Titre: Title:	Identification de données suspectes par apprentissage profond dans les précipitations journalières mesurées au Québec
Auteur: Author:	Delhio Calves
Date:	2025
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Calves, D. (2025). Identification de données suspectes par apprentissage profond dans les précipitations journalières mesurées au Québec [Mémoire de maîtrise, Polytechnique Montréal]. PolyPublie. https://publications.polymtl.ca/66534/

URL de PolyPublie: PolyPublie URL:	https://publications.polymtl.ca/66534/
Directeurs de recherche: Advisors:	Jonathan Jalbert, & Camélia Dadouchi
Programme: Program:	Maîtrise en mathématiques appliquées

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Identification de données suspectes par apprentissage profond dans les précipitations journalières mesurées au Québec

DELHIO CALVES
Département de mathématiques et de génie industriel

Mémoire présenté en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*
Mathématiques

Juin 2025

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Ce mémoire intitulé :

Identification de données suspectes par apprentissage profond dans les précipitations journalières mesurées au Québec

présenté par **Delhio CALVES**
en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*
a été dûment accepté par le jury d'examen constitué de :

Bruno AGARD, président
Jonathan JALBERT, membre et directeur de recherche
Camélia DADOUCHI, membre et codirectrice de recherche
Sara-Ann PISCOPO, membre

DÉDICACE

À Papo

REMERCIEMENTS

Pour commencer, je tiens à remercier chaleureusement mes directeurs de recherche, Jonathan Jalbert et Camélia Dadouchi, grâce à qui j'ai découvert l'univers de la recherche dans d'excellentes conditions. Leur expertise, leur expérience et leur dévouement m'ont guidé tout au long de mon parcours, et je leur en suis grandement reconnaissant. Avoir un accompagnement de si grande qualité est une chance que je mesure aujourd'hui à sa rareté, et travailler à leurs côtés va indéniablement me manquer tant leur sollicitude envers leurs étudiants est à nulle autre pareille. Merci également aux personnes impliquées dans le projet, qu'elles travaillent chez Hydro-Québec ou chez Solutions Mesonet : Pierre-Olivier Caron-Périgny, Alexandre Vanasse, Joël Papineau, Vincent Graveline, Charles Mathieu, Fanny Payette et Samer Alghabra. Leur investissement dans les réunions de travail m'ont permis de m'orienter dans certaines directions plus prometteuses, et m'ont encouragé. Merci aussi à Vincent Fortin, Nicolas Gasset, Dominik Jacques, Luc Perreault et Philippe Roy pour leurs propositions et leur aide sur le projet lors des réunions à Ouranos (et même en dehors). Je remercie mes collègues et amis du LID et de l'équipe d'*Extremes*, avec qui je me suis épanoui et me suis senti à mon aise pendant ma recherche.

Merci à Mitacs et Hydro-Québec pour leur soutien financier.

Je tiens également à remercier Côme, qui m'a appris à écrire des *logs* dans un fichier, me renseignant sur l'avancement de mes programmes, ainsi que Clément, grâce à qui j'ai découvert une erreur lourde de conséquences dans mon jeu de données, découverte qui m'a permis de réellement propulser mon projet. Merci aussi à Kat et Chloé pour leur soutien moral et bromatologique. Enfin, merci à mes proches qui m'ont soutenu tout au long de la maîtrise : ma famille pour leur accompagnement depuis l'autre côté de l'océan, Anaïs pour son soutien indéfectible et ses relectures minutieuses et attentives.

Pour terminer, je souhaite remercier les membres du jury, Bruno Agard et Sara-Ann Piscopo, qui ont accepté de relire consciencieusement mon travail.

RÉSUMÉ

L'identification des valeurs suspectes dans les données météorologiques est une étape importante pour de nombreuses applications. Ce processus de contrôle qualité passe habituellement par des tests statistiques puis par l'analyse d'experts dans le domaine, ce qui peut être chronophage et coûteux. Ainsi, ce projet de recherche s'inscrit dans ce contexte, et traite donc de données météorologiques, plus particulièrement les précipitations mesurées. L'objectif est d'automatiser l'identification des données suspectes provenant de capteurs. Les données suspectes peuvent en réalité être des données réelles représentant des phénomènes rares ou des valeurs erronées. Ce contrôle est nécessaire pour s'assurer de la fiabilité des données alimentant les analyses réalisées pour la prévision météorologique, la planification de la production d'énergie hydro-électrique, ou encore en soutien à l'agriculture. Le modèle proposé dans cette étude se base sur les réseaux de neurones convolutifs et intègre différents types de données de précipitations, ponctuelles et surfaciques, et les combine afin d'effectuer une classification binaire, les deux classes étant celle des anomalies et celle des observations authentiques. Il intègre également une notion de dépendance spatiale notamment puisque les observations mesurées aux stations voisines sont intégrées au modèle de classification. Le modèle démontre d'excellentes performances de classification sur des données réelles, puisque les données suspectes sont correctement détectées dans plus de 95 % des cas. Il paraît donc pertinent pour une possible implémentation pour l'assurance qualité à l'opérationnelle.

ABSTRACT

Identifying abnormal values in meteorological data is an important step for a variety of applications. This quality assurance starts with statistical tests, and then continues with an expert analysis, which is costly and time-consuming. This research project is part of this context, and focuses especially on anomaly detection in precipitation records. The goal is to automize the quality assurance process of automatic sensor records that come from meteorological stations, since these sensors may provide abnormal data. This quality assurance process is of high stake as it enables reliability in datasets that are useful in various applications: meteorological prediction, hydroelectricity production planning, and support for agriculture, among others. The proposed model includes convolutional neural networks and integrates various types of precipitation data: point records and gridded data. It combines them in order to achieve a binary classification: authentic observations against suspect ones, i.e. anomalies. The method also includes a spatial dependency notion since it includes a second channel integrating records from neighbouring stations. The model shows interesting performance on a real dataset, since more than 95 % of the abnormal data are correctly detected. Henceforth, this model seems relevant to be implemented for an operational quality assurance.

TABLE DES MATIÈRES

DÉDICACE iii

REMERCIEMENTS iv

RÉSUMÉ v

ABSTRACT vi

TABLE DES MATIÈRES vii

LISTE DES TABLEAUX x

LISTE DES FIGURES xi

LISTE DES SIGLES ET ABRÉVIATIONS xii

LISTE DES ANNEXES xiii

CHAPITRE 1 INTRODUCTION 1

CHAPITRE 2 REVUE DE LITTÉRATURE 4

2.1 Tests statistiques classiques 5

2.2 Méthodes géostatistiques (basées sur les stations voisines) 5

2.3 Méthodes spatio-temorelles 6

2.4 Méthodes d'apprentissage automatique 7

2.5 Réseaux de neurones convolutifs 10

2.5.1 Classification binaire en apprentissage supervisé 10

2.5.2 Filtres convolutifs 10

2.5.3 Hyperparamètres des réseaux de neurones convolutifs 13

2.5.4 *Convolve & pool* 14

2.5.5 *Batch normalization* et fonctions d'activation 14

2.5.6 Fonction de perte et descente de gradient 14

2.5.7 Ensemble de validation, *dropout* et *early stopping* 16

2.5.8 Sortie du réseau et décision 17

2.5.9 Évaluation de la classification 17

2.6 Synthèse de la revue de la littérature 19

CHAPITRE 3 MÉTHODOLOGIE 20
3.1 Problématique 20
3.2 Objectif et sous-objectifs 20
3.3 Données nécessaires 23
3.3.1 Précipitations observées (validées par des experts) 23
3.3.2 Précipitations observées (brutes et non validées) 23
3.3.3 Précipitations réanalysées 23
3.3.4 Données d'altitude 24
3.4 Prétraitement 24
3.4.1 Extraction des précipitations positives 24
3.4.2 Mise à l'échelle des précipitations observées 25
3.4.3 Génération de valeurs suspectes artificielles 25
3.5 Création du jeu de données 26
3.5.1 Fusion des données d'observation et du produit sur grille 26
3.5.2 Deuxième canal 28
3.6 Modélisation 30
3.6.1 Séparation en ensembles d'entraînement, de validation et de test . . . 33
3.6.2 Ajustement des poids 33
3.6.3 Optimisation des hyperparamètres 33
3.6.4 Classification 34
3.7 Choix du meilleur modèle 34
3.8 Validation 34

CHAPITRE 4 CAS D'ÉTUDE 36
4.1 Données nécessaires 37
4.1.1 Précipitations observées (validées par des experts) 37
4.1.2 Précipitations observées (brutes et non validées) 38
4.1.3 Précipitations réanalysées 39
4.1.4 Données d'altitude 39
4.2 Prétraitement 39
4.2.1 Extraction des précipitations positives 40
4.2.2 Mise à l'échelle des précipitations observées 40
4.2.3 Génération de valeurs suspectes artificielles 42
4.3 Création du jeu de données 42
4.3.1 Fusion des données d'observation et du produit sur grille 42
4.3.2 Deuxième canal 42

4.4 Modélisation 43
4.4.1 Séparation en ensembles d'entraînement, de validation et de test . . . 44
4.4.2 Ajustement des poids 44
4.4.3 Optimisation des hyperparamètres 44
4.4.4 Classification 44
4.5 Choix du meilleur modèle 44
4.6 Validation 45
4.7 Résultats 45
4.7.1 Erreurs artificielles 46
4.7.2 Performances du modèle sur un jeu de données réel 49

CHAPITRE 5 DISCUSSION ET CONCLUSION 50
5.1 Choix de la dimension de la grille et du nombre de cellules centrales remplacées 50
5.2 Pertinence de l'entraînement station par station 50
5.3 Inclusion des valeurs aux stations voisines 51
5.4 Ajustement du seuil de classification avec le jeu de données réel 52
5.5 Synthèse des travaux 52
5.6 Limites de la solution proposée 53
5.7 Améliorations futures 54

RÉFÉRENCES 56

ANNEXES 60

LISTE DES TABLEAUX

Tableau 3.1 Récapitulatif des scénarios 30

Tableau 4.1 Métriques p, FPR et FNR pour les différentes versions des modèles pour des erreurs radicales (en %) 47

Tableau 4.2 Seuils optimisés et métrique p pour les différentes versions des modèles pour des erreurs radicales (en %) 47

Tableau 4.3 Métriques p, FPR et FNR pour les différentes versions des modèles pour différentes amplitudes d'erreurs plus modérées (en %) 49

LISTE DES FIGURES

Figure 2.1	Processus de convolution d'une image.	11
Figure 3.1	Schéma général de la méthodologie utilisée.	22
Figure 3.2	Incorporation des observations dans la réanalyse sur grille.	27
Figure 3.3	Schéma de la paire d'image obtenue grâce aux deux canaux (scénario B).	29
Figure 3.4	Exemple d'image du premier canal pour (a) une observation cohérente et (b) un observation suspecte.	31
Figure 3.5	Architecture du modèle.	32
Figure 4.1	Précipitomètre.	36
Figure 4.2	Stations météorologiques considérées dans l'étude.	38
Figure 4.3	Régression linéaire pour la station de La Tuque.	41
Figure 4.4	Inclusion des valeurs des stations voisines.	43
Figure A.1	Faux positifs	60
Figure A.2	Faux négatifs	61

LISTE DES SIGLES ET ABRÉVIATIONS

RMCQ	Réseau Météorologique Coopératif du Québec
OMM	Organisation Météorologique Mondiale
CNN	*Convolutional Neural Network*
MLP	*Multi-Layer Perceptron*
TP	*True positive*
TN	*True negative*
FP	*False positive*
FN	*False negative*
FPR	*False positive rate*
FNR	*False negative rate*
QR	*Quality rate*
AR	*Anomaly rate*
RDRS	Système Régional de Réanalyse Déterministe
RDPS	Système Régional Déterministe de Précipitations
ECCC	Environnement et Changement climatique Canada
ETOPO2022	*Earth TOPOgraphy 2022*
CaSPAr	*Canadian Surface Prediction Archive*

LISTE DES ANNEXES

Annexe A Faux positifs et faux négatifs 60
Annexe B Génération de valeurs suspectes artificielles 62
Annexe C Définition des labels pour le jeu de données réel 65

CHAPITRE 1 INTRODUCTION

Disposer de données météorologiques de qualité est indispensable dans de nombreux domaines comme la climatologie, la météorologie et l'hydrologie. En effet, ces données sont essentielles pour alimenter la prise de décisions dans le domaine de l'énergie, notamment dans la gestion et la production hydrolélectrique. Prendre des décisions éclairées nécessite donc de détenir des informations précises, pertinentes et fiables. Or, la récolte de données météorologiques s'effectue au moyen d'une instrumentation très spécifique, que sont les capteurs comme les précipitomètres, les thermomètres, les radiomètres ou encore les anémomètres. Seulement, toute mesure est entachée d'erreur, et ces capteurs ne font pas exception, d'autant plus qu'ils sont exposés à des conditions météorologiques qui peuvent causer des défaillances du système de mesure. Ceci explique en partie pourquoi les données enregistrées aux stations météorologiques contiennent des anomalies. Pour pallier cette difficulté, des procédures d'assurance qualité des données météorologiques ont été développées, en vue d'identifier les valeurs suspectes parmi les observations, susceptibles d'être des données aberrantes. Ces méthodes nécessitent généralement une intervention humaine, afin de valider les résultats de procédures automatiques basées sur des statistiques. Ce processus est donc coûteux en temps, et bénéficierait d'outils de détection d'anomalies plus avancés que les tests statistiques classiques actuellement utilisés.

Hydro-Québec est une société d'État, leader de la production, du transport et de la distribution d'électricité au Québec. Elle produit 99 % de son énergie par la filière hydro-électrique, qui fait partie de celles émettant le moins de gaz à effet de serre par kilowattheure produit. Seulement, cette filière nécessite d'être en mesure d'acquérir des données météorologiques de qualité, notamment pour la prévision de consommation, la planification de la production ou encore la réduction du temps des pannes. Pour ce faire, Hydro-Québec requiert des mesures météorologiques fiables autour de ses installations et de ses clients afin notamment d'estimer la demande et planifier la production en fonction des conditions sur les bassins versants exploités. Les stations météorologiques sont généralement munies de différents types de capteurs, dédiés à différentes variables météorologiques. Ces capteurs, qui sont aujourd'hui automatiques, peuvent parfois produire des données erronées, en raison de défaillances ou de conditions météorologiques particulières. En effet, pour les précipitations notamment, le vent peut notamment entraîner une sous-captation de la quantité réelle de précipitations survenues, puisque celle-ci peut être déviée du réceptacle (Goodison *et al.*, 1998). Les précipitations solides souffrent elles aussi souvent de problèmes de mesure (Groisman *et al.*, 1991). Par exemple, la neige peut s'accumuler en bordure du capteur, puis fondre, soit en tombant

d'un coup dans le capteur, augmentant la quantité mesurée, soit en tombant à l'extérieur du capteur. Ce type d'erreur est appelé *snow capping* (Rasmussen *et al.*, 2012). D'autres types d'erreurs sont également spécifiques à la mesure des précipitations solides, comme le phénomène de pourdrerie qui peut entraîner une captation additionnelle de neige provenant du sol (Rasmussen *et al.*, 2012) ou encore les pertes en évaporation (Leeper et Kochendorfer, 2015). Enfin, en hiver certains types de capteurs sont bloqués pour la saison, puisqu'ils ne sont pas adaptés aux précipitations solides comme la neige. C'est le cas des pluviomètres à auget basculant, qui sont dédiés aux précipitations liquides uniquement. Des erreurs de transmission peuvent également survenir entre la station d'observation et le serveur qui centralise les données, ainsi que des problèmes de maintenance ou des inconsistances dans les standards utilisés.

Pour limiter ces problèmes, le Réseau Météorologique Coopératif du Québec (RMCQ) suit les normes de gestion et d'exploitation de l'Organisation Météorologique Mondiale (OMM) pour ses stations. Ce réseau rassemble, met en commun et permet l'échange des données des stations d'observation de ses membres en temps réel. Les membres du RMCQ sont :

- le Ministère de l'Environnement, de la Lutte contre les changements climatiques, de la Faune et des Parcs ;
- Rio Tinto Alcan ;
- Environnement et Changement climatique Canada ;
- Hydro-Québec ;
- le Ministère de l'Énergie et des Ressources naturelles ;
- la Société de protection des forêts contre le feu (SOPFEU).

Malgré l'existence de ces normes, il est nécessaire de prendre en compte le fait que les données provenant des capteurs sont susceptibles d'être anormales, et peuvent donc fausser les analyses si les anomalies y sont intégrées. Dans ce cadre, l'organisme à but non lucratif Solutions Mesonet collecte et contrôle la qualité des données de ses membres, qui sont ceux du RMCQ auxquels on ajoute la Financière agricole du Québec. Une équipe est dédiée à l'assurance qualité des observations. Ce processus d'assurance qualité inclut des analyses préliminaires automatisées, basées sur des seuils et des règles définies, et des analyses manuelles effectuées à l'aide de différentes cartes météorologiques, de séries temporelles et de données radar. Si une défaillance matérielle est détectée, alors une notification est envoyée aux propriétaires de la station en cause. Enfin, les données validées sont diffusées aux membres pour consultation sous forme de tables, de graphiques ou de cartes après avoir été archivées sur des plateformes infonuagiques.

Parmi toutes les variables météorologiques contrôlées par Solutions Mesonet, celle qui représente le plus gros défi en termes de difficulté d'identification des données suspectes est

probablement celle des précipitations. En effet, les précipitations sont caractérisées par une grande variabilité spatiale et temporelle. De plus, cette variable météorologique est d'un intérêt tout particulier pour de nombreux utilisateurs tels Hydro-Québec, Rio Tinto, la SOPFEU, etc. Les données de précipitations sont validées et corrigées quotidiennement par les équipes en charge de l'assurance qualité. Ces données représentent un volume très important, puisque plusieurs centaines de stations sont en service au Québec. De plus, la résolution temporelle des observations est horaire, donc chacune d'entre elles envoie 24 observations de précipitations par jour. Ainsi, l'identification des données météorologiques suspectes constitue une tâche exigeante en ressources humaines.

Un outil permettant la détection automatique des valeurs suspectes constituerait une façon d'augmenter l'efficacité des équipes chargées de l'assurance qualité. Ce premier tri automatique permettrait d'identifier rapidement les données qui doivent être validées par les experts. De cette façon, l'équipe d'assurance qualité pourrait se concentrer sur un nombre plus restreint de données à contrôler et plus largement améliorer son efficacité opérationnelle. Ainsi, l'objectif de ce mémoire est d'améliorer l'identification des données suspectes de précipitations journalières enregistrées sur l'ensemble du territoire québécois

Le présent mémoire se décompose en cinq parties, qui débutent par cette partie introductive. Le chapitre 2 est dédié à une revue de la littérature en matière d'identification de données suspectes dans le domaine météorologique, afin de dresser le portrait des dernières avancées dans ce domaine ainsi que des potentielles avenues possibles pour ce projet de recherche. Le chapitre 3 est dédié à la méthodologie utilisée pour automatiser l'identification de données de précipitations suspectes. Ensuite vient le chapitre 4 qui expose le cas d'étude spécifique du projet en collaboration avec le partenaire industriel, avec les données utilisées ainsi que les résultats obtenus. Enfin, le chapitre 5 discute de ces résultats, souligne les limites de l'outil, questionne les perspectives envisagées pour des travaux futurs et établit la conclusion de ce projet de recherche.

CHAPITRE 2 REVUE DE LITTÉRATURE

L'identification des données météorologiques suspectes est un processus clé à plusieurs égards. En effet, les données brutes provenant des capteurs (thermomètres, précipitomètres) sont parfois entachées d'erreurs. Ces imprécisions peuvent provenir de différentes causes : instrumentales (défaillance de capteur, maintenance inadéquate) ou météorologiques (pertes en évaporation, sous-captation due au vent, *snow capping*)(Leeper et Kochendorfer, 2015; Goodison *et al.*, 1998; Steinacker, 2011; Rasmussen *et al.*, 2012). Ainsi, les informations recueillies par les capteurs sur diverses variables météorologiques d'intérêt, comme la température et les précipitations, sont imparfaites et contiennent une proportion de valeurs aberrantes. Ces valeurs aberrantes peuvent avoir un impact négatif sur plusieurs activités. En particulier, l'intégration de ces valeurs erronées peut nuire à la prise de décision pour la planification ou la sécurité des infrastructures.

Par ailleurs, l'expansion des réseaux de stations météorologiques a mené à l'augmentation du volume des données à traiter. Cela oblige les organismes intéressés par l'identification des données suspectes à réfléchir à des méthodes pouvant être déployées à grande échelle. En effet, l'identification des données suspectes est chronophage et coûteuse en ressources lorsqu'elle est effectuée par des experts. Ainsi, des méthodes de validation automatiques peuvent s'avérer utiles dans le cas où le volume de données est élevé : c'est pourquoi certaines sont déjà implémentées dans des organismes responsables de l'assurance qualité des données météorologiques. Nous proposons de présenter les études sélectionnées selon 4 catégories. Cette catégorisation synthétise les principales méthodes que nous avons jugées pertinentes pour l'identification de données météorologiques suspectes.

- tests statistiques classiques ;
- méthodes géostatistiques (basées sur les stations voisines) ;
- méthodes spatio-temporelles ;
- méthodes d'apprentissage automatique.

Nous passerons en revue ces différentes méthodes et analyserons leurs avantages et leurs inconvénients respectifs. Dans la suite du document, nous nous référerons à la station dont nous voulons identifier les données suspectes par la station d'intérêt et l'observation particulière actuellement soumise à ce contrôle, par l'observation d'intérêt.

