Modèles d'impact statistiques en agriculture : de la prévision saisonnière à la prévision à long terme, en passant par les estimations annuelles

Jordane Mathieu

THÈSE DE DOCTORAT

de l'Université de recherche Paris Sciences & Lettres
PSL Research University

Préparée à l'École Normale Supérieure

Modèles d'impact statistiques en agriculture :
des prévisions saisonnières aux prévisions à long terme,
en passant par les estimations annuelles.

École doctorale n°129

ÉCOLE DOCTORALE DES SCIENCES DE L'ENVIRONNEMENT D'ILE-DE-FRANCE

Spécialité MATHÉMATIQUES APPLIQUÉES

COMPOSITION DU JURY :

M. David Makowski
Agro Paritech, INRA, Paris,
Rapporteur

M. Jean-Christophe Calvet
CNRM, Toulouse
Rapporteur

M. Filipe Aires
Observatoire de Paris, CNRS
Directeur de thèse

M. Philippe CIAIS
LSCE, Paris
Examinateur

M. Eric Parent
Agro Paritech, Paris
Examinateur

M. Robert Vautard
LSCE, Paris
Prsident du jury

Soutenue par Jordane MATHIEU
le 29.03.2018

Dirigée par **Filipe AIRES**

l'Observatoire de Paris

ENS ÉCOLE NORMALE SUPÉRIEURE | PSL RESEARCH UNIVERSITY PARIS

Résumé

En agriculture, la météo est le principal facteur de variabilité d'une année sur l'autre. Cette thèse vise à construire des modèles statistiques à grande échelle qui estiment l'impact des conditions météorologiques sur les rendements agricoles. Le peu de données agricoles disponibles impose de construire des modèles simples avec peu de prédicteurs, et d'adapter les méthodes de sélection de modèles pour éviter le sur-apprentissage. Une grande attention a été portée sur la validation des modèles statistiques. Des réseaux de neurones et modèles à effets mixtes (montrant l'importance des spécificités locales) ont été comparés.

Les estimations du rendement de maïs aux États-Unis en fin d'année ont montré que les informations de températures et de précipitations expliquent en moyenne 28% de la variabilité du rendement. Dans plusieurs états davantage météo-sensibles, ce score passe à près de 70%. Ces résultats sont cohérents avec de récentes études sur le sujet.

Les prévisions du rendement au milieu de la saison de croissance du maïs sont possibles à partir de juillet : dès juillet, les informations météorologiques utilisées expliquent en moyenne 25% de la variabilité du rendement final aux États-Unis et près de 60% dans les états plus météo-sensibles comme la Virginie. Les régions du nord et du sud-est des États-Unis sont les moins bien prédites. Le rendements extrêmement faibles ont nécessité une méthode particulière de classification : avec seulement 4 prédicteurs météorologiques, 71% des rendements très faibles sont bien détectés en moyenne.
L'impact du changement climatique sur les rendements jusqu'en 2060 a aussi été étudié : le modèle construit nous informe sur la rapidité d'évolution des rendements dans les différents cantons des États-Unis et localisent ceux qui seront le plus impactés. Pour les états les plus touchés (au sud et sur la côte Est), et à pratique agricole constante, le modèle prévoit des rendements près de deux fois plus faibles que ceux habituels, en 2060 sous le scénario RCP 4.5 du GIEC. Les états du nord seraient peu touchés.

Les modèles statistiques construits peuvent aider à la gestion sur le cours terme (prévisions saisonnières) ou servent à quantifier la qualité des récoltes avant que ne soient faits les sondages post-récolte comme une aide à la surveillance (estimation en fin d'année). Les estimations pour les 50 prochaines années participent à anticiper les conséquences du changement climatique sur les rendements agricoles, pour définir des stratégies d'adaptation ou d'atténuation. La méthodologie utilisée dans cette thèse se généralise aisément à d'autres cultures et à d'autres régions du monde.

Publications

Mathieu, J.A. and Aires, F. (2018) *Using Neural Network classifier approach for statistically forecasting extreme corn yield losses in Eastern USA*, Earth and Space Sciences, in review.

Mathieu, J.A. and Aires, F. (2018) *Impact of agro-climatic indices to improve crop yield forecasting*, Agricultural and forest Meteorology, 15 (30).

Mathieu, J.A. and Aires, F. (2016) *Statistical weather impact models : an application of neural network and mixed-effects forn corn production over the United-States*, Journal of applied Meteorology and Climatology, 55 (11)

Mathieu, J.A., Hatté, C., Balesdent, J. and Parent, É. (2015) *Deep soil carbon dynamics are driven more by soil type than by climate : a worldwide meta-analysis of radiocarbon profiles*, Global Change Biology, 21 (11)

Table des matières

Remerciements

Quatre ans ont passés depuis cette sortie à cheval où, discutant avec ma famille, je parlais de projets post-master, et de la probabilité quasi-nulle de continuer sur une thèse. Je ne me doutais pas que quelques mois plus tard, une discussion avec mon directeur de master et fantastique enseignant, Christophe Giraud, allait faire basculer le déroulement des quatre années qui ont suivi, et probablement, de celles qui suivront. Je ne l'oublierai pas.

