786 ± 1246	1493 ± 835	434 ± 373
2021	1813 ± 1962	5903 ± 4371	3462 ± 1900
2022	779 ± 802	2566 ± 2192	586 ± 667
2023	1595 ± 2317	1843 ± 2156	881 ± 1568

TABLE 3.4 – Concentrations en *E. coli* en NPP/100 mL durant la période estivale par année et par site en Seine selon le suivi réglementaire.

Année	Alma	Tolbiac RD	Tolbiac RG
2015	—	4380 ± 3622	–
2017	1750 ± 2276	—	14224 ± 18055
2018	3079 ± 8862	4965 ± 8619	6690 ± 9836
2019	1825 ± 4016	4568 ± 7123	4476 ± 7858
2020	710 ± 954	1910 ± 3105	2210 ± 4478
2021	5733 ± 11218	5436 ± 8083	7737 ± 10665
2022	4190 ± 8546	6778 ± 12096	6096 ± 11340
2023	1292 ± 1914	2275 ± 3274	4140 ± 8408

Les concentrations moyennes de temps sec avant les événements pluvieux étaient statistiquement similaires entre les différents sites, à l'exception entre les sites SMV10 et SMV14, la concentration moyenne étant la plus faible à SMV14 (Test de Kruskal-Wallis, suivi des tests post-hoc, p=0,041, n=22, Tableau 3.5). Chaque pluie était associée à une augmentation significative de la concentration par rapport au temps sec précédant (Test de Wilcoxon apparié, p<0,001, n=46, Tableau 3.5). Les pics de contamination étaient statistiquement similaires entre les différents sites pour chaque épisode de pluie (Test de Kruskal-Wallis, p=0,196, n=46, Tableau 3.5). Une diminution significative de la concentration en *E. coli* était ensuite constatée durant les 3 jours de temps sec suivant chaque pluie (Test de Wilcoxon apparié, p<0,001, n=46, Tableau 3.5). En ce qui concerne les concentrations à partir du $3^{ème}$ jour de temps sec après la pluie, une différence significative a été constatée également uniquement entre SMV10 et SMV14 (Test de Kruskal-Wallis suivi de tests post-hoc, p=0,031, n=22).

TABLE 3.5 – Concentrations en *E. coli* en NPP/100 mL par site pour les événements sélectionnés (temps sec avant la pluie (avant), concentration initiale au pic de pollution (pic) et temps sec après la pluie (après).

Site	**Avant**	**Pic**	**Aprés**
Alma	643 ± 268	5366 ± 3895	1121 ± 164
Tolbiac RG	1502 ± 742	18910 ± 14053	1883 ± 586
Tolbiac RD	1543 ± 1967	15619 ± 17577	1793 ± 3336
SMV1	1080 ± 1679	6334 ± 6918	360 ± 2507
SMV10	1192 ± 850	7278 ± 10214	1511 ± 756
SMV14	488 ± 437	3990 ± 5698	260 ± 341

3.3.2.1.2. Caractéristiques des données de l'analyseur ColiMinder

L'analyse des mesures en *E. coli* par le système ColiMinder durant toute la période estivale (de juin à septembre entre 2020-2023), montre une contamination significativement plus élevée en 2022 comparé aux autres années, que ce soit au pont de l'Alma ou au pont de Tolbiac (Tableau 3.6), à l'exception de 2021 et 2022 au pont de Tolbiac dont les concentrations en *E. coli* ne sont pas significativement différentes (Test de Kruskal-Wallis, p<0,001, n=6625 et 4681 respectivement pour les deux sites, suivi de tests post-hoc, p>0,340).

TABLE 3.6 – Concentrations en *E. coli* en NPP/100 mL estimées par l'analyseur ColiMinder durant toute la période estivale par année et par site

Année	**Alma**	**Tolbiac**
2020	560 ± 104	—
2021	4897 ± 2580	5036 ± 4156
2022	9285 ± 10402	18497 ± 18436
2023	3836 ± 4080	7543 ± 5047

Pour l'ensemble des événements pluvieux sélectionnés dans la base de données du système ColiMinder, la concentration en *E. coli* augmente après la pluie et dépasse le seuil de qualité suffisante de 1800 NPP/100 mL (selon l'instruction N°DGS/EA4/2020/111 du 2 juillet 2020 pour la gestion en cours de saison) dans 79% des pluies sélectionnées. Un total de 69% des pluies sélectionnées dépassait ce seuil au niveau du pont de l'Alma et 85% au niveau du pont de Tolbiac.

Avant la pluie par temps sec, la concentration en *E. coli* était en moyenne de 525 ± 288 NPP/100 mL au pont de l'Alma et de 937 ± 960 NPP/100 mL au pont de Tolbiac, et les deux sites ne différaient pas significativement (Test de Wilcoxon, p=0,456, n=34). L'événement pluvieux

entraînait une augmentation significative de la concentration en *E. coli* quel que soit le site (Test de Wilcoxon apparié, p<0,001, n=34). Le pic de concentration au niveau des 2 sites présentait des concentrations moyennes statistiquement similaires (39769 ± 71990 NPP/100 mL au pont de l'Alma et 35334 ± 38322 NPP/100 mL au pont de Tolbiac) (Test de Wilcoxon, p=0,178, n=34). Une diminution significative de la concentration était observée ensuite pendant les 3 jours de temps sec suivant la pluie (Test de Wilcoxon apparié, p<0,001, n=34). La concentration moyenne en *E. coli* le 3ème et le 4ème jour après l'événement pluvieux était de 1226 ± 2067 NPP/100 mL au pont de l'Alma et 1275 ± 1466 NPP/100 mL au pont de Tolbiac.

3.3.2.2. Estimation du taux de disparition

3.3.2.2.1. Taux de disparition estimés avec les bases de données réglementaires

Quelque soit le site en Seine ou en Marne, les modèles exponentiels permettant une estimation du taux de disparition n'étaient significatifs que pour 24% des événements sélectionnés (Tableau 3.7). En effet, il y avait généralement 2 à 5 mesures par événement et les modèles les plus significatifs ont été observés pour des événements avec des mesures successives avec au minimum un intervalle de 24 h. La constante de cinétique (K_1) a été calculée pour chaque événement pluvial sélectionné et chaque site. Les résultats ont montré que les valeurs de la constante de cinétique étaient comprises entre 0,18 et 1,55 jr^{-1} (Tableau 3.7).

TABLE 3.7 – Constante de cinétique (K_1, jr^{-1}) obtenues par le modèle linéaire exponentiel (p-valeur et R^2), le taux de disparition (K_2, jr^{-1}). Moyenne ± écart type ou [Min : Max]. p-valeur significative (S), non significative (NS) au seuil 0,05. NA non applicable.

Site	**K_1**	**p-valeur**	**R^2**	**K_2**
SMV1 (n=12)	0,60 ± 0,31	S(2) NS(7) NA(3) [0,027 : 0,401]	0,89 ± 0,13	0,36 ± 0,32
SMV10 (n=11)	0,76 ± 0,51	NA(3) NS(8) [0,054 : 0,256]	0,87 ± 0,15	0,46 ± 0,49
SMV14 (n=10)	0,78 ± 0,22	S(2) NS(7) NA(1) [0,010 : 0,388]	0,82 ± 0,18	0,49 ± 0,22
Alma (n=5)	0,82 ± 0,46	S(3) NS(2) [0,002 : 0,498]	0,81 ± 0,21	0,52 ± 0,50
Tolbiac RD (n=5)	0,73 ± 0,31	S(2) NS(3) [0,001 : 0,245]	0,77 ± 0,23	0,44 ± 0,23
Tolbiac RG (n=3)	0,60 ± 0,19	S(2) NS(1) [0,004 : 0,059]	0,88 ± 0,12	0,43 ± 0,12

Les courbes modèles ont ainsi pu être obtenues pour chaque événement observé et ont permis d'affiner la valeur du taux de disparition. Un intervalle de confiance à 95% a été estimé pour chaque courbe moyenne (Figure 3.4).

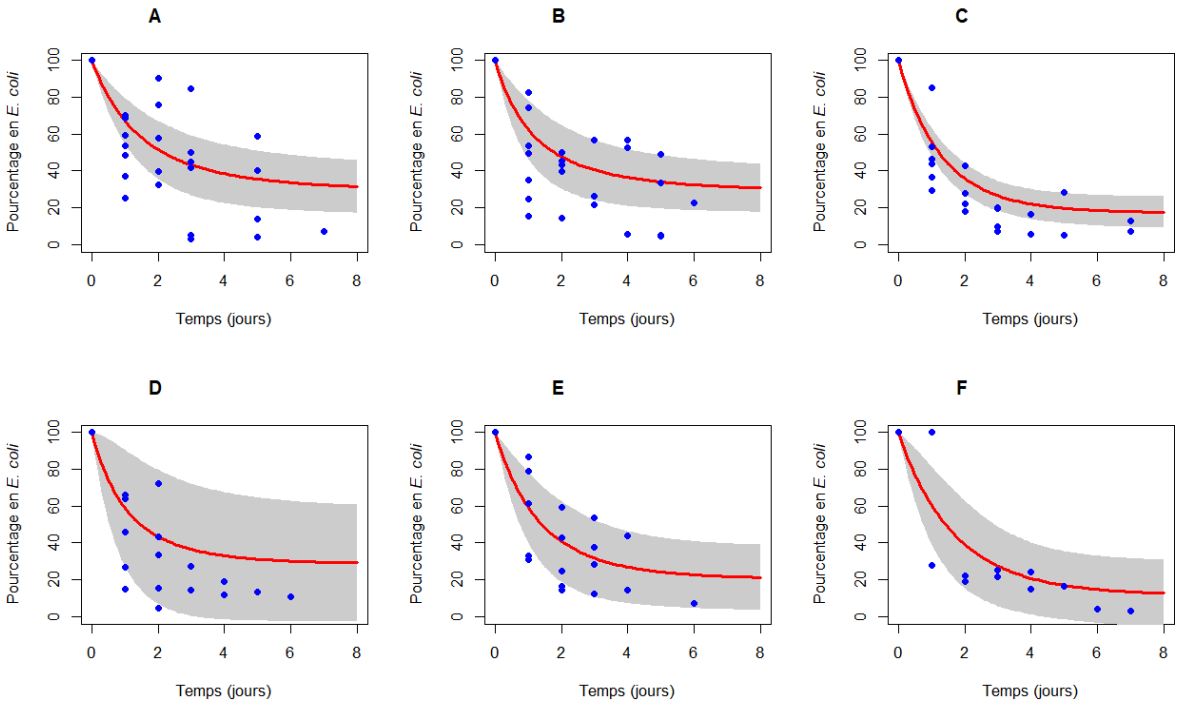

FIGURE 3.4 – Courbe modèle moyenne (ligne rouge) ajustée aux concentrations relatives en *E. coli* pour chaque site. (A) site SMV1, (B) site SMV10, (C) site SMV14, (D) pont de l'Alma, (E) pont de Tolbiac RD et (F) pont de Tolbiac RG. Les valeurs mesurées relatives exprimées en pourcentage de la concentration initiale sont représentées en bleu, les intervalles de confiance à 95% des courbes sont représentés en gris.

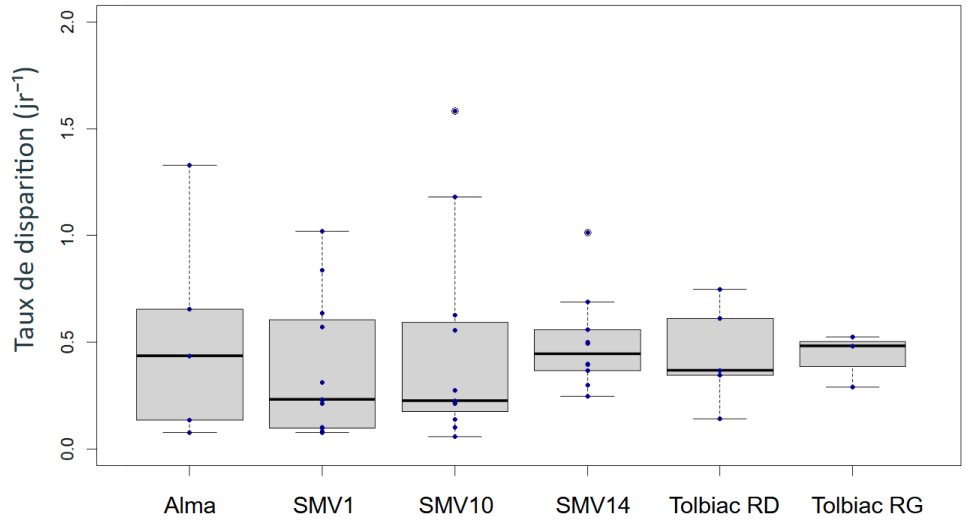

FIGURE 3.5 – Comparaison des taux de disparition (jr^{-1}) entre les 6 sites.

Pour l'ensemble des événements, le taux de disparition ne variait pas significativement d'un site à l'autre (Test de Kruskal-Wallis, n= 46, p=0,69, Figure 3.5). En effet, le taux était très proche entre les différents sites, il était le plus faible à SMV1 (0,36 ± 0,32 jr^{-1}) et le plus élevé au pont de l'Alma (0,52 ± 0,50 jr^{-1}). Cela indiquerait une vitesse de disparition relativement

similaire au niveau des 6 sites analysés que ce soit en Seine (0,47 ± 0,32 jr^{-1}) ou en Marne (0,44 ± 0,35 jr^{-1}).

3.3.2.2.2. Taux de disparition estimés avec les données du ColiMinder

La constante de cinétique (K_1) a été calculée pour chaque événement pluvial sélectionné en utilisant les données mesurées toutes les 2 h. Les modèles étaient validés avec une p-valeur <0,05, sauf pour 2/13 modèles au pont de l'Alma et 4/21 modèles au pont de Tolbiac. Les valeurs K_1 étaient estimées entre 0,26 et 37,72 jr^{-1} (Tableau 3.8). Les taux de disparition estimés étaient en moyenne de 5,50 ± 9,50 jr^{-1}(entre 0,12 et 30,53 jr^{-1}) pour le pont de l'Alma et de 5,39 ± 7,94 jr^{-1} (entre 0,09 et 34,82 jr^{-1}) pour le pont de Tolbiac.

TABLE 3.8 – Constante de cinétique (K_1, jr^{-1}) dérivées du modèle linéaire exponentiel (p-valeur et R^2) et estimation du taux de disparition (K_2, jr^{-1}). Moyenne ± écart type ou [Min : Max]. p-valeur significative (S), non significative (NS) au seuil 0,05.

Site et intervalle	**K_1**	**p-valeur**	**R^2**	K_2
Alma 2 h (n=13)	6,25 ± 9,81	S(11) NS(2) [<0,001 : 0,250]	0,71 ± 0,25	5,50 ± 9,50
Tolbiac 2 h (n=21)	6,05 ± 8,44	S(17) NS(4) [<0,001 : 0,280]	0,79 ± 0,22	5,39 ± 7,94

3.3.2.2.3. Intervalle de temps minimum entre chaque mesure

Afin de savoir quel est l'effet du nombre de mesures sur la capacité du modèle à estimer correctement les taux de disparition, différents intervalles de temps entre deux mesures (2 h, 4 h, 6 h, 8 h, 12 h et 24 h) ont été testés. Il est attendu que le modèle sera moins significatif au fur et à mesure que l'intervalle de temps entre les mesures augmente, mais s'agit-il d'une relation continue ou existe-t-il un seuil ? L'objectif était de déterminer si une mesure réglementaire par jour est suffisante pour suivre et modéliser la décroissance bactérienne après un événement pluvieux. La constante de cinétique (K_1) a été calculée pour chaque pluie sélectionnée et chaque intervalle de temps analysé. Les résultats ont montré que le pourcentage de modèles significatifs diminue fortement de 81% et 85% (pour les ponts de Tolbiac et de l'Alma respectivement) à 2 h d'intervalle, pour atteindre seulement 5 et 15% lorsque l'intervalle était de 24 h (Tableau S1). Ceci indique que plus l'intervalle entre les mesures est réduit, meilleure est l'estimation de la constante de cinétique. Le nombre d'heures entre chaque mesure avait un impact significatif sur les constantes de cinétique K_1 (Test de Friedman, p<0,010, n=204, Tableau S1).

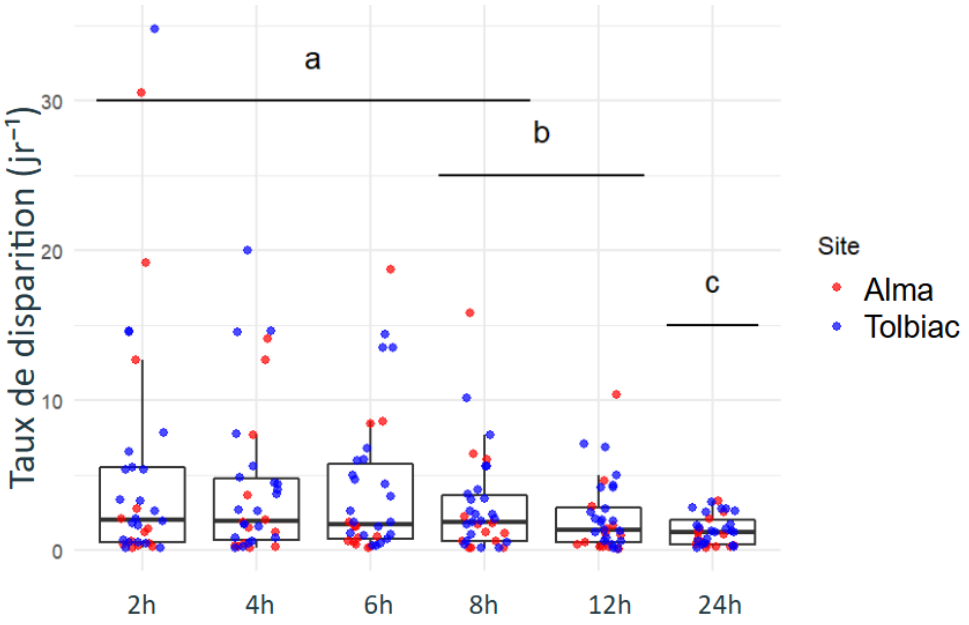

FIGURE 3.6 – Comparaison des taux de disparition (jr^{-1}) selon l'interval de temps entre 2 mesures (de 2 h à 24 h) pour *E. coli* pour les ponts de l'Alma et de Tolbiac. Les barres horizontales et les lettres (a, b, c) regroupent les intervalles sans différence significative.

À la suite des modélisations effectuées pour chaque événement, le taux de disparition a été calculé pour chaque intervalle de temps analysé. L'intervalle de temps impactait significativement le calcul du taux de disparition (Test de Friedman, p<0,010, n=204, Figure 3.6). Toutefois, aucune différence significative n'a été détectée pour les intervalles allant de 2 h à 8 h (Test post-hoc, p=0,182, n=68). Pour l'intervalle de 12 h, les taux de disparition étaient significativement plus faibles que ceux estimés pour les intervalles de 2 h à 6 h (Test post-hoc, p=0,004, n=68). Enfin, pour l'intervalle de 24 h, les taux de disparition étaient significativement plus faibles que ceux estimés avec tous les autres intervalles (Test post-hoc, p<0,001, n=68). Les taux de disparition estimés avec une mesure par jour (24 h) étaient en moyenne de 1,04 ± 1,00 jr^{-1} (entre 0,13 et 3,22 jr^{-1}) pour le pont de l'Alma et de 1,44 ± 0,98 jr^{-1} (entre 0,10 et 3,15 jr^{-1}) pour le pont de Tolbiac (Tableau S1). Ces résultats suggèrent une diminution significative de l'estimation du taux de disparition à partir d'un intervalle de 12 h. Une ou deux mesures par jour sembleraient donc insuffisantes pour estimer correctement le taux de disparition.

De plus, lorsque les données du dispositif Colimindre étaient analysées avec un intervalle de 24 h, le taux de disparition moyen de 1,29 ± 0,99 jr^{-1} estimé en Seine était significativement supérieur à celui estimé avec les mesures réglementaires de 0,47 ± 0,32 jr^{-1} (Test de Wilcoxon, n=47, p<0,01).

3.3.2.3. Impact des conditions environmentales sur les taux de disparition

Différents modèles ont été testés afin de déterminer les paramètres influençant les taux de disparition, que ce soit avec la base de données réglementaires ou les données estimées par le système ColiMinder pour un intervalle de 2 h. Le tableau 3.9 présente les résultats du meilleur modèle sélectionné.

TABLE 3.9 – Relation entre les facteurs environnementaux (site, concentration en*E. coli* initiale au pic de contamination en NPP/100 mL, pluviométrie cumulée de l'événement en mm et taux de disparition d'*E. coli*) en jr^{-1}. Les p-valeurs (p) sont indiquées pour chaque paramètre, les interactions significatives et le modèle global. NA : les paramètres non retenus, lm : modèle linéaire, lmm : modèle linéaire mixte.

Base de données	**ColiMinder**	**Reglementaire**
Effectif	28	46
r^2	0,51	-0,60
p-valeur globale	0,001	0,008
p Intercept	0,001	<0,001
p Pluviométrie	0,009	NA
p Concentration initiale	0,005	0,005
p Site	0,042	NA
p Interaction Pluviométrie-Concentration	0,018	NA
Modèle	lm	lmm

Pour la base de données réglementaires, le modèle linéaire mixte incluant les sites en effet aléatoire (r^2=-0,60, p=0,008, n=46), montre que les taux de disparition sont significativement influencés par la concentration au pic de pollution (p= 0,005). Cela indique que plus la contamination est élevée pendant la pluie, plus la vitesse de disparition des *E. coli* est ensuite lente.

Pour la base de données ColiMinder, le modèle linéaire significatif (r^2= 0,51, p=0,001, n=28) incluait la pluviométrie (discrétisées en pluies $<$ ou $\geq$ à 10 mm), la concentration initiale au pic de pollution et le site (respectivement, p=0,009, p=0,005, et p=0,042). De plus, l'interaction entre la pluviométrie et la concentration initiale en *E. coli* était également significative (p <0,018). Que ce soit avec les données réglementaires ou les données ColiMinder, les résultats soulignent l'importance de tenir compte des caractéristiques spécifiques à chaque site dans l'analyse, de même que la concentration en *E. coli* au pic de pollution.

