Quantification et réduction de l'incertitude dans un modèle de transfert de pesticides à l'échelle du bassin versant

Émilie Rouzies

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

École doctorale : MSTII - Mathématiques, Sciences et technologies de l'information, Informatique
Spécialité : Mathématiques Appliquées
Unité de recherche : Laboratoire Jean Kuntzmann

Quantification et réduction de l'incertitude dans un modèle de transfert de pesticides à l'échelle du bassin versant

Quantifying and reducing the uncertainties in a pesticide transfer model at the catchment scale

Présentée par :

Emilie ROUZIES

Direction de thèse :

Arthur VIDARD Chargé de recherche HDR, INRIA CENTRE GRENOBLE RHONE-ALPES	Directeur de thèse
Claire Lauvernet INRAE	Co-encadrante de thèse

Rapporteurs :

ROGER MOUSSA
Directeur de recherche, INRAE CENTRE OCCITANIE- MONTPELLIER
SOPHIE RICCI
Ingénieur docteur, CERFACS

Thèse soutenue publiquement le **23 février 2023**, devant le jury composé de :

ROGER MOUSSA Directeur de recherche, INRAE CENTRE OCCITANIE- MONTPELLIER	Rapporteur
SOPHIE RICCI Ingénieur docteur, CERFACS	Rapporteure
EMMANUEL COSME Maître de conférences HDR, UNIVERSITE GRENOBLE ALPES	Examinateur
CELINE HELBERT Maître de conférences HDR, ECOLE CENTRALE LYON	Examinatrice
FLORENCE HABETS Directeur de recherche, CNRS DELEGATION PARIS CENTRE	Examinatrice
ROBERT FAIVRE Directeur de recherche, INRAE CENTRE OCCITANIE- TOULOUSE	Président
CLAUDIO PANICONI Professeur, Institut National Recherche Scientifique	Examinateur
ELISE ARNAUD Maître de conférences, UNIVERSITE GRENOBLE ALPES	Examinatrice

Invités :

ARTHUR VIDARD
Chargé de recherche HDR, INRIA CENTRE GRENOBLE RHONE-ALPES
CLAIRE Lauvernet
Chargé de recherche, INRAE LYON-GRENOBLE AUVERGNE-RHONE-ALPES

Quantification et réduction de l'incertitude dans un modèle de transfert de pesticides à l'échelle du bassin versant

Emilie Rouzies

14 octobre 2022

Chapitre 1

Introduction

Sommaire

1.1 Contexte général

La question de l'usage des pesticides est un enjeu de société majeur, souvent clivant entre consommateurs et riverains inquiets pour leur santé et agriculteurs tributaires d'une agriculture productiviste. C'est dans les années 30 que l'utilisation des pesticides a débuté et aujourd'hui, le recours à ces subtances soutient encore largement un modèle d'agriculture intensive. La France est le 9ième plus gros consommateur de pesticides en Europe avec en moyenne 4.46 kg/ha en 2019 (FAO, 2022). Cependant, une telle utilisation engendre de nombreux risques liés à des enjeux diversifiés. Sur la période 2017-2019, près de 20% des points de mesure du réseau de surveillance des eaux de surface françaises dépassaient les concentrations maximales ou les concentrations moyennes annuelles admissibles pour au moins un pesticide. Pour les eaux souterraines, c'est près de 80% des points de mesure qui révélaient la présence de pesticides. Pour 35 % de ces points, les concentrations totales, toutes substances confondues dépassaient la norme de 0.5 μg/L et pour 47% d'entre eux, elles dépassaient la norme de 0.1 μg/L pour au moins une substance individuelle (SDES, 2020). Ces normes, fixées par la directive européenne 98/83/CE définissent les seuils de potabilité de l'eau et mettent ainsi en lumière les risques que fait peser l'utilisation de pesticides sur l'accès à l'eau potable. Les enjeux sont également économiques puisque de telles contaminations impliquent la mise en place de mesures de dépollution voire la création de nouveaux captages. D'autre part, cette contamination généralisée est aussi impactante pour les organismes (animaux et végétaux) des cours d'eau et elle entraîne de profonds déséquilibres des écosystèmes aquatiques.
Les leviers d'action permettant de limiter de telles contaminations sont multiples et se situent à plusieurs niveaux. Le processus d'autorisation de mise sur le marché (homologation) permet d'abord d'évaluer les risques de la substance pour la santé humaine et l'environnement. La procédure d'homologation est mise en oeuvre par des agences au niveau européen (comité phytosanitaire permanent) puis national (ministère chargé de l'agriculture et agence nationale de sécurité sanitaire, de l'alimentation, de l'environnement et du travail en France) qui évaluent à la fois l'efficacité du pesticide et son écotoxicité, en examinant notamment son impact sur la dégradation de la qualité des eaux. Pour cela, son potentiel de transfert vers les masses d'eau est évalué, notamment grâce à des outils de modélisation. Ces outils sont développés à partir de l'expertise scientifique concernant les voies de transferts des pesticides et sont ensuite appliqués sur des scénarios de "pires cas réalistes". Un second levier d'action passe ensuite par la réduction des usages par diverses actions incluant le changement des pratiques agricoles, l'affinement des critères de décision de traitement et des dosages associés ou le recours à des méthodes alternatives inspirées par exemple de la protection agroécologique des cultures. Enfin, un dernier levier d'action consiste à atténuer les transferts après l'application des pesticides sur les parcelles cibles. Pour cela, on s'appuie entre autre sur l'aménagement du territoire et sur le rôle tampon de certains éléments du paysage. Cependant, l'utilisation du paysage comme levier d'action nécessite une connaissance fine des voies d'écoulements des pesticides et du rôle des différentes structures paysagères vis à vis de ces écoulements. Là encore, en complément de la connaissance du terrain, l'utilisation d'outils

de modélisation peut contribuer à cette démarche pour quantifier l'impact du paysage sur les transferts et comparer plusieurs scénarios d'aménagement du territoire.

1.2 Devenir des pesticides à l'échelle d'un bassin versant

Une partie des pesticides appliqués dans les zones agricoles atteint des espaces non cibles, notamment des milieux aquatiques sensibles. Au fil de leurs déplacements dans le bassin versant, ils interagissent aussi avec ce milieu : une partie des pesticides peut être piégée dans le sol, une autre peut être dégradée. Comprendre le devenir des pesticides et les concentrations résiduelles dans les espaces aquatiques nécessite ainsi de connaître les voies de transferts et les processus qui les affectent. Les principaux processus régissant le devenir des pesticides à l'échelle d'un bassin versant sont résumés sur la Figure 1.1 et présentés dans les paragraphes suivants.

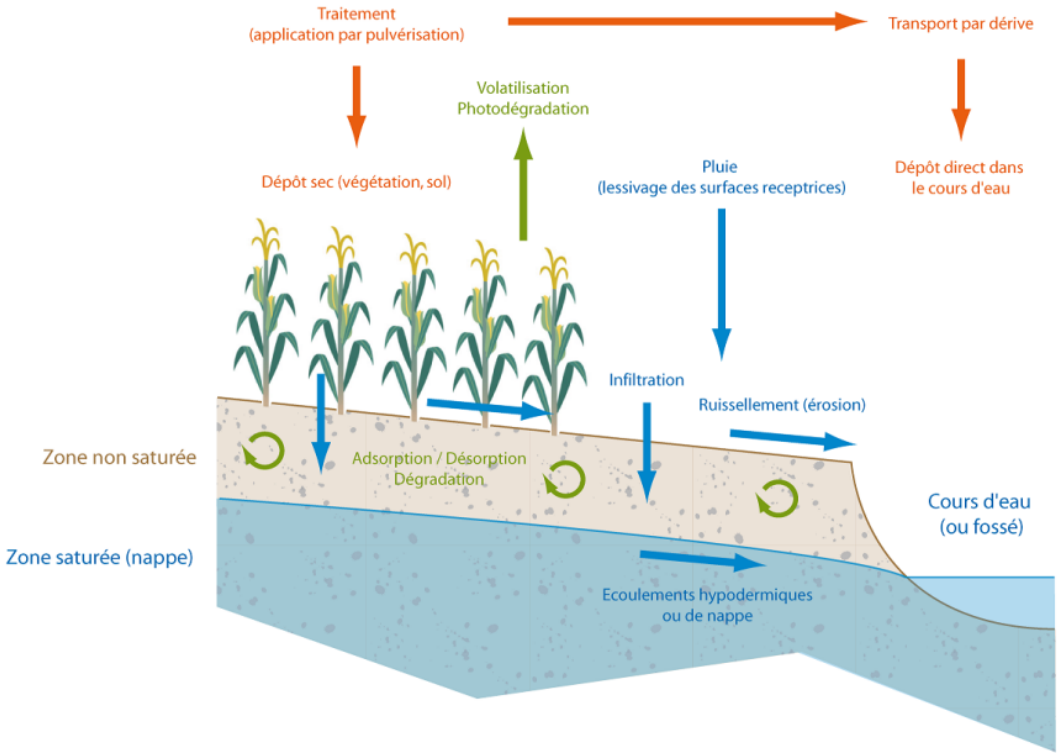


Figure 1.1 – Principaux processus de transfert et de transformation des pesticides à l'échelle d'un bassin versant (Catalogne et Le Hénaff, 2016).

1.2.1 Les voies de transfert

Au moment de leur application sur une parcelle agricole, une partie des pesticides est dispersée dans l'atmosphère et transportée par voie aérienne. On parle alors de dérive de

pulvérisation. A celle-ci vient parfois s'ajouter le phénomène de volatilisation qui entraîne une remobilisation des molécules déposées en surface du sol ou des plantes en phase gazeuse, et ce, parfois plusieurs semaines après l'application. L'ampleur des phénomènes de dérive et de pulvérisation dépend principalement du matériel utilisé, des conditions météorologiques lors de l'application et des propriétés physico-chimiques de la solution appliquée. Ainsi, les valeurs reportées dans la littérature sont très disparates, allant de 2 % à 90 % des pesticides appliqués transportés par dérive et volatilisation (Taylor et Spencer, 1990 ; Carlsen et al., 2006 ; Druart et al., 2011 ; Lefrancq et al., 2013) .

Une portion des pesticides qui atteignent la surface du sol peut ensuite être déplacée vers le reste du bassin versant par le biais des flux d'eau qui adviennent en surface et en subsurface. Le ruissellement de surface est l'une des voies majeures de transfert, responsable d'écoulements parfois intenses et pouvant contenir de fortes concentrations en pesticides. Il est donc potentiellement très dommageable pour les milieux aquatiques environnants. On distingue généralement deux sortes de ruissellement qui peuvent se produire de manière concomittante sur un bassin malgré des origines bien distinctes : le ruissellement hortonien (Horton, 1933) advient lorsque l'intensité des précipitations est supérieure à la capacité d'infiltration du sol. Le ruissellement par saturation (Dunne et Black, 1970) intervient quant à lui, lorsque le sol est saturé à cause de la présence d'une nappe, ne pouvant donc infiltrer qu'un volume d'eau fortement limité. Les molécules de pesticides sont ainsi transportées dans le flux ruisselé en phase dissoute. A cela s'ajoute la présence de pesticides adsorbés sur des matières en suspension plus ou moins fines, arrachées à la surface du sol et transportées elles aussi par ruissellement. L'intensité des flux d'eau et de matière ruisselés dépend entre autre de la nature pédologique et géologique du milieu, de la topographie mais aussi du travail du sol (sens du travail par rapport à la pente, mise en place de techniques de conservation des sols, etc.).

Le volume d'eau disponible à la surface d'une parcelle et qui ne ruisselle pas finit alors par s'infiltrer. L'eau et les pesticides s'infiltrent verticalement dans les pores du sol et peuvent ainsi percoler jusqu'à la nappe. Cette infiltration, lente si elle a lieu dans la matrice du sol, peut être accélérée par la présence de macropores (racines, galerie d'animaux, micro fissures, etc.) qui constituent des voies d'écoulements préférentiels. Le volume d'eau et les molécules qu'il transporte qui empruntent ces macropores rejoignent plus rapidement la nappe. Si le rôle essentiel des écoulements préférentiels dans la compréhension des pollutions des nappes phréatiques est connu depuis plusieurs décennies (Beven et Germann, 1982 ; Barbash et Resek, 1996), il n'en reste pas moins que ces écoulements restent difficiles à quantifier.

Finalement, de l'eau peut s'accumuler tout le long de la colonne de sol, à la faveur par exemple de changements d'horizons pédologiques. En effet, de tels changements peuvent entraîner des ruptures de perméabilité et l'apparition de zones saturées en fond de profil ou plus superficiellement (on parle alors de nappes perchées). Si la pente est suffisante, des écoulements latéraux saturés apparaissent alors, lesquels peuvent finir par rejoindre le cours d'eau aval. Les concentrations en pesticides que contiennent ces écoulements dépendent alors de la pente, du contexte pédologique et des interactions avec le sol (Peyrard, 2016).

1.2.2 Les processus de transformation et de rétention

En plus des processus de transport décrits dans le paragraphe précédent, les pesticides sont également sujets à des processus de réaction avec le milieu. Rétention et dégradation influent ainsi de manière non négligeable sur les concentrations résultantes dans les milieux aquatiques.

Tout d'abord, la majorité des pesticides se caractérisent par leur capacité d'adsorption, c'est-à-dire leur capacité à se fixer aux particules de sol. L'adsorption est fonction des propriétés du milieu (pH, minéralogie du sol, taux de matière organique) mais aussi des propriétés physico-chimiques des molécules. Le phénomène d'adsorption est parfois réversible et on observe dans certain cas la remobilisation des molécules adsorbées.

D'autre part, une part importante des pesticides appliqués finit par se dégrader au contact du milieu environnant. Une multitude de réactions, à la fois à la surface et en subsurface du sol, entraîne la transformation du produit de base. Parmi les réactions abiotiques les plus fréquentes, on compte l'oxydoréduction, l'hydrolyse et la photolyse. D'autre part, de nombreuses réactions biotiques font intervenir les microorganismes du sol (en particulier champignons et bactéries) préférentiellement dans les zones du sol riches en matière organique. Ces transformations successives produisent des métabolites aux caractéristiques différentes du produit initial, parfois plus nocives pour l'environnement. Quand la dégradation est complète, elle aboutit à la minéralisation du pesticide (Alletto et al., 2006 ; Madrigal et al., 2007 ; Alletto et al., 2013).

1.3 Impact du paysage sur les transferts

Le paragraphe précédent a permis de montrer que le devenir des pesticides à l'échelle d'un bassin versant résulte de nombreux processus physiques complexes, qui interagissent fortement les uns avec les autres. La contribution de chacun de ces processus à la contamination des milieux aquatiques dépend fortement des propriétés du sol et de l'organisation du paysage. Ainsi, cette organisation peut être optimisée afin de réduire les phénomènes de transfert, en atténuant et ralentissant les flux d'eau et de matière en suspension. Pour cela, on peut notamment s'appuyer sur l'utilisation de zones tampons. Il s'agit d'éléments du paysage, le plus souvent des espaces interparcellaires ayant la capacité d'intercepter les flux d'eau de surface et/ou de subsurface et de réduire les flux de substances associés afin de protéger les espaces aquatiques récepteurs de ces flux. Une zone tampon favorise les processus de dissipation en s'appuyant sur trois leviers d'action : l'allongement des temps de transfert, l'accroissement de l'activité biologique et des processus de dégradation associés et l'accroissement de l'adsorption grâce à la présence de matière organique.

1.3.1 Variété et implantation des zones tampons

Il existe une grande variété de zones tampons qui se distinguent par leurs structures et leurs impacts sur les différentes voies de transfert d'eau et de matières. Parmi elles, on compte des dispositifs enherbés (bandes enherbées situées le long d'une rivière, prairies, dispositifs dérivés type enherbement inter-rangs,etc.) qui ralentissent le ruissellement de surface et favorisent la dégradation et l'adsorption des pesticides. Les dispositifs ligneux comme les haies ou les zones boisées le long des cours d'eau (ripisylves) constituent également d'efficaces zones tampons en agissant notamment sur l'interception de la dérive de pulvérisation et de l'érosion. C'est également le cas des fascines, des structures artificielles constituées de l'enchevêtrement de branchages formant des murets et agissant aussi comme barrage à l'érosion (voir la Figure 1.2 pour une illustration). Enfin, on notera l'intérêt particulier des Zones Tampons Humides Artificielles (ZTHA) qui constituent des espaces privilégiés de dégradation notamment grâce à la présence d'une faible lame d'eau et d'espèces végétales diversifiées. Les ZTHA sont particulièrement plébiscitées à l'aval d'un réseau de collecte des écoulements constitués de fossés ou de drains agricoles.
Des exemples de ces différents dispositifs sont illustrés sur la Figure 1.2. Cette même figure présente un exemple d'aménagements variés combinés à l'échelle d'un bassin versant pour protéger efficacement l'intégralité des milieux aquatiques.

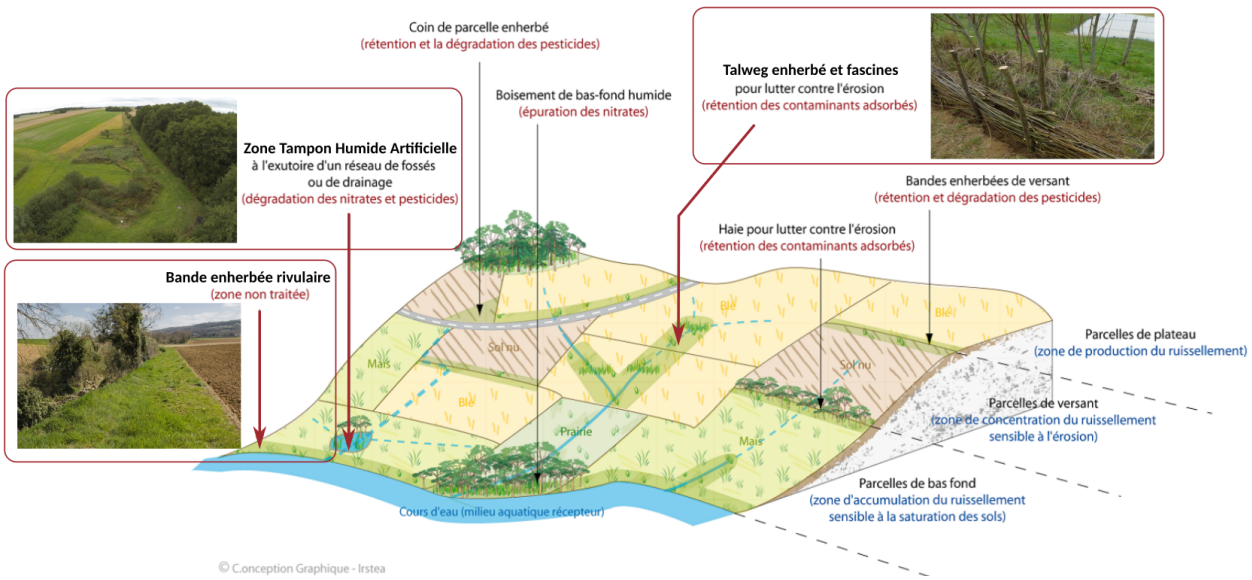


Figure 1.2 – Exemple d'implantation de zones tampons diversifiées à l'échelle d'un bassin versant à dominante agricole (adapté de Catalogne et Le Hénaff 2016).

Ainsi, les zones tampons sont des outils d'aménagement du territoire efficaces qui, en complément de bonnes pratiques agricoles, permettent de limiter les transferts de pesticides vers les milieux aquatiques. Cependant, leur implantation doit être précédée d'un diagnostic approfondi. Celui-ci permet de mieux comprendre le fonctionnement du bassin versant d'étude, d'identifier les voies de transfert majoritaires et de localiser précisément les chemins

de l'eau. Ce diagnostic est indispensable afin de bien positionner et bien dimensionner les différentes zones tampons. Il fait notamment l'object d'une sensibilitation importante dans l'équipe pollutions diffuses d'INRAE avec la mise en place d'un guide d'implantation des zones tampons (Catalogne et Le Hénaff, 2016) et le développement de formations à destination des professionnels concernés (Carluer et al., 2019).

1.3.2 Zoom sur les bandes enherbées

Les bandes enherbées sont les zones tampons les plus emblématiques puisque la réglementation française impose d'implanter de telles structures pour séparer les parcelles agricoles des cours d'eau. Ces zones tampons ont été largement étudiées (Lacas et al., 2005 ; Corpen, 2007 ; Arora et al., 2010) ce qui permet de décrire précisément leur fonctionnement. L'impact des bandes enherbées pour atténuer les transferts de pesticides est multiple. Tout d'abord, le couvert végétal herbacé qui les caractérise permet une atténuation efficace du ruissellement et l'interception puis la sédimentation des matières en suspension. Le réseau racinaire de ce couvert végétal dense génère également une zone de conductivité importante qui favorise l'infiltration. La végétation est aussi associée à un taux de matière organique élevé dans les premiers centimètres du sol, favorisant l'adsorption et la dégradation des pesticides.
Les principales propriétés d'une bande enherbée qui déterminent son potentiel d'interception sont sa capacité d'infiltration, sa capacité de sédimentation et sa capacité d'adsorption. La capacité d'infiltration d'une bande enherbée est fortement influencée par la structuration du sol et le niveau de nappe sous-jacent (Dosskey et al., 2011 ; Muñoz-Carpena et al., 2018 ; Lauvernet et Muñoz-Carpena, 2018 ; Fox et al., 2018) et de manière générale, le potentiel de piégeage varie fortement selon le type de sol, le climat, la végétation et les pratiques agricoles locales (Reichenberger et al., 2007 ; Poletika et al., 2009 ; Mander et Tournebize, 2015 ; Carluer et al., 2017). Ainsi, on constate des niveaux d'efficacité très disparates, allant de 0 à 99 % pour ce type de dispositif (Lacas, 2005).

1.4 Approches de modélisation pour comprendre et atténuer les transferts de pesticides

Il existe de nombreux outils de modélisation permettant de simuler les transferts et le devenir des pesticides. Ceux-ci se caractérisent par leur échelle, leur niveau de complexité et les processus qu'ils intègrent, lesquels varient selon l'objectif d'application du modèle.
D'une part, de nombreux modèles ont été développés à l'échelle de l'élément. Parmi eux, certains sont basés sur une approche conceptuelle comme c'est le cas pour PRZM (Carsel et Baldwin, 2000) qui met l'accent sur la représentation du ruissellement dans les parcelles. D'autres, sont basés sur des approches plus mécanistes. C'est entre autre le cas de MACRO (Larsbo et Jarvis, 2003), un modèle 1D dédié à la représentation du lessivage des pesti-

cides, notamment en présence de drains. Aussi suivant une approche mécaniste, VFSMOD (Muñoz-Carpena et al., 1999 ; Lauvernet et Muñoz-Carpena, 2018) permet de simuler les bandes enherbées et leur rôle d'interception du ruissellement et d'infiltration accrue de l'eau, des sédiments et des contaminants. TOXSWA (Adriaanse, 1997) quant à lui, simule l'évolution de la concentration en pesticides dans un fossé.
Si ces modèles à l'échelle locale permettent de représenter finement le pouvoir d'atténuation potentiel d'un élément du paysage vis-à-vis des transferts de pesticides, ils sont limités pour prendre en compte leur position dans le bassin et leurs interactions avec le reste du paysage. Pour cela, il faut plutôt se tourner vers la classe des modèles 3D à l'échelle du bassin versant. Cependant, qu'ils soient semi-conceptuels et semi-distribués comme SWAT (Arnold et al., 1998) ou entièrement distribués et à base physique comme CATHY (Paniconi et Putti, 1994 ; Camporese et al., 2010) ou ParFlow (Ashby et Falgout, 1996 ; Kollet et Maxwell, 2006), ces modèles sont en général basés sur un maillage fixe. Modifier un élément du paysage ou en intégrer un nouveau implique donc de reconstruire l'intégralité du maillage. Ainsi, de tels outils s'avèrent être peu adaptés pour explorer différentes configurations du paysage avec des éléments constitutifs aux géométries, propriétés et processus physiques dominants contrastés.
Pour atteindre cet objectif, on peut se tourner vers la classe émergente des modèles intégrés et modulaires. Ces outils sont élaborés à partir du couplage de plusieurs codes simulant chacun un processus physique ou un élément du paysage. Il en résulte des modèles hybrides, pouvant combiner des codes à base physique et conceptuels, aux discrétisations et complexité contrastées. Cette conception de la modélisation, souvent plébiscitée dans la littérature (Buytaert et al., 2008 ; Fatichi et al., 2016) est déjà utilisée pour construire des modèles hydrologiques (voir les plateformes JAMS, (Kralisch et Krause, 2006), WaterCAST (Argent et al., 2009), LIQUID (Branger et al., 2010) ou OpenFLUID (Fabre et al., 2010)). Les initiatives sont plus rares pour la modélisation des transferts de pesticides. On compte parmi elles les modèles construits sur la plateforme CMF (Djabelkhir et al., 2016) ou plus récemment le modèle PESHMELBA (Rouzies et al., 2019) développé dans l'équipe pollutions diffuses pour aider à identifier une configuration du paysage optimale vis-à-vis de l'atténuation des transferts de pesticides.

Parmi les nombreux outils de modélisation existants, nombreux sont ceux qui font l'objet d'une utilisation opérationnelle. Par exemple, PRZM, MACRO ou TOXSWA sont largement utilisés au niveau européen pour l'homologation. VFSMOD est lui aussi utilisé pour l'homologation par l'Environment Protection Agency (EPA) aux Etats-Unis et en France, il est intégré dans l'outil BUVARD (Catalogne et al., 2018) pour le dimensionnement de bandes enherbées. De même, le modèle PESHMELBA se positionne entre autre comme un modèle d'aide à la décision pour l'aménagement du territoire. Quelle que soit l'utilisation opérationnelle qui est considérée, celle-ci devrait toujours se faire en prenant en compte les incertitudes liées aux sorties du modèle. Si les résultats des simulations impliquent des actions des acteurs du terrain, il est notamment indispensable d'assortir ces résultats d'un indicateur sur

leur niveau de fiabilité. A minima, l'incertitude devrait donc être rigoureusement quantifiée (ce qui n'est en pratique que rarement réalisé). Au mieux, elle devrait être minimisée pour assurer une utilisation pertinente de l'outil dans un cadre d'aide à la décision.

1.5 Objectifs et structure du manuscrit

Avant d'envisager l'utilisation opérationnelle du modèle PESHMELBA (par exemple dans un contexte d'aide à la décision), il est indispensable de quantifier et de réduire les incertitudes liées à ses sorties car il s'agit d'un modèle récent, encore peu exploré. Quantification et réduction de ces incertitudes constituent ainsi les deux axes principaux de ce travail de thèse avec pour objectif de préparer son utilisation opérationnelle. L'articulation de ces deux axes ainsi que les différentes problématiques abordées pour chacun d'entre eux sont résumées dans la Figure 1.3. Ces thématiques étant traitées pour la première fois dans PESHMELBA, l'objectif est de proposer une méthodologie adaptée aux spécificités du modèle, en particulier liées à son aspect modulaire et distribué. Les développements se font sur un scénario virtuel, compromis entre simplification et réalisme, pour faciliter l'évaluation des outils utilisés tout en restant aussi représentatif que possible des défis d'une application réelle.

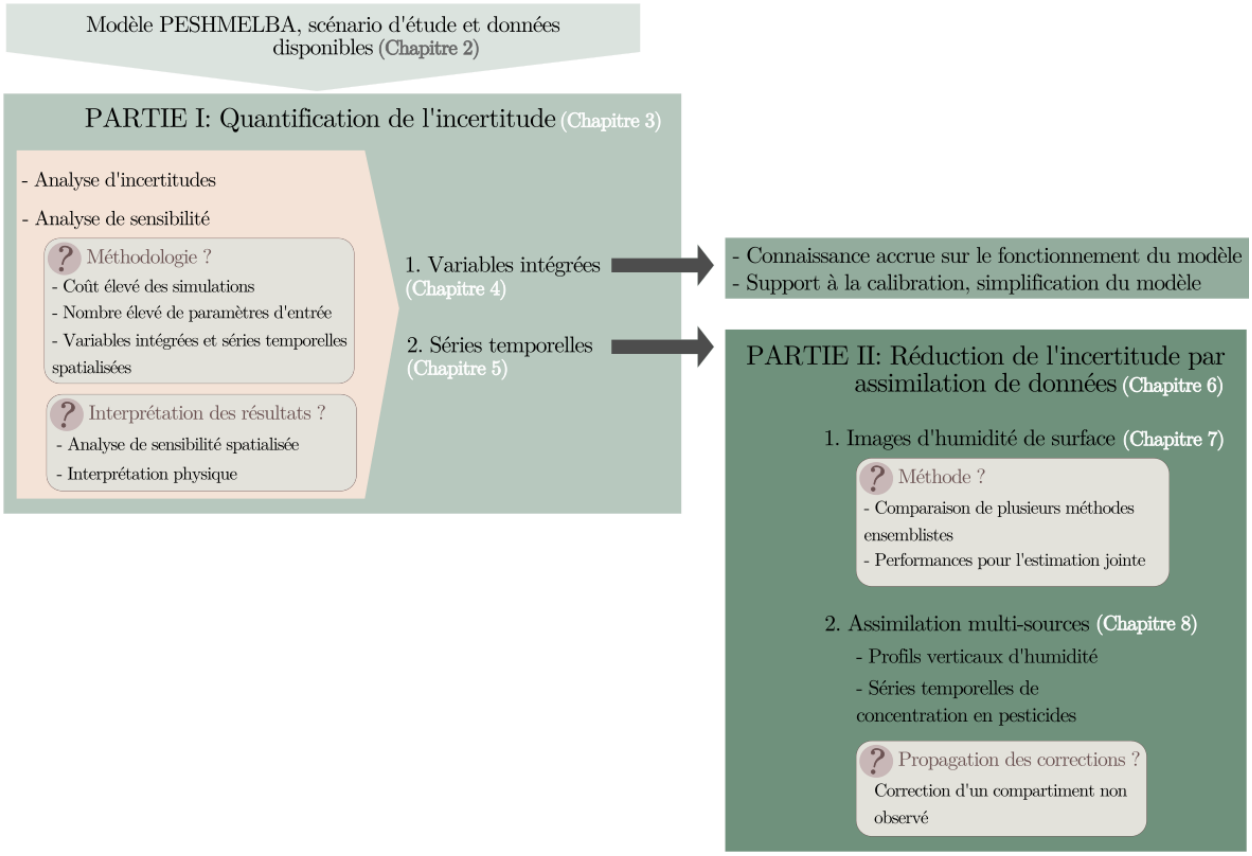


Figure 1.3 – Structure générale de la thèse : articulation des différentes parties et lien avec les chapitres du manuscrit.

Quantification d'incertitude La quantification de l'incertitude dans PESHMELBA est explorée à l'aide de deux outils complémentaires, l'analyse d'incertitude et l'analyse de sensibilité. L'analyse d'incertitude permet d'abord de quantifier le niveau d'incertitude des sorties du modèle qui est causé par des incertitude sur ses entrées. L'analyse de sensibilité permet ensuite de déterminer dans quelles proportions l'incertitude des sorties peut être attribuée à celle des entrées. Au delà de leur intérêt pour cibler les efforts de calibration, ces outils peuvent être utilisés pour mieux comprendre le fonctionnement du modèle. Pour cela, le choix de variables les plus informatives possible est crucial. Pour gagner en connaissances sur le fonctionnement de PESHMELBA, on s'intéresse particulièrement à des variables intégrées représentatives des transferts latéraux d'eau et de pesticides en surface et en subsurface. On étudie également la sensibilité de séries temporelles d'humidité du sol, en surface et en profondeur, et de concentration en pesticide puisque ces variables sont les variables cibles pour la réduction d'incertitude. En ce sens, ce travail de quantification de l'incertitude est un prérequis nécessaire pour aborder le deuxième axe de la thèse.
Un tel travail est une première pour PESHMELBA et il est plus généralement assez novateur pour les modèles modulaires simulant le devenir des pesticides. Il existe de nombreux outils, notamment pour l'analyse de sensibilité et le choix de la méthode est critique, notamment dans le cas de variables spatiales et spatiotemporelles comme c'est le cas ici. Cet aspect est largement exploré dans ce manuscript avec pour objectif d'apporter des éléments méthodologiques transposables à d'autres modèles.

Réduction de l'incertitude Dans une deuxième partie, la question de la réduction des incertitudes dans PESHMELBA est abordée par le biais des méthodes d'assimilation de données. L'assimilation de données propose un cadre mathématique rigoureux pour combiner un modèle numérique, des observations et leurs incertitudes respectives. Le problème est tout d'abord posé en définissant quelles sont les variables à estimer. En l'occurence on s'intéresse à un problème d'estimation jointe visant à estimer simultanément des variables de sortie et des paramètres d'entrée de PESHMELBA. Dans un premier temps, on fait l'hypothèse que l'on dispose seulement d'observations d'humidité de surface. Là encore, il existe plusieurs approches d'assimilation et le choix de la méthode, en fonction du problème posé, de la structure du modèle et des conclusions de l'axe 1 constitue le premier objectif de cette partie. Pour cela, on compare plusieurs méthodes ensemblistes issues de l'approche bayésienne de l'assimilation.
Dans un second temps, on suppose que l'on dispose de plusieurs types d'observations : images d'humidité de surface, profils verticaux d'humidité du sol et séries temporelles de concentration en pesticides dans la rivière. L'enjeu de cette partie consiste alors à identifier quel type d'observation permet de corriger quelles variables et quels paramètres. La question se pose d'autant plus que PESHMELBA est un modèle modulaire et multi-compartiments. On cherche donc à évaluer dans quelle mesure il est possible de propager les corrections d'un compartiment aux autres compartiments du modèle. Pour cela, les possibilités de l'assimilation multi-sources sont largement explorées.

Structure du manuscrit Le Chapitre 2 présente le modèle PESHMELBA, le cas d'étude mis en oeuvre dans la thèse ainsi que les différentes sources de données dont on dispose pour l'assimilation de données.
La Partie I, composée des Chapitres 3, 4 et 5 regroupe les résultats relatifs à la quantification d'incertitude dans PESHMELBA. Les outils utilisés et la méthodologie mise en oeuvre pour l'analyse d'incertitude et l'analyse de sensibilité sont d'abord présentés (Chapitre 3). Ensuite, le Chapitre 4 présente les résultats relatifs aux variables intégrées alors que le Chapitre 5 regroupe les résultats pour les séries temporelles.
La Partie II (Chapitres 6, 7 et 8) regroupe ensuite les résultats relatifs à la réduction d'incertitude par assimilation de données dans PESHMELBA. Là encore, les méthodes mises en oeuvre dans la thèse sont d'abord présentées dans le Chapitre 6. Ensuite, sont présentés les résultats relatifs à l'assimilation d'images d'humidité de surface (Chapitre 7) puis à l'assimilation multi-sources (Chapitre 8).
Finalement, le Chapitre 9 regroupe des éléments de conclusion sur l'intérêt et les limites de ce travail de thèse ainsi que des perspectives envisageables.

Chapitre 2

Comprendre et limiter les pollutions diffuses par les pesticides : apport de la modélisation et des données expérimentales

Sommaire

2.1 Le modèle PESHMELBA

2.1.1 Présentation générale

Le modèle PESHMELBA (PESticides et Hydrologie: Modélisation à l'échELle du BAssin versant, Rouzies et al. 2019) est développé dans l'équipe pollutions diffuses depuis 2016 pour simuler les transferts et le devenir des pesticides à l'échelle de petits bassins versants agricoles (quelques dizaines de km^2). Les motivations pour le développement d'un tel modèle sont multiples. L'objectif est d'abord de fournir un outil permettant de comprendre le devenir des pesticides en représentant de manière physique et spatialisée les processus impliqués dans les différents éléments qui composent un bassin versant agricole. Utilisé dans un contexte opérationnel, PESHMELBA peut ainsi participer à identifier des zones particulièrement exposées aux contaminants agricoles, comprendre pourquoi elles le sont et ainsi servir de point de départ pour réfléchir l'implantation de solutions correctives (zones tampons en particulier). D'autre part, la structure de PESHMELBA met l'accent sur une représentation explicite de l'organisation du paysage. L'objectif est ainsi de quantifier le rôle, positif ou négatif, des différents éléments du paysage sur les transferts de pesticides. Le but plus opérationnel est ensuite de simuler plusieurs scénarios d'aménagement du territoire et de les comparer en termes d'efficacité vis-à-vis de l'atténuation des transferts de pesticides.

La démarche de modélisation adoptée dans PESHMELBA est la suivante : tout d'abord, pour modéliser un paysage, un maillage hétérogène qui représente explicitement ses éléments constitutifs est généré. L'outil permettant un tel découpage (geoMELBA) ainsi que le maillage résultant sont décrits dans la Section 2.1.2. Chaque élément du découpage est discrétisé selon sa nature et un ensemble de processus physiques le caractérisant est ensuite simulé. On représente également les processus de transfert entre les différents éléments du paysage. Les processus représentés sont décrits dans la Section 2.1.3. Chacun d'entre eux est intégré au modèle sous la forme d'un code indépendant permettant la réutilisation de modèles pré-existants. Il en résulte une structure finale composée de codes basés sur des discrétisations et des complexités hétérogènes. Finalement, pour coupler ces unités de code et aboutir à une représentation complète du bassin versant, on se base sur une gestion différenciée des pas de temps décrite en Section 2.1.4 et un outil de couplage adapté décrit en Section 2.1.4. La version actuelle de PESHMELBA permet de représenter le fonctionnement de parcelles, bandes enherbées, fossés, haies, haies sur talus et rivière. Dans le cadre de la thèse, on se base sur une version simplifiée intégrant seulement des parcelles, des bandes enherbées et des rivières.

2.1.2 Structure du maillage : l'outil geoMELBA

Contrairement à des modèles comme ParFlow ou CATHY qui sont basés sur un maillage régulier, le maillage de PESHMELBA est constitué d'éléments irréguliers, surfaciques ou

linéaires qui représentent les différentes structures composant un paysage. Chacun de ces éléments élémentaires est caractérisé par un type (parcelle, bande enherbée ou rivière dans notre cas), un type de sol unique et potentiellement un itinéraire cultural et calendrier de pratiques. Ces éléments sont ainsi homogènes en termes de processus physiques. On les appelle **Unités Homogènes** (UH) si ce sont des éléments surfaciques et **Tronçons Elémentaires** (TE) si ce sont des éléments linéaires. Ainsi, les parcelles et les bandes enherbées sont représentées par des UH alors que les rivières, les fossés et les haies sont représentés par des TE.

Le maillage de PESHMELBA est obtenu avec l'outil geoMELBA (Grillot et al., 2022) en faisant l'intersection du parcellaire cadastral, d'une carte des sols et des réseaux de linéaires comme représenté sur la Figure 2.1. Il est composé des unités homogènes et des tronçons élémentaires qui composent le bassin versant ainsi que des connexions entre ces éléments. Ces connexions sont unidirectionnelles et se basent sur les différences d'altitudes entre centroïdes. Ainsi, un élément ne peut recevoir de flux que des éléments voisins dont les centroïdes sont situés plus haut que son propre centroïde. Chaque UH peut être la source et la cible de plusieurs connexions latérales avec d'autres UH ou TE. En longitudinal, les TE se structurent en réseaux.

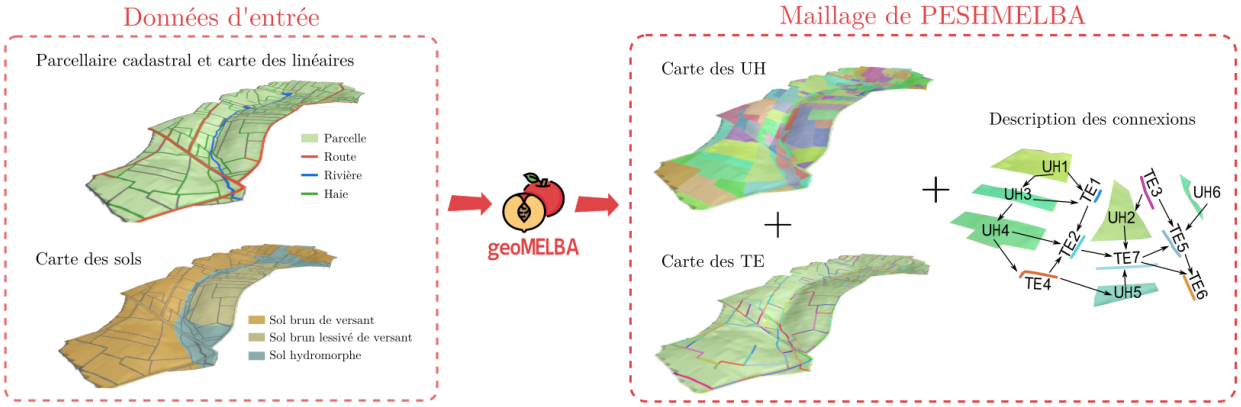


Figure 2.1 – Principe de découpage d'un paysage par l'outil geoMELBA.

Ainsi, grâce à la définition des connexions, le maillage de PESHMELBA permet de représenter facilement le rôle d'interception totale ou partielle des flux par des éléments linéaires. Par exemple, une connexion latérale entre deux parcelles peut être remplacée par une connexion latérale entre la parcelle et un tronçon de fossé puis une connexion longitudinale vers un autre tronçon de fossé pour rendre compte de l'effet de court-circuit et de redirection de ce dernier.

2.1.3 Processus physiques représentés

Le modèle PESHMELBA est un modèle orienté processus, spatialisé, à base physique. Il simule les processus dominant le devenir des pesticides à l'échelle d'un bassin versant sur

chacune des unités homogènes et des tronçons élémentaires du maillage. PESHMELBA se limite cependant à la représentation des flux d'eau et de contaminants en surface et en subsurface du sol. Le compartiment aérien n'est pas représenté et le modèle n'inclut donc pas de représentation des phénomènes de dérive et de volatilisation. Les processus en jeu entre et dans les éléments dépendent de leur nature, tout comme le type de discrétisation adopté.

La Figure 2.2 illustre les processus intégrés dans la version de PESHMELBA utilisée dans la thèse ainsi que les discrétisations associées. Dans ces travaux, chaque parcelle ou bande enherbée est représentée par une unique colonne de sol de 4 m de haut divisée en cellules numériques d'épaisseur allant de 0.5 cm en surface jusqu'à 1 m en profondeur. Chaque tronçon de rivière est représenté par un réservoir unique à section trapézoïdale. Les processus intégrés sont :

- l'infiltration verticale dans les parcelles et les bandes enherbées ;
- l'extraction racinaire dans les parcelles et les bandes enherbées ;
- le ruissellement entre les parcelles, les bandes enherbées et son interception par la rivière ;
- les échanges latéraux saturés de subsurface entre les parcelles et les bandes enherbées ;
- les échanges latéraux saturés entre la nappe et la rivière ;
- les écoulements dans la rivière ;
- l'advection des pesticides avec les flux d'eau représentés ;
- l'adsorption et la dégradation des pesticides sur tous les éléments.

Il est important de noter que de manière générale, et dans ces travaux de thèse, on considère que les mêmes forçages climatiques (précipitations et évapotranspiration potentielle) sont appliqués sur tous les éléments du bassin versant. Cette hypothèse se justifie par la taille limitée des bassins versants modélisés par PESHMELBA (inférieure à la dizaine de km^2).

Les paragraphes suivants détaillent la représentation de chacun de ces processus. Les équations et paramètres d'entrée sont explicités afin de fournir une vue d'ensemble des variables et paramètres d'entrée d'intérêt dans ces travaux de thèse.

Définitions des principales propriétés hydrodynamiques d'un sol

Dans ces travaux, les principales variables qui représentent le comportement hydrodynamique d'un sol et qui interviennent donc pour modéliser les processus physiques intégrés dans PESHMELBA sont la teneur en eau, le potentiel matriciel et la conductivité hydraulique.

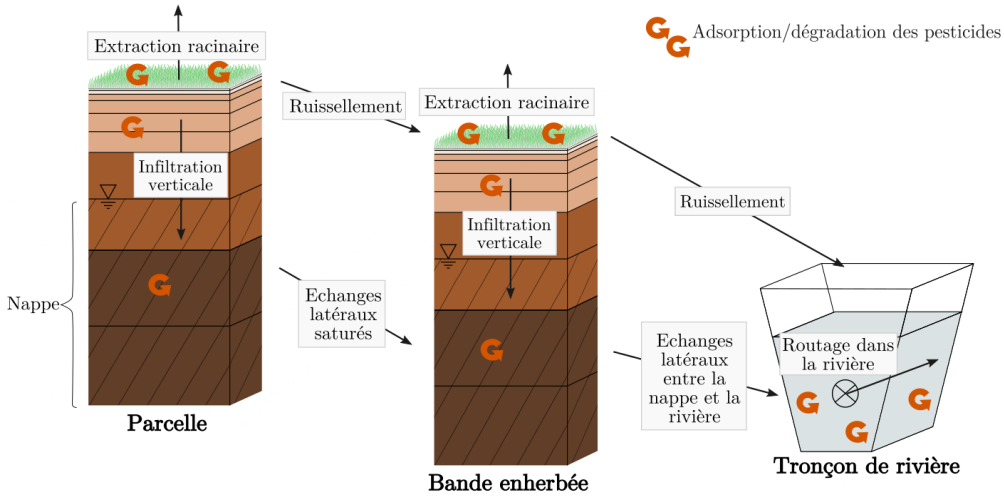


Figure 2.2 – Illustration des processus physiques et des discrétisations considérées dans la version de PESHMELBA utilisée dans la thèse.

Teneur en eau La teneur en eau θ [L^3L^{-3}] (aussi appelée *humidité*) correspond au rapport du volume occupé par de l'eau V_w [L^3] sur le volume total V_T [L^3] des 3 phases (liquide, solide, gazeuse) du sol :

$$\theta = \frac{V_w}{V_T} \tag{2.1}$$

Le volume d'eau maximal que peut contenir un volume de sol définit ainsi la teneur en eau à saturation θ_s [L^3L^{-3}]. De même, la teneur en eau résiduelle θ_r définit le volume d'eau minimal que peut contenir un sol. La teneur en eau est aussi communément remplacée par son expression normalisée, la saturation S [-] :

$$S = \frac{\theta - \theta_r}{\theta_s - \theta_r} \tag{2.2}$$

Potentiel matriciel Le potentiel matriciel h [L] correspond à la pression nécessaire pour extraire l'eau du sol. Il est positif lorsque le milieu est saturé et négatif sinon. On en déduit également le potentiel total (ou *charge hydraulique*) H [L] qui s'exprime en fonction de la profondeur z [L] et du potentiel matriciel :

$$H = z + h \tag{2.3}$$

Conductivité hydraulique La conductivité hydraulique K [LT^{-1}] correspond à l'aptitude du sol à se laisser traverser par de l'eau. Elle est pour la circulation de l'eau dans un sol ce que représente la conductance électrique pour la circulation des électrons dans un corps conducteur et dépend fortement du degré de saturation du sol (Bruand et Coquet, 2005).

Infiltration verticale

L'infiltration verticale dans les parcelles et les bandes enherbées est simulée à partir d'une résolution 1D de l'équation de Richards (Richards, 1931) avec un terme puits/source qui permet d'intégrer l'impact de l'extraction racinaire et des échanges latéraux entre UH :

$$\frac{\partial \theta}{\partial t} = \frac{\partial}{\partial z}\left[K_v(h)(\frac{\partial h}{\partial z} - 1)\right] + R \tag{2.4}$$

où θ est la teneur en eau [L^3L^{-3}], h est le potentiel matriciel [L], K_v est la conductivité hydraulique verticale [LT^{-1}] et R le terme puits/source [$L^3L^{-3}T^{-1}$]. L'équation de Richards permet ainsi de mettre à jour les profils verticaux d'humidité et de calculer les hauteurs de ponding, correspondant aux hauteurs d'eau surfacique non infiltrée ou exfiltrée pendant un pas de temps. La résolution utilisée dans PESHMELBA est basée sur le schéma numérique proposé par Ross (2003).

Pour résoudre cette équation, il est nécessaire d'expliciter les relations qui lient les propriétés hydrodynamiques du sol. Pour se faire, on définit communément $S = f(h)$, la courbe de rétention qui relie saturation (ou teneur en eau) et potentiel matriciel et $K_v = f(h)$ la courbe de conductivité qui relie la conductivité hydraulique au potentiel matriciel.

La **courbe de rétention** est décrite ici avec la relation de Van Genuchten :

$$S(h) = (1 + (\alpha|h|)^n)^{-m} \tag{2.5}$$

où α [L^{-1}] est un paramètre empirique lié à la pression d'entrée d'air, n [-] est un paramètre de forme lié à la distribution de la taille des pores et m [-] est un paramètre de forme, communément lié à n par la relation $m = 1 - \frac{1}{n}$ [1]. Dans PESHMELBA, la formulation de van Genuchten est remplacée par une expression quadratique près de la saturation pour éviter les difficultés numériques (Ross, 2006).

La **courbe de conductivité** est décrite par la formulation de Schaap-Van Genuchten (Schaap et van Genuchten, 2006) laquelle se base sur la formulation de Mualem-Van Genuchten (Mualem, 1976) avec une modification près de la saturation. Une telle modification permet non seulement d'éliminer les pentes trop fortes mais aussi d'intégrer l'influence des écoulements préférentiels dans les mésopores et les macropores. Pour cela, la gamme de potentiel matriciel est divisée en trois domaines définissant différents régimes d'écoulements :

- un régime dominé par les écoulements dans la matrice de sol pour des potentiels matriciels inférieurs à $hmac_2$;
- un régime dominé par les écoulements dans les mésopores de taille intermédiaire pour

1. Dans la suite, on utilisera le paramètre n ou le paramètre $mn = n - 1$

des potentiels matriciels compris entre $hmac_2$ et $hmac_1$;

- un régime dominé par les écoulements dans les macropores pour les potentiels matriciels entre $hmac_1$ et zéro.

L'expression de la conductivité dans chacun de ces domaines est résumée dans la Figure 2.3. Les potentiels h_{mac_1} et h_{mac_2} qui partitionnent les différents régimes d'écoulements sont fixées à h_{mac_1}=−4 cm et h_{mac_2}=−40 cm. R_{mac_1} [-] est fixé à 0.25. Les paramètres empiriques n, m et α sont les mêmes que ceux de la courbe de rétention (Eq. 2.5) et dépendent du type de sol. p [-] est un paramètre empirique lié à l'interaction des pores entre elles. K_x et K_s représentent respectivement la conductivité hydraulique à saturation de la matrice et la conductivité hydraulique à saturation totale [LT^{-1}].

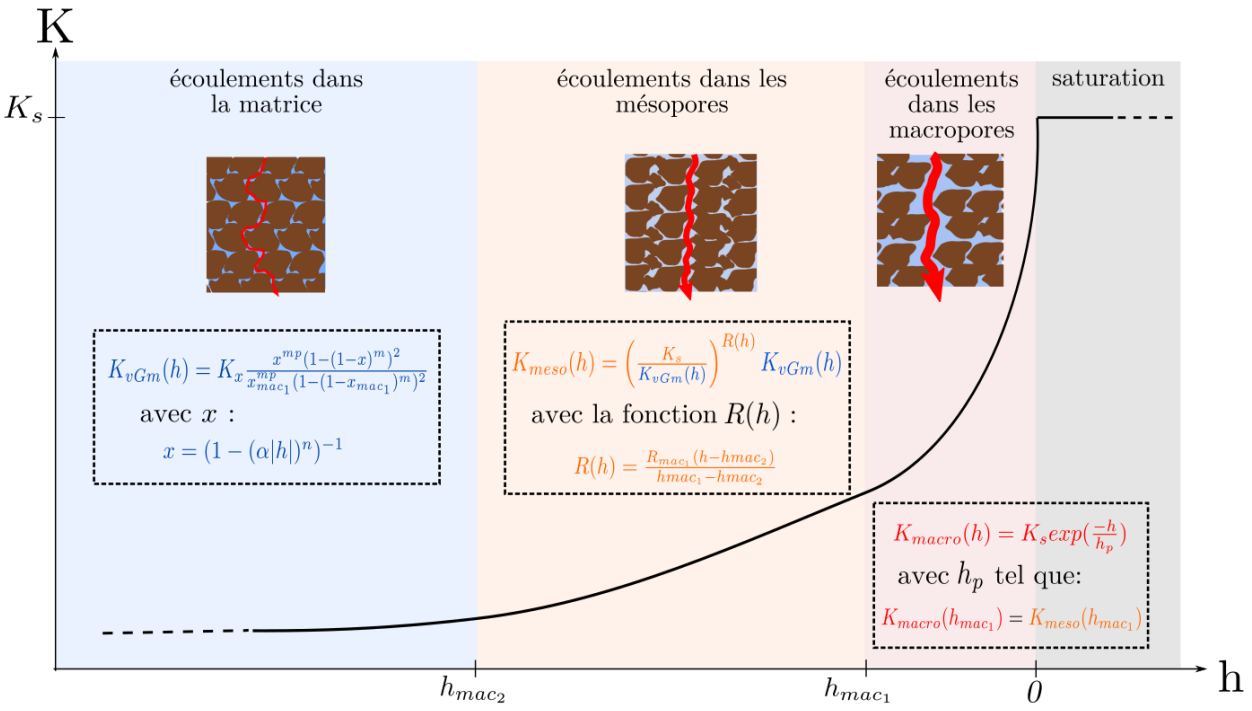


Figure 2.3 – Description de la courbe de conductivité utilisée dans PESHMELBA pour la résolution de l'équation de Richards selon la formulation de Schaap et van Genuchten (2006).

Processus liés à la végétation

Les processus liés à la végétation incluent dans cette version de PESHMELBA sont le calcul de l'indice de surface foliaire, de la transpiration et de l'extraction racinaire. La représentation de ces processus est illustrée sur la Figure 2.4 et reprend largement les travaux de Varado (2004).

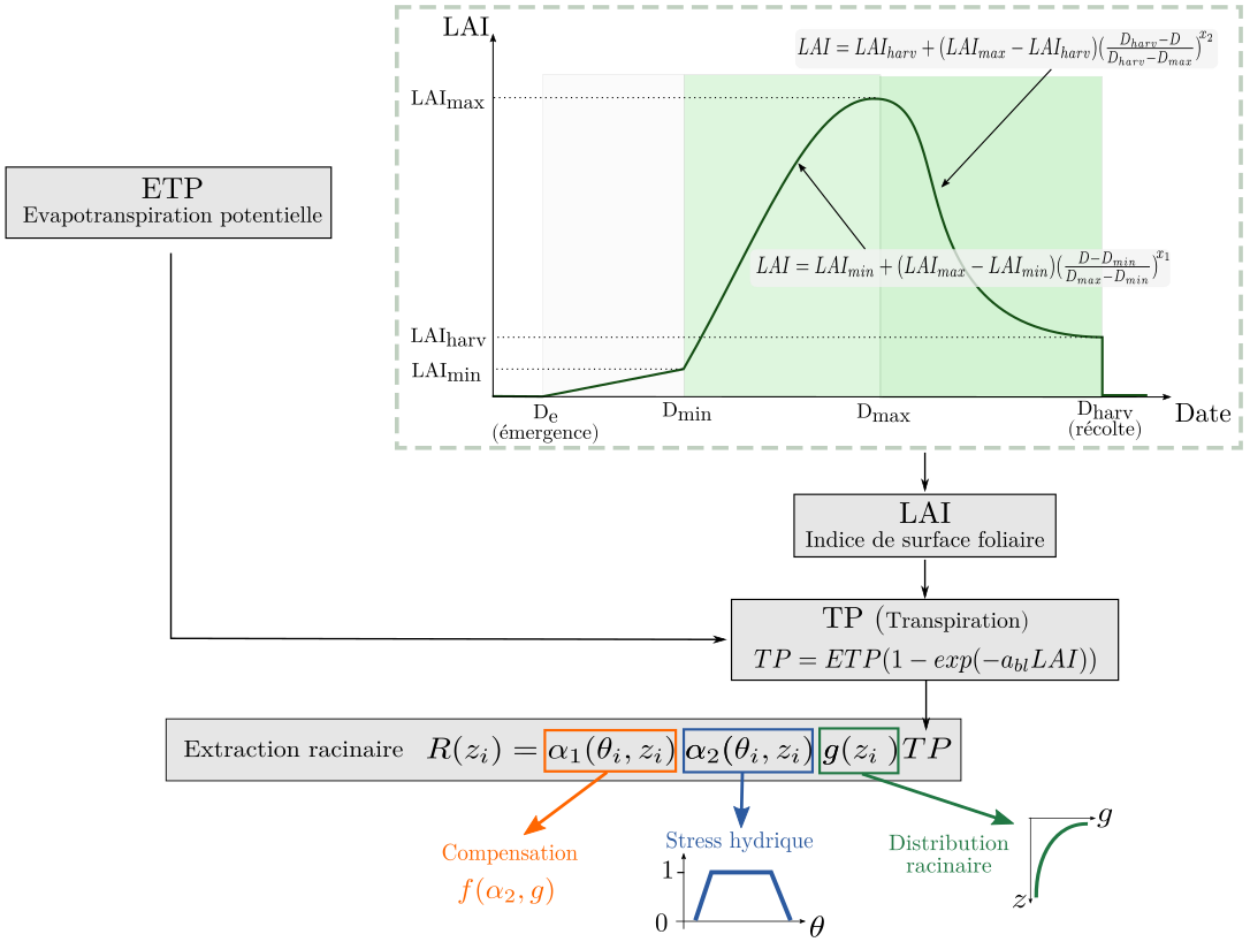


Figure 2.4 – Modélisation des processus liés à la végétation dans PESHMELBA : évolution du LAI basée sur les équations de Larsbo et Jarvis (2003), calcul de la transpiration et calcul de l'extraction racinaire selon le modèle de Li et al. (2001).

Transpiration A chaque instant, la transpiration potentielle TP est calculée à partir d'un partitionnement de l'évapotranspiration potentielle (ETP) selon une loi de Beer-Lambert :

$$TP = ETP(1 - exp(-a_{bl}LAI)) \tag{2.6}$$

avec ETP [LT^{-1}] l'évapotranspiration potentielle, a_{bl} [-] un coefficient d'extinction et LAI [-] la valeur de l'indice de surface foliaire [2] à l'instant donné. Conformément à Varado (2004), le coefficient d'extinction a_{bl} est fixé à 0.5.

Calcul du LAI Le LAI qui intervient dans le calcul de la transpiration potentielle est supposé constant sur les prairies enherbées (FOCUS, 2001). Pour les cultures annuelles, on modélise son évolution au cours du temps suivant le modèle proposé dans MACRO (Larsbo

2. L'indice de surface foliaire correspond au ratio de la surface totale supérieure des feuilles sur la surface du sol sur laquelle la végétation se développe.

et JARVIS, 2003). Il suppose que le sol est nu avant l'émergence et après la récolte. Une phase de développement linéaire est modélisée entre la date d'émergence D_e et la date D_{min} où le LAI atteint sa valeur minimale. Le pic de croissance est ensuite décrit en faisant intervenir x_1 et x_2 deux paramètres de forme empiriques et les dates et valeurs de LAI associées au stade de développement maximal et au moment de la récolte. Dans cette application, et comme proposé dans FOCUS (2001), les paramètres x_1 et x_2 sont fixés à 2 et 0.2 respectivement.

Extraction racinaire La transpiration potentielle intervient ensuite dans le calcul de l'extraction racinaire qui constitue un terme puits dans l'équation de Richard (Eq. 2.4). Elle est simulée à partir du modèle de LI et al. (2001) qui la décrit comme fonction de la transpiration potentielle TP et d'une fonction empirique comprise entre 0 et 1. Cette dernière intègre les phénomènes de compensation α_1, de stress hydrique α_2 et la distribution racinaire g :

$$R(z_i) = \alpha_1(\theta_i, z_i)\alpha_2(\theta_i, z_i)g(z_i)TP \tag{2.7}$$

avec θ_i et z_i la teneur en eau et la profondeur de la cellule numérique i.

La distribution racinaire utilisée est celle proposée par LI et al. (2001). Elle s'exprime en fonction de Z_r [L] la profondeur maximale racinaire et F_{10} [-], la fraction de densité racinaire présente dans les 10% supérieurs de la zone racinaire (LI et al., 1999). La fonction de stress hydrique α_2 est basée sur le modèle de FEDDES et al. (1978) qui tient compte de la diminution d'extraction racinaire lorsqu'il y a stress hydrique ou étouffement de la plante par excès d'eau. La fonction de compensation α_1 s'écrit en fonction de g et α_2 et permet de considérer l'extraction de l'eau dans les couches profondes et humides lorsque la surface est sèche.

Ruissellement

Le ruissellement est représenté en utilisant l'approximation de l'onde cinématique qui consiste à assimiler l'écoulement à une succession d'états permanents uniformes. Pour faciliter le couplage avec la subsurface, l'intégration du ruissellement est faite par une approche type bilan de masse qui permet de calculer des volumes ruisselés sortants et entrants par UH sans rediscrétiser la surface. Le volume ruisselé sortant d'une UH est calculé à partir de h_p, la hauteur de ponding disponible en surface en utilisant la formule empirique de Manning-Strickler reliant hauteur d'eau et vitesse moyenne :

$$V = \frac{h_p^{\frac{2}{3}}}{Manning}\sqrt{S_0} \tag{2.8}$$

avec V [LT^{-1}] la vitesse de l'écoulement, h_{pond} [L] la hauteur de ponding fournie par le module d'infiltration verticale, $Manning$ [LT$^{\frac{-1}{3}}$] le paramètre de Manning-Strickler caractérisant la rugosité du sol et S_0 [LL^{-1}] la pente.

La formule de Manning-Strickler permet de calculer un volume d'eau sortant de chaque UH à chaque pas de temps. Ce volume est ensuite réparti entre les éléments situés à l'aval de l'UH émettrice (que ce soient des UH ou TE) en fonction des gradients d'altitude et des longueurs d'interface avec chaque élément. Le calcul du ruissellement n'est activé que lorsque la hauteur d'eau disponible en surface de la parcelle dépasse une hauteur de ponding limite *hpond* [L].

Echanges latéraux de subsurface entre UH

Dans PESHMELBA, les transferts latéraux non saturés sont supposés négligeables par rapport aux transferts saturés. Ainsi, seuls les échanges latéraux saturés sont simulés en résolvant l'équation de Darcy (Darcy, 1857). A chaque pas de temps, le flux saturé $Q_{1\to2}$ [L^3T^{-1}] entre deux UH, $UH1$ et $UH2$ est calculé comme suit :

$$Q_{1\to2} = K_{int}\, S_{int}\, \nabla H_{1\to2} \tag{2.9}$$

avec K_{int} [LT^{-1}] la conductivité hydraulique horizontale à l'interface entre les deux UH, S_{int} [L^2] la surface d'échange à l'interface et $\nabla H_{1\to2} = \frac{H_1 - H_2}{d_{1-2}}$ [-] le gradient de charge entre les deux UH calculé à partir de H_1 et H_2 les charges respectives dans UH1 et UH2 et d_{1-2} la distance entre leurs centroïdes. La relation de Darcy permet de calculer un flux scalaire qui est ensuite réparti entre les cellules numériques de l'UH amont et aval en fonction de leur épaisseur et de leur conductivité hydraulique. Le lecteur peut se référer à Rouzies et al. (2019) pour plus de détails sur le calcul du flux latéral et sa répartition entre les UH amont et aval.

Echanges latéraux nappe-rivière

Les échanges en subsurface entre la nappe d'une UH et la rivière sont simulés en utilisant la relation de Miles (Miles, 1985) :

$$Q_{nappe-riviere} = C_m K s_{river} \Delta H \tag{2.10}$$

avec $\Delta H = H_{riv} - H_{UH}$ la perte de charge entre la rivière et la nappe, Ks_{river} [LT^{-1}] la conductivité hydraulique du fond de la rivière et C_m le coefficient de Miles:

$$C_m = \frac{5[0.25(w_s + w) + H_{riv} + s]}{D_i + h + s} \tag{2.11}$$

où w_s et w sont respectivement la largeur au miroir [L] et la largeur au radier [L] du tronçon de rivière, H_{riv} la charge dans la rivière (correspondant au tirant d'eau) [L], D_i est la distance entre le fond de la rivière et l'imperméable de la zone saturée [L] et s la distance de suintement [L]. Ces différentes grandeurs sont illustrées sur la Figure 2.5.

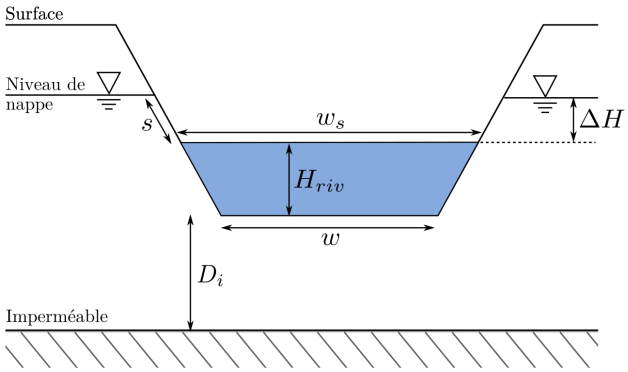


Figure 2.5 – Description des grandeurs géométriques impliquées dans le calcul de la relation de Miles (Miles, 1985).

Ecoulements dans la rivière

Comme pour le ruissellement de surface, le calcul du routage de l'eau et des pesticides dans le réseau de rivières se base sur l'approximation de l'onde cinématique. La vitesse de l'écoulement est calculée dans chaque TE en fonction du rayon hydraulique à l'aide de la formule de Manning (Eq. 2.8). A partir de l'équation de conservation de la masse, les hauteur d'eau sont ensuite mises à jour dans l'intégralité du réseau en intégrant des termes puits-sources correspondant par exemple aux échanges nappe-rivière, à l'interception du ruissellement ou aux précipitations. Le routage dans le réseau ne s'active que lorsque le tirant d'eau dans un tronçon est supérieur à une hauteur de ponding limite *hpond* [L].

Advection, adsorption et dégradation des pesticides

Les sections précédentes présentent les processus de transferts d'eau dans le bassin versant. Les transferts de pesticides sont représentés conjointement et hormis le module d'infiltration verticale qui permet de résoudre l'équation d'advection-dispersion, les autres modules considèrent que le transport est exclusivement convectif, n'incluant pas de composante de dispersion ou de diffusion.

Les modules relatifs aux compartiments de surface et de subsurface du sol ainsi qu'aux tronçons de rivière incluent également une représentation des processus de transformation (adsorption et dégradation) des pesticides. L'adsorption est représentée par un isotherme d'équilibre qui représente la variation de la masse adsorbée sur le solide en fonction de la concentration à l'équilibre dans la phase liquide. Dans ces travaux, on suppose que l'équilibre d'adsorption s'établit instantanément et qu'il s'agit d'un mécanisme totalement réversible. L'équilibre entre les pesticides dissous et adsorbés s'exprime grâce à un isotherme de Freundlich (Freundlich, 1909) :

$$s = \theta c_w + bdK_d c_w^{n_f} \tag{2.12}$$

où s est la masse de pesticides par volume de sol [ML^{-3}], θ est la teneur en eau [-], c_w est la

concentration en pesticides dans la phase dissoute [ML^{-3}], *bd* est la densité volumique du sol [ML^{-3}], K_d le coefficient d'adsorption [$L^{3n}M^{-n}$] et n_f le coefficient de Freundlich. Le coefficient d'adsorption dépendant fortement de la teneur en carbone organique du susbstrat *moc* [%] et on utilise souvent une expression normalisée de ce dernier :

$$K_{oc} = \frac{100 K_d}{moc} \tag{2.13}$$

L'adsorption est calculée dans toute la colonne de sol. En surface, elle a lieu dans une couche de mélange entre les pesticides du sol et ceux de la lame ruisselée. Son épaisseur est renseignée au travers du paramètre *adsorpthick* [L].

D'autre part, l'ensemble des processus de dégradation, biotiques et abiotiques sont modélisés de manière simplifiée en représentant seulement la vitesse de dégradation via une loi cinétique de premier ordre :

$$c_w(t) = c_{w_0} e^{-tk} \tag{2.14}$$

où c_{w_0} est la concentration en pesticides à l'instant initial [ML^{-3}] et k la constante de vitesse de premier ordre [T^{-1}] liée au temps de demi-vie $DT50$ [T] par la relation $k = \frac{ln(2)}{DT50}$.

Conclusion : paramètres d'entrée considérés dans la thèse

Pour conclure sur la représentation des différents processus physiques dans PESHMELBA, l'ensemble des paramètres d'entrée du modèle qui sont considérés dans la suite de ces travaux (ce qui n'inclut pas les paramètres géométriques et certains paramètres de forme) ainsi que les notations utilisées sont regroupés dans le Tableau 2.1.

Paramètre [unité]	Description
Infiltration verticale/échanges latéraux de subsurface	
θ_s ou *thetas* [$cm^3.cm^{-3}$]	Teneur en eau à saturation
θ_r ou *thetar* [$cm^3.cm^{-3}$]	Teneur en eau résiduelle
K_s [$m.s^{-1}$]	Conductivité hydraulique à saturation totale
α ou *alpha* [m^{-1}]	Inverse de la pression d'entrée d'air
mn [-]	Paramètre de forme dans la courbe de rétention de van Genuchten
K_x [$m.s^{-1}$]	Conductivité hydraulique à saturation de la matrice
p [-]	Paramètre empirique de connectivité des pores
Croissance de la végétation	
LAI [-]	Indice de surface foliaire si LAI supposé constant
LAI_{max} [-]	Indice de surface foliaire maximal si on considère un LAI dynamique

LAI_{min} [-]	Indice de surface foliaire minimal si on considère un LAI dynamique
LAI_{harv} [-]	Indices de surface foliaire à la récolte sion considère un LAI dynamique
D_e [-]	Numéro du jour d'émergence si on considère un LAI dynamique
D_{min} [-]	Numéro du jour correspondant au stade de développement intermédiaire du couvert végétal si on considère un LAI dynamique
D_{max} [-]	Numéro du jour correspondant au stade de développement maximal du couvert végétal si on considère un LAI dynamique
D_{harv} [-]	Numéro du jour de récolte si on considère un LAI dynamique
	Extraction racinaire
Z_r [m]	Profondeur racinaire maximale
F_{10}	Densité racinaire dans les 10% supérieurs de la zone racinaire
	Ruissellement
$Manning$ [m.s$^{\frac{-1}{3}}$]	Rugosité de Manning-Strickler
$hpond$ [m]	Hauteur d'eau minimale avant activation du ruissellement (ponding)
	Echanges nappe-rivière
Ks_{river} [m.s^{-1}]	Conductivité hydraulique du fond de la rivière
D_i [m]	Distance entre le fond de la rivière et l'imperméable de la zone saturée
	Routage dans la rivière
$hpond$ [m]	Hauteur d'eau minimale avant activation du routage dans la rivière (ponding)
	Advection et transformation des pesticides
$adsorpthick$ [m]	Epaisseur de la couche de mélange
K_{oc} [mL.g^{-1}]	Coefficient d'adsorption
moc [%]	Teneur en matière organique du sol
$DT50$ [jours]	Temps de demi-vie

Tableau 2.1 – Variables et paramètres d'entrée du modèle PESHMELBA utilisés pour simuler les principaux processus physiques.

2.1.4 Gestion du couplage

La gestion du temps dans PESHMELBA

Compte tenu de la différence de temps caractéristique des dynamiques de surface et de subsurface, PESHMELBA est basé sur deux pas de temps distincts : un pas de temps de base du coupleur **dt_PALM** qui est utilisé pour simuler les processus de subsurface (infiltration et échanges latéraux saturés) et un sous pas de temps **dt_RO** utilisé pour les processus de surface (ruissellement et routage dans le réseau). Pour représenter plus finement la dynamique de l'eau et des pesticides en période de pluie ou de ressuyage de nappe (période "humide"), le pas de temps de base dt_PALM peut être raffiné sur ces périodes. Le sous pas de temps de surface dt_RO quant à lui ne varie pas. La Figure 2.6 (gauche) illustre l'adaptation du pas de temps dt_PALM en fonction du hyétogramme d'entrée. Dans ces travaux, on choisit pour dt_PALM, une valeur de 1 h en période sèche et 30 min en période humide. Le sous pas de temps dt_RO est quant à lui fixé à 6 min et la période de ressuyage de nappe est fixée à 3 h.

L'utilisation conjointe de deux pas de temps doit être prise en compte pour coupler temporellement les différents modules car cela implique des contraintes supplémentaires. Le couplage entre infiltration et ruissellement illustre de telles contraintes : les hauteurs d'eau en surface utilisées pour calculer le ruissellement au sous pas de temps dt_RO correspondent aux hauteurs de ponding disponibles après calcul de l'infiltration au pas de temps dt_PALM. En contre-partie, la mise à jour des hauteurs d'eau en surface tous les dt_RO impose de nouvelles conditions limites surfaciques pour le module d'infiltration qui ne s'exécute qu'au pas de temps dt_PALM.
Pour gérer cette difficulté, le couplage entre surface et subsurface a lieu au pas de temps dt_PALM. Pour faciliter la synchronisation et limiter les rétroactions permettant ainsi à la simulation d'avancer dans le temps, les modules sont lancés de manière séquentielle. La Figure 2.6 (droite) résume l'ordre de lancement des modules, les principaux échanges entre variables internes et les pas de temps utilisés. Les modules d'infiltration et d'extraction racinaire sont d'abord lancés simultanément sur toutes les UH, entre t et t+dt_PALM. A la fin du pas de temps dt_PALM, on dispose des hauteurs de ponding en surface (eau non infiltrée ou exfiltrée) sur chaque UH. Ces variables sont ensuite utilisées pour calculer le ruissellement entre t et t+dt_PALM. Comme le ruissellement est calculé au sous pas de temps dt_RO, le module est appelé plusieurs fois. A la fin de chaque sous pas de temps, le volume de ruissellement intercepté par chaque tronçon de rivière est calculé et stocké. Le routage dans la rivière est alors lancé entre t et t+dt_PALM, plusieurs fois, au sous pas de temps dt_RO. A chaque sous pas de temps dt_RO, le niveau d'eau est mis à jour dans chaque tronçon en prenant en compte le routage dans le réseau et en intégrant le volume provenant des échanges nappes-rivière calculé au pas de temps précédent et le volume de ruissellement intercepté. Enfin, à t+dt_PALM les flux latéraux entre UH et entre UH et rivières sont calculés et ces flux seront appliqués au pas de temps suivant.

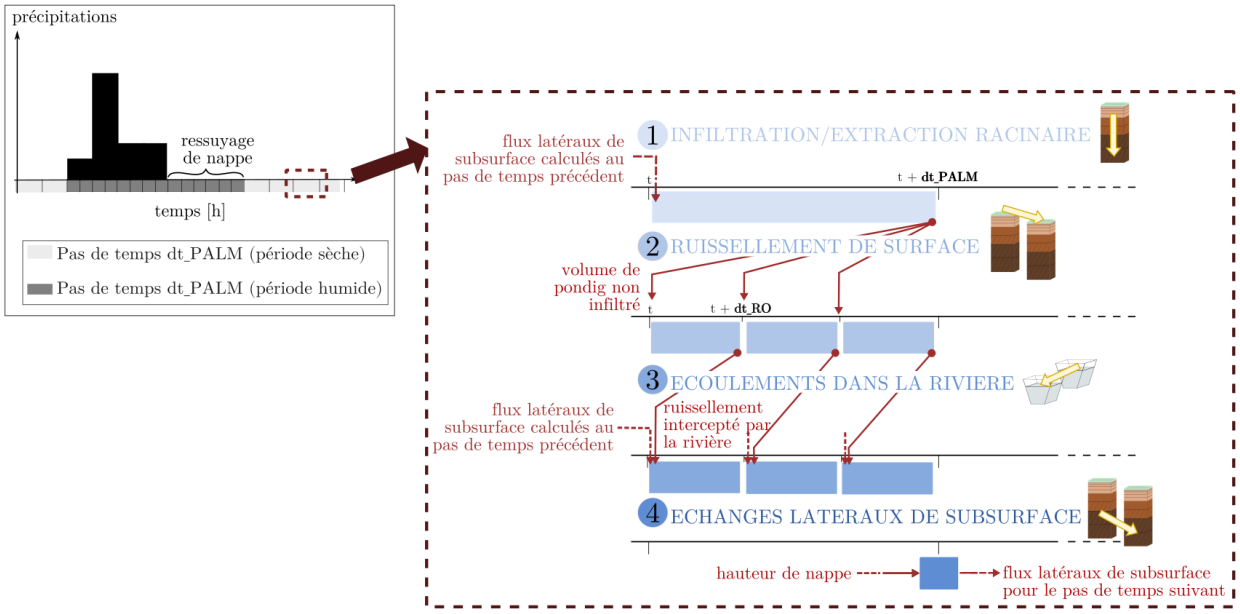


Figure 2.6 – Ligne de temps avec emboîtement des pas de temps et échanges de variables dans PESHMELBA. Gauche : hyétogramme schématique illustrant le raffinement de dt_PALM pendant et après la période de pluie. Droite : ordre de lancement des différents modules pendant un pas de temps dt_PALM et échanges de variables associés (adapté de Rouzies et al. 2019.)

Utilisation du coupleur OpenPALM

Présentation de l'outil Le couplage entre les différents codes indépendants qui constituent PESHMELBA est permis grâce à l'outil OpenPALM (Fouilloux et Piacentini, 1999 ; Buis et al., 2006). OpenPALM est un coupleur dynamique, open-source co-développé par le CERFACS et l'ONERA depuis 1996. A l'origine développé pour les besoins de la modélisation océanographique, ses champs d'application se sont maintenant diversifiés, allant de la modélisation environnementale (météorologie, feux de forêt, hydraulique...) à des applications industrielles notamment dans le domaine de l'aéronautique (interactions fluide/structure par exemple).
OpenPALM se présente sous la forme d'une bibliothèque permettant de programmer l'exécution de codes existants, de manière séquentielle ou parallèle ainsi que l'échange de données entre ces codes. L'utilisation de wrappers permet d'intégrer des codes en Fortran, C, C++ et Python. L'intégration d'un nouveau code se fait de manière relativement peu intrusive puisqu'elle implique seulement d'ajouter une carte d'identitée en début de code et des commandes d'envoi ou de réception de variables. OpenPALM dispose également de divers outils de synchronisation et d'ordonnancement qui permettent de construire des schémas de couplage complexes et notamment de gérer l'emboîtement des pas de temps.
De plus, les communications entre unités de code et le schéma d'exécution peuvent se construire depuis une interface graphique, rendant l'outil accessible à des utilisateurs thématiciens sans connaissances informatiques avancées, notamment en termes de calcul parallèle.

Limites de l'utilisation d'OpenPALM Durant la thèse, l'utilisation d'OpenPALM s'est souvent avérée problématique. La bibliothèque PALM se base entre autre sur plusieurs compilateurs et sur une bibliothèque MPI pour le calcul parallèle et assurer la compatibilité des différents éléments a été un véritable casse-tête tout au long de ces travaux. Le transfert de PESHMELBA sur le supercalculateur d'INRAE et plus largement sur toute machine récente s'est révélé particulièrement problématique. OpenPALM n'étant plus maintenu par le CERFACS depuis 2020, les mises à jour des différents éléments sur lesquels s'appuie la bibliothèque ont plus d'une fois destabilisé l'installation. Pour pallier les nombreux problèmes de versions qui ont affecté ces trois années de travail, l'application a finalement été containerisée grâce à l'outil Singularity. L'utilisation d'une telle méthode commence à se développer dans les domaines de la recherche appliquée car elle permet de figer des versions du système de manière plus flexible qu'une machine virtuelle ou un Container classique. L'image Singularity est disponible sur le dépôt git suivant : `https://forgemia.inra.fr/singularity-mesolr/peshmelba-singularity-mpichv-3-1-4`.
Néanmoins, compte tenu des difficultés rencontrées et de l'avenir incertain d'OpenPALM, il a été décidé de transférer PESHMELBA vers un autre outil de couplage spatial et dynamique. L'équipe pollutions diffuses travaille donc actuellement à transférer le modèle vers la plateforme OpenFLUID (Fabre et al., 2010) plus particulièrement dédiée à la modélisation environnementale spatialisée et qui est bien maintenue et qui dispose d'une communauté d'utilisateurs actifs.

2.2 Cas d'étude : le bassin versant de la Morcille

2.2.1 Présentation du bassin versant

Le bassin versant d'application ciblé pour la thèse est celui de La Morcille situé dans le Beaujolais, au Nord du département du Rhône (voir Figure 2.7). Il s'agit d'un site expérimental suivi par INRAE depuis 1986 (Lacas, 2005 ; Peyrard, 2016 ; Gouy et al., 2021) notamment du fait de la forte activité viticole qui le caractérise (70% de sa surface occupée par des vignes) entraînant un recours non négligeable aux pesticides. Le bassin versant de la Morcille a une surface d'environ 4.8 km^2 et est marqué par des pentes assez fortes (12% en moyenne). Il repose sur un socle granitique peu profond et le climat est continental, d'influence méditerranéeenne.

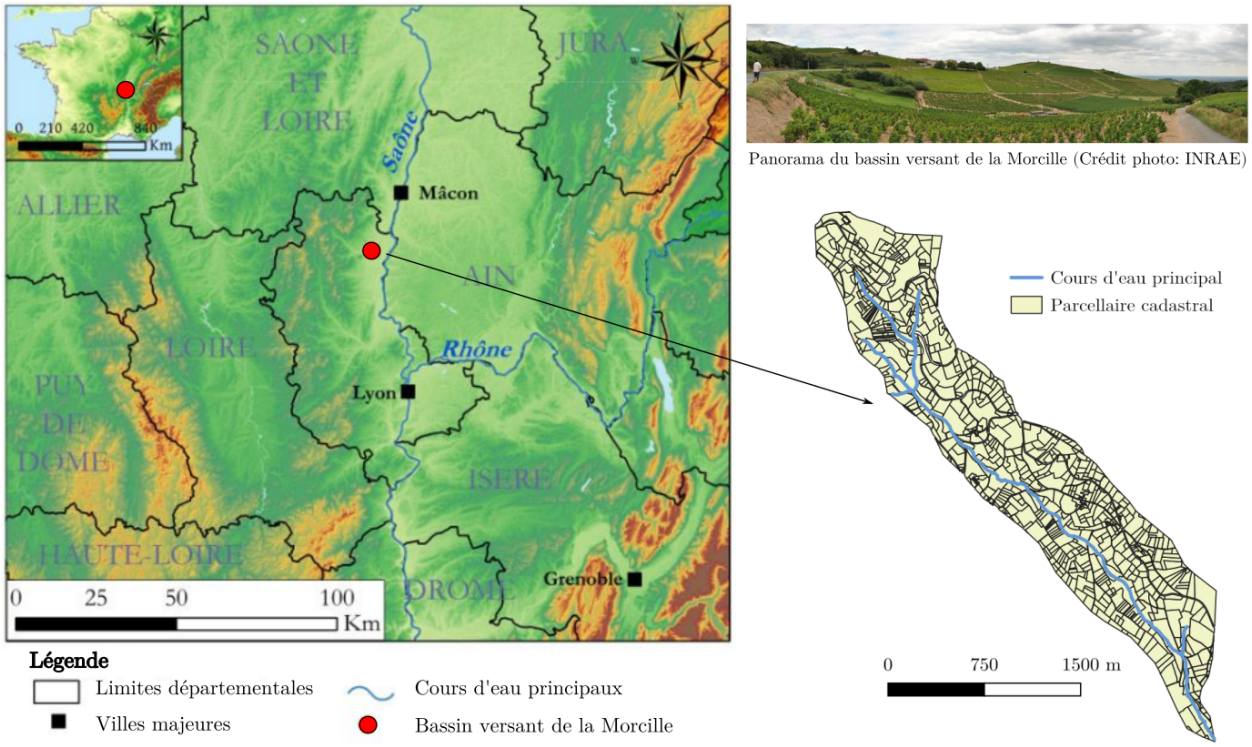


Figure 2.7 – Situation géographique (gauche) et composition (droite) du bassin versant de la Morcille (adapté de Peyrard 2016).

2.2.2 Mise en place du cas d'étude

Les travaux présentés dans ce manuscrit se basent sur une application de PESHMELBA à un bassin versant virtuel inspiré du bassin versant de la Morcille. L'objectif est ainsi de développer tous les outils méthodologiques sur un cas simplifié avant de passer à l'échelle du bassin entier. Pour cela, une portion du bassin de la Morcille a été sélectionnée et paramétrée à l'image du bassin réel. L'objectif est ainsi de simplifier le développement des outils et

l'analyse des résultats tout en étant aussi représentatif que possible des défis que représentera l'application finale.

Le scénario d'étude est un mini bassin versant de 0.12 km^2, composé de deux versants et d'une portion de rivière. Il contient 14 UH dont 10 parcelles de vigne et 4 bandes enherbées. Les pentes des UH, entre 10% et et 25%, sont également représentatives de la distribution de pentes du bassin d'origine. Les différents éléments, les connexions entre eux et la position de l'exutoire sont illustrés Figure 2.8. Les paramètres utilisés pour calibrer ce scénario proviennent en majorité des données du bassin versant réel de la Morcille.

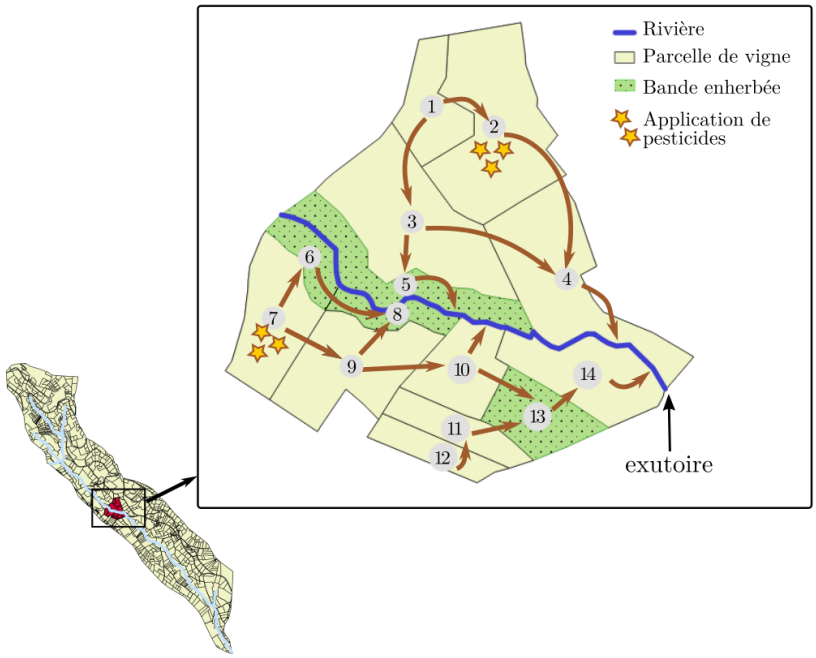


Figure 2.8 – Position et composition du mini bassin virtuel constituant le scénario d'étude. Les UH jaunes sont des parcelles de vigne alors que les vertes sont des bandes enherbées. Les flèches marrons indiquent les principales connexions entre les éléments et les étoiles sur les UH2 et 7 symbolisent les applications de pesticides.

Types de sol

Les types de sol utilisés dans ce cas d'étude reprennent les trois Unités Cartographiques de Sol (UCS) présentes sur le bassin versant de la Morcille et reportées par Frésard (2010). On ne considère que l'Unité Typologique de Sol (UTS) majoritaire parmi toutes celles pouvant exister au sein d'une UCS pour en représenter la variabilité. Une UTS est définie par la succession d'horizons de sol aux épaisseurs variables et caractérisés par des comportements hydrodynamiques distincts. Les UCS considérées pour ce scénario, la profondeur des horizons associés et leur répartition dans le bassin d'étude sont représentés Figure 2.9. On y retrouve les UCS suivantes :

- UCS1 : sols sableux sur altérites ;
- UCS2 : sols sableux sur argile ;
- UCS3 : sols sableux hétérogènes de bas de pentes et de thalwegs.

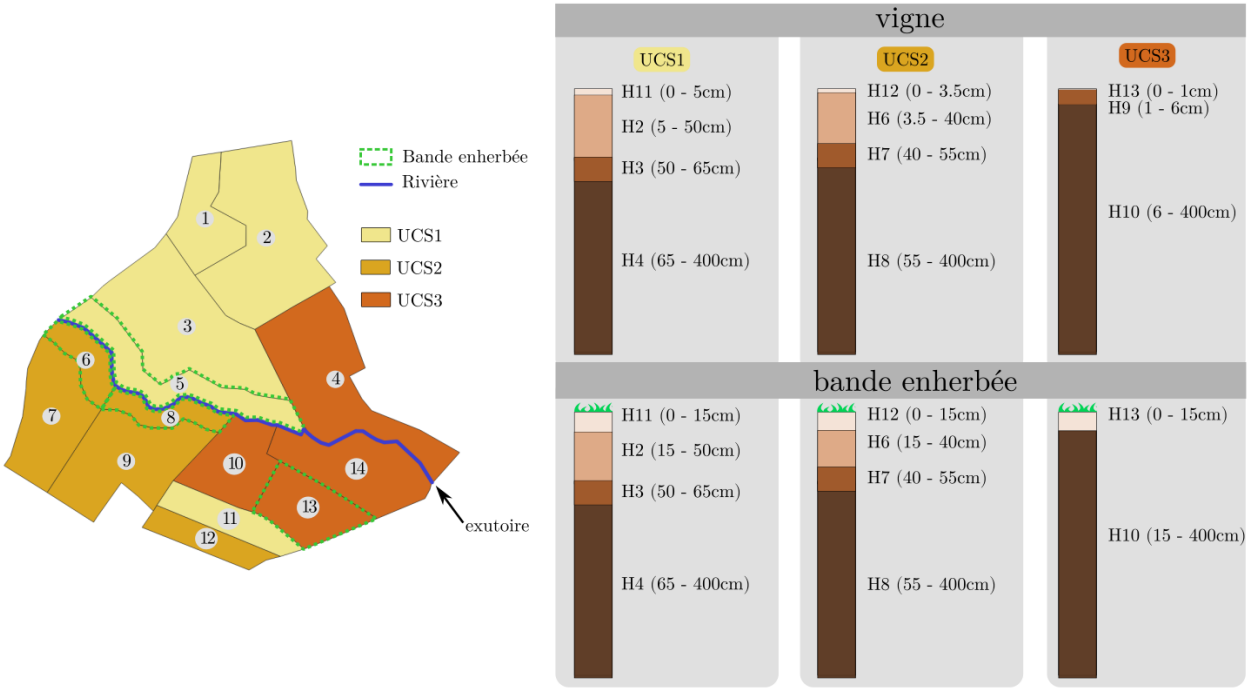


Figure 2.9 – Distribution spatiale des UCS sur le bassin virtuel d'étude et description de leur composition en termes de succession d'horizons de sol (voir Tableau 2.2).