2.1 Tests statistiques classiques

En identification des données suspectes, certains tests sont généralement appliqués pour effectuer un tri préliminaire dans les données météorologiques. Ils permettent d'éviter que des mesures contenant des erreurs grossières soient conservées dans les produits finaux et transmises aux utilisateurs. Certains d'entre eux sont basés sur les séries temporelles des variables, comme les tests de cohérence (Artz *et al.*, 2023), les tests d'amplitude (Boulanger *et al.*, 2010) ou encore les tests de pics (Øgland, 1993). La limite principale de ces tests est qu'ils nécessitent l'intégralité de la série temporelle et ne gèrent pas les valeurs manquantes. Par ailleurs, des tests d'homogénéité (Dyck, 1976; Pearson, 1900; Berger et Zhou, 2014; Wallace, 1959) peuvent être menés pour détecter un changement dans les phénomènes physiques sous-jacents aux précipitations, et même combinés entre eux (Hofmeister *et al.*, 2023). Ces derniers ne peuvent néanmoins pas caractériser quantitativement ce changement, et cette limite est prise en compte par les tests de tendance (Mann, 1945; Kendall, 1975; Hofmeister *et al.*, 2023; Schönwiese, 2000), ou encore les tests de détection de rupture (Bernier, 1994; Pettitt, 1979; Xiong et Guo, 2004). Ces tests permettent donc de détecter des données suspectes qui sont des erreurs évidentes, mais ne sont pas suffisants pour effectuer une identification des données suspectes complète et approfondie, c'est pourquoi ils sont souvent suivis par des vérifications manuelles par des experts. Dans le cadre de ce projet, ces tests ne sont pas très pertinents, puisque des tests similaires sont déjà implémentés dans le processus d'assurance qualité de Solutions Mesonet.

Néanmoins, bien que les derniers tests mentionnés indiquent qu'un changement a eu lieu dans le comportement des données, ils n'indiquent pas si cela est dû à une anomalie de capteur ou à un réel changement dans le phénomène de précipitation. Ainsi, prendre en compte l'information provenant d'autres stations météorologiques à proximité pourrait aider à répondre à cette question, et c'est l'objet de la section suivante, dédiée aux méthodes géostatistiques.

2.2 Méthodes géostatistiques (basées sur les stations voisines)

La géostatistique est une branche de la statistique qui se concentre sur la modélisation spatiale des variables (Rivoirard, 2005). Celle-ci contient notamment la méthode de pondération inverse à la distance (Di Piazza *et al.*, 2011), ou encore le krigeage (Matheron, 1963). Cette dernière méthode est la plus populaire en géostatistique : elle repose sur la modélisation spatiale d'un champ gaussien à partir d'observations ponctuelles. En identification des données suspectes, on peut estimer la quantité de précipitations survenues à une station d'intérêt via

l'une ou l'autre de ces méthodes à l'aide des stations voisines, et comparer cette valeur estimée à la valeur réellement mesurée. Si la mesure tombe en dehors d'un certain intervalle entourant la valeur estimée, alors l'observation peut être considérée comme suspecte. Ces deux méthodes performent généralement moins bien dans des régions où les stations d'observation sont moins denses, et subissent la forte variabilité spatiale du phénomène de précipitation. Ainsi, les méthodes géostatistiques ne sont pas très intéressantes pour ce projet car la densité des stations météorologiques varie fortement au Québec : le sud de la province est beaucoup plus densément doté que le nord.

Après avoir étudié les tests prenant en compte uniquement la série temporelle des stations individuellement, puis les tests incorporant les données d'autres stations mais uniquement pour un instant donné, il est temps d'évoquer l'étape suivante, à savoir les approches qui intègrent à la fois de la dépendance spatiale et de la dépendance temporelle : les méthodes spatio-temporelles.

2.3 Méthodes spatio-temorelles

Les méthodes spatio-temporelles ont l'avantage de regrouper l'information sur différentes stations et sur différents instants, intégrant donc une notion de dépendance spatiale et de dépendance temporelle. Le test de régression spatiale (You et Hubbard, 2006) en fait partie, tout comme la méthode *Multiple intervals gamma distribution* (You *et al.*, 2007). Le test de régression spatiale prend en compte des observations des stations voisines, mais intègre également les données qui précèdent et succèdent l'observation d'intérêt (voisins temporels), ce qui permet une identification de valeurs suspectes plus riche que lorsqu'on se concentre sur un instant unique. Cette méthode est en effet supérieure à la pondération inverse à la distance (Hubbard et You, 2005). Une version probabiliste de ce test (Xu *et al.*, 2014) permet de quantifier le niveau de confiance associé à une observation. D'autres méthodes plus avancées comme le P-BSHADE (Xu *et al.*, 2013) incorporent davantage de données et sont donc plus performantes pour détecter les valeurs suspectes. Enfin, le cadre bayésien peut être adopté pour traiter rigoureusement l'incertitude (Ingleby et Lorenc, 1993; Gelfand *et al.*, 2005), et d'autres sources de données peuvent être utilisées pour enrichir l'analyse, comme des données radar (Yan *et al.*, 2024).

Ainsi, ces méthodes spatio-temporelles combinent l'information sur la dépendance spatiale et la dépendance temporelle des processus à l'œuvre, ce qui améliore les performances de détection de valeurs suspectes. Malgré tout, ces méthodes ne sont pas utilisables si on est en présence de valeurs manquantes, puisqu'elles nécessitent de disposer des séries temporelles complètes. Dans le contexte de ce projet de recherche, il peut arriver que certaines valeurs

soient manquantes, donc ces méthodes ne sont pas nécessairement utilisables. Ainsi il serait intéressant de trouver une méthode qui incorpore seulement les voisins spatiaux (et non les voisins temporels), mais qui soit plus souple et plus performante que les méthodes géostatistiques telles que le krigeage. En effet, la majorité des tests passés en revue jusqu'à présent relèvent d'une modélisation paramétrique, où l'on suppose que les observations suivent une loi de probabilités appartenant à une certaine famille de lois. Ceci nous permet de tirer profit des outils probabilistes qui en découlent, comme les intervalles de confiance, qui servent de critère pour effectuer l'identification des données suspectes. Cependant, d'autres méthodes relèvent du cadre non-paramétrique, où aucune hypothèse préalable sur la distribution des observations n'est requise. Ces méthodes sont parfois utiles lorsque la modélisation statistique n'est pas aisée, mais le coût est au niveau de la complexité de ces approches, qui sont par ailleurs souvent moins facilement interprétables que leurs homologues paramétriques. Ces techniques font l'objet de la prochaine section.

2.4 Méthodes d'apprentissage automatique

L'apprentissage automatique est une branche de l'intelligence artificielle qui se base sur les statistiques pour permettre à des algorithmes d'améliorer leurs performances dans la réalisation de tâches sans être explicitement programmés pour chacune. Des méthodes de détection d'anomalies existent donc dans cette branche, et utilisent des algorithmes génétiques (Xiong *et al.*, 2017), des arbres de décision (Xiong *et al.*, 2022) ou encore des données fonctionnelles spatiales (Burbano-Moreno et Mayrink, 2024). Ces méthodes sont performantes mais elles impliquent un certain coût lié à leur complexité et à la subtilité de leur paramétrage.

Par ailleurs, l'apprentissage profond est un sous-ensemble de l'apprentissage automatique qui fait intervenir des réseaux de neurones. Cette discipline a connu un essor considérable avec l'amélioration de la capacité des ordinateurs et la disponibilité de données massives, et connaît depuis lors un fort engouement dans la communauté scientifique et plus largement dans la société.

L'apprentissage profond a bouleversé bon nombre d'usages et a eu un très fort impact sur la société du XXI^{e} siècle. Cet outil est très puissant, car il permet d'effectuer des prédictions d'une précision impressionnante, rendant caducs bon nombre de méthodes utilisées historiquement dans un très large éventail de domaines. En particulier, un certain type d'architecture nous intéresse singulièrement pour notre sujet : il s'agit des réseaux de neurones convolutifs (ou CNN, pour *Convolutional Neural Network* en anglais). Ce type de modèles est particulièrement adapté à l'analyse des structures en deux dimensions. Par exemple, il peuvent servir aux analyses d'images, qui sont généralement composées de trois matrices

représentant l'intensité des pixels dans les trois canaux de couleurs (rouge, vert, bleu) (LeCun *et al.*, 2015; Goodfellow *et al.*, 2016). Ces réseaux convolutifs sont notamment utilisés en apprentissage supervisé pour effectuer de la classification binaire (Konda *et al.*, 2019) ou multi-classes (Li *et al.*, 2016). Cette méthode est basée sur des couches convolutives, et fait partie de l'apprentissage supervisé. Ainsi, il est possible d'utiliser cette architecture de modèles pour effectuer l'identification de données météorologiques suspectes, comme décrit dans le prochain paragraphe.

Les données météorologiques sont parfois représentées par le biais d'analyses, qui sont fournies sous forme de matrice de cellules, chaque cellule correspondant à une zone géographique définie en fonction de la résolution spatiale du produit. Ainsi, ces matrices de cellules peuvent être considérées comme des images, où chaque valeur de cellule correspond à une intensité de pixel. Ainsi, utiliser des outils d'apprentissage automatique (et notamment d'analyse d'images comme les réseaux de neurones convolutifs) pour identifier des données météorologiques suspectes paraît pertinent. Seulement, encore faut-il savoir quels types de données utiliser en entrée, et comment les assembler ou les intégrer afin d'établir une méthode efficace. Sha *et al.* (2021) proposent une méthode pour identifier des données de précipitation suspectes qui intègre différents types de données : observations aux stations météorologiques et données surfaciques d'analyse, sous forme de grille (prévisions de précipitations par un modèle météorologique et altitude). C'est la seule étude, à notre connaissance, à avoir utilisé des réseaux de neurones convolutifs pour identifier des mesures de précipitations suspectes.

Dans leur étude, Sha *et al.* (2021) considèrent le voisinage de chaque station d'intérêt, pour les données en grille, de dimensions 64×64 pixels, pour tous les instants d'intérêt. Pour chaque observation provenant d'une station, on dispose donc de son voisinage en termes de grilles d'analyse de précipitation et d'altitude. Ensuite, cette observation est intégrée dans la grille d'analyse de précipitation en remplaçant ses 4 valeurs centrales par la valeur de l'observation d'intérêt. En parallèle, ce procédé est répété pour différentes résolutions spatiales, chacune permettant de représenter des motifs spatiaux plus ou moins grossiers de champs de précipitations. Puis, un réseau de neurones convolutifs est entraîné sur une partie du jeu de données étiquetées (étiquettes de qualité : donnée authentique ou donnée suspecte) manuellement par des équipes d'assurance qualité. Cela est en adéquation avec notre projet, puisque le processus d'assurance qualité de Solutions Mesonet (basé sur des analyses préliminaires automatisées et sur des analyses manuelles effectuées à l'aide de différentes cartes météorologiques, de séries temporelles et de données radar) produit également des étiquettes de qualité sur les observations de précipitation.

L'entraînement du réseau de neurones constitue donc un apprentissage supervisé pour une

classification binaire. Les étiquettes de qualité définies par les experts sont considérées comme la vérité, et servent de base à l'apprentissage du modèle de classification. En effet, la confiance accordée à ces étiquettes est grande puisque le processus d'assurance qualité est rigoureux, approfondi et fait intervenir des spécialistes expérimentés qui vérifient manuellement la qualité des observations. L'entraînement est réalisé pour chaque résolution spatiale, puis les réseaux sont couplés grâce à un perceptron multi-couches (MLP, pour *Multi-Layer Perceptron* en anglais) qui tire profit des différentes résolutions pour réaliser la meilleure prédiction possible. Cette méthode apparaît comme étant particulièrement pertinente, dans la mesure où elle tire profit de la capacité des CNN à extraire des motifs significatifs pour prédire la qualité d'une observation en fonction de son environnement décrit par des produits d'analyse. L'orographie ayant également un impact sur les précipitations, l'intégration des données d'altitude est également un point permettant d'améliorer la justesse des prédictions.

En revanche, une critique qui peut être faite à cette méthode est l'interprétabilité des résultats, qui est plus difficile que pour des méthodes statistiques plus classiques, puisque les réseaux de neurones sont en quelque sorte une «boîte noire» qui prend des décisions sur des éléments qui ne sont pas évidents à déterminer. Cependant, les auteurs de l'article sont conscients de cette limite et dédient d'ailleurs une section à ce sujet, et implémentent une méthode qui utilise l'*Empirical Orthogonal Function*, afin de produire des cartes représentant les zones sur lesquelles le modèle se concentre le plus pour classer les observations. Malgré tout, la méthode décrite ci-dessus est pertinente dans le cadre de notre étude, car les données requises correspondent à celles disponibles dans le projet de recherche décrit dans ce mémoire, à une subtilité près. En effet, bien qu'elle repose sur un produit de réanalyse intégrant l'assimilation de données, les observations des stations météorologiques utilisées dans l'étude de Sha *et al.* (2021) ne sont pas incluses dans ce processus. Cela permet d'éviter que ces observations influencent à la fois la réanalyse et l'évaluation. Ainsi, pour garantir cette même rigueur dans notre cas, il sera nécessaire d'utiliser un produit de réanalyse sans assimilation de données. En outre, les étiquettes de qualité produites par Solutions Mesonet peuvent servir de base d'apprentissage à un modèle de classification, et les différentes sources de données nécessaires peuvent être récupérées pour le contexte du Québec. En conclusion, cette méthode apparaît innovante et pertinente, en plus d'obtenir des performances très intéressantes.

Nous proposons donc de nous intéresser plus précisément à la notion de réseaux de neurones convolutifs, qui sont basés sur une théorie reposant sur des outils spécifiques. Ces outils sont présentés brièvement dans la section suivante de façon à mieux appréhender les mécanismes à l'œuvre dans ces réseaux.

2.5 Réseaux de neurones convolutifs

Cette section est dédiée à une brève introduction aux éléments essentiels pour comprendre la classification par réseau de neurones convolutifs. Une compréhension générale du fonctionnement de ces réseaux permet de mieux appréhender les mécanismes à l'œuvre dans la méthode proposée.

2.5.1 Classification binaire en apprentissage supervisé

L'apprentissage supervisé est une branche de l'apprentissage automatique dans laquelle le modèle a accès à un jeu de données qui contient des exemples annotés, qui lui servent de base d'apprentissage. Par exemple, un modèle réalisant une classification entre images de chiens et images de chats, aurait donc accès à un jeu de données d'entraînement contenant des images de chiens et des images de chats, auxquelles seraient associées des étiquettes identifiant la classe chien ou chat de chaque image. Le modèle se baserait sur cet ensemble annoté pour apprendre des données durant la phase d'entraînement. Ensuite, on viendrait mesurer ses performances sur un autre jeu de données, l'ensemble de test, également composé d'images de chiens et de chats. Seulement, cette fois on ne dévoilerait pas au modèle les étiquettes indiquant la classe de chaque image, mais on lui laisserait faire les prédictions en fonction de son apprentissage sur l'ensemble d'entraînement. Cela permet d'évaluer sa capacité de généralisation, c'est-à-dire son habileté à apprendre une fonction permettant de réaliser des prédictions cohérentes sur des données non-présentes au moment de l'apprentissage.

Dans cet exemple, la classification est qualifiée de binaire, car elle a pour objectif d'effectuer une décision entre deux classes : la classe chien et la classe chat. D'autres cas de figure, comme la classification d'un jeu de données de chiffres manuscrits entre 0 et 9, n'est pas binaire mais multi-classes.

Dans notre étude, on propose de classer les précipitations mesurées aux stations en deux classes : la classe des données authentiques et celle des anomalies. La classe des données authentiques est dénotée par 0, tandis que la classe des données suspectes par 1. À présent, voyons comment le modèle de classification se sert des couches convolutives pour apprendre des motifs de l'ensemble d'entraînement afin d'estimer la classe des précipitations.

2.5.2 Filtres convolutifs

Un filtre convolutif, autrement appelé noyau de convolution, est une strate de neurones artificiels qui partagent tous les mêmes paramètres. Ces neurones s'appliquent sur une zone spécifique de l'image, et puisqu'ils appliquent tous la même fonction, on peut le voir comme

un filtre qui comporte des poids. Ce filtre peut être de dimension variable, et s'applique à des zones de dimension correspondante. Il parcourt toute l'image et s'applique à chaque zone, en transmettant à la couche suivante le résultat de la transformation du filtre sur chaque zone de l'image. Ce résultat dépend des poids du filtre, qui seront ajustés au cours de la phase d'entraînement. À la figure 2.1 se trouve un schéma explicatif du mécanisme de convolution.

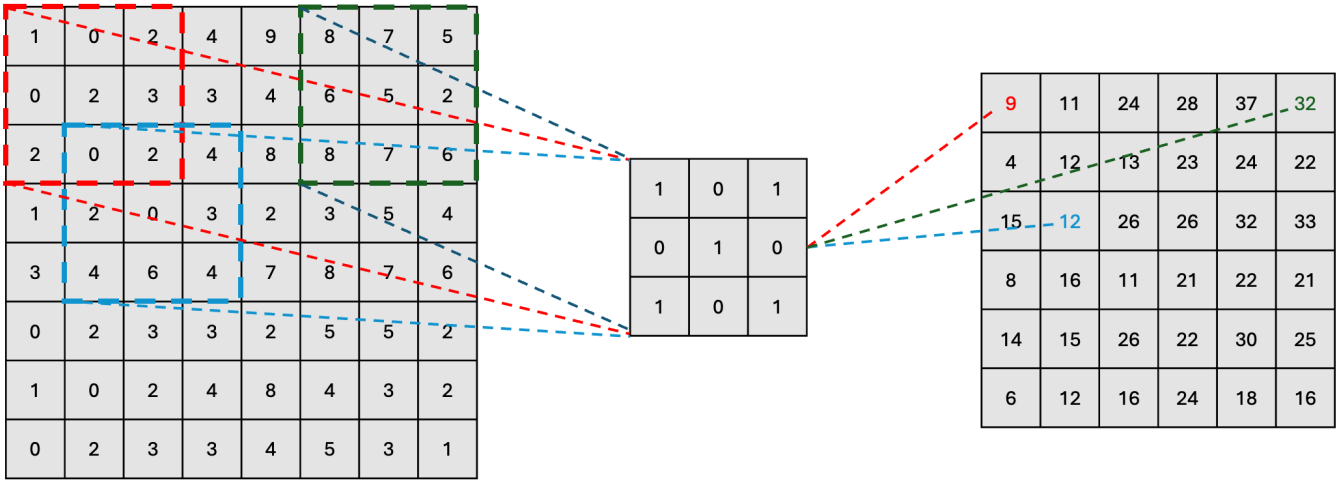


FIGURE 2.1 Processus de convolution d'une image.

À gauche de la figure, on note l'image qui est fournie en entrée au filtre convolutif, ce dernier est ici de dimension 3×3, et il est représenté au centre. L'image de droite représente la sortie de la couche convolutive, qui comme on peut le remarquer est de dimension inférieure à l'image d'entrée. Ces images sont matérialisées par des matrices de pixels, chacun d'entre eux comportant une valeur représentant une intensité. Ainsi, le filtre convolutif va parcourir l'image intégralement, et à chacune de ses positions il va réaliser la convolution 2D avec la zone de l'image sur laquelle il est présentement positionné. Le résultat de cette convolution sera ensuite donné en sortie, et composera un pixel de l'image appelée *feature map*. Une fois toutes les convolutions réalisées sur l'image d'entrée, l'image de sortie comportera tous les résultats des convolutions. Dans l'exemple de la figure, l'image de sortie est une matrice carrée de dimension 6 quand l'image d'entrée est elle de dimension 8. Ceci est logique puisque le filtre de convolution est de dimension 3, donc peut seulement se déplacer de 6 pixels verticalement et horizontalement, d'où la dimension de la *feature map*. L'expression générale de la convolution 2D d'une image est donnée par

$$g_{x,y} = \omega \ * \ f_{x,y} = \sum_{i=-a}^{a} \sum_{j=-b}^{b} \omega_{i,j} f_{x-i,y-j}$$

où $g_{x,y}$ représente le pixel de position (x, y) de la *feature map*, ω représente le filtre convolutif et f l'image d'entrée.

Dans le cas de la figure 2.1, $a = b = 1$, puisque le filtre est de dimension 3 : ses indices sont donc $-1, 0, 1$ pour les axes vertical et horizontal. De façon imagée, on peut représenter une convolution comme la somme de termes obtenus par la multiplication pixel à pixel de la zone de l'image au filtre, que l'on aurait préalablement retourné selon les axes horizontaux et verticaux. Le filtre de la figure de l'exemple étant symétrique selon ces deux axes, un tel double retournement est sans effet, mais il est nécessaire dans le cas général.

Des modifications de natures différentes peuvent être apportées au processus de convolution, notamment les notions appelées *stride* et *padding*. Le *stride* représente le nombre de pixels duquel le filtre va se décaler en parcourant l'image. Un *stride* de 2 par exemple signifie que le filtre va, d'une convolution à l'autre, se déplacer de 2 pixels. Augmenter le *stride* a donc pour effet de réduire la dimension de la *feature map*, l'image qui résulte de la convolution, puisque le filtre effectue de plus grands bonds sur l'image, entraînant une analyse plus grossière, moins détaillée. Le *padding*, quant à lui, concerne la gestion des bordures de l'image (voir section 2.5.3). Ces deux notions, en plus de la dimension du filtre, qui elle aussi peut être modifiée, rendent le processus de convolution assez ajustable et flexible. Cela permet aux modèles associés à ces processus de convolution de traiter un large éventail de types d'images différents.