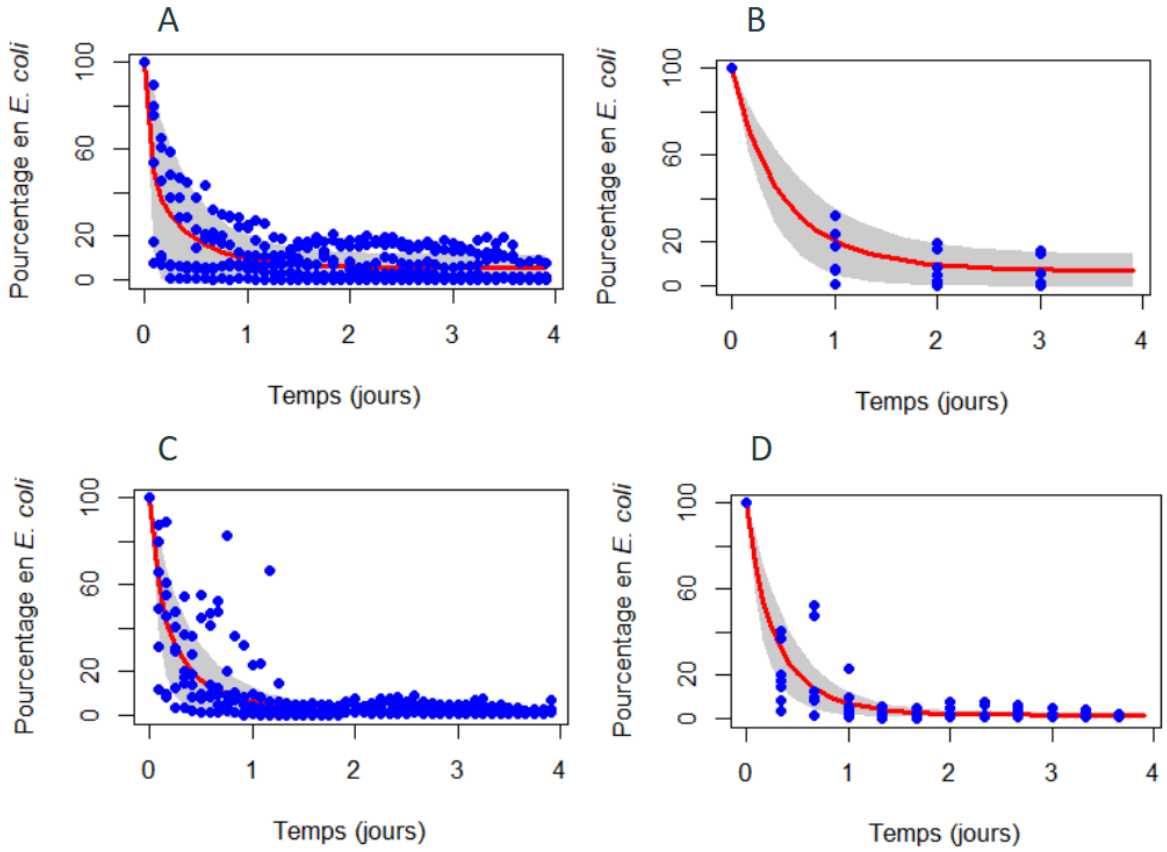

FIGURE 3.7 – Courbes modèles moyennes (ligne rouge) de la variation temporelle des concentrations relatives en *E. coli* pour les pluies ≥10 mm, selon le site. (A) Alma 2 h, (B) Alma 24 h, (C) Tolbiac 2 h et (D) Tolbiac 24 h. Les concentrations en *E. coli* relatives exprimées en pourcentage de la concentration au pic de pollution sont en bleu, les intervalles de confiance à 95% sont représentés en gris.

Afin d'identifier si l'intensité de la pluie influe fortement sur le taux de disparition pour les données obtenues par le système ColiMinder, les pluies ont été subdivisées en deux catégories : faibles (<10 mm cumulés) et les pluies élevées (≥10 mm cumulés). Les courbes modèles, ajustées aux suivis au cours du temps des *E. coli* (concentrations relatives par rapport au pic de pollution, exprimées en %) après une pluie, ont été tracées pour chaque épisode pluvial sélectionné (Figure 3.7 et 3.8).

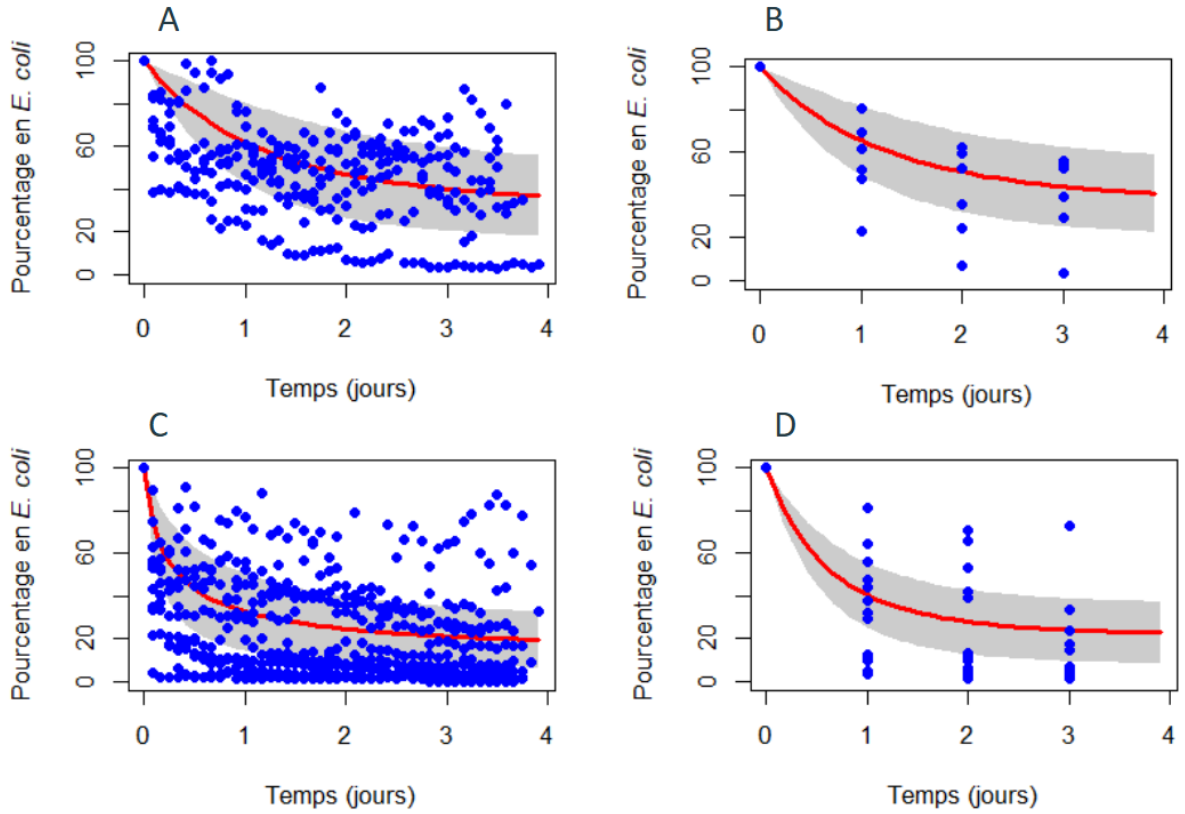

FIGURE 3.8 – Courbes modèles moyennes (ligne rouge) de la variation temporelle des concentrations relatives en *E. coli* pour les pluies <10 mm pour chaque site.(A) Alma 2 h, (B) Alma 24 h, (C) Tolbiac 2 h et (D) Tolbiac 24 h.Les concentrations en *E. coli* relatives exprimées en pourcentage de la concentration au pic de pollution sont en bleu, les intervalles de confiance à 95% sont représentés en gris.

3.3.3. Résilience et résistance

Une analyse de la résilience et de la résistance des sites a été effectuée en estimant 3 paramètres : le T_{90}, et les amplitudes de variation de la pollution et de la récupération après le pic de pollution.

3.3.3.1. Estimation de la résilience et de la résistance avec les mesures réglementaires

Aucune différence significative n'a été observée entre les sites de la Marne et de la Seine au niveau du T_{90} et des amplitudes de variation de la pollution et de la récupération (Anova, p=0,660 pour le T_{90}, p=0,465 pour l'amplitude de pollution, p=0,355 pour l'amplitude de récupération, n=46). La résilience était donc similaire entre les deux rivières, avec un T_{90} de 3,50 ± 1,43 jr en Seine et de 3,68 ± 1,18 jr en Marne, avec amplitude moyenne de récupération respectivement de 77 ± 19% et 73 ± 18% (Anova, n=46, respectivement p=0,102, p=0,444, Tableau 3.10). La résistance était également similaire avec une amplitude moyenne de pollution de 75 ± 18% en Seine et de 83 ± 12% en Marne (Anova, n=46, p=0,662, Tableau 3.10).

TABLE 3.10 – Valeurs moyennes du temps de retour (T_{90}, jr) et des amplitudes de variation de la pollution (AV_{avant}, %) et de récupération ($AV_{aprés}$, %) lors des campagnes réglementaires en Seine et en Marne.

Site	**T_{90} (jr)**	**AV_{avant} (%)**	**$AV_{aprés}$ (%)**
SMV1	4,11 ± 1,54	75 ± 20	68 ± 22
SMV10	3,73 ± 1,68	70 ± 21	69 ± 19
SMV14	3,12 ± 0,78	81 ± 12	82 ± 11
Alma	3,32 ± 1,45	79 ± 16	70 ± 25
Tolbiac RD	3,46 ± 1,22	82 ± 12	78 ± 14
Tolbiac RG	3,87 ± 0,94	90 ± 4,18	87 ± 7

3.3.3.2. Intervalle de temps minimum entre chaque mesure

Un effet significatif de l'intervalle de temps entre deux mesures par le système ColiMinder (2 h, 4 h, 6 h, 8 h, 12 h et 24 h) a été observé, que ce soit pour le T_{90}, l'amplitude de variation de la pollution ou l'amplitude de récupération après la pluie (Kruskal-Wallis, p<0,001, p<0,001, p<0,001, n=204) (Tableau 3.11). La différence était significative entre l'intervalle de 24 h et les intervalles de 2 h à 8 h pour le T_{90} (Test post-hoc, p<0,011, n=68, Tableau 3.11). Cela indique qu'une mesure par jour est insuffisante pour bien évaluer la résilience et la résistance des sites. En effet, le T_{90} moyen variait de 1,58 ± 1,29 jr à 1,18 ± 1,04 jr (respectivement pour les ponts de l'Alma et de Tolbiac) lorsque l'intervalle était de 2 h. Il augmentait à 1,96 ± 0,79 jr et 1,56 ± 0,62 jr (respectivement pour les ponts de l'Alma et de Tolbiac) lorsque l'intervalle est de 24 h (Tableau 3.11). De plus, le temps de retour pour un intervalle de 24 h toutes stations confondues (1,71 ± 0,71 jr), était significativement plus faible lorsqu'il était estimé avec les données du système ColiMinder comparé aux estimations avec les mesures réglementaires (3,50 ± 1,43 jr) (Test de Student, n=47, p<0,01).

Pour l'amplitude de variation de la pollution, une différence significative a été observée entre l'intervalle de 2 h avec les intervalles de 8 h à 24 h et aussi entre l'intervalle de 4 h et les intervalles de 12 h et de 24 h (Test de Kruskall-Wallis suivi de tests post-hoc, p<0,027, n=68, Tableau 3.11). Avec un intervalle de 2 h, le pourcentage moyen de variation entre le temps sec avant la pluie et le pic de contamination était 75 ± 23% et 86 ± 19% (respectivement pour les ponts de l'Alma et de Tolbiac) et avec un intervalle de 24 h il présentait des valeurs moyennes de 70 ± 28% et 84 ± 20% (respectivement pour les ponts de l'Alma et de Tolbiac) (Tableau 3.11).

L'intervalle de 24 h sous-estimait significativement l'amplitude de récupération après la pluie, (test de Kruskall-Wallis suivi de tests post-hoc, p<0,045, n=68). Le pourcentage de récupération était en moyenne de 77 ± 22% (Pont de l'Alma) et de 86 ± 20% (pont de Tolbiac)

avec l'intervalle de 2 h, alors qu'il était de 74 ± 22% (pont de l'Alma) et 84 ± 22% (pont de Tolbiac) avec un intervalle de 24 h (Tableau 3.11).

TABLE 3.11 – Valeurs moyennes du temps de retour (T_{90}) et des amplitudes de variation de la pollution (AV_{avant}) et de récupération ($AV_{aprés}$) en fonction de l'intervalle de temps entre chaque mesure lors des campagnes ColiMinder en Seine.

Intervalle	**T_{90} (jr)**		**AV_{avant} (%)**		**$AV_{aprés}$ (%)**	
	Alma	**Tolbiac**	**Alma**	**Tolbiac**	**Alma**	**Tolbiac**
2 h	1,58 ± 1,29	1,18 ± 1,04	75 ± 23	86 ± 19	77 ± 22	86 ± 20
4 h	1,52 ± 1,18	1,07 ± 0,93	74 ± 24	86 ± 19	76 ± 24	87 ± 18
6 h	1,39 ± 0,85	0,92 ± 0,66	73 ± 26	85 ± 20	77 ± 21	87 ± 18
8 h	1,52 ± 1,08	1,17 ± 0,82	72 ± 26	84 ± 22	76 ± 23	86 ± 19
12 h	1,80 ± 0,97	1,28 ± 0,81	73 ± 26	83 ± 22	75 ± 24	85 ± 22
24 h	1,96 ± 0,79	1,56 ± 0,62	70 ± 28	84 ± 20	74 ± 22	84 ± 22

Nous allons nous focaliser par la suite sur l'intervalle de 2 h pour les données obtenues avec le système ColiMinder car il permettait une estimation plus précise des métriques. Globalement, ces résultats indiquent une résilience en moyenne plus faible au pont de Tolbiac ainsi qu'une amplitude de récupération légèrement plus élevée, avec un temps de retour moyen plus lent. Cependant, aucune différence significative n'a été observée entre les deux sites pour les 3 métriques analysées (Test de Wilcoxon, p=0,320 pour le T_{90}, p=0,256 pour l'amplitude de pollution, p=0,246 pour l'amplitude de récupération, n=34, Tableau 3.11).

3.3.3.3. Impact des conditions environnementales sur la résilience et la résistance des sites en Seine et en Marne

Différents modèles linéaires ont été testés, afin de déterminer les paramètres environnementaux influençant les 3 métriques de résilience et de résistance que ce soit avec la base de données réglementaires ou les données du système ColiMinder avec une acquisition toutes les 2 h. Le tableau 3.12 présente les résultats du meilleur modèle sélectionné pour chaque base de données.

Pour la base de données réglementaires, les modèles expliquent faiblement la variation des métriques, avec des r^2 compris entre 0,10 et 0,23. Le modèle significatif expliquant les variations du temps de retour (T_{90}) (n=46, p=0,029) comportait la pluviométrie (catégories < et $\geq$ à 10 mm) (p=0,030). Pour l'amplitude de variation de la pollution, le modèle était significatif (n=46, p<0,001). L'effet du débit était une tendance (p=0,058), avec une interaction significative entre le débit et la concentration initiale au pic de pollution (p<0,001). Pour l'amplitude de récupération après la pluie, le modèle le plus significatif (n=46, p=<0,001) comportait la concentration initiale

TABLE 3.12 – Relation entre les facteurs environnementaux (débit en m^3/s), site, concentration en *E. coli* initiale au pic de contamination en NPP/100 mL, pluviométrie et taux de disparition d'*E. coli* en jr^{-1}). Les p-valeurs sont indiquées pour chaque paramètre, les interactions significatives et le modèle global. NA : les paramètres non retenus, lm : modèle linéaire, glm : modèle linéaire généralisé, ^{a} : pluviométrie en catégories, ^{b} : pluviométrie cumulée de l'événement en mm.

Base de données	**Réglementaire**			**ColiMinder**		
	T_{90}	AV_{avant}	$AV_{aprés}$	T_{90}	AV_{avant}	AV_{apres}
Effectif	46	46	46	34	34	34
r²	0,10	0,23	0,17	0,56	0,42	0,45
p globale	0,029	<0,001	<0,001	0,002	<0,001	<0,001
p intercept	<0,001	0,072	0,322	0,108	0,274	0,158
p débit	NA	0,058	NA	0,894	NA	NA
p pluviométrie	0,030^a	NA	0,146^b	0,061^a	0,005^a	0,006^b
p concentration initiale	NA	<0,001	0,038	0,001	<0,001	<0,001
Site	NA	NA	NA	0,121	NA	NA
p Interaction débit-site	NA	NA	NA	0,360	NA	NA
p Interaction débit-concentration	NA	0,032	NA	0,002	NA	NA
p Interaction concentration-site	NA	NA	NA	0,046	NA	NA
p Interaction pluviométrie-concentration	NA	NA	NA	NA	0,001	0,001
Modèle	lm	glm	glm	glm	glm	glm

au pic de pollution (p= 0,038), contrôlée par la pluviométrie en cumul (non significatif, p=0,140).

Pour les données ColiMinder mesurées toutes les 2 h, les modèles montraient un meilleur ajustement pour les 3 métriques (r² variant de 0,42 à 0,56) par comparaison aux données réglementaires. Pour les 3 paramètres analysés, les modèles globaux étaient significatifs (n=34, p=0,002 pour le T_{90}, p<0,001 pour l'amplitude de pollution, p<0,001 pour l'amplitude de récupération). Le T_{90} était significativement influencé par la concentration initiale au pic de pollution (p=0,001) et tendait à être influencé par la pluviométrie (p=0,061), les interactions entre le débit et la concentration initiale au pic de pollution (p=0,002) et entre la concentration initiale et le site (p=0,046) étant significatives. Concernant les amplitudes de récupération, le modèle final montrait des contributions significatives de la concentration initiale (p<0,001) et de la pluviométrie (p=0,006), avec une interaction significative entre ces deux paramètres (p=0,001).

Ces résultats soulignent l'importance de la pluviométrie (parfois en cumul et parfois en catégories) et de la concentration initiale comme facteurs clés dans la dynamique de résilience et de résistance d'*E. coli*, renforcée par les interactions avec d'autres variables. Ces résultats

sont confirmés avec les deux bases de données.

3.4. Discussion

Ces dernières décennies, des efforts importants ont été déployés pour améliorer la qualité des cours d'eau (Kistemann et al., 2016) et l'Île-de-France n'y fait pas exception avec la construction d'infrastructures (bassins de rétention, stations de dépollution des eaux pluviales), l'amélioration des réseaux, la résolution de mauvais branchements, le raccordement des bateaux et la désimperméabilisation (Bouleau et al., 2024). Ces efforts ont bénéficié de la volonté politique d'organiser à Paris les Jeux Olympiques et Paralympiques (JOP) en 2024 et de l'ouverture envisagée de plusieurs baignades urbaines sur les bords de la Seine et de la Marne à l'horizon 2025 en héritage des JOP (Noury et al., 2018). Selon la directive baignade 2006/7/CE, le classement d'un site de baignade nécessite un suivi de 4 ans de la qualité microbiologique et d'établir un profil de baignade. Notre étude vient en complément en déterminant la dynamique de décroissance d'*E. coli* en Seine et en Marne lors d'événements polluants en vue d'aider à mieux comprendre la dynamique temporelle des pollutions microbiologiques sur des sites emblématiques (JOP, futures baignades) et de fournir des paramètres pour la modélisation des pollutions. Il existe encore peu d'études sur la décroissance d'*E. coli* en Île-de-France dans la Seine et la Marne (Passerat et al., 2011; Menon et al., 2003; Servais et al., 1999; Van et al., 2022). Notre étude élargit cette analyse en incluant des données provenant de différents sites en Seine et en Marne et en comparant la dynamique mesurée à l'aide du dispositif ColiMinder à celle obtenue avec les suivis réglementaires par culture. Par ailleurs, grâce à la fréquence élevée des mesures du système ColiMinder, nous avons pu analyser l'effet du nombre de mesures sur l'estimation des taux de disparition.

3.4.1. Taux de disparition et temps de retour

La disparition d'*E. coli* dans les milieux aquatiques résulte de l'action combinée de divers paramètres environnementaux liés d'une part à l'hydrologie et l'hydromorphologie de la rivière (dilution et diffusion des effluents, sédimentation et resuspension des bactéries dans le sédiment) et d'autre part liés à la capacité de survivre et de croître des bactéries dans cet habitat secondaire (prédation, compétition, stress physiologique, épuisement ou disponibilité de nutriments et des sources de carbone, rayonnement solaire, température) (Barcina et al., 1997;

Dick et al., 2010; Carneiro et al., 2018; Korajkic et al., 2014; Mattioli et al., 2017; Brooks and Field, 2016). Les bactéries d'origine fécale dont l'habitat primaire est le tube digestif des organismes homéothermes, une fois rejetées dans la rivière vont y subir des conditions environnementales favorables ou défavorables qui peuvent entraîner une mortalité ou une perte de capacité à croître, toutefois certaines populations peuvent éventuellement s'acclimater et survivre voire croître dans les sédiments, les végétaux et les biofilms (Liu et al., 2006; Passerat et al., 2011; Gonzales-Siles and Sjöling, 2016). Nos résultats quant aux taux de mortalité et de disparition sont dans la gamme des taux rapportés dans la littérature. Ainsi un taux d'inactivation de 0,672 ± 0,11 jr^{-1} a été mesuré pour *E. coli* avec une eau de rivière subissant des rejets d'eaux usées (Blaustein et al., 2013). Une autre étude a identifié un taux de mortalité en laboratoire de 0,72 jr^{-1} (Servais et al., 2007b). Ces taux sont proches du taux de mortalité de 0,97 ± 0,48 jr^{-1} estimé par les expériences dans les sacs à dialyse disposés au site SMV14 (Marne). La comparaison avec la littérature peut toutefois être un exercice difficile car les conditions expérimentales diffèrent. Les taux de disparition mesurés sur le même site SMV14 avec les données de suivi réglementaire étaient relativement plus rapides (0,49 ± 0,22 jr^{-1}). Ceci est cohérent car d'autres facteurs que la mortalité entrent en jeu dans la disparition, incluant la dilution du rejet, la dispersion, l'advection et la sédimentation des bactéries (Jalliffier-Verne et al., 2017). En effet, quand une certaine quantité d'un contaminant est rejetée dans une rivière, il est transporté en aval par le mouvement de l'eau et continuellement mélangé et redistribué dans l'eau. Ce processus dépend des caractéristiques hydrologiques et hydrodynamiques du bief de la rivière, et donc de la géométrie et de la morphométrie de la rivière qui vont déterminer la vitesse et la turbulence du courant (Rowiński et al., 2022). Ainsi, notre analyse statistique a montré que la rivière et la pluviométrie impactaient significativement les taux de disparition. Passerat et al. (2011) ont constaté qu'une estimation prenant en compte la "dilution + mortalité + sédimentation" permettait de mieux modéliser les concentrations en *E. coli* dans les eaux de la Seine. Cette approche suggère que la sédimentation joue un rôle notable dans le sort des BIF attachées dans les eaux affectées par le déversement d'eaux pluviales. Nos résultats suggèrent également un taux de disparition et un temps de retour similaires en Seine (respectivement de 0,47 ± 0,32 jr^{-1} et de 3,50 ± 1,43 jr) et en Marne (respectivement de 0,44 ± 0,35 jr^{-1} et de 3,68 ± 1,18 jr) avec l'analyse des données réglementaires. Une étude menée sur les coliformes fécaux a estimé un taux de disparition moyen de 0,428 jr^{-1} dans les eaux de rivière (Chigbu et al., 2005). Il convient cependant de noter que ces deux paramètres ne sont pas directement

comparables, étant donné que *E. coli* est une bactérie qui fait partie des coliformes fécaux. Il a également été montré que des eaux de mêmes catégories présentaient une inactivation d'*E. coli* similaire (Blaustein et al., 2013). L'ensemble de ces résultats n'exclut donc pas la possibilité de généraliser la valeur moyenne du taux de disparition obtenue à l'ensemble des sites de baignades potentielles avec des profils de baignade similaires permettant de fournir un paramètre utile pour les modèles hydrodynamiques.

