

Faculté des bioingénieurs

Correction a posteriori de la cartographie des polluants issue du modèle SIRANE

Techniques d'assimilation de données appliquées
sur la Région de Bruxelles-Capitale

Auteur
Promoteur
Lecteurs

GELLENS CLAIRE
BOGAERT PATRICK
BRASSEUR OLIVIER
VANCLOOSTER MARNIK

Année académique 2024 - 2025

Mémoire de fin d'études présenté en vue de l'obtention du diplôme de
Bioingénieur en Sciences et technologies de l'environnement

Remerciements

Je souhaite tout d'abord remercier toutes les personnes qui, de près ou de loin, ont contribué à la réalisation de ce mémoire. Rien de cela n'aurait été possible sans vous.

Je tiens à exprimer ma profonde gratitude au Pr. Patrick Bogaert, pour le temps qu'il m'a accordé, sa compréhension et son accompagnement constant. Ses conseils et remarques ont été d'une aide inestimable.

Un grand merci également à Axel Briffault, toujours présent lors de nos réunions hebdomadaires. Sa disponibilité, son accessibilité et son écoute ont rendu nos échanges particulièrement enrichissants et agréables.

Je suis aussi très reconnaissante envers Olivier Brasseur, dont l'expertise, les conseils et la bienveillance ont constitué un soutien précieux tout au long de ce travail.

Je remercie par ailleurs l'École Centrale de Lyon, à l'origine du développement de SIRANE et de nombreuses recherches sur la modélisation des polluants en milieu urbain. Merci aussi pour vos réponses et votre disponibilité face à nos questions.

Enfin, je n'aurai jamais assez de mots pour remercier mes parents : merci pour votre temps, vos relectures, votre soutien et votre accompagnement constant. Je ne sais que trop bien la chance que j'ai.

Table des matières

Table des figures	iii
Liste des tableaux	iv
Liste des acronymes	v
1 Introduction	1
2 Etat de l'art - Évaluation de la qualité de l'air	4
2.1 Modélisation de la pollution de l'air	4
2.1.1 Modélisation des émissions liées au trafic	4
2.1.2 Modélisation de la dispersion des polluant	6
2.2 L'assimilation de données	10
2.2.1 Types d'assimilation de données	10
2.2.2 Méthodes d'assimilation de données	11
3 Matériel et méthodes	18
3.1 Le modèle SIRANE	19
3.1.1 Origines et spécificités du modèle	19
3.1.2 Structure générale du modèle	19
3.1.3 Le module de canopée urbaine	20
3.1.4 Le module de couche atmosphérique externe	21
3.1.5 Intégration de la chimie atmosphérique	21
3.1.6 Données d'entrée du modèle SIRANE	22
3.2 Correction du modèle par assimilation de données	24
3.2.1 Formalisme général	24
3.2.2 Technique d'ajustement du biais (Bias Adjustment Technique — BAT)	25
3.2.3 Best Linear Unbiased Estimator (BLUE)	26
3.2.4 Données nécessaires pour l'assimilation de données	30
3.3 Critères d'évaluation de l'assimilation de données	31
4 Résultats et discussions	33
4.1 Techniques d'ajustement du biais (BAT)	33
4.1.1 Analyse statistique	36
4.1.2 Discussion des résultats de l'assimilation de données avec la méthode BAT	38
4.2 Best Linear Unbiased Estimator (BLUE)	40
4.2.1 Matrice R	40
4.2.2 Matrice B	42
4.2.3 Analyse statistique	45
4.2.4 Discussion des résultats de l'assimilation de données avec la méthode BLUE	52
4.3 Discussion générale	53
4.4 Validation	54
4.4.1 Stratégie de validation mise en place	54
4.4.2 Contexte et intérêt des stations de validation	55

4.4.3	Séries temporelles des stations de validation	55
4.4.4	Résultats de la validation	56
4.4.5	Discussion	60
5	Conclusion	61
6	Annexes	67
	Annexe 1 – Implémentation de l’assimilation de données et utilisation de YODA	67
	Annexe 2 – BAT – Statistiques par capteur pour les 4 trimestres	69

Table des figures

1	Principales sources de pollution pour les NO _x , les SO _x et PM _{2.5} (BRUXELLES ENVIRONNEMENT, 2024)	2
2	Tendances de concentrations des différents polluants en RBC de 1990 à 2022 – adapté à partir des données de BRUXELLES ENVIRONNEMENT (2024)	2
3	Récapitulatif des méthodes d’assimilation de données - Adapté et traduit de ASCH et al. (2016)	16
4	Schéma du workflow général de la méthodologie mise en place	18
5	Schéma conceptuel du modèle SIRANE tiré de SOULHAC et al. (2011)	20
6	Localisation des 10 stations utilisées dans les différentes techniques d’assimilation de données – Région de Bruxelles-Capitale	30
7	Évolution des concentrations de NO ₂ simulées par SIRANE et corrigées selon les méthodes BAT additive et multiplicative (1/2)	34
7	Évolution des concentrations de NO ₂ simulées par SIRANE et corrigées selon les méthodes BAT additive et multiplicative (2/2)	35
8	Champs de concentrations de NO ₂ corrigées par la méthode BLUE pour le premier trimestre 2021, selon les six formulations de la matrice B	43
9	Champs de concentrations de NO ₂ corrigées par la méthode BLUE pour le deuxième trimestre 2021, selon les six formulations de la matrice B	44
10	Champs de concentrations de NO ₂ corrigées par la méthode BLUE pour le troisième trimestre 2021, selon les six formulations de la matrice B	44
11	Champs de concentrations de NO ₂ corrigées par la méthode BLUE pour le quatrième trimestre 2021, selon les six formulations de la matrice B	45
12	Comparaison des biais moyens entre les stations, avec et sans assimilation (méthode BLUE, $B(LCr)$)	48
13	Comparaison du RMSE entre les stations, avec et sans assimilation	49
14	Comparaison du NMSE entre les stations, avec et sans assimilation	50
15	Comparaison du coefficient de corrélation entre les stations, avec et sans assimilation	51
16	Emplacements des stations ; en orange, les stations utilisées pour la validation	55
17	Séries temporelles des stations de validation	55
18	Diagramme de dispersion comparant les observations mesurées avec celles extraites à l’emplacement de la station 41CHA1 par le modèle SIRANE et l’assimilation de données BLUE	56
19	Diagramme de dispersion comparant les observations mesurées avec celles extraites à l’emplacement de la station 41REG1 par le modèle SIRANE et l’assimilation de données BLUE	57
20	Comparaison des indicateurs statistiques des stations de validation et ceux des stations utilisées par SIRANE et BLUE	58

Liste des tableaux

1	Seuils d'acceptabilité des indicateurs en milieu urbain	32
2	Facteurs de correction obtenus pour les BAT additives et multiplicatives par période en 2021	33
3	Comparaison des performances du modèle SIRANE corrigé par les méthodes BAT additive et multiplicative pour chaque trimestre de 2021	36
4	Statistiques par capteur pour la correction BAT additive – Janvier à Mars 2021	37
5	Statistiques par capteur pour la correction BAT multiplicative – Janvier à Mars 2021	37
6	Matrice R – janvier à mars 2021	41
7	Paramètres optimisés pour le modèle $B(LCr)$ selon les 4 trimestres de l'année 2021	42
8	Statistiques pour la période Janvier - Mars 2021	46
9	Statistiques pour la période Avril - Juin 2021	46
10	Statistiques pour la période Juillet - Septembre 2021	46
11	Statistiques pour la période Octobre - Décembre 2021	46
12	Comparaison des performances du modèle SIRANE corrigé par les méthodes BAT et BLUE pour chaque trimestre de 2021	53
13	Stations les plus proches de la station 41REG1	59
14	Comparaison des statistiques de performance pour la station 41REG1 selon différentes méthodes	59
15	Statistiques par capteur pour la correction BAT additive – Janvier à Mars 2021	69
16	Statistiques par capteur pour la correction BAT multiplicative – Janvier à Mars 2021	69
17	Statistiques par capteur pour la correction BAT additive – Avril à Juin 2021	69
18	Statistiques par capteur pour la correction BAT multiplicative – Avril à Juin 2021	69
19	Statistiques par capteur pour la correction BAT additive – Juillet à Sep- tembre 2021	70
20	Statistiques par capteur pour la correction BAT multiplicative – Juillet à Septembre 2021	70
21	Statistiques par capteur pour la correction BAT additive – Octobre à Décembre 2021	70
22	Statistiques par capteur pour la correction BAT additive – Octobre à Décembre 2021	70

Liste des acronymes

3D-Var Three-Dimensional Variational assimilation. 12, 13, 15, 53

4D-Var Four-Dimensional Variational assimilation. 12, 13, 17, 24

AD assimilation de données. 10, 17

BAT Bias Adjustment Technique. 18, 24, 25, 33

BC *Black Carbon*. 1

BLUE Best Linear Unbiased Estimator. 14, 15, 18, 24, 33

CFD dynamique des fluides numériques. 9, 19

CO monoxyde de carbone. 1

COV composés organiques volatils. 1

EDA Ensemble of Data Assimilation. 17

EKF filtre de Kalman étendu. 15, 24

EnKF filtre d'ensemble de Kalman. 15, 16, 24, 53

FB biais fractionnel. 31

IEnKS Iterative Ensemble Kalman Smoother. 17

IRM Institut Royal Météorologique. 22

KF filtre de Kalman. 15

NMSE erreur quadratique normalisée. 31, 36

NO monoxyde d'azote. 5, 21

NO₂ dioxyde d'azote. 5, 21

NO_x oxydes d'azote. 1

O₃ ozone. 1, 21

OI interpolation optimale. 15

OMS Organisation Mondiale de la Santé. 1

PM particules fines. 1

RBC Région de Bruxelles-Capitale. 1

RMSE erreur quadratique moyenne. 31, 36

SO₂ dioxyde de soufre. 1

SO_x oxydes de soufre. 2

1 Introduction

La pollution atmosphérique constitue l'une des menaces majeures pour la santé publique et l'environnement de notre époque. Selon l'Organisation Mondiale de la Santé (OMS), en 2019, la pollution de l'air ambiant a été responsable de 4,2 millions de décès prématurés (WHO, 2024). La qualité de l'air est influencée par 4 facteurs principaux : les sources d'émissions de polluants, la météorologie, la physico-chimie des polluants et leur transport (CELINE, s.d.). Parmi les polluants les plus préoccupants figurent les particules fines (PM), les composés organiques volatils (COV), les oxydes d'azote (NOx), le dioxyde de soufre (SO₂), le monoxyde de carbone (CO), et divers métaux lourds tels que le plomb (MANISALIDIS et al., 2020).

Ces polluants ont des impacts sanitaires multiples. Le système respiratoire est particulièrement vulnérable, avec une augmentation des cas d'asthme, de bronchopneumopathie chronique obstructive (BPCO), de cancers pulmonaires et d'infections respiratoires aiguës. Ces pathologies sont souvent exacerbées par l'inhalation de particules fines et de COV, provoquant des inflammations et des lésions des tissus pulmonaires (MAIO et al., 2023).

Certaines maladies cardiovasculaires sont également en lien avec cette pollution. Des pathologies telles que l'hypertension, les infarctus, l'insuffisance cardiaque et les arythmies peuvent résulter du stress oxydatif, de l'inflammation systémique et du dysfonctionnement endothélial induits par les polluants (LEDERER et al., 2021). De plus, le système nerveux central peut être affecté, favorisant le déclin cognitif, les maladies neurodégénératives (Alzheimer, Parkinson) et diverses formes de démence (KAMPA et CASTANAS, 2008).

Cette pollution de l'air ne se limite pas qu'aux impacts sur la santé humaine. Elle entraîne également des effets environnementaux majeurs, comme l'acidification des sols et des eaux due aux pluies acides, la dégradation des écosystèmes aquatiques par l'eutrophisation, le dépérissement des forêts, ainsi que la détérioration des bâtiments et des monuments (SPORTISSE, 2008).

Les conséquences de la pollution de l'air sont aggravées pour les populations vulnérables, notamment les enfants, les personnes âgées, les femmes enceintes et les individus souffrant de maladies chroniques (MAUNG et al., 2022). Les inégalités sociales et économiques jouent également un rôle, les communautés à faibles revenus étant souvent plus exposées aux sources de pollution (LAURIKS et al., 2022).

La situation de Bruxelles

En Région de Bruxelles-Capitale (RBC), la gestion de la pollution atmosphérique est un enjeu sanitaire, environnemental et politique majeur, comme en témoignent la mise en place des zones de basses émissions ou le plan GoodMove (BRUXELLES MOBILITÉ, 2025). Pour quantifier le degré de pollution, Bruxelles Environnement, organisme en charge de la surveillance et de l'information sur la qualité de l'air dans la Région de Bruxelles-Capitale, exploite en 2025 un réseau de 13 stations télémétriques réparties sur le territoire régional. Ce réseau mesure divers polluants réglementés : 11 stations surveillent les NOx, 6 les particules fines (PM_{2.5} et PM₁₀), 5 l'ozone (O₃), et un nombre plus restreint pour le *Black Carbon* (BC), le CO ou encore le SO₂.

Comme l'illustre la Figure 1, la répartition des sources d'émissions varie fortement selon le polluant. En 2022, les émissions de NO_x provenaient principalement du transport routier (47 %), suivi de la combustion dans les bâtiments résidentiels et tertiaires (34 %) et d'autres sources plus marginales (10 % pour diverses activités et 9 % pour la production d'énergie). Les émissions d'oxydes de soufre (SO_x) étaient, quant à elles, largement dominées par la combustion dans les bâtiments (93 %), la production d'énergie et les autres sources représentant ensemble moins de 7 %. Les particules fines (ici, PM_{2.5} plus particulièrement, donc celles de taille inférieure à 2.5 µm) provenaient majoritairement de la combustion dans les bâtiments (49 %), suivie du transport routier (24 %), des activités industrielles (11 %) et de la gestion des déchets (11 %), les autres contributions restant marginales.

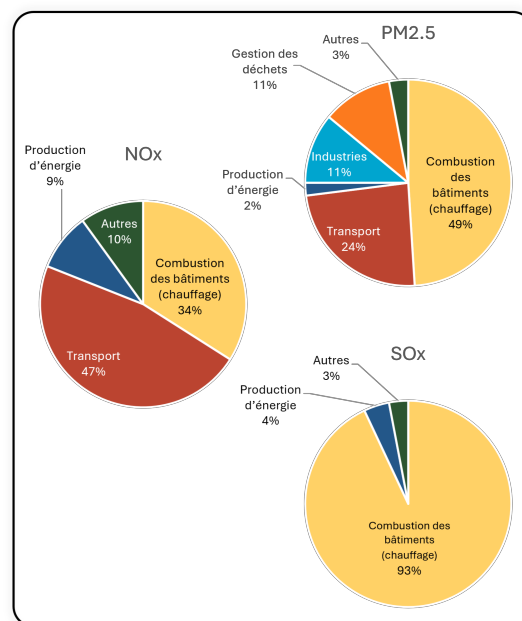


FIGURE 1 – Principales sources de pollution pour les NO_x, les SO_x et PM_{2.5} (BRUXELLES ENVIRONNEMENT, 2024)

Depuis 1990, les concentrations annuelles moyennes de la plupart des polluants ont diminué de plus de 50 %, certaines, comme les SO_x, chutant de près de 80 %. Cette tendance, visible à la Figure 2, s'explique par les progrès technologiques (filtres à particules, chaudières plus performantes), l'évolution des carburants, des réglementations plus strictes et, dans une moindre mesure, par des changements de comportement. Toutefois, les dépassements ponctuels des recommandations de l'OMS persistent, en particulier pour le NO₂ et les particules fines, ce qui justifie la poursuite d'efforts pour suivre l'évolution de la qualité de l'air et améliorer sa gestion.

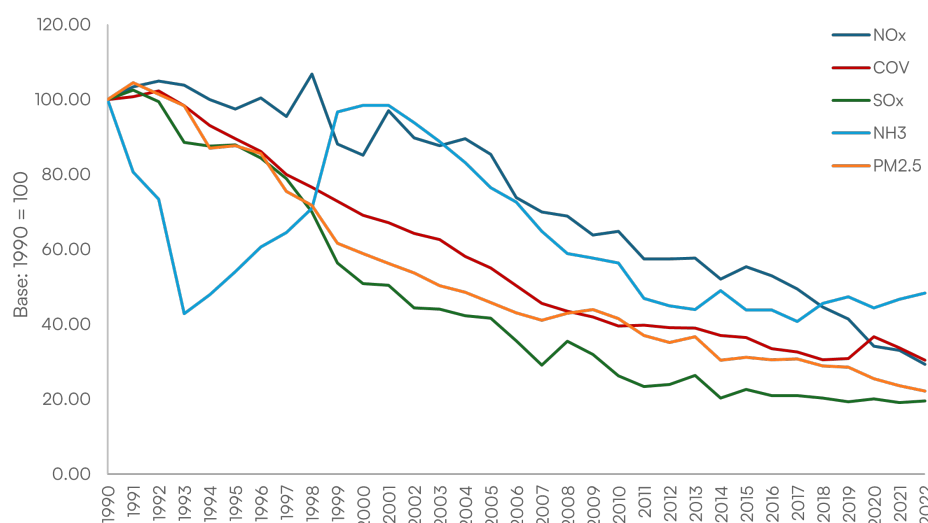


FIGURE 2 – Tendances de concentrations des différents polluants en RBC de 1990 à 2022 – adapté à partir des données de BRUXELLES ENVIRONNEMENT (2024)

Limites de la modélisation et intérêt de l'assimilation de données

Dans ce contexte, la modélisation de la dispersion des polluants constitue un outil indispensable pour comprendre les mécanismes de transport et d'accumulation des polluants, évaluer l'impact des émissions et soutenir les politiques publiques. Cependant, les modèles, même avancés, présentent des limites : incertitudes liées aux inventaires d'émissions, simplifications dans la représentation des processus atmosphériques, erreurs de paramétrisation météorologique, ou encore limites de résolution spatiale et temporelle. Ces biais peuvent conduire à des écarts significatifs entre les concentrations simulées et les mesures effectuées sur le terrain.

Un traitement supplémentaire a posteriori (*post-processing*) par assimilation de données apparaît comme une solution pour réduire ces écarts. En combinant les sorties des modèles avec les observations issues de stations de mesure, cette approche améliore la précision et la représentativité des cartes de qualité de l'air. Elle permet non seulement de corriger les valeurs ponctuelles, mais aussi de générer des champs spatiaux et temporels cohérents, utilisables pour l'analyse, la communication au public et la prise de décision.

Objectifs

Ce mémoire s'inscrit dans le cadre d'une collaboration entre l'Université catholique de Louvain (UCLouvain) et Bruxelles Environnement, organisme opérant le modèle urbain de dispersion des polluants SIRANE, pour la Région de Bruxelles-Capitale. Or, des écarts sont attendus entre les concentrations modélisées par SIRANE et celles mesurées par le réseau officiel de télémesures. Dans la perspective d'une production régulière de cartographies horaires issues du modèle urbain SIRANE, Bruxelles Environnement souhaite disposer de cartes corrigées, scientifiquement robustes et intégrant l'information issue des mesures in situ.

L'objectif principal de ce travail est donc de mettre au point et d'évaluer des méthodes de post-traitement par assimilation de données permettant de corriger ces cartes en tenant compte des observations, afin de fournir une représentation plus fidèle de la répartition spatiale de la pollution. Il convient de préciser que, dans l'ensemble de ce mémoire, l'expression *assimilation de données* renvoie exclusivement à ce cadre de post-traitement. En effet, de manière générale, l'assimilation peut être distinguée entre son intégration directe au sein du modèle numérique et son application en phase de post-traitement ; c'est cette seconde approche qui est privilégiée ici. Deux approches sont explorées et comparées :

- La technique d'ajustement du biais (Bias Adjustment Technique, BAT) qui applique une correction uniforme du biais entre le modèle et les observations
- Le Best Linear Unbiased Estimator (BLUE) qui combine les informations issues du modèle et des observations en tenant compte de leurs incertitudes respectives.

Les performances des deux approches sont évaluées de manière trimestrielle sur l'ensemble de l'année 2021. L'évaluation s'appuie sur les concentrations de NO₂ modélisées par SIRANE et sur les mesures issues du réseau téléométrique officiel, en appliquant une validation croisée de type leave-one-out pour tester leur robustesse.

Après cette introduction, ce mémoire présente d'abord un état de l'art sur la modélisation de la qualité de l'air et les principales méthodes d'assimilation de données. La section suivante décrit le modèle SIRANE, les données et les méthodologies BAT et BLUE. La partie résultats analyse les performances des deux approches, les discute et détaille la validation de la méthode la plus performante. Enfin, ce document se conclut par un résumé des résultats, des limites et des pistes d'amélioration.

2 Etat de l'art - Évaluation de la qualité de l'air

L'évaluation de la qualité de l'air repose sur une combinaison de mesures, de modélisation et, de plus en plus, sur des approches hybrides intégrant l'assimilation de données. Cette section présente un panorama des principales techniques utilisées pour modéliser la pollution atmosphérique, en allant des approches les plus simples aux modèles les plus complexes.

Dans un second temps, les méthodes d'assimilation de données seront passées en revue, avec une attention particulière portée aux approches variationnelles et aux méthodes statistiques. Cette analyse permettra de situer les techniques retenues dans le cadre de ce travail et d'en évaluer la pertinence au regard des objectifs poursuivis.

2.1 Modélisation de la pollution de l'air

La modélisation de la pollution de l'air est un outil numérique essentiel pour analyser les relations complexes entre les émissions, les conditions météorologiques, les concentrations atmosphériques et les dépôts (DALY et ZANNETTI, 2007). Contrairement aux mesures de qualité de l'air, qui fournissent des valeurs ponctuelles et statiques des concentrations ambiantes, la modélisation permet une représentation déterministe et détaillée des processus sous-jacents aux problématiques de pollution atmosphérique (MODI et al., 2013). Pour être efficace, cette approche exige une compréhension approfondie des phénomènes en jeu et des équations mathématiques qui relient les processus physiques aux données observées.

En général, la modélisation de la qualité de l'air prend en considération les sources d'émissions, les transformations physiques et chimiques des polluants dans l'atmosphère, ainsi que les conditions météorologiques. Ce cadre d'analyse permet de mieux comprendre le comportement des polluants dans l'environnement et d'identifier les principaux facteurs influençant leur répartition.

Les modèles de pollution de l'air s'appliquent à une large gamme d'échelles d'application, depuis le niveau continental jusqu'à celui des quartiers, en fonction des besoins et des spécificités de chaque étude. Ces modèles offrent une large gamme d'applications, parmi lesquelles on trouve notamment l'identification des principales sources de pollution, l'estimation de l'exposition de la population à des résolutions spatiales et temporelles plus fines que les mesures directes, ainsi que la prévision de l'évolution des concentrations de polluants dans le temps. Ces fonctionnalités sont cruciales pour cibler les actions de réduction sur les sources les plus polluantes et pour évaluer l'impact des mesures d'atténuation sur la qualité de l'air (TIWARY et WILLIAMS, 2018).

Dans les sections suivantes, deux composantes principales de la modélisation de la pollution de l'air en milieu urbain seront explorées : la modélisation des émissions de trafic, qui est souvent la principale cause des concentrations élevées de polluants en ville (SINI et MESTAYER, 1998), et la modélisation de la dispersion des polluants, qui décrit la manière dont les substances se déplacent et se diluent dans l'atmosphère.

2.1.1 Modélisation des émissions liées au trafic

La modélisation des émissions liées au trafic est un outil essentiel pour évaluer l'impact environnemental des transports routiers, en particulier dans les zones urbaines. Les

véhicules, notamment les voitures, constituent souvent la principale source de pollution atmosphérique en milieu urbain, en particulier pour les oxydes d'azote (monoxyde d'azote (NO) et dioxyde d'azote (NO₂)). Cette modélisation permet d'obtenir des estimations précises des émissions et joue un rôle crucial dans la mise en place de politiques de réduction de la pollution. Elle facilite également la compréhension des contributions spécifiques du trafic à la dégradation de la qualité de l'air et aide les autorités à définir des mesures correctives adaptées pour atteindre des objectifs environnementaux et sanitaires.

2.1.1.1 Les modèles basés sur la vitesse moyenne

Les modèles d'émissions basés sur la vitesse moyenne (Average Speed Models) évaluent les émissions en utilisant des facteurs d'émission (g/véhicule-kilomètre) qui dépendent de la vitesse moyenne des véhicules. Simples d'utilisation, ces modèles s'appuient sur des données accessibles telles que le nombre de kilomètres parcourus et la vitesse moyenne (SMIT et al., 2010). Ils prennent en compte les caractéristiques des véhicules, le type de carburant et le contexte de conduite (urbain, rural ou autoroute).

Parmi les modèles couramment utilisés, COPERT (NTZIACHRISTOS et al., 2009) calcule les émissions à partir de profils de conduite standardisés et d'équations adaptées à chaque catégorie de véhicule. Intégré à la méthodologie EMEP/EEA, il est utilisé pour les inventaires nationaux d'émissions dans de nombreux pays européens (BORGE et al., 2012). Aux États-Unis, des modèles similaires tels que MOBILE et EMFAC sont employés à des fins similaires.

Ces modèles présentent cependant certaines limites. Ils ne capturent pas les variations spécifiques des opérations des véhicules, comme les arrêts fréquents en zone urbaine, et bien que les cycles de conduite standardisés incluent des variations de vitesse, ils ne reproduisent que partiellement les situations de trafic, notamment en cas de congestion (GROTE et al., 2016).

2.1.1.2 Les modèles de situation de trafic

Les modèles d'émissions basés sur les situations de circulation (Traffic Situation Models) estiment les émissions en fonction de descriptions détaillées de différentes conditions de trafic, comme le mode "stop-and-go" ou la circulation fluide sur autoroute. Contrairement aux modèles basés sur la vitesse moyenne, ces modèles utilisent des scénarios spécifiques liés aux types de routes et à des niveaux qualitatifs de congestion (SMIT et al., 2010).

Ces modèles simulent des conditions de circulation réalistes en associant des profils d'émissions moyens à des situations spécifiques. Par exemple, les émissions générées lors d'une situation de congestion diffèrent de celles produites par un trafic fluide sur autoroute. Un modèle représentatif de cette catégorie est HBEFA (Handbook Emission Factors for Road Transport), qui fournit des facteurs d'émission en fonction de situations de circulation particulières (BORGE et al., 2012). Bien que ces modèles soient plus précis pour capturer les effets de la congestion, leur applicabilité est souvent limitée aux régions pour lesquelles les données de calibration sont disponibles.

2.1.1.3 Les modèles variables de trafic

Les modèles d'émissions dits « Traffic-variable » estiment les facteurs d'émission (g/véhicule-kilomètre) en tenant compte de variables telles que la vitesse moyenne, la densité de trafic, la longueur des files d'attente et les paramètres des feux de signalisation. Ils exploitent

des données issues de modèles de trafic macroscopiques ou microscopiques, permettant ainsi une meilleure prise en compte des effets de la congestion par rapport aux modèles basés uniquement sur la vitesse moyenne (GROTE et al., 2016 ; SMIT et al., 2010).

Un exemple notable est le modèle TEE-KCF, qui utilise un facteur de correction cinématique (KCF) pour ajuster les émissions calculées avec COPERT en fonction de la densité de trafic et du temps de feu vert. D'autres études ont amélioré la précision de ces modèles en intégrant des distributions détaillées des vitesses des véhicules (GROTE et al., 2016). Bien que prometteurs, ces modèles demeurent en développement et ne sont pas encore largement disponibles.

2.1.1.4 Les modèles de cycle variables

Les modèles d'émissions « Cycle-variable » (par exemple MEASURE, VERSIT+ et NEMO) estiment les facteurs d'émission en fonction de variables associées au cycle de conduite, telles que le temps d'arrêt, la vitesse moyenne et l'énergie cinétique positive. Ces modèles, d'une résolution temporelle élevée (de la seconde à la minute), nécessitent des informations détaillées sur les mouvements des véhicules, comme la vitesse instantanée, l'accélération et la pente de la route. Ces données peuvent être obtenues grâce à des modèles de trafic microscopiques ou à des équipements GPS (GROTE et al., 2016 ; SMIT et al., 2010).

Par exemple, le modèle VERSIT+ LD a été développé pour les véhicules légers. Il repose sur des régressions linéaires appliquées à des données de tests d'émissions pour établir une relation empirique entre les facteurs d'émission et les variables du cycle de conduite. Le modèle NEMO, élaboré dans le cadre du projet ARTEMIS, simplifie la collecte des données en utilisant un cycle représentatif de l'ensemble des véhicules circulant sur un même segment, réduisant ainsi le temps de calcul.

2.1.1.5 Les modèles modaux

Les modèles modaux, tels que PHEM et CMEM, calculent les facteurs d'émission (g/s ou g/mode) en fonction des différents modes de fonctionnement du véhicule. Ces modèles offrent une résolution temporelle très fine (de l'ordre de la seconde) et nécessitent des informations précises sur le comportement de conduite, comme la vitesse, l'accélération et les changements de vitesse (SMIT et al., 2010).

Le modèle MOVES (Motor Vehicle Emission Simulator) illustre cette approche. Il évalue les facteurs d'émission en fonction de la distribution des modes de fonctionnement, définie par une combinaison de vitesse et de puissance spécifique du véhicule. MOVES se distingue par sa capacité à calculer des émissions instantanées, bien qu'il puisse être assimilé à un modèle « Traffic-variable » lorsqu'il utilise une distribution agrégée pour tout le trafic sur un segment donné. Le modèle PHEM suit une approche similaire en intégrant des données sur le comportement de conduite et les caractéristiques des véhicules pour produire des estimations d'émissions à haute résolution (GROTE et al., 2016).

2.1.2 Modélisation de la dispersion des polluants

Les modèles de dispersion des polluants permettent de prédire comment les substances émises dans l'atmosphère se déplacent, se diluent et se concentrent. En tenant compte des paramètres météorologiques, de la géométrie et de l'intensité des sources, ces modèles

permettent d'estimer quantitativement les concentrations de polluants en fonction de leur localisation et du temps (TIWARY et WILLIAMS, 2018).

2.1.2.1 Les modèles en boîte

Les modèles en boîte constituent l'une des premières et des plus simples approches de la modélisation de la dispersion des polluants atmosphériques. Basés sur le principe de conservation de la masse, ces modèles considèrent le site étudié comme un volume fixe, ou "boîte", dans lequel les polluants sont uniformément distribués (HOLMES et MORAWSKA, 2006). Les émissions sont supposées se mélanger instantanément dans la boîte, produisant ainsi une concentration homogène dans l'ensemble de l'espace défini (NICOLAS et CRAFFE, 2003).

Ces modèles prennent en compte des paramètres simples, tels que le taux d'émission de polluants, la vitesse moyenne du vent traversant la boîte, ainsi que les dimensions de cette dernière. Grâce à cette approche simplifiée, ils nécessitent peu de puissance de calcul et permettent des simulations rapides, ce qui les rend particulièrement adaptés aux évaluations préliminaires et aux environnements peu complexes.