Pour les images en couleur, qui sont en réalité la superposition de 3 canaux d'intensité (canaux RGB, pour les couleurs rouge, vert et bleu), des filtres indépendants peuvent s'appliquer sur les différents canaux, ces derniers étant recombinés par la suite. Pour les images en noir et blanc (i.e. niveaux de gris), seul un canal est nécessaire, ce qui réduit le coût computationnel. Dans le cas de notre étude, nous nécessiterons, à l'instar des images RGB, plusieurs canaux en entrée de notre réseau de neurones convolutif.

Ces filtres de convolution agissent comme des détecteurs de motifs, qui agissent sur une entrée et font donc ressortir certaines caractéristiques de l'image (contrastes, bordures, motifs). De cette façon, la sortie indique où et à quel degré ces caractéristiques visées sont présentes dans l'image. Comme ces filtres sont appris durant la phase d'entraînement, ils sont ajustés automatiquement de façon à cibler les caractéristiques pertinentes en fonction de la tâche à effectuer (classification par exemple). Ainsi, la sortie d'un filtre de convolution est une version transformée de l'entrée qui met en évidence ce qui est important pour réaliser la tâche d'intérêt, et qui gomme ce qui ne l'est pas.

2.5.3 Hyperparamètres des réseaux de neurones convolutifs

Les CNN peuvent être ajustés par le biais de plusieurs paramètres modifiables, appelés hyperparamètres, dont la fonction est explicitée dans la section courante.

Nombre de couches convolutives et dimension des filtres

Le nombre de couches de convolutions implémentées est un hyperparamètre important des CNN : augmenter celui-ci accroît le niveau d'abstraction du modèle, qui peut donc traiter des motifs plus complexes, en plus de traiter l'image de façon plus globale. Néanmoins, cette complexité accrue entraîne un coût computationnel plus élevé : davantage de paramètres doivent être entraînés, ce qui allonge le temps de calcul. Par ailleurs, la dimension des filtres convolutifs joue un rôle complémentaire au nombre de couches. En effet, des filtres de petite dimension permettent une analyse locale et fine, tandis que de grands filtres offrent davantage de couverture, incorporent donc plus de contexte, mais sont moins adaptés à la détection de détails fins.

Padding

La notion de *padding* est répandue en apprentissage profond, car elle permet à la sortie d'une couche de convolution de conserver la dimension de l'image d'entrée. En effet, le padding consiste en l'ajout d'une couche de pixels autour de l'image d'entrée, ce qui a pour effet bénéfique de faire en sorte que tous les pixels de l'*input* soient pris en compte par les filtres convolutifs le même nombre de fois. La politique d'attribution de la valeur de ces pixels peut donc être déterminée soit de façon fixe, comme le *zero-padding*, qui consiste simplement en l'ajout de pixels nuls sur le pourtour de l'image et pour une largeur donnée, soit de façon adaptative en fonction de l'image, comme le *reflection padding* ou encore le *replication padding*, qui se basent sur les pixels bordant l'image pour l'ajout des nouveaux pixels entourant l'image originale. En réalité, en l'absence de *padding*, les pixels dans les coins de l'image sont moins souvent pris en compte par les filtres convolutifs que ceux sur les bords, qui sont eux-mêmes moins souvent pris en compte que les pixels strictement à l'intérieur de l'image. Ce phénomène entraîne un différentiel dans l'importance relative des pixels, d'où l'intérêt du *padding* qui règle le problème en ajoutant artificiellement des pixels supplémentaires sur les bords de l'image, rendant ainsi tous les pixels de l'image originale équitablement importants dans le traitement convolutif appliqué. Cette technique peut permettre d'obtenir des modèles plus robustes, dont les prévisions sont plus fiables. En particulier, nous allons implémenter un certain type de *padding* appelé *zero-padding*, qui consiste simplement en l'ajout d'une couche

de pixels dont les valeurs sont toutes égales à 0.

Après avoir mentionné les hyperparamètres des CNN, il est utile de s'intéresser à certaines notions inhérentes au fonctionnement même de ces réseaux de neurones : structure générale, apprentissage, séparation en ensembles d'entraînement et de test et sortie du modèle sont le sujet des prochaines sections.

2.5.4 *Convolve & pool*

Généralement, les filtres convolutifs sont utilisés conjointement avec des couches de *max-pooling*. Le principe du *maxpooling* est relativement simple, puisqu'il s'agit de réduire la dimension de l'image en conservant uniquement le pixel qui présente la valeur maximale pour chaque carré de dimension 2×2 qui constitue l'image. Ainsi, la dimension est réduite de moitié en appliquant une couche de *maxpooling*, et il est courant d'alterner couches convolutives et couches de *maxpooling* : on parle du schéma *convolve and pool*. Il est également courant d'utiliser une architecture de *multilayer perceptron* en sortie des couches de convolution et de *pooling* de façon à ce que la sortie du réseau soit un réel entre 0 et 1, ce qui est souhaitable pour une tâche de classification.

2.5.5 *Batch normalization* et fonctions d'activation

Il est courant d'utiliser le principe de *batch normalization* en apprentissage profond, afin de rendre l'entraînement plus rapide et plus stable. Il consiste en une mise à l'échelle, puisque chaque *batch* de données, c'est-à-dire chaque sous-ensemble d'entraînement, est normalisé de façon à avoir une moyenne et une variance prédéfinie. Les fonctions d'activation, quant à elles, sont utilisées en sortie des couches de convolution afin d'assurer que les sorties se situent dans un intervalle souhaité. Une fonction d'activation répandue est la fonction *Rectified Linear Unit*, qui assure que les sorties soient positives. Elle est définie sur les réels par :

$$\mathrm{ReLU}(x) = \max(0, x)$$

2.5.6 Fonction de perte et descente de gradient

Les convolutions successives permettent au réseau d'apprendre des motifs et des caractéristiques signifiantes des images, grâce à l'algorithme de rétro-propagation qui permet aux poids des filtres convolutifs de s'ajuster tout au long de la phase d'entraînement. En effet, les poids des filtres sont des neurones à part entière du réseau, qui s'adaptent pour minimiser la fonction de perte choisie. La fonction de perte choisie dans notre cadre de classification

binaire est celle de l'**entropie croisée binaire**, définie par

$$l(x, y) = -\{y \ln x + (1 - y) \ln(1 - x)\}$$

où x est la classe prédite de l'observation et y est l'étiquette réelle de l'observation.

Dans notre cas d'étude, x et y appartiennent à l'ensemble $\{0, 1\}$, ce qui théoriquement peut entraîner des valeurs de perte infinies. En pratique, dans les librairies informatiques dédiées à l'apprentissage profond, le logarithme est tronqué inférieurement, de façon à éviter que la fonction de perte prenne des valeurs infinies, et ce afin de rendre la valeur de la fonction de perte finie et d'être plus adaptée pour effectuer la méthode de rétro-propagation. En particulier, dans la librairie PyTorch de Python, la fonction `BCELoss` tronque son logarithme à -100, qui ne pourra donc pas descendre en-dessous de cette valeur.

En général, en apprentissage profond, les jeux de données sont partitionnés en sous-ensembles appelés *batchs*. La perte est calculée pour chaque *batch*, que l'on suppose de taille n, en prenant simplement la moyenne des pertes des prédictions individuelles :

$$\ell(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{n} \sum_{i=1}^{n} l(x_i, y_i)$$

où $\boldsymbol{x}$ et $\boldsymbol{y}$ représentent les vecteurs de classes prédites et des étiquettes réelles correspondant à l'intégralité du *batch*, et où x_i et y_i pour $i = 1, \ldots, n$ sont les composantes du *batch*, à savoir la classe prédite, respectivement la vraie étiquette de la i^{e} donnée du *batch*.

Forts de cette fonction de perte, le principe est ensuite de la minimiser en utilisant des algorithmes d'optimisation, comme l'algorithme Adam, dont l'utilisation est très répandue en apprentissage profond. En effet, ce procédé permet, en calculant la fonction de perte pour chaque *batch*, d'effectuer une descente de gradient en utilisant le principe de rétro-propagation. Ce dernier permet de partir du gradient des couches de sortie pour le propager vers l'arrière, à savoir en direction des couches d'entrée, mettant ainsi au passage les poids ajustables du réseau à jour en fonction de la direction de descente du gradient de la fonction de perte. Ainsi, en effectuant plusieurs itérations au cours desquelles le réseau va se confronter aux mêmes données, la fonction de perte va être calculée puis les poids s'ajusteront automatiquement de façon à faire décroître la perte. À chaque itération la perte va donc diminuer, augmentant ainsi la précision et la pertinence du modèle, dont les prédictions s'améliorent à chaque itération, souvent appelée *epoch* en apprentissage profond.

D'une *epoch* à l'autre, les poids sont ajustés grâce à l'optimiseur de notre choix, à savoir l'algorithme Adam (Kingma et Ba, 2017), qui réalise une optimisation stochastique efficace

en se basant sur les moments d'ordre 1 et 2 du gradient. Il se base sur les avancées de deux autres algorithmes antérieurs, à savoir AdaGrad et RMSProp. Cet optimiseur est crucial car il est responsable de la mise à jour des poids des neurones, rôle central dans l'apprentissage supervisé.

2.5.7 Ensemble de validation, *dropout* et *early stopping*

Comme mentionné précédemment, les modèles d'apprentissage automatique apprennent sur un jeu de données d'entraînement et leurs performances sont évaluées sur un jeu de données de test, disjoint du premier, pour également s'assurer de sa capacité de généralisation à des données non encore rencontrées. En effet, il peut parfois arriver que le modèle soit trop flexible et s'approche excessivement des données d'entraînement : c'est le phénomène de surajustement, qui mène souvent à des performances plus mitigées sur l'ensemble de test. Pour se prémunir de ce fléau, on peut agir dans un premier temps directement sur les neurones du réseau. En effet, en désactiver certains est une manière d'empêcher le réseau de se rapprocher excessivement de l'ensemble d'entraînement. Ainsi, on peut envisager le *dropout*, une notion qui consiste à mettre à 0, suivant une certaine probabilité, les neurones d'une certaine couche du réseau. Même si d'autres valeurs peuvent être efficaces, une valeur communément utilisée pour cette probabilité est de 0.5, on propose donc d'utiliser cette valeur pour notre méthode.

En outre, il est possible de scinder l'ensemble d'entraînement en deux sous-ensembles : l'un sera dédié à l'entraînement, tandis que l'autre servira d'ensemble de validation. Ce dernier ensemble a pour vocation d'évaluer la capacité de généralisation du modèle au cours même de la phase d'apprentissage, puisque ses données ne servent pas à ajuster les neurones mais sont simplement utilisées pour caractériser les performances du réseau. Contrairement à l'ensemble de test, qui lui ne sert qu'après la fin de la phase d'apprentissage, l'ensemble de validation sert à chaque *epoch* pour donner un indicateur de la tendance du modèle, notamment pour éviter le surajustement. En effet, si la fonction de perte associée à l'ensemble de validation baisse, alors le modèle s'améliore et il mérite donc de continuer son entraînement, mais si au contraire celle-ci recommence à augmenter, alors c'est le signe que le réseau généralise moins bien et qu'il vaut mieux stopper l'entraînement pour se prémunir du surajustement : c'est la méthode dite d'*early stopping*.

En pratique, on peut décider d'un nombre maximal d'*epochs* au cours desquelles la fonction de perte sur l'ensemble de validation n'a pas baissé. Au-delà de ce nombre, l'entraînement s'interrompra, et les performances du modèle seront évaluées sur l'ensemble de test. Dénotons par K ce paramètre correspondant au nombre maximal d'*epochs* sans baisse de la fonction de perte.

2.5.8 Sortie du réseau et décision

Pour s'assurer que la sortie de notre modèle sera un réel compris entre 0 et 1, on utilise comme dernière couche du réseau une fonction d'activation appelée sigmoïde, dont l'utilisation est très répandue en apprentissage profond. Cette fonction a pour ensemble d'arrivée l'intervalle $[0, 1]$, ainsi quel que soit le réel résultant de la succession des couches convolutives et du *multilayer perceptron*, son image par la fonction sigmoïde sera comprise entre 0 et 1. Cette sortie peut donc être interprétée comme une classe (0 ou 1) en ayant fixé un seuil de classification : ce dernier est par défaut fixé à 0.5. Ainsi, dans notre contexte, toute observation dont la sortie par le réseau est supérieure à 0.5 est considérée comme une anomalie, et au contraire considérée comme donnée de qualité si la sortie correspondante est inférieure à ce seuil. Plus formellement, définissons l'ensemble des décisions possibles, dénoté $\mathcal{A}$:

$$\mathcal{A} = \{d_0, d_1\},$$

avec d_i : l'observation appartient à la classe i, $\quad i = 0, 1$.

On peut donc définir la règle de décision d concernant la classe estimée de l'observation (0 ou 1) de la façon suivante :

$$d = d_{\mathbb{1}\{x>0.5\}} \tag{2.1}$$

où x est la sortie du réseau de neurones auquel on a donné en entrée l'observation dont on veut estimer la classe.

Ainsi, si $d = d_1$, l'observation est considérée comme appartenant à la classe 1 (anomalie), et si $d = d_0$ alors l'observation est prédite comme appartenant à la classe 0 (bonne qualité).

2.5.9 Évaluation de la classification

Pour réaliser l'évaluation des performances d'un outil de classification binaire, il est nécessaire de se baser sur des métriques objectives, adaptées au contexte d'études. Dans notre cas, il s'agit d'une classification binaire, et il est donc pertinent de tenir compte du nombre de vrais positifs, de vrais négatifs, de faux positifs et de faux négatifs. En effet, chaque élément est assigné à une classe, qui est soit la classe 0 (pour les données de bonne qualité), soit la classe 1 (pour les valeurs suspectes). Pour évaluer les performances des modèles implémentés, une métrique naturelle est la proportion d'observations correctement classées. Soit TP (*true positive*), TN (*true negative*), FP (*false positive*), respectivement FN (*false negative*) le nombre de vrais positifs, de vrais négatifs, de faux positifs, respectivement de faux négatifs. Alors, la proportion de données correctement classées, parfois appelée précision et notée p

est définie par :

$$p = \frac{TP + TN}{TP + TN + FP + FN}$$

Cette proportion p est un compromis entre deux métriques qui donnent davantage d'informations sur la propension du modèle à produire des faux positifs et des faux négatifs : ces deux métriques sont le taux de faux positifs (noté FPR, pour *false positive rate*), et le taux de faux négatifs (noté FNR, pour *false negative rate*), définis par :

$$FPR = \frac{FP}{FP + TN} \qquad\qquad FNR = \frac{FN}{FN + TP}$$

En particulier, on peut retrouver la quantité p grâce aux index FPR et FNR, puisque l'on a :

$$p = 1 - (QR \times FPR + AR \times FNR)$$

avec QR pour *quality rate* et AR pour *anomaly rate* définis par :

$$QR = \frac{TN + FP}{TP + TN + FP + FN} \qquad\qquad AR = \frac{TP + FN}{TP + TN + FP + FN}$$

Ces deux quantités de taux de faux positifs et de faux négatifs renseignent donc plus précisément sur le comportement du modèle, et peuvent être utiles selon l'usage que l'on souhaite faire de notre algorithme de détection d'anomalies.

Ces trois métriques, que sont les quantités p, FPR et FNR, seront utilisées pour évaluer les performances des différentes versions proposées du modèle.

Il existe une métrique supplémentaire, appelée score F_1. Ce score mesure la performance prédictive d'un modèle. Il introduit un compromis entre taux de faux positifs et taux de faux négatifs qui diffère de celui utilisé par la métrique p, il est défini par :

$$F_1 = \frac{2TP}{2TP + FP + FN}$$

Après avoir étudié plusieurs méthodes permettant d'identifier des données météorologiques suspectes, il est désormais temps d'effectuer la synthèse des travaux évoqués dans cette revue de littérature.

2.6 Synthèse de la revue de la littérature

Nous avons donc passé en revue différentes méthodes utilisées pour effectuer l'identification de données météorologiques suspectes, du simple test statistique à l'apprentissage profond en passant par les méthodes spatio-temporelles. Rappelons que ces méthodes soutiennent le processus d'assurance qualité, qui est crucial et chronophage, d'où l'intérêt de l'automatiser pour réduire la charge de travail des experts. Il ressort que les méthodes les plus simplistes comme les tests statistiques classiques et les méthodes géostatistiques permettent d'éliminer des erreurs grossières, mais peinent à détecter des erreurs moins flagrantes. Ces tests sont déjà implémentés à l'opérationnel dans le cadre d'assurance qualité effectué par Solutions Mesonet. Ainsi, nous nous employons plutôt à la détection d'erreurs moins flagrantes. De ce fait, il paraît pertinent de se tourner vers des méthodes plus sophistiquées, à l'instar des méthodes spatio-temporelles et des méthodes d'apprentissage automatique. Malgré tout, les méthodes spatio-temporelles nécessitent de disposer de l'ensemble des séries chronologiques, tandis que l'utilisation de réseaux de neurones convolutifs permet l'intégration de diverses sources de données. Ainsi, ces réseaux de neurones présentent un intérêt certain puisqu'ils sont très adaptés à la classification d'images et sont capables de détecter automatiquement des motifs complexes et cruciaux pour la tâche imposée. Leur fonctionnement a été présenté en détail dans la section 2.5 car l'étude de Sha *et al.* (2021), qui repose sur cette notion, est prometteuse et a constitué la base de ce travail. Par ailleurs, au meilleur de nos connaissances, aucune recherche portant sur l'identification de données de précipitations journalières suspectes par réseaux de neurones convolutifs au Québec n'a encore été menée. En particulier, aucune méthode ne constitue un jeu de données basé sur l'intégration de différentes sources d'informations météorologiques journalières au Québec. De plus, aucun étude de ce genre n'a proposé d'automatiser la détection des valeurs suspectes pour l'assurance qualité des précipitations journalières au Québec.Par conséquent, la revue de littérature nous encourage à adapter la méthode de Sha *et al.* (2021) initialement réalisée en Colombie-Britannique, pour nos données du Québec.

Après avoir rappelé le cadre théorique des réseaux de neurones convolutifs, qui concerne leurs fondements et leur application à l'identification de données de précipitations suspectes, il est temps de passer à la méthode proposée proprement dite. En effet, comment identifier les données de précipitations suspectes grâce à un modèle de réseaux de neurones ? C'est l'objet du prochain chapitre, qui décrira le cadre méthodologique général de notre outil.

CHAPITRE 3 MÉTHODOLOGIE

Tel que démontré dans la section précédente, aucune méthode ne permet d'identifier les données suspectes dans les mesures de précipitations journalières au Québec.

La méthodologie proposée dans ce projet a pour but de proposer une solution aux enjeux de qualité des données dans le contexte météorologique, où les mesures peuvent être erronées et où le volume de données à traiter est conséquent, et donc impossible à traiter manuellement uniquement (voir section 2.6). En particulier, les données journalières de précipitations au Québec sont étudiées dans ce projet.

3.1 Problématique

Nous tentons donc à travers ce travail de répondre à la problématique suivante : comment améliorer l'identification des données suspectes dans les mesures de précipitations journalières au Québec ?

Pour répondre à cette problématique, nous proposons d'atteindre l'objectif et les sous-objectifs présentés à la section 3.2.

3.2 Objectif et sous-objectifs

L'objectif du projet de recherche consiste à améliorer l'identification des données suspectes de précipitations journalières enregistrées sur l'ensemble du territoire québécois. Les sous-objectifs sont les suivants :

- Constituer un jeu de données basé sur l'intégration de différentes sources d'informations météorologiques journalières au Québec.
- Automatiser la détection des valeurs suspectes pour l'assurance qualité des précipitations journalières au Québec.

En effet, la revue de littérature effectuée a permis de mettre en évidence l'absence d'études permettant d'intégrer automatiquement les différentes sources de données nécessaires à l'identification de données de précipitations suspectes au Québec. En effet, une étude de Sha *et al.* (2021) a traité d'une problématique semblable sur des données différentes en Colombie-Britannique, mais au meilleur de nos connaissances cette intégration n'a jamais été réalisée au Québec. De plus, l'usage de l'apprentissage profond permettrait d'ingérer un grand volume de données et d'en tirer profit grâce à la puissance des algorithmes modernes pour obtenir des routines plus performantes que les approches classiques en identification des données

suspectes. Actuellement, le processus d'identification de données météorologiques suspectes au Québec est chronophage et coûteux en ressources humaines, le deuxième sous-objectif que nous proposons contribuerait à l'atténuation de ces coûts.

La méthodologie proposée est itérative et sera appliquée dans un premier temps à des données simulées afin de contrôler l'amplitude des erreurs dans les données ainsi que de déterminer la meilleure version du modèle de classification. Dans un second temps, une étape de validation sur des données réelles sera réalisée, afin de déterminer les performances du modèle de classification dans une situation similaire au contexte opérationnel.

Dans cette section, nous allons présenter la méthodologie générale de ce projet de recherche. La figure 3.1 présente les étapes principales de celle-ci et seront détaillés au fil de la section.