3.4.2. Comparaison des bases de données

On ce qui concerne le dispositif ColiMinder, nous avons pu constater qu'à partir d'un intervalle de 24 heures (soit une mesure par jour) qu'il y avait un biais significatif dans l'estimation du taux de disparition, du temps de retour et des amplitudes de variation de pollution et de récupération. En effet, le temps de retour minimal était de 1,92 h pour un intervalle de 2 h, celui-ci augmente à 18 h avec un intervalle de 24 h. Cela indique que pour les sites en Seine, il faudrait au minimum deux mesures par jour pour pouvoir bien suivre la disparition d'*E. coli* dans la rivière afin de pouvoir bien évaluer la résilience d'un site. Une étude antérieure sur la Seine a mesuré une décroissance des BIF par mortalité et sédimentation de 66% après un rejet urbain de temps de pluie intense (39 mm) qui a entraîné le déversement du déversoir d'orage de Clichy (Passerat et al., 2011). Nos propres résultats indiquent une forte capacité de résilience avec un retour rapide voire très rapide à des concentrations de temps sec pour certains des événements analysés. Nos résultats montrent bien que pour 91% des événements sélectionnés, les pollutions étaient de court terme, c'est à dire affectant la qualité de l'eau moins de 3 jours selon la définition de l'agence française de sécurité sanitaire de l'environnement (Duboudin et al., 2007).

Les résultats obtenus avec les deux bases de données n'étaient pas totalement en accord, mais ceci peut s'expliquer d'une part par le fait que la régularité et la quantité des données différaient entre les données du système ColiMinder toutes les 24 h, et les données réglementaires pas systématiquement mesurées tous les jours. De plus, le dispositif ColiMinder mesure l'activité de la β-D-glucuronidase, donnant une estimation indirecte mais rapide des concentrations en *E. coli*. Cette méthode vise toutes les bactéries cibles, qu'elles soient cultivables, non cultivables ou mortes, ainsi que les enzymes libres (Cazals et al., 2020; Garcia-Armisen and Servais, 2009). Par contre, la méthode ISO 9308-3 de culture en microplaques quantifie les bactéries cultivables et viables qui sont thermotolérantes et possèdent la β-D-glucuronidase, avec une estimation statistique basée sur la loi de Poisson (Cazals et al., 2020; Garcia-Armisen and Servais, 2009).

Ces différences méthodologiques peuvent conduire à des écarts dans les résultats. Les méthodes de culture sous-estiment souvent le nombre de bactéries dans des environnements fortement contaminés en raison de la diminution de la quantité d'enzyme par bactérie cultivable. De ce fait, dans ce type d'échantillon, il existe une fraction plus importante des BIF à l'état viable mais non cultivable (Carneiro et al., 2018). En effet, il a été constaté une diminution du ratio enzyme/*E. coli* lorsque la contamination devenait plus importante (Garcia-Armisen and Servais, 2009; Cazals et al., 2020).

Avec suffisamment de données (3 à 4 points de mesure par événement pluvial), nos modèles exponentiels étaient significatifs, donc l'estimation des taux de disparition et des taux de mortalité faite peut être considérée comme valide dans ce cas. Toutefois, l'utilisation des données acquises en temps quasi réel montre que l'évaluation des taux de disparition en utilisant les données réglementaires peut être biaisée. En effet, il est nécessaire d'avoir a minima 2 à 3 mesures par jour durant un événement pluvial et les jours qui suivent pour estimer les taux de disparition de manière précise. Un plan d'échantillonnage de 1 prélèvement par jour entraîne donc un biais, avec des valeurs de taux de disparition jusqu'à 3 fois plus faibles comparées à une mesure toutes les 2 à 8 h. Le dispositif ColiMinder est donc particulièrement adapté pour le suivi en quasi temps réel, et permet une observation fine des dynamiques rapides comme celles observées en temps de pluie lors des orages d'été.

3.4.3. Pics de pollution

Les valeurs seuil européennes pour le classement des baignades et les valeurs guides pour la gestion quotidienne, sont souvent dépassées dans les rivières fortement urbanisées comme la Marne et la Seine, en particulier lors des fortes précipitations (Kistemann et al., 2016; Bouleau et al., 2024). Durant ces événements pluvieux, les rejets de microorganismes provenant des sources de pollution fécale ponctuelles ou diffuses augmentent considérablement (Ahmed et al., 2018). En effet, les précipitations entraînent le transfert de la contamination fécale du sol aux cours d'eau (Jardé et al., 2018). De plus, par temps de pluie, les réseaux d'assainissement peuvent déborder et apporter des eaux usées non-traitées (Passerat et al., 2011), entrainant une augmentation des concentrations en BIF dans les eaux de surface pouvant dépasser jusqu'à 100 à 1000 fois les concentrations de temps sec (Salmore et al., 2006; Krometis et al., 2007). Ainsi, les mesures bactériologiques réalisées en Seine et en Marne au niveau des 6 sites ont montré qu'après une pluie, une augmentation de la concentration en *E. coli* était constatée

par rapport au temps sec pouvant aller jusqu'à 22 fois en Seine et 32 fois en Marne et même plus de 100 fois avec le dispositif ColiMinder. Ainsi le rejet d'eau usée non traitée durant les temps de pluie (déversoirs d'orage, by-pass des stations d'épuration, mauvais branchements dans les réseaux séparatifs) contribue fortement à la dégradation de la qualité des eaux de surface (Cyterski et al., 2022). La remise en suspension des sédiments dans les rejets pluviaux et les eaux de ruissellement lors d'événements pluviaux peut également contribuer à l'augmentation des concentrations en BIF au niveau des eaux de surface urbaines (Lee et al., 2006; Wu et al., 2009). Cette augmentation rapide était suivie d'une diminution dans les 2 à 3 jours après la pluie. Au niveau de notre étude, au bout de 3 jours, une diminution de la concentration de 34 à 95% en Seine et de 33 à 96% en Marne a été constatée avec les mesures réglementaires, et pour les mesures avec le dispositif ColiMinder de 29 à 99% (mesures toutes les 2 h) et de 27 à 99% (mesures toutes les 24 h). L'impact de la pluie sur la qualité microbiologique de la rivière peut être variable, ce qui se traduit par une large variabilité parmi les pics de concentration mesurés. En effet, nous avons observé des amplitudes de variation de la pollution allant de 24 à 97% des valeurs de temps sec précédant la pluie pour différents événements sélectionnés dans la base de données réglementaires. Les données du dispositif ColiMinder donnaient des amplitudes de pollution de 31 à 100% pour les mesures toutes les 2 h et de 26 à 100% pour les mesures toutes les 24 h. Cette variabilité de l'amplitude de pollution était partiellement expliquée par la pluviométrie, avec une forte significativité pour les estimations avec les données du système ColiMinder.

3.4.4. Temps et niveau de récupération après la pollution

La résilience de la qualité de l'eau d'une rivière est liée à sa capacité à absorber l'apport en polluants (perturbation) et à rapidement restaurer ou améliorer la qualité de l'eau au cours du temps (Park et al., 2025). Il existe encore peu d'études qui s'intéressent à la résilience de la qualité de l'eau dans les rivières, et celles-ci se focalisent la plupart du temps sur des polluants chimiques, les pollutions microbiologiques étant rarement prises en compte (Hoque et al., 2012; Li et al., 2016; Mirauda et al., 2021; Park et al., 2025). Nous avons abordé la résistance et la résilience des sites potentiels de baignade ou d'organisation d'événements sportifs dans les rivières franciliennes sous l'angle de la réponse "écologique" des systèmes aux perturbations (Mirauda et al., 2021), à l'aide de 4 métriques : le taux de disparition qui mesure une facette de la résilience du système face à la pollution, l'amplitude de pollution qui mesure la robustesse du

système, le temps de retour qui mesure la rapidité de récupération, et l'amplitude de récupération qui mesure la capacité à revenir à l'équilibre antérieur. Dans la littérature, il est rare que différents aspects de la résilience soient mesurés concernant la qualité de l'eau (Park et al., 2025). Nous avons aussi testé l'impact de plusieurs facteurs hydrométéorologiques et physico-chimiques sur les métriques de la résilience. Ainsi, les temps de retour et l'amplitude de récupération étaient expliqués partiellement par les hauteurs de pluie (en particulier avec les données du système ColiMinder), tout comme les taux de disparition. En effet, les rivières sont des systèmes dynamiques dont la qualité dépend de relations complexes entre les caractéristiques du bassin versant et la variabilité du climat. De plus, le niveau de pollution atteint (concentration initiale au pic de pluie) avait aussi un impact sur le temps de retour T_{90} et sur l'amplitude de récupération (en particulier pour les métriques estimées avec les données du système ColiMinder). Il est donc clair que le niveau de dégradation de la qualité conditionnait la capacité de retour au niveau de base avant la pollution. Des interactions avec le débit ou la concentration au pic de pollution étaient significatives. Ces interactions doivent être prises en compte pour mieux comprendre et prédire la dynamique des contaminants dans les systèmes aquatiques impactés par les rejets d'effluents (Carneiro et al., 2018). Il serait également intéressant d'inclure des caractéristiques du bassin versant telles que l'usage des sols, la densité de population, le taux d'imperméabilisation, le nombre de rejets et leurs volumes déversés, car ces variables influencent les apports en pollution fécale (Paule-Mercado et al., 2016).

Pour les données réglementaires, le temps de retour T_{90} était en moyenne de 87 ± 32 h tous sites confondus, et pour les données issues du système ColiMinder l'estimation du T_{90} était en moyenne de 31 ± 22 h en Seine pour les mesures toutes les 2 h. Ces temps de retour étaient situés le plus souvent dans la limite de 72 h ce qui est spécifié par la directive 2006/7/EC pour la gestion des pollutions temporaires. Toutefois, il est noté que certaines pollutions (notamment sur les stations SMV1 et SMV10), pouvaient durer plus de 72 h. Il est donc recommandé de vérifier le niveau des *E. coli* avant la réouverture. Les mesures rapides basées sur la PCR quantitative, ou sur les mesures enzymatiques peuvent compléter les mesures réglementaires effectuées sur des échantillons collectés après la pluie. De plus, la modélisation et les systèmes de suivi en temps quasi réel ou réel (comme le ColiMinder, ou les capteurs de fluorescence 3D) peuvent alors aider à avoir une gestion réactive en cas de pluie ou d'incident sur le réseau (Burnet et al., 2019; Offenbaume et al., 2020; Angelotti de Ponte Rodrigues et al., 2024).

Cette approche de la résilience permet de mieux prendre en compte les contaminations

microbiennes dans une rivière face à des perturbations aléatoires comme les rejets de temps de pluie, en focalisant sur la dynamique et la variabilité des changements de concentration en BIF. Cette approche sur l'adaptabilité du site de baignade offre un cadre conceptuel pour le gestionnaire qui peut ainsi prendre en compte la vulnérabilité du site de baignade face aux événements polluants. Ces résultats pourront également alimenter les modèles déterministes qui sont développés en Marne et en Seine pour prédire la contamination des eaux de surface, comme le modèle ProSe (Poulin et al., 2013) ou le modèle Telemac (Van et al., 2022) pour la gestion active des futurs sites de baignade.

3.5. Conclusion

L'ouverture de sites de baignade en Marne et en Seine nécessite une mise en place et une gestion des futurs sites. Une fois l'ouverture des sites, il faudra une gestion active de la pollution par un suivi *in situ* automatisé ou semi-automatisé, couplé à des modèles prédictifs. L'analyse des jeux de données de la Ville de Paris et du Syndicat Marne Vive (suivi réglementaire et système ColiMinder) a permis de développer une évaluation de la résistance et de la résilience de plusieurs futurs sites de baignade ou de sites ayant servi pour les JOP 2024. Cette approche dynamique a démontré la robustesse et l'adaptabilité de ces sites face aux événements polluants temporaires liés au temps de pluie. Les résultats des campagnes de mesure ont montré que le site SMV14 était très réactif avec un temps de retour relativement court, permettant une réouverture au bout de 75 h en moyenne (entre 56 et 94 h selon l'événement pluvial), ce qui est quasi conforme avec la directive 2006/7/CE (fermeture des baignades 72 h après une pollution ponctuelle) suivi par le pont de l'Alma avec un temps de retour moyen de 80 h (entre 36 et 117 h selon l'événement pluvial). Par contre, pour SMV10, des apports importants en amont (rejet de l'usine de traitement des eaux usées Marne Aval qui n'était pas encore équipée de la désinfection, nombreux rejets pluviaux polluants en amont) semblent contribuer à dégrader la qualité de ce site en temps sec comme en temps de pluie, ce qui explique les amplitudes moyennes de variation de la pollution légèrement plus faibles (Petrucci and Vaury, 2018). En Seine, le temps de retour calculé avec les données du système Colimінder est beaucoup plus faible (2 à 80 h). Cet équipement permet un suivi toutes les 2 h et donc avec une détermination plus précise du temps nécessaire pour un retour à une qualité de temps sec. Le suivi en temps réel permettrait au gestionnaire d'adapter ses fermetures et réouvertures à chaque pluie et ainsi

de maximiser les ouvertures sur la saison. En effet, une étude a montré que pour les plages du lac Michigan, 12% du temps, les fermetures des plages n'étaient pas nécessaires, ce qui pouvait potentiellement représenter une perte de 1274 à 37030 dollars par jour (Rabinovici et al., 2004). Avoir un suivi avec des résultats dans la journée permettrait d'éviter ce problème.

Cette étude a permis d'apporter une analyse de la mortalité et de la disparition d'*E. coli* dans les rivières franciliennes. D'autres analyses complémentaires pourraient enrichir ces résultats. En effet, l'analyse de la dynamique de différents indicateurs comme les entérocoques intestinaux et des indicateurs de sources de contamination (humaine et animales) pourrait représenter une approche complémentaire intéressante afin de comparer la résistance et la résilience des sites avec différents marqueurs. En effet, une diminution plus élevée des entérocoques intestinaux dans la Seine après un rejet du déversoir d'orage de Clichy dans la Seine (région parisienne) a été constatée dans une étude antérieure (Passerat et al., 2011). Une étude qui a comparé le taux de décroissance d'*E. coli* à celui du marqueur humain HF183 a montré que le temps de retour du marqueur humain était plus élevé qu'*E. coli*, sans que cette différence soit significative (Dick et al., 2010). Il serait aussi intéressant de calculer la résilience des sites pour quelques pathogènes ou marqueurs humains viraux, car ils n'ont sûrement pas la même dynamique temporelle en Seine que les BIF. Un indicateur de résilience multimétrique pourrait être proposé, intégrant plusieurs microorganismes à l'instar des index de résilience développés pour la pollution chimique.

Remerciements : Nous remercions les Conseils Départementaux du Val-de-Marne et de la Seine-Saint-Denis pour leur contribution au jeu de données. Nous remercions aussi Vincent Rocher (SIAAP) de nous avoir autorisées à prélever de l'eau à la sortie de l'usine Marne Aval et Frédéric Van Delanoote (Conseil Départemental de la Seine-Saint-Denis) de nous avoir fourni des échantillons d'une reprise de temps sec. Nous remercions Olivier Monfort (Voies Navigables de France) de nous avoir permis de faire les expérimentations *in situ* en Marne, Sandrine Michot et Christin Lenief de nous avoir autorisées à circuler sur la voirie afin de pouvoir transporter notre matériel jusqu'au site d'expérimentation.

3.6. Annexe

TABLE S1 – Valeurs moyennes de la constante de cinétique (K1, jr^{-1}) obtenues par le modèle linéaire exponentiel (p-valeur et R²) et le taux de disparition K_2 en jr^{-1}. Moyenne ± écart type ou [Min : Max]. p-valeur significative (S), non significative (NS) au seuil 0,05.

Station et intervalle	**K_1**	**p-valeur**	**R^2**	**K_2**
Alma 2 h (n=13)	6,25 ± 9,81	S(11) NS(2) [<0,001 : 0,250]	0,71 ± 0,25	5,50 ± 9,50
Alma 4 h (n=13)	4,24 ± 5,28	S(11) NS(2) [<0,001 : 0,370]	0,75 ± 0,26	3,51 ± 4,84
Alma 6 h (n=13)	4,09 ± 5,48	S(9) NS(4) [<0,001 : 0,280]	0,78 ± 0,23	3,39 ± 5,44
Alma 8 h(n=13)	3,48 ± 4,39	S(6) NS(7) [<0,001 : 0,330]	0,64 ± 0,29	2,90 ± 4,39
Alma 12 h (n=13)	2,27 ± 2,82	S(2) NS(11) [<0,001 : 0,500]	0,79 ± 0,19	1,88 ± 2,85
Alma 24 h (n=13)	1,39 ± 0,89	S(2) NS(11) [<0,001 : 0,440]	0,86 ± 0,16	1,04 ± 1,00
Tolbiac 2 h (n=21)	6,05 ± 8,43	S(17) NS(4) [<0,001 : 0,280]	0,79 ± 0,22	5,39 ± 7,94
Tolbiac 4 h (n=21)	5,52 ± 5,70	S(13) NS(8) [<0,001 : 0,550]	0,75 ± 0,23	4,61 ± 5,40
Tolbiac 6 h (n=21)	4,88 ± 4,39	S(12) NS(9) [<0,001 : 0,740]	0,69 ± 0,30	4,30 ± 4,47
Tolbiac 8 h(n=21)	3,35 ± 2,58	S(10) NS(11) [<0,001 : 0,790]	0,82 ± 0,15	2,89 ± 2,57
Tolbiac 12 h (n=21)	2,71 ± 2,06	S(0) NS(21) [0,060 : 0,770]	0,73 ± 0,29	2,42 ± 2,08
Tolbiac 24 h (n=21)	1,76 ± 0,82	S(1) NS(20) [0,020 : 0,710]	0,79 ± 0,24	1,44 ± 0,98

4. Conclusion

La gestion de la qualité des eaux de surface dans des régions fortement urbanisées, comme l'Île-de-France, pose des défis complexes liés à la variabilité temporelle et spatiale de la contamination microbiologique et aux incertitudes associées aux méthodes de mesure. La directive 2006/7/CE a établi des normes spécifiques pour les eaux de baignade, mais leur application nécessite de mieux comprendre la dynamique des bactéries indicatrices fécales, comme *E. coli*, et les facteurs environnementaux influençant leur décroissance. Une approche plus précise, tenant compte de cette variabilité, permettrait non seulement de renforcer la robustesse des décisions de gestion (ouverture/fermeture des sites de baignade) mais aussi de soutenir les efforts de reconquête des rivières pour des usages récréatifs.

Notre étude offre une analyse intégrée de la dynamique spatiale et temporelle des indicateurs de qualité microbiologique des eaux de rivière, en explorant à la fois les incertitudes liées à la méthodologie dans des contextes variés et en fournissant des outils pour l'intégration de cette incertitude dans la prise de décision mais également par l'analyse du processus de décroissance et de disparition d'*E. coli*. Une harmonisation des pratiques d'échantillonnage et d'analyse permettrait de réduire les incertitudes et d'améliorer la comparabilité des données, notamment grâce à l'intégration de la logique floue et de dispositifs de suivi en quasi temps réel comme le système ColiMinder. Ces outils facilitent une prise de décision plus réactive et nuancée afin de savoir le jour même s'il faut ouvrir ou fermer la baignade.

Nous avons mis en évidence la pertinence des taux de disparition pour une compréhension plus approfondie de la dynamique d'*E. coli* dans la Seine et la Marne. Les résultats montrent des variations significatives des temps de retour selon les sites, soulignant l'impact des apports en amont et des rejets ponctuels, mais également de l'interaction entre différents paramètres sur cette évolution temporelle suite à un événement pluvial impactant la qualité microbiologique. L'approche expérimentale, couplée à des suivis en temps réel ou quasi-réel sur le terrain, offre des perspectives intéressantes pour optimiser la gestion des baignades, en particulier dans des conditions de pollution après une pluie.

L'analyse de marqueurs bactériens ou viraux supplémentaires, incluant des indicateurs fécaux spécifiques de sources de contaminations humaines ou animales et des pathogènes, pourrait enrichir la compréhension des dynamiques des différents marqueurs microbiens et de la résilience des sites de baignade face à une pollution microbienne ponctuelle. Ces données

contribueraient à une évaluation plus globale et à une gestion renforcée de la qualité des eaux en milieu urbain.

En tenant compte des taux de disparition d'*E. coli* et de la position du système de mesure en continu (Coliminder) par rapport au site de baignade et du débit de la rivière, il est possible de définir des intervalles temporels de prise de décision adaptés à chaque site. Par exemple, sur un site avec un temps de retour rapide et un débit élevé, un intervalle de mesure d'*E.coli* de quelques heures peut être pertinent pour capturer une dynamique représentative de la qualité de l'eau. Cet intervalle tient compte du temps nécessaire pour que l'eau atteigne la zone de baignade par rapport à la position du ColiMinder. En combinant ces informations dans un modèle de logique floue, les décisions de gestion peuvent être ajustées en fonction des caractéristiques spécifiques de chaque site, garantissant ainsi une meilleure précision dans l'évaluation des risques et la protection de la santé publique.

Conclusion générale

Ce travail a permis de mettre en lumière les défis liés à la gestion de la qualité des eaux de surface, en particulier dans des environnements fortement urbanisés. Les efforts se sont concentrés sur la mise en place d'approches innovantes, combinant des outils technologiques avancés, des modèles prédictifs robustes et le développement de guides méthodologiques, pour répondre aux exigences croissantes de surveillance et de gestion de la qualité des eaux de surface et développer des réseaux de surveillance intelligents (smart water). La figure 4.1 propose un cadre synthétique et structurant des différents aspects liés à la gestion des baignades en ville qui ont été abordés au cours de cette thèse. Les paragraphes suivants commentent la figure 4.1.

La gestion des baignades dans les rivières nécessite une approche intégrée pour optimiser la surveillance et réduire les incertitudes. Le processus débute avec la surveillance des sites de baignade pour mesurer le risque sanitaire lié à la présence potentielle de pathogènes dans les eaux de surface (Avila et al., 2018; Visser et al., 2022). Pour ce faire, des paramètres physico-chimiques et microbiologiques sont habituellement suivis, utilisant des outils comme des mesures réglementaires et des systèmes de mesure en (quasi) temps réel comme les capteurs à haute résolution (Cazals et al., 2020), les capteurs à bas coût (Farouk et al., 2022) et les dispositifs de mesure enzymatique ou microbiologiques automatisés (comme le système ColiMinder). Positionnés de manière stratégique, ces systèmes permettent un suivi en temps réel de paramètres physico-chimiques ou microbiologiques clés. Cependant, la quantité et la fiabilité des données envoyées sont cruciales (de Camargo et al., 2023). De ce fait, avant l'installation de ces systèmes automatisés *in situ*, différentes actions doivent être menées pour réduire l'incertitude sur la mesure (calibration, vérification de la stabilité du signal, correction du signal en fonction de variables influentes comme la température ou la luminosité). L'envoi des données doit être aussi optimisé, afin de s'assurer qu'il n'y aura pas de perte ou de dégradation de la qualité des données (Wang et al., 2019a). Nous avons ainsi développé un guide pour l'installation et la validation des capteurs physico-chimiques, tout en réduisant les coûts via l'utilisation de technologies IdO. Cependant, la quantité importante de données produites et les maintenances nécessaires sur les capteurs déployés *in situ*, posent la question de l'optimisation de leur installation, de leur entretien sur le long terme et du traitement des données. L'étape suivante serait de réduire la

dépendance à la supervision humaine. Pour ce faire, la semi-automatisation des tâches pourrait être facilitée par l'utilisation d'algorithmes permettant de détecter lorsqu'un capteur appartenant à un réseau devient défectueux, dérive ou cesse d'émettre (Chen and Han, 2018). Ainsi, la théorie des jeux peut offrir un cadre théorique à la création d'algorithmes de détection d'anomalies, ce qui peut aider à diminuer les coûts et le temps de gestion (Casado-Vara et al., 2018). Cependant, ce type d'approche pour gérer les réseaux de capteurs repose sur les données des capteurs voisins pour corriger les anomalies, rendant les résultats potentiellement biaisés en cas de défaillance généralisée ou dans des environnements dynamiques ou imprévisibles.