Pour pouvoir simuler le fonctionnement de ces UCS, il est indispensable de renseigner pour chacune d'entre elles les propriétés hydrodynamiques reportés dans le Tableau 2.1 (paramètres relatifs à l'infiltration verticale). Pour cela, les teneurs en eau à saturation θ_s et résiduelles θ_r ainsi que les paramètres α et n décrivant la courbe de rétention de van Genuchten sont calées à partir des données de Van den Bogaert (2011) grâce à l'outil SWRCfit (Seki, 2007). Les mesures de conductivité provenant de ce même rapport permettent également de définir directement les conductivités hydrauliques à saturation totales K_s puis de caler les paramètres restants (conductivité hydraulique à saturation de la matrice K_x et paramètre empirique de connextivité de pore p) par méthodes des moindres carrés. Les densités apparentes *bd* proviennent également des données de Van den Bogaert (2011). Les taux de carbone organique *moc* pour les parcelles de vigne proviennent quant à eux des mesures effectuées par Randriambololohasinirina (2012).

Les parcelles et les bandes enherbées sont caractérisées par les mêmes horizons et les mêmes paramètres hydrodynamiques hormis pour l'horizon de surface. Sur les bandes enherbées, ce dernier est supposé s'étendre sur les 15 premiers cm, quelque soit l'UCS considérée (voir Figure 2.9). Pour rendre compte de l'impact de ces zones végétalisées sur l'infiltration

	H	*thetas* [$m^3.m^{-3}$]	*thetar* [$m^3.m^{-3}$]	*hg* [m]	*n* [-]	*Ks* [$m.s^{-1}$]	*Ko* [$m.s^{-1}$]	*L* [-]	*bd* [$g.cm^{-3}$]	*moc* [%]
UCS1	11	0.34	0.04	$-9.69\cdot10^{-2}$	1.27	$3.93\cdot10^{-5}$	$2.86\cdot10^{-7}$	-8.43	1.34	0.91
	2	0.34	0.05	$-3.29\cdot10^{-2}$	1.20	$8.64\cdot10^{-5}$	$2.28\cdot10^{-7}$	-6.52	1.47	0.39
	3	0.32	0.08	$-2.09\cdot10^{-2}$	1.20	$5.39\cdot10^{-5}$	$7.47\cdot10^{-7}$	-4.24	1.57	0.10
	4	0.28	0.07	$-5.99\cdot10^{-2}$	1.23	$3.11\cdot10^{-5}$	$1.47\cdot10^{-6}$	-0.14	1.53	0.07
UCS2	12	0.34	0.04	$-9.69\cdot10^{-2}$	1.27	$3.93\cdot10^{-5}$	$2.86\cdot10^{-7}$	-8.43	1.34	1.15
	6	0.35	0	$-6.60\cdot10^{-2}$	1.13	$2.16\cdot10^{-5}$	$3.19\cdot10^{-7}$	9.66	1.59	0.68
	7	0.32	0	$-7.18\cdot10^{-2}$	1.08	$9.60\cdot10^{-6}$	$1.67\cdot10^{-7}$	-10	1.66	0.35
	8	0.42	0	$-0.30\cdot10^{-2}$	1.08	$3.98\cdot10^{-6}$	$9.72\cdot10^{-8}$	10	1.54	0.28
UCS3	13	0.34	0.04	$-9.69\cdot10^{-2}$	1.27	$3.93\cdot10^{-5}$	$2.86\cdot10^{-7}$	-8.43	1.34	0.75
	9	0.33	0.08	$-6.72\cdot10^{-2}$	1.26	$3.05\cdot10^{-5}$	$3.36\cdot10^{-7}$	0.42	1.46	0.37
	10	0.32	0.06	$-3.56\cdot10^{-2}$	1.18	$2.38\cdot10^{-5}$	$3\cdot10^{-7}$	1.05	1.62	0.40

Tableau 2.2 – Propriétés hydrodynamiques des horizons de sol composant les UCS1, 2 et 3 pour les parcelles de vigne, selon les données de Van den Bogaert (2011). La deuxième colonne regroupe les indices des horizons de sol.

et l'adsorption des pesticides, et d'après Catalogne et al. (2018) et Randriambololohasinirina (2012), on suppose une conductivité hydraulique à saturation et un taux de carbone organique accrus par rapport aux parcelles de vigne : K_s=155 $mm.h^{-1}$ (=$4.31\cdot10^{-5}$ $m.s^{-1}$) et *moc*=2.8%.

Forçages et conditions initiales

PESHMELBA a été développé pour pouvoir évaluer l'impact du paysage sur les transferts à l'échelle de plusieurs mois, voire de plusieurs années. Cependant, des simulations de longue durée n'ayant jamais été réalisées auparavant, on considère pour cette application une durée de simulation de 3 mois ce qui suffit pour permettre l'activation et la contribution significative de tous les processus physiques dans le modèle. Les chroniques de précipitations d'un pluviomètre installé sur le bassin sont disponibles dans la base BDOH (Gouy et al., 2015). Les données d'ETP proviennent de données Météo France correspondant au site voisin de Liergues puis moyenées par décades et corrigées par un facteur de -11 % pour correspondre au site de la Morcille (Durand, 2014 ; Caisson, 2019). On sélectionne deux périodes aux régimes climatiques distincts : une période hivernale caractérisée par des évènements pluvieux assez longs et intenses et une ETP faible (pluie=667 mm cumulés, ETP=33 mm cumulés), et une période estivale caractérisée par des événements pluvieux courts et peu intenses et une ETP élevée (pluie=369 mm cumulés, ETP=74 mm cumulés). Les chroniques correspondantes sont représentées sur la Figure 2.10. Dans la suite, on fera référence à ces deux scénarios avec les dénominations *scénario hivernal* et *scénario estival*.

Les conditions initiales de pression dans les UH sont calculées en supposant un équilibre hydrostatique et des profils partiellement saturés. Bien que le scénario soit virtuel, des niveaux de nappes initiaux aussi plausibles que possible ont été utilisés pour l'initialiser. Ces

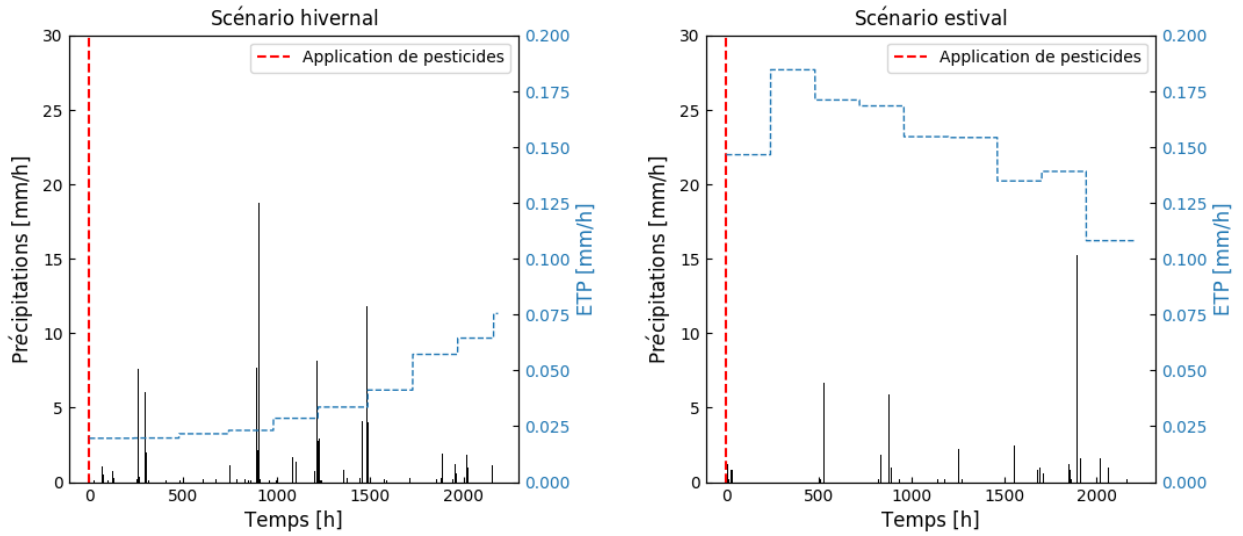


Figure 2.10 – Chroniques de pluie et d'évapotranspiration potentielle pour le scénario hivernal (gauche) et le scénario estival (droite).

niveaux de nappes sont déduits de niveaux piézométriques mesurés sur un versant instrumenté à proximité du site utilisé pour constituer le bassin virtuel (voir Figure 2.11). Les données piézométriques étant disponibles sur un transect perpendiculaire à la rivière, elles sont extrapolées le long du versant virtuel pour reconstituer un profil de nappe pour les périodes de simulation sélectionnées. Le niveau de nappe dans chaque UH est ensuite initialisé en fonction de la distance entre le centroïde et la rivière, et ce sur les deux versants.

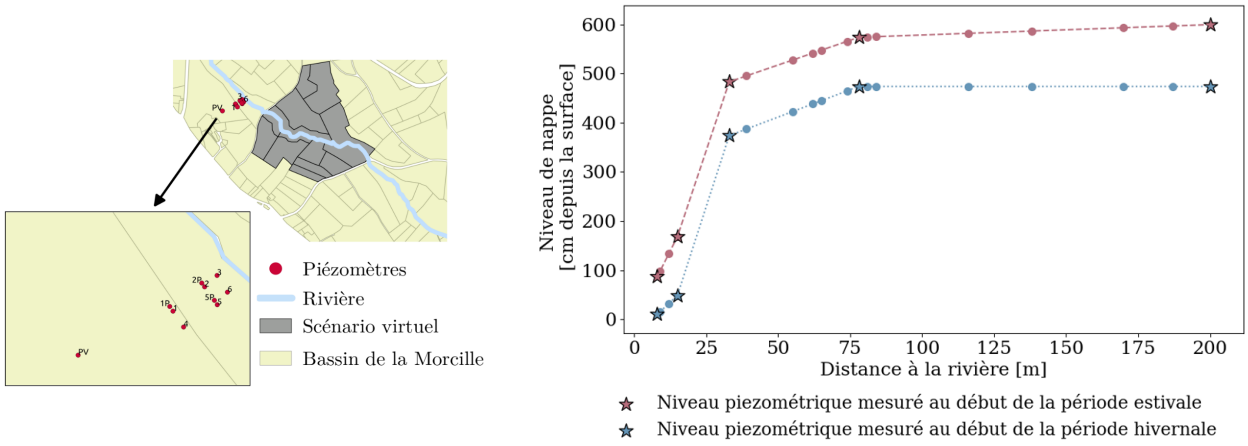


Figure 2.11 – Gauche : position des différents piézomètres disponibles. Droite : profil de nappe interpolé à partir des données mesurées. Les étoiles correspondent aux points de mesure alors que les ronds correspondent aux positions des centroïdes des différentes UH composant le bassin virtuel.

Végétation

Dans ce scénario, on simule un couvert végétal type vigne sans enherbement inter-rang sur les parcelles et type prairie permanente sur les bandes enherbées. Compte tenu de la durée envisagée des simulations, on suppose constantes la profondeur racinaire maximale Z_r et la densité racinaire dans les 10% supérieurs du sol F_{10}. Pour la vigne, ces valeurs sont tirées de Smart et al. (2006) et confirmées à dire d'experts pour la région du Beaujolais. Elles sont fixées à Z_r=2.62 m et F_{10}=37%. Pour la prairie permanente Z_r est fixé à 0.9 m et F_{10} à 33.5% (Brown et al., 2007).
Pour la vigne, le modèle de croissance du LAI proposé dans MACRO (voir Figure 2.4) et initialement développé pour des cultures annuelles est adapté au cycle de développement de la vigne. Les dates et valeurs du cycle sont tirées de Brown et al. (2007). On ne considère pas de phase de développement linéaire et le LAI évolue à partir d'une valeur minimale $LAI_{min} = 0.01$ atteinte le 1er Février jusqu'à $LAI_{max} = 2.5$ atteint le 1er Mai avant de redécroître jusqu'à $LAI_{harv} = 0.01$ le 1er Novembre. Le LAI est supposé constant et égal à 5 sur les bandes enherbées.
Les rugosités de Manning sont tirées de Arcement et Schneider (1989). On choisit un coefficient de culture pérenne en rangs pour la vigne (0.033 $s.m^{-1/3}$) et un coefficient de prairie pour la bande enherbée (0.2 $s.m^{-1/3}$).

Pesticide

On simule les transferts et le devenir du tébuconazole, un fongicide communément utilisé sur la Morcille. Les paramètres caractéristiques de l'isotherme de Freundlich sont obtenus de la PPDB (Lewis et al., 2016) (Koc=769 $mL.g^{-1}$, n^f=0.84). Le coefficient de dégradation surfacique est également tiré de la PPDB ($DT50$=47.1 jours) et on considère une loi de dégradation exponentielle en fonction de la profondeur (FOCUS, 2001). Pour les deux scénarios, on considère une application de 0.5 $kg.ha^{-1}$ de tébuconazole en début de simulation (voir Figure 2.10) sur les parcelles de vigne 2 et 7 situées à l'amont du bassin. On suppose que la majorité des processus de dégradation et d'adsorption qui affectent le tébuconazole ont lieu sur les parcelles et dans les bandes enherbées. Ainsi, ni dégradation ou adsorption ne sont simulées dans la rivière.

Description des UH et de la rivière

La hauteur de ponding limite *hpond* est fixée à 0.01 m sur les parcelles de vigne et à 0.05 m sur les bandes enherbées (Gatel, 2018). L'épaisseur de la couche de mélange pour le calcul de l'adsorption *adsorpthick* est fixée à 0.01 m (Gao et al., 2004 ; Walter et al., 2007) sur toutes les UH.
Dans la rivière, la distance entre le fond du lit et l'imperméable de la zone saturée D_i est fixée à 1.5 m (mesures de tomographie de resistivité électrique locales, communication personnelle) et la conductivité hydraulique à saturation du fond du lit Ks_{river} est fixée à $2.38 \cdot 10^{-5}$ $m.s^{-1}$

conformément aux valeurs mesurées dans le sol avoisinant. La hauteur de ponding limite dans la rivière *hpond* est fixée à 0.01 m et la rugosité de Manning est choisie égale à 0.079 $m.s^{-1}$ conformément à Arcement et Schneider (1989), pour des canaux à obstruction limitée.

2.2.3 Incertitudes liées aux variables d'entrée et paramètres

Les valeurs nominales des paramètres d'entrée détaillées dans les sections précédentes sont définies à partir de mesures terrain, de la littérature ou à dires d'experts. Il en résulte qu'elles sont entâchées d'incertitudes parfois importantes, pouvant affecter grandement les résultats des simulations. De telles incertitudes seront prises en compte tout au long de cette thèse par l'analyse de sensibilité et l'assimilation de données. Pour les décrire, on définit la distribution associée à chaque paramètre d'entrée considéré incertain. Afin de limiter la complexité finale de l'application, les paramètres d'entrée considérés comme incertains sont ceux reportés dans le Tableau 2.1, excepté les dates relatives au cycle de développement du LAI. En tenant compte des différents horizons de sol, il en résulte 145 paramètres d'entrée dont les distributions doivent être définies. Une fois de plus, les formes des distributions ainsi que les paramètres qui les caractérisent sont issues de mesures terrain ou de valeurs reportées dans la littérature.
Une distribution lognormale est classiquement attribuée à la conductivité hydraulique à saturation K_s (Coutadeur et al., 2002 ; Fox et al., 2010 ; Schwen et al., 2011 ; Dairon, 2015 ; Dubus et Brown, 2002 ; Dubus et al., 2003). Un coefficient de variation de 20% est utilisé afin de rester représentatif du comportement hydrodynamique de chaque horizon de sol. Les distributions pour les paramètres de Schaap-Van Genuchten n'étant pas disponibles dans la littérature, on attribue au paramètre empirique de connectivité des pores p une distribution uniforme s'étendant à +/- 20% autour de la valeur moyenne (Zajac, 2010). Comme la conductivité à saturation dans la matrice K_x a le même sens physique que K_s, on lui attribue également une distribution lognormale et un coefficient de variation de 20%. Les teneurs en eau à saturation *thetas*, les teneurs en eau résiduelles *thetar* ainsi que les paramètres de van Genuchten *mn* et *alpha* suivent des distributions normales (Schwen et al., 2011 ; Alletto et al., 2015 ; Gatel et al., 2019). Les coefficients de variation correspondants sont fixés à 10% pour *thetas*, *mn* et *alpha* et à 25% pour *thetar* (Schwen et al., 2011 ; Alletto et al., 2015 ; Gatel et al., 2019 ; Lauvernet et Muñoz-Carpena, 2018). On attribue une distribution uniforme au taux de carbone organique *moc*, une distribution triangulaire au coefficient d'adsorption K_{oc} et une distribution normale au temps de demie-vie $DT50$ (Lauvernet et Muñoz-Carpena, 2018). On considère un coefficient de variation de 60% pour le K_{oc} et le $DT50$ car ces paramètres sont généralement considérés comme très incertains (Dubus et al., 2003). Des distributions triangulaires sont attribuées aux rugosités de Manning sur les UH et dans la rivière (Lauvernet et Muñoz-Carpena, 2018 ; Gatel et al., 2019). Pour les paramètres restants pour lesquels aucune information permettant de caractériser leurs incertitudes n'a pu être trouvée, on utilise des distributions uniformes à +/- 20% autour de la valeur nominale (Zajac, 2010). Les fonctions de densité de probabilité

(pdf) sont regroupées dans le tableau de l'Annexe A.

Dans ces travaux, tous les paramètres sont supposés être indépendants et aucune structure de corrélation n'est considérée. Bien qu'elle simplifie le problème, cette hypothèse est très probablement fausse, notamment pour les propriétés hydrodynamiques du sol qui doivent définir des courbes de conductivité et de rétention qui ont un sens. Pour s'assurer d'une telle cohérence, les points qui ont été échantillonnés dans les distributions des paramètres hydrodynamiques pour l'analyse de sensibilité ont été utilisés pour tracer les courbes de rétention et de conductivité correspondant à chaque échantillon. L'objectif était alors d'identifier des jeux de paramètres potentiellement problématiques car aboutissant à des formes de courbe irréalistes. Cependant, les distributions des paramètres hydrodynamiques ayant été définies avec des coefficients de variation assez limités, cette étape de vérification n'a abouti à l'élimination d'aucun jeu de paramètres.

2.3 Observations virtuelles

Cette section décrit les différents types d'observations utilisées pour l'assimilation de données dans la thèse. On rappelle que l'on travaille sur un scénario virtuel, les observations utilisées sont donc elles aussi virtuelles. On s'attache cependant à reproduire au mieux les caractéristiques des vraies données correspondantes (erreur, résolution spatiale et temporelle) pour préparer une future application sur un bassin versant réel. Les différents types de données utilisés ainsi que leurs caractéristiques sont illustrés sur la Figure 2.12. Les premières observations considérées sont les images d'humidité de surface obtenues à partir de données satellite (Section 2.3.1). On suppose ensuite disposer de mesures in-situ : profils verticaux d'humidité du sol et séries temporelles de concentration en pesticides dans la rivière (Section 2.3.2).

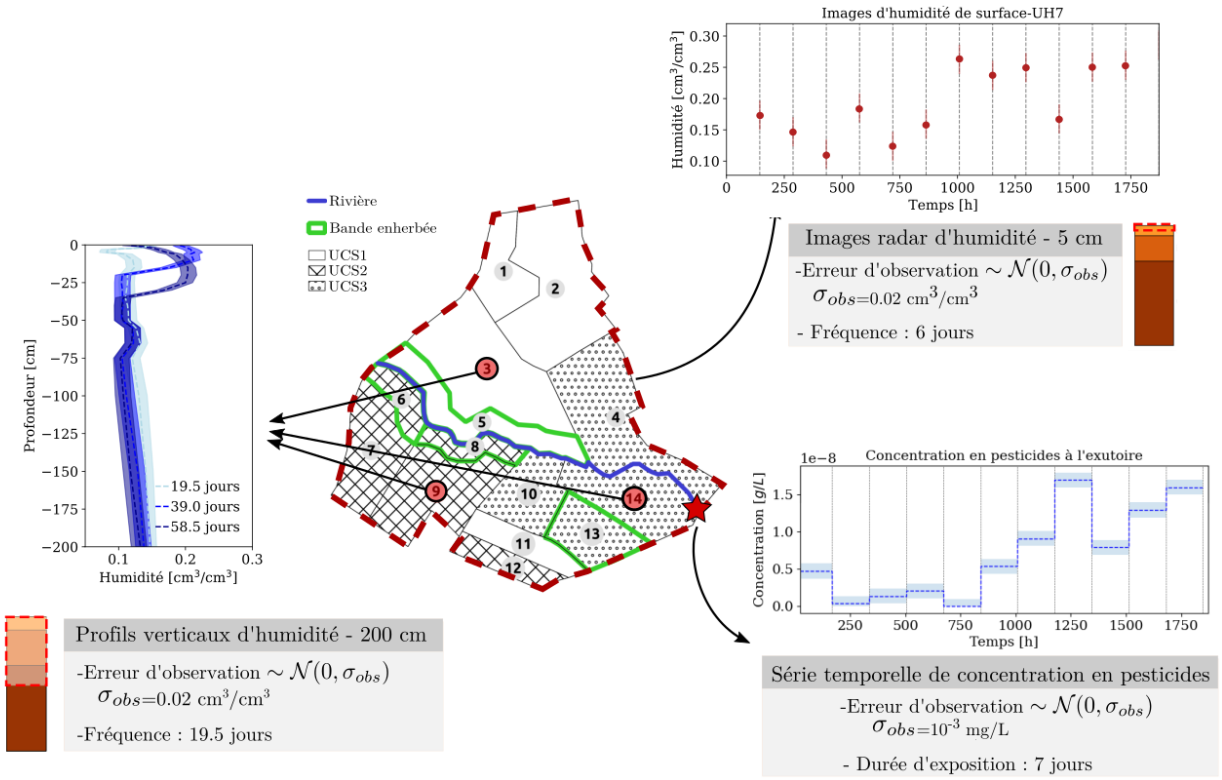


Figure 2.12 – Description des différents types de données disponibles pour le scénario d'étude : images satellite d'humidité de surface, profils verticaux d'humidité et concentration en pesticides à l'exutoire intégrée sur une semaine. Sur les 3 exemples de séries temporelles ou de profil verticaux, l'erreur d'observation ($\pm\sigma_{obs}$) est représentée par une barre ou une enveloppe.

2.3.1 Images satellite

Il est d'abord possible de tirer profit des produits satellite que les missions spatiales produisent en grande quantité pour obtenir des observations à un moindre coût. Dans cette thèse,

on considère disponibles des images d'humidité de surface obtenues à partir de l'utilisation couplée de données radar Sentinel 1 et de données optiques Sentinel 2. Cette approche décrite dans El Hajj et al. (2017) et Bousbih et al. (2018), fait maintenant l'objet d'une chaîne opérationnelle développée et utilisée à la maison de la télédétection à Montpellier. Les cartes d'humidité sont notamment accessibles à `https://thisme.cines.teledetection.fr/home`. Du fait de la haute résolution spatiale de Sentinel 1 et 2, cette approche permet de fournir des cartes d'humidité jusqu'à l'échelle de la parcelle. De plus, la haute fréquence de revisite des deux constellations (Sentinel 1A + Sentinel 1B et Sentinel 2A + Sentinel 2B) permet en principe de disposer de données tous les 6 jours.

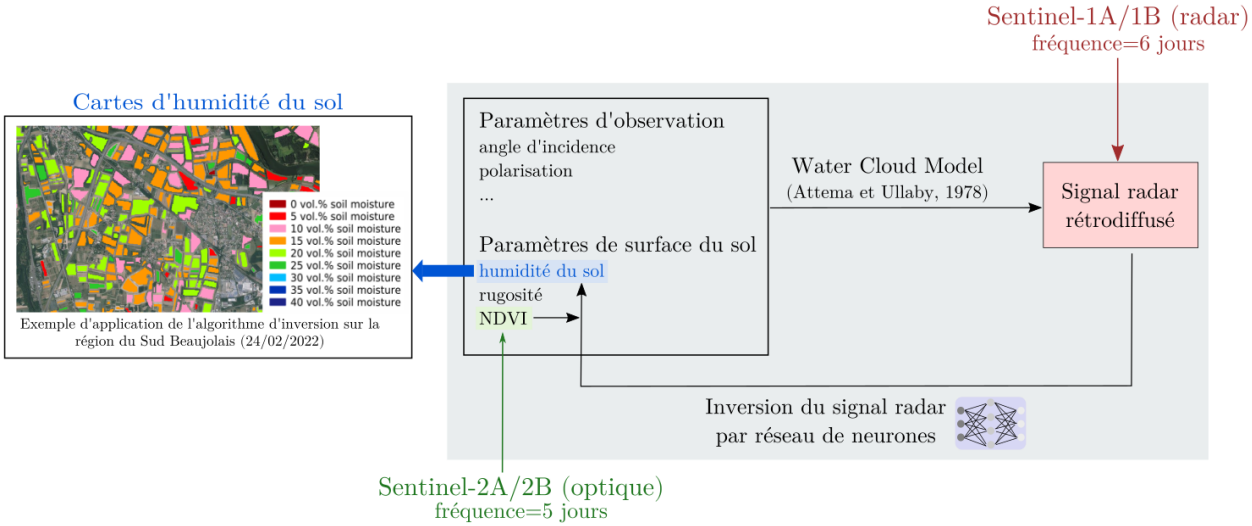


Figure 2.13 – Principe de l'algorithme d'inversion du signal radar pour l'obtention de cartes d'humidité du sol (adapté de Lozac'h et al. 2020).

Le principe global est décrit Figure 2.13. Le signal radar rétrodiffusé mesuré par la constellation Sentinel 1 dépend non seulement de paramètres d'observation intrinsèques à l'appareil de mesure (bande de fréquence radar, polarisation, angle d'incidence) mais aussi de l'état de la surface observée (rugosité, humidité du sol, végétation). La relation directe entre état de surface et signal radar rétrodiffusé est en général modélisée par des modèles semi-empiriques ou physiques. Ici, c'est le Water Cloud Model (WCM, Attema et Ulaby 1978), un modèle semi-empirique, qui est utilisé. Les travaux de Baghdadi et al. (2012) proposent alors de modéliser la relation inverse à l'aide d'un réseau de neurones pour déduire l'humidité de surface à partir du signal radar rétrodiffusé. Or, le WCM prend en compte la contribution du couvert végétal sur le signal radar rétrodiffusé. Le réseau de neurones permettant l'inversion de ce signal utilise donc le NDVI (Normalized Differential Vegetation Index) qui est estimé directement à partir des images optiques de la constellation Sentinel 2. Finalement, les observations d'humidité obtenues sont moyennées à l'échelle de la parcelle. Compte tenu du faible pouvoir de pénétration du signal radar dans le sol, la valeur de l'humidité obtenue est représentative des 5 premiers centimètres du sol.

Cette approche a été développée puis validée pour des couverts végétaux type grande culture d'hiver ou prairie (El Hajj et al., 2017). Dans ce cadre, l'erreur d'estimation de l'humidité (RMSE) est évaluée à environ 0.05 cm^3cm^{-3}. Cependant, il est important de noter que ces performances sont probablement très variables selon les couverts végétaux considérés. A l'heure actuelle, le réseau de neurones permettant l'inversion du signal radar n'a pas été entraîné sur des couverts végétaux type vigne. Pour préparer une application réelle sur le bassin versant de la Morcille, caractérisé par son intense activité viticole, il sera nécessaire d'entraîner le réseau de neurones dans ce contexte particulier. A noter que ceci sera probablement à l'origine de difficultés supplémentaires puisque la présence de structures métalliques dans les rangs de vigne perturbe le signal radar. Les fortes pentes qui caractérisent certaines portions de la Morcille pourraient également limiter la mesure du signal radar rétrodiffusé. Dans le cadre de ces travaux, on fait l'hypothèse que ce travail amont a été réalisé et que l'on dispose d'images d'humidité moyenne dans les 5 premiers cm de sol sur les parcelles de vigne et dans les bandes enherbées, tous les 6 jours. L'erreur d'observation est supposée gaussienne et non biaisée. Dans un premier temps, l'écart-type est fixé de manière arbitraire à une valeur suffisamment faible (σ_{obs}=0.02 cm^3cm^{-3}) pour assurer un impact significatif du processus d'assimilation. L'effet de l'amplitude de cette erreur sur les performances de l'assimilation sera exploré dans un second temps. De plus, on fait l'hypothèse que les erreurs d'observation ne sont pas corrélées spatialement et temporellement. Cette hypothèse, très forte, est probablement fausse compte tenu de la nature du capteur (tous les pixels sont mesurés par le même appareil) mais simplifie fortement l'application des méthodes d'assimilation de données.

2.3.2 Mesures in-situ

En plus des images d'humidité disponibles à haute fréquence temporelle et spatiale, on fait l'hypothèse que l'on dispose de données in-situ. Bien que ces données soient potentiellement plus coûteuses et plus complexes à obtenir, elles permettent de disposer d'observations sur d'autres compartiments que la surface. On suppose disponibles des profils verticaux d'humidité du sol et des données de concentration en pesticides dans la rivière. Ces deux types de données sont décrits dans les paragraphes suivants.

Profils verticaux d'humidité du sol

On suppose que des profils verticaux d'humidité du sol sont disponibles ponctuellement sur le bassin. En pratique, de tels profils sont obtenus à partir de sondes TDR ou de mesures de résistivité électrique (ERT). S'il s'agit de sondes TDR, on obtient un unique profil valable localement. Compte tenu de la discrétisation adoptée dans le modèle PESHMELBA, on suppose alors qu'un profil est représentatif à l'échelle d'une UH. Dans le cas de mesures ERT, on obtient une cartographie 3D de la resistivité (Brunet et al., 2010). La résistivité est ensuite reliée à l'humidité du sol à l'aide d'une relation empirique qui est fonction du type de sol. L'ERT a ainsi un intérêt particulier pour des UH présentant de fortes hétérogénéités

et où un profil obtenu par sonde n'est plus représentatif. Cependant, pour être utilisée dans PESHMELBA, une telle information doit être moyennée pour obtenir un profil moyen d'humidité par UH à un instant donné. On note également que l'erreur d'observation associée à ce type de données, et notamment la composante liée au processus d'inversion de la résistivité électrique, peut être particulièrement difficile à estimer. Un exemple de cartographie de résistivité obtenue sur une parcelle de la Morcille est présenté sur la Figure 2.14.

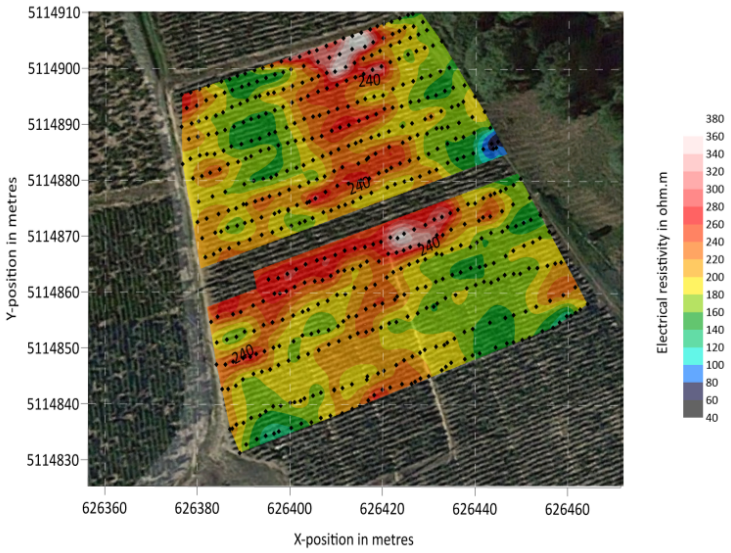


Figure 2.14 – Champ de résistivité électrique surfacique reconstruit à partir de mesures réalisées sur une parcelle de la Morcille. Les points noirs représentent la position des sondes.

Dans le cadre des expériences virtuelles menées pour cette thèse, on suppose disponibles des profils verticaux d'humidité du sol couvrant les 2 premiers mètres de sol ponctuellement sur le bassin virtuel. L'impact du nombre et de la position de ces profils sur les performances de l'assimilation sera notamment étudié dans ces travaux. L'erreur d'observation est supposée gaussienne, non biaisée et en première approximation, on fixe l'écart-type à 0.02 $cm^3.cm^{-3}$. Une telle valeur devra être affinée lors du passage à des observations réelles selon le capteur considéré. Là encore, on suppose que les erreurs d'observation ne sont pas corrélées spatialement et temporellement.

Suivi de concentration dans la rivière

Enfin, on suppose qu'une série temporelle de concentration en pesticides provenant d'un échantillonneur intégré passif (EIP) est disponible tout au long de la simulation. L'EIP est un dispositif constitué d'une membrane et d'une phase réceptrice sur lesquels les pesticides présents dans l'eau vont s'accumuler (Le Dreau et al., 2018). Il est immergé dans le milieu aquatique et permet d'échantillonner en continu tout au long de la période d'immersion. Il en résulte des valeurs de concentrations moyennées sur la durée d'exposition de l'appareil.

La durée d'immersion de l'EIP varie de 7 jours à 3 mois. L'incertitude des concentrations moyennes mesurées varie considérablement selon la molécule suivie (notamment sa capacité d'adsorption), les propriétés de l'EIP (volume, taux d'échantillonnage,...) et celles du milieu (vitesse du courant notamment). Les erreurs de mesures constatées dans Martin (2016) pour le suivi de tébuconazole sur l'Ardières, près du site modélisé et dont la Morcille est un affluent sont de l'ordre de la dizaine de ng/L (voir Figure 2.15). Cependant, les concentrations simulées dans notre cas d'étude présenté en Section 2.2 sont bien souvent éloignées de celles mesurées par Martin (2016) et les valeurs des erreurs estimées sont ainsi difficilement transposables. Pour assurer les développements méthodologiques de cette thèse, on fait l'hypothèse d'un capteur relativement fiable et d'une erreur gaussienne, non biaisée, d'écart-type 10^3 ng/L(=10^{-3} mg/L). On suppose également que le temps d'exposition du capteur est de 7 jours.

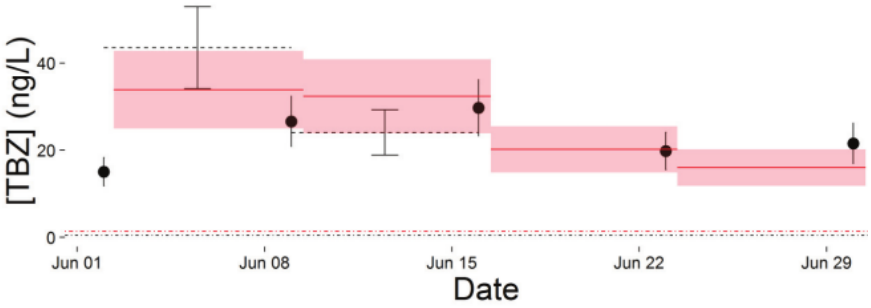


Figure 2.15 – Comparaison des concentrations en tébuconazole mesurées sur l'Ardières en Juin 2015 avec un échantillonneur ponctuel (points), un échantillonneur moyenné au temps (lignes pointillées) et un échantillonneur passif (ligne rouge) et incertitudes associées (intervalle de confiance à 95 %). La ligne rouge (resp. noire) en pointillés le long de l'axe des abscisses représente la limite de quantification pour l'échantillonneur ponctuel (resp. l'échantillonur passif). Tiré de Martin (2016).

2.4 Conclusion

Ce chapitre a permis de présenter le modèle et les observations qui font l'objet de ces travaux de thèse.
D'un côté le modèle PESHMELBA, développé pour simuler un système complexe, caractérisé par de nombreux processus qui interagissent fortement. Comme pour tout modèle, son développement se base sur de nombreuses hypothèses et simplifications, parfois fortes, dont il faut avoir conscience pour en faire une utilisation éclairée. Cependant, malgré un niveau de détails relativement modéré dans la représentation des processus qui régissent le devenir des pesticides, PESHMELBA nécessite de renseigner de nombreux paramètres d'entrée. Sa structure complexe rend opaque la relation entre ces paramètres d'entrée et les variables de sortie. La première partie de ces travaux de thèse permettra de caractériser cette relation.
De l'autre côté, les observations disponibles sur le bassin versant : des images satellite d'humidité de surface, des profils ponctuels d'humidité dans le sol et des séries temporelles de

concentration en pesticides à l'exutoire. Elles ont des fréquences temporelles et des emprises spatiales hétérogènes. Dans la seconde partie de ces travaux, nous verrons comment ces observations peuvent être fusionnées avec le modèle pour améliorer au mieux la connaissance du système physique modélisé par PESHMELBA.

Première partie

Comment quantifier les incertitudes dans le modèle PESHMELBA ?

Chapitre 3

Présentation des outils de quantification d'incertitude utilisés dans la thèse

Sommaire

3.1 Introduction

La quantification d'incertitude est une étape clé pour faire de PESHMELBA un modèle d'aide à la décision robuste et utile dans la gestion des risques liés aux transferts de pesticides. En effet, des incertitudes sur les valeurs des paramètres et/ou des variables d'entrée [1] et sur les équations implémentées peuvent aboutir à des erreurs conséquentes sur les prédictions du modèle. Pour envisager l'utilisation opérationnelle de PESHMELBA, il est

1. Dans la suite, on regroupe sous le terme générique de **facteurs d'entrée** les **paramètres d'entrée** qui sont statiques et les **variables d'entrée** qui sont dynamiques. La distinction entre paramètre et variable d'entrée diffère selon les communautés et on choisit ici la distinction qui fait le plus de sens dans ces travaux.

indispensable d'être capable d'identifier ses principales sources d'incertitude et d'en analyser les conséquences pour la prédiction et l'aide à la décision. Dans ces travaux de thèse, on s'appuie pour cela sur deux outils complémentaires : l'analyse d'incertitude et l'analyse de sensibilité.
L'analyse d'incertitude est en général utilisée dans un premier temps. Elle permet d'évaluer le niveau d'incertitude sur les prédictions du modèle qui est induit par l'incertitude sur les facteurs d'entrée. Cette information peut ainsi contribuer à évaluer la qualité prédictive du modèle mais aussi à mettre en place une stratégie de réduction de l'incertitude adéquate.
L'analyse de sensibilité consiste quant à elle à caractériser l'influence des facteurs d'entrée sur une ou plusieurs variables de sortie du modèle. Elle permet, entre autre, d'identifier et de hiérarchiser les facteurs d'entrée qui ont une forte (ou une moindre) influence sur les sorties du modèle. Ses résultats permettent ainsi de mieux cibler les efforts de calibration sur les facteurs les plus influents ou, au contraire, de simplifier le modèle en fixant ou en supprimant les facteurs peu ou non influents.
Ce chapitre présente la démarche d'analyse d'incertitude (Section 3.2) et les différents outils d'analyse de sensibilité (Section 3.3) utilisés dans la thèse. La méthodologie résultante est ensuite décrite dans la Section 3.4. On y distingue variables intégrées et séries temporelles puisque l'approche d'analyse de sensibilité correspondante diffère légèrement.

3.2 Analyse d'incertitude

Dans ces travaux de thèse, l'analyse d'incertitude vise à déterminer la fonction de densité de probabilité de variables de sortie du modèle. Dans des cas simples, celle-ci peut être calculée de manière analytique à partir des densités de probabilité des facteurs incertains. PESHMELBA étant un modèle trop complexe pour envisager un tel calcul, la densité de probabilité des variables cibles doit ici être estimée. Pour cela, on utilise une méthode de propagation de l'incertitude par simulations de Monte Carlo basée sur des tirages aléatoires (de Rocquigny, 2006). Une telle méthode est notamment décrite dans Faivre et al. (2013) et on en rappelle ici les principales étapes en se basant sur cet ouvrage. La Figure 3.1 en résume le principe et les 4 étapes qui la définissent (Da Veiga et al., 2021).

1. Définition des distributions des facteurs d'entrée incertains Cette étape est en général réalisée en se basant sur des mesures terrains, sur la littérature scientifique ou l'expertise disponible. Il n'est pas rare de supposer que les facteurs d'entrée sont indépendants mais si ce n'est pas le cas, on définit à cette étape les structures de corrélation entre eux. Cette étape est cruciale mais aussi souvent subjective. Elle dépend également du cadre d'application du modèle qui est considéré (en l'occurence le contexte agropédoclimatique pour PESHMELBA) ce qui implique que les résultats de l'analyse d'incertitude sont seulement valides dans ce contexte.

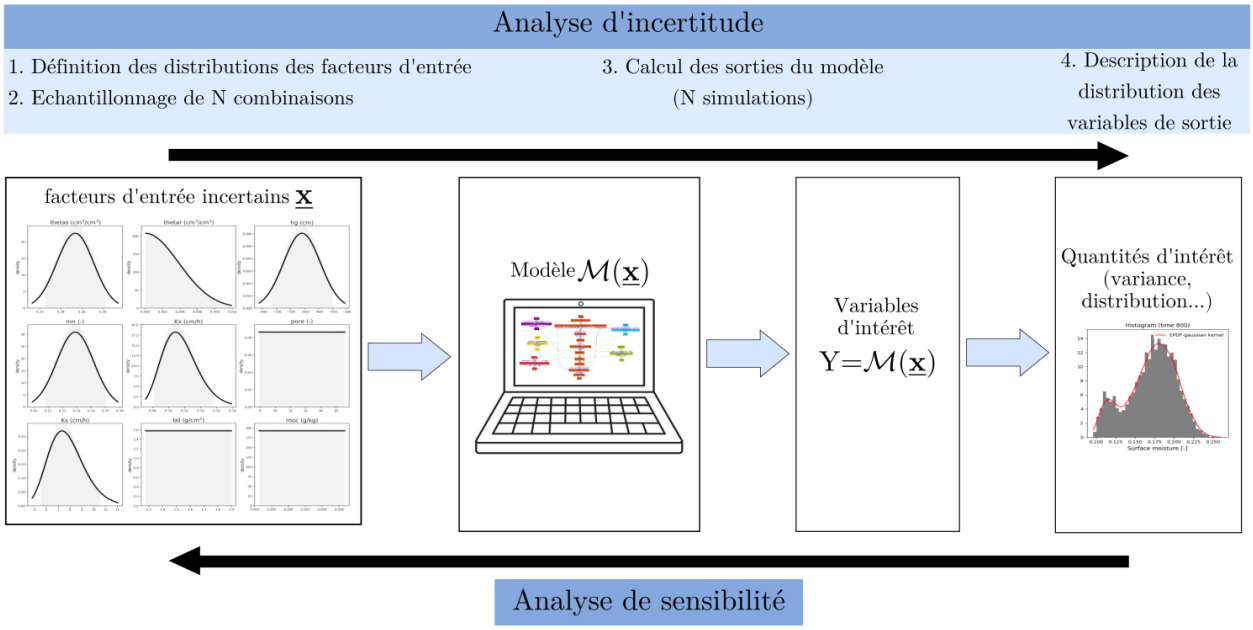


Figure 3.1 – Vue d'ensemble de la méthodologie d'analyse d'incertitude utilisée dans la thèse (adapté de Da Veiga et al. 2021).

2. Echantillonnage dans les distributions On tire ensuite N combinaisons de facteurs d'entrée selon les distributions définies à l'étape précédente. La valeur de N doit être suffisante pour garantir la stabilité des résultats mais une valeur trop importante peut entraîner un coût de calcul irréaliste. Il existe de nombreux algorithmes pour générer des valeurs aléatoires dans les distributions et on échantillonne ici par hypercube latin (LHS, McKay et al. 1979) en alternative à la méthode de Monte Carlo. Pour construire un hypercube latin, la gamme de variation de chaque facteur d'entrée est divisée en intervalles de probabilité égale. Des points sont ensuite échantillonnés dans ces intervalles de manière à ce que chaque échantillon soit seul dans un intervalle selon chaque dimension. Un exemple en 2 dimensions est présenté sur la Figure 3.2. Le LHS permet ainsi d'assurer une bonne exploration de toutes les dimensions de l'espace des facteurs d'entrée à partir d'un échantillon de taille limitée. Toutefois, il est à noter que si l'exemple de la Figure 3.2 montre que la construction d'un LHS est aisée en 2 dimensions, celle-ci devient plus difficile dans le cas d'un espace de grande dimension. D'autre part, pour améliorer l'exploration de l'espace des facteurs d'entrée, on peut imposer des critères supplémentaires lors de la construction du LHS, comme les critères minimax ou maximin (Johnson et al., 1990) pour permettre de disperser les points au mieux.

3. Calcul des sorties du modèles Le modèle est utilisé pour obtenir les N valeurs des variables cibles correspondant aux N combinaisons de facteurs d'entrée déterminés à l'étape 2.

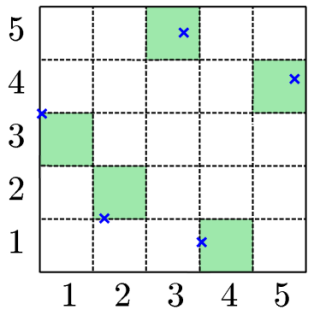


Figure 3.2 – Exemple d'échantillonnage dans un espace à 2 dimensions selon le principe de l'hypercube latin (Sudret et Marelli, 2020). Dans chaque intervalle défini en ligne et en colonne, il n'y a qu'un seul point échantillonné (indiqué par une croix bleue).

4. Description de la distribution de la variable de sortie Finalement, on décrit la distribution des variables de sortie d'intérêt par diverses quantités d'intérêt (moyenne, min, max, variance, quantiles, etc.) ou graphiquement avec un histogramme et une fonction de densité de probabilité (pdf) empirique estimée par méthode des noyaux (Rosenblatt, 1956).

3.3 Analyse de sensibilité

3.3.1 Introduction

L'analyse de sensibilité permet d'affecter chaque portion de la variabilité d'une variable de sortie aux différents facteurs d'entrée ou à leurs interactions. Les objectifs d'une telle démarche sont multiples (Saltelli et al., 2004) :

- La corroboration du modèle : explorer la relation entrée/sortie du modèle pour le valider, gagner en connaissances sur son comportement et sa structure et vérifier qu'il n'est pas trop dépendant d'hypothèses fragiles.
- La priorisation de la recherche : identifier des facteurs d'entrée qui méritent d'être analysés ou mesurés de manière plus approfondie.
- La simplification du modèle : identifier des facteurs d'entrée ou des compartiments du modèle qui peuvent être fixés ou simplifiés.
- L'identification de régions critiques ou autrement intéressantes dans l'espace des facteurs d'entrée : identifier des facteurs d'entrée qui intéragissent et qui pourraient ainsi générer des valeurs extrêmes (par exemple dans des études de fiabilité des systèmes).
- Avant l'estimation de paramètres ou de variables d'entrée, pour aider à mettre en place une expérience où la sensibilité de la sortie au facteur à estimer est la plus grande.

Selon les raisons qui motivent l'analyse de sensibilité, les méthodes utilisées ne sont pas les mêmes. Dans ces travaux, on fera exclusivement référence aux deux familles de méthodes

suivantes : les méthodes de **criblage** qui permettent de distinguer les facteurs influents des non influents et les méthodes de **classement** qui permettent de hiérarchiser les facteurs par ordre d'influence sur la sortie.

Dans le domaine de la modélisation des transferts de pesticides, Dubus et al. (2003) réalisait le premier une analyse de sensibilité locale, "One-At-a-Time" (OAT) sur quatre modèles de pesticides à l'échelle de la parcelle dans un objectif de classement. Dans ce type d'approche, l'influence de chaque facteur d'entrée est étudiée individuellement, autour d'une valeur nominale alors que les autres facteurs restent fixes. Ces méthodes ont le mérite d'être souvent simples mais elles supposent aussi que la relation entrée/sortie est linéaire (Saltelli et al., 2004 ; Nossent et Bauwens, 2012) ce qui est généralement faux pour des modèles environnementaux complexes. D'autre part, elles ne permettent pas de prendre en compte les effets des interactions entre facteurs d'entrée, ni d'explorer l'ensemble de l'espace multi-dimensionnel (Saltelli et al., 2019). Par la suite, pour dépasser les limitations de ces méthodes, les efforts se sont plutôt focalisés sur les méthodes d'analyse de sensibilité globales (GSA) "all-at-time". Celles-ci permettent de faire varier simultanément tous les facteurs d'entrée et d'explorer l'intégralité de leur gamme de variation. Comme elles sont souvent coûteuses à mettre en place et difficilement applicables quand le modèle comporte de nombreux facteurs d'entrée, une première étape de criblage qualitative et peu coûteuse est souvent réalisée afin de distinguer facteurs influents et non influents (Saltelli et al., 2004). La méthode des effets élémentaires, ou méthode de Morris (Morris, 1991) est classiquement utilisée (Fox et al., 2010 ; Vanuytrecht et al., 2014 ; Gatel et al., 2018 ; Lauvernet et Muñoz-Carpena, 2018). Dans un second temps, une méthode de classement plus coûteuse est appliquée. Parmi elles, les méthodes de décomposition de la variance comme les méthodes de Sobol ou extended Fast ont été largement exploitées dans le domaine de la modélisation hydrologique et des transferts de pesticides (Fox et al., 2010 ; Muñoz-Carpena et al., 2010 ; Hong et Purucker, 2018 ; Lauvernet et Muñoz-Carpena, 2018 ; Gatel et al., 2019). Elles fournissent notamment une information intéressante sur les effets interactifs entre facteurs d'entrée. Cependant, même si ces méthodes sont largement répandues, leur mise en oeuvre nécessite classiquement de gros échantillons et un tel coût de calcul ne peut pas toujours être assumé. D'autre part, elles supposent que la variance de la variable de sortie constitue un résumé pertinent de sa variabilité ce qui n'est pas toujours le cas. Pour dépasser ces limitations, il existe d'autres approches beaucoup plus récentes, avec d'autres définitions de la sensibilité encore peu appliquées aux modèles environnementaux. On citera parmi elles les mesures de dépendance (Da Veiga, 2015 ; De Lozzo et Marrel, 2014 ; De Lozzo et Marrel, 2016) ou les mesures d'importance basées sur une forêt aléatoire (Harper et al., 2011 ; Aulia et al., 2019 ; Bénard, 2021 ; Antoniadis et al., 2021).

Ainsi, pour mettre en place une analyse de sensibilité, il est indispensable de prendre en compte l'objectif et les caractéristiques de l'application. Pour PESHMELBA et notamment le scénario d'étude mis en place pour la thèse, on en rappelle les principales spécificités :

- Le scénario d'étude compte 145 paramètres d'entrée incertains.
- Le temps de calcul élevé d'une simulation PESHMELBA limite à quelques milliers la taille de tout échantillon.
- PESHMELBA est spatialisé, les variables de sortie d'intérêt (des variables intégrées temporellement ou des séries temporelles) le sont aussi.
- Compte tenu de la structure du modèle, les variables de sorties résultent souvent de l'interaction de nombreux processus physiques et des non linéarités et effets de seuil peuvent rendre complexe la caractérisation de la variabilité de la sortie.

Dans ces travaux, pour réduire le nombre élevé de paramètres d'entrée, on réalise d'abord une étape de criblage. Puis, pour garantir une analyse de sensibilité aussi exhaustive et robuste que possible, on explore et combine plusieurs méthodes de classement définissant différemment la notion de sensibilité : méthode de Sobol, mesure de dépendance et forêts aléatoires. Les différentes méthodes de criblage et de classement utilisées sont présentées dans la Section 3.3.2. Dans cette section, elles sont appliquées à des variables scalaires puis la Section 3.3.3 présente les outils utilisés pour les étendre à des variables de sortie spatialisées et temporelles.

3.3.2 Analyse de sensibilité globale de variables scalaires

Dans ce qui suit, on note $\mathrm{Y} \in \mathbb{R}$ la variable aléatoire scalaire d'intérêt provenant de PESHMELBA. On a $\mathrm{Y} = M(\underline{\mathbf{X}})$ où $M(.)$ est le modèle PESHMELBA et où $\underline{\mathbf{X}} = (\mathrm{X}_1, ..., \mathrm{X}_\mathrm{M}) \in \mathbb{R}^M$ est le vecteur aléatoire qui contient les facteurs d'entrée considérés pour ce cas d'étude. Dans ces travaux, on considère dans les facteurs d'entrée incertains la majorité des paramètres d'entrée du modèle mais pas les forçages climatiques. Le vecteur $\underline{\mathbf{X}}$ contient ainsi les 145 paramètres décrits dans le Tableau 2.1.

Criblage par effets élémentaires : la méthode de Morris

La méthode de Morris (Morris, 1991 ; Campolongo et al., 2007 ; Saltelli et al., 2008) est une méthode de criblage qui permet de détecter plus ou moins qualitativement les facteurs d'entrée importants à un coût très faible. Elle se base sur le calcul de variations de la variable de sortie entre deux points successifs (un effet élémentaire). Un tel calcul d'effet élémentaire est local mais les mesures sont ensuite moyennées sur plusieurs points pour s'extraire de la dépendance à un seul point d'échantillonnage, donnant ainsi un aspect plus global à la méthode.
Pour construire un échantillon de Morris, la gamme de variation de chaque facteur d'entrée est divisée en p niveaux avec un pas de discrétisation Δ. Le croisement de ces niveaux définit alors un ensemble de p^M noeuds. On définit ensuite R trajectoires qui parcourent chacune $M+1$ noeuds et qui sont construites de telle sorte que chaque facteur ne varie qu'une seule fois par trajectoire. Un exemple de trajectoire dans un hypercube de dimension $M = 3$, avec $p = 5$ niveaux est illustré Figure 3.3.

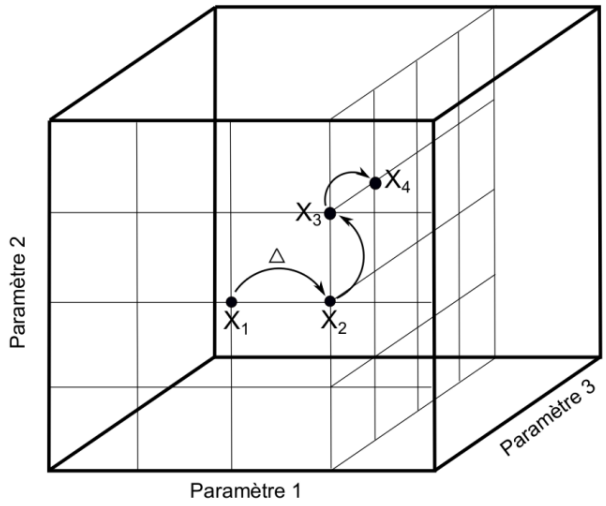


Figure 3.3 – Exemple de trajectoire de Morris échantillonnée dans un espace à trois dimensions avec 5 niveaux et un pas de discrétisation $\Delta = 0.25$ (tiré de Dairon 2015).

Au final, un échantillon de Morris contient ainsi les $N = R(M + 1)$ points parcourus par toutes les trajectoires. Pour garantir l'uniformité des points échantillonnés et ne privilégier aucune zone de l'espace d'entrée, le pas de discrétisation Δ est généralement choisi égal à $\frac{p}{2(p-1)}$ et Campolongo et al. (2007) suggère de choisir R supérieur à 10.

Les effets élémentaires sont ensuite calculés pour chaque facteur d'entrée X_i et pour chaque trajectoire r :

$$EE_i^r = \frac{M(X_1, X_2, ...X_i + \Delta, .., X_M) - M(X_1, X_2, ...X_i, .., X_M)}{\Delta} \tag{3.1}$$

Pour chaque facteur d'entrée, on calcule ensuite la moyenne absolue des effets élémentaires μ_i^* et son écart type σ_i. Finalement, les valeurs de μ_i^* et σ sont reportées respectivement en abscisse et en ordonné d'un même graphique (voir Figure 3.4) dont l'analyse permet de déterminer rapidement :

- les facteurs dont les effets sont négligeables : points proches de l'origine ;
- les facteurs dont l'effet linéaire est important en moyenne : poins situés à droite sur l'axe des abscisses ;
- les facteurs dont les effets sont non-linéaires et/ou incluent de l'interaction avec d'autres facteurs : points situés en haut à droite.

La méthode de Morris permet ainsi de réaliser un criblage visuel en éliminant les facteurs situés près de l'origine dans le graphe $\mu^* - \sigma$. Cependant, les résultats dépendent fortement du nombre de trajectoires et de niveaux fixés (Yang, 2011) et la distinction visuelle entre les différents domaines peut être floue s'il y a beaucoup de facteurs d'entrée considérés.
Par ailleurs, la calcul des mesures de sensibilité proposé dans la méthode originale conduisent

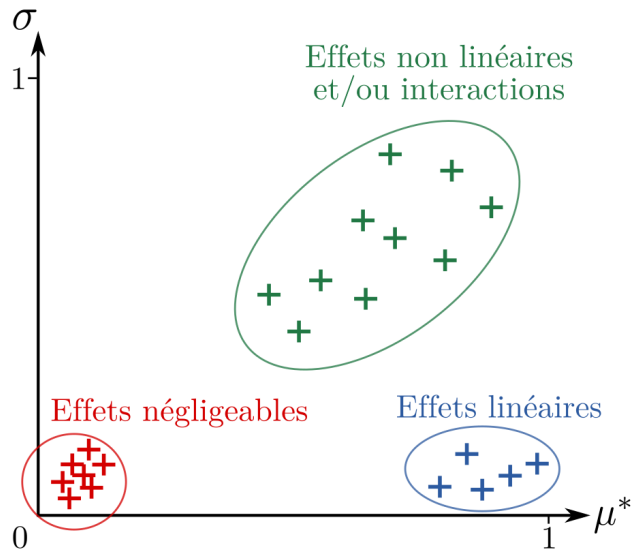


Figure 3.4 – Interprétation des clusters de paramètres dans un graphe de Morris (ici normé entre 0 et 1). μ^* est la moyenne absolue et σ est l'écart type des effets élémentaires.

parfois à une sélection peu fiable des paramètres importants (Campolongo et al., 2007) et l'échantillonnage est limité en termes d'efficacité. Cette méthode reste donc encore largement explorée afin d'en garantir la robustesse et l'efficacité, notamment dans des applications environnementales complexes (Campolongo et al., 2007 ; Saltelli et al., 2009 ; Campolongo et al., 2011 ; Ruano et al., 2012 ; Khare et al., 2015) .

Décomposition de la variance

Définition Les méthodes d'analyse de sensibilité basées sur la décomposition de la variance visent à déterminer dans quelles proportions les facteurs d'entrée contribuent à la variance de la sortie. Une des méthodes les plus populaires est la méthode de Sobol (Sobol, 1993). Elle est basée sur la décomposition ANOVA de la variable de sortie Y (ANalysis Of VAriances, Archer et al. 1997) en une somme de termes de dimensions croissantes :

$$\mathrm{Y} = M(x_1, ..., x_M) = m_0 + \sum_{s=1}^{M} \sum_{i_1<...<i_s}^{M} m_{i_1,...i_s}(x_{i_1}, ..., x_{i_s}) \tag{3.2}$$

où les termes $m_{i_1,...i_s}(x_{i_1}, ..., x_{i_s})$ sont définis tels que l'intégrale selon chacune de leur variable indépendante soit égale à zéro. Cette propriété implique que les termes sont orthogonaux, rendant la décomposition ANOVA unique. Si $\mathrm{Y} = M(\underline{\mathbf{X}})$ est de carré intégrable, tous les termes $m_{i_1,...,i_s}$ sont aussi de carré intégrable et en élévant au carré puis en intégrant l'Eq. (3.2), on obtient :

$$\int M^2(\underline{\mathbf{X}}) \mathrm{d}\underline{\mathbf{X}} = m_0^2 + \sum_{s=1}^{M} \sum_{i_1<...<i_s}^{M} \int m_{i_1,...,i_s}^2 \mathrm{d}x_{i_1} ... \mathrm{d}x_{i_s} \tag{3.3}$$

$\int M^2(\underline{\mathbf{X}})\mathrm{d}\underline{\mathbf{X}} - m_0^2$ est la variance totale de $M(\underline{\mathbf{X}})$ notée Var[Y] dans la suite. Si on introduit la notion de variance partielle $\mathrm{V}_{i_1,...,i_s}$ telle que :

$$\mathrm{V}_{i_1,...i_s} = \int m^2_{i_1,...,i_s}(x_{i_1},...x_{i_s})\mathrm{d}x_{i_1},...,\mathrm{d}x_{i_s} \text{ pour } 1 \leq i_1 < ... < i_s \leq M, s = 1,...,M \quad (3.4)$$

alors, l'Eq. (3.3) se lit :

$$\mathrm{Var[Y]} = \sum_{s=1}^{M} \sum_{i_1<...<i_s}^{M} \mathrm{V}_{i_1,...,i_s} \quad (3.5)$$

Dans cette formulation, $\mathrm{V}_{i_1,...i_s}$ indique la portion de variance qui peut être attribuée aux interactions entre les facteurs d'entrée $X_i, i \in i_1,...i_s$. La variance partielle d'indice unique V_i indique la portion de variance attribuée à l'effet direct de X_i, pris individuellement. A partir de là, on peut définir les indices de Sobol tels que :

$$\mathrm{S}_{i_1,...,i_s} = \frac{\mathrm{V}_{i_1,...i_s}}{\mathrm{Var}[Y]} \quad (3.6)$$

Par définition, $0 \leq \mathrm{S}_{i_1,...,i_s} \leq 1$. En particulier, les indices de premier ordre $\mathrm{S}_i = \frac{\mathrm{V}_i}{\mathrm{Var[Y]}}$ ne prennent en compte que les effets directs. Ils peuvent être interprétés comme la diminution de la variance totale qui pourrait être obtenue si X_i n'était pas incertain (Tarantola et al., 2002). Ces indices sont généralement calculés en premier lieu puisqu'ils expliquent souvent une portion importante de la variance de la sortie (Faivre et al., 2013).
Ensuite, les indices de Sobol totaux évaluent l'effet total d'un facteur d'entrée X_i sur la sortie en intégrant à la fois son effet direct S_i et tous les termes interactifs dans lesquels il intervient :

$$\mathrm{S}_{\mathrm{T}_i} = \sum_{\mathcal{I}_i} \mathrm{S}_{i_1,...,i_s}, \quad \mathcal{I}_i = \{(i_1,...,i_s) \mid \exists k, 1 \leq k \leq s, i_k = i\} \quad (3.7)$$

L'indice de sensibilité $\mathrm{S}_{\mathrm{T}_i}$ représente ainsi la portion de la variance de sortie qui subsiste tant que le facteur X_i demeure incertain.

Calcul des indices de Sobol par décomposition en polynômes du chaos Dans son papier original, Sobol (1993) proposait de calculer ses indices de sensibilité à partir d'une méthode de Monte Carlo. Cette méthode est basée sur deux échantillons de taille N. Seulement, une telle estimation des indices de sensibilité requiert de lancer le modèle 2^N fois. Saltelli et al. (2008) recommandent en particulier de choisir $N = k(M + 2)$ où k un nombre entre 500 et 1000. Dans le cas d'un modèle coûteux à lancer comprenant beaucoup de facteurs d'entrée comme c'est le cas pour PESHMELBA, un tel échantillon est impossible à obtenir. L'utilisation d'un métamodèle plus rapide pour remplacer le modèle initial est alors une alternative intéressante. C'est particulièrement le cas de la décomposition en Polynômes du Chaos (PC) puisque leur expression analytique permet de déduire directement les indices de Sobol. Le paragraphe suivant présente brièvement les PC et leur lien avec les indices de Sobol, en grande partie basé sur Le Gratiet et al. (2017).

Pour toute variable de sortie de carré intégrable $\mathrm{Y} = M(\underline{\mathbf{X}})$, $\mathrm{Y} \in \mathbb{R}$, sa décomposition en PC fournit une approximation basée sur sa projection dans une base adéquate composée de fonctions polynomiales dans les entrées aléatoires (Ghanem et Spanos, 1991) :

$$\mathrm{Y} = \sum_{\boldsymbol{\alpha} \in \mathbb{N}^M} \gamma_{\boldsymbol{\alpha}} \Psi_{\boldsymbol{\alpha}}(\underline{\mathbf{X}}) \tag{3.8}$$

où les $\Psi_{\boldsymbol{\alpha}}$ sont les polynômes multivariés orthonormaux qui constituent la base et les $\gamma_{\boldsymbol{\alpha}}$ sont les coordonnées correspondantes dans cette base. Les polynômes multivariés sont construits comme le produit tensoriel de polynômes univariés :

$$\Psi_{\boldsymbol{\alpha}}(\boldsymbol{x}) = \prod_{i=1}^{M} \psi_{\alpha_i}(x_i) \tag{3.9}$$

où $\psi_{\alpha_i}(x_i)$ est le polynôme univarié orthonormal de degré α_i associé à x_i. Ces polynômes univariés sont choisis parmi des familles classiques de polynômes (Legendre, Hermite, etc.), en fonction de la forme de la densité de probabilité marginale du facteur d'entrée considéré (Xiu et Karniadakis, 2002).

La décomposition en PC peut être réécrite en termes d'ordres croissants. Cette formulation se base sur les ensembles d'indices $\mathcal{A}_{i_1,...,i_s}$ contenant tous les tuples $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_M)$ où seuls les indices $(i_1, ..., i_s)$ sont non nuls :

$$\mathcal{A}_{i_1,...,i_s} = \left\{ \boldsymbol{\alpha} : \begin{array}{l} \alpha_k > 0,\ \forall k = 1, ..., M \ \mid\ k \in (i_1, ..., i_s) \\ \alpha_k = 0,\ \forall k = 1, ..., M \ \mid\ k \notin (i_1, ...i_s) \end{array} \right\} \tag{3.10}$$

Comme détaillé dans Sudret (2008), l'Eq. (3.8) peut être réécrite comme suit :

$$\begin{aligned} \mathrm{Y} = \gamma_0 + \sum_{i=1}^{M} \sum_{\boldsymbol{\alpha} \in \mathcal{A}_i} \gamma_{\boldsymbol{\alpha}} \psi_{\boldsymbol{\alpha}}(x_i) + \sum_{1 \leq i_1 < i_2 \leq M} \sum_{\boldsymbol{\alpha} \in \mathcal{A}_{i_1,i_2}} \gamma_{\boldsymbol{\alpha}} \psi_{\boldsymbol{\alpha}}(x_{i_1}, x_{i_2}) + ... + \\ \sum_{1 \leq i_1 < ... < i_s \leq M} \sum_{\boldsymbol{\alpha} \in \mathcal{A}_{i_1,...,i_s}} \gamma_{\boldsymbol{\alpha}} \psi_{\boldsymbol{\alpha}}(x_{i_1}, ..., x_{i_s}) + ... + \\ \sum_{\boldsymbol{\alpha} \in \mathcal{A}_{1,2,...,M}} \gamma_{\boldsymbol{\alpha}} \psi_{\boldsymbol{\alpha}}(x_1, ..., x_M) \end{aligned} \tag{3.11}$$

Par unicité de la décomposition ANOVA, l'Eq. (3.11) est la décomposition ANOVA de Y. Les indices de Sobol peuvent donc être déduits des coefficients de la décomposition en PC. Les polynômes multivariés impliqués dans cette dernière étant orthonormaux, la variance partielle n'est autre que la somme des carrés des coefficients des PC. En normalisant cette somme, on obtient :

$$\mathrm{S}_{i_1,...,i_s} = \sum_{\boldsymbol{\alpha} \in \mathcal{A}_{i_1,...,i_s}} \frac{\gamma_{\boldsymbol{\alpha}}^2}{\mathrm{Var[Y]}} \tag{3.12}$$

où $\mathrm{Var[Y]} = \sum_{\boldsymbol{\alpha} \in \mathcal{A}} \gamma_{\boldsymbol{\alpha}}^2$ est la variance totale. En particulier, les indices de premier ordre s'écrivent :

$$\mathrm{S}_i = \sum_{\boldsymbol{\alpha} \in \mathcal{A}_i} \frac{\gamma_{\boldsymbol{\alpha}}^2}{\mathrm{Var[Y]}}, \quad \mathcal{A}_i = \{\boldsymbol{\alpha} \in \mathbb{N}^M \mid \alpha_i > 0, \alpha_j = 0, \forall j \neq i\} \tag{3.13}$$

et les indices d'ordre total s'écrivent :

$$S_i^T = \sum_{\boldsymbol{\alpha} \in \mathcal{A}_i^T} \frac{\gamma_{\boldsymbol{\alpha}}^2}{\mathrm{Var[Y]}}, \quad \mathcal{A}_i^T = \{\boldsymbol{\alpha} \in \mathbb{N}^M \mid \alpha_i > 0\}. \tag{3.14}$$

Estimation Dans ces travaux, la décomposition en polynômes du chaos ainsi que les indices de Sobol sont calculés à l'aide du logiciel Matlab UQLab (Marelli et Sudret, 2014). Pour permettre le calcul pratique, l'Eq. (3.8) est tronquée en une somme finie :

$$\mathrm{Y} \approx \sum_{\boldsymbol{\alpha} \in \mathcal{A}} \gamma_{\boldsymbol{\alpha}} \Psi_{\boldsymbol{\alpha}}(\underline{\mathbf{X}}) \tag{3.15}$$

où $\mathcal{A} \subset \mathbb{N}^M$ est le sous-ensemble des tuples sélectionnés. Plusieurs schémas de troncature existent et l'un des plus populaires consiste à plafonner le degré maximal avec lequel chaque facteur d'entrée X_i est représenté dans la base. On fixe alors p tel que pour tout i, $\alpha_i < p$. Pour limiter encore le nombre d'éléments de la base de décompostion, on utilise ici un schéma de troncature hyperbolique qui privilégie les termes polynomiaux de haut degré, inférieur à p, en une variable puis des termes mixtes incluant peu d'interactions. Ce schéma de troncature se base sur le constat récurrent que seul les effets interactifs d'ordre les plus faibles (c'est-à-dire impliquant peu de facteurs) contribuent à la variance de la sortie et doivent donc

être représentés (on parle de *sparsity of effets*, Le Gratiet et al. 2017). Ceci est fait en introduisant $q \in (0, 1)$ et en imposant que la norme q du vecteur $\boldsymbol{\alpha}$ soit inférieure à p :

$$\mathcal{A}^{M,p} = \{\boldsymbol{\alpha} \in \mathbb{N}^M : ||\boldsymbol{\alpha}||_q \leq p\}, \text{ avec } ||\boldsymbol{\alpha}||_q = \left(\sum_{i=1}^{M} \alpha_i^q\right)^{\frac{1}{q}} \tag{3.16}$$

Plusieurs approches sont proposées dans UQLab pour évaluer les coefficients $\gamma_{\boldsymbol{\alpha}}$ de la décomposition en PC et on utilise dans ces travaux une procédure de régression linéaire sur un échantillon d'apprentissage adaptée à l'hypothèse de *sparsity of effects* (LARS, voir Blatman et Sudret 2011 pour plus de détails sur cette procédure). Plus précisément, c'est une version de la procédure LARS avec degré p et norme q adaptatifs qui est utilisée.

Mesures de dépendance

Bien qu'extrêmement répandue, l'utilisation des indices de Sobol suppose que la variance de la sortie d'un modèle suffit pour résumer sa variabilité. Or, cette hypothèse peut s'avérer particulièrement limitante dans certains cas (par exemple dans le cas de variables multi-modales ou non symétriques...). Ainsi, des méthodes récentes proposent plutôt de mesurer de manière plus globale la dépendance entre les variables aléatoires $\mathrm{X_i}$ et Y. Il existe une grande variété de mesures de dépendance et on s'intéresse ici particulièrement au Critère d'Indépendance d'Hilbert-Schmidt (HSIC). L'idée générale derrière la mesure HSIC est de calculer la covariance entre n'importe quelle transformation non linéaire d'un facteur d'entrée X_i et n'importe quelle transformation non linéaire de la sortie Y (De Lozzo et Marrel, 2016). Ainsi, une telle mesure de dépendance capture simultanément un très large spectre de formes de dépendance entre les variables (Meynaoui et al., 2018). Au delà de l'analyse de sensibilité, l'utilisation des mesures HSIC est donc plébiscitée dans de nombreux problèmes de classification, régression, identification d'images, sélection de gènes, etc. (Gangeh et al., 2017 ; Wan-Duo et al., 2018 ; Novello et al., 2022) .

Définition Le principe de la mesure HSIC est résumé de manière schématique sur la Figure 3.5. Celui-ci repose sur la théorie des espaces de Hilbert à noyaux reproduisants (RKHS) et sur l'utilisation de fonctions noyaux. Ces notions sont détaillées dans l'Annexe B.

Notons $\mathcal{F}_i$ le RKHS composé de toutes les fonctions continues, bornées prenant X_i comme entrée, à valeur dans $\mathbb{R}$ et $\mathcal{G}$ le RKHS composé des fonctions bornées de la sortie Y également à valeur dans $\mathbb{R}$. $\langle \cdot, \cdot \rangle_{\mathcal{F}_i}$ (resp. $\langle \cdot, \cdot \rangle_{\mathcal{G}}$) est le produit scalaire sur $\mathcal{F}_i$ (resp. $\mathcal{G}$) et κ_{X_i} (resp. κ_Y) est la fonction noyau correspondante. Cela signifie que les éléments de $\mathcal{F}_i$ et $\mathcal{G}$ permettent de décrire l'ensemble des transformations non linéaires de X_i et Y. On peut alors exprimer la covariance entre n'importe quelles fonctions f et g dans ces RKHS, soit $\mathrm{cov}(f(\mathrm{X}_i), g(\mathrm{Y}))$. Pour calculer ces covariances, on définit l'opérateur de covariance $C[\mathcal{G}\mathcal{F}_i] : \mathcal{G} \to \mathcal{F}_i$ comme

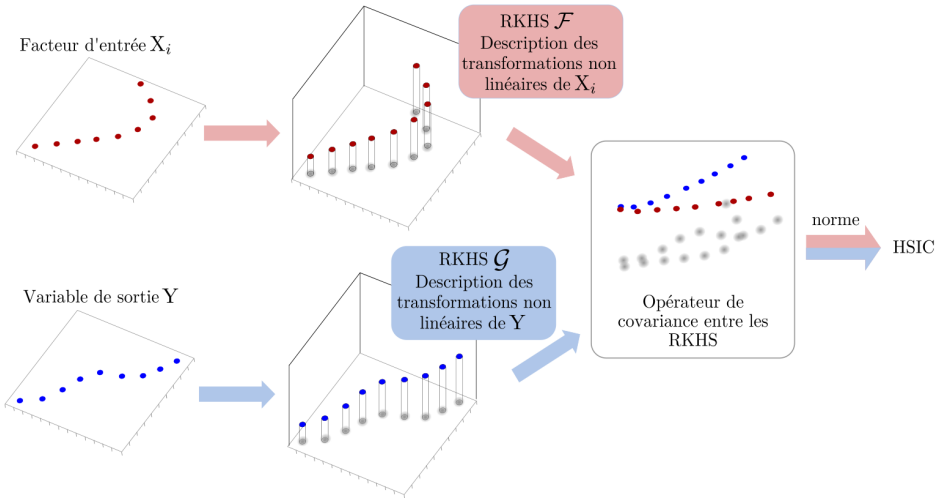


Figure 3.5 – Illustration du principe de la mesure HSIC calculée entre 2 variables X_i et Y. On note que ces dernières sont représentées dans le plan $\mathbb{R}^2$ pour simplifier la représentation mais qu'elles peuvent en réalité appartenir à n'importe quel espace (adapté de Jimenez et al. 2018).

étant l'unique opérateur de $\mathcal{G}$ à $\mathcal{F}_i$ tel que :

$$\forall f_i \in \mathcal{F}_i, \forall g \in \mathcal{G}, \langle f, C[\mathcal{G}\mathcal{F}_i](g)\rangle_{\mathcal{F}_i} = \text{cov}(f_i(X_i), g(Y)) \tag{3.17}$$

La mesure HSIC correspond alors au carré de la norme d'Hilbert-Schmidt de l'opérateur de covariance $C[\mathcal{G}\mathcal{F}_i]$. Par définition, la norme d'Hilbert-Schmidt d'un opérateur linéaire $C : \mathcal{F}_i \to \mathcal{G}$ s'écrit :

$$||C||^2_{HS} = \sum_{j,k} \langle u^i_j, C(v_k)\rangle_{\mathcal{F}_i} \tag{3.18}$$

où $(u^i_j)_{j\geq 0}$ et $(v_k)_{k\geq 0}$ sont les bases orthogonales de $\mathcal{F}_i$ et $\mathcal{G}$ respectivement. La norme d'Hilbert-Schmidt généralise ainsi la notion de norme de Frobenius [2] définie pour les matrices.

La mesure HSIC s'écrit alors :

$$HSIC(X_i, Y)_{\mathcal{F}_i,\mathcal{G}} = ||C[\mathcal{G}\mathcal{F}_i]||^2_{HS} = \sum_{j,k} \langle u^i_j, C[\mathcal{G}\mathcal{F}_i](v_k)\rangle_{\mathcal{F}_i} = \sum_{j,k} \text{cov}(u^i_j(X_i), v_k(Y)) \tag{3.19}$$

La mesure de dépendance HSIC dépend fortement du choix des RKHS $\mathcal{F}_i$ et $\mathcal{G}$ et plus précisément du produit scalaire définissant la relation entre les éléments de ces RKHS. C'est la fonction noyau qui définit un tel produit scalaire. Comme proposé dans De Lozzo et Marrel (2014) et De Lozzo et Marrel (2016) et Da Veiga (2015), on choisit ici un noyau gaussien, souvent associé à de bonnes performances à la fois pour des variables scalaires

2. Pour une matrice $\mathbf{A} \in M_{mn}(\mathbb{R})$, sa norme de Frobénius est définie par $||A||_F = (\sum_{i=1,j=1}^{m,n} |a_{ij}|)^{\frac{1}{2}}$.

ou vectorielles (Gretton et al., 2005). Dans le cas d'une variable vectorielle, $\mathbf{x} \in \mathbb{R}^d$, il s'exprime comme suit :

$$\kappa(\mathbf{x}, \mathbf{x}') = \exp(-\lambda||\mathbf{x} - \mathbf{x}'||_2^2), \tag{3.20}$$

où $||.||_2$ représente la norme euclidienne dans $\mathbb{R}^d$ [3] et où l'hyperparamètre λ est appelé largeur de bande du noyau. Dans cette étude, la largeur de bande λ est estimée à partir de l'inverse de l'écart-type de l'échantillon.
On note enfin qu'il existe de nombreux autres noyaux (linéaire, polynomial, Laplacien, etc.) et l'une des forces des méthodes à noyaux est que selon le noyau considéré, il est possible de considérer des variables scalaires, fonctionnelles, catégoriques, voire des densités de probabilité (Da Veiga et al., 2021).

Estimation de la mesure de dépendance HSIC La mesure de dépendance HSIC peut être écrite en termes de noyaux (Gretton et al., 2005) :

$$\begin{aligned} HSIC(\mathrm{X}_i, \mathrm{Y})_{\mathcal{F}_i,\mathcal{G}} = &\mathbb{E}[\kappa_{\mathrm{X}_i}(\mathrm{X}_i, \mathrm{X}_i')\kappa_{\mathrm{Y}}(\mathrm{Y}, \mathrm{Y}')] \\ &+ \mathbb{E}[\kappa_{\mathrm{X}_i}(\mathrm{X}_i, \mathrm{X}_i')]\mathbb{E}[\kappa_{\mathrm{Y}}(\mathrm{Y}, \mathrm{Y}')] \\ &- 2\mathbb{E}[\mathbb{E}[\kappa_{\mathrm{X}_i}(\mathrm{X}_i, \mathrm{X}_i')|\mathrm{X}_i]\mathbb{E}[\kappa_{\mathrm{Y}}(\mathrm{Y}, \mathrm{Y}')|\mathrm{Y}]] \end{aligned} \tag{3.21}$$

où κ_{X_i} (resp. κ_Y) est la fonction noyau définissant le RKHS associé à X_i (resp Y), $\underline{\mathbf{X}}' = (\mathrm{X}_1', .., \mathrm{X}_i', ..., \mathrm{X}_M')$ est une copie indépendante et identiquement distribuée de $\underline{\mathbf{X}} = (\mathrm{X}_1, .., \mathrm{X}_i, ..., \mathrm{X}_M)$ et Y' est la sortie associée à $\underline{\mathbf{X}}'$. L'Eq. (3.21) signifie qu'un estimateur $\widehat{HSIC}(\mathrm{X}_i, \mathrm{Y})$ de la mesure HSIC qui nécessite de calculer une norme sur un espace de dimension possiblement infinie peut être calculé uniquement à partir des fonctions noyaux appliquées à un échantillon fini de N points $(x_i^j, y^j), j \in \{1, .., N\}$ de $(\mathrm{X}_i, \mathrm{Y})$. L'estimateur résultant $\widehat{HSIC}(\mathrm{X}_i, \mathrm{Y})$ s'exprime ainsi (Gretton et al., 2005) :

$$\widehat{HSIC}(\mathrm{X}_i, \mathrm{Y})_{\mathcal{F}_i,\mathcal{G}} = \frac{1}{(N-1)^2} Tr(KHLH) \tag{3.22}$$

où $H \in \mathbb{R}^{N \times N}$ est la matrice de centrage $H_{jk} = \delta_{jk} - \frac{1}{N}$ et où $K \in \mathbb{R}^{N \times N}$ et $L \in \mathbb{R}^{N \times N}$ sont les matrices de Gram définies par $K_{jk} = \kappa_{X_i}(x_i^j, x_i^k)$ et $L_{jk} = \kappa_Y(y^j, y^k)$.
La mesure de dépendance HSIC peut ainsi être calculée facilement et rapidement à partir d'un échantillon préexistant. Toutefois, De Lozzo et Marrel (2014) font remarquer qu'elle peut être fortement biaisée si la taille de l'échantillon est trop petite.
Dans ces travaux, le script R fourni dans De Lozzo et Marrel (2016) (voir supplementary material) a été adapté en Python pour le calcul des mesures de dépendance HSIC.

Utilisation des mesures HSIC pour le classement et le criblage Les mesures HSIC peuvent être utilisées dans un contexte d'analyse de sensibilité, avec un objectif de criblage

3. $\forall \mathbf{x} \in \mathbb{R}^d, ||\mathbf{x}||_2 = \sqrt{\sum_{i=1}^d |x_i^2|}$

ou de classement :

1. L'utilisation de la mesure HSIC pour le **criblage** impose de choisir un noyau universel (cf Annexe B), lequel permet de caractériser pleinement l'indépendance de deux variables. Si les noyaux définissant les RKHS $\mathcal{F}_i$ et $\mathcal{G}$ sont universels, la mesure HSIC(X_i,Y) est égale à 0 si et seulement si X_i et Y sont indépendantes (Gretton et al., 2005 ; Da Veiga, 2015). Le noyau gaussien utilisé dans cette étude fait notamment partie des noyaux universels.
   Pour réaliser le criblage, De Lozzo et Marrel (2014) proposent une approche basée sur un test statistique d'indépendance. Pour chaque facteur d'entrée X_i, on considère un échantillon de N points $(x_i^1, ...x_i^N)$ et les sorties correspondantes $(y^1, ..., y^N)$ et on teste l'hypothèse suivante *H0: "*X_i *et Y sont indépendants"*. Pour cela, on calcule d'abord un estimateur $\widehat{HSIC}(\mathrm{X}_i, \mathrm{Y})$ de la mesure $HSIC(\mathrm{X}_i, \mathrm{Y})$. Ensuite, B versions bootstrappées de l'échantillon original $(y^1, ..., y^N)$ sont générées. Ces B ensembles d'échantillons $\underline{\mathbf{y}}^{[1]}$,...,$\underline{\mathbf{y}}^{[B]}$ sont générés avec remplacement de manière à contenir chacun le même nombre de points que l'échantillon original. Pour chaque $\underline{\mathbf{y}}^{[B]}$, les points de l'échantillon associés à X_i ne sont pas rééchantillonnés. Une illustration sur un exemple avec $N = 5$ et $B = 3$ est présenté sur la Figure 3.6 Ainsi, en considérant l'hypothèse d'indépendance, n'importe quelle valeur de Y peut être associée à X_i. Pour chaque bootstrap b, un estimateur $\widehat{HSIC}^{[b]}(\mathrm{X}_i, \mathrm{Y})$ est alors calculé. On calcule ensuite la probabilité critique (p-val) correspondante comme suit :

$$p\text{-val}_B = \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}_{\widehat{HSIC}^{[b]}(\mathrm{X}_i,\mathrm{Y}) > \widehat{HSIC}(\mathrm{X}_i,\mathrm{Y})} \tag{3.23}$$

   En notant α le niveau de risque, si p-val$_B$<α, l'hypothèse d'indépendance est rejetée, sinon elle est acceptée.

2. Pour le **classement**, les indices de sensibilité proposés par Da Veiga (2015) sont définis pour chaque facteur d'entrée $\mathrm{X}_i, i \in \{1, ..., M\}$ en normalisant la mesure HSIC comme suit :

$$S^2_{\mathrm{X}_i} = \frac{HSIC(\mathrm{X}_i, \mathrm{Y})_{\mathcal{F}_i,\mathcal{G}}}{\sqrt{HSIC(\mathrm{X}_i, \mathrm{X}_i)_{\mathcal{F}_i,\mathcal{F}_i} HSIC(\mathrm{Y}, \mathrm{Y})_{\mathcal{G},\mathcal{G}}}} \tag{3.24}$$

   Là encore, les mesures HSIC intervenant dans l'Eq. (3.24) sont estimées à partir de l'Eq. (3.22).

Ainsi, les indices de sensibilité basés sur la mesure HSIC permettent de mesurer des niveaux de dépendance globale mais il est important de noter qu'ils ne fournissent pas une information précise sur le ou les type(s) de dépendance(s) prédominante(s) dans la relation entre X_i et Y. Ils ne donnent pas non plus d'information sur les effets interactifs entre paramètres d'entrée comme dans la méthode de Sobol. Les travaux très

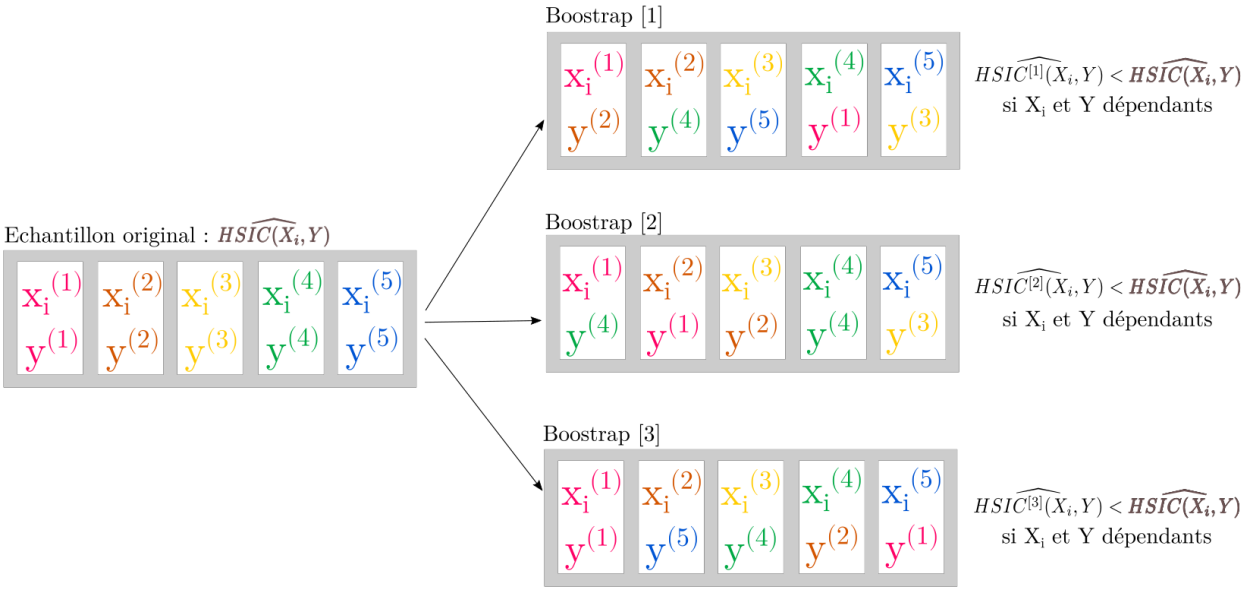


Figure 3.6 – Exemple de rééchantillonnage par bootstraps dans le cas $N = 5$ et $B = 3$ pour le test d'indépendance basée sur la mesure HSIC.

récents de Da Veiga (2021) ouvrent cependant cette perspective en proposant une décomposition ANOVA de ces indices HSIC. Ces travaux pourraient ainsi permettre de se rapprocher de l'interprétation classique des indices de Sobol en voyant les indices de sensibilité HSIC comme des pourcentages de dépendance HSIC expliqués par les différents sous-groupes de paramètres.

Mesures d'importance obtenues par forêts aléatoires

Formulation générale Les forêts aléatoires ou random forests (RF, Breiman 2001) constituent une méthode de métamodélisation issue du machine learning. Construire un métamodèle par forêt aléatoire consiste à moyenner les résultats de K arbres de décision créés indépendamment comme illustré sur la Figure 3.7.