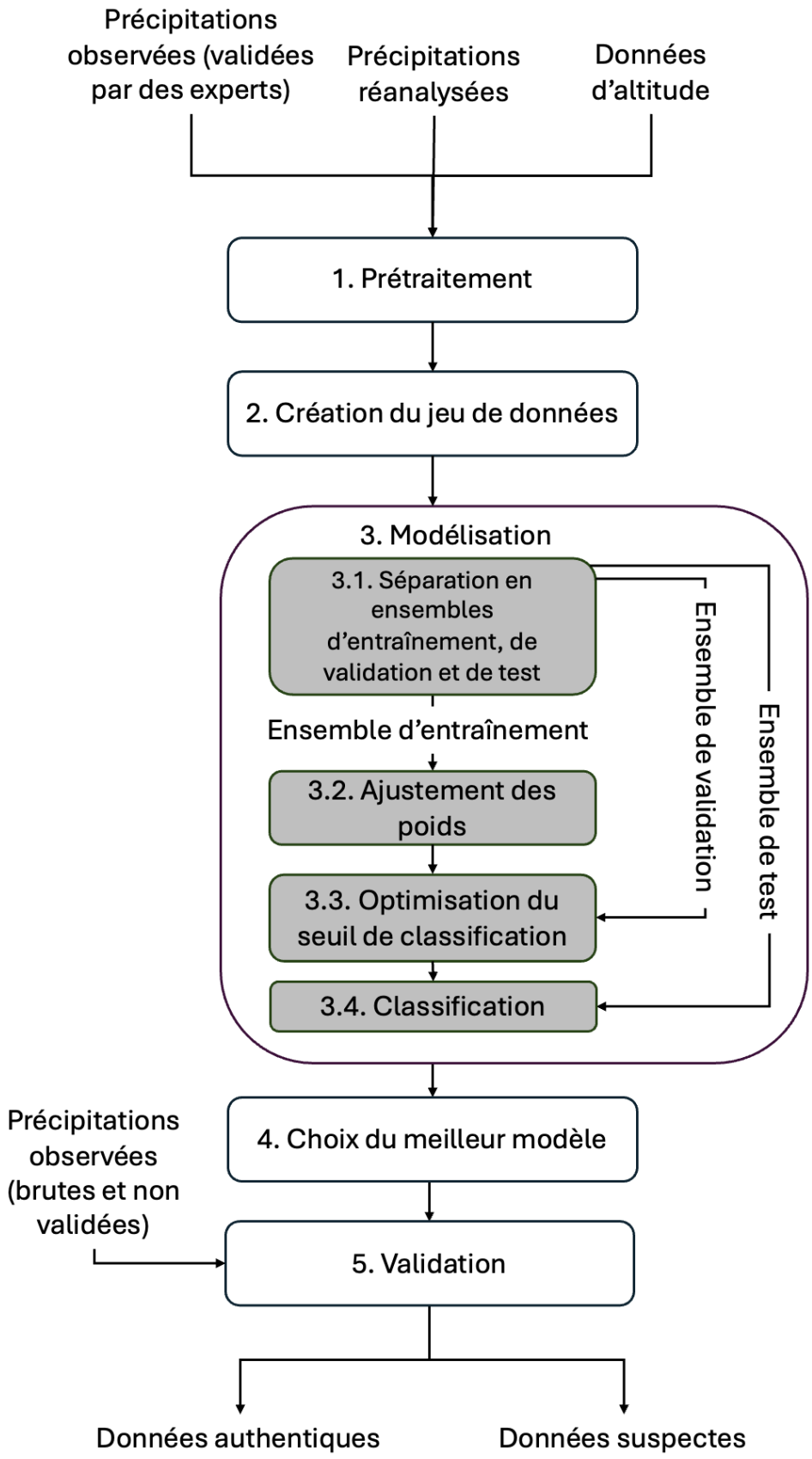


FIGURE 3.1 Schéma général de la méthodologie utilisée.

3.3 Données nécessaires

La méthodologie proposée nécessite quatre types de données, soit les données sur les précipitations observées validées par des experts, les précipitations observées non validées, les données sur les précipitations réanalysées et les données d'altitude. Chaque type de donnée sera détaillé dans les sous-sections suivantes.

3.3.1 Précipitations observées (validées par des experts)

Dans notre étude, les mesures de précipitations d'intérêt sont celles relevant du pas de temps journalier. Des accumulations sur 24h sont considérées, accumulations qui proviennent de capteurs hébergés par des stations météorologiques. Ces capteurs doivent être soumis à certaines normes de gestion et de maintenance, de façon à garantir un minimum de fiabilité dans les mesures. De plus, les données doivent avoir été validées par des experts en assurance qualité de façon à s'assurer que toutes les données considérées sont authentiques et supposées de bonne qualité. Pour faire fonctionner l'outil, des données de précipitations journalières reliées à leur date d'occurrence sont nécessaires, et aux coordonnées GPS (longitude, latitude) de la station météorologique d'où elles proviennent. Ces données complémentaires seront utiles pour faire le lien avec le type de données suivant, qui est décrit dans le prochain paragraphe.

Par ailleurs, il est d'ores et déjà utile de préciser que ces données seront utilisées pour entraîner un modèle d'apprentissage profond, ce qui signifie que le volume de données traité serait dans l'idéal relativement élevé en raison du grand nombre de paramètres à apprendre dans cette modélisation.

3.3.2 Précipitations observées (brutes et non validées)

Des données de précipitations non validées par des experts sont également nécessaires dans cette méthode, toujours pour la résolution temporelle journalière. Ces données contiennent donc des données authentiques, mais également des données suspectes. Ces observations doivent être étiquetées : la classe de chaque observation (donnée authentique ou suspecte) doit être définie dans le jeu de données.

3.3.3 Précipitations réanalysées

En plus des précipitations observées pour lesquelles l'objectif consiste à identifier les données suspectes, la méthode requiert des données de réanalyse de précipitations. En particulier, le type de données de réanalyse requises est un produit de reprévision dans le passé qui n'utilise

pas l'assimilation des données. Ces produits sont généralement des données sur grille : les réanalyses sont faites pour une matrice de cellules, chacune d'entre elles représentant une certaine zone géographique. Ces réanalyses sont parfois issues de modèles numériques météorologiques pour lesquels la résolution des équations de la mécanique des fluides est discrétisée spatialement et temporellement. Ces réanalyses reposent sur des conditions initiales concernant l'état de l'atmosphère. Dépendamment de la résolution spatiale, la zone couverte par un point de grille peut être plus ou moins étendue ; cela varie en fonction du modèle de prévision utilisé (modèle global, modèle régional). Les métadonnées à obtenir sont les mêmes que pour les précipitations mesurées, à savoir la date d'occurrence et les coordonnées GPS du point de grille.

3.3.4 Données d'altitude

Des données d'élévation provenant d'un modèle numérique sont également nécessaires dans les versions provisoires de notre méthodologie. À l'instar des données de réanalyse, ce sont des données sur grille, et elles représentent l'altitude moyenne de celles-ci.

Une fois ces données obtenues, il est temps de passer à l'étape qui permet d'opérer un prétraitement sur ces données, puis d'intégrer et de fusionner ces éléments afin de créer le jeu de données pertinent pour réaliser la détection de valeurs suspectes. C'est l'objet de la section suivante.

3.4 Prétraitement

L'étape 1 de notre méthodologie présentée à la figure 3.1 est le prétraitement des données. Elle comporte trois sous-étapes : extraction des précipitations positives (section 3.4.1), mise à l'échelle des précipitations observées (section 3.4.2) et génération de valeurs suspectes artificielles (section 3.4.3). Cette étape est nécessaire pour s'assurer d'avoir des données valides et exploitables. Cette étape dépend des données recueillies et du format de celles-ci. Les différents traitements qui peuvent être appliqués aux données sont détaillés dans cette section. Ces traitements sont appliqués pour deux modes d'entraînement : toutes les stations sont prises en compte conjointement de façon globale (scénario A), ou bien chacune d'entre elles est considérée de façon individuelle (scénario B).

3.4.1 Extraction des précipitations positives

Cette étape a pour but de supprimer les données négatives ou les données nulles du jeu de données à étudier. En effet, dans le contexte des précipitations, il est rare que des données

nulles soient erronées. Cette étape permet donc de filtrer les données pour ne conserver que les données de précipitation strictement positives.

3.4.2 Mise à l'échelle des précipitations observées

Pour pallier les difficultés liées à la différence d'échelle des observations et des données de réanalyse, les précipitations doivent être mise à l'échelle des précipitations réanalysées. Pour ce faire, les précipitations réanalysées correspondant aux cellules contenant chaque station doivent être extraites. Ensuite, les données doivent être divisées en deux sous-ensembles : précipitations strictement inférieures à 5 mm et précipitations supérieures à 5 mm. Ceci permet de différencier la caractérisation des faibles précipitations de celle des fortes valeurs d'accumulation. Une régression linéaire simple entre les mesures de précipitations et la valeur de la cellule qui contient la station sur chacun des sous-ensembles doit être réalisée. Enfin, au moyen de la droite de régression, les précipitations observées doivent être projetées pour les faire correspondre à l'échelle des précipitations réanalysées. Selon le mode d'entraînement, cette opération doit être répétée de façon globale (scénario A) ou station par station (scénario B).

3.4.3 Génération de valeurs suspectes artificielles

La régression linéaire de la section 3.4.2 est utilisée pour générer des observations erronées. Tout d'abord, un résidu studentisé extrême est défini comme ayant une valeur supérieure à 3. Pour générer les valeurs suspectes, une portion de chaque sous-ensemble du jeu de données (précipitation inférieure ou supérieure à 5 mm) est sélectionnée au hasard, puis deux vecteurs aléatoires dont les marginales sont indépendantes et identiquement distribuées (i.i.d.), uniformément distribuées sur l'intervalle $[3, 4]$ sont échantillonnés, de dimension égale au nombre de valeurs suspectes désirées dans chaque sous-ensemble. Cette portion devrait correspondre à la part de données suspectes présentes dans les jeux de données réels du partenaire industriel. Ensuite, chacun de ces vecteurs est associé à un des sous-ensembles. Enfin, ces résidus studentisés extrêmes sont implantés dans les sous-ensembles correspondants pour simuler des observations suspectes (voir l'annexe B pour plus de détails de calcul). Selon le mode d'entraînement, l'opération est répétée de façon globale (scénario A) ou station par station (scénario B).

En outre, ces opérations sont répétées pour des vecteurs aléatoires dont les marginales sont indépendantes et identiquement distribuées (i.i.d.), uniformément distribuées sur les intervalles suivants :

$$[2.5, 3.5] \quad ; \quad [2, 3] \quad ; \quad [1.5, 2.5] \quad ; \quad [1.25, 2.25]$$

À présent, les données provenant des mesures issues de capteurs et mises à l'échelle seront intégrées dans un jeu de données surfacique, présenté sous forme de matrice de cellules dont chacune représente une zone géographique. La façon dont cette intégration sera réalisée est l'objet de la prochaine section.

3.5 Création du jeu de données

Après l'étape du prétraitement vient celle de la création du jeu de données en tant que tel, qui permettra ensuite d'alimenter le modèle de classification. Elle est représentée à l'étape 2 de la figure 3.1, et comporte deux sous-étapes : fusion des données d'observation et du produit sur grille (section 3.5.1) et deuxième canal avec les stations voisines (section 3.5.2).

3.5.1 Fusion des données d'observation et du produit sur grille

Cette sous-étape se décompose elle-même en deux parties : découpage du voisinage de chaque station (section 3.5.1) puis remplacement de la cellule centrale (section 3.5.1). Afin de fusionner les données mises à l'échelle provenant de capteurs et les précipitations réanalysées, le voisinage des précipitations réanalysées de chaque station sera découpé puis la cellule centrale sera remplacée. Ces étapes sont détaillées dans les prochaines sections. Dans la suite, la station météorologique dont les données suspectes doivent être identifiées est dénotée par station d'intérêt. L'instant d'intérêt est défini par le moment d'occurrence des précipitations soumises au processus d'identification de valeurs suspectes.

Découpage du voisinage de chaque station

Considérons un instant d'intérêt quelconque. À partir des coordonnées GPS (longitude, latitude) d'une station météorologique, le point de grille du produit de réanalyse sur lequel se situe ladite station peut être déterminé. Une fois ce point de grille déterminé, il est aisé d'accéder à son voisinage. Ainsi, un voisinage carré de ce point de grille est sauvegardé, de dimension 16×16 cellules, pour cet instant d'intérêt. Plus précisément, la cellule sur laquelle se situe la station se trouve à être celle en position $(8, 8)$ sur la grille de cellules.

Ce processus est répété avec toutes les stations d'observation qui sont à l'origine des données analysées, et ce pour tous les instants d'intérêt de notre étude. Cette étape est effectuée pour les scénarios A et B.

En outre, plusieurs autres dimensions de grilles sont testées : 64×64, 32×32, et 8×8.

Remplacement de la cellule centrale

Disposant du voisinage de chaque station en termes de réanalyse de précipitation, il est possible d'opérer la fusion des mesures d'observation mises à l'échelle dans les produits sur grille. La méthode prévoit de procéder comme illustré sur la figure 3.2 ci-contre.

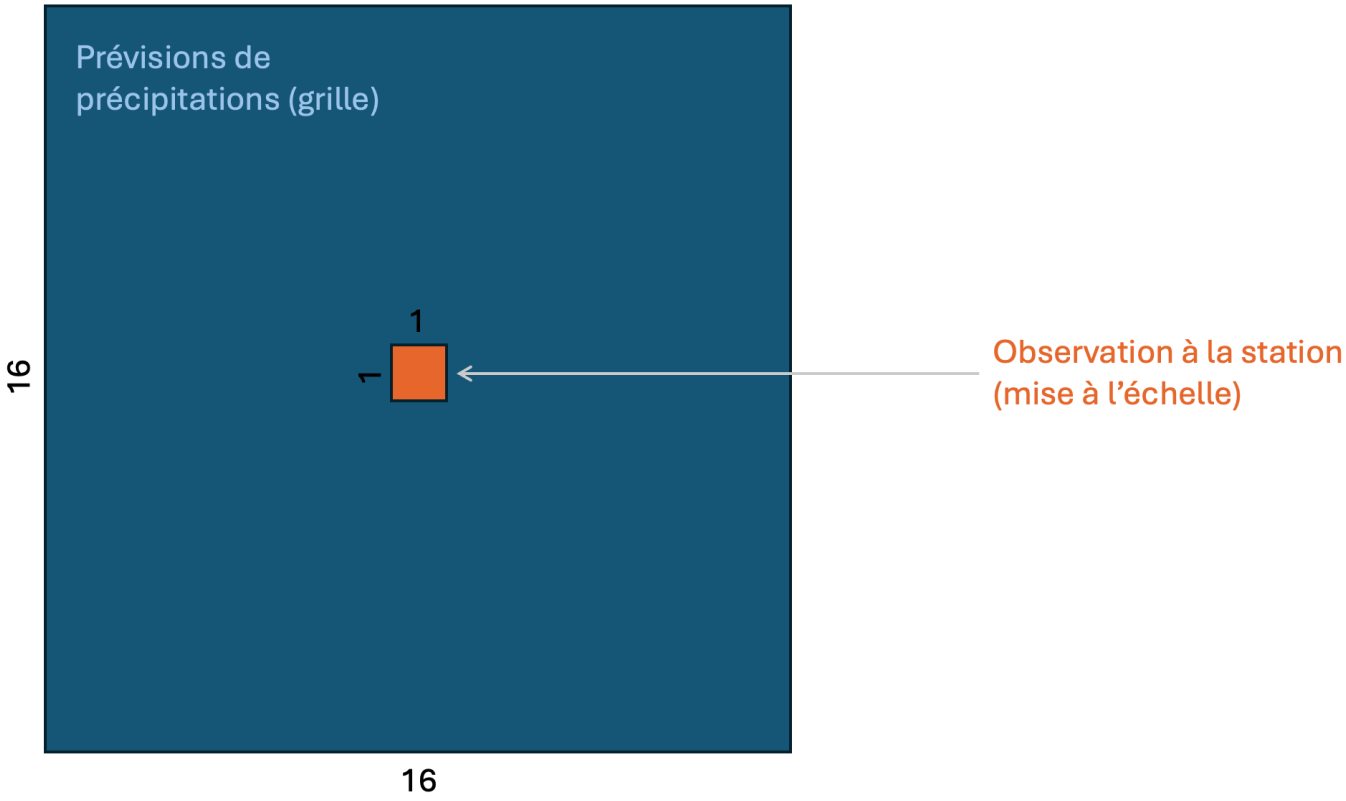


FIGURE 3.2 Incorporation des observations dans la réanalyse sur grille.

Ainsi, comme illustré ci-dessus, la cellule centrale est remplacée par la valeur de l'observation (correctement mise à l'échelle) à la station météorologique. Cette étape est effectuée pour les scénarios A et B.

De plus, le modèle est également testé avec 4 cellules centrales remplacées par la valeur de l'observation mise à l'échelle (carré de dimension 2×2). L'intérêt de tester ces différentes versions est de permettre de déterminer la version qui démontre les meilleures performances pour l'identification des mesures de précipitation.

3.5.2 Deuxième canal

Parallèlement aux images déjà créées, qui intègrent des mesures de précipitations mises à l'échelle dans des grilles de données de réanalyse, un deuxième canal doit être créé.

Pour le scénario A, soit la version où le mode d'entraînement du modèle est global (toutes les stations considérées conjointement), le deuxième canal est le voisinage en termes d'altitude autour de la station, avec le même nombre de cellules que le premier canal. Ce deuxième canal permet de mieux contextualiser les précipitations réanalysées et les observations en tenant compte du relief de la zone concernée.

Pour le scénario B (chaque station est considérée individuellement), le deuxième canal sera uniquement basé sur les précipitations observées projetées dans l'espace du produit de réanalyse : il n'utilisera ni le produit de réanalyse ni les données d'altitude. Pour créer les images de ce canal, le voisinage de chaque station pour tous les instants d'intérêt est considéré, comme dans la section précédente. Une grille de dimension 16×16 cellules est donc découpée, centrée sur la station et avec la résolution spatiale du produit de réanalyse. Seulement, toutes les cellules sont mises à 0, sauf celles où se situent des stations actives à cet instant d'intérêt. Ainsi, les stations qui se situent dans ce voisinage de la station d'intérêt voient leurs mesures incluses dans ce canal. En particulier, la mesure de la station d'intérêt sera intégrée sur la cellule en position $(8, 8)$. Si plusieurs stations se trouvent sur le même point de grille, la valeur de cette cellule est définie par la moyenne des mesures de ces stations. Ce deuxième canal permet d'inclure une notion de dépendance spatiale entre les observations aux stations météorologiques.

Ce processus est répété pour toutes les autres versions de jeux de données, en l'occurrence d'autres dimensions de grille (voir section 3.5.1) et de nombres différents de cellules centrales remplacées (voir section 3.5.1).

Ainsi, après avoir créé ce deuxième canal, pour chaque station d'intérêt, chaque instant d'intérêt et chaque scénario, deux images sont disponibles pour une observation d'une version du jeu de données.

Pour le scénario A, l'image du premier canal intègre une mesure de précipitation projetée dans une grille de données de réanalyse, et l'image du deuxième canal incorpore les données d'altitude entourant la station d'intérêt.

Pour le scénario B, l'image du premier canal intègre une mesure de précipitation projetée dans une grille de réanalyse, et l'image du deuxième canal incorpore les mesures de précipitation des stations voisines.

Chaque paire d'image est associée à une étiquette de qualité : authentique si aucune erreur

artificielle n'a été insérée, et suspecte si au contraire une erreur a été insérée dans la mesure (voir section 3.4.3). À ce stade, cette paire d'image est schématisée par la figure 3.3 ci-dessous pour le scénario B.

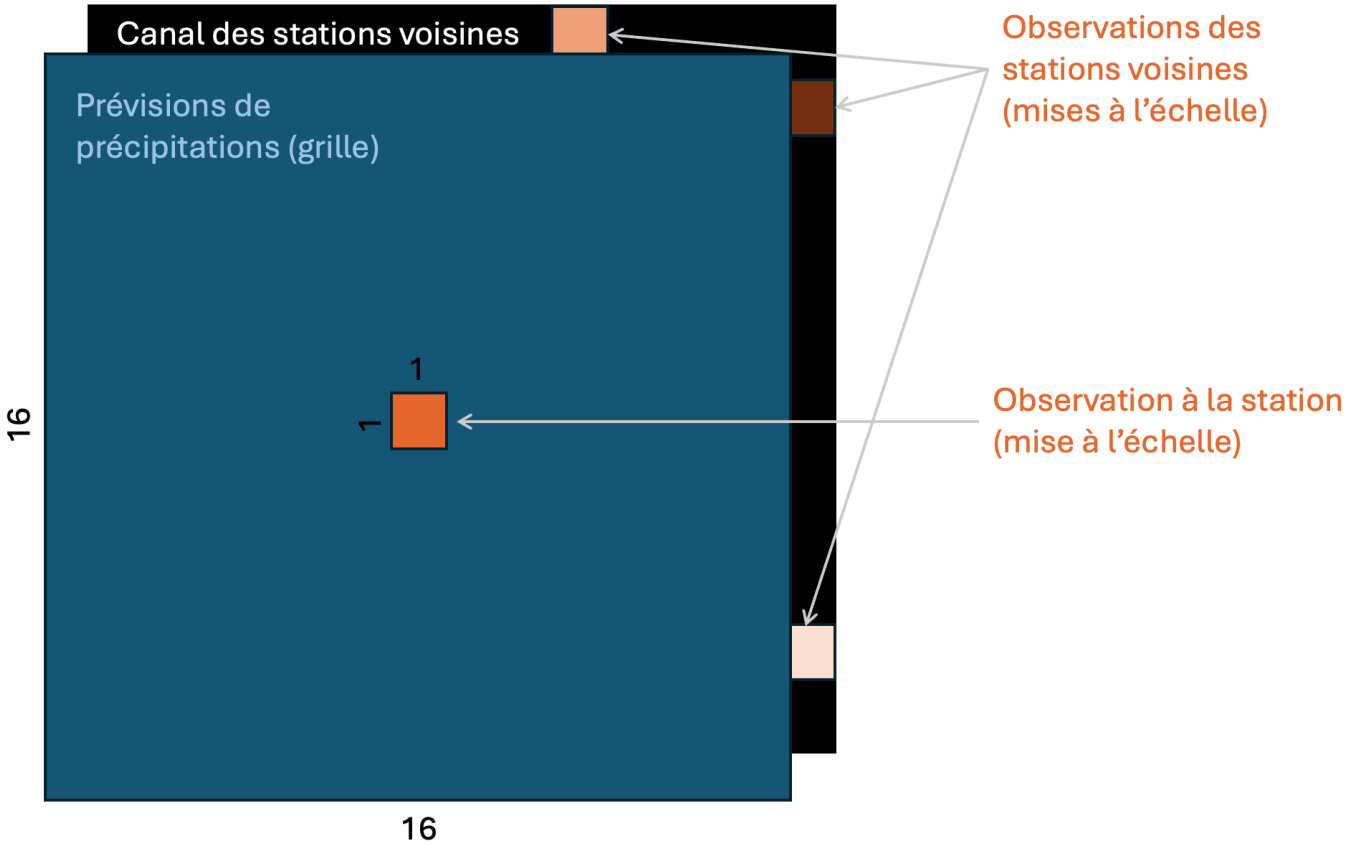


FIGURE 3.3 Schéma de la paire d'image obtenue grâce aux deux canaux (scénario B).