Les rivières sont justement des systèmes complexes très dynamiques et les pollutions microbiologiques présentent un aspect aléatoire qui les rend difficiles à prédire. En effet, la qualité des eaux de surface est influencée par des facteurs complexes, incluant l'hydromorphologie et l'hydrodynamique de la rivière, les caractéristiques du bassin versant, les événements météorologiques, les flux de rejets urbains et les caractéristiques propres à chaque espèce microbienne suivie (Zhu et al., 2022; Jia et al., 2021). L'approche que nous avons utilisée, basée sur l'analyse de la résistance et la résilience des sites de baignade, permet une caractérisation dynamique des contaminations microbiennes affectant ces sites. Afin d'affiner l'analyse du risque microbiologique associé à l'ouverture d'un site de baignade, un profilage des sources de contamination en amont du site est requis par la directive 2006/7/CE. Ces informations, combinées au suivi en temps réel, permettraient une gestion plus effective des plages urbaines pendant la saison de baignade et une meilleure sélection des stratégies pour améliorer la qualité des eaux. Il existe désormais un ensemble d'outils permettant de détecter l'origine des contaminations, tels que la recherche de bactéries ou de virus intestinaux spécifiques de leur hôte et la comparaison des communautés bactériennes (Ahmed et al., 2019b). Une meilleure compréhension de la dynamique de ces marqueurs spécifiques lors des événements polluants serait essentielle pour renforcer la précision des décisions de gestion mais fait encore largement défaut. Nous avons réalisé une évaluation de l'incertitude des méthodes d'échantillonnage et de mesure microbiologique des indicateurs bactériens de contamination fécale (indicateurs réglementaires et indicateurs spécifiques de sources). Diminuer l'incertitude sur la mesure des indicateurs permettrait de renforcer la robustesse des décisions de gestion tout en s'alignant sur les exigences de la directive 2006/7/CE. Ainsi, nous avons estimé l'incertitude de l'étape d'échantillonnage jusqu'à la mesure des BIF, ainsi que l'incertitude d'échantillonnage et de stockage de 3 indicateurs de sources animales (marqueurs moléculaires Gull2 pour les mouettes et goélands, BacCan pour les chiens

et CGOF1 pour les oies bernaches), d'un indicateur de sources humaines (marqueur moléculaire HF183) et de 2 pathogènes du genre *Campylobacter* (*C. jejuni* et *C. lari*). Cette incertitude sur la mesure et l'échantillonnage pourrait être intégrée au processus de prise de décision quant à la classe de qualité d'un échantillon d'eau en cours de saison de baignade pour savoir si la baignade peut être autorisée. En effet, étant données les incertitudes analysées, lorsque la valeur mesurée, ajoutée à son incertitude, est proche de la valeur seuil, la classification devient plus complexe (Brandão et al., 2022). Classer correctement la qualité microbiologique pour fermer une baignade peut permettre de prévenir 42% des maladies liées à la baignade dans des eaux de surface urbaines (Rabinovici et al., 2004; Ross, 2005). À l'aide d'un processus de logique floue intégrant l'incertitude de la mesure, nous avons démontré qu'il est possible d'utiliser les données acquises toutes les 2 h par un système ColiMinder pour classer correctement les échantillons et aider à la décision de fermeture d'une baignade le matin en s'appuyant sur les 4 à 24 heures de suivi précédentes.

Les données acquises en temps réel ou quasi-réel par les systèmes de mesure automatisée et les capteurs pourraient alimenter une base de données structurée permettant d'évaluer les dynamiques spatiales et temporelles des contaminations microbiologiques et chimiques, facilitant ainsi une surveillance via la prédiction intégrant une combinaison de modèles (les modèles hydrodynamiques et les modèles de machine learning) (Eregno et al., 2018; Qiu et al., 2017). Ce dernier peut être optimisé via des approches comme l'apprentissage par transfert et l'apprentissage fédéré permettant le transfert et le partage des connaissances des modèles, tout en assurant la confidentialité des données. L'association de ces modèles gouvernés par un méta-modèle pourrait permettre une amélioration de la prédiction en sélectionnant, pour chaque événement polluant sur chaque site surveillé, le modèle qui donne la prédiction la plus juste. Cette architecture peut augmenter l'adaptabilité de la modélisation à de nouveaux sites, de nouvelles conditions météorologiques ou des modifications du bassin versant, tout en prenant en compte les spécificités de chaque site de baignade. Les concentrations en BIF prédites permettraient d'alimenter un système d'alerte jouant un rôle clé, en intégrant une surveillance en temps réel de l'état actuel en utilisant des modèles comme le Random Forest qui se basent sur les données historiques de la base de données (nowcasting) et une planification en anticipant les tendances sur des périodes plus longues pour produire des prévisions (forecasting). Le forecasting comme par exemple avec les modèles LSTM, présente plusieurs avantages, notamment sa capacité à établir des relations non linéaires entre les variables de qualité de l'eau et à fournir des prévisions

fiables avec des structures simples (Liu et al., 2019; Shinde and Shah, 2018). Cependant, ses performances dépendent fortement de la qualité et de la quantité des données disponibles. Il est généralement recommandé d'avoir des données à intervalles réguliers afin de capturer les dépendances temporelles et les relations entre les observations passées et futures de manière plus précise (Liang et al., 2020). L'ensemble de ces informations sont essentielles pour prendre des décisions éclairées, telles que la planification des ouvertures ou la fermeture des sites de baignade en cas de pollution. Ce système offrirais également un mécanisme d'alerte pour guider les prélèvements manuels lorsque des incertitudes persistent dans la base de données lors des prédictions et cela par des approches d'apprentissage actif. Cela permettrait d'augmenter la base de données efficacement tout en minimisant les coûts et en réduisant les incertitudes des modèles de prédiction (Bouneffouf, 2016). Nous avons proposé une stratégie pour utiliser les modèles comme le Random Forest pour identifier les classes de données minoritaires parmi les paramètres predictifs du modèle. À l'aide de cet outil, il est possible d'augmenter la base de données soit par échantillonnage ciblé des classes minoritaires, soit par génération de données synthétiques lorsqu'il n'est pas possible d'obtenir les données manquantes qui déséquilibrent la distribution des données dans la base de données pour certains paramètres.

Les résultats de ce travail de thèse offrent des perspectives prometteuses pour la gestion durable des ressources en eau dans des environnements urbains. Ils mettent en avant l'intérêt de combiner des outils technologiques avancés avec des pratiques opérationnelles adaptées pour répondre aux défis environnementaux et réglementaires que pose l'ouverture de sites de baignade dans les rivières urbaines en période post-industrielle. En tenant compte de l'ensemble de ces outils et des connaissances disponibles, une décision plus avisée peut être prise. À long terme, l'intégration de ces dispositifs dans des réseaux intelligents de surveillance à l'échelle régionale pourrait non seulement améliorer la sécurité des usagers des rivières mais également renforcer les efforts de préservation des écosystèmes aquatiques urbains en permettant une gestion ciblée du site.

En conclusion, ce travail illustre la nécessité d'une approche transversale et interdisciplinaire pour relever les défis complexes liés à la gestion de la qualité de l'eau de surface dans les environnements urbains. L'amélioration de la surveillance des eaux de surface et la prise de décision d'ouverture ou fermeture des sites de baignade peut bénéficier de la combinaison des connaissances scientifiques, de l'innovation technologique, et de la validation sur le terrain. Améliorer la qualité des eaux de surface en vue de permettre la baignade et les activités

récréatives et sportives dans les rivières et canaux urbains, est un levier politique et sociétal puissant qui contribuera en même temps à l'amélioration de la qualité écologique de ces milieux aquatiques fortement impactés par l'urbanisation et l'activité humaine.

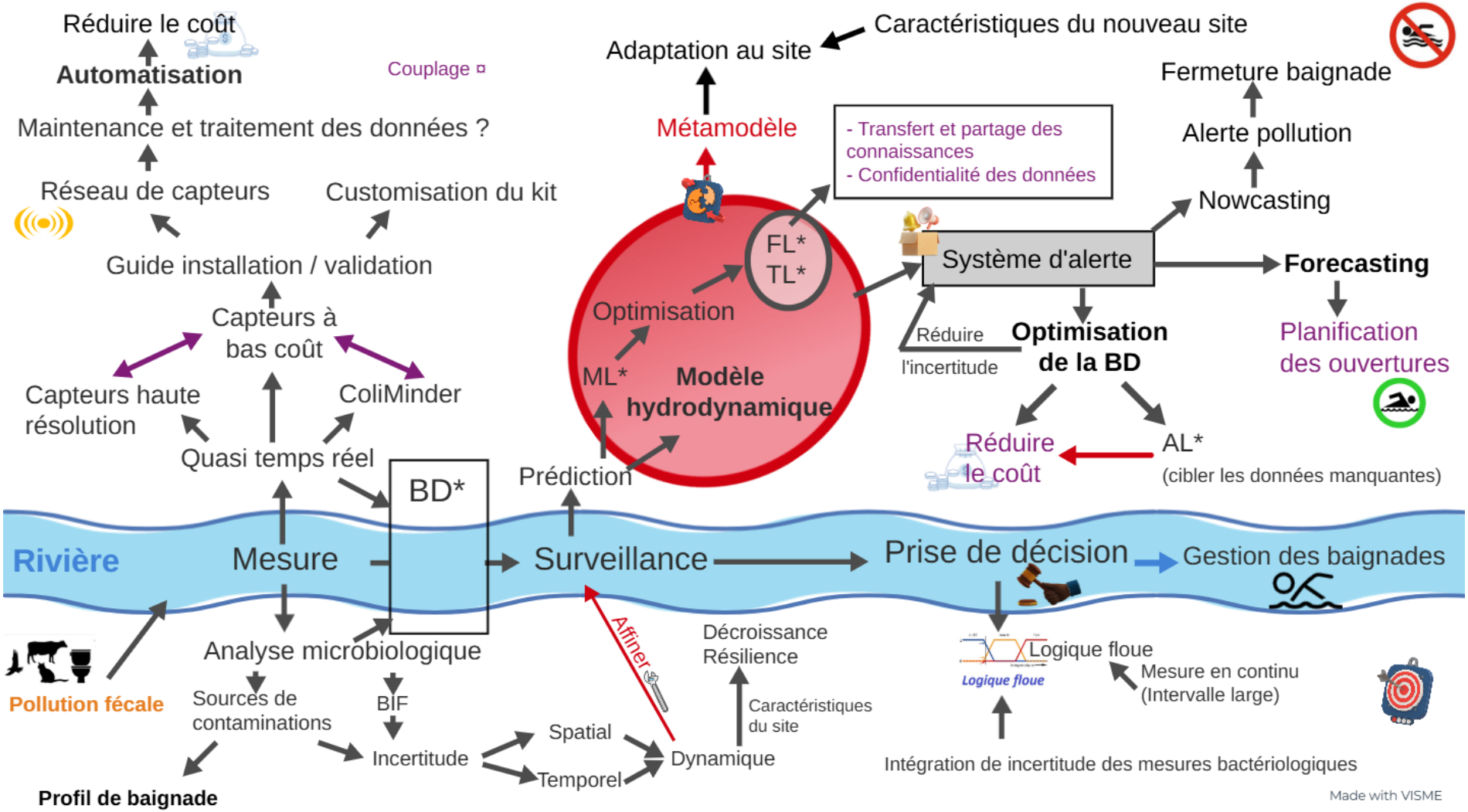

FIGURE 4.1 – Cadre général pour la gestion de la qualité des rivières et la prise de décision en matière de baignade. * : (ML : machine learning ; BD : base de donnée ; TL : transfert learning, FL : federate learning ; AL : active learning), ¤ : Couplage, rapport coût/bénéfice, en violet : les avantages des méthodes, en gras : les futures perspectives.

Références bibliographiques

Brook W Abegaz, Tania Datta, and Satish M Mahajan. Sensor technologies for the energy-water nexus–a review. *Applied energy*, 210 :451–466, 2018.

Mostafa Abotaleb. Authenticated wifi-based wireless data transmission from multiple sensors through a laboratory stand based on collaboration between atmega2560 and esp32 microcontrollers. *Scientific Journal of Gdynia Maritime University*, (127) :27–41, 2023.

AEE. European bathing water quality in 2023, 2024. URL `https://www.eea.europa.eu/publications/european-bathing-water-quality-in-2022`. Accessed : 2025-01-28.

AELB. Agence de l'eau loire-bretagne, le prélèvement d'échantillons en rivière. cours d'eau, surveillance eaux et milieux aquatiques. *ISBN 10 : 2-916869-00-X*, 2006.

AFNOR. FD T90-523-1 Qualité de l'eau - Guide de prélèvement pour le suivi de qualité des eaux dans l'environnement - Partie 1 : prélèvement d'eau superficielle. Technical report.

Agence Régionale de Santé Bretagne. Guide de recommandations sanitaires pour bases nautiques, 2017. Consulté le 12 novembre 2024.

Environmental Protection Agency. Assessment of the effects of holding time and enterococci concentrations in fresh and marine recreational waters and escherichia coli concentrations in fresh recreational waters. 2006.

Umair Ahmed, Rafia Mumtaz, Hirra Anwar, Asad A Shah, Rabia Irfan, and José García-Nieto. Efficient water quality prediction using supervised machine learning. *Water (Basel)*, 11 (11) :2210, 2019a. ISSN 2073-4441.

Warish Ahmed, C Staley, MJ Sadowsky, P Gyawali, Jatinder PS Sidhu, A Palmer, DJ Beale, and S Toze. Toolbox approaches using molecular markers and 16s rrna gene amplicon data sets for identification of fecal pollution in surface water. *Applied and environmental microbiology*, 81(20) :7067–7077, 2015.

Warish Ahmed, Aldo Lobos, Jacob Senkbeil, Jayme Peraud, Javier Gallard, and Valerie J Harwood. Evaluation of the novel crassphage marker for sewage pollution tracking in storm drain outfalls in tampa, florida. *Water Research*, 131 :142–150, 2018.

Warish Ahmed, Kerry Hamilton, Simon Toze, Stephen Cook, and Declan Page. A review on microbial contaminants in stormwater runoff and outfalls : Potential health risks and mitigation strategies. *Science of the Total Environment*, 692 :1304–1321, 2019b.

Warish Ahmed, Sudhi Payyappat, Michele Cassidy, Nathan Harrison, and Colin Besley. Inter-laboratory accuracy and precision among results of three sewage-associated marker genes in urban environmental estuarine waters and freshwater streams. *Science of the Total Environment*, 741 :140071, 2020.

Sedat Akkurt, Gokmen Tayfur, and Sever Can. Fuzzy logic model for the prediction of cement compressive strength. *Cement and concrete research*, 34(8) :1429–1433, 2004.

Fhranz Marc Lou S Alimorong, Haziel Anne D Apacionado, and Jocelyn Flores Villaverde. Arduino-based multiple aquatic parameter sensor device for evaluating ph, turbidity, conductivity and temperature. In *2020 IEEE 12th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM)*, pages 1–5. IEEE, 2020.

Elizabeth Wheeler Alm, Janice Burke, and Anne Spain. Fecal indicator bacteria are abundant in wet sand at freshwater beaches. *Water research*, 37(16) :3978–3982, 2003.

D.E Angelescu, V Huynh, A Hausot, G Yalkin, V Plet, J.-M Mouchel, S Guérin-Rechdaoui, S Azimi, and V Rocher. Autonomous system for rapid field quantification of Escherichia coli in surface waters. *Journal of applied microbiology*, 126(1) :332–343, 2019. ISSN 1364-5072.

Natalia Angelotti, Arthur Guillot-Le Goff, Rémi Arthur Carmigniani, and Vinçon-Leite Brigitte. Open water swimming in urban areas e. coli distribution with telemac-3d. In *XXVIIIth TELEMAC User Conference*, 2022.

Natália Angelotti de Ponte Rodrigues, Rémi Carmigniani, Arthur Guillot-Le Goff, Françoise S Lucas, Claire Therial, Manel Naloufi, Aurélie Janne, Francesco Piccioni, Mohamed Saad, Philippe Dubois, et al. Fluorescence spectroscopy for tracking microbiological contamination in urban waterbodies. *Frontiers in Water*, 6 :1358483, 2024.

APE États-Unis. Basic information about nonpoint source (nps) pollution, 2022. URL `https://www.epa.gov/nps/basic-information-about-nonpoint-source-nps-pollution`. Disponible sur : `https://www.epa.gov/nps/basic-information-about-nonpoint-source-nps-pollution`.

Arduino. Product description and specifications,analog turbidity sensor for arduino. https://www.dfrobot.com/product-1394.html, 2023. Accessed : November 2023.

Arduino Boards. What is the operating temperature range for arduino boards ? https://support.arduino.cc/hc/en-us/articles/360016076980-What-is-the-operating-temperature-range-for-Arduino-boards, 2023. Accessed : April 2023.

Arduino Pro Gateway Documentation. Arduino pro gateway : Product documentation. https://docs.arduino.cc/retired/kits/pro-gateway, 2023. Accessed : November 2023.

Henry-Joseph Audéoud, Martin Heusse, and Andrzej Duda. Single reception estimation of wireless link quality. In *2020 IEEE 31st Annual International Symposium on Personal, Indoor and Mobile Radio Communications*, pages 1–7. IEEE, 2020.

Martin T Auer and Stephen L Niehaus. Modeling fecal coliform bacteria—i. field and laboratory determination of loss kinetics. *Water Research*, 27(4) :693–701, 1993.

Rodelyn Avila, Beverley Horn, Elaine Moriarty, Roger Hodson, and Elena Moltchanova. Evaluating statistical model performance in water quality prediction. *Journal of environmental management*, 206 :910–919, 2018.

Tharsana Balachandran, Thiago Abreu, Manel Naloufi, Sami Souihi, Françoise Lucas, and Aurélie Janne. Iot and transfer learning based urban river quality prediction. In *GLOBECOM 2022-2022 IEEE Global Communications Conference*, pages 257–262. IEEE, 2022.

Talent Banda and Muthukrishnavellaisamy Kumarasamy. Application of multivariate statistical analysis in the development of a surrogate water quality index (wqi) for south african watersheds. *Water (Basel)*, 12(6) :1584, 2020. ISSN 2073-4441.

Flavio Barboza, Herbert Kimura, and Edward Altman. Machine learning models and bankruptcy prediction. *Expert systems with applications*, 83 :405–417, 2017. ISSN 0957-4174.

I Barcina, P Lebaron, and J Vives-Rego. Survival of allochthonous bacteria in aquatic systems : a biological approach. *FEMS Microbiology Ecology*, 23(1) :1–9, 1997.

Pascal Beaudeau, Nicolas Tousset, Franck Bruchon, Amélie Lefèvre, and Huw D Taylor. In situ measurement and statistical modelling of escherichia coli decay in small rivers. *Water Research*, 35(13) :3168–3178, 2001.

Brian L Benham, Claire Baffaut, Rebecca Winfrey Zeckoski, Kyle R Mankin, Yakov A Pachepsky, AM Sadeghi, Kevin M Brannan, ML Soupir, and MJ Habersack. Modeling

bacteria fate and transport in watersheds to support tmdls. *Transactions of the ASABE*, 49 (4) :987–1002, 2006.

P Bergeron, H Oujati, V Catalán Cuenca, JM Huguet Mestre, and S Courtois. Rapid monitoring of escherichia coli and enterococcus spp. in bathing water using reverse transcription-quantitative pcr. *International journal of hygiene and environmental health*, 214(6) : 478–484, 2011.

Eugen Betke and Julian Kunkel. Real-time I/O-monitoring of HPC applications with SIOX, elasticsearch, Grafana and FUSE. In *International Conference on High Performance Computing*, pages 174–186. Springer : Cham, Switzerland, 2017.

Ana Maria Bianco, M Garcia Ben, EJ Martinez, and Vıctor J Yohai. Outlier detection in regression models with arima errors using robust estimates. *Journal of Forecasting*, 20(8) : 565–579, 2001.

RA Blaustein, Y Pachepsky, RL Hill, DR Shelton, and G Whelan. Escherichia coli survival in waters : temperature dependence. *Water research*, 47(2) :569–578, 2013.

Alexandria B Boehm. Enterococci concentrations in diverse coastal environments exhibit extreme variability, 2007.

Alexandria B Boehm and Lauren M Sassoubre. Enterococci as indicators of environmental fecal contamination. 2014.

Alexandria B Boehm, SB Grant, JH Kim, SL Mowbray, CD McGee, CD Clark, DM Foley, and DE Wellman. Decadal and shorter period variability of surf zone water quality at huntington beach, california. *Environmental science & technology*, 36(18) :3885–3892, 2002.

Razvan Bogdan, Camelia Paliuc, Mihaela Crisan-Vida, Sergiu Nimara, and Darius Barmayoun. Low-cost internet-of-things water-quality monitoring system for rural areas. *Sensors*, 23 (8) :3919, 2023.

Angel Borja, Mathew P White, Elisa Berdalet, Nikolaj Bock, Claire Eatock, Peter Kristensen, Anne Leonard, Josep Lloret, Sabine Pahl, Mariluz Parga, et al. Moving toward an agenda on ocean health and human health in europe. *Frontiers in Marine Science*, 7 :37, 2020.

Gabrielle Bouleau, Françoise Lucas, Jean-Marie Mouchel, Sam Azimi, Sabine Barles, Marion Delarbre, Agathe Euzen, Angélique Goffin, Sabrina Guérin, Jean-Paul Haghe, Arthur

Guillot-Le Goff, Vincent Jauzein, Paul Kennouche, Laurence Lestel, Laurent Moulin, Julia Moutiez, Manel Naloufi, Vincent Rocher, Gaële Rouillé-Kielo, Gilles Varrault, Brigitte Vinçon-Leite, and Sébastien Wurtzer. *La baignade en Seine et en Marne*. Piren-Seine, Paris, France, 2024.

Djallel Bouneffouf. Exponentiated gradient exploration for active learning. *Computers (Basel)*, 5(1) :1–12, 2016. ISSN 2073-431X.

Andrew J. Bramburger, R. Stephen Brown, Jennifer Haley, and Jeffrey J. Ridal. A new, automated rapid fluorometric method for the detection of Escherichia coli in recreational waters. *Journal of Great Lakes Research*, 41(1) :298–302, 2015. ISSN 0380-1330.

Luciene Pires Brandão, Vanilson Fragoso Silva, Marcelo Bassi, and Elcio Cruz de Oliveira. Risk assessment in monitoring of water analysis of a brazilian river. *Molecules*, 27(11) : 3628, 2022.

L Breiman. Random forests. *Machine Learning*, 45 :5–32, 10 2001.

Leo Breiman. Bagging predictors. *Machine learning*, 24(2) :123–140, 1996. ISSN 0885-6125.

Wolfram Bremser, F-K Lücke, C Urmetzer, E Fuchs, and U Leist. An approach to integrated data assessment in a proficiency test on the enumeration of escherichia coli. *Journal of applied microbiology*, 110(1) :128–138, 2011.

Ciprian Briciu-Burghina, Brendan Heery, Gillian Duffy, Dermot Brabazon, and Fiona Regan. Demonstration of an optical biosensor for the detection of faecal indicator bacteria in freshwater and coastal bathing areas. *Analytical and bioanalytical chemistry*, 411 :7637–7643, 2019.