Un arbre de décision est composé d'un ensemble de conditions binaires discriminantes dénommées noeuds de l'arbre. Ces conditions sont appliquées de manière hiérarchique depuis un noeud racine jusqu'à un noeud terminal (feuille de l'arbre). A chaque noeud, l'espace des facteurs d'entrée est partitionné en deux groupes plus restreints selon les valeurs de la variable de réponse Y. Ces partitionnements se succèdent jusqu'à atteindre un nombre seuil minimal de membres à un noeud donné.
Plusieurs arbres sont ensuite combinés pour construire la forêt aléatoire. On atténue ainsi la forte sensibilité d'un arbre de décision individuel à l'échantillon utilisé pour l'entraîner. Pour éviter les corrélations entre les points d'un tel échantillon et assurer la stabilité du modèle final, les RF se basent également sur le principe de "bagging". Le bagging consiste à entraîner chaque arbre de décision à partir d'un échantillon différent plus petit que l'échantillon d'origine (voir Figure 3.7). Ces sous-échantillons sont construits à partir de l'original en ré-

échantillonnant avec remplacement, faisant ainsi que certains points sont utilisés plusieurs fois alors que d'autres ne le sont potentiellement pas du tout. Cette méthode est plébiscitée par Breiman (1996) et Breiman (2001), notamment pour rendre les RF plus robustes en cas de légères variations dans l'espace d'entrée exploré et globalement plus précis.
Les échantillons qui ne sont pas utilisés pour construire chaque arbre constituent l'échantillon "out-of-bag" (OOB) et peuvent être réutilisés comme des sous-échantillons de test pour chaque arbre entraîné sur l'échantillon "InBag" correspondant.
A noter qu'on ne considère ici que des problèmes de régression où les RF sont utilisés pour des variables de réponse continues.

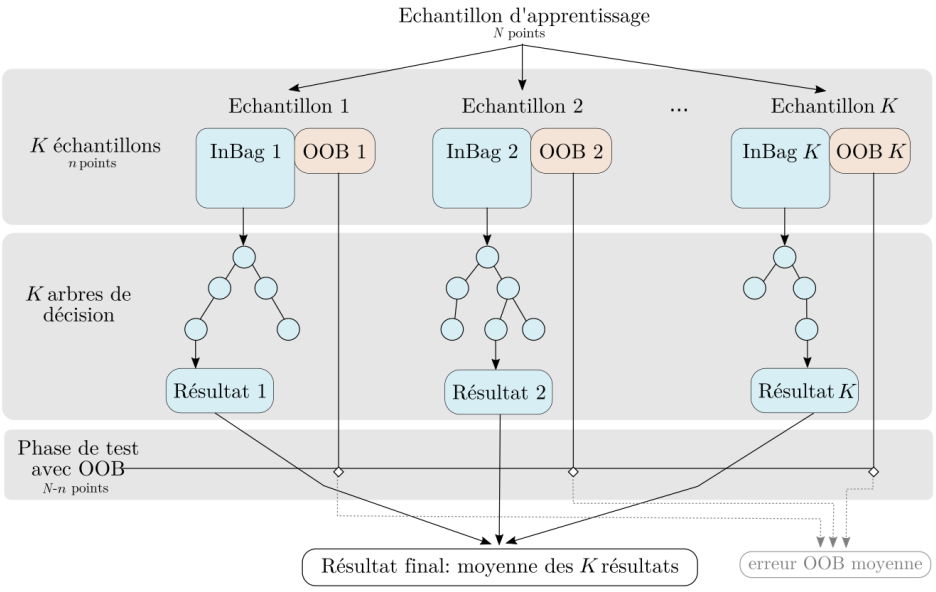


Figure 3.7 – Structure et éléments constitutifs d'un métamodèle construit par forêt aléatoire (adapté de Rodriguez-Galiano et al. 2014).

Utilisation des mesures d'importance pour l'analyse de sensibilité Il est possible d'extraire de l'information sur l'importance de chaque facteur d'entrée pour la construction de la forêt aléatoire et une telle information peut être interprétée comme un indice de sensibilité (Gregorutti et al., 2017 ; Antoniadis et al., 2021). Pour obtenir une telle *mesure d'importance*, la forêt aléatoire est d'abord entraînée à partir d'un échantillon d'apprentissage de taille N $(\underline{\mathbf{x}}^j, \mathrm{y}^j)$ for $j \in \{1, ..., N\}$. Une fois le RF construit, pour chaque facteur d'entrée X_i, les valeurs correspondantes sont permutées individuellement dans l'échantillon d'apprentissage de manière à supprimer le lien entre X_i et Y dans l'échantillon. L'effet d'une telle permutation sur la précision de la forêt aléatoire est alors quantifié : une forte diminution de la précision indique que le facteur X_i est fortement influent sur la variable de sortie alors qu'un faible impact sur cette précision indique une faible influence. La métrique utilisée pour quantifier l'impact sur la prédiction de la forêt aléatoire est le Mean Decrease in Accuracy (MDA, voir Bénard et al. (2021) pour une revue complète des différentes formulations de

ce MDA dans les packages R et Python). Le MDA utilisé ici correspond à la formulation originale décrite dans l'article de Breiman (2001). Celui-ci est calculé comme l'écart quadratique entre les prédictions des points de l'OOB avant et après permutation, moyenné sur tous les arbres de la forêt aléatoire. Dans ces travaux, le package R randomForestSRC (Ishwaran et Kogalur, 2020) est utilisé pour calculer les mesures d'importance. L'algorithme correspondant est largement décrit dans la littérature (voir par exemple Soleimani 2021 ou Bénard et al. 2021) et il est brièvement rappelé ci-dessous :

1. Pour chaque arbre k :
   - Estimer $\hat{\epsilon}_{OOB_k}$ l'erreur associée à l'échantillon OOB_k:

$$\hat{\epsilon}_{OOB_k} = \frac{1}{|\mathrm{OOB}_k|} \sum_{j \setminus (\underline{\mathbf{x}}^j, y^j) \in \mathrm{OOB}_k} (y^j - \hat{M}_{RF}(\underline{\mathbf{x}}^j))^2 \tag{3.25}$$

     où $\hat{M}_{RF}$ est le métamodèle estimé par forêt aléatoire.
   - Pour chaque facteur d'entrée X_i :
     - Permuter aléatoirement x_i dans $\{\underline{\mathbf{x}}^j \in \mathrm{OOB}_k\}$ pour générer un nouvel échantillon $\{\underline{\mathbf{x}}^{j*} \in \mathrm{OOB}_k\}$.
     - Estimer $\hat{\epsilon}^*_{OOB_k}(i)$ à partir de l'échantillon permuté :

$$\hat{\epsilon}^*_{OOB_k}(i) = \frac{1}{|\mathrm{OOB}_k|} \sum_{j \setminus (\underline{\mathbf{x}}^j, y^j) \in \mathrm{OOB}_k} (y^j - \hat{M}_{RF}(\underline{\mathbf{x}}^{j*}))^2 \tag{3.26}$$

2. Pour chaque facteur d'entrée X_i :
   - Calculer le Mean Decrease in Accuracy MDA_i :

$$MDA_i = \frac{1}{K} \sum_{k=1}^{K} \hat{\epsilon}_{OOB_k} - \hat{\epsilon}^*_{OOB_k}(i) \tag{3.27}$$

     où K est le nombre total d'arbres dans la forêt aléatoire

Malgré l'aspect boîte noire de la structure des forêts aléatoires, des travaux récents établissent un lien théorique entre Mean Decrease in Accuracy et indices de Sobol totaux lorsque les facteurs d'entrée sont indépendants. En particulier, les travaux de Gregorutti et al. (2017) établissent la relation suivant pour chaque facteur d'entrée X_i :

$$ST_i = \frac{MDA_i}{2\mathrm{Var}[\mathrm{Y}]} \tag{3.28}$$

En conclusion, les forêts aléatoires sont capables de métamodéliser tout type de relation entrées/sortie sans présupposer de la forme de cette relation, ce qui les rend particulièrement attractives. Par ailleurs, la construction de ce métamodèle fournit pour un moindre coût

supplémentaire des indices de sensibilité potentiellement interprétables comme une décomposition de la variance. Ceci en fait une méthode qui vaut la peine d'être explorée dans ce cas d'étude.

3.3.3 Extension aux variables spatiotemporelles

Les variables considérées dans PESHMELBA sont généralement des variables intégrées ou des séries temporelles spatialisées. Les différentes approches introduites dans la Section 3.3.2 doivent ainsi être étendues pour considérer des sorties vectorielles $\underline{\mathbf{Y}} \in \mathbb{R}^d$. Dans ce qui suit, on traite séparément aspect temporel et spatial en considérant que la principale différence entre les deux est la dimension d de l'espace de sortie considéré. L'accent est mis sur l'extension des indices de Sobol au cas vectoriel car ces derniers sont largement utilisés dans la communauté hydrologique. Leur interprétation transparente et l'évaluation fine des interactions qu'ils proposent restent en effet des arguments largement avancés pour justifier leur utilisation plutôt que celle de mesures de dépendance ou d'importance dont l'interprétation reste encore souvent mal comprise.

Variables spatialisées

Pour les indices de Sobol, Gamboa et al. (2013) propose des indices de Sobol aggrégés[4] ASI_u pour des sorties vectorielles, pour tout sous-ensemble de facteurs d'entrée $u = \{i_1, .., i_s\}$. Cette expression consiste en une moyenne des indices de Sobol calculés sur chaque composante du vecteur et pondérée par les variances locales :

$$ASI_u = \frac{\sum_{j=1}^{d} \mathrm{Var}[\mathrm{Y}_j] S_u^{(j)}}{\sum_{j=1}^{d} \mathrm{Var}[\mathrm{Y}_j]} \tag{3.29}$$

où $\mathrm{Var}[\mathrm{Y}_j]$ est la variance de la $j^{\text{ème}}$ composante scalaire de $\underline{\mathbf{Y}}$ et $S_u^{(j)}$ et l'indice de Sobol de Y_j associé au groupe de facteurs u. Cette formulation s'applique notamment au calcul des indices de premier ordre et des indices totaux. Cependant, l'Eq. (3.29) suppose de calculer des indices de Sobol associés à chaque composante pour calculer les indices aggrégés, ce qui peut se révéler coûteux en grande dimension. Cette difficulté est abordée dans Gamboa et al. (2013) qui proposent un estimateur pick-freeze permettant de calculer directement de tels indices aggrégés.

Variables temporelles

Pour les variables dynamiques $Y(t) = M(\mathbf{X}, t)$, $t \in \tau$, la même approche que Gamboa et al. (2013) peut être appliquée mais la taille du vecteur $\underline{\mathbf{Y}}$ résultant est souvent bloquante

4. On utilise ici la dénomination *indices aggrégés* par cohérence avec Radišić et al. (2022) et pour différencier ces indices de ceux proposés par Lamboni et al. (2011) mais les travaux initiaux de Gamboa et al. (2013) utilisent plutôt le terme d'*indices généralisés*.

pour calculer les indices aggrégés, que ça soit à partir des indices locaux ou avec l'estimateur pick-freeze. Pour contourner cette difficulté, une méthode classique, entre autre recommandée par Lamboni et al. (2011) et Da Veiga et al. (2021) consiste à réaliser une Analyse en Composantes Principales fonctionnelle (ACPf) de la série temporelle. Celle-ci consiste à trouver un sous espace de représentation qui permet de décrire de manière parcimonieuse la variabilité de la sortie dynamique. Pour toute variable aléatoire dynamique de carré intégrable, on a :

$$Y(t) = \mu(t) + \sum_{j \in \mathbb{N}} H_j \mathbf{v}_j(t) \tag{3.30}$$

où $\mu(t) = \mathbb{E}[Y(t)]$, $\mathbf{v}_j(t)$ est la $j$ième composante principale et H_j est le score sur la $j$ième composante principale (ou encore H_j est la coordonnée selon le vecteur $\mathbf{v}_j(t)$ dans la base $\{\mathbf{v}_j(t)\}_j$) (Ramsay et Silverman, 2005). La somme de l'Eq. (3.30) est généralement tronquée aux J premières composantes principales pour représenter un pourcentage fixe de l'inertie de Y ou ne conserver qu'un nombre précis de composantes principales. Ainsi, l'analyse de sensibilité de la sortie Y(t) est remplacée par l'analyse de sensibilité des scores $\{H_j\}_{1 \leq j \leq J}$. Dans ces travaux, l'ACPf est réalisée à partir de la librarie R *fda* et de la fonction *pca.fd*
Pour obtenir une information synthétisée sur la sensibilité de la série temporelle, Lamboni et al. (2011) proposent alors de calculer des indices généralisés pour tout groupe de facteurs d'entrée u :

$$GSI_u = \frac{\sum_{j=1}^{J} \lambda_j S_u^{(j)}}{\sum_{j=1}^{J} \lambda_j} \tag{3.31}$$

où $S_u^{(j)}$ est l'indice de Sobol de H_j associé au groupe de facteurs d'entrée u et λ_j la part d'inertie de la sortie expliquée par la composante principale $\mathbf{v}_j$.

3.4 Méthologie mise en place dans la thèse

3.4.1 Variables cibles

Les variables cibles considérées dans ces travaux sont de deux types : variables intégrées et séries temporelles et la quantification de leur incertitude répond à deux objectifs différents. On s'intéresse tout d'abord à des variables intégrées représentatives des processus physiques simulés dans PESHMELBA. Parmi elles, les flux cumulés correspondant aux échanges latéraux entre UH, en surface par le ruissellement et en subsurface par les échanges saturés sont particulièrement ciblés. L'intégration des échanges latéraux de surface et de subsurface constituent en effet une des spécificités de PESHMELBA et l'approche de modélisation, relativement originale, mérite d'être explorée. L'analyse de sensibilité est ici appliquée pour mieux comprendre et pour valider le comportement du modèle, en particulier le couplage des différents processus physiques.
On considère donc 4 variables scalaires intégrées temporellement :

- les volumes d'eau cumulés transférés par échanges latéraux saturés de subsurface (variable **WaterLateralFlow**) ;
- les volumes d'eau cumulés transférés par ruissellement de surface (variable **WaterSurfaceRunoff**) ;
- les masses de pesticides cumulées transférées par échanges latéraux saturés de subsurface (variable **SoluteLateralFlow**) ;
- les masses de pesticides cumulées transférées par ruissellement de surface (variable **SoluteSurfaceRunoff**).

Ces variables sont intégrées temporellement mais leur variabilité spatiale est préservée pour explorer l'aspect distribué du modèle. On dispose donc de 4 variables pour chaque UH du bassin versant.

En plus des variables représentatives du comportement du modèle, on cible également des variables dont on souhaite réduire l'incertitude dans la deuxième partie de ces travaux. Cette étape vise à les caractériser au mieux pour choisir une méthode d'assimilation adaptée et pour construire un vecteur d'état pertinent. Pour cela, on cherche en particulier à identifier si ces variables sont globalement gaussiennes (analyse d'incertitude) puis à identifier quels sont les principaux paramètes d'entrée auxquels elles sont sensibles (analyse de sensibilité). On s'intéresse ici aux séries temporelles suivantes :

- humidité de surface ;
- humidité de subsurface (entre autre à 10 cm et à 2 m de profondeur) ;
- concentration journalière en pesticides dans la rivière.

Là encore, les séries temporelles d'humidité sont spatialisées. Par contre, la concentration journalière en pesticides n'est analysée qu'à l'exutoire.

3.4.2 Echantillonnage et analyse d'incertitude

Pour l'analyse d'incertitude, on reprend les 4 étapes présentées dans la Section 3.2 et on considère incertains les 145 paramètres d'entrée définis dans le Chapitre 2. Leurs distributions sont celles qui sont définies dans le Chapitre 2 et détaillées dans l'Annexe A. On rappelle que pour cette application, les facteurs d'entrée sont échantillonnés de manière indépendante bien que cette hypothèse soit en réalité discutable, notamment pour les caractéristiques hydrodynamiques des sols. Un hypercube latin est utilisé pour permettre un échantillonage efficace et les simulations sont lancées sur le cluster HIICS, à INRAE Montpellier (26 noeuds, 692 coeurs, 64 à 256 GO de RAM par serveur). Compte tenu du temps de calcul d'une simulation PESHMELBA (temps apparent entre 45 min et 120 min selon les noeuds considérés), on limite à 4000 points la taille maximale des échantillons considérés. Enfin, pour décrire l'incertitude des variables de sortie, on décrit leurs premiers moments (moyenne et écart-type) ainsi que la forme de leurs distributions, décrites par un histogramme et éventuellement une pdf empirique estimée par noyaux. Ces résultats sont décrits dans la Section 4.1 pour les variables intégrées et dans la Section 5.1 pour les séries temporelles.

3.4.3 Analyse de sensibilité des variables intégrées

La méthodologie globale pour l'analyse de sensibilité des variables intégrées est résumée sur la Figure 3.8. A partir du LHS de 4000 points utilisé pour l'analyse d'incertitude, une première étape de criblage est réalisée avec un test statistique d'indépendance basé sur la mesure HSIC (étape 1). Le criblage est réalisé indépendamment sur chaque UH et on retient l'union des différents ensembles de facteurs d'entrée sélectionnés afin d'être le plus conservatif possible. Cette étape permet d'identifier pour chaque variable cible les facteurs influents et de fixer les autres à une valeur nominale.
Ensuite, pour chaque variable, un nouveau LHS de 1000 points est généré en n'échantillonnant que les facteurs d'entrée retenus à l'étape de criblage. Pour les classer, on calcule les indices de Sobol à partir d'une décomposition en polynômes du chaos, les indices de sensibilité issus de la mesure de dépendance HSIC et ceux provenant d'un métamodèle par random forest. De tels indices sont d'abord calculés de manière locale sur chaque UH du bassin versant (étape 2.a). Pour chaque méthode, les intervalles de confiance à 95% sur les indices sont également calculés. Ensuite, des indices de Sobol aggrégés sont calculés (étape 2.b) à différentes échelles (le versant puis le bassin versant) permettant d'avoir pour chaque variable, une information synthétisée de la sensibilité. Cette approche reprend notamment les travaux de Saint-Geours et al. (2014) et on utilise dans la suite la même terminologie : indices de sensibilité par *sites* si le calcul est fait à l'échelle d'une UH et indices de sensibilité par *blocs* si les indices sont aggrégés.

Les résultats de l'application de cette méthodologie aux variables intégrées seront décrits dans la Section 4.2 et ont fait l'objet de la publication suivante soumise dans la revue *Geos-*

cientific Model Development et actuellement en cours de révision :

Rouzies, E.; Lauvernet, C.; Sudret, B. et Vidard, A. How to perform global sensitivity analysis of a catchment-scale, distributed pesticide transfer model? Application to the PESH-MELBA model *Geoscientific Model Development Discussions*, 2021, 1-44 (soumis)

3.4.4 Analyse de sensibilité des séries temporelles

La méthodologie utilisée pour l'analyse de sensibilité des séries temporelles est résumée sur la Figure 3.9. L'aspect temporel et l'aspect spatial y sont traités de manière séquentielle : le criblage est précédé d'une ACPf (étape 1). Pour chaque série temporelle et pour chaque UH, le sous-espace de représentation est déterminé à partir d'un échantillon de Morris. On ne retient que les J premières composantes principales qui permettent d'expliquer plus de 90% l'inertie totale. Les scores selon ces J composantes principales deviennent les nouvelles variables cibles et on réalise un criblage de Morris pour chacun d'entre eux, sur chaque UH (étape 2). À nouveau, on conserve l'union des facteurs d'entrée identifiés par chaque criblage. Le classement se fait ensuite localement sur chaque UH en calculant d'abord les indices de Sobol pour chaque composante principale à partir d'une décomposition en polynômes du chaos sur un LHS de 3000 points. Ces derniers sont ensuite généralisés selon l'Eq. (3.31) pour obtenir les indices par site synthétisant ainsi la sensibilité des séries temporelles sur chaque UH (étape 3.a). Finalement, les indices par bloc sont aggrégés à l'échelle du bassin versant à partir de l'Eq. (3.29) appliquée aux indices par site (étape 3.b).

Cette méthodologie pour l'analyse de sensibilité globale spatiotemporelle dans PESH-MELBA a été développée et en partie appliquée par Katarina Radišić dans le cadre de son stage de fin d'études de 6 mois que j'ai encadré. Ces travaux seront décrits dans la Section 5.2 et ont été soumis dans une *special issue (sensitivity analysis of model output)* de la revue *Socio-Environmental Systems Modelling* :

Radišić, K.; Rouzies, E.; Lauvernet, C. et Vidard, A. Global sensitivity analysis of a distributed hydrological model at the catchment scale *Socio-Environmental Systems Modelling*, 2022 (soumis)

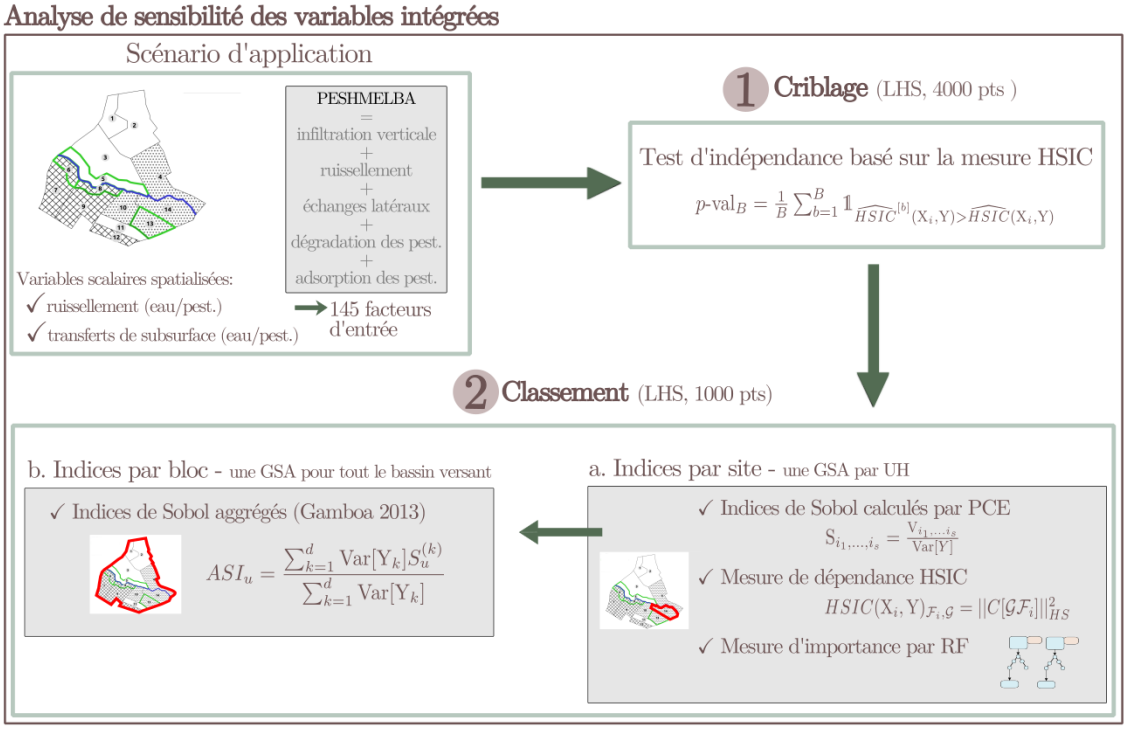


Figure 3.8 – Methodologie utilisée dans la thèse pour l'analyse de sensibilité des variables intégrées.

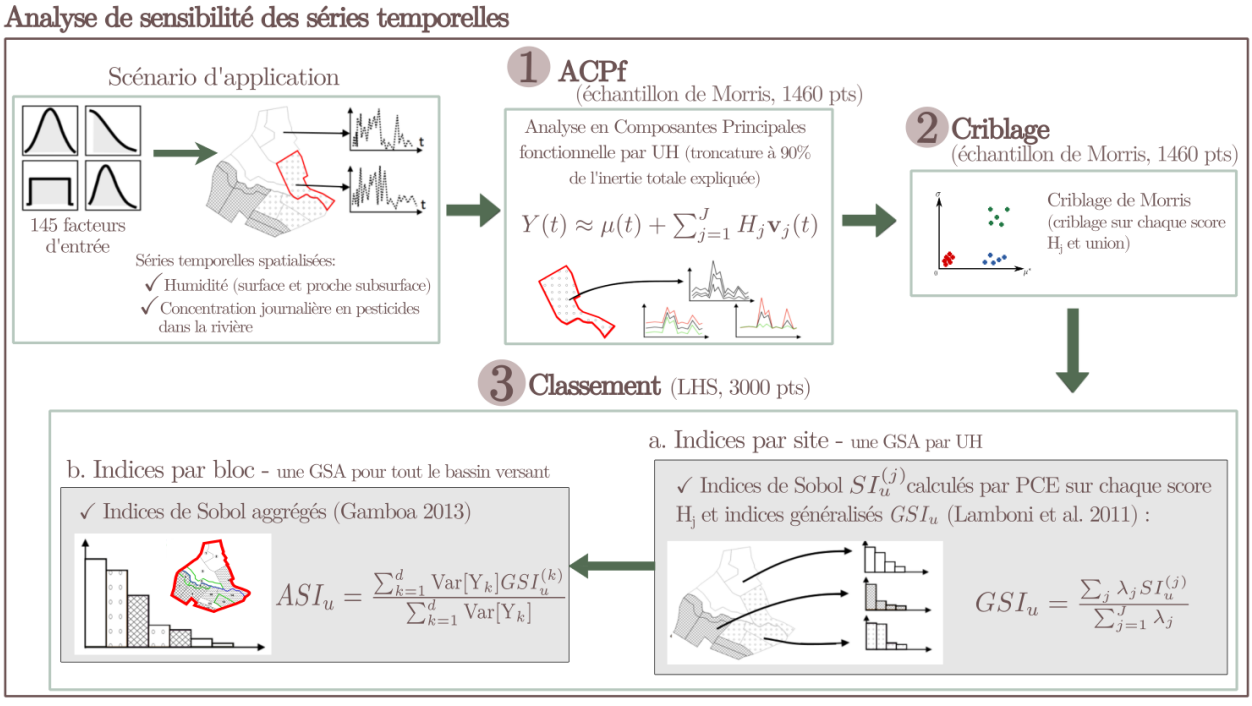


Figure 3.9 – Methodologie utilisée dans la thèse pour l'analyse de sensibilité des variables temporelles.

Chapitre 4

Quantification d'incertitude pour les variables intégrées

Sommaire

Ce chapitre regroupe les résultats relatifs à la quantification d'incertitude pour 4 variables intégrées de PESHMELBA. Ces variables (les volumes d'eau et les masses de pesticides cumulés transférés par échanges latéraux de surface et de subsurface sur chaque UH) ont été choisies car elles sont particulièrement représentatives du fonctionnement de PESHMELBA. Dans ce chapitre, l'objectif de la quantification d'incertitude est ainsi de gagner en connaissances sur le fonctionnement du modèle. Pour cela, on réalise d'abord une analyse d'incertitude dont les résultats sont présentés dans la Section 4.1. Dans un second temps, une analyse de sensibilité est menée dans la Section 4.2. Outre une interprétation des résultats en lien avec les processus physiques simulés dans PESHMELBA, cette section fournit également des éléments de réflexion méthodologiques concernant l'analyse de sensibilité de modèles spatialisés caractérisés par de nombreux couplages et un nombre important de paramètres.

L'intégralité des résultats de ce chapitre est présentée pour le scénario **hivernal** (voir Figure 2.10, gauche). En effet, seul ce scénario aboutit à une quantité de ruissellement significative permettant d'analyser les cumuls d'eau et de pesticides transférés en surface.

4.1 Analyse d'incertitude

L'analyse d'incertitude est menée à partir d'un LHS de 4000 points suivant les distributions des paramètres d'entrée décrites dans la Section 2.2.3 et rappelées en Annexe A.

Volumes d'eau transférés

La Figure 4.1 présente les histogrammes pour les volumes d'eau cumulés transférés en subsurface (variable WaterLateralFlow) et en surface (variable WaterSurfaceRunoff) pour différentes UH. Pour ces deux variables, un cumul positif indique que l'UH est globalement réceptrice (volume entrant cumulé supérieur au volume sortant) alors qu'un cumul négatif indique que l'UH est globalement émettrice. Pour aider à l'interprétation des résultats, on rappelle que la Figure 2.8 présente la configuration du bassin versant d'étude ainsi que la position des différents types de sol et éléments constitutifs.

La variable WaterLateralFlow suit une distribution approchant une gaussienne mais ses caractéristiques (moyenne et écart-type) varient significativement d'une UH à l'autre. Les volumes cumulés transférés en subsurface sont en moyenne plus importants sur le versant gauche que sur le versant droit (moyenne absolue entre 23 m^3 et 3665 m^3 sur le versant droit et entre 1574 m^3 et 8604 m^3 sur le versant gauche). Le versant gauche est caractérisé par des pentes plus importantes que le versant droit induisant des gradients hydrauliques plus forts pouvant expliquer une telle différence. Les écart-types sont aussi en moyenne plus élevés sur ce versant (max 500 m^3 sur le versant droit contre 681 m^3 sur le versant gauche). On ne note pas de tendance dans l'évolution de la dispersion entre l'amont et l'aval et la position par rapport à l'exutoire n'influe pas sur le niveau d'incertitude. Il est toutefois possible que cette conclusion ne soit pas généralisable à des cas d'étude réels impliquants des bassins

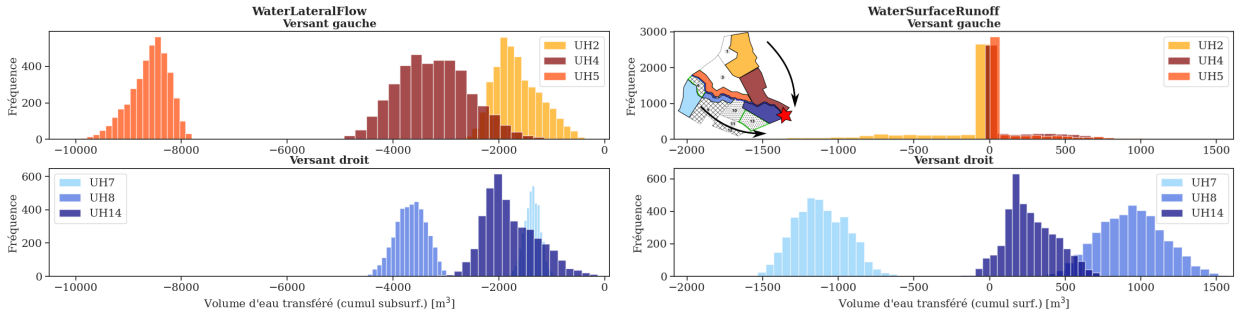


Figure 4.1 – Histogrammes du volume d'eau cumulé transféré par la subsurface (gauche) et la surface (droite) pour 6 UH sur toute la durée de la simulation (78 jours). Le pictogramme dans la figure en haut à droite indique la position de chaque UH dans le bassin versant. Sur ce pictogramme, l'étoile rouge indique la position de l'exutoire et les flèches rappellent le sens principal des écoulements.

versants plus grands. Dans ces cas, le découpage pourrait avoir un impact significatif sur la propagation de l'incertitude.
Pour la variable WaterSurfaceRunoff, les volumes cumulés transférés sont en moyenne inférieurs à la subsurface. La forte perméabilité des horizons de sol, majoritairement sableux, explique que peu de ruissellement soit émis sur le bassin. Là aussi, les distributions approchent des gaussiennes mais on note la présence d'un mode prépondérant autour de zéro pour les UH du versant gauche. Celui-ci est probablement dû à la présence d'un fort effet de seuil dans le déclenchement du ruissellement. En effet, dans PESHMELBA le ruissellement n'est activé qu'en présence d'une hauteur d'eau supérieure à la hauteur de ponding limite *hpond*. On rappelle que les valeurs nominales des hauteurs de ponding (*hpond*=1 cm pour les parcelles de vigne et *hpond*=5 cm pour les bandes enherbées) sont ensuite échantillonnées contribuant ainsi en partie à la dispersion et aux effets de seuil observés dans les distributions. On note également que, contrairement à la subsurface, certaines UH sont globalement réceptrices (support de distribution positif). C'est notamment le cas des UH 5 et 8 qui sont des bandes enherbées et qui ont donc un rôle d'interception du ruissellement.

Masses de pesticide transférées

De la même manière, la Figure 4.2 présente les histogrammes pour les masses de pesticides cumulées transférées en subsurface (variable SoluteLateralFlow) et en surface (variable SoluteSurfaceRunoff) pour différentes UH (sur des axes différents compte tenu des différences d'ordre de grandeur). Comme pour les volumes d'eau, on distingue les UH globalement réceptrices et émettrices selon le signe de la variable. On rappelle que pour les deux variables, seules les UH 2 et 7 peuvent être globalement émettrices puisque les pesticides ne sont appliqués que sur ces UH.

La variable SoluteLateralFlow suit une distribution globalement lognormale et un second mode en zéro se distingue pour les UH du versant gauche démontrant là aussi l'existence d'un effet de seuil. Sur ce versant, les pesticides ne sont appliqués que sur l'UH2 et une

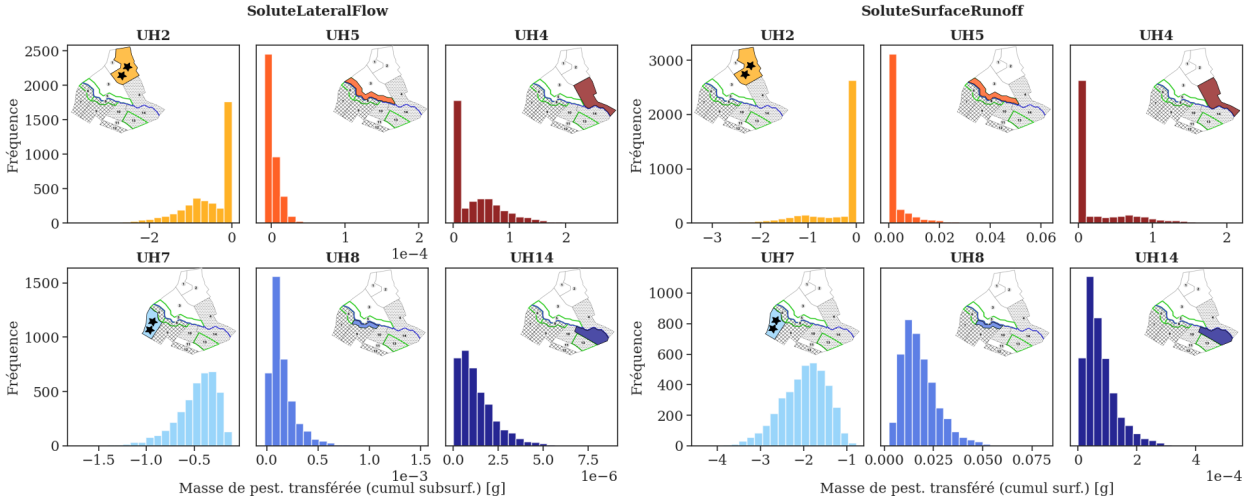


Figure 4.2 – Histogrammes des masses de pesticides cumulées transférées par la subsurface (gauche) et la surface (droite) pour 6 UH. Pour chaque figure, le pictogramme indique la position de chaque UH dans le bassin versant avec un remplissage coloré. Les étoiles noires rappellent les UH sur lesquelles sont appliqués les pesticides.

faible portion est donc propagée depuis cette UH vers les UH situées à l'aval. C'est ce qui explique la présence du mode en zéro pour toutes les UH de ce versant. D'autre part, la Figure 4.1 (gauche, ligne du haut) n'indique pas de tel mode en zéro pour les transferts d'eau en subsurface. C'est donc exclusivement les processus d'adsorption et de dégradation des molécules qui expliquent cette faible portion transférée. De plus, on note que même si la portion transférée est probablement faible sur ce versant, les masses cumulées transférées vers l'UH la plus à aval restent bien plus importantes que sur le versant droit (jusqu'à 2 g pour l'UH4 contre $5.1 \cdot 10^{-6}$ g sur l'UH14). Ceci s'explique par le découpage du versant droit, caractérisé par plus d'UH qui jouent un rôle tampon sur les transferts, et ce qu'il s'agisse de parcelles de vigne ou de bandes enherbées.
En surface, les distributions sont également lognormales sur le versant droit et largement marquées par un mode en zéro pour le versant gauche. Ce dernier est cohérent avec le mode en zéro identifié pour les transferts d'eau en surface. On note ici que, contrairement aux volumes d'eau, les ordres de grandeur des masses transférées sont plus grands en surface qu'en subsurface, confirmant ainsi l'existence d'écoulements plus concentrés dans ce compartiment.

Conclusion

L'analyse d'incertitude permet ainsi de tirer des premières conclusions quant au comportement du modèle. Elle permet notamment d'analyser les processus physiques qui interviennent dans les différentes parties du bassin ainsi que les interactions spatiales entre UH. Cependant, ces conclusions restent parcellaires et doivent être complétées par une analyse de sensibilité pour mieux comprendre la contribution des différents paramètres d'entrée aux variables de sortie cibles.

A retenir

- ✓ Les variables liées à l'eau sont globalement gaussiennes alors que celles liées aux pesticides sont globalement lognormales.
- ✓ Les niveaux d'incertitudes sont hétérogènes d'un versant à l'autre, en partie lié à des différences de pentes et de composition des sols.
- ✓ Les distributions sont souvent caractérisées par un mode autour de zéro lié à l'adsorption et à la dégradation ainsi qu'à la présence d'effets de seuil dans la représentation du ruissellement.

4.2 Analyse de sensibilité

La Figure 4.3 rappelle la méthodologie suivie pour l'analyse de sensibilité des variables intégrées. Dans la Section 4.2.1, une étape de criblage est d'abord appliquée pour diminuer la dimension du problème. Une étape de classement est ensuite réalisée et pour cela, différentes méthodes plus ou moins novatrices et caractérisées par des définitions de la sensibilité distinctes sont explorées. L'objectif est ici d'évaluer leur intérêt pour un modèle avec de nombreux couplages et un coût de calcul élevé comme PESHMELBA lorsque l'objectif de l'analyse de sensibilité est de mieux comprendre le fonctionnement du modèle. Les résultats présentés dans la Section 4.2.2 comparent donc les indices de sensibilité de Sobol, les indices de sensibilité obtenus avec la mesure de dépendance HSIC et les mesures d'importance calculées à partir de la construction d'un métamodèle par forêt aléatoire (RF, pour Random Forest) à l'échelle d'une UH. Dans la suite, et par souci de concision, on emploiera les dénominations *indices de Sobol*, *indices HSIC* et *indices RF* pour évoquer chacun de ces types d'indices de sensibilité. Après une analyse à l'échelle de chaque UH, les indices de sensibilité sont calculés à l'échelle du bassin versant dans la Section 4.2.3.

4.2.1 Criblage

Le criblage est réalisé à partir du LHS de 4000 points généré pour l'analyse d'incertitude avec un test d'indépendance basé sur la mesure HSIC. On effectue un criblage par UH puis les paramètres influents à l'échelle du bassin versant sont déduits à partir de l'union des paramètres influents sur chaque UH. Pour le test statistique d'indépendance, on utilise 100 bootstraps et le niveau de risque α est fixé à 1%. Il s'agit là d'un seuil arbitraire, fixé après plusieurs tests pour aboutir à des résultats raisonnables en termes de dimension d'espace résultant du criblage. On note qu'un tel arbitrage est souvent nécessaire dans les méthodes de criblage et dépend fortement de l'expérience de l'utilisateur sur le fonctionnement du modèle ainsi que des objectifs de l'analyse de sensibilité.
Après criblage et union, 42 facteurs d'entrée influents sont sélectionnés pour la variable WaterLateralFlow (subsurface) et 43 sont sélectionnés pour WaterSurfaceRunoff (surface). Pour

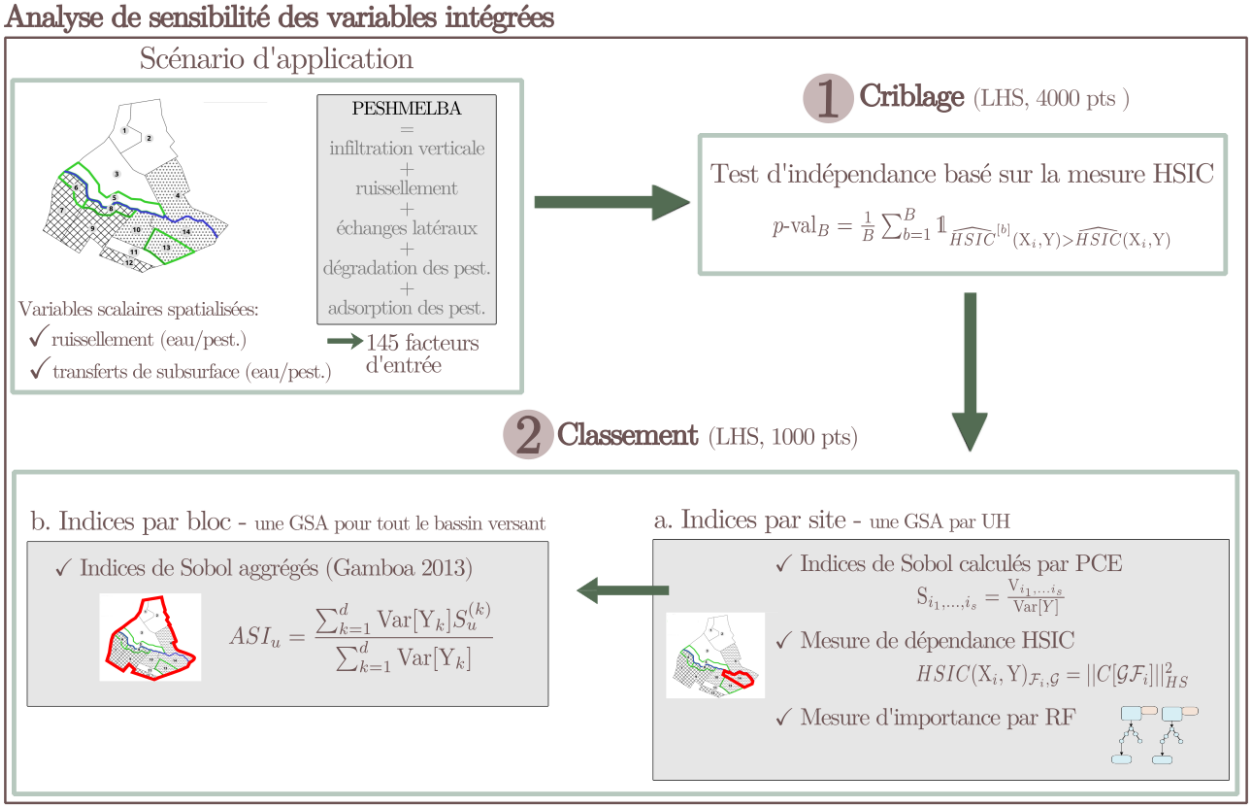


Figure 4.3 – Rappel de la méthodologie utilisée dans la thèse pour l'analyse de sensibilité des variables intégrées.

les pesticides, le criblage aboutit à 54 facteurs sélectionnés pour SoluteLateralFlow (subsurface) et 45 sont sélectionnés pour SoluteSurfaceRunoff (surface). Les facteurs d'entrée sélectionnés sont détaillés dans l'Annexe C. Parmi les paramètres retenus, figurent majoritairement les propriétés hydrodynamiques des différents horizons de surface et de subsurface (en particulier les teneurs en eau résiduelles et à saturation, les conductivités hydrauliques à saturation et les paramètres de Van Genuchten), quelques paramètres liés à la rivière, à la végétation ainsi que les hauteurs de ponding.
Le nombre de paramètres retenus après criblage demeure relativement élevé. Une première explication est que la méthodologie (ou ses paramètres, en l'occurence le niveau de risque α) n'est pas assez discriminante. Toutefois, cela peut aussi être la conséquence des nombreux processus physiques simulés dans PESHMELBA, chacun avec son propre jeu de paramètres d'entrée, qui contribuent à expliquer une variable de sortie.

D'autre part, les résultats montrent des hétérogénéités spatiales en termes de nombre de paramètres retenus, comme illustré sur la Figure 4.4. Contrairement à ce que l'on aurait pu imaginer, aucun gradient amont-aval ne se dessine clairement. Par contre, quelque soit la variable de sortie considérée, plus de paramètres influents sont retenus sur le versant droit. Ce contraste est particulièrement visible pour les variables relatives aux pesticides (colonnes 3 et 4) pour lesquelles un plus grand nombre de paramètres influents est notamment retenu

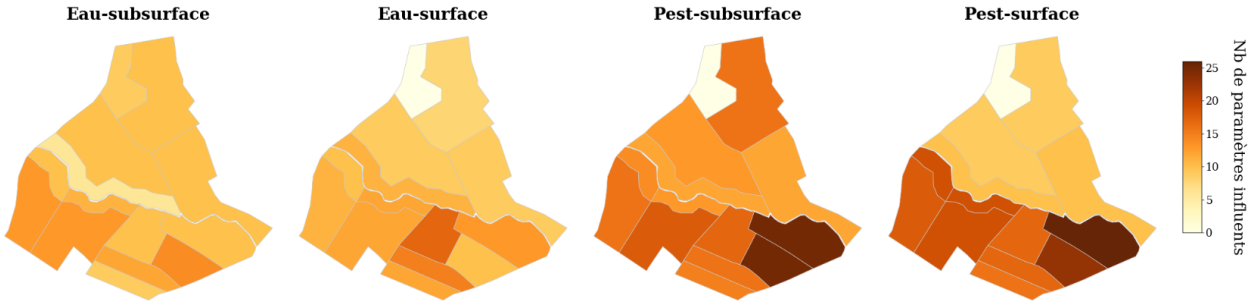


Figure 4.4 – Nombre de facteurs d'entrée retenus après criblage sur chaque UH pour les différentes variables cibles.

sur le versant droit, près de l'exutoire. Les hétérogénéités spatiales sont probablement liées à des contrastes dans les processus physiques activés dans les différentes parties du bassin en lien avec les types de sol, les pentes locales et les zones d'applications des pesticides (seulement sur 2 UH en amont). Cependant, cette hypothèse est à confirmer par un classement quantitatif afin d'identifier à quels processuss physiques sont reliés les principaux paramètres influents.

A retenir

- ✓ Le criblage permet d'éliminer près de deux tiers des paramètres pour chaque variable.
- ✓ Le nombre de paramètres retenus est fortement hétérogène d'une UH à l'autre.

4.2.2 Classement par UH

Les 3 méthodes de classement (indices de Sobol, indices HISC et indices RF) sont appliquées sur chaque UH à partir d'un LHS de 1000 points généré avec les paramètres retenus après criblage. Pour la décomposition en polynômes du chaos permettant le calcul des indices de Sobol, le degré maximal est fixé à p=3 et la norme q est choisie dans $[0.1, 0.2, ..., 1]$. Pour les forêts aléatoires, on compte entre 50 et 100 arbres par forêt selon la variable considéréee.

La Figure 4.6 regroupe les indices de sensibilité calculés pour toutes les variables avec chacune de ces méthodes pour l'UH14 située près de l'exutoire (voir légende correspondante sur la Figure 4.5). Sur cette figure, apparaissent seulement les paramètres classés comme étant parmi les plus influents d'après les indices de Sobol totaux. La comparaison des classements non tronqués (non montrés ici) permet de s'assurer qu'au sein des 10 paramètres aux indices de Sobol les plus élevés on retrouve toujours les 5 paramètres aux indices les plus élevés avec les autres méthodes. On garantit ainsi que la Figure 4.6 ne manque aucun paramètre dominant d'après les indices HSIC et RF.

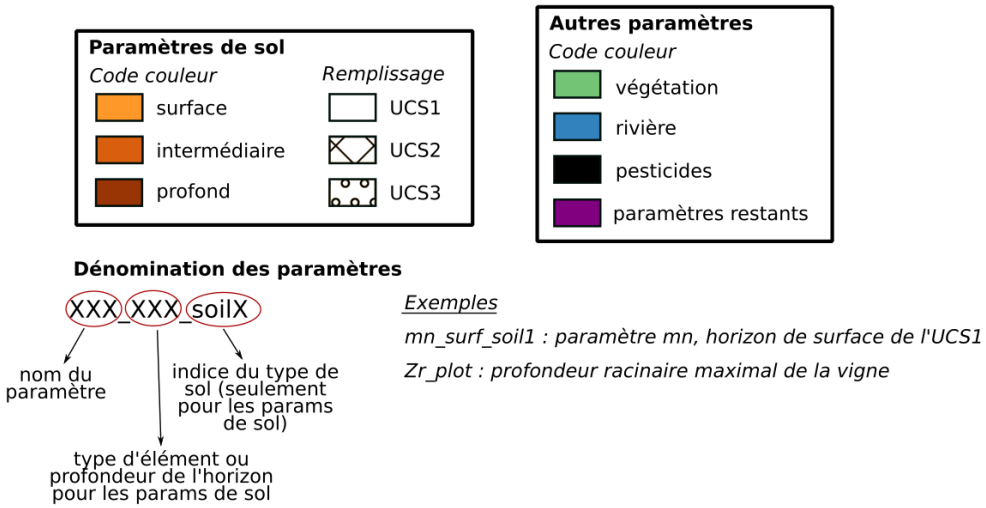


Figure 4.5 – Couleurs, motifs de remplissage et syntaxe utilisés pour présenter les résultats d'analyse de sensibilité dans les figures des Chapitres 4 et 5.

Comparaison des méthodes

Les classements obtenus d'après les 3 méthodes sont globalement cohérents même si l'interprétation des indices de Sobol doit se faire avec prudence puisque les résultats sont assortis de barres d'erreur supérieures à celles des autres méthodes. Si on compare les classements en détail, on ne relève que de légères différences pour WaterLateralFlow et WaterSurfaceRunoff. Par contre, les différences sont plus marquées pour les variables relatives aux pesticides. Dans les paragraphes suivants, on analyse et discute ces différences en comparant séparement les indices HSIC aux indices de Sobol puis les indices RF aux indices de Sobol.

Comparaison des indices de Sobol et des indices HSIC Avant de comparer les indices de Sobol et les indices HSIC, il faut se souvenir qu'ils reposent sur des définitions de la sensibilité intrinsèquement différentes. L'information fournie par les indices de Sobol ne concerne que des pourcentages de variance expliquée alors que les indices HSIC mesurent un niveau de dépendance globale et aggrègent théoriquement toutes les formes de relation non linéaire entre le paramètre d'entrée et la variable de sortie. D'autre part, contrairement aux indices de Sobol, les indices HSIC ne sont pas intrinsèquement construits pour capturer les interactions entre paramètres. Dans notre cas, les différences de classement entre les deux méthodes sont les plus marquées pour les variables relatives aux pesticides, caractérisées par plus de paramètres dont les effets directs sur la variance (indices de Sobol de premier ordre) sont nuls ou très faibles, démontrant ainsi l'existence d'effets quasi purement interactifs. Pour ces paramètres, les différences de classement sont donc attribuables à la fois à la non-prise en compte des effets interactifs des indices HSIC et aux différences de définition de la sensibilité des deux méthodes (le paramètre peut être influent sur une autre quantité que la variance). Toutefois, compte tenu du caractère global des indices HSIC, on ne peut pas avoir

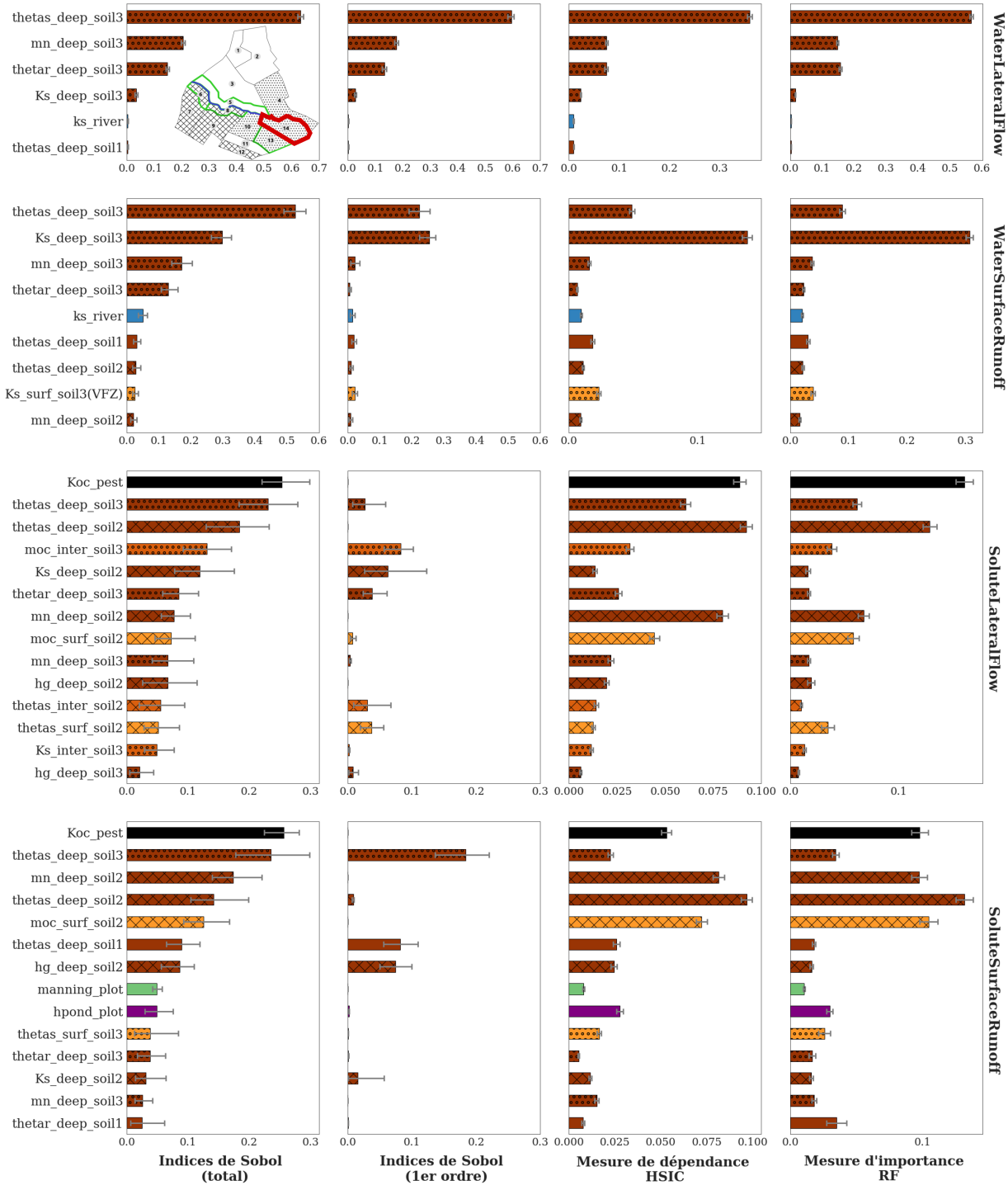


Figure 4.6 – Indices de Sobol totaux et de premier ordre, indices HISC et indices RF et intervalles de confiance à 95% associés calculés sur l'UH14. Les indices RF sont normalisés par 2Var[Y]. L'UH14 est entourée en rouge sur le pictogramme de la figure en haut à gauche. Pour chaque méthode, les paramètres affichés sont les plus influents d'après les indices de Sobol totaux. Les couleurs, le remplissage des barres et la définition des noms des paramètres sont détaillés dans la Figure 4.5.

une information plus détaillée sur la forme de la (ou des) dépendance(s) prédominante(s).

Comparaison des indices de Sobol et des indices RF Les écarts entre les indices de Sobol totaux et les indices RF pourraient également être interprétés en termes de différences dans les définitions de la sensibilité. Cependant, on rappelle que les travaux de Gregorutti et al. (2017) ont prouvé qu'indices RF et indices de Sobol sont liés. Cette relation (voir Eq. 3.28) n'est pourtant pas respectée dans notre cas. Pour mieux comprendre ces écarts, la qualité des métamodèles construits par RF et par PCE est évaluée en calculant les coefficients de prédictivité Q^2 sur un échantillon de test indépendant (200 points) :

$$Q^2 = 1 - \frac{\sum_{i=1}^{N}(M(\underline{\mathbf{X}}^i) - Y^i)^2}{\sum_{i=1}^{N}(Y^i - \overline{Y})^2} \tag{4.1}$$

où $\overline{\mathrm{Y}} = \frac{1}{N}\sum_{i=1}\mathrm{Y}^i$ est la moyenne empirique de l'échantillon. Or, les résultats montrent que l'utilisation d'une forêt aléatoire aboutit à un métamodèle de qualité parfois faible (voir Table 4.1). Ces performances limitées peuvent être à l'origine des écarts entre les indices RF et les indices de Sobol totaux. Pour expliquer pourquoi les forêts aléatoires aboutissent à un métamodèle de moins bonne qualité que les polynômes du chaos, on rappelle que contrairement à ces derniers, les forêts aléatoires ne formulent aucune hypothèse sur la forme de la relation entre X_i et Y. Cette relation est donc construite sans connaissance *a priori*, ce qui peut nécessiter plus de points.

	PCE	RF
WaterLateralFlow	0.98	0.85
WaterSurfaceRunoff	0.80	0.44
PesticideLateralFlow	0.75	0.55
PesticideSurfaceRunoff	0.75	0.51

Tableau 4.1 – Q^2 calculés pour toutes les variables sur l'UH14 à partir d'un échantillon de test indépendant de 200 points.

A propos des erreurs sur les indices de sensibilité

Malgré les performances limitées des RF reportées dans le Tableau 4.1, leurs barres d'erreur sont assez faibles, contrairement à celles associées aux indices de Sobol. Or, les barres d'erreur calculées sur les indices de Sobol incluent une évaluation de la qualité de la décomposition en polynômes du chaos (Marelli et Sudret, 2018) alors que la technique de sous-échantillonnage utilisée pour calculer les barres d'erreur des RF ne cible que la précision des mesures d'importance. Là encore, il faut donc rester prudent sur la comparaison de ces erreurs puisqu'elles ne sont pas calculées selon la même procédure.

Processus physiques liés aux paramètres influents

Compte tenu de leur bonne interprétabilité, on se concentre ici sur les indices de Sobol totaux et de premier ordre pour analyser les résultats d'un point de vue physique.
Tout d'abord, les indices de Sobol de premier ordre (colonne 2) indiquent que les effets directs expliquent près de 95 % de la variance de WaterLateralFlow et que les interactions (définies comme $S_T - S_i$) n'interviennent que très peu. Les conclusions pour les autres variables sont plus contrastées puisque les interactions expliquent près de 40% de la variance de WaterSurfaceRunoff, plus de 70% de la variance de SoluteLateralFlow et atteignent 80% de variance expliquée pour SoluteSurfaceRunoff. L'importance des effets interactifs est liée au fait que de nombreux paramètres sont nécessaires pour simuler chacun des processus physiques intégrés dans PESHMELBA mais ils reflètent aussi probablement des interactions entre processus physiques dans le modèle.

Les paramètres les plus influents diffèrent largement selon la variable cible considérée. Ils sont liés à différents processus physiques, susceptibles d'interagir les uns avec les autres, et sont porteurs d'informations quant au comportement hydrodynamique du bassin versant :

- Le transfert d'eau par subsurface (WaterLateralFlow, ligne 1) est majoritairement régi par les paramètres hydrodynamiques de l'horizon profond, lesquels sont liés à l'infiltration verticale et aux échanges latéraux saturés.
- Le transfert d'eau par ruissellement (WaterSurfaceRunoff, ligne 2) est également sensible à des paramètres de sol profond. On identifie ainsi que le ruissellement est probablement généré par saturation de la colonne de sol plutôt que suite à un dépassement de la capacité d'infiltration du sol en surface. Les échanges de subsurface avec la rivière contribuent également à expliquer cette variable puisque le paramètre Ks_river est classé parmi les plus influents. Ceci est cohérent avec la position de l'UH14, connectée à de nombreuses autres UH amonts ainsi qu'à un tronçon de rivière.
- Les variables relatives aux pesticides (lignes 3 et 4) sont influencées par des paramètres plus diversifiés caractérisant des processus contrastés. Elles sont toutes deux majoritairement sensibles au coefficient d'adsorption Koc. De plus, contrairement aux variables relatives à l'eau, les paramètres hydrodynamiques des horizons superficiels et intermédiaires ressortent aussi dans les classements. Ces paramètres sont liés à l'infiltration mais aussi à l'adsorption des pesticides. C'est notamment le cas des teneurs en carbone organique *moc* et des teneurs en eau à saturation *thetas* qui interviennent dans le calcul de l'équilibre d'adsorption (voir Eq. 2.12). On retrouve ainsi que la simulation de l'adsorption dans PESHMELBA permet de faire ressortir l'influence conjointe des propriétés intrinsèques de la molécule et des caractéristiques du milieu d'application, tel que largement observé sur le terrain. De plus, le coefficient de rugosité de Manning (*manning_plot*) et la hauteur de ponding (*hpond_plot*) liés au calcul du flux ruisselé sont aussi largement influents sur la variable SoluteSurfaceRunoff.

Discussion sur l'approche de classement

Les indices de Sobol, HSIC et RF ont été calculés pour classer les paramètres par niveau d'influence sur les variables de sortie cibles. En explorant plusieurs méthodes, l'objectif était d'analyser leurs spécificités et l'intérêt de chacune pour un exercice d'analyse de sensibilité sur un modèle complexe, caractérisé par de nombreux paramètres et couplages comme PESHMELBA. Au delà de leur comparaison, on cherchait à évaluer s'il était possible et utile de combiner plusieurs de ces méthodes.

Concernant les indices de Sobol et HSIC, l'utilisation combinée de ces deux types d'indices reste peu informative car les différences de classement qu'elles produisent sont difficilement interprétables. S'il faut choisir entre l'une des deux approches, ce choix doit se faire en fonction du modèle, de l'objectif visé avec l'analyse de sensibilité et des moyens de calcul dont on dispose :

1. Si l'analyse de sensibilité est utilisée pour gagner en connaissances sur le modèle en analysant finement son comportement, la méthode de Sobol reste attractive car elle fournit une interprétation claire des indices calculés (des pourcentages de variance expliquée) et une information explicite sur les interactions entre paramètres. Ces éléments sont particulièrement précieux lorsqu'on souhaite utiliser l'analyse de sensibilité dans un objectif de compréhension du fonctionnement du modèle et c'est d'ailleurs pour ça que cette approche est encore largement utilisée dans la communauté hydrologique. Pour estimer ces indices, l'utilisation de polynômes du chaos est intéressante puisqu'elle permet d'utiliser un échantillon déjà existant, de taille raisonnable (cette notion serait toutefois à préciser en menant des tests de convergence des indices calculés en fonction de la taille de l'échantillon).
2. Si l'analyse de sensibilité vise à simplifier le modèle ou focaliser des efforts de calibration, que l'interprétation physique des résultats n'est pas prioritaire et que l'on dispose d'un échantillon de points de taille très limitée (là encore, des tests supplémentaires pour évaluer la convergence des indices HSIC devraient être menés dans ce cas d'étude), l'utilisation des indices HSIC est une bonne option. On note toutefois que le choix du noyau peut influer sur les résultats de classement car chaque noyau spécifique est susceptible de donner plus ou moins d'importance à l'infinité de formes de dépendance qui sont capturées par HSIC. La question du choix du noyau est délicate mais encore assez peu abordée dans la littérature. Si quelques papiers proposent de choisir le type et le paramétrage des noyaux de manière à maximiser la possible dépendance entre X_i et Y (Fukumizu et al., 2009 ; Balasubramanian et al., 2013), l'interprétation des résultats semble être moins claire. D'autre part, il y a encore relativement peu de travaux qui appliquent cette solution et les limitations ne sont pas encore forcément toutes identifiées. HSIC est une méthode encore récente et son utilisation en hydrologie pour un exercice de classement restera donc délicate tant qu'il n'y aura pas de consensus sur le choix du noyau ni l'interprétation des résultats. Toutefois, ces pro-

blèmes ne se posent pas dans le cadre de l'utilisation de HSIC pour du criblage (si $HSIC(X_i, Y) = 0$, toutes les covariances de l'Eq. (3.19) de définition de HSIC sont égales à 0 et on s'assure qu'il n'y a aucune dépendance entre X_i et Y) et HSIC est donc peut-être à privilégier pour ce type d'exercice.

Concernant les indices RF, il n'y a pas d'intérêt à les combiner aux indices de Sobol puisqu'ils sont supposés fournir la même information que les indices de Sobol totaux (à un facteur près). L'utilisation de forêts aléatoires peut plutôt servir à fournir un estimateur de indices de Sobol totaux si l'on ne dispose d'aucune autre méthode pour estimer ces derniers. Cependant, dans cette étude le métamodèle construit par RF est de moins bonne qualité que celui construit par PCE. D'autre part, les PCE restent à privilégier puisqu'ils fournissent une information plus complète incluant non seulement les indices de Sobol totaux mais aussi les indices à tous les ordres.

A retenir

- ✓ Sur l'UH14, les 3 méthodes produisent des classements très proches pour les variables liées à l'eau alors qu'il y a des différences significatives pour les variables liées aux pesticides.
- ✓ Les différences entre indices de Sobol et indices HSIC correspondent à des différences de définition de sensibilité alors que les différences entre indices de Sobol et indices RF sont probablement liées à la qualité limitée du métamodèle calculé par RF.
- ✓ Pour cette application, l'utilisation des indices de Sobol calculés à partir d'un métamodèle par polynômes du chaos est l'approche de classement la plus adaptée.
- ✓ Les effets directs sont majoritaires pour les variables liées à l'eau alors que les effets interactifs sont majoritaires pour les variables liées aux pesticides.
- ✓ Les variables liées à l'eau sont majoritairement influencées par des paramètres hydrodynamiques de l'horizon profond local.
- ✓ Les variables liées aux pesticides sont dominées par l'effet du coefficient d'adsorption puis influencées par des paramètres hydrodynamiques à différentes profondeurs.

4.2.3 Analyse spatialisée

Cartes de sensibilité

On se concentre sur les indices de Sobol totaux et de premier ordre pour étendre l'analyse de sensibilité à l'échelle du bassin versant. Les indices de sensibilité calculés pour toutes les UH comme présentés sur la Figure 4.6 pour l'UH14 sont regroupés sous forme de cartes de sensibilité pour les variables WaterSurfaceRunoff et SoluteSurfaceRunoff (voir Figure 4.7). Les résultats pour la subsurface ne sont pas présentés ici puisque les conclusions sont sensiblement les mêmes que pour les variables de surface.

De manière générale, les cartes montrent de fortes hétérogénéités spatiales dans les valeurs des indices de sensibilité et notamment un comportement contrasté entre les versants gauche et droite. Pour les deux variables, les paramètres hydrodynamiques (*thetas*, *thetar* et *mn*) de l'horizon profond du type de sol 1 (resp. 2 et 3) sont seulement influents sur les UH appartenant au type de sol 1 (resp. 2 et 3). Les paramètres hydrodynamiques locaux sont ainsi dominants pour expliquer la variance de la variable de sortie. Un cas particulier est l'UH4 (indiquée avec une flèche sur la Figure 4.7) qui appartient au type de sol 3 alors que des paramètres caractéristiques du sol 1 expliquent une grande partie de la variance des 2 variables de sortie considérées. La localisation de l'UH4, près de l'exutoire, à l'aval de plusieurs UH de sol 1 peut expliquer de telles interactions spatiales. Pour SoluteSurfaceRunoff, en plus de paramètres de sol spécifiques, d'autres paramètres tels que la rugosité de Manning sur les parcelles de vigne (*manning_plot*) ou le coefficient d'adsorption (Koc_pest) ont une plus grande influence sur les UH du versant droit.

Enfin, la comparaison des cartes d'indices de premier ordre et d'indices totaux montre des résultats relativement similaires pour WaterSurfaceRunoff. Cela montre que les effets directs sont prédominants pour la plupart des paramètres et sur la plupart des UH. La conclusion est bien plus nuancée pour SoluteSurfaceRunoff. Dans ce cas, la plupart des paramètres sont influents presque seulement au travers d'effets interactifs puisque les indices de premier ordre sont très faibles comparés aux indices totaux.

Indices de sensibilité par bloc

Les indices de Sobol sont ensuite aggrégés selon Gamboa et al. (2013) comme décrit en Section 3.3.3, pour les variables WaterSurfaceRunoff et SoluteSurfaceRunoff (voir Figure 4.8). Comme les cartes d'indices de sensibilité ont mis en lumière des comportements contrastés sur les deux versants, on calcule des indices par bloc non seulement à l'échelle du bassin versant mais aussi à l'échelle intermédiaire des versants gauche et droite. Ces derniers permettent de fournir une information aggrégée sans masquer les différences de sensibilité des deux versants. Comme à l'échelle locale, ces indices fournissent une information condensée des processus physiques impliqués pour expliquer la variable de sortie. Pour WaterSurfaceRunoff, les paramètres hydrodynamiques liés à l'infiltration verticale (*thetas*, *thetar* et *mn*) dominent. Pour SoluteSurfaceRunoff, les paramètres hydrodynamiques de l'horizon profond du sol 1 et le coefficient d'adorption Koc expliquent une majeure portion de la variance sur le versant gauche alors que le temps de demi-vie $DT50$ et le paramètre de surface *hpond_plot* n'y ont que peu ou pas d'influence. Au contraire, les paramètres de surface (*manning_plot* et *hpond_plot*) sont plus influents sur le versant droit. Sur ce versant, les horizons de sol sont sensiblement moins perméables et les gradients hydrauliques plus faibles. Ceci résulte en plus de ruissellement généré sur le versant droit comme déjà constaté à partir des résultats de l'analyse d'incertitude. De plus, les paramètres de pesticides ($DT50$ et Koc) sont aussi

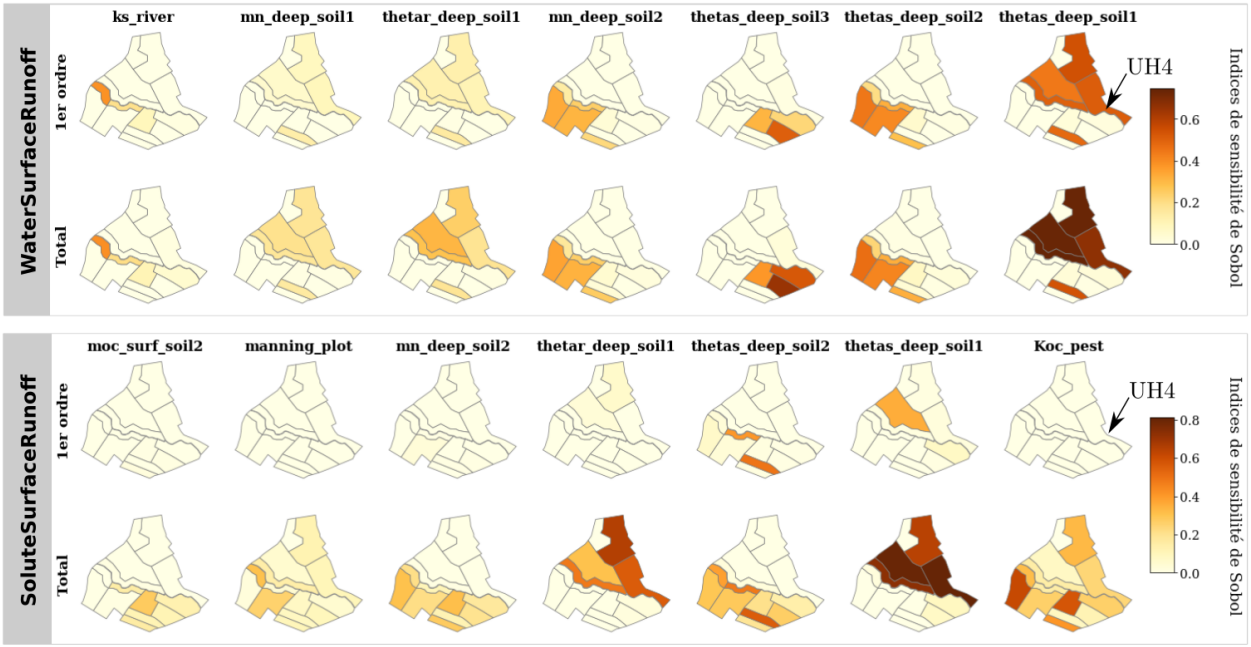


Figure 4.7 – Haut : carte des indices de Sobol par site pour les paramètres les plus influents pour le volume d'eau cumulé transféré par ruissellement de surface. Bas : carte des indices de Sobol par site pour les paramètres les plus influents pour la masse de pesticides cumulée transférée par ruissellement de surface.

plus influents sur ce versant. Plus généralement, les résultats montrent que les masses de pesticides transférées par ruissellement de surface résultent probablement de l'activation et de l'interaction de plus de processus physiques sur le versant droit que sur le versant gauche.

Conclusion

Les cartes d'indices de sensibilité fournissent une information détaillée et locale à propos des paramètres influents. Cependant, elles sont potentiellement coûteuses à réaliser puisqu'elles supposent de réaliser une GSA par UH. Cette approche peut être difficile voire impossible à transposer à l'échelle de vrais bassins versants composés de plusieurs centaines d'éléments. Les indices par bloc fournissent quant à eux une information aggrégée. Dans cette étude, de tels indices ont été calculés à partir des indices par site mais il est important de rappeler qu'un estimateur pick-freeze existe (Gamboa et al., 2013 ; De Lozzo et Marrel, 2016) pour calculer les indices de Sobol ASI servant à l'aggrégation sans passer par le calcul de tous les indices locaux. Ces estimateurs peuvent ainsi permettre de contourner la difficulté du coût de calcul si la taille du problème est limitante.
Ici, on peut déduire de ces indices par bloc un besoin de concentrer les efforts de calibration sur les paramètres hydrodynamiques des horizons les plus profonds et sur le coefficient d'adsorption du tebuconazole pour améliorer la qualité des simulations. Comme précisé dans Marrel et al. (2015), les deux approches (indices par site et par bloc) sont complémentaires et permettent de mieux comprendre le fonctionnement du modèle spatialisé. Si les indices

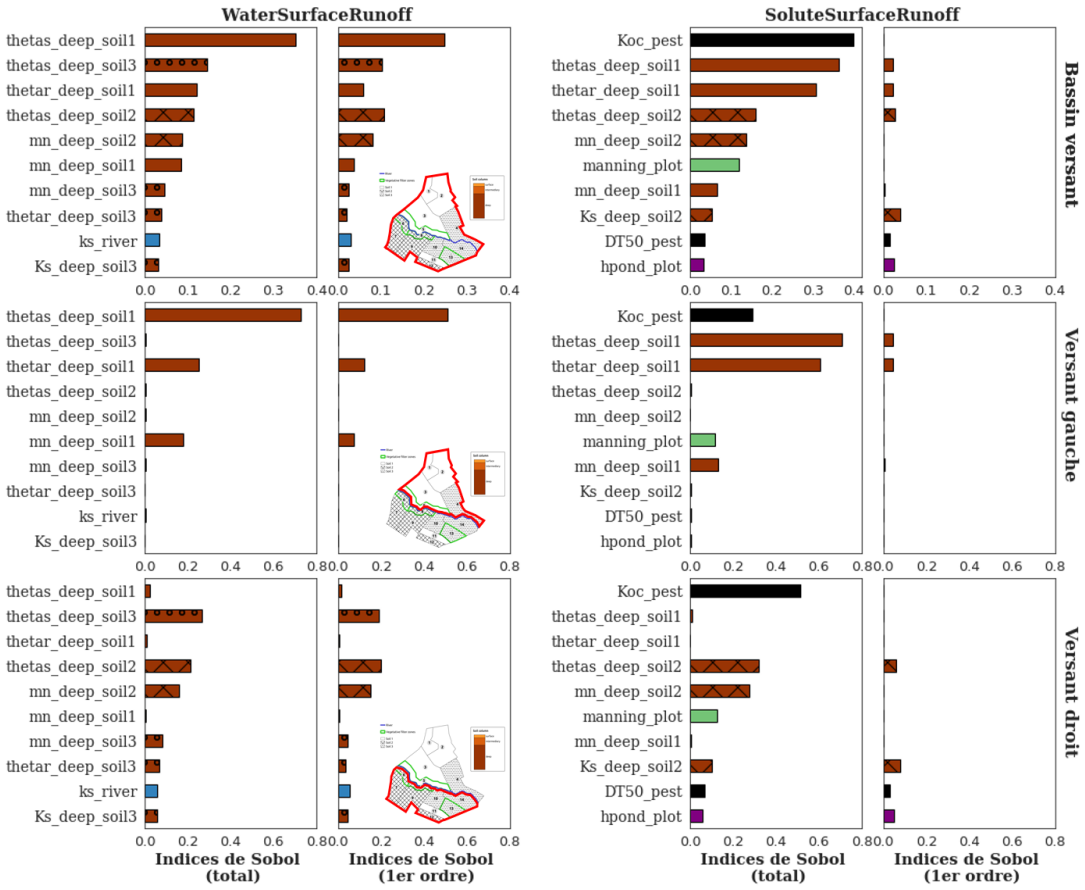


Figure 4.8 – Indices de Sobol totaux et de premier ordre pour les variables WaterSurfaceRunoff (gauche) et SoluteSurfaceRunoff (droite) calculés à l'échelle du bassin (haut), du versant gauche (milieu) et du versant droit (bas). Les paramètres affichés sont les 11 les plus influents d'après les indices de Sobol totaux. Les couleurs, le remplissage des barres et la définition des noms des paramètres sont détaillés dans la Figure 4.5.

par site ne peuvent pas être calculés, l'échelle du versant constitue une échelle intermédiaire d'intérêt, compromis entre indices locaux détaillés permettant une interprétation physique et coût de calcul limité.

A retenir

- ✓ Les valeurs des indices de sensibilité sont très hétérogènes d'une UH à l'autre et plus généralement d'un versant à l'autre.
- ✓ A l'échelle du bassin versant, les effets directs dominent pour les variables liées à l'eau alors que les effets interactifs dominent pour les variables liées aux pesticides.
- ✓ Pour avoir une vision globale de la sensibilité, le mieux est de produire des cartes d'indices de sensibilité et des indices aggrégés à l'échelle du bassin versant.
- ✓ Si on ne peut pas produire de cartes d'indices de sensibilité, on peut calculer des indices aggrégés à l'échelle du versant pour limiter le coût de calcul et conserver l'hétérogénéité spatiale.

4.3 Conclusion : apport de l'analyse d'incertitude et de l'analyse de sensibilité pour la compréhension du modèle

Dans ce chapitre, l'incertitude relative à des variables intégrées du modèle PESHMELBA a été quantifiée en réalisant successivement une analyse d'incertitude et une analyse de sensibilité globale. L'objectif de ces deux analyses était d'améliorer notre compréhension de PESHMELBA et de participer à sa validation. Pour cela, elles ont été appliqués à des variables cibles choisies de manière a être représentatives du fonctionnement du modèle : des volumes d'eau et des masses de pesticides cumulés, transférés entre chaque UH en surface et en subsurface. Pour l'analyse de sensibilité, compte tenu du nombre important de paramètres d'entrée incertains, une première étape de criblage a été réalisée à partir d'un test d'indépendance basé sur la mesure HSIC. Dans un second temps, les paramètres restants ont été classés par ordre d'influence en utilisant 3 méthodes : une décomposition de la variance, des mesures de dépendance HSIC et des mesures d'importance obtenues à partir d'un métamodèle par forêt aléatoire. La quantification de l'incertitude a été abordée de manière à prendre en compte l'aspect spatialisé de PESHMELBA. Pour cela, nous avons combiné plusieurs outils : histogrammes spatialisés, indice de sensibilité par site regroupés en carte et indices de sensibilité calculés à l'échelle du bassin versant. Les objectifs, la méthodologie et les principaux résultats de ce chapitre sont résumés sur la Figure 4.9.

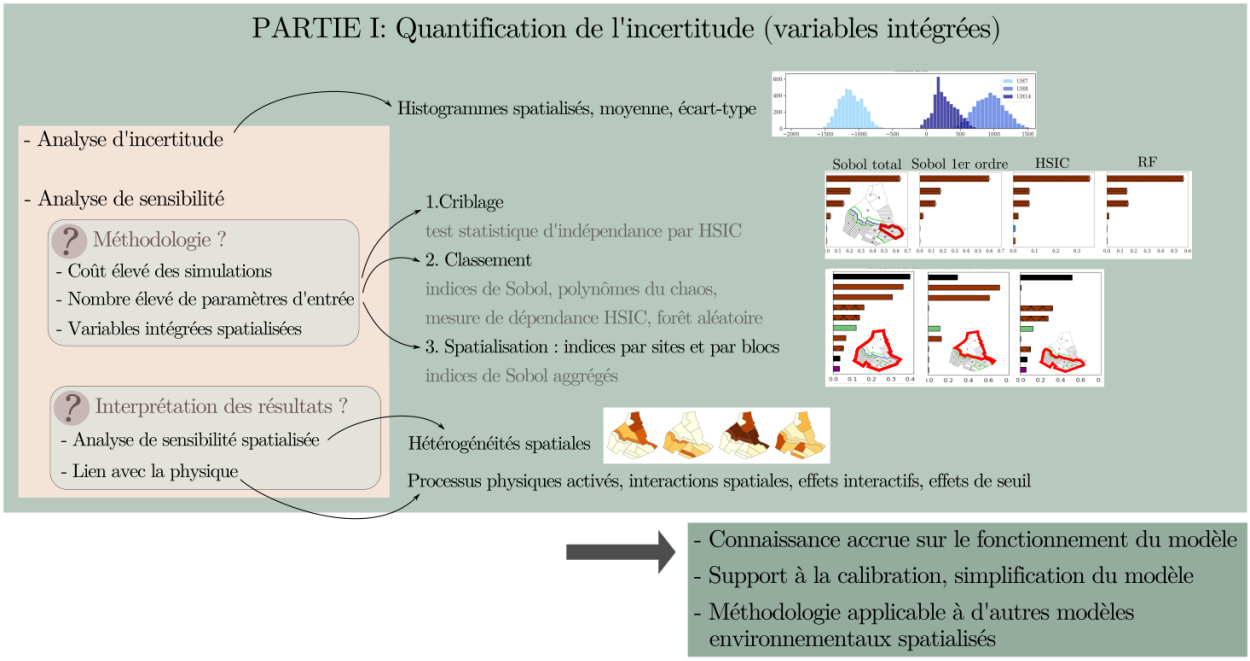


Figure 4.9 – Résumé visuel du Chapitre 4: problématiques de départ, outils utilisés et principaux résultats obtenus pour la quantification d'incertitude des variables scalaires.

L'interprétation physique des résultats fournis par ces différents outils a permis de ré-

pondre aux objectifs de meilleure compréhension du modèle, appliqué au scénario d'étude. L'analyse d'incertitude et de sensibilité ont notamment permis de caractériser certains processus physiques comme le ruissellement ou l'adsorption et d'en valider leur représentation dans le modèle. Ils ont également permis de mettre en lumière la complexité du modèle (forts effets de seuil) ainsi que les différences de comportements selon les zones du bassin, en fonction des pentes, des types de sol locaux et des interactions spatiales entre éléments du paysage.
D'un point de vue méthodologique, bien que la méthodologie globale soit classique (criblage puis classement) les méthodes utilisées pour chacune de ces étapes sont innovantes dans le domaine de la modélisation hydrologique et des transferts de contaminants. Parmi les méthodes de classement utilisées, les indices de Sobol calculés à partir d'une décomposition en polynômes du chaos restent les plus adaptés pour ce cas d'étude compte tenu de leur interprétabilité. L'utilisation des indices HSIC est prometteuse mais leur interprétation parfois floue et leur sensibilité au choix du noyau constituent des inconvénients non négligeables. L'utilisation des mesures d'importance obtenues à partir de la construction d'une forêt aléatoire est également une approche prometteuse mais qui reste à consolider, notamment en utilisant des échantillons plus grands.
Pour prendre en compte l'aspect spatialisé, l'utilisation combinée d'indices locaux et d'indices aggrégés est intéressante. Les résultats ont notamment permis de montrer l'intérêt d'une échelle d'aggrégation intermédiaire comme compromis entre coût de calcul et information détaillée. L'échelle du versant a été identifiée comme une telle échelle d'intérêt bien que ce choix dépende probablement du site d'étude.

Les résultats, spatialement hétérogènes, varient également de manière importante selon les sorties du modèle considérées et montrent que les conclusions relatives à un processus physique ou plus généralement à un compartiment du sol, sont difficilement transposables. Plus généralement, on rappelle que les résultats de l'analyse d'incertitude et de l'analyse de sensibilité sont valides pour une variable d'intérêt et dans un contexte d'application particulier (climat, sol, topographie,...) (Alves Ferreira et al., 1995 ; Lauvernet et Muñoz-Carpena, 2018 ; Saltelli et al., 2019). Les conclusions sont ainsi susceptibles de varier grandement sur d'autres bassins versants d'application s'ils sont caractérisés par un contexte agropédoclimatique et/ou une taille différente. Pour aller plus loin, il serait intéressant de réaliser la même étude sur le scénario estival puis sur d'autres bassins versants, en commençant par le bassin versant complet de la Morcille, pour mettre en lumière les différences de comportement de PESHMELBA.
D'autre part, l'hypothèse d'indépendance des paramètres formulée dans ces travaux est probablement contestable, notamment pour les paramètres hydrodynamiques de Van Genuchten. Ainsi, il serait intéressant d'aborder l'analyse de sensibilité de PESHMELBA en considérant des paramètres d'entrée dépendants. La formulation d'indices de Sobol dans ce cadre a déjà été explorée (Chastaing et al., 2015) mais la question reste ouverte dans le cas de l'utilisation des mesures de dépendance HSIC et des mesures d'importance par RF.

Enfin, les variables de sortie considérées dans ce chapitre sont spatialisées mais intégrées dans le temps afin de simplifier le problème. En s'intéressant à de variables spatialisées et dynamiques, l'analyse de sensibilité devient plus complexe mais cela peut aussi permettre d'aller plus loin dans la compréhension du fonctionnement de PESHMELBA. Une telle analyse fait donc l'objet du chapitre suivant en s'intéressant à des séries temporelles d'humidité et de concentration.

Chapitre 5

Quantification d'incertitude pour les séries temporelles

Sommaire

Ce chapitre présente les résultats de la quantification d'incertitude de séries temporelles issues de PESHMELBA. Les variables cibles sont l'humidité de surface (et plus largement l'humidité dans la colonne de sol) et la concentration moyenne journalière en pesticides à l'exutoire. Outre leur pouvoir informatif sur le fonctionnement du modèle, ces variables sont sélectionnées car elles font l'objet du processus d'assimilation de données présenté dans la deuxième partie de ce travail. L'objectif de ce chapitre est ainsi de les caractériser au mieux pour pouvoir mettre en place un cadre d'assimilation le plus adapté possible, en termes de choix de méthode et de définition du vecteur d'état. Pour cela, on souhaite répondre à deux questions spécifiques:

1. La distribution des variables cibles est-elle gaussienne ?
2. Quels sont les paramètres d'entrée les plus influents sur les variables cibles ?

Pour répondre à la première question, on réalise une analyse d'incertitude, présentée dans la Section 5.1. La deuxième question est abordée par une analyse de sensibilité globale dont les résultats sont présentés dans la Section 5.2.

Pour l'humidité, l'analyse est menée sur les deux scénarios climatiques (estival et hivernal) décrits dans la Section 2.2. Cependant, aucun de ces scénarios n'aboutit à un signal significatif en termes de concentration en pesticides à l'exutoire. Pour mener l'analyse sur cette variable, on utilise dans la suite des travaux un scénario hybride dont les caractéristiques sont les suivantes :

- mêmes paramètres d'entrée que pour les scénarios précédents ;
- application de tebuconazole sur **toutes** les parcelles de vigne du bassin à t=1 h et t=1104 h (=46 jours) (voir Figure 5.1) ;
- conditions initiales (niveaux de nappes) identiques au scénario hivernal ;
- chronique de pluie virtuelle, caractérisée par des intensités 10 fois supérieures au scénario hivernal.

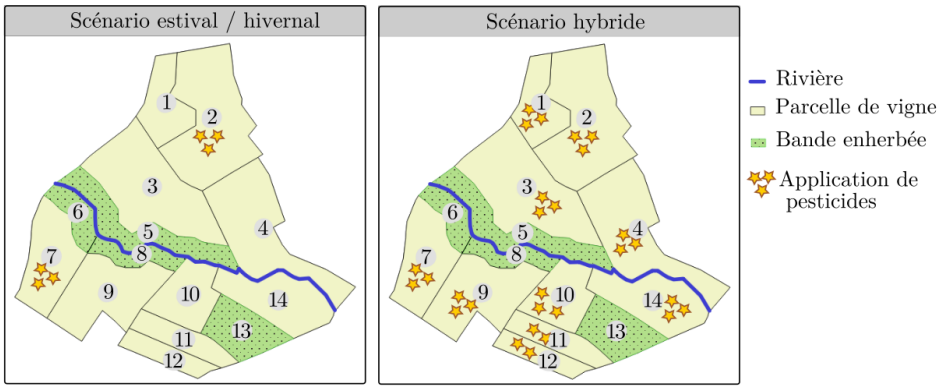


Figure 5.1 – Comparaison des différents scénarios en termes de zones d'application des pesticides.

Ainsi, ce scénario hybride s'éloigne d'une application réaliste car les précipitations sont largement surestimées. Les résultats devront donc être interprétés avec précaution d'un point de vue de la physique. Dans ce scénario, on se concentrera plutôt sur les conclusions méthodologiques afin de valider la faisabilité de l'analyse d'incertitude, de sensibilité puis de l'assimilation de données sur la concentration.

5.1 Analyse d'incertitude

5.1.1 Humidité de surface

Comme pour les variables scalaires, l'analyse d'incertitude est menée à partir d'un LHS de 4000 points généré en perturbant les 145 paramètres d'entrée de la simulation.

Scénario estival

La Figure 5.2 présente l'ensemble des séries temporelles d'humidité de surface sur l'UH2 pour le scénario estival ainsi que les histogrammes à plusieurs instants de la simulation.