Cependant, l'hypothèse d'homogénéité du mélange induit des limites importantes pour ce type de modèle, notamment en ce qui concerne la précision des variations spatiales des concentrations à l'intérieur de la boîte, et donc dans l'environnement local. Des variations locales de paramètres tels que la vitesse du vent ou l'intensité des émissions influencent la dispersion des polluants de manière différente, mais ces effets ne sont pas capturés par le modèle (HOLMES et MORAWSKA, 2006). Ce type de modèle n'est donc pas vraiment adapté aux environnements complexes, avec des obstacles comme peuvent l'être par exemple les espaces urbains.

2.1.2.2 Les modèles gaussiens

Les modèles gaussiens, largement utilisés pour modéliser la dispersion atmosphérique, supposent que les polluants se distribuent selon une courbe normale dans les directions horizontale et verticale. Ces modèles sont souvent appliqués pour estimer la dispersion dans des environnements simples et ouverts, tels que des terrains plats (HOLMES et MORAWSKA, 2006 ; JOHNSON, 2022).

Ces modèles représentent un panache de polluants émis de manière continue, avec des taux d'émission et des conditions météorologiques constants. Le panache se dilate en fonction de la stabilité atmosphérique et de la vitesse du vent, qui déterminent son élargissement (HOLMES et MORAWSKA, 2006). Un concept clé dans ce cadre est celui de « panache gaussien » : chaque panache est assimilé à un nuage de polluants qui se disperse en se déplaçant dans la direction du vent, incorporant de l'air environnant. À mesure qu'il se dilate, la concentration de polluants diminue. Ce panache peut être modélisé comme une série infinie de « puffs » ou petits nuages, permettant de calculer la concentration de polluants à différentes distances de la source (TIWARY et WILLIAMS, 2018).

Cependant, cette approche présente des limites importantes. Elle ne prend pas en compte les transformations chimiques au sein du panache, ce qui réduit sa précision dans des contextes complexes ou urbains, où les interactions avec les bâtiments et les phénomènes de recirculation sont fréquents. En conséquence, les modèles gaussiens peuvent surestimer les concentrations dans des conditions de faible vent ou à proximité immédiate de la

source (moins de 100 m) (HOLMES et MORAWSKA, 2006 ; LATEB et al., 2016). Pour ces environnements, des modèles hybrides, intégrant des "puffs" successifs, apportent des améliorations en introduisant davantage de variabilité dans la dispersion (TIWARY et WILLIAMS, 2018).

2.1.2.3 Les modèles eulériens

Les modèles eulériens de dispersion atmosphérique reposent sur la résolution numérique d'équations de transport atmosphérique, généralement formulées à partir des équations d'advection-diffusion. Ils se concentrent sur les variations de concentration au sein d'un cadre fixe, ou "grille" eulérienne (LEELŐSSY et al., 2014). Ce cadre divise l'espace d'étude en cellules tridimensionnelles, dans lesquelles la concentration moyenne des polluants est calculée à chaque instant.

Le modèle prend en compte deux processus principaux : le transport par advection, c'est-à-dire le déplacement des polluants sous l'influence du vent, et la diffusion turbulente, qui représente la dispersion aléatoire due aux mouvements atmosphériques. La résolution de ces équations permet de simuler les variations spatiales et temporelles de la concentration des polluants dans chaque cellule. Une des forces majeures des modèles eulériens est leur capacité à intégrer directement les processus de dépôt et de réaction chimique, ce qui permet de prendre en compte les mécanismes de disparition ou de transformation des particules (HOLMES et MORAWSKA, 2006).

Cependant, ce niveau de complexité nécessite des ressources de calcul importantes, particulièrement lorsque les cellules doivent être affinées pour capturer des gradients de concentration prononcés autour d'obstacles ou dans des zones de forte pollution. Bien que des simplifications soient possibles, telles que l'approximation d'un champ de pression constant, ces compromis peuvent influencer la précision des résultats (NICOLAS et CRAFFE, 2003).

2.1.2.4 Les modèles lagrangiens

Les modèles lagrangiens offrent une approche alternative en suivant les trajectoires de particules de polluants dans un champ de vent tridimensionnel, contrairement aux modèles eulériens qui calculent des concentrations moyennes sur une grille fixe. Ces modèles suivent les particules soumises à l'effet du vent et à la turbulence locale, modélisée par une vitesse aléatoire dérivée d'un processus de Markov (NICOLAS et CRAFFE, 2003).

Ils sont particulièrement adaptés aux environnements complexes, permettant de modéliser la sédimentation, les transformations chimiques et des caractéristiques comme les bâtiments ou la végétation (HOLMES et MORAWSKA, 2006). Pour des distances plus grandes, des « puffs » successifs, suivant des trajectoires stochastiques, simulent les variations locales du vent et de la dispersion (JOHNSON, 2022).

L'un de leurs avantages majeurs est l'indépendance vis-à-vis de la résolution de la grille, réduisant les erreurs de diffusion numérique. Cependant, ces modèles nécessitent un nombre élevé de trajectoires, ce qui peut augmenter les coûts de calcul. Pour pallier à ce problème, des approches hybrides, combinant modèles lagrangiens et eulériens, permettent d'améliorer l'efficacité de la modélisation (LEELŐSSY et al., 2014).

2.1.2.5 Les modèles de dynamique des fluides numériques

Les modèles de dynamique des fluides numériques (CFD) reposent sur les équations de Navier-Stokes, qui décrivent le mouvement des fluides en tenant compte de la masse, de la quantité de mouvement et des forces de viscosité. Ces modèles sont utilisés pour simuler l'écoulement de l'air et la dispersion des polluants, en particulier dans les environnements urbains où les bâtiments influencent fortement le flux d'air (LEELŐSSY et al., 2014). Bien qu'ils offrent une grande précision pour modéliser la pollution à petite échelle, ils nécessitent des ressources de calcul importantes.

Les CFD se déclinent en trois grandes catégories : RANS (Reynolds-Averaged Navier-Stokes), LES (Large Eddy Simulation) et DNS (Direct Numerical Simulation). Les modèles RANS, qui moyennent les écoulements dans le temps, sont souvent utilisés pour les simulations à grande échelle en raison de leur coût de calcul plus modéré (HOLMES et MORAWSKA, 2006). Les modèles LES, qui résolvent directement les grandes échelles de turbulence et modélisent les petites échelles, permettent une meilleure précision des fluctuations temporelles de la pollution (JOHNSON, 2022). Enfin, les modèles DNS, qui résolvent les équations sans simplification, sont rarement utilisés en raison de leur coût informatique très élevé, mais sont adaptés aux études à microéchelle dans des environnements complexes (LATEB et al., 2016). Ces modèles nécessitent toutefois une validation par des données expérimentales pour garantir la fiabilité des résultats (MODI et al., 2013).

2.1.2.6 Modélisation urbaine de la dispersion des polluants

La modélisation urbaine de la dispersion des polluants pose des défis en termes de précision, de coût et de vitesse. La complexité des flux autour des bâtiments, avec des effets comme la recirculation des vents et le confinement de polluants dans les rues, impose un niveau élevé de précision. Les modèles doivent être assez rapides pour être utiles en situations d'urgence, ce qui limite souvent le recours aux simulations à haute résolution comme les CFD, qui demandent un temps de calcul élevé (CLEMENTS et al., 2024).

Les approches disponibles varient en précision et en coût. Les modèles les plus simples, tels que les modèles gaussiens (AERMOD, ADMS-Urban), sont largement utilisés pour leur rapidité et leur faible coût en ressources, tout en offrant des options de paramétrisation pour des environnements urbains. Les modèles lagrangiens, comme NAME et HYSPLIT, permettent une adaptation à différentes échelles, tandis que les modèles de réseau de rues (par exemple SIRANE, MUNICH) se spécialisent dans la simulation des concentrations de polluants dans les canyons urbains. Les modèles LES, comme PALM, offrent une grande précision dans la résolution des flux de l'air et des polluants, mais leur coût de calcul reste un obstacle à leur utilisation opérationnelle (HERTWIG et al., 2018).

Le modèle SIRANE (SOULHAC et al., 2011), qui sera utilisé dans ce rapport, s'inscrit dans la catégorie des modèles de réseau de rues. Il permet de simuler la dispersion au sein d'un réseau urbain tout en tenant compte des effets induits par les rues et intersections, un compromis intéressant entre précision et rapidité pour l'étude des polluants à l'échelle locale.

2.2 L'assimilation de données

Les techniques d'assimilation de données (AD) sont étudiées et utilisées depuis plus d'un demi-siècle, avec un développement continu des méthodes et un élargissement progressif des domaines d'application. Initialement développées dans les années 1960 pour les prévisions météorologiques numériques, elles ont été progressivement adaptées à d'autres champs tels que l'océanographie, la climatologie, les géosciences, l'hydrologie, mais également à des disciplines variées comme l'agriculture, l'imagerie médicale ou encore la finance (PANDYA et al., 2022). L'évolution et le développement de ces méthodes ont été étroitement liés aux progrès des modèles numériques, dont la complexité n'a cessé d'augmenter depuis les années 1980 (PARK et ZUPANSKI, 2022).

L'objectif principal de l'assimilation de données est de combiner de manière optimale les informations issues des observations et celles issues des modèles numériques, afin de fournir la meilleure estimation possible de l'état d'un système. Autrement dit, il s'agit d'ajuster les prévisions d'un modèle aux observations disponibles pour rapprocher ce modèle de la réalité, en améliorant simultanément la représentation du phénomène étudié et les capacités de prévision (ASCH et al., 2016). Une telle approche est indispensable, car ni les observations ni les modèles ne suffisent à eux seuls : les premières sont souvent incomplètes, bruitées et spatialement/temporalement irrégulières, tandis que les seconds souffrent d'incertitudes liées aux conditions initiales, aux paramètres empiriques ou aux processus physiques mal représentés (BUIZZA et al., 2022).

D'un point de vue théorique, l'assimilation de données s'inscrit dans le cadre des problèmes inverses, où l'on cherche à retrouver les causes (paramètres du modèle, état initial) à partir des effets observés. L'assimilation de données peut être vue comme un cas particulier d'optimisation, visant soit à minimiser l'écart entre le modèle et les observations (approche variationnelle), soit à minimiser l'incertitude des estimations (approche statistique) (ASCH et al., 2016).

2.2.1 Types d'assimilation de données

Les méthodes d'assimilation de données peuvent être classées selon plusieurs critères, en fonction des approches conceptuelles et opérationnelles mises en place. Dans la littérature scientifique, trois distinctions majeures sont couramment utilisées : la distinction entre les méthodes séquentielles et non séquentielles, l'opposition entre les approches intermittentes et continues, ainsi que la séparation entre les méthodes variationnelles et statistiques.

Pour la première distinction, soit l'opposition entre les méthodes séquentielles et non-séquentielles, il s'agit principalement de distinguer la manière dont les observations sont intégrées dans le temps. Ainsi, les méthodes séquentielles mettent à jour l'état du système au fur et à mesure de l'arrivée des observations, généralement à chaque pas de temps ou à chaque cycle d'assimilation. Elles exploitent principalement les données disponibles à l'instant de l'assimilation et sont donc particulièrement adaptées aux applications en temps réel, telles que les prévisions météorologiques opérationnelles (BOUTTIER et COURTIER, 2002). Les méthodes non-séquentielles, également appelées rétrospectives, assimilent simultanément un ensemble d'observations couvrant une fenêtre temporelle définie, incluant des observations postérieures à l'instant considéré. Elles sont davantage utilisées pour les réanalyses ou pour reconstruire des états passés de manière cohérente (NGUYEN, 2017).

La seconde distinction, soit celle opposant les méthodes intermittentes à celles continues, se réfère à la longueur des fenêtres temporelles utilisées pour l'assimilation de données. Les méthodes dites intermittentes réalisent l'analyse sur des fenêtres temporelles relativement courtes, en traitant les observations par petits lots successifs. Elles sont plus simples à mettre en œuvre et particulièrement adaptées aux applications opérationnelles nécessitant des mises à jour rapides. À l'inverse, les approches continues considèrent des fenêtres temporelles plus longues, permettant d'obtenir une analyse plus lisse et plus cohérente avec la dynamique du système. Ces dernières sont particulièrement utiles pour les études où la continuité temporelle joue un rôle majeur, comme en climatologie ou en océanographie (NGUYEN, 2017).

Enfin, la distinction entre les techniques variationnelles et les techniques statistiques repose sur les fondements théoriques employés pour ajuster le modèle aux observations. Les méthodes variationnelles s'appuient sur la théorie du contrôle optimal et le calcul des variations. Elles visent à minimiser une fonction de coût représentant l'écart pondéré entre l'état du modèle et les observations (PARK et ZUPANSKI, 2022). À l'opposé, les méthodes statistiques reposent sur un cadre bayésien et sur la théorie de l'estimation statistique. Elles visent à estimer la distribution de probabilité a posteriori de l'état du système en minimisant l'incertitude associée. Alors que les premières relèvent d'une logique déterministe d'optimisation, les secondes adoptent une approche probabiliste (ASCH et al., 2016).

Dans la littérature, ces distinctions sont parfois confondues, bien que basées sur des critères bien distincts. La distinction séquentielle contre non-séquentielle est avant tout temporelle, car elle décrit la manière dont les observations sont assimilées au fil du temps. La distinction variationnelle contre statistique est, quant à elle, théorique puisqu'elle décrit les principes mathématiques employés pour réaliser l'assimilation. Dans la pratique, les deux dimensions se recoupent et sont fréquemment confondues. En effet, les méthodes séquentielles adoptent généralement une approche statistique, tandis que les méthodes non-séquentielles reposent le plus souvent sur une approche variationnelle, d'où l'apparition de ces confusions.

2.2.2 Méthodes d'assimilation de données

2.2.2.1 Méthodes variationnelles

Les méthodes variationnelles constituent une classe majeure d'approches en assimilation de données, largement utilisée dans les systèmes de prévision opérationnels en météorologie, océanographie et au-delà. Elles ont émergé dans les années 1980 avec les travaux fondateurs de LE DIMET et TALAGRAND (1986), en s'appuyant sur la théorie du contrôle optimal, dérivée elle-même du calcul des variations (ASCH et al., 2016, PARK et ZUPANSKI, 2022).

Contrairement aux méthodes statistiques, fondées sur l'estimation probabiliste et la mise à jour incrémentale de l'état du système (si elles sont aussi séquentielles), les approches variationnelles cherchent à déterminer un état optimal qui minimise une fonction de coût, représentant en général une distance quadratique entre les observations disponibles et les équivalents simulés par le modèle (ASCH et al., 2016). Ce problème d'optimisation s'effectue sous contrainte, la contrainte étant ici l'équation dynamique du modèle lui-même. Ainsi, la solution doit à la fois satisfaire la dynamique du système et s'ajuster au mieux aux observations. Ces approches sont dites rétrospectives, car elles intègrent l'ensemble

des observations d'une fenêtre temporelle pour corriger un état initial unique.

Mathématiquement, ce type de problème peut être formulé comme un problème de moindres carrés non linéaire sous contrainte, que l'on peut résoudre à l'aide des équations d'Euler-Lagrange issues du calcul des variations, ou encore par le principe du minimum de Pontryagin dans le cadre du contrôle optimal (PARK et ZUPANSKI, 2022). La mise en œuvre numérique de cette minimisation repose sur des techniques d'optimisation, dont le calcul du gradient est souvent effectué à l'aide de la méthode adjointe, un outil central de l'assimilation variationnelle (ASCH et al., 2016). Celle-ci permet de remonter efficacement les sensibilités des observations par rapport aux états du modèle, même dans des systèmes complexes ou de grande dimension.

Les méthodes variationnelles sont particulièrement puissantes dans les contextes où il est nécessaire d'assimiler des observations sur une fenêtre temporelle étendue, tout en respectant strictement les équations dynamiques du modèle. C'est notamment le cas des approches dites Three-Dimensional Variational assimilation (3D-Var), qui assimilent les données à un instant donné mais sur tout l'espace, et Four-Dimensional Variational assimilation (4D-Var), qui intègrent l'information sur une période de temps entière en utilisant explicitement l'évolution du modèle.

Dans les sections suivantes, les techniques d'assimilation de données variationnelles les plus emblématiques sont présentées : la formulation générale du contrôle optimal appliqué à l'assimilation, ainsi que les approches 3D-Var et 4D-Var.

Le contrôle optimal

Les méthodes variationnelles en assimilation de données s'inscrivent dans le cadre du contrôle optimal, une généralisation moderne du calcul des variations. Elles visent à déterminer l'état initial d'un système dynamique de façon à minimiser une fonction de coût, tout en respectant les équations du modèle. Ce cadre théorique, développé notamment par Pontryagin dans les années 1950 et généralisé aux équations aux dérivées partielles par Lions dans les années 1970, fournit une base rigoureuse à l'assimilation variationnelle (PARK et ZUPANSKI, 2022).

La minimisation s'effectue sous la contrainte du modèle dynamique. Pour résoudre ce problème, on construit une fonction de Lagrange, introduisant des multiplicateurs associés aux équations du modèle. Cela conduit à un système d'optimalité composé :

- de l'équation d'état (la dynamique du modèle),
- de l'équation adjointe, qui rétro-propage l'information des observations vers l'état initial,
- et d'une condition d'optimalité, donnant l'expression du contrôle optimal.

Ce système représente une condition nécessaire d'optimalité du premier ordre. Dans la pratique, la résolution repose sur des méthodes d'optimisation (type gradient ou quasi-Newton), le gradient étant calculé de manière efficace via le modèle adjoint (ASCH et al., 2016). Celui-ci permet d'accéder aux sensibilités du coût par rapport aux variables de contrôle, sans coût additionnel dépendant du nombre d'observations.

Ainsi, l'assimilation variationnelle peut être vue comme un problème de moindres carrés non linéaire sous contrainte, fondé sur des outils analytiques robustes. Cette structure sert de base à ses deux formulations les plus répandues : 3D-Var et 4D-Var, détaillées dans les

sections suivantes.

3D-Var

La méthode 3D-Var estime l'état optimal à un instant donné en assimilant les observations disponibles dans une fenêtre temporelle restreinte, sans tenir compte explicitement de l'évolution temporelle du système. Le modèle n'intervient donc que pour définir l'état de fond, et les observations sont considérées comme simultanées.

Mathématiquement, le problème se ramène à la minimisation d'une fonction coût quadratique visible à l'équation 1 :

$$J(x) = \frac{1}{2}(x - x^b)^T B^{-1}(x - x^b) + \frac{1}{2}(Hx - y)^T R^{-1}(Hx - y) \quad (1)$$

avec x l'état à estimer, x^b l'état de fond, y le vecteur d'observation, H l'opérateur des observations, B la matrice de covariance des erreurs du modèle et R la matrice de covariance des erreurs d'observations (ASCH et al., 2016, MARTIN et al., 2025).

Cette technique a été adoptée dans les années 1990 par de nombreux centres de prévision météorologique (NCEP, ECMWF, Météo-France, UKMO...), en remplacement progressif des méthodes d'interpolation optimale (JUNG et al., 2024), qui seront décrites dans la section consacrée aux méthodes statistiques. Sa structure relativement simple et son faible coût en ont fait une solution opérationnelle efficace, encore utilisée dans certains environnements contraints (comme JEDI-MPAS ou WRFDA) (MELINC et ZAPLOTNIK, 2024).

4D-Var

La 4D-Var étend naturellement la 3D-Var en intégrant la dynamique temporelle du modèle sur une fenêtre d'assimilation. Il consiste à déterminer l'état initial qui, une fois propagé par le modèle, permet de reproduire au mieux les observations réparties dans le temps. Le modèle intervient alors comme contrainte forte dans l'optimisation, et le problème devient fondamentalement un contrôle optimal (PARK et ZUPANSKI, 2022).

Cette approche est beaucoup plus coûteuse en ressources que la 3D-Var, car elle nécessite l'utilisation du modèle direct, du modèle tangent et du modèle adjoint, mais elle offre une meilleure cohérence dynamique et une meilleure assimilation des observations asynchrones. Elle est aujourd'hui largement utilisée dans les centres opérationnels météorologiques (MARTIN et al., 2025).

Notons enfin que de nombreux systèmes 3D-Var ont été progressivement adaptés en 4D-Var, en réutilisant notamment leur structure logicielle et leurs modèles de covariance. Ainsi, le 4D-Var constitue non seulement une extension naturelle mais aussi une évolution logique des systèmes variationnels existants (JUNG et al., 2024).

2.2.2.2 Méthodes statistiques

Les méthodes statistiques constituent l'autre grande famille d'approches en assimilation de données, aux côtés des méthodes variationnelles. Elles se fondent sur un cadre bayésien dans lequel l'état du système est modélisé comme une variable aléatoire, et l'assimilation consiste à combiner de manière optimale l'information contenue dans les observations

et celle issue du modèle pour estimer une distribution a posteriori de l'état du système (PANDYA et al., 2022 ; PARK et ZUPANSKI, 2022 ; REICH, 2019).

Historiquement, ces méthodes ont été introduites à travers les travaux sur le Best Linear Unbiased Estimator (BLUE), et plus concrètement sur l'interpolation optimale (Optimal Interpolation, OI), qui représente un cas particulier de BLUE appliqué à des observations ponctuelles avec un état de fond fixe (PARK et ZUPANSKI, 2022). Ces approches ont ensuite été étendues aux systèmes dynamiques avec l'introduction du filtre de Kalman (Kalman, 1960), qui fournit une solution récursive exacte dans le cas linéaire et gaussien, et qui constitue la base de nombreuses méthodes statistiques modernes (ASCH et al., 2016 ; REICH, 2019).

Dans ces cadres statistiques, l'assimilation s'effectue séquentiellement dans le temps : à chaque pas, l'état du système est mis à jour à l'aide des nouvelles observations disponibles, en tenant compte de l'incertitude issue de la dynamique du modèle. Cela rend ces méthodes particulièrement adaptées à l'assimilation en temps réel et à des systèmes très évolutifs. De plus, elles offrent en sortie une estimation explicite des incertitudes, ce qui constitue un atout majeur par rapport à certaines approches variationnelles (PANDYA et al., 2022).

Dans les sections suivantes, les méthodes statistiques classiques sont présentées, en suivant une progression naturelle : le Best Linear Unbiased Estimator, l'interpolation optimale puis les filtres de Kalman et ses extensions modernes.

Le Best Linear Unbiased Estimator

Le Best Linear Unbiased Estimator (BLUE) constitue une référence fondamentale dans le cadre de l'assimilation de données statistique. Il désigne l'estimateur linéaire qui, sous hypothèses de bruit gaussien et de modèle linéaire, est à la fois non biaisé et optimal au sens de la variance minimale. Autrement dit, parmi tous les estimateurs linéaires non biaisés, le BLUE est celui minimisant la variance de l'erreur d'estimation (ASCH et al., 2016 ; PARK et ZUPANSKI, 2022).

Le problème se pose dans un contexte où un état de fond (issu, par exemple, d'une simulation antérieure) et un ensemble d'observations bruitées sont disponibles, chacun étant associé à une matrice de covariance représentant son incertitude. Le BLUE fournit alors une estimation de l'état réel du système sous la forme d'une combinaison linéaire pondérée de ces deux sources d'information. Les poids attribués à chaque contribution dépendent de leur incertitude respective : plus une source est jugée fiable (variance faible), plus elle influence l'estimation finale.

Dans le cas gaussien linéaire, cette approche coïncide avec l'estimation bayésienne au sens de l'espérance a posteriori, et se révèle équivalente à la solution du problème de minimisation d'une fonction de coût quadratique. Cette observation établit un lien direct avec les méthodes variationnelles, qui reposent sur une formulation mathématiquement similaire dans un cadre plus général.

Le BLUE constitue également le socle théorique de nombreuses techniques d'assimilation, telles que l'interpolation optimale ou les premières formulations linéaires du filtre de Kalman.

L'interpolation optimale

L'interpolation optimale (OI), développée dans les années 1960 par Gandin, constitue une application opérationnelle du cadre général du BLUE à l'assimilation atmosphérique. Elle repose sur le même principe de minimisation de l'erreur quadratique sous contraintes d'impartialité et de linéarité, mais introduit des hypothèses simplificatrices (stationnarité spatiale des erreurs de fond, indépendance des observations, et traitement monovariable), qui facilitent sa mise en œuvre sur des domaines restreints autour des points d'analyse (PANDYA et al., 2022, BOUTTIER et COURTIER, 2002). Cette simplification permet une assimilation efficace et peu coûteuse, adaptée aux capacités computationnelles de l'époque.

Historiquement, l'OI s'est imposée comme la méthode d'assimilation statistique de référence avant l'introduction des techniques variationnelles, telles que la 3D-Var, ou l'intégration explicite de la dynamique du modèle dans les méthodes statistiques.

Les filtres de Kalman et ses extensions modernes

Le filtre de Kalman (KF), introduit par Kalman et Bucy en 1961, constitue l'une des pierres angulaires de l'assimilation de données moderne, notamment dans les systèmes dynamiques linéaires. Il s'agit d'un estimateur récursif optimal au sens de la variance minimale dans un cadre linéaire-gaussien, combinant un modèle de prévision et des observations bruitées selon une approche bayésienne séquentielle. À chaque pas de temps, une étape de *prévision* met à jour l'état du système et sa covariance d'erreur à partir du modèle, suivie d'une étape d'*analyse* qui corrige cet état à l'aide des observations via un gain de Kalman dérivé analytiquement (ASCH et al., 2016 ; PARK et ZUPANSKI, 2022).

Le filtre de Kalman peut être interprété comme une généralisation dynamique du Best Linear Unbiased Estimator (BLUE) : alors que le BLUE fournit un estimateur optimal pour un état inconnu à partir d'un ensemble d'observations fixes, le KF intègre explicitement la dynamique du système. Il conserve la structure linéaire et gaussienne du BLUE, mais l'étend à un cadre séquentiel dans lequel les données sont assimilées progressivement au fil du temps (ASCH et al., 2016). Le filtre repose sur plusieurs hypothèses : linéarité des opérateurs dynamiques et d'observation, erreurs non corrélées et centrées, ainsi que distributions gaussiennes. Sous ces conditions, il fournit une estimation optimale de l'état du système à chaque pas de temps, ainsi que de son incertitude. Cependant, dans de nombreuses applications géophysiques, les systèmes sont non linéaires et de grande dimension, ce qui limite l'applicabilité du KF classique. Deux extensions majeures ont été développées :

- Le **filtre de Kalman étendu (EKF)**, qui repose sur une linéarisation locale des opérateurs non linéaires à l'aide de leur développement de Taylor. Le modèle est ainsi approché linéairement autour de l'estimation courante, permettant d'appliquer les équations du KF de manière approximative (ASCH et al., 2016 ; PARK et ZUPANSKI, 2022).
- Le **filtre d'ensemble de Kalman (EnKF)**, introduit par EVENSEN (1994), qui utilise un ensemble d'états du système pour approximer les moments de la distribution d'erreur. Le gain de Kalman est calculé à partir de la covariance empirique issue de l'ensemble, ce qui permet une prise en compte dynamique des incertitudes. Cette approche est particulièrement adaptée aux systèmes de grande dimension, bien qu'elle repose encore sur l'hypothèse de gaussienneté (PANDYA et al., 2022 ; PARK et ZUPANSKI, 2022).

Ces méthodes forment aujourd'hui une famille centrale des approches statistiques d'assimilation de données. Elles permettent non seulement d'estimer l'état optimal, mais aussi de caractériser l'incertitude associée à cette estimation, ce qui les rend incontournables dans les systèmes de prévision probabiliste et dans les chaînes de simulation en sciences de l'environnement (BOUTTIER et COURTIER, 2002 ; MARTIN et al., 2025).

La Figure 3 recense toutes les techniques d'assimilation de données déjà vues dans cet état de l'art, ainsi que les méthodes hybrides qui sont introduites au point suivant.

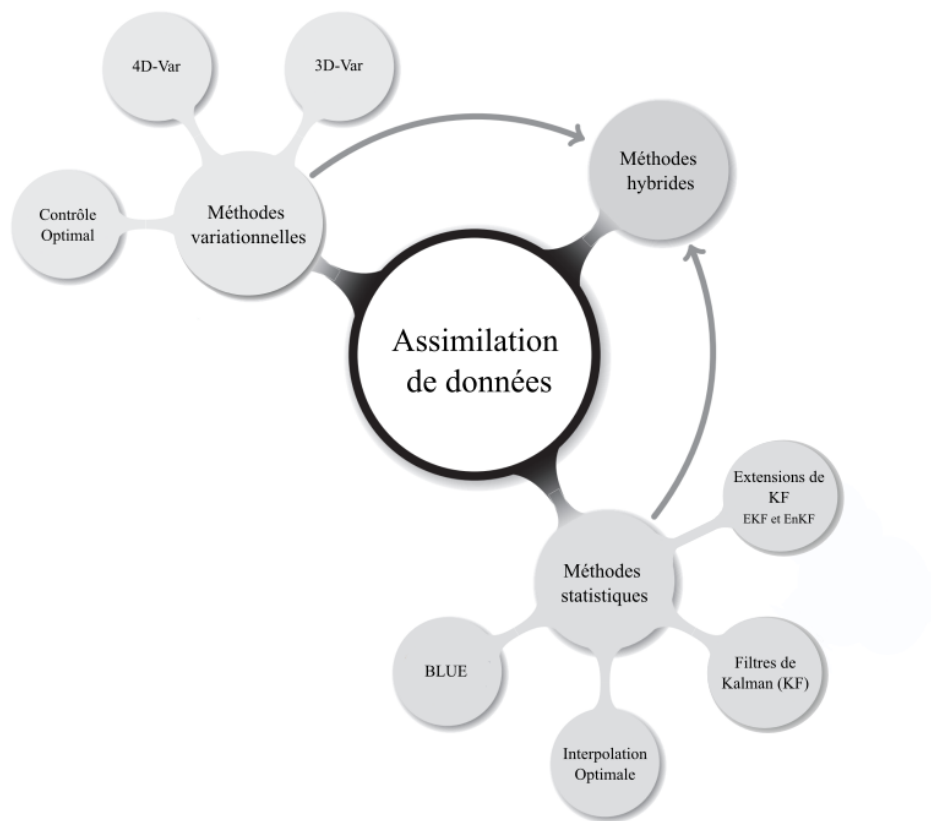


FIGURE 3 – Récapitulatif des méthodes d'assimilation de données - Adapté et traduit de ASCH et al. (2016)

2.2.2.3 Les modèles hybrides

Les approches hybrides visent à combiner les avantages des méthodes variationnelles (VAR), qui offrent une estimation précise du maximum *a posteriori* via l'optimisation d'une fonction de coût, et des méthodes statistiques de type ensemble (ENS), qui permettent une représentation dynamique et dépendante de l'écoulement des erreurs. Ces méthodes hybrides répondent à la difficulté intrinsèque de chaque approche prise isolément : les VAR reposent souvent sur des erreurs statiques et nécessitent des modèles adjoints coûteux à maintenir, tandis que les méthodes de type EnKF présentent des erreurs dépendantes de l'écoulement, mais réduites en rang et sensibles aux non-linéarités.