Un tel schéma serait similaire pour le scénario A, à la seule différence que le canal des stations voisines serait remplacé par le canal des données d'altitude.

Dans le tableau 3.1 ci-dessous se trouve un récapitulatif des scénarios avec leurs canaux respectifs et leur utilité. Cela permet de clarifier les spécificités de chacun et d'éviter les confusions.

Tableau 3.1 Récapitulatif des scénarios

	Scénario A	Scénario B
Mode d'entraînement	Global	Station par station
Premier canal	Fusion de l'observation dans la grille RDRS	Fusion de l'observation dans la grille RDRS
Deuxième canal	Données d'altitude	Valeurs d'observation aux stations voisines
Utilité du deuxième canal	Prise en compte du relief dans les motifs de précipitation	Ajout d'une dépendance spatiale entre stations

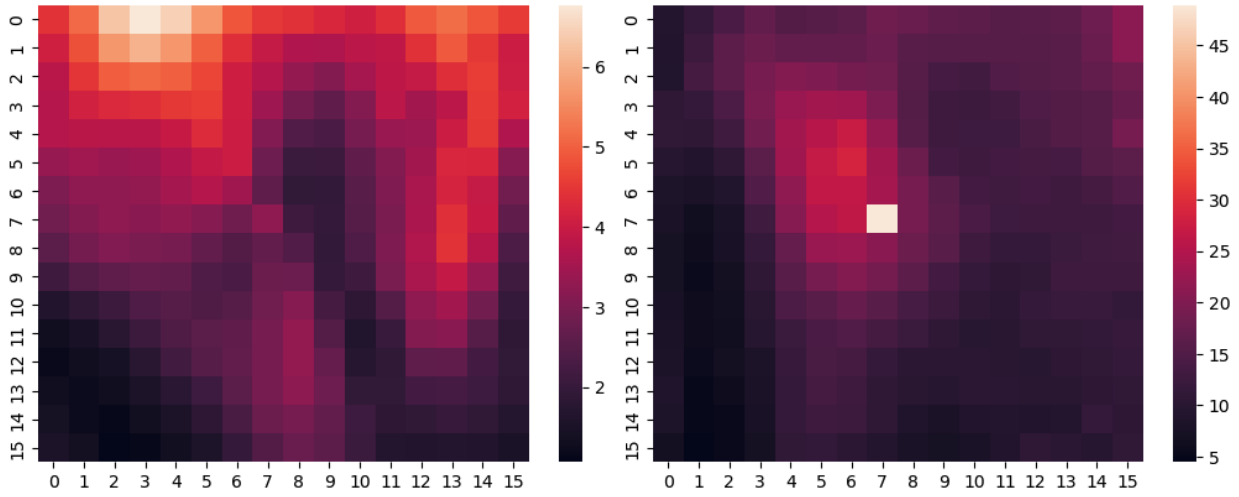
Une fois le jeu de données construit, il est désormais temps de passer à la modélisation, qui va servir à effectuer le contrôle de la qualité proprement dit. Cette modélisation repose sur l'apprentissage profond et réalise une classification binaire dans le cadre de l'apprentissage supervisé (voir section 2.5.1). La section suivante est dédiée à cette modélisation.

3.6 Modélisation

Cette étape de modélisation constitue l'étape 3 de la méthodologie et comporte 4 sous-étapes : la séparation en ensembles d'entraînement, de validation et de test (voir section 3.6.1), l'ajustement des poids (voir section 3.6.2), l'optimisation du seuil de classification (voir section 3.6.3) et la classification (voir section 3.6.4).

Illustration et principe intuitif

Notre méthode repose sur une analyse d'images qui compare une observation provenant d'une station météorologique mise à l'échelle avec son environnement provenant d'un produit sur grille d'analyse de précipitations. Pour comprendre l'intuition derrière ce procédé, une illustration simplifiée est exposée dans la figure 3.4 ci-contre. Celle-ci représente le premier canal pour les scénarios A et B : il s'agit donc de l'analyse de précipitations à laquelle les observations aux stations mises à l'échelle ont été incorporées.



(a) (b)

FIGURE 3.4 Exemple d'image du premier canal pour (a) une observation cohérente et (b) un observation suspecte.

Dans cet exemple, la figure de gauche est homogène : l'observation à la station mise à l'échelle s'insère de façon cohérente dans son environnement donné par l'analyse de précipitation. À l'inverse, la figure de droite est plus suspecte, puisque la cellule centrale présente une dissimilarité marquée comparativement à son voisinage, il y a donc fort à parier que l'observation à la station est une anomalie. Cette analyse d'image peut être réalisée automatiquement par les réseaux de neurones convolutifs : c'est l'objet de la prochaine section.

Réseau de neurones convolutif pour l'identification de données suspectes

Comme le précise la section 2.4, les réseaux de neurones convolutifs sont particulièrement adaptés aux structures en 2D, comme les images. En effet, grâce à la succession de filtres convolutifs, ces modèles sont capables d'apprendre des motifs signifiants et des caractéristiques des images qui leur permettent d'effectuer la tâche qui leur est demandée. Dans notre cas, il s'agit de réaliser une classification binaire en associant une étiquette de qualité à chaque observation : valeur suspecte ou donnée authentique. Ainsi, il convient de mettre en place un réseau de neurones convolutif qui réalise une classification binaire.

Chaque élément du jeu de données étant composé de deux images, deux canaux séparés seront nécessaires pour notre réseau de neurones convolutif. À l'instar des images en couleur qui sont encodées sur 3 canaux comme expliqué dans le paragraphe sur les filtres convolutifs de la section 2.5.2, notre méthode prévoit l'utilisation de multiples canaux. Dans les deux

scénarios, le premier canal correspond à la grille de réanalyse de précipitations avec la cellule centrale remplacée.

Pour le scénario A, le deuxième canal est celui des données d'altitude. De plus, un seul réseau de neurones intégrant les données de toutes les stations de façon conjointe sera entraîné. Un seul modèle global est suffisant dans le scénario A puisque des données d'altitude ont été incorporées, ce qui permet au modèle de tenir compte de l'impact du relief dans l'apparition de données de précipitation suspectes.

Pour le scénario B, le deuxième canal représente les valeurs des observations des stations voisines. Étant donné que certaines stations présentent plus de valeurs suspectes que d'autres, il est pertinent d'entraîner un modèle distinct pour chaque station : cela fait partie de nos contributions. Ainsi, il convient de **séparer le jeu de données en sous-parties dont chacune correspond à une station**. Chaque sous-partie servira à entraîner un **réseau de neurones différent**.

Architecture implémentée du réseau de neurones convolutif

Après avoir expliqué l'intérêt de l'utilisation des réseaux de neurones convolutifs (voir chapitre 2), il est temps de mettre en place celui qui servira à notre projet. Un schéma récapitulatif de son architecture est présenté à la figure 3.5, et reprend les différents éléments introduits dans la section 2.5.

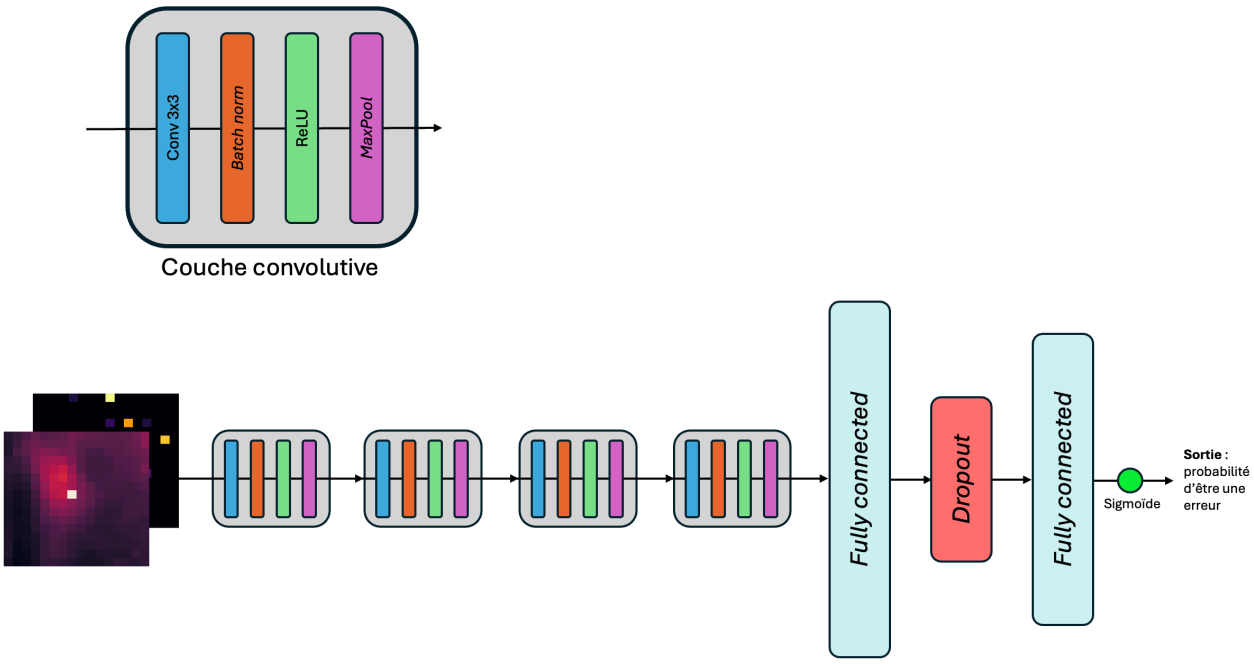


FIGURE 3.5 Architecture du modèle.

3.6.1 Séparation en ensembles d'entraînement, de validation et de test

De façon assez classique en apprentissage profond, chaque sous-ensemble est séparé en trois parties disjointes : un ensemble d'entraînement, un ensemble de validation et un ensemble de test. Les fonctions respectives de ces ensembles sont bien distinctes, et sont détaillées dans les paragraphes suivants.

3.6.2 Ajustement des poids

Durant la phase d'entraînement, l'ensemble d'entraînement est utilisé pour ajuster les poids des neurones du modèle de façon à minimiser la fonction de perte. La section 2.5.6 de la revue de littérature propose plus de détails quant à la minimisation de la fonction de perte.

3.6.3 Optimisation des hyperparamètres

L'ensemble de validation est utilisé pour ajuster les hyperparamètres que sont le seuil de classification et le nombre d'itérations réalisées pour l'entraînement du modèle en évaluant la performance du modèle au cours de l'entraînement.

Tout d'abord, cet ensemble de validation sert à empêcher le surajustement via le processus d'*early stopping* (voir section 2.5.7). Cela contrôle le nombre d'*epochs* réalisées durant la phase d'entraînement. Dans notre méthode, K est fixé égal à 2, afin d'éviter le surajustement tout en permettant au modèle de faire augmenter sa fonction de perte pendant une *epoch*, pour peu qu'il parvienne à l'améliorer à la suivante. En effet, une fonction de perte qui croît peut arriver occasionnellement, et n'est pas nécessairement synonyme de surajustement, c'est pourquoi l'entraînement n'est pas interrompu à la première hausse de la fonction de perte sur l'ensemble de validation.

L'ensemble de validation sert également à ajuster le seuil de classification, jusqu'ici fixé à 0.5 pour des raisons pratiques liées à l'implémentation : un paramètre est ajouté à notre règle de décision définie par l'équation (2.1), de la façon suivante

$$\tilde{d} = d_{\mathbb{1}\{x > s_{opt}\}}$$

où s_{opt} est le seuil qui maximise la proportion d'observations bien classées sur l'ensemble de validation.

Cette optimisation du seuil a été menée via un algorithme de *grid search*, c'est-à-dire que le paramètre optimal s_{opt} est trouvé en faisant varier un paramètre s dans un ensemble fini de valeurs et en retenant celle qui maximise la proportion de bonnes classifications sur l'ensemble

de validation. Ce seuil peut donc être ajusté globalement (scénario A), ou bien station par station, gracieuseté de l'entraînement d'un modèle distinct pour chaque station individuelle (scénario B).

3.6.4 Classification

Le modèle entraîné qui a donné les meilleures performances sur l'ensemble de validation, et qui a donc les meilleurs hyperparamètres est donc utilisé pour prédire la qualité des données de l'ensemble de test. Le modèle effectue donc la classification binaire de ces données, en les séparant en deux classes : données authentiques et valeurs suspectes.

Cette classification est effectuée pour chaque dimension de grille proposée, et pour chaque nombre de cellules centrales remplacées (voir section 3.5.1).

3.7 Choix du meilleur modèle

Basé sur les performances de classification des différentes versions de modèle obtenues à la section 3.6.4, le meilleur modèle est choisi et utilisé pour être testé sur des données réelles, et non des données simulées.

À présent, il est temps d'évaluer la performance du modèle choisi sur un jeu de données réel. La prochaine section est justement dédiée à cette fin.

3.8 Validation

La validation constitue l'étape 5 de la méthodologie. Pour réaliser l'évaluation des performances de l'outil proposé, il est nécessaire de se baser sur des métriques objectives, adaptées au contexte d'étude. Dans notre cas, il s'agit d'une classification binaire, et il est donc pertinent de tenir compte du nombre de vrais positifs, de vrais négatifs, de faux positifs et de faux négatifs. Rappelons que la classe négative (0) correspond aux données authentiques, de bonne qualité, tandis que la classe positive (1) correspond aux valeurs suspectes. Pour réaliser la validation, les étapes 1 à 3 de la méthodologie sont réitérées, avec une variante. En effet, ces étapes sont initialement réalisées avec des précipitations observées validées par des experts, mais cette fois-ci elles seront effectuées avec un autre type d'observations : des mesures de précipitation non validées. Ensuite, le modèle choisi à la section 3.7 effectue la classification binaire. Enfin, les outils définis à la section 2.5.9 sont employés afin d'évaluer les performances du modèle sur des données réelles. Ce test va permettre d'évaluer les performances de la méthode dans un contexte qui se rapproche de l'opérationnel.

Afin d'adapter le modèle au jeu de données réel, qui est plus complexe à classer que les jeux de données simulés, comportant des erreurs artificielles, l'architecture et les seuils de classification du modèle sont légèrement modifiées de telle sorte à ne pas excessivement augmenter le taux de faux négatifs. En particulier, 5 couches de convolution sont à présent employées. En effet, les 3 premières couches de convolution ont des filtres de dimension 3×3, tandis que ceux des 2 dernières sont de dimension 2×2. De plus, basé sur le score F_1 introduit dans la section 2.5.9, le seuil de classification est fixé pour chaque station comme étant celui qui maximise le score F_1. En particulier, si un nouveau seuil proposé fait en sorte que pour une station le score F_1 sur l'ensemble de validation est strictement meilleur que le seuil précédent, et que de plus il fait baisser le nombre de faux négatifs de 5 %, alors celui-ci est conservé. Si le nouveau seuil proposé mène à un score F_1 égal à celui du seuil précédent, alors il sera choisi comme nouveau seuil si la proportion de bonnes classifications est strictement supérieure à celle du seuil précédent et si le nombre de faux négatifs baisse. Ces modifications sont motivées par l'envie de conserver un taux de faux négatifs cohérent, de façon à ce que le modèle soit capable de bien détecter les données suspectes.

Après avoir posé le cadre de notre méthodologie, il est à présent temps de l'appliquer à notre cas d'étude, à savoir la détection de valeurs suspectes dans les jeux de données de précipitations journalières au Québec.

CHAPITRE 4 CAS D'ÉTUDE

Notre cas d'étude se concentre sur les mesures de précipitations journalières au Québec. Il vise à identifier les données suspectes dans ces données, dont on va justement détailler les caractéristiques dans la suite.

Le territoire étudié s'étend sur le territoire québécois, puisque les stations météorologiques vont du sud du Québec, là où elles sont très densément réparties, jusqu'au nord de la province, où la densité des stations est beaucoup plus faible. Les données considérées dans cette étude proviennent de 194 stations. Les capteurs considérés sont les précipitomètres, qui sont dédiés à la mesure des précipitations totales (précipitations sous forme liquide ou nivale). Un précipitomètre est affiché à la figure 4.1 à des fins d'illustration.

FIGURE 4.1 Précipitomètre.

Ce type de capteurs est parfois plus robuste que les pluviomètres à auget basculeur, qui sont plus sensibles au gel. Malgré tout, les mesures peuvent être entachées par le vent, qui peut entraîner une sous-captation de la quantité de précipitations survenues. Elles peuvent

également être erronées en cas de précipitations solides (forme nivale), qui présentent une plus forte adhérence aux parois du capteur, et peuvent donc être captées avec un temps de retard. Enfin, un défaut de maintenance ou de calibration du capteur peut entraîner l'occurrence de valeurs suspectes.

Pour minimiser le risque d'anomalies dues aux précipitations solides, on propose de se concentrer sur les précipitations estivales (de mai à octobre). Dans un premier temps, conformément à la partie 3.4.3 de la méthodologie, on génère des valeurs suspectes sur 28.3 % des mesures. Cette proportion précise provient du fait que l'article de référence de Sha *et al.* (2021) présentait cette proportion dans son jeu de données : nous l'avons donc initialement reproduite dans un but de comparaison. Seulement, dans ce projet on se concentre sur les données journalières, ce qui rend cette comparaison caduque, l'étude de référence se focalisant sur une résolution temporelle plus fine.

Après avoir présenté le contexte général du cas d'étude, passons maintenant aux données de façon plus précise, à commencer par les mesures de précipitation.

4.1 Données nécessaires

4.1.1 Précipitations observées (validées par des experts)

Dans ce projet, on se concentre sur les précipitations journalières enregistrées au Québec par les précipitomètres du réseau exploité par l'organisme Solutions Mesonet. Dans un premier temps, on utilise les données de la table *hourly_weather_qa*, qui sont le résultat du processus de contrôle qualité effectué par Solutions Mesonet. En effet, ces données passent systématiquement par une succession de tests statistiques, suite à quoi les responsables du contrôle qualité passent en revue manuellement, à l'aide de cartes, de séries temporelles et d'outils visuels avancés afin de confirmer ou d'infirmer quotidiennement la sortie des tests automatiques. Ainsi, les données de cette table sont fiables, puisque seules les données considérées de bonne qualité sont conservées dans la table *hourly_weather_qa*. Elles sont donc toutes considérées comme des données fiables et authentiques dans la méthodologie que nous adoptons. Cette table contient également d'autres informations essentielles, comme l'identifiant de la station qui a enregistré la donnée et la date de l'observation. On considère les pluies estivales (de mai à octobre) entre 2012 et 2018 inclusivement, soit 7 saisons de précipitations. Les stations météorologiques considérées dans ce projet sont réparties sur le Québec, mais ne sont pas présentes sur la partie Nord de la province. Les stations prises en compte sont affichées sur la carte de la figure 4.2 ci-dessous.

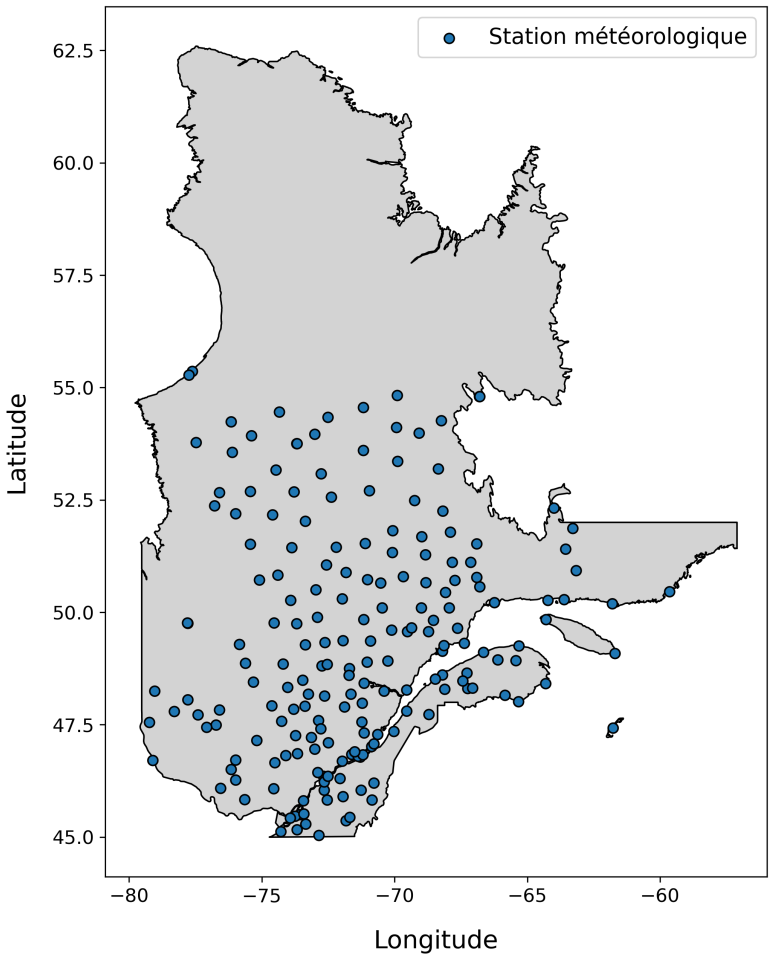


FIGURE 4.2 Stations météorologiques considérées dans l'étude.