Lauren E Brooks and Katharine G Field. Bayesian meta-analysis to synthesize decay rate constant estimates for common fecal indicator bacteria. *Water research*, 104 :262–271, 2016.

Duie Tien Bui, Khabat Khosravi, John Tiefenbacher, Hoang Nguyen, and Nerantzis Kazakis. Improving prediction of water quality indices using novel hybrid machine-learning algorithms. *The Science of the total environment*, 721 :137612–137612, 2020. ISSN 0048-9697.

Jonas Bunsen, Markus Berger, and Matthias Finkbeiner. Planetary boundaries for water–a review. *Ecological Indicators*, 121 :107022, 2021.

Jean-Baptiste Burnet, Quoc Tuc Dinh, Sandra Imbeault, Pierre Servais, Sarah Dorner, and

Michèle Prévost. Autonomous online measurement of β-d-glucuronidase activity in surface water : is it suitable for rapid e. coli monitoring ? *Water research*, 152 :241–250, 2019.

Jean-Baptiste Burnet, Marc Habash, Mounia Hachad, Zeinab Khanafer, Michèle Prévost, Pierre Servais, Emile Sylvestre, and Sarah Dorner. Automated targeted sampling of waterborne pathogens and microbial source tracking markers using near-real time monitoring of microbiological water quality. *Water*, 13(15) :2069, 2021.

Muruleedhara N Byappanahalli, Meredith B Nevers, Asja Korajkic, Zachery R Staley, and Valerie J Harwood. Enterococci in the environment. *Microbiology and Molecular Biology Reviews*, 76(4) :685–706, 2012.

Davide Cacciarelli, Murat Kulahci, and John Sølve Tyssedal. Stream-based active learning with linear models. *Knowledge-Based Systems*, 254 :109664, 2022. ISSN 0950-7051. doi : https://doi.org/10.1016/j.knosys.2022.109664. URL `https://www.sciencedirect.com/science/article/pii/S0950705122008425`.

Marcos Tavares Carneiro, Myriam Bandeira Vianna Cortes, and Julio Cesar Wasserman. Critical evaluation of the factors affecting escherichia coli environmental decay for outfall plume models. *Revista Ambiente & Água*, 13(4) :e2106, 2018.

Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. Machine learning interpretability : A survey on methods and metrics. *Electronics*, 8(8) :832, 2019.

Roberto Casado-Vara, Francisco Prieto-Castrillo, and Juan M Corchado. A game theory approach for cooperative control to improve data quality and false data detection in wsn. *International Journal of Robust and Nonlinear Control*, 28(16) :5087–5102, 2018.

Margot Cazals. *Application d'une nouvelle technologie de détection enzymatique pour le suivi en quasi-temps réel de la dynamique d'Escherichia coli dans des eaux récréatives*. Ecole Polytechnique, Montreal (Canada), 2019.

Margot Cazals, Rebecca Stott, Carole Fleury, François Proulx, Michèle Prévost, Pierre Servais, Sarah Dorner, and Jean-Baptiste Burnet. Near real-time notification of water quality impairments in recreational freshwaters using rapid online detection of -D-glucuronidase activity as a surrogate for Escherichia coli monitoring. *The Science of the total environment*, 720 :137303–137303, 2020. ISSN 0048-9697.

YoonKyung Cha, Mi-Hyun Park, Sang-Hyup Lee, Joon Ha Kim, and Kyung Hwa Cho. Mo-

deling spatiotemporal bacterial variability with meteorological and watershed land-use characteristics. *Water research*, 100 :306–315, 2016.

Kangyang Chen, Hexia Chen, Chuanlong Zhou, Yichao Huang, Xiangyang Qi, Ruqin Shen, Fengrui Liu, Min Zuo, Xinyi Zou, Jinfeng Wang, Yan Zhang, Da Chen, Xingguo Chen, Yongfeng Deng, and Hongqiang Ren. Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data. *Water research (Oxford)*, 171 :115454–115454, 2020. ISSN 0043-1354.

Yiheng Chen and Dawei Han. Water quality monitoring in smart city : A pilot project. *Automation in construction*, 89 :307–316, 2018. ISSN 0926-5805.

Mohamed Cheniti, Belaout Abdesslam, and Ramdane Kaouthar. An arduino-based water quality monitoring system using ph, temperature, turbidity, and tds sensors. 2023.

P Chigbu, S Gordon, and TR Strange. Fecal coliform bacteria disappearance rates in a north-central gulf of mexico estuary. *Estuarine, Coastal and Shelf Science*, 65(1-2) :309–318, 2005.

Kyung Hwa Cho, Sung Min Cha, Joo-Hyon Kang, Seung Won Lee, Yongeun Park, Jung-Woo Kim, and Joon Ha Kim. Meteorological effects on the levels of fecal indicator bacteria in an urban stream : a modeling approach. *Water research*, 44(7) :2189–2202, 2010.

Soo Yeon Choi and Il Won Seo. Prediction of fecal coliform using logistic regression and tree-based classification models in the north han river, south korea. *Journal of Hydro-environment Research*, 21 :96–108, 2018.

Carlo Ciaponi, Enrico Creaco, Armando Di Nardo, Michele Di Natale, Carlo Giudicianni, Dino Musmarra, and Giovanni Francesco Santonastaso. Optimal sensor placement in a partitioned water distribution network for the water protection from contamination. In *Multidisciplinary Digital Publishing Institute Proceedings*, volume 2, page 670, 2018.

Lenore S Clesceri. Standard methods for examination of water and wastewater. *American public health association*, 9, 1998.

Commission européenne. Directive européenne 2006/7/ce. `https://baignades.sante.gouv.fr/baignades/editorial/fr/controle/directive2006_7_CE.pdf`, 2006.

Conductivity Meter. Gravity : Analog electrical conductivity sensor pro (k=1). 2024. Accessed : May 2024.

Conductivity Meter V2. Gravity : Analog electrical conductivity sensor meter v2 (k=1). 2023. Accessed : November 2023.

T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1) :21–27, 1967.

by SR Crane and JA Moore. Modeling enteric bacterial die-off : a review. *Water, Air, and Soil Pollution*, 27 :411–439, 1986.

Gunther F Craun, Rebecca L Calderon, and Michael F Craun. Outbreaks associated with recreational water in the united states. *International journal of environmental health research*, 15(4) :243–262, 2005.

Frank C Curriero, Jonathan A Patz, Joan B Rose, and Subhash Lele. The association between extreme precipitation and waterborne disease outbreaks in the united states, 1948–1994. *American journal of public health*, 91(8) :1194–1199, 2001.

Mike Cyterski, Orin C. Shanks, Pauline Wanjugi, Brian McMinn, Asja Korajkic, Kevin Oshima, and Rich Haugland. Bacterial and viral fecal indicator predictive modeling at three great lakes recreational beach sites. *Water Research*, 223 :118970, 2022. ISSN 0043-1354. doi : https://doi.org/10.1016/j.watres.2022.118970. URL `https://www.sciencedirect.com/science/article/pii/S0043135422009174`.

Amine Dahane, Rabaie Benameur, Manel Naloufi, Sami Souihi, Thiago Abreu, Francoise S. Lucas, and Abdelhamid Mellouk. Iot urban river water quality system using federated learning via knowledge distillation. In *ICC 2024 - IEEE International Conference on Communications*, pages 1515–1520, 2024. doi : 10.1109/ICC51166.2024.10622491.

Cheryl M Davies, Julian A Long, Margaret Donald, and Nicholas J Ashbolt. Survival of fecal microorganisms in marine and freshwater sediments. *Applied and Environmental Microbiology*, 61(5) :1888–1896, 1995.

Robert Davies-Colley, Amanda Valois, and Juliet Milne. Faecal pollution and visual clarity in new zealand rivers : Correlation of key variables affecting swimming suitability. *Journal of Water and Health*, 16(3) :329–339, 2018. doi : 10.2166/wh.2018.214.

Edson Tavares de Camargo, Fabio Alexandre Spanhol, Juliano Scholz Slongo, Marcos Vinicius Rocha da Silva, Jaqueline Pazinato, Adriana Vechai de Lima Lobo, Fábio Rizental Coutinho, Felipe Walter Dafico Pfrimer, Cleber Antonio Lindino, Marcio Seiji Oyamada,

and Leila Droprinchinski Martins. Low-cost water quality sensors for iot : A systematic review. *Sensors*, 23(9), 2023. ISSN 1424-8220. doi : 10.3390/s23094424. URL https://www.mdpi.com/1424-8220/23/9/4424.

Johan F. De Jonckheere. Origin and evolution of the worldwide distributed pathogenic amoeboflagellate naegleria fowleri. *Infection, Genetics and Evolution*, 11(7) :1520–1528, 2011. ISSN 1567-1348. doi : https://doi.org/10.1016/j.meegid.2011.07.023. URL https://www.sciencedirect.com/science/article/pii/S1567134811002784.

AM de Roda Husman and FM Schets. Climate change and recreational water-related infectious diseases. *RIVM rapport 330400002*, 2010.

Kara Dean and Jade Mitchell. Identifying water quality and environmental factors that influence indicator and pathogen decay in natural surface waters. *Water Research*, 211 :118051, 2022. ISSN 0043-1354. doi : https://doi.org/10.1016/j.watres.2022.118051. URL https://www.sciencedirect.com/science/article/pii/S0043135422000148.

Hugues Delamare, Alexandra Septfons, Serge Alfandari, and Alexandra Mailles. Freshwater sports and infectious diseases : A narrative review. *Infectious Diseases Now*, page 104883, 2024.

Alexander T Demetillo, Michelle V Japitana, and Evelyn B Taboada. A system for monitoring water quality in a large aquatic area using wireless sensor network technology. *Sustainable Environment Research*, 29(1) :1–9, 2019.

Jamie E DeNizio and David A Hewitt. Infection from outdoor sporting events—more risk than we think ? *Sports Medicine-Open*, 5(1) :37, 2019.

Anuradha M Desai and Hanadi S Rifai. Escherichia coli concentrations in urban watersheds exhibit diurnal sag : Implications for water-quality monitoring and assessment. *JAWRA Journal of the American Water Resources Association*, 49(4) :766–779, 2013.

Celly Desir. Étude de la qualité microbiologique des eaux naturelles urbaines, 2024. Mémoire d'apprentissage, École Nationale des Ponts et Chaussées, année 2023/2024. Soutenance le 2 septembre. Maître d'apprentissage : Marion Delarbre, Tuteur d'apprentissage : Gilles Varrault.

Megan L Devane, Beth Robson, Fariba Nourozi, Paula Scholes, and Brent J Gilpin. A pcr marker for detection in surface waters of faecal pollution derived from ducks. *Water research*, 41 (16) :3553–3560, 2007.

Megan L Devane, Louise Weaver, Shailesh K Singh, and Brent J Gilpin. Fecal source tracking methods to elucidate critical sources of pathogens and contaminant microbial transport through new zealand agricultural watersheds–a review. *Journal of environmental management*, 222 :293–303, 2018.

Megan L Devane, Elaine Moriarty, Louise Weaver, Adrian Cookson, and Brent Gilpin. Fecal indicator bacteria from environmental sources ; strategies for identification to improve water quality monitoring. *Water Research*, 185 :116204, 2020.

DFROBOT. Dfrobot official website. `https://www.dfrobot.com/`, 2023. Accessed : September 2023.

Zhenzhen Di, Miao Chang, and Peikun Guo. Water quality evaluation of the yangtze river in china using machine learning techniques and data monitoring on different time scales. *Water (Basel)*, 11(2) :339–, 2019. ISSN 2073-4441.

Linda K Dick, Erin A Stelzer, Erin E Bertke, Denise L Fong, and Donald M Stoeckel. Relative decay of bacteroidales microbial source tracking markers and cultivated escherichia coli in freshwater microcosms. *Applied and environmental microbiology*, 76(10) :3255–3262, 2010.

Sarkar Dipanjan. *Hands-on transfer learning with Python : implement advanced deep learning and neural network models using TensorFlow and Keras*. Packt Publishing, Birmingham, England, 2018. ISBN 978-17-8883-905-1.

DO. Gravity : Analog dissolved oxygen sensor. 2023. Accessed : November 2023.

Alba Balmaseda Dominguez. Bathing waters as urban activators : Observing ongoing practices in inland european cities. *PERFORMING SPACE*, page 241.

Maxime Dooms, Abalo Chango, and AL Abdel-Nour. Pcr quantitative (qpcr) et le guide de bonnes pratiques miqe : Adaptation et pertinence dans le contexte de la biologie clinique. *Ann. Biol. Clin*, 72(3) :265–269, 2014.

Dragino LoRa Shield. Lora shield : Product description and specifications. `https://wiki1.dragino.com/index.php/Lora_Shield`, 2023. Accessed : November 2023.

Ian G Droppo, Steven N Liss, Declan Williams, Tara Nelson, Chris Jaskot, and Brian Trapp. Dynamic existence of waterborne pathogens within river sediment compartments. implications for water quality regulatory affairs. *Environmental science technology*, 43(6) : 1737–1743, 2009. ISSN 0013-936X.

IG Droppo, BG Krishnappan, SN Liss, C Marvin, and J Biberhofer. Modelling sediment-microbial dynamics in the south nation river, ontario, canada : Towards the prediction of aquatic and human health risk. *Water Research*, 45(12) :3797–3809, 2011.

DS18B20. Waterproof ds18b20 digital temperature sensor. `https://wiki.dfrobot.com/Waterproof_DS18B20_Digital_Temperature_Sensor__SKU_DFR0198_`, 2023. Accessed : November 2023.

C. Duboudin, M. Legeas, P.-J. Cabillic, Y. Levi, and J. Lesne. *Qualité microbiologique des eaux de baignade. Valeurs seuils échantillon unique pour les eaux de baignade : étude de faisabilité méthodologique*. Afsset, 2007.

M. Elias Dueker, Gregory O'Mullan, Joaquín Martínez Martínez, Andrew Juhl, and Kathleen Weathers. Onshore wind speed modulates microbial aerosols along an urban waterfront. *Atmosphere*, 8(12) :215–, 2017. ISSN 2073-4433.

Al Dufour. A short history of methods used to measure bathing beach water quality. *Journal of microbiological methods*, 181 :106134, 2021.

Darcy L Ebentier, Kaitlyn T Hanley, Yiping Cao, Brian D Badgley, Alexandria B Boehm, Jared S Ervin, Kelly D Goodwin, Michèle Gourmelon, John F Griffith, Patricia A Holden, et al. Evaluation of the repeatability and reproducibility of a suite of qpcr-based microbial source tracking methods. *Water Research*, 47(18) :6839–6848, 2013.

Françoise Elbaz-Poulichet, Jean-Luc Seidel, Corinne Casiot, and Marie-Hélène Tusseau-Vuillemin. Short-term variability of dissolved trace element concentrations in the marne and seine rivers near paris. *Science of the Total Environment*, 367(1) :278–287, 2006.

C Elmas. Fuzzy logic inspections (theory, application, neural fuzzy logic). *Seckin, Ankara*, 230 : 1878, 2003.

Amber A Enns, Laura J Vogel, Amir M Abdelzaher, Helena M Solo-Gabriele, Lisa RW Plano, Maribeth L Gidley, Matthew C Phillips, James S Klaus, Alan M Piggot, Zhixuan Feng, et al. Spatial and temporal variation in indicator microbe sampling is influential in beach management decisions. *Water research*, 46(7) :2237–2246, 2012.

EPA. Method 1696 : characterization of human fecal pollution in water by hf183/bacr287 taqman quantitative polymerase chain reaction (qpcr) assay. 2019. *US EPA Washington, DC*, 2019.

Fasil Ejigu Eregno, Ingun Tryland, Torulv Tjomsland, Magdalena Kempa, and Arve Heistad.

Hydrodynamic modelling of recreational water quality using escherichia coli as an indicator of microbial contamination. *Journal of Hydrology*, 561 :179–186, 2018.

Kim H Esbensen and Claas Wagner. Theory of sampling (tos) versus measurement uncertainty (mu)–a call for integration. *TrAC Trends in Analytical Chemistry*, 57 :93–106, 2014.

European Commission. Quality of europe's bathing waters remains high. `https://ec.europa.eu/commission/presscorner/detail/en/ip_23_3041`, June 9 2023. Accessed : 2024-11-11.

Muhammad Izz Hakimi Zaidi Farouk, Mohd Fuad Abdul Latip, and Zadariana Jamil. Integrated water quality monitoring system and iot technology for surface water monitoring. In *AIP Conference Proceedings*, volume 2532, page 050002. AIP Publishing LLC, 2022.

Maeve Louise Farrell, Aoife Joyce, Sinead Duane, Kelly Fitzhenry, Brigid Hooban, Liam P Burke, and Dearbháile Morris. Evaluating the potential for exposure to organisms of public health concern in naturally occurring bathing waters in europe : A scoping review. *Water Research*, 206 :117711, 2021.

Christobel Ferguson, Ana Maria de Roda Husman, Nanda Altavilla, Daniel Deere, and Nicholas Ashbolt. Fate and transport of surface water pathogens in watersheds. 2003.

CM Ferguson. Refrigerated autosampling for the assessment of bacteriological water quality. *Water Research*, 28(4) :841–847, 1994.

Katie Fisher and Carol Phillips. The ecology, epidemiology and virulence of enterococcus. *Microbiology*, 155(6) :1749–1757, 2009.

James M. Fleisher, David Kay, Richard L. Salmon, Fiona Jones, Mark D. Wyer, and Andrew F. Godfree. Marine waters contaminated with domestic sewage : nonenteric illnesses associated with bather exposure in the united kingdom. *American Journal of Public Health*, 86 (9) :1228–1234, September 1996. doi : 10.2105/ajph.86.9.1228.

Donna S Francy, Amie MG Brady, Jessica R Cicale, Harrison D Dalby, and Erin A Stelzer. Nowcasting methods for determining microbiological water quality at recreational beaches and drinking-water source waters. *Journal of Microbiological Methods*, 175 :105970, 2020.

B Fremaux, T Boa, and Chris K Yost. Quantitative real-time pcr assays for sensitive detection of canada goose-specific fecal pollution in water sources. *Applied and Environmental Microbiology*, 76(14) :4886–4889, 2010.

Y. Freund and R. Schapire. Experiments with a new boosting algorithm. In *In Proceedings of the Thirteenth International Conference on International Conference on Machine Learning, Bari, Italy*, page 148–156. Morgan Kaufmann Publishers Inc. : San Francisco, CA, USA, 1996.

J Stephen Fries, Gregory W Characklis, and Rachel T Noble. Attachment of fecal indicator bacteria to particles in the neuse river estuary, nc. *Journal of Environmental Engineering*, 132(10) :1338–1345, 2006.

J Stephen Fries, Gregory W Characklis, and Rachel T Noble. Sediment–water exchange of vibrio sp. and fecal indicator bacteria : implications for persistence and transport in the neuse river estuary, north carolina, usa. *Water research*, 42(4-5) :941–950, 2008.

Helen Galfi, Kerstin Nordqvist, Monica Sundelin, Godecke-Tobias Blecken, Jiri Marsalek, and Maria Viklander. Comparison of indicator bacteria concentrations obtained by automated and manual sampling of urban storm-water runoff. *Water, Air, & Soil Pollution*, 225 :1–12, 2014.

Tamara Garcia-Armisen and Pierre Servais. Respective contributions of point and non-point sources of e. coli and enterococci in a large urbanized watershed (the seine river, france). *Journal of environmental management*, 82(4) :512–518, 2007.

Tamara Garcia-Armisen and Pierre Servais. Partitioning and fate of particle-associated e. coli in river water. *Water environment research : a research publication of the Water Environment Federation*, 81 :21–8, 02 2009.

Mekonnen Gebremichael, Menberu M Bitew, Feyera A Hirpa, and Gebrehiwot N Tesfay. Accuracy of satellite rainfall estimates in the b lue n ile b asin : L owland plain versus highland mountain. *Water Resources Research*, 50(11) :8775–8790, 2014.

AH Geeraerd, VP Valdramidis, and JF Van Impe. Ginafit, a freeware tool to assess non-log-linear microbial survivor curves. *International journal of food microbiology*, 102(1) :95–105, 2005.

Zoubin Ghahramani. Probabilistic machine learning and artificial intelligence. *Nature*, 521 : 452–459, 2015.

Hamed Gharibi, Amir Hossein Mahvi, Ramin Nabizadeh, Hossein Arabalibeik, Masud Yunesian, and Mohammad Hossein Sowlat. A novel approach in water quality assessment based on fuzzy logic. *Journal of Environmental Management*, 112 :87–95, 2012.

Subhasis Giri. Water quality prospective in twenty first century : Status of water quality in major river basins, contemporary strategies and impediments : A review. *Environmental Pollution*, 271 :116332, 2021.

Youdi Gong, Guangzhen Liu, Yunzhi Xue, Rui Li, and Lingzhong Meng. A survey on dataset quality in machine learning. *Information and Software Technology*, 162 :107268, 2023.

Lucia Gonzales-Siles and Åsa Sjöling. The different ecological niches of enterotoxigenic e scherichia coli. *Environmental microbiology*, 18(3) :741–751, 2016.

David M Gordon. The ecology of escherichia coli. In *Escherichia coli*, pages 3–20. Elsevier, 2013.

Michele Gourmelon. Contamination des eaux de baignade et des coquillages par des matières fécales : comment identifier les sources. *The Conversation*, 2023.

D Gowri, S SHIVA Prasad, O Kiran, and Mr M Mysaiah. Water quality level monitoring system using arduino. *J. Eng. Sci*, 14 :226–229, 2023.

Gravity. Product description and specifications,i2c ads1115 16-bit adc module. `https://www.dfrobot.com/product-1730.html`, 2023. Accessed : November 2023.

Hyatt C Green, Richard A Haugland, Manju Varma, Hana T Millen, Mark A Borchardt, Katharine G Field, William A Walters, R Knight, Mano Sivaganesan, Catherine A Kelty, et al. Improved hf183 quantitative real-time pcr assay for characterization of human fecal pollution in ambient surface water samples. *Applied and environmental microbiology*, 80 (10) :3086–3094, 2014.

Dale W Griffin, Erin K Lipp, Molly R Mclaughlin, and Joan B Rose. Marine recreation and public health microbiology : Quest for the ideal indicator : This article addresses the historic, recent, and future directions in microbiological water quality indicator research. *Bioscience*, 51(10) :817–825, 2001.

Andrew D Gronewold and Robert L Wolpert. Modeling the relationship between most probable number (mpn) and colony-forming unit (cfu) estimates of fecal coliform concentration. *Water research*, 42(13) :3327–3334, 2008.

Andrew D. Gronewold, Luke Myers, Jenise L. Swall, and Rachel T. Noble. Addressing uncertainty in fecal indicator bacteria dark inactivation rates. *Water Research*, 45(2) :652–664, 2011. ISSN 0043-1354. doi : https://doi.org/10.1016/j.watres.2010.08.029. URL `https://www.sciencedirect.com/science/article/pii/S0043135410005932`.

Hélène Guérineau, Sarah Dorner, Annie Carrière, Natasha McQuaid, Sébastien Sauvé, Khadija Aboulfadl, Mariam Hajj-Mohamad, and Michèle Prévost. Source tracking of leaky sewers : a novel approach combining fecal indicators in water and sediments. *Water Research*, 58 : 50–61, 2014.