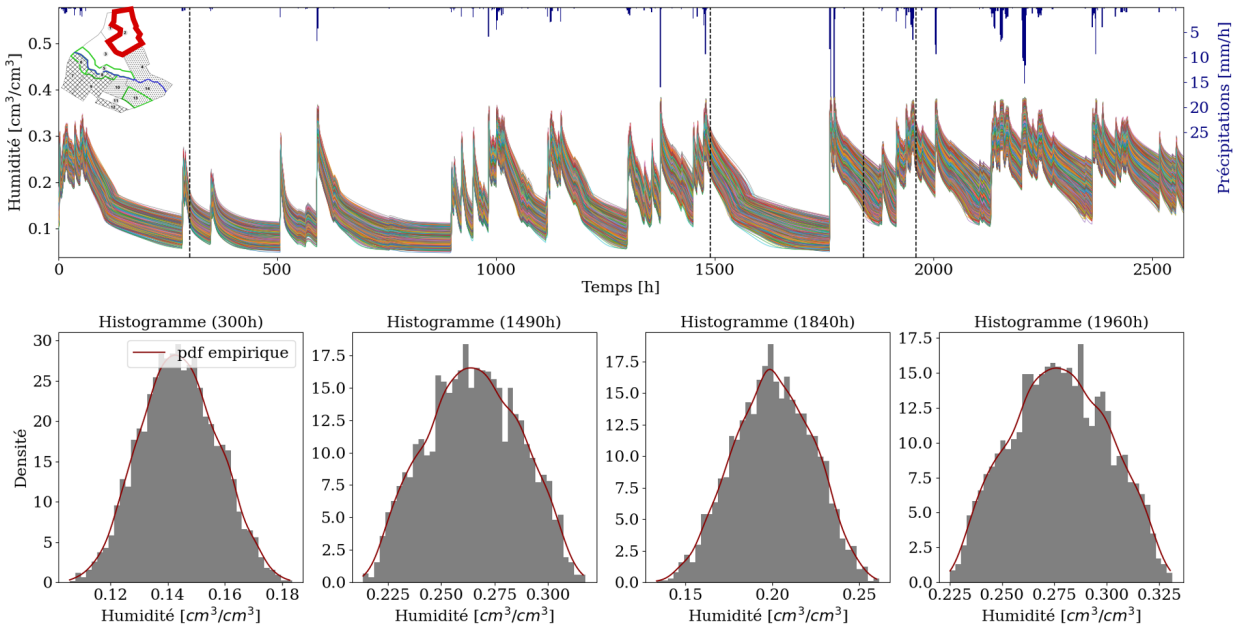


Figure 5.2 – Haut : séries temporelles d'humidité de surface sur l'UH2 (scénario estival). Les traits verticaux en pointillés indiquent les différents instants de la simulation pour lesquels sont tracés les histogrammes. Bas : histogrammes ponctuels et pdf empiriques estimées par noyau gaussien.

L'humidité de surface augmente instantanément en réaction aux précipitations alors que les périodes sans pluie sont caractérisées par un assèchement marqué de la surface dû à une forte évaporation. Les histogrammes ponctuels montrent que l'humidité de surface est unimodale, symétrique et d'allure globalement gaussienne tout au long de la simulation. Toutefois,

moyenne et écart-type varient selon les instants observés.

La Figure 5.3 montre l'évolution de l'écart-type de l'humidité de surface sur l'UH2 et généralise ce constat. Les événements pluvieux sont associés à une dispersion importante, liée à une réponse du modèle plus marquée et à l'activation de plus de processus physiques. La dispersion se maintient quelques heures après chaque événement pluvieux avant de diminuer rapidement pour atteindre les niveaux les plus faibles en période sèche. On note également que les écart-types sont globalement plus importants après 1760 h de simulation (≈ 73 jours), correspondant à une période caractérisée par des évènements pluvieux plus fréquents et plus intenses.

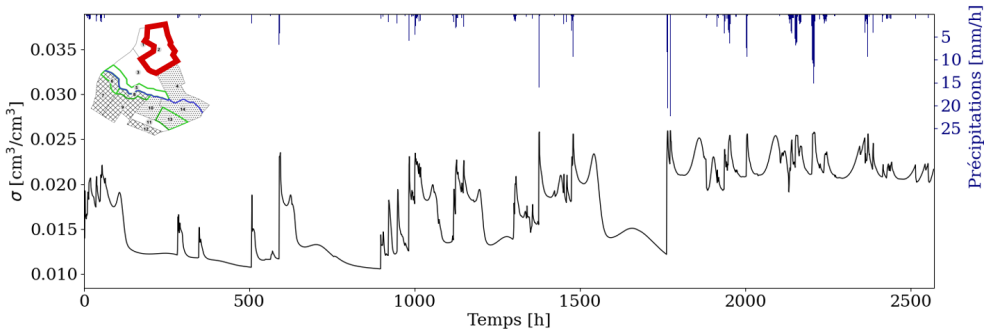


Figure 5.3 – Evolution de l'écart type de l'humidité de surface sur l'UH2 (scénario estival).

Les résultats présentés pour l'UH2 pour le scénario estival se généralisent ensuite aux autres UH du bassin versant. Les séries temporelles et distributions correspondantes ne sont donc pas montrées ici.

Scénario hivernal

La Figure 5.4 présente les séries temporelles et les histogrammes ponctuels de l'humidité de surface sur l'UH13 pour le scénario hivernal. La trajectoire d'humidité répond directement aux précipitations durant les 1840 premières heures de simulation (≈ 75 jours). Ensuite, une partie des simulations présente un plateau jusqu'à la fin de la simulation. Ce plateau correspond au seuil de saturation du sol : la teneur en eau maximale est atteinte et l'humidité ne peut plus augmenter. Plus globalement, les UH 2, 3, 7, 9 et 13 atteignent la saturation au moins à un instant pour le scénario hivernal et présentent donc le même type de dynamique (figures non montrées ici). Les plateaux qui en résultent sont en général transitoires, de durées variables. Les différences de type de sol, de hauteur de nappe initiale, de position dans le bassin ainsi que de nature des UH (parcelles de vigne ou bandes enherbées) peuvent expliquer de tels contrastes entre UH. La divergence de comportement entre simulations avec ou sans plateau de saturation est également visible sur les histogrammes ponctuels. Pour l'UH13, les histogrammes ponctuels à 1840 h et 1960 h montrent la présence de deux modes caractérisant l'état saturé ou non saturé. Le mode le plus humide (le plus à droite)

a le même support que la distribution du paramètre *thetas* (teneur en eau à saturation) de l'horizon de surface sur cette UH. L'humidité dépend donc probablement très fortement de ce paramètre lorsque la saturation est atteinte ce qui est cohérent avec sa définition. En dehors des périodes potentielles de saturation et comme pour le scénario estival, l'humidité suit une distribution globalement gaussienne dont l'écart-type reste fortement dépendant des forçages climatiques.

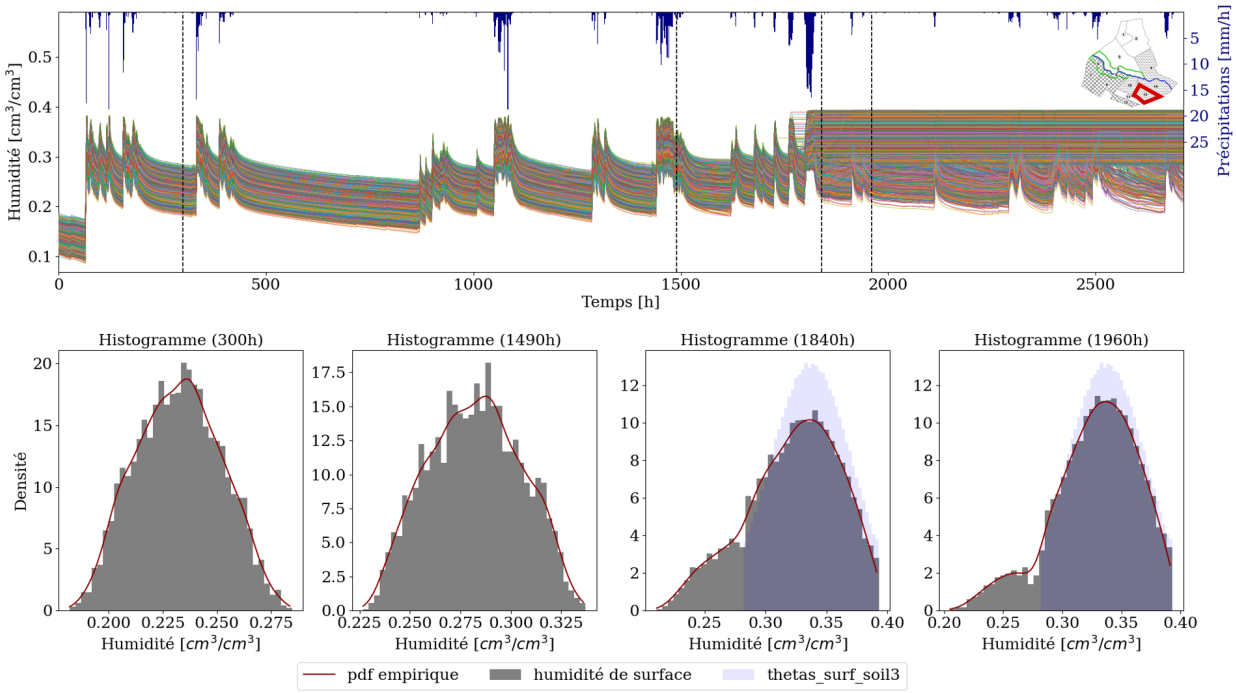


Figure 5.4 – Haut : séries temporelles d'humidité de surface sur l'UH13 (scénario hivernal). Les traits verticaux en pointillés indiquent les différents instants de la simulation pour lesquels sont tracés les histogrammes. Bas : histogrammes ponctuels et pdf empiriques estimées par noyau gaussien. L'histogramme du paramètre *thetas_surf_soil3(VFS)* est représenté en bleu.

D'autre part, les UH ne présentant pas de plateau de saturation sont caractérisées par le même type de dynamique que le début de simulation de l'UH13 (réaction rapide aux précipitations et distribution gaussienne). Les résultats ne sont donc pas montrés ici.

5.1.2 Humidité dans la colonne de sol

Ces résultats sont ensuite généralisés à l'humidité dans la colonne de sol en traçant à plusieurs profondeurs (surface, 10cm et 2m) les mêmes ensembles de séries temporelles et d'histogrammes. Les résultats sont présentés dans l'Annexe D pour le scénario hivernal sur l'UH13. On note que la réponse aux précipitations est légèrement atténuée en profondeur (pas de réaction aux événements pluvieux les plus faibles et amplitude des pics plus limitée). De plus, la teneur en eau est plus importante en profondeur qu'en surface, notamment à cause de la proximité de la nappe et d'une extraction racinaire plus faible. D'un point de

vue des distributions, on retient que les conclusions sont globalement les mêmes que pour la surface : en dehors des périodes de saturation visibles sur quelques UH sur le scénario hivernal, l'humidité suit une distribution s'approchant d'une gaussienne pour l'ensemble des profondeurs.

Pour le scénario estival, les conclusions concernant les formes de distributions sont très similaires (distributions gaussiennes à toutes les profondeurs, pour toutes les UH et pendant toute la durée de la simulation) et les résultats ne sont donc pas présentés ici.

5.1.3 Concentration à l'exutoire

La Figure 5.5 présente les résultats de l'analyse d'incertitude pour la concentration journalière en pesticides à l'exutoire sur le scénario hybride.

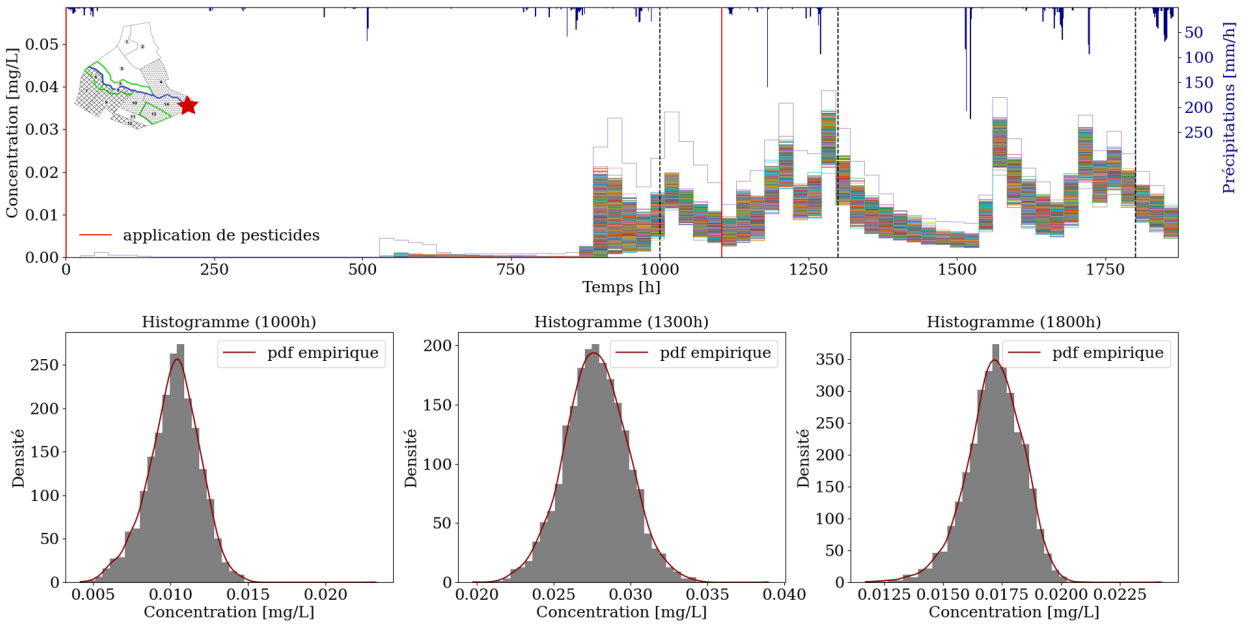


Figure 5.5 – Haut : séries temporelles de concentration journalière moyenne à l'exutoire. Les traits verticaux en pointillés indiquent les différents instants de la simulation pour lesquels sont tracés les histogrammes. Bas : histogrammes ponctuels et pdf empiriques estimées par noyau gaussien.

Les simulations montrent qu'un signal significatif n'est détecté dans la rivière qu'à partir de 500 h ($\approx$ 21 jours), soit après le premier évènement pluvieux intense et bien après la première application de pesticides. La corrélation avec la chronique de précipitations est moins évidente que pour l'humidité et marquée par un décalage temporel. L'impact de la seconde application n'est notamment pas immédiat dans la rivière. Ceci confirme que la concentration dépend de nombreux facteurs qui interagissent et dont l'impact est soumis à des décalages temporels et probablement spatiaux. Le rôle des bandes enherbées comme zones tampons au pouvoir de de rétention ainsi que des phénomènes d'adsorption et désorption

successives peuvent notamment être à la source des décalages constatés. La distribution de la concentration semble toutefois masquer une telle complexité puisque les histogrammes et les pdfs empiriques montrent qu'elle peut être assimilée à une gaussienne tout au long de la simulation.
D'autre part, l'ordre de grandeur des concentrations exportées à l'exutoire simulées, de l'ordre de la dizaine de μg.L^{-1}, semble surestimé par rapport aux mesures dont on dispose sur la Morcille (voir Figure 2.15). On rappelle toutefois que le scénario hybride est caractérisé par une chronique de précipitations et un calendrier d'applications loin d'être réalistes et il n'est donc pas pertinent d'analyser les résultats d'un point de vue quantitatif.

5.1.4 Conclusion

L'analyse d'incertitude a ainsi permis de répondre à la première question formulée en introduction en montrant que les variables considérées sont pour la plupart gaussiennes. La présence de plateaux de saturation pour l'humidité de proche surface nuance cependant cette conclusion avec l'apparition d'un second mode. Le lien avec la distribution des *thetas* donne une première information sur les paramètres influents mais celle-ci reste limitée à un contexte particulier (saturation). Pour la concentration en pesticides à l'exutoire, le lien avec de potentiels paramètres d'entrée influents est impossible à déduire de cette analyse. Pour les deux variables, il est donc indispensable de compléter ces résultats par une analyse de sensibilité globale.

A retenir

- ✓ L'humidité et la concentration suivent majoritairement des distributions gaussiennes.
- ✓ L'humidité de surface est parfois bimodale dans le scénario hivernal et le second mode correspond à l'apparition d'un plateau de saturation.
- ✓ Le signal d'humidité de surface est directement lié à la chronique de précipitations.
- ✓ Le signal de concentration à l'exutoire est désynchronisé par rapport aux forçages climatiques et aux applications de pesticides.
- ✓ Dans le cas de la concentration à l'exutoire, l'apport du scénario virtuel est limité pour analyser le réalisme des résultats.

5.2 Analyse de sensibilité

La Figure 5.6 rappelle la méthodologie suivie pour l'analyse de sensibilité des séries temporelles comme décrite en Section 3.4.4 : une première étape consiste à réaliser une analyse en composantes principales fonctionnelle pour réduire la dimension des variables cibles (Section 5.2.1). Un criblage est ensuite réalisé sur les scores selon chaque composante (Section 5.2.2) et permet de diminuer la taille de l'espace des paramètres d'entrée. Le classement des

paramètres retenus comme influents est finalement réalisé en calculant les indices de Sobol par décomposition en polynômes du chaos (Section 5.2.3). Dans cette dernière section, les indices de Sobol sont d'abord calculés pour chaque score, sur chaque UH puis ils sont aggrégés à l'échelle de l'UH (équivalent d'aggrégation temporelle), puis à l'échelle du bassin versant (aggrégation spatiale).
On rappelle que les sections relatives à l'humidité de surface pour le scénario hivernal reprennent les résultats du stage de Katarina Radišić.

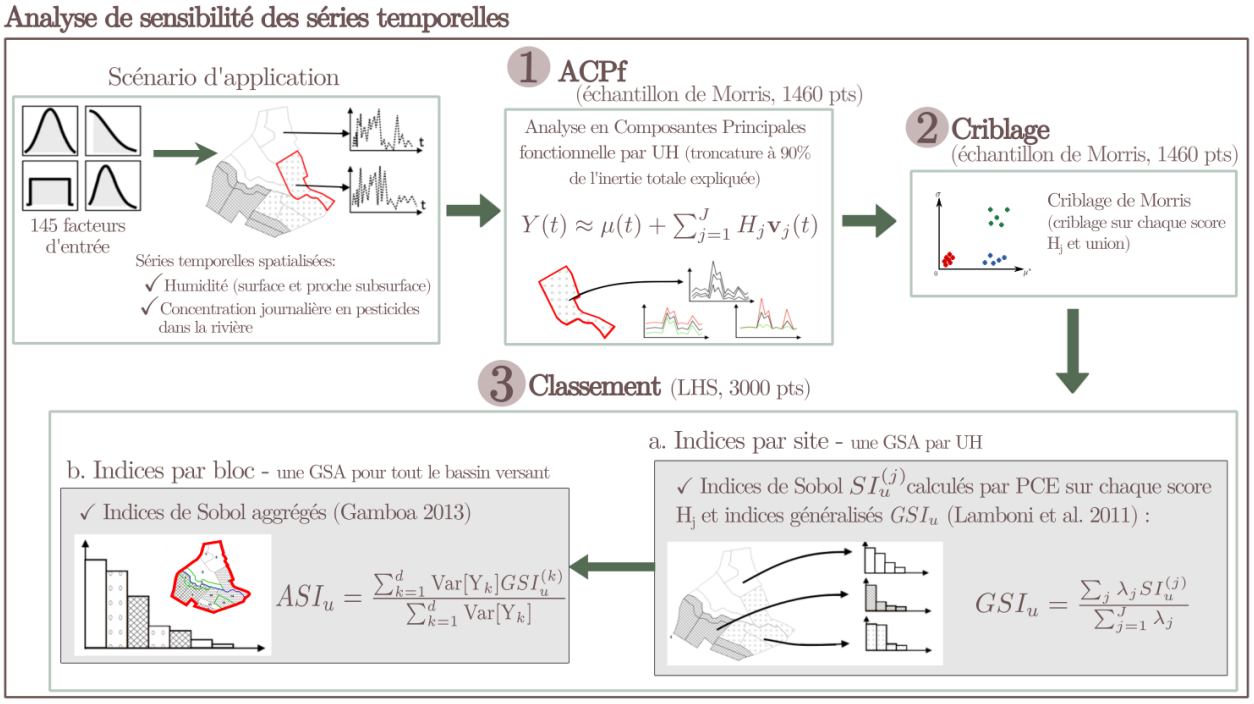


Figure 5.6 – Rappel de la méthodologie utilisée dans la thèse pour l'analyse de sensibilité des variables temporelles (adapté de Radišić et al. 2021).

5.2.1 Analyse en composantes principales fonctionnelle

On rappelle que l'ACPf est réalisée sur l'échantillon de Morris utilisé pour le criblage. Bien qu'un tel échantillon ne permette pas une couverture de l'espace aussi dense que le LHS, la variabilité des différentes séries temporelles est bien représentée. Des résultats préliminaires ont permis de vérifier que la base de décomposition déterminée est équivalente à celle déterminée sur un LHS.

Humidité de surface

Scénario estival Le pourcentage d'inertie expliqué en fonction du nombre de composantes retenues pour le scénario estival est présenté sur la Figure 5.8. On montre ainsi que 2 com-

posantes principales suffisent pour représenter plus de 90% de l'inertie sur toutes les UH.

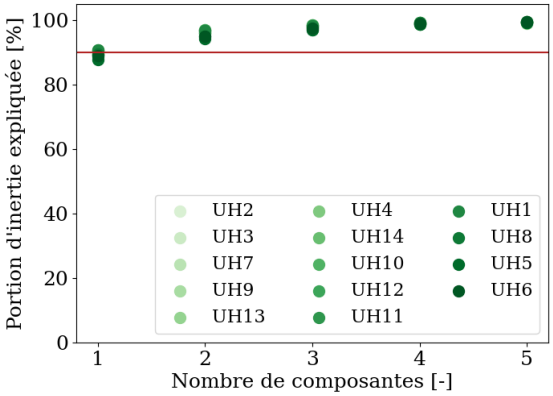


Figure 5.7 – Pourcentage de variabilité d'humidité expliqué en fonction du nombre de composantes retenues (scénario estival). Le trait rouge horizontal indique le seuil des 90 % de variabilité expliquée qui est visé.

La Figure 5.8 présente ces 2 premières composantes principales pour l'UH2 en représentant en ordonnée les perturbations par rapport à la moyenne. Sur cette figure, la première composante (PC1) caractérise le décalage vertical de la série d'humidité. Un score H_1 faible selon cette composante indique une humidité de surface en moyenne faible alors qu'un score élevé indique une humidité en moyenne élevée. La seconde composante (PC2) caractérise quant à elle l'amplitude des variations d'humidité (score H_2 d'autant plus faible que le contraste entre humidité maximale et minimale est important). On note que pour le scénario estival, l'interprétation des deux premières composantes est transposable à toutes les autres UH du bassin. En moyenne, PC1 explique 87% d'inertie et PC2 environ 7% de l'inertie.

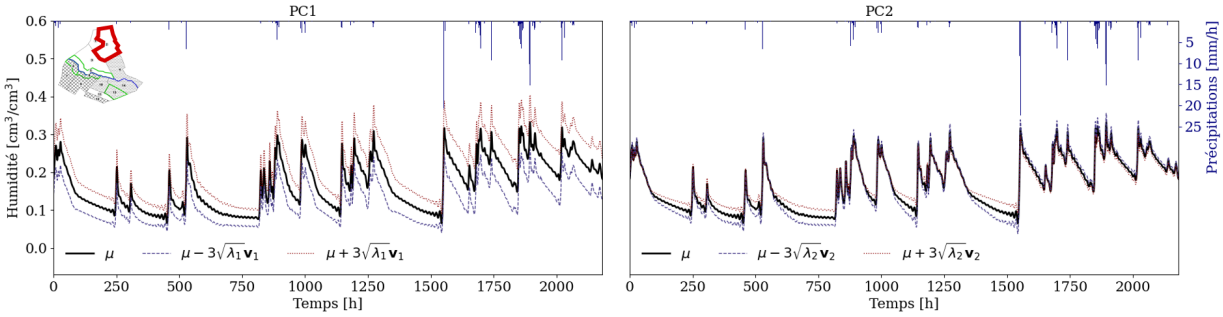


Figure 5.8 – Première (PC1, gauche) et deuxième (PC2, droite) composante principale représentées comme des perturbations positives et négatives par rapport à la moyenne μ pour l'humidité de surface sur l'UH2 (scénario estival).

Scénario hivernal Pour le scénario hivernal, une seule composante suffit pour représenter plus de 90% d'inertie sur les UH sans plateau de saturation. Pour les UH présentant un plateau, même temporaire, deux composantes principales sont nécessaires (voir Figure 5.9).

La Figure 5.10 présente ces deux premières composantes sur l'UH13 caractérisée par un plateau de saturation (voir Figure 5.4 pour un rappel des caractéristiques du plateau). L'interprétation de la première composante est identique au cas estival mais la seconde composante caractérise plutôt le contraste d'humidité pendant la période de saturation et hors de cette période. Là encore, on ne représente que les résultats sur l'UH13 mais ces interprétations restent valides sur toutes les UH avec plateau de saturation.

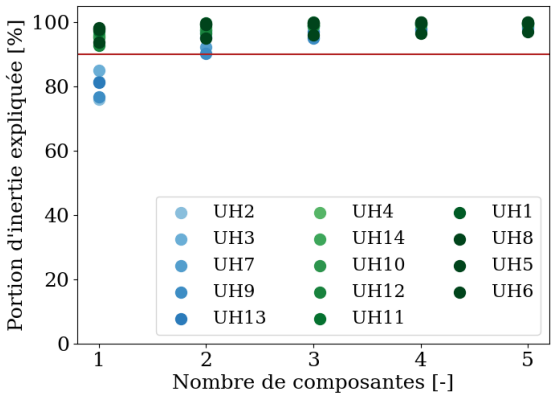


Figure 5.9 – Pourcentage de variabilité d'humidité expliqué en fonction du nombre de composantes retenues (scénario hivernal). Le trait rouge horizontal indique le seuil des 90 % de variabilité expliquée qui est visé. Les UH représentées en bleu correspondent aux UH avec plateau alors que les vertes sont les UH sans plateau.

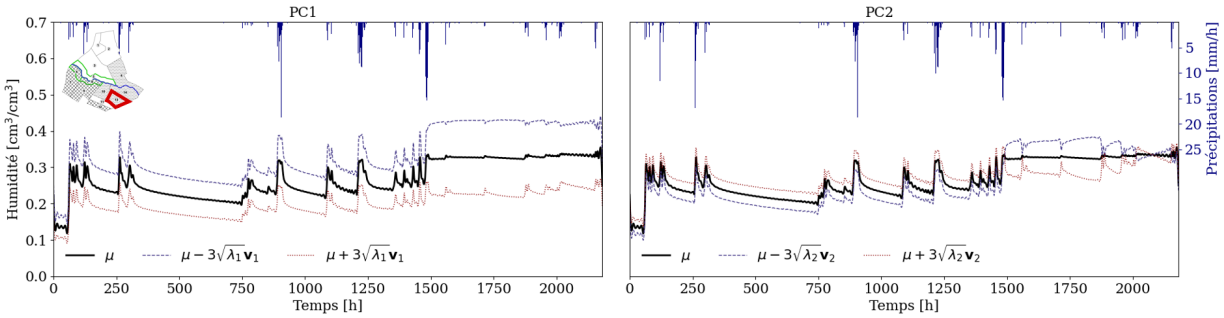


Figure 5.10 – Première (PC1, gauche) et deuxième (PC2, droite) composante principale représentées comme des perturbations positives et négatives par rapport à la moyenne μ pour l'humidité de surface sur l'UH13 (scénario hivernal).

Dans le cas des UH sans plateau, l'interprétation de la première composante reste la même que dans le cas estival (décalage vertical de la série temporelle).

Humidité dans la colonne de sol

Comme pour l'analyse d'incertitude, les résultats relatifs à l'humidité ne sont présentés qu'en surface mais les conclusions pour les profondeurs intermédiaires sont fortement similaires (nombre et interprétation des modes). Plus généralement, même si dans certains cas

un seul mode suffit pour atteindre les 90% de variabilité de l'humidité expliquée, pour plus de cohérence on en retient systématiquement deux pour réaliser l'analyse de sensibilité.

Concentration à l'exutoire

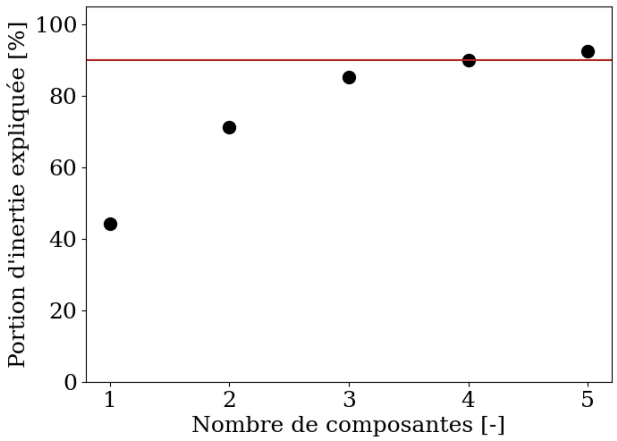


Figure 5.11 – Pourcentage de variabilité de concentration à l'exutoire expliqué en fonction du nombre de composantes retenues (scénario hybride). Le trait rouge horizontal indique le seuil des 90 % de variabilité expliquée qui est visé.

Pour la concentration à l'exutoire, 4 composantes principales sont nécessaires pour représenter plus de 90% de la variabilité (voir Figure 5.11) attestant de la plus grande complexité de la série temporelle. Contrairement à l'humidité pour laquelle la première composante est largement plus explicative, les deux premières composantes expliquent ici des pourcentages d'inertie assez proches (44.2% pour PC1 et 27.1% pour PC2). Les 4 premières composantes sont représentées sur la Figure 5.12 et leur interprétation est moins évidente que pour l'humidité reflétant la plus grande complexité de cette variable de sortie. PC2 caractérise de manière assez univoque le décalage vertical de la chronique de concentration alors que PC1, PC3 et PC4 semblent chacune caractériser les contrastes d'amplitude pour certains pics de la chronique.

Conclusion

L'ACPf permet de réduire significativement la dimension des variables d'intérêt. Pour l'humidité, 2 composantes suffisent à résumer en grande partie la dynamique de la sortie. L'analyse de sensibilité s'applique donc à une variable spatiotemporelle de taille maximale 14UH $\times$ 2PC = 28 éléments. Pour la concentration moyenne journalière à l'exutoire, bien que la dynamique soit plus complexe, sa variabilité peut aussi être synthétisée en un nombre raisonnable de composantes principales garantissant l'applicabilité des méthodes d'analyse de sensibilité.

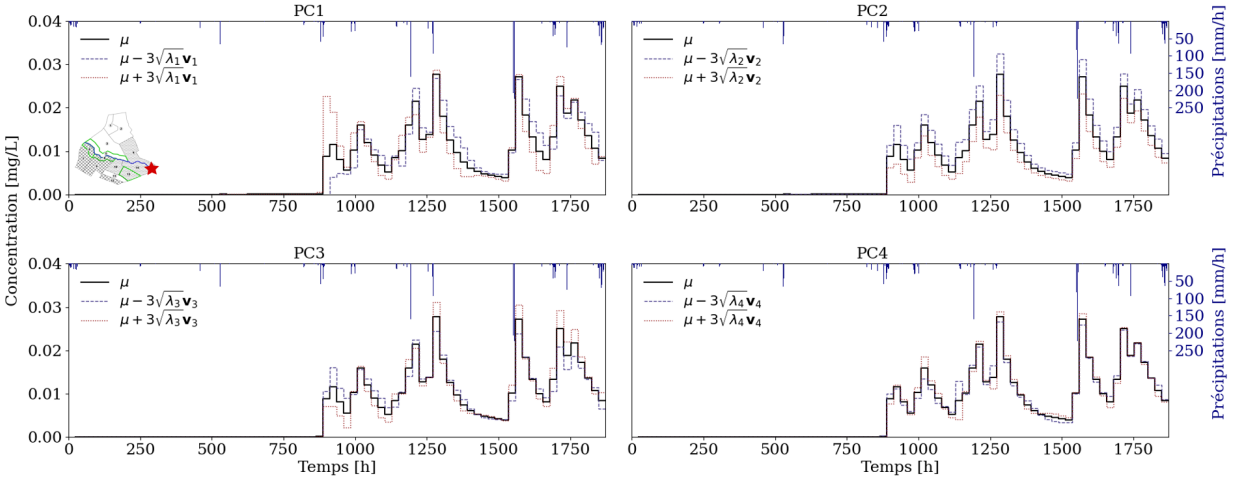


Figure 5.12 – Premières composantes principales PCi, $i \in \{1, 2, 3, 4\}$ représentées comme des perturbations positives et négatives par rapport à la moyenne μ pour la concentration en pesticides journalière moyenne à l'exutoire.

A retenir

- ✓ Deux composantes principales suffisent à représenter 90% de l'inertie des séries d'humidité.
- ✓ La première composante représente le décalage vertical de la série (humidité moyenne), la deuxième n'a pas la même interprétation selon qu'il y ait ou non un plateau de saturation.
- ✓ Quatre composantes principales sont nécessaires pour représenter 90% de l'inertie de la série de concentration
- ✓ L'interprétation des composantes est moins claire pour la concentration que pour l'humidité.

5.2.2 Criblage

Le criblage est réalisé avec la méthode des effets élémentaires de Morris décrite en Section 3.3.2. On utilise p=4 niveaux et R=10 trajectoires. Compte tenu du nombre élevé de paramètres d'entrée et de variables de sortie impliqués, la délimitation visuelle de clusters dans le graphe μ^*-σ peut s'avérer difficile. Pour pallier cette difficulté, on identifie comme influents les paramètres X_i dont les effets normalisés vérifient $\mu_i^* \geq \mu_{min}^*$ ou $\sigma_i \geq \sigma_{min}$ avec μ_{min}^* et σ_{min} des seuils fixés de manière à retenir un nombre raisonnable de paramètres. On note que cette approche reste arbitraire, comme souvent dans les méthodes de criblage.

Humidité de surface

Scénario estival Pour l'humidité de surface sur le scénario estival, on sélectionne les paramètres $\{X_i,\ \mu_i^* \geq 0.1$ ou $\sigma_i \geq 0.1\}$ dans le graphe de Morris normalisé. La Figure 5.13 présente les résultats sur l'UH2 pour ce scénario.

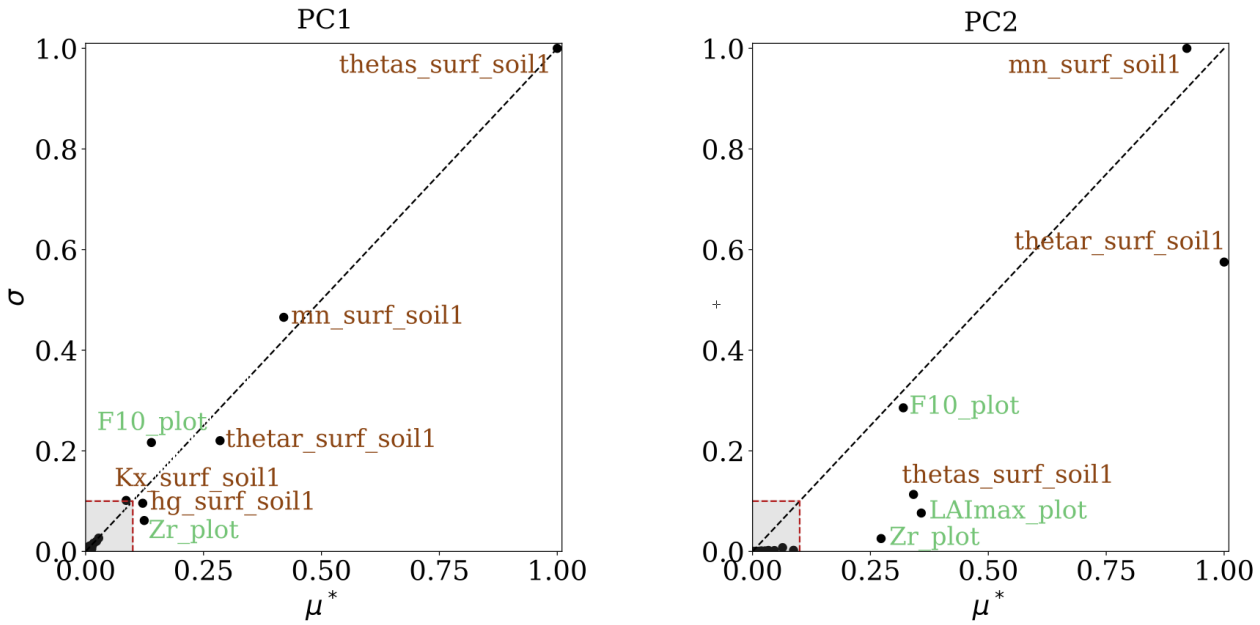


Figure 5.13 – Graphe de Morris normalisé pour les deux premières composantes de l'humidité de surface sur l'UH2 (scénario estival). La partie grisée délimite l'espace du graphe correspondant aux paramètres qui sont exclus. Les couleurs et la définition des noms des paramètres sont détaillés dans la Figure 4.5.

Les clusters de paramètres influents et non influents s'y distinguent assez facilement visuellement, et on retient respectivement 7 et 6 paramètres sur PC1 et PC2. Les paramètres retenus sont majoritairement des paramètres hydrodynamiques de l'horizon de surface (teneur en eau à saturation *thetas*, paramètre de Van Genuchten *mn* et teneur en eau résiduelle *thetar*) qui interviennent dans l'équation de Richards. Il semble ainsi cohérent que ces paramètres influent de manière dominante sur l'humidité. D'autre part, des paramètres de végétation ressortent aussi comme influents. Ces paramètres sont associés à l'extraction racinaire, laquelle est également fortement influente sur l'humidité dans le scénario estival caractérisé par une évapotranspiration marquée. Après union des paramètres retenus pour chacune des composantes principales et chacune des UH, on aboutit à un espace réduit à 44 paramètres d'entrée pour l'ensemble du bassin versant, dont la liste est détaillée dans l'Annexe E. A l'image de l'UH2, les paramètres retenus sont majoritairement des teneurs en eau à saturation, des teneurs en eau résiduelles et des paramètres de Van Genuchten pour les horizons de surface des différents types de sol ainsi que des paramètres de végétation.

Scénario hivernal Pour PC1, on sélectionne les paramètres $\{X_i,\ \mu_i^* \geq 0.08$ ou $\sigma_i \geq 0.2\}$ et pour PC2, l'ensemble $\{X_i,\ \mu_i^* \geq 0.22$ ou $\sigma_i \geq 0.13\}$. La Figure 5.14 présente les résultats du criblage sur l'UH13. 7 paramètres sont retenus pour PC1 et 4 pour PC2. Comme pour le scénario estival, les paramètres retenus pour PC1 sont principalement des paramètres hydrodynamiques de surface intervenant dans l'équation de Richards (*thetas*, *mn* et *hg*). Aucun paramètre de végétation n'est retenu sur PC1 et PC2 montrant que la contribution de l'extraction racinaire est moins marquée que sur le scénario estival, ce qui est cohérent avec une demande évaporative plus limitée. Pour PC2 qui caractérise le plateau de saturation pour l'UH13, des paramètres hydrodynamiques de l'horizon le plus profond sont retenus, suggérant l'existence d'une rétroaction des couches les plus profondes sur la surface. Au final, après union pour les deux composantes, sur toutes les UH, on conserve 52 paramètres, dont la liste est détaillée dans l'Annexe E.

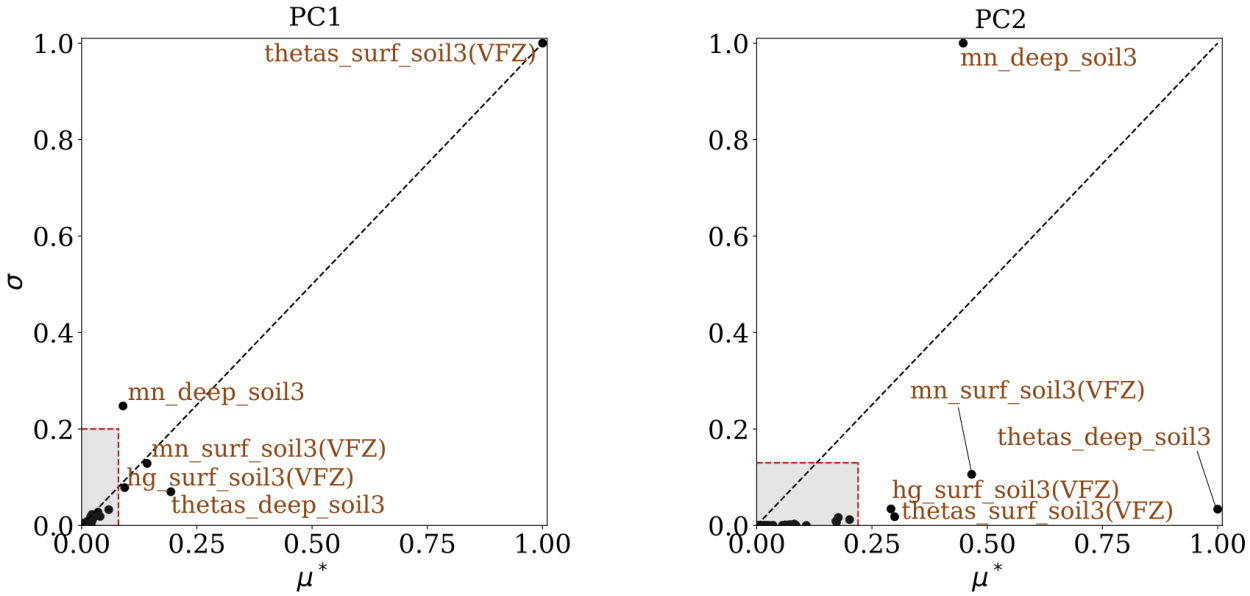


Figure 5.14 – Graphe de Morris normalisé pour les deux premières composantes de l'humidité de surface sur l'UH13 (scénario hivernal). La partie grisée délimite l'espace du graphe correspondant aux paramètres qui sont exclus. Les couleurs et la définition des noms des paramètres sont détaillés dans la Figure 4.5.

Concentration à l'exutoire

La Figure 5.15 illustre les paramètres retenus sur chaque composante principale pour la concentration en pesticides à l'exutoire.

Contrairement à l'humidité, il est difficile de discriminer visuellement les différents clusters de paramètres comme décrit en Figure 3.4 (Chapitre 3), notamment pour PC4. On retient pour chaque mode les paramètres $\{X_i,\ \mu_i^* \geq 0.1$ ou $\sigma_i \geq 0.1\}$. Là encore, le choix de ce seuil est arbitraire, déterminé de sorte à ne retenir qu'un nombre raisonnable de paramètres par composante principale, mais il se trouve qu'il est cohérent avec celui choisi pour

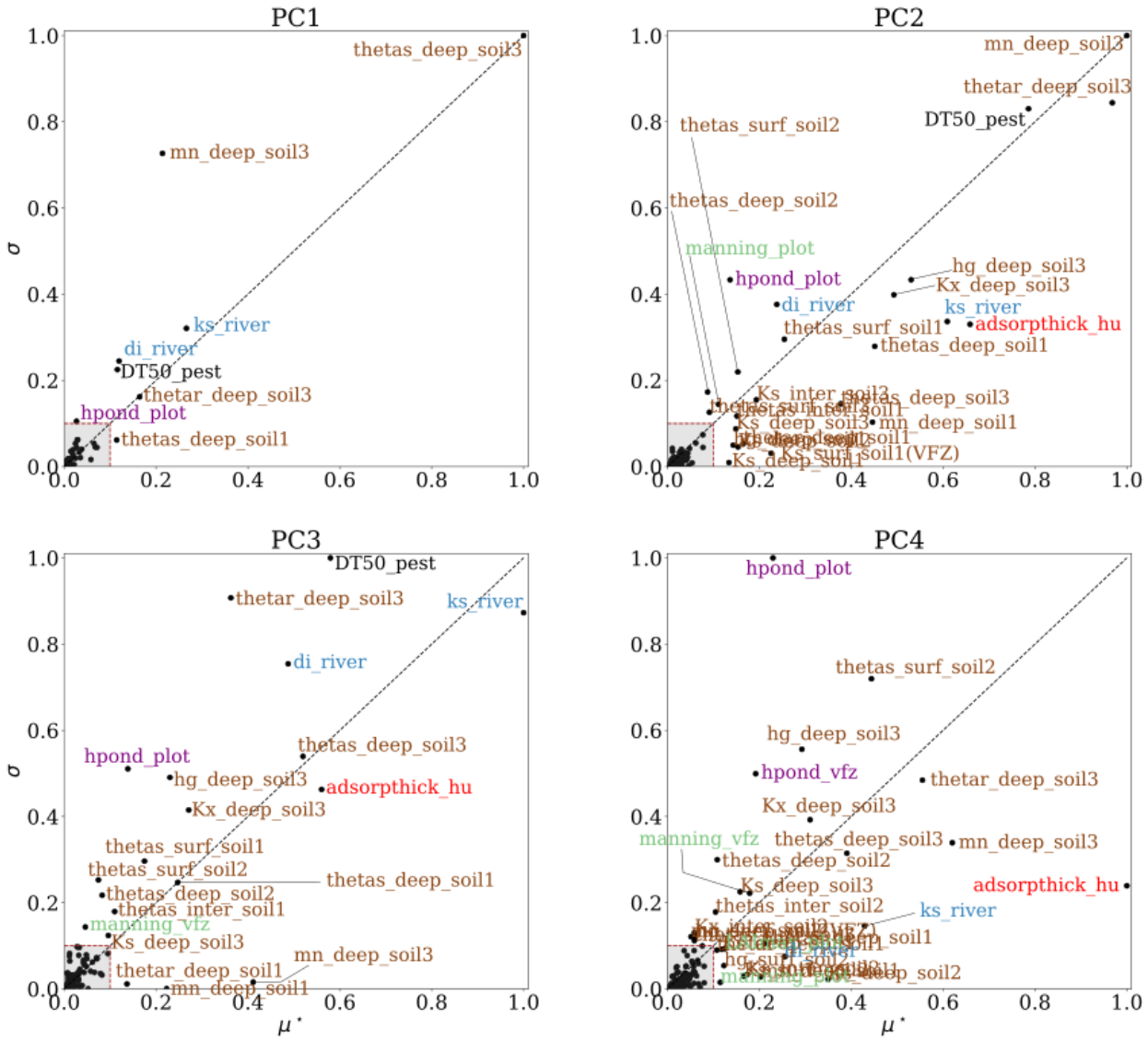


Figure 5.15 – Graphe de Morris normalisé pour les 4 premières composantes de la concentration en pesticides à l'exutoire. La partie grisée délimite l'espace du graphe correspondant aux paramètres qui sont exclus. Les couleurs et la définition des noms des paramètres sont détaillés dans la Figure 4.5.

l'humidité de surface. On retient ainsi 7 paramètres sur PC1, 23 sur PC2, 15 sur PC3 et 25 sur PC4. Comme le montre la Figure 5.15, ces derniers sont souvent redondants d'un mode à l'autre et l'union n'aboutit qu'à 31 paramètres dont la liste est détaillée dans l'Annexe E. Parmi ces paramètres, on retrouve une majorité de paramètres hydrodynamiques pour les différents types de sol, à différentes profondeurs. Les teneurs en eau à saturation *thetas* relatives à l'infiltration verticale mais aussi au calcul de l'adsorption et les conductivités hydrauliques à saturation Ks paramétrant les transferts latéraux d'eau et de pesticides sont en particulier retenus pour plusieurs horizons. Ensuite, des paramètres liés aux échanges nappe-rivière (*ks_river* et *di_river*) sont sans surprise retenus sur les 4 modes. Des paramètres

liés au ruissellement (*hpond*), à l'adsorption (*adsorpthick*) et à la dégradation des pesticides ($DT50$) ressortent également comme influents pour quasiment toutes les composantes.

Conclusion

La méthode de Morris permet de discriminer relativement efficacement les ensembles de paramètres influents, surtout pour l'humidité de surface. En effet, les clusters qui caractérisent classiquement le graphe μ^*-σ se distinguent facilement pour l'humidité de surface, mais c'est bien moins évident pour la concentration à l'exutoire. Cette dernière est probablement impactée par de nombreuses interactions et/ou non linéarités dans le modèle expliquant de telles difficultés. Compte tenu de ces difficultés ainsi que du nombre d'UH et de composantes considérées, il était plus aisé d'automatiser le criblage en fixant des seuils μ^*_{min} et σ_{min}. Le choix de tels seuils dépend fortement de la variable, du scénario considéré mais aussi de l'expérience du modélisateur et de son objectif avec l'analyse de sensibilité. Cela rend l'exercice plus délicat et peut affecter la robustesse des conclusions. Bien qu'arbitraire, l'utilisation de tels seuils pour la sélection est une approche classique dans les applications de la méthode de Morris (Vanuytrecht et al., 2014 ; Garcia et al., 2019). Pour aller plus loin, la procédure de sélection basée sur la combinaison de critères visuels et numériques proposée dans Garcia et al. (2019) pourrait permettre une sélection quasi-automatique, plus rigoureuse des paramètres influents.
D'autre part, on note que pour sélectionner des sous ensembles de paramètres influents, la méthode de Morris fournit des indicateurs quantitatifs μ^* et σ. Les indicateurs μ^* sont souvent considérés comme une bonne approximation des indices de Sobol totaux (Campolongo et al., 2007 ; Herman et al., 2013 ; Saltelli et al., 2004). Toutefois, il est important de rappeler qu'ils doivent être interprétés avec précaution et qu'ils ne sauraient remplacer un classement quantitatif réalisé à l'aide d'une méthode adaptée. Un tel classement fait l'objet de la section suivante.

A retenir

- ✓ Le criblage de Morris permet d'éliminer quasiment deux tiers des paramètres pour chaque variable.
- ✓ Pour l'humidité, on retient principalement des paramètres hydrodynamiques.
- ✓ Pour la concentration, on retient des paramètres de nature variée.
- ✓ Les clusters se distinguent facilement visuellement dans le graphe de Morris pour l'humidité mais sont beaucoup plus flous pour la concentration.
- ✓ La méthode de Morris reste une méthode pratique, rapide et peu coûteuse mais elle elle ne permet pas toujours de dégager des groupes de paramètres à éliminer selon la complexité de la variable considérée.

5.2.3 Classement

Après le criblage, le classement des paramètres influents est réalisé en calculant leurs indices de Sobol totaux et de premier ordre. Pour estimer ces derniers, on utilise une décomposition en polynômes du chaos comme décrit en Section 3.3.2. Celle-ci est tronquée en fixant le degré maximal *pmax* à 5 et la norme q à 0.1. Les intervalles de confiance à 95% associés sont également calculés à partir de 100 bootstraps.

Humidité de surface

Qualité du métamodèle construit par PCE Pour l'humidité, le métamodèle obtenu est de bonne, voire de très bonne qualité, en attestent les valeurs des Q^2 calculées sur un échantillon de test de 1000 points (voir Tableau 5.1 pour les valeurs sur le scénario hivernal). Sur PC1, les Q^2 sont excellents, tous supérieurs à 0.98. Sur PC2, les performances restent satisfaisantes, bien qu'inférieures dans le scénario hivernal, avec des valeurs de Q^2 toujours supérieures à 0.77. Dans le scénario hivernal, ce sont sur les UH 2, 3, 7, 9 et 13 que les valeurs des Q^2 sont les plus faibles (voir Tableau 5.1). Cela correspond aux UH présentant un plateau de saturation et dont l'ensemble se divise parfois en deux faisceaux, un avec plateau et l'autre sans comme illustré sur la Figure 5.4. Pour ces cas-là, l'utilisation d'un seul métamodèle ne peut pas permettre de modéliser correctement les deux comportements simultanés.

UH	1	2	3	4	5	6	7	8	9	10	11	12	13	14
PC1	1.00	0.99	0.99	0.99	0.99	0.99	0.99	1.00	0.99	0.99	1.00	0.99	0.98	0.99
PC2	NA	0.80	0.77	0.99	NA	0.97	0.82	NA	0.80	0.98	NA	0.95	0.78	0.99

Tableau 5.1 – Valeurs de Q^2 associées aux deux premières composantes principales de l'humidité de surface calculées sur un échantillon de test de 1000 points pour toutes les UH (scénario hivernal).

Indices de Sobol par site et par composante principale Pour le scénario **estival**, la Figure 5.16 présente les indices de Sobol totaux sur les scores selon PC1 et PC2. Pour les deux composantes, les indices de Sobol de premier ordre ne sont pas représentés car quasiment égaux aux indices totaux. L'humidité de surface est ainsi exclusivement soumise à des effets directs des paramètres sans interaction notable.
Pour la première composante (PC1, Figure 5.17, haut), le *thetas* de l'horizon de surface local est de loin le paramètre le plus influent sur toutes les UH. Il explique au minimum 75% de la variance de H_1. Parmi les autres paramètres influents, la teneur en eau résiduelle *thetar* et le paramètre de Van Genuchten *mn* de l'horizon de surface ont également une influence modérée. Les intervalles de confiance à 95% sont très faibles, montrant que les indices de Sobol sont calculés avec une grande précision. Pour la PC2, le score est majoritairement sensible aux paramètres *mn*, *thetas* et *thetar* de l'horizon de surface local. Certains paramètres de végétation apparaissent aussi dans ces classements en accord avec la demande évapotranspirative plus marquée dans ce scénario estival.

La Figure 5.17 présente les indices de Sobol totaux sur les scores selon PC1 et PC2 pour chaque UH du scénario **hivernal**. Là encore, les effets interactifs sont négligeables et seuls les indices totaux sont représentés. Les résultats sont très semblables au scénario estival pour PC1 et les *thetas* ont une influence largement majoritaire. Pour la seconde composante, on ne calcule pas les indices de Sobol pour les UH 11, 5 et 8 car PC1 explique déjà plus de 99% de l'inertie de la sortie. Sur les UH restantes, on note une différence marquée de classement entre les UH sans plateau de saturation (12, 6, 4, 10, 14) et les UH présentant un plateau de saturation (2, 3, 7, 9 et 13). Pour les premières, comme dans le scénario estival, le score selon PC2 est majoritairement sensible aux paramètres *mn*, *thetas* et *thetar* de l'horizon de surface local. Par contre, ce sont les paramètres de l'horizon le plus profond qui sont les plus influents pour les UH avec plateau de saturation. Une telle rétroaction de la profondeur sur la surface s'explique car la colonne de sol est entièrement saturée pendant près d'un tiers de la simulation. On retrouve ainsi l'interprétation des résultats du Chapitre 4 (Section 4.2.2) qui reliaient ruissellement de surface et saturation de l'intégralité de la colonne de sol. Contrairement à PC1, les intervalles de confiance montrent une plus grande incertitude sur les indices de Sobol pour PC2. Celle-ci s'explique probablement par les moins bonnes performances du métamodèle sur PC2 que sur PC1 (voir Tableau 5.1).

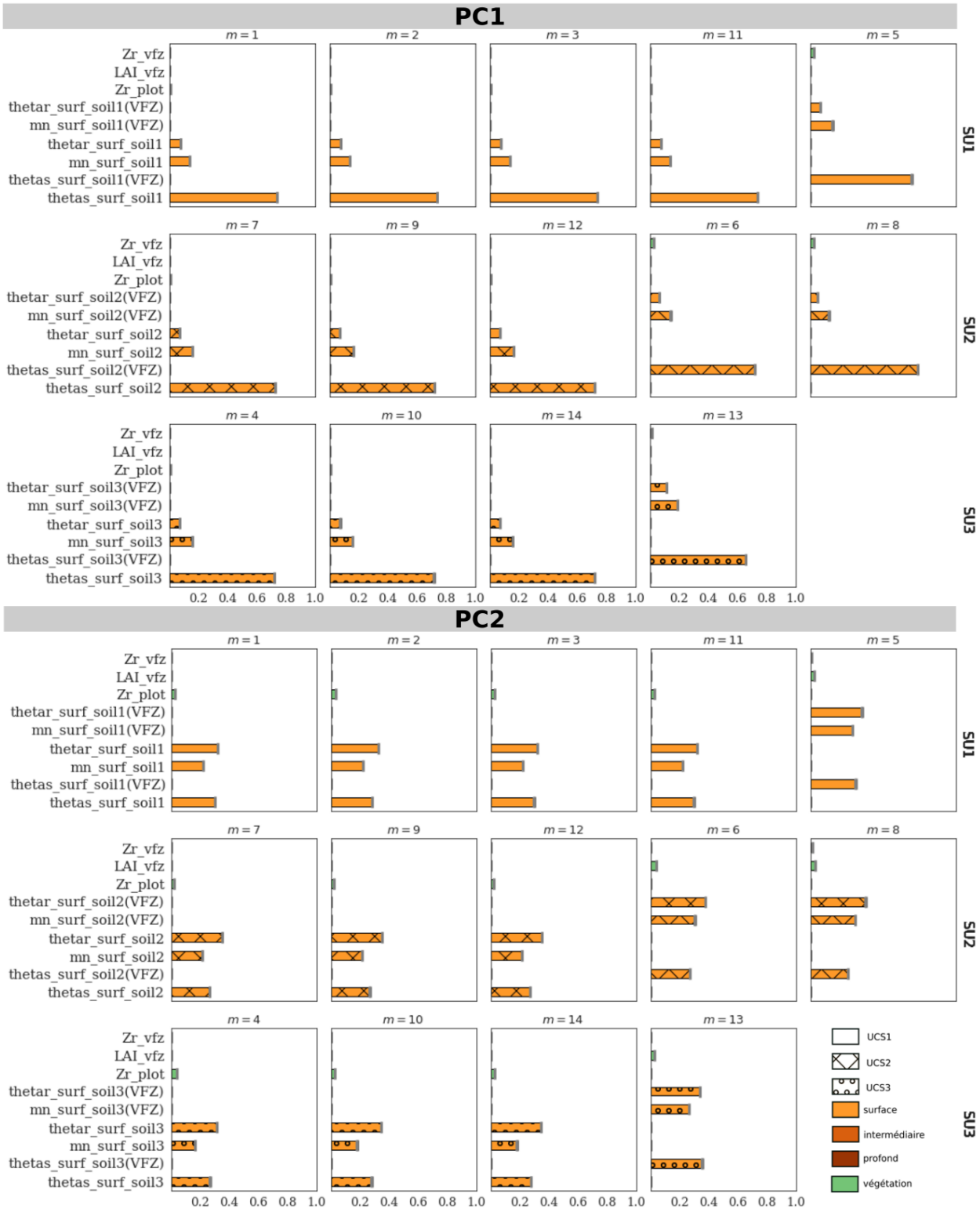


Figure 5.16 – Indices de Sobol totaux calculé sur les scores selon la PC1 (haut) et la PC2 (bas) de l'humidité de surface pour chaque UH et intervalles de confiance à 95% associés (scénario estival). L'indice de l'UH m est indiqué au dessus de chaque classement. Les différentes lignes regroupents respectivement les UH appartenant au types de sol 1, 2 et 3. Dans chaque classement, sont reportés les 10 paramètres aux indices de Sobol les plus élevés. Le type de hachure et la couleur de chaque barre caractérise l'horizon auquel se rapporte chaque paramètre, comme décrit dans la Figure 4.5.

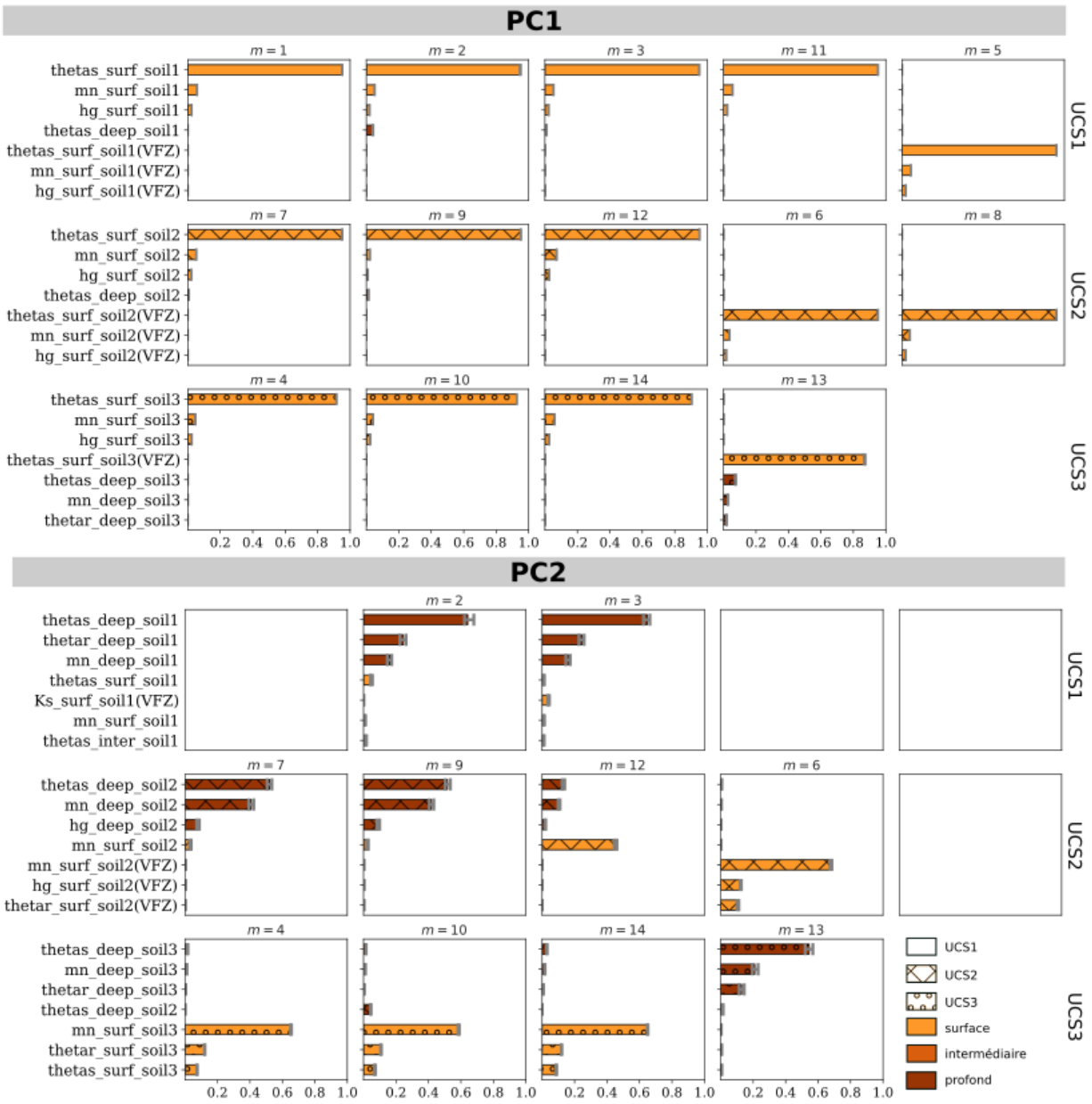


Figure 5.17 – Indices de Sobol totaux calculé sur les scores selon la PC1 (haut) et la PC2 (bas) de l'humidité de surface pour chaque UH et intervalles de confiance à 95% associés (scénario hivernal). L'indice de l'UH m est indiqué au dessus de chaque classement. Les différentes lignes regroupents respectivement les UH appartenant au types de sol 1, 2 et 3. Dans chaque classement, sont reportés les 7 paramètres aux indices de Sobol les plus élevés. Le type de hachure et la couleur de chaque barre caractérise l'horizon auquel se rapporte chaque paramètre, comme décrit dans la Figure 4.5. Pour la PC2, les UH 1, 11, 5 et 8 ne sont pas représentées car la variabilité de l'humidité de surface est essentiellement expliquée par PC1 (tiré de Radišić et al. 2022).

Indices de Sobol aggrégés Une fois que les indices de Sobol ont été calculés pour les deux premières composantes principales, sur toutes les UH, on en déduit des indices aggrégés spatiotemporellement en combinant les Eq. 3.31 et 3.29. Dans l'Eq. 3.29, on considère que pour chaque UH m, $\mathrm{Var}[\mathrm{Y}_m] = \lambda_1^{(m)} + \lambda_2^{(m)}$ et $S_u^{(m)} = GSI_u^{(m)}$. Les indices de Sobol totaux aggrégés sont présentés sur la Figure 5.18 pour l'humidité de surface pour le scénario hivernal et estival. Sans surprise, ces résultats résument les mêmes conclusions que les Figures 5.17, 5.16. En effet, on rappelle que PC1 explique près de 80% d'inertie dans le scénario hivernal et près de 90% d'inertie dans le scénario estival et que les *thetas* des horizons de surface ont les indices de Sobol les plus élevés sur cette dernière. Les indices aggrégés reflètent ainsi ces conclusions avec des *thetas* de surface majoritairement influents pour tous les types de sol, ordonnés par nombre d'UH caractérisées par ce type de sol dans le bassin versant. Dans ce classement, les paramètres liés aux plateaux de saturation ou à la végétation n'apparaissent pas du tout, illustrant comment les aggrégations masquent les hétérogénéités spatiales et temporelles.

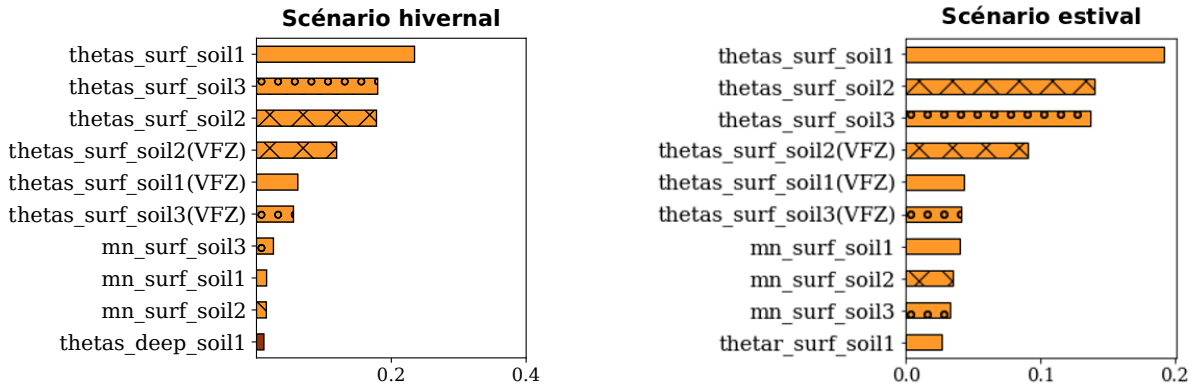


Figure 5.18 – Indices de Sobol totaux aggrégés temporellement et spatialemement GSI pour l'humidité de surface pour le scénario hivernal (gauche) et le scénario estival (droite). Seuls les 10 paramètres avec les plus grandes valeurs de GSI sont représentés. Le type de hachure et la couleur de chaque barre caractérise l'horizon auquel se rapporte chaque paramètre, comme décrit dans la légende de la Figure 4.6.

Humidité dans la colonne de sol

Les conclusions pécédentes pour l'humidité de surface se généralisent à l'humidité dans la colonne de sol et les résultats ne sont pas montrés ici. Celle-ci est majoritairement influencée par le *thetas* de l'horizon local (humidité dans l'horizon intermédiaire influencée par le *thetas* de l'horizon intermédiaire et humidité de l'horizon profond influencée par le *thetas* de l'horizon profond).

Concentration à l'exutoire

Qualité du métamodèle construit par PCE Comme pour l'humidité de surface, la qualité du métamodèle construit par PCE pour la concentration à l'exutoire est évaluée sur un échantillon indépendant de test de 1000 points. Les valeurs des Q^2 calculées (Tableau 5.2) attestent de sa très bonne qualité pour les 4 composantes principales considérées.

	Q^2
PC1	0.99
PC2	0.95
PC3	0.97
PC4	0.89

Tableau 5.2 – Valeurs de Q^2 associées aux 4 premières composantes principales de la concentration en pesticides à l'exutoire calculées sur un échantillon de test de 1000 points (scénario hybride).

Indices de Sobol par composante principale La Figure 5.19 présente les indices de Sobol de premier ordre et totaux pour les 11 paramètres les plus influents sur les scores selon les 4 composantes principales retenues.

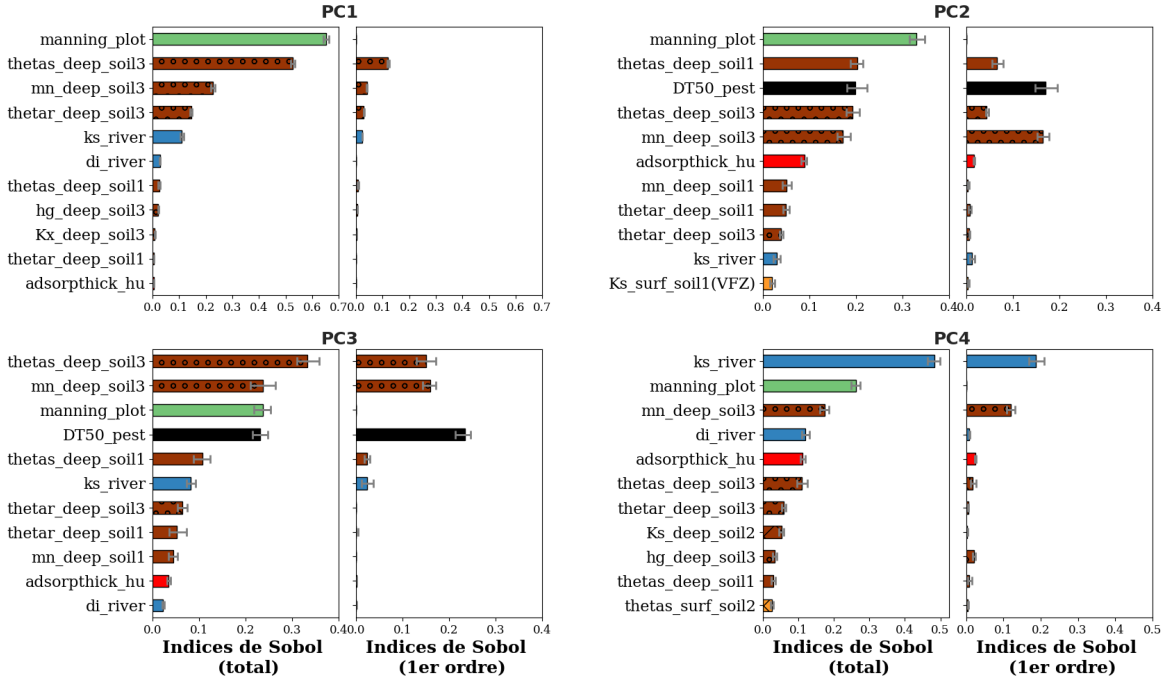


Figure 5.19 – Indices de Sobol totaux et de premier ordre calculés sur les scores selon les PC 1 à 4 pour la concentration à l'exutoire et intervalles de confiance à 95% associés. Le type de hachure et la couleur de chaque barre caractérise le type de paramètre comme décrit dans la Figure 4.5.

Contrairement à l'humidité de surface, les indices de premier ordre sont souvent inférieurs aux indices totaux montrant l'existence de forts effets d'interaction entre les paramètres. La rugosité de Manning des parcelles de vigne *manning_plot* compte parmi les paramètres les plus influents sur les 4 PC. Comme déjà constaté pour les variables intégrées, ce paramètre est influent seulement au travers d'effets interactifs. De la même manière, le *thetas* de l'horizon profond de l'UCS3, les paramètres de la rivière *ks_river* et *di_river* ainsi que l'épaisseur de la couche d'adsorption *adsorpthick_hu* sont aussi largement influents au travers d'effets interactifs. Les intervalles de confiance sont très faibles pour PC1 et limités pour les autres PC, donnant là encore confiance dans les classement obtenus. D'autre part, le top 10 des paramètres les plus influents d'après les indices de Sobol totaux inclut généralement les mêmes paramètres d'une composante principale à l'autre, ce qui était déjà visible dans le graphe de Morris sans être forcément intuitif. Toutefois, concernant la comparaison avec la méthode de Morris, on note que certains paramètres parmi les plus influents, n'étaient pas assortis des valeurs μ^* les plus grandes sur la Figure 5.15. C'est notamment le cas pour la rugosité de Manning. Il semble donc que le calcul des valeurs μ^* ne fournisse pas toujours une bonne approximation des indices de Sobol totaux. Ceci confirme que l'utilisation de la méthode de Morris pour approximer les indices de Sobol doit donc se faire avec la plus grande précaution, notamment en présence de paramètres caractérisés par des effets interactifs marqués.

Indices de Sobol aggrégés Les indices de Sobol aggrégés présentés sur la Figure 5.20 résument globalement les résultats précédents. Comme attendu, c'est le coefficient de Manning qui ressort comme le plus influent, et seulement au travers d'effets interactifs. Viennent ensuite des paramètres hydrodynamiques des horizons profonds, le temps de demi-vie du tebuconazole $DT50$, des paramètres de la rivière et l'épaisseur d'adsorption. Bien que la transposition avec la physique ne soit probablement pas directe, la diversité des types de paramètres retenus illustre la diversité des processus physiques qui contribuent à expliquer le signal de concentration à l'exutoire. Ceci est cohérent avec la position du point étudié, qui intègre des contributions de tous les compartiments et de tous les points du bassin versant.

A retenir

- ✓ A chaque profondeur, l'humidité est surtout sensible à la teneur en eau à saturation *thetas* de l'horizon local.
- ✓ Pour l'humidité, les *thetas* ne sont influents qu'au travers d'effets directs.
- ✓ La concentration à l'exutoire est sensible à une plus grande variété de paramètres, notamment la rugosité de Manning des parcelles de vigne.
- ✓ Les effets interactifs entre paramètres sont majoritaires pour expliquer la sensibilité de la concentration.

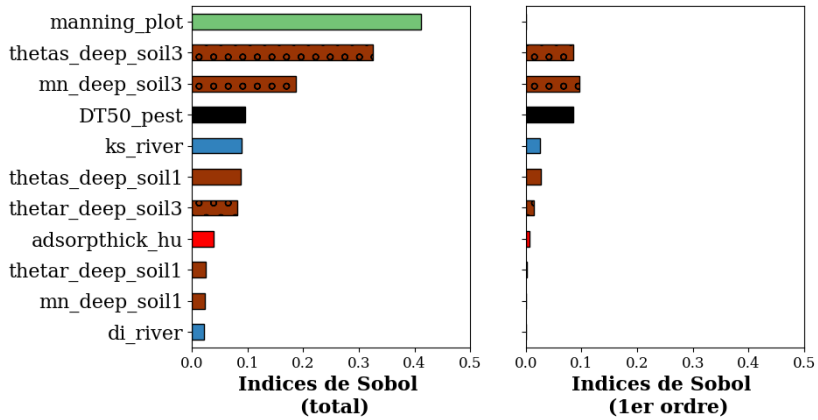


Figure 5.20 – Indices de Sobol totaux (gauche) et de premier ordre (droite) aggrégés temporellement pour la concentration journalière à l'exutoire. Seuls les 11 paramètres avec les plus grandes valeurs d'indice total sont représentés. Le type de hachure et la couleur de chaque barre caractérise le type de paramètre comme décrit dans la Figure 4.5.

5.3 Conclusion

L'objectif de ce chapitre était de caractériser des séries temporelles issues du modèle PESHMELBA afin de développer un cadre d'assimilation de données adapté à ces variables. On s'intéressait particulièrement à l'humiditié de surface et plus généralement à l'humidité dans la colonne de sol ainsi qu'à la concentration en pesticides moyenne journalière à l'exutoire. Pour caractériser ces variables, on cherchait notamment à determiner 1/ si les variables suivent une distribution gaussienne et 2/ quels sont les paramètres les plus influents sur ces variables. Pour cela, une analyse d'incertitude et une analyse de sensibilité ont été successivement réalisées. Pour l'analyse de sensibilité, compte tenu de la nature des variables (spatiotemporelle pour l'humidité et temporelle pour la concentration), une analyse en composantes principales fonctionnelle a d'abord été réalisée pour réduire la dimension temporelle des variables. L'analyse de sensibilité a ensuite été appliquée sur les scores obtenus selon chaque composante principale. Comme pour les variables intégrées, l'analyse de sensibilité s'est faite en deux étapes : une première étape de criblage pour diminuer la taille de l'espace des paramètres d'entrée avec une méthode de Morris puis une étape de classement en calculant les indices de Sobol par décomposition en polynômes du chaos sur les paramètres retenus. Les indices de Sobol ont d'abord été calculés sur chaque score, pour chaque UH puis aggrégés pour obtenir un résumé de la sensibilité spatiotemporelle.
Les résultats ont montré que l'humidité suit en grande partie une distribution gaussienne mais qu'un second mode apparaît pour certaines UH sur le scénario hivernal lorsque le sol se sature entièrement. Malgré des différences de dynamiques et de processus physiques activés selon les scénarios climatiques, l'humidité est toujours influencée majoritairement par la teneur en eau à saturation associée à la profondeur considérée. La concentration moyenne journalière à l'exutoire est elle aussi gaussienne malgré une dynamique plus complexe et plus

difficile à interpréter. Cette complexité s'illustre cependant dans les paramètres influents qui la caractérisent. Leur diversité (rugosité de Manning, temps de demi-vie, paramètres hydrodynamiques profonds...) ainsi que l'importance des effets interactifs mis en évidence par les indices de Sobol, illustrent la diversité des processus physiques et des interactions qui contribuent à cette variable.
Les objectifs, la méthodologie et les principaux résultats de ce chapitre sont résumés sur la Figure 5.21. Les prochains chapitres permettront de détailler comment ces informations sont utilisées pour l'assimilation de données.

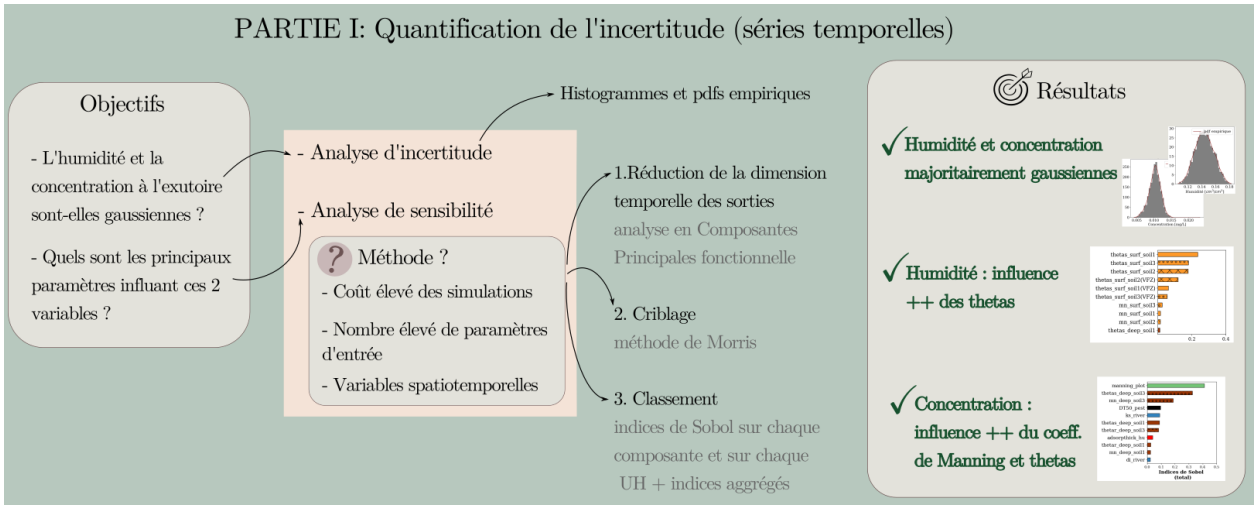


Figure 5.21 – Résumé visuel du Chapitre 5 : objectifs de départ, outils utilisés et principaux résultats obtenus pour la quantification d'incertitude des séries temporelles.

Au delà des informations fournies pour l'assimilation, ce travail de quantification d'incertitude a permis d'acquérir une solide connaissance du fonctionnement de PESHMELBA. Ces connaissances permettent par exemple d'envisager des pistes de simplification du modèle. Au vu des résultats combinés de l'analyse de sensibilité de variables intégrées et dynamiques de natures variées, on peut par exemple envisager une simplification de la courbe de conductivité utilisée. En effet, le modèle de Schaap-Van Genuchten utilisé propose de différencier écoulements préférentiels et matriciels en introduisant notamment les paramètres p (paramètre de pore) et Kx (conductivité hydraulique à saturation de la matrice). Or, il s'avère que ces paramètres, bien que parfois retenus au criblage, ne ressortent comme influents sur aucune des variables avec les indices de Sobol. On pourrait ainsi envisager de simplifier la courbe de conductivité utilisée dans PESHMELBA en s'en tenant à une fonctionnelle de Mualem (Mualem, 1976) qui nécessite moins de paramètres que le modèle actuel.

D'un point de vue méthodologique, la démarche ACPf → criblage → classement → classement aggrégé a permis de réaliser efficacement l'analyse de sensibilité sur des variables spatiotemporelles. Pour aller plus loin, l'étape de criblage pourrait être améliorée en utilisant un test statistique d'indépendance basé sur la mesure HSIC en alternative à la méthode de

Morris. Cette méthode pourrait notamment permettre d'éviter l'utilisation de seuils μ^*_{min} et σ_{min}, parfois délicats à déterminer compte tenu du nombre élevé de paramètres d'entrée considéré. D'autre part, pour gérer le cas de variables bimodales comme rencontrées ici, il serait intéressant d'intégrer une analyse de clusters à l'analyse de sensibilité, comme proposé dans Roux et al. (2021). Cette approche permettrait d'explorer plus finement les différents comportements du modèle et le basculement de l'un à l'autre. En l'occurence, elle pourrait contribuer à déterminer pourquoi certaines simulations d'humidité présentent un plateau de saturation et pas d'autres.

Deuxième partie

Comment réduire les incertitudes dans le modèle PESHMELBA ?

Chapitre 6

Présentation de l'assimilation de données

Sommaire

La quantification d'incertitude présentée dans la première partie de ce travail est une étape indispensable pour préparer l'utilisation opérationnelle de PESHMELBA. Dans la continuité de ces travaux, cette seconde partie aborde la question de la réduction de cette incertitude à l'aide de méthodes d'assimilation de données.
L'objectif de ce chapitre est de présenter le principe général de l'assimilation ainsi que les grandes familles de méthodes existantes. On introduit notamment les "ingrédients" de l'assimilation et son formalisme puis la formulation du problème d'assimilation d'un point de vue bayésien dans la Section 6.2. La Section 6.3 présente plus spécifiquement les méthodes utilisées dans ces travaux. Finalement, la Section 6.4 décrit le problème abordé dans la thèse, les ingrédients dont on dispose et la méthodologie adoptée.

6.1 Principe général et ingrédients de l'assimilation de données

Historiquement, la communauté de la météorologie a été la première à s'emparer de l'assimilation de données et à développer les outils permettant son utilisation opérationnelle (Le Dimet et Talagrand, 1986 ; Houtekamer et Mitchell, 2001). Elle s'est ensuite étendue au domaine de l'océanographie où les contraintes des applications (fortes non linéarités, taille importante du problème considéré, etc.) ont nécessité le développement de méthodes adaptées (Evensen, 1992 ; Blayo et al., 2003 ; Brankart et al., 2003). Aujourd'hui, l'assimilation de données est utilisée dans de nombreux domaines des géosciences, notamment en hydrologie (Le Dimet et al., 2009 ; Crow et Ryu, 2009 ; Devers et al., 2020) ou en modélisation environnementale (Lauvernet et al., 2008 ; Rochoux et al., 2014). Son efficacité ainsi que la disponibilité croissante d'observations, notamment avec le développement des missions satellites, lui garantissent un avenir radieux.

Asch et al. (2016) proposent la définition suivante de l'assimilation de données : *une analyse qui consiste à estimer au mieux l'état vrai d'un système en combinant de manière optimale des observations distribuées dans le temps et l'espace, un modèle dynamique et leurs incertitudes respectives.*

Cette définition fait intervenir les différents ingrédients de l'assimilation : **état**, **observations**, **modèle** et **incertitudes** (ou **erreurs**). Ces différentes notions et les notations associées sont illustrées sur la Figure 6.1 et présentées dans les paragraphes suivants en se basant largement sur Blayo et al. (2019) et Asch et al. (2016).

L'état du système

L'état du système est en général représenté sous la forme discrète d'un vecteur d'état $\mathbf{x} \in \mathbb{R}^n$. Le vecteur d'état peut contenir les valeurs d'une ou plusieurs variables physiques

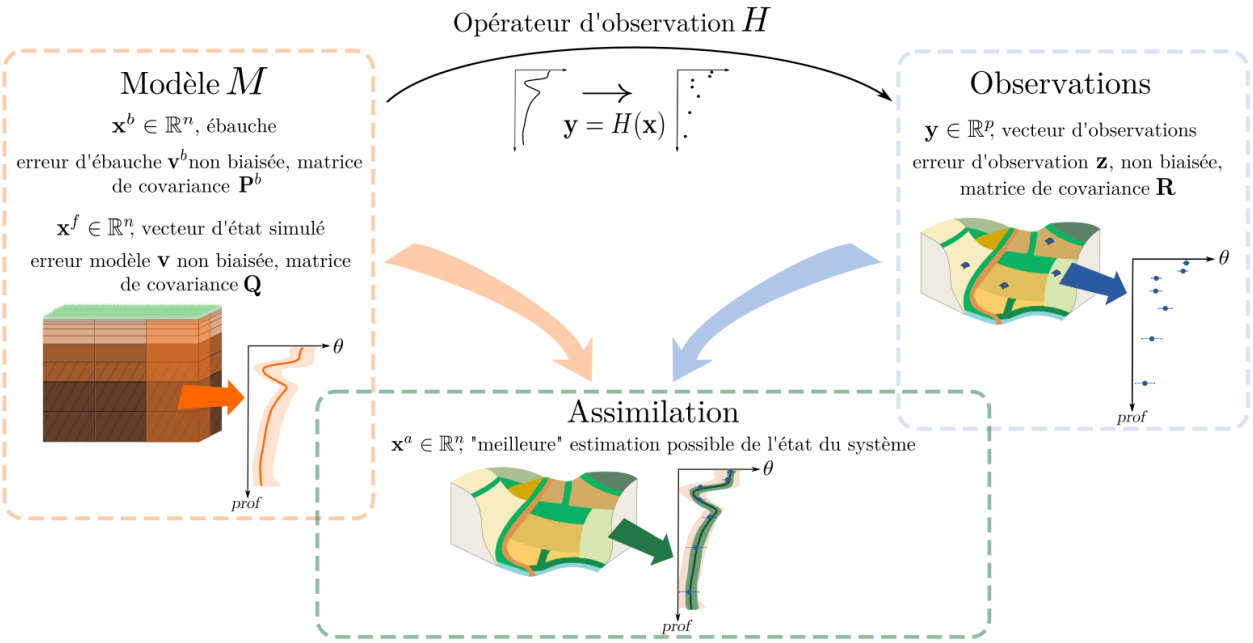


Figure 6.1 – Principe de l'assimilation de données, ingrédients et notations utilisées illustrés sur un exemple d'estimation de profils verticaux d'humidité.

d'intérêt pour différents points de l'espace. Les valeurs des conditions initiales, des conditions limites ou de certains paramètres d'entrée du modèle peuvent aussi apparaître dans le vecteur d'état.
Dans la suite, on distinguera si nécessaire $\mathbf{x}^t$ l'état vrai (*true*) du système aux points discrétisés d'intérêt, $\mathbf{x}^b$ la première estimation de l'état dont on dispose, provenant de connaissance *a priori* ou d'une simulation préalable (*ébauche*, *prior* ou *background*), $\mathbf{x}^f$ l'état du système estimé par le modèle seulement (*prévision* ou *forecast*) et $\mathbf{x}^a$ l'état analysé (*a posteriori* ou *analyse*) obtenu après assimilation de données. Que ce soit pour l'état vrai, l'état prédit par le modèle ou l'état analysé, on note $\mathbf{x}_k$ la valeur du vecteur $\mathbf{x}$ au temps t_k.

Le modèle et ses incertitudes

Le modèle discret $M_{k\to k+1} : \mathbb{R}^n \to \mathbb{R}^n$ permet de propager le vecteur d'état du temps t_k au temps t_{k+1}. On a ainsi la loi d'évolution suivante :

$$\mathbf{x}^t_{k+1} = M_{k\to k+1}(\mathbf{x}^t_k) + \mathbf{v}_k \tag{6.1}$$

où le terme $\mathbf{v}_k$ désigne l'erreur commise par le modèle à l'instant t_k lors de la propagation de l'état du système. Une telle erreur provient des différents hypothèses et approximations utilisées dans la construction du modèle : discrétisation spatiale et temporelle, équations utilisées représentant de manière parcellaire ou inadaptée la réalité, méconnaissance des conditions limites et initiales et de certains paramètres d'entrée du modèle.
En général, et dans ces travaux, on suppose que l'erreur modèle est non biaisée et on note

$\mathbf{Q}_k \in \mathbb{R}^{n\times n}$ sa matrice de covariance à l'instant t_k. L'erreur d'ébauche est également supposée non biaisée et on note $\mathbf{P}^b \in \mathbb{R}^{n\times n}$ sa matrice de covariance.

Les observations et leurs incertitudes

Les observations sont contenues dans un vecteur d'observations $\mathbf{y} \in \mathbb{R}^p$. Elles n'ont pas la même résolution spatiotemporelle que le vecteur d'état et en général, on dispose de beaucoup moins d'observations que de variables à corriger ($p << n$). A chaque temps t_k pour lequel on dispose d'observations, on définit $H_k : \mathbb{R}^n \to \mathbb{R}^p$, l'opérateur d'observation discret qui permet de passer de l'espace de l'état à l'espace des observations, et on a :

$$\mathbf{y}_k = H_k(\mathbf{x}_k^t) + \mathbf{z}_k \tag{6.2}$$

où le terme $\mathbf{z}_k$ désigne l'erreur d'observation à l'instant t_k. L'erreur d'observation est souvent délicate à définir, notamment car elle peut intégrer plusieurs contributions : erreur de mesure, erreur liée au preprocessing des données brutes (par exemple lors d'inversion ou d'interpolation), erreur de représentativité, etc.
L'erreur d'observation est supposée non biaisée et on note $\mathbf{R}_k \in \mathbb{R}^{p\times p}$ sa matrice de covariance à l'instant t_k.