Parmi les méthodes hybrides les plus répandues, on distingue :

- **Hybrid EnVar (Ensemble Variational)** : cette approche repose sur une combinaison linéaire de matrices d'erreurs de fond issues d'un système variationnel

statique et d'un ensemble de prévisions. Elle est formulée comme une minimisation variationnelle augmentée, incorporant les statistiques d'erreurs dynamiques issues de l'ensemble (PARK et ZUPANSKI, 2022). L'approche proposée initialement par LORENC (2003) a été étendue dans divers systèmes opérationnels, et permet d'exploiter la dépendance à l'écoulement des erreurs tout en conservant la capacité d'optimisation variationnelle. Des implémentations typiques incluent 3DEnVar ou 4DEnVar selon la nature temporelle de la fonction de coût (ASCH et al., 2016).

- **Ensemble of Data Assimilation (EDA)** : plutôt que d'hybrider directement deux approches, l'EDA consiste à exécuter plusieurs systèmes d'assimilation variationnelle (souvent 4D-Var) sur des observations perturbées, générant ainsi une estimation statistique des erreurs d'analyse. Cette technique permet d'introduire de la variabilité dynamique dans des systèmes variationnels autrement statiques (ASCH et al., 2016).
- **Iterative Ensemble Kalman Smoother (IEnKS)** : l'IEnKS étend le filtre de Kalman d'ensemble en y intégrant une optimisation variationnelle non linéaire à chaque pas d'analyse, réalisée dans l'espace réduit de l'ensemble. Ce formalisme permet d'améliorer la cohérence temporelle et la prise en compte des non-linéarités, tout en générant un ensemble mis à jour de manière cohérente (FREI et KÜNSCH, 2013).

Ces approches hybrides marquent une évolution importante vers des systèmes plus flexibles et performants. Elles sont aujourd'hui au cœur des systèmes opérationnels d'assimilation de données météorologiques et océaniques, mais sont de plus en plus souvent utilisées dans des domaines différents.

2.2.2.4 Tendances actuelles et perspectives

Initialement conçue pour la météorologie et l'océanographie, l'assimilation de données (AD) s'est largement étendue à de nouveaux domaines d'application, tels que l'hydrologie, les sciences environnementales, les systèmes énergétiques, la dynamique des fluides industrielle, l'agriculture ou encore la médecine et la cybersécurité (ARCUCCI et al., 2021 ; CAMPORESE et GIOTTO, 2022 ; MARTIN et al., 2025). Cette diffusion repose sur sa capacité à améliorer les prévisions dans des contextes incertains, à partir d'observations hétérogènes ou incomplètes.

Sur le plan méthodologique, une évolution marquante réside dans l'intégration croissante de l'apprentissage automatique au sein des systèmes d'assimilation. Les réseaux de neurones, notamment profonds, sont mobilisés pour approximer des dynamiques complexes, modéliser les erreurs ou accélérer certains calculs (ARCUCCI et al., 2021 ; CHENG et al., 2023). Des cadres comme la *Deep Data Assimilation* (DDA) cherchent à coupler modèles physiques et apprentissage statistique pour construire des prédicteurs plus efficaces, en s'appuyant sur des architectures récurrentes ou convolutionnelles entraînées sur les sorties d'assimilation.

Enfin, l'avenir de l'AD semble résider dans des approches hybrides, à l'interface entre techniques variationnelles, méthodes statistiques à base d'ensembles, et intelligence artificielle. Ces combinaisons visent à renforcer la représentation des incertitudes, la robustesse face aux non-linéarités, et l'adaptabilité à de nouveaux contextes opérationnels, tout en conservant l'assise physique des modèles dynamiques. L'assimilation de données occupe désormais une place centrale dans de nombreux champs scientifiques, en articulant modélisation physique, traitement des données et méthodes d'apprentissage.

3 Matériel et méthodes

La méthodologie générale de ce travail s'inspire directement des travaux de NGUYEN et SOULHAC (2021), qui ont publié un article se nommant : *Data assimilation methods for urban air quality at the local scale*. Cette étude s'inscrivait dans le cadre plus large de la thèse de Chi Vuong Nguyen, consacrée à "l'assimilation de données et au couplage d'échelles pour la modélisation de la dispersion atmosphérique en milieu urbain". Leurs expérimentations initiales portaient plus particulièrement sur la ville de Lyon.

L'objectif principal de ce mémoire est d'adapter et d'appliquer cette méthodologie à un nouveau contexte urbain, en l'occurrence la Région de Bruxelles-Capitale. L'ensemble des développements repose sur l'intégration de données multiples dans le modèle urbain de dispersion SIRANE, couplée à des techniques d'assimilation de données visant à corriger les biais systématiques du modèle.

La Figure 4 résume l'organisation générale du travail effectué. Plusieurs types de données sont mobilisées : la géométrie urbaine, les conditions météorologiques, les émissions linéaires, ponctuelles et surfaciques, ainsi que la pollution de fond. Ces informations alimentent le modèle SIRANE, qui fournit des grilles de concentration pour différents polluants atmosphériques (NO_2 , NO , PM_{10} , $\text{PM}_{2.5}$ et O_3). Dans cette étude, seules les concentrations de dioxyde d'azote (NO_2) sont exploitées, en raison de leur intérêt pour leur importante variabilité spatiale, ainsi que leur lien direct avec les émissions de trafic.

À partir des champs de concentration produits par SIRANE, deux méthodes d'assimilation de données ont été sélectionnées et intégrées pour améliorer les estimations modélisées : la Bias Adjustment Technique (BAT) et le Best Linear Unbiased Estimator (BLUE). Ces techniques permettent de générer des cartes corrigées, plus proches des observations mesurées aux stations.

L'analyse est menée par trimestre sur l'année 2021, avec une résolution horaire, et couvre l'ensemble du territoire de la Région de Bruxelles-Capitale.

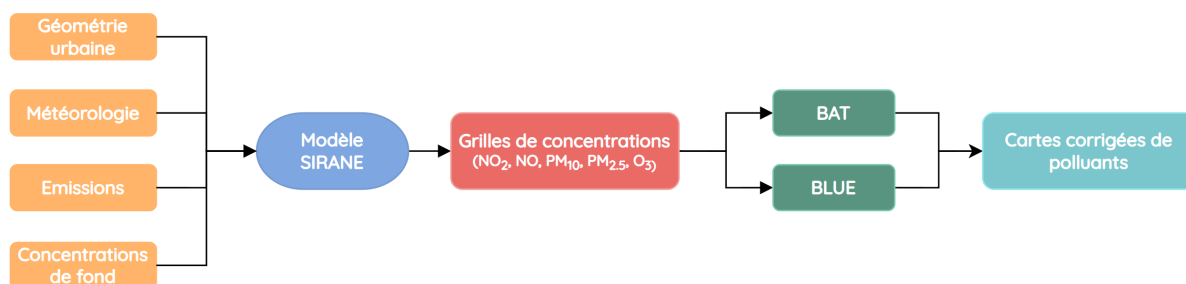


FIGURE 4 – Schéma du workflow général de la méthodologie mise en place

Cette section présentera d'abord le modèle SIRANE, son fonctionnement et les données nécessaires à son utilisation. Elle détaillera ensuite la méthodologie des deux méthodes d'assimilation de données mises en œuvre, ainsi que les jeux de données spécifiques qu'elles mobilisent.

3.1 Le modèle SIRANE

3.1.1 Origines et spécificités du modèle

Le modèle SIRANE est un modèle de dispersion atmosphérique en milieu urbain développé par le Laboratoire de Mécanique des Fluides et d'Acoustique (LMFA) de l'École Centrale de Lyon au début des années 2010. Son développement repose sur plus d'une décennie de recherches sur la modélisation de la dispersion dans les rues canyons, initiées par SOULHAC (2000). La première description complète et formalisée du modèle a été publiée en 2011 (SOULHAC et al., 2011).

Depuis, SIRANE est devenu un outil opérationnel utilisé par plusieurs agences de surveillance de la qualité de l'air, dont Bruxelles Environnement, notamment pour évaluer l'impact de mesures de gestion du trafic telles que l'instauration de la zone à faibles émissions (LEZ).

Ce modèle se distingue des modèles gaussiens classiques (type ADMS-Urban, AERMOD) et des modèles de canyon de rue isolé (type OSPM) par sa représentation explicite de l'ensemble du réseau de rues. Chaque rue est représentée comme un « canyon », soit une boîte de rue interconnectée aux autres via leurs intersections, ce qui permet de modéliser les transferts de polluants d'une rue à l'autre plutôt que de traiter chaque tronçon indépendamment. Cette approche en réseau de boîtes est plus détaillée que les modèles purement empiriques, tout en restant beaucoup moins coûteuse en calcul que les simulations CFD 3D complètes. Il s'agit ainsi d'un bon compromis méthodologique : il paramètre les principaux processus physiques de dispersion à l'échelle du quartier, rendant le modèle suffisamment rapide et opérationnel pour simuler des centaines de rues à l'échelle d'une ville à pas de temps horaires.

3.1.2 Structure générale du modèle

Par définition, le modèle SIRANE repose sur le concept de réseau de rues dans lequel l'ensemble du domaine urbain est discrétisé en tronçons de rue interconnectés par des intersections. Ce choix de représentation permet de modéliser explicitement les transferts de masse d'une rue à l'autre, tout en conservant un niveau de complexité compatible avec une utilisation opérationnelle à l'échelle d'une ville entière. Le modèle est structuré en deux modules distincts, chacun traitant un milieu physique différent : la canopée urbaine, qui correspond à l'espace compris entre le sol et le sommet des bâtiments, et la couche externe atmosphérique, située au-dessus des toits. Ces deux modules fonctionnent de manière indépendante pour le calcul des flux et de la dispersion dans leur domaine respectif, mais sont couplés via des échanges de masse : la canopée urbaine alimente la couche externe en polluants par advection et diffusion verticale, tandis que cette dernière peut, en retour, injecter de la pollution de fond dans les rues.

La Figure 5 illustre cette architecture au travers de quatre étapes successives : (a) la représentation d'un quartier urbain comme un réseau de tronçons de rue, (b) la modélisation de chaque tronçon par un modèle de boîtes bien mélangées et l'établissement du bilan de masse, (c) la prise en compte des échanges de polluants aux intersections en fonction de la direction du vent extérieur, et (d) la dispersion au-dessus des toits à l'aide d'un modèle de panache gaussien.

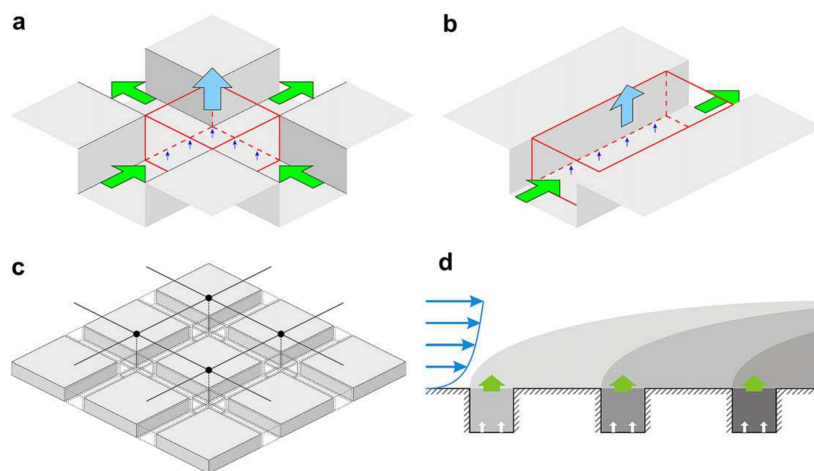


FIGURE 5 – Schéma conceptuel du modèle SIRANE tiré de SOULHAC et al. (2011)

Ce schéma permet aussi de comprendre le lien entre les échelles : les processus locaux au sein de chaque rue interagissent avec le transport à plus grande échelle dans la couche externe, garantissant une représentation cohérente de la dispersion urbaine. La description détaillée de chacun de ces modules est présentée dans les sections suivantes.

3.1.3 Le module de canopée urbaine

La canopée urbaine, dans SIRANE, correspond à l'ensemble des volumes compris entre le sol et la hauteur moyenne des toits. Elle est modélisée comme un réseau de boîtes de rue interconnectées, chacune représentant un canyon de rue. Dans cette représentation, on suppose que la concentration de polluants est uniforme à l'intérieur de chaque boîte. Le débit d'air longitudinal dans chaque rue est déterminé par la composante du vent extérieur parallèle à son axe : plus cette composante est importante, plus l'advection le long de la rue est efficace. Dans les rues perpendiculaires au vent, la ventilation est dominée par les échanges turbulents verticaux.

Trois processus principaux sont pris en compte dans ce module. Premièrement, l'advection longitudinale qui entraîne les polluants depuis l'amont vers l'aval du canyon, où ils peuvent être exportés vers la couche externe ou vers une intersection. Deuxièmement, le transfert vertical à l'interface avec la couche externe permet l'évacuation d'une partie des polluants au-dessus des toits, ou au contraire l'injection de pollution de fond dans la rue. Ce flux est paramétré à partir de coefficients de diffusion turbulente dépendant des conditions météorologiques (stabilité atmosphérique, vitesse de friction) et de la morphologie urbaine (hauteur des bâtiments). Troisièmement, les échanges aux intersections assurent la redistribution des masses entre rues adjacentes : une partie du flux issu d'une rue peut être transférée dans une rue perpendiculaire ou sortante, en fonction de la direction du vent et de la configuration géométrique du carrefour.

Ce traitement en réseau interconnecté permet de représenter des phénomènes essentiels, tels que la propagation de la pollution d'un axe fortement émetteur vers des rues voisines moins exposées, ou encore l'accumulation locale en conditions de faible ventilation. Les paramètres utilisés dans ce module sont dérivés d'études expérimentales en soufflerie et de simulations numériques, garantissant un réalisme suffisant tout en conservant une structure calculatoire légère adaptée à un usage opérationnel.

3.1.4 Le module de couche atmosphérique externe

La couche externe atmosphérique représente l'air circulant au-dessus de la canopée urbaine, depuis le niveau des toits jusqu'à la couche limite atmosphérique. Dans SIRANE, elle constitue le second module principal et traite le transport à plus grande échelle des polluants évacués des rues. Une fois libérés au-dessus des toits - soit à l'issue d'un canyon, soit par transfert vertical - ces polluants sont transportés par le vent synoptique et se dispersent horizontalement et verticalement selon un schéma de panache gaussien.

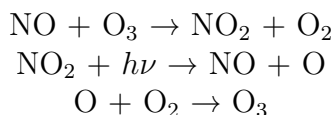
Les paramètres de diffusion du panache sont calculés à partir de la théorie de similarité de Monin-Obukhov, qui relie l'intensité de la turbulence et la stabilité atmosphérique aux conditions météorologiques observées (vitesse de vent, rayonnement solaire, température). Cette approche permet de quantifier l'élargissement du panache au fil de son parcours au-dessus de la ville, et donc d'estimer l'impact potentiel d'une source située en amont sur des zones plus éloignées.

Le vent dans la couche externe est supposé uniforme en vitesse et en direction sur l'ensemble du domaine modélisé, et constant durant chaque pas de temps horaire. Cette hypothèse, couramment adoptée dans les modèles urbains opérationnels, simplifie le calcul tout en capturant la dynamique principale du transport. Le module de couche externe gère également l'apport de pollution de fond provenant de l'extérieur du domaine simulé : ces concentrations régionales entrantes sont imposées aux frontières et peuvent ensuite être injectées dans les rues via les transferts verticaux, influençant la qualité de l'air locale même en l'absence de sources internes.

Il s'agit ainsi d'un module qui agit comme un intermédiaire entre la micro-échelle urbaine (rues et intersections) et la méso-échelle régionale. Il assure la cohérence entre les émissions locales, leur dispersion dans la ville, et l'influence des masses d'air extérieures, permettant à SIRANE de reproduire la variabilité spatiale des concentrations de fond tout en conservant un lien physique avec le réseau de rues simulé.

3.1.5 Intégration de la chimie atmosphérique

Un dernier point non négligeable du modèle SIRANE est qu'il intègre un module chimique simplifié destiné à représenter les réactions rapides entre oxydes d'azote (NO et NO₂) et ozone (O₃) dans l'atmosphère urbaine. Les réactions qu'il prend en compte correspondent au cycle photochimique rapide, ou cycle photo-stationnaire, qui assure la conversion réversible entre NO et NO₂ sous l'influence de la lumière solaire :



Ce cycle tend vers un équilibre dans lequel les concentrations relatives de NO, NO₂ et O₃ dépendent du taux de photolyse (lié à l'ensoleillement), des constantes de réaction et de la concentration d'ozone disponible en arrière-plan. Dans SIRANE, il est supposé que cet équilibre est atteint sur l'échelle du pas de temps horaire, ce qui est justifié par la rapidité des réactions en conditions diurnes. Les réactions plus lentes, telles que l'oxydation du NO₂ par OH ou la formation d'acide nitrique (HNO₃), sont négligées, car elles se déroulent sur des échelles de temps supérieures à l'horizon de modélisation de SIRANE.

3.1.6 Données d'entrée du modèle SIRANE

Le modèle SIRANE repose sur quatre catégories de données pour simuler la dispersion des polluants en milieu urbain : (i) la géométrie urbaine, (ii) la météorologie, (iii) la concentration de fond et (iv) les émissions. Dans le cadre de ce travail centré sur le post-traitement, l'ensemble de ces données d'entrée a été fourni par Bruxelles Environnement et l'UCLouvain, et utilisé telles quelles.

3.1.6.1 Géométrie urbaine

La géométrie urbaine est représentée par un réseau de rues : chaque rue est modélisée comme un segment rectiligne relié aux autres au niveau des intersections. Les rues courbes sont découpées en tronçons linéaires afin de satisfaire les contraintes géométriques du modèle et de faciliter la distinction entre rues-canyons, où les bâtiments limitent la dispersion, et rues ouvertes, davantage connectées à l'atmosphère libre. Dans ce travail, le fichier shapefile du réseau a été adapté et fourni par Axel Briffault, dans le prolongement du projet ExpAIR.

Outre la topologie du réseau, SIRANE requiert des paramètres de surface homogènes sur l'ensemble du domaine. Ceux-ci régissent les échanges énergétiques et aérodynamiques dans la canopée urbaine (GRIMMOND et OKE, 1999 ; OKE, 1988). Les paramètres retenus sont :

- l'albédo $[-]$ (0.275), la fraction du rayonnement solaire réfléchi par les surfaces urbaines ;
- l'émissivité $[-]$ (0.9805), la capacité des surfaces à émettre un rayonnement infrarouge ;
- le coefficient de Priestley–Taylor $[-]$ (1.0), utilisé pour estimer l'évapotranspiration potentielle ;
- la rugosité aérodynamique $[m]$ (0.9 m), soit la hauteur caractéristique des aspérités urbaines contrôlant la turbulence et le cisaillement ;
- l'épaisseur de déplacement $[m]$ (13 m), représentant la hauteur à laquelle le vent est ralenti significativement dû aux obstacles.

Ces paramètres relient directement la géométrie urbaine au régime de turbulence et conditionnent la capacité de dispersion des polluants dans les rues.

3.1.6.2 Données météorologiques

La météorologie contrôle à la fois le transport, la dispersion et certaines transformations chimiques. Dans SIRANE, elle est représentée par une description unidimensionnelle de la couche limite atmosphérique, supposée homogène horizontalement, à partir d'observations en un point de référence (SOULHAC et al., 2012). Dans ce travail, les données horaires de la station d'Uccle (fournies par l'Institut Royal Météorologique (IRM)) ont été utilisées : vitesse et direction du vent, température de l'air, précipitations et rayonnement solaire. Ces données sont considérées comme représentatives de la région bruxelloise aux échelles étudiées. Les paramètres de site (rugosité, épaisseur de déplacement, albédo, émissivité, coefficient de Priestley–Taylor) ont été harmonisés avec ceux du domaine urbain, conformément aux recommandations pour les environnements urbains tempérés (GRIMMOND et OKE, 1999 ; OKE, 1988).

3.1.6.3 Concentration de fond

La concentration de fond correspond à la pollution d'origine régionale et transfrontalière, indépendante des sources locales. Dans SIRANE, elle est ajoutée à chaque pas de temps en tout point du domaine (SOULHAC et al., 2017). Pour Bruxelles, les mesures horaires de NO_2 de la station d'Uccle (41R012, cf. Figure 6) ont été retenues, cette station étant éloignée des influences directes du trafic. La concentration de fond est supposée uniformément répartie sur l'ensemble du domaine d'étude.

3.1.6.4 Données d'émissions

Dans SIRANE, les émissions sont représentées selon trois classes : (i) linéaires (trafic routier), (ii) ponctuelles (sources localisées, telles que les sorties de tunnels) et (iii) surfaciques (chauffage résidentiel et tertiaire).

Émissions linéaires

Au niveau des émissions linéaires, liées au trafic routier, celles-ci sont obtenues par la combinaison de deux outils : (a) le modèle de trafic MuSTi (Modèle Urbain Stratégique de Transport Intermodal), développé par Bruxelles Environnement, qui fournit les flux et distances parcourues par segment et par classe de véhicules (calibrés sur des comptages datant de 2018) ; (b) le modèle d'émissions COPERT (Computer Programme to Calculate Emissions from Road Transport), développé par le Joint Research Centre (JRC) pour l'Agence européenne de l'environnement. COPERT prend en compte les principaux régimes d'émissions (à chaud, à froid, non-échappement : évaporation, abrasion pneus/freins) et fournit des facteurs détaillés selon le type de véhicule, la norme Euro, le carburant, la vitesse et le style de conduite.

En pratique, les kilomètres parcourus issus de MuSTi sont croisés avec les facteurs COPERT afin d'obtenir des émissions segmentées, ensuite réparties temporellement à l'échelle horaire.

Émissions ponctuelles

Concernant les émissions ponctuelles, il s'agit principalement des sorties de tunnels routiers, qui constituent des sources localisées de pollution. Chaque exutoire est représenté comme une source fixe, proportionnelle au trafic interne et au nombre de voies, calculés à partir de MuSTi et COPERT. Les rejets sont injectés au-dessus de la canopée et dispersés sous forme de panaches gaussiens, conformément aux pratiques décrites dans la littérature (de MÉVERGNIES, 2023 ; SOULHAC et al., 2012).

Émissions surfaciques

Pour les émissions surfaciques, la principale source correspond au chauffage des bâtiments résidentiels et tertiaires. Leur répartition spatiale est basée sur la surface bâtie issue des données cadastrales, tandis que leur modulation temporelle repose sur les degrés-jours équivalents, intégrant l'inertie thermique pour reproduire la saisonnalité du chauffage (de MÉVERGNIES, 2023). Les niveaux sectoriels hors trafic sont cohérents avec l'inventaire régional de Bruxelles Environnement.

3.2 Correction du modèle par assimilation de données

Parmi la large gamme de méthodes d'assimilation de données existantes, deux approches statistiques ont été retenues dans ce travail : la Bias Adjustment Technique (BAT) et le Best Linear Unbiased Estimator (BLUE). Cette sélection s'appuie sur la méthodologie développée par NGUYEN et SOULHAC (2021), qui les ont mises en œuvre pour la ville de Lyon.

La BAT constitue une méthode simple et peu coûteuse à mettre en place, appliquée a posteriori aux sorties brutes du modèle. Elle ne nécessite ni recalcul complet des champs de concentration ni modification du cœur physique du modèle, ce qui la rend particulièrement adaptée pour une première évaluation des performances de SIRANE, une prise en main des données et pour identifier la structure des écarts entre modèle et observations.

La méthode BLUE, quant à elle, est un schéma d'assimilation statistique optimal au sens des moindres carrés généralisés. Bien qu'issue du même formalisme que le 3D-Var, elle est ici appliquée dans sa formulation purement statistique, sans étape d'intégration temporelle. Ce choix se justifie par la nature des cartes produites — trimestrielles — qui ne requièrent pas de schéma dynamique comme un filtre de Kalman ou ses extensions (EnKF, EKF). La méthode BLUE présente aussi une complexité numérique bien plus faible que les méthodes variationnelles complètes (4D-Var), tout en offrant un cadre robuste pour combiner les prévisions du modèle et les observations en fonction de leurs incertitudes respectives.

3.2.1 Formalisme général

Dans le cadre de ce travail, l'assimilation de données consiste à combiner les concentrations issues du modèle SIRANE avec les observations fournies par le réseau de stations fixes, afin d'obtenir un champ corrigé des concentrations de NO_2 . Le problème à résoudre peut être formulé comme l'estimation d'un état atmosphérique vrai $\mathbf{x}^t$ (inconnu), à partir de deux sources d'information :

- L'ébauche $\mathbf{x}^b$ (background), de taille n , correspondant aux concentrations simulées par le modèle sur l'ensemble des n mailles du domaine.
- Les observations $\mathbf{y}$, de taille m , issues des m stations de mesure disponibles ($m \ll n$).

L'assimilation vise à produire un état analysé $\mathbf{x}^a$ (analysis), de taille $n \times 1$, qui combine ces deux sources de manière cohérente au regard de leurs incertitudes respectives.

Le lien entre l'espace du modèle et celui des observations est assuré par l'opérateur d'observation $\mathbf{H}$, de dimension $m \times n$. Cet opérateur extrait ou interpole les valeurs du modèle aux emplacements des stations de mesure. Ici, $\mathbf{H}$ est supposé linéaire, c'est-à-dire que cette projection s'effectue par combinaison linéaire des valeurs du modèle. On note ainsi $\mathbf{H}\mathbf{x}^b$ et $\mathbf{H}\mathbf{x}^t$ les concentrations issues respectivement de l'ébauche et de l'état réel, projetées aux emplacements des stations.

Le problème de l'assimilation de données consiste donc à ajuster $\mathbf{x}^b$ pour se rapprocher de $\mathbf{x}^t$, en exploitant l'information contenue dans $\mathbf{y}$ et en tenant compte du fait que ni le modèle ni les observations ne sont exempts d'erreurs.

3.2.2 Technique d'ajustement du biais (Bias Adjustment Technique — BAT)

La première méthode d'assimilation considérée est la *technique d'ajustement du biais* (Bias Adjustment Technique, BAT), ou débiaisement. Elle repose sur l'hypothèse que les observations sont non biaisées par rapport à l'état réel, soit :

$$\overline{y - Hx^t} = 0 \quad (2)$$

où x^t est l'état réel, y le vecteur d'observations et H l'opérateur d'observation. L'objectif de la BAT est de produire une analyse x^a telle que cette propriété soit vérifiée également pour l'analyse, c'est-à-dire :

$$\overline{x^t - x^a} = 0 \Leftrightarrow \overline{y - Hx^a} = 0 \quad (3)$$

Dans ce travail, deux variantes sont considérées : la BAT additive et la BAT multiplicative. Dans les deux cas, le biais est supposé spatialement uniforme sur le domaine et sur la période considérée, ce qui permet de le calculer à partir des seules stations disponibles. Les paramètres d'ajustement (ϵ pour l'additif et η pour le multiplicatif) sont choisis de manière à minimiser, sur l'ensemble des stations, le biais moyen de l'analyse vis-à-vis des observations.

Débiaisement additif

Pour l'ajustement du biais additif, la correction appliquée est de la sorte :

$$x^a = x^b + \epsilon = x^b + \frac{\sum_i^m w_i (y_i - (Hx^b)_i)}{\sum_i^m w_i} \quad (4)$$

Avec ϵ la variable de débiaisement, y_i la $i^{\text{ème}}$ mesure au $i^{\text{ème}}$ capteur et w_i la pondération associée. ϵ est ainsi choisi comme la moyenne pondérée du biais $y_i - (Hx^b)_i$, de sorte que l'analyse minimise le biais moyen par rapport aux observations.

Débiaisement multiplicatif

Dans le cas du débiaisement multiplicatif, l'assimilation de données suit l'équation suivante :

$$x^a = x^b + \eta = x^b + \frac{\sum_i^m w_i y_i}{\sum_i^m w_i (Hx^b)_i} \quad (5)$$

Dans l'Equation (5), η correspond à la variable d'ajustement et est aussi choisie de manière à minimiser l'erreur moyenne relative entre les observations et l'ébauche aux points d'observations.

Un avantage notable de cette approche est qu'elle préserve la positivité des concentrations dans l'analyse, contrairement à la BAT additive qui, dans certains cas de faible concentration, pourrait conduire à des valeurs négatives si ϵ est trop grand en valeur absolue.

3.2.3 Best Linear Unbiased Estimator (BLUE)

La seconde méthode d'assimilation de données utilisée dans ce travail est le *Best Linear Unbiased Estimator* (BLUE), une technique statistique visant à combiner de manière optimale une ébauche x^b et un ensemble d'observations y afin d'obtenir une analyse x^a minimisant la variance des erreurs d'analyse, tout en respectant deux contraintes fondamentales : la linéarité et l'absence de biais.

On définit les erreurs associées à l'ébauche, aux observations et à l'analyse comme :

$$e^b = x^b - x^t, \quad e^o = y - Hx^t \quad \text{et} \quad e^a = x^a - x^t \quad (6)$$

où x^t désigne l'état réel, H l'opérateur des observations, et y le vecteur des mesures. Dans ce cadre, l'hypothèse d'absence de biais s'exprime comme :

$$\overline{e^b} = 0, \quad \overline{e^o} = 0, \quad \overline{e^a} = 0 \quad (7)$$

Les deux premières relations supposent que l'ébauche et les observations ne présentent pas de biais moyen vis-à-vis de l'état réel. La troisième traduit la volonté de produire une analyse elle aussi non biaisée.

Sous ces hypothèses, BLUE exprime l'analyse sous la forme :

$$x^a = x^b + K(y - Hx^b) \quad (8)$$

où K est le gain de Kalman optimisant la pondération entre l'ébauche et les observations. Cette matrice est déterminée de manière à minimiser la variance des erreurs d'analyse A :

$$A = \overline{e^a(e^a)^T} \quad (9)$$

En tenant compte des hypothèses précédentes, le gain de Kalman optimal s'écrit :

$$K = BH^T (HBH^T + R)^{-1} \quad (10)$$

avec :

$$B = \overline{e^b(e^b)^T} \quad (11)$$

$$R = \overline{e^o(e^o)^T} \quad (12)$$

La matrice B représente la covariance des erreurs de l'ébauche, tandis que R correspond à la covariance des erreurs d'observation, incluant à la fois l'incertitude instrumentale et l'erreur associée à l'opérateur H .

Le rôle de K est donc de moduler la correction apportée à l'ébauche en fonction de la confiance accordée respectivement aux données modélisées et aux observations. Une faible incertitude sur les observations (valeurs faibles de R) conduit à une correction plus marquée, alors qu'une incertitude élevée réduit l'impact des observations sur l'analyse.

Lien avec la méthode variationnelle 3D-Var

Lorsque l'opérateur des observations H est supposé linéaire, la solution optimale fournie par BLUE est strictement équivalente à celle obtenue par la méthode variationnelle 3D-Var. Dans ce cas, la minimisation de la fonction coût variationnelle et la formulation BLUE aboutissent à la même expression pour le gain K et donc pour l'analyse x^a .

Matrice de covariance des erreurs d'observation R

La matrice R quantifie l'incertitude associée aux mesures issues des stations de surveillance de la qualité de l'air. Dans le cadre de la méthode BLUE, elle est essentielle pour pondérer la correction apportée à l'ébauche.