Amir Guidara, Ghofrane Fersi, Maher Ben Jemaa, and Faouzi Derbel. A new deep learning-based distance and position estimation model for range-based indoor localization systems. *Ad Hoc Networks*, 114 :102445, 2021. ISSN 1570-8705. doi : https://doi.org/10.1016/j.adhoc.2021.102445. URL `https://www.sciencedirect.com/science/article/pii/S1570870521000214`.

Guide Îles de loisirs. Région Île-de-france. `https://www.iledefrance.fr/sites/files/2024-03`, March 2024. Accessed : January 2025.

Nathalie Guigues, Bénédicte Lepot, Michèle Desenfant, and Jacky Durocher. Estimation of the measurement uncertainty, including the contribution arising from sampling, of water quality parameters in surface waters of the loire-bretagne river basin, france. *Accreditation and Quality Assurance*, 25 :281–292, 2020.

Arthur Guillot-Le Goff, Natalia Angelotti, Rémi Carmigniani, Guilherme Calabro Souza, Mohamed Saad, Philippe Dubois, and Brigitte Vinçon-Leite. Prévision de la qualité microbiologique des milieux aquatiques : modélisation hydrodynamique pour anticiper des épisodes de contamination microbiologique sur des sites de baignade urbaine. In *Novatech 2023 : 11e Conférence internationale sur l'eau dans la ville*, 2023.

Abdul Jailani Gusri and Harmadi Harmadi. Rancang bangun alat penguras air pada wadah penampungan berbasis turbidity sensor sen0189. *Jurnal Fisika Unand*, 10(3) :330–336, 2021.

Aurélien Géron. *Machine Learning avec Scikit-Learn*. Dunod, 2019. ISBN 9782100797820.

Josiah Hacker. Assessing the ability of arduino-based sensor systems to monitor changes in water quality. 2023.

WL Hakim, L Hasanah, B Mulyanti, and A Aminudin. Characterization of turbidity water sensor sen0189 on the changes of total suspended solids in the water. In *Journal of Physics : Conference Series*, volume 1280, page 022064. IOP Publishing, 2019.

Mohammed Hameed, Saadi Shartooh Sharqi, Zaher Mundher Yaseen, Haitham Abdulmohsin Afan, Aini Hussain, and Ahmed Elshafie. Application of artificial intelligence (ai) tech-

niques in water quality index prediction : a case study in tropical region, malaysia. *Neural computing applications*, 28(S1) :893–905, 2017. ISSN 0941-0643.

William P Hamilton, Moonil Kim, and Edward L Thackston. Comparison of commercially available escherichia coli enumeration tests : Implications for attaining water quality standards. *Water Research*, 39(20) :4869–4878, 2005.

RD Harmel, JM Hathaway, KL Wagner, JE Wolfe, R Karthikeyan, Wendy Francesconi, and DT McCarthy. Uncertainty in monitoring e. coli concentrations in streams and stormwater runoff. *Journal of Hydrology*, 534 :524–533, 2016.

Masihullah Hasanyar. *High Frequency Data Assimilation in the ProSe-PA Water Quality Model : Focus on the drivers of river metabolism under low flow conditions*. PhD thesis, Université Paris sciences et lettres, 2023.

Trevor. Hastie. *The Elements of Statistical Learning Data Mining, Inference, and Prediction, Second Edition*. Springer Series in Statistics. Springer New York, New York, NY, 2nd ed. 2009. edition, 2009. ISBN 9780387848587.

JM Hathaway, WF Hunt, and OD Simmons III. Statistical evaluation of factors affecting indicator bacteria in urban storm-water runoff. *Journal of Environmental Engineering*, 136(12) : 1360–1368, 2010.

Jon M Hathaway, William F Hunt, Robert M Guest, and David Thomas McCarthy. Residual indicator bacteria in autosampler tubing : A field and laboratory assessment. *Water Science and Technology*, 69(5) :1120–1126, 2014.

Masaki Hayashi. Temperature-electrical conductivity relation of water for environmental monitoring and geophysical data inversion. *Environmental monitoring and assessment*, 96 : 119–128, 2004.

Yiping He, Xiaomin Yao, Nereus W Gunther, Yanping Xie, Shu-I Tu, and Xianming Shi. Simultaneous detection and differentiation of campylobacter jejuni, c. coli, and c. lari in chickens using a multiplex real-time pcr assay. *Food analytical methods*, 3 :321–329, 2010.

Serge Hébert and Stéphane Légaré. Suivi de la qualité de l'eau des rivières et des petits cours d'eau. 2000.

John David Hem. *Study and interpretation of the chemical characteristics of natural water*, volume 2254. Department of the Interior, US Geological Survey, 1985.

Ilona Herrig, Wolfgang Seis, Helmut Fischer, Julia Regnery, Werner Manz, Georg Reifferscheid, and Simone Böer. Prediction of fecal indicator organism concentrations in rivers : the shifting role of environmental factors under varying flow conditions. *Environmental Sciences Europe*, 31 :1–16, 2019.

Michael L Hitchman. Measurement of dissolved oxygen. *(No Title)*, 1978.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9 (8) :1735–1780, 1997.

Wong Jun Hong, Norazanita Shamsuddin, Emeroylariffion Abas, Rosyzie Anna Apong, Zarifi Masri, Hazwani Suhaimi, Stefan Herwig Gödeke, and Muhammad Nafi Aqmal Noh. Water quality monitoring with arduino based sensors. *Environments*, 8(1) :6, 2021.

Yamen M Hoque, Shivam Tripathi, Mohamed M Hantush, and Rao S Govindaraju. Watershed reliability, resilience and vulnerability analysis under uncertainty using water quality data. *Journal of Environmental Management*, 109 :101–112, 2012.

Corine J Houtman. Emerging contaminants in surface waters and their relevance for the production of drinking water in europe. *Journal of Integrative Environmental Sciences*, 7(4) : 271–295, 2010.

Juan Huan, Hui Li, Fan Wu, and Weijian Cao. Design of water quality monitoring system for aquaculture ponds based on nb-iot. *Aquacultural Engineering*, 90 :102088, 2020.

Hydrolab DS5X. User manual for hydrolab ds5x, ds5, and ms5 water quality multiprobes. `https://www.ott.com/download/user-manual-hydrolab-ds5x-ds5-and-ms5-water-quality-multiprobes-1/`, 2024. Accessed January 2024.

Hydrolab Sensor. Hydrolab ms5/ds5/ds5x sensor : Product datasheet. `https://www.hemmis.com/docs/hydrolab/MS5_DS5_DS5X_low_F.pdf`, 2024. Accessed : January 2024.

ElMahdy Mohamed ElMahdy Ibrahim, Mohamed Azab El-Liethy, Akebe Luther King Abia, Bahaa Ahmed Hemdan, and Mohamed Nasr Shaheen. Survival of e. coli o157 : H7, salmonella typhimurium, hadv2 and mnv-1 in river water under dark conditions and varying storage temperatures. *Science of the total environment*, 648 :1297–1304, 2019.

Yilmaz Icaga. Fuzzy evaluation of water quality classification. *Ecological Indicators*, 7(3) : 710–718, 2007.

Maryam Imani, Md Mahmudul Hasan, Luiz Fernando Bittencourt, Kent McClymont, and Zoran Kapelan. A novel machine learning application : Water quality resilience prediction model. *Science of the Total Environment*, 768 :144459, 2021.

Satoshi Ishii and Michael J Sadowsky. Escherichia coli in the environment : implications for water quality and human health. *Microbes and environments*, 23(2) :101–108, 2008.

MM Majedul Islam, Muhammad Shahid Iqbal, Rik Leemans, and Nynke Hofstra. Modelling the impact of future socio-economic and climate change scenarios on river microbial water quality. *International Journal of Hygiene and Environmental Health*, 221(2) :283–292, 2018.

MS Islam, A Akbar, AYSHA Akhtar, MM Kibria, and MS Bhuyan. Water quality assessment along with pollution sources of the halda river. *Journal of the Asiatic Society of Bangladesh, Science*, 43(1) :61–70, 2017.

Vidhatri Iyer. Forecasting urban water escherichia coli contamination using machine learning models. 2024.

Isabelle Jalliffier-Verne, Robert Leconte, Uriel Huaringa-Alvarez, Mourad Heniche, Anne-Sophie Madoux-Humery, Laurène Autixier, Martine Galarneau, Pierre Servais, Michèle Prévost, and Sarah Dorner. Modelling the impacts of global change on concentrations of escherichia coli in an urban river. *Advances in Water Resources*, 108 :450–460, 2017.

Cheng-Shin Jang. Using probability-based spatial estimation of the river pollution index to assess urban water recreational quality in the tamsui river watershed. *Environmental monitoring and assessment*, 188(1) :1–17, 2016. ISSN 0167-6369.

J. Jantzen. Design of fuzzy controllers, technical report. *Department of Automation*, No :98-E864, 1999.

Emilie Jardé, Laurent Jeanneau, Loïc Harrault, Emmanuelle Quenot, Olivia Solecki, Patrice Petitjean, Solen Lozach, Julien Chevé, and Michèle Gourmelon. Application of a microbial source tracking based on bacterial and chemical markers in headwater and coastal catchments. *Science of The Total Environment*, 610 :55–63, 2018.

B Jarvis, Cordula Wilrich, and P-T Wilrich. Reconsideration of the derivation of most probable numbers, their standard deviations, confidence bounds and rarity values. *Journal of applied microbiology*, 109(5) :1660–1667, 2010.

P Jeroschewski and D Zur Linden. A flow system for calibration of dissolved oxygen sensors. *Fresenius' journal of analytical chemistry*, 358 :677–682, 1997.

Xiaowei Jia, Beiyu Lin, Jacob Zwart, Jeffrey Sadler, Alison Appling, Samantha Oliver, and Jordan Read. *Graph-based Reinforcement Learning for Active Learning in Real Time : An Application in Modeling River Networks*, pages 621–629. 2021.

Jiping Jiang, Sijie Tang, Dawei Han, Guangtao Fu, Dimitri Solomatine, and Yi Zheng. A comprehensive review on the design and optimization of surface water quality monitoring networks. *Environmental modelling & software : with environment data news*, 132 :104792, 2020. ISSN 1364-8152.

Guang Jin, Andrew J Englande, Henry Bradford, and Huei-Wang Jeng. Comparison of e. coli, enterococci, and fecal coliform as indicators for brackish water quality assessment. *Water environment research*, 76(3) :245–255, 2004.

Dusan Jovanovic, Simone Gelsinari, Louise Bruce, Matthew Hipsey, Ian Teakle, Matthew Barnes, Rhys Coleman, Ana Deletic, and David T Mccarthy. Modelling shallow and narrow urban salt-wedge estuaries : Evaluation of model performance and sensitivity to optimise input data collection. *Estuarine, coastal and shelf science*, 217 :9–27, 2019. ISSN 0272-7714.

Slaven Jozić, Mira Morović, Mladen Šolić, Nada Krstulović, and Marin Ordulj. Effect of solar radiation, temperature and salinity on the survival of two different strains of escherichia coli. *Fresenius Environ. Bull*, 23(8) :1852–1859, 2014.

Slaven Jozić, Arijana Cenov, Marin Glad, Danijela Peroš-Pucar, Katarina Kurić, Tatjana Puljak, Marin Ordulj, Ana Vrdoljak Tomaš, Nikolina Baumgartner, Damir Ivanković, et al. The effect of sampling frequency and spatial and temporal variation in the density of fecal indicator bacteria on the assessment of coastal bathing water quality. *Water research*, 264 : 122192, 2024.

Syun-suke Kadoya, Osamu Nishimura, Hiroyuki Kato, and Daisuke Sano. Predictive water virology : hierarchical bayesian modeling for estimating virus inactivation curve. *Water*, 11(10) :2187, 2019.

Prakash Raj Kannel, Seockheon Lee, Young-Soo Lee, Sushil Raj Kanel, and Siddhi Pratap Khan. Application of water quality indices and dissolved oxygen as indicators for river water

classification and urban impact assessment. *Environmental monitoring and assessment*, 132 :93–110, 2007.

David Kay, John Crowther, Carl M Stapleton, Mark D Wyer, Lorna Fewtrell, A Edwards, CA Francis, Adrian T McDonald, John Watkins, and J Wilkinson. Faecal indicator organism concentrations in sewage and treated effluents. *Water Research*, 42(1-2) :442–454, 2008.

George V Keller and Conrad Frank. Electrical methods in geophysical prospecting. *(No Title)*, 1966.

Beverly J Kildare, Christian M Leutenegger, Belinda S McSwain, Dustin G Bambic, Veronica B Rajal, and Stefan Wuertz. 16s rrna-based assays for quantitative detection of universal, human-, cow-, and dog-specific fecal bacteroidales : a bayesian approach. *Water research*, 41(16) :3701–3715, 2007.

Thomas Kistemann, Alexandra Schmidt, and Hans-Curt Flemming. Post-industrial river water quality—fit for bathing again ? *International Journal of Hygiene and Environmental Health*, 219(7) :629–642, 2016.

Public Lab KnowFLow. 2021. URL `url{https://www.eea.europa.eu/publications/european-bathing-water-quality-in-2018}`. Accessed on 30 June 2021.

Asja Korajkic, Brian R McMinn, Orin C Shanks, Mano Sivaganesan, G Shay Fout, and Nicholas J Ashbolt. Biotic interactions and sunlight affect persistence of fecal indicator bacteria and microbial source tracking genetic markers in the upper mississippi river. *Applied and environmental microbiology*, 80(13) :3952–3961, 2014.

Asja Korajkic, Brian R. McMinn, and Valerie J. Harwood. Relationships between microbial indicators and pathogens in recreational water settings. *International Journal of Environmental Research and Public Health*, 15(12) :2842, December 2018. doi : 10.3390/ijerph15122842.

Asja Korajkic, Brian R McMinn, Nicholas J Ashbolt, Mano Sivaganesan, Valerie J Harwood, and Orin C Shanks. Extended persistence of general and cattle-associated fecal indicators in marine and freshwater environment. *Science of the Total Environment*, 650 :1292–1302, 2019.

Hana Krakovská, Christian Kuehn, and Iacopo P Longo. Resilience of dynamical systems. *European Journal of Applied Mathematics*, 35(1) :155–200, 2024.

S Krishnan, R Manikandan, et al. Water quality prediction : a data-driven approach exploiting

advanced machine learning algorithms with data augmentation. *Journal of Water and Climate Change*, 2024.

Leigh-Anne H Krometis, Gregory W Characklis, Otto D Simmons III, Mackenzie J Dilts, Christina A Likirdopulos, and Mark D Sobsey. Intra-storm variability in microbial partitioning and microbial loading rates. *Water Research*, 41(2) :506–516, 2007.

Peter Kruse. Review on water quality sensors. *Journal of Physics D : Applied Physics*, 51(20) : 203002, 2018.

Molly J Lane. The implementation of qpcr beach monitoring methods : Analysis of a multi-lab validation study and the role of environmental parameters on a comparison of colilert and qpcr methods. 2019.

Hach Lange. *Manuel d'instruction Préleveur d'échantillons BÜHLER 2000*, 2012. Disponible sur : `www.hach-lange.com`.

Blythe A Layton, Yiping Cao, Darcy L Ebentier, Kaitlyn Hanley, Elisenda Balleste, João Brandão, Muruleedhara Byappanahalli, Reagan Converse, Andreas H Farnleitner, Jennifer Gentry-Shields, et al. Performance of human fecal anaerobe-associated pcr-based assays in a multi-laboratory method evaluation study. *Water Research*, 47(18) :6897–6908, 2013.

Julian Le Deunf, Nathalie Debese, Thierry Schmitt, and Romain Billot. A review of data cleaning approaches in a hydrographic framework with a focus on bathymetric multibeam echosounder datasets. *Geosciences*, 10(7) :254, 2020.

Dae-Young Lee, Kelly Shannon, and Lee A Beaudette. Detection of bacterial pathogens in municipal wastewater using an oligonucleotide microarray and real-time quantitative pcr. *Journal of microbiological methods*, 65(3) :453–467, 2006.

C. D. Lewis. *Industrial and business forecasting methods : a practical guide to exponential smoothing and curve fitting / Colin D. Lewis*. Butterworth Scientific London, 1982. ISBN 0408005599.

Junnan Li, Qingsheng Zhu, Quanwang Wu, and Zhu Fan. A novel oversampling technique for class-imbalanced learning based on smote and natural neighbors. *Information Sciences*, 565 :438–455, 2021.

Yi Li, Jan Degener, Matthew Gaudreau, Yangfan Li, and Martin Kappas. Adaptive capacity based water quality resilience transformation and policy implications in rapidly urbanizing landscapes. *Science of the Total Environment*, 569 :168–178, 2016.

Zhongyao Liang, Rui Zou, Xing Chen, Tingyu Ren, Han Su, and Yong Liu. Simulate the forecast capacity of a complicated water quality model using the long short-term memory approach. *Journal of Hydrology*, 581 :124432, 2020.

Cindy M Liu, Maliha Aziz, Sergey Kachur, Po-Ren Hsueh, Yu-Tsung Huang, Paul Keim, and Lance B Price. Bactquant : an enhanced broad-coverage bacterial quantitative real-time pcr assay. *BMC microbiology*, 12 :1–13, 2012.

Hancong Liu, Sirish Shah, and Wei Jiang. On-line outlier detection and data cleaning. *Computers & chemical engineering*, 28(9) :1635–1647, 2004.

Lubo Liu. Application of a hydrodynamic and water quality model for inland surface water systems. *Applications in water systems management and modeling*, 10, 2018.

Lubo Liu, Mantha S Phanikumar, Stephanie L Molloy, Richard L Whitman, Dawn A Shively, Meredith B Nevers, David J Schwab, and Joan B Rose. Modeling the transport and inactivation of e. coli and enterococci in the near-shore region of lake michigan. *Environmental science & technology*, 40(16) :5022–5028, 2006.

Ping Liu, Jin Wang, Arun Sangaiah, Yang Xie, and Xinchun Yin. Analysis and prediction of water quality using lstm deep neural networks in iot environment. *Sustainability (Basel, Switzerland)*, 11(7) :2058–, 2019. ISSN 2071-1050.

Sin Kit Lo, Qinghua Lu, Chen Wang, Hye-Young Paik, and Liming Zhu. A systematic literature review on federated machine learning : From a software engineering perspective. *ACM computing surveys*, 54(5) :1–39, 2021. ISSN 0360-0300.

Stelina Loiodice. Rapport d'apprentissage, master ii, mention biologie-santé, finalité microbiologie & génie biologique, 2023-2024. Rapport, Eau de Paris – DRDQE, R&D Biologie, 33 Avenue Jean Jaurès, 94200 Ivry-sur-Seine, 2024.

Iago López, César Álvarez, José L Gil, and José A Revilla. Does the bathing water classification depend on sampling strategy ? a bootstrap approach for bathing water quality assessment, according to directive 2006/7/ec requirements. *Journal of environmental management*, 111 :236–242, 2012.

LoRa. Gps lora hat for raspberry pi : Product description. `https://fr.farnell.com/seeed-studio/113990254/gps-lora-hat-for-raspberry-pi/dp/3498581?gad_source=5&cjevent=6c2f6bd47f0811ee816c1cda0a18ba74&cjdata=`

MXxZfDB8WXww&CMP=AFC-CJ-FR-1765328&gross_price=true&source=CJ, 2023. Accessed : November 2023.

Hongfang Lu and Xin Ma. Hybrid decision tree-based machine learning models for short-term water quality prediction. *Chemosphere (Oxford)*, 249 :126169–126169, 2020. ISSN 0045-6535.

F Lucas and P Servais. Etude de la qualité bactériologique de la zone aval de la marne. *Qualité bactériologique de l'aval de la Marne*, pages ,1–17, 2016.

Françoise Lucas, Bernard de Gouvello, Jean-Marie Mouchel, Laurent Moulin, Pierre Servais, and Sébastien Wurtzer. The microbiological quality of the seine river : Is it compatible with open water olympic competitions ? In *Hosting the Olympic Games*, pages 163–177. Routledge, 2019.

Françoise S Lucas, Claire Therial, Alexandre Gonçalves, Pierre Servais, Vincent Rocher, and Jean-Marie Mouchel. Variation of raw wastewater microbiological quality in dry and wet weather conditions. *Environmental Science and Pollution Research*, 21 :5318–5328, 2014.

Françoise Lucas, Pierre Servais, and Aurélie Janne. *Qualité bactériologique de la zone aval de la Marne, Synthèse des campagnes estivales 2015-2019*. Rapport 2020, 2020.

Diana Di Luccio, Angelo Riccio, Ardelio Galletti, Giuliano Laccetti, Marco Lapegna, Livia Marcellino, Sokol Kosta, and Raffaele Montella. Coastal marine data crowdsourcing using the internet of floating things : Improving the results of a water quality model. *IEEE access*, 8 :101209–101223, 2020. ISSN 2169-3536.

C Mahabir, FE Hicks, and A Robinson Fayek. Application of fuzzy logic to forecast seasonal runoff. *Hydrological processes*, 17(18) :3749–3762, 2003.

Konstantinos F Makris, Bas Hoefeijzers, Laura Seelen, Remy Schilperoort, and Jeroen G Langeveld. The potential of near real-time monitoring of β-d-glucuronidase activity to establish effective warning systems in urban recreational waters. *Environmental Science : Water Research & Technology*, 9(12) :3257–3268, 2023.

Michael A. Mallin, Kathleen E. Williams, E. Cartier Esham, and R. Patrick Lowe. Effect of human development on bacteriological water quality in coastal watersheds. *Ecological Applications*, 10(4) :1047–1056, 2000. doi : 10.1890/1051-0761(2000)010[1047:EOHDOB]2.0.CO;2. URL https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1890/1051-0761%282000%29010%5B1047%3AEOHDOB%5D2.0.CO%3B2.

Libu Manjakkal, Srinjoy Mitra, Yvan R Petillot, Jamie Shutler, E Marian Scott, Magnus Willander, and Ravinder Dahiya. Connected sensors, innovative sensor deployment, and intelligent data analysis for online water quality monitoring. *IEEE Internet of Things Journal*, 8(18) : 13805–13824, 2021.

Peter A Maraccini, Mia Catharine M Mattioli, Lauren M Sassoubre, Yiping Cao, John F Griffith, Jared S Ervin, Laurie C Van De Werfhorst, and Alexandria B Boehm. Solar inactivation of enterococci and escherichia coli in natural waters : effects of water absorbance and depth. *Environmental science & technology*, 50(10) :5068–5076, 2016.

F Eduardo Martinez and Andrew J Hooper. Drowning and immersion injury. *Anaesthesia & Intensive Care Medicine*, 15(9) :420–423, 2014.

Mia Catharine Mattioli, Lauren M Sassoubre, Todd L Russell, and Alexandria B Boehm. Decay of sewage-sourced microbial source tracking markers and fecal indicator bacteria in marine waters. *Water Research*, 108 :106–114, 2017.

Graham B McBride, Judith L McWhirter, and Matthew H Dalgety. Uncertainty in most probable number calculations for microbiological assays. *Journal of AOAC International*, 86(5) : 1084–1088, 2003.

David Thomas McCarthy, Ana Deletic, Valerie Grace Mitchell, Timothy David Fletcher, and Clare Diaper. Uncertainties in stormwater e. coli levels. *Water Research*, 42(6-7) :1812–1824, 2008.

Scott J McGrane. Impacts of urbanisation on hydrological and water quality dynamics, and urban water management : a review. *Hydrological Sciences Journal*, 61(13) :2295–2311, 2016.

Jean ET McLain, Channah M Rock, Kathleen Lohse, and James Walworth. False-positive identification of escherichia coli in treated municipal wastewater and wastewater-irrigated soils. *Canadian journal of microbiology*, 57(10) :775–784, 2011.