"Meilleure estimation possible"

L'objectif de l'assimilation de données est de déterminer la "meilleure estimation possible" $\mathbf{x}^a$ de l'état du système. On parle de meilleure estimation possible car la connaissance exacte du système physique est inatteignable. Par exemple, l'assimilation de données ne peut pas directement réduire les erreurs de représentativité (erreurs dans la discrétisation ou dans la représentation de la physique). Ainsi, la "meilleure estimation possible" que l'on peut obtenir en pratique est une approximation raisonnable de l'estimation optimale de l'état.

Il existe une grande variété de méthodes qui diffèrent dans la manière dont elles définissent cette "meilleure estimation". Une classification commune des méthodes d'assimilation est présentée sur la Figure 6.2 et distingue notamment les méthodes variationnelles comme le 4D-Var (*4 Dimensional VARiational assimilation*, Le Dimet et Talagrand 1986) des méthodes statistiques comme l'EnKF (*Ensemble Kalman filter*, Evensen 1994). Les méthodes variationnelles se basent sur une approche déterministe et sur la théorie du contrôle optimal. Elles cherchent à minimiser une fonction coût représentant l'écart entre l'état du système et les deux sources d'informations disponibles : l'ébauche et les observations. Les méthodes statistiques considèrent quant à elles l'état du système comme un vecteur aléatoire dont on observe partiellement une réalisation. L'objectif est alors de caractériser au mieux ce vecteur aléatoire en déterminant son espérance, sa variance, voire sa densité de probabilité toute entière.
Cette distinction est cependant arbitraire et dans certains cas particuliers (modèle linéaire,

erreurs gaussiennes, etc.), les deux approches sont équivalentes et convergent vers la même solution. Si l'approche statistique est souvent plus complexe et coûteuse en temps de calcul que l'approche variationnelle, elle fournit aussi une information plus complète sur l'état optimal (solution moyenne et variabilité avec potentiellement l'ensemble des membres possibles). Ainsi, plus récemment, de nombreuses méthodes hybrides tendent à combiner les deux approches pour tirer profit à la fois de la robustesse et de la relative rapidité des méthodes variationnelles et de l'information complète fournie par le point de vue statistique.

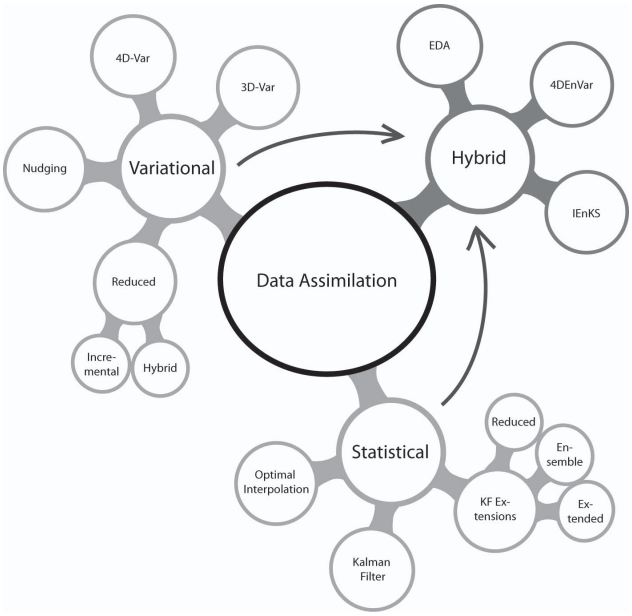


Figure 6.2 – Classification possible des méthodes d'assimilation de données (Asch et al., 2016).

Les différents objectifs de l'assimilation

L'assimilation de données peut être mise en oeuvre pour atteindre conjointement un ou plusieurs objectifs :

- corriger l'état du système présent et/ou passé (réanalyse) ;
- estimer l'état du système à des pas de temps futurs (prévision) ;
- estimer les paramètres d'entrée du modèle ou ses conditions initiales (calibration) ;
- étudier l'impact du réseau d'observations (optimisation).

Dans le cadre de cette thèse, l'assimilation est utilisée pour corriger simultanément des variables du modèle PESHMELBA et estimer certains de ses paramètres d'entrée. On parle alors d'**estimation jointe**. Le vecteur d'état $\mathbf{x}$ est qualifié d'**augmenté** puisqu'il contient à la fois les variables d'état et les paramètres d'entrée cibles.

6.2 Zoom sur l'approche bayésienne de l'assimilation de données

La structure complexe du modèle PESHMELBA le rend difficilement différentiable et limite ainsi l'application de méthodes variationnelles utilisant un modèle adjoint. Ainsi, dans ces travaux, on se base plutôt sur la famille des méthodes statistiques d'assimilation de données. Pour cela, cette section présente la formulation bayésienne du problème d'estimation qui en définit le cadre général. Le filtre de Kalman, la méthode la plus connue qui en découle est ensuite décrite.

6.2.1 Formulation bayésienne du problème d'estimation

D'un point de vue statistique, l'état $\mathbf{x}$ est considéré comme un vecteur aléatoire noté X et l'objectif de l'assimilation de données est de déterminer la densité de probabilité de l'état analysé $p(\mathbf{x}_k^a)$ [1] à un ou plusieurs pas de temps t_k avec $k \in \{1, ..., K\}$, à partir d'une ébauche au temps t_0, notée $p(\mathbf{x}^b)$ (ou $p(\mathbf{x}_0)$) et de la connaissance fournie par les observations $\mathbf{y}_k$, $k \in \{1, ..., K\}$.

La forme de la densité de probabilité conditionnée aux observations à estimer (densité *a posteriori*) dépend du problème formulé :

- Dans un problème de **filtrage**, on cherche à obtenir la meilleure estimation de l'état actuel du système à partir de l'ébauche $p(\mathbf{x}^b)$ et des observations présentes et passées. La densité de probabilité que l'on cherche à calculer s'écrit alors :

$$p(\mathbf{x}_K|\mathbf{y}_1, ..., \mathbf{y}_K) \text{ aussi notée } p(\mathbf{x}_K|\mathbf{y}_{1:K}) \tag{6.3}$$

- Dans un problème de **prévision**, on cherche à estimer un état postérieur aux observations disponibles. La densité de probabilité à estimer s'exprime :

$$p(\mathbf{x}_{K+l}|\mathbf{y}_{1:K}) \text{ avec } l > 0 \tag{6.4}$$

- Dans un problème de **lissage**, on cherche à estimer un ou plusieurs états passés à partir d'observations ultérieures. On évoquera notamment le lissage à point fixe qui correspond à l'estimation de l'état à un instant donné à partir d'observations ultérieures :

$$p(\mathbf{x}_k|\mathbf{y}_{1:K}) \text{ avec } k < K \tag{6.5}$$

et le lissage à intervalle fixe où l'on recherche l'intégralité d'une trajectoire sur un

1. Pour alléger les notations, on abrège la notation $p_{\mathbf{X}}(\mathbf{x})$ qui désigne la densité de probabilité du vecteur aléatoire $\mathbf{X}$, en $p(\mathbf{x})$. De même la densité de probabilité de $\mathbf{X}$ conditionnée à $\mathbf{Y}$, $p_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y})$ est abrégée en $p(\mathbf{x}|\mathbf{y})$.

intervalle :

$$p(\mathbf{x}_{0:K}|\mathbf{y}_{1:K}) \tag{6.6}$$

Le lissage est particulièrement intéressant dans le cadre d'exercices de réanalyse qui visent à obtenir la meilleure estimation de l'état d'un système de manière retrospective, en utilisant toute les informations disponibles (VAN LEEUWEN et EVENSEN, 1996 ; CROW et RYU, 2009 ; COSME et al., 2010).

Dans les paragraphes suivants, on se concentre sur la description générale d'un problème de **filtrage**.

Dans la suite, l'évolution du vecteur d'état au cours du temps décrite par l'Eq. (6.1) est représentée de façon équivalente selon une **densité de transition** notée $p(\mathbf{x}_k|\mathbf{x}_{k-1})$. De même, la relation entre le vecteur d'état $\mathbf{x}_k$ et le vecteur d'observations $\mathbf{y}_k$ exprimée dans l'Eq. (6.2) est représentée par la **fonction de vraisemblance** $p(\mathbf{y}_k|\mathbf{x}_k)$.

Hypothèses et théorème de Bayes

La formulation du problème d'estimation s'appuie généralement sur une série d'hypothèses concernant la séquence des états et des observations. On suppose que la séquence des états définit une chaîne de Markov cachée et celle des observations une chaîne de Markov. De telles hypothèses permettent de simplifier les densités de probabilité conditionnelles comme suit :

1. La densité de probabilité du vecteur $\mathbf{x}_k$ au temps t_k est seulement déterminée par sa valeur la plus récente au temps t_{k-1} :

$$p(\mathbf{x}_k|\mathbf{x}_1, ..., \mathbf{x}_{k-1}) = p(\mathbf{x}_k|\mathbf{x}_{k-1}) \tag{6.7}$$

2. La densité de probabilité des observations au temps t_k est seulement déterminée par l'état du système au temps t_k :

$$p(\mathbf{y}_k|\mathbf{x}_1, ..., \mathbf{x}_k) = p(\mathbf{y}_k|\mathbf{x}_k) \tag{6.8}$$

3. La densité de probabilité du vecteur $\mathbf{x}_k$ au temps t_k ne dépend des observations passées qu'au travers de son propre état passé :

$$p(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathbf{y}_1, ..., \mathbf{y}_{k-1}) = p(\mathbf{x}_k|\mathbf{x}_{k-1}) \tag{6.9}$$

La formulation bayésienne du problème d'estimation se base ensuite sur le théorème de Bayes qui permet à un instant donné de déterminer la densité *a posteriori* $p(\mathbf{x}_k|\mathbf{y}_k)$ à partir d'une information *a priori* sur l'état $p(\mathbf{x}_k)$ et de la vraisemblance $p(\mathbf{y}_k|\mathbf{x}_k)$:

$$p(\mathbf{x}_k|\mathbf{y}_k) = \frac{p(\mathbf{x}_k)p(\mathbf{y}_k|\mathbf{x}_k)}{p(\mathbf{y}_k)} \tag{6.10}$$

Le terme $p(\mathbf{y}_k)$ n'est en général pas calculé et comme il n'intervient qu'en tant que terme de normalisation, l'Equation (6.10) peut être écrite sous la forme générale suivante :

$$p(\mathbf{x}_k|\mathbf{y}_k) \propto p(\mathbf{x}_k)p(\mathbf{y}_k|\mathbf{x}_k) \tag{6.11}$$

Filtrage séquentiel

Grâce aux hypothèses formulées sur les probabilités conditionnelles et au théorème de Bayes, la résolution du problème de filtrage et donc le calcul de la densité $p(\mathbf{x}_K|\mathbf{y}_{1:K})$ peut être effectué de manière récursive. Pour cela, le filtrage séquentiel se base sur la succession de cycles d'assimilation qui alternent chacun une étape de prévision fondée sur l'utilisation du modèle et une étape d'analyse qui permet d'intégrer l'observation reçue à l'instant courant, comme illustré sur la Figure 6.3.

1. L'étape de **prévision** permet de propager la densité de probabilité de l'état entre deux instants où sont disponibles des observations. Elle consiste à calculer la densité de probabilité prédite à l'instant t_k, notée $p(\mathbf{x}_k|\mathbf{y}_{1:k-1})$ à partir de la densité de probabilité *a posteriori* précédente $p(\mathbf{x}_{k-1}|\mathbf{y}_{1:k-1})$ et de la densité de transition $p(\mathbf{x}_k|\mathbf{x}_{k-1})$ représentant l'utilisation du modèle grâce à l'équation de Chapman-Kolmogorov :

$$p(\mathbf{x}_k|\mathbf{y}_{1:k-1}) = \int_{\mathbb{R}^n} p(\mathbf{x}_k|\mathbf{x}_{k-1})p(\mathbf{x}_{k-1}|\mathbf{y}_{1:k-1})d\mathbf{x}_{k-1} \tag{6.12}$$

   Cette densité de probabilité reflète toutes les incertitudes des variables considérées lorsque l'information transmise par les observations courantes n'est pas prise en compte.

2. L'étape d'**analyse** se base ensuite sur le théorème de Bayes pour mettre à jour la densité de probabilité de l'état à partir des observations disponibles au pas de temps actuel et de leur incertitude exprimée au travers de la vraisemblance :

$$\begin{aligned} p(\mathbf{x}_k|\mathbf{y}_{1:k}) &= p(\mathbf{x}_k|\mathbf{y}_k, \mathbf{y}_{1:k-1}) \\ &= \frac{p(\mathbf{y}_k|\mathbf{x}_k, \mathbf{y}_{1:k-1})p(\mathbf{x}_k|\mathbf{y}_{1:k-1})}{p(\mathbf{y}_k|\mathbf{y}_{1:k-1})} \\ &= \frac{p(\mathbf{y}_k|\mathbf{x}_k)p(\mathbf{x}_k|\mathbf{y}_{1:k-1})}{p(\mathbf{y}_k|\mathbf{y}_{1:k-1})} \end{aligned} \tag{6.13}$$

   Cet état analysé peut alors servir de point de départ pour un nouveau cycle d'assimilation en permettant d'initialiser une nouvelle étape de prévision entre t_k et t_{k+1}.

Ainsi, à partir d'une connaissance sur l'état initial du système $p(\mathbf{x}^b)$, il est possible en appliquant les Eq. (6.12) et (6.13) de calculer récursivement la densité de probabilité de l'état du système en intégrant l'information contenue dans les observations à chaque fois qu'une

nouvelle mesure est disponible.
Cependant, ces équations ne peuvent pas toujours être résolues explicitement. Si l'on suppose que le système est linéaire et gaussien et que l'ébauche $p(\mathbf{x}^b)$ est aussi gaussienne, alors la densité de probabilité *a posteriori* est aussi gaussienne et l'on dispose d'une solution explicite aux Eq. (6.12) et (6.13). Ce cas particulier correspond au filtre de Kalman (Kalman, 1960) présenté dans la Section 6.2.2. Par contre dans des cas plus généraux, on ne dispose pas de solution exacte pour les équations du filtre séquentiel. Il faut alors recourir à des techniques d'approximation, permettant par exemple de linéariser le modèle localement comme dans le filtre de Kalman étendu (EKF, Jazwinski 1970). Les méthodes d'approximation particulaires comme le filtre de Kalman d'ensemble (EnKF, Evensen 1994) ou le filtre SIR (*sampling importance resampling*, Gordon et al. 1993) basées sur des méthodes de Monte Carlo sont également largement utilisées. Chacune de ces techniques d'approximation présente des avantages et des inconvénients et leur pertinence dépend grandement des spécificités de l'application (densité de probabilité de l'état gaussienne ou non, degré de non-linéarité, taille du vecteur d'état, etc.).

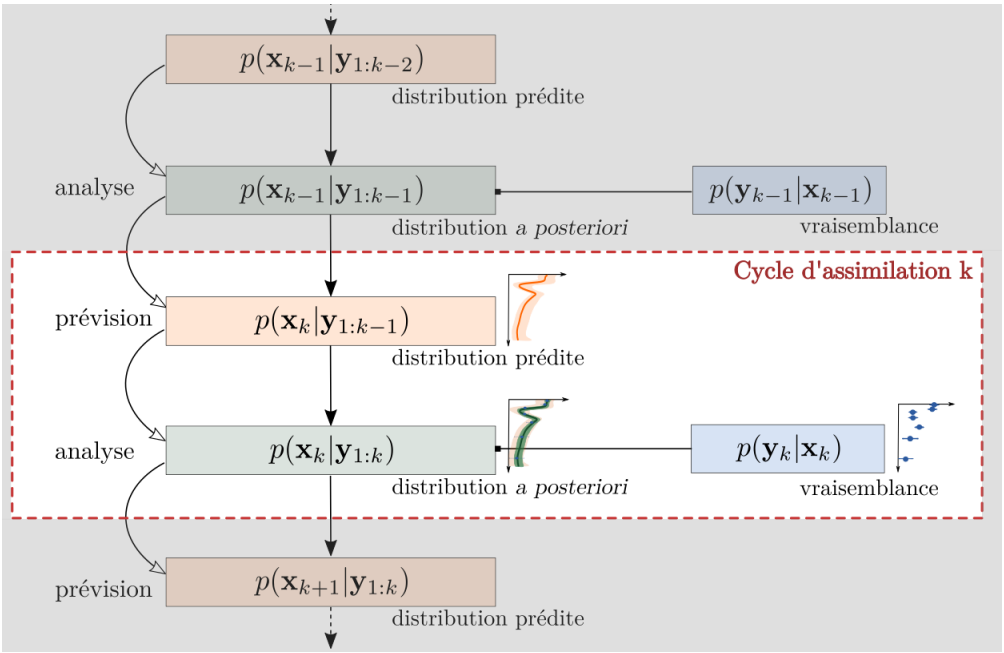


Figure 6.3 – Diagramme du filtrage séquentiel bayésien. Chaque cycle d'assimilation comprend une étape de prévision et une étape d'analyse.

6.2.2 Le filtre de Kalman

Les équations du filtre de Kalman (Kalman, 1960) correspondent à la résolution exacte des équations du filtre séquentiel qui vient d'être présenté dans le cas où le système est *linéaire* et *gaussien*. Considérer un système linéaire et gaussien revient à formuler les hypothèses

suivantes sur le modèle, l'opérateur d'observation, l'état et les observations (en plus de celles présentées dans le paragraphe 6.1) :

- L'ébauche suit une distribution gaussienne : $\mathbf{x}^b \sim \mathcal{N}(\mathbf{0}, \mathbf{P}^b)$.

- L'erreur modèle et l'erreur d'observation sont indépendantes et suivent des distributions gaussiennes :

$$\mathbf{v}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_k) \tag{6.14}$$
$$\mathbf{z}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_k) \tag{6.15}$$

- Le modèle est linéaire et l'Eq. (6.1) s'écrit donc :

$$\mathbf{x}^t_{k+1} = \mathbf{M}_{k+1}\mathbf{x}^t_k + \mathbf{v}_k \tag{6.16}$$

où $\mathbf{M}_k$ désigne la matrice du modèle.

- L'opérateur d'observation est linéaire et l'Eq. (6.2) s'écrit donc :

$$\mathbf{y}_k = \mathbf{H}_k\mathbf{x}^t_k + \mathbf{z}_k \tag{6.17}$$

où $\mathbf{H}_k$ est la matrice de l'opérateur d'observation.

Le filtre de Kalman permet de propager les deux premiers moments de la densité de probabilité $p(\mathbf{x}_k|\mathbf{y}_{1:k})$ en implémentant l'alternance des étapes de prévision et d'analyse qui constituent le filtre séquentiel. Le principe du filtre de Kalman appliqué à la correction d'une trajectoire est illustré sur la Figure 6.4.
Dans la suite, on note $\hat{\mathbf{x}}^f_k = \mathbb{E}[\mathbf{X}_k|\mathbf{Y}_{1:k-1}]$ et $\hat{\mathbf{x}}^a_k = \mathbb{E}[\mathbf{X}_k|\mathbf{Y}_{1:k}]$ et $\mathbf{P}^f_k = \mathbb{E}[(\mathbf{X}_k - \hat{\mathbf{X}}^f_k)(\mathbf{X}_k - \hat{\mathbf{X}}^f_k)]$ et $\mathbf{P}^a_k = \mathbb{E}[(\mathbf{X}_k - \hat{\mathbf{X}}^a_k)(\mathbf{X}_k - \hat{\mathbf{X}}^a_k)]$ ces moments.

Prévision

D'une part, on suppose qu'à t_{k-1}, la densité de probabilité *a posteriori* est définie comme suit :

$$p(\mathbf{x}_{k-1}|\mathbf{y}_{1:k-1}) = \mathcal{N}(\hat{\mathbf{x}}^a_{k-1}, \mathbf{P}^a_{k-1}) \tag{6.18}$$

D'autre part, compte tenu des hypothèses (6.16) et (6.14), la densité de transition est définie comme :

$$p(\mathbf{x}_k|\mathbf{x}_{k-1}) = \mathcal{N}(\mathbf{M}_k\mathbf{x}_{k-1}, \mathbf{Q}_k) \tag{6.19}$$

où l'on rappelle que $\mathbf{Q}_k$ désigne la matrice d'erreur modèle au temps t_k.

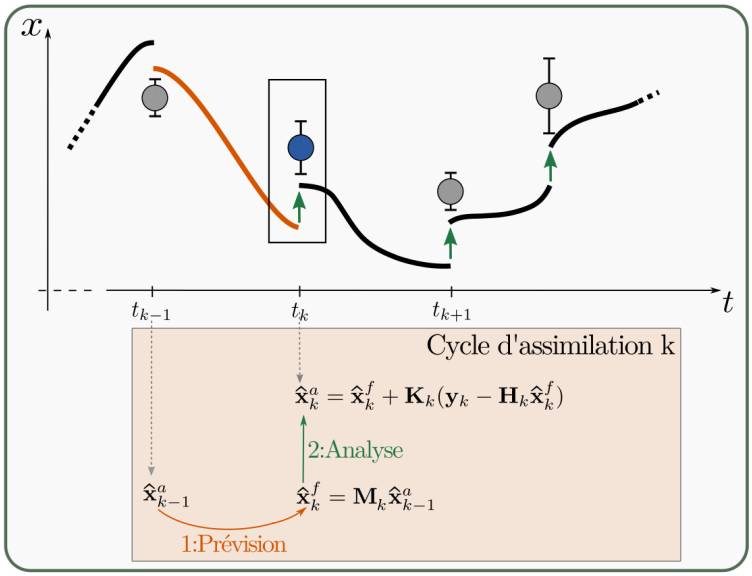


Figure 6.4 – Schéma d'assimilation séquentielle implémenté par le filtre de Kalman. Le point bleu dans la fenêtre noire schématise les observations qui sont utilisées pour le cycle d'assimilation actuel. Les observations non utilisées sont schématisées en gris.

En injectant les expressions analytiques de ces distributions gaussiennes dans l'équation de prévision du filtre séquentiel (6.12), on trouve que la densité de probabilité *prédite* à l'instant t_k s'exprime :

$$p(\mathbf{x}_k|\mathbf{y}_{1:k-1}) = \mathcal{N}(\hat{\mathbf{x}}_k^f, \mathbf{P}_k^f) \tag{6.20}$$

avec :

$$\hat{\mathbf{x}}_k^f = \mathbf{M}_k\hat{\mathbf{x}}_{k-1}^a \tag{6.21}$$
$$\mathbf{P}_k^f = \mathbf{M}_k\mathbf{P}_{k-1}^a\mathbf{M}_k^T + \mathbf{Q}_k \tag{6.22}$$

Analyse

L'état est ensuite mis à jour pendant l'étape d'analyse à partir du théorème de Bayes comme exprimé dans (6.13).
Compte tenu des hypothèses (6.15) et (6.17), la vraisemblance s'écrit d'abord :

$$p(\mathbf{y}_k|\mathbf{x}_k) = \mathcal{N}(\mathbf{H}_k\mathbf{x}_k, \mathbf{R}_k) \tag{6.23}$$

où l'on rappelle que $\mathbf{R}_k$ désigne la matrice d'erreur d'observation au temps t_k.
En injectant (6.20) et (6.23) dans (6.13), on obtient :

$$p(\mathbf{x}_k|\mathbf{y}_{1:k}) \propto \mathcal{N}(\mathbf{H}_k\mathbf{x}_k, \mathbf{R}_k)\mathcal{N}(\hat{\mathbf{x}}_k^f, \mathbf{P}_k^f) \tag{6.24}$$

En remplaçant les distributions gaussiennes par leurs expressions analytiques respectives puis en développant les formes quadratiques en argument des exponentielles (le développement est par exemple détaillé dans MAGNANT 2016, annexe D.1), on montre que la densité de probabilité *a posteriori* à l'instant t_k est gaussienne et satisfait :

$$p(\mathbf{x}_k|\mathbf{y}_{1:k}) = \mathcal{N}(\hat{\mathbf{x}}_k^a, \mathbf{P}_k^a) \tag{6.25}$$

avec :

$$\hat{\mathbf{x}}_k^a = \hat{\mathbf{x}}_k^f + \mathbf{K}_k(\mathbf{y}_k - \mathbf{H}_k\hat{\mathbf{x}}_k^f) \tag{6.26}$$

$$\mathbf{P}_k^a = (\mathbf{I}_n - \mathbf{K}_k\mathbf{H}_k)\mathbf{P}_k^f \tag{6.27}$$

$$\mathbf{K}_k = \mathbf{P}_k^f\mathbf{H}_k^T(\mathbf{H}_k\mathbf{P}_k^f\mathbf{H}_k^T + \mathbf{R}_k)^{-1} \tag{6.28}$$

L'état *a posteriori* $\hat{\mathbf{x}}_k^a$ est ainsi estimé à partir de la combinaison linéaire entre la prévision du modèle $\hat{\mathbf{x}}_k^f$ et un terme d'innovation qui représente l'information nouvelle apportée par les observations à l'instant t_k. Dans cette combinaison linéaire, la matrice $\mathbf{K}_k$ est le gain de Kalman qui joue le rôle de poids pesant plus vers le prior ou les observations selon la valeur du ratio entre leurs incertitudes respectives. En général, les observations fournies sont moins incertaines que la prévision du modèle, menant à une forte correction du vecteur d'état vers les points observés. D'autre part, l'état analysé aux points non observés bénéficie aussi des corrections provenant des observations au travers des corrélations avec les valeurs du modèle aux points observés. L'information des points observés peut donc se propager à tout le domaine ainsi qu'aux conditions initiales, conditions limites et paramètres du modèle non observés, éventuellement contenus dans le vecteur d'état.

Quelques remarques

Les équations (6.21)-(6.22) et (6.26)-(6.27)-(6.28) constituent les équations du filtre de Kalman. Dans un cadre gaussien et linéaire, celui-ci résout le problème de filtrage séquentiel de manière exacte. En effet, à chaque instant les densités de probabilité *prédite* et *a posteriori* sont gaussiennes. Le filtre de Kalman permet donc de les caractériser entièrement en décrivant l'évolution de leur espérances et de leur covariances.
Toutefois, il est important de rappeler que formuler l'hypothèse de linéarité du modèle et/ou de l'opérateur d'observation est souvent impossible dans les applications environnementales qui impliquent des modèles complexes. D'autre part, le stockage des matrices de covariance $\mathbf{P}_k^f \in \mathbb{R}^{n\times n}$ $\mathbf{P}_k^a \in \mathbb{R}^{n\times n}$ peut également s'avérer difficile dans ces applications qui impliquent souvent des vecteurs d'état de grande dimension. De même, le calcul du gain de Kalman implique d'effectuer des produits entre matrices de grandes tailles, ce qui peut aussi s'avérer difficile, voire impossible.
Pour toutes ces raisons, le filtre de Kalman reste assez peu utilisé dans le domaine de la modélisation environnementale. On se tourne plutôt vers des solutions approchées du filtre

séquentiel qui permettent notamment de contourner les difficultés liées au coût de calcul et de dépasser (plus ou moins) le cadre restreint d'un modèle et opérateur d'observation linéaires et gaussiens. Les extensions ensemblistes du filtre de Kalman présentées dans la section suivante font partie des approximations du filtre de Kalman largement utilisées.

6.3 Extensions ensemblistes du filtre de Kalman

Le filtre de Kalman d'ensemble (EnKF, Evensen 1994 ; Evensen 2003) est l'extension ensembliste du filtre de Kalman la plus connue. D'abord utilisé pour des systèmes linéaires et gaussiens pour contourner le problème de coût numérique du filtre de Kalman, l'EnKF a ensuite été étendu aux systèmes non linéaires (Evensen, 1997) où ses très bonnes performances ont été montrées dans de nombreuses applications.

L'EnKF consiste à approximer la densité de probabilité de l'état par celle d'un ensemble fini de M membres, représentant chacun un état possible $\mathbf{x}^{(i)}, i \in \{1, ..., M\}$. Les étapes de prévision et d'analyse qui caractérisent l'approche séquentielle du filtrage peuvent ensuite être appliquées à chaque membre de l'ensemble. Ainsi, chaque membre est propagé à l'étape de prévision en appliquant le modèle avant d'être mis à jour à l'étape d'analyse. L'erreur modèle est représentée par la dispersion de l'ensemble obtenue d'une part au moment de la génération de l'ensemble, à partir des incertitudes (sur les conditions limites, conditions initiales et paramètres d'entrée) et d'autre part, à partir du modèle lui-même, lors de sa propagation.
L'EnKF permet ainsi de traiter des problèmes de grande dimension. En plus d'être conceptuellement simple, il est bien adapté à la programmation parallèle. Comme pour le filtre de Kalman, l'EnKF est optimal lorsque les distributions manipulées sont gaussiennes. Malgré cette hypothèse forte, l'EnKF a prouvé son efficacité dans un très large éventail d'applications, tant académiques qu'opérationnelles (en particulier pour la prévision océanique) et il a été largement utilisé au cours des 30 dernières années. De plus, le principe implémenté par l'EnKF peut facilement être adapté pour résoudre des problèmes de lissage.

Dans cette thèse, en plus de l'EnKF, deux autres méthodes d'assimilation ensemblistes sont utilisées. Le cadre de l'étude correspondant à un problème de réanalyse, deux lisseurs qui diffèrent dans la manière dont ils prennent en compte les observations et dans la réalisation de l'analyse sont implémentés : le lisseur d'ensemble avec assimilation multiple (ES-MDA, Emerick et Reynolds 2013) et le lisseur de Kalman itératif (iEnKS, Bocquet et Sakov 2014). Leurs principes et la manière dont les observations sont intégrées sont comparés sur la Figure 6.5. Comme présenté précemment, dans l'approche de filtrage séquentielle de l'EnKF, la densité de probabilité de l'état est corrigée à chaque mise à jour en utilisant seulement les observations du pas de temps courant (voir Figure 6.5, gauche). Dans l'ES-MDA, l'ensemble est d'abord intégré sur toute la fenêtre d'assimilation lors d'une unique étape de prévision (Figure 6.5, milieu). La même analyse que pour l'EnKF est alors réalisée, en utili-

sant cette fois-ci toutes les observations en même temps. Enfin, l'iEnKS se situe entre l'EnKF et l'ES-MDA en termes d'utilisation des observations puisqu'il utilise toutes les observations disponibles entre t_{k+1} et t_{k+L} où L est la longueur de la fenêtre d'assimilation pour corriger l'état au pas de temps t_k (voir Figure 6.5, droite). Les algorithmes correspondants à chaque méthode sont présentés dans les paragraphes suivants.

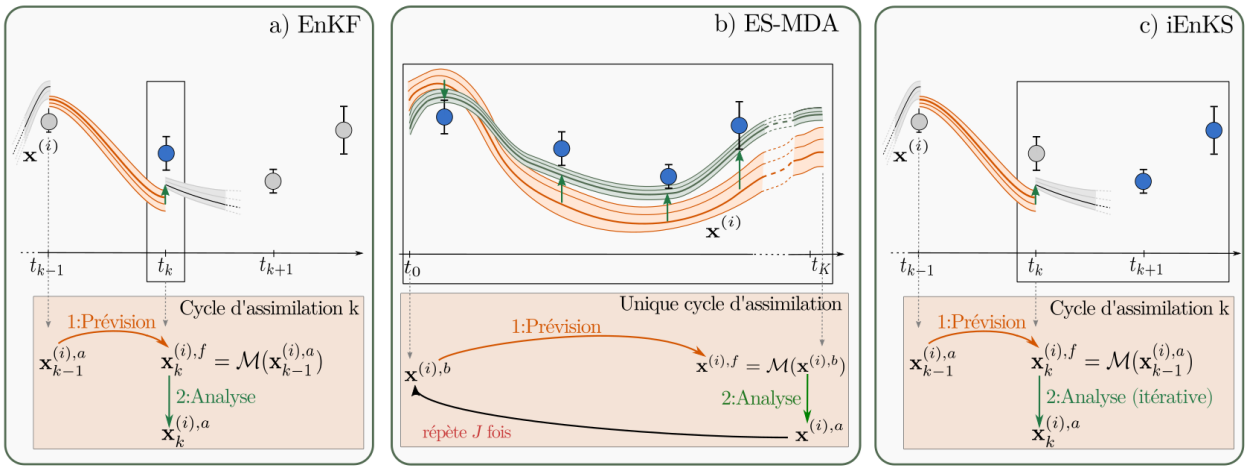


Figure 6.5 – Vue schématique d'un cycle d'assimilation pour l'EnKF (gauche), l'ES-MDA (milieu) et l'iEnKS (droite). $\mathbf{x}^{(i)}$ désigne le i-ième membre de l'ensemble. Les lignes représentent les trajectoires de certains membres de l'ensemble et l'enveloppe colorée représente l'incertitude de l'ensemble. Les points représentent les observations disponibles : les points bleus dans la fenêtre noire sont les observations qui sont utilisées pour le cycle d'assimilation actuel, tandis que les points gris sont les observations non utilisées.

On note que ces méthodes font toutes l'hypothèse que les densités de probabilité manipulées sont gaussiennes pour être optimales, ce que la première partie de ces travaux de thèse a permis de vérifier. Cependant, une telle hypothèse n'est pas justifiée dans tous les cas. On rappelle notamment que sur le scénario hivernal, l'humidité de surface suit dans certains cas une distribution bimodale. Pour ces cas là, une approche type filtre particulaire peut être une alternative intéressante car elle ne se base sur aucune hypothèse de gaussianité. Cette autre approche du filtrage d'ensemble n'est pas implémentée dans ces travaux mais elle est brièvement présentée dans l'Annexe F.

6.3.1 Filtre de Kalman d'ensemble

Le filtre de Kalman d'ensemble est la méthode d'assimilation statistique la plus communément utilisée, notamment dans les domaines de la modélisation hydrologique et des transferts de pesticides à l'échelle du versant et du bassin versant (e.g. Hendricks Franssen et Kinzelbach, 2008 ; Baatz et al., 2017 ; Botto et al., 2018). L'EnKF permet non seulement de résoudre des problèmes de correction de variables mais aussi d'aborder des problèmes d'estimation jointe. C'est notamment le cas de Pasetto et al. (2015) qui examinent le potentiel de l'EnKF pour estimer le champ de conductivité hydraulique à saturation dans

le modèle CATHY appliqué sur un versant virtuel. De la même manière, Xie et Zhang (2010) ont également démontré les capacités de l'EnKF à estimer le Curve Number, un paramètre empirique qui décrit le ruissellement en plus de diverses variables pronostiques dans le modèle conceptuel hydrologique SWAT (Arnold et al., 1998).

Il existe plusieurs versions de l'EnKF avec plusieurs stratégies de mise à jour des membres au moment de l'analyse et de perturbation (ou non) des observations (e.g. Burgers et al., 1998 ; Anderson, 2001 ; Bishop et al., 2001 ; Houtekamer et Mitchell, 2001 ; Whitaker et Hamill, 2002). La version utilisée dans ces travaux est l'*Ensemble Transform Kalman Filter* (ETKF, Bishop et al., 2001 ; Hunt et al., 2007) permettant d'éviter d'introduire un bruit stochastique qui peut impacter les performances du filtre comme dans la version proposée par Burgers et al. (1998).

On considère un ensemble de M membres $\mathbf{x}^{(i)}, i \in \{1, ..., M\}$ et on note $\mathbf{X}_k^f \in \mathbb{R}^{n \times M}$, la matrice des perturbations normalisées au temps t_k dont les colonnes s'écrivent :

$$[\mathbf{X}_k^f]_i = \frac{\mathbf{x}_k^{(i),f} - \overline{\mathbf{x}}_k^f}{\sqrt{M-1}} \tag{6.29}$$

avec

$$\overline{\mathbf{x}}_k^f = \frac{1}{M} \sum_{i=1}^{M} \mathbf{x}_k^{(i),f} \tag{6.30}$$

L'ETKF décrit le vecteur d'état analysé $\mathbf{x}_k^a$ comme appartenant au sous-espace affine défini par les anomalies: $\overline{\mathbf{x}}_k^f + \text{Vec}\{\mathbf{x}_k^{(1),f} - \overline{\mathbf{x}}_k^f, ..., \mathbf{x}_k^{(M),f} - \overline{\mathbf{x}}_k^f\}$. L'analyse s'exprime ainsi :

$$\mathbf{x}_k^a = \overline{\mathbf{x}}_k^f + \mathbf{X}_k^f \mathbf{w}_k \tag{6.31}$$

où $\mathbf{w}_k \in \mathbb{R}^M$ est le vecteur contenant les coordonnées de l'état analysé dans le sous-espace des anomalies. Le vecteur de coordonnées optimales est obtenu à partir de l'équation du filtre de Kalman :

$$\mathbf{x}_k^a = \overline{\mathbf{x}}_k^f + \mathbf{K}_k[\mathbf{y}_k - H_k(\overline{\mathbf{x}}_k^f)] \tag{6.32}$$

où le gain de Kalman $\mathbf{K}_k = \mathbf{P}_k^f \mathbf{H}_k^\top (\mathbf{H}_k \mathbf{P}_k^f \mathbf{H}_k^\top + \mathbf{R}_k)^{-1}$ est calculé à partir de la matrice de covariance d'erreur de prévision exprimée à partir de l'ensemble $\mathbf{P}_k^f = \mathbf{X}_k^f \mathbf{X}_k^{f^\top}$. En identifiant les termes et en utilisant la formule de Sherman-Morrison-Woodbury, on peut ainsi exprimer le vecteur de coordonnées optimal $\mathbf{w}_k$:

$$\mathbf{w}_k = (\mathbf{I}_M + \mathbf{Y}_k^{f^\top} \mathbf{R}_k^{-1} \mathbf{Y}_k^f)^{-1} \mathbf{Y}_k^{f^\top} \mathbf{R}_k^{-1} \delta_k \tag{6.33}$$

où δ_k est le vecteur d'innovation : $\delta_k = \mathbf{y}_k - H_k(\overline{\mathbf{x}}_k^f)$, qui contient les observations et où

$\mathbf{Y}_k^f \in \mathbb{R}^{p\times M}$ contient les anomalies d'observation normalisées :

$$[\mathbf{Y}_k^f]_i = \frac{H_k(\mathbf{x}_k^{(i),f}) - \overline{\mathbf{y}}_k^f}{\sqrt{M-1}} \tag{6.34}$$

avec :

$$\overline{\mathbf{y}}_k^f = \frac{1}{M}\sum_{i=1}^{M} H_k(\mathbf{x}_k^{(i),f}) \tag{6.35}$$

L'ensemble des perturbations *a posteriori* est généré de manière à être représentatif de l'erreur *a posteriori*, laquelle peut être factorisée sous la forme suivante: $\mathbf{P}_k^a = \mathbf{X}_k^a \mathbf{X}_k^{a^\top}$. Les anomalies normalisées de l'analyse s'expriment ainsi :

$$\mathbf{X}_k^a = \mathbf{X}_k^f (\mathbf{I}_M + \mathbf{Y}_k^{f^\top} \mathbf{R}^{-1} \mathbf{Y}_k^f)^{-\frac{1}{2}} \tag{6.36}$$

On obtient ainsi l'expression suivante pour les membres analysés $\mathbf{x}_k^{(i),a}$, $\forall i \in \{1, ..., M\}$:

$$\mathbf{x}_k^{(i),a} = \overline{\mathbf{x}}_k^f + \mathbf{X}_k^f (\mathbf{w}_k + \sqrt{M-1}[(\mathbf{I}_M + \mathbf{Y}_k^{f^\top} \mathbf{R}^{-1} \mathbf{Y}_k^f)^{-\frac{1}{2}}]_i) \tag{6.37}$$

où l'on rappelle que la notation $[.]_i$ désigne la i$^{\text{ème}}$ colonne de la matrice.
Ainsi, la plupart des calculs algébriques ont lieu dans le sous espace de l'ensemble comme le montre l'Algorithme 1. Cet espace est généralement de taille largement inférieure à celle de l'espace de l'état ou des observations. L'ETKF permet ainsi de limiter le coût de calcul de l'analyse (les produits et inversions de matrices comme dans l'étape 7 sont effectués dans un espace de taille réduite) et est donc plébiscité pour les problèmes de grande dimension.

Algorithme 1 Algorithme de l'ETKF : pseudo-code pour le k-ième cycle d'analyse+prévision et taille des objets manipulés

Input: $\mathbf{E}_k \in \mathbb{R}^{n\times M}$ est l'ensemble au temps t_k et $\mathbf{y}_k \in \mathbb{R}^p$ est le vecteurs des observations disponibles à t_k. M_{k+1} est le modèle entre le temps t_k et t_{k+1}, H et $\mathbf{R}_k$ sont respectivement l'opérateur d'observation et la matrice de covariance des erreurs d'observation au temps t_k.

1: $\bar{\mathbf{x}}_k = \mathbf{E}_k \mathbf{1}_M / M$ ▷ $\in \mathbb{R}^n$
2: $\mathbf{X}_k = (\mathbf{E}_k - \bar{\mathbf{x}}_k \mathbf{1}_M^\intercal)/\sqrt{M-1}$ ▷ $\in \mathbb{R}^{n\times M}$
3: $\mathbf{Z}_k = H_k(\mathbf{E}_k)$ ▷ $\in \mathbb{R}^{p\times M}$
4: $\bar{\mathbf{y}}_k = \mathbf{Z}_k \mathbf{1}_M / M$ ▷ $\in \mathbb{R}^p$
5: $\mathbf{S} = \mathbf{R}_k^{-\frac{1}{2}}(\mathbf{Z}_k - \bar{\mathbf{y}}_k \mathbf{1}_M^\intercal)/\sqrt{M-1}$ ▷ $\in \mathbb{R}^{p\times M}$
6: $\mathbf{d} = \mathbf{R}_k^{-\frac{1}{2}}(\mathbf{y}_k - \bar{\mathbf{y}}_k)$ ▷ $\in \mathbb{R}^p$
7: $\mathbf{T} = (I_M + \mathbf{S}^\intercal \mathbf{S})^{-1}$ ▷ $\in \mathbb{R}^{M\times M}$
8: $\mathbf{w} = \mathbf{T}\mathbf{S}^\intercal \delta$ ▷ $\in \mathbb{R}^M$
9: $\mathbf{E}_k = \bar{\mathbf{x}}_k \mathbf{1}_M^\intercal + \mathbf{X}_k(\mathbf{w}\mathbf{1}_M^\intercal + \sqrt{M-1}\mathbf{T}^{-\frac{1}{2}})$ ▷ $\in \mathbb{R}^{n\times M}$
10: $\mathbf{E}_{k+1} = M_{k+1}(\mathbf{E}_k)$ ▷ $\in \mathbb{R}^{n\times M}$

6.3.2 Lisseur d'ensemble et lisseur d'ensemble avec assimilation multiple

Le lisseur d'ensemble (*Ensemble Smoother* ou ES, van Leeuwen et Evensen 1996) est un lisseur à intervalle fixe qui vise à estimer la distribution *a posteriori* du vecteur d'état dans une fenêtre d'assimilation $[t_1, ..., t_K]$ en se basant sur toutes les observations disponibles dans cette fenêtre. L'ES a souvent été appliqué dans le domaine de la modélisation hydrologique (Dunne et Entekhabi, 2005 ; Bailey et Baù, 2010 ; Todaro et al., 2021), notamment pour résoudre des problèmes d'estimation jointe. Dans Crestani et al. (2013), les performances de l'EnKF et de l'ES sont par exemple comparées pour déduire la distribution spatiale de conductivité hydraulique à partir d'un modèle d'écoulements et de transport souterrain. De la même manière, Bailey et Baù (2012) extraient une distribution de conductivité à partir de l'ES, utilisé à la fois dans sa version initiale et dans une variante itérative.
Dans le cas d'estimation jointe, le vecteur d'état $\mathbf{x}$ contient les trajectoires temporelles pour chaque variable d'état, à chaque point de la grille du modèle ainsi que les paramètres d'entrée. On peut ainsi aboutir à un vecteur d'état de très grande taille dont la dimension peut rapidement poser des difficultés sur le plan numérique. Le vecteur d'observation contient quant à lui toutes les observations disponibles sur la fenêtre temporelle et l'opérateur d'observation est construit de manière à faire correspondre chaque observation disponible au bon pas de temps de la trajectoire temporelle simulée. La même analyse que pour l'EnKF est alors réalisée, en utilisant cette fois-ci, toutes les observations en même temps. Comme pour le filtre de Kalman, les valeurs de $\mathbf{x}$ non observées dans le temps ou dans l'espace sont également corrigées, au travers des covariances spatiotemporelles estimées à partir de l'ensemble.

Plus récemment, Emerick et Reynolds (2013) ont proposé le lisseur d'ensemble avec assimilation multiple (*Ensemble Smoother with Multiple Data Assimilation* ou ES-MDA), une version itérative du lisseur d'ensemble, particulièrement adaptée à l'estimation jointe. L'ES-MDA est notamment appliqué dans Cui et al. (2020) pour estimer des paramètres hydrodynamiques à partir du modèle Hydrus-1D (Šimůnek et al., 1998). L'ES-MDA consiste à réaliser la séquence *intégration du modèle sur toute la fenêtre temporelle* puis *analyse* de manière itérative (voire Figure 6.5, milieu). A noter que le nombre d'itérations est fixé avant de lancer la procédure d'assimilation et ne dépend pas d'un critère de convergence. L'algorithme de l'analyse n'est pas détaillé ici puisque celle-ci est réalisée selon la même procédure que l'ETKF (voir Algorithme 1, lignes 1 à 9) avec un vecteur d'état qui inclut la dimension temporelle. Dans cette procédure, seule la matrice d'erreur d'observation $\mathbf{R}$ est remplacée à chaque itération (j) pour prendre en compte l'impact de l'assimilation multiple. L'étape d'analyse de l'Eq. (6.36) est ainsi remplacée par :

$$\mathbf{X}^a_{(j)} = \mathbf{X}^f_{(j)}(\mathbf{I}_M + \mathbf{Y}^{f^\top}\alpha_{(j)}\mathbf{R}^{-1}\mathbf{Y}^f)^{-\frac{1}{2}} \tag{6.38}$$

où la notation $\mathbf{X}^a_k$ (resp. $\mathbf{X}^f_k$ et $\mathbf{Y}^f_k$) au temps t_k a été simplifiée en $\mathbf{X}^a$ (resp. $\mathbf{X}^f$ et

$\mathbf{Y}^f$) et où le poids $\alpha_{(j)}$ est un terme d'inflation appliqué à la matrice de covariance d'erreur d'observation $\mathbf{R}$ utilisé pour compenser le fait que chaque observation est assimilée plusieurs fois. En notant J, le nombre d'itérations prédéfini dans le schéma de l'ES-MDA, les poids $\alpha_{(j)}, j \in \{1, ..., J\}$ doivent vérifier :

$$\sum_{j=1}^{J} \alpha_{(j)} = 1. \tag{6.39}$$

Dans ces travaux, on fixe $\alpha_{(j)} = \frac{1}{J}, \forall\, j \in \{1, ..., J\}$ (Emerick et Reynolds, 2013 ; Cui et al., 2020). Dans les applications de l'ES-MDA rencontrées dans la littérature, le nombre d'itérations J varie en général entre 2 et 10. Dans ces travaux, cette valeur sera déterminée après une première série de tests préliminaires, de manière à garantir la convergence de l'analyse tout en limitant le coût de calcul.

L'intégration du modèle à l'itération $(j + 1)$ est ensuite initialisée en utilisant la distribution postérieure des paramètres du modèle résultant de l'analyse à l'itération (j). Un tel schéma remplace l'unique mise à jour parfois abrupte de l'ES par plusieurs mises à jour qui intègrent progressivement le bénéfice d'une meilleure estimation des paramètres. On s'attend notamment à ce que l'ES-MDA estime mieux la densité de probabilité *a posteriori* dans le cas de modèles fortement non linéaires (Emerick et Reynolds, 2013).

6.3.3 Lisseur de Kalman d'ensemble itératif

Comme pour l'EnKF, le lisseur de Kalman d'ensemble itératif (*iterative Ensemble Kalman Smoother*, iEnKS, Bocquet et Sakov 2014) alterne des étapes de prévision et d'analyse pour réaliser des mises à jour incrémentales de l'état. Il s'agit d'un lisseur à fenêtre glissante et L observations sont utilisées à chaque analyse. Le modèle est ensuite propagé sur l'intervalle $S\Delta t$ (où Δt correspond à l'intervalle de temps entre 2 observations successives) de telle sorte que les fenêtres d'assimilation peuvent éventuellement se superposer.
Contrairement à l'EnKF ou l'ES-MDA qui sont des méthodes purement statistiques, l'iEnKS est une méthode variationnelle d'ensemble. L'analyse consiste en effet à minimiser une fonction coût $\mathcal{I}$ déduite de la formulation bayésienne du problème d'estimation. Les paragraphes suivants détaillent la dérivation d'une telle fonction coût et sa minimisation dans le sous-espace de l'ensemble. Le fonctionnement global de l'iEnKS est illustré sur la Figure 6.6.

Expression de la fonction coût dans l'espace de l'ensemble

En faisant l'hypothèse que l'état et la vraisemblance suivent des distributions gaussiennes, l'Eq. (6.11) de l'étape d'analyse d'un problème d'estimation bayésienne s'exprime sous la forme :

$$p(\mathbf{x}_k|\mathbf{y}_k) \propto \exp(-\mathcal{I}(\mathbf{x}_k^a)) \tag{6.40}$$

où $\mathcal{I}$ est la fonction coût pour l'estimation de l'état $\mathbf{x}_k^a$:

$$\mathcal{I}(\mathbf{x}_k^a) = \frac{1}{2}\|\mathbf{x}_k^a - \mathbf{x}_k^f\|^2_{\mathbf{P}_k^f} + \sum_{l=1}^{L} \frac{1}{2}\|\mathbf{y}_{k+l} - H_{k+l} \circ M_{k\to k+l}(\mathbf{x}_k^a)\|^2_{\mathbf{R}_{k+l}}, \tag{6.41}$$

avec $\|\mathbf{x}\|^2_{\mathbf{A}} = \mathbf{x}^\top \mathbf{A}^{-1}\mathbf{x}$ avec $\mathbf{x} \in \mathbb{R}^n$ et $\mathbf{A} \in \mathbb{R}^{n\times n}$.

Comme l'EnKF ou l'ES-MDA, l'iEnKS est une méthode d'ensemble. On note donc :

- l'ensemble de M membres: $\mathbf{x}^{(i)}$, $i \in \{1, ..., M\}$;
- $\mathbf{E} = [\mathbf{x}^{(1)}, ..., \mathbf{x}^{(M)}] \in \mathbb{R}^{n\times M}$ la matrice qui regroupe les membres ;
- $\mathbf{X}$ la matrice des perturbations comme formulée dans l'Eq. (6.29) ;
- la matrice de covariance d'erreur de prévision exprimée à partir de l'ensemble : $\mathbf{P}_k^f = \mathbf{X}_k^f \mathbf{X}_k^{f^\top}$.

Les calculs sont alors effectués dans le sous-espace de l'ensemble puisque sa taille est en général largement inférieure à celle de l'état. Au temps t_k, on cherche à exprimer le vecteur d'état analysé comme une combinaison linéaire des membres de l'ensemble $\mathbf{x}_k^a = \overline{\mathbf{x}}_k^f + \mathbf{X}_k^f \mathbf{w}$. En réinjectant cette expression dans l'Eq. (6.41), on aboutit à l'expression suivante de la fonction coût à minimiser exprimée dans l'espace de l'ensemble :

$$\widetilde{\mathcal{I}}(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2 + \sum_{l=1}^{L} \frac{1}{2}\|\mathbf{y}_{k+l} - H_{k+l} \circ M_{k\to k+l}(\overline{\mathbf{x}}_k^f + \mathbf{X}_k^f \mathbf{w})\|^2_{\mathbf{R}_{k+l}} \tag{6.42}$$

où $\|\mathbf{w}\|^2 = \mathbf{w}^{\mathrm{T}}\mathbf{w}$.

Une fois la minimisation terminée et l'expression du vecteur d'état analysé déterminé (au temps t_0 sur le cycle représenté sur la Figure 6.6), l'ensemble est propagé jusqu'à t_{k+S} dans une étape de prévision. Dans ces travaux, on fixe $S = 1$ et l'ensemble est donc propagé jusqu'à la prochaine observation disponible. Comme le montre la Figure 6.6, la k-ième analyse utilise donc les observations aux temps t_{k+1} à t_{k+L} alors que la $k+1$-ème analyse utilise les observations t_{k+2} à t_{k+1+L}. Les fenêtres d'assimilation se superposant, chaque observation est utilisée plusieurs fois ce qui est incohérent statistiquement. Pour faire face à cette incohérence, des coefficients $\alpha_l \in [0, 1]$ qui pondèrent le poids des termes d'innovation dans la fonction coût (6.42) sont introduits. Cette dernière s'exprime alors :

$$\widetilde{\mathcal{I}}(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2 + \sum_{l=1}^{L} \frac{\alpha_l}{2}\|\mathbf{y}_{k+l} - H_{k+l} \circ M_{k\to k+l}(\overline{\mathbf{x}}_k^f + \mathbf{X}_k^f \mathbf{w})\|^2_{\mathbf{R}_{k+l}} \tag{6.43}$$

Pour chaque cycle d'assimilation, leur somme vérifie :

$$\sum_{l=1}^{L} \alpha_l = 1 \tag{6.44}$$

Une justification heuristique pour un tel schéma est donnée dans Bocquet et Sakov, 2014 et le lecteur intéressé est invité à se référer à ce papier pour plus de détails sur ce schéma. Dans ces travaux, on choisit un schéma uniforme tel que $\alpha_l = \frac{1}{L}$, $\forall\, l = 1, .., L$.

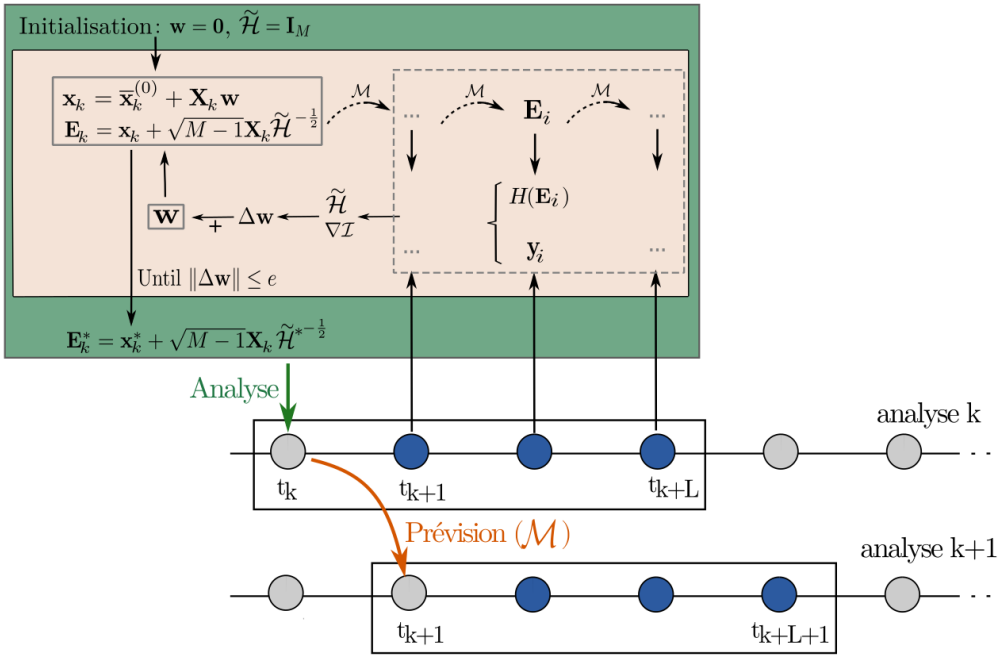


Figure 6.6 – Détails d'un cycle d'assimilation effectué par l'iEnKS ($\Delta\mathbf{w}$ désigne l'incrément $\mathbf{w}_{(j+1)} - \mathbf{w}_{(j)}$). Dans cet exemple, L=3. Les points bleus schématisent les observations utilisées pendant le cycle d'assimilation courant alors que les cercles gris illustrent les observations disponibles mais non utilisées.

Minimisation de la fonction coût dans l'espace de l'ensemble

Dans ces travaux, la minimisation de la fonction coût $\widetilde{\mathcal{I}}$ est réalisée dans l'espace de l'ensemble à l'aide d'un algorithme de Gauss-Newton, ce qui est en théorie équivalent à d'autres schémas de minimisation type Levenberg-Marquardt ou quasi-Newton (Bocquet et Sakov, 2012 ; Asch et al., 2016).
L'analyse est réalisée en utilisant l'approximation linéaire tangente de l'opérateur qui permet de passer de l'espace de l'ensemble à l'espace des observations $H \circ M$: $\mathbf{Y}_{k+l,(j)} = [H_{k+l} \circ M_{k\to k+l}]'_{|\mathbf{x}_{k,(j)}} \mathbf{X}_k^f$. Dans ces travaux, et comme proposé dans Bocquet et Sakov (2013), cet opérateur linéaire tangent est estimé par différences finies (*bundle* iEnKS). L'ensemble est contracté autour de la moyenne par un facteur ε, propagé grâce au modèle puis remis à l'échelle par le facteur inverse ε^{-1} :

$$\mathbf{Y}_{k+l,(j)} \approx \frac{1}{\varepsilon} H_{k+l} \circ M_{k\to k+l}(\overline{\mathbf{x}}_{k,(j)} + \varepsilon \mathbf{X}_{k,(j)})(\mathbf{I}_M - \frac{\mathbf{1}_M \mathbf{1}_M^\intercal}{M}) \tag{6.45}$$

L'Eq. (6.45) montre donc que la matrice $\mathbf{Y}_{k+l,(j)}$ peut être directement estimée à par-

tir de l'ensemble, sans avoir besoin d'estimer $[H_{k+l} \circ M_{k\to k+l}]'$ ce qui constitue un avantage conséquent par rapport à une méthode variationnelle type 4D-Var. D'autre part, la matrice $\mathbf{Y}_{k+l,(j)}$ est de taille $p \times M$. Puisque p et M sont en général de tailles limitées, le calcul de $\mathbf{Y}_{k+l,(j)}$ et $\mathbf{Y}^{\top}_{k+l,(j)}$ n'est pas trop coûteux.

La matrice Hessienne approximée de la fonction coût $\widetilde{\mathcal{H}}$ et son gradient $\nabla\mathcal{I}$ peuvent ainsi être calculés à partir de l'ensemble :

$$\nabla\mathcal{I}_{(j)} = \mathbf{w}_{(j)} - \sum_{l=1}^{L} \alpha_l \mathbf{Y}^{\top}_{k+l,(j)} \mathbf{R}^{-1}_{k+l}[\mathbf{y}_{k+l} - H_{k+l} \circ M_{k\to k+l}(\mathbf{x}^a_{k,(j)})] \tag{6.46}$$

$$\widetilde{\mathcal{H}}_{(j)} = \mathbf{I}_M + \sum_{l=1}^{L} \alpha_l \mathbf{Y}^{\top}_{k+l,(j)} \mathbf{R}^{-1}_{k+l} \mathbf{Y}_{k+l,(j)} \tag{6.47}$$

L'algorithme de Gauss-Newton s'écrit alors :

$$\mathbf{w}_{(j+1)} = \mathbf{w}_{(j)} - \widetilde{\mathcal{H}}^{-1}_{(j)} \nabla\mathcal{I}_{(j)}(\mathbf{w}_{(j)}), \tag{6.48}$$

où j est l'indice d'itération, $\widetilde{\mathcal{H}}$ est l'approximation de la matrice Hessienne de $\mathcal{I}$ et où ∇ désigne l'opérateur gradient. La minimisation a alors lieu jusqu'à ce qu'un critère de convergence ou un nombre maximal d'itérations soit atteint.

Comme pour l'EnKF, l'ensemble *a posteriori* est généré de manière à être représentatif de l'incertitude *a posteriori* (voir Eq. 6.36 pour l'EnKF). La Hessienne obtenue à la fin de la minimisation $\widetilde{\mathcal{H}}^*$ est utilisée pour approximer l'inverse de la matrice de covariance d'erreur et les membres analysés $\mathbf{x}^{(i),a}_k$, $\forall i \in \{1, ..., M\}$ s'expriment :

$$\mathbf{x}^{(i),a}_k = \mathbf{x}^a_k + \sqrt{M-1}[\mathbf{X}^f_k \widetilde{\mathcal{H}^*}^{-\frac{1}{2}}]_i \tag{6.49}$$

avec

$$\mathbf{x}^a_k = \overline{\mathbf{x}}^f_k + \mathbf{X}^f_k \mathbf{w}^* \tag{6.50}$$

Ainsi, grâce à son approche ensembliste, l'iEnKS permet de résoudre un problème de taille raisonnable à chaque analyse (voir taille des objets manipulés dans l'Algorithme 2, en particulier pour les étapes 12 à 14) et de contourner l'utilisation du modèle adjoint, dont la détermination est souvent une étape délicate.

6.3.4 Comparaison des coûts de calcul associés à chaque méthode

Les différentes méthodes d'assimilation mises en oeuvre dans ces travaux reposent toutes sur une approche ensembliste. Cependant, le problème résolu (filtrage pour l'EnKF, lissage à intervalle fixe pour l'ES-MDA et lissage à point fixe pour l'iEnKS) et la stratégie utilisée pour

Algorithme 2 Algorithme de l'iEnKS-MDA version bundle avec minimisation par algorithme de Gauss-Newton : pseudo-code associé au k-ième cycle d'analyse+prévision et taille des objets manipulés. Pour simplifier les notations, le modèle et l'opérateur d'observation sont supposés être les mêmes à tous les pas de temps : $M_k = M$ et $H_k = H$.

Input: $\mathbf{E}_k \in \mathbb{R}^{n\times M}$ est l'ensemble au temps t_k et $\mathbf{y}_k \in \mathbb{R}^p$ est le vecteurs des observations disponibles à t_k. M_{k+1} est le modèle entre le temps t_k et t_{k+1}, H et $\mathbf{R}_k$ sont respectivement l'opérateur d'observation et la matrice de covariance des erreurs d'observation au temps t_k et les α_l, $0 \leq l \leq L$ sont les poids des observations dans la fenêtre d'assimilation. Paramètres de l'algorithme : $jmax$, ε, e.

$j = 0$, $\mathbf{w} = 0$
$\overline{\mathbf{x}}_k^{(0)} = \mathbf{E}_k \mathbf{1}_M / M$ ▷ $\in \mathbb{R}^n$
$\mathbf{X}_k = (\mathbf{E}_k - \bar{\mathbf{x}}_k^{(0)} \mathbf{1}_M^\intercal)/\sqrt{M-1}$ ▷ $\in \mathbb{R}^{n\times M}$
while $||\Delta \mathbf{w}|| \geq e$ **or** $j \leq jmax$ **do**
    $\overline{\mathbf{x}}_k = \overline{\mathbf{x}}_k^{(0)} + \mathbf{X}_k \mathbf{w}$ ▷ $\in \mathbb{R}^n$
    $\mathbf{E}_k = \overline{\mathbf{x}}_k \mathbf{1}_M^\intercal + \varepsilon \mathbf{X}_k$ ▷ $\in \mathbb{R}^{n\times M}$
    for l = 1,...,L **do**
        $\mathbf{E}_{k+l} = M(\mathbf{E}_{k+l-1})$ ▷ $\in \mathbb{R}^{n\times M}$
        $\overline{\mathbf{y}}_{k+l} = H(\mathbf{E}_{k+l}) \mathbf{1}_M / M$ ▷ $\in \mathbb{R}^p$
        $\mathbf{Y}_{k+l} = (H(\mathbf{E}_{k+l}) - \overline{\mathbf{y}}_{k+l} \mathbf{1}_M^\intercal)/\varepsilon$ ▷ $\in \mathbb{R}^{p\times M}$
    end for
    $\nabla \mathcal{I} = \mathbf{w} - \sum_{l=1}^{L} \alpha_l \mathbf{Y}_{k+l}^T \mathbf{R}_{k+l}^{-1} [\mathbf{y}_{k+l} - \overline{\mathbf{y}}_{k+l}]$ ▷ $\in \mathbb{R}^M$
    $\widetilde{\mathcal{H}} = \mathbf{I}_M + \sum_{l=1}^{L} \alpha_l \mathbf{Y}_{k+l}^T \mathbf{R}_{k+l}^{-1} \mathbf{Y}_{k+l}$ ▷ $\in \mathbb{R}^{M\times M}$
    Solve $\widetilde{\mathcal{H}} \Delta \mathbf{w} = \nabla \mathcal{I}$ ▷ $\in \mathbb{R}^M$
    $\mathbf{w} \leftarrow \mathbf{w} - \Delta \mathbf{w}$ ▷ $\in \mathbb{R}^M$
    $j \leftarrow j + 1$
end while
$\mathbf{E}_k = \bar{\mathbf{x}}_k \mathbf{1}_M^\intercal + \mathbf{X}_k \widetilde{\mathcal{H}}^{-\frac{1}{2}}$ ▷ $\in \mathbb{R}^{n\times M}$
$\mathbf{E}_{k+1} = M(\mathbf{E}_k)$ ▷ $\in \mathbb{R}^{n\times M}$

réaliser l'analyse (méthode itérative ou pas) diffèrent de l'une à l'autre impliquant des coûts de calcul contrastés. Sans aller jusqu'à un calcul formel de complexité, on peut comparer leur coût numérique à partir du nombre et de la durée des simulations et de l'emprunte mémoire nécessaires pour chacune.

En termes d'intégrations du modèle, l'EnKF est la méthode la plus économe (Tableau 6.1, colonnes 1 et 2). Le modèle est intégré entre chaque observation disponible lors des étapes de prévisions, pour chaque membre et sans itération. Pour l'ES-MDA, le modèle est intégré en une fois sur toute la durée de la simulation, pour chaque membre et ce autant de fois que le nombre d'itérations J l'impose. Le temps de simulation total par membre est plus long mais il y a moins d'interruptions et d'opérations de lecture/écriture ce qui peut constituer un gain de temps non négligeable. L'iEnKS avec son approche itérative est la méthode la plus coûteuse en termes d'intégrations du modèle.

Pour les 3 méthodes, l'analyse est réalisée dans le sous-espace de l'ensemble impliquant majoritairement le stockage et la manipulation d'objets de taille M et $M \times M$. Par contre, la taille n du vecteur d'état et celle p du vecteur d'observations peuvent être beaucoup plus importantes dans l'ES-MDA que dans l'EnKF et l'iEnKS puisque la dimension temporelle est intégrée (Tableau 6.1, colonne 3). L'emprunte mémoire et le coût de calcul des étapes d'analyse (stockage et mise à jour de l'ensemble $\mathbf{E}$, projection dans l'espace des observations $H(\mathbf{E})$) sont ainsi susceptibles d'être fortement impactés.

	Nombre total de simulations	intervalle d'intégration d'UNE simulation	Taille du vecteur d'état
EnKF	MC *50 × 13*	Δt *144h*	variables d'état : à un instant donné sur tout le **domaine spatial** + paramètres d'entrée *364*
ES-MDA	MJ *50 × 3*	$C\Delta t$ *13 × 144h*	variables d'état : sur tout le **domaine spatial et temporel** + paramètres d'entrée *899864*
iEnKS	$M(C + Ljmax)$ *50 × (13 + 5 ×3)*	Δt *144h*	variables d'état : à un instant donné sur tout le **domaine spatial** + paramètres d'entrée *364*

Tableau 6.1 – Comparaison du nombre de simulations requises et de la taille du vecteur d'état pour l'EnKF, l'ES-MDA et l'iEnKS. M désigne le nombre de membres de l'ensemble, C le nombre de cycles d'assimilation (prévision+analyse) dans la simulation, Δt l'intervalle de temps entre 2 observations successives (supposé constant pour simplifier), J le nombre d'itérations de l'ES-MDA, L la taille de la fenêtre d'assimilation de l'iEnKS et $jmax$ le nombre d'itérations maximal autorisé pour l'iEnKS. Les valeurs en gris correspondent aux valeurs nominales des premiers tests réalisés dans ces travaux et présentés dans la Section 6.4.2.

6.4 Méthodologie et définition du problème abordé dans la thèse

6.4.1 Les outils utilisés

Expériences jumelles

Dans ce travail, l'utilisation de méthodes d'assimilation de données pour le modèle PESHMELBA est testée pour la première fois. Pour cela, on utilise des expériences jumelles permettant d'explorer de manière intensive les performances du système mis en place.

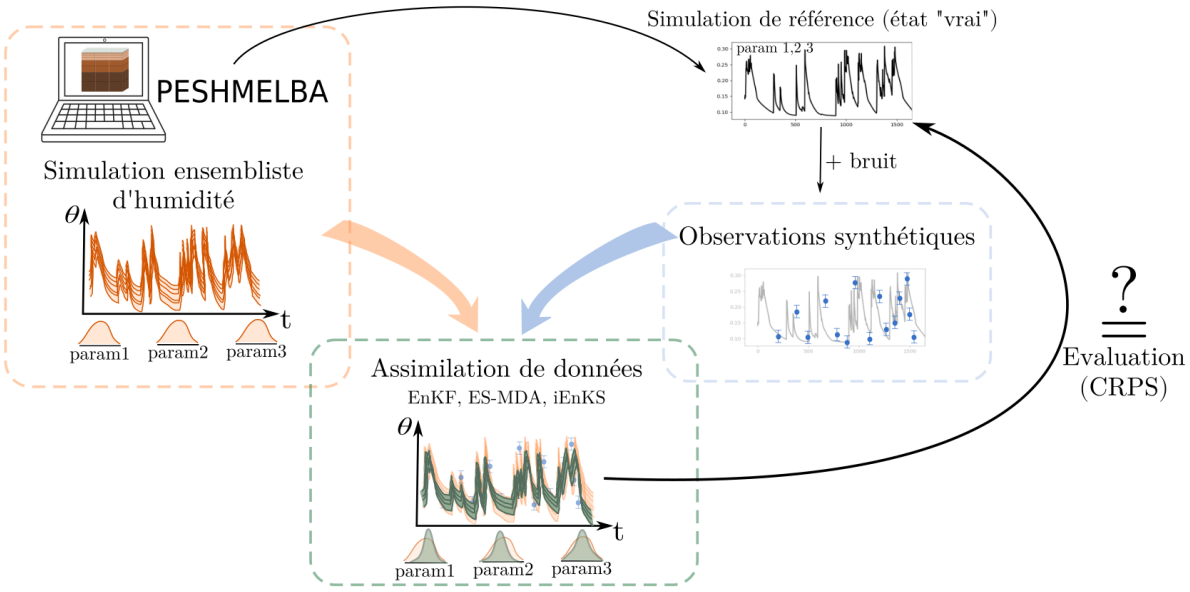


Figure 6.7 – Principe des expériences jumelles appliqué à l'estimation de l'humidité dans le modèle PESHMELBA. La métrique utilisée pour comparer l'état analysé à l'état de référence est le CRPS présenté dans le paragraphe suivant.

Le principe des expériences jumelles est illustré sur la Figure 6.7. Elles consistent à fabriquer une expérience d'assimilation où l'état vrai du système $\mathbf{x}^t$ que l'on cherche à estimer est parfaitement connu. Dans ces travaux, cet état vrai est obtenu à partir d'une simulation PESHMELBA paramétrisée à partir de valeurs de référence des paramètres d'entrée et des conditions initiales. On fait l'hypothèse que ces valeurs sont les valeurs "vraies" des paramètres, et une simulation PESHMELBA avec ces paramètres aboutit à des trajectoires spatialisées d'humidité et une série temporelle de concentration à l'exutoire, elles aussi considérées comme la référence vraie. Des observations synthétiques sont ensuite créées à partir de ces trajectoires de référence en leur ajoutant un bruit qui respecte les caractéristiques de l'erreur d'observation définie pour chacun des types de données disponibles. L'objectif de l'expérience d'assimilation est alors de retrouver les valeurs vraies de certains des paramètres et les trajectoires vraies d'humidité et de concentration, à partir des observations synthé-

tiques.

La solution exacte étant connue, les performances du système d'assimilation peuvent ainsi être explorées dans un grand nombre de configurations et évaluées de façon fixe. Les expériences jumelles permettent par exemple d'explorer l'impact du nombre d'observations disponibles, de leur fréquence temporelle, résolution et répartition spatiale.

Localisation

Dans les méthodes ensemblistes, la taille de l'ensemble utilisé pour représenter l'état est souvent de taille très inférieure à celle du vecteur d'état. Il peut en résulter des erreurs d'échantillonnage importantes qui affectent notamment la matrices de covariance d'erreur de prévision $\mathbf{P}$ et entraînent l'apparition de corrélations douteuses dans cette dernière. Ces corrélations connectent des domaines physiques qui sont en réalité entièrement décorrélés et peuvent ainsi dégrader la qualité de l'analyse. Pour pallier ces difficultés, il est possible de localiser l'analyse et de garantir ainsi que des domaines physiques distants sont entièrement indépendants. On parle ainsi de *localisation* (Houtekamer et Mitchell, 2001 ; Hamill et al., 2001 ; Evensen, 2003 ; Ott et al., 2004). Celle-ci peut être mise en oeuvre en effectuant une *localisation par domaines* ou une *localisation des covariances* (Asch et al., 2016).

La *localisation par domaines* consiste à découper le vecteur d'état en portions indépendantes (en général des domaines spatiaux indépendants) et à réaliser une analyse par portion en utilisant seulement les observations disponibles sur, ou à proximité immédiate, de cette portion (la notion de proximité immédiate étant à définir). La taille du problème défini par chaque analyse est ainsi considérablement réduite (moins d'observations, moins de variables d'état) mais il faut réaliser plusieurs analyses. Si celles-ci peuvent être réalisées en parallèle, le gain en temps de calcul peut être significatif (Hunt et al., 2007).
Dans le cas de la *localisation des covariances*, une seule analyse est réalisée mais la matrice $\mathbf{P}$ calculée à partir de l'ensemble est d'abord filtrée afin d'atténuer ou de mettre à zéro certaines valeurs de corrélation. On s'appuie en général pour cela sur une matrice de localisation qui est appliquée à $\mathbf{P}$ par un produit de Schur avant de réaliser l'analyse. Pour définir la matrice de localisation, on utilise classiquement la distance euclidienne entre la variable d'état et les observations (Anderson, 2012) ou des fonctions décroissantes avec la distance (Gaspari et Cohn, 1999).

En pratique, la localisation est souvent indispensable même si elle n'est pas toujours explicitée dans les applications. Dans le domaine de la modélisation environnementale et hydrologique spatialisée, la localisation des covariances reste plus utilisée (El Gharamti et al., 2021 ; Devers et al., 2020) que la localisation par domaines (Waller et al., 2017).

Diagnostic des performances du système d'assimilation

Continuous Ranked Probability Score Dans ces travaux, les capacités du système d'assimilation à produire un ensemble qui se rapproche au mieux de l'état "vrai" de référence sont évaluées grâce au *Continuous Ranked Probability Score* (CRPS, Brown, 1974). Le CRPS est une métrique, largement utilisée dans l'évaluation des performances d'un ensemble et qui permet notamment de comparer un ensemble à une valeur déterministe (ici la valeur vraie de référence). Le CRPS exprime la distance entre la fonction cumulative de distribution (cdf) de la prévision ensembliste et la cdf d'une référence. Si on considère la référence comme déterministe, sa cdf s'exprime à partir de la fonction de Heaviside. Le CRPS représente finalement la distance quadratique entre les deux fonctions (Figure 6.8).

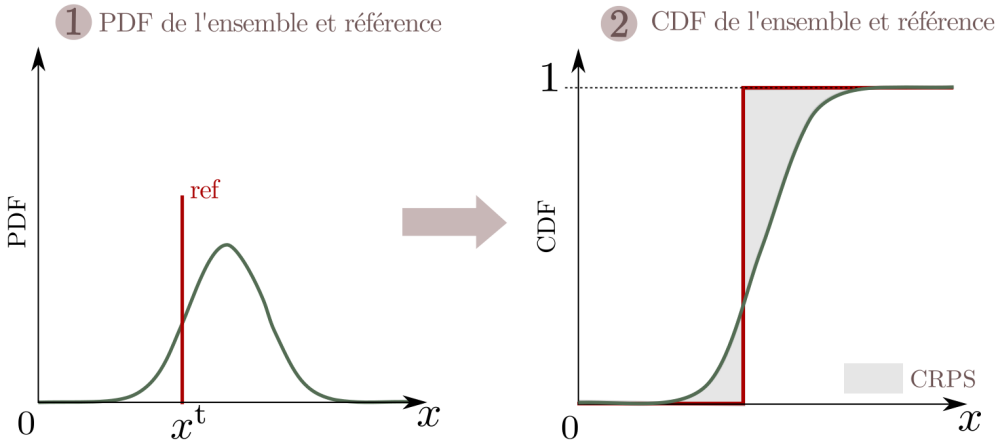


Figure 6.8 – Principe schématique du CRPS.

Le CRPS s'exprime ainsi :

$$CRPS(x, x^t) = \int_{-\infty}^{\infty} [F_x(s) - H(s - x^t)]^2 ds \tag{6.51}$$

où F_x est la cdf de l'ensemble 1D associé à la variable x, x^t est la valeur de référence déterministe et $H(s)$ est la fonction de Heaviside. En pratique, la cdf de l'ensemble est estimée à partir des membres et assimilée à une fonction en escaliers. La formulation qui en découle est détaillée dans l'Annexe G (Hersbach, 2000). Le CRPS est toujours positif et plus il est proche de 0, plus l'ensemble s'approche de la référence déterministe. Il est exprimé dans la même unité que la variable évaluée ce qui en facilite l'interprétation. Dans le cas d'une variable $\mathbf{Y}$ multidimensionnelle (spatiotemporelle par exemple), le CRPS est moyenné dans le temps et/ou l'espace.

Comme le soulignent Bocquet et Sakov (2014), il est crucial de rappeler que les performances d'un schéma d'assimilation dépendent fortement de la métrique choisie pour évaluer sa qualité. En hydrologie, la *Root Mean Square Error* (RMSE) est souvent choisie pour

évaluer les performances d'un système d'assimilation (e.g. Camporese et al., 2009 ; Camporese et al., 2010 ; Botto et al., 2018) mais celle-ci ne fournit qu'un résumé restrictif de la distribution puisqu'elle n'utilise que la moyenne de l'ensemble. Ici, on utilise plutôt le CRPS car il permet de généraliser rigoureusement la notion d'erreur absolue moyenne aux prédictions stochastiques. D'autre part, il peut être décomposé pour fournir une information plus détaillée sur la fiabilité et la résolution de l'ensemble *a posteriori* (Hersbach, 2000).

Décomposition du CRPS Le CRPS est un score synthétique qui peut être décomposé pour évaluer séparément la composante de fiabilité (*indice de fiabilité*) et d'incertitude (*CRPS potentiel*) de l'ensemble (Hersbach, 2000) :

$$CRPS(x, x^t) = Indice\,de\,fiabilité + CRPS\,potentiel \tag{6.52}$$

L'indice de fiabilité peut être interprété en termes d'histogramme de rang (Hamill, 2001) et représente qualitativement la capacité de l'ensemble à contenir la référence. L'indice de fiabilité est donc fortement sensible aux biais. Plus l'ensemble est fiable, plus l'indice de fiabilité est proche de 0.
Le CRPS potentiel correspond quant à lui à la valeur de CRPS qui serait calculée si l'ensemble était entièrement fiable (*Indice de fiabilité*=0). Il est sensible à la dispersion de l'ensemble et plus l'incertitude de l'ensemble est importante plus sa valeur est élevée.
Comme pour le CRPS, l'indice de fiabilité et le CRPS potentiel sont exprimés dans la même unité que la variable évaluée et sont toujours positifs.

Continuous Ranked Probability Skill Score Dans ces travaux, on utilise également le *Continuous Ranked Probability Skill Score* (CRPSS) pour évaluer les performances de l'assimilation par rapport à une simulation sans assimilation. Le CRPSS est défini comme le ratio entre le CRPS de l'état assimilé $CRPS_{DA}$ et le CRPS de l'état non assimilé $CRPS_{free}$:

$$CRPSS = 1 - \frac{CRPS_{DA}}{CRPS_{free}} \tag{6.53}$$

Le CRPSS est positif et d'autant plus proche de 1 que l'assimilation améliore l'estimation de l'état. Si le CRPSS est négatif, cela signifie que le processus d'assimilation dégrade l'estimation par rapport à l'état non assimilé.

6.4.2 Mise en oeuvre dans la thèse

Définition du problème

L'objectif de ces travaux est de réduire l'incertitude associée aux simulations spatialisées d'humidité de surface et de subsurface ainsi qu'aux simulations de concentration moyenne journalière à l'exutoire. On formule pour cela un problème d'estimation jointe où l'on cherche

à corriger ces variables ainsi que les paramètres d'entrée du modèle les plus influents sur ces dernières. La première partie de ces travaux a permis de montrer que pour les variables d'humidité, il s'agit des teneurs en eau à saturation (*thetas*) des horizons de surface et de subsurface. Pour la concentration à l'exutoire, il s'agit principalement du coefficient de rugosité de Manning des parcelles de vigne ($manning\_1$) ainsi que des paramètres hydrodynamiques ($thetas\_10$ et mn_10) de l'horizon le plus profond du type de sol 3.

Observations et opérateurs d'observations

On rappelle que l'on dispose des sources d'observations suivantes : des images satellites d'humidité de surface couvrant les 5 premiers centimètres du sol (une image par UH), des profils verticaux d'humidité ponctuels couvrant les 2 premiers mètres du sol et des séries de concentration moyenne hebdomadaires de pesticides à l'exutoire. Les résolutions spatiotemporelles et les erreurs associées à chaque type de données sont décrites dans la Section 2.3 (Chapitre 2).

- Pour les images d'humidité de surface, l'opérateur d'observation H est exprimé sous forme matricielle de manière à relier chaque observation à la moyenne pondérée d'humidité simulée dans les 6 cellules numériques qui composent les 5 premiers centimètres de la colonne de sol :

$$\theta_{5cm} = \frac{\sum_{j=1}^{6} \theta_j dz_j}{\sum_{j=1}^{6} dz_j}, \tag{6.54}$$

  où θ_j et dz_j sont respectivement les valeurs de l'humidité et de l'épaisseur de la cellule numérique j.

- Pour les profils verticaux d'humidité, on suppose observer directement l'état simulé et l'opérateur d'observation H associé est donc l'identité pour les 2 premiers mètres de sol.

- Pour la concentration dans la rivière, l'opérateur d'observation est également exprimé sous forme matricielle et permet de prendre en compte l'intégration temporelle qu'implique l'utilisation d'observations de concentration moyenne hebdomadaire. Chaque observation est ainsi reliée à la moyenne hebdomadaire de la concentration simulée :

$$c_{hebdo} = \frac{\sum_{j=1}^{7} c_{jour}}{7}, \tag{6.55}$$

  où c_{hebdo} est la concentration moyenne hebdomadaire observée et c_{jour} la concentration moyenne journalière simulée.

Génération de l'ensemble

Dans un premier temps, l'erreur modèle est supposée être contenue intégralement dans la définition des 145 paramètres d'entrée du modèle et dans les conditions initiales.

Pour perturber les paramètres, les distributions utilisées sont les mêmes que celles utilisées pour l'analyse de sensibilité. Toutefois, un biais supplémentaire est ajouté aux paramètres *thetas* ciblés par l'estimation jointe afin de tester les capacités du système d'assimilation à réduire à la fois le biais et la dispersion de l'ensemble. La valeur du biais varie entre 2 à 10% de la valeur de référence selon les horizons considérés et les distributions résultantes sont décrites dans l'Annexe H.
Pour générer une incertitude sur les conditions initiales, les niveaux de nappes mesurés par les piézomètres et utilisés pour initialiser les profils de pression des différentes UH sont supposés incertains. En première approximation, l'erreur est supposée gaussienne, non biaisée et d'écart-type σ=50 cm dans toutes les configurations.
On considère dans un premier temps, un ensemble de 50 membres. Toutefois, les performances du système d'assimilation pouvant varier grandement selon la taille de l'ensemble considéré, la sensibilité à ce paramètre sera explorée.

Configurations des expériences réalisées

Plusieurs configurations incluant différentes variables/paramètres dans le vecteur d'état et différentes sources d'observations combinées sont testées. Ces expériences font l'objet des Chapitres 7 et 8. Elles sont décrites ci-dessous et résumées dans le Tableau 6.3 :

- ✓ **Expérience 1** : on cherche à estimer l'humidité à toutes les profondeurs ainsi que les paramètres *thetas* de tous les horizons uniquement à partir des images satellites d'humidité de surface. Pour l'EnKF et l'iEnKS, à chaque analyse le vecteur d'état augmenté $\mathbf{x}$ contient les profils d'humidité pour toutes les UH et les valeurs des 14 *thetas*, soit une taille de n=14UH×25 cellules numériques+14 paramètres = 364. Pour l'ES-MDA, les variables corrigées sont les trajectoires temporelles et les paramètres. Le vecteur d'état augmenté résultant a ainsi une taille de $n = 899864$.

  - ■ **Comparaison des méthodes et sélection de la plus efficace**
    Dans un premier temps, l'EnKF, l'ES-MDA et l'iEnKS sont comparées de manière à déterminer quelle approche est la plus performante dans ce cas d'étude. Pour évaluer l'efficacité et la robustesse des méthodes d'assimilation testées, leur sensibilité à l'erreur d'observation et à la fréquence d'observation sont explorées. Pour les paramétrer au mieux, leur sensibilité à la taille de l'ensemble est également explorée. Chacun de ces facteurs est exploré de manière indépendante à partir de l'Expérience 1. Les valeurs testées sont reportées dans le Tableau 6.2. Une fois que la méthode la plus performante pour ce cas d'étude est identifiée, les expériences suivantes sont réalisées avec cette dernière et avec les valeurs nominales de fréquence et d'erreur d'observation. Quant à la taille de l'ensemble, est elle réglée de manière à optimiser les performances du système d'assimilation tout en garantissant un coût de calcul raisonnable.

écart type de l'erreur d'observation [cm^3cm^{-3}]	0.001 **0.020** 0.040 0.060 0.080 0.100 0.120 0.140 0.160 0.180 0.200 0.300 0.400
fréquence d'observation [jours]	1 2 3 4 5 **6** 8 9 10
taille de l'ensemble [-]	10 20 **50** 100 200 500

Tableau 6.2 – Liste des valeurs testées pour évaluer la sensibilité des méthodes d'assimilation à l'erreur d'observation, la fréquence d'observation et la taille de l'ensemble. Les valeurs nominales sont indiquées en gras.

La comparaison des 3 méthodes d'assimilation et l'exploration de leurs performances fait l'objet de la publication suivante soumise à la revue *Hydrology and Earth System Sciences* et actuellement en cours de révision:
Rouzies, E., Lauvernet, C. and Vidard, A. (2022) *Comparison of different ensemble assimilation methods for joint estimation in a pesticide transfer model* [document de travail]

- **Incertitudes sur les forçages climatiques**
  Dans l'Expérience 1, l'erreur modèle est supposée être contenue intégralement dans la définition des paramètres d'entrée et des conditions initiales. Dans un second temps, on considère que les forçages climatiques (précipitations uniquement) sont également incertains. L'impact d'une incertitude sur les précipitations est étudié dans l'Expérience 1rain en utilisant la méthode d'assimilation la plus adaptée identifiée dans l'Expérience 1.

✓ **Expérience 2** : on cherche à estimer l'humidité à toutes les profondeurs ainsi que les paramètres *thetas* de tous les horizons à partir des images satellites d'humidité de surface et des profils ponctuels d'humidité. Le vecteur d'état considéré reste le même que pour l'Expérience 1 et pour cette expérience, seuls les paramètres d'entrée et les conditions initiales sont supposés incertains pour générer l'ensemble (pas d'incertitude sur les forçages climatiques).

- **Intégration des profils verticaux d'humidité**
  Dans l'Expérience 2single, on suppose que des profils verticaux d'humidité sont disponibles sur une seule parcelle. Dans l'Expérience 2vineyard, des profils sont disponibles sur 3 parcelles de vigne répartis sur les différents types de sol et dans l'Expérience 2VFS, on dispose de profils sur 3 bandes enherbées des différents types de sol. Pour chaque configuration, on compare les performances du système en assimilant seulement les images de surface (configuration "Images"), seulement les profils verticaux (configuration "Profils") ou les 2 sources de données combinées (configuration "Multi").
- **Apport de la localisation**
  Les expériences issues de l'Expérience 2 sont réalisées sans puis avec localisation.

Une localisation par domaines est appliquée afin d'évaluer son potentiel pour diminuer le temps de calcul et atténuer de potentielles corrélations douteuses liées aux erreurs d'échantillonnage.

✓ **Expérience 3** : on cherche à estimer l'humidité, la concentration journalière en pesticide à l'exutoire et les paramètres influents sur les deux variables avec une assimilation multi-sources de données d'humidité de surface, de profils verticaux et de concentrations moyennées à l'exutoire. Le vecteur d'état contient ainsi les variables d'humidité et de concentration (1 valeur par jour pour la concentration, soit une trajectoire temporelle de taille 78) ainsi que l'intégralité des paramètres influents sur ces dernières (14 pour l'humidité + 2 pour la concentration à l'exutoire). Là encore, seuls les paramètres d'entrée et les conditions initiales sont considérés incertains pour la génération de l'ensemble.

Expérience	Variables/paramètres à corriger	Observations disponibles	Méthode	Sources d'incertitudes
Exp. 1 (Chapitre 7)	➢ Variables de sortie : - Humidité de surface et subsurface ➢ Paramètres d'entrée : - *thetas* pour tous les horizons	- Images d'humidité de surface	- EnKF - ES-MDA - iEnKS	- Paramètres d'entrée - Conditions initiales
Exp. 1rain (Chapitre 7)	➢ Variables de sortie : - Humidité de surface et subsurface ➢ Paramètres d'entrée : - *thetas* pour tous les horizons	- Images d'humidité de surface	- EnKF - ES-MDA - iEnKS	- Paramètres d'entrée - Conditions initiales - Forçages climatiques
Exp. 2 (Chapitre 8)	➢ Variables de sortie : - Humidité de surface et subsurface ➢ Paramètres d'entrée : - *thetas* pour tous les horizons	- Images d'humidité de surface - Profils verticaux ponctuels d'humidité	Meilleure méthode identifiée par l'Exp. 1	- Paramètres d'entrée - Conditions initiales

Expérience	Variables/paramètres à corriger	Observations disponibles	Méthode	Sources d'incertitudes
Exp. 3 (Chapitre 8)	➢ Variables de sortie : - Humidité de surface et subsurface - Concentration journalière à l'exutoire ➢ Paramètres d'entrée : - *thetas* pour tous les horizons - Rugosité de Manning et paramètres de VG horizon 10	- Images d'humidité de surface - Profils verticaux ponctuels d'humidité - Concentration moyenne hebdomadaire à l'exutoire	Meilleure méthode identifiée par l'Exp. 1	- Paramètres d'entrée - Conditions initiales

Tableau 6.3 – Description des expériences d'assimilation de données réalisées dans la thèse. Pour les Expériences 2 et 3, plusieurs configurations ont été testées et sont listées dans les Sections 8.1 et 8.2.