On suppose ici que les erreurs d'observation ne sont pas corrélées entre capteurs distincts, conformément à l'hypothèse :

$$\overline{e_i^o e_{j \neq i}^o} = 0 \quad (13)$$

Cette hypothèse est justifiée par le fait que chaque station utilise un appareil indépendant, avec des sources d'incertitudes propres. Ainsi, R est diagonale et seuls les termes de variance apparaissent sur la diagonale.

Contrairement à certaines approches qui imposent une incertitude uniforme pour toutes les stations, nous considérons ici que le niveau d'incertitude varie d'un capteur à l'autre, proportionnellement aux concentrations mesurées. Cette modélisation repose sur la directive européenne 2008/50/CE, qui fixe un seuil maximal d'incertitude de 15 % pour les mesures fixes de NO_2 . En supposant que les erreurs suivent une loi normale centrée, 95 % des valeurs se trouvent dans l'intervalle $[\mu - 1.96\sigma, \mu + 1.96\sigma]$. On en déduit que l'écart-type associé au capteur i est :

$$1.96 \sigma_i = 0.15 \bar{y}_i \quad (14)$$

où $\bar{y}_i = \frac{1}{T} \sum_{t=1}^T y_{i,t}$ est la concentration moyenne mesurée sur T pas de temps.

La variance associée est donc :

$$\sigma_i^2 = \left(\frac{0.15}{1.96} \bar{y}_i \right)^2 \quad (15)$$

et la matrice R s'écrit :

$$R = \text{diag}(\sigma_1^2, \sigma_2^2, \dots, \sigma_m^2) \quad (16)$$

Cette représentation, stationnaire dans le temps, permet de tenir compte à la fois des incertitudes propres à chaque station et de leur indépendance statistique.

Matrice de covariance des erreurs d'ébauche B

La matrice B , quant à elle, représente la covariance des erreurs d'ébauche :

$$B = \overline{e^b (e^b)^T} \quad (17)$$

où $e^b = x^b - x^t$ désigne l'erreur entre l'ébauche et l'état réel. Elle joue un rôle central dans la diffusion spatiale de l'information, en déterminant comment les corrections issues des observations sont propagées vers les zones non couvertes par les capteurs, selon les structures de corrélation supposées entre les erreurs du modèle.

Six formulations distinctes de B ont été testées dans cette étude, inspirées de NGUYEN et SOULHAC (2021), chacune reposant sur une hypothèse spécifique quant à la manière dont les erreurs sont corrélées.

1. Modèle $B(L)$ — Distance euclidienne

Ce modèle suppose que la corrélation entre les erreurs diminue exponentiellement avec la distance physique directe l_{ij} séparant deux points :

$$B_{ij} = \nu_0 \exp\left(-\frac{l_{ij}}{L_l}\right) \quad (18)$$

où ν_0 est la variance des erreurs en un point et L_l la longueur de corrélation spatiale.

2. Modèle $B(D)$ — Distance le long du réseau

Pour le modèle $B(D)$, la corrélation est fonction de la distance d_{ij} parcourue en suivant le réseau de rues, ce qui tient compte des contraintes géométriques imposées par le milieu urbain :

$$B_{ij} = \nu_0 \exp\left(-\frac{d_{ij}}{L_d} - \frac{|P_i - P_j|}{L_p + \beta \min(P_i, P_j)}\right) \quad (19)$$

où P_i et P_j sont les périmètres urbains locaux (paramètres décrivant la morphologie), L_p est une longueur de corrélation associée à ce périmètre et β un coefficient d'ajustement. Cette formulation devrait être plus réaliste pour les environnements urbains denses que le modèle $B(L)$.

3. Modèle $B(C)$ — Écart de concentration

Le modèle $B(C)$ repose sur l'hypothèse que les erreurs sont d'autant plus corrélées que les concentrations modélisées en deux points sont proches. On définit :

$$B_{ij} = \sqrt{\sigma_i^{2,z} \sigma_j^{2,z} \rho_{ij}^z} \quad (20)$$

avec

$$\begin{cases} \sigma_i^{2,z} = \alpha \sigma_i^{2,b} \\ \rho_{ij}^z = \rho_0 \exp\left(-\frac{\Delta C_{ij}}{L_c}\right) \end{cases} \quad (21)$$

où $\sigma_i^{2,b}$ est la variance d'ébauche au point i , α un facteur de pondération, ΔC_{ij} la différence absolue de concentration entre i et j (deux points), L_c une longueur de corrélation en concentration, et ρ_0 une corrélation maximale.

4. Modèle $B(r)$ — Corrélation structurelle

Dans ce cas, la corrélation est estimée à partir de la similarité temporelle des variations simulées sur une période donnée (corrélation de Pearson ρ_{ij}). On conserve la même forme générale que pour $B(C)$ avec :

$$B_{ij} = \sqrt{\sigma_i^{2,z} \sigma_j^{2,z} \rho_{ij}^z} \quad (22)$$

mais ρ_{ij}^z est exprimée différemment :

$$\begin{cases} \sigma_i^{2,z} = \alpha \sigma_i^{2,b} \\ \rho_{ij}^z = \rho_0 \exp\left(-\frac{1-\rho_{ij}}{L_p}\right) \end{cases} \quad (23)$$

où L_ρ contrôle la décroissance en fonction de la dissimilarité temporelle.

5. Modèle $B(LCr)$ — Hybride distance–concentration–structure

Ce cinquième modèle combine 3 des sous-modèles précédents ($B(L)$, $B(C)$ et $B(r)$), intégrant la distance euclidienne l_{ij} , l'écart de concentration ΔC_{ij} et la corrélation structurelle ρ_{ij} :

$$B_{ij} = \sqrt{\sigma_i^{2,z} \sigma_j^{2,z} \rho_{ij}^z} \quad (24)$$

avec

$$\begin{cases} \sigma_i^{2,z} = \alpha \sigma_i^{2,b} \\ \rho_{ij}^z = \rho_0 \exp \left(-\frac{\Delta C_{ij}}{L_c} - \frac{1-\rho_{ij}}{L_\rho} - \frac{l_{ij}}{L_l} \right) \end{cases} \quad (25)$$

6. Modèle $B(DCr)$ — Hybride réseau–concentration–structure

De même que $B(LCr)$, ce modèle combine 3 sous-modèles ($B(D)$, $B(C)$ et $B(r)$) mais en utilisant la distance le long du réseau d_{ij} au lieu de la distance euclidienne l_{ij} :

$$B_{ij} = \sqrt{\sigma_i^{2,z} \sigma_j^{2,z} \rho_{ij}^z} \quad (26)$$

avec

$$\begin{cases} \sigma_i^{2,z} = \alpha \sigma_i^{2,b} \\ \rho_{ij}^z = \rho_0 \exp \left(-\frac{\Delta C_{ij}}{L_c} - \frac{1-\rho_{ij}}{L_\rho} - \frac{d_{ij}}{L_d} - \frac{|P_i - P_j|}{L_p + \alpha \min(P_i, P_j)} \right) \end{cases} \quad (27)$$

Ces formulations permettent de tester différentes hypothèses de corrélation spatiale, depuis des modèles purement géométriques jusqu'à des approches hybrides intégrant la dynamique et les niveaux de concentration. Les paramètres (L_l , L_d , L_c , L_ρ , ν_0 , α , ρ_0 , etc.) sont optimisés pour chaque trimestre à l'aide du *diagnostic du χ^2* , qui permet de vérifier la cohérence entre les covariances d'erreurs supposées et les innovations. À chaque pas de temps t , on définit l'innovation comme :

$$d_t = y_t - H_t x_t^b \quad (28)$$

et on calcule :

$$\chi_t^2 = d_t^T (H_t B_t H_t^T + R_t)^{-1} d_t \quad (29)$$

où y_t est le vecteur d'observation, x_t^b l'ébauche, H_t l'opérateur d'observation, B_t la matrice de covariance des erreurs d'ébauche et R_t celle des erreurs d'observation au temps t . Sous l'hypothèse que B_t et R_t décrivent correctement les erreurs et que celles-ci suivent une loi gaussienne centrée, l'espérance mathématique de χ_t^2 est égale au nombre de capteurs m_t disponibles à l'instant t .

Pour juger la cohérence sur l'ensemble de la période, on calcule :

$$V_\chi = \frac{1}{T} \sum_{t=1}^T \frac{\chi_t^2}{m_t} \quad (30)$$

Une valeur $V_\chi \approx 1$ indique que la modélisation des erreurs est cohérente avec les innovations. Un $V_\chi \gg 1$ traduit une sous-estimation des erreurs, tandis qu'un $V_\chi \ll 1$ suggère une surestimation.

Dans ce travail, les paramètres des différents sous-modèles de B sont déterminés en minimisant l'écart $|V_\chi - 1|$, de façon à garantir que la structure et l'amplitude de B soient statistiquement cohérentes avec les observations.

3.2.4 Données nécessaires pour l'assimilation de données

Dans le cadre de cette méthodologie, l'assimilation de données repose sur les éléments suivants :

- Les **concentrations aux stations** : pour chaque heure, on dispose des valeurs de NO_2 simulées par SIRANE et des mesures correspondantes issues des stations fixes. Ces données sont organisées par station, chacune possédant un fichier horaire regroupant ses observations et les concentrations modélisées.
- Les **caractéristiques des stations** : pour chaque capteur sont renseignés l'identifiant, les coordonnées (x, y) en mètres, l'incertitude associée aux observations (15 %, utilisée dans la construction de la matrice R), ainsi que le poids attribué lors de la correction.
- Les **champs horaires de concentration** : issus directement des simulations SIRANE, ils fournissent les concentrations du polluant étudié (NO_2) sur l'ensemble du domaine, avec une résolution spatiale de 10 m.
- Le **réseau de voirie** : représenté sous forme de *shapefile*, constitué de segments de lignes (et non de polygones), chacun étant connecté aux autres pour assurer la continuité du réseau.

La Figure 6 illustre la localisation des 10 stations retenues pour l'assimilation par les méthodes BAT et BLUE sur la Région de Bruxelles-Capitale.



FIGURE 6 – Localisation des 10 stations utilisées dans les différentes techniques d'assimilation de données – Région de Bruxelles-Capitale

3.3 Critères d'évaluation de l'assimilation de données

L'évaluation de la performance de l'assimilation de données repose sur la comparaison des concentrations simulées (C_p) avec celles mesurées (C_m) aux stations de référence. Au-delà d'une simple comparaison visuelle, elle nécessite l'utilisation d'indicateurs statistiques permettant de caractériser l'existence d'une erreur systématique, l'ampleur moyenne des écarts ou encore la cohérence des variations temporelles.

Dans ce travail, cinq indicateurs couramment employés en modélisation de la qualité de l'air ont été retenus (BLOND et al., 2003 ; CHANG et HANNA, 2004 ; HANNA et CHANG, 2012 ; NGUYEN et SOULHAC, 2021) : le biais, le biais fractionnel, le *Root Mean Square Error* (RMSE), le *Normalized Mean Square Error* (NMSE) et le coefficient de corrélation linéaire.

Biais (erreur systématique moyenne)

$$\text{Biais} = \overline{C_m - C_p} \quad (31)$$

Le biais mesure la différence moyenne entre concentrations observées et simulées. Un $\text{Biais} > 0$ traduit une **sous-estimation** des concentrations par le modèle (les observations sont globalement supérieures aux prédictions), tandis qu'un $\text{Biais} < 0$ indique une **surestimation**. La valeur optimale est 0, ce qui correspond à l'absence de biais global. Simple et direct, cet indicateur a l'inconvénient qu'il ne renseigne pas sur la variabilité des erreurs : un biais nul peut masquer des erreurs ponctuelles importantes si elles se compensent en moyenne.

Biais fractionnel (biais normalisé et symétrique)

$$FB = \frac{2 \overline{C_m - C_p}}{\overline{C_m} + \overline{C_p}} \quad (32)$$

Le biais fractionnel (FB) est une forme normalisée du biais, variant entre -2 et $+2$. Une valeur positive traduit une sous-estimation, une valeur négative une surestimation. Sa normalisation par la moyenne des concentrations mesurées et simulées permet la comparaison entre stations présentant des gammes de concentrations différentes, ce qui le rend adapté aux réseaux de mesure hétérogènes.

Erreur quadratique moyenne (précision globale)

$$RMSE = \sqrt{\overline{(C_m - C_p)^2}} \quad (33)$$

L'erreur quadratique moyenne (RMSE) quantifie l'amplitude moyenne des écarts entre les valeurs observées et simulées. En élevant les écarts au carré avant de les moyenniser, il accorde davantage de poids aux erreurs importantes. Il combine les erreurs systématiques et aléatoires ; l'optimum est 0. Cet indicateur est pertinent pour apprécier la performance globale, mais ne distingue pas la nature de l'erreur.

Erreur quadratique normalisée (version adimensionnelle)

$$NMSE = \frac{\overline{(C_m - C_p)^2}}{\overline{C_m C_p}} \quad (34)$$

L'erreur quadratique normalisée (NMSE) est une version adimensionnelle du RMSE, normalisée par le produit moyen des concentrations mesurées et simulées. Il permet une

comparaison directe entre stations ou périodes aux niveaux de concentration différents. Une valeur faible indique un bon accord global, tandis qu'une valeur élevée peut résulter d'un biais important et/ou d'une dispersion élevée des écarts.

Coefficient de corrélation linéaire (cohérence des variations)

$$Corr = \frac{(\overline{C_m} - \overline{C_m})(\overline{C_p} - \overline{C_p})}{\sqrt{(\overline{C_m} - \overline{C_m})^2 (\overline{C_p} - \overline{C_p})^2}} \quad (35)$$

Le coefficient de corrélation évalue la capacité du modèle à reproduire les variations temporelles des concentrations observées. Un coefficient de corrélation égal à 1 indique une corrélation linéaire parfaite, s'il est égal à 0, il indique l'absence de relation linéaire et s'il vaut -1, une corrélation négative parfaite. Cet indicateur ne reflète pas l'existence d'un biais, mais il est essentiel pour juger de la fidélité de la dynamique simulée.

Seuils de qualité

En contexte urbain, les recommandations de HANNA et CHANG (2012), reprises par NGUYEN et SOULHAC (2021), fixent les repères suivants pour évaluer la pertinence d'un modèle de dispersion :

TABLE 1 – Seuils d'acceptabilité des indicateurs en milieu urbain

Indicateur	Seuil urbain	Optimum	Interprétation
Biais	$\leq 0.33 \times \overline{C_m}$	0	Faible erreur systématique moyenne
FB	≤ 0.67	0	Biais relatif limité
RMSE	$\leq \overline{C_m}$	0	Ampleur moyenne des écarts limitée
NMSE	≤ 6	0	Dispersion moyenne faible des écarts
Corr	≥ 0.60	1	Bonne reproduction des variations temporelles

Ces seuils doivent être interprétés comme des repères pratiques et non comme des valeurs absolues car il n'y a pas encore de consensus scientifique réel pour ces seuils. L'évaluation finale doit reposer sur une lecture conjointe des indicateurs, afin d'identifier la nature des écarts et la capacité du modèle à reproduire fidèlement la dynamique temporelle observée.

Validation croisée Leave-One-Out (LOO)

Toutes les statistiques présentées dans ce travail sont calculées après une validation croisée de type *Leave-One-Out* (LOO). Cette approche consiste à exclure successivement chaque station du jeu d'assimilation, exécuter la correction avec les stations restantes, puis calculer les indicateurs pour la station exclue. L'opération est répétée pour l'ensemble des stations, et les résultats finaux sont obtenus en moyennant les indicateurs. Cette méthode permet d'évaluer la capacité des corrections à prédire les concentrations sur un site qui n'a pas contribué à l'assimilation, assurant ainsi une estimation plus robuste des performances prédictives.

4 Résultats et discussions

Cette section présente les résultats obtenus avec les deux techniques d'assimilation de données mises en œuvre - la *Bias Adjustment Technique* (BAT) et le *Best Linear Unbiased Estimator* (BLUE) - ainsi que leur analyse comparative. Les performances respectives de ces approches sont discutées, avant de passer à la validation visant à évaluer leur capacité à améliorer les estimations issues du modèle SIRANE.

Des précisions sur l'implémentation pratique de ces méthodes, les contraintes computationnelles rencontrées, et les raisons ayant conduit au recours à l'outil YODA, développé par l'Ecole Centrale de Lyon, sont fournies en Annexe 1.

4.1 Techniques d'ajustement du biais (BAT)

La première méthode d'assimilation de données testée dans ce mémoire a été la technique d'ajustement du biais, aussi appelée débiaisement. De nature assez simple, elle consiste à supprimer le biais généré par l'ébauche, soit le biais du modèle SIRANE, vis-à-vis des observations prises aux stations de mesures. L'objectif devient ainsi de fournir une analyse qui ne soit pas biaisée par rapport à l'état réel. Pour faire cela, deux techniques ont été comparées : la BAT additive et la BAT multiplicative. Pour rappel, leurs équations simplifiées respectives sont les suivantes :

- Pour la BAT additive : $x^a = x^b + \epsilon$
- Pour la BAT multiplicative : $x^a = x^b \eta$

Avec x^a le vecteur d'analyse, x^b le vecteur d'ébauche, ϵ la variable de débiaisement et η la variable d'ajustement.

Dans le cas de cette méthode, le biais est supposé être spatialement uniforme, et les variables d'ajustement et de débiaisement sont sélectionnées comme étant celles garantissant un biais nul par rapport à l'état réel. Calculés initialement avec les cartes horaires extraites du modèle SIRANE, les résultats sont ensuite moyennés et agrégés pour obtenir une carte de concentration de NO₂ pour la période étudiée. Dans notre cas, il s'agit des 4 trimestres de l'année 2021. En fonction de la nature de la correction - additive ou multiplicative - son application se manifestera de manière distincte sur les cartes. Dans le cas de la BAT additive, le facteur de débiaisement ϵ sera ajouté ou soustrait en tous points de la carte. Pour la BAT multiplicative, tous les points de la carte seront multipliés par le facteur d'ajustement η . La Table 2 résume les facteurs de correction obtenus selon la méthode appliquée, en fonction des périodes étudiées en 2021.

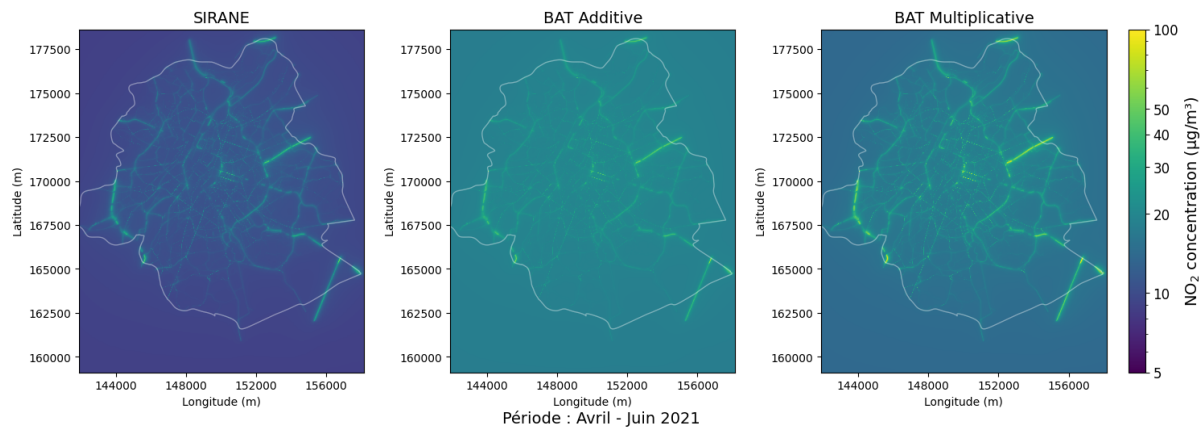
Période	Coefficient BAT Additive	Coefficient BAT Multiplicative
Janvier–Mars 2021	7.396	1.780
Avril–Juin 2021	8.748	2.532
Juillet–Septembre 2021	6.256	2.131
Octobre–Décembre 2021	3.680	1.647

TABLE 2 – Facteurs de correction obtenus pour les BAT additives et multiplicatives par période en 2021

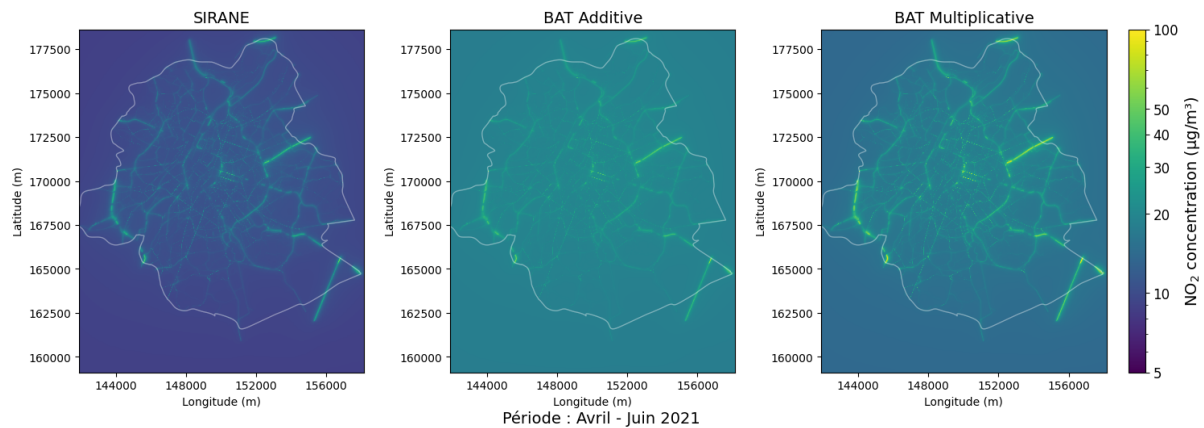
Au regard des coefficients moyens obtenus, plusieurs remarques s'imposent. Dans le cas de la correction additive, on ajoute uniformément entre 3.680 et 8.748 $\mu\text{g}/\text{m}^3$ de NO_2 sur l'ensemble des points de la carte. En revanche, la correction multiplicative (BAT multiplicative) consiste à multiplier chaque pixel par un facteur compris entre 1.647 et 2.532. Cette différence a des implications importantes sur la distribution spatiale des concentrations corrigées.

Dans le cas de la correction multiplicative, comme les résultats sont supérieurs à 1, elle amplifie en particulier les concentrations les plus élevées : les zones déjà fortement polluées voient ainsi leurs valeurs croître de manière significative. À l'inverse, la correction additive affecte davantage les zones faiblement polluées, car une même quantité de NO_2 est ajoutée indépendamment du niveau initial. L'impact relatif de cette correction sera donc plus marqué dans les secteurs les moins exposés de Bruxelles.

Les résultats obtenus par trimestres singuliers pour les deux méthodes sont données aux Figures 7a à 7d.

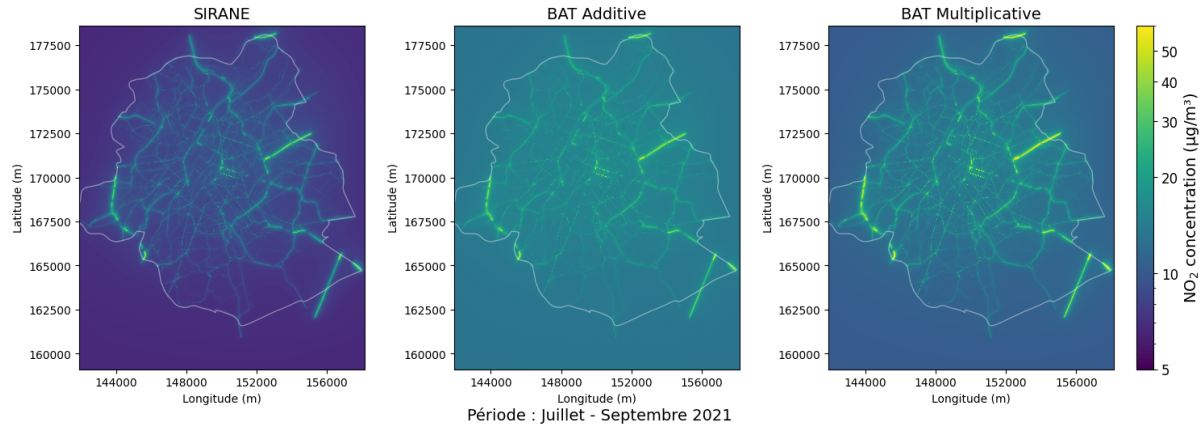


(a) Janvier – Mars 2021

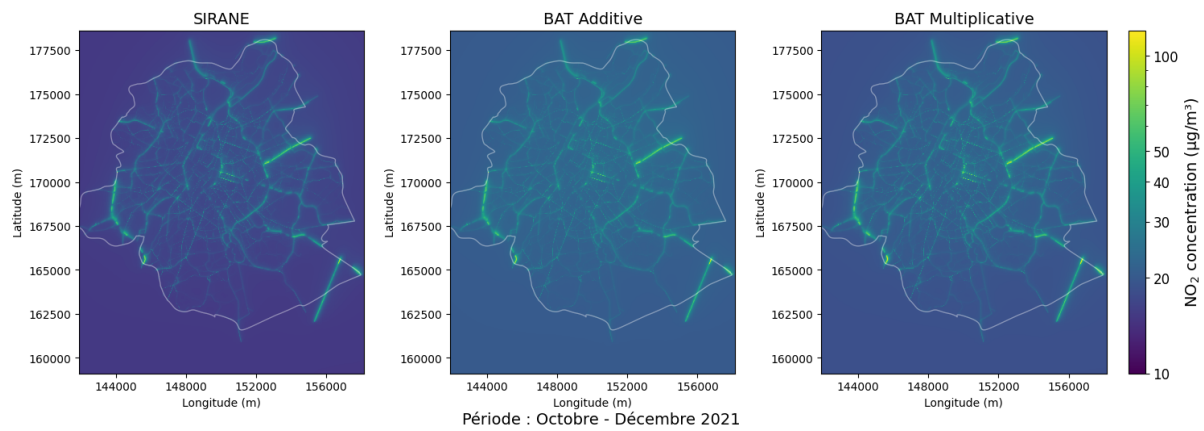


(b) Avril – Juin 2021

FIGURE 7 – Évolution des concentrations de NO_2 simulées par SIRANE et corrigées selon les méthodes BAT additive et multiplicative (1/2)



(c) Juillet – Septembre 2021



(d) Octobre – Décembre 2021

FIGURE 7 – Évolution des concentrations de NO_2 simulées par SIRANE et corrigées selon les méthodes BAT additive et multiplicative (2/2)

Bien que les différences entre les corrections additive et multiplicative soient visuellement atténuées par l'échelle logarithmique utilisée dans les figures 7a à 7d, des tendances nettes se dégagent. Comme attendu, la correction multiplicative amplifie fortement les concentrations le long des axes routiers, où les niveaux étaient déjà élevés dans les simulations initiales produites par le modèle SIRANE.

À l'inverse, la correction additive produit, quelle que soit la période, un rehaussement uniforme du champ de concentration. Son effet est donc plus marqué dans les zones initialement peu polluées, où l'ajout d'une valeur constante a un impact relatif plus important.

Il convient également de souligner que, contrairement à la correction multiplicative, la correction additive ne garantit pas la positivité des concentrations simulées. Néanmoins, dans notre cas, aucune valeur négative simulée n'a été relevée. Si cela avait été le cas, un seuillage à zéro aurait été appliqué, conformément aux pratiques courantes visant à éviter les incohérences physiques.

Ainsi, les deux méthodes testées pour débiaiser le modèle SIRANE présentent des comportements distincts, en accord avec les principes théoriques qui les sous-tendent. Les effets

observés sur la répartition spatiale de la pollution illustrent clairement comment le choix de la correction influe sur les concentrations simulées.

4.1.1 Analyse statistique

Statistiques globales

La Table 3 présente les performances statistiques du modèle SIRANE corrigé par les méthodes BAT additive et multiplicative pour les quatre trimestres de l'année 2021. Comme déjà énoncé dans la méthodologie, l'ensemble de ces statistiques reposent sur une validation croisée, effectuée en excluant successivement un capteur à chaque pas de temps, dans le but de tester la robustesse de l'assimilation sur des points non utilisés lors de l'ajustement.

TABLE 3 – Comparaison des performances du modèle SIRANE corrigé par les méthodes BAT additive et multiplicative pour chaque trimestre de 2021

Période	Méthode	$\overline{C_{mes}}$ [$\mu\text{g}/\text{m}^3$]	$\overline{C_{mod}}$ [$\mu\text{g}/\text{m}^3$]	Biais [$\mu\text{g}/\text{m}^3$]	Biais fract	RMSE [$\mu\text{g}/\text{m}^3$]	NMSE	Corr
Janvier – Mars 2021	SIRANE		21.767	-7.420	-0.281	13.947	0.311	0.760
	BAT Add	29.185	29.159	-0.027	0.018	10.781	0.152	0.836
	BAT Mult		29.464	0.278	0.013	11.735	0.163	0.836
Avril – Juin 2021	SIRANE		14.186	-8.630	-0.436	14.014	0.629	0.717
	BAT Add	22.817	22.884	0.066	0.056	10.890	0.285	0.794
	BAT Mult		23.557	0.740	0.038	13.594	0.353	0.748
Juillet – Septembre 2021	SIRANE		12.652	-6.190	-0.384	12.561	0.710	0.669
	BAT Add	18.843	18.921	0.076	0.049	10.455	0.376	0.750
	BAT Mult		19.620	0.778	0.027	12.389	0.422	0.732
Octobre – Décembre 2021	SIRANE		23.657	-3.713	-0.143	13.623	0.293	0.787
	BAT Add	27.371	27.335	-0.034	0.012	13.144	0.251	0.796
	BAT Mult		28.224	0.854	0.015	14.366	0.274	0.793

De manière générale, les deux méthodes permettent une nette amélioration des performances du modèle par rapport à la version brute, obtenue par SIRANE. On observe systématiquement une réduction significative du biais moyen, du biais fractionnel, de l'erreur quadratique moyenne (RMSE), et de l'erreur quadratique normalisée (NMSE). La corrélation entre les concentrations simulées et mesurées est également améliorée dans tous les cas, ce qui confirme la pertinence d'une correction a posteriori.

L'analyse comparative des deux méthodes met toutefois en évidence que la BAT additive performe légèrement mieux, de manière globale, que la BAT multiplicative. En effet, celle-ci présente systématiquement des biais moyens plus proches de zéro, un RMSE inférieur, et un NMSE plus faible que la méthode multiplicative, et ce pour l'ensemble des périodes analysées. Par exemple, au second trimestre (avril–juin 2021), la BAT additive ramène le biais moyen à 0.066 $\mu\text{g}/\text{m}^3$ contre 0.740 $\mu\text{g}/\text{m}^3$ pour la correction multiplicative, avec un RMSE réduite de 13.6 à 10.9 $\mu\text{g}/\text{m}^3$.

Par ailleurs, les coefficients de corrélation sont soit équivalents, soit légèrement supérieurs dans le cas de la correction additive. Cette méthode semble donc plus efficace pour améliorer simultanément la précision moyenne et la structure spatiale des concentrations simulées, sans introduire de sur-correction dans les zones à forte pollution. Cela est cohérent avec le fait que la correction additive applique une translation constante sur l'ensemble

du champ de concentration, ce qui compense efficacement les biais systématiques tout en limitant les distorsions locales.