Cynthia L Meays, Klaas Broersma, Rick Nordin, and Asit Mazumder. Source tracking fecal bacteria in water : a critical review of current methods. *Journal of environmental management*, 73(1) :71–79, 2004.

Kais Mekki, Eddy Bajic, Frederic Chaxel, and Fernand Meyer. A comparative study of LPWAN technologies for large-scale IoT deployment. *ICT express*, 5(1) :1–7, 2019.

LA Méndez-Barroso, JA Rivas-Márquez, I Sosa-Tinoco, and A Robles-Morúa. Design and implementation of a low-cost multiparameter probe to evaluate the temporal variations of water quality conditions on an estuarine lagoon system. *Environmental Monitoring and Assessment*, 192(11) :710, 2020.

Patricia Menon, Gilles Billen, and Pierre Servais. Mortality rates of autochthonous and fecal bacteria in natural aquatic ecosystems. *Water research*, 37(17) :4151–4158, 2003.

Jimmy Millican and Larry Hauck. Texas commission on environmental quality. 2008.

Domenica Mirauda, Donatella Caniani, Maria Teresa Colucci, and Marco Ostoich. Assessing the fluvial system resilience of the river bacchiglione to point sources of pollution in northeast italy : a novel water resilience index (wri) approach. *Environmental Science and Pollution Research*, 28(27) :36775–36792, 2021.

Jean-Marie Mouchel, Françoise S Lucas, Laurent Moulin, Sébastien Wurtzer, Agathe Euzen, Jean-Paul Haghe, Vincent Rocher, Sam Azimi, and Pierre Servais. Bathing activities and microbiological river water quality in the paris area : A long-term perspective. *The Seine River Basin, the handbook of environmental chemistry*, 90 :323–354, 2020.

Laurent Moulin, Fanny Richard, Sabrina Stefania, Marion Goulet, Sylvie Gosselin, Alexandre Gonçalves, Vincent Rocher, Catherine Paffoni, and Aurélien Dumètre. Contribution of treated wastewater to the microbiological quality of seine river in paris. *water research*, 44 (18) :5222–5231, 2010.

Julia Moutiez. Se baigner à nouveau dans la seine : l'héritage promis par les jeux olympiques et paralympiques de paris 2024. *Projets de paysage. Revue scientifique sur la conception et l'aménagement de l'espace*, (25), 2021.

Cristian William Moyón Rivera and Dayana Karina Ordóñez Berrones. Construcción de un prototipo de red de nodos inteligentes para supervisar la calidad y niveles del agua potable en los tanques de reserva de ep-emapar. B.S. thesis, Escuela Superior Politécnica de Chimborazo, 2019.

Claire M Murphy, Daniel L Weller, Reza Ovissipour, Renee Boyer, and Laura K Strawn. Spatial versus nonspatial variance in fecal indicator bacteria differs within and between ponds. *Journal of Food Protection*, 86(3) :100045, 2023.

I Muslea, S Minton, and C. A Knoblock. Active learning with multiple views. *The Journal of artificial intelligence research*, 27 :203–233, 2006. ISSN 1076-9757.

Hans-Joachim Mälzer, Tim aus der Beek, Silke Müller, and Jörg Gebhardt. Comparison of different model approaches for a hygiene early warning system at the lower ruhr river, germany. *International journal of hygiene and environmental health*, 219(7) :671–680, 2016. ISSN 1438-4639.

Nabila Nafsin and Jin Li. Prediction of total organic carbon and e. coli in rivers within the milwaukee river basin using machine learning methods. *Environmental Science : Advances*, 2(2) :278–293, 2023.

Paty Nakhle, Laurie Boithias, Anne Pando-Bahuon, Chanthamousone Thammahacksa, Nicolas Gallion, Phabvilay Sounyafong, Norbert Silvera, Keooudone Latsachack, Bounsamay Soulileuth, Emma J Rochelle-Newall, et al. Decay rate of escherichia coli in a mountainous tropical headwater wetland. *Water*, 13(15) :2068, 2021.

Manel Naloufi, Françoise S Lucas, Sami Souihi, Pierre Servais, Aurélie Janne, and Thiago Wanderley Matos De Abreu. Evaluating the performance of machine learning approaches to predict the microbial quality of surface waters and to optimize the sampling effort. *Water*, 13(18) :2457, 2021.

Méry Ndione. *Dynamique et identification des sources de contamination fécale dans un espace littoral connaissant des pratiques de tourisme et de loisirs : l'exemple de la baie d'Aytré*. PhD thesis, Université de La Rochelle, 2022.

Meredith B. Nevers and Richard L. Whitman. Nowcast modeling of escherichia coli concentrations at multiple urban beaches of southern lake michigan. *Water Research*, 39(20) :5250–5260, 2005. ISSN 0043-1354. doi : https://doi.org/10.1016/j.watres.2005.10.012. URL `https://www.sciencedirect.com/science/article/pii/S0043135405005841`.

Phu Nguyen, Nicolas Ferry, Gencer Erdogan, Hui Song, Stéphane Lavirotte, Jean-Yves Tigli, and Arnor Solberg. Advances in deployment and orchestration approaches for IoT-a systematic review. In *2019 IEEE International Congress on Internet of Things (ICIOT)*, pages 53–60. IEEE, 2019.

Daniel Ekane Nnane, James Edward Ebdon, and Huw David Taylor. Integrated analysis of water quality parameters for cost-effective faecal pollution management in river catchments. *Water research (Oxford)*, 45(6) :2235–2246, 2011. ISSN 0043-1354.

Segev Noam. *Transfer Learning Using Decision Forests*. Technion - Computer Science Department, 2016.

Rachel T Noble, Ioannice M Lee, and Kenneth C Schiff. Inactivation of indicator micro-organisms from various sources of faecal contamination in seawater and freshwater. *Journal of applied microbiology*, 96(3) :464–472, 2004.

Rachel T Noble, A Denene Blackwood, John F Griffith, Charles D McGee, and Stephen B Weisberg. Comparison of rapid quantitative pcr-based and conventional culture-based methods for enumeration of enterococcus spp. and escherichia coli in recreational waters. *Applied and environmental microbiology*, 76(22) :7437–7443, 2010.

A Noury, P Peloux, and D ALBA. *Sites de baignade en Seine et en Marne héritage JO Paris 2024*. APUR, 2018.

Kane L Offenbaume, Edoardo Bertone, and Rodney A Stewart. Monitoring approaches for faecal indicator bacteria in water : Visioning a remote real-time sensor for e. coli and enterococci. *Water*, 12(9) :2591, 2020.

Leslie Ogorzaly, Isabelle Bertrand, Myriam Paris, Armand Maul, and Christophe Gantzer. Occurrence, survival, and persistence of human adenoviruses and f-specific rna phages in raw groundwater. *Applied and Environmental Microbiology*, 76(24) :8019–8025, 2010.

David M Oliver, Kenneth DH Porter, A Louise Heathwaite, Ting Zhang, and Richard S Quilliam. Impact of low intensity summer rainfall on e. coli-discharge event dynamics with reference to sample acquisition and storage. *Environmental monitoring and assessment*, 187 :1–13, 2015.

OMS. Organisation mondiale de la santé. `https://www.who.int/docs/default-source/wash-documents/who-recommendations-on-ec-bwd-august-2018.pdf`, 2018. Accessed on 16 July 2021.

Gregory D O'Mullan, M Elias Dueker, and Andrew R Juhl. Challenges to managing microbial fecal pollution in coastal environments : extra-enteric ecology and microbial exchange among water, sediment, and air. *Current Pollution Reports*, 3 :1–16, 2017.

Y. A. Pachepsky, A. Allende, L. Boithias, K. Cho, R. Jamieson, N. Hofstra, and M. Molina. Microbial water quality : Monitoring and modeling. *Journal of Environmental Quality*, 47 (5) :931–938, 2018.

LS Pakasi. Health risks associated with recreational water activities. In *IOP Conference Series : Materials Science and Engineering*, volume 434, page 012329. IOP Publishing, 2018.

Annik Pardailhé-Galabrun. Les déplacements des parisiens dans la ville aux xviième et xviiième siecles : Un essai de problématique. *Histoire, économie et société*, pages 205–253, 1983.

Soohyun Park, Soyi Jung, Haemin Lee, Joongheon Kim, and Jae-Hyun Kim. Large-scale water quality prediction using federated sensing and learning : A case study with real-world sensing big-data. *Sensors*, 21(4), 2021. doi : 10.3390/s21041462.

Yujin Park, Se-Rin Park, Sang-Woo Lee, and Junga Lee. Impacts of watershed and meteorological characteristics on stream water quality resilience. *Journal of Hydrology*, page 132663, 2025.

Adam M Paruch and Trond Mæhlum. Specific features of escherichia coli that distinguish it from coliform and thermotolerant coliform bacteria and define it as the most accurate indicator of faecal contamination in the environment. *Ecological Indicators*, 23 :140–142, 2012.

Julien Passerat, Nouho Koffi Ouattara, Jean-Marie Mouchel, Vincent Rocher, and Pierre Servais. Impact of an intense combined sewer overflow event on the microbiological water quality of the seine river. *Water research*, 45(2) :893–903, 2011.

MA Paule-Mercado, JS Ventura, SA Memon, D Jahng, J-H Kang, and C-H Lee. Monitoring and predicting the fecal indicator bacteria concentrations from agricultural, mixed land use and urban stormwater runoff. *Science of the Total Environment*, 550 :1171–1181, 2016.

Pierre Payment and Annie Locas. Pathogens in water : value and limits of correlation with microbial indicators. *Groundwater*, 49(1) :4–11, 2011.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, 12 :2825–2830, 2011.

David Pendergrass, Anne McFarland, and Larry Hauck. Instream bacteria influences from bird habitation of bridges. *JAWRA Journal of the American Water Resources Association*, 51 (6) :1519–1533, 2015.

Lin Peng, Huan Wu, Min Gao, Hualing Yi, Qingyu Xiong, Linda Yang, and Shuiping Cheng. Tlt : Recurrent fine-tuning transfer learning for water quality long-term prediction. *Water Research*, 225 :119171, 2022.

Pierluigi Penna, Elisa Baldrighi, Mattia Betti, Luigi Bolognini, Alessandra Campanelli, Samuela Capellacci, Silvia Casabianca, Christian Ferrarin, Giordano Giuliani, Federica Grilli, et al.

Water quality integrated system : A strategic approach to improve bathing water management. *Journal of Environmental Management*, 295 :113099, 2021.

Fritz Petersen and Jason A Hubbart. Physical factors impacting the survival and occurrence of escherichia coli in secondary habitats. *Water*, 12(6) :1796, 2020.

Adrian I Petrariu, Alexandru Lavric, and Eugen Coca. Lorawan gateway : Design, implementation and testing in real environment. In *2019 IEEE 25th International Symposium for Design and Technology in Electronic Packaging (SIITME)*, pages 49–53. IEEE, 2019.

G Petrucci and I Vaury, E andPoulleau. Identification de sites potentiels de baignade en marne, au regard de la qualité bactériologique et des travaux d'assainissement à réaliser. *Prolog ingenierie, Semeru*, page 229, 2018.

pH V2. Gravity : Analog ph sensor meter kit v2. `https://wiki.dfrobot.com/Gravity__Analog_pH_Sensor_Meter_Kit_V2_SKU_SEN0161-V2`, 2023a. Accessed : November 2023.

pH V2. Analog ph meter pro. `https://wiki.dfrobot.com/Analog_pH_Meter_Pro_SKU_SEN0169`, 2023b. Accessed : November 2023.

Gregory S Piorkowski, Rob C Jamieson, Lisbeth Truelstrup Hansen, Greg S Bezanson, and Chris K Yost. Characterizing spatial structure of sediment e. coli populations to inform sampling design. *Environmental monitoring and assessment*, 186 :277–291, 2014.

PME. Minidot logger : Dissolved oxygen sensor for oceanography. `https://www.cascoantiguopro.com/fr/oceanographie/capteurs-oceanographie/minidot-logger.html`, 2023. Accessed : May 2023.

Kathy Pond. Water recreation and disease : plausibility of associated infections : acute effects, sequelae, and mortality. 2005.

Michel Poulin, Pierre Servais, Jean-Marie Mouchel, Claire Thérial, Ludivine Lesage, Vincent Rocher, Alexandre Goncalves, Sophie Masnada, Françoise Lucas, Nicolas Flipo, et al. Modélisation de la contamination fécale en seine : impact des rejets de temps de pluie. *Programme PIREN-Seine Rapport Modélisation de La Contamination Fécale Par Temps de Pluie*, 2013.

Benoit Prevost, Françoise S Lucas, Alexandre Goncalves, Fanny Richard, Laurent Moulin, and Sébastien Wurtzer. Large scale survey of enteric viruses in river and waste water underlines the health status of the local population. *Environment international*, 79 :42–50, 2015.

Rahul Priyadarshi, Bharat Gupta, and Amulya Anurag. Deployment techniques in wireless sensor networks : a survey, classification, challenges, and future research issues. *The Journal of Supercomputing*, pages 1–41, 2020.

Antonis Protopsaltis, Panagiotis Sarigiannidis, Dimitrios Margounakis, and Anastasios Lytos. Data visualization in internet of things : tools, methodologies, and challenges. In *Proceedings of the 15th International Conference on Availability, Reliability and Security*, pages 1–11, 2020.

Annette Prüss. Review of epidemiological studies on health effects from exposure to recreational water. *International journal of epidemiology*, 27(1) :1–9, 1998.

Mompoloki Pule, Abid Yahya, and Joseph Chuma. Wireless sensor networks : A survey on monitoring water quality. *Journal of applied research and technology*, 15(6) :562–570, 2017.

Pengjiang Qian, Yangyang Chen, Jung-Wen Kuo, Yu-Dong Zhang, Yizhang Jiang, Kaifa Zhao, Rose Al Helo, Harry Friel, Atallah Baydoun, Feifei Zhou, Jin Uk Heo, Norbert Avril, Karin Herrmann, Rodney Ellis, Bryan Traughber, Robert S. Jones, Shitong Wang, Kuan-Hao Su, and Raymond F. Muzic. mdixon-based synthetic ct generation for pet attenuation correction on abdomen and pelvis jointly using transfer fuzzy clustering and active learning-based classification. *IEEE Transactions on Medical Imaging*, 39(4) :819–832, 2020.

Hua-Peng Qin, Qiong Su, and Soon-Thiam Khu. An integrated model for water management in a rapidly urbanizing catchment. *Environmental modelling & software*, 26(12) :1502–1514, 2011.

Xueheng Qiu, Ye Ren, Ponnuthurai Nagaratnam Suganthan, and Gehan A.J Amaratunga. Empirical mode decomposition based ensemble deep learning for load demand time series forecasting. *Applied soft computing*, 54 :246–255, 2017. ISSN 1568-4946.

Richard S Quilliam, Katie Clements, Caroline Duce, Simon B Cottrill, Shelagh K Malham, and Davey L Jones. Spatial variation of waterborne escherichia coli–implications for routine water quality monitoring. *Journal of water and health*, 9(4) :734–737, 2011.

R-Core-Team. R : A language and environment for statistical computing. r foundation for statistical computing, vienna, austria. `https://www.R-project.org/`, 2018.

R Core Team. R : A language and environment for statistical computing. `https://www.R-project.org/`, 2021.

Sharyl JM Rabinovici, Richard L Bernknopf, Anne M Wein, Don L Coursey, and Richard L Whitman. Economic and health risk trade-offs of swim closures at a lake michigan beach, 2004.

Hamed Rahimi, Ali Zibaeenejad, and Ali Akbar Safavi. A novel IoT architecture based on 5G-IoT and next generation technologies. In *In Proceedings of the 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON),Vancouver, Canada*, pages 81–88, 2018.

Mushtaque Ahmed Rahu, Muhammad Mujtaba Shaikh, Sarang Karim, Sarfaraz Ahmed Soomro, Deedar Hussain, and Sayed Mazhar Ali. Water quality monitoring and assessment for efficient water resource management through internet of things and machine learning approaches for agricultural irrigation. *Water Resources Management*, pages 1–42, 2024.

Maneesha V Ramesh, KV Nibi, Anupama Kurup, Renjith Mohan, A Aiswarya, A Arsha, and PR Sarang. Water quality monitoring and waste management using iot. In *IEEE Global Humanitarian Technology Conference (GHTC), San Jose, CA, USA*, pages 1–7, 2017.

MA RANDIKA, AAGD AMARASOORIYA, and SK WERAGODA. Development of an arduino-based low-cost turbidity and electric conductivity meter for wastewater characterization. *LARHYSS Journal P-ISSN 1112-3680/E-ISSN 2521-9782*, (51) :115–127, 2022.

Garreta Raul. *Scikit-learn : machine learning simplified*. Packt Publishing, Birmingham, England, 2017. ISBN 978-17-8883-152-9.

Michael Rode, Andrew J. Wade, Matthew J. Cohen, Robert T. Hensley, Michael J. Bowes, James W. Kirchner, George B. Arhonditsis, Phil Jordan, Brian Kronvang, Sarah J. Halliday, Richard A. Skeffington, Joachim C. Rozemeijer, Alice H. Aubert, Karsten Rinke, and Seifeddine Jomaa. Sensors in the stream : The high-frequency wave of the present. *Environmental Science & Technology*, 50(19) :10297–10307, 2016.

Adélaïde Roguet. *Characterization of anthropogenic and environmental pressures influencing the bacterial compartment in shallow lakes*. PhD thesis, Université Paris-Est, 2015.

Timothy J Ross. *Fuzzy logic with engineering applications*. John Wiley & Sons, 2005.

Alessandra Rossi, Bernabas T Wolde, Lee H Lee, and Meiyin Wu. Prediction of recreational water safety using escherichia coli as an indicator : case study of the passaic and pompton rivers, new jersey. *Science of the Total Environment*, 714 :136814, 2020.

Gaële Rouillé-Kielo and Gabrielle Bouleau. *Rendre les cours d'eau urbains baignables, une comparaison Paris-Berlin*. PhD thesis, PIREN Seine phase 8, 2021.

Paweł M Rowiński, Tomasz Okruszko, and Artur Radecki-Pawlik. Environmental hydraulics research for river health : recent advances and challenges. *Ecohydrology & Hydrobiology*, 22(2) :213–225, 2022.

Imam Abdul Rozaq, Noor Yulita Dwi Setyaningsih, and Bud Gunawan. Pengkondisian sinyal sensor salinitas dfr0300 menggunakan arduino due. 2020.

Stefania Russo, Moritz Lürig, Wenjin Hao, Blake Matthews, and Kris Villez. Active learning for anomaly detection in environmental data. *Environmental Modelling & Software*, 134 : 104869, 2020.

Hodon Ryu, John F Griffith, Izhar UH Khan, Stephen Hill, Thomas A Edge, Carlos Toledo-Hernandez, Joel Gonzalez-Nieves, and Jorge Santo Domingo. Comparison of gull feces-specific assays targeting the 16s rrna genes of catellicoccus marimammalium and streptococcus spp. *Applied and Environmental Microbiology*, 78(6) :1909–1916, 2012.

Vasit Sagan, Kyle T Peterson, Maitiniyazi Maimaitijiang, Paheding Sidike, John Sloan, Benjamin A Greeling, Samar Maalouf, and Craig Adams. Monitoring inland water quality using remote sensing : Potential and limitations of spectral indices, bio-optical simulations, machine learning, and cloud computing. *Earth-Science Reviews*, 205 :103187, 2020.

Sajal Saha, Rakibul Hasan Rajib, and Sumaiya Kabir. Iot based automated fish farm aquaculture monitoring system. In *2018 International Conference on Innovations in Science, Engineering and Technology (ICISET)*, pages 201–206. IEEE, 2018.

Abdul Salam. *Internet of Things for Water Sustainability*, pages 113–145. 01 2020. ISBN 978-3-030-35290-5. doi : 10.1007/978-3-030-35291-2_4.

Sadia Salam, Rachel McDaniel, Bruce Bleakley, Louis Amegbletor, and Sara Mardani. Variability of e. coli in streambed sediment and its implication for sediment sampling. *Journal of Contaminant Hydrology*, 242 :103859, 2021.

Alissa K Salmore, Erika J Hollis, and Sandra L McLellan. Delineation of a chemical and biological signature for stormwater pollution in an urban river. *Journal of water and health*, 4(2) :247–262, 2006.

Reynee W Sampson, Sarah A Swiatnicki, Vicki L Osinga, Jamie L Supita, Colleen M McDer-

mott, and GTl Kleinheinz. Effects of temperature and sand on e. coli survival in a northern lake water microcosm. *Journal of Water and Health*, 4(3) :389–393, 2006.

Randy Erfa Saputra, Budhi Irawan, and Yakub Eka Nugraha. System design and implementation automation system of expert system on hydroponics nutrients control using forward chaining method. In *2017 IEEE Asia Pacific Conference on Wireless and Mobile (APWiMob)*, pages 41–46. IEEE, 2017.

Monika Schaffner, Hans-Peter Bader, and Ruth Scheidegger. Modeling the contribution of point sources and non-point sources to thachin river water pollution. *Science of the Total Environment*, 407(17) :4902–4915, 2009.

FM Schets, AM De Roda Husman, and AH Havelaar. Disease outbreaks associated with untreated recreational water use. *Epidemiology & Infection*, 139(7) :1114–1125, 2011.

JF Schijven and AM de Roda Husman. Effect of climate changes on waterborne disease in the netherlands. *Water Science and Technology*, 51(5) :79–87, 2005.

Siegfried Schloissnig, Manimozhiyan Arumugam, Shinichi Sunagawa, Makedonka Mitreva, Julien Tap, Ana Zhu, Alison Waller, Daniel R Mende, Jens Roat Kultima, John Martin, et al. Genomic variation landscape of the human gut microbiome. *Nature*, 493(7430) : 45–50, 2013.

Christiane Schreiber, Andrea Rechenburg, Esther Rind, and Thomas Kistemann. The impact of land use on microbial surface water pollution. *International journal of hygiene and environmental health*, 218(2) :181–187, 2015.

Claudia Schultz-Fademrecht, Marc Wichern, and Harald Horn. The impact of sunlight on inactivation of indicator microorganisms both in river water and benthic biofilms. *Water research*, 42(19) :4771–4779, 2008.

Leela Sedaghat, John Hersey, and Michael P McGuire. Detecting spatio-temporal outliers in crowdsourced bathymetry data. In *Proceedings of the Second ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information*, pages 55–62, 2013.

Noam Segev, Maayan Harel, Shie Mannor, Koby Crammer, and Ran El-Yaniv. Learn on source, refine on target : A model transfer learning framework with random forests. *CoRR*, abs/1511.01258, 2015.

Wolfgang Seis, Malte Zamzow, Nicolas Caradot, and Pascale Rouault. On the implementation of reliable early warning systems at european bathing waters using multivariate bayesian regression modelling. *Water Research*, 143 :301–312, 2018.

Patrick Sejkora, Mary Jo Kirisits, and Michael Barrett. Colonies of cliff swallows on highway bridges : A source of escherichia coli in surface waters 1. *JAWRA Journal of the American Water Resources Association*, 47(6) :1275–1284, 2011.

K Chandra Sekhar, Byna Venkatesh, Korchipati Sudeep Reddy, Guntimadugu Giridhar, Koppulu Nithin, and Kinnera Eshwar. Iot-based realtime water quality management system using arduino microcontroller. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, 14(2) :783–792, 2023.