Chapitre 7

Assimilation d'images d'humidité de surface

Sommaire

Ce chapitre regroupe les résultats relatifs à l'assimilation d'images satellite d'humidité de surface dans le modèle PESHMELBA (Expérience 1 et Expérience 1rain). L'objectif est ici de réduire l'incertitude sur les profils verticaux d'humidité simulés et sur les teneurs en eau à saturation figurant parmi les paramètres d'entrée du modèle. L'utilisation de l'assimilation de données étant explorée pour la première fois dans PESHMELBA, 3 méthodes ensemblistes issues du filtre de Kalman sont comparées sur des expériences jumelles : filtre de Kalman d'ensemble (EnKF), lisseur d'ensemble avec assimilation multiple (ES-MDA) et lisseur de Kalman d'ensemble itératif (iEnKS). Dans ces expériences, on considère uniquement le scénario **estival** pour s'assurer que l'état suit une distribution gaussienne tout au long de la simulation, garantissant ainsi que les méthodes choisies peuvent être utilisées dans un cadre le plus optimal possible.
Le package Python développé pour mettre en oeuvre l'assimilation de données avec PESHMELBA est présenté Section 7.1. La comparaison des méthodes ainsi que l'exploration de leurs performances et de leur robustesse sont ensuite présentées dans la Section 7.2. Une fois qu'une méthode a été identifiée comme la plus performante pour assimiler les images d'humidité, la prise en compte d'incertitudes sur les forçages climatiques dans le système d'assimilation est envisagée dans la Section 7.3.

7.1 Présentation du package Python PASHA

De nombreux outils, notamment écrits en Python, existent pour l'assimilation de données, comme DAPPER (Raanes et al., 2018) ou PyDA (Ahmed et al., 2020). Cependant, adapter ces derniers aux contraintes qu'impose PESHMELBA pour le lancement, la surveillance d'une simulation et la mise à jour des paramètres d'entrée s'est avéré difficile. De plus, les outils existants mettent souvent l'accent sur l'optimisation numérique des calculs rendant parfois opaque la compréhension des algorithmes qu'ils implémentent. Pour toutes ces raisons, les méthodes d'assimilation utilisées dans ces travaux de thèse ont été recodées et regroupées dans le package Python PASHA (PedAgogic StocHastic data Assimilation). Celui-ci permet d'appliquer de manière relativement transparente et pédagogique des méthodes stochastiques d'assimilation à des modèles complexes comme PESHMELBA.

La structure globale du package est illustrée sur la Figure 7.1. Elle reprend la structure du package CASSIS développé par Agnès Printemps lors de son stage et se compose principalement d'une classe **Model** qui implémente le modèle et d'une classe **DAscheme** qui implémente la méthode d'assimilation utilisée. La classe **Model** est composée d'un ensemble de méthodes permettant d'initialiser le modèle, de le propager sur un ou plusieurs pas de temps consécutifs et de mettre à jour les conditions initiales et/ou les paramètres pour le cycle d'assimilation suivant. La classe **DAscheme** permet de piloter l'intégralité de l'expérience d'assimilation en réalisant l'alternance d'étapes de prévision et d'analyse. Différentes variantes du filtre de Kalman sont intégrées sous la forme de classes héritant de **DAscheme** : EnKF (deux variantes intégrées : Ensemble Transform Kalman Filter et version stochastique

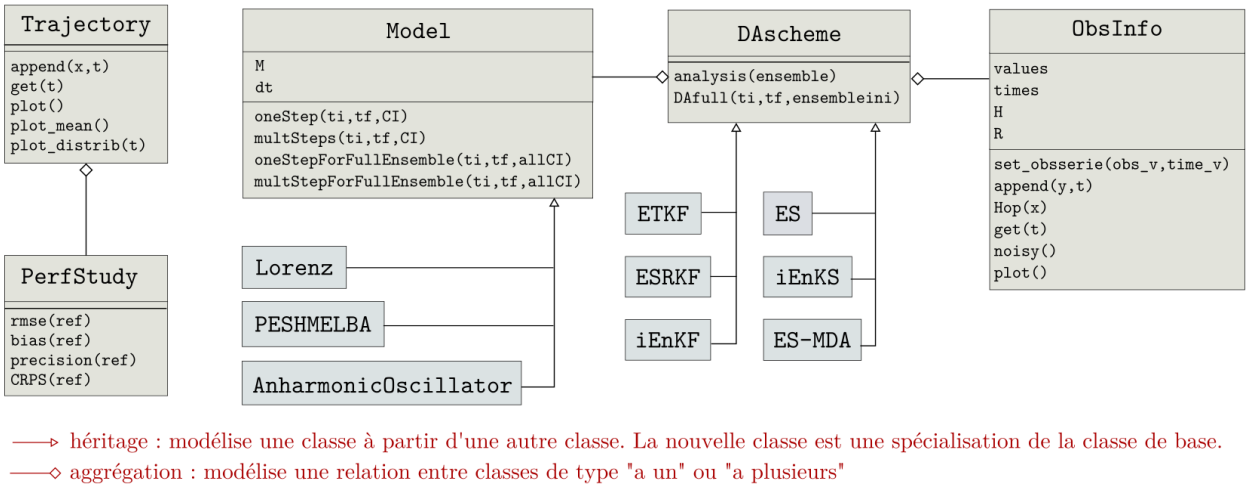


Figure 7.1 – Diagramme de classes partiel du package PASHA. Seuls les méthodes et attributs principaux sont détaillés. La composition des classes en gris n'est pas détaillée car leurs attributs et méthodes sont globalement les mêmes que ceux des classes Model et DAscheme dont elles héritent.

de l'EnKF comme proposée dans Burgers et al. 1998), ES, ES-MDA et iEnKS. Pour chacune d'entre elles, les méthodes *analysis* et *DAfull* sont redéfinies selon la stratégie utilisée pour l'analyse et selon le type d'assimilation considérée (filtre, lisseur à point fixe ou lisseur à intervalle fixe.).
Par ailleurs, une classe **ObsInfo** permet de fournir toutes les informations relatives aux observations (valeurs, fréquence, opérateur d'observation, matrice de covariance d'erreur,) et une classe **Trajectory** permet d'extraire certaines valeurs du vecteur d'état pour constituer des trajectoires temporelles. Les classes **ObsInfo** et **Trajectory** comportent également un ensemble de méthodes facilitant la visualisation de l'évolution temporelle de l'ensemble, des observations et de leurs erreurs. Enfin, la classe **PerfStudy** permet de calculer un ensemble de métriques permettant d'évaluer la qualité des résultats de l'assimilation par rapport à une trajectoire de référence (biais, RMSE, dispersion, CRPS).

L'utilisation de PASHA pour un nouveau modèle nécessite ainsi de créer une nouvelle classe qui hérite de la classe **Model** et où la méthode *oneStep* permettant de lancer le modèle sur un pas de temps est redéfinie en fonction des spécificités de ce dernier. Les méthodes de mise à jour des conditions initiales et des paramètres du modèle doivent aussi être redéfinies. Dans ces travaux, la classe PESHMELBA a ainsi été créée et sa méthode *oneStep* permet de gérer le lancement d'une simulation sur une machine classique ou sur un noeud de calcul d'un supercalculateur. De même, la méthode *oneStepForFullEnsemble* permettant de lancer PESHMELBA pour tous les membres d'un ensemble a été redéfinie de manière à permettre un lancement séquentiel ou en parallèle selon la machine utilisée.

Le package PASHA est disponible sur le dépôt `https://forgemia.inra.fr/emilie`.

`rouzies/pasha.git`. En plus de PESHMELBA, un modèle de Lorenz et un oscillateur anharmonique ont été implémentés pour une première prise en main et permettent de reproduire les codes et exemples fournis dans Asch et al. (2016). Aucune optimisation numérique n'a été réalisée et les algorithmes ont été gardés les plus compacts et transparents possible, ce qui permet une meilleure vision d'ensemble de la théorie mathématique qu'ils implémentent. Ainsi PASHA est un package simple qui peut constituer un outil pédagogique intéressant pour appréhender les méthodes stochastiques d'assimilation de données.

7.2 Comparaison des 3 méthodes

Dans cette section, les performances des 3 méthodes d'assimilation sont comparées pour la correction de l'humidité et l'estimation des teneurs en eau à saturation à partir d'images d'humidité moyenne dans les 5 premiers cm du sol sur le cas d'étude synthétique inspiré de la Morcille (scénario estival).
Un travail préliminaire a consisté à paramétrer l'iEnKS et l'ES-MDA pour cette application. Pour l'iEnKS, il était nécessaire de régler la taille L de la fenêtre d'assimilation (voir Figure 6.6) alors que l'ES-MDA nécessitait de renseigner le nombre d'itérations J. Pour cela, plusieurs valeurs de paramètres ont été testées sur le scénario nominal ([1, 3, 5, 8, 10, 12] pour L et [1, 2, 3, 5, 10, 15] pour J) et les valeurs retenues sont celles aboutissant au CRPS moyenné spatiotemporellement le plus faible sur les séries d'humidité. Les résultats ne sont pas montrés ici mais on retient que ces tests ont permis de fixer $L = 5$ pour l'iEnKS et $J = 3$ pour l'ES-MDA. D'autre part, pour l'iEnKS les valeurs du nombre d'itérations maximal $jmax$ et du facteur de contraction de l'ensemble ε sont respectivement fixées à $jmax = 3$ et $\varepsilon = 0.1$.

7.2.1 Correction des variables d'humidité

Pour les 3 méthodes, le CRPS est calculé à chaque pas de temps et à chaque profondeur pour évaluer l'impact de l'assimilation sur la correction des profils d'humidité. Dans ce paragraphe, les résultats sont d'abord présentés pour l'UH10 à 3 profondeurs (surface, 20 cm et 4 m) à titre d'exemple avant d'être généralisés à l'ensemble du bassin.

Afin de disposer d'une référence pour l'interprétation des résultats, la Figure 7.2 présente les trajectoires temporelles d'humidité sur l'UH10 aux 3 profondeurs pour la simulation de référence (état "vrai" utilisé dans les expériences jumelles). En surface, la dynamique de l'humidité est fortement dépendante des précipitations tout au long de la simulation. A la profondeur intermédiaire (20 cm), l'humidité ne varie significativement qu'après 1200 h de simulation, lorsque des événements pluvieux plus intenses se produisent. Enfin, en fond de profil (4 m), l'humidité n'est pas affectée par les forçages climatiques et reste constante tout au long de cette simulation.

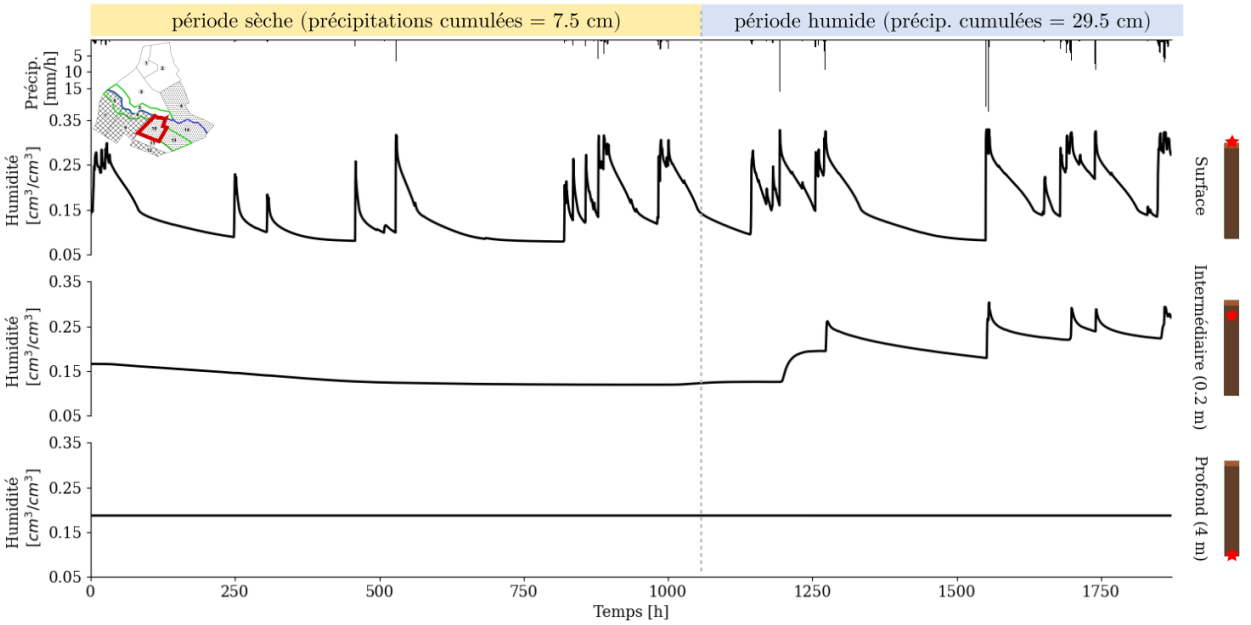


Figure 7.2 – Série temporelle d'humidité en surface (haut), à 20 cm de profondeur (milieu) et en fond de profil (bas) sur l'UH10 pour la simulation de référence (état "vrai" synthétique). L'histogramme noir inversé en haut de la figure illustre la chronique de précipitations et les profondeurs représentées sont situées par une étoile rouge sur les colonnes de sol à droite de chaque série temporelle.

Correction de l'humidité de surface par UH

La Figure 7.3 présente les séries temporelles de CRPS calculé pour l'humidité de surface sur l'UH10 sans assimilation (free run) et avec assimilation en utilisant les 3 méthodes.
Pour le free run, l'évolution temporelle du CRPS est largement corrélée avec la série temporelle de précipitations. Les événements pluvieux et les périodes de récession qui les suivent sont associés à des pics dans la série de CRPS. Ceux-ci illustrent un niveau d'incertitude plus élevé dans le système en conditions humides, lié à une dynamique plus marquée et impliquant plus de processus physiques.
Au cours des 3 premiers cycles d'assimilation (jusqu'à 576 h), l'iEnKS et l'EnKF ne permettent qu'une diminution limitée de l'erreur par rapport au free run. Dans cette première partie de la simulation, les corrections effectuées pendant les étapes d'analyse sont quasiment oubliées à partir de l'événement pluvieux suivant. Jusqu'à 576 h, l'ES-MDA est donc la seule méthode qui garantit une réduction claire du CRPS, à la fois pendant et entre les événements pluvieux. Par contre, à partir de 1000 h toutes les méthodes permettent une réduction significative du CRPS. Pour expliquer de tels écarts, on rappelle d'abord que seul l'ES-MDA effectue une correction globale et intègre l'information provenant de toutes les observations à la fois ainsi que la dynamique du système pour corriger l'état. L'ES-MDA garantit ainsi que les corrections peuvent se propager des instants observés vers les instants non observés et notamment des périodes de pluie aux périodes séparant les évènements (à condition qu'il existe des corrélations temporelles suffisantes). D'autre part, on peut supposer que pendant

la première partie de simulation caractérisée par un régime hydrologique plutôt sec, les teneurs en eau à saturation *thetas* ne sont pas observables, limitant aussi les performances de l'EnKF et de l'iEnKS qui corrigent le vecteur d'état augmenté à un instant donné. Par contre, la correction globale et itérative de l'ES-MDA permet de répercuter l'impact d'une bonne estimation des paramètres rendue possible pendant le régime humide observable vers le régime sec non observable.

Comme pressenti par les séries temporelles de CRPS, l'ES-MDA aboutit ainsi à la valeur de CRPSS moyenné temporellement la plus élevée sur l'UH10 et cette meilleure performance se généralise ensuite quasiment à l'intégralité du bassin versant (voir Tableau 7.1). Comme sur l'UH10, les performances de l'EnKF et de l'iEnKS sont assez similaires.

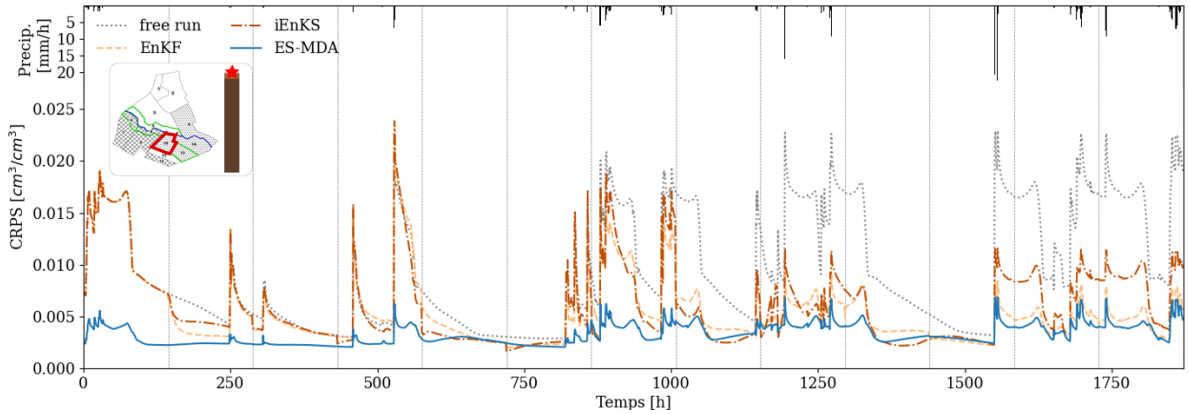


Figure 7.3 – Comparaison des séries temporelles de CRPS pour le free run et pour les différentes méthodes d'assimilation pour le compartiment de surface sur l'UH10. L'histogramme noir inversé en haut de la figure illustre la chronique de précipitations et les lignes verticales grises indiquent les instants auxquels des observations d'humidité de surface sont disponibles.

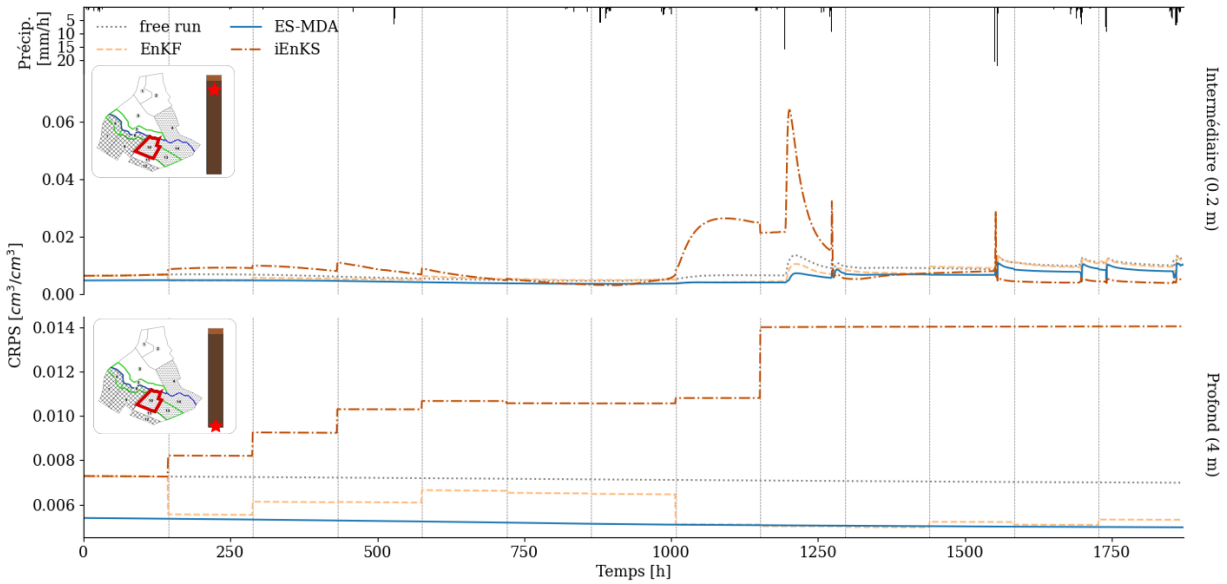


Figure 7.4 – Comparaison des séries temporelles de CRPS pour le free run et pour les différentes méthodes d'assimilation sur l'UH10 aux profondeurs de 20 cm (haut) et de 4 m (bas). L'histogramme noir inversé en haut de la figure illustre la chronique de précipitations et les lignes verticales grises infiquent les instants auxquels des observations d'humidité de surface sont disponibles.

	Surface			Intermédiaire (0.2 m)			Profond (4 m)		
UH	EnKF	iEnKS	ES-MDA	EnKF	iEnKS	ES-MDA	EnKF	iEnKS	ES-MDA
1	0,501	0,400	**0,713**	-0,029	-0,199	**0,106**	**-0,136**	-0,671	-0,362
2	**0,501**	0,400	0,424	-0,489	**0,399**	-0,334	0,172	0,497	**0,549**
3	0,503	0,400	**0,627**	0,099	-0,272	**0,268**	0,188	-0,627	**0,277**
4	0,389	0,342	**0,665**	0,097	-0,228	**0,265**	0,179	-0,537	**0,277**
5	0,265	0,402	**0,627**	0,111	**0,315**	0,253	0,179	-0,717	**0,261**
6	0,495	0,363	**0,664**	0,096	-0,223	**0,264**	0,186	-0,603	**0,278**
7	0,103	0,364	**0,714**	-0,029	-0,195	**0,108**	**-0,129**	-0,693	-0,361
8	**0,506**	0,399	0,424	-0,499	**0,404**	-0,332	0,170	0,492	**0,544**
9	0,104	0,367	**0,713**	-0,027	-0,206	**0,105**	**-0,134**	-0,678	-0,362
10	0,416	0,361	**0,713**	-0,025	-0,216	**0,104**	**-0,130**	-0,689	-0,361
11	0,502	0,400	**0,781**	-0,442	**0,075**	-0,193	0,262	0,604	**0,660**
12	0,105	0,365	**0,424**	-0,496	**0,397**	-0,335	0,175	0,503	**0,553**
13	0,470	0,338	**0,757**	-0,010	-0,229	**0,142**	-0,286	**-0,147**	-0,387
14	0,416	0,360	**0,777**	-0,383	**0,305**	-0,142	0,263	0,641	**0,662**

Tableau 7.1 – Comparaison des CRPSS moyennés temporellement aux 3 profondeurs pour l'estimation de l'humidité (Expérience 1). La ligne en rouge correspond aux résultats de l'UH10 présentés sur les Figures 7.3 et7.4. Pour chaque ligne et dans chaque compartiment, la valeur en gras indique la meilleure estimation.

Correction de l'humidité de subsurface par UH

Contrairement à l'estimation de l'humidité de surface, les 3 méthodes présentent des performances bien plus limitées en subsurface comme le montre la Figure 7.4 pour l'UH10. Sur cette UH, à 20 cm de profondeur le free run est associé à des valeurs de CRPS plus élevées après 1200 h, instant à partir duquel les événements pluvieux deviennent plus intenses. Là encore, l'ES-MDA est la méthode la plus performante pour réduire l'erreur. Toutefois, pour toutes les méthodes appliquées, le gain par rapport au free run est beaucoup plus limité qu'en surface. A l'échelle du bassin versant, l'assimilation a plutôt tendance à dégrader l'estimation de l'humidité dans ce compartiment (voir valeurs de CRPSS souvent négatives dans le Tableau 7.1, colonne 2).
On note que sur l'UH10 l'iEnKS dégrade l'estimation de l'humidité, en particulier entre 1000 h et 1250 h. On rappelle que l'iEnKS est la seule méthode qui est caractérisée par une fenêtre d'assimilation mouvante. A chaque analyse, l'état actuel est corrigé en utilisant les L observations suivantes ($L = 5$ dans ce cas). Pour l'analyse ayant lieu à 1008 h, le changement de dynamique du système entre l'instant de l'analyse et les instants des observations utilisées (entre 1008 h et 1728 h) peut expliquer une si mauvaise correction de l'état. Cependant, si les performances de l'iEnKS sont particulièrement mauvaises pour l'UH10, les valeurs des CRPSS reportées dans le Tableau 7.1 pour les autres UH nuancent ces conclusions. En effet, l'iEnKS aboutit à des valeurs de CRPSS positives pour les UH2, 5, 8, 11, 12 et 13 alors que l'EnKF dégrade l'estimation de l'humidité sur quasiment toutes les UH. Le comportement reporté sur la Figure 7.4 correspond ainsi probablement à un phénomène local qui doit être interprété avec prudence.
En fond de profil, le CRPS reste quasiment constant entre chaque étape d'analyse car l'humidité ne varie que très peu dans ce compartiment. L'EnKF et l'ES-MDA aboutissent à une diminution limitée du CRPS par rapport au free run, tandis que l'iEnKS a plutôt tendance à diverger et à dégrader l'estimation jusqu'à 1152 h, date de la dernière analyse (après 1152 h, il n'y a plus assez d'observations disponibles dans la fenêtre d'assimilation). Là encore, sur les autres UH, l'ES-MDA aboutit le plus souvent aux valeurs de CRPSS les plus élevées alors que les performances de l'EnKF sont très limitées et que l'iEnKS dégrade l'estimation de l'humidité sur près de deux tiers des UH (Tableau 7.1, colonne 3).
Plus généralement, on rappelle que dans cette expérience, la subsurface n'est pas observée directement puisque les observations se concentrent en surface. Malgré l'existance de rétroactions de la subsurface vers la surface, c'est majoritairement la dynamique en surface qui régit celle de la subsurface. Le problème inverse ainsi formulé est donc peut-être mal choisi puisqu'on souhaite plutôt que la variable à corriger ait un impact sur les observations.

Généralisation des résultats à l'échelle du bassin versant

Afin d'identifier quantitativement la méthode la plus performante pour la correction des profils d'humidité, les CRPSS moyennés sur toute la simulation et sur tout le bassin versant sont calculés pour les 3 méthodes. Les résultats sont regroupés sur la Figure 7.5 et montrent

que toutes les méthodes améliorent de plus de 38% l'estimation de l'humidité en surface. L'ES-MDA aboutit à de meilleures performances que l'EnKF et l'iEnKS en surface mais aussi en subsurface. A 20 cm de profondeur, l'ES-MDA et l'iEnKS améliorent légèrement l'estimation de l'humidité alors que l'EnKF dégrade significativement l'estimation de l'état (valeur de CRPSS négative). En fond de profil, l'ES-MDA est également plus performante que l'EnKF tandis que l'iEnKS conduit à une dégradation significative de l'estimation de l'état par rapport au free run. Néanmoins, il est important de noter que l'ordre de grandeur de la réduction de l'erreur atteint par l'ES-MDA en subsurface est bien inférieur à celui de la surface.

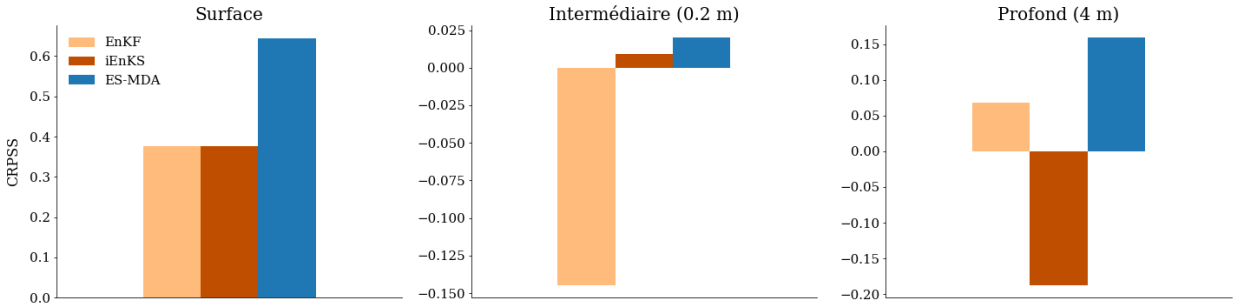


Figure 7.5 – CRPSS moyens pour les différentes méthodes d'assimilation en surface (gauche), à 20 cm de profondeur (milieu) et en fond de profil (droite). Une valeur positive indique que l'assimilation diminue l'erreur par rapport au free run alors qu'une valeur négative indique une augmentation de l'erreur par rapport au free run.

A retenir

- ✓ En surface, toutes les méthodes améliorent significativement l'estimation de l'humidité.
- ✓ En surface, l'ES-MDA aboutit à de meilleures performances que l'EnKF et l'iEnKS.
- ✓ En subsurface, l'assimilation n'améliore que très peu (voire dégrade) l'estimation de l'humidité.
- ✓ En subsurface, l'ES-MDA aboutit aux meilleurs résultats malgré des performances très limitées par rapport à la surface.

7.2.2 Estimation des teneurs en eau à saturation

Pour examiner les performances des 3 méthodes pour l'estimation de paramètres, la Figure 7.6 compare les distributions *a priori* et *a posteriori* pour les teneurs en eau à saturation puis les valeurs de CRPSS associées sont regroupées dans le Tableau 7.2.
L'estimation du paramètre *thetas* pour les horizons de **surface** (colonnes 1 et 2) est nettement améliorée par les 3 méthodes puisque les distributions à posteriori montrent visuellement une réduction claire du biais et de la dispersion et les valeurs de CRPSS dépassent 0.58 dans tous les cas. D'après les valeurs de CRPSS, l'EnKF et l'ES-MDA sont identifiées

comme les méthodes les plus performantes.
Les conclusions sont moins claires pour l'estimation des paramètres de **subsurface** (colonnes 3, 4 et 5). En effet, les valeurs de CRPSS sont bien plus faibles que pour la surface et l'assimilation dégrade même l'estimation de *thetas* pour les horizons 4, 7 et 6 (hormis pour l'iEnKS). Par contre, dans ce compartiment, l'iEnKS aboutit étonnamment à de meilleures performances que pour la correction des variables d'humidité. Finalement, que ça soit pour les paramètres de surface ou de subsurface, on note que la distribution à posteriori et les CRPSS associés obtenus par l'EnKF et l'ES-MDA sont assez proches. Cela montre que l'assimilation des observations une par une ou toutes à la fois en utilisant une méthode basée sur un filtre de Kalman d'ensemble, conduit à des performances comparables pour l'estimation des paramètres alors que des différences significatives sont notées pour la correction des variables.

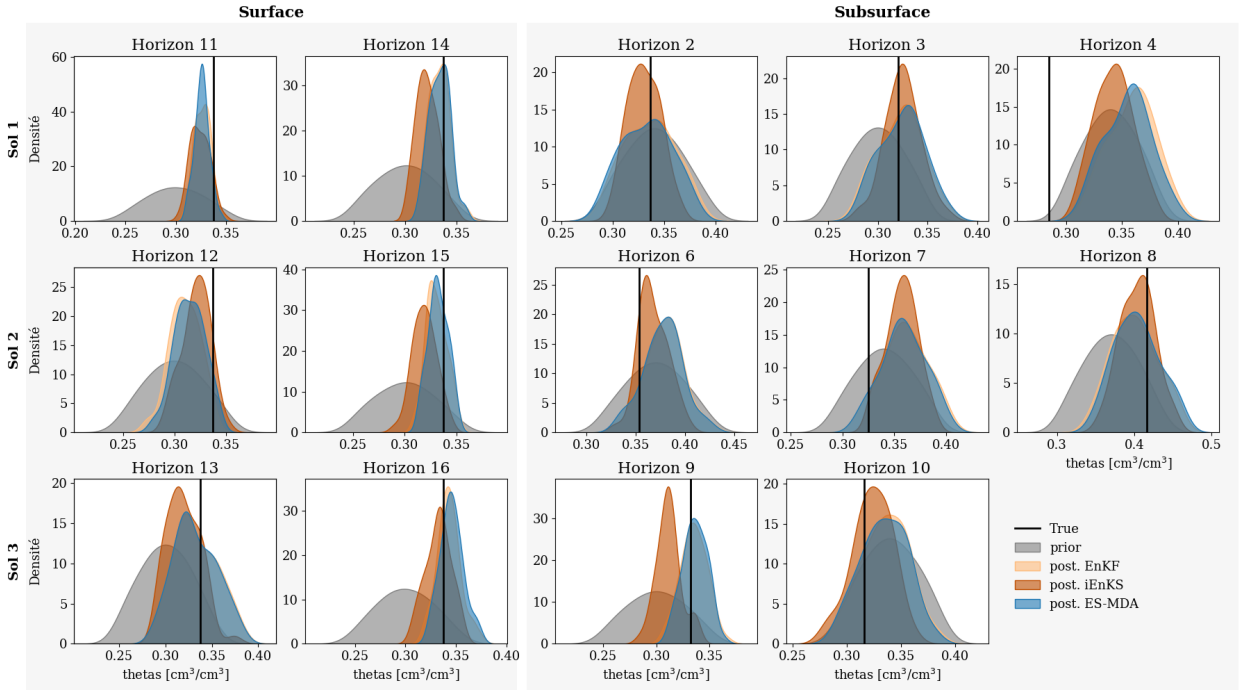


Figure 7.6 – Pdf empiriques pour les teneurs en eau à saturation estimées pour les horizons de sol de l'UCS1 (haut), l'UCS2 (milieu) et l'UCS3 (bas). La première colonne regroupe les pdf des *thetas* pour les horizons de surface des parcelles de vigne alors que la seconde colonne regroupe les pdf des *thetas* définis en surface des bandes enherbées. Les colonnes suivantes regroupent les paramètres relatifs aux horizons de subsurface. La valeur utilisée pour le scénario de référence (valeur "vraie") est indiquée par une ligne noire pour chaque horizon. Pour les horizons où seules 2 pdf sont visibles, les pdf associées à l'EnKF et l'ES-MDA se superposent.

	Horizon	EnKF	ES-MDA	iEnKS
Surface	11	**0.68**	0.65	0.62
	12	0.18	0.36	**0.58**
	13	**0.70**	**0.70**	0.52
	14	0.87	**0.88**	0.54
	15	0.80	**0.85**	0.47
	16	**0.86**	0.80	0.85
Subsurface	2	0.06	0.07	**0.26**
	3	0.51	0.48	**0.64**
	4	-0.47	-0.39	**-0.14**
	6	-0.52	-0.48	**0.28**
	7	-1.44	**-1.35**	-1.42
	8	0.62	0.66	**0.74**
	9	**0.84**	0.83	0.23
	10	0.16	0.20	**0.59**

Tableau 7.2 – Valeurs des CRPSS associées à l'estimation de chaque *thetas* par les 3 méthodes d'assimilation. Une valeur positive indique une diminution de l'erreur par rapport à l'état non assimilé (free run) alors que les valeurs négatives indiquent une augmentation de l'erreur par rapport au free run. Pour chaque ligne, la valeur en gras indique la meilleure estimation obtenue par assimilation.

A retenir

- ✓ Les teneurs en eau à saturation soint mieux estimées en surface qu'en subsurface.
- ✓ Les performances de l'EnKF et de l'ES-MDA sont les meilleures et assez proches en surface.
- ✓ En subsurface, l'iEnKS mène aux meilleures performances.

7.2.3 Exploration du comportement des 3 méthodes

Dans ce paragraphe, les performances et la robustesse des différentes méthodes sont explorées en évaluant leur sensibilité à :

1. l'amplitude de l'erreur d'observation ;
2. la fréquence d'observation ;
3. la taille de l'ensemble.

Chacun de ces facteurs est exploré de manière indépendante à partir des valeurs regroupées dans le Tableau 6.2. Compte tenu du coût de calcul très élevé associé à l'utilisation de l'iEnKS et à des difficultés techniques liées au lancement de PESHMELBA, l'étude de la sensibilité de l'iEnKS n'a pas pu être menée à bien et les résultats ne sont présentés que pour l'EnKF et l'ES-MDA.
Pour rappel, la configuration de la simulation nominale est la suivante : erreur d'observation de 0.02 $cm^3.cm^{-3}$, fréquence d'observation de 6 jours (144 h) et ensemble composé de 50

membres. Les paragraphes précédents ayant montré des performances très limitées des 3 méthodes en subsurface, l'analyse n'est menée qu'à partir des valeurs de CRPSS associées à l'humidité de surface.

Sensibilité à l'erreur d'observation

La Figure 7.7 regroupe les valeurs de CRPSS moyennés spatiotemporellement pour l'estimation de l'humidité de surface en fonction de l'écart-type de l'erreur d'observation. Dans cette expérience, la fréquence d'observation est fixée à 144 h et la taille de l'ensemble à 50. Comme attendu, que ça soit pour l'EnKF ou l'ES-MDA, le CRPSS est le plus haut pour les erreurs d'observation les plus faibles et décroît régulièrement à mesure que l'erreur d'observation augmente. Pour l'EnKF, il n'y a plus de correction significative de l'état pour des erreurs d'observation supérieures à 0.1 cm^3cm^{-3}. Pour l'ES-MDA, l'état est significativement corrigé pour des erreurs jusqu'à 0.2 cm^3cm^{-3}.
Or, on rappelle que des observations d'humidité de surface obtenues à partir de l'inversion synergique de données Sentinel-1/Sentinel-2 ne sont pas encore disponibles sur les vignes. Ces résultats fournissent ainsi des informations précieuses sur la qualité requise de ces observations lorsqu'on souhaite les utiliser pour de l'assimilation de données. En effet, pour obtenir une amélioration significative de l'estimation de l'humidité de surface (fixée par exemple à plus de 20% par rapport au free run), on retiendra qu'il faut considérer une erreur d'observation inférieure à 0.05 cm^3cm^{-3} pour l'EnKF (resp. 0.1 cm^3cm^{-3} pour l'ES-MDA).

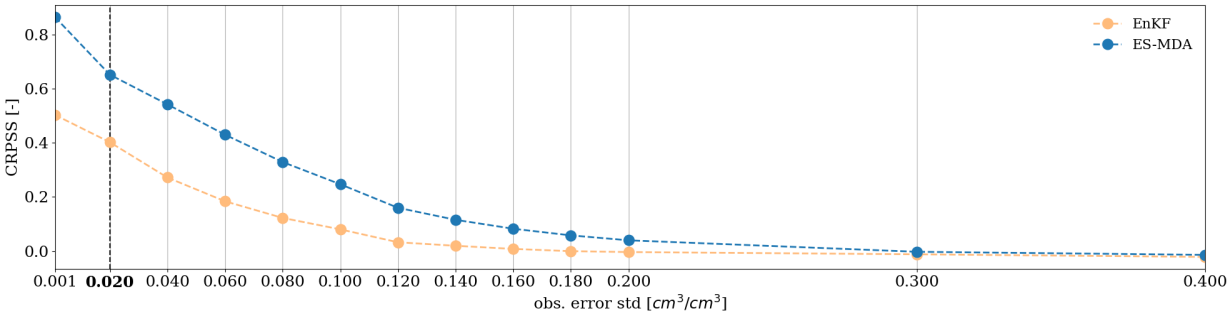


Figure 7.7 – Sensibilité du CRPSS de l'humidité de surface à l'écart-type de l'erreur d'observation (le CRPSS est moyenné spatiotemporellement à l'échelle du bassin). La ligne verticale en pointillés et le label en gras indiquent la valeur nominale d'erreur d'observation utilisée dans le scénario nominal.

Sensibilité à la fréquence d'observation

La Figure 7.7 regroupe les valeurs de CRPSS moyennés spatiotemporellement pour l'estimation de l'humidité de surface en fonction de la fréquence d'observation. L'écart-type de l'erreur d'observation est fixé à 0.02 cm^3cm^{-3} et la taille de l'ensemble à 50. Pour l'EnKF, le CRPSS est relativement stable pour des fréquences d'observation (et donc d'analyse) jusqu'à 72 h puis elle décroît régulièrement. Dans le cas de l'ES-MDA, ses performances sont

relativement stables pour des fréquences d'observation inférieures à 144 h. Elles diminuent légèrement pour des fréquences de 192 h et 216 h avant de s'effondrer lorsque les observations sont disponibles toutes les 240 h (10 jours). Comme l'ES-MDA utilise toutes les observations à la fois, la quantité d'information utilisée pour corriger le système est intrinséquement plus grande pour des fréquences faibles (*i.e.* lorsque les observations sont disponibles le plus fréquemment). Compte tenu de la longueur de la fenêtre d'assimilation (78 jours), augmenter la fréquence d'observation n'a donc pas d'intérêt pour l'ES-MDA contrairement aux autres méthodes. En effet, l'intégration d'information supplémentaire sur l'humidité de surface ne contribue pas à mieux contraindre le système.
En moyenne, la fréquence des images d'humidité de surface fournies par Sentinel-1/Sentinel-2 est de 6 jours. Elle est optimale pour l'ES-MDA ce qui n'est pas le cas pour l'EnKF qui nécessiterait d'intégrer des observations toutes les 72 h pour optimiser ses performances.

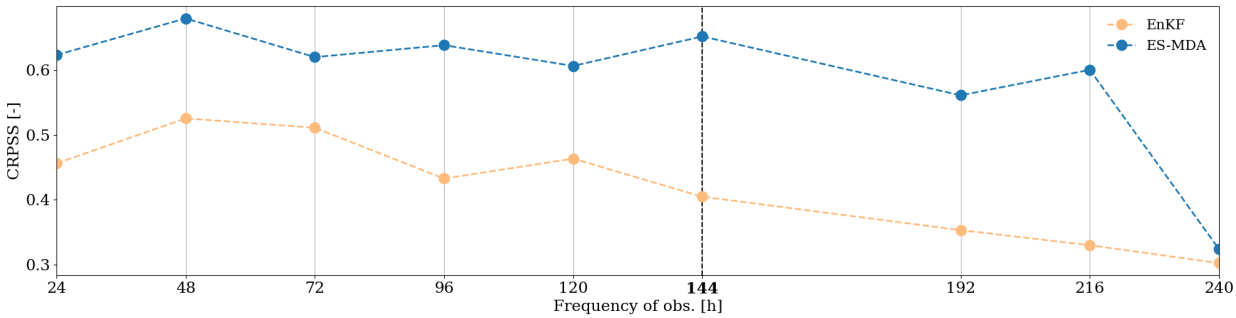


Figure 7.8 – Sensibilité du CRPSS de l'humidité de surface à la fréquence d'observation (le CRPSS est moyenné spatiotemporellement à l'échelle du bassin). La ligne verticale en pointillés et le label en gras indiquent la valeur nominale de fréquence d'observation utilisée dans le scénario nominal.

Sensibilité à la taille de l'ensemble

Alors que l'amplitude de l'erreur d'observation et la fréquence d'observation sont des propriétés intrinsèques du jeu d'observations, la taille de l'ensemble est un paramètre qui peut être réglé par l'utilisateur. Ce choix impacte d'ailleurs de manière critique le coût numérique de l'expérience d'assimilation et il faut chercher un compromis entre coût de calcul limité et précision suffisante de l'analyse.
Dans cette expérience, la sensibilité des différentes méthodes à la taille de l'ensemble est testée sur un scénario incluant une erreur d'observation σ_{obs}=0.02 cm^3cm^{-3} et une fréquence d'observations de 144 h. Compte tenu des erreurs d'échantillonnage qui peuvent grandement affecter les performances de l'assimilation, les expériences pour les ensembles de petite taille devraient être reproduites plusieurs fois en rééchantillonnant de nouveaux membres *a priori*. Dans cette expérience, c'est le cas pour l'ES-MDA pour lequel les expériences ont été reproduites 10 fois pour les ensembles contenant 10, 20 et 50 membres. Pour l'EnKF, une seule expérience a pu être réalisée pour chaque taille d'ensemble. La Figure 7.9 montre l'évolution du CRPSS de l'humidité de surface en fonction de la taille de l'ensemble pour l'ES-MDA

et l'EnKF. Pour l'EnKF, la valeur de CRPSS la plus faible (0.11) est atteinte pour une taille d'ensemble de 20 membres puis le CRPSS se stabilise autour de 0.35 pour des tailles d'ensemble supérieures à 100 membres. Pour l'ES-MDA, le CRPSS moyen est le plus faible pour une ensemble de 10 membres. Comme attendu, c'est aussi pour cette taille d'ensemble que l'on constate la plus grande variabilité de performances démontrant qu'une très petite taille d'ensemble ne permet pas d'améliorer de manière fiable l'estimation de l'humidité. Les valeurs de CRPSS augmentent ensuite et la variabilité associée diminue pour les ensembles de plus grande taille. Le CRPSS se stabilise autour de 0,56 pour des tailles d'ensemble supérieures à 100.
Comme pour l'EnKF, 100 membres semble alors un compromis acceptable pour garantir à la fois l'exactitude des résultats et un coût de calcul limité. Cette taille de 100 est d'ailleurs souvent choisie dans les études qui décrivent leur dispositif expérimental d'assimilation (Camporese et al., 2009 ; Nie et al., 2011 ; Lei et al., 2020).

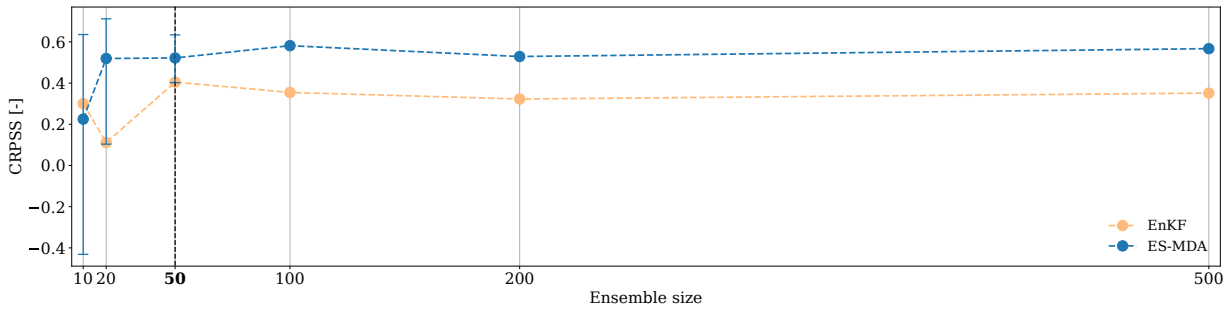


Figure 7.9 – Sensibilité du CRPSS de l'humidité de surface à la taille de l'ensemble (le CRPSS est moyenné spatiotemporellement à l'échelle du bassin). La ligne verticale en pointillés et le label en gras indiquent la valeur nominale de taille d'ensemble utilisée dans le scénario nominal. Pour l'ES-MDA, pour les tailles 10, 20 et 50, les points bleus correspondent aux valeurs moyennes tandis que les extrémités des barres d'incertitudes indiquent les valeurs minimales et maximales.

A retenir

- ✓ La sensibilité à l'erreur d'observation permet de déterminer la qualité nécessaire des images radar provenant de Sentinel-1/Sentinel-2 pour pouvoir utiliser chaque méthode d'assimilation.
- ✓ La fréquence d'observation de 144 h (6 jours) des images radar provenant de Sentinel-1/Sentinel-2 est optimale pour l'ES-MDA mais pas pour l'EnKF.
- ✓ Pour l'EnKF et l'ES-MDA, un ensemble de 100 membres ou plus est nécessaire pour limiter les erreurs d'échantillonnage.

7.2.4 Discussion sur le choix de la méthode

Les paragraphes précédents ont permis de montrer que l'ES-MDA qui intègre toutes les observations à la fois et prend en compte la dynamique du système pour effectuer l'analyse

est le plus efficace pour corriger les trajectoires d'humidité, particulièrement en surface. Par contre, pour l'estimation de paramètres *thetas* en surface, l'intégration des observations une par une est suffisante et l'EnKF aboutit à des performances semblables à l'ES-MDA.
Les résultats ont également montré que l'iEnKS est une approche prometteuse, notamment pour l'estimation de paramètres. Cependant, le coût de calcul moyen de l'iEnKS (2143 hCPU) demeure largement supérieur à celui de l'EnKF (277 hCPU) et de l'ES-MDA (558 hCPU). En effet, on rappelle que l'utilisation d'une fenêtre d'assimilation de taille L et d'une approche itérative impliquent qu'il faut intégrer le modèle jusqu'à $M \times L \times jmax$ fois (où $jmax$ est le nombre d'itérations maximal fixé) pour effectuer chaque analyse (voir Tableau 6.1) alors que cette étape ne fait pas intervenir d'intégration supplémentaire dans le cas de l'EnKF ou de l'ES-MDA. Or, comme c'est souvent le cas, l'intégration de PESHMELBA est l'étape la plus coûteuse du processus d'assimilation, et ce même sur des fenêtres temporelles courtes (même si le modèle est assez rapide, la lecture et l'écriture de fichiers extérieurs augmentent significativement le temps de calcul d'une intégration). Ainsi, pour être capable de mettre en oeuvre l'iEnKS, le nombre d'itérations autorisées pour la minimisation de la fonction coût a été limité à 3. Cependant, en fixant arbitrairement ce nombre d'itérations, on ne garantit pas la convergence de l'analyse comme le montre la Figure 7.10.

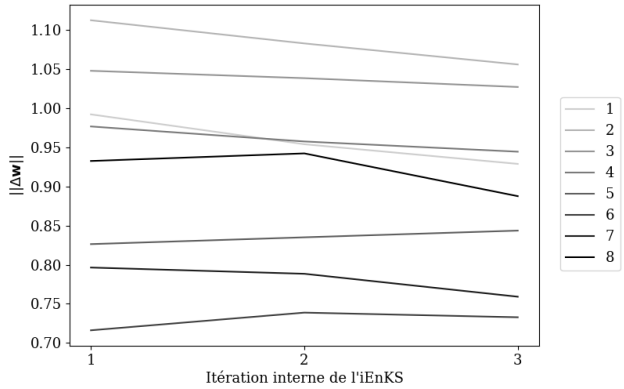


Figure 7.10 – Evolution de la norme de l'incrément du vecteur de coordonnées $||\mathbf{\Delta w}||$ à chaque itération interne de l'iEnKS pour les 8 analyses réalisées.

Ainsi, des tests supplémentaires incluant plus d'itérations devraient idéalement être conduits. Dans le cas de PESHMELBA, de tels tests ne sont pas envisageables et le coût de l'iEnKS en termes d'intégrations du modèle justifie de l'abandonner pour cette application. Pour pouvoir explorer les performances d'une telle méthode plus en détails, des modifications de la structure de PESHMELBA pourront être envisagées à l'avenir afin d'en limiter le coût d'intégration : utilisation du coupleur OpenFLUID en remplacement d'OpenPALM, intégration du module d'assimilation à l'intérieur du coupleur, etc.

Au vu des résultats précédents, l'ES-MDA est identifié comme le meilleur compromis entre performances et temps de calcul pour l'estimation jointe de profils d'humidité et de paramètres *thetas*. Cette méthode est facile à implémenter et ne nécessite pas d'interrompre

régulièrement le code de calcul pour réaliser ses analyses. De plus, bien qu'il s'agisse d'une méthode itérative, son coût de calcul reste raisonnable et on a montré que seulement quelques itérations sont nécessaires pour optimiser ses performances. L'analyse de la sensibilité de ses performances à l'erreur et à la fréquence d'observation attestent de sa robustesse et de son intérêt pour la configuration mise en place dans ces travaux. Enfin, l'analyse de sa senbilité à la taille de l'ensemble a permis d'identifier une taille d'ensemble optimale pour ces travaux (100 membres) garantissant à la fois la précision des calculs et un coût numérique limité.

Cependant, même s'il a été montré que l'ES-MDA est performant pour l'estimation des variables et paramètres de surface, il est important de rappeler qu'il échoue pour estimer variables et paramètres en subsurface. Plus généralement, aucune des méthodes testées n'a permis d'estimer correctement les variables profondes à partir de l'assimilation d'observations couvrant les 5 cm supérieurs de la colonne de sol.
Or, l'examen des matrices de corrélations à différents instants de la simulation (voir Figure 7.11 pour un exemple) montre qu'il n'existe quasiment aucune corrélation entre la surface et la subsurface.

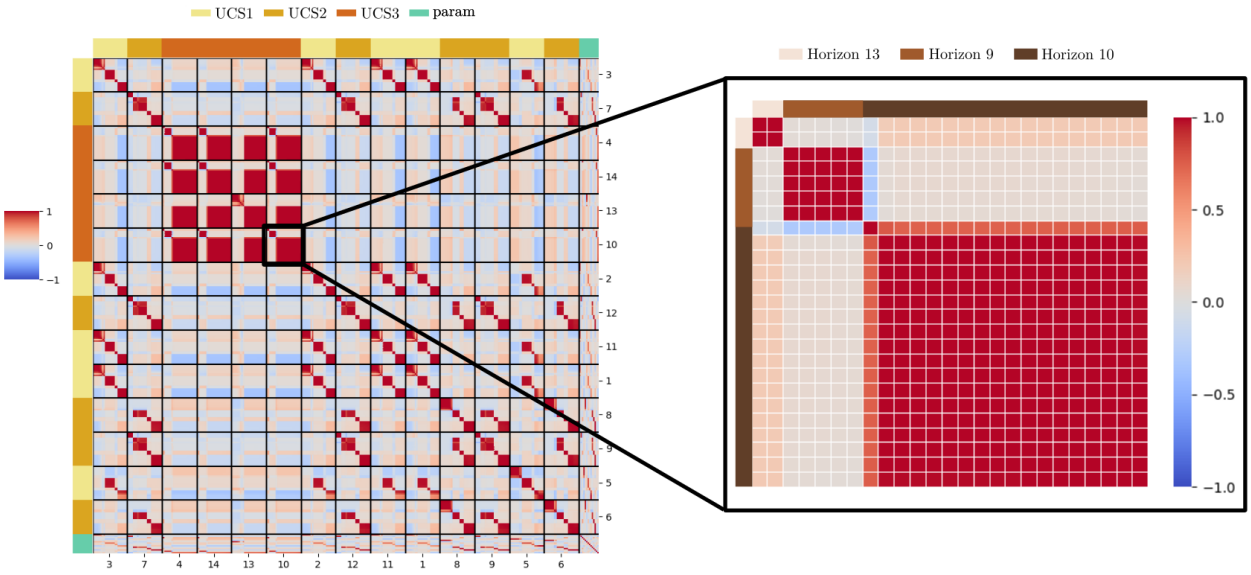


Figure 7.11 – Gauche : matrice de corrélation de l'ensemble après 144 h de simulation pour le free run sur l'ensemble des UH. Les lignes noires horizontales et verticales délimitent les cellules numériques d'une même UH. Les couleurs en haut et à gauche indiquent le type de sol de chaque UH alors que les numéros en bas et à droite indiquent les indices des UH. Droite : extrait de la matrice de corrélation correspondant à l'UH10. Chaque pixel délimite une cellule numérique du profil vertical de l'UH10. Les couleurs en haut et à gauche indiquent l'horizon de sol auquel appartient chaque cellule numérique.

De même, les matrices de corrélation entre trajectoires temporelles en surface et en subsurface (non montrées ici) montrent qu'il n'existe que très peu de corrélations entre la surface et la subsurface, même en incluant des potentiels décalages temporels. Or, dans les méthodes

type filtre de Kalman d'ensemble, la correction des compartiments et des paramètres non observés se fait au travers des corrélations avec les points spatiotemporels observés. En l'absence de telles corrélations, les corrections fournies par l'assimilation d'images d'humidité de surface ne peuvent pas se propager aux variables et paramètres de subsurface. Dans ce cas d'étude, la présence d'horizons distincts dans chaque type de sol représenté ainsi que la discrétisation verticale, beaucoup plus raffinée en surface qu'en subsurface peuvent expliquer une telle absence de corrélation. Une telle absence de corrélation a d'ailleurs déjà été constatée dans Bonan et al. (2020) (dans des contextes arides seulement) et la discrétisation de l'équation de Richards est mise en cause pour expliquer une telle décorrélation (B. Bonan, communication personnelle).
Toutefois, la Figure 7.11 montre aussi qu'à toutes les profondeurs, deux cellules numériques de deux UH différentes sont fortement corrélées si elles appartiennent au même horizon de sol (couleurs autour de la matrice de droite sur la Figure 7.11). Cela suppose que des observations de chaque horizon de sol, sur quelques UH du bassin versant pourraient être suffisantes pour corriger l'humidité dans toutes les autres UH caractérisées par le même type de sol. Cette configuration mérite d'être explorée et correspond à l'Expérience 2 décrite dans la Section 6.4 et mise en oeuvre dans le Chapitre 8.

7.3 Impact d'incertitudes sur les forçages climatiques

Expérience	Variables/paramètres à corriger	Observations disponibles	Sources d'incertitudes
Exp. 1	- Humidité de surface et subsurface - *thetas* pour tous les horizons	- Images d'humidité de surface	- Paramètres d'entrée - Conditions initiales
Exp. 1rain	- Humidité de surface et subsurface - *thetas* pour tous les horizons	- Images d'humidité de surface	- Paramètres d'entrée - Conditions initiales - **Forçages climatiques**

Tableau 7.3 – Rappel des caractéristiques des Expérience 1 et 1rain.

Dans l'Expérience 1, seules les valeurs des paramètres d'entrée et les conditions initiales sont considérées comme sources d'incertitudes du modèle. Dans ce paragraphe, l'impact d'incertitudes supplémentaires sur les forçages climatiques, et plus précisément sur la chronique de précipitations est pris en compte dans l'Expérience 1rain. Compte tenu des conclusions des paragraphes précédents concernant notamment l'impact de la taille de l'ensemble, on se concentre ici sur l'utilisation de l'ES-MDA pour l'estimation de l'humidité et des *thetas* de

surface à partir d'un ensemble de M=100 membres et d'observations d'humidité de surface disponibles à la fréquence de 6 jours avec une erreur d'observation σ=0.02 cm^3cm^{-3}. Les caractéristiques des Expériences 1 et 1rain sont rappelées dans le Tableau 7.3.

7.3.1 Définition de l'incertitude liée aux précipitations

L'incertitude liée aux précipitations est en général difficile à caractériser car elle inclut de nombreuses composantes (erreur de mesure liée au fonctionnement de l'appareil, erreur de représentativité lié à son emplacement puis erreur de représentativité liée à l'utilisation d'une même pluie ponctuelle sur tout le bassin...). La base de données BDOH dont sont extraites les chroniques de précipitations utilisées dans ces travaux (Gouy et al., 2015) ne fournit pas d'information sur les erreurs de mesure ou de représentativité des pluies. La quantification précise de ces erreurs n'étant pas l'objectif principal de cette thèse, elle n'est pas abordée plus en détail ici. En première approximation pour cette expérience synthétique, on considère ainsi que la série temporelle d'erreur liée aux précipitations suit une distribution normale non corrélée dans le temps et d'écart-type σ=2 mm. Cela signifie qu'à chaque pas de temps, on échantillonne une nouvelle valeur d'erreur dans la distribution $\mathcal{N}(0, 2)$ et qu'on l'applique à l'ensemble du bassin versant. Cette forme de perturbation n'est à priori pas réaliste mais servira de point de départ pour explorer la prise en compte d'incertitudes sur les forçages dans le système d'assimilation.

De plus, la gestion du temps dans PESHMELBA impose des contraintes supplémentaires d'un point de vue pratique pour la définition de cette erreur. En effet, on rappelle que le modèle permet de raffiner le pas de temps en période humide (voir Figure 2.6). Ce raffinement implique que le nombre de pas de temps total de la simulation peut varier d'un membre à l'autre selon la chronique de pluie considérée (plus de pas de temps pour une chronique caractérisée par beaucoup de précipitations, moins de pas de temps pour une chronique caractérisée par peu de précipitations). Or, pour pouvoir appliquer l'ES-MDA il est indispensable de s'assurer que les vecteurs d'état associés à chaque membre sont de même taille. Une première solution consiste à fixer le même pas de temps en période sèche et en période humide mais ceci se fait au prix d'une perte de résolution si l'on choisit le pas de temps le plus grossier, ou d'un temps de calcul élevé si l'on choisit le pas de temps raffiné. Ainsi, pour ces travaux et dans une première approche simplifiée, on modifie la chronique de pluie associée à chaque membre en s'assurant que les événements pluvieux qu'elle simule ont lieu sur les mêmes pas de temps que dans la simulation de référence.

Les chroniques résultantes sont donc composées d'événements pluvieux et de périodes sèches ayant tous lieu sur les mêmes périodes mais dont les formes et les cumuls peuvent varier comme illustré sur la Figure 7.12. On note cependant que cette forme de perturbations entraîne une surestimation du volume de précipitations cumulé (Figure 7.12, vignette b) dont les conséquences sur les performances du système d'assimilation devront être examinées.

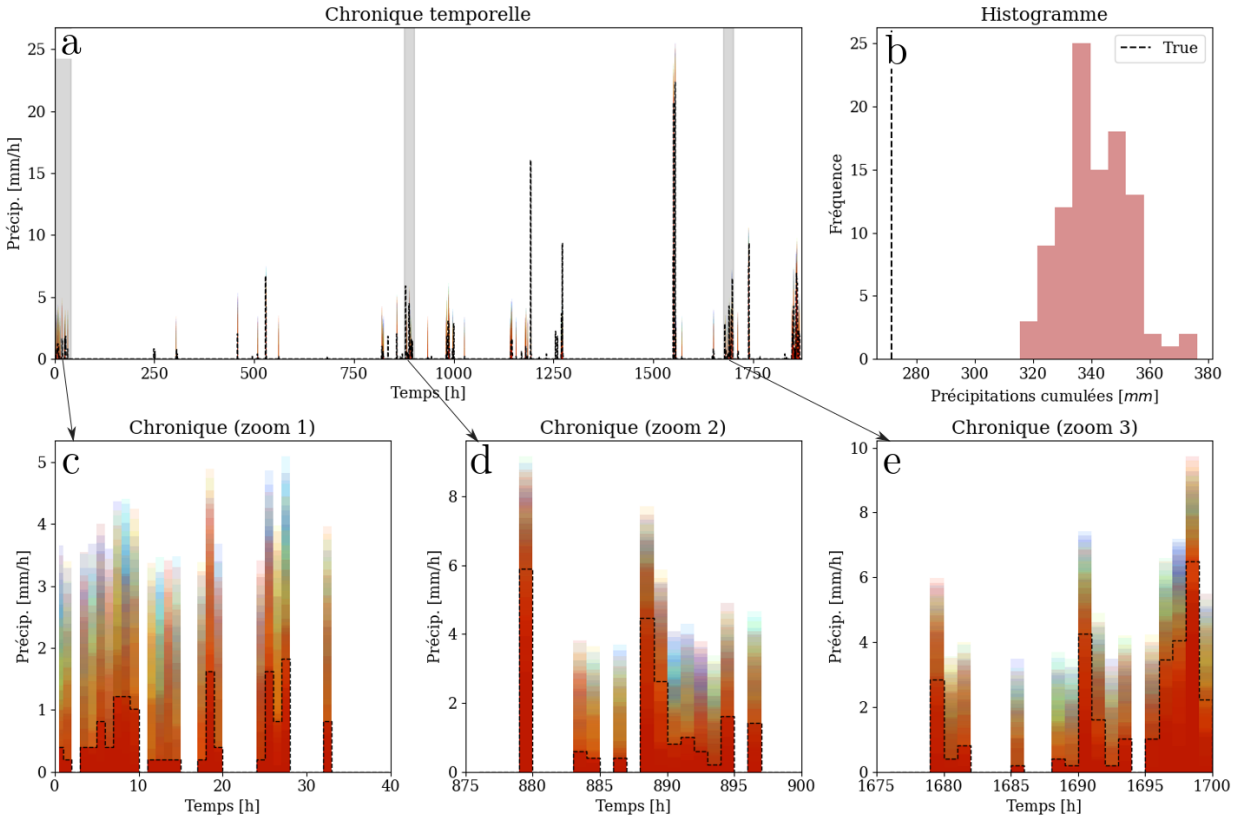


Figure 7.12 – Ensemble de séries temporelles de précipitations utilisées dans l'Expérience 1rain. a) Superposition des séries de précipitations sur l'intégralité de la durée de la simulation. Les bandes grises délimitent les périodes temporelles qui font l'objet d'un zoom sur la ligne du bas (vignettes c, d et e). Dans ces vignettes, chaque couleur représente un membre. b) Histogramme des cumuls de précipitations pour l'ensemble de l'Expérience 1rain. Pour toutes les figures, la ligne noire en pointillés représente la chronique de précipitations de la simulation de référence "True".

7.3.2 Résultats

L'impact de la prise en compte d'incertitudes sur les précipitations est évalué en comparant les CRPS moyennés spatiotemporellement pour l'humidité de **surface** avant (free run équivalent à l'ébauche pour l'ES-MDA) et après assimilation (analyse) pour l'Expérience 1 et 1rain.

Les CRPS sont regroupés sur la Figure 7.13 et ceux-ci fournissent deux résultats principaux développés dans les paragraphes suivants :

Comparaison des CRPS avant assimilation La comparaison des CRPS des free runs des deux expériences montre que l'ajout d'une nouvelle source d'incertitude n'augmente pas l'erreur d'ébauche. Au contraire, l'erreur est sensiblement plus faible lorsque l'ensemble est

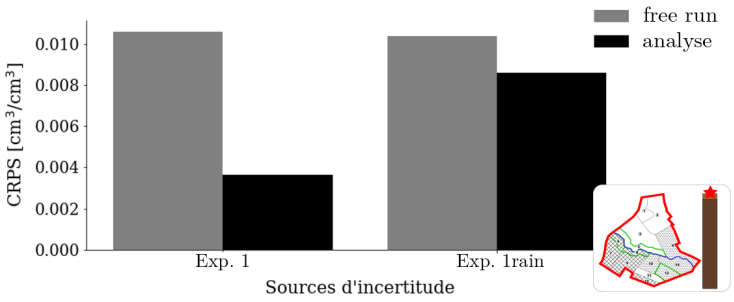


Figure 7.13 – Comparaison des CRPS associés à l'estimation de l'humidité de surface moyennés spatiotemporellement à l'échelle du bassin versant avant assimilation (free run) et après assimilation (analyse) pour l'Expérience 1 et l'Expérience 1rain.

généré en perturbant paramètres, conditions initiales et précipitations attestant probablement de la présence de compensations.
L'analyse des trajectoires des free runs correspondant à chaque expérience ainsi que de la décomposition du CRPS en fiabilité et en CRPS potentiel permettent cependant de nuancer ce résultat. Pour cela, la Figure 7.14 regroupe les trajectoires d'humidité de surface des free runs sur l'UH10 pour les deux expériences ainsi que les moyennes temporelles de CRPS total, de CRPS potentiel et de fiabilité. Les résultats étant similaires sur les autres UH, ils ne sont pas présentés dans cette section.

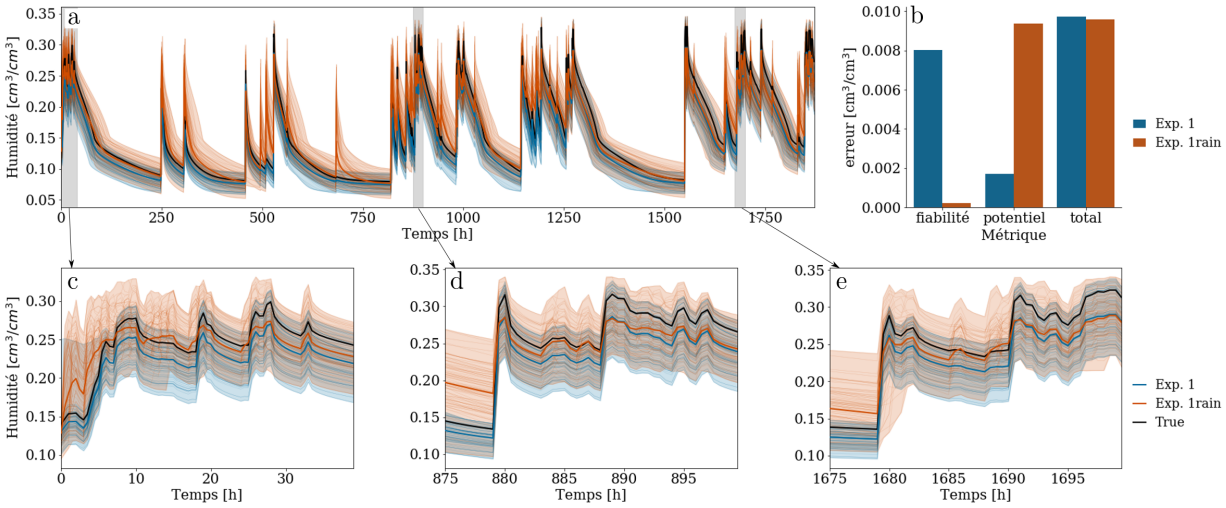


Figure 7.14 – a) Ensemble des trajectoires d'humidité de surface des free runs sur l'UH10 pour l'Expérience 1 (bleu) et l'Expérience 1rain (orange). Les traits épais colorés représentent les trajectoires moyennes de chaque ensemble, le trait noir représente la trajectoire de référence "True" et les bandes grises délimitent les périodes temporelles qui font l'objet d'un zoom sur la ligne du bas (vignettes c, d et e). b) Moyennes temporelles de CRPS total, de CRPS potentiel et de fiabilité pour chaque expérience.

D'une part, les valeurs de CRPS potentiel (directement sensible à la dispersion) ainsi que les faisceaux des ensembles attestent d'une dispersion plus importante dans l'Expérience

1rain que dans l'Expérience 1. D'autre part, les valeurs de fiabilité issues de la décomposition du CRPS montrent que le free run de l'Expérience 1 est nettement moins fiable que celui de l'Expérience 1rain [1]. La comparaison de la trajectoire moyenne avec la référence montre d'ailleurs que le free run de l'Expérience 1 sous-estime systématiquement l'humidité. Ce biais est probablement en partie responsable de la fiabilité limitée de l'ensemble et il est une conséquence directe des biais existants (volontairement ajoutés) dans les distributions des *thetas* utilisées pour générer l'ébauche (voir Figure 7.6 pour une comparaison des distributions des paramètres à priori avec la valeur de référence de ces paramètres). Dans l'Expérience 1rain, les perturbations appliquées aux précipitations surestiment le cumul de précipitations sur la durée de la simulation (Figure 7.12, b) et compensent ainsi probablement un tel biais négatif.

En conclusion, l'ajout d'incertitudes sur les précipitations modifie profondément la structure de l'ensemble (sa fiabilité et sa dispersion). Ce comportement, illustré et analysé sur l'UH10 se généralise ensuite à l'intégralité du bassin versant.
On retiendra également que ce résultat a été obtenu en décomposant le CRPS dont les valeurs totales semblaient au premier abord similaires. Bien qu'il soit souvent plébiscité et utilisé pour l'analyse d'ensembles (Baudin, 2015 ; Devers et al., 2020 ; Xu et al., 2022), le CRPS n'apporte donc qu'une information partielle s'il n'est utilisé que dans sa forme condensée. Plus généralement, pour avoir une évaluation la plus complète possible de la qualité d'un ensemble, il est toujours judicieux de combiner plusieurs métriques, en témoigne la diversité de scores existants, qu'ils soient déterministes (biais, RMSE, dispersion, corrélation, etc.) ou probabilistes (histogramme de rang, score de Brier, CRPS, etc.).

Comparaison des CRPS après assimilation Les CRPS des états analysés montrent que l'intégration de perturbations sur les précipitations aboutit à une correction plus limitée de l'humidité de surface pour l'Expérience 1rain que pour l'Expérience 1.
Comme déjà constaté précédemment, une hypothèse possible pour expliquer une telle dégradation des performances de l'ES-MDA s'appuie sur l'analyse des matrices de corrélation de l'ensemble. La Figure 7.15 présente les matrices de corrélations temporelles des free runs pour l'humidité de surface sur l'UH10. On constate que l'ensemble des trajectoires d'humidité générées dans l'Expérience 1rain présentent beaucoup moins de corrélations temporelles que celles de l'Expérience 1. Or, on rappelle que l'humidité de surface est fortement dépendante des précipitations appliquées. Si l'ensemble des précipitations se caractérise par des corrélations temporelles limitées, comme c'est le cas ici, ceci a pour conséquence moins de corrélations temporelles dans la série d'humidité. Comme pour l'analyse des corrélations verticales, on peut alors supposer que l'absence de corrélations temporelles fortes dans l'Expérience 1rain empêche la propagation des corrections de l'ES-MDA des instants observés vers les instants non observés.

1. on rappelle qu'un ensemble parfaitement fiable a une composante de fiabilité égale à 0

Ainsi, malgré une ébauche moins fiable et moins dispersée, l'assimilation aboutit à de meilleures performances dans l'Expérience 1 que dans l'Expérience 1rain. Ce résultat était d'ailleurs prévisible puisque l'Expérience 1rain consistait à rajouter une source d'incertitude dans le modèle sans lui apporter de correction avec l'assimilation. L'analyse des matrices de corrélation a également permis de montrer que les corrélations temporelles jouent un rôle prépondérant sur les performances de l'assimilation. Ces corrélations temporelles étant fortement liées à la forme de la perturbation appliquée aux précipitations, des expériences supplémentaires, intégrant des perturbations plus réalistes (cumul de précipitation non biaisé, ordre de grandeur affiné, structure d'autocorrélation définie, etc.) devront être menées pour explorer au mieux les capacités du système d'assimilation avant d'aborder une application réaliste.

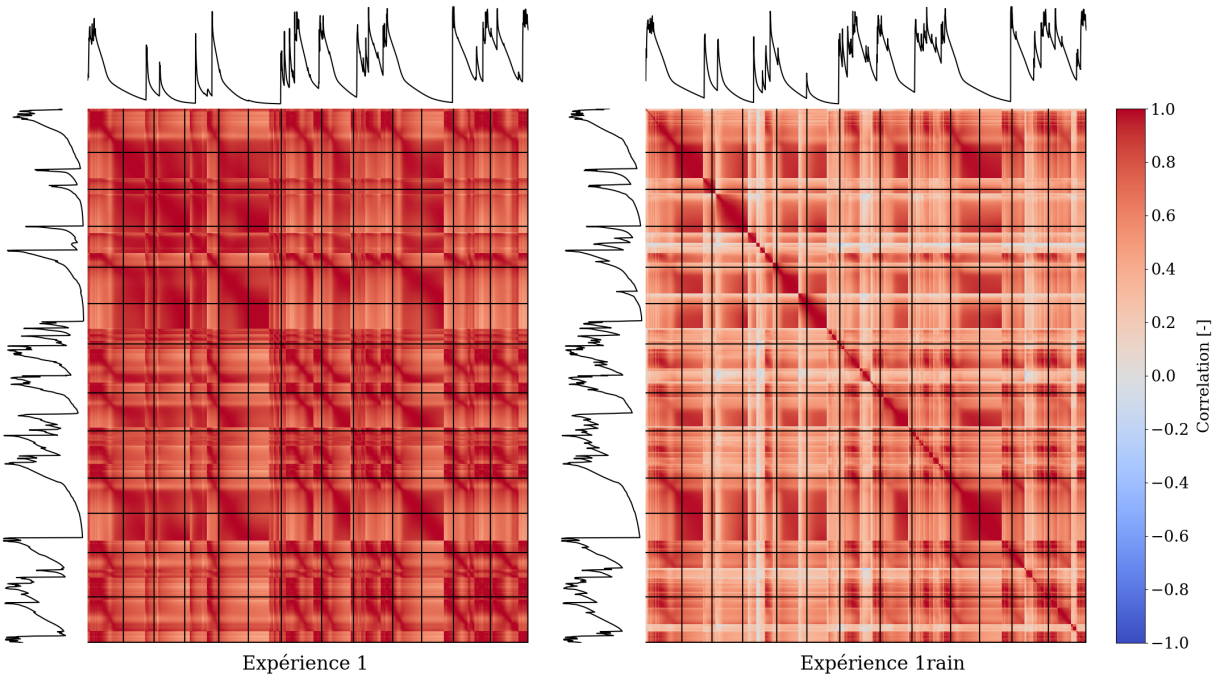


Figure 7.15 – Portions des matrices de corrélation des free runs correspondant à la trajectoire d'humidité de surface sur l'UH10 pour l'Expérience 1 (gauche) et l'Expérience 1rain (droite). Les trajectoires noires en haut et à gauche de chaque matrice représentent les trajectoires moyennes de chaque ensemble et les traits noirs verticaux et horizontaux indiquent les instants auxquels des observations sont disponibles.

A retenir

- ✓ En prenant en compte une incertitude sur les précipitations, on obtient un ensemble plus dispersé mais caractérisé par moins de corrélations temporelles.
- ✓ L'absence de corrélations temporelles marquées altère les performances de l'ES-MDA.
- ✓ Attention à l'interprétation du CRPS ! Sa décomposition est indispensable pour comprendre la structure et les sources de l'erreur qu'il quantifie.

7.4 Conclusion : apports et limites de l'assimilation de données d'humidité de surface

Dans ce chapitre, l'apport de l'assimilation d'images satellite d'humidité de surface dans le modèle PESHMELBA a été évalué. La mise en place d'expériences jumelles sur le scénario virtuel inspiré de la Morcille a permis d'explorer de manière intensive les performances de 3 méthodes d'assimilation ensemblistes : EnKF, ES-MDA et iEnKS. Les objectifs, la méthodologie et les principaux résultats de ce chapitre sont résumés sur la Figure 7.16.

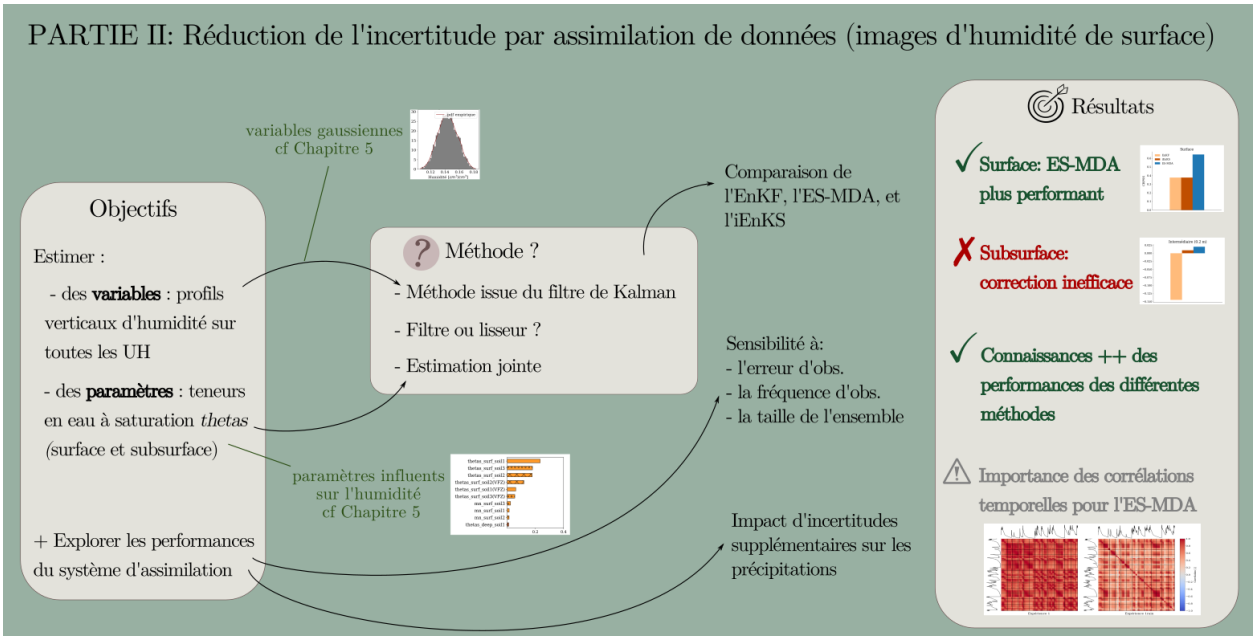


Figure 7.16 – Résumé visuel du Chapitre 7: objectifs de départ, outils utilisés et principaux résultats obtenus pour l'assimilation d'images satellites d'humidité de surface.

Les performances des 3 méthodes pour corriger l'humidité de surface et les teneurs en eau à saturation ont d'abord été comparées sur un scénario nominal. Cette expérience a permis d'identifier l'ES-MDA comme la méthode d'assimilation la plus efficace, notamment car elle permet d'intégrer l'information provenant de toutes les observations à la fois ainsi que la dynamique du système. L'ES-MDA est aussi un très bon compromis pratique puisqu'elle est facile à implémenter et relativement économe en termes de coût de calcul. Toutefois, cette première expérience a aussi montré que si l'ES-MDA (et les autres méthodes) permettent une correction satisfaisante en surface, l'apport de l'assimilation est bien plus limité dans le compartiment de subsurface qui n'est pas observé. Le manque de corrélations verticales entre la surface et la subsurface a notamment été mis en lumière pour expliquer ces performances limitées.

Les expériences jumelles ont également été exploitées pour explorer la sensibilité des différentes méthodes à l'erreur et la fréquence d'observation ainsi qu'à la taille de l'ensemble. Ces expériences ont permis d'évaluer la robustesse des différentes méthodes et ont également

permis de choisir une taille d'ensemble adéquate (100 membres) pour les futures expériences menées avec l'ES-MDA. On a également pu conclure que la fréquence d'observation de 6 jours des images Sentinel est optimale dans cette configuration. Quant à l'erreur d'observation, elle est évidemment très impactante sur les performances de l'ES-MDA. Il sera donc nécessaire de quantifier au plus juste une telle erreur avant d'envisager une application réaliste, et ce, malgré les difficultés que cela représente.
Finalement, des incertitudes sur les précipitations ont été intégrées dans une dernière expérience avec l'objectif de générer un ensemble le plus représentatif possible des erreurs du modèle. Cette expérience a permis de montrer que la prise en compte d'incertitudes supplémentaires permet de générer un ensemble à priori de meilleure qualité mais pour lequel l'ES-MDA peut aboutir à de moins bonnes performances car les corrélations temporelles au sein de l'ensemble sont fortement affectées. Toutefois, on rappelle que les incertitudes sur les précipitations ont été représentées de manière assez rudimentaire dans ce premier test. Ainsi, il serait nécessaire de les caractériser plus physiquement dans de prochaines expériences, ce qui n'est à priori pas évident.

La série d'expériences présentée dans ce chapitre constitue ainsi une première base de connaissances solides sur les apports et les limites de l'assimilation de données dans PESH-MELBA. Jusqu'à présent, seules des images d'humidité de surface ont été utilisées. Pour aller plus loin, le chapitre suivant s'attachera notamment à évaluer l'impact de l'assimilation lorsqu'on considère à la fois des observations de surface et des profils verticaux ponctuels d'humidité.

Chapitre 8

Assimilation multi-sources

Sommaire

Le chapitre précédent a permis de montrer que l'assimilation d'images satellite de surface est efficace pour la correction des variables et paramètres relatifs à l'humidité dans le compartiment de surface. Cependant, il a aussi été montré que l'effet de l'assimilation de données de surface ne se propage pas à la correction des variables et paramètres de subsurface. Dans ce chapitre, on tente de dépasser de telles limitations en intégrant de nouvelles sources de données dans le système d'assimilation. Dans la Section 8.1, images d'humidité de surface et profils verticaux ponctuels d'humidité sont assimilés conjointement. On évalue alors l'intérêt de ces 2 sources de données conjuguées pour la correction de variables et paramètres relatifs à l'humidité dans les compartiments de surface et de subsurface (Expérience 2).
Dans la Section 8.2, on évalue cette fois la possibilité de corriger une variable liée aux pesticides avec l'assimilation. Ainsi, la série temporelle de concentration moyenne journalière à l'exutoire est corrigée, d'abord à partir d'observations d'humidité puis à partir d'observations de concentration dans la rivière (Expérience 3).

Compte tenu des conclusions du Chapitre 7, la méthode d'assimilation utilisée pour mener les Expériences 2 et 3 est l'ES-MDA. Le nombre d'itérations est fixé à 3 et on considère un ensemble de 100 membres généré en perturbant les paramètres d'entrée et les conditions initiales.

8.1 Assimilation multi-sources pour l'estimation de l'humidité

8.1.1 Intégration de profils verticaux d'humidité

Dans cette section, on considère l'assimilation conjointe d'images d'humidité de surface et de profils verticaux d'humidité. On rappelle que ces deux sources de données n'ont ni la même fréquence temporelle ni la même résolution spatiale. Les images satellite fournissent une valeur d'humidité moyenne sur les 5 premiers cm du sol, sur toutes les UH, tous les 6 jours. Les profils verticaux couvrent les 2 premiers m du sol et sont obtenus soit par sonde TDR, soit par mesures d'ERT. Ils sont plus ponctuels dans le temps et l'espace compte tenu de leur coût technique (opérateur sur place, installation des sondes, etc.). Dans cette section, plusieurs configurations différant par le nombre de profils assimilés et leurs positions sont testées. Leurs caractéristiques sont résumées dans le Tableau 8.1. Dans toutes les configurations, les profils verticaux sont supposés disponibles à une fréquence de 19.5 jours (soit 3 profils sur la durée de la simulation). Pour aider à l'interprétation des résultats, la Figure 8.1 rappelle la répartition et la composition des UCS utilisés dans le scénario virtuel ainsi que les portions de colonne de sol observées avec les images radar et les profils verticaux d'humidité.

Expérience	Données assimilées
Expérience 1	Images radar
Expérience 2single/Profils	Profils verticaux sur l'UH9 (parcelle)
Expérience 2single/Multi-sources	Images radar + Profils verticaux sur l'UH9 (parcelle)
Expérience 2vineyard/Profils	Profils verticaux sur les UH8, 13 et 5 (parcelles)
Expérience 2vineyard/Multi-sources	Images radar + Profils verticaux sur les UH8, 13 et 5 (parcelles)
Expérience 2VFS/Profils	Profils verticaux sur les UH9, 3 et 14 (VFS)
Expérience 2VFS/Multi-sources	Images radar + Profils verticaux sur les UH9, 3 et 14 (VFS)

Tableau 8.1 – Description des expériences réalisées pour l'assimilation conjointe d'images radar et de profils verticaux d'humidité. L'abbréviation VFS désigne les bandes enherbées (*Vegetative Filter Strips*).

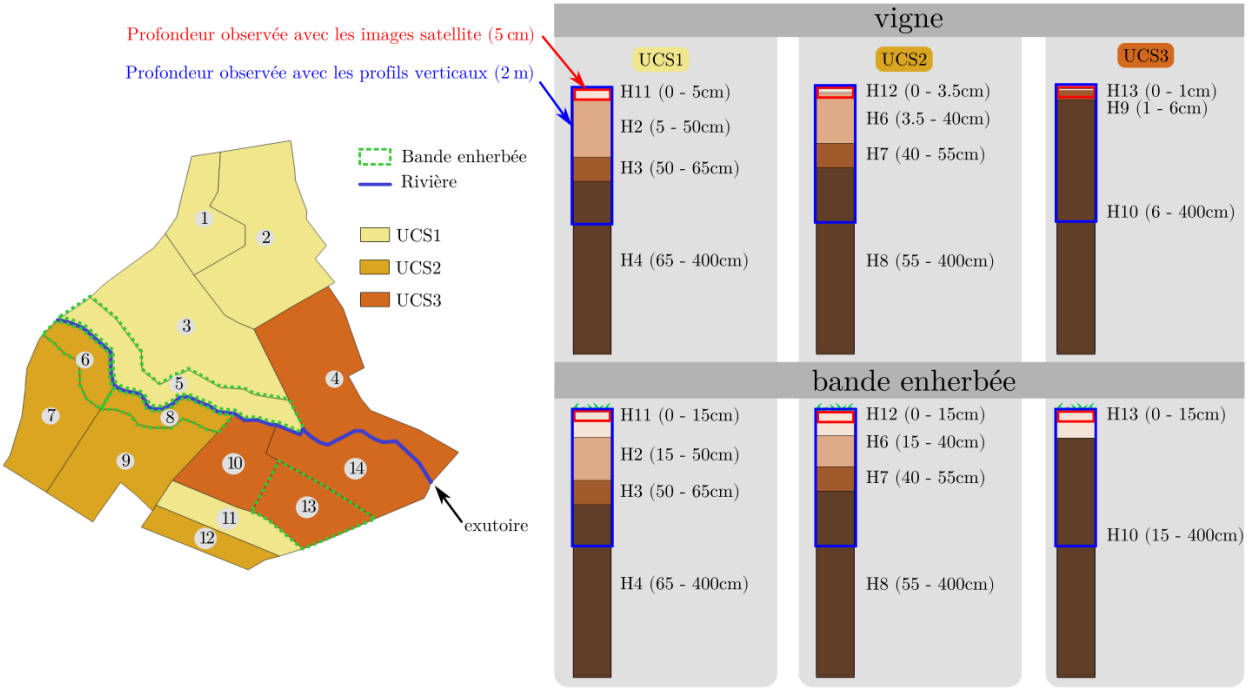


Figure 8.1 – Rappel de la distribution et de la composition des UCS sur le bassin d'étude. Les rectangles rouges et bleus sur les colonnes de sol indiquent les portions qui sont observées par les images satellite et les profils verticaux d'humidité.

Profils sur une unique UH

Dans un premier temps, on considère disponibles des profils verticaux d'humidité sur une unique UH du bassin (UH9), avec une fréquence temporelle de 19.5 jours. On compare l'Expérience 1 où seules les images radar sont assimilées avec l'Expérience 2single/Profils où seuls les profils verticaux sont assimilés et l'Expérience 2single/Multi où les 2 sources d'observations sont assimilées conjointement. La Figure 8.2 regroupe les cartes de CRPSS relatifs à l'estimation de l'humidité moyennés temporellement pour les Expériences 1 et 2single et le Tableau 8.2 regroupe les valeurs de CRPSS correspondant à l'estimation des *thetas*.