La correction multiplicative, quant à elle, tend à accentuer les contrastes spatiaux en amplifiant davantage les valeurs déjà élevées. Elle peut ainsi sur-corriger les zones fortement polluées tout en restant moins efficace dans les zones à faible concentration initiale. Cette propriété se traduit par un biais légèrement positif résiduel et une erreur quadratique plus élevée.

Statistiques au niveau des capteurs

Afin d'approfondir l'analyse statistique, les mêmes indicateurs de performance ont été calculés individuellement pour chaque station de mesure. Cette approche permet d'évaluer plus finement l'efficacité des corrections, en fonction du type d'environnement dans lequel sont réalisées les observations (milieu urbain, industriel, etc.), ainsi que l'influence du trafic sur les mesures et les corrections. Les Tables 4 et 5 présentent ces résultats pour les deux types de correction appliqués au cours du premier trimestre de 2021, et permettent ainsi de comparer la performance des méthodes selon les contextes locaux. Les 6 autres tableaux, pour les autres périodes, sont consultables en Annexe 2.

TABLE 4 – Statistiques par capteur pour la correction BAT additive – Janvier à Mars 2021

Capteur	Milieu	Trafic	$\overline{C_{mes}}$ [$\mu\text{g}/\text{m}^3$]	$\overline{C_{mod}}$ [$\mu\text{g}/\text{m}^3$]	Biais [$\mu\text{g}/\text{m}^3$]	Biais fract	RMSE [$\mu\text{g}/\text{m}^3$]	NMSE	Corr
41R001	Urbain	Modéré	29.85	28.19	-1.66	-0.06	8.15	0.08	0.89
41R002	Urbain	Fort	34.04	29.27	-4.77	-0.15	9.67	0.09	0.89
41B001	Urbain	Très fort	41.84	39.02	-2.82	-0.07	22.08	0.30	0.61
41B004	Urbain	Modéré	30.71	25.01	-5.70	-0.20	8.30	0.09	0.93
41B011	Urbain	Très faible	19.96	25.20	5.24	0.23	9.43	0.18	0.83
41R012	Urbain	Très faible	16.39	25.95	9.56	0.45	12.86	0.39	0.76
41N043	Industriel	Modéré	28.75	29.29	0.54	0.02	8.20	0.08	0.88
41MEU1	Urbain	Faible	25.13	25.10	-0.03	0.00	8.11	0.10	0.86
41B006	Urbain	Faible	27.89	25.75	-2.15	-0.08	7.47	0.08	0.89
41B008	Urbain	Très fort	37.29	38.81	1.52	0.04	13.54	0.13	0.82

TABLE 5 – Statistiques par capteur pour la correction BAT multiplicative – Janvier à Mars 2021

Capteur	Milieu	Trafic	$\overline{C_{mes}}$ [$\mu\text{g}/\text{m}^3$]	$\overline{C_{mod}}$ [$\mu\text{g}/\text{m}^3$]	Biais [$\mu\text{g}/\text{m}^3$]	Biais fract	RMSE [$\mu\text{g}/\text{m}^3$]	NMSE	Corr
41R001	Urbain	Modéré	29.85	27.70	-2.15	-0.07	8.65	0.09	0.87
41R002	Urbain	Fort	34.04	29.57	-4.46	-0.14	9.65	0.09	0.88
41B001	Urbain	Très fort	41.84	45.87	4.03	0.09	29.83	0.46	0.54
41B004	Urbain	Modéré	30.71	23.68	-7.03	-0.26	9.45	0.12	0.92
41B011	Urbain	Très faible	19.96	23.27	3.31	0.15	7.88	0.13	0.86
41R012	Urbain	Très faible	16.39	24.10	7.70	0.38	10.29	0.27	0.84
41N043	Industriel	Modéré	28.75	29.72	0.97	0.03	8.29	0.08	0.88
41MEU1	Urbain	Faible	25.13	23.56	-1.58	-0.06	8.23	0.11	0.86
41B006	Urbain	Faible	27.89	24.40	-3.49	-0.13	7.38	0.08	0.92
41B008	Urbain	Très fort	37.29	42.77	5.48	0.14	17.70	0.20	0.79

Sans surprise, les deux méthodes améliorent les résultats provenant du modèle SIRANE, mais les effets ne se traduisent pas de la même manière sur tous les sites.

Lorsque les deux techniques sont comparées en regardant les statistiques au niveau des capteurs, la BAT additive semble fournir des résultats globalement plus qualitatifs que la BAT multiplicative, fournissant des coefficient de corrélation quasi systématiquement plus haut que pour la seconde méthode. Certains capteurs montrent une réduction marquée

du RMSE et une amélioration du coefficient de corrélation, tout particulièrement dans les zones où le trafic a une influence faible à modérée, comme c'est le cas pour 41MEU1, 41B006 ou encore 41B004. Ce comportement est cohérent avec la nature de la correction additive, qui induit un rehaussement uniforme des concentrations et tend à mieux corriger les sous-estimations dans les zones de fond urbain.

À l'inverse, la méthode BAT multiplicative présente des performances plus contrastées. Elle améliore légèrement les estimations sur certains capteurs (par exemple 41R002 ou 41N043), mais tend à surestimer les concentrations dans les zones déjà fortement polluées. C'est notamment le cas des stations 41B001 et 41B008, pour lesquelles le biais devient fortement positif, avec une augmentation sensible du RMSE (jusqu'à 29.83 $\mu\text{g}/\text{m}^3$ pour 41B001). De nouveau, ce comportement est attendu, car la correction multiplicative accentue les fortes valeurs initiales du modèle, ce qui peut conduire à des surestimations localisées.

4.1.2 Discussion des résultats de l'assimilation de données avec la méthode BAT

Les résultats obtenus à l'aide des méthodes de débiaisement (additive et multiplicative) révèlent l'intérêt d'une correction a posteriori des champs simulés par le modèle SIRANE. Ces techniques, bien que simples, permettent une amélioration notable de la performance statistique du modèle, notamment en ce qui concerne le biais moyen, l'erreur quadratique moyenne (RMSE) et la corrélation avec les observations. Cependant, plusieurs éléments méritent d'être discutés en profondeur afin de cerner les limites inhérentes à ces approches.

Une efficacité globale, mais contrastée selon les méthodes et les contextes locaux

Sur l'ensemble de l'année 2021, la correction additive s'est montrée légèrement plus efficace que la correction multiplicative pour améliorer les scores globaux (voir Table 3). Cette tendance est également visible dans l'analyse par station : la BAT additive fournit généralement des résultats plus équilibrés, avec une réduction plus homogène du biais et du RMSE, notamment dans les zones présentant un niveau de trafic faible à modéré. Cela suggère que cette méthode est bien adaptée pour corriger les sous-estimations systématiques du modèle dans les milieux urbains peu pollués, où les valeurs simulées sont souvent trop faibles. Cette performance tient à la nature même de la correction : en appliquant un ajustement constant, elle agit prioritairement sur les zones sous-évaluées, sans trop distordre la structure spatiale du champ de concentration.

À l'inverse, la BAT multiplicative, bien qu'efficace dans certains cas (comme pour les capteurs 41R002 ou 41N043), tend à surcorriger les zones déjà fortement polluées. C'est notamment le cas des stations 41B001 et 41B008, situées en zone de trafic intense, où l'on observe une inversion du biais : la correction additive maintenait une légère sous-estimation, tandis que la correction multiplicative conduit à une surestimation marquée, accompagnée d'une hausse significative du RMSE.

Limites conceptuelles des corrections uniformes

De manière plus générale, ces résultats soulignent les limites des approches BAT fondées sur une correction uniforme - qu'elle soit additive ou multiplicative. Ces techniques ne tiennent pas compte des spécificités locales du modèle, des erreurs spatialisées, etc. Elles corrigent l'ébauche de manière globale, mais sans distinguer les mécanismes sous-jacents à l'erreur du modèle. Or, les erreurs de SIRANE sont dues à une combinaison de facteurs : la complexité de la topographie urbaine, la précision des émissions en entrée, les modèles utilisés initialement, la dispersion atmosphérique, etc.

Dès lors, appliquer un filtre correcteur unique à l'ensemble du domaine revient à ignorer cette complexité. Cela explique pourquoi certaines zones sont sur-corrigées tandis que d'autres restent sous-estimées. Pour surmonter ces limitations, des méthodes d'assimilation plus avancées, capables d'adapter localement la correction en fonction de la structure de l'erreur et des observations disponibles, apparaissent nécessaires. Une de ces méthodes sera explorée, avec l'utilisation de la méthode BLUE.

.

4.2 Best Linear Unbiased Estimator (BLUE)

La seconde méthode d'assimilation de données implémentée dans ce travail est le Best Linear Unbiased Estimator (BLUE). Comme son nom l'indique, cette approche vise à produire le meilleur estimateur linéaire non biaisé, en minimisant la variance de l'erreur d'analyse. Autrement dit, l'analyse résultante combine linéairement l'ébauche (simulation du modèle) et les observations, en pondérant leur contribution respective en fonction de leurs incertitudes.

Le principe fondamental de cette méthode repose sur l'hypothèse que les erreurs du modèle (ébauche) et celles des observations sont centrées (non biaisées), et statistiquement caractérisées par leurs matrices de covariance respectives. L'équation 36 formalise cette combinaison linéaire :

$$x^a = x^b + K(y - Hx^b) = x^b + BH^T(HBH^T + R)^{-1}(y - Hx^b) \quad (36)$$

Avec, x^a le vecteur d'analyse, x^b le vecteur d'ébauche, K le gain de Kalman, y le vecteur des observations, H l'opérateur d'observation, R la matrice de covariance des erreurs d'observations et B la matrice de covariance des erreurs d'ébauche.

Comme pour la méthode BAT, l'assimilation est effectuée à pas de temps horaire, puis les résultats sont agrégés et moyennés par trimestre pour l'année 2021. Cependant, contrairement aux corrections simples utilisées précédemment, la méthode BLUE repose sur une modélisation explicite de l'incertitude, à travers les matrices R et B . Ces deux matrices jouent un rôle central : elles conditionnent le poids attribué aux observations et à l'ébauche via le calcul du gain de Kalman, soit la matrice K . Leur définition et leur impact sur la correction seront donc examinés dans les sections suivantes.

4.2.1 Matrice R

La matrice R représente la matrice de covariance des erreurs d'observation. Dans le cadre de la méthode BLUE, elle joue un rôle central en quantifiant l'incertitude associée aux mesures issues des stations de surveillance de la qualité de l'air.

Dans ce travail, nous faisons l'hypothèse que les erreurs d'observation ne sont pas corrélées entre capteurs. Cette hypothèse est justifiée par le fait que chaque station de mesure utilise un appareil distinct, avec des sources d'incertitudes propres. Il en résulte que R est une matrice diagonale, où seuls les termes de variance (diagonaux) sont non nuls.

Contrairement à certaines approches qui supposent une incertitude uniforme pour toutes les stations, nous considérons ici que les erreurs de mesure varient selon les capteurs, en fonction du niveau de pollution mesuré, et donc des concentrations des polluants étudiés.

Cette modélisation repose également sur la directive européenne 2008/50/CE, qui impose un seuil maximal d'incertitude de 15 % pour les mesures fixes de NO_2 . En supposant que les erreurs suivent une loi normale centrée, 95 % des erreurs doivent ainsi être inférieures à 15 % de la valeur moyenne mesurée. Or, pour une distribution normale, 95 % des valeurs se situent dans l'intervalle $[\mu - 1.96\sigma, \mu + 1.96\sigma]$, ce qui permet de relier l'écart-type σ_i à la concentration moyenne mesurée par le capteur i :

$$1.96 \sigma_i = 0.15 \cdot \left(\frac{1}{T} \sum_{t=1}^T y_{i,t} \right) = 0.15 \cdot \bar{y}_i \quad (37)$$

où $y_{i,t}$ est la concentration mesurée par le capteur i au temps t , et T le nombre total de pas de temps.

La variance associée au capteur i est donc donnée par :

$$\sigma_i^2 = \left(\frac{0.15}{1.96} \bar{y}_i \right)^2 \quad (38)$$

Ainsi, la matrice R est définie comme :

$$R = \text{diag}(\sigma_1^2, \sigma_2^2, \dots, \sigma_m^2) \quad (39)$$

Cette modélisation permet de tenir compte à la fois du niveau d'incertitude propre à chaque capteur, et du fait que les erreurs de mesure sont non corrélées. Elle garantit une représentation plus réaliste des erreurs d'observation et permet un ajustement plus pertinent lors de l'assimilation des données par la méthode BLUE.

À titre d'exemple, la Table 6 détaille la matrice de covariance des erreurs d'observation R estimée pour le premier trimestre de 2021.

TABLE 6 – Matrice R – janvier à mars 2021

Capteur	41R001	41R002	41B001	41B004	41B011	41R012	41N043	41MEU1	41B006	41B008
41R001	5.22	0	0	0	0	0	0	0	0	0
41R002	0	6.78	0	0	0	0	0	0	0	0
41B001	0	0	10.25	0	0	0	0	0	0	0
41B004	0	0	0	5.52	0	0	0	0	0	0
41B011	0	0	0	0	2.33	0	0	0	0	0
41R012	0	0	0	0	0	1.57	0	0	0	0
41N043	0	0	0	0	0	0	4.84	0	0	0
41MEU1	0	0	0	0	0	0	0	3.70	0	0
41B006	0	0	0	0	0	0	0	0	4.56	0
41B008	0	0	0	0	0	0	0	0	0	8.15

Découlant des hypothèses formulées concernant la nature des erreurs d'observation (voir Equations (38) et (39)) les stations fortement influencées par le trafic, comme 41B001 et 41B008, présentent les plus fortes variances d'erreur (respectivement 10.25 et 8.15). Ce résultat reflète la dépendance directe de l'incertitude à l'intensité de la pollution observée, conduisant à une pondération plus faible de ces stations dans l'analyse finale. À l'inverse, des capteurs situés dans des zones moins exposées, tels que 41R012 (1.57) ou 41B011 (2.33), sont associés à des erreurs moindres et exercent ainsi une influence plus importante dans l'assimilation.

Cette configuration est également cohérente avec l'idée que les stations de fond - situées dans des environnements moins pollués et caractérisés par des conditions plus stables (comme 41R012) - sont généralement moins affectées par des effets locaux complexes et constituent ainsi des points de référence plus fiables pour contraindre l'analyse.

4.2.2 Matrice B

Dans le cadre de la méthode BLUE, la matrice de covariance des erreurs d'ébauche B joue un rôle central dans la diffusion spatiale de l'information. Elle permet d'étendre les corrections issues des observations vers les zones non couvertes, en fonction des structures de corrélation supposées entre les erreurs du modèle.

Dans cette étude, six formulations distinctes de la matrice de covariance des erreurs d'ébauche B ont été testées, chacune reposant sur une hypothèse spécifique concernant la structure spatiale des erreurs. Ces formulations sont notées comme suit :

- **$B(L)$ – Distance euclidienne** : les corrélations décroissent exponentiellement avec la distance physique directe entre deux points.
- **$B(D)$ – Distance le long du réseau de rue** : similaire au modèle précédent, mais la distance utilisée est celle suivie dans le réseau de voirie, ce qui reflète mieux les contraintes urbaines.
- **$B(C)$ – Écart de concentration** : les corrélations diminuent avec l'écart absolu entre les concentrations modélisées en deux points, supposant que des niveaux de pollution proches impliquent des erreurs similaires.
- **$B(r)$ – Corrélation structurelle** : les corrélations d'erreur sont basées sur la similitude statistique (corrélation de Pearson) entre les variations des concentrations simulées sur une période donnée. Ce modèle prend en compte la dynamique conjointe des concentrations.
- **$B(LCr)$ – Hybride distance–concentration–structure** : combinaison du modèle euclidien L , de l'écart de concentration C et de la corrélation structurelle r , avec des pondérations spécifiques.
- **$B(DCr)$ – Hybride réseau–concentration–structure** : même principe que le précédent, mais la distance est ici calculée selon le réseau de voirie (modèle D).

Les expressions mathématiques associées à ces six formulations sont énoncées dans la section 3.2.3. Les paramètres spécifiques à chaque formulation, (longueurs de corrélation, pondérations, etc.) ont été déterminés pour chaque trimestre de 2021 par validation croisée. Cette procédure vise à minimiser l'erreur quadratique moyenne (RMSE) sur les capteurs exclus à chaque pas de temps, via l'optimisation du critère χ^2 . Ainsi, les paramètres optimisés en utilisant le sous-modèle $B(LCr)$, définis dans l'Equation (24) pour les différentes périodes de 2021, sont donnés à la Table 7.

TABLE 7 – Paramètres optimisés pour le modèle $B(LCr)$ selon les 4 trimestres de l'année 2021

Période	L_d [km]	L_c [$\mu\text{g}/\text{m}^3$]	ν	L_r
Janvier – Mars 2021	6	50	20	0.40
Avril – Juin 2021	8	25	200	0.35
Juillet – Septembre 2021	20	20	150	0.30
Octobre – Décembre 2021	2	50	250	0.70

Afin d'illustrer l'influence des différentes formulations de la matrice de covariance des erreurs d'ébauche B sur le comportement spatial de la correction apportée par la méthode BLUE, les Figures 9 à 11 présentent, pour chacun des quatre trimestres de l'année 2021,

les champs de concentration de NO_2 corrigés selon les six modèles testés $B(L)$, $B(D)$, $B(C)$, $B(r)$, $B(LCr)$, $B(DCr)$.

Ces cartes permettent de visualiser concrètement la manière dont les structures de corrélation imposées par chaque formulation influencent la diffusion spatiale des corrections. Alors que les modèles basés sur la distance ($B(L)$, $B(D)$) tendent à propager les corrections de manière plus uniforme autour des points d'observation, les modèles fondés sur les caractéristiques du champ simulé ($B(C)$, $B(r)$) produisent des champs plus hétérogènes, avec des corrections souvent localisées le long des axes de forte pollution. Les formulations hybrides ($B(LCr)$ et $B(DCr)$) cherchent quant à elles un compromis entre la diffusion spatiale et le respect des structures du champ d'ébauche.

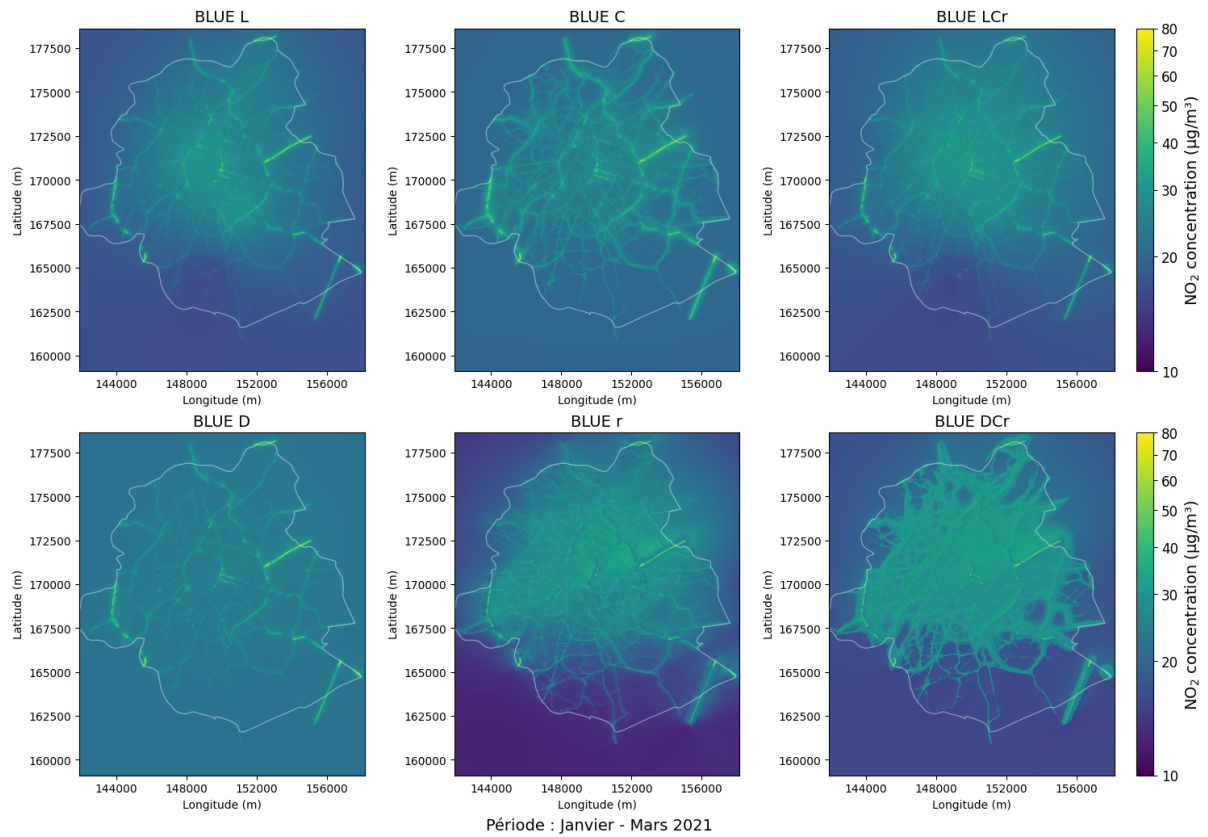


FIGURE 8 – Champs de concentrations de NO_2 corrigées par la méthode BLUE pour le premier trimestre 2021, selon les six formulations de la matrice B

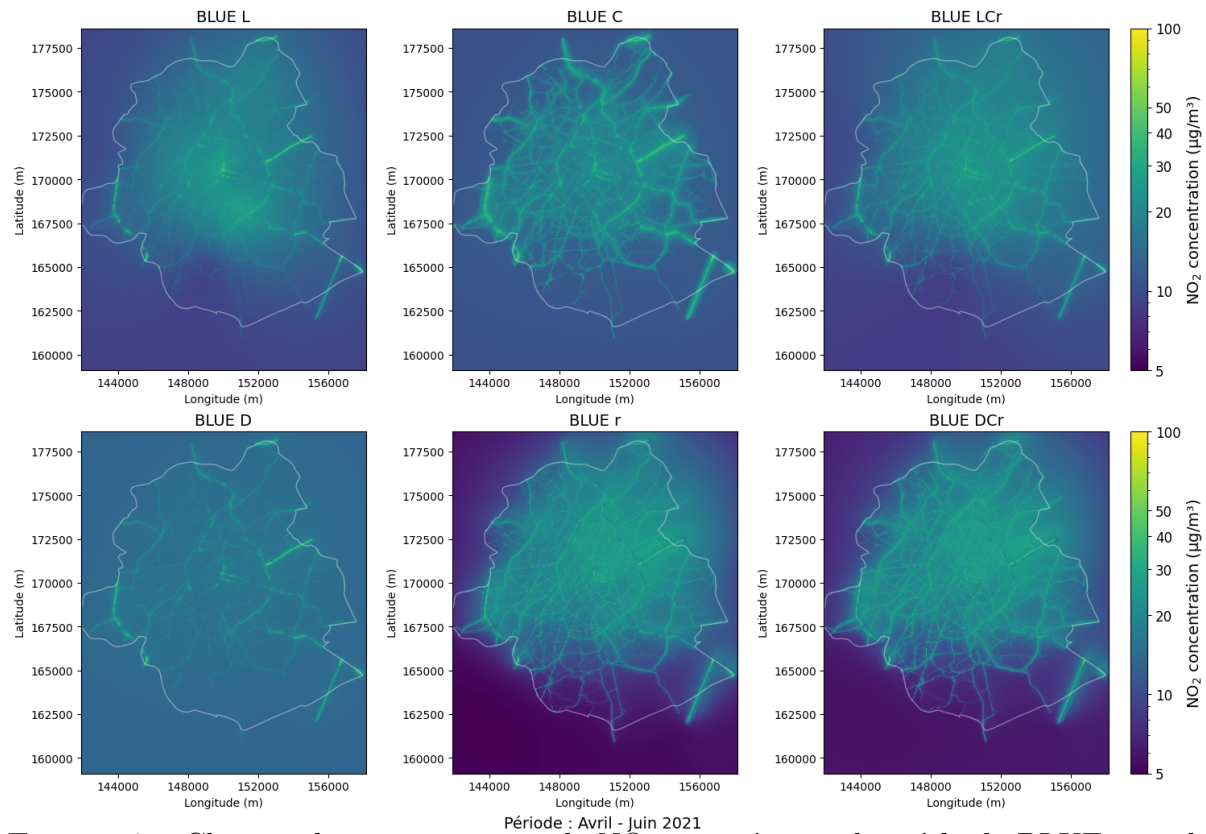


FIGURE 9 – Champs de concentrations de NO_2 corrigées par la méthode BLUE pour le deuxième trimestre 2021, selon les six formulations de la matrice B

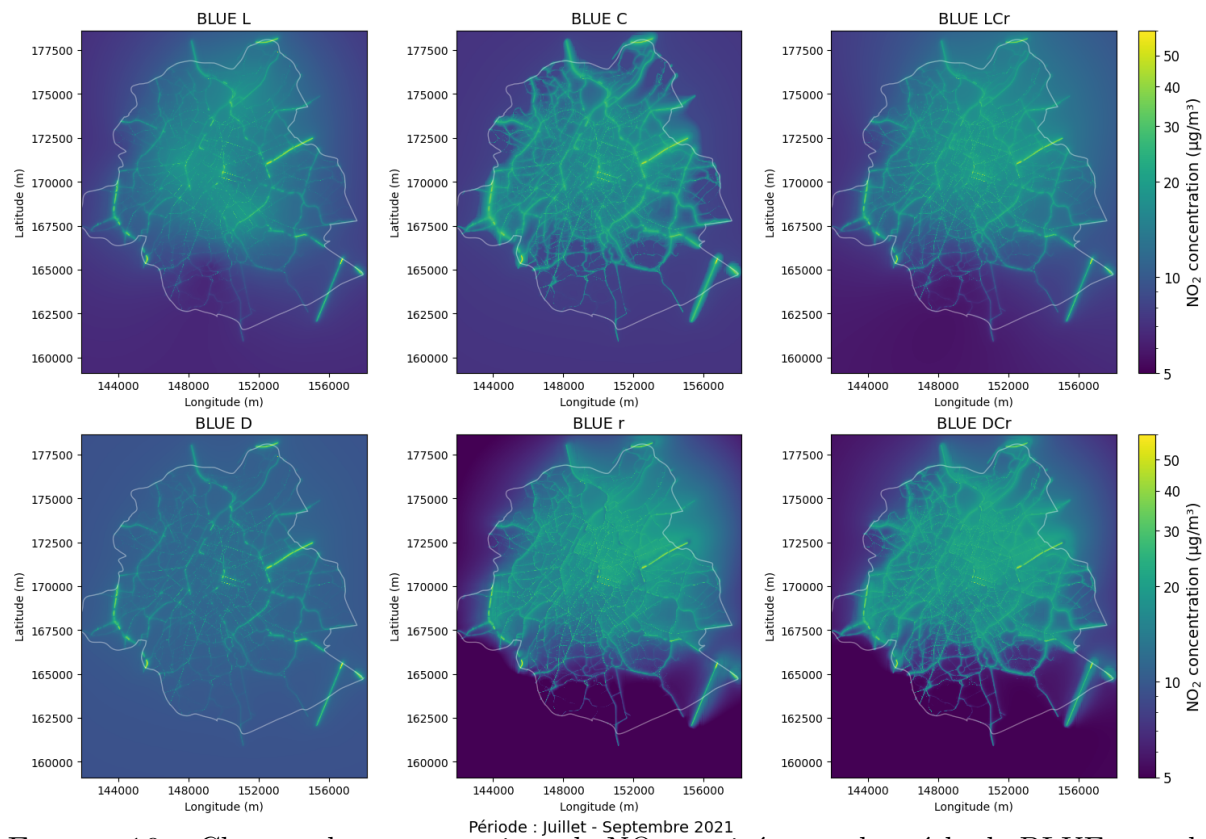


FIGURE 10 – Champs de concentrations de NO_2 corrigées par la méthode BLUE pour le troisième trimestre 2021, selon les six formulations de la matrice B

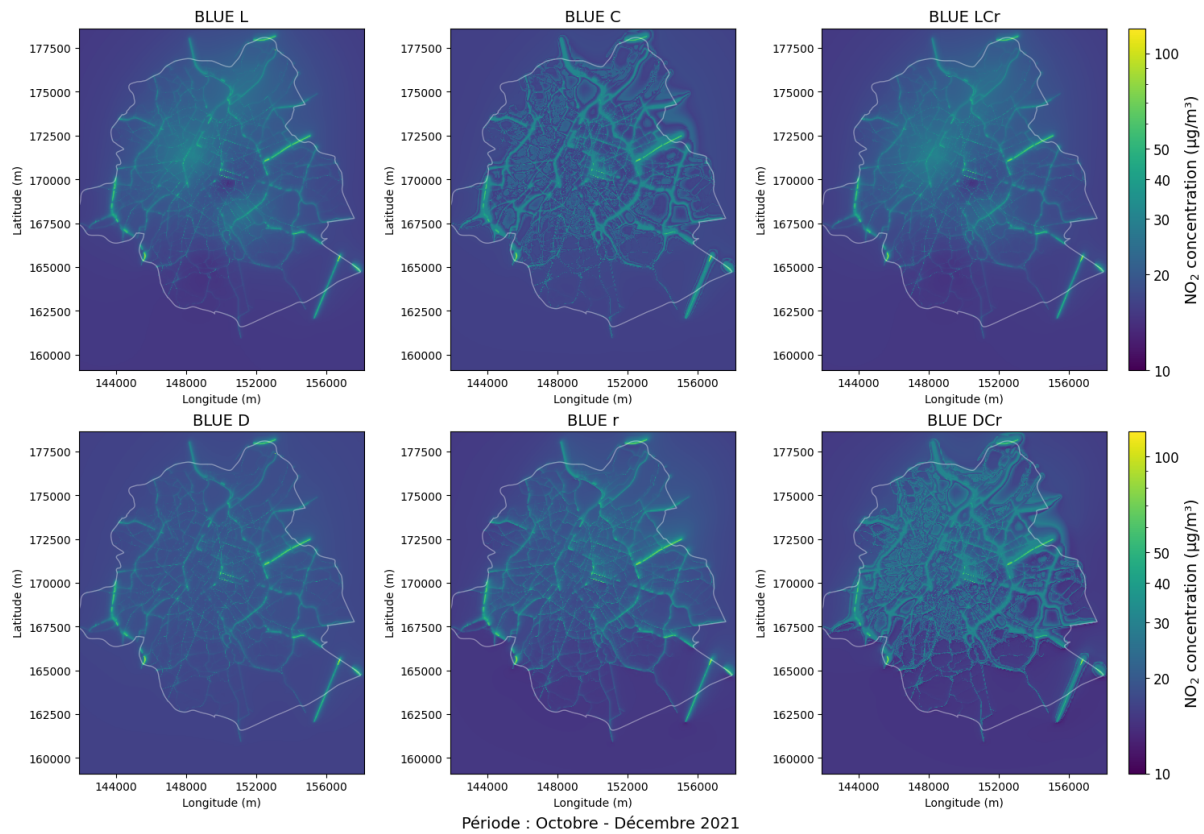


FIGURE 11 – Champs de concentrations de NO_2 corrigées par la méthode BLUE pour le quatrième trimestre 2021, selon les six formulations de la matrice B

Tout comme la méthode BAT additive, la méthode BLUE, en tant qu'estimateur linéaire, ne garantit pas la positivité des concentrations corrigées. Ainsi, dans certaines situations, notamment en présence de fortes corrections négatives, des valeurs de concentration inférieures à zéro peuvent apparaître. Conformément aux contraintes physiques, ces valeurs ont été remplacées par zéro a posteriori, afin d'assurer la cohérence des résultats.