4.1.2 Précipitations observées (brutes et non validées)

Pour ce projet, on dispose d'un accès à la table *hourly_weather_noqa*, qui contient les données non validées par l'équipe d'assurance qualité de Solutions Mesonet. En parallèle, on dispose de la table *hourly_weather_qaflags*, qui contient les étiquettes de qualité pour les données de la table *hourly_weather_noqa*. Ces étiquettes sont représentées par des entiers, chacun ayant une définition spécifique. Quelques exemples d'étiquettes possibles pour une donnée sont "bonne", "suspecte", "*warning*", ou encore "*failure*". En combinant ces deux tables et en suivant la méthodologie, un jeu de données réel qui contient de vraies erreurs de capteur peut être créé.

4.1.3 Précipitations réanalysées

Le produit considéré dans cette étude s'appelle Système Régional de Réanalyse Déterministe (RDRS) et est une reprévision dans le passé qui correspond à la sortie brute du modèle de prévision Système Régional de Prévision Déterministe (RDPS) (Fortin *et al.*, 2018), qui se base sur le modèle *Global Environment Multiscale* développé par Environnement et Changement climatique Canada (ECCC). En effet, cette dernière n'intègre pas d'assimilation de données (des précipitations observées sont incorporées à l'analyse pour «corriger» la prévision). La résolution horizontale d'une cellule est de 0.09° (soit environ 10 km). On récupère également les coordonnées spatiales du point de grille, ainsi que la date d'échéance de la prévision. Ces informations permettront de faire le lien avec les mesures provenant des stations météorologiques.

4.1.4 Données d'altitude

La méthode intègre également des données d'altitude, et nous avons opté pour le modèle numérique d'élévation *Earth TOPOgraphy 2022* (ETOPO2022) (NOAA National Centers for Environmental Information, 2022), qui possède une résolution spatiale d'une minute d'arc. Cela signifie que pour une cellule située à 50° Nord, la longueur de son côté est d'environ 1,2 km. Ces données présentent une résolution spatiale différente de celle de la prévision de précipitations, mais cela n'est pas dérangeant pour la méthode que nous utilisons. Ces informations orographiques ont une influence sur les précipitations, c'est pourquoi elles sont prises en compte dans la méthode de référence, même si dans celle-ci le jeu de données utilisé pour l'altitude est différent, puisqu'il s'agit d'ETOPO1, qui est moins récent. L'élévation n'est pas utilisée lorsque le modèle est ajusté individuellement sur chacune des stations, mais elle l'est lorsque le modèle est estimé conjointement sur toutes les stations dans un même CNN à l'instar de l'étude de Sha *et al.* (2021).

Une fois les données utiles rassemblées, il est temps de leur appliquer un prétraitement conformément à la section 3.4 de la méthodologie.

4.2 Prétraitement

Dans cette section, on présente certains détails spécifiques à notre cas d'étude, tout en se conformant à la méthodologie décrite dans le chapitre 3. Ainsi, certains éléments de la méthodologie sont repris dans cette section pour apporter des précisions sur notre jeu de données spécifiques des précipitations au Québec, mais d'autres ne le sont pas pour ne pas alourdir inutilement le document. Voyons donc maintenant les spécificités relatives à notre cas

d'étude, à commencer par la conversion des mesures de précipitations horaires en données journalières.

4.2.1 Extraction des précipitations positives

Conformément à la section 3.4.1, on extrait les valeurs strictement positives de précipitation. Contrairement à la méthode de Sha *et al.* (2021), dans laquelle la résolution temporelle considérée est la demi-heure, ce sont les précipitations journalières qui nous intéressent dans ce projet. Les précipitations journalières sont plus lisses dans le temps et l'espace car elles sont le résultat d'un cumul de 24 observations horaires, mais demeurent néanmoins une variable caractérisée par une grande variabilité naturelle. Ainsi, les précipitations horaires observées ont été agrégées afin d'obtenir des cumuls journaliers, et ce de la façon suivante : chaque période d'accumulation commence à midi de chaque jour (12z), et s'étend jusqu'à la fin de la 24^{e} heure d'accumulation. Les données sur grille proviennent de l'archive *Canadian Surface Prediction Archive* (CaSPAr), qui recense les prévisions historiques. On dispose pour chaque jour d'une prévision à midi, avec des horizons de prévision de 1 à 24h, c'est-à-dire qu'en additionnant ces 24 valeurs de précipitations on obtient la valeur de précipitation journalière, dont la fenêtre temporelle correspond avec nos données journalières d'observation aux stations. En effet, pour les données ponctuelles comme pour les données sur grille, on obtient les données journalières (de chaque jour à midi au lendemain à midi).

4.2.2 Mise à l'échelle des précipitations observées

Les précipitations observées sont des mesures ponctuelles, tandis que les prévisions de précipitations sont des données surfaciques, qui représentent une moyenne sur 100 km^2. Il y a donc une différence d'échelle entre précipitations mesurées et prévisions de précipitations. Certains produits tentent de réduire ce biais en effectuant de l'assimilation, mais nous n'utilisons pas un tel produit (voir section 3.4.2 pour plus d'explications). Conformément à la section 3.4.2, on effectue une régression linéaire simple entre les mesures de précipitations et la valeur du point de grille le plus proche donnée par le produit de prévision. À ce stade il convient d'anticiper légèrement sur la suite : le modèle de classification présenté à la section 3.6 va être testé sur le jeu de données global, mais sera aussi testé sur chaque station indépendamment (voir le dernier paragraphe de la section 3.6). En effet, initialement le modèle a été implémenté sur toutes les données d'un coup, puis l'idée d'entraîner un modèle spécifique pour chaque station a paru pertinente. Ainsi, lorsque le modèle de classification est entraîné de façon globale, la régression linéaire est effectuée sur deux sous-ensembles formant une partition : les données où l'accumulation de précipitations mesurées sur 24 heures est inférieure à 5 mm, et celles

où l'accumulation est au moins égale à 5 mm. Ceci permet de différencier la caractérisation des faibles précipitations de celle des fortes valeurs d'accumulation. Lorsque le modèle est entraîné station par station, alors la régression linéaire est également réalisée station par station, avec toujours cette même division en pluies faibles et pluies fortes au sein des données de chaque station. Détaillons à présent seulement la version station par station, puisque c'est celle qui a été conservée pour la version finale du projet.

À titre d'exemple, figure 4.3 illustre la régression linéaire effectuée entre les précipitations observées à la station de La Tuque et les prévisions de la cellule correspondante.

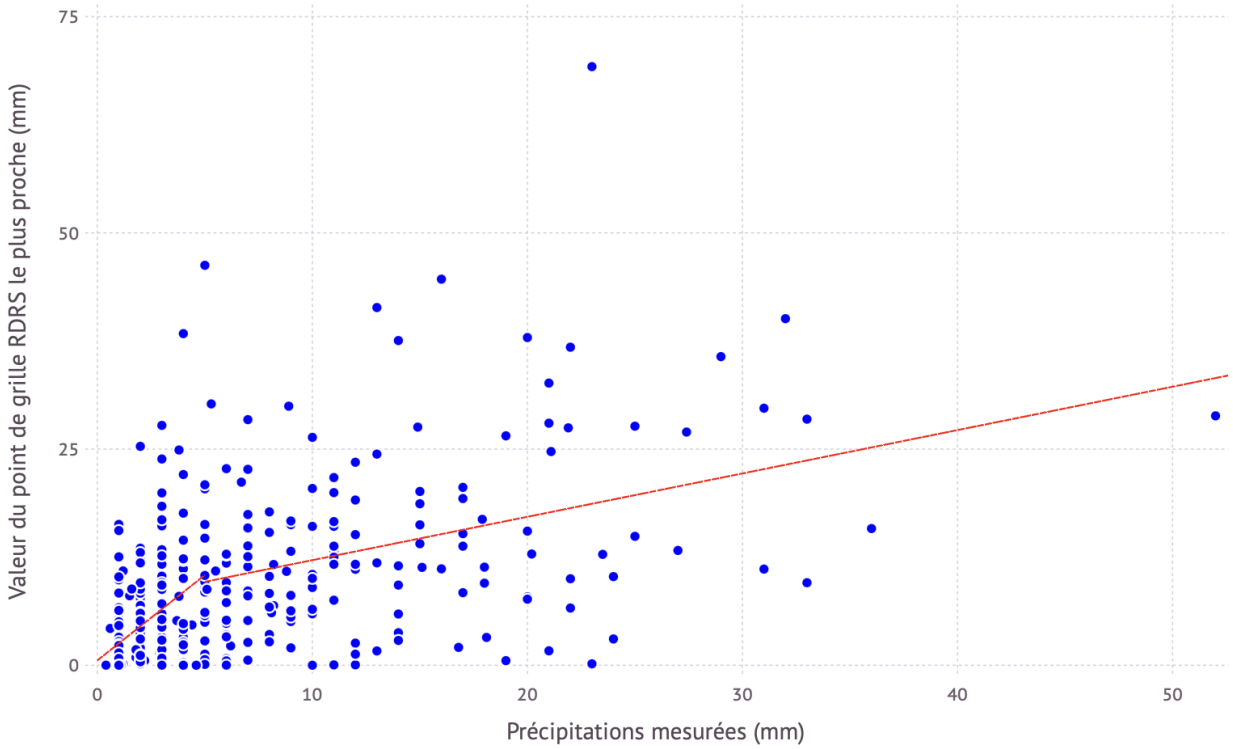


FIGURE 4.3 Régression linéaire pour la station de La Tuque.

La régression linéaire est également réalisée sur chacune des autres stations. Les valeurs suspectes artificielles sont générées de façon similaire à la section 3.4.3 de la méthodologie.

Remarque importante : Dans notre cas d'étude, nous nous sommes restreints aux stations recensant un minimum de 200 mesures journalières de précipitations, et ce afin de permettre aux modèles de classification de disposer d'un nombre suffisant d'observations pour s'ajuster correctement.

4.2.3 Génération de valeurs suspectes artificielles

Conformément à la section 3.4.3, on génère des erreurs artificielles dans notre jeu de données de précipitations. Dans notre cas, nous avons généré 28.3 % de données suspectes. Cette proportion précise provient de la part de données suspectes présente dans l'étude de Sha *et al.* (2021).

4.3 Création du jeu de données

4.3.1 Fusion des données d'observation et du produit sur grille

Découpage du voisinage de chaque station

Pour donner une idée de l'étendue géographique recouverte par le voisinage de 16×16 cellules pour le canal des précipitations, cette zone représente un peu moins de 26 000 km^2. Plus de détails sont donnés dans la sous-section 3.5.1 concernant le découpage du voisinage de la station d'intérêt.

Remplacement de la cellule centrale

Concernant le remplacement de la cellule centrale, on procède exactement comme dans la sous-section 3.5.1.

4.3.2 Deuxième canal

Dans notre cas, nous avons procédé comme expliqué dans la section 3.5.2. Concernant le jeu de données d'altitude ETOPO2022, nous l'avons utilisé pour le scénario A : un deuxième canal représentant l'altitude voisine de la station est intégré au jeu de données considérant toutes les stations conjointement. Pour le scénario B : sans le deuxième canal des stations voisines, le modèle compare les valeurs d'observation aux stations météorologiques avec leur voisinage en termes d'analyse de précipitations sur grille. Les stations ne s'influencent pas mutuellement, puisque seuls les cellules centrales de l'image sont remplacées. En d'autres termes, certaines stations d'observation se situent dans le voisinage d'une station d'intérêt, mais leurs données n'apparaissent pas dans les images relatives à cette station d'intérêt. Notre idée était donc de pouvoir prendre en compte la dépendance spatiale entre les observations de différentes stations, puisque celle-ci peut enrichir l'analyse et permettre de détecter davantage d'anomalies. En effet, une valeur de précipitations très élevée pour une station est plus suspecte si les stations voisines présentent des faibles précipitations que si ses voisines ont

aussi des valeurs élevées. Pour ce faire, on ajoute un canal supplémentaire, de même dimension que celle de l'analyse de précipitations, conformément à la section 3.5.2. Pour chaque image du jeu de données original, et donc pour chaque station et pour chaque instant d'intérêt on construit ce canal supplémentaire, que l'on vient ajouter au jeu de données. À la figure 4.4 se trouve un exemple d'une image de ce canal.

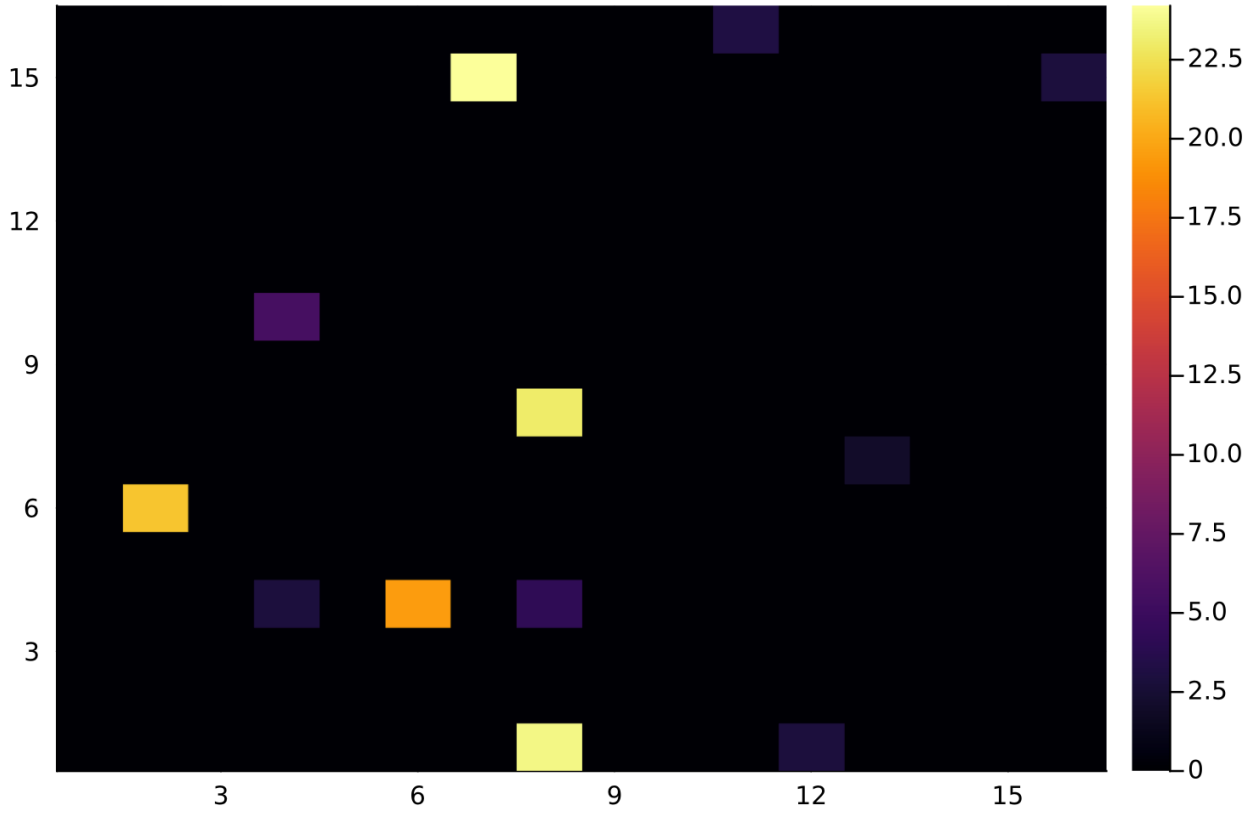


FIGURE 4.4 Inclusion des valeurs des stations voisines.

Ainsi, notre modèle est à présent en capacité de bénéficier d'informations provenant de différentes stations, en plus des informations du produit d'analyse de précipitations.

4.4 Modélisation

La modélisation utilisée est celle de la section 3.6. Dans notre cas, nous avons implémenté un réseau de neurones convolutifs pour réaliser une classification binaire des données de précipitations journalières. Pour le scénario A, un seul réseau de neurones global a été entraîné, sur un jeu de données comportant deux canaux : un qui fusionne mesures de précipitations mises à l'échelle et données de prévision RDRS, et l'autre qui représente l'altitude. Pour le scénario B, un réseau de neurones à été entraîné pour chaque station individuellement, sur des jeux

de données comportant deux canaux : un qui fusionne mesures de précipitations mises à l'échelle et données de prévision RDRS, et l'autre qui intègre les données de précipitations aux stations voisines. L'architecture utilisée pour l'entraînement sur des données simulées est celle décrite à la section 3.6.

4.4.1 Séparation en ensembles d'entraînement, de validation et de test

La séparation en sous-ensembles d'entraînement, de validation et de test est effectuée en accord avec la section 3.6.1. En particulier, dans notre cas, l'ensemble d'entraînement représente 49 % du jeu de données, l'ensemble de validation, quant à lui, en représente 21 %. Enfin, l'ensemble de test compte pour les 30 % restants.

4.4.2 Ajustement des poids

Les poids du réseau de neurones sont ajustés comme à la section 3.6.2. Dans notre cas, nous avons utilisé l'optimiseur Adam (Kingma et Ba, 2017), très répandu en apprentissage profond. Il permet aux neurones du réseau de s'ajuster de façon efficace pendant la phase d'entraînement, en suivant des directions de descente du gradient adaptatives, ce qui permet une convergence rapide et stable.

4.4.3 Optimisation des hyperparamètres

Les hyperparamètres sont ajustés comme décrit dans la section 3.6.3. Nous avons donc optimisé le nombre d'itérations réalisées pour l'entraînement du modèle afin d'éviter le surajustement, ainsi que le seuil de classification, de façon à maximiser la part de bonnes classifications sur l'ensemble de validation.

4.4.4 Classification

Les observations de précipitation ont été classées comme défini dans la section 3.6.4. En effet, une fois le réseau de neurones entraîné, la prédiction des classes des observations journalières de précipitations contenues dans l'ensemble de test peut être réalisée. Par la suite, les métriques reliées aux performances de classification peuvent être calculées.

4.5 Choix du meilleur modèle

Conformément à la section 3.7, le meilleur modèle est déterminé pour être choisi et testé sur des données réelles. Dans notre cas d'étude, le modèle choisi est la version où la grille est de

dimension 16×16, avec un mode d'entraînement station par station (scénario B) et une seule cellule centrale remplacée. Ce modèle a été choisi, puisque ses performances étaient similaires à d'autres versions de plus grande dimension, et donc plus coûteux à entraîner que la version 16×16 cellules (voir tableau 4.2). La version 8×8 n'a pas été privilégiée puisque la zone d'intérêt semblait trop restreinte pour identifier correctement les données suspectes.

4.6 Validation

On procède à la validation du modèle grâce à la méthodologie décrite à la section 3.8. Le modèle avec une grille de dimension 16×16 et un mode d'entraînement station par station est testé sur un jeu de données réel. Pour ce faire, les étiquettes de qualité sont définies par rapport à ceux utilisé à l'opérationnel par Solutions Mesonet (voir l'annexe C pour plus de détails techniques sur ces étiquettes).

Ce nouveau jeu de données contient donc des erreurs de capteurs réelles clairement identifiées au moyen de la présence de deux classes 0 et 1. Il est désormais possible d'entraîner notre modèle puis d'évaluer ses performances en vue d'avoir une bonne idée du comportement de ce dernier dans un contexte plus opérationnel.

Les changements introduits dans la section 3.8 sont réalisés : l'architecture du modèle ainsi que l'ajustement des seuils de classification sont modifiés par rapport au modèle appliqué à des données simulées. En effet, si le seuil avait été ajusté pour maximiser la proportion de bonnes classifications, alors le modèle aurait privilégié la bonne détection des données authentiques, étant donné qu'elles sont très majoritaires dans le jeu de données réel (92.47 %). En effet, le jeu de données contenant des erreurs artificielles était moins débalancé, puisque l'on s'était basé sur l'article de référence dont la méthode introduisait des erreurs dans 28.3 % des données. Ainsi, voulant conserver un taux de faux négatifs relativement bas, on se base également sur cette métrique.

Après avoir mis en place nos jeux de données et entraîné nos modèles de classification, il est temps de passer à l'exploration des résultats de chacune de ces versions.

4.7 Résultats

Après avoir explicité la méthodologie de notre étude, et posé le cadre de notre étude, on se concentre à présent sur les performances de classification de notre modèle. Cette section

est dédiée à la présentation des résultats des versions de la méthode implémentée. En particulier, on propose d'utiliser certaines métriques adaptées à notre cadre d'étude, à savoir une classification binaire (voir section 2.5.9). Dans cette section, les résultats de deux modes d'entraînement sont exposés : l'entraînement global, qui prend en compte toutes les stations conjointement (scénario A) et l'entraînement station par station, qui prend en compte chaque station individuellement (scénario B).

Les différentes versions du modèle ont été décrites dans la section 3, et sont à présent évaluées grâce aux trois critères définis à la section précédente. Pour plus de clarté, les quantités p, FPR et FNR sont exprimées en pourcentage.

4.7.1 Erreurs artificielles

Erreurs radicales

Les résultats suivants sont ceux correspondant à l'application des versions du modèle sur un jeu de données contenant des erreurs artificielles radicales (résidus studentisés dans l'intervalle $[3, 4]$).