Cette thèse n'aurait pu voir le jour sans l'aide précieuse de mon directeur de thèse Filipe Aires, qui m'a guidée et encouragée, et aux côtés de qui j'ai beaucoup appris. Son encadrement, très justement équilibré, m'a laissé de grandes libertés d'organisation et de prises de décisions. M. Makowski et M. Calvet ont accepté d'être les rapporteurs de cette thèse, et je les en remercie, de même que pour leur participation au jury. Ils ont également contribué par leurs nombreuses remarques et suggestions à améliorer la qualité de ce manuscrit.

Je remercie également les deux membres de mon comité de thèse, Jan Polcher et Wouter Dorigo pour leur vision extérieure et leurs conseils. A l'étranger, je tiens sincèrement à remercier l'accueil très chaleureux des membres de la conférence « Applied Statistics in Agriculture » et particulièrement, Matthew Kramer avec qui j'ai eu le plaisir d'échanger, moi avec mon pauvre anglais, et lui avec ses très bons souvenirs de français.

Je tiens également à remercier mes collègues et co-bureau du LERMA avec qui j'ai eu grand plaisir de parler d'autre chose que de ma thèse. Merci à Catherine, Victor, Lise, Samuel, Binh, Carlos, sans oublier deux anciens thésards Dié et Fabien, avec qui j'ai partagé de très bons moments à la chorale. Je n'oublie pas les administrateurs pour leur impressionnante disponibilité et leur sympathie, en particulier Laurent Giraud avec les sorties natation.

Merci également à Alain pour m'avoir formée, conseillée et encouragée pour les observations avec la lunette Arago : ces soirées d'observations resteront toujours des moments privilégiés, hors du temps. En ce qui concerne la partie informatique de ce travail, je remercie Benoit Albert pour m'avoir fourni du matériel de qualité qui m'a permis de travailler hors de l'Observatoire, rendant ainsi possible mon parcours d'enseignement.

Enseignement et recherche ont été deux composantes indissociables de ma thèse : alors que le premier ne devait durer que deux ans, c'est lui qui va désormais perdurer. Je suis à ce sujet extrêmement reconnaissante envers Pascal Combemale qui, me faisant confiance, m'a ouvert les portes du CPES et de PSL pour enseigner durablement dans des conditions exceptionnelles que je n'aurais espérées. Le cours d'une vie tient parfois à peu de choses. C'est toute ma carrière qui en sera marquée.

Au terme de ce parcours, je remercie ma famille et mes amis, pour leurs attentions et encouragements qui m'ont accompagnée tout au long de ces années. Enfin et surtout, je suis profondément reconnaissante envers mon cher compagnon Aymeric, pour son immense soutien et ses précieux conseils tant sur le plan scientifique que personnel. Merci de m'avoir encouragée en répétant que même si le sujet était très appliqué et loin des formules et des raisonnements théoriques de la recherche mathématique, c'était un sujet compliqué. Cette thèse n'aurait pas vue le jour sans toi. C'est bien plus qu'un projet de recherche qui s'est construit ces quatre dernières années.

CHAPITRE I

Introduction

Table des matières

« 70% de l'économie mondiale est météo-sensible. »

Allianz Risk Transfer Group
[Alli13]

1 La météo-sensibilité et les modèles d'impact

Les façons dont la météorologie [1] actuelle et future affecte les activités socio-économiques sont largement décrites dans la littérature, que ce soit dans le domaine de l'environnement [Sult09], de l'agriculture [Thom88, Kotl07a, Kand02], de l'énergie [Adam98, Kayl91], de la logistique dans les ventes [Star00, Lazo11], des assurances contre les catastrophes, ou encore du tourisme [Jews05, Mart04]. L'analyse de ces impacts météorologiques sur les activités humaines nécessite de construire des modèles - appelés modèles d'impact - qui décrivent les liens et les effets entre la météo et les données socio-économiques [Aire12].

Ces modèles sont basés sur la météo et conçus pour être aussi corrélés que possible à l'activité humaine concernée. Par exemple, [Yu09] définissent un indice pour évaluer l'impact de la météo sur l'activité touristique.

1. Par impact "météo" nous entendons l'impact de toutes variables atmosphériques, qu'elles soient brutes (comme les températures moyennes mensuelles ou journalières) ou transformées (comme les indices agro-climatiques).

On peut utiliser les modèles d'impact à deux niveaux. Comme les activités humaines peuvent être directement affectées par les catastrophes ou encore par la variabilité naturelle de la météo, les modèles d'impact peuvent être utilisés pour améliorer la gestion de l'activité sur le court ou moyen terme. Un exemple d'adaptation est celui de la mise en vente de produits plus adaptés, en logistique. La production et la consommation d'énergie peuvent aussi être adaptées. Des contrats d'assurance peuvent aussi utiliser les modèles d'impact : pour des raisons pratiques, il peut être plus facile et moins couteux de mesurer les variables météorologiques que, par exemple, la production agricole dans une région. Cette production agricole peut alors être assurée contre les conditions météorologiques défavorables.