Sandra Sendra, Lorena Parra, Jose M Jimenez, Laura Garcia, and Jaime Lloret. Lora-based network for water quality monitoring in coastal areas. *Mobile Networks and Applications*, 28(1) :65–81, 2023.

Mustapha Reda Senouci and Abdelhamid Mellouk. *Deploying wireless sensor networks : theory and practice*. Elsevier, 2016.

Pierre Servais, Martin Seidl, and Jean-Marie Mouchel. Comparison of parameters characterizing organic matter in a combined sewer during rainfall events and dry weather. *Water Environment Research*, 71(4) :408–417, 1999.

Pierre Servais, Tamara Garcia-Armisen, Anne Sophie Lepeuple, and Philippe Lebaron. An early warning method to detect faecal contamination of river waters. *Annals of Microbiology*, 55(2) :151–156, 2005.

Pierre Servais, Gilles Billen, A Goncalves, and Tamara Garcia-Armisen. Modelling microbiological water quality in the seine river drainage network : past, present and future situations. *Hydrology and Earth System Sciences*, 11(5) :1581–1592, 2007a.

Pierre Servais, Tamara Garcia-Armisen, Isabelle George, and Gilles Billen. Fecal bacteria in the rivers of the seine drainage network (france) : sources, fate and modelling. *Science of the Total Environment*, 375(1-3) :152–167, 2007b.

Muhammad Shahid Iqbal, Luo Bin, Tamoor Khan, Rashid Mehmood, and Muhammad Sadiq. Heterogeneous transfer learning techniques for machine learning. *Iran Journal of Computer Science*, 1, 04 2018. doi : 10.1007/s42044-017-0004-z.

Meixia Shi, Jingbo Ma, and Kai Zhang. The impact of water temperature on in-line turbidity detection. *Water*, 14(22), 2022. ISSN 2073-4441. doi : 10.3390/w14223720. URL https://www.mdpi.com/2073-4441/14/22/3720.

Pramila P Shinde and Seema Shah. A review of machine learning and deep learning applications. In *2018 Fourth international conference on computing communication control and automation (ICCUBEA)*, pages 1–6. IEEE, 2018.

D. Shrestha and D. Solomatine. Experiments with adaboost.rt, an improved boosting scheme for regression. *Neural Computation*, 18 :1678–1710, 2006.

Hillel Shuval. Estimating the global burden of thalassogenic diseases : human infectious diseases caused by wastewater pollution of the marine environment. *Journal of water and health*, 1 (2) :53–64, 2003.

Nuno Silva, Gilberto Igrejas, Alexandre Gonçalves, and Patricia Poeta. Commensal gut bacteria : distribution of enterococcus species and prevalence of escherichia coli phylogenetic groups in animals and humans in portugal. *Annals of microbiology*, 62 :449–459, 2012.

Joyce M Simpson, Jorge W Santo Domingo, and Donald J Reasoner. Microbial source tracking : state of the science. *Environmental science & technology*, 36(24) :5279–5288, 2002.

Lester W Sinton, Rochelle K Finlay, and Philippa A Lynch. Sunlight inactivation of fecal bacteriophages and bacteria in sewage-polluted seawater. *Applied and environmental microbiology*, 65(8) :3605–3613, 1999.

Ekaterina Sokolova, Oscar Ivarsson, Ann Lilliesträm, Nora K Speicher, Henrik Rydberg, and Mia Bondelind. Data-driven models for predicting microbial water quality in the drinking water source using e. coli monitoring and hydrometeorological data. *Science of the Total Environment*, 802 :149798, 2022.

Jeffrey A Soller, Mary E Schoen, Timothy Bartrand, John E Ravenscroft, and Nicholas J Ashbolt. Estimated human health risks from exposure to recreational waters impacted by human and non-human sources of faecal contamination. *Water research*, 44(16) :4674–4691, 2010.

Helena M Solo-Gabriele, Melinda A Wolfert, Timothy R Desmarais, and Carol J Palmer. Sources of escherichia coli in a coastal subtropical environment. *Applied and environmental microbiology*, 66(1) :230–237, 2000.

Christopher Staley, Gary M Dunny, and Michael J Sadowsky. Environmental and animal-associated enterococci. *Advances in applied microbiology*, 87 :147–186, 2014.

Ian Stewart, Penelope M Webb, Philip J Schluter, and Glen R Shaw. Recreational and occupational field exposure to freshwater cyanobacteria–a review of anecdotal and case reports, epidemiological studies and the challenges for epidemiologic assessment. *Environmental Health*, 5 :1–13, 2006.

Curtis H Stumpf, Michael F Piehler, Suzanne Thompson, and Rachel T Noble. Loading of fecal indicator bacteria in north carolina tidal creek headwaters : hydrographic patterns and terrestrial runoff relationships. *Water Research*, 44(16) :4704–4715, 2010.

Arief Dhany Sutadian, Nitin Muttil, Abdullah Gokhan Yilmaz, and BJC Perera. Development of river water quality indices—a review. *Environmental monitoring and assessment*, 188 : 1–29, 2016.

Scott Sutton. Accuracy of plate counts. *Journal of validation technology*, 17(3) :42–46, 2011.

Philip H. Swain and Hans Hauska. The decision tree classifier : Design and potential. *IEEE Transactions on Geoscience Electronics*, 15(3) :142–147, 1977.

Émile Sylvestre, Jean-Baptiste Burnet, Patrick Smeets, Gertjan Medema, Michèle Prévost, and Sarah Dorner. Can routine monitoring of e. coli fully account for peak event concentrations at drinking water intakes in agricultural and urban rivers ? *Water research (Oxford)*, 170 : 115369–115369, 2020. ISSN 0043-1354.

Karolina Tatari, Charlotte B. Corfitzen, Hans-Jørgen Albrechtsen, and Sarah Christine Boesgaard Christensen. *Sensors for microbial drinking water quality*. Technical University of Denmark, DTU Environment, 2016.

Marjolijn Tijdens, Dedmer B Van de Waal, Hana Slovackova, Hans L Hoogveld, and Herman J Gons. Estimates of bacterial and phytoplankton mortality caused by viral lysis and microzooplankton grazing in a shallow eutrophic lake. *Freshwater Biology*, 53(6) :1126–1141, 2008.

Ananda Tiwari, Seppo I Niemelä, Asko Vepsäläinen, Jarkko Rapala, Seija Kalso, and Tarja Pitkänen. Comparison of colilert-18 with miniaturised most probable number method for monitoring of escherichia coli in bathing water. *Journal of Water and Health*, 14(1) : 121–131, 2016.

James Topping. *Errors of Observation and their Treatment*, volume 62. Springer Science & Business Media, 2012.

Andreas Tornevi, Olof Bergstedt, and Bertil Forsberg. Precipitation effects on microbial pollution in a river : lag structures and seasonal effect modification. *PloS one*, 9(5) :e98546, 2014.

Jarrod Trevathan, Wayne Read, and Simon Schmidtke. Towards the development of an affordable and practical light attenuation turbidity sensor for remote near real-time aquatic monitoring. *Sensors*, 20(7), 2020. ISSN 1424-8220. doi : 10.3390/s20071993. URL `https://www.mdpi.com/1424-8220/20/7/1993`.

Jarrod Trevathan, Simon Schmidtke, Wayne Read, Tony Sharp, and Abdul Sattar. An iot general-purpose sensor board for enabling remote aquatic environmental monitoring. *Internet of Things*, 16 :100429, 2021.

Ingun Tryland, Fasil Ejigu Eregno, Henrik Braathen, Goran Khalaf, Ingrid Sjølander, and Marie Fossum. On-line monitoring of Escherichia coli in raw water at Oset drinking water treatment plant, Oslo (Norway). *International journal of environmental research and public health*, 12(2) :1788–1802, 2015. ISSN 1660-4601.

Athina Tsanousa, Vasileios-Rafail Xefteris, Georgios Meditskos, Stefanos Vrochidis, and Ioannis Kompatsiaris. Combining rssi and accelerometer features for room-level localization. *Sensors*, 21(8), 2021. ISSN 1424-8220. doi : 10.3390/s21082723. URL `https://www.mdpi.com/1424-8220/21/8/2723`.

Shahadat Uddin, Arif Khan, Md Ekramul Hossain, and Mohammad Ali Moni. Comparing different supervised machine learning algorithms for disease prediction. *BMC medical informatics and decision making*, 19(1) :281–281, 2019. ISSN 1472-6947.

Lan-Anh Van, Kim-Dan Nguyen, François Le Marrec, and Aïcha Jairy. Development of a tool for modeling the fecal contamination in rivers with turbulent flows—application to the seine et marne rivers (parisian region, france). *Water*, 14(8) :1191, 2022.

Ilse A van Asperen, Gertjan Medema, Martien W Borgdorff, Macr JW Sprenger, and Arie H Havelaar. Risk of gastroenteritis among triathletes in relation to faecal pollution of fresh waters. *International journal of epidemiology*, 27(2) :309–315, 1998. ISSN 0300-5771.

ES van der Meulen, A Tertienko, AN Blauw, NB Sutton, FHM van de Ven, HHM Rijnaarts, and PR van Oel. A review of prediction models for e. coli in urban surface waters. *Urban Water Journal*, pages 1–10, 2024.

Vladimir Naumovich Vapnik. *The nature of statistical learning theory*. Springer, New York, 1995 - 1995. ISBN 1-4757-2440-3.

J Vellingiri, K Kalaivanan, MP Gopinath, C Gobinath, Prabhakar Rontala Subramaniam, and Sarathkumar Rangarajan. Strategies for classifying water quality in the cauvery river using a federated learning technique. *International Journal of Cognitive Computing in Engineering*, 4 :187–193, 2023.

Tony Venelinov. Comparison of iso 21748 and iso 11352 standards for measurement uncertainty estimation in water analysis. *Ovidius University Annals Series : Civil Engineering*, 18 : 187–192, 2016.

Valérie Villeneuve, Stéphane Légaré, Jean Painchaud, and Warwick Vincent. Dynamique et modélisation de l'oxygène dissous en rivière. *Revue des sciences de l'eau*, 19(4) :259–274, 2006.

Hans Visser, Niels Evers, Arjan Bontsema, Jasmijn Rost, Arie de Niet, Paul Vethman, Sido Mylius, Annelotte van der Linden, Joost van den Roovaart, Frank van Gaalen, et al. What drives the ecological quality of surface waters ? a review of 11 predictive modeling tools. *Water Research*, 208 :117851, 2022.

Lucie Vondrakova, Jarmila Pazlarova, and Katerina Demnerova. Detection, identification and quantification of campylobacter jejuni, coli and lari in food matrices all at once using multiplex qpcr. *Gut pathogens*, 6 :1–9, 2014.

Timothy J Wade, Nitika Pai, Joseph NS Eisenberg, and John M Colford Jr. Do us environmental protection agency water quality guidelines for recreational waters prevent gastrointestinal illness ? a systematic review and meta-analysis. *Environmental health perspectives*, 111 (8) :1102–1109, 2003.

Xi Wang and Chen Wang. Time series data cleaning : A survey. *Ieee Access*, 8 :1866–1881, 2019.

Xiaoping Wang, Fei Zhang, and Jianli Ding. Evaluation of water quality based on a machine learning algorithm and water quality index for the ebinur lake watershed, china. *Scientific reports*, 7(1) :12858, 2017.

Zhaohui Aleck Wang, Hassan Moustahfid, Amy V. Mueller, Anna P. M. Michel, Matthew Mowlem, Brian T. Glazer, T. Aran Mooney, William Michaels, Jonathan S. McQuillan, Julie C. Robidart, James Churchill, Marc Sourisseau, Anne Daniel, Allison Schaap, Sam Monk, Kim Friedman, and Patrice Brehmer. Advancing observation of ocean biogeochemistry,

biology, and ecosystems with cost-effective in situ sensing technologies. *Frontiers in Marine Science*, 6 :519, 2019a. ISSN 2296-7745. doi : 10.3389/fmars.2019.00519. URL https://www.frontiersin.org/article/10.3389/fmars.2019.00519.

Zirui Wang, Zihang Dai, Barnabás Póczos, and Jaime Carbonell. Characterizing and avoiding negative transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11293–11302, 2019b.

Chelsea J Weiskerger and Mantha S Phanikumar. Numerical modeling of microbial fate and transport in natural waters : Review and implications for normal and extreme storm events. *Water (Basel)*, 12(7) :1876, 2020. ISSN 2073-4441.

Chelsea J Weiskerger and Richard L Whitman. Monitoring e. coli in a changing beachscape. *Science of The Total Environment*, 619 :1236–1246, 2018.

Daniel L Weller, Tanzy MT Love, Alexandra Belias, and Martin Wiedmann. Predictive models may complement or provide an alternative to existing strategies for assessing the enteric pathogen contamination status of northeastern streams used to provide water for produce production. *Frontiers in sustainable food systems*, 4 :561517, 2020.

Michael J Whelan, Conor Linstead, Fred Worrall, Steve J Ormerod, Isabelle Durance, Andrew C Johnson, David Johnson, Mark Owen, Emma Wiik, Nicholas JK Howden, et al. Is water quality in british rivers "better than at any time since the end of the industrial revolution" ? *Science of the Total Environment*, 843 :157014, 2022.

MJ Whelan, A Ramos, R Villa, I Guymer, B Jefferson, and M Rayner. A new conceptual model of pesticide transfers from agricultural land to surface waters with a specific focus on metaldehyde. *Environmental Science : Processes & Impacts*, 22(4) :956–972, 2020.

Paul G Whitehead, Robert L Wilby, Richard W Battarbee, Martin Kernan, and Andrew John Wade. A review of the potential impacts of climate change on surface water quality. *Hydrological sciences journal*, 54(1) :101–123, 2009.

Richard L Whitman and Meredith B Nevers. Summer e. coli patterns and responses along 23 chicago beaches. *Environmental science & technology*, 42(24) :9217–9224, 2008.

WHO. World Health Organization. https://www.who.int/docs/default-source/wash-documents/who-recommendations-on-ec-bwd-august-2018.pdf, 2018. Accessed on 16 July 2021.

Jared D Willard, Jordan S Read, Alison P Appling, Samantha K Oliver, Xiaowei Jia, and Vipin Kumar. Predicting water temperature dynamics of unmonitored lakes with meta-transfer learning. *Water Resources Research*, 57(7) :e2021WR029579, 2021.

Timothy P Wilson, Cherie V Miller, and Evan A Lechner. Guidelines for the use of automatic samplers in collecting surface-water quality and sediment data. Technical report, US Geological Survey, 2024.

Yong Jie Wong, Rei Nakayama, Yoshihisa Shimizu, Akinori Kamiya, Shang Shen, Idlan Zarizi Muhammad Rashid, and Nik Meriam Nik Sulaiman. Toward industrial revolution 4.0 : Development, validation, and application of 3d-printed iot-based water quality monitoring system. *Journal of Cleaner Production*, 324 :129230, 2021. ISSN 0959-6526. doi : https://doi.org/10.1016/j.jclepro.2021.129230. URL `https://www.sciencedirect.com/science/article/pii/S0959652621034168`.

Mary E Wright, Helena M Solo-Gabriele, Samir Elmir, and Lora E Fleming. Microbial load from animal feces at a recreational beach. *Marine pollution bulletin*, 58(11) :1649–1656, 2009.

Jianyong Wu, Paula Rees, Sara Storrer, Kerri Alderisio, and Sarah Dorner. Fate and transport modeling of potential pathogens : The contribution from sediments 1. *JAWRA Journal of the American Water Resources Association*, 45(1) :35–44, 2009.

Yipeng Wu, Shuming Liu, and Zoran Kapelan. Addressing data limitations in leakage detection of water distribution systems : Data creation, data requirement reduction, and knowledge transfer. *Water Research*, page 122471, 2024.

Susanne Wuijts, Marit De Vries, Wilma Zijlema, Judith Hin, Lewis R Elliott, Liesbet Dirven-van Breemen, Enrico Scoccimarro, Ana Maria de Roda Husman, Mart Külvik, Ilias S Frydas, et al. The health potential of urban water : Future scenarios on local risks and opportunities. *Cities*, 125 :103639, 2022a.

Susanne Wuijts, Lieke Friederichs, Judith A Hin, Franciska M Schets, Helena FMW Van Rijswick, and Peter PJ Driessen. Governance conditions to overcome the challenges of realizing safe urban bathing water sites. *International Journal of Water Resources Development*, 38 (4) :554–578, 2022b.

Sebastien Wurtzer, Benoit Prevost, Francoise S Lucas, and Laurent Moulin. Detection of

enterovirus in environmental waters : a new optimized method compared to commercial real-time rt-qpcr kits. *Journal of virological methods*, 209 :47–54, 2014.

Mark D Wyer, David Kay, Huw Morgan, Sam Naylor, Simon Clark, John Watkins, Cheryl M Davies, Carol Francis, Hamish Osborn, and Sarah Bennett. Within-day variability in microbial concentrations at a uk designated bathing water : Implications for regulatory monitoring and the application of predictive modelling based on historical compliance data. *Water research X*, 1 :100006, 2018.

L Wymer, A Dufour, and C McGee. Temporal variability of microbial indicators of faecal contamination of marine and freshwater beaches. In *American Society for Microbiology General 107th Meeting, Toronto, CANADA, May*, 2007.

Fangnan Xiao, Huapeng Qin, and Taotao Sun. Effects of low impact development on runoff pollution and water quality resilience in an urbanized estuary area. *Journal of Hydrology*, 644 :132129, 2024.

Shu Xu, Bo Lu, Michael Baldea, Thomas F Edgar, Willy Wojsznis, Terrence Blevins, and Mark Nixon. Data cleaning in the process industries. *Reviews in Chemical Engineering*, 31(5) : 453–490, 2015.

Jianzhuo Yan, Ya Gao, Yongchuan Yu, Hongxia Xu, and Zongbao Xu. A prediction model based on deep belief network and least squares svr applied to cross-section water quality. *Water (Basel)*, 12(7) :1929, 2020. ISSN 2073-4441.

Jiawei Yang, Susanto Rahardja, and Pasi Fränti. Mean-shift outlier detection and filtering. *Pattern Recognition*, 115 :107874, 2021. ISSN 0031-3203. doi : https://doi.org/10.1016/j.patcog.2021.107874. URL `https://www.sciencedirect.com/science/article/pii/S0031320321000613`.

Irina Yaroshenko, Dmitry Kirsanov, Monika Marjanovic, Peter A Lieberzeit, Olga Korostynska, Alex Mason, Ilaria Frau, and Andrey Legin. Real-time water quality monitoring with chemical sensors. *Sensors*, 20(12) :3432, 2020.

Lotfi A Zadeh. Fuzzy sets. *Information and Control*, 1965.

Fang Zhou and Ting-Yu Chen. A hybrid group decision-making approach involving pythagorean fuzzy uncertainty for green supplier selection. *International Journal of Production Economics*, 261 :108875, 2023.

Junxing Zhu, Jiawei Zhang, Quanyuan Wu, Yan Jia, Bin Zhou, Xiaokai Wei, and Philip S Yu. Constrained active learning for anchor link prediction across multiple heterogeneous social networks. *Sensors (Basel, Switzerland)*, 17(8) :1786, 2017. ISSN 1424-8220.

Mengyuan Zhu, Jiawei Wang, Xiao Yang, Yu Zhang, Linyu Zhang, Hongqiang Ren, Bing Wu, and Lin Ye. A review of the application of machine learning in water quality evaluation. *Eco-Environment & Health*, 1(2) :107–116, 2022.

Qingchuan Zhu, Frederic Cherqui, and Jean-Luc Bertrand-Krajewski. End-user perspective of low-cost sensors for urban stormwater monitoring : a review. *Water Science & Technology*, 87(11) :2648–2684, 2023.

Amity G Zimmer-Faust, Vanessa Thulsiraj, Catalina Marambio-Jones, Yiping Cao, John F Griffith, Patricia A Holden, and Jennifer A Jay. Effect of freshwater sediment characteristics on the persistence of fecal indicator bacteria and genetic markers within a southern california watershed. *Water Research*, 119 :1–11, 2017.

Alireza Zourmand, Andrew Lai Kun Hing, Chan Wai Hung, and Mohammad AbdulRehman. Internet of things (iot) using lora technology. In *2019 IEEE international conference on automatic control and intelligent systems (I2CACIS)*, pages 324–330. IEEE, 2019.

Alain F Zuur, Elena N Ieno, Neil J Walker, Anatoly A Saveliev, Graham M Smith, et al. *Mixed effects models and extensions in ecology with R*, volume 574. Springer, 2009.

Résumé

Ce travail a permis de mettre en lumière les défis liés à la gestion de la qualité microbiologique des eaux de surface, en particulier dans des environnements fortement urbanisés. Les efforts se sont concentrés sur la mise en place d'approches innovantes, combinant des outils technologiques avancés, des modèles prédictifs robustes, et le développement de guides pratiques méthodologiques, pour répondre aux exigences croissantes de surveillance et de gestion de la qualité des eaux de surface. Nous avons développé une méthodologie intégrant des outils d'apprentissage automatique et des dispositifs de mesure en quasi temps réel pour la surveillance et la prédiction de la qualité de l'eau. Cette approche souligne le potentiel des réseaux de capteurs continus combinant des capteurs à bas coût et des capteurs de haute précision pour améliorer les prises de décision. Les tests et validations sur le terrain ont démontré la faisabilité et l'efficacité de ces dispositifs pour une gestion durable et précise. De plus, l'évaluation de l'incertitude, de l'échantillonnage à la mesure s'est révélée cruciale pour garantir la robustesse des données collectées. L'intégration de l'incertitude sur la mesure d'*E. coli* dans le processus de classement des échantillons à l'aide de la logique floue s'est également révélée être une approche intéressante pour améliorer la prise de décision pour l'ouverture ou la fermeture des sites de baignade. En complément une meilleure compréhension de la dynamique temporelle des pollutions microbiologiques est essentielle pour renforcer la surveillance et pour étudier la résistance ainsi que la résilience des sites de baignade face aux événements polluants liés au temps de pluie ou aux accidents sur le réseau d'assainissement. Ces approches ont pour objectif de diminuer le risque sanitaire lié à la baignade dans des eaux soumises à une forte pression anthropique.

Mots-clés : baignades, rivière urbaine, qualité microbiologique, contamination, *E. coli*, prédiction, incertitude, dynamique

Abstract

This work has highlighted the challenges associated with managing the microbiological quality of surface waters, particularly in highly urbanized environments. Efforts have focused on implementing innovative approaches that combine advanced technological tools, robust predictive models, and the development of practical methodological guidelines to meet the growing demands for surface water quality monitoring and management. We developed a methodology integrating machine learning tools and near real-time measurement devices for water quality monitoring and prediction. This approach underscores the potential of continuous sensor networks combining low-cost sensors with high-precision ones to enhance decision-making processes. Field tests and validations demonstrated the feasibility and effectiveness of these devices for sustainable and accurate management. Furthermore, evaluating the uncertainty from sampling to measurement proved crucial in ensuring the robustness of collected data. The integration of *E. coli* uncertainty into the sample classification process using fuzzy logic also emerged as a promising approach to improve decision-making regarding the opening or closing of bathing sites. Additionally, a better understanding of the temporal dynamics of microbiological pollution is essential for strengthening monitoring efforts and studying the resistance and resilience of bathing sites to pollution events caused by rainfall or accidents in the sanitation network. hese approaches aim to reduce the health risks associated with swimming in waters subjected to high anthropogenic pressure.

Keywords : bathing, urban river, microbiological quality, contamination, *E. coli*, prediction, uncertainty, dynamics