Seuil de classification original Dans la première version du modèle, le modèle de classification est fixé à 0.5. Cela signifie que si la sortie du modèle est inférieure à 0.5, alors l'observation est classée comme étant de bonne qualité, et dans le cas contraire elle est classée comme une anomalie. Les résultats sont compilés dans le tableau 4.1 : les nombres en gras représentent les meilleurs scores associés à des modèles non triviaux (qui ne prédisent pas systématiquement une des deux classes seulement), et qui sont appliqués à toutes les stations.

TABLEAU 4.1 Métriques p, FPR et FNR pour les différentes versions des modèles pour des erreurs radicales (en %)

Mode d'entraînement	Cellules centrales	Dim. grille	p	FPR	FNR
Station La Tuque	4	64	94.59	0.00	16.67
Station par station	4	64	93.66	**1.47**	18.74
Global	4	64	96.38	3.04	5.07
		32	96.71	2.47	5.56
		16	96.93	1.54	6.84
		8	**97.10**	2.45	**4.17**
	1	64	71.86	0.00	100.00
		32	96.47	2.26	6.96
		16	95.77	1.95	10.74
		8	96.18	2.52	7.04

Seuil de classification optimisé Comme décrit dans la section 3.6.3, on a optimisé le seuil de classification du modèle, noté s_{opt}. Dans un premier temps, le seuil a été optimisé mais restait partagé entre toutes les stations. Ensuite, on a permis au seuil d'être optimisé pour chaque station indépendamment. Pour cette expérimentation, le modèle a été entraîné station par station, et les erreurs artificielles restent les mêmes qu'à la section précédente. Les résultats sont exposés dans le tableau 4.2 : le nombre en gras correspond au modèle le plus performant d'après la métrique p.

TABLEAU 4.2 Seuils optimisés et métrique p pour les différentes versions des modèles pour des erreurs radicales (en %)

Mode d'entraînement	s_{opt}	Cellules centrales	Dim. grille	p
Global	0.35	4	64	94.10
	0.5	1	8	91.83
Station par station	/	1	16	**95.24**

Remarque : Pour les modèles entraînés station par station (scénario B), il faut noter que le jeu de données utilisé comporte un seul canal : celui présentant la fusion entre mesures de précipitations mises à l'échelle et les données RDRS. Ces résultats sont donc provisoires, dans l'attente de l'ajout d'un deuxième canal incluant les stations voisines (prochain paragraphe de la section 4.7.1).

Inclusion de la valeur des stations voisines

Comme décrit à la section 4.3.2, les données des stations voisines ont été incluses dans le modèle (scénario B). Cette expérimentation a été menée sur la version du modèle comportant des grilles de dimension 16×16, pour lesquelles seul une cellule centrale a été remplacée par la valeur d'observation à la station mise à l'échelle. Le résultat de cet ajout donne la métrique suivante sur l'ensemble de test :

$$p = 96.14\ \% \qquad FPR = 1.17\ \% \qquad FNR = 10.89\ \%$$

soit 0.90 % de plus pour p par rapport au modèle précédent qui n'incorpore pas le canal supplémentaire des observations aux stations voisines.

Ajout de la notion de *padding*

Comme expliqué à la section 2.5.3, en ajoutant une couche de cellules autour des images, les résultats ont été améliorés. Ceci permet également d'ajouter deux couches de convolution, ce qui rend le modèle plus flexible. On obtient effectivement 0.68 % supplémentaire pour le taux de bonnes classifications par rapport au modèle précédent, ce qui nous amène à :

$$p = 96.82\ \% \qquad FPR = 0.92\ \% \qquad FNR = 9.07\ \%$$

Erreurs plus modérées

Le meilleur modèle des versions précédentes a également été évalué sur des jeux de données contenant des erreurs moins radicales. En effet, comme décrit dans la section 3.4.3, les résidus studentisés ont été progressivement réduits afin de générer des erreurs moins flagrantes, et testés sur la version du modèle contenant des grilles de dimension 16×16 avec une cellule centrale remplacée. Les résultats sont regroupés dans le tableau ci-dessous. Les bornes inférieures et supérieures données correspondent aux valeurs définissant l'intervalle du support de la loi uniforme utilisée pour échantillonner des valeurs de résidus studentisés (voir section 3.4.3 pour plus de détails sur la méthode de génération d'erreurs artificielles).

TABLEAU 4.3 Métriques p, FPR et FNR pour les différentes versions des modèles pour différentes amplitudes d'erreurs plus modérées (en %)

Borne inférieure	Borne supérieure	p	FPR	FNR
2.5	3.5	93.63	1.71	18.54
2	3	90.06	2.64	28.67
1.5	2.5	85.87	4.03	41.28
1.25	2.25	81.12	3.57	59.26

4.7.2 Performances du modèle sur un jeu de données réel

Conformément à la méthode décrite à la section 3.8, on se propose de tester notre modèle sur des données suspectes, non plus générées artificiellement, mais provenant d'un jeu de données réel. Avec une légère adaptation de l'architecture originale, comme décrite à cette même section 3.8, on obtient :

$$p = 74.69\ \% \qquad FPR = 27.00\ \% \qquad FNR = 4.29\ \%$$

Ainsi les résultats des différentes versions du modèle ont été exposés. Dans la section suivante, nous allons les interpréter et les discuter.

CHAPITRE 5 DISCUSSION ET CONCLUSION

Après avoir évoqué en détail la méthode proposée ainsi que les résultats qui en découlent, on peut à présent souligner les contributions de ce projet. Cette étude a permis d'améliorer l'identification de données suspectes dans les mesures de précipitations journalières au Québec, en intégrant diverses sources de données et en automatisant un processus coûteux et chronophage. Pour ce faire, des réseaux de neurones convolutifs ont été mis en place, ce qui à notre connaissance n'avait jamais été utilisé pour les précipitations au Québec. Les avantages et les limites de la solution proposée seront évoqués dans cette section. Cette partie comporte également une discussion autour des modèles implémentés ainsi que de leurs performances sur les différents jeux de données testés. Celle-ci se conclura par une synthèse des travaux réalisés et par l'évocation des possibilités d'amélioration et de travaux futurs envisageables.

5.1 Choix de la dimension de la grille et du nombre de cellules centrales remplacées

Après avoir effectué des tests concernant les dimensions définissant notre matrice de cellules, il est apparu que le modèle était relativement robuste aux hyperparamètres, que ce soit la dimension de la grille utilisée ou le nombre de cellules centrales remplacées (voir tableau 4.1). En effet, les performances sur l'ensemble de test étaient stables. Ainsi, la version prévoyant une grille de 16×16 cellules, avec une cellule centrale remplacée, a été choisie. Cette version avait l'avantage d'être bien plus légère à entraîner en termes de coût computationnel comparée à son homologue de 64×64 cellules, donc plus rapide, tout en conservant des performances comparables. La raison pour laquelle la version 8×8 cellules n'a pas été retenue est qu'il fallait s'assurer de couvrir une surface suffisante pour tenir compte de motifs significatifs permettant une bonne identification des données suspectes.

5.2 Pertinence de l'entraînement station par station

Dans notre méthode, un réseau de neurones convolutif différent a été entraîné pour chaque station (scénario B, voir section 3.6). D'une part, il est clair que les phénomènes météorologiques qui surviennent varient de façon significative d'une station à l'autre, en raison du climat local, qui ne sera pas le même selon si la station se trouve au bord du Saint-Laurent ou dans le grand nord québécois. La quantité de précipitations s'en trouvera donc impactée, et les distributions sous-jacentes seront donc difficilement comparables entre stations. D'autre

part, les stations météorologiques peuvent présenter des variations en termes de proportion de données suspectes, car certaines d'entre elles sont plus sujettes à une surexposition au vent, au gel ou à un plus fort taux de défaillance dû à des maintenances moins fréquentes, ce qui impacte la probabilité qu'une observation soit anormale ou non. Ainsi, le fait d'entraîner un modèle différent par station permet de prendre en compte cette variation de la part de données suspectes en ajustant le seuil de classification en conséquence. Enfin, traditionnellement en contrôle de la qualité, il est d'usage de prendre en compte les stations individuellement, ce qui a également influencé le choix d'un ajustement séparé des modèles pour chaque station.

5.3 Inclusion des valeurs aux stations voisines

Le fait d'inclure les observations aux stations voisines sur un canal additionnel permet d'inclure une notion de dépendance spatiale à l'intérieur du réseau de stations (scénario B). En effet, l'ajout de ce canal permet une comparaison entre les observations (mises à l'échelle, voir section 3.4.2) provenant des stations, et non uniquement une comparaison d'une observation avec son voisinage donné en termes du produit de prévision (provenant du modèle RDPS, voir section 4.1.3, dans le cas d'étude). L'ajout de ce canal est bénéfique par rapport aux performances du modèle sur le jeu de données contenant des erreurs artificielles, et l'est donc vraisemblablement aussi pour le jeu de données réel. Cela s'explique par le fait que le modèle incorpore davantage de données, *a fortiori* des données vraisemblablement de bonne qualité. En effet, ces données sont potentiellement plus fiables que les données de prévision du premier canal, puisque seulement 7 % des observations du jeu de données réel sont anormales. Cependant, cet argument peut être critiqué par le fait que ces 7 % d'anomalies vont être également pris en compte pour l'analyse des observations pour les stations qui se situeront dans le voisinage d'observations suspectes, rendant ce canal moins fiable. Par ailleurs, cet ajout bénéficiera davantage aux stations situées dans des zones de forte densité de stations qu'à celles situées notamment dans le nord du Québec, puisque ces dernières n'auront pas ou peu d'information additionnelle *via* ce second canal, compte tenu de leur fort éloignement des stations les plus proches. Malgré tout, cet ajout ayant amélioré les performances sur les erreurs artificielles, on peut raisonnablement considérer qu'il est pertinent et donc à conserver pour la version finale du modèle, d'autant que les performances sur le jeu de données réel sont intéressantes.

Ainsi, ces précédents points révèlent les contributions de ce projet vis-à-vis du premier sous-objectif, qui est de constituer un jeu de données basé sur l'intégration de différentes sources d'informations météorologiques journalières au Québec. En effet, différentes sources d'information sont combinées : des données provenant de capteurs sont fusionnées avec un produit

de prévision de façon astucieuse pour obtenir un ensemble cohérent permettant d'alimenter un modèle d'apprentissage profond.

5.4 Ajustement du seuil de classification avec le jeu de données réel

Lorsqu'est venu le temps de tester notre modèle sur un jeu de données qui contient non plus des erreurs générées artificiellement mais des vraies erreurs provenant de capteurs, les résultats sur l'ensemble de test n'étaient pas aussi bons qu'escomptés. Cela serait dû au fait que les erreurs sont plus subtiles que celles du jeu de données simulées, et il fallait donc trouver un moyen d'améliorer ces performances. En ajoutant une couche de convolution, en effectuant quelques essais avec différentes tailles de noyaux convolutifs (voir section 2.5.2) et en utilisant la méthode de la section 3.8 pour ajuster le seuil de classification, les performances étaient effectivement plus intéressantes. En particulier, un compromis pertinent a été obtenu, qui permet de maintenir un taux de faux négatifs très bas tout en conservant un taux de faux positifs modéré. De cette façon, on s'assure que les données suspectes sont bien détectées, tout en tâchant de ne pas rejeter à tort trop de données authentiques. La façon de trouver cette nouvelle méthode est majoritairement empirique, avec plusieurs itérations d'essais/erreurs. Malgré tout, la conviction qu'il fallait contrôler d'une certaine façon le nombre de faux négatifs a entraîné l'inclusion de cet indicateur dans le choix du seuil de classification de chaque station.

5.5 Synthèse des travaux

Ainsi, après avoir effectué une revue de la littérature sur les différentes façons d'identifier les données météorologiques suspectes, une méthode en particulier s'est démarquée par son ingéniosité et son caractère innovant. De cette façon, la méthode de Sha *et al.* (2021) utilisée sur des données de Colombie-Britannique a été adaptée pour les mesures de précipitations au Québec. Des modifications ainsi que des ajouts de canaux supplémentaires ont été réalisés. En particulier, une identification des données dans les précipitations journalières via des réseaux de neurones convolutifs a été effectuée. Ces réseaux effectuent une analyse d'images, lesquelles combinent différentes sources de données : ponctuelles provenant des capteurs aux stations météorologiques, et surfaciques pour le produit de réanalyse RDRS. Sur un jeu de données contenant des données non validées, et donc présentant des anomalies réelles provenant des capteurs, le modèle proposé classe correctement près de 75 % des observations, et les données suspectes sont correctement détectées dans plus de 95 % des cas. En parallèle, les observations de bonne qualité sont injustement rejetées par le modèle 1 fois sur 4 en moyenne. Ainsi, cette

étude a contribué à l'automatisation de la détection des valeurs suspectes pour l'assurance qualité des précipitations journalières au Québec, s'acquittant au passage du deuxième sous-objectif de ce projet de recherche.

5.6 Limites de la solution proposée

La limite principale de la solution développée est le fait que les efforts ont été concentrés sur la résolution temporelle journalière. En effet, les données validées par les équipes d'assurance qualité chez Solutions Mesonet, organisme en charge de l'assurance qualité d'un vaste réseau de stations météorologiques, sont à une résolution horaire. Or, le transfert d'une méthode automatique comme celle que l'on a développée de l'échelle journalière à l'échelle horaire ne se fait pas simplement. En effet, les sorties des modèles de prévision utilisées dans la méthode sont liées à une incertitude inévitable, due à l'extrême complexité des phénomènes météorologiques, comme par exemple la convection profonde qui résulte en des précipitations dont la prévision est excessivement difficile. Lorsque des données journalières sont traitées, une accumulation de 24 données horaires est effectuée, ce qui a pour effet de lisser ces incertitudes et de réduire la variance. Néanmoins, pour les données horaires ces incertitudes représentent un enjeu crucial, qui demandera probablement une adaptation sérieuse de la méthode proposée dans ce projet.

Par ailleurs, une limite supplémentaire est la difficulté d'analyser les raisons des faux positifs et faux négatifs engendrés par le modèle de classification (voir annexe A). Quelques pistes sont bien sûr privilégiées : en effectuant une régression linéaire, une certaine incertitude liée à cette modélisation est ajoutée, ce qui peut entraîner des erreurs de classification. Le modèle de prévision peut également être mis en cause, car il comporte nécessairement une part d'incertitude inévitable. De plus, certaines stations peuvent être déclarées comme étant en erreur en raison d'anomalies récurrentes, leurs mesures sont donc systématiquement reliées à un *flag* de mauvaise qualité alors pourtant que la donnée produite à une certaine journée pour diverses raisons peut s'avérer correcte, ce qui peut entraîner un faux négatif. Cela est dû au fait que la référence utilisée pour représenter la réalité est celle donnée par les *flags* des experts en assurance qualité, ce qui n'est pas parfait car il existe certaines contraintes liées aux stations les plus problématiques, et l'erreur est humaine. Malgré tout, c'est le mieux qui puisse être faire car une grande confiance est donnée en le processus d'assurance qualité actuellement implémenté, et c'est pour cela qu'il sert de base pour entraîner le modèle d'apprentissage profond.

5.7 Améliorations futures

Une avenue pour une possible amélioration future est l'inclusion d'un canal supplémentaire correspondant à des données radar. En effet, ces données supplémentaires enrichiraient le modèle puisque ce nouveau canal correspondrait précisément à la réflectivité des précipitations détectée par le radar. Cette réflectivité est en fait la puissance retournée au radar lorsque le faisceau rencontre des hydrométéores (neige, pluie, verglas, grêle). Ainsi, elle renseigne sur la vraisemblance de l'occurrence de précipitations. À titre d'exemple, un capteur détectant des précipitations à un moment où le radar ne détecte aucun hydrométéore paraît suspect, et ce cas de figure serait justement géré par ce nouveau canal. Ce canal était censé être inclus dans la méthodologie, mais en raison d'une incompatibilité temporelle il a été impossible de l'inclure. En effet, les données radar disponibles s'étendaient sur une période débutant en 2019, alors que les données RDRS utilisées, à savoir la version 2, disponibles sur le portail d'archives s'arrêtent en 2018. Cet arrêt est dû au fait que la version 3 de RDRS a tout récemment remplacé la version 2. Cette nouvelle version a été rendue disponible très peu de temps avant la fin de la période d'expérimentation de ce projet, c'est pourquoi il n'était pas réaliste de l'inclure dans ce mandat. En outre, dans l'optique d'ajouter ce canal de données radar, il faut garder en tête que certaines stations ne seront couvertes par aucun radar. Ainsi, cet ajout sera impossible, mais pour les stations qui le sont, cela sera sûrement bénéfique pour les performances d'identification de données suspectes. Les modèles étant entraînés individuellement station par station, rien n'empêche d'ajouter un canal de données radar pour les stations couvertes par ce type de données, et de laisser les autres avec le nombre de canaux original.

Par ailleurs, une autre amélioration possible serait de considérer d'autres types de capteurs : en plus des observations provenant des précipitomètres, il serait bénéfique d'incorporer celles issues des pluviomètres à auget basculeur. En effet, leur densité est plus élevée que celle des précipitomètres dans le réseau de stations météorologiques. Leur apport sur le volume de données utilisé pour entraîner le modèle de classification serait probablement bénéfique, d'autant que la période étudiée, à savoir les précipitations estivales, fait partie du cycle de fonctionnement de ces capteurs à augets.

Enfin, une voie pour ce projet qui représenterait son accomplissement serait qu'il soit déployé en temps réel. En effet, dès l'instant où un capteur envoie une mesure, grâce à la prévision RDRS correspondante, le modèle serait capable de prédire l'étiquette de qualité de l'observation. Cela pourrait donc permettre aux équipes en charge du contrôle qualité de prioriser leur travail, voire d'en alléger la charge en ayant en temps quasi-réel une classification automatisée capable de détecter des données suspectes plus subtiles que celles détectées par les tests

statistiques déjà implémentés. Malgré tout, cela nécessiterait la mise en place d'un pipeline automatisé qui réalise l'intégration des différentes sources de données, en plus d'ajustements pour l'entraînement des modèles, qui seraient alimentés au fur et à mesure et pourraient donc bénéficier de l'apprentissage incrémental (*online learning*). Néanmoins, cela pourrait tout à fait constituer l'objet de futurs travaux, puisque ce projet a démontré la pertinence de l'outil proposé pour détecter les valeurs suspectes dans les mesures de précipitations journalières au Québec.

RÉFÉRENCES

Artz, R., Ball, G., Behrens, K., Bonnin, G., Bower, C., Canterford, R., Childs, B., Claude, H., Crum, T., Dombrowsky, R., Edwards, M., Evans, R., Feister, E., Forgan, B., Hilger, D., Holleman, I., Hoogendijk, K., Johnson, M., Klapheck, K.-H., Klausen, J., Koehler, U., Ledent, T., Luke, R., Nash, J., Oke, T., Painting, D., Pannett, R., Qixian, Q., Rudel, E., Saffle, R., Schmidlin, F., Sevruk, B., Srivastava, S., Steinbrecht, W., Stickland, J., Stringer, R., Sturgeon, M., Thomas, R., Van der Meulen, J., Vanicek, K., Wieringa, J., Winkler, P., Zahumensky, I. et Weixin, Z. (2023). Guide to instruments and methods of observation (WMO-No. 8) | World Meteorological Organization.

Berger, V. W. et Zhou, Y. (2014). Kolmogorov–Smirnov Test : Overview. *In Wiley StatsRef : Statistics Reference Online*. John Wiley & Sons, Ltd.

Bernier, J. (1994). Statistical detection of changes in geophysical series. *In* Duckstein, L. et Parent, E., éditeurs : *Engineering Risk in Natural Resources Management : With Special References to Hydrosystems under Changes of Physical or Climatic Environment*, pages 159–176. Springer Netherlands, Dordrecht.

Boulanger, J.-P., Aizpuru, J., Leggieri, L. et Marino, M. (2010). A procedure for automated quality control and homogenization of historical daily temperature and precipitation data (APACH) : Part 1 : Quality control and application to the Argentine weather service stations. *Climatic Change*, 98(3):471–491.

Burbano-Moreno, A. A. et Mayrink, V. D. (2024). Spatial Functional Data analysis : Irregular spacing and Bernstein polynomials. *Spatial Statistics*, 60:100832.

Di Piazza, A., Conti, F. L., Noto, L. V., Viola, F. et La Loggia, G. (2011). Comparative analysis of different techniques for spatial interpolation of rainfall data to create a serially complete monthly time series of precipitation for sicily, italy. *International Journal of Applied Earth Observation and Geoinformation*, 13(3):396–408.

Dyck, S. (1976). *Angewandte Hydrologie : Teil 1 ; Berechnung und Regelung des Durchflusses der Flüsse : Teil 2 ; Der Wasserhaushalt der Flüsse*. VEB f. Bauwesen.

Fortin, V., , G., R., , T., S., , K., K., , N., G., et Mahidjiba, A. (2018). Ten years of science based on the Canadian Precipitation Analysis : A CaPA system overview and literature review. *Atmosphere-Ocean*, 56(3):178–196.

Gelfand, A., Banerjee, S. et Gamerman, D. (2005). Spatial process modelling for univariate and multivariate dynamic spatial data. *Environmetrics*, 16(5):465–479.

Goodfellow, I., Bengio, Y. et Courville, A. (2016). *Deep Learning*. MIT Press.

Goodison, B., Louie, P. et Yang, D. (1998). Wmo solid precipitation measurement intercomparison. *World Meteorological Organization-Publications-WMO TD*, 67.