4.2.3 Analyse statistique

Statistiques globales

L'examen visuel des champs corrigés permet de mettre en évidence certaines tendances générales et, dans certains cas, des comportements très différents selon les méthodes ou les trimestres. Toutefois, une évaluation purement qualitative reste insuffisante pour juger de manière rigoureuse la performance des différentes techniques d'assimilation. Il est donc indispensable de compléter cette analyse par une comparaison statistique approfondie.

Les Tables 8 à 11 présentent, pour chacun des quatre trimestres de l'année 2021, les principaux indicateurs d'évaluation : biais moyen, biais fractionnel, RMSE, NMSE et coefficient de corrélation. Ces métriques sont reportées pour le modèle non corrigé SIRANE ainsi que pour les six formulations de la matrice B testées dans le cadre de la méthode BLUE.

TABLE 8 – Statistiques pour la période Janvier - Mars 2021

Méthode	$\overline{C_{mes}}$ [$\mu\text{g}/\text{m}^3$]	$\overline{C_{mod}}$ [$\mu\text{g}/\text{m}^3$]	Biais [$\mu\text{g}/\text{m}^3$]	Biais fract	RMSE [$\mu\text{g}/\text{m}^3$]	NMSE	Corr
SIRANE	29.185	21.767	-7.420	-0.281	13.947	0.311	0.760
B(L)		21.767	-7.420	-0.281	13.947	0.311	0.760
B(D)		27.090	-2.095	-0.056	11.001	0.159	0.845
B(C)		27.901	-1.285	-0.036	10.987	0.157	0.833
B(r)		27.799	-1.387	-0.032	11.168	0.167	0.828
B(LCr)		28.653	-0.532	-0.001	9.665	0.119	0.869
B(DCr)		27.722	-1.463	-0.035	11.050	0.163	0.824

TABLE 9 – Statistiques pour la période Avril - Juin 2021

Méthode	$\overline{C_{mes}}$ [$\mu\text{g}/\text{m}^3$]	$\overline{C_{mod}}$ [$\mu\text{g}/\text{m}^3$]	Biais [$\mu\text{g}/\text{m}^3$]	Biais fract	RMSE [$\mu\text{g}/\text{m}^3$]	NMSE	Corr
SIRANE	22.817	14.186	-8.630	-0.436	14.014	0.629	0.717
B(L)		21.376	-1.439	-0.033	11.002	0.264	0.778
B(D)		18.550	-4.263	-0.159	11.668	0.344	0.802
B(C)		21.234	-1.581	-0.057	11.078	0.291	0.789
B(r)		21.972	-0.847	-0.001	10.640	0.267	0.792
B(LCr)		21.862	-0.956	0.005	9.803	0.221	0.827
B(DCr)		21.724	-1.094	-0.012	10.635	0.268	0.793

TABLE 10 – Statistiques pour la période Juillet - Septembre 2021

Méthode	$\overline{C_{mes}}$ [$\mu\text{g}/\text{m}^3$]	$\overline{C_{mod}}$ [$\mu\text{g}/\text{m}^3$]	Biais [$\mu\text{g}/\text{m}^3$]	Biais fract	RMSE [$\mu\text{g}/\text{m}^3$]	NMSE	Corr
SIRANE	18.843	12.652	-6.190	-0.384	12.561	0.710	0.669
B(L)		17.977	-0.868	-0.014	10.409	0.338	0.725
B(D)		15.384	-3.459	-0.169	11.072	0.455	0.765
B(C)		18.002	-0.841	-0.040	10.389	0.383	0.733
B(r)		18.258	-0.587	0.003	9.796	0.324	0.762
B(LCr)		18.182	-0.661	0.012	9.268	0.277	0.790
B(DCr)		18.163	-0.680	-0.006	9.675	0.317	0.757

TABLE 11 – Statistiques pour la période Octobre - Décembre 2021

Méthode	$\overline{C_{mes}}$ [$\mu\text{g}/\text{m}^3$]	$\overline{C_{mod}}$ [$\mu\text{g}/\text{m}^3$]	Biais [$\mu\text{g}/\text{m}^3$]	Biais fract	RMSE [$\mu\text{g}/\text{m}^3$]	NMSE	Corr
SIRANE	27.371	23.657	-3.713	-0.143	13.623	0.293	0.787
B(L)		25.642	-1.728	-0.050	11.886	0.221	0.778
B(D)		24.801	-2.568	-0.090	13.149	0.264	0.816
B(C)		26.868	-0.504	-0.006	13.173	0.278	0.810
B(r)		26.127	-1.244	-0.055	13.202	0.265	0.820
B(LCr)		25.869	-1.501	-0.039	11.709	0.205	0.829
B(DCr)		27.646	0.274	0.026	13.418	0.285	0.809

L'analyse de ces statistiques globales permet de comparer la capacité des différentes formulations de la matrice B à améliorer les résultats du modèle SIRANE. Sans surprise, ce dernier présente systématiquement les erreurs les plus élevées : les biais moyens dépassent souvent $-6 \mu\text{g}/\text{m}^3$, et les RMSE atteignent jusqu'à $14 \mu\text{g}/\text{m}^3$.

En comparaison, les méthodes d'assimilation basées sur une modélisation plus fine des erreurs d'ébauche permettent une amélioration significative des performances. Les formulations hybrides $B(LCr)$ et $B(DCr)$ s'avèrent globalement les plus efficaces. Le modèle $B(LCr)$, qui combine la distance euclidienne, l'écart de concentration et la corrélation

structurelle, se démarque particulièrement. Il présente des biais très faibles, des RMSE réduits (souvent inférieurs à $10 \mu\text{g}/\text{m}^3$) et les coefficients de corrélation les plus élevés (jusqu'à 0.87 au premier trimestre). Le modèle $B(DCr)$, basé notamment sur la distance via le réseau de voirie, donne également de bons résultats, bien que légèrement en retrait.

Les formulations simplifiées $B(C)$ et $B(r)$, basés sur les concentrations, permettent une amélioration notable, mais inférieure à celle des modèles hybrides. Enfin, les approches $B(L)$ et $B(D)$, reposant uniquement sur la distance spatiale, montrent une efficacité variable selon les périodes, et restent globalement moins performantes.

Le modèle $B(LCr)$, qui apparaît comme le plus performant, surpasse nettement SIRANE sur l'ensemble de l'année. Le biais moyen, par exemple, passe de $-8.63 \mu\text{g}/\text{m}^3$ (avril-juin) à $-0.96 \mu\text{g}/\text{m}^3$ avec $B(LCr)$, soit une réduction de près de 89 %. Le RMSE diminue de 31 % au premier trimestre (de 13.95 à $9.67 \mu\text{g}/\text{m}^3$), et est inférieur de 20–30 % pour les trimestres suivants. De même, le coefficient de corrélation augmente de manière sensible : de 0.76 à 0.87 en janvier-mars, ou de 0.669 à 0.79 en été, représentant une amélioration relative de 15 à 18 %. Ces résultats confirment que la formulation $B(LCr)$, en combinant plusieurs dimensions de la structure d'erreur, permet une assimilation plus précise, robuste et cohérente avec les observations, que le modèle SIRANE et les autres sous-modèle de matrice B .

Statistiques au niveau des capteurs

Si les résultats agrégés permettent de dégager des tendances globales sur la qualité des corrections, ils ne suffisent pas à rendre compte de la variabilité spatiale observée dans les performances du modèle BLUE.

L'analyse globale précédente met en évidence les performances moyennes des différentes formulations de la matrice B sur l'ensemble des stations disponibles. Toutefois, cette vue d'ensemble masque les disparités spatiales importantes qui peuvent exister d'un site à l'autre. Pour mieux évaluer les effets locaux des différentes techniques d'assimilation, une analyse détaillée est menée pour chaque station de mesure.

Les statistiques retenues pour cette évaluation sont identiques à celles de l'analyse globale (biais, RMSE, NMSE, coefficient de corrélation), mais calculées individuellement pour chaque capteur. À ce titre, les performances locales de chaque station sont comparées selon deux configurations : d'une part, la simulation brute issue du modèle SIRANE, et d'autre part, la simulation corrigée à l'aide de la matrice $B(LCr)$, identifiée précédemment comme la plus performante. Cette comparaison permet de quantifier plus précisément l'apport de l'assimilation sur différents types de sites (urbain, industriel), en mettant en évidence les améliorations, ou les éventuelles limites, propres à chaque environnement de mesure.

Analyse du biais

Le biais reflète les écarts systématiques entre les simulations et les observations. Une valeur proche de zéro est idéale. La Figure 12 reprend les diagrammes de comparaison des biais respectifs de SIRANE et de la correction BLUE B(LCr) pour chaque station. Les zones roses indiquent les cas où le modèle SIRANE est plus performant que les résultats du modèle BLUE. La figure montre que l'assimilation permet une réduction marquée du biais sur la majorité des stations et de manière très claire pour la majorité des trimestres, sauf le 4^{ième}, où l'apport de la correction est moins clair. Les stations 41R001, 41R002, 41B001 et 41B004 présentent des corrections importantes ramenant le biais à des niveaux proches de zéro. Certaines stations comme 41B011, 41B008 ou 41R012 présentent des résultats moins satisfaisants, voire une inversion du biais (sur-correction), signalant une mauvaise représentation locale des erreurs.

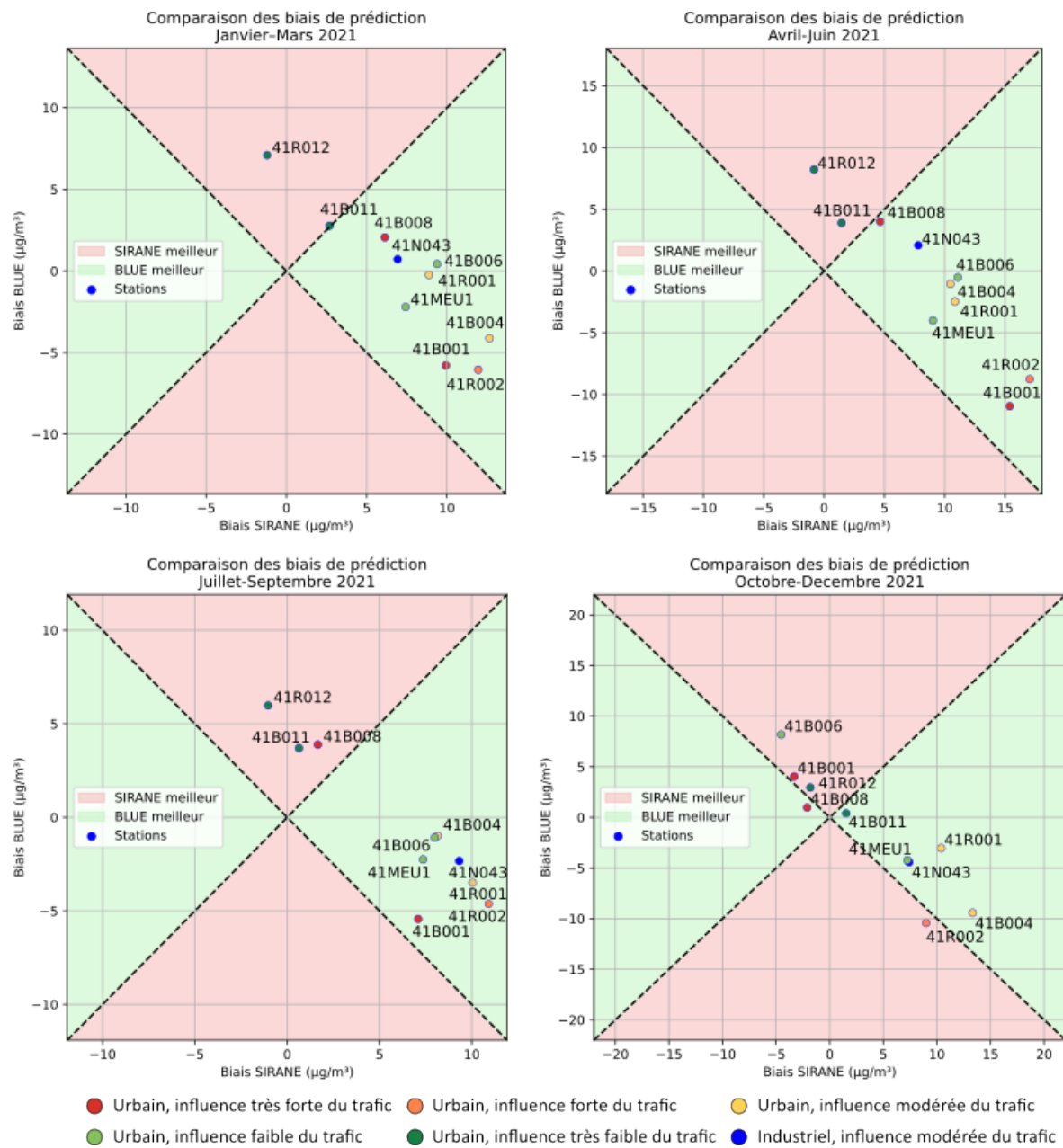


FIGURE 12 – Comparaison des biais moyens entre les stations, avec et sans assimilation (méthode BLUE, $B(LCr)$)

Analyse du RMSE

A l'instar de la Figure 12, la Figure 13 reprend les diagrammes de comparaisons des RMSE pour les différentes stations. Le RMSE mesure les écarts globaux entre les simulations et les observations, en pénalisant fortement les grandes erreurs. L'assimilation réduit significativement cet indicateur pour la majorité des stations, avec des réductions allant jusqu'à 50 % (ex. : 41R001, de 14.91 à 6.47 $\mu\text{g}/\text{m}^3$). Les stations 41B004, 41MEU1 ou 41N043 montrent également des gains notables.

À l'inverse, 41B001 conserve des RMSE très élevés et 41R012 connaît une forte dégradation après correction, indiquant des sur-corrections ou des erreurs mal modélisées localement.

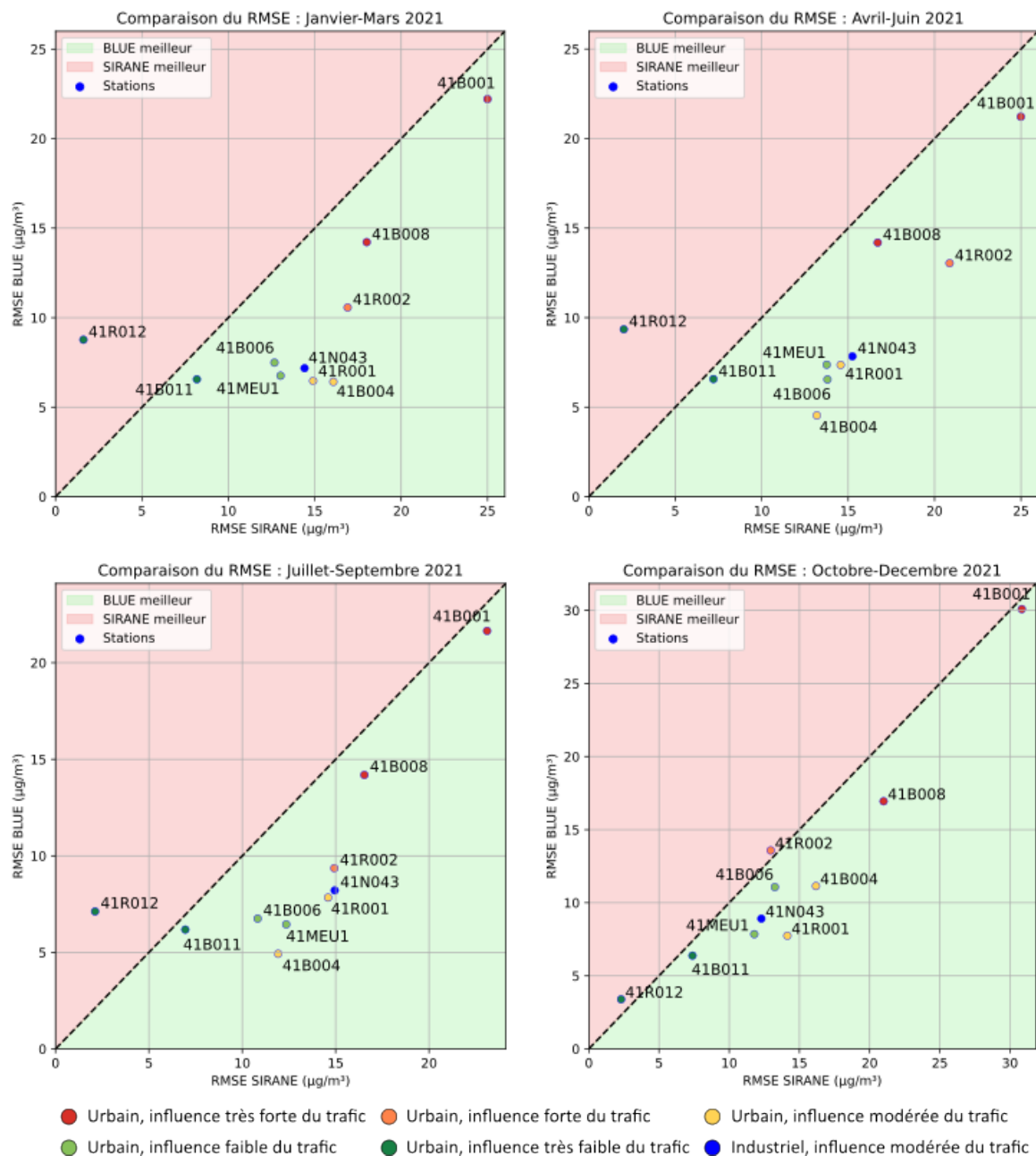


FIGURE 13 – Comparaison du RMSE entre les stations, avec et sans assimilation

Analyse du NMSE

À la Figure 14, le NMSE est comparé selon les stations. L'assimilation améliore sensiblement le NMSE sur la majorité des stations (ex. : 41R001, de 0.355 à 0.05). Les stations 41B004, 41N043 ou 41MEU1 montrent également de fortes réductions.

Toutefois, certaines stations comme 41B001 et 41B008 conservent des valeurs élevées, et 41R012 présente une nette dégradation, confirmant les limites déjà observées avec le RMSE.

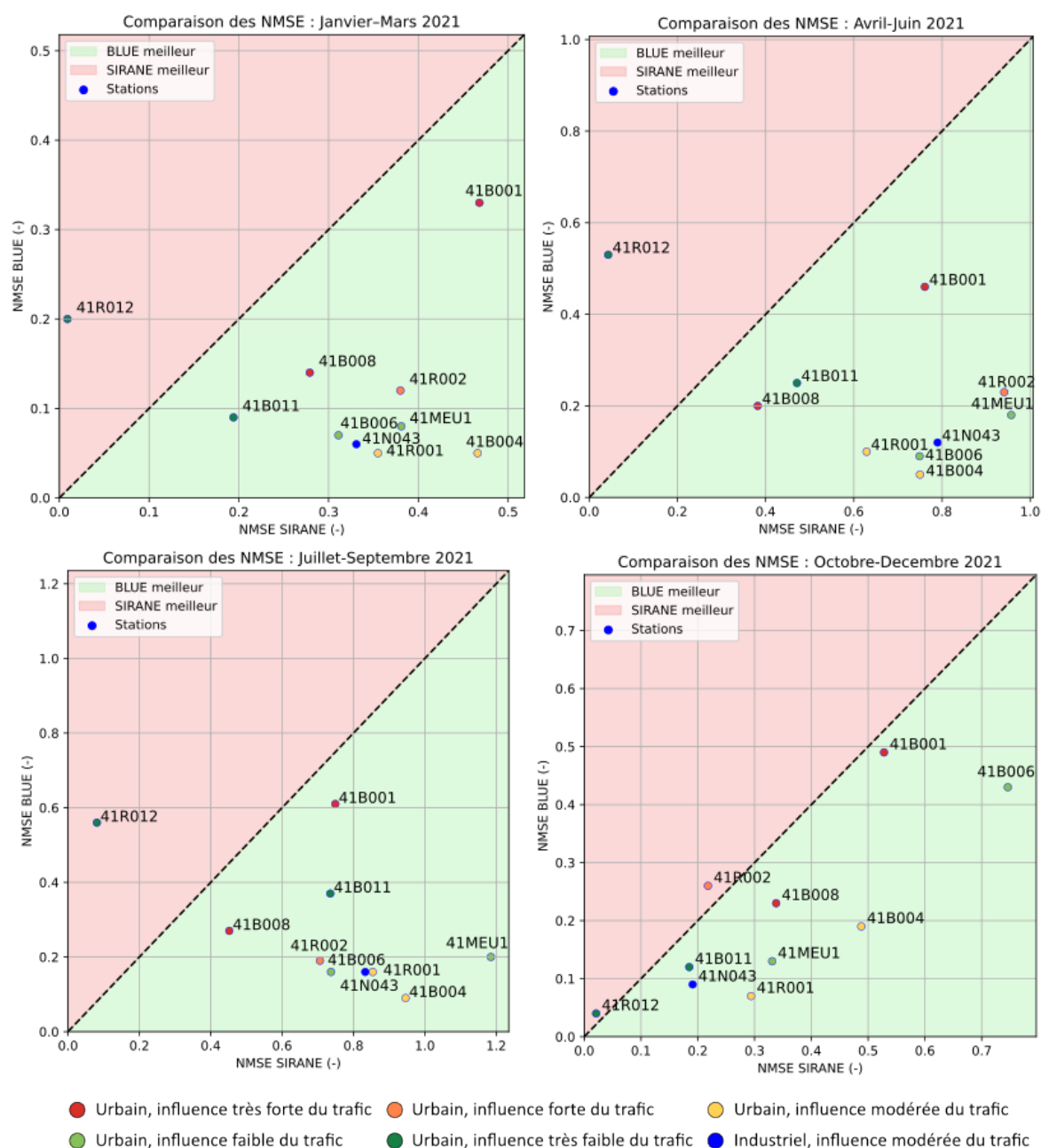


FIGURE 14 – Comparaison du NMSE entre les stations, avec et sans assimilation

Analyse du coefficient de corrélation

Pour finir, le dernier indicateur statistique évalué est le coefficient de corrélation, qui évalue la fidélité de la dynamique temporelle simulée. La Figure 15 reprend les comparaisons de ce coefficient de corrélation entre les différentes stations. De manière générale, une amélioration nette est observée pour la grande majorité des capteurs. Certaines stations comme 41R001 ou 41B004 voient leur corrélation atteindre 0.93–0.95 après assimilation, signalant une restitution bien plus fidèle des variations temporelles.

Pour les stations déjà bien corrélées avec SIRANE (41R012, 41B011), les gains sont moindres, mais la qualité est maintenue. Les stations 41B001, 41B008 et 41B006, avec des corrélations plus faibles, confirment les difficultés d’assimilation sur ces sites spécifiques.

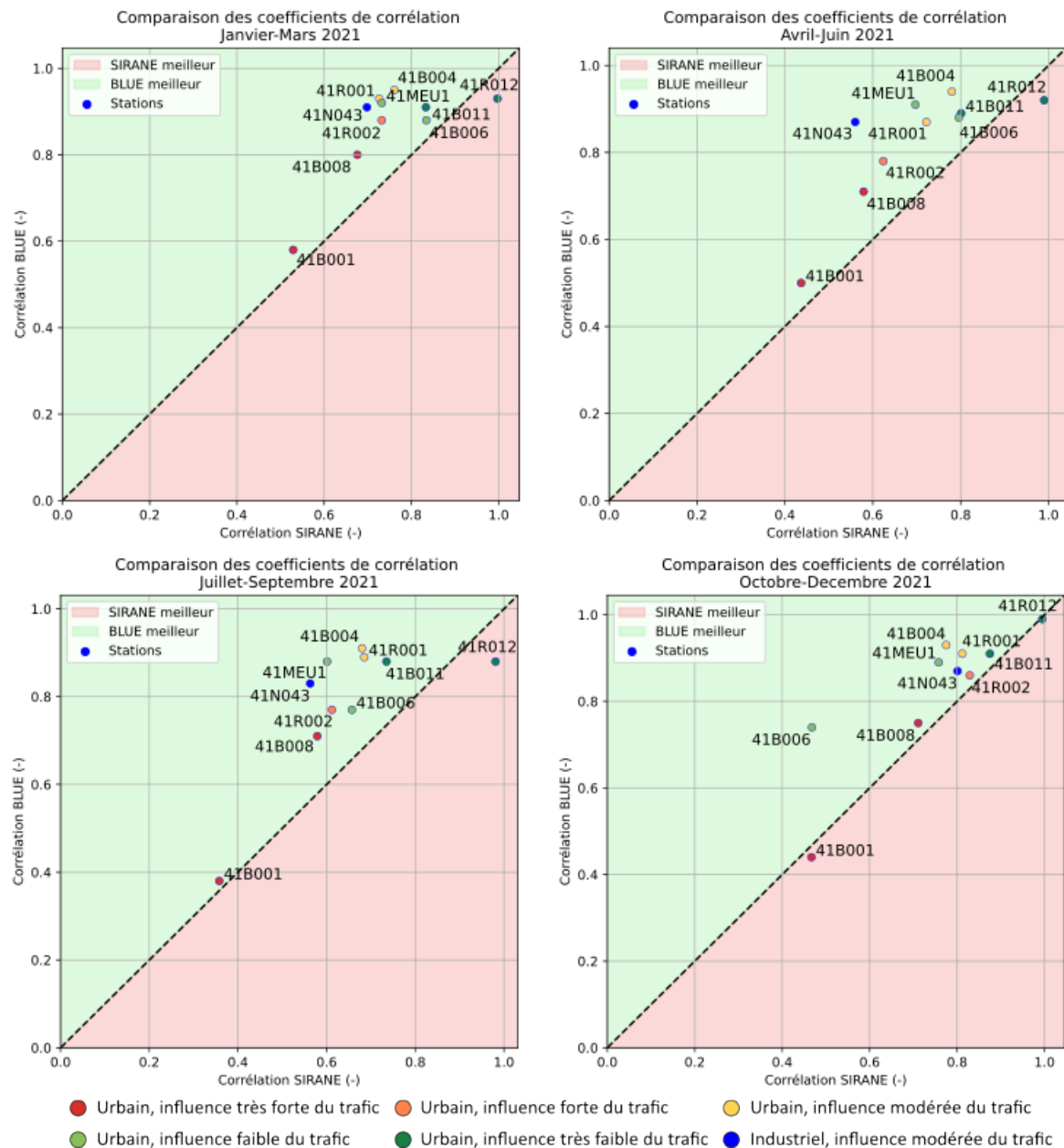


FIGURE 15 – Comparaison du coefficient de corrélation entre les stations, avec et sans assimilation

4.2.4 Discussion des résultats de l’assimilation de données avec la méthode BLUE

L’évaluation statistique de la méthode BLUE, tant globale que par station, met en évidence une amélioration significative des performances du modèle SIRANE, en particulier lorsqu’elle repose sur la formulation hybride $B(LCr)$. Néanmoins, plusieurs limites persistent, notamment sur certaines stations spécifiques, révélant des points d’attention sur la structure des erreurs et les hypothèses du système d’assimilation.

Stations difficilement modélisables

Parmi les stations les plus problématiques, 41B001 et 41B008, situées en zone urbaine dense et fortement influencées par le trafic, se distinguent par des résultats de moins bonne qualité, même après assimilation. Ces stations présentent systématiquement des RMSE et NMSE élevés, ainsi que des corrélations faibles. Ces performances limitées s’expliquent probablement par une mauvaise représentation initiale des émissions de trafic dans le modèle SIRANE. Le module de flux de trafic utilisé par la Région bruxelloise, MuSti, semble peu adapté à la réalité du réseau bruxellois, comme suggéré dans d’autres travaux (BOIGELOT, 2025), et les biais importants persistants dans les résultats corrigés reflètent l’incapacité de l’assimilation à compenser des erreurs structurelles profondes. Un changement de modèle de flux de trafic, plus précis, réaliste et mieux calibré à l’échelle urbaine, constituerait une piste à explorer pour améliorer tant la modélisation, que l’assimilation sur ces sites.

Sur-correction des stations de fond

Le cas de la station 41R012, située en fond périurbain, est différent mais tout aussi révélateur. Contrairement aux autres stations, les performances se dégradent après assimilation : le RMSE et le NMSE augmentent sensiblement, et le biais s’inverse. Ce comportement atypique suggère une situation de sur-correction : les faibles erreurs initiales du modèle brut (SIRANE) sont amplifiées par l’assimilation, probablement en raison d’une inadéquation entre la structure de l’erreur modélisée (matrice B) et la réalité du site. Cela souligne la nécessité de mécanismes de pondération adaptative, évitant des ajustements inutiles là où le modèle est déjà performant.

Sensibilité au choix de la matrice B

Les résultats confirment aussi l’importance cruciale de la matrice B , qui structure la manière dont les erreurs de l’ébauche sont corrélées dans l’espace. Le modèle $B(LCr)$, en combinant des critères de distance, de concentration et de corrélation structurelle (calculée sur les séries horaires), fournit les meilleurs résultats globaux. Il permet une correction plus ciblée et adaptée aux structures spatiales et temporelles du champ simulé, tout en tenant compte des similarités statistiques observées sur les séries horaires. Toutefois, même avec cette formulation optimisée, certaines stations montrent des limites. Cela suggère que d’autres combinaisons ou pondérations pourraient encore être explorées, voire des modèles de covariance entièrement empiriques, pour mieux caractériser les erreurs, notamment dans les environnements hétérogènes comme les zones urbaines.

Modélisation des erreurs d'observation

La matrice R , fondée sur une hypothèse de 15 % d'incertitude maximale (directive européenne 2008/50/CE), reflète l'incertitude associée à chaque station en fonction de ses concentrations moyennes. Ce choix a l'avantage de rester physiquement cohérent, mais il induit une pondération plus faible des stations à forte pollution, et donc potentiellement un moindre impact de ces stations (souvent urbaines) dans la correction finale. Par exemple, les stations 41B001 et 41B008, déjà mal modélisées, sont en plus peu influentes dans le gain de Kalman, ce qui peut aggraver les lacunes sur ces sites. Une amélioration possible serait d'envisager une modélisation empirique ou contextuelle de R , mieux adaptée aux conditions locales (type de site, niveau de pollution, qualité de maintenance des capteurs), et non simplement proportionnelle aux moyennes trimestrielles.

Limites de la méthode BLUE

En tant qu'approche linéaire, la méthode BLUE ne permet pas de représenter les relations non linéaires entre émissions, dispersion et concentrations, et reste sensible aux erreurs inhérentes au modèle. De plus, l'apparition de valeurs négatives dans les champs corrigés, bien que corrigées a posteriori, illustre l'absence de contraintes physiques dans l'assimilation. L'introduction d'approches variationnelles (3D-Var) ou bayésiennes non linéaires (EnKF par exemple) pourrait permettre de dépasser ces contraintes, au prix d'une complexité accrue.