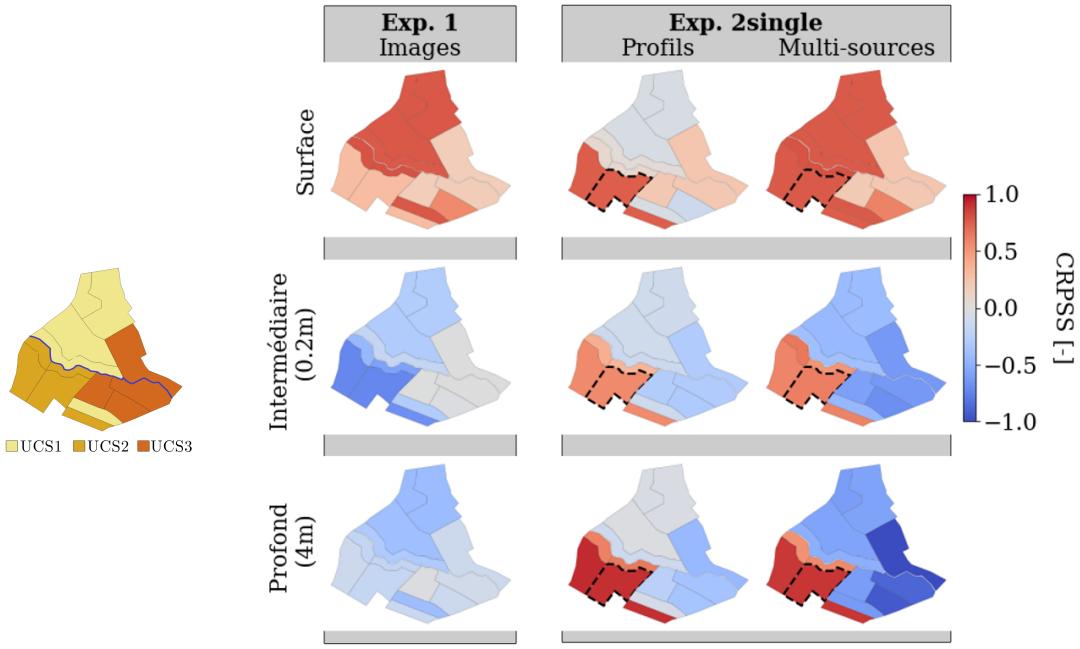


Figure 8.2 – Cartes de CRPSS relatifs à l'estimation de l'humidité obtenus pour les Expériences 1 et 2single. L'UH entourée de pointillés (UH9) indique la position des profils verticaux d'humidité disponibles.

Expérience 1 Dans l'Expérience 1, on retrouve les conclusions du Chapitre 7. En **surface**, les valeurs de CRPSS positives de la Figure 8.3 indiquent que l'humidité est très bien estimée par l'assimilation. Les performances sont homogènes pour toutes les UH d'un même type de sol, en accord avec les fortes corrélations entre elles. On note cependant des valeurs différentes d'un type de sol à l'autre, possiblement dûes à la composition du sol sur les 5 cm observés. En effet, les meilleures performances sont obtenues pour le sol 1 pour lequel le premier horizon s'étend sur 5 cm et est donc entièrement observé alors que le sol 3 caractérisé par un horizon de surface très fin (1 cm) aboutit à l'amélioration la plus limitée. Dans ce dernier cas, l'information contenue dans l'humidité moyenne sur les 5 premiers cm ne représente donc que très peu l'horizon superficiel.

	Horizon	Exp. 1 Images	Exp. 2single Profils	Multi
Surface	11	**0.81**	-0.08	0.75
	12	0.31	**0.76**	0.70
	13	0.05	**0.33**	0.10
	14	**0.85**	0.03	**0.85**
	15	**0.87**	0.07	0.86
	16	0.79	-0.23	**0.85**
Subsurface	2	-0.84	**-0.02**	-0.68
	3	**0.03**	-0.13	-0.07
	4	-0.31	**-0.09**	-0.48
	6	-0.66	0.47	**0.67**
	7	-0.06	0.67	**0.73**
	8	-0.10	**0.61**	0.53
	9	0.55	-0.31	**0.65**
	10	**-0.21**	-0.23	-0.75

Tableau 8.2 – Valeurs de CRPSS associées à l'estimation des *thetas* pour les Expériences 1 et 2single. Pour chaque ligne, la valeur en gras indique la meilleure estimation.

En **subsurface**, l'effet de l'assimilation est très faible, il dégrade même l'estimation de l'humidité sur les UH de l'UCS 1 et 2.

Les conclusions sont similaires pour l'estimation des paramètres (Tableau 8.2). Les paramètres de **surface** des parcelles de vigne (horizons 11, 12 et 13) aboutissent à des CRPSS positifs mais celui du sol 1 (horizon 11) est le mieux estimé alors que celui du sol 3 (horizon 13) n'est quasiment pas corrigé. Pour les bandes enherbées, l'estimation des *thetas* des horizons 14, 15 et 16 est plus homogène car dans tous les cas, le premier horizon s'étend sur les 15 premiers cm de sol et est donc entièrement observé.
En **subsurface**, les CRPSS sont négatifs montrant que l'assimilation dégrade l'estimation des paramètres. La seule exception est le *thetas* du second horizon du sol 3 (horizon 9) qui obtient un CRPSS de 0.55 car il compose la majorité des 5 premiers cm de sol qui sont observés.

Expérience 2single/Profils Dans l'Expérience 2single/Profils, l'assimilation de profils verticaux sur l'UH9 3 fois pendant la simulation permet une correction significative de l'humidité à toutes les profondeurs, sur toutes les parcelles de vigne appartenant au même UCS (UCS2). Sur les UH appartenant à d'autres UCS, l'estimation de l'humidité n'est que très peu améliorée. L'absence de corrélations entre les différents UCS mise en lumière au chapitre précédent (voir la matrice de corrélation de la Figure 7.11 pour un rappel) explique de tels contrastes de performances. En effet, les corrections sur une UCS observée ne peuvent pas se propager aux UCS non observées. Sur les UCS non observées, l'assimilation dégrade même

significativement l'estimation de l'humidité dans certains cas, notamment en subsurface. On peut supposer ici que des erreurs d'échantillonnage liées à la taille de l'ensemble génèrent des corrélations douteuses qui connectent de manière artificielle les différents types de sol. Dans ce type de situation, la localisation peut potentiellement être un outil pertinent.
On note également qu'en **surface**, les UH 6 et 8 ne bénéficient pas des corrections de l'assimilation bien qu'elles appartiennent aussi à l'UCS2. Or, ces UH sont des bandes enherbées et la différence de dynamique entre vigne et bande enherbée implique une absence de corrélations fortes entre leurs horizons de surface et explique un tel effet limité. Par contre, les horizons de **subsurface** des parcelles de vigne et des bandes enherbées étant les mêmes, toutes les UH de l'UCS2, vignes ou bandes enherbées, sont bien corrigées à la profondeur intermédiaire (20 cm) et la plus profonde (4 m).

Expérience 2single/Multi Dans l'Expérience 2single/Multi, l'humidité de **surface** est mieux estimée sur tout le bassin versant et l'assimilation des deux sources de données permet une correction plus efficace qu'avec chaque source de données prise séparément.
En **subsurface**, l'effet positif des profils verticaux sur l'UH9 est maintenu sur les UH de l'UCS2 mais l'intégration des images de surface dégrade l'estimation de l'humidité sur les UH des autres types de sol par rapport aux expériences précédentes.

Pour l'estimation des paramètres (Tableau 8.2), que ce soit dans l'Expérience 2single/Profils ou l'Expérience 2single/Multi, l'assimilation de profils verticaux sur une UH de l'UCS2 permet d'améliorer l'estimation des *thetas* de cet UCS (horizons 12, 6, 7 et 8) mais n'a que très peu d'effet, voire dégrade l'estimation des *thetas* des horizons de subsurface pour les autres UCS. Là encore, l'absence de corrélation (voire l'apparition de corrélations douteuses) entre les UCS explique ces contre performances.

En conclusion, l'intégration de profils verticaux d'humidité est prometteuse pour pallier les limites de l'assimilation d'images d'humidité de surface en permettant de corriger l'humidité de subsurface et les *thetas* des horizons associés. Cependant, l'impact de tels profils est significatif seulement sur les UH caractérisées par le même UCS que l'UH observée. Pour garantir une correction homogène des variables/paramètres à l'échelle du bassin versant, il faut donc envisager d'intégrer à minima un profil par UCS. C'est la configuration qui est testée dans le paragraphe suivant.

Profils sur une UH par UCS

On considère maintenant disponibles des profils verticaux d'humidité sur 3 parcelles de vigne (Expérience 2vineyard) ou 3 bandes enherbées (Expérience 2VFS) du bassin, répartis sur chaque UCS. La Figure 8.3 regroupe les cartes de CRPSS relatif à l'estimation de l'humidité moyennés temporellement pour les Expériences 1, 2vineyard et 2VFS et le Tableau 8.3 regroupe les valeurs de CRPSS correspondant à l'estimation des *thetas*.

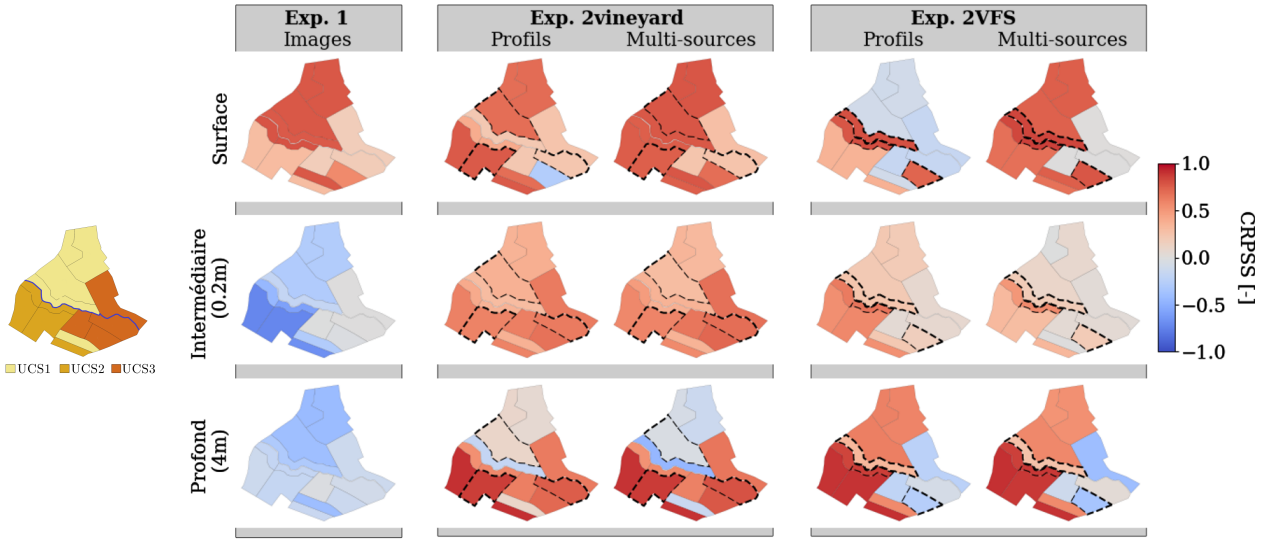


Figure 8.3 – Cartes de CRPSS relatifs à l'estimation de l'humidité obtenus pour les différentes expériences impliquant 3 profils d'humidité sur des parcelles de vigne (Exp. 2vineyard) ou des bandes enherbées (Exp. 2VFS). Les UH entourées de pointillés indiquent la position des profils verticaux d'humidité disponibles pour chaque expérience.

Expérience 2vineyard/Profils et Expérience 2VFS/Profils En assimilant uniquement des profils verticaux sur des parcelles de vigne (Expérience 2vineyard/Profils), les valeurs de CRPSS sont positives à la fois en surface (Figure 8.3, ligne 1) et à 20 cm de profondeur (Figure 8.3, ligne 2). Bien que des observations soient seulement disponibles sur 3 parcelles de vigne, cela suffit à corriger toutes les autres parcelles, là encore grâce aux corrélations spatiales existants entre UH du même UCS. On note cependant qu'en **surface**, les UH de l'UCS 3 (orange foncé sur la carte de la Figure 8.1) sont globalement moins bien corrigées que celles des autres UCS, du fait de la finesse de l'horizon de surface. D'autre part, comme dans l'Expérience 2single, l'humidité est moins bien corrigée (voire dégradée pour l'UH13) sur les bandes enherbées malgré la présence d'observations sur des parcelles de vignes voisines. Cet effet est encore plus visible sur l'Expérience 2VFS/Profils qui concentre les observations sur les bandes enherbées. Dans cette configuration, l'humidité de surface est bien corrigée localement sur les UH observées mais les corrections ne se propagent que très peu au reste du bassin par manque de corrélation entre bandes enherbées et vignes.
Par contre, à **20 cm de profondeur** les valeurs de CRPSS restent majoritairement positives puisque en dessous de 15 cm, parcelles de vigne et bandes enherbées ont la même composition et sont donc à nouveau corrélées.
En **fond de profil**, l'impact de l'assimilation de profils est plus aléatoire dans les 2 expériences. Dans l'Expérience 2VFS/Profils, on obtient même une valeur négative de CRPSS sur l'UH13 alors que celle-ci est observée. Les observations verticales d'humidité ne s'étendant que sur les 2 premiers mètres de sol, il est probable que la correction n'arrive pas à se propager jusqu'au fond de la colonne de sol.

		Exp. 1	Exp. 2vineyard		Exp. 2VFS	
	Horizon	Images	Profils	Multi	Profils	Multi
Surface	11	**0.81**	0.00	0.19	-0.09	0.77
	12	0.31	**0.94**	0.82	0.34	0.66
	13	**0.05**	-0.78	-0.62	-0.13	-0.28
	14	0.85	0.80	0.61	0.87	**0.92**
	15	0.87	0.38	0.70	0.85	**0.91**
	16	0.79	-2.01	-0.75	**0.82**	0.81
Subsurface	2	-0.84	-2.25	**0.46**	0.31	0.03
	3	0.03	-2.23	-0.23	0.13	**0.50**
	4	-0.31	-0.46	-0.44	**0.37**	0.25
	6	-0.66	0.25	**0.70**	0.65	0.51
	7	-0.06	-0.75	-0.97	0.70	**0.71**
	8	-0.10	0.91	**0.93**	0.86	0.90
	9	0.55	0.40	**0.96**	0.21	0.67
	10	**-0.21**	-0.54	-0.35	-0.37	-0.61

Tableau 8.3 – Valeurs de CRPSS associées à l'estimation des *thetas* pour les différentes expériences impliquant 3 profils d'humidité sur des parcelles de vigne (Exp. 2vineyard) ou des bandes enherbées (Exp. 2VFS). Pour chaque ligne, la valeur en gras indique la meilleure estimation.

Pour l'estimation des paramètres (Tableau 8.3), l'Expérience 2vineyard/Multi aboutit globalement à de bonnes performances en surface et aux meilleures performances en subsurface ce qui est logique puisque cette configuration conjugue les 2 types de données. Les conclusions sont toutefois contrastées d'un horizon à l'autre attestant probablement de la présence de compensations d'erreurs, de corrélations douteuses ou d'une défaillance de l'ensemble. Pour tirer des conclusions robustes, il serait nécessaire de reproduire ces expériences, peut-être avec un ensemble plus grand et une configuration plus simple à analyser dans un premier temps.

Expérience 2vineyard/Multi et Expérience 2VFS/Multi En assimilant conjointement les images radar et les profils verticaux, l'estimation de l'humidité en **surface** est largement améliorée (Figure 8.3). A nouveau, l'assimilation des deux sources de données permet une correction plus efficace qu'avec chaque source de données prise séparément. Comme déjà exposé précédemment, la composition verticale de l'UCS 3 implique que les UH appartenant à ce sol sont moins bien corrigées que celles des UCS 1 et 2.
A **20 cm de profondeur**, l'estimation de l'humidité est améliorée de manière assez homogène par les Expériences 2vineyard/Multi et 2VFS/Multi. A cette profondeur, l'Expérience 2vineyard/Multi aboutit à de meilleurs performances que l'Expérience 2VFS/Multi.
Finalement, en **fond de profil**, l'assimilation conjointe des deux types d'observation aboutit à une amélioration significative mais hétérogène d'un type de sol à l'autre de l'estimation de l'humidité.

Pour l'estimation des paramètres (Tableau 8.3), l'Expérience 2vineyard/Multi aboutit aux valeurs de CRPSS les plus élevées en subsurface et les valeurs en surface restent globalement satisfaisantes bien qu'inférieures à celles obtenus en assimilant seulement les images de surface.

Conclusion

L'intégration de profils ponctuels dans le système d'assimilation permet de pallier les limites des images satellite pour l'estimation de l'humidité et des *thetas* en surface et en subsurface. Cependant, une correction satisfaisante de tous les horizons nécessite de disposer au minimum d'un point d'observation par type de sol. D'autre part, les apports de l'assimilation sont plus significatifs si ces derniers sont disposés sur des parcelles de vigne plutôt que sur des bandes enherbées.
Ces expériences fournissent ainsi des éléments précieux pour la mise en place d'une stratégie de collecte de données lors du passage à une application réaliste sur la Morcille. On note toutefois que la présence de structures métalliques dans les parcelles de vigne peut perturber l'acquisition de données ERT. Pour une telle application, il faudra ainsi plutôt envisager de se tourner vers une sonde TDR ou accepter des performances limitées en échantillonnant sur une bande enherbée.

A retenir

- ✓ L'intégration de profils verticaux ponctuels d'humidité permet d'améliorer l'estimation des variables/paramètres relatifs à l'humidité en subsurface.
- ✓ Les performances des différentes expériences sont hétérogènes d'un type de sol à l'autre en fonction de leur composition verticale.
- ✓ La configuration aboutissant globalement aux meilleurs corrections consiste à assimiler à la fois des images d'humidité de surface et des profils verticaux d'humidité sur des parcelles de vigne appartenant aux différents types de sol présents sur le bassin.

8.1.2 Apport de la localisation

Motivation et mise en oeuvre

Malgré la taille limitée du scénario considéré dans cette étude, la dimension du vecteur d'état augmente rapidement puisqu'il intègre les trajectoires temporelles d'humidité sur toutes les UH, à toutes les profondeurs. Si la taille résultante est gérable numériquement dans cette application ($n = 899864$), elle pourrait devenir rapidement un problème lors d'applications à plus grande échelle (par exemple, il faudra compter environ 500 UH à modéliser sur le bassin versant de la Morcille). Au delà de la manipulation de matrices de grandes tailles qui peut être problématique, les performances de l'ES-MDA peuvent être grandement

affectées si l'ensemble est de taille très inférieure à celle du vecteur d'état ($M << n$). Les résultats de la section précédente ont ainsi montré que l'assimilation d'observations sur un type de sol est susceptible de dégrader l'estimation de l'humidité sur les autres types de sol s'ils sont non observés, et ce, probablement dû à l'apparition de corrélations douteuses que la localisation permettra de limiter.

Dans cette section, pour anticiper ces difficultés lors du passage à des applications à plus grande échelle, on explore l'impact de la localisation pour alléger le coût numérique de l'analyse et atténuer l'impact des corrélations douteuses. Compte tenu de la structure des corrélations spatiales déjà existante entre types de sol, la localisation par domaines est naturellement utilisée : une analyse par type de sol est réalisée en utilisant pour chacun seulement les observations disponibles sur les UH de ce type de sol. De cette façon, les corrélations entre portions du bassin appartenant à des UCS différents sont intégralement supprimées et la taille du vecteur d'état est grossièrement divisée par 3. Les analyses sont ensuite réalisées en parallèle pour limiter le temps de calcul.
Dans un premier temps, on compare les résultats de l'Expériences 2single/Multi impliquant d'assimiler les images satellite d'humidité de surface et les profils disponibles uniquement sur l'UH9, sans et avec localisation (Figure 8.4, gauche). On compare ensuite les résultats de l'Expérience 2vineyard/Multi intégrant les images d'humidité de surface et des profils verticaux d'humidité sur les parcelles de vigne 9, 3 et 14, sans et avec localisation (Figure 8.4, droite).

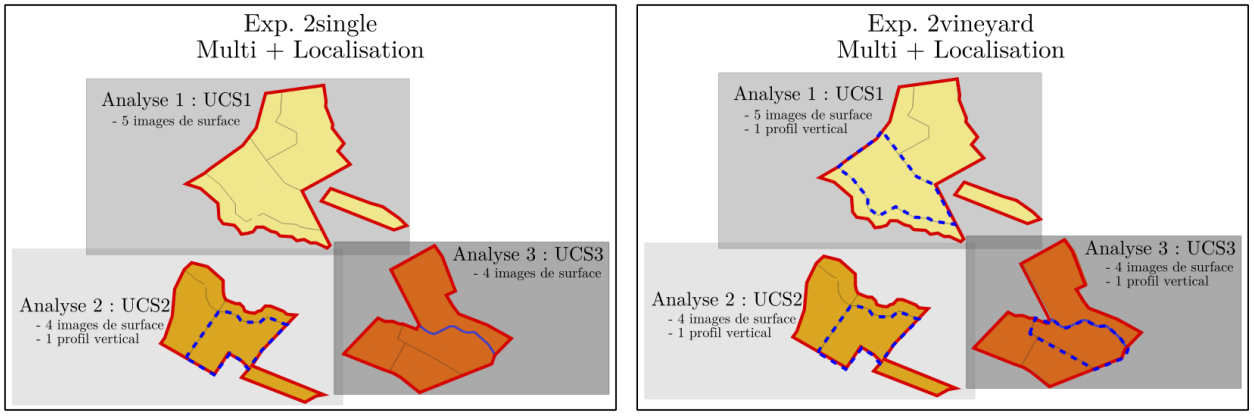


Figure 8.4 – Découpage de l'analyse mis en oeuvre avec la localisation par domaines dans les expériences 2single/Multi et 2vineyard/Multi. Les UH entourées en bleu sont celles où l'on dispose de profils verticaux.

Résultats

Tout d'abord, concernant les temps de calcul, une analyse localisée est réalisée en 26 sCPU par domaine ($3 \times 26 = 78$ sCPU au total) contre 46 sCPU pour une analyse non localisée. La localisation permet de gagner en efficacité en termes de temps apparent mais l'apport n'est

probablement pas le plus flagrant à cette petite échelle. Son intérêt sera probablement plus marqué en passant à une application à une plus grande échelle.
La Figure 8.5 regroupe ensuite les cartes de CRPSS pour les expériences 2single/Multi et 2vineyard/Multi menées sans, puis avec localisation.

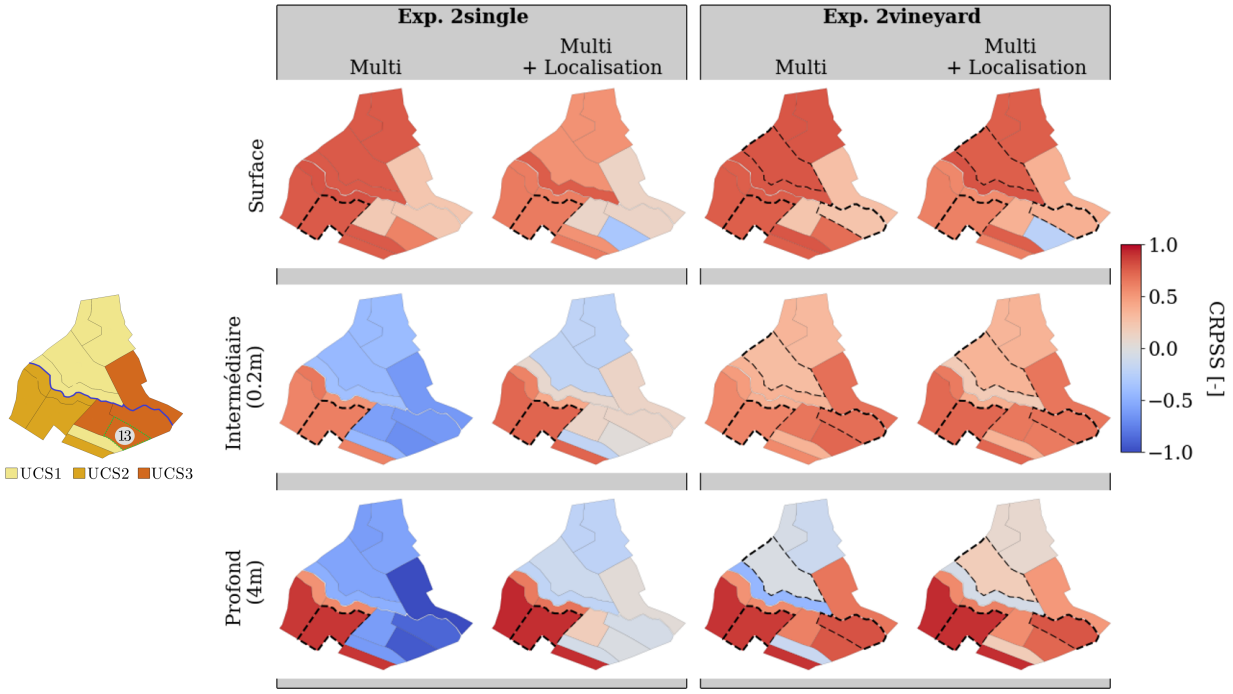


Figure 8.5 – Cartes de CRPSS relatifs à l'estimation de l'humidité obtenues pour les expériences 2single/Multi et 2vineyard/Multi avec et sans localisation. Les UH entourées de pointillés indiquent la position des profils verticaux d'humidité disponibles pour chaque expérience.

Pour les 2 expériences, la localisation a un impact légèrement négatif sur l'estimation de l'humidité de **surface** et ce, de manière plus ou moins marquée selon l'UCS. La bande enherbée 13 est particulièrement impactée par la localisation puisque l'estimation de l'humidité qui était améliorée dans les 2 expériences sans localisation est dégradée lorsque l'analyse est localisée. Compte tenu de la dégradation généralisée de l'analyse lorsqu'elle est localisée, on peut suspecter l'existence de corrélations réelles en surface qui connectent les différents domaines que la localisation sépare ensuite. Ces corrélations pourraient résulter des flux latéraux de ruissellement qui traversent les versants et lient des UH spatialement éloignées. Pour le cas particulier de l'UH13, on peut par exemple imaginer que le rôle d'interception du ruissellement de cette bande enherbée la connecte de manière marquée à d'autres UH plus en amont.
En **subsurface**, la localisation à un effet plus positif. Dans l'Expérience 2single, aux 2 profondeurs étudiées, la localisation permet de limiter fortement la dégradation de l'estimation de l'humidité dans les UCS non observées. L'humidité est même améliorée dans certaines UH qui étaient nettement dégradées sans localisation. Dans l'Expérience 2vineyard, l'estimation de l'humidité, déjà très satisfaisante à 20 cm est sensiblement améliorée, notamment dans

	Horizon	Exp. 2single		Exp. 2vineyard	
		Multi	Multi + Localisation	Multi	Multi + Localisation
Surface	11	0.75	**0.90**	0.75	0.89
	12	0.70	0.44	**0.72**	0.37
	13	0.10	0.22	0.15	**0.28**
	14	0.85	0.86	0.84	**0.87**
	15	0.86	**0.89**	0.84	0.88
	16	0.85	**0.86**	0.85	0.85
Subsurface	2	-0.68	-0.59	**0.33**	0.31
	3	-0.07	0.01	**0.36**	-0.33
	4	-0.48	-0.19	-0.38	**-0.04**
	6	0.67	**0.73**	0.68	0.66
	7	**0.73**	0.67	0.71	0.61
	8	0.53	**0.61**	0.51	0.58
	9	**0.65**	0.56	0.41	0.40
	10	-0.75	-0.09	**0.80**	0.79

Tableau 8.4 – Valeurs de CRPSS associées à l'estimation des *thetas* pour les expériences 2single/-Multi et 2vineyard/Multi sans et avec localisation. Pour chaque ligne, la valeur en gras indique la meilleure estimation.

l'UCS 2. A 4 m, alors que l'assimilation avait un effet négatif sur les UH de l'UCS 3, l'effet devient positif en localisant l'analyse. Ainsi, en subsurface, la localisation a un effet positif et il semble donc que les transferts latéraux dans ce compartiment ont un rôle moins marqué qu'en surface pour connecter les UH de différents types de sol.

Les conclusions sont assez contrastées pour l'estimation des teneurs en eau à saturation. Le Tableau 8.4 réunit les valeurs de CRPSS associées à l'estimation des *thetas* pour les expériences 2single/Multi et 2vineyard/Multi, sans et avec localisation. En surface, et contrairement aux résultats obtenus pour l'estimation de l'humidité, la localisation améliore légèrement l'estimation des *thetas* pour les 2 expériences. En subsurface, la localisation dégrade l'estimation des paramètres alors qu'elle améliorait l'estimation de l'humidité. Ces résultats restent donc difficiles à interpréter et l'on peut supposer qu'ils résultent de l'effet (éventuellement cumulé) de compensations d'erreurs, d'erreurs d'échantillonnage et d'une structure de corrélation non adaptée. Là encore, il serait utile de reproduire ces tests et de les prolonger sur des configurations comprenant moins d'hétérogénités spatiales pour qu'elles soient plus simples à analyser.

Conclusion

En conclusion, les résultats montrent que la localisation a un impact assez limité sur les performances du système d'assmilation. L'effet n'est pas le même sur l'estimation de l'humidité et des *thetas* puisqu'il est plutôt bénéfique sur les premiers et plutôt négatif sur les

seconds. D'un point de vue technique, la localisation permet de diminuer la taille du problème et d'alléger le temps de calcul puisque les analyses sont réalisées en parallèle. Sur ce point, l'intérêt de la localisation n'est pas flagrant dans ce cas d'étude de taille limitée mais il peut être beaucoup plus net, voire essentiel lors d'une application à plus grande échelle.
Toutefois, il est important de rappeler que la quasi indépendance des 3 domaines spatiaux que les matrices de corrélations et la comparaison des résultats avec et sans localisation ont mis en lumière n'est pas forcément cohérente avec les processus physiques régissant les transferts d'eau à l'échelle du bassin ni avec la construction de PESHMELBA. En effet, à l'échelle du bassin versant, les transferts latéraux peuvent avoir un rôle prédominant et ce sont justement ces transferts latéraux que PESHMELBA s'attache à reproduire. Si dans ce cas d'étude, la dynamique verticale semble prépondérante, il n'en sera probablement pas de même avec d'autres forçages climatiques ou à une autre échelle spatiale et/ou temporelle. Dans ce cas, localiser par type de sol ne sera peut-être plus pertinent, voire contre-productif, et il sera probablement intéressant d'explorer d'autres manières de fractionner le bassin (par exemple par versants). On peut aussi envisager d'utiliser plutôt une matrice de localisation permettant de localiser en fonction de la distance comme cela est fait classiquement, voire de localiser en fonction des directions des flux d'eau.

A retenir

- ✓ La localisation par domaines améliore légèrement la correction de l'humidité et dégrade légèrement l'estimation des *thetas*.
- ✓ La localisation par domaines peut être intéressante pour réduire le coût de calcul sur des applications à plus grande échelle.
- ✓ Dans une application à l'échelle d'un bassin versant réel, diviser l'analyse par types de sol ne sera peut être pas la solution la plus adaptée.

8.2 Assimilation multi-sources pour l'estimation de la concentration

Pour disposer d'un signal de concentration journalière à l'exutoire permettant d'évaluer l'impact de l'assimilation sur la concentration, on rappelle que le scénario nominal a été modifié. Alors que toutes les expériences jumelles précédentes ont été réalisées sur le scénario climatique estival, on considère ici le scénario climatique hybride dont on rappelle les caractéristiques :

- Mêmes paramètres d'entrée que pour les scénarios précédents ;
- Application de tebuconazole sur **toutes** les parcelles de vigne du bassin à t=1 h et t=1104 h (=46 jours) ;
- Conditions initiales (niveaux de nappes) identiques au scénario hivernal ;

- Chronique de pluie virtuelle, caractérisée par des intensités 10 fois supérieures au scénario hivernal.

L'impact de l'assimilation sur les variables relatives aux pesticides est exploré à partir de ce scénario hybride. L'objectif de cette dernière partie est de corriger la série temporelle de concentration journalière à l'exutoire ainsi que les paramètres d'entrée influents sur cette dernière (rugosité de Manning sur les parcelles de vigne, teneur en eau à saturation *thetas* et paramètre de forme de Van Genuchten *mn* de l'horizon 10 d'après les résultats du Chapitre 5) en plus de l'humidité sur toutes les UH et à toutes les profondeurs et des teneurs en eau à saturation. Dans un premier temps, on suppose disposer pour cela d'observations de concentration moyenne à l'exutoire obtenues par échantillonneurs passifs avec une durée d'exposition de 7 jours (voir Section 2.3 pour un rappel des caractéristiques des observations de concentration).
Pour cela, deux expériences successives sont réalisées (voir Tableau 8.5). Dans l'Expérience 3hydro (Section 8.2.1), on évalue d'abord la possibilité de corriger la concentration à l'exutoire seulement à partir d'images radar et de profils verticaux d'humidité sur les UH9, 3 et 14 (configuration aboutissant aux meilleures performances d'assimilation multi-sources pour l'humidité d'après la section précédente). Dans ce cas, la concentration fait partie du vecteur mais puisqu'elle n'est pas observée, elle ne peut être corrigée qu'au travers ses corrélations spatiales et temporelles avec l'humidité et une meilleure estimation des paramètres hydrodynamiques du système (*thetas*). On évalue ainsi par ce biais l'impact de l'assimilation d'observations accessibles et peu coûteuses pour corriger/estimer des variables et paramètres peu accessibles sur le terrain. Cette expérience illustre particulièrement le principe de *strongly-coupled assimilation* qui consiste à corriger l'état d'un ou plusieurs composants d'un modèle couplé en assimilant des observations sur d'autres compartiments (Penny et Hamill, 2017). L'Expérience 3hydro est également l'occasion d'évaluer les performances du système d'assimilation pour l'estimation de l'humidité et des paramètres *thetas* sur un autre scénario climatique.
Dans l'Expérience 3pest (Section 8.2.2), des observations de concentration moyenne à l'exutoire sont assimilées en plus des images radar et des profils d'humidité. Dans cette configuration, on cherche à corriger les profils d'humidité sur toutes les UH, la concentration journalière à l'exutoire ainsi que les paramètres *thetas* de tous les horizons, et les paramètres $Manning\_1$ (vigne) et mn_10.

Expérience	Variables de sortie / Paramètres d'entrée estimés	Observations assimilées
Expérience 3hydro	➢ Variables de sortie : - Humidité sur toutes les UH, à toutes les profondeurs - Concentration moyenne journalière à l'exutoire ➢ Paramètres d'entrée : - *thetas* pour tous les horizons	- Images d'humidité de surface - Profils verticaux d'humidité sur les UH9, 3 et 14
Expérience 3pest	➢ Variables de sortie : - Humidité sur toutes les UH, à toutes les profondeurs - Concentration moyenne journalière à l'exutoire ➢ Paramètres d'entrée : - *thetas* pour tous les horizons - Rugosité de Manning de la vigne et paramètre de VG *mn* dans l'horizon 10	- Images d'humidité de surface - Profils verticaux d'humidité sur les UH9, 3 et 14 - Concentration moyenne hebdomadaire à l'exutoire

Tableau 8.5 – Description des expériences réalisées pour l'estimation de la concentration à l'exutoire.

8.2.1 Propagation des corrections du compartiment hydrologique sur les variables de qualité de l'eau

Les paragraphes suivants regroupent les résultats de l'Expérience 3hydro. On y évalue d'abord les performances du système d'assimilation pour la correction de l'humidité (en comparaison avec le scénario estival). Puis, la capacité du système d'assimilation à corriger une variable qualité (i.e. relative aux pesticides) en observant seulement une variable hydrologique est évaluée.

Estimation de l'humidité

Comme pour le scénario estival, l'assimilation conjointe d'images satellite et de profils verticaux d'humidité permet une correction efficace de l'humidité à toutes les profondeurs (voir Tableau 8.6, gauche). A toutes les profondeurs, les performances sont supérieures à celles de l'Expérience 2Multi/vineyard qui inclut les mêmes observations, aux mêmes emplacement

et à la même fréquence mais avec des forçages climatiques différents (scénario estival dans l'Expérience 2Multi/vineyard, voir Tableau 8.6, droite).

	Expérience 3hydro	Expérience 2vineyard/Multi
Surface	0.81	0.66
Intermédiaire	0.61	0.51
Profond	0.72	0.42

Tableau 8.6 – Comparaison des CRPSS obtenus pour l'estimation de l'humidité moyennés temporellement et sur toutes les UH aux différentes profondeurs dans l'Expérience 3hydro (scénario hybride) et dans l'Expérience 2Multi/vineyard (scénario estival).

Dans le cas du scénario hybride, on peut supposer que l'apparition de plateaux de saturation plus ou moins longs (par exemple pour l'UH2 sur la Figure 8.6) participe à ces meilleures performances. En effet, pendant les périodes saturées qui sont plus fréquentes dans le scénario hybride, l'humidité ne prend qu'une seule valeur. La corrélation temporelle est donc maximale au sein de ces périodes, permettant une propagation plus efficace des corrections depuis les points observés vers les points non observés.

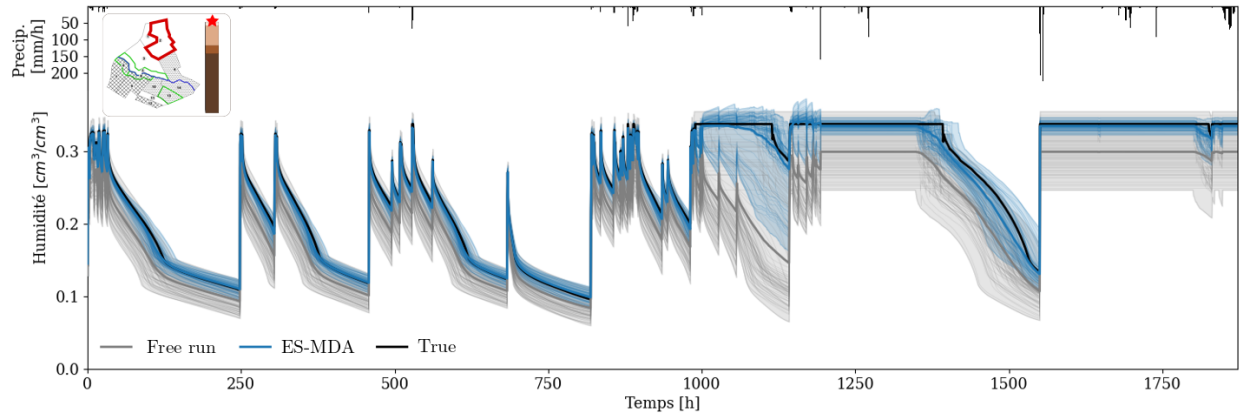


Figure 8.6 – Comparaison des trajectoires d'humidité de surface de la référence ("True", noir), de l'ensemble non assimilés ("Free run", gris) et de l'ensemble analysé (bleu) sur l'UH2 (Expérience 3hydro).

Cependant, même si les performances de l'assimilation sont globalement très bonnes, l'ES-MDA présente localement quelques difficultés. C'est notamment le cas pour l'humidité de surface sur l'UH2 entre 1000 h et 1200 h (voir Figure 8.6). Dans cet intervalle de temps, l'analyse reproduit difficilement le plateau de saturation. Or, l'ensemble de l'ébauche est caractérisé par 2 faisceaux (un saturé, un non saturé) sur cet intervalle de temps, ce qui affecte probablement les performances de l'ES-MDA puisque l'hypothèse de gaussianité n'est plus vérifiée.

	Horizon	Expérience 3hydro	Expérience 2vineyard/Multi
Surface	11	**0.93**	0.75
	12	**0.91**	0.72
	13	**0.82**	0.15
	14	**0.87**	0.84
	15	**0.93**	0.84
	16	**0.87**	0.85
Subsurface	2	**0.67**	0.33
	3	**0.75**	0.36
	4	**0.81**	-0.38
	6	**0.81**	0.68
	7	0.66	**0.71**
	8	**0.94**	0.51
	9	**0.90**	0.41
	10	0.67	0.80

Tableau 8.7 – Valeurs de CRPSS associées à l'estimation des *thetas* pour les expériences 3hydro et 2vineyard/Multi. our chaque ligne, la valeur en gras indique la meilleure estimation.

D'autre part, les performances pour l'estimation des paramètres *thetas* sont également meilleures dans l'Expérience 3hydro que dans l'Expérience 2Multi/vineyard (voir Tableau 8.7). Or, on a montré dans le Chapitre 5, que la distribution de l'humidité en période de saturation est identique à celle du paramètre *thetas* de l'horizon considéré (voir Figure 5.4 pour un rappel). Ainsi, si on dispose d'observations des plateaux de saturations, on dispose aussi directement d'observations des teneurs en eau à saturation *thetas*, expliquant ces meilleures performances.

Estimation de la concentration à l'exutoire

L'impact sur l'estimation de la série temporelle de concentration est aussi significatif puisque l'ES-MDA aboutit à un CRPSS de **0.28** (voir Figure 8.7 pour une comparaison des trajectoires de l'ébauche et de l'analyse).

L'humidité spatialisée et la concentration à l'exutoire semblent donc suffisamment corrélées pour permettre la propagation des corrections d'un compartiment à l'autre. De plus, la concentration à l'exutoire étant partiellement influencée par la teneur en eau à saturation de l'horizon 10 d'après les résultats du Chapitre 5, l'estimation de ce paramètre participe probablement à la correction de la série temporelle de concentration.

Il faut cependant rester prudent avant de conclure à l'intérêt systématique de l'assimilation d'observations d'humidité pour corriger une variable qualité. En effet, on rappelle que

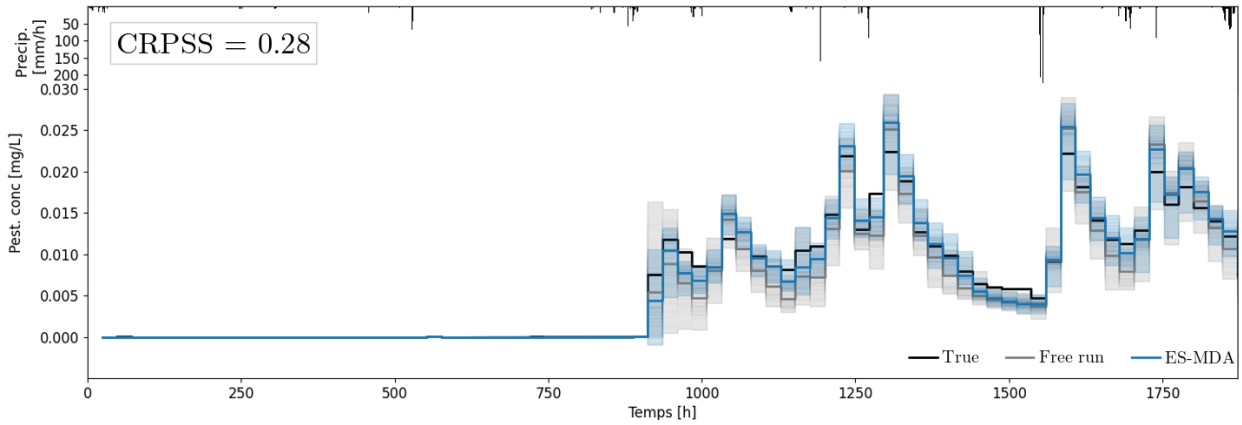


Figure 8.7 – Comparaison des trajectoires de concentration journalière moyenne à l'exutoire de la référence ("True", noir), de l'ensemble non assimilé ("Free run", gris) et de l'ensemble analysé (bleu) (Expérience 3hydro).

le scénario hybride se caractérise par des précipitations intenses entraînant la saturation de nombreuses colonnes de sol et des flux de ruissellement importants. Ces phénomènes seront probablement moins marqués dans des conditions moins humides et l'impact de l'assimilation devrait être étudié dans d'autres configurations.

8.2.2 Assimilation multi-sources : intégration d'observations de concentration

On considère maintenant l'assimilation d'observations d'humidité de surface et de subsurface ainsi que de concentration moyenne hebdomadaire à l'exutoire (Expérience 3pest). Les valeurs de CRPSS pour la correction de la concentration à l'exutoire et des paramètres *manning*_1 et *mn*_10 sont regroupés dans le Tableau 8.8. Contre toute attente, l'intégration d'observations de concentration n'aboutit pas à de meilleures performances que dans l'Expérience 3hydro pour la correction de la concentration. Il semble donc que les corrélations de l'ébauche apportent plus d'informations que les observations directes de concentration. Pour l'estimation des paramètres, l'impact sur la rugosité de Manning est positif alors que l'assimilation dégrade l'estimation du paramètre *mn*_10. Là encore, des compensations d'erreur et des niveaux de corrélation différents avec la concentration dans l'ensemble peuvent expliquer ces contrastes.

Pour mieux comprendre ces résultats, l'Expérience 3pest est renouvelée en faisant varier simultanément l'amplitude de l'erreur et la durée d'exposition de l'échantillonneur passif (donc la fréquence d'observation). Pour limiter le temps de calcul, le nombre d'itérations est fixé à 1 et on retrouve ainsi la configuration du lisseur d'ensemble (ES) classique. Cette configuration donne donc une première idée sur l'intérêt de l'intégration d'observations de concentration même si elle ne permet pas de prendre en compte l'impact d'une meilleure estimation des paramètres.

	CRPSS
Concentration à l'exutoire	0.28
Rugosité de Manning	0.22
mn - Horizon 10	-1.33

Tableau 8.8 – CRPSS moyen obtenus pour l'estimation de la concentration à l'exutoire et des paramètre *Manning*_1 et *mn*_10 dans l'Expérience 3pest.

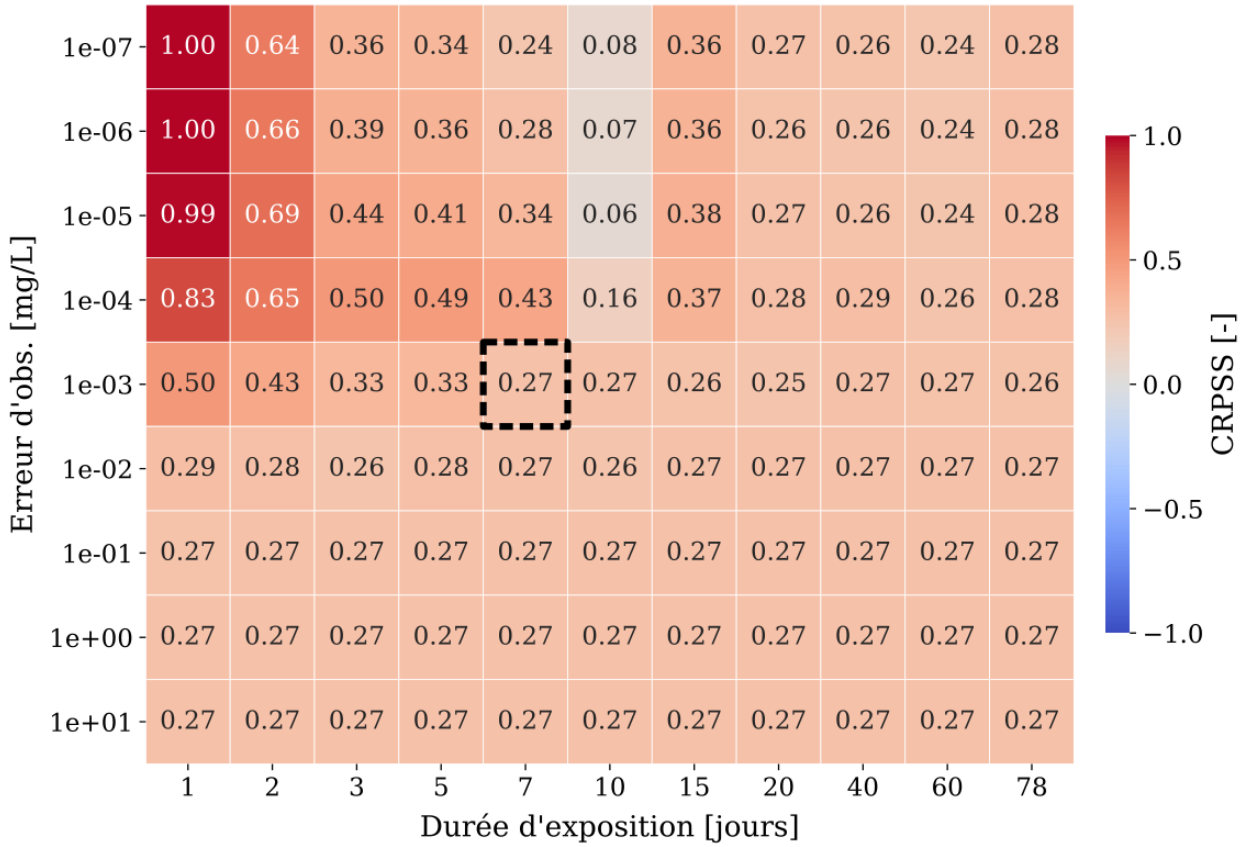


Figure 8.8 – Comparaison des CRPSS de la concentration à l'exutoire quand la durée d'exposition et l'erreur d'observation de l'échantillonneur passif varient. Le pixel entouré en pointillés indique les caractéristiques des observations utilisée dans les expériences précédentes.

Pour des durées d'exposition supérieures à 7 jours et des erreurs d'observation supérieures à 10^{-3} mg.L^{-1}, le CRPSS de la concentration est constant, égal à 0.27 (voir Figure 8.8). On retrouve quasiment la valeur de l'Expérience 3hydro (0.28, Figure 8.7), montrant que les observations de concentration n'apportent aucune information supplémentaire par rapport à l'ébauche et aux corrélations qu'elle contient. Pour constater une amélioration de l'estimation de la concentration significative (par exemple supérieure à 0.6) il est nécessaire de fournir des observations caractérisées par une durée d'expostion inférieure à 2 jours (équivalent à une fréquence d'assimilation de 2 jours) et une erreur d'observation inférieure à 10^{-4} mg.L^{-1}. Sur le

terrain, l'acquisition de données à cette résolution temporelle peut être coûteuse à envisager car elle nécessite la présence régulière d'un technicien. D'autre part, des échantillonneurs de grande précision sont également nécessaires. A titre d'exemple, les observations de la Figure 2.15 obtenues avec des matériaux innovants pour l'échantillonnage passif (Martin, 2016) à proximité de la Morcille sont associées à une erreur d'environ $7 \cdot 10^{-6}$ mg.L^{-1} ce qui serait ici une erreur largement acceptable pour que l'assimilation ait un impact conséquent.
Sur la Figure 8.8, on note également que les observations caractérisées par une erreur inférieure à 10^{-4} mg.L^{-1} avec une durée d'exposition de 10 jours aboutissent à des performances particulièrement mauvaises par rapport aux fréquences de 7 et 15 jours voisines. L'origine d'une telle contre performance reste à explorer et une explication précise n'a pas pu être formulée. On peut imaginer qu'à cette fréquence particulière, les observations sont affectées par des phénomènes type repliement de spectre qui affectent la nature du signal enregistré mais cette hypothèse reste à explorer.
D'autre part, il est important de noter que la concentration à l'exutoire est particulièrement intéressante à estimer pour évaluer globalement les quantités de pesticides transférées à l'échelle du bassin versant. Toutefois, observer et estimer directement cette même concentration comme présenté dans cette partie n'est pas un choix judicieux en termes de formulation du problème inverse. Pour effectivement corriger le modèle, dans de futurs tests on pourrait plutôt envisager d'estimer les quantités de pesticides sur les différentes UH du bassin, voire les doses appliquées, à partir d'observations de concentration à l'exutoire.

8.2.3 Conclusion

Les expériences réalisées montrent que les corrections apportées à l'humidité par l'assimilation d'images et de profils d'humidité peuvent se propager pour corriger la série temporelle de concentration à l'exutoire. Dans ce cas, ce sont les corrélations entre les deux compartiments qui contribuent à cette correction. De telles corrélations sont mêmes plus efficaces que les observations de concentration si la résolution temporelle et la qualité de ces dernières n'est pas suffisante.

On note toutefois que ces corrélations dépendent des choix de modélisation ainsi que des conditions climatiques et que leur amplitude et structure pourront grandement varier dans un autre cas d'étude. Les expériences menées sur le scénario hybride devront notamment être prolongées à l'échelle de la Morcille avant de tirer des conclusions sur l'intérêt de l'assimilation dans une application à cette échelle. En effet, on atteint ici les limites du scénario hybride qui permet d'observer un signal de concentration à l'exutoire mais qui n'est pas physiquement entièrement réaliste. En effet, si l'assimilation s'est avérée efficace dans cette configuration, les corrélations spatiales et temporelles qui existent entre humidité et concentration seront probablement moins marquées dans un scénario plus réaliste en termes de précipitations, car moins de ruissellement sera généré. Pour obtenir des conclusions plus robustes, il semble donc indispensable de s'en tenir à une configuration physiquement accep-

table. On rappelle cependant que de telles configurations n'aboutissaient à aucun signal à l'exutoire dans notre cas, notamment car trop peu de ruissellement était généré. Pour pallier cette difficulté, on pourrait envisager d'utiliser un bassin plus grand incluant plus de zones de génération de ruissellement. Une autre solution serait d'intégrer une représentation de la croûte de battance dans les parcelle de vigne puisque ce phénomène est souvent observé, notamment sur la Morcille. Cette croûte couvrant les premiers centimètres de sol et résultant souvent de pluies intenses est très compacte et peu perméable, favorisant ainsi l'apparition de ruissellement de surface.

A retenir

- ✓ En assimilant seulement des observations d'humidité pour corriger l'humidité, il est aussi possible de corriger la concentration à l'exutoire.
- ✓ Pour que les observations de concentration aient un intérêt, l'échantillonneur passif doit être caractérisé par une durée d'exposition inférieures à 2 jours et une erreur inférieure à 0.1 mg/L.
- ✓ Ces conclusions doivent être prises avec des pincettes car elles ont été obtenues sur un scénario physiquement peu réaliste.

8.3 Conclusion : apports de l'assimilation multi-sources

L'objectif de ce chapitre était d'évaluer l'intérêt de l'assimilation multi-sources pour le modèle PESHMELBA. Les objectifs, la méthodologie et les principaux résultats sont résumés sur la Figure 8.9.

Pour l'humidité

Le chapitre précédent a montré que l'assimilation d'images satellite d'humidité dans les 5 premiers centimètres du sol permet de corriger l'humidité et les *thetas* de l'horizon de surface mais que cette correction ne se propage pas à la subsurface. Dans ce chapitre, les images d'humidité de surface ont été assimilées conjointement à des profils verticaux ponctuels d'humidité sur les 2 premiers mètres de sol pour tenter de dépasser cette limitation. Plusieurs tests faisant varier le nombre et la position des profils ont été menés sur des expériences jumelles pour évaluer quelle configuration aboutit aux meilleures performances. Les résultats ont montré que l'assimilation des images de surface et des profils verticaux ponctuels permet de corriger efficacement l'humidité et les *thetas* à toutes les profondeurs. Les performances du système sont maximales si l'on dispose à minima d'un profil par type de sol et que ces profils sont situés sur des parcelles de vigne plutôt que sur des bandes enherbées.
D'autre part, compte tenu des corrélations douteuses qui apparaissaient dans certains tests et dégradaient les performances de l'ES-MDA, des tests supplémentaires ont été menés en testant l'apport de la localisation par domaine sur l'analyse (1 domaine par type de sol). A

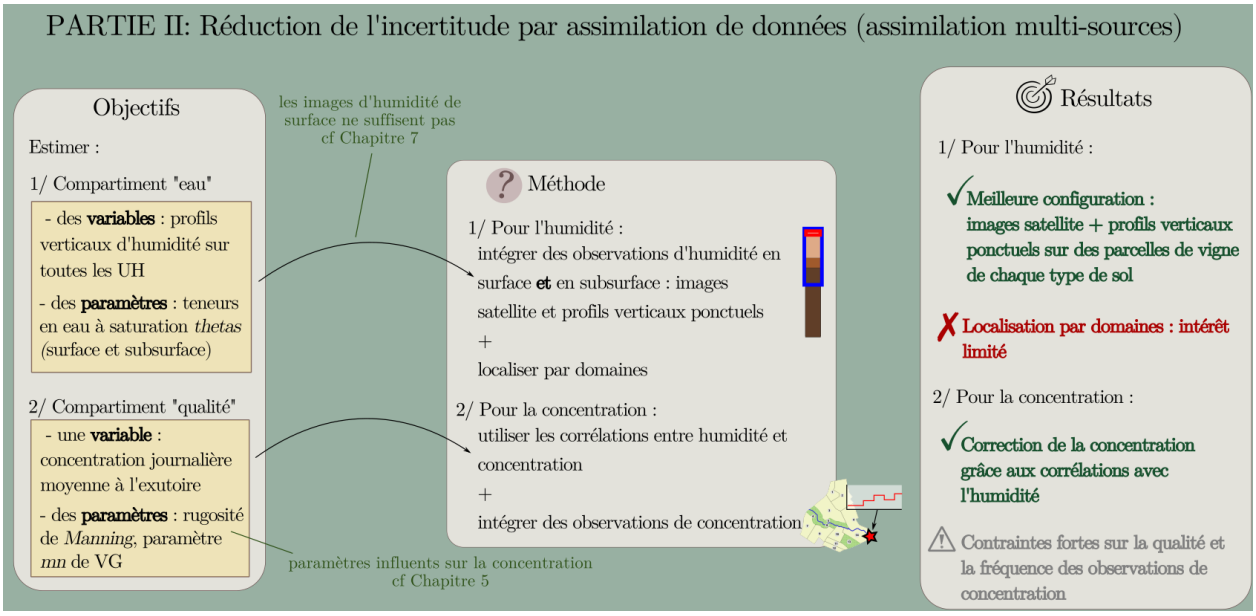


Figure 8.9 – Résumé visuel du chapitre 8 : objectifs de départ, outils utilisés et principaux résultats obtenus pour l'assimilation multi-sources.

l'échelle du bassin versant virtuel, son impact sur les performances de l'assimilation était très limité. Pour un problème à l'échelle du bassin versant de la Morcille, la localisation mérite cependant d'être plus amplement explorée, peut-être sous une autre forme, car son intérêt peut être réel, notamment pour diminuer le coût de calcul de l'analyse.

Pour la concentration

Dans ce chapitre, l'intérêt de l'assimilation multi-sources a également été exploré pour corriger la concentration journalière moyenne à l'exutoire et 2 paramètres d'entrée influents sur cette dernière. Dans un premier temps, aucune observation de concentration n'a été intégrée et la propagation des corrections de l'humidité spatialisée vers la concentration à l'exutoire a été testée. Les résultats ont montré qu'en corrigeant l'humidité, l'assimilation a un impact positif significatif sur la concentration si cette dernière est aussi corrigée. Dans un second temps, des observations de concentration ont été assimilées en complément de l'humidité et l'impact de leur qualité et résolution a été évalué. On a alors montré que les observations de concentration doivent être de très bonne qualité (erreur d'observation et fréquence temporelle) pour que leur assimilation ait un impact significatif.

Limites des tests sur un bassin versant virtuel

Comme dans les chapitres précédents, l'ensemble des tests d'assimilation multi-sources ont été réalisés sur le mini bassin versant virtuel inspiré de la Morcille et défini dans le Chapitre 2. Si ce bassin virtuel permet de tirer des premières conclusions méthodologiques quant à l'intérêt des différentes sources d'observation, la transposition des conclusions à une

application à l'échelle de la Morcille devient difficile. La localisation illustre particulièrement cette problématique. L'intérêt d'une localisation par type de sol semble en effet spécifique à ce scénario de petite taille où la dynamique verticale prédomine sur les transferts latéraux. Il est impossible de prédire que le même type de conclusion s'appliquera à l'échelle de la Morcille. De même, pour la correction de la concentration, il a été nécessaire de modifier les forçages climatiques de manière peu physique pour pallier les limitations qu'imposait la petite taille du bassin en termes de génération de ruissellement. Là encore, le passage à l'échelle d'un véritable bassin semble nécessaire. La réalisation d'expériences jumelles à cette échelle impliquera très probablement d'autres défis techniques mais est une étape nécessaire pour le développement d'un système d'assimilation opérationnel sur la Morcille.

Chapitre 9

Conclusion et perspectives

Sommaire

9.1 Rappel des objectifs et principaux résultats

Les risques que fait peser l'utilisation intensive de pesticides sur les écosystèmes non cibles et sur la santé humaine sont nombreux. L'existance de tels risques a motivé le développement de modèles numériques qui visent à simuler le devenir des pesticides après leur application. Au delà d'un objectif de meilleure compréhension des processus de transfert et de transformation des pesticides, de tels modèles ont souvent vocation à être utilisés dans un cadre opérationnel par exemple pour limiter la contamination des zones aquatiques sensibles. Le modèle PESHMELBA est développé dans ce contexte, afin d'évaluer l'impact du paysage sur les transferts de pesticides et à terme, servir de support à la prise de décision concernant l'implantation de solutions correctives telles que les bandes enherbées ou les haies. Cependant, pour que PESHMELBA puisse être utilisé comme un outil opérationnel pertinent, il est indispensable d'être capable de quantifier et de réduire ses incertitudes.
Le développement d'une méthodologie adaptée pour quantifier puis réduire les incertitudes liées à PESHMELBA constituait les deux axes principaux de ces travaux de thèse. Ces thématiques étant explorées pour la première fois dans ce modèle, un scénario virtuel inspiré du bassin versant de la Morcille, l'un des premiers bassins versants ciblé pour l'application de PESHMELBA, a été mis en place. Malgré plusieurs simplifications, notamment en termes de taille ($< 1\,\mathrm{km}^2$) et de composition (parcelles de vigne et bandes enherbées uniquement), ce scénario a été pensé pour rester aussi représentatif que possible des challenges que représentera une application à l'échelle d'un bassin versant réel. Il a été paramétré à partir des connaissances disponibles sur la Morcille et les incertitudes relatives aux paramètres d'entrée (formes et paramètres des distributions) ont aussi été définies de manière à refléter au mieux la réalité.
A partir de ce scénario, nous avons ensuite exploré la thématique de la quantification d'incertitude à partir d'une analyse d'incertitude et d'une analyse de sensibilité. Puis, la réduction d'incertitude a été abordée par le biais des méthodes d'assimilation de données. Les objectifs et principaux résultats obtenus pour chacun de ces axes sont résumés dans les paragraphes suivants.

9.1.1 Quantification de l'incertitude

La quantification des incertitudes relatives aux sorties de PESHMELBA a été mise en oeuvre avec plusieurs objectifs :

1. Améliorer notre compréhension du modèle ;
2. Participer à sa validation ;
3. Proposer éventuellement des pistes de simplification de certains processus représentés par le modèle ;
4. Caractériser au mieux les variables ciblées pour l'assimilation de données dans la deuxième partie de ces travaux.

Pour cela, plusieurs variables intégrées temporellement et représentatives du fonctionnement du modèle (volumes d'eau et masses de pesticides cumulés transférés en surface puis en subsurface) et plusieurs séries temporelles ciblées pour l'assimilation de données (humidité à plusieurs profondeurs de la colonne de sol et concentration moyenne en pesticides à l'exutoire) ont été analysées. La quantification de l'incertitude dans PESHMELBA a ensuite été réalisée à partir d'une analyse d'incertitude et d'une analyse de sensibilité.

L'**analyse d'incertitude** a été réalisée par une méthode classique de propagation de l'incertitude par simulations de Monte Carlo. En plus de participer à caractériser certains processus physiques, en mettant notamment en lumière la présence d'effets de seuils, l'analyse d'incertitude a aussi permis de conclure que l'humidité et la concentration à l'exutoire suivent toutes deux une distribution gaussienne ce qui constitue une information importante pour le choix de la méthode d'assimilation.

Quant à l'**analyse de sensibilité**, elle est en général plus complexe à mettre en oeuvre et il existe de nombreuses méthodes différant par le type d'information qu'elles fournissent et leur complexité de mise en oeuvre. N'ayant jamais été réalisée auparavant pour le modèle PESHMELBA, la question de la méthodologie s'est avérée critique dans ces travaux. En effet, les contraintes suivantes s'imposaient compte tenu de la nature et de la structure du modèle :

- nombre élevé de paramètres d'entrée (145 dans le cas d'étude considéré) ;
- variables de sortie spatialisées ;
- coût élevé d'une simulation PESHMELBA.

Pour prendre en compte ces contraintes, deux méthodologies différentes ont été utilisées pour les variables intégrées et pour les séries temporelles.

Pour les variables intégrées Une première étape de criblage permettant de distinguer paramètres influents et non influents a été appliquée afin de réduire la dimension du problème. Compte tenu des difficultés rencontrées dans ces travaux pour mettre en oeuvre des méthodes de criblage classiques comme la méthode de Morris, une méthode récente (test d'indépendance basé sur la mesure de dépendance HSIC) a été appliquée avec succès. Celle-ci a permis de réduire de près de deux tiers le nombre de paramètres impliqués dans l'analyse de sensibilité pour chaque variable. Dans un second temps, plusieurs méthodes de classement différant éventuellement par leur définition de la sensibilité ont été testées : indices de Sobol calculés à partir d'une décomposition en polynômes du chaos, mesures de dépendance HSIC et mesures d'importance obtenues par construction d'une forêt aléatoire. La comparaison de ces méthodes a permis de montrer que le calcul des indices de Sobol par PCE reste la méthode la plus adaptée à ce cas d'étude, de par la robustesse des indices qu'elle permet de calculer à partir d'un échantillon de taille limitée et l'interprétabilité de l'information que ces indices fournissent. Les indices de Sobol calculés à la fois à l'échelle locale de la

parcelle et de manière aggrégée à l'échelle du versant puis du bassin versant ont ainsi permis de répondre aux objectifs de meilleure compréhension du modèle. La quantification de l'incertitude a notamment permis de caractériser certains processus comme le ruissellement ou l'adsorption et d'en valider la représentation dans PESHMELBA. Certaines pistes de simplification ont également été proposées, comme par exemple pour la représentation de la courbe de conductivité. En effet, celle-ci inclue actuellement une représentation des écoulements dans les macropores qui implique de renseigner des paramètres supplémentaires ne ressortant influents dans aucune des analyses de sensiblité du modèle.

Pour les séries temporelles Compte tenu de la dimension des séries temporelles considérées ($>10^3$), une décomposition en composantes principales fonctionnelle a d'abord été réalisée pour réduire cette dimension. On a ainsi retenu 2 composantes principales pour l'humidité et 4 pour la concentration à l'exutoire, réduisant ainsi drastiquement la dimension des variables de sortie ciblées tout en conservant l'information relative à leur variabilité. Une étape de criblage a ensuite été réalisée. Ces variables étant moins complexes que les variables intégrées analysées, la méthode de Morris a pu être appliquée, réduisant le nombre de paramètres influents sur chaque mode de chaque variable à moins de 50. Finalement, pour chaque mode (et chaque parcelle pour l'humidité), les indices de Sobol ont été calculés à partir d'une décomposition en Polynômes du Chaos permettant d'identifer que les teneurs en eau à saturation de tous les horizons (resp. la rugosité de Manning des parcelles de vignes, la teneur en eau à saturation et le paramètre de Van Genuchten *mn* d'un horizon profond) influent majoritairement sur l'humidité dans la colonne de sol (resp. la concentration en pesticides à l'exutoire).

Ce premier axe de travail a ainsi permis de proposer et de mettre en oeuvre une méthodologie adaptée pour caractériser l'incertitude liée aux sorties du modèle PESHMELBA. Ce travail est informatif en soi mais fournit aussi un ensemble d'éléments précieux pour construire un problème d'assimilation de données qui fait sens.

9.1.2 Réduction de l'incertitude par assimilation de données

Dans une deuxième partie, l'assimilation de données a été mise en oeuvre avec plusieurs objectifs :

1. Corriger les variables d'humidité et estimer les paramètres influents sur ces dernières (identifiés grâce à l'analyse de sensibilité) ;
2. Corriger la concentration en pesticides à l'exutoire et les paramètres influents sur celle-ci (également identifiés à partir de l'analyse de sensibilité).

Pour cela, on supposait disponibles plusieurs types d'observations :

- des images radar d'humidité de surface disponibles sur chaque parcelle ;

- des profils verticaux d'humidité sur 2 m de profondeur disponibles ponctuellement sur quelques parcelles ;
- des relevés de concentration hebdomadaire moyenne en pesticides à l'exutoire.

Choix de la méthode L'utilisation de l'assimilation de données pour le modèle PESH-MELBA a débuté par une réflexion sur la méthode la plus adaptée aux spécificités du modèle et aux objectifs formulés. Plusieurs méthodes ensemblistes dérivant du filtre de Kalman (EnKF, ES-MDA et iEnKS) ont été testées sur un problème de réanalyse impliquant l'estimation jointe de variables (humidité à plusieurs profondeurs) et de paramètres (teneurs en eau à saturation) à partir d'images d'humidité de surface. Ces méthodes ont été mises en oeuvre sur des expériences jumelles et leurs performances pour corriger variables de sortie et paramètres d'entrée ont été comparées. La robustesse des 3 méthodes a également été étudiée en explorant leur sensibilité à l'amplitude de l'erreur d'observation, à la fréquence d'observation ainsi qu'à la taille de l'ensemble. Les résultats ont permis d'identifier l'ES-MDA comme le meilleur compromis pour ce cas d'étude, notamment puisqu'elle permet d'intégrer les informations concernant la dynamique temporelle du système pour effectuer ses analyses. A partir de ces expériences, on a également pu formuler des recommandations quant à la qualité des images de surface à utiliser pour que l'assimilation ait un impact significatif sur les simulations d'humidité. Ces expériences ont finalement permis de déterminer une taille d'ensemble optimale pour que la précision des analyses de l'ES-MDA soit suffisante sans trop affecter son coût numérique.
Toutefois, que cela soit pour l'ES-MDA ou les autres méthodes testées, cette première étude a également montré qu'en assimilant des observations d'humidité de surface, il est seulement possible de corriger l'humidité de surface et les teneurs en eau à saturation de surface. En effet, la correction ne se propage pas vers les horizons de subsurface par manque de corrélation entre les horizons de surface et de subsurface.

Assimilation multi-sources Pour dépasser les limitations identifiées concernant la correction de l'humidité de subsurface, des expériences incluant plusieurs sources de données ont ensuite été réalisées. Les profils verticaux d'humidité (plus coûteux et plus difficiles à obtenir que les images radar) ont été intégrés au système d'assimilation et leur impact sur la correction de l'humidité et sur l'estimation des teneurs en eau à saturation en surface et en subsurface a été évalué. Pour cela, plusieurs configurations ont été comparées en faisant varier le nombre de profils intégrés et leurs positions dans le bassin là encore à partir d'expériences jumelles. Les résultats ont montré que l'assimilation de ces deux sources de données conjuguées permet une correction efficace en surface et en subsurface. Cette correction est la meilleure lorsqu'on assimile à la fois les images d'humidité de surface et des profils verticaux d'humidité mesurés à minima sur une parcelle de chaque type de sol constituant le bassin versant. Des expériences supplémentaires consistant à effectuer une analyse localisée (localisation par domaines avec un domaine par type de sol) ont été réalisées. Les résultats ne

montrent pas un impact clair de la localisation sur la qualité de la correction mais l'approche reste à explorer pour limiter le coût de calcul dans un scénario plus grand.
Finalement, l'apport de l'assimilation multi-sources pour la correction de la concentration à l'exutoire et des paramètres d'entrée influents sur cette dernière a été exploré. Une première expérience a permis d'évaluer la capacité du système d'assimilation à corriger la concentration à partir d'observations concernant seulement un autre compartiment (images radar et profils verticaux d'humidité). Cette configuration aboutit à une correction significative de la concentration validant ainsi l'efficacité de l'assimilation "fortement couplée" (*strongly-coupled assimilation*) sur ce scénario. Dans un deuxième temps, l'intégration d'observations de concentration a permis d'évaluer quelles devront être les caractéristiques de ce type de données (fréquence et amplitude d'erreur d'observation) pour qu'elles puissent améliorer de manière significative les simulations de concentration journalière.

Ainsi, ce deuxième axe de travail constitue la première application de méthodes d'assimilation de données dans le modèle PESHMELBA. Les possibilités qu'offrent ce type de méthodes ont été explorées de manière intensive et cette étude ouvre ainsi la voie au développement d'un tel système conjointement à toute nouvelle application du modèle.

9.2 Perspectives : vers une application au bassin versant de la Morcille

Ces travaux de thèse se sont appuyés sur des expériences jumelles réalisées à partir d'un bassin versant virtuel simplifié en termes de taille et de composition. Il est naturel d'envisager que les différentes méthodes mises en oeuvre dans cette thèse soient transposées dans un deuxième temps sur un bassin versant réel, en commençant par la Morcille, à partir d'observations elles aussi réelles.
Une telle transposition ouvre de nouvelles perspectives et implique de nouveaux défis qui sont abordés dans les paragraphes suivants.

9.2.1 Application du modèle et fonctionnement du bassin versant

La mise en place de PESHMELBA à l'échelle du bassin versant de la Morcille constitue un prérequis à l'application des méthodes utilisées dans ces travaux. Toutefois, une telle application est loin d'être évidente et constituera un premier défi. L'utilisation intensive de PESHMELBA sur le scénario virtuel s'est avérée complexe et a été fortement perturbée par des difficultés techniques qu'elles soient numériques, liées à l'utilisation du coupleur OpenPALM (qui n'a pas été maintenu pendant la thèse) sur le supercalculateur d'INRAE ou à la nécessité d'interrompre le code pour réaliser les analyses de l'assimilation. Certaines de ces difficultés ne sont toujours pas résolues et devront être contournées lors du passage à

l'échelle de la Morcille.

D'autre part, on peut supposer que le fonctionnement de la Morcille sera autrement plus complexe que celui du cas d'étude utilisé dans la thèse. Par exemple, la topographie, le nombre de parcelles et la répartition spatiale des pentes sont autant de facteurs pouvant influer sur les contributions des écoulements de surface et de subsurface. Il en découle que les paramètres influents identifiés pour chaque variable dans les Chapitre 4 et 5 sont probablement amenés à changer tout comme l'échelle d'aggrégation des indices de Sobol proposée dans le Chapitre 4. Il sera ainsi indispensable de réaliser une nouvelle analyse de sensibilité, notamment pour identifier quels paramètres il serait raisonnable d'envisager estimer avec l'assimilation de données. De même, on peut supposer que l'application à un bassin autrement plus complexe est susceptible d'affecter l'observabilité de certains processus ou compartiments ainsi que les corrélations entre les différents horizons de surface et de subsurface ce qui peut altérer les performances de l'assimilation.

9.2.2 Acquisition et caractérisation des observations

Le passage à une application réaliste implique d'effectuer des mesures et de caractériser au plus juste les erreurs associées à chaque type d'observation. En termes d'acquisition, les résultats de ces travaux ont permis de fournir un certain nombre de recommandations en termes de position des profils verticaux d'humidité ou de durée d'exposition des échantillonneurs passifs pouvant aider à dimensionner les campagnes de mesures à réaliser.

Concernant la caractérisation de ces observations, on rappelle que dans les expériences jumelles mises en oeuvre dans ces travaux, les erreurs d'observation étaient entièrement connues et maîtrisées. En réalité, de telles erreurs sont souvent difficiles à définir puisqu'elles résultent de plusieurs facteurs combinés. Lors d'une application au bassin versant de la Morcille, ces erreurs devront être soigneusement caractérisées pour tous les types d'observation dont on dispose.

De plus, dans ces travaux on a formulé l'hypothèse que les erreurs d'observation associées aux différents types de capteurs ne sont pas corrélées spatialement. Cette hypothèse est souvent formulée puisqu'elle permet d'exprimer la matrice d'erreur d'observation $\mathbf{R}$ sous la forme d'une matrice diagonale, ce qui allège considérablement le coût numérique de son inversion. Or, cette hypothèse est probablement fausse, surtout pour les images radar d'humidité de surface car tous les pixels sont mesurés par le même appareil. Ignorer de telles corrélations peut fortement affecter les performances de l'assimilation (Stewart et al., 2013 ; Chabot et al., 2015) et une perspective intéressante à ces travaux serait de prendre en compte de telles corrélations dans le processus d'assimilation. Deux défis se présentent alors :

1. Il faut être capable de caractériser les structures de corrélation existantes. Ceci peut être délicat pour les images de surface pour lesquelles l'impact du processus d'inversion

du signal radar sur des structures de corrélation existantes est jusqu'à présent inconnu.

2. Il faut être capable de manipuler une matrice d'erreur d'observation $\mathbf{R}$ non diagonale sans faire exploser les coûts de calcul. Pour cela, une approche intéressante pourrait être de représenter cette erreur d'observation dans un espace multi-échelles par exemple à l'aide d'ondelettes ou de curvelettes comme proposé dans Chabot et al., 2015. Ce type de transformation et la procédure proposée dans Chabot et al. (2015) et Chabot et al. (2020) permet en effet de conserver l'information importante quant aux structures spatiales des erreurs tout en utilisant une matrice diagonale dans le coeur de l'algorithme.

9.2.3 Impacts sur le coût numérique des méthodes

Le changement d'échelle que représente l'application au bassin versant de la Morcille affectera grandement le coût de calcul des méthodes à mettre en oeuvre pour l'analyse de sensibilité et l'assimilation de données. D'une part, l'impact sera significatif sur le temps de calcul d'une simulation PESHMELBA. D'autre part, la taille du problème considéré va être grandement augmentée. En effet, le scénario virtuel utilisé dans ces travaux comptait 14 UH alors qu'on en compte plus de 500 sur le bassin de la Morcille.
Pour l'analyse de sensibilité, il semble difficile d'envisager que des indices de Sobol puissent être calculés sur chaque UH à partir d'une décomposition en polynômes du chaos. Le recours à des indices aggrégés directement estimés à partir d'une méthode *pick-freeze* sera probablement indispensable mais celle-ci devra être précédée d'une réflexion sur l'échelle d'aggrégation la plus adéquate.
Pour l'assimilation de données, l'intégration de toutes les trajectoires temporelles dans le vecteur d'état considéré pour l'ES-MDA aboutira à un problème de dimension de l'ordre de 10^8 éléments. Si la localisation ne s'est pas avérée essentielle dans ces travaux, elle sera probablement incontournable et plus efficace lors d'une application sur la Morcille. Là encore, l'identification de domaines indépendants est une étape cruciale mais délicate.

9.2.4 Estimation robuste des paramètres

La mise en place d'un système d'assimilation à l'échelle de la Morcille pourra être l'occasion de proposer des valeurs de paramètres corrigées et utilisables pour l'exploration de scénarios d'aménagement du territoire. Or dans ces travaux, l'utilisation de l'assimilation de données pour corriger les paramètres ne prend pas en compte les incertitudes "subies" comme l'incertitude liée aux doses et dates des applications de pesticides, une information souvent très difficile à obtenir ou les incertitudes liées à la variabilité intrinsèque des forçages climatiques comme l'ETP ou la pluie. Par exemple, lorsqu'on considère des pluies différentes, on peut aboutir à des estimations différentes des paramètres comme l'illustre la Figure 9.1. Dans ce cas, il est impossible de statuer sur l'estimation du paramètre *thetas11* qui est correcte. L'estimation des paramètres considérée dans ces travaux est en quelque sorte *localisée*

vis-à-vis des conditions climatiques et il serait souhaitable de s'extraire de cette localité pour aboutir à une estimation robuste des paramètres.

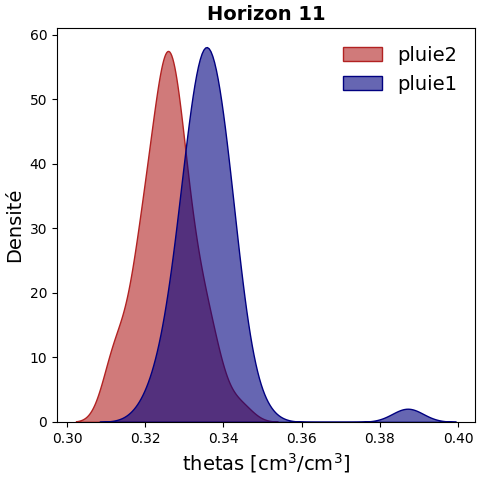


Figure 9.1 – Exemples des distributions *a posteriori* obtenues pour l'estimation du paramètre *thetas11* en considérant 2 chroniques de pluie différentes.

Dans des travaux futurs, il serait intéressant de prendre en compte les incertitudes aléatoires dans l'estimation des paramètres d'entrée de PESHMELBA (on parle d'optimisation robuste ou optimisation sous incertitude). Pour cela, il sera notamment nécessaire d'identifier des critères de robustesse adaptés et de mettre en oeuvre une méthodologie compatible avec les contraintes imposées par PESHMELBA.

Bibliographie

Adriaanse, P. I. (1997). « Exposure assessment of pesticides in field ditches : The TOXSWA model ». *Pesticide Science* 49.2, p. 210-212. doi : 10.1002/(SICI)1096-9063(199702)49:2<210::AID-PS496>3.0.CO;2-1.

Ahmed, S. E., S. Pawar et O. San (2020). « PyDA : A Hands-On Introduction to Dynamical Data Assimilation with Python ». *fluids* 5.4. doi : 10.3390/fluids5040225.

Alletto, L., P. Benoit, B. Bolognési, M. Couffignal, V. Bergheaud, V. Dumény, C. Longueval et E. Barriuso (2013). « Sorption and mineralisation of S-metolachlor in soils from fields cultivated with different conservation tillage systems ». *Soil and Tillage Research* 128, p. 97-103. doi : 10.1016/j.still.2012.11.005.

Alletto, L., Y. Coquet, P. Benoit et V. Bergheaud (2006). « Effects of temperature and water content on degradation of isoproturon in three soil profiles ». *Chemosphere* 64.7, p. 1053-1061. doi : 10.1016/j.chemosphere.2005.12.004.

Alletto, L., V. Pot, S. Giuliano, M. Costes, F. Perdrieux et E. Justes (2015). « Temporal variation in soil physical properties improves the water dynamics modeling in a conventionally-tilled soil ». *Geoderma* 243-244, p. 18-28. doi : 10.1016/j.geoderma.2014.12.006.

Alves Ferreira, V., G. Weesies, D. Yoder, G.R. Foster et K. Renard (1995). « The site and condition specific nature of sensitivity analysis ». *Journal of Soil and Water Conservation* 50, p. 493-497.

Anderson, J. L. (2001). « An ensemble adjustment Kalman filter for data assimilation ». *Monthly weather review* 129.12, p. 2884-2903. doi : 10.1175/1520-0493(2001)129<2884:AEAKFF>2.0.CO;2.

Anderson, J. L. (2012). « Localization and sampling error correction in ensemble Kalman filter data assimilation ». *Monthly Weather Review* 140, p. 2359-2371. doi : 10.1175/MWR-D-11-00013.1.

Antoniadis, A., S. Lambert-Lacroix et J. M. Poggi (2021). « Random forests for global sensitivity analysis : A selective review ». *Reliability Engineering and System Safety* 206, p. 107312. doi : 10.1016/j.ress.2020.107312.

Arcement, G. J. et V. R. Schneider (1989). *Guide for selecting Manning's roughness coefficients for natural channels and flood plains*. Rapp. tech. U.S. G.P.O.

Archer, G. E. B., A. Saltelli et I. M. Sobol (1997). « Sensitivity measures, ANOVA-like Techniques and the use of bootstrap ». *Journal of Statistical Computation and Simulation* 58.2, p. 99-120. doi : 10.1080/00949659708811825.

Argent, R. M., J. M. Perraud, J. M. Rahman, R. B. Grayson et G. M. Podger (2009). « A new approach to water quality modelling and environmental decision support systems ». *Environmental Modelling & Software* 24.7, p. 809-818. doi : 10.1016/j.envsoft.2008.12.010.

Arnold, J. G., R Srinivasan, R. S. Muttiah et J. R. Williams (1998). « Large area hydrologic modeling and assessment - Part 1 : Model development ». *Journal of the American Water Resources Association* 34.1, p. 73-89. doi : 10.1111/j.1752-1688.1998.tb05961.x.

Arora, K., S. K. Mickelson, M. J. Helmers et J. L. Baker (2010). « Review of Pesticide Retention Processes Occurring in Buffer Strips Receiving Agricultural Runoff1 ». *Journal of the American Water Resources Association* 46.3, p. 618-647. doi : 10.1111/j.1752-1688.2010.00438.x.

Asch, M., M. Bocquet et M. Nodet (2016). *Data assimilation : methods, algorithms, and applications*. Fundamentals of Algorithms. SIAM, p. xviii + 306.

Ashby, S. F. et R. D. Falgout (1996). « A parallel multigrid preconditioned conjugate gradient algorithm for groundwater flow simulations ». *Nuclear Science and Engineering* 124.1, p. 145-159. doi : 10.13182/NSE96-A24230.

Attema, E. P. W. et F. T. Ulaby (1978). « Vegetation modeled as a water cloud ». *Radio Science* 13.2, p. 357-364. doi : 10.1029/RS013i002p00357.

Aulia, A., D. Jeong, I. Mohd Saaid, D. Kania, M. Taleb Shuker et N. A. El-Khatib (2019). « A Random Forests-based sensitivity analysis framework for assisted history matching ». *Journal of Petroleum Science and Engineering* 181, p. 106237. doi : 10.1016/j.petrol.2019.106237.

Baatz, D., W. Kurtz, H. J. Hendricks Franssen, H. Vereecken et S. J. Kollet (2017). « Catchment tomography - An approach for spatial parameter estimation ». *Advances in Water Resources* 107, p. 147-159. doi : 10.1016/j.advwatres.2017.06.006.

Baghdadi, N., R. Cresson, E. Pottier, M. Aubert, M. Zribi, A. Jacome et S. Benabdallah (2012). « A Potential Use for the C-Band Polarimetric SAR Parameters to Characterize the Soil Surface Over Bare Agriculture Fields ». *IEEE Transactions on Geoscience and Remote Sensing* 50.10, p. 3844-3858. doi : 10.1109/TGRS.2012.2185934.

Bailey, R. T. et D. Baù (2010). « Ensemble smoother assimilation of hydraulic head and return flow data to estimate hydraulic conductivity distribution ». *Water Resources Research* 46.12. doi : 10.1029/2010WR009147.

Bailey, R. T. et D. Baù (2012). « Estimating geostatistical parameters and spatially-variable hydraulic conductivity within a catchment system using an ensemble smoother ». *Hydrology and Earth System Sciences* 16.2, p. 287-304. doi : 10.5194/hess-16-287-2012.

Balasubramanian, K., B. K. Sriperumbudur et G. Lebanon (2013). « Ultrahigh dimensional feature screening via RKHS embeddings ». *16th International Conference on Artificial Intelligence and Statistics, Scottsdale, AZ, USA*. T. 31. Microtome Publishing, p. 126-134.

Barbash, J. E. et E. A. Resek (1996). « Pesticides in ground water—Distribution, trends, and governing factors ». *Ann Arbor Press*, p. 418-419.

Baudin, P. (2015). « Prévision séquentielle par agrégation d'ensemble : application à des prévisions météorologiques assorties d'incertitudes ». Thèse de doct. Université Paris 11.

Bénard, C. (2021). « Forêts aléatoires et interprétabilité des algorithmes d'apprentissage ». Thèse de doct. Sorbonne Université.

Bénard, C., S. Da Veiga et E. Scornet (2021). « MDA for random forests : inconsistency, and a practical solution via the Sobol-MDA ». working paper or preprint.

Berlinet, A. et C. Thomas-Agnan (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. doi : 10.1007/978-1-4419-9096-9.

Beven, K. et P. Germann (1982). « Macropores and water flow in soils ». *Water Resources Research* 18.5, p. 1311-1325. doi : 10.1029/WR018i005p01311.

Bishop, C. H., B. J. Etherton et S. J. Majumdar (2001). « Adaptive sampling with the ensemble transform Kalman filter. Part I : Theoretical aspects ». *Monthly weather review* 129.3, p. 420-436.

Blatman, G. et B. Sudret (2011). « Adaptive sparse polynomial chaos expansion based on least angle regression ». *Journal of Computational Physics* 230, p. 2345-2367. doi : 10.1016/j.jcp.2010.12.021.

Blayo, E., E. Cosme et A. Vidard (2019). *Introduction to data assimilation*. Sous la dir. d'Université Grenoble Alpes. Cours.

Blayo, E., S. Durbiano, A. Vidard et F. X. Le Dimet (2003). « Reduced order strategies for variational data assimilation in oceanic models ». *Data Assimilation for Geophysical Flows*. Springer-Verlag.

Bocquet, M. et P. Sakov (2012). « Combining inflation-free and iterative ensemble Kalman filters for strongly nonlinear systems ». *Nonlinear Processes in Geophysics* 19.3, p. 383-399.

Bocquet, M. et P. Sakov (2013). « Joint state and parameter estimation with an iterative ensemble Kalman smoother ». *Nonlinear Processes in Geophysics* 20.5, p. 803-818. doi : 10.5194/npg-20-803-2013.

Bocquet, M. et P. Sakov (2014). « An iterative ensemble Kalman smoother ». *Quarterly Journal of the Royal Meteorological Society* 140.682, p. 1521-1535. doi : 10.1002/qj.2236.

Bonan, B., C. Albergel, Y. Zheng, A. L. Barbu, D. Fairbairn, S. Munier et J. C. Calvet (2020). « An ensemble square root filter for the joint assimilation of surface soil moisture and leaf area index within the Land Data Assimilation System LDAS-Monde : application over the Euro-Mediterranean region ». *Hydrology and Earth System Sciences* 24.1, p. 325-347. doi : 10.5194/hess-24-325-2020.

Botto, A., E. Belluco et M. Camporese (2018). « Multi-source data assimilation for physically based hydrological modeling of an experimental hillslope ». *Hydrology and Earth System Sciences* 22.8, p. 4251-4266. doi : 10.5194/hess-22-4251-2018.

Bousbih, S., M. Zribi, M. El Hajj, N. Baghdadi, Z. Lili-Chabaane, Q. Gao et P. Fanise (2018). « Soil moisture and irrigation mapping in a semi-arid region, based on the synergetic use of Sentinel-1 and Sentinel-2 data ». *Remote Sensing* 10.12. doi : 10.3390/rs10121953.