Groisman, P. Y., Koknaeva, V. V., Belokrylova, T. A. et Karl, T. R. (1991). Overcoming biases of precipitation measurement : A history of the ussr experience. *Bulletin of the American Meteorological Society*, 72(11):1725–1733.

Hofmeister, F., Graziano, F., Marcolini, G., Willems, W., Disse, M. et Chiogna, G. (2023). Quality assessment of hydrometeorological observational data and their influence on hydrological model results in Alpine catchments. *Hydrological Sciences Journal*, 68(4):552–571.

Hubbard, K. G. et You, J. (2005). Sensitivity analysis of quality assurance using the spatial regression approach — A case study of the maximum/minimum air temperature. *Journal of Atmospheric and Oceanic Technology*, 22(10):1520–1530.

Ingleby, N. B. et Lorenc, A. C. (1993). Bayesian quality control using multivariate normal distributions. *Quarterly Journal of the Royal Meteorological Society*, 119(513):1195–1225.

Kendall, M. G. (1975). *Rank correlation methods*. Griffin, London, 4th ed., 2d impression édition.

Kingma, D. P. et Ba, J. (2017). Adam : a method for stochastic optimization.

Konda, S., Rani, B., Mangu, V., Gunda, M. et Ramana, B. (2019). Convolution neural networks for binary classification. *Journal of Computational and Theoretical Nanoscience*, 16:4877–4882.

LeCun, Y., Bengio, Y. et Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

Leeper, R. D. et Kochendorfer, J. (2015). Evaporation from weighing precipitation gauges : impacts on automated gauge measurements and quality assurance methods. *Atmospheric Measurement Techniques*, 8(6):2291–2300.

Li, J., Xu, H., Deng, J. et Sun, X. (2016). Hyperbolic linear units for deep convolutional neural networks. *In 2016 International Joint Conference on Neural Networks, IJCNN 2016, July 24, 2016 - July 29, 2016*, volume 2016-October de *Proceedings of the International*

Joint Conference on Neural Networks, pages 353–359, Vancouver, BC, Canada. Institute of Electrical and Electronics Engineers Inc.

Mann, H. B. (1945). Nonparametric tests against trend. *Econometrica*, 13(3):245–259.

Matheron, G. (1963). *Traité de geostatistique appliquée. Tome II. Le Krigeage*. Mémoires du Bureau de recherches géologiques et minières. Ed. B.R.G.M., Paris.

NOAA National Centers for Environmental Information (2022). ETOPO 2022 15 Arc-Second Global Relief Model.

Pearson, K. (1900). X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175.

Pettitt, A. N. (1979). A non-parametric approach to the change-point problem. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(2):126–135.

Rasmussen, R., Baker, B., Kochendorfer, J., Meyers, T., Landolt, S., Fischer, A. P., Black, J., Thériault, J. M., Kucera, P., Gochis, D., Smith, C., Nitu, R., Hall, M., Ikeda, K. et Gutmann, E. (2012). How well are we measuring snow : The noaa/faa/ncar winter precipitation test bed. *Bulletin of the American Meteorological Society*, 93(6):811 – 829.

Rivoirard, J. (2005). *Concepts and Methods of Geostatistics*, pages 17–37. Springer New York, New York, NY.

Schönwiese, C.-D. (2000). *Praktische Statistik für Meteorologen und Geowissenschaftler*. Schweizerbart Science Publishers, Stuttgart, Germany.

Sha, Y., Ii, D. J. G., West, G. et Stull, R. (2021). Deep-learning-based precipitation observation quality control. *Journal of Atmospheric and Oceanic Technology*, 38(5):1075–1091.

Steinacker, R. (2011). Data quality control based on self-consistency. *Monthly Weather Review*, 139.

Wallace, D. L. (1959). Simplified beta-approximations to the Kruskal-Wallis H test. *Journal of the American Statistical Association*, 54(285):225–230.

Xiong, L. et Guo, S. (2004). Trend test and change-point detection for the annual discharge series of the yangtze river at the yichang hydrological station. *Hydrological Sciences Journal*, 49(1):99–112.

Xiong, X., Jiang, Z., Tang, H., Zhang, Y. et Ye, X. (2022). Research on quality control methods for surface temperature observations via spatial correlation analysis. *International Journal of Climatology*, 42(16):10268–10284.

Xiong, X., Ye, X. et Zhang, Y. (2017). A quality control method for surface hourly temperature observations via gene-expression programming. *International Journal of Climatology*, 37(12):4364–4376.

Xu, C.-D., Wang, J.-F., Hu, M.-G. et Li, Q.-X. (2013). Interpolation of missing temperature data at meteorological stations using p-bshade. *Journal of Climate*, 26(19):7452–7463.

Xu, C.-D., Wang, J.-F., Hu, M.-G. et Li, Q.-X. (2014). Estimation of uncertainty in temperature observations made at meteorological stations using a probabilistic spatiotemporal approach. *Journal of Applied Meteorology and Climatology*, 53(6):1538–1546.

Yan, Q., Zhang, B., Jiang, Y., Liu, Y., Yang, B. et Wang, H. (2024). Quality control of hourly rain gauge data based on radar and satellite multi-source data. *Journal of Hydroinformatics*, 26(5):1042–1058.

You, J. et Hubbard, K. G. (2006). Quality control of weather data during extreme events. *Journal of Atmospheric and Oceanic Technology*, 23(2):184–197.

You, J., Hubbard, K. G., Nadarajah, S. et Kunkel, K. E. (2007). Performance of quality assurance procedures on daily precipitation. *Journal of Atmospheric and Oceanic Technology*, 24(5):821–834.

Øgland, P. (1993). Theoretical analysis of the dip-test in quality control of geophysical observations. *Klima*, 10:1–18.

ANNEXE A FAUX POSITIFS ET FAUX NÉGATIFS

Dans cette annexe, on présente quelques exemples de faux positifs et de faux négatifs tirés aléatoirement du jeu de données comportant des valeurs suspectes provenant des capteurs, et non pas générées artificiellement.

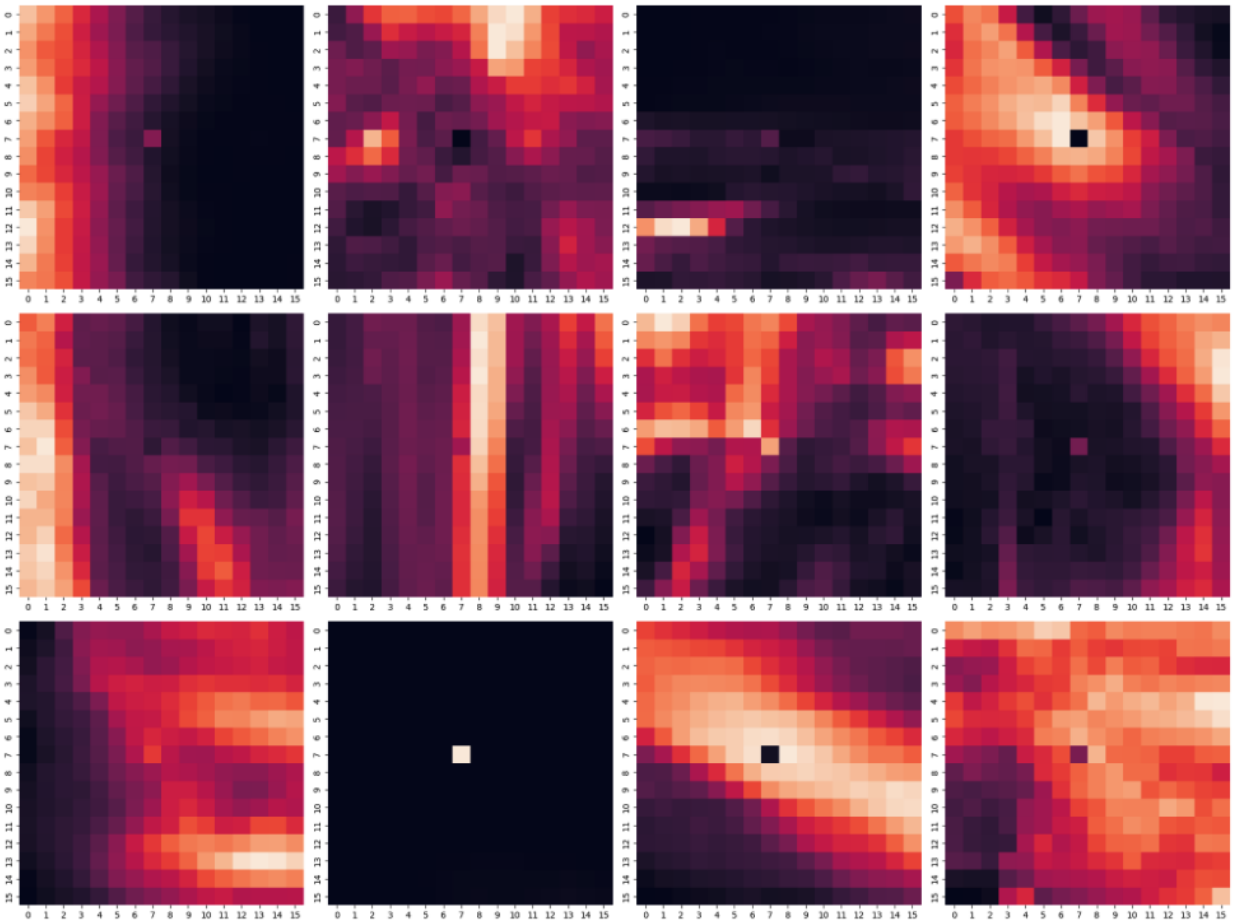


FIGURE A.1 Faux positifs

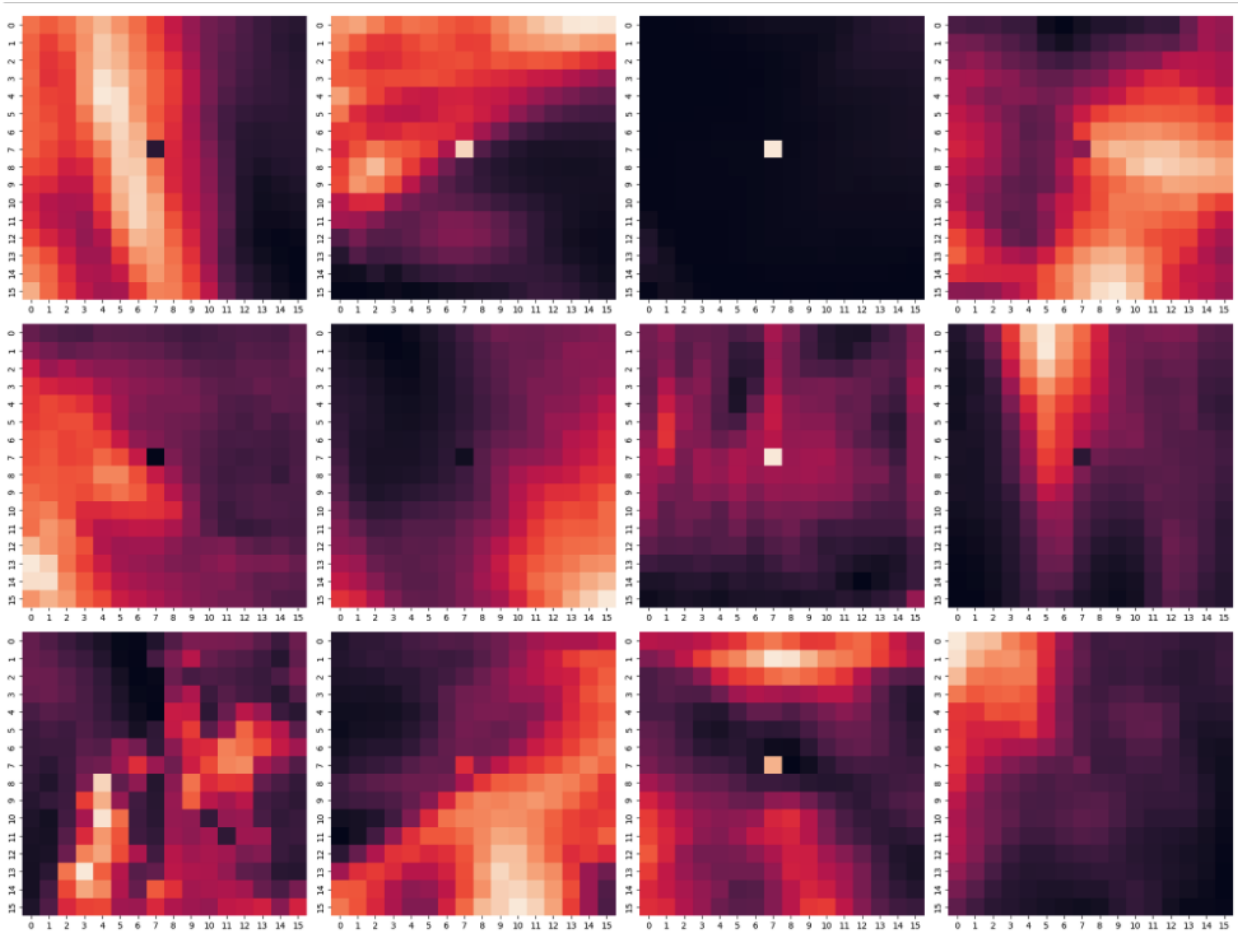


FIGURE A.2 Faux négatifs

ANNEXE B GÉNÉRATION DE VALEURS SUSPECTES ARTIFICIELLES

Comme expliqué dans la section 3.4.2, afin de faire correspondre les précipitations observées avec les précipitations réanalysées, on va venir remettre à l'échelle les précipitations mesurées via la droite de régression linéaire que nous avons déterminée (pour chaque station indépendamment ou de façon globale). En effet, on se munit de l'estimateur $\hat{\boldsymbol{\beta}}$ et de la matrice de structure X, que l'on multiplie ensemble pour obtenir le vecteur de prédictions $\hat{\boldsymbol{y}}$, qui correspond aux mesures de précipitations remises à l'échelle pour correspondre aux précipitations réanalysées :

$$\hat{\boldsymbol{y}} = X\hat{\boldsymbol{\beta}} \tag{B.1}$$

D'une certaine façon, on a projeté les précipitations observées dans l'espace des précipitations réanalysées, afin de les rendre comparables, ce qui revêt une grande importance dans la section 3.5.1. Forts de ces prédictions, il est possible d'évaluer leur précision à l'aide du calcul des résidus, définis par :

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}}$$

où $\boldsymbol{y}$ est le vecteur des données de prévision de précipitations du point de grille le plus proche de la station (c'est-à-dire la variable expliquée).

Plus les composantes de ce vecteur sont proches de 0, plus les prédictions sont justes. En outre, grâce aux résidus on peut donner un estimateur non biaisé de la variance de l'erreur σ^2 noté $\hat{\sigma}^2$. Il est défini dans notre cas avec une seule variable explicative par :

$$\hat{\sigma}^2 = \frac{\|\boldsymbol{e}\|^2}{n-2}$$

Afin de reproduire la situation de l'article de Sha *et al.* (2021), on propose de générer des anomalies sur une portion des observations, de façon artificielle et contrôlée, pour rester maître de l'amplitude des erreurs introduites. Afin de s'assurer de créer des erreurs qui puissent dans un premier temps être détectées assez facilement, on veut éviter que celles-ci soient trop discrètes. Ainsi, on utilise la modélisation statistique de la régression linéaire qui caractérise le lien entre précipitations mesurées et prévision sur grille pour générer des observations erronées basées sur un indice appelé résidu studentisé, qui quantifie la distance de l'observation à la droite de régression. Le résidu studentisé associé à une observation est défini par :

$$s_i = \frac{e_i}{\sqrt{\hat{\sigma}^2(1-h_i)}} = \frac{y_i - \hat{y}_i}{\sqrt{\hat{\sigma}^2(1-h_i)}} \tag{B.2}$$

où e_i est la i^{e} composante du vecteur $\boldsymbol{e}$, y_i est la i^{e} composante du vecteur $\boldsymbol{y}$, $\hat{y}_i$ est la i^{e} composante du vecteur $\hat{\boldsymbol{y}}$ et h_i est le i^{e} élément de la diagonale de la matrice $H = X(X^\top X)^{-1}X^\top$.

Généralement, on considère les observations ayant un résidu studentisé dont la valeur absolue est supérieure à 3 comme étant suspectes. Ainsi, pour générer nos valeurs suspectes, on sélectionne au hasard une portion du jeu de données, puis on échantillonne deux vecteurs aléatoires dont les marginales sont indépendantes et identiquement distribuées (i.i.d.), uniformément distribuées sur l'intervalle $[3, 4]$. Chacun d'entre eux correspond à un sous-ensemble (faibles ou fortes précipitations) et est donc de dimension égale au nombre de valeurs suspectes que l'on désire implanter dans chaque sous-ensemble. Cette étape correspond à la génération de résidus studentisés extrêmes, qui s'associent à une observation suspecte. La valeur de l'observation en question peut être retrouvée grâce à l'équation (B.2), en isolant y_i :

$$y_i = \hat{y}_i + s_i\sqrt{\hat{\sigma}^2(1 - h_i)} \tag{B.3}$$

Ainsi, pour les observations sélectionnées, nous remplaçons l'observation par le résultat de l'équation précédente. Celle-ci a pour effet de projeter l'observation à la station dans l'espace de la prévision du point de grille le plus proche, puis d'ajouter une quantité correspondant à une erreur de captation de la précipitation. Ces données comportent donc une erreur artificielle, ainsi elles sont considérées comme des anomalies et marquées comme telles. Or, comme nous avons généré des résidus studentisés positifs uniquement, nous allons, lorsque cela est possible, appliquer la formule suivante, de façon à simuler la conséquence d'un résidu studentisé négatif :

$$y_i = \hat{y}_i - s_i\sqrt{\hat{\sigma}^2(1 - h_i)} \tag{B.4}$$

Puisque nous utilisons des jeux de données de précipitations, cette nouvelle valeur de y_i doit demeurer positive (une précipitation négative n'a pas de sens). Si ce n'est pas le cas, alors on utilise l'équation (B.3), de façon à s'assurer que toutes les précipitations résultant de l'insertion d'erreurs artificielles sont bien positives (ou nulles). En réalité, pour le sous-ensemble des faibles précipitations, aucune observation n'a pu être modifiée selon l'équation (B.4), ce qui semble logique, car comme les précipitations sont déjà faibles, il est normal qu'on ne puisse pas les diminuer encore davantage tout en maintenant une valeur positive. En revanche, pour le sous-ensemble des plus fortes précipitations, certaines observations ont pu être transformées en anomalies *via* l'équation (B.4), car les valeurs des observations initiales étaient assez élevées pour permettre une baisse sans pour autant passer en-dessous de 0.

Les observations n'ayant pas été désignées pour accueillir une erreur artificielle sont considérées de bonne qualité, car provenant d'un jeu de données contrôlé et validé. Afin de faire cor-

respondre précipitations observées et précipitations réanalysées, on projette les observations aux stations dans l'espace des prévisions sur grille grâce à la droite de régression, séparément pour les deux sous-ensembles. On utilise pour cela l'équation (B.1), pour les valeurs de $\hat{\boldsymbol{\beta}}$ respectives des deux sous-parties.

ANNEXE C DÉFINITION DES LABELS POUR LE JEU DE DONNÉES RÉEL

Les labels employés par Solutions Mesonet sont les labels $0, 1, 2, 3, 4, 8$ et 9. Les labels $4, 8$ et 9 ne sont pas pris en compte, car ils correspondent respectivement à un capteur temporairement bloqué, un capteur non installé et une station qui n'envoie pas de données. Ces observations sont donc retirées du jeu de données, car elles ne concernent pas spécifiquement la qualité des données, mais plutôt leur disponibilité ou leur période de fonctionnement (certains capteurs sont bloqués en hiver car non adaptés aux précipitations solides). Les labels restants sont donc les labels $0, 1, 2$ et 3, qui définissent respectivement une observation de bonne qualité, une donnée douteuse (*suspect*) qui est visible dans les produits destinés aux partenaires, une donnée douteuse (*warning*) mais qui n'est pas visible dans les produits destinés aux partenaires, et enfin une donnée jugée en erreur (*failure*). Au total, les labels 0 et 1 sont considérés comme des données de bonne qualité, puisque les données correspondantes sont transmises aux partenaires, et les labels 2 et 3 sont considérés comme des anomalies, puisque les observations associées ne sont pas visibles par les partenaires. En outre, les labels dont on dispose correspondent à des observations de précipitations horaires, donc il convient de convertir ces labels horaires en labels journaliers. Pour ce faire, on adopte une approche conservative, puisqu'on considère qu'une observation de précipitations journalières est une anomalie si au moins un des labels horaires contenus dans la période d'accumulation témoigne d'une anomalie (label 2 ou 3). En particulier, le label journalier est défini par :

$$l_{journalier} = \max_{1 \leq i \leq 24} \left\{ l_{horaire}^{(i)} \right\}$$

où $l_{horaire}^{(i)}$ correspond au label de la i$^{\text{e}}$ observation de précipitation horaire de la période d'accumulation. Cette définition nous permet de ne négliger aucune observation anormale, car chaque journée contenant au moins une de ses observations horaires présentant une anomalie est considérée comme de mauvaise qualité. Ensuite, comme dans la section 2.5.1, on crée deux classes pour nos observations, qui sont construites sur le même principe. En effet, les observations journalières présentant un label de qualité valant 0 ou 1 relèveront de la classe 0 (bonne qualité), tandis que celles dont le label est égal à 2 ou 3 seront de la classe 1 (anomalie).

Dans ce cas, les valeurs suspectes ne sont pas générées *via* les résidus studentisés de la régression linéaire comme dans la section 3.4.3, mais proviennent directement de vraies erreurs

de mesure des précipitomètres.