4.3 Discussion générale

Afin de confronter les deux techniques de corrections exploitées en 4.1 et 4.2, la Table 12 compare les performances des méthodes BAT (additive) et BLUE ($B(LCr)$) sur les quatre trimestres de l'année 2021. Dans les deux cas, une nette amélioration des performances est observée, tout particulièrement sur le biais, le RMSE et le coefficient de corrélation. Hormis pour le biais, la méthode BLUE fournit systématiquement de meilleurs résultats que la méthode de débiaisement additive, avec des RMSE et NMSE plus faibles, ainsi que des coefficients de corrélation plus élevés.

TABLE 12 – Comparaison des performances du modèle SIRANE corrigé par les méthodes BAT et BLUE pour chaque trimestre de 2021

Période	Méthode	$\overline{C_{mes}}$ [$\mu\text{g}/\text{m}^3$]	$\overline{C_{mod}}$ [$\mu\text{g}/\text{m}^3$]	Biais [$\mu\text{g}/\text{m}^3$]	Biais fract	RMSE [$\mu\text{g}/\text{m}^3$]	NMSE	Corr
Janvier – Mars 2021	SIRANE		21.767	-7.420	-0.281	13.947	0.311	0.760
	BAT Add	29.185	29.159	-0.027	0.018	10.781	0.152	0.836
	BLUE B(LCr)		28.653	-0.532	-0.001	9.665	0.119	0.869
Avril – Juin 2021	SIRANE		14.186	-8.630	-0.436	14.014	0.629	0.717
	BAT Add	22.817	22.884	0.066	0.056	10.890	0.285	0.794
	BLUE B(LCr)		21.862	-0.956	0.005	9.803	0.221	0.827
Juillet – Septembre 2021	SIRANE		12.652	-6.190	-0.384	12.561	0.710	0.669
	BAT Add	18.843	18.921	0.076	0.049	10.455	0.376	0.750
	BLUE B(LCr)		18.182	-0.661	0.012	9.268	0.277	0.790
Octobre – Décembre 2021	SIRANE		23.657	-3.713	-0.143	13.623	0.293	0.787
	BAT Add	27.371	27.335	-0.034	0.012	13.144	0.251	0.796
	BLUE B(LCr)		25.869	-1.501	-0.039	11.709	0.205	0.829

D'une part, le biais plus faible observé avec la BAT additive découle directement de

la nature même de cette méthode, qui vise explicitement à corriger le biais moyen de l'ébauche. Cette performance, bien que attendue, ne reflète donc pas nécessairement une amélioration plus robuste ou plus générale du champ simulé.

D'autre part, la supériorité globale de BLUE s'explique par sa capacité à intégrer une modélisation plus fine des erreurs, et à adapter spatialement la correction en fonction des observations disponibles. Telle qu'elle est construite, la BAT, en appliquant une correction uniforme à tout le domaine, ne tient pas compte des spécificités locales, ce qui limite sa précision, notamment dans les environnements hétérogènes.

Néanmoins, même avec les meilleures configurations d'assimilation, les performances restent limitées par la qualité des simulations initiales produites par SIRANE. Les erreurs structurelles liées aux sous-modèles utilisés (en particulier le modèle de trafic MuSti) ne peuvent être entièrement compensées par l'assimilation. Certaines stations, notamment celles fortement influencées par le trafic, conservent des erreurs importantes, quelles que soient les corrections appliquées.

Perspectives

Parmi les approches testées dans ce travail, l'assimilation BLUE avec la matrice $B(LCr)$ s'est révélée la plus performante pour améliorer la qualité des simulations issues de SIRANE. Toutefois, cette supériorité reste relative au cadre méthodologique exploré ici. Des méthodes plus avancées, telles que l'assimilation variationnelle (3D-Var) ou les filtres bayésiens (EnKF), pourraient permettre de mieux capturer les dynamiques non linéaires et d'intégrer des contraintes physiques plus strictes. En parallèle, l'amélioration des sous-modèles utilisés dans SIRANE, notamment le module de trafic, demeure une condition essentielle pour tirer pleinement parti des techniques d'assimilation a posteriori.

4.4 Validation

4.4.1 Stratégie de validation mise en place

Pour évaluer la pertinence des corrections apportées par l'assimilation de données, la validation se révèle être une étape clé indispensable. L'objectif est ici de vérifier si la méthode BLUE, désignée comme la plus efficace, permet d'améliorer la qualité des simulations. Cette validation, pour être pertinente, se doit d'être effectuée sur un set de données indépendantes, et donc non-utilisée pour l'assimilation.

Toutefois, en raison du nombre limité de stations disponibles en Région de Bruxelles-Capitale pour l'année 2021 (10 stations), retirer deux stations du jeu d'assimilation pour constituer un jeu de validation indépendant aurait compromis la robustesse des champs corrigés. Une telle approche, bien que classique, n'était donc pas envisageable ici sans affaiblir significativement la qualité de la correction.

Afin de contourner cette difficulté, une autre opportunité a été exploitée : deux stations supplémentaires - la station 41REG1 et la station 41CHA1 - ont été mises en service durant le dernier trimestre de 2021. Ces stations n'ont donc pas été incluses dans l'assimilation et ont été utilisées comme points de validation indépendants. La validation se restreindra ainsi à la période couvrant octobre à décembre 2021.

4.4.2 Contexte et intérêt des stations de validation

Pour la validation des résultats, deux stations supplémentaires ont été utilisées pour les comparer aux données d'analyse. Elles présentent des caractéristiques distinctes, particulièrement intéressantes pour évaluer la robustesse des corrections apportées :

- **Station 41REG1 - Régent** : située en bordure du Pentagone bruxellois, à proximité immédiate d'un axe routier très fréquenté, cette station est fortement influencée par les émissions locales de trafic. Elle permet d'évaluer la capacité de l'assimilation à corriger les concentrations dans des contextes urbains critiques, où le modèle SIRANE montre généralement ses limites.
- **Station 41CHA1 - Ganshoren** : également localisée en milieu urbain, cette station se trouve dans une zone à trafic plus modéré. Sa position un peu excentrée par rapport au cœur de la zone d'étude est également intéressante, car elle permet de tester la qualité des corrections plus en bordure du domaine spatial.



FIGURE 16 – Emplacements des stations; en orange, les stations utilisées pour la validation

Les emplacements des nouvelles stations sont visibles à la Figure 16.

Grâce à leur éloignement géographique et à leurs contextes urbains distincts, ces deux stations constituent deux cas de validation particulièrement intéressants. Elles permettent de vérifier si les corrections apportées par assimilation restent efficaces dans des zones non utilisées initialement, et donc d'apprécier le pouvoir de généralisation de la méthode BLUE.

4.4.3 Séries temporelles des stations de validation

Avant d'évaluer la performance des champs corrigés, il est essentiel de caractériser les nouvelles données de mesure mises à disposition. Une inspection des séries temporelles des concentrations mesurées par les deux stations de validation permet de poser un premier diagnostic qualitatif sur leur comportement respectif (Figure 17).

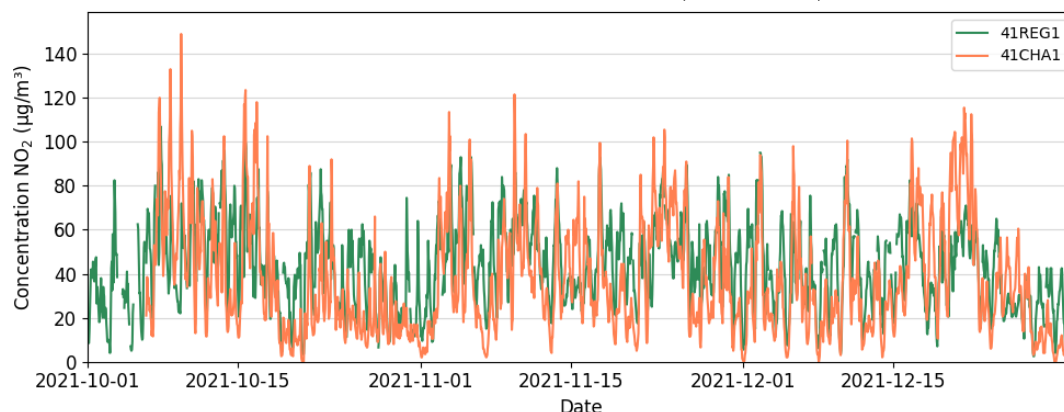


FIGURE 17 – Séries temporelles des stations de validation

À noter que les observations de la station 41CHA1 ne sont disponibles qu'à partir du 6 octobre 2021 à 11h. Cette légère différence de couverture temporelle reste toutefois négligeable à l'échelle du trimestre considéré.

Cette première analyse, visuelle, met en évidence deux comportements contrastés. La station 41REG1, située en bordure d'un axe très fréquenté, enregistre des concentrations moyennes généralement plus élevées que 41CHA1, ce qui est conforme aux attentes. Toutefois, la station 41CHA1 présente ponctuellement des pics de concentration particulièrement marqués, dépassant parfois ceux de 41REG1. Ces pics pourraient résulter de phénomènes locaux spécifiques ou suggérer que la station n'a pas été correctement classée en termes de typologie de site par Bruxelles-Environnement.

Malgré ces différences ponctuelles, les profils temporels des deux stations restent globalement cohérents avec une dynamique saisonnière typique du NO_2 en contexte urbain : des niveaux moyens déjà assez élevés, entrecoupés de hausses transitoires plus intenses.

4.4.4 Résultats de la validation

Évaluation des performances sur la station 41CHA1

La Figure 18 illustre la comparaison entre les concentrations horaires modélisées par SIRANE, les valeurs corrigées par la méthode BLUE, et les observations mesurées à la station 41CHA1 sur la période d'octobre à décembre 2021.

Pour les faibles à moyennes concentrations (jusqu'aux environs de $60 \mu\text{g}/\text{m}^3$), les points simulés par SIRANE et ceux corrigés par BLUE se répartissent globalement autour de la droite $y = x$, indiquant une absence de biais systématique marqué dans cette gamme. Une dispersion non négligeable subsiste néanmoins, traduisant d'incertitudes sur la représentation de la variabilité des observations.

Au-delà de ce seuil ($60 \mu\text{g}/\text{m}^3$), une sous-estimation quasi systématique est clairement visible pour les deux méthodes. Les concentrations simulées et corrigées se situent très majoritairement en dessous des valeurs mesurées, ce qui reflète une certaine incapacité du modèle SIRANE et de la correction BLUE à reproduire correctement les épisodes de pollution plus intenses.

Si l'on compare les deux méthodes, une légère amélioration semble être apportée par la méthode BLUE dans la restitution des fortes concentrations, en comparaison du modèle SIRANE. La sous-estimation, bien que persistant, est légèrement moins forte. Cette atténuation partielle du biais indique que la correction introduite par l'assimilation

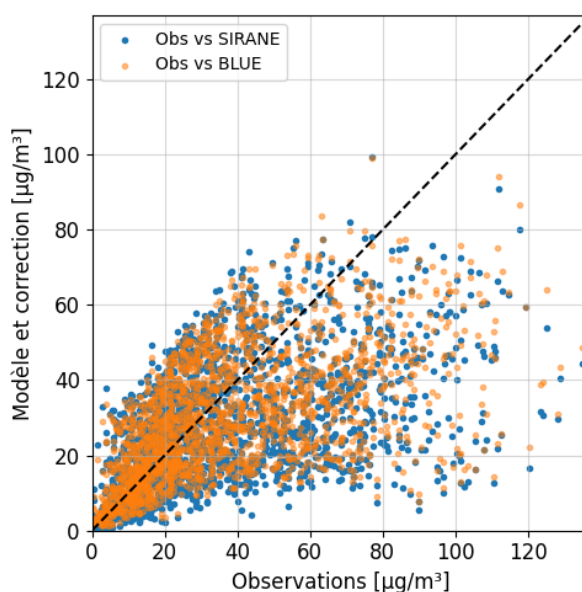


FIGURE 18 – Diagramme de dispersion comparant les observations mesurées avec celles extraites à l'emplacement de la station 41CHA1 par le modèle SIRANE et l'assimilation de données BLUE

n'améliore que légèrement les performances locales du modèle dans les régimes de pollution élevée, sans toutefois les rendre satisfaisantes.

Évaluation des performances sur la station 41REG1

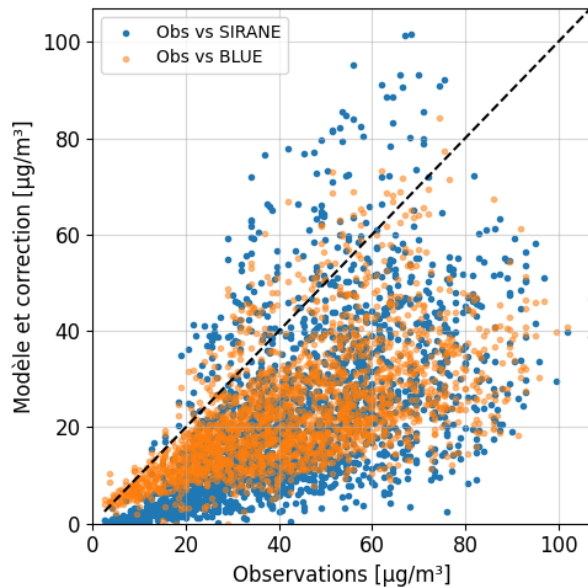


FIGURE 19 – Diagramme de dispersion comparant les observations mesurées avec celles extraites à l'emplacement de la station 41REG1 par le modèle SIRANE et l'assimilation de données BLUE

La Figure 19 montre les résultats obtenus pour la validation à l'emplacement de la station 41REG1, selon le même principe que pour la station 41CHA1.

Dans ce cas, les deux configurations (SIRANE et BLUE) présentent une sous-estimation nette et systématique des concentrations mesurées, visible sur l'ensemble de la gamme de valeurs. Le nuage de points est globalement localisé sous la droite $y = x$, ce qui traduit un biais négatif.

La méthode BLUE permet néanmoins une amélioration significative car les points corrigés (orange) se rapprochent de la droite $y = x$ par rapport à ceux fournis par SIRANE (bleu). Cette réduction du biais est particulièrement marquée pour les concentrations modérées (entre 20 et 60 $\mu\text{g}/\text{m}^3$), qui constituent la majorité des observations. À l'inverse, pour les concentrations les plus élevées ($> 60 \mu\text{g}/\text{m}^3$), les deux configurations continuent de sous-estimer fortement

les observations, bien que BLUE parvienne à réduire partiellement cette erreur. De plus, BLUE permet aussi une réduction globale de la dispersion des points et donc de l'incertitude.

De manière globale, la dispersion reste importante, indiquant une variabilité spatiale partiellement mal représentée. Toutefois, l'amélioration obtenue par BLUE confirme sa capacité à corriger localement les biais du modèle, même en l'absence d'observations directes à proximité immédiate de la station.

Analyse comparative des performances aux stations de validation

La Figure 20 permet de situer les stations de validation (41CHA1 et 41REG1) par rapport à l'ensemble des stations du réseau, en comparant les performances du modèle brut SIRANE et celles obtenues après assimilation avec la méthode BLUE. Quatre indicateurs statistiques sont présentés : le biais, le RMSE, le NMSE et le coefficient de corrélation.

Pour le biais, BLUE améliore légèrement les résultats pour les deux stations. Toutefois, la station 41REG1 présente un biais très élevé (plus de 20 $\mu\text{g}/\text{m}^3$, largement supérieur au seuil de $0.33 \times \overline{C_{mod}} = 8.53 \mu\text{g}/\text{m}^3$, cf. Table 1), signalant une correction insuffisante dans ce contexte local spécifique. À l'inverse, le biais de la station 41CHA1 reste plus modéré, comparable à d'autres stations ayant servis à la modélisation et à l'assimilation.

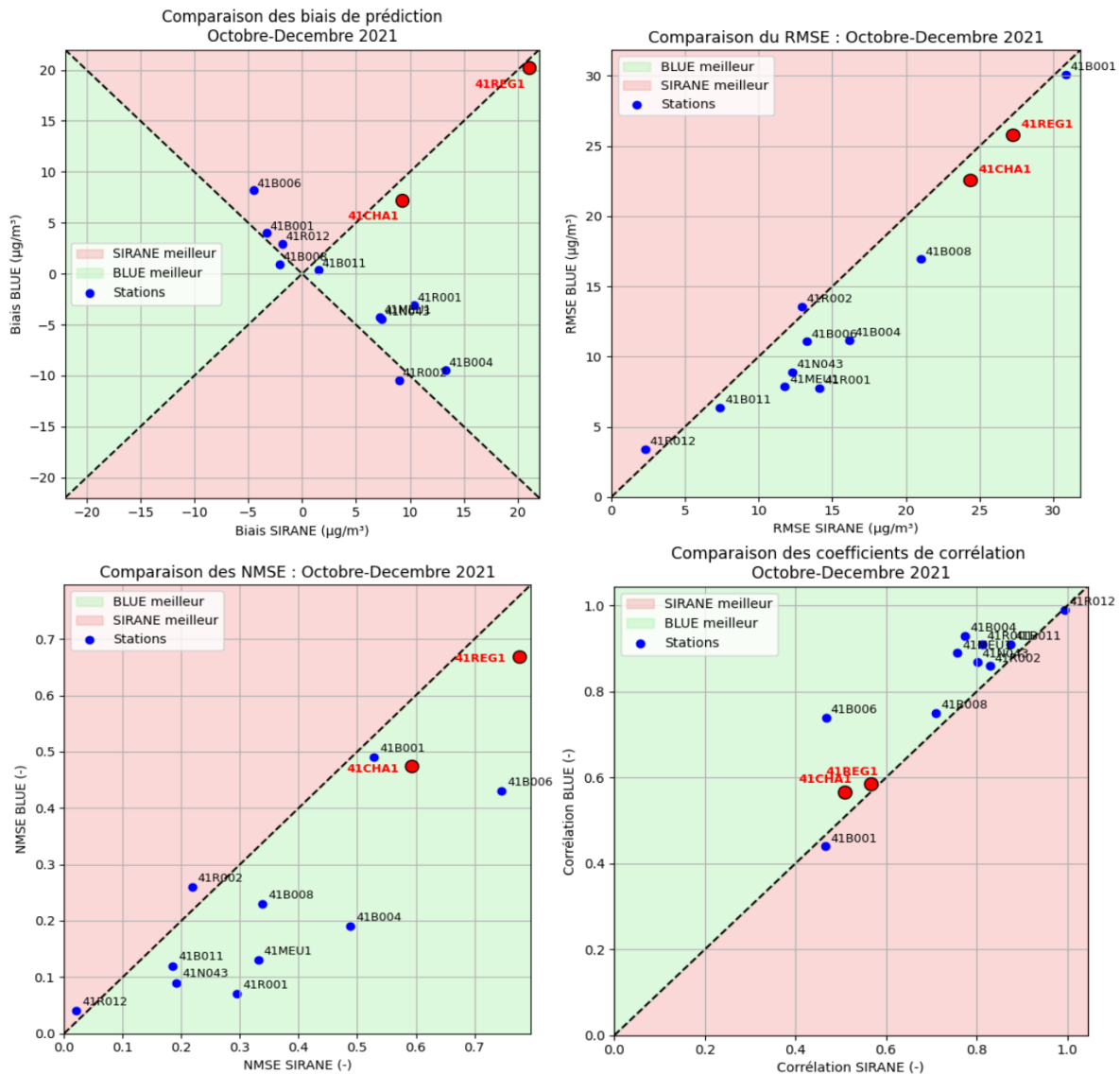


FIGURE 20 – Comparaison des indicateurs statistiques des stations de validation et ceux des stations utilisées par SIRANE et BLUE

Concernant le RMSE, les deux stations de validation affichent des erreurs parmi les plus élevées du réseau, bien que BLUE réduise légèrement ces écarts par rapport à SIRANE. Il en va de même pour le NMSE, où la station de Régent (41REG1) présente les pires performances du jeu, et la station de Ganshoren (41CHA1) se situe également parmi les stations les moins bien corrigées.

Enfin, le coefficient de corrélation reste relativement faible pour les deux stations (autour de 0.6), avec une amélioration marginale grâce à BLUE, mais qui reste en deçà des performances obtenues sur la majorité des stations assimilées.

Ainsi, si l'assimilation BLUE améliore systématiquement les indicateurs par rapport à SIRANE, les performances sur ces stations indépendantes restent limitées, soulignant les difficultés de généralisation spatiale dans des contextes mal représentés par le réseau initial.

Analyse approfondie de la station 41REG1

Comme montré ci-dessus, l'une des deux stations de validation, la station 41REG1, présente des performances particulièrement faibles, avec notamment le biais moyen le plus élevé de toutes les stations étudiées et un RMSE important. Ce constat soulève la question suivante : cette mauvaise performance est-elle due à une faiblesse intrinsèque de la méthode d'assimilation sur ce site, ou bien à une difficulté plus générale à modéliser cette station, même de manière simple ?

Pour explorer cette question, une approche complémentaire a été mise en place. L'idée était de tester la capacité à reconstruire les concentrations mesurées à 41REG1 à partir de celles de stations voisines, en utilisant une régression linéaire multiple. Deux cas de figure ont été considérés : (1) utiliser les concentrations modélisées aux stations voisines comme variables explicatives, et (2) utiliser les concentrations observées à ces mêmes stations. Dans les deux cas, la régression linéaire ordinaire (OLS) a été appliquée à partir des trois puis des cinq stations les plus proches de 41REG1, listées dans le tableau ci-dessous :

TABLE 13 – Stations les plus proches de la station 41REG1

Capteur	Distance à 41REG1 (m)
41B001	255.15
41B008	667.59
41B006	736.39
41B004	1679.68
41R002	2382.92

Les résultats de ces régressions sont présentés dans le tableau suivant, pour les différentes configurations :

TABLE 14 – Comparaison des statistiques de performance pour la station 41REG1 selon différentes méthodes

Statistique	SIRANE	BLUE LCr	3 stat. mod.	5 stat. mod.	3 stat. mes.	5 stat. mes.
Biais [$\mu\text{g}/\text{m}^3$]	21.045	20.203	21.037	21.035	7.465	7.240
RMSE [$\mu\text{g}/\text{m}^3$]	27.270	25.803	25.052	25.111	13.207	12.641
NMSE	0.776	0.669	0.654	0.658	0.113	0.104
Corr	0.566	0.585	0.729	0.725	0.844	0.862

Ce tableau permet de mettre en évidence plusieurs éléments :

- Utiliser 3 ou 5 stations ne modifie que marginalement les résultats.
- Les régressions basées sur les données modélisées aux stations voisines donnent des résultats proches, voire légèrement moins bons, que ceux obtenus via l'assimilation BLUE. Il n'y a que le coefficient de corrélation qui soit amélioré en effectuant la régression multiple à partir des données modélisées de SIRANE.
- En revanche, les régressions utilisant les observations réelles des stations voisines permettent de réduire drastiquement le biais, le NMSE et le RMSE, tout en améliorant significativement le coefficient de corrélation.

Un autre élément à prendre en compte pour expliquer la faible performance de la régression basée sur les données modélisées est la qualité elle-même des simulations issues de SIRANE

pour les stations voisines. En effet, plusieurs des stations utilisées dans la régression - notamment 41B001 et 41B008 - sont connues pour être fortement influencées par le trafic routier et présentent, tout au long de l'étude, des indicateurs statistiques médiocres, même après assimilation. Il n'est donc pas surprenant que les données modélisées en ces points soient peu informatives pour prédire les concentrations de la station 41REG1, conduisant à des résultats insatisfaisants en régression multiple. En revanche, les mêmes stations, lorsqu'elles sont prises avec leurs observations réelles (mesurées), fournissent des estimations bien plus fiables, confirmant que la colinéarité entre stations existe bel et bien, mais qu'elle est mal capturée par le modèle SIRANE.

Ce constat renforce l'idée que les difficultés observées à la station 41REG1 ne résultent pas d'un comportement atmosphérique atypique ou imprévisible, mais plutôt de limites structurelles du modèle, en particulier dans la représentation des émissions liées au trafic urbain. Cela ouvre la voie à des pistes d'amélioration concrètes, comme une modélisation plus fine du trafic à partir d'observations plus précises ou complémentaires visant à agrandir l'échantillonnage. Celles-ci pourraient provenir de campagnes de comptage collaboratif citoyen, ou encore de réseaux de capteurs participatifs à bas coût, tels que Telraam, déjà déployés sur certains axes routiers bruxellois.

4.4.5 Discussion

La stratégie de validation mise en place, bien que limitée par les disponibilités de données, a permis de tester la capacité de généralisation de la méthode d'assimilation BLUE dans un contexte réaliste. Les deux stations sélectionnées, 41CHA1 et 41REG1, ont offert des profils urbains contrastés et suffisamment éloignés du réseau initial pour constituer des points de contrôle indépendants pertinents. Les résultats de la validation effectuée, confirment que, même en dehors des points d'assimilation, la méthode BLUE améliore systématiquement la qualité des simulations par rapport au modèle SIRANE seul. Cela se traduit par une réduction globale du biais, du RMSE et du NMSE, ainsi qu'une légère augmentation du coefficient de corrélation.

Toutefois, les performances restent limitées, en particulier pour la station 41REG1, où les erreurs demeurent importantes malgré l'assimilation. L'analyse complémentaire par régression multiple a mis en évidence que cette station n'est pas isolée dans son comportement : ses concentrations sont fortement corrélées à celles de ses voisines. Ce serait donc les défauts du modèle SIRANE - et non un comportement atmosphérique atypique - qui expliqueraient les mauvaises performances à l'emplacement de cette station.

Il semble donc crucial d'améliorer, en priorité, les résultats produits par SIRANE et ses sous-modèles, ou à défaut, d'envisager des stratégies de correction spécifiques, en post-processing, selon le contexte local des stations. En effet, l'efficacité des techniques d'assimilation reste en grande partie conditionnée par la qualité de l'ébauche initiale. Dans cette optique, une approche différenciée, fondée sur la typologie des stations (urbaine, industrielle, trafic, etc.) et leur degré d'exposition aux sources locales, pourrait permettre de mieux adapter les corrections et de pallier certaines limites structurelles du modèle. Par ailleurs, l'intégration de données issues de dispositifs participatifs tels que Telraam pourrait contribuer à une meilleure représentation des flux de trafic locaux et, in fine, à une amélioration de la modélisation et de l'assimilation en post-processing.

5 Conclusion

Ce mémoire a pour objectif d'évaluer l'apport de l'assimilation de données pour corriger les champs de NO_2 simulés par le modèle urbain SIRANE sur la Région de Bruxelles-Capitale, en comparant une approche simple de débiaisement, la Bias Adjustment Technique (BAT) sous ses formes additive et multiplicative, et un estimateur statistique optimal, le Best Linear Unbiased Estimator (BLUE). La question initiale porte à la fois sur la possibilité d'améliorer significativement la qualité des cartes issues de SIRANE et sur l'identification de la méthode la plus adaptée à un usage opérationnel dans un contexte urbain complexe.

L'analyse menée de manière trimestrielle sur l'année 2021 montre que les deux méthodes testées apportent une amélioration nette par rapport aux champs bruts du modèle, aussi bien en termes de réduction du biais que de diminution des erreurs globales et d'augmentation des corrélations avec les mesures. Toutefois, la méthode BLUE, associée à une formulation optimisée de la matrice de covariance des erreurs de fond, surpasse systématiquement la BAT sur la plupart des indicateurs, confirmant ainsi son intérêt pour une correction robuste et spatialement cohérente des cartes de concentration. La BAT additive conserve néanmoins un atout important : sa simplicité de mise en œuvre et sa capacité à neutraliser efficacement le biais moyen, ce qui en fait une option pertinente pour un débiaisement rapide et homogène. À l'inverse, la BAT multiplicative, bien qu'efficace sur certaines stations, tend à amplifier les contrastes et à dégrader les performances sur les zones déjà fortement polluées, en particulier les stations fortement influencées par le trafic.

La validation indépendante effectuée sur deux stations non utilisées dans l'assimilation confirme la capacité des méthodes à améliorer les estimations, mais met aussi en évidence des performances plus modestes que sur les stations intégrées à l'analyse. Cette différence souligne la difficulté de généraliser spatialement les corrections lorsque le réseau d'observation ne capture pas toute la diversité des environnements urbains. La réponse à la question initiale se dessine ainsi clairement : l'assimilation améliore la qualité des cartes SIRANE, et, dans le cadre étudié, BLUE associé à la formulation B(LCr) constitue la solution la plus performante, tandis que la BAT additive demeure une alternative simple et efficace dans des contextes où la rapidité prime sur l'optimisation fine.

Ces résultats doivent toutefois être replacés dans le contexte des limites identifiées. La qualité de l'ébauche fournie par SIRANE reste un facteur déterminant : certaines erreurs structurelles, notamment liées à la modélisation du trafic, ne peuvent être complètement corrigées par l'assimilation. La performance de BLUE est en outre fortement dépendante de la qualité des matrices de covariance, dont la construction repose ici sur des hypothèses simplificatrices, tant pour les erreurs d'ébauche que pour les erreurs d'observation. Enfin, les méthodes testées, par leur nature linéaire, restent très simples vis-à-vis de la réalité qu'elles tentent de reproduire et ne garantissent pas, pour certaines d'entre elles, la positivité des concentrations sans post-traitement complémentaire.

Ces constats ouvrent plusieurs perspectives. L'amélioration préalable de la modélisation du trafic et des inventaires d'émission apparaît comme une étape essentielle pour renforcer l'efficacité des corrections. L'affinement des matrices de covariance, en intégrant mieux la variabilité spatiale et les différentes catégories de stations, pourrait également accroître la précision des estimations. L'élargissement du réseau d'observation, notamment par l'apport de données issues de dispositifs complémentaires, permettrait d'améliorer la représentativité spatiale des corrections. Enfin, l'exploration de schémas d'assimilation

plus avancés, capables de mieux appréhender la complexité du système atmosphérique et d'intégrer des contraintes physiques, pourrait constituer une suite logique à ce travail, dans la perspective de déploiements opérationnels plus ambitieux.

En définitive, ce travail a pu montrer que l'assimilation de données offre une plus-value réelle pour la production de cartes urbaines de qualité de l'air, en renforçant la cohérence spatiale et en réduisant les écarts avec les observations. Les méthodes évaluées, et en particulier BLUE avec la formulation B(LCr), apportent un compromis intéressant entre précision et faisabilité, tout en mettant en lumière les conditions nécessaires à leur pleine efficacité. Les améliorations méthodologiques et instrumentales envisagées constituent la voie à suivre pour consolider ces résultats et répondre plus largement aux enjeux de surveillance et de gestion de la qualité de l'air en milieu urbain.