Branger, F., I. Braud, S. Debionne, P. Viallet, J. Dehotin, H. Henine, Y. Nedelec et S. Anquetin (2010). « Towards multi-scale integrated hydrological models using the LIQUID® framework. Overview of the concepts and first application examples ». *Environmental Modelling and Software* 25, p. 1672-1681. doi : 10.1016/j.envsoft.2010.06.005.

Brankart, J. M., C. E. Testut, P. Brasseur et J. Verron (2003). « Implementation of a multivariate data assimilation scheme for isopycnic coordinate ocean models : Application to a 1993–1996 hindcast of the North Atlantic Ocean circulation ». *Journal of Geophysical Research : Oceans* 108.C3. doi : doi.org/10.1029/2001JC001198.

Breiman, L. (1996). « Bagging predictors ». *Machine Learning* 24.2, p. 123-140. doi : 10.1007/BF00058655.

Breiman, L. (2001). « Random forests ». *Machine Learning* 45.1, p. 5-32. doi : 10.1023/A:1010933404324.

Brown, C., A. Alix, J. L. Alonso-Prados, D. Auteri, J. J. Gril, R. Hiederer, C. Holmes, A. Huber, F. de Jong, M. Liess, S. Loutseti, N. Mackay, W. M. Maier,

S. Maund, C. Pais, W. Reinert, M. Russell, T. Schad, R. Stadler, M. Streloke, M. Styczen et J. van de Zande (2007). *Landscape and mitigation factors in aquatic risk assessment. Volume 2 : detailed technic*. Rapp. tech. European Commission.

Brown, T. A. (1974). *Admissible scoring systems for continuous distributions*. Rapp. tech. The Rand Corporation.

Bruand, A. et Y. Coquet (2005). « Les sols et le cycle de l'eau ». *Science du Sol et Environnement*.

Brunet, P., R. Clément et C. Bouvier (2010). « Monitoring soil water content and deficit using Electrical Resistivity Tomography (ERT) – A case study in the Cevennes area, France ». *Journal of Hydrology* 380.1, p. 146-153. doi : 10.1016/j.jhydrol.2009.10.032.

Buis, S., A. Piacentini et D. Déclat (2006). « PALM : a computational framework for assembling high-performance computing applications ». *Concurrency and Computation : Practice and Experience* 18.2, p. 231-245. doi : 10.1002/cpe.914.

Burgers, G., J. P. van Leeuwen et G. Evensen (1998). « Analysis scheme in the ensemble Kalman filter ». *Monthly weather review* 126.6, p. 1719-1724.

Buytaert, W., D. Reusser, S. Krause et Renaud J. P. (2008). « Why can't we do better than Topmodel ? » *Hydrological Processes* 22.20, p. 4175-4179. doi : 10.1002/hyp.7125.

Caisson, A. (2019). « Prise en main et application d'un modèle spatialisé à base physique (CATHY) sur un versant expérimental pour la mise en place d'un système d'assimilation de données ». Mém. de mast. ENGEES.

Campolongo, F., J. Cariboni et A. Saltelli (2007). « An effective screening design for sensitivity analysis of large models ». *Environmental Modelling and Software* 22.10. Modelling, computer-assisted simulations, and mapping of dangerous phenomena for hazard assessment, p. 1509-1518. doi : 10.1016/j.envsoft.2006.10.004.

Campolongo, F., A. Saltelli et J. Cariboni (2011). « From screening to quantitative sensitivity analysis. A unified approach ». *Computer Physics Communications* 182.4, p. 978-988. doi : 10.1016/j.cpc.2010.12.039.

Camporese, M., C. Paniconi, M. Putti et M. Orlandini (2010). « Surface-subsurface flow modeling with path-based runoff routing, boundary condition-based coupling, and assimilation of multisource observation data ». *Water Resources Research* 46.2. doi : 10.1029/2008WR007536.

Camporese, M., C. Paniconi, M. Putti et P. Salandin (2009). « Ensemble Kalman filter data assimilation for a process-based catchment scale model of surface and subsurface flow ». *Water Resources Research* 45.10. doi : 10.1029/2008WR007031.

Carlsen, S. C. K., N. H. Spliid et B. Svensmark (2006). « Drift of 10 herbicides after tractor spray application. 2. Primary drift (droplet drift) ». *Chemosphere* 64.5, p. 778-786. doi : 10.1016/j.chemosphere.2005.10.060.

Carluer, N., C. Lauvernet, D. Noll et R. Muñoz-Carpena (2017). « Defining context-specific scenarios to design vegetated buffer zones that limit pesticide transfer via surface runoff ». *Science of The Total Environment* 575, p. 701-712. doi : 10.1016/j.scitotenv.2016.09.105.

Carluer, N., J. Tournebize, L. Liger, C. Margoum, C. Morbois, A. L. Achard, J. F. Ouvry, F. Pierlot, H. Dubaele et C. Catalogne (2019). « Formation Zones tampons : limiter les transferts de contaminants. Formation complète (3 modules) ». Cours. France.

Carsel E., J. et J. E. Baldwin (2000). *PRZM-3, A model for predicting Pesticide and nitrogen fate in the crop root and unsaturated soil zones*. Rapp. tech. Environmental Protection Agency.

Catalogne, C., C. Lauvernet et N. Carluer (2018). *Guide d'utilisation de l'outil BUVARD pour le dimensionnement des bandes tampons végétalisées destinées à limiter les transferts de pesticides par ruissellement*. Rapp. tech. Agence française pour la biodiversité.

Catalogne, C. et G. Le Hénaff (2016). *Guide d'aide à l'implantation des zones tampons pour l'atténuation des transferts de contaminants d'origine agricole*. Rapport Irstea-ONEMA élaboré dans le cadre du Groupe Technique Zones Tampons. irstea, p. 69.

Chabot, V., M. Nodet, N. Papadakis et A. Vidard (2015). « Accounting for observation errors in image data assimilation ». *Tellus A* 67. doi : 10.3402/tellusa.v67.23629.

Chabot, V., M. Nodet et A. Vidard (2020). « Multiscale Representation of Observation Error Statistics in Data Assimilation ». *Sensors* 20.5, p. 1-20. doi : 10.3390/s20051460.

Chastaing, G., F. Gamboa et C. Prieur (2015). « Generalized Sobol sensitivity indices for dependent variables : numerical methods ». *Journal of Statistical Computation and Simulation* 85.7, p. 1306-1333. doi : 10.1080/00949655.2014.960415.

Cornuéjols, A., L. Miclet et V. Barra (2021). *Apprentissage artificiel : concepts et algorithmes - De Bayes et Hume au Deep Learning*. Sous la dir. d'Editions Eyrolles.

Corpen (2007). *Les fonctions environnementales des zones tampons -Les bases scientifiques et techniques des fonctions de protection des eaux, première édition, Comité d'orientation pour des pratiques agricoles respectueuses de l'environnement.* Sous la dir. de Ministère de l'Ecologie,de l'Environnement, du Développement Durable et de l'Aménagement du Territoire.

Cosme, E., J. M. Brankart, J. Verron, P. Brasseur et M. Krysta (2010). « Implementation of a reduced rank square-root smoother for high resolution ocean data assimilation ». *Ocean Modelling* 33.1, p. 87-100. doi : 10.1016/j.ocemod.2009.12.004.

Coutadeur, C., Y. Coquet et J. Roger-Estrade (2002). « Variation of hydraulic conductivity in a tilled soil ». *European Journal of Soil Science* 53.4, p. 619-628. doi : 10.1046/j.1365-2389.2002.00473.x.

Crestani, E., M. Camporese, D. Baú et P. Salandin (2013). « Ensemble Kalman filter versus ensemble smoother for assessing hydraulic conductivity via tracer test data assimilation ». *Hydrology and Earth System Sciences* 17.4, p. 1517-1531. doi : 10.5194/hess-17-1517-2013.

Crow, W. T. et D. Ryu (2009). « A new data assimilation approach for improving runoff prediction using remotely-sensed soil moisture retrievals ». *Hydrology and Earth System Sciences* 13.1, p. 1-16. doi : 10.5194/hess-13-1-2009.

Cui, F., J. Bao, Z. Cao, L. Li et Q. Zheng (2020). « Soil hydraulic parameters estimation using ground penetrating radar data via ensemble smoother with multiple data assimilation ». *Journal of Hydrology* 583, p. 124552. doi : 10.1016/j.jhydrol.2020.124552.

Da Veiga, S (2015). « Global sensitivity analysis with dependence measures ». *Journal of Statistical Computation and Simulation* 85.7, p. 1283-1305. doi : 10.1080/00949655.2014.945932.

Da Veiga, S. (2021). « Kernel-based ANOVA decomposition and Shapley effects - Application to global sensitivity analysis ». working paper or preprint.

Da Veiga, S., F. Gamboa, B. Iooss et C. Prieur (2021). *Basics and Trends in Sensitivity Analysis*. Sous la dir. de Society for Industrial et Applied Mathematics, p. 291. doi : 10.1137/1.9781611976694.

Dairon, R. (2015). « Identification des processus dominants de transfert des produits phytosanitaires dans le sol et évaluation de modèles numériques pour des contextes agro-pédo-climatiques variés ». Thèse de doct. Université Claude Bernard - Lyon 1.

Darcy, H. (1857). *Recherches expérimentales relatives au mouvement de l'eau dans les tuyaux*. Sous la dir. de Mallet - Bachelier.

De Lozzo, M. et A. Marrel (2014). « New improvements in the use of dependence measures for sensitivity analysis and screening ». *Journal of Statistical Computation and Simulation*. doi : 10.1080/00949655.2016.1149854.

De Lozzo, M. et A. Marrel (2016). « Sensitivity analysis with dependence and variance-based measures for spatio-temporal numerical simulators ». *Stochastic Environmental Research and Risk Assessment* 31. doi : 10.1007/s00477-016-1245-3.

de Rocquigny, E. (2006). « La maîtrise des incertitudes dans un contexte industriel. 1re partie : une approche méthodologique globale basée sur des exemples ». *Journal de la société française de statistique* 147.3, p. 33-71.

Devers, A., J. P. Vidal, C. Lauvernet, B. Graff et O. Vannier (2020). « A framework for high-resolution meteorological surface reanalysis through offline data assimilation in an ensemble of downscaled reconstructions ». *Quarterly Journal of the Royal Meteorological Society* 146.726, p. 153-173. doi : 10.1002/qj.3663.

Djabelkhir, K., C. Lauvernet, P. Kraft et N. Carluer (2016). « Development of a dual permeability model within a hydrological catchment modeling framework : 1D application ». *Science of the Total Environment* 575.

Dosskey, M. G., M. J. Helmers et D. E. Eisenhauer (2011). « A design aid for sizing filter strips using buffer area ratio ». *Journal of Soil and Water Conservation* 66.1, p. 29-39. doi : 10.2489/jswc.66.1.29.

Doucet, A., N. De Freitas et N. J. Gordon (2001). *Sequential Monte Carlo methods in practice*. T. 1. 2. Springer. doi : 10.1007/978-1-4757-3437-9.

Druart, C., M. Millet, R. Scheifler, O. Delhomme, C. Raeppel et A. De Vaufleury (2011). « Snails as indicators of pesticide drift, deposit, transfer and effects in the vineyard ». *Science of the total environment* 409.20, p. 4280-4288. doi : 10.1016/j.scitotenv.2011.07.006.

Dubus, I. G. et C. D. Brown (2002). « Sensitivity and First-Step Uncertainty Analyses for the Preferential Flow Model MACRO ». *Journal of Environmental Quality* 31.1, p. 227-240. doi : 10.2134/jeq2002.2270.

Dubus, I. G., C. D. Brown et S. Beulke (2003). « Sensitivity analyses for four pesticide leaching models ». *Pest Management Science* 59.9, p. 962-982. doi : 10.1002/ps.723.

Dunne, S. et D. Entekhabi (2005). « An ensemble-based reanalysis approach to land data assimilation ». *Water resources research* 41.2. doi : 10.1029/2004WR003449.

Dunne, T. et R. Black (1970). « An Experimental Investigation of Runoff Production in Permeable Soils ». *Water Resources Research* 6, p. 478-490. doi : 10.1029/WR006i002p00478.

Durand, C. (2014). « Modélisation du transfert de pesticides à l'échelle de la parcelle. Application au bassin versant de la Morcille (Nord Beaujolais, 69) et analyse de sensibilité du modèle ». Mém. de mast. ENGEES.

El Gharamti, M., J. L. McCreight, S. J. Noh, T. J. Hoar, A. RafieeiNasab et B. K. Johnson (2021). « Ensemble streamflow data assimilation using WRF-Hydro and DART : novel localization and inflation techniques applied to Hurricane Florence floo-

ding ». *Hydrology and Earth System Sciences* 25.9, p. 5315-5336. DOI : 10.5194/hess-25-5315-2021.

El Hajj, M., N. Baghdadi, M. Zribi et H. Bazzi (2017). « Synergic use of Sentinel-1 and Sentinel-2 images for operational soil moisture mapping at high spatial resolution over agricultural areas ». *Remote Sensing* 9.12, p. 1292. DOI : 10.3390/rs9121292.

Emerick, A. et A. Reynolds (2013). « Ensemble Smoother with Multiple Data Assimilation ». *Computers & Geosciences* 55, p. 3-15. DOI : 10.1016/j.cageo.2012.03.011.

Evensen, G. (1992). « Using the extended Kalman filter with a multilayer quasi-geostrophic ocean model ». *Journal of Geophysical Research : Oceans* 97.C11, p. 17905-17924. DOI : doi.org/10.1029/92JC01972.

Evensen, G. (1994). « Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics ». *Journal of Geophysical Research : Oceans* 99.C5, p. 10143-10162. DOI : 10.1029/94JC00572.

Evensen, G. (1997). « Advanced data assimilation for strongly nonlinear dynamics ». *Monthly Weather Review* 125.6, p. 1342-1354. DOI : 10.1175/1520-0493(1997)125<1342:ADAFSN>2.0.CO;2.

Evensen, G. (2003). « The ensemble Kalman filter : Theoretical formulation and practical implementation ». *Ocean dynamics* 53.4, p. 343-367. DOI : 10.1007/s10236-003-0036-9.

Fabre, J. C., X. Louchart, F. Colin, C. Dagès, R. Moussa, M. Rabotin, D. Raclot, P. Lagacherie et M. Voltz (2010). « OpenFLUID : a software environment for modelling fluxes in landscapes ». *LandMod 2010 : International Conference on Integrative Landscape Modelling . 2010 ; International Conference on Integrative Landscape Modelling, Montpellier, France, 1-13*.

Faivre, R., B. Iooss, S. Mahévas, D. Makowski et H. Monod (2013). *Analyse de sensibilité et exploration de modèles*. Collection Savoir-Faire. Editions Quae, 352 p.

FAO (2022). *Agri-environmental Indicators / Pesticides*. Sous la dir. de FAOSTAT.

Fatichi, S., E. Vivoni, F. Ogden, V. Ivanov, B. Mirus, D. Gochis, C. Downer, M. Camporese, J. Davison, B. Ebel, N. Jones, J. Kim, G. Mascaro, R Niswonger, P. Restrepo, R. Rigon, C. Shen, M. Sulis et D. Tarboton (2016). « An overview of current applications, challenges, and future trends in distributed process-based models in hydrology ». *Journal of Hydrology* 537, p. 45-60. DOI : 10.1016/j.jhydrol.2016.03.026.

Feddes, R. A, P. J. Kowalik et H. Zaradny (1978). *Simulation of field water use and crop yield*. English. Wageningen : Pudoc for the Centre for Agricultural Publishing et Documentation.

FOCUS (2001). *FOCUS Surface Water Scenarios in the EU Evaluation Process under 91/414/EEC*. Report of the FOCUS Working Group on Surface Water Scenarios, EC Document Reference SANCO/4802/2001. European commission.

Fouilloux, A. et A. Piacentini (1999). « The PALM Project : MPMD Paradigm for an Oceanic Data Assimilation Software ». *Euro-Par'99 Parallel Processing : 5th International Euro-Par Conference Toulouse, France, August 31 - September 3, 1999 Proceedings*. Berlin, Heidelberg : Springer Berlin Heidelberg, p. 1423-1430.

Fox, G. A., R. Muñoz-Carpena et R. A. Purvis (2018). « Controlled laboratory experiments and modeling of vegetative filter strips with shallow water tables ». *Journal of Hydrology* 556, p. 1-9. DOI : 10.1016/j.jhydrol.2017.10.069%7D.

Fox, G. A., R. Muñoz-Carpena et G. J. Sabbagh (2010). « Influence of flow concentration on parameter importance and prediction uncertainty of pesticide trapping by vegetative filter strips ». *Journal of Hydrology* 384.1, p. 164-173. DOI : 10.1016/j.jhydrol.2010.01.020.

Frésard, F. (2010). « Cartographie des sols d'un petit bassin versant en Beaujolais viticole, en appui à l'évaluation du risque de contamination des eaux par les pesticides ». Mém. de mast. Université de Franche Comté.

Freundlich, H. (1909). *Kapillarchemie, eine Darstellung der Chemie der Kolloide und verwandter Gebiete*. akademische Verlagsgesellschaft.

Fukumizu, K., A. Gretton, G. Lanckriet, B. Schölkopf et B. K. Sriperumbudur (2009). « Kernel Choice and Classifiability for RKHS Embeddings of Probability Distributions ». *Advances in Neural Information Processing Systems*. T. 22. Curran Associates, Inc.

Gamboa, F., A. Janon, T. Klein et A. Lagnoux (2013). « Sensitivity indices for multivariate outputs ». *Comptes Rendus Mathématiques* 351.7, p. 307-310. DOI : 10.1016/j.crma.2013.04.016.

Gangeh, M. J., H. Zarkoob et A. Ghodsi (2017). « Fast and Scalable Feature Selection for Gene Expression Data Using Hilbert-Schmidt Independence Criterion ». *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 14.1, p. 167-181. DOI : 10.1109/TCBB.2016.2631164.

Gao, B., M. T. Walter, T. S. Steenhuis, W. L. Hogarth et J. Y. Parlange (2004). « Rainfall induced chemical transport from soil to runoff : theory and experiments ». *Journal of Hydrology* 295.1, p. 291-304. DOI : 10.1016/j.jhydrol.2004.03.026.

Garcia, D., I. Arostegui et R. Prellezo (2019). « Robust combination of the Morris and Sobol methods in complex multidimensional models ». *Environmental Modelling and Software* 122, p. 104517. DOI : 10.1016/j.envsoft.2019.104517.

Gaspari, G. et S. E. Cohn (1999). « Construction of correlation functions in two and three dimensions ». *Quarterly Journal of the Royal Meteorological Society* 125.554, p. 723-757. doi : 10.1002/qj.49712555417.

Gatel, L. (2018). « Construction et évaluation d'un modèle de transport de contaminants r éactifs coupl é surface-subsurface à l' échelle du versant ». Thèse de doct. Université Grenoble Alpes.

Gatel, L., C. Lauvernet, N. Carluer, S. Weill et C. Paniconi (2019). « Sobol Global Sensitivity Analysis of a Coupled Surface/Subsurface Water Flow and Reactive Solute Transfer Model on a Real Hillslope ». *Water* 12, p. 121. doi : 10.3390/w12010121.

Gatel, L., C. Lauvernet, N. Carluer, S. Weill, J. Tournebize et C. Paniconi (2018). « Global evaluation and sensitivity analysis of a physically based flow and reactive transport model on a laboratory experiment ». *Environmental Modelling and Software* 113, p. 73-83. doi : 10.1016/j.envsoft.2018.12.006.

Geist, M., O. Pietquin et G. Fricout (jan. 2010). « Astuce du Noyau & Quantification Vectorielle ». *Actes du 17ème colloque sur la Reconnaissance des Formes et l'Intelligence Artificielle*.

Ghanem, R. G. et P. D. Spanos (1991). « Spectral stochastic finite-element formulation for reliability analysis ». *Journal of Engineering Mechanics* 117.10, p. 2351-2372.

Gordon, N. J., D. J. Salmond et Smith A. F. M. (1993). « Novel approach to nonlinear/non-Gaussian Bayesian state estimation ». *IEE Proceedings F (Radar and Signal Processing)* 140 (2), p. 107-113. doi : 10.1049/ip-f-2.1993.0015.

Gouy, V., L. Liger, S. Ahrouch, C. Bonnineau, N. Carluer, A. Chaumot, M. Coquery, A. Dabrin, C. Margoum et S. Pesce (2021). « Ardières-Morcille in the Beaujolais, France : A research catchment dedicated to study of the transport and impacts of diffuse agricultural pollution in rivers ». *Hydrological Processes* 35.10, e14384. doi : 10.1002/hyp.14384.

Gouy, V., L. Liger, N. Carluer et C. Margoum (2015). *BDOH, Site Atelier Ardières Morcille*. Sous la dir. d'Irstea.

Gregorutti, B., B. Michel et P. Saint-Pierre (2017). « Correlation and variable importance in random forests ». *Statistics and Computing* 27, p. 659-678. doi : 10.1007/s11222-016-9646-1.

Gretton, A., R. Herbrich, A. Smola, O. Bousquet et B. Schölkopf (2005). « Kernel Methods for Measuring Independence ». *Journal of Machine Learning Research* 6, p. 2075-2129.

Grillot, J., M. Rabotin, V. Gouy, N. Carluer et C. Lauvernet (2022). « GEOMELBA - outil pédagogique pour la visualisation des transferts de produits phytosanitaires à la surface d'un bassin versant ». *50e congrès du Groupe Français de recherche sur les Pesticides*.

Hamill, T. M. (2001). « Interpretation of Rank Histograms for Verifying Ensemble Forecasts ». *Monthly Weather Review* 129.3, p. 550-560. doi : 10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2.

Hamill, T. M., J. S. Whitaker et C. Snyder (2001). « Distance-Dependent Filtering of Background Error Covariance Estimates in an Ensemble Kalman Filter ». *Monthly Weather Review* 129.11, p. 2776-2790. doi : 10.1175/1520-0493(2001)129<2776:DDFOBE>2.0.CO;2.

Harper, E. B., J. C. Stella et A. K. Fremier (2011). « Global sensitivity analysis for complex ecological models : a case study of riparian cottonwood population dynamics ». *Ecological Applications* 21.4, p. 1225-1240. doi : 10.1890/10-0506.1.

Hendricks Franssen, H. J. et W. Kinzelbach (2008). « Real-time groundwater flow modeling with the Ensemble Kalman Filter : Joint estimation of states and parameters and the filter inbreeding problem ». *Water Resources Research* 44.9. doi : 10.1029/2007WR006505.

Herman, J. D., J. B. Kollat, P. M. Reed et T. Wagener (2013). « Technical Note : Method of Morris effectively reduces the computational demands of global sensitivity analysis for distributed watershed models ». *Hydrology and Earth System Sciences* 17.7, p. 2893-2903. doi : 10.5194/hess-17-2893-2013.

Hersbach, H. (2000). « Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems ». *Weather and Forecasting* 15.5, p. 559-570. doi : 10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2.

Hong, T. et S. T. Purucker (2018). « Spatiotemporal sensitivity analysis of vertical transport of pesticides in soil ». *Environmental Modelling and Software*, p. 24-38. doi : 10.1016/j.envsoft.2018.03.018.

Horton, R. E. (1933). « The role of infiltration in the hydrologic cycle ». *Eos, Transactions American Geophysical Union* 14.1, p. 446-460. doi : 10.1029/TR014i001p00446.

Houtekamer, P. L. et H. L. Mitchell (2001). « A sequential ensemble Kalman filter for atmospheric data assimilation ». *Monthly Weather Review* 129.1, p. 123-137.

Hunt, B. R., E. K. Kostelich et I. Szunyogh (2007). « Efficient data assimilation for spatiotemporal chaos : A local ensemble transform Kalman filter ». *Physica D : Nonlinear Phenomena* 230.1, p. 112-126. doi : 10.1016/j.physd.2006.11.008.

Ishwaran, H. et U. Kogalur (2020). *Fast Unified Random Forests for Survival, Regression,and Classification (RF-SRC). R package version 2.9.3.*

Jazwinski, A. H. (1970). *Stochastic processes and filtering theory*. eng. Mathematics in science and engineering ; v. 64. New York : Academic Press.

Jimenez, S., G. Dueñas, A. Gelbukh, C. Rodríguez et S. Mancera (2018). « Automatic Detection of Regional Words for Pan-Hispanic Spanish on Twitter : » *16th Ibero-American Conference on AI, Trujillo, Peru*, p. 404-416. doi : 10.1007/978-3-030-03928-8_33.

Johnson, M. E., L. M. Moore et D. Ylvisaker (1990). « Minimax and maximin distance designs ». *Journal of Statistical Planning and Inference* 26.2, p. 131-148. doi : 10.1016/0378-3758(90)90122-B.

Jurić, D. (p. d.). *Object Tracking : Particle Filter with Ease*. Sous la dir. de Code Project. url : https://www.codeproject.com/Articles/865934/Object-Tracking-Particle-Filter-with-Ease#opticalFlow.

Kalman, R. E. (1960). « A New Approach to Linear Filtering and Prediction Problems ». *Transactions of the ASME–Journal of Basic Engineering* 82.Series D, p. 35-45.

Khare, Y. P., R. Muñoz-Carpena, R. W. Rooney et C. J. Martinez (2015). « A multi-criteria trajectory-based parameter sampling strategy for the screening method of elementary effects ». *Environmental Modelling and Software* 64, p. 230-239. doi : 10.1016/j.envsoft.2014.11.013.

Kollet, S. J. et R. M. Maxwell (2006). « Integrated surface-groundwater flow modeling : A free-surface overland flow boundary condition in a parallel groundwater flow model ». *Advances in Water Resources* 29.7, p. 945-958. doi : 10.1016/j.advwatres.2005.08.006.

Kralisch, S. et P. Krause (2006). « JAMS - A framework for natural resource model development and application ».

Lacas, J. G. (2005). « Processus de dissipation des produits phytosanitaires dans les zones tampons enherbées. Etude expérimentale et modélisation en vue de limiter la contamination des eaux de surface ». Thèse de doct. CEMAGREF.

Lacas, J. G., M. Voltz, V. Gouy, N. Carluer et J. J. Gril (2005). « Using grassed strips to limit pesticide transfer to surface water : a review ». *Agronomy for Sustainable Development* 25.2, p. 253-266.

Lamboni, M., H. Monod et D. Makowski (2011). « Multivariate sensitivity analysis to measure global contribution of input factors in dynamic models ». *Reliability Engineering and System Safety* 96.4, p. 450-459. doi : 10.1016/j.ress.2010.12.002.

Larsbo, M. et N. Jarvis (2003). *MACRO 5.0 : A Model of Water Flow and Solute Transport in Macroporous Soil : Technical Description*. Emergo (Uppsala). Department of Soil Sciences, Swedish University of Agricultural Sciences.

Lauvernet, C., F. Baret, L. Hascoet, S. Buis et F. X. Le Dimet (2008). « Multitemporal-patch ensemble inversion of coupled surface-atmosphere radiative transfer models for land surface characterization ». *Remote Sens. Environ.* 112.3, p. 851-861.

Lauvernet, C. et R. Muñoz-Carpena (2018). « Shallow water table effects on water, sediment, and pesticide transport in vegetative filter strips – Part 2 : model coupling, application, factor importance, and uncertainty ». *Hydrology and Earth System Sciences* 22.1, p. 71-87. doi : 10.5194/hess-22-71-2018.

Le Dimet, F. X., W. Castaings, P. Ngnepieba et B. Vieux (2009). « Data Assimilation in hydrology : Variational Approach ». *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications*. Sous la dir. de Seon Park et Liang Xu. Springer, p. 367-407. doi : 10.1007/978-3-540-71056-1_20.

Le Dimet, F. X. et O. Talagrand (1986). « Variational algorithms for analysis and assimilation of meteorological observations : theoretical aspects ». *Tellus A : Dynamic Meteorology and Oceanography* 38.2, p. 97-110.

Le Dreau, M., C. Margoum, C. Guillemain, A. Martin, L. Liger, N. Mazzella et V. Gouy (2018). « Les échantillonneurs passifs pour évaluer la contamination en pesticides des eaux de surface : intérêts et limites actuelles au transfert opérationnel vers les gestionnaires ». *48eme Congrès du Groupe Français des Pesticides*. Limoges, France.

Le Gratiet, L. S., S. Marelli et B. Sudret (2017). « Metamodel-based sensitivity analysis : polynomial chaos expansions and Gaussian processes ». *Handbook on Uncertainty Quantification*. Cham, Switzerland. Springer International Publishing. Chap. 38, p. 1289-1325.

Lefrancq, M., G. Imfeld, S. Payraudeau et M. Millet (2013). « Kresoxim methyl deposition, drift and runoff in a vineyard catchment ». *Science of The Total Environment* 442, p. 503-508. doi : /10.1016/j.scitotenv.2012.09.082.

Lei, F., W. T. Crow, W. P. Kustas, J. Dong, Y. Yang, K. R. Knipper, M. C. Anderson, F. Gao, C. Notarnicola, F. Greifeneder, L. M. McKee, J. G. Alfieri, C. Hain et N. Dokoozlian (2020). « Data assimilation of high-resolution thermal and radar remote sensing retrievals for soil moisture monitoring in a drip-irrigated vineyard ». *Remote Sensing of Environment* 239, p. 111622. doi : https://doi.org/10.1016/j.rse.2019.111622.

Lewis, K. A., J. Tzilivakis, D. J. Warner et A. Green (2016). « An international database for pesticide risk assessments and management ». *Human and Ecological Risk*

Assessment : An International Journal 22.4, p. 1050-1064. DOI : 10.1080/10807039.2015.1133242.

Li, K. Y., J. B. Boisvert et R. De Jong (1999). « An exponential root-water-uptake model ». *Canadian Journal of Soil Science* 79.2, p. 333-343. DOI : 10.4141/S98-032.

Li, K. Y., R De Jong et J. B. Boisvert (2001). « An exponential root-water-uptake model with water stress compensation ». *Journal of Hydrology* 252, p. 189-204. DOI : 10.1016/S0022-1694(01)00456-5.

Lozac'h, L., H. Bazzi, N. Baghdadi, M. El Hajj et M. Zribi (2020). « Sentinel-1/Sentinel-2 derived soil moisture product at plot scale (S2MP) ». *Mediterranean and Middle-East Geoscience and Remote Sensing Symposium*.

Madrigal, I., P. Benoit, E. Barriuso, B. Real, A. Dutertre, M. Moquet, M. Trejo et L. Ortiz (2007). « Pesticide degradation in vegetative buffer strips : Grassed and tree barriers : Case of isoproturon ». *Agrociencia* 41, p. 205-217.

Magnant, C. (2016). « Approches bayésiennes pour le pistage radar de cibles de surface potentiellement manoeuvrantes ». Thèse de doct. Université de Bordeaux.

Mander, Ü. et J. Tournebize (2015). « Riparian Buffer Zones : functions and dimensioning ». *Reference Module in Earth Systems and Environmental Sciences*. Elsevier. DOI : 10.1016/B978-0-12-409548-9.09304-0.

Marelli, S. et B. Sudret (2014). « UQLab : A framework for uncertainty quantication in Matlab ». *Proc. 2nd Int. Conf. on Vulnerability, Risk Analysis and Management (ICVRAM2014)*.

Marelli, S. et B. Sudret (2018). « An active-learning algorithm that combines sparse polynomial chaos expansions and bootstrap for structural reliability analysis ». *Structural Safety* 75, p. 67-74. DOI : 10.1016/j.strusafe.2018.06.003.

Marrel, A., N. Marie et M. De Lozzo (2015). « Advanced surrogate model and sensitivity analysis methods for sodium fast reactor accident assessment ». *Reliability Engineering and System Safety* 138, p. 232-241. DOI : 10.1016/j.ress.2015.01.019.

Martin, A. (2016). « Développement de matériaux innovants à base d'élastomère de silicone pour l'échantillonnage passif de pesticides dans les eaux de surface et de subsurface ». Thèse de doct. Université de Lyon.

McKay, M. D., R. J. Beckman et W. J. Conover (1979). « A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code ». *Technometrics* 21.2, p. 239-245.

Meynaoui, A., A. Marrel et B. Laurent-Bonneau (2018). « Méthodologie basée sur les mesures de dépendance HSIC pour l'analyse de sensibilité de second niveau ». *50èmes Journées de Statistique (JdS2018)*. Palaiseau, France.

Miles, J. C. (1985). « The representation of flows to partially penetrating rivers using groundwater flow models ». *Journal of Hydrology* 82, p. 341-355. doi : 10.1016/0022-1694(85)90026-5.

Moradkhani, H., K. L. Hsu, H. Gupta et S. Sorooshian (2005). « Uncertainty assessment of hydrologic model states and parameters : Sequential data assimilation using the particle filter ». *Water Resources Research* 41.5. doi : 10.1029/2004WR003604.

Morris, M. D. (1991). « Factorial Sampling Plans for Preliminary Computational Experiments ». *Technometrics* 33.2, p. 161-174. doi : 10.1080/00401706.1991.10484804.

Mualem, Y. (1976). « A new model for predicting the hydraulic conductivity of unsaturated porous media ». *Water Resources Research* 12, p. 513-522.

Muñoz-Carpena, R., G. Fox et G. Sabbagh (2010). « Parameter Importance and Uncertainty in Predicting Runoff Pesticide Reduction with Filter Strips ». *Journal of environmental quality* 39, p. 630-41. doi : 10.2134/jeq2009.0300.

Muñoz-Carpena, R., C. Lauvernet et N. Carluer (2018). « Shallow water table effects on water, sediment, and pesticide transport in vegetative filter strips – Part 1 : nonuniform infiltration and soil water redistribution ». *Hydrology and Earth System Sciences* 22.1, p. 53-70. doi : 10.5194/hess-22-53-2018.

Muñoz-Carpena, R., J. E. Parsons et J. W. Gilliam (1999). « Modeling hydrology and sediment transport in vegetative filter strips ». *Journal of Hydrology* 214.1, p. 111-129. doi : 10.1016/S0022-1694(98)00272-8.

Nie, S., J. Zhu et Y. Luo (2011). « Simultaneous estimation of land surface scheme states and parameters using the ensemble Kalman filter : identical twin experiments ». *Hydrology and Earth System Sciences* 15.8, p. 2437-2457. doi : 10.5194/hess-15-2437-2011.

Nossent, J. et W. Bauwens (2012). « Multi-variable sensitivity and identifiability analysis for a complex environmental model in view of integrated water quantity and water quality modeling ». *Water Science and Technology* 65.3, p. 539-549. doi : 10.2166/wst.2012.884.

Novello, P., T. Fel et D. Vigouroux (2022). « Making Sense of Dependence : Efficient Black-box Explanations Using Dependence Measure ». *Advances in Neural Information Processing Systems (NeurIPS)*. New Orleans, United States.

Ott, E., B. Hunt, I. Szunyogh, A. Zimin, E. Kostelich, M. Corazza, E. Kalnay, D. Patil et J. Yorke (2004). « A Local Ensemble Kalman Filter for Atmospheric Data Assimilation ». *Tellus*. doi : 10.1111/j.1600-0870.2004.00076.x.

Paniconi, C. et M. Putti (1994). « A comparison of Picard and Newton iteration in the numerical solution of multidimensional variably saturated flow problems ». *Water Resources Research* 30.12, p. 3357-3374–. doi : 10.1029/94WR02046.

Pasetto, D., M. Camporese et M. Putti (2012). « Ensemble Kalman filter versus particle filter for a physically-based coupled surface–subsurface model ». *Advances in Water Resources* 47, p. 1-13. doi : 10.1016/j.advwatres.2012.06.009.

Pasetto, D., G. Y. Niu, L. Pangle, C. Paniconi, M. Putti et P. A. Troch (2015). « Impact of sensor failure on the observability of flow dynamics at the Biosphere 2 LEO hillslopes ». *Advances in Water Resources* 86, p. 327-339. doi : 10.1016/j.advwatres.2015.04.014.

Penny, S. G. et T. M. Hamill (2017). « Coupled Data Assimilation for Integrated Earth System Analysis and Prediction ». *Bulletin of the American Meteorological Society* 98.7, ES169-ES172. doi : 10.1175/BAMS-D-17-0036.1.

Peyrard, X. (2016). « Assessment of pesticide transfer in subsurface lateral flow on a sloping vineyard in Beaujolais : field monitoring, tracing experiment and modeling ». Thèse de doct. Université de Lyon.

Pham, D. T. (2001). « Stochastic Methods for Sequential Data Assimilation in Strongly Nonlinear Systems ». *Monthly Weather Review* 129.5, p. 1194-1207. doi : 10.1175/1520-0493(2001)129<1194:SMFSDA>2.0.CO;2.

Poletika, N., P. Coody, G. Fox, J. G. Sabbagh, D. S. Dolder et J. White (2009). « Chlorpyrifos and Atrazine Removal from Runoff by Vegetated Filter Strips : Experiments and Predictive Modeling ». *Journal of Environmental Quality* 38, p. 1042-52. doi : 10.2134/jeq2008.0404.

Raanes, P. N., Y. Chen, C. Grudzien, M. Tondeur et R. Dubois (2018). *DAPPER*. Sous la dir. de Nansen Environmental et Remote Sensing Center (NERSC). doi : 10.5281/zenodo.2029296.

Radišić, K., E. Rouzies, C. Lauvernet et A. Vidard (2021). « Sensitivity Analysis of a Spatio-Temporal Hydrological Model for Pesticide Transfers ». *Rencontres MEXICO 2021*. Toulouse, France.

Radišić, K., E. Rouzies, C. Lauvernet et A. Vidard (2022). « Global sensitivity analysis of a distributed hydrological model at the catchment scale ». working paper or preprint.

Ramsay, J. O. et B. W. Silverman (2005). *Principal components analysis for functional data*. Springer New York, NY. 429 p. doi : 10.1007/b98888.

Randriambololohasinirina, P. (2012). « Pesticide dissipation properties in soils of a wine-growing watershed. » Mém. de mast. Université Pierre et Marie Curie (Paris 6) ; Institut des Sciences et Industries du Vivant et de l'Environnement.

Reichenberger, S., M. Bach, A. Skitschak et H. G. Frede (2007). « Mitigation strategies to reduce pesticide inputs into ground- and surface water and their effectiveness ; A review ». *Science of The Total Environment* 384.1-3, p. 1-35. doi : 10.1016/j.scitotenv.2007.04.046.

Richards, A. L. (1931). « Capillary conduction of liquids in porous mediums ». *Physics* 1, p. 318-333. doi : 10.1063/1.1745010.

Rochoux, M. C., S. Ricci, D. Lucor, C. Benedicte et A. Trouve (2014). « Towards predictive data-driven simulations of wildfire spread – Part I : Reduced-cost Ensemble Kalman Filter based on a Polynomial Chaos surrogate model for parameter estimation ». *Natural Hazards and Earth System Sciences* 14, p. 1-23. doi : 10.5194/nhess-14-2951-2014.

Rodriguez-Galiano, V., M. P. Mendes, M. J. Garcia-Soldado, M. Chica-Olmo et L. Ribeiro (2014). « Predictive modeling of groundwater nitrate pollution using Random Forest and multisource variables related to intrinsic and specific vulnerability : A case study in an agricultural setting (Southern Spain) ». *Science of The Total Environment* 476-477, p. 189-206. doi : 10.1016/j.scitotenv.2014.01.001.

Rosenblatt, M. (1956). « Remarks on some nonparametric estimates of a density function ». *The annals of mathematical statistics*, p. 832-837.

Ross, P. J. (2003). « Modeling Soil Water and Solute Transport - Fast, Simplified Numerical Solutions ». *Agronomy Journal* 95.6, p. 1352-1361. doi : 10.2134/agronj2003.1352.

Ross, P. J. (2006). *Fast Solution of Richards' Equation for Flexible Soil Hydraulic Property Descriptions*. Rapp. tech. CSIRO.

Roux, S., S. Buis, F. Lafolie et M. Lamboni (2021). « Cluster-based GSA : Global sensitivity analysis of models with temporal or spatial outputs using clustering ». *Environmental Modelling and Software* 140, p. 105046. doi : 10.1016/j.envsoft.2021.105046.

Rouzies, E., C. Lauvernet, C. Barachet, T. Morel, F. Branger, I. Braud et N. Carluer (2019). « From agricultural catchment to management scenarios : A modular tool to assess effects of landscape features on water and pesticide behavior ». *Science of The Total Environment* 671, p. 1144-1160. doi : 10.1016/j.scitotenv.2019.03.060.

Ruano, M. V., J. Ribes, A. Seco et J. Ferrer (2012). « An improved sampling strategy based on trajectory design for application of the Morris method to systems with many input factors ». *Environmental Modelling and Software* 37, p. 103-109. doi : 10.1016/j.envsoft.2012.03.008.

Saint-Geours, N., J. S. Bailly, F. Grelot et C. Lavergne (2014). « Multi-scale spatial sensitivity analysis of a model for economic appraisal of flood risk management policies ». en. *Environmental Modelling and Software* 60, p. 153-166. doi : 10.1016/j.envsoft.2014.06.012.

Saltelli, A., K. Aleksankina, W. Becker, P. Fennell, F. Ferretti, N. Holst, S. Li et Q. Wu (2019). « Why so many published sensitivity analyses are false : A systematic review of sensitivity analysis practices ». *Environmental Modelling and Software* 114, p. 29-39. doi : 10.1016/j.envsoft.2019.01.012\%7D.

Saltelli, A., F. Campolongo et J. Cariboni (2009). « Screening important inputs in models with strong interaction properties ». *Reliability Engineering and System Safety* 94.7. Special Issue on Sensitivity Analysis, p. 1149-1155. doi : 10.1016/j.ress.2008.10.007.

Saltelli, A., M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana et S. Tarantola (2008). *Global sensitivity analysis : the primer*. John Wiley & Sons.

Saltelli, A., S. Tarantola, F. Campolongo et M. Ratto (2004). *Sensitivity Analysis in Practice : A Guide to Assessing Scientific Models*. Wiley. doi : 10.1002/0470870958.

Schaap, M. G. et M. T. van Genuchten (2006). « A modified Mualem-van Genuchten formulation for improved description of the hydraulic conductivity near saturation ». *Vadose Zone Journal* 5, p. 27-34. doi : 10.2136/vzj2005.0005.

Schwen, A., G. Bodner, P. Scholl, G. Buchan et W. Loiskandl (2011). « Temporal dynamics of soil hydraulic properties and the water-conducting porosity under different tillage ». *Soil and Tillage Research* 113, p. 89-98. doi : 10.1016/j.still.2011.02.005.

SDES (2020). *Eau et milieux aquatiques. Les chiffres clés. Édition 2020*. Rapp. tech. Ministère de la transition écologique.

Seki, K. (2007). « SWRC fit &ndash ; a nonlinear fitting program with a water retention curve for soils having unimodal and bimodal pore structure ». *Hydrology and Earth System Sciences Discussions* 4, p. 407-437. doi : 10.5194/hessd-4-407-2007.

Šimůnek, J., M. T. Van Genuchten et M. Šejna (1998). *HYDRUS-1D - Simulating the one-dimensional movement of water, heat, and multiple solutes in variably-saturated media*. Rapp. tech. U.S. Salinity Lab., Riverside, CA.

Smart, D. R., E. Schwass, A. Lakso et L. Morano (2006). « Grapevine rooting patterns : A comprehensive analysis and a review ». *American Journal of Enology and Viticulture* 57, p. 89-104.

Sobol, I. M. (1993). « Sensitivity estimates for nonlinear mathematical models ». *Mathematical Modelling and Computational Experiments* 1.4, p. 407-414.

Soleimani, F. (2021). « Analytical seismic performance and sensitivity evaluation of bridges based on random decision forest framework ». *Structures* 32, p. 329-341. doi : 10.1016/j.istruc.2021.02.049.

Stewart, L. M., S. L. Dance et N. K. Nichols (2013). « Data assimilation with correlated observation errors : experiments with a 1-D shallow water model ». *Tellus A : Dynamic Meteorology and Oceanography* 65.1, p. 19546. doi : 10.3402/tellusa.v65i0.19546.

Sudret, B. (2008). « Global sensitivity analysis using polynomial chaos expansions ». *Reliability Engineering and System Safety* 93.7, p. 964-979. doi : 10.1016/j.ress.2007.04.002.

Sudret, B. et S. Marelli (2020). « Course on Uncertainty Quantification and Data Analysis in Applied Sciences ».

Tarantola, S., N. Giglioli, J. Jesinghaus et A. Saltelli (2002). « Can global sensitivity analysis steer the implementation of models for environmental assessments and decision-making ? » *Stochastic Environmental Research and Risk Assessment* 16, p. 63-76. doi : 10.1007/s00477-001-0085-x.

Taylor, A. W. et W. F. Spencer (1990). « Volatilization and Vapor Transport Processes ». *Pesticides in the Soil Environment : Processes, Impacts and Modeling*. John Wiley & Sons, Ltd. Chap. 7, p. 213-269. doi : 10.2136/sssabookser2.c7.

Todaro, V., M. D'Oria, M. G. Tanda et J. J. Gómez-Hernández (2021). « Ensemble smoother with multiple data assimilation to simultaneously estimate the source location and the release history of a contaminant spill in an aquifer ». *Journal of Hydrology* 598, p. 126215. doi : 10.1016/j.jhydrol.2021.126215.

Van den Bogaert, R. (2011). « Typologie des sols du bassin versant de la Morcille, caractérisation de leur propriétés hydrauliques et test de fonctions de pédotransfert ». Mém. de mast. Université Pierre et Marie Curie, AgroParisTech.

van Leeuwen, P. J. et G. Evensen (1996). « Data Assimilation and Inverse Methods in Terms of a Probabilistic Formulation ». *Monthly Weather Review* 124.12, p. 2898-2913. doi : 10.1175/1520-0493(1996)124<2898:DAAIMI>2.0.CO;2.

Vanuytrecht, E., D. Raes et P. Willems (2014). « Global sensitivity analysis of yield output from the water productivity model ». *Environmental Modelling and Software* 51, p. 323-332. doi : 10.1016/j.envsoft.2013.10.017.

Varado, N. (2004). « Contribution to the development of a distributed hydrological modelling. Application to the Donga catchment, in Benin ». Thèse de doct. Institut national polytechnique de Grenoble, p. 320-.

Waller, J., S. Dance et N. Nichols (2017). « On diagnosing observation error statistics with local ensemble data assimilation ». *Quarterly Journal of the Royal Meteorological Society* 143. doi : 10.1002/qj.3117.

Walter, M. T., B. Gao et J. Y. Parlange (2007). « Modeling soil solute release into runoff with infiltration ». *Journal of Hydrology* 347.3, p. 430-437. doi : 10.1016/j.jhydrol.2007.09.033.

Wan-Duo, K. M., J. P. Lewis et W. Bastiaan Kleijn (2018). « Blind Facial Basis Discovery Using the Hilbert-Schmidt Independence Criterion ». *2018 International Conference on Image and Vision Computing New Zealand (IVCNZ)*, p. 1-5. doi : 10.1109/IVCNZ.2018.8634682.

Whitaker, J. S. et T. M. Hamill (2002). « Ensemble data assimilation without perturbed observations ». *Monthly weather review* 130.7, p. 1913-1924. doi : 10.1175/1520-0493(2002)130<1913:EDAWPO>2.0.CO;2.

Xie, X. et D. Zhang (2010). « Data assimilation for distributed hydrological catchment modeling via ensemble Kalman filter ». *Advances in Water Resources* 33.6, p. 678-690. doi : 10.1016/j.advwatres.2010.03.012.

Xiu, D. et G. E. Karniadakis (2002). « The Wiener–Askey polynomial chaos for stochastic differential equations ». *SIAM Journal on Scientific Computing* 24.2, p. 619-644. doi : 10.1137/S1064827501387826.

Xu, J., F. Anctil et M. A. Boucher (2022). « Exploring hydrologic post-processing of ensemble streamflow forecasts based on affine kernel dressing and non-dominated sorting genetic algorithm II ». *Hydrology and Earth System Sciences* 26.4, p. 1001-1017. doi : 10.5194/hess-26-1001-2022.

Yang, J. (2011). « Convergence and uncertainty analyses in Monte-Carlo based sensitivity analysis ». *Environmental Modelling and Software* 26.4, p. 444-457. doi : 10.1016/j.envsoft.2010.10.007.

Zajac, Z. B. (2010). « Global sensitivity and uncertainty analysis of spatially distributed watershed models ». Thèse de doct. University of Florida.

Annexe A

Densités de probabilité des paramètres d'entrée

Horizons de sol		
soilhorizon_thetas_2	[m^3m^{-3}]	N(0.3362, 0.0336)
soilhorizon_thetar_2	[m^3m^{-3}]	TN(0.0510, 0.0128, 0, 1)
soilhorizon_Ks_2	[ms^{-1}]	LN(-9.38,-1.88)
soilhorizon_hg_2	[m]	N(-0.0329,0.00329)
soilhorizon_mn_2	[-]	N(0.1988, 0.0199)
soilhorizon_Ko_2	[ms^{-1}]	LN(-15.31,-3.06)
soilhorizon_L_2	[-]	U(-7.8216, -5.2144)
soilhorizon_bd_2	[gcm^{-3}]	U(1.1768, 1.7652)
soilhorizon_moc_2	[gg^{-1}]	U(0.0024, 0.0054)
soilhorizon_thetas_3	[m^3m^{-3}]	N(0.3202, 0.0320)
soilhorizon_thetar_3	[m^3m^{-3}]	TN(0.0812, 0.0203, 0, 1)
soilhorizon_Ks_3	[ms^{-1}]	LN(-9.85, -1.97)
soilhorizon_hg_3	[m]	N(-0.0209, 0.00209)
soilhorizon_mn_3	[-]	N(0.2046, 0.0205)
soilhorizon_Ko_3	[ms^{-1}]	LN(-14.13,-2.83)
soilhorizon_L_3	[-]	U(-5.0844, -3.3896)
soilhorizon_bd_3	[gcm^{-3}]	U(1.2536, 1.8804)
soilhorizon_moc_3	[gg^{-1}]	U(0.0006, 0.0014)
soilhorizon_thetas_4	[m^3m^{-3}]	N(0.2844, 0.0284)
soilhorizon_thetar_4	[m^3m^{-3}]	TN(0.0661, 0.0165, 0, 1)
soilhorizon_Ks_4	[ms^{-1}]	LN(-10.40,-2.08)
soilhorizon_hg_4	[m]	N(-0.0599,0.00599)
soilhorizon_mn_4	[-]	N(0.2274, 0.0227)
soilhorizon_Ko_4	[ms^{-1}]	LN(-13.45,-2.69)
soilhorizon_L_4	[-]	U(-0.1716, -0.1144)

soilhorizon_bd_4	[gcm^{-3}]	U(1.2240, 1.8360)
soilhorizon_moc_4	[gg^{-1}]	U(4.3840 10^{-4},9.6160 10^{-3})
soilhorizon_thetas_6	[m^{3}m^{-3}]	N(0.3537, 0.0354)
soilhorizon_thetar_6	[m^{3}m^{-3}]	TN(0, 0.0093, 0, 1)
soilhorizon_Ks_6	[ms^{-1}]	LN(-10.77,-2.15)
soilhorizon_hg_6	[m]	N(0.066,0.0066)
soilhorizon_mn_6	[-]	N(0.1289, 0.0129)
soilhorizon_Ko_6	[ms^{-1}]	LN(-14.97,-3.00)
soilhorizon_L_6	[-]	U(7.7240, 19.3100)
soilhorizon_bd_6	[gcm^{-3}]	U(1.2704, 1.9056)
soilhorizon_moc_6	[gg^{-1}]	U(0.0042, 0.0094)
soilhorizon_thetas_7	[m^{3}m^{-3}]	N(0.3247, 0.0325)
soilhorizon_thetar_7	[m^{3}m^{-3}]	TN(0, 0.0093, 0, 1)
soilhorizon_Ks_7	[ms^{-1}]	LN(-11.57,-2.31)
soilhorizon_hg_7	[m]	N(-0.0718,0.00718)
soilhorizon_mn_7	[-]	N(0.0751, 0.0075)
soilhorizon_Ko_7	[ms^{-1}]	LN(-15.63,-3.13)
soilhorizon_L_7	[-]	U(-12, -8)
soilhorizon_bd_7	[gcm^{-3}]	U(1.3256, 1.9884)
soilhorizon_moc_7	[gg^{-1}]	U(0.0019, 0.0051)
soilhorizon_thetas_8	[m^{3}m^{-3}]	N(0.4162, 0.0416)
soilhorizon_thetar_8	[m^{3}m^{-3}]	TN(0, 0.0093, 0, 1)
soilhorizon_Ks_8	[ms^{-1}]	LN(-12.45,-2.49)
soilhorizon_hg_8	[m]	N(-0.3018, 0.03018)
soilhorizon_mn_8	[-]	N(0.10000, 0.0100)
soilhorizon_Ko_8	[ms^{-1}]	LN(-16.17,-3.23)
soilhorizon_L_8	[-]	U(8, 12)
soilhorizon_bd_8	[gcm^{-3}]	U(1.2304, 1.8456)
soilhorizon_moc_8	[gg^{-1}]	U(0.0018, 0.0037)
soilhorizon_thetas_9	[m^{3}m^{-3}]	N(0.3322, 0.0332)
soilhorizon_thetar_9	[m^{3}m^{-3}]	TN(0.0770, 0.0192, 0, 1)
soilhorizon_Ks_9	[ms^{-1}]	LN(-10.41,-2.08)
soilhorizon_hg_9	[m]	N(-0.0671,0.00671)
soilhorizon_mn_9	[-]	N(0.2582, 0.0258)
soilhorizon_Ko_9	[ms^{-1}]	LN(-14.93,-2.99)
soilhorizon_L_9	[-]	U(0.3376, 0.8440)
soilhorizon_bd_9	[gcm^{-3}]	U(1.1664, 1.7496)
soilhorizon_moc_9	[gg^{-1}]	U(0.0023, 0.0051)
soilhorizon_thetas_10	[m^{3}m^{-3}]	N(0.3160, 0.0316)
soilhorizon_thetar_10	[m^{3}m^{-3}]	TN(0.0612, 0.0153, 0, 1)

*soilhorizon_Ks*_10	[ms^{-1}]	LN(-10.67,-2.13)
*soilhorizon_hg*_10	[m]	N(-0.0356, 0.00356)
*soilhorizon_mn*_10	[-]	N(0.1791, 0.0179)
*soilhorizon_Ko*_10	[ms^{-1}]	LN(-15.04,-3.01)
*soilhorizon_L*_10	[-]	U(0.8376, 2.0940)
*soilhorizon_bd*_10	[gcm^{-3}]	U(1.2984, 1.9476)
*soilhorizon_moc*_10	[gg^{-1}]	U(0.0025, 0.0055)
*soilhorizon_thetas*_11	[m^{3}m^{-3}]	N(0.3375, 0.0338)
*soilhorizon_thetar*_11	[m^{3}m^{-3}]	TN(0.0372, 0.0093, 0, 1)
*soilhorizon_Ks*_11	[ms^{-1}]	LN(-10.16,-2.03)
*soilhorizon_hg*_11	[m]	N(-0.0969,0.00969)
*soilhorizon_mn*_11	[-]	N(0.2685, 0.0268)
*soilhorizon_Ko*_11	[ms^{-1}]	LN(-15.09,-3.02)
*soilhorizon_L*_11	[-]	U(-10.1124, -6.7416)
*soilhorizon_bd*_11	[gcm^{-3}]	U(1.0752, 1.6128)
*soilhorizon_moc*_11	[gg^{-1}]	U(0.0049, 0.0050)
*soilhorizon_thetas*_14	[m^{3}m^{-3}]	N(0.3375, 0.0338)
*soilhorizon_thetar*_14	[m^{3}m^{-3}]	TN(0.0372, 0.0093, 0, 1)
*soilhorizon_Ks*_14	[ms^{-1}]	LN(-10.11,-2.02)
*soilhorizon_hg*_14	[m]	N(-0.0969,0.00969)
*soilhorizon_mn*_14	[-]	N(0.2685, 0.0268)
*soilhorizon_Ko*_14	[ms^{-1}]	LN(-15.09,-3.02)
*soilhorizon_L*_14	[-]	U(-10.1124, -6.7416)
*soilhorizon_bd*_14	[gcm^{-3}]	U(1.0752, 1.6128)
*soilhorizon_moc*_14	[gg^{-1}]	U(0.0175, 0.0385)
*soilhorizon_thetas*_12	[m^{3}m^{-3}]	N(0.3375, 0.0338)
*soilhorizon_thetar*_12	[m^{3}m^{-3}]	TN(0.0372, 0.0093, 0, 1)
*soilhorizon_Ks*_12	[ms^{-1}]	LN(-10.16,-2.03)
*soilhorizon_hg*_12	[m]	N(-0.0969,0.00969)
*soilhorizon_mn*_12	[-]	N(0.2685, 0.0268)
*soilhorizon_Ko*_12	[ms^{-1}]	LN(-15.09,-3.02)
*soilhorizon_L*_12	[-]	U(-10.1124, -6.7416)
*soilhorizon_bd*_12	[gcm^{-3}]	U(1.0752, 1.6128)
*soilhorizon_moc*_12	[gg^{-1}]	U(0.0072, 0.0158)
*soilhorizon_thetas*_15	[m^{3}m^{-3}]	N(0.3375, 0.0338)
*soilhorizon_thetar*_15	[m^{3}m^{-3}]	TN(0.0372, 0.0093, 0, 1)
*soilhorizon_Ks*_15	[ms^{-1}]	LN(-10.11,-2.02)
*soilhorizon_hg*_15	[m]	N(-0.0969,0.00969)
*soilhorizon_mn*_15	[-]	N(0.2685, 0.0268)
*soilhorizon_Ko*_15	[ms^{-1}]	LN(-15.09,-3.02)

soilhorizon_L_15	[-]	U(-10.1124, -6.7416)
soilhorizon_bd_15	[gcm^{-3}]	U(1.0752, 1.6128)
soilhorizon_moc_15	[gg^{-1}]	U(0.0175, 0.0385)
soilhorizon_thetas_13	[m^{3}m^{-3}]	N(0.3375, 0.0338)
soilhorizon_thetar_13	[m^{3}m^{-3}]	TN(0.0372, 0.0093, 0, 1)
soilhorizon_Ks_13	[ms^{-1}]	LN(-10.16,-2.03)
soilhorizon_hg_13	[m]	N(-0.0969,0.00969)
soilhorizon_mn_13	[-]	N(0.2685, 0.0268)
soilhorizon_Ko_13	[ms^{-1}]	LN(-15.09,-3.02)
soilhorizon_L_13	[-]	U(-10.1124, -6.7416)
soilhorizon_bd_13	[gcm^{-3}]	U(1.0752, 1.6128)
soilhorizon_moc_13	[gg^{-1}]	U(0.0067, 0.0080)
soilhorizon_thetas_16	[m^{3}m^{-3}]	N(0.3375, 0.0338)
soilhorizon_thetar_16	[m^{3}m^{-3}]	TN(0.0372, 0.0093, 0, 1)
soilhorizon_Ks_16	[ms^{-1}]	LN(-10.11,-2.02)
soilhorizon_hg_16	[m]	N(-0.0969,0.00969)
soilhorizon_mn_16	[-]	N(0.2685, 0.0268)
soilhorizon_Ko_16	[ms^{-1}]	LN(-15.09,-3.02)
soilhorizon_L_16	[-]	U(-10.1124, -6.7416)
soilhorizon_bd_16	[gcm^{-3}]	U(1.0752, 1.6128)
soilhorizon_moc_16	[gg^{-1}]	U(0.0175, 0.0385)
Végétation		
veget_manning_1	[sm$^{-1/3}$]	T(0.0250, 0.0330, 0.041)
veget_Zr_1	[m]	U(2.096,3.144)
veget_F10_1	[-]	U(0.2960, 0.4440)
veget_LAImin_1	[-]	U(0.0080, 0.0120)
veget_LAImax_1	[-]	U(2, 3)
veget_LAIharv_1	[-]	U(0.0080, 0.0120)
veget_manning_2	[sm$^{-1/3}$]	T(0.1000, 0.2000, 0.3000)
veget_Zr_2	[m]	U(0.72,1.08)
veget_F10_2	[-]	U(0.2680, 0.4020)
veget_LAI_2	[-]	U(4, 6)
Rivière		
river_hpond	[m]	U(0.008,0.012)
river_di	[m]	U(1.2, 1.8)
river_Ks	[ms^{-1}]	LN(-10.67, -2.13)
river_manning	[sm$^{-1/3}$]	T(0.061, 0.079, 0.097)
UH		
plot_hpond	[m]	U(0.008, 0.012)
vfz_hpond	[m]	U(0.04, 0.06)

hu_adsorpthick	[m]	U(0.005, 0.015)
	Pesticide	
pest_Koc	[mLg^{-1}]	T(461.4000, 538.3000, 769.0000)
*pest_DT*50	[d]	N(47.1, 28.26)

Tableau A.1 – Densités de probabilité associées aux 145 paramètres d'entrée. Les fonctions sont décrites par leur moyenne μ et leur écart-type σ pour les lois LN (lognormale) et N (normale), par le minimum, le centre et le maximum pour la loi T (triangulaire) et par le minimum et le maximum pour la loi U (Uniforme). Les noms de paramètres suivent la syntaxe suivante : XXX_XXX_XXX où le premier bloc indique le type d'élément auquel le paramètre fait référence (horizon de sol, type de végétation, rivière, parcelle de vigne, bande enherbée ou type de pesticide), le second bloc indique le nom du paramètre et le troisième bloc, s'il existe, indique l'indice de l'horizon de sol ou du type de végétation considéré.

Annexe B

Fonctions noyaux et RKHS

Dans cette annexe, les définitions des fonctions noyaux et des espaces de Hilbert à noyaux reproduisants (RKHS) sont introduites de manière aussi qualitative que possible, largement basé sur Cornuéjols et al. (2021). Pour une présentation formelle approfondie de ces concepts, le lecteur est invité à se référer à Berlinet et Thomas-Agnan (2004).

On rappelle pour commencer la définition d'un espace de Hilbert :

Espace de Hilbert

Un espace de Hilbert est un espace vectoriel sur $\mathbb{R}$ ou $\mathbb{C}$ muni d'un produit scalaire et qui est complet [a] pour la norme associée.

a. un espace de fonctions $\mathcal{F}$ est *complet* si toute suite de Cauchy $\{f_n\}_{n\geq 1}$ d'éléments de $\mathcal{F}$ converge vers un élément $f \in \mathcal{F}$.

Les espaces de Hilbert sont ainsi des espaces vectoriels de dimension **infinie** les plus simples. Dans ces travaux, on s'intéresse aux espaces de Hilbert fonctionnels.

On considère ensuite $\mathcal{X}$ un ensemble énumérable et on définit comme suit une fonction noyau :

Fonction noyau

Une fonction $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ est une fonction noyau si :

1. elle est symétrique i.e. pour tout $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$, $\kappa(\mathbf{x}, \mathbf{x}') = \kappa(\mathbf{x}', \mathbf{x})$
2. elle est positive semi-définie i.e. pour tout $n \in \mathbb{N}$, $\mathbf{x}_1, ..., \mathbf{x}_n \in \mathcal{X}$ et $\alpha_1, ..., \alpha_n \in \mathbb{R}$,

$$\sum_{i,j=1}^{n} \alpha_i \alpha_j \kappa(\mathbf{x}_i, \mathbf{x}_j) \geq 0$$

Une autre définition des fonctions noyaux est donnée ci-dessous. Celle-ci fait intervenir la notion d'espace de redescription, c'est-à-dire un espace, en général de plus grande dimension que l'espace des données $\mathcal{X}$ qui préserve le groupement inhérent des données tout en simplifiant la structure associée (Geist et al., 2010).

fonction noyau - un autre point de vue

Une fonction noyau est une fonction $\kappa : \mathbf{x}, \mathbf{x}' \in \mathcal{X}^2 \to \mathbb{R}$ satisfaisant :

$$\kappa(\mathbf{x}, \mathbf{x}') = \langle \Phi(\mathbf{x}), \Phi(\mathbf{x}') \rangle$$

où Φ est une fonction de $\mathcal{X}$ vers un espace de redescription $\mathcal{F}$ doté d'un produit scalaire :

$$\Phi : \mathbf{x} \to \Phi(\mathbf{x}) \in F$$

Cette définition montre que l'utilisation d'une fonction noyau permet implicitement de remplacer un produit scalaire dans l'espace de redescription de grande dimension (voire de dimension infinie) par un calcul n'impliquant qu'un nombre fini de termes. Dans notre cas, cet espace est un espace de Hilbert de fonctions à valeurs réelles $f : \mathcal{X} \to \mathbb{R}$ associé au produit scalaire $\langle \cdot, \cdot \rangle_{\mathcal{F}}$.

On peut ainsi définir un espace de Hilbert à noyau reproduisant :

Espace de Hilbert à noyau reproduisant (RKHS)

L'espace de Hilbert $\mathcal{F}$ de fonctions réelles définies sur $\mathcal{X}$ et doté du produit scalaire $\langle \cdot, \cdot \rangle_{\mathcal{F}}$ est un espace de Hilbert à noyau reproduisant (RKHS) s'il existe une fonction $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ ayant les propriétés suivantes :

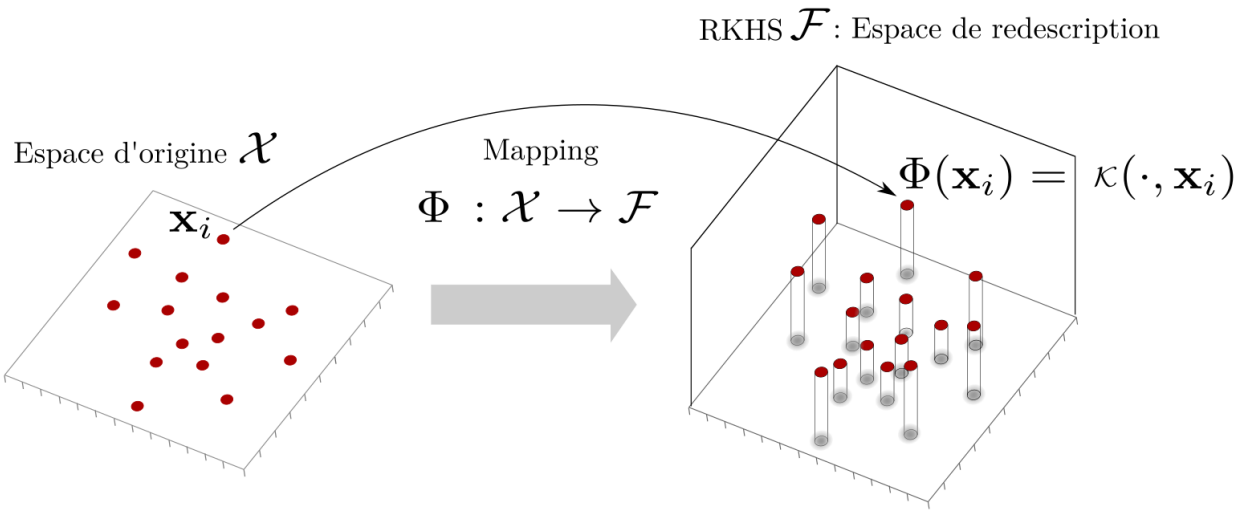
1. pour tout élément $\mathbf{x} \in \mathcal{X}$, $\kappa(\mathbf{x}, \cdot)$ appartient à $\mathcal{F}$,
2. la fonction κ est une fonction noyau reproduisante, c'est-à-dire telle que pour toute fonction $f \in \mathcal{F}$, on a $\langle f, \kappa(\mathbf{x}, \cdot) \rangle_{\mathcal{F}} = f(\mathbf{x})$ (propriété de reproduction).

Ainsi, le fait que la fonction noyau soit *reproduisante* signifie que toute fonction $f \in \mathcal{F}$ peut être exprimée sous forme de produit scalaire.

Théorème 1 : Bijection entre noyau et RKHS associé

Chaque fonction positive semi-définie $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ définit un unique RKHS dont elle est le noyau reproduisant.

Les propriétés principales d'un RKHS sont résumées sur la Figure B.1.



Produit scalaire $\langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle_{\mathcal{F}} = \kappa(\mathbf{x}_i, \mathbf{x}_j)$

Propriété de reproduction $f(\mathbf{x}) = \langle f, \kappa(\mathbf{x}, \cdot) \rangle_{\mathcal{F}}, \forall f \in \mathcal{F}$

Figure B.1 – Illustration schématique du lien entre espace des données et RKHS associé et principales propriétés de ce dernier.

Annexe C

Résultats du criblage (variables scalaires)

WaterLateralFlow	SoluteLateralFlow	WaterSurfaceRunoff	SoluteSurfaceRunoff
soilhorizon_thetas_2	*soilhorizon_thetas_12*	*soilhorizon_thetas_11*	*soilhorizon_thetas_11*
soilhorizon_thetas_4	*soilhorizon_thetas_15*	*soilhorizon_thetas_15*	*soilhorizon_thetas_12*
soilhorizon_thetas_6	*soilhorizon_thetas_2*	*soilhorizon_thetas_2*	*soilhorizon_thetas_13*
soilhorizon_thetas_7	*soilhorizon_thetas_4*	*soilhorizon_thetas_4*	*soilhorizon_thetas_15*
soilhorizon_thetas_8	*soilhorizon_thetas_6*	*soilhorizon_thetas_6*	*soilhorizon_thetas_2*
soilhorizon_thetas_10	*soilhorizon_thetas_7*	*soilhorizon_thetas_7*	*soilhorizon_thetas_4*
soilhorizon_thetar_3	*soilhorizon_thetas_8*	*soilhorizon_thetas_8*	*soilhorizon_thetas_6*
soilhorizon_thetar_4	*soilhorizon_thetas_10*	*soilhorizon_thetas_10*	*soilhorizon_thetas_7*
soilhorizon_thetar_8	*soilhorizon_thetar_2*	*soilhorizon_thetar_15*	*soilhorizon_thetas_8*
soilhorizon_thetar_10	*soilhorizon_thetar_4*	*soilhorizon_thetar_2*	*soilhorizon_thetas_10*
soilhorizon_moc_13	*soilhorizon_thetar_8*	*soilhorizon_thetar_4*	*soilhorizon_thetar_15*
soilhorizon_mn_3	*soilhorizon_thetar_10*	*soilhorizon_thetar_8*	*soilhorizon_thetar_2*
soilhorizon_mn_4	*soilhorizon_pore_6*	*soilhorizon_thetar_10*	*soilhorizon_thetar_4*
soilhorizon_mn_6	*soilhorizon_moc_12*	*soilhorizon_pore_9*	*soilhorizon_thetar_8*
soilhorizon_mn_8	*soilhorizon_moc_15*	*soilhorizon_mn_11*	*soilhorizon_thetar_10*
soilhorizon_mn_10	*soilhorizon_moc_2*	*soilhorizon_mn_2*	*soilhorizon_moc_6*
soilhorizon_Kx_3	*soilhorizon_moc_6*	*soilhorizon_mn_4*	*soilhorizon_moc_12*
soilhorizon_Kx_4	*soilhorizon_moc_9*	*soilhorizon_mn_6*	*soilhorizon_mn_11*
soilhorizon_Kx_8	*soilhorizon_mn_11*	*soilhorizon_mn_7*	*soilhorizon_mn_16*
soilhorizon_Kx_10	*soilhorizon_mn_16*	*soilhorizon_mn_8*	*soilhorizon_mn_2*
soilhorizon_Ks_11	*soilhorizon_mn_4*	*soilhorizon_mn_10*	*soilhorizon_mn_4*
soilhorizon_Ks_13	*soilhorizon_mn_6*	*soilhorizon_Kx_8*	*soilhorizon_mn_6*
soilhorizon_Ks_14	*soilhorizon_mn_8*	*soilhorizon_Kx_10*	*soilhorizon_mn_7*
soilhorizon_Ks_15	*soilhorizon_mn_10*	*soilhorizon_Ks_12*	*soilhorizon_mn_8*
soilhorizon_Ks_16	*soilhorizon_Kx_12*	*soilhorizon_Ks_13*	*soilhorizon_mn_10*

soilhorizon_Ks_3	*soilhorizon_Kx_9*	*soilhorizon_Ks_15*	*soilhorizon_Ks_15*
soilhorizon_Ks_4	*soilhorizon_Kx_10*	*soilhorizon_Ks_16*	*soilhorizon_Ks_2*
soilhorizon_Ks_6	*soilhorizon_Ks_12*	*soilhorizon_Ks_4*	*soilhorizon_Ks_4*
soilhorizon_Ks_7	*soilhorizon_Ks_14*	*soilhorizon_Ks_6*	*soilhorizon_Ks_8*
soilhorizon_Ks_8	*soilhorizon_Ks_15*	*soilhorizon_Ks_8*	*soilhorizon_Ks_9*
soilhorizon_Ks_9	*soilhorizon_Ks_16*	*soilhorizon_Ks_9*	*soilhorizon_hg_3*
soilhorizon_Ks_10	*soilhorizon_Ks_2*	*soilhorizon_Ks_10*	*soilhorizon_hg_4*
soilhorizon_hg_16	*soilhorizon_Ks_4*	*soilhorizon_hg_4*	*soilhorizon_hg_8*
soilhorizon_hg_2	*soilhorizon_Ks_6*	*soilhorizon_hg_6*	*soilhorizon_bd_6*
soilhorizon_hg_4	*soilhorizon_Ks_8*	*soilhorizon_hg_8*	*soilhorizon_bd_13*
soilhorizon_hg_8	*soilhorizon_Ks_9*	*soilhorizon_hg_10*	*soilhorizon_bd_12*
soilhorizon_hg_9	*soilhorizon_Ks_10*	*soilhorizon_bd_3*	*river_ks*
soilhorizon_bd_2	*soilhorizon_hg_4*	*river_ks*	*river_di*
river_ks	*soilhorizon_hg_6*	*river_di*	*plot_hpond*
river_di	*soilhorizon_hg_8*	*plot_hpond*	*pest_Koc_1*
plot_hpond	*soilhorizon_hg_10*	*vfz_hpond*	*pest_DT50_1*
VFS_hpond	*soilhorizon_bd_11*	*veget_LAIharv_1*	*HU_adsorpthick*
	soilhorizon_bd_12	*veget_F10_1*	*veget_Zr_1*
	soilhorizon_bd_13		*veget_manning_1*
	soilhorizon_bd_15		*veget_F10_1*
	soilhorizon_bd_2		
	soilhorizon_bd_6		
	soilhorizon_bd_9		
	river_ks		
	river_di		
	plot_hpond		
	veget_Zr_1		
	pest_Koc_1		
	pest_DT50_1		

Tableau C.1 – Paramètres restants après criblage sur chaque variable scalaire. Dans la syntaxe XXX_XXX_XXX des noms de paramètres, le premier bloc indique le type d'élément auquel le paramètre fait référence (horizon de sol, rivière, végétation, pesticide, HU), la deuxième partie est le nom du paramètre tandis que la troisième est l'indice de l'élément auquel le paramètre fait référence (horizon de sol ou type de végétation).

Annexe D

Analyse d'incertitude de l'humidité en profondeur

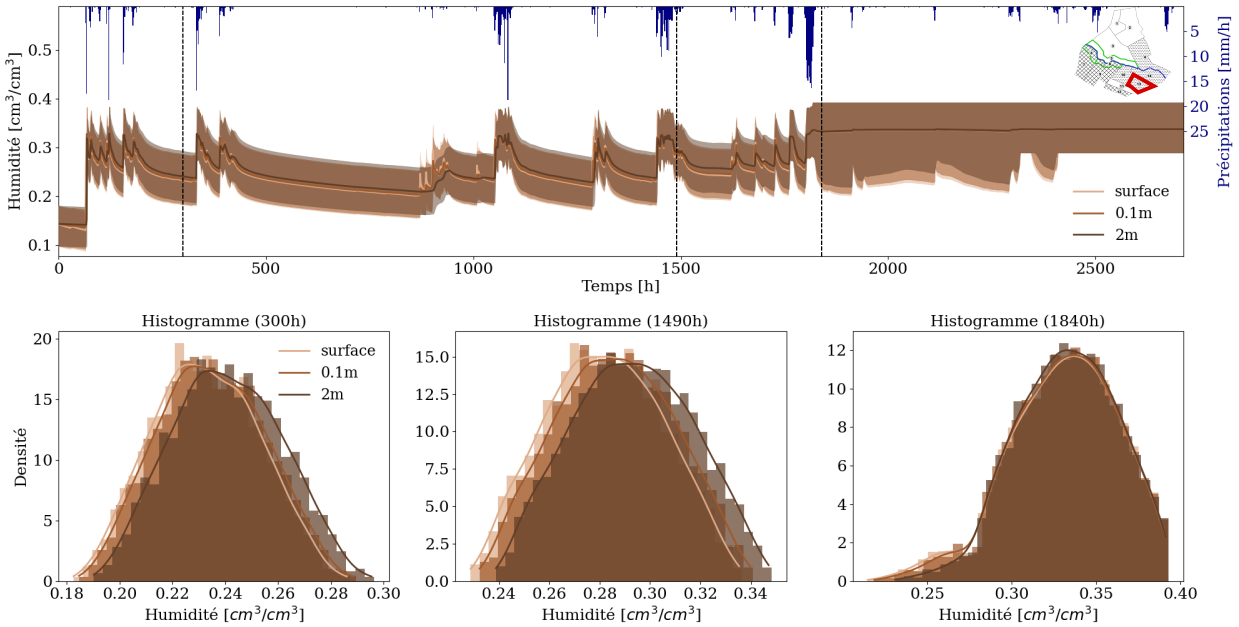


Figure D.1 – Haut : séries temporelles moyennes d'humidité en surface, à 0.1 m et à 2 m de profondeur et enveloppes entre les valeurs min et max associées pour l'UH13 (scénario hivernal). Les traits verticaux en pointillés indiquent les différents instants de la simulation pour lesquels sont tracés les histogrammes. A partir de 1800 h, les valeurs moyennes et les bornes max aux 3 profondeurs sont les mêmes d'où la quasi-superposition observée. Bas : histogrammes ponctuels et pdf empiriques estimées par noyau gaussien.

Annexe E

Résultats du criblage (variables temporelles)

Humidité de surface (été)	Humidité de surface (hiver)	Concentration à l'exutoire
soilhorizon_Ks_13	soilhorizon_Ks_12	soilhorizon_Ks_10
soilhorizon_Kx_11	soilhorizon_Ks_14	soilhorizon_Ks_12
soilhorizon_Kx_12	soilhorizon_Ks_15	soilhorizon_Ks_14
soilhorizon_Kx_13	soilhorizon_Ks_16	soilhorizon_Ks_4
soilhorizon_Kx_14	soilhorizon_Ks_6	soilhorizon_Ks_7
soilhorizon_Kx_15	soilhorizon_Ks_7	soilhorizon_Ks_8
soilhorizon_Kx_16	soilhorizon_Ks_8	soilhorizon_Ks_9
soilhorizon_Kx_6	soilhorizon_Ks_9	soilhorizon_Kx_10
soilhorizon_Kx_9	soilhorizon_Kx_12	soilhorizon_Kx_6
soilhorizon_hg_11	soilhorizon_Kx_7	soilhorizon_hg_10
soilhorizon_hg_12	soilhorizon_hg_11	soilhorizon_hg_12
soilhorizon_hg_13	soilhorizon_hg_12	soilhorizon_hg_4
soilhorizon_hg_14	soilhorizon_hg_13	soilhorizon_hg_8
soilhorizon_hg_15	soilhorizon_hg_14	soilhorizon_mn_10
soilhorizon_hg_16	soilhorizon_hg_15	soilhorizon_mn_15
soilhorizon_hg_9	soilhorizon_hg_16	soilhorizon_mn_4
soilhorizon_mn_11	soilhorizon_hg_8	soilhorizon_thetar_10
soilhorizon_mn_12	soilhorizon_mn_10	soilhorizon_thetar_4
soilhorizon_mn_13	soilhorizon_mn_11	soilhorizon_thetas_10
soilhorizon_mn_14	soilhorizon_mn_12	soilhorizon_thetas_11
soilhorizon_mn_15	soilhorizon_mn_13	soilhorizon_thetas_12
soilhorizon_mn_16	soilhorizon_mn_14	soilhorizon_thetas_13
soilhorizon_thetar_11	soilhorizon_mn_15	soilhorizon_thetas_2
soilhorizon_thetar_12	soilhorizon_mn_16	soilhorizon_thetas_4
soilhorizon_thetar_13	soilhorizon_mn_4	soilhorizon_thetas_7

soilhorizon_thetar_14	soilhorizon_mn_6	soilhorizon_thetas_8
soilhorizon_thetar_15	soilhorizon_mn_8	hu_adsorpthick
soilhorizon_thetar_16	soilhorizon_thetar_10	pest_DT50_1
soilhorizon_thetar_9	soilhorizon_thetar_11	plot_hpond
soilhorizon_thetas_10	soilhorizon_thetar_12	river_di
soilhorizon_thetas_11	soilhorizon_thetar_13	river_ks
soilhorizon_thetas_12	soilhorizon_thetar_14	veget_LAImax_1
soilhorizon_thetas_13	soilhorizon_thetar_15	veget_manning_1
soilhorizon_thetas_14	soilhorizon_thetar_4	veget_manning_2
soilhorizon_thetas_15	soilhorizon_thetar_6	vfz_hpond
soilhorizon_thetas_16	soilhorizon_thetar_8	
soilhorizon_thetas_2	soilhorizon_thetas_10	
soilhorizon_thetas_9	soilhorizon_thetas_11	
veget_F10_1	soilhorizon_thetas_12	
veget_F10_2	soilhorizon_thetas_13	
veget_LAI_2	soilhorizon_thetas_14	
veget_LAImax_1	soilhorizon_thetas_15	
veget_Zr_1	soilhorizon_thetas_16	
veget_Zr_2	soilhorizon_thetas_2	
	soilhorizon_thetas_4	
	soilhorizon_thetas_6	
	soilhorizon_thetas_8	
	soilhorizon_thetas_9	
	plot_hpond	
	river_di	
	river_ks	
	veget_LAImax_1	

Tableau E.1 – Paramètres restants après criblage sur chaque variable temporelle (voir légende du Tableau C.1 pour la description de la syntaxe des paramètres).

Annexe F

Présentation du filtre particulaire

Les performances de l'EnKF et ses variantes présentées dans ce manuscrit peuvent se trouver fortement affectées en présence d'un système fortement non linéaire/non gaussien. Dans ce cas, il est possible de se tourner vers une autre classe de méthodes dont le filtre particulaire est la plus largement utilisée.

Les filtres particulaires (Gordon et al., 1993 ; Doucet et al., 2001 ; van Leeuwen et Evensen, 1996) sont des méthodes ensemblistes qui implémentent également l'approche séquentielle du filtrage bayésien en alternant étapes de prévision et d'analyse. L'analyse diffère cependant de celle de l'EnKF puisqu'elle ne modifie pas les membres. Chaque membre de l'ensemble est comparé aux observations et pondéré en fonction de sa ressemblance à ces dernières. Les pondérations sont calculées à partir du théorème de Bayes et permettent d'affecter des poids élevés aux membres qui sont proches des observations et des poids faibles aux membres qui en sont loin. Dans un second temps, l'ensemble est rééchantillonné, en sélectionnant majoritairement les membres avec les poids les plus élevés. Le nouvel ensemble est ensuite éventuellement perturbé pour augmenter sa dispersion puis propagé par le modèle jusqu'à la prochaine analyse. Le principe du filtre particulaire est résumé sur la Figure F.1.

Le filtre particulaire est une méthode facile à implémenter et qui a l'avantage de ne pas modifier les membres de l'ensemble, garantissant que ceux-ci respectent la physique du modèle. Il permet également de résoudre le problème d'estimation bayésienne complet puisqu'il ne fait pas d'hypothèse sur la forme de la distribution de l'état. Cependant, la flexibilité d'une telle approche de rééchantillonnage se fait au prix d'une extrême sensibilité à la dégénérescence, c'est-à-dire que tous les membres sont susceptibles de finir par se ressembler. Dans ce cas, le filtre ne peut plus prendre en compte les observations car l'information provenant du modèle manque de dispersion. Plusieurs méthodes existent pour empêcher une telle dégénérescence du filtre (Gordon et al., 1993 ; Pham, 2001 ; Doucet et al., 2001), mais elles impliquent avant tout de considérer des ensembles de taille importante, ce qui constitue souvent un obstacle à l'application pratique du filtre particulaire. C'est pour cette raison que l'EnKF reste majoritairement utilisé en pratique et que les exemples d'applications du

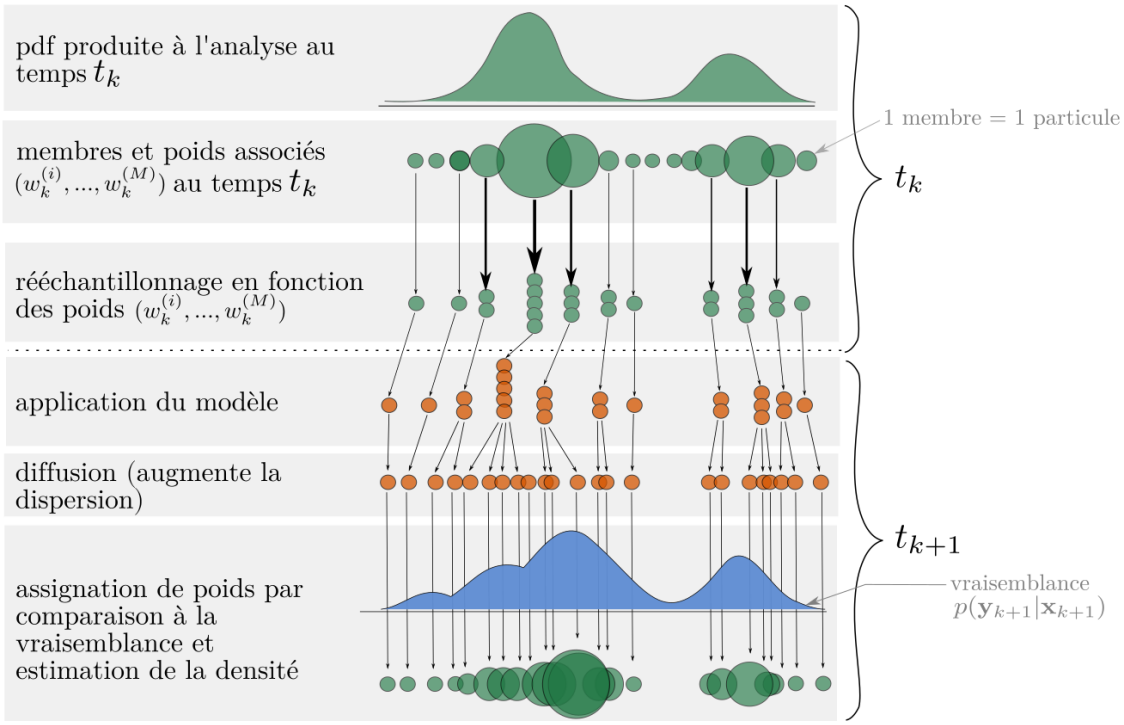


Figure F.1 – Principe du filtre particulaire illustré sur un cycle d'assimilation. Les ronds colorés représentent les membres de l'ensemble (vert pendant l'analyse et orange pendant la prévision) et leurs tailles illustrent les poids associés à chacun (adapté de Jurić p. d.).

filtre particulaire à l'hydrologie sont assez limités (e.g. Moradkhani et al., 2005 ; Pasetto et al., 2012). Il n'en reste pas moins que le filtre particulaire reste une méthode attractive, qu'il peut être intéressant d'explorer en cas d'échec du filtre de Kalman d'ensemble.

Annexe G

Evaluation du CRPS à partir d'un ensemble

Soient un ensemble de M valeurs x_1,...,x_M classées de la plus petite à la plus grande telles que :

$$x_i \leq x_j, \text{ pour } i < j, \tag{G.1}$$

et x^{t} la valeur déterministe de référence.
Le CRPS pour l'ensemble est calculé comme suit (Hersbach, 2000) :

$$CRPS = \sum_{i=0}^{M} \alpha_i p_i^2 + \beta_i (1 - p_i)^2, \tag{G.2}$$

où $p_i = \frac{1}{M}$ et où α_i et β_i sont déterminés comme suit :

$0 < i < M$	α_i	β_i
$x^{\mathrm{t}} > x_i$	$x_{i+1} - x_i$	0
$x_{i+1} > x^{\mathrm{t}} > x_i$	$x^{\mathrm{t}} - x_i$	$x_{i+1} - x^{\mathrm{t}}$
$x^{\mathrm{t}} < x_i$	0	$x_{i+1} - x_i$

. (G.3)

Les cas $i = 0$ et $i = M$ participent seulement au CRPS quand la valeur de référence x^{t} est une valeur extrême, c'est-à-dire qu'elle est inférieure à x_1 ou supérieure à x_M. Dans ce cas, le Tableau G.3 doit être modifié comme suit :

Outlier	α_i	β_i
$x^{\mathrm{t}} < x_1$	0	$x_1 - x^{\mathrm{t}}$
$x_M < x^{\mathrm{t}}$	$x^{\mathrm{t}} - x_M$	0

. (G.4)

Annexe H

Densités de probabilité des teneurs en eau à saturation biaisées

soilhorizon_thetas_2	N(0.34, 0.0336)
soilhorizon_thetas_3	N(0.30, 0.0320)
soilhorizon_thetas_4	N(0.34, 0.0284)
soilhorizon_thetas_6	N(0.37, 0.0354)
soilhorizon_thetas_7	N(0.34, 0.0325)
soilhorizon_thetas_8	N(0.37, 0.0416)
soilhorizon_thetas_9	N(0.3, 0.0332)
soilhorizon_thetas_10	N(0.34, 0.0316)
soilhorizon_thetas_11	N(0.3, 0.0338)
soilhorizon_thetas_14	N(0.3, 0.0338)
soilhorizon_thetas_12	N(0.3, 0.0338)
soilhorizon_thetas_15	N(0.3, 0.0338)
soilhorizon_thetas_13	N(0.3, 0.0338)
soilhorizon_thetas_16	N(0.3, 0.0338)

Tableau H.1 – Description des fonctions de densité de probabilité (pdf) des teneurs en eau à saturation utilisées pour l'assimilation de données. Toutes les pdf considérées sont normales (N) et décrites par leurs moyennes μ et écarts-types σ.