Références

- ARCUCCI, R., ZHU, J., HU, S., & GUO, Y.-K. (2021). Deep data assimilation : integrating deep learning with data assimilation. *Applied Sciences*, 11(3), 1114.
- ASCH, M., BOCQUET, M., & NODÉ, M. (2016). *Data Assimilation*. Society for Industrial ; Applied Mathematics. <https://doi.org/10.1137/1.9781611974546>
- BLOND, N., BEL, L., & VAUTARD, R. (2003). Three-dimensional ozone data analysis with an air quality model over the Paris area. *Journal of Geophysical Research : Atmospheres*, 108(D23).
- BOIGELOT, B. (2025). *Validation du modèle SIRANE par données de crowdsourcing : Comparaison des concentrations de NO₂ modélisées et mesurées dans la Région de Bruxelles-Capitale* [Mémoire de master]. Université catholique de Louvain. <https://hdl.handle.net/2078.2/43219>
- BORGE, R., de MIGUEL, I., de la PAZ, D., LUMBRERAS, J., PÉREZ, J., & RODRÍGUEZ, E. (2012). Comparison of road traffic emission models in Madrid (Spain). *Atmospheric Environment*, 62, 461-471.
- BOUTTIER, F., & COURTIER, P. (2002). Data Assimilation Concepts and Methods.
- BRUXELLES ENVIRONNEMENT. (2024). *Émissions de polluants*. Récupérée octobre 27, 2024, à partir de <https://environnement.brussels/citoyen/outils-et-donnees/etat-des-lieux-de-lenvironnement/emissions-de-polluants>
- BRUXELLES MOBILITÉ. (2025). *Le plan régional de mobilité, c'est quoi ?* Récupérée août 13, 2025, à partir de <https://www.bruxelles.be/le-plan-regional-de-mobilite-cest-quoi>
- BUIZZA, C., CASAS, C. Q., NADLER, P., MACK, J., MARRONE, S., TITUS, Z., LE CORNEC, C., HEYLEN, E., DUR, T., RUIZ, L. B., et al. (2022). Data learning : Integrating data assimilation and machine learning. *Journal of Computational Science*, 58, 101525.
- CAMPORESE, M., & GIROTTO, M. (2022). Recent advances and opportunities in data assimilation for physics-based hydrological modeling. *Frontiers in Water*, 4, 948832.
- CELINE. (s.d.). *Modèles de qualité de l'air* [Consulté le 17 décembre 2024]. <https://www.irceline.be/fr/documentation/modeles#:~:text=Les%20mod%C3%A8les%20de%20qualit%C3%A9%20de,polluants%20pr%C3%A9sents%20dans%20l'air>.
- CHANG, J. C., & HANNA, S. R. (2004). Air quality model performance evaluation. *Meteorology and Atmospheric Physics*, 87(1), 167-196.
- CHENG, S., QUILODRÁN-CASAS, C., OUALA, S., FARCHI, A., LIU, C., TANDEO, P., FABLET, R., LUCOR, D., IOOSS, B., BRAJARD, J., XIAO, D., JANJIC, T., DING, W., GUO, Y., CARRASSI, A., BOCQUET, M., & ARCUCCI, R. (2023). Machine Learning With Data Assimilation and Uncertainty Quantification for Dynamical Systems : A Review. *IEEE/CAA Journal of Automatica Sinica*, 10(6), 1361-1387. <https://doi.org/10.1109/JAS.2023.123537>
- CLEMENTS, D., COBURN, M., COX, S., BULOT, F. M., XIE, Z.-T., & VANDERWEL, C. (2024). Comparing large-eddy simulation and Gaussian plume model to sensor measurements of an urban smoke plume. *Atmosphere*, 15(9).
- DALY, A., & ZANNETTI, P. (2007). Air pollution modeling—An overview. *Ambient air pollution*, 15-28.
- de MÉVERGNIES, D. N. (2023). *Évaluation de l'effet de la mise en sens unique de l'avenue E. Cambier sur les concentrations de NO₂ dans le quartier des Azalées à Bruxelles* [Mémoire de master]. Université catholique de Louvain. <https://hdl.handle.net/2078.2/33794>

- EVENSEN, G. (1994). Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research : Oceans*, 99(C5), 10143-10162.
- FREI, M., & KÜNSCH, H. R. (2013). Bridging the ensemble Kalman and particle filters. *Biometrika*, 100(4), 781-800.
- GRIMMOND, C. S. B., & OKE, T. R. (1999). Aerodynamic properties of urban areas derived from analysis of surface form. *Journal of applied meteorology*, 38(9), 1262-1292.
- GROTE, M., WILLIAMS, I., PRESTON, J., & KEMP, S. (2016). Including congestion effects in urban road traffic CO2 emissions modelling : Do Local Government Authorities have the right options? *Transportation Research Part D : Transport and Environment*, 43, 95-106.
- HANNA, S., & CHANG, J. (2012). Acceptance criteria for urban dispersion model evaluation. *Meteorology and Atmospheric Physics*, 116(3), 133-146.
- HERTWIG, D., SOULHAC, L., FUKA, V., AUERSWALD, T., CARPENTIERI, M., HAYDEN, P., ROBINS, A., XIE, Z.-T., & COCEAL, O. (2018). Evaluation of fast atmospheric dispersion models in a regular street network. *Environmental Fluid Mechanics*, 18, 1007-1044.
- HOLMES, N. S., & MORAWSKA, L. (2006). A review of dispersion modelling and its application to the dispersion of particles : An overview of different dispersion models available. *Atmospheric environment*, 40(30), 5902-5928.
- JOHNSON, J. B. (2022). An Introduction to Atmospheric Pollutant Dispersion Modelling. *Environmental Sciences Proceedings*, 19(1), 18.
- JUNG, B.-J., MÉNÉTRIÉ, B., SNYDER, C., LIU, Z., GUERRETTE, J. J., BAN, J., BAÑOS, I. H., YU, Y. G., & SKAMAROCK, W. C. (2024). Three-dimensional variational assimilation with a multivariate background error covariance for the Model for Prediction Across Scales – Atmosphere with the Joint Effort for Data assimilation Integration (JEDI-MPAS 2.0.0-beta). *Geoscientific Model Development*, 17(9), 3879-3895. <https://doi.org/10.5194/gmd-17-3879-2024>
- KAMPA, M., & CASTANAS, E. (2008). Human health effects of air pollution. *Environmental pollution*, 151(2), 362-367.
- LATEB, M., MERONEY, R. N., YATAGHENE, M., FELLOUAH, H., SALEH, F., & BOUFADEL, M. (2016). On the use of numerical modelling for near-field pollutant dispersion in urban environments- A review. *Environmental Pollution*, 208, 271-283.
- LAURIKS, F., JACOBS, D., & MEYSMAN, F. J. (2022). CurieuzenAir : Data collection, data analysis and results. *Report, University of Antwerp*.
- LE DIMET, F.-X., & TALAGRAND, O. (1986). Variational algorithms for analysis and assimilation of meteorological observations : Theoretical aspects. *Tellus*, 38A, 97-110. <https://doi.org/10.3402/tellusa.v38i2.11706>
- LEDERER, A. M., FREDRIKSEN, P. M., NKEH-CHUNGAG, B. N., EVERSON, F., STRIJDOM, H., DE BOEVER, P., & GOSWAMI, N. (2021). Cardiovascular effects of air pollution : current evidence from animal and human studies. *American Journal of Physiology-Heart and Circulatory Physiology*.
- LEELÖSSY, Á., MOLNÁR, F., IZSÁK, F., HAVASI, Á., LAGZI, I., & MÉSZÁROS, R. (2014). Dispersion modeling of air pollutants in the atmosphere : a review. *Open Geosciences*, 6(3), 257-278.
- LORENC, A. C. (2003). The potential of the ensemble Kalman filter for NWP—A comparison with 4D-Var. *Quarterly Journal of the Royal Meteorological Society : A*

- journal of the atmospheric sciences, applied meteorology and physical oceanography*, 129(595), 3183-3203.
- MAIO, S., SARNO, G., TAGLIAFERRO, S., PIRONA, F., STANISCI, I., BALDACCI, S., & VIEGI, G. (2023). Outdoor air pollution and respiratory health. *The International Journal of Tuberculosis and Lung Disease*, 27(1), 7-12.
- MANISALIDIS, I., STAVROPOULOU, E., STAVROPOULOS, A., & BEZIRTZOGLU, E. (2020). Environmental and health impacts of air pollution : a review. *Frontiers in public health*, 8, 14.
- MARTIN, M. J., HOTEIT, I., BERTINO, L., & MOORE, A. M. (2025). Data assimilation schemes for ocean forecasting : state of the art. *State of the Planet*, 5, 1-12.
- MAUNG, T. Z., BISHOP, J. E., HOLT, E., TURNER, A. M., & PFRANG, C. (2022). Indoor air pollution and the health of vulnerable groups : a systematic review focused on particulate matter (PM), volatile organic compounds (VOCs) and their effects on children and people with pre-existing lung disease. *International journal of environmental research and public health*, 19(14), 8752.
- MELINC, B., & ZAPLOTNIK, Ž. (2024). 3D-Var data assimilation using a variational autoencoder. *Quarterly Journal of the Royal Meteorological Society*, 150(761), 2273-2295.
- MODI, M., SK, L. A., HUSSAIN, Z., et al. (2013). A review on theoretical air pollutants dispersion models. *International Journal of Pharmaceutical, Chemical & Biological Sciences*, 3(4).
- NGUYEN, C. V. (2017). *Assimilation de données et couplage d'échelles pour la simulation de la dispersion atmosphérique en milieu urbain* [thèse de doct.] [2017LYSEC018]. <http://www.theses.fr/2017LYSEC018/document>
- NGUYEN, C. V., & SOULHAC, L. (2021). Data assimilation methods for urban air quality at the local scale. *Atmospheric Environment*, 253, 118366.
- NICOLAS, J., & CRAFT, F. (2003). La modélisation de la dispersion atmosphérique. Cas particulier des odeurs.
- NTZIACHRISTOS, L., GKATZOFLIAS, D., KOURIDIS, C., & SAMARAS, Z. (2009). COPERT : a European road transport emission inventory model. *Information Technologies in Environmental Engineering : Proceedings of the 4th International ICSC Symposium Thessaloniki, Greece, May 28-29, 2009*, 491-504.
- OKE, T. R. (1988). The urban energy balance. *Progress in Physical geography*, 12(4), 471-508.
- PANDYA, D., VACHHARAJANI, B., & SRIVASTAVA, R. (2022). A review of data assimilation techniques : Applications in engineering and agriculture. *Materials Today : Proceedings*, 62, 7048-7052.
- PARK, S. K., & ZUPANSKI, M. (2022). *Principles of data assimilation*. Cambridge University Press.
- REICH, S. (2019). Data assimilation : the Schrödinger perspective. *Acta Numerica*, 28, 635-711.
- SINI, J.-F., & MESTAYER, P. G. (1998). Traffic-induced urban pollution : A numerical simulation of street dispersion and net production. *Air Pollution Modeling and Its Application XII*, 369-377.
- SMIT, R., NTZIACHRISTOS, L., & BOULTER, P. (2010). Validation of road vehicle and traffic emission models—A review and meta-analysis. *Atmospheric environment*, 44(25), 2943-2953.

- SOULHAC, L. (2000). *Modélisation de la dispersion atmosphérique à l'intérieur de la canopée urbaine* [thèse de doct., Ecole Centrale de Lyon].
- SOULHAC, L., NGUYEN, C. V., VOLTA, P., & SALIZZONI, P. (2017). The model SIRANE for atmospheric urban pollutant dispersion. PART III : Validation against NO₂ yearly concentration measurements in a large urban agglomeration. *Atmospheric environment*, 167, 377-388.
- SOULHAC, L., SALIZZONI, P., CIERCO, F.-X., & PERKINS, R. (2011). The model SIRANE for atmospheric urban pollutant dispersion ; part I, presentation of the model. *Atmospheric environment*, 45(39), 7379-7395.
- SOULHAC, L., SALIZZONI, P., MEJEAN, P., DIDIER, D., & RIOS, I. (2012). The model SIRANE for atmospheric urban pollutant dispersion ; PART II, validation of the model on a real case study. *Atmospheric environment*, 49, 320-337.
- SPORTISSE, B. (2008). *Pollution atmosphérique. Des processus à la modélisation*. Springer-Verlag France. <https://inria.hal.science/inria-00581172>
- TIWARY, A., & WILLIAMS, I. (2018). *Air pollution : measurement, modelling and mitigation*. Crc Press.
- WHO. (2024). *Ambient (outdoor) air pollution* [Consulté le 17 décembre 2024]. [https://www.who.int/news-room/fact-sheets/detail/ambient-\(outdoor\)-air-quality-and-health](https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health)

6 Annexes

Annexe 1 – Implémentation de l’assimilation de données et utilisation de YODA

Passer de la théorie à la pratique en assimilation de données est souvent plus ardu qu’il n’y paraît. Dès les premières tentatives d’implémentation des méthodes BAT et BLUE, un obstacle majeur est apparu : la puissance de calcul requise. En théorie, ces deux techniques sont relativement légères, mais dans les faits, la méthode BLUE s’est révélée difficile à faire tourner sur l’ensemble du domaine d’étude, soit la Région de Bruxelles-Capitale.

Les premiers tests ont été effectués localement, sur une zone restreinte, afin de prendre en main les données et vérifier que les codes fonctionnaient. Ces essais ont été concluants à petite échelle. Mais dès que l’assimilation était exécutée sur le domaine complet, l’exécution devenait impossible, que ce soit localement ou sur les serveurs de l’UCLouvain. La cause principale de cette incapacité est liée à la taille colossale des matrices manipulées, en particulier la matrice de covariance des erreurs d’ébauche B .

Pour rappel, BLUE est définie par :

$$x^a = x^b + K(y - Hx^b)$$

où le gain de Kalman K s’écrit :

$$K = BH^T(HBH^T + R)^{-1}$$

Si n est le nombre de points de grille et m le nombre de stations, alors :

- R est de dimension $m \times m$,
- H est de dimension $m \times n$,
- B est de dimension $n \times n$.

Dans notre cas, avec une résolution spatiale de 10 m, n dépasse les 3,16 millions de points compris dans le domaine, ce qui signifie que la matrice B serait théoriquement de taille $3,165,176 \times 3,165,176$ - soit plus de 10^{13} coefficients. Une telle matrice, d’une telle ampleur, a été impossible à générer ou stocker avec les ressources mises à disposition.

Deux options étaient alors envisageables :

1. Découper le domaine en sous-domaines et recoller les résultats - au prix, notamment, d’incertitudes sur la validité des hypothèses aux frontières.
2. Diminuer la résolution spatiale - ce qui est regrettable quand on dispose de sorties du modèle SIRANE, d’une résolution spatiale de 10 m.

Après quelques échanges avec l’École Centrale de Lyon, nous avons pu utiliser YODA, un outil interne capable de faire tourner BAT et BLUE selon la même méthodologie que celle décrite ici, mais sur des domaines étendus. Bien que les détails de mise en œuvre ne soient pas publics, il est vraisemblable que l’outil contourne le problème de taille de la matrice B en utilisant un opérateur de covariance, aussi appelé modèle de covariance implicite. Ce procédé consiste non pas à stocker B dans son intégralité, mais à définir une loi paramétrique (par exemple, une décroissance exponentielle avec la distance) permettant de calculer à la volée chaque coefficient B_{ij} à partir d’un petit nombre de paramètres (L_l , L_d , ν_0 , etc.). Une telle approche permettrait de réduire drastiquement la mémoire requise tout en préservant la structure spatiale des corrélations.

Si cette approche transforme à nouveau l'outil en "boîte noire" (comme SIRANE), elle permet néanmoins de travailler sur tout le domaine à la résolution souhaitée, ce qui était inenvisageable avec une implémentation brute.

Mise en application de YODA

L'implémentation de YODA pour la région bruxelloise nécessite la mise à disposition d'un ensemble précis de données d'entrée. Celles-ci sont organisées dans le répertoire **INPUT**, qui doit contenir quatre sous-dossiers principaux :

- **Grilles** : contient les champs horaires de concentration issus de SIRANE. Par défaut, SIRANE ne fournit que les grilles moyennées ; il est donc nécessaire de demander explicitement la sortie horaire dans ses paramètres.
- **Groupe_Sources** : regroupe les différentes catégories d'émissions modélisées (industrielles, trafic, urbain). Ces données ne sont pas utilisées dans ce travail, mais sont requises pour une autre méthode d'assimilation de données (Source Apportionment Least Square), non implémentée ici.
- **Recepteurs** : contient un fichier par station de mesure, listant pour chaque date les concentrations mesurées et modélisées. SIRANE ne produisant pas ces fichiers directement, un script a été développé pour assembler ces données. Ce dossier contient également un fichier récapitulatif des caractéristiques des stations (identifiant, coordonnées (x, y) , incertitude pour la matrice R , poids dans l'assimilation, etc.).
- **Reseau** : fichier shapefile représentant le réseau de voirie (en projection WGS84), constitué de segments de ligne interconnectés.

Une fois ces données préparées, l'ensemble des paramètres nécessaires à l'assimilation (période d'étude, méthode utilisée, sous-modèles de B , chemins d'accès aux fichiers d'entrée et de sortie, polluants considérés, etc.) est renseigné dans le fichier de configuration `Donnees_template_linux.dat`.

Annexe 2 – BAT - Statistiques par capteur pour les 4 trimestres

TABLE 15 – Statistiques par capteur pour la correction BAT additive – Janvier à Mars 2021

Capteur	Milieu	Trafic	$\overline{C_{mes}}$ [$\mu\text{g}/\text{m}^3$]	$\overline{C_{mod}}$ [$\mu\text{g}/\text{m}^3$]	Biais [$\mu\text{g}/\text{m}^3$]	Biais fract	RMSE [$\mu\text{g}/\text{m}^3$]	NMSE	Corr
41R001	Urbain	Modéré	29.85	28.19	-1.66	-0.06	8.15	0.08	0.89
41R002	Urbain	Fort	34.04	29.27	-4.77	-0.15	9.67	0.09	0.89
41B001	Urbain	Très fort	41.84	39.02	-2.82	-0.07	22.08	0.30	0.61
41B004	Urbain	Modéré	30.71	25.01	-5.70	-0.20	8.30	0.09	0.93
41B011	Urbain	Très faible	19.96	25.20	5.24	0.23	9.43	0.18	0.83
41R012	Urbain	Très faible	16.39	25.95	9.56	0.45	12.86	0.39	0.76
41N043	Industriel	Modéré	28.75	29.29	0.54	0.02	8.20	0.08	0.88
41MEU1	Urbain	Faible	25.13	25.10	-0.03	0.00	8.11	0.10	0.86
41B006	Urbain	Faible	27.89	25.75	-2.15	-0.08	7.47	0.08	0.89
41B008	Urbain	Très fort	37.29	38.81	1.52	0.04	13.54	0.13	0.82

TABLE 16 – Statistiques par capteur pour la correction BAT multiplicative – Janvier à Mars 2021

Capteur	Milieu	Trafic	$\overline{C_{mes}}$ [$\mu\text{g}/\text{m}^3$]	$\overline{C_{mod}}$ [$\mu\text{g}/\text{m}^3$]	Biais [$\mu\text{g}/\text{m}^3$]	Biais fract	RMSE [$\mu\text{g}/\text{m}^3$]	NMSE	Corr
41R001	Urbain	Modéré	29.85	27.70	-2.15	-0.07	8.65	0.09	0.87
41R002	Urbain	Fort	34.04	29.57	-4.46	-0.14	9.65	0.09	0.88
41B001	Urbain	Très fort	41.84	45.87	4.03	0.09	29.83	0.46	0.54
41B004	Urbain	Modéré	30.71	23.68	-7.03	-0.26	9.45	0.12	0.92
41B011	Urbain	Très faible	19.96	23.27	3.31	0.15	7.88	0.13	0.86
41R012	Urbain	Très faible	16.39	24.10	7.70	0.38	10.29	0.27	0.84
41N043	Industriel	Modéré	28.75	29.72	0.97	0.03	8.29	0.08	0.88
41MEU1	Urbain	Faible	25.13	23.56	-1.58	-0.06	8.23	0.11	0.86
41B006	Urbain	Faible	27.89	24.40	-3.49	-0.13	7.38	0.08	0.92
41B008	Urbain	Très fort	37.29	42.77	5.48	0.14	17.70	0.20	0.79

TABLE 17 – Statistiques par capteur pour la correction BAT additive – Avril à Juin 2021

Capteur	Milieu	Trafic	$\overline{C_{mes}}$ [$\mu\text{g}/\text{m}^3$]	$\overline{C_{mod}}$ [$\mu\text{g}/\text{m}^3$]	Biais [$\mu\text{g}/\text{m}^3$]	Biais fract	RMSE [$\mu\text{g}/\text{m}^3$]	NMSE	Corr
41R001	Urbain	Modéré	24.57	22.27	-2.30	-0.10	8.09	0.12	0.84
41R002	Urbain	Fort	31.66	22.35	-9.32	-0.35	13.41	0.25	0.78
41B001	Urbain	Très fort	37.35	29.96	-7.39	-0.22	20.02	0.36	0.52
41B004	Urbain	Modéré	21.34	19.65	-1.69	-0.08	5.69	0.08	0.91
41B011	Urbain	Très faible	11.25	19.39	8.14	0.53	10.37	0.49	0.85
41R012	Urbain	Très faible	9.34	20.06	10.72	0.73	12.87	0.88	0.80
41N043	Industriel	Modéré	21.49	22.27	0.79	0.04	8.63	0.16	0.83
41MEU1	Urbain	Faible	19.28	19.02	-0.26	-0.01	8.25	0.19	0.82
41B006	Urbain	Faible	22.43	19.87	-2.57	-0.12	7.64	0.13	0.85
41B008	Urbain	Très fort	29.46	34.00	4.54	0.14	13.93	0.19	0.74

TABLE 18 – Statistiques par capteur pour la correction BAT multiplicative – Avril à Juin 2021

Capteur	Milieu	Trafic	$\overline{C_{mes}}$ [$\mu\text{g}/\text{m}^3$]	$\overline{C_{mod}}$ [$\mu\text{g}/\text{m}^3$]	Biais [$\mu\text{g}/\text{m}^3$]	Biais fract	RMSE [$\mu\text{g}/\text{m}^3$]	NMSE	Corr
41R001	Urbain	Modéré	24.57	22.48	-2.08	-0.09	9.59	0.17	0.77
41R002	Urbain	Fort	31.66	22.91	-8.76	-0.32	13.16	0.24	0.77
41B001	Urbain	Très fort	37.35	37.03	-0.32	-0.01	27.70	0.55	0.35
41B004	Urbain	Modéré	21.34	17.11	-4.23	-0.22	7.10	0.14	0.90
41B011	Urbain	Très faible	11.25	15.55	4.29	0.32	7.84	0.35	0.85
41R012	Urbain	Très faible	9.34	17.34	8.00	0.60	10.17	0.64	0.87
41N043	Industriel	Modéré	21.49	22.53	1.04	0.05	9.58	0.19	0.79
41MEU1	Urbain	Faible	19.28	16.05	-3.22	-0.18	8.75	0.25	0.83
41B006	Urbain	Faible	22.43	17.90	-4.53	-0.22	8.36	0.17	0.86
41B008	Urbain	Très fort	29.46	46.67	17.21	0.45	33.69	0.83	0.49

TABLE 19 – Statistiques par capteur pour la correction BAT additive – Juillet à Septembre 2021

Capteur	Milieu	Trafic	$\overline{C_{mes}}$ [$\mu\text{g}/\text{m}^3$]	$\overline{C_{mod}}$ [$\mu\text{g}/\text{m}^3$]	Biais [$\mu\text{g}/\text{m}^3$]	Biais fract	RMSE [$\mu\text{g}/\text{m}^3$]	NMSE	Corr
41R001	Urbain	Modéré	21.59	17.38	-4.21	-0.22	8.98	0.21	0.85
41R002	Urbain	Fort	24.04	18.66	-5.38	-0.25	9.57	0.20	0.79
41B001	Urbain	Très fort	30.48	29.51	-0.97	-0.03	21.33	0.51	0.41
41B004	Urbain	Modéré	16.99	14.90	-2.10	-0.13	6.31	0.16	0.87
41B011	Urbain	Très faible	8.43	14.71	6.29	0.54	8.72	0.61	0.81
41R012	Urbain	Très faible	6.94	15.42	8.47	0.76	10.89	1.11	0.72
41N043	Industriel	Modéré	21.69	18.22	-3.46	-0.17	9.19	0.21	0.80
41MEU1	Urbain	Faible	15.61	14.46	-1.15	-0.08	7.92	0.28	0.78
41B006	Urbain	Faible	17.22	15.28	-1.95	-0.12	7.45	0.21	0.72
41B008	Urbain	Très fort	25.44	30.67	5.22	0.19	14.19	0.26	0.75

TABLE 20 – Statistiques par capteur pour la correction BAT multiplicative – Juillet à Septembre 2021

Capteur	Milieu	Trafic	$\overline{C_{mes}}$ [$\mu\text{g}/\text{m}^3$]	$\overline{C_{mod}}$ [$\mu\text{g}/\text{m}^3$]	Biais [$\mu\text{g}/\text{m}^3$]	Biais fract	RMSE [$\mu\text{g}/\text{m}^3$]	NMSE	Corr
41R001	Urbain	Modéré	21.59	17.03	-4.56	-0.24	9.53	0.25	0.82
41R002	Urbain	Fort	24.04	19.05	-4.99	-0.23	9.47	0.20	0.78
41B001	Urbain	Très fort	30.48	38.82	8.34	0.24	32.78	0.91	0.27
41B004	Urbain	Modéré	16.99	13.05	-3.94	-0.26	7.25	0.24	0.86
41B011	Urbain	Très faible	8.43	11.96	3.54	0.35	6.72	0.45	0.83
41R012	Urbain	Très faible	6.94	13.16	6.22	0.62	8.15	0.73	0.84
41N043	Industriel	Modéré	21.69	18.50	-3.18	-0.16	9.63	0.23	0.76
41MEU1	Urbain	Faible	15.61	12.41	-3.20	-0.23	8.51	0.37	0.78
41B006	Urbain	Faible	17.22	13.74	-3.49	-0.23	7.58	0.24	0.75
41B008	Urbain	Très fort	25.44	38.48	13.04	0.41	24.27	0.60	0.63

TABLE 21 – Statistiques par capteur pour la correction BAT additive – Octobre à Décembre 2021

Capteur	Milieu	Trafic	$\overline{C_{mes}}$ [$\mu\text{g}/\text{m}^3$]	$\overline{C_{mod}}$ [$\mu\text{g}/\text{m}^3$]	Biais [$\mu\text{g}/\text{m}^3$]	Biais fract	RMSE [$\mu\text{g}/\text{m}^3$]	NMSE	Corr
41R001	Urbain	Modéré	31.79	24.91	-6.88	-0.24	11.01	0.15	0.86
41R002	Urbain	Fort	32.60	26.68	-5.91	-0.20	10.01	0.12	0.87
41B001	Urbain	Très fort	40.82	48.57	7.75	0.17	31.13	0.49	0.47
41B004	Urbain	Modéré	30.77	20.01	-10.76	-0.42	13.38	0.29	0.86
41B011	Urbain	Très faible	17.95	20.35	2.40	0.13	8.99	0.22	0.82
41R012	Urbain	Très faible	14.98	21.10	6.12	0.34	9.28	0.27	0.83
41N043	Industriel	Modéré	32.06	27.92	-4.14	-0.14	8.34	0.08	0.87
41MEU1	Urbain	Faible	24.42	20.81	-3.61	-0.16	9.18	0.17	0.81
41B006	Urbain	Faible	13.24	21.68	8.45	0.48	11.88	0.49	0.77
41B008	Urbain	Très fort	35.08	41.32	6.24	0.16	18.24	0.23	0.80

TABLE 22 – Statistiques par capteur pour la correction BAT additive – Octobre à Décembre 2021

Capteur	Milieu	Trafic	$\overline{C_{mes}}$ [$\mu\text{g}/\text{m}^3$]	$\overline{C_{mod}}$ [$\mu\text{g}/\text{m}^3$]	Biais [$\mu\text{g}/\text{m}^3$]	Biais fract	RMSE [$\mu\text{g}/\text{m}^3$]	NMSE	Corr
41R001	Urbain	Modéré	31.79	24.64	-7.15	-0.25	11.03	0.16	0.87
41R002	Urbain	Fort	32.60	26.68	-5.92	-0.20	10.01	0.12	0.87
41B001	Urbain	Très fort	40.82	59.70	18.88	0.38	44.76	0.82	0.28
41B004	Urbain	Modéré	30.77	19.34	-11.43	-0.46	13.43	0.30	0.89
41B011	Urbain	Très faible	17.95	19.09	1.14	0.06	7.49	0.16	0.87
41R012	Urbain	Très faible	14.98	19.77	4.79	0.28	7.02	0.17	0.91
41N043	Industriel	Modéré	32.06	28.31	-3.75	-0.12	8.21	0.07	0.87
41MEU1	Urbain	Faible	24.42	19.96	-4.46	-0.20	8.91	0.16	0.84
41B006	Urbain	Faible	13.24	20.46	7.23	0.43	11.37	0.48	0.77
41B008	Urbain	Très fort	35.08	44.29	9.21	0.23	21.43	0.30	0.76

Correction a posteriori de la cartographie des polluants issue du modèle SIRANE

Ce mémoire s'inscrit dans le contexte de la surveillance de la qualité de l'air en milieu urbain, où la production de cartes fiables constitue un enjeu majeur pour l'évaluation de l'exposition des populations. L'objectif était d'évaluer l'apport de l'assimilation de données pour corriger les champs de dioxyde d'azote (NO_2) simulés par le modèle urbain SIRANE sur la Région de Bruxelles-Capitale. Deux approches ont été comparées : une méthode simple de débiaisement, la *Bias Adjustment Technique* (BAT, sous formes additive et multiplicative), et un estimateur statistique optimal, le *Best Linear Unbiased Estimator* (BLUE).

L'évaluation conduite par trimestre sur l'année 2021 à partir des stations de référence montre que l'assimilation améliore nettement les performances de SIRANE (réduction du biais, baisse des erreurs globales, corrélations accrues). Cette amélioration est observée sur la majorité des stations, mais elle reste plus limitée sur les sites fortement influencés par le trafic, où les erreurs de modélisation persistent malgré la correction. BLUE, qui combine de façon optimale l'information issue des observations et celle de l'ébauche en fonction de leurs incertitudes respectives, fournit des corrections robustes et spatialement cohérentes. Avec une matrice de covariance des erreurs de fond formulée comme B(LCr), il offre systématiquement les meilleurs résultats. La BAT additive reste une alternative simple et efficace pour un débiaisement rapide, tandis que la BAT multiplicative tend à amplifier les contrastes dans les zones fortement polluées.

Une validation indépendante sur deux stations non utilisées dans l'assimilation confirme l'amélioration, tout en révélant des performances plus modestes que sur les stations intégrées, ce qui souligne la difficulté de généraliser lorsque le réseau d'observation ne couvre pas toute la diversité urbaine.

En conclusion, l'assimilation de données apporte une plus-value réelle pour la production de cartes urbaines de NO_2 . Dans le cadre étudié, BLUE (B(LCr)) apparaît comme la solution la plus performante, tandis que la BAT additive constitue une option opérationnelle lorsque la rapidité prime. Des gains additionnels sont attendus via une meilleure modélisation du trafic et des émissions, un affinement des matrices de covariance, l'élargissement du réseau d'observation et l'exploration de schémas d'assimilation plus avancés.