En seconde utilisation, les modèles d'impact peuvent servir à anticiper des évolutions à long terme qui résulteraient du changement climatique [Leck02, GIEC14, Raub05]. Pour ce second type d'utilisation, les modèles d'impact à long terme peuvent être associés aux modèles de prévision climatiques (à partir des modèles de climat globaux - GCM) pour estimer les conséquences du changement climatique sur les activités humaines [Schi96]. Ce type d'étude permet d'optimiser les investissements à long terme comme le choix des pratiques agricoles [Lewa99], la construction de champs d'éoliennes ou de barrages hydrauliques, ou encore la mise en place de politiques de santé publiques [Gree01]. Ils renforcent et épaulent les stratégies d'adaptation au changement climatique.

Dans cette étude, nous nous focaliserons sur **la création de modèles d'impact météorologiques liés à l'agriculture**, et sur la compréhension de la sensibilité des rendements agricoles à diverses variables météorologiques. Les rendements extrêmes seront aussi étudiés. Les utilisations possibles des modèles d'impact en agriculture sont développées aux sections suivantes.

2 Application à l'agriculture

2.1 Comment s'explique la variabilité du rendement agricole ?

Avant d'être semées, les graines disposent d'un certain potentiel de rendement. Des champs expérimentaux bénéficiant de conditions optimales permettent d'obtenir une approximation de ce potentiel. Dans la réalité, le rendement sera conditionné par des paramètres relativement stables, tels que les propriétés physiques et chimiques des sols, la gestion agricole ou la variété utilisée. Les pratiques culturales vont également conditionner le rendement et dépendront de l'habileté et des compétences des cultivateurs. L'utilisation d'engrais et la protection contre les maladies sont liées à ces compétences, mais également aux conditions économiques.
La variabilité des conditions météorologiques dans les limites des conditions climatiques habituelles permet souvent d'expliquer la plus grande partie de la variabilité annuelle sur une brève période, alors que les pratiques culturales et l'introduction de nouvelles variétés rendent compte, quant à elles, de l'essentiel de la variabilité (sous forme de tendance) sur des périodes allant de dix à vingt ans. Pour des périodes plus longues, les changements climatiques et l'amélioration ou la dégradation des sols pourraient bien constituer d'autres facteurs importants influençant le rendement [Boum94]. Ces divers éléments ne sont pas indépendants les uns des autres ; certaines pratiques agricoles atténuent ou, au contraire, renforcent la variabilité due aux conditions météorologiques. La compréhension de la variabilité des rendements est aussi nécessaire à la compréhension des prévisions.

Il est donc important de distinguer les types de facteurs influençant le rendement ainsi que leur temps caractéristique d'impact : certains facteurs ont un effet à court terme tandis que d'autres agissent sur le long terme. Il faut aussi remarquer que les pratiques agricoles dépendent en partie des conditions météorologiques, et donc que les effets de la météo et ceux des pratiques agricoles ne peuvent pas être toujours facilement distingués. Dans cette étude, l'évolution à long terme sera considérée sous la forme d'une tendance, permettant d'introduire la variabilité inter-annuelle des rendements : **la météo reste le plus grand facteur incontrôlable influençant le développement des cultures** [Tayl97]. Celle-ci peut varier considérablement d'un champ à l'autre et d'une année à l'autre.

2.2 Obtention des rendements par observations directes in situ

La prévision des rendements peut être vue comme "*l'art d'identifier les facteurs qui déterminent la variabilité spatiale et inter-annuelle des rendements agricoles*" [Gomm03].

Pour les stratégies nationales de sécurité alimentaire, il est déterminant d'avoir des prévisions de rendements agricole précises, fiables, ponctuelles et objectives : elles touchent les politiques d'importation et d'exportation de plants, les prix, mais aussi les décisions des cultivateurs pour l'utilisation d'engrais chimiques ou de l'irrigation. Les prévisions des rendements à l'échelle de l'exploitation sont aussi importantes pour une gestion à petite échelle et aident les cultivateurs dans des choix pouvant avoir une implication économique ou environnementale significative [Lian04].

On distingue la prévision de l'estimation des rendements agricoles. La première donne souvent lieu à des applications à court terme, tandis que la seconde donne lieu à des applications à moyen ou long terme (Section 3.2).
La prévision des rendements repose encore dans la plupart des pays, sur des rapports effectués à intervalles réguliers, par culture, et pendant la période de croissance : ils renseignent sur les conditions de croissance de la culture, par le bias d'indicateurs portant sur l'apparence observée in situ. Les conditions de culture sont par exemple décrites par "excellentes", "bonnes", "mauvaises" ou par des numéros [Spin56a].

Cependant, estimer les rendements par des observations in situ ne permet que très rarement d'avoir des informations à large échelle car les mesures sont faites localement. Elles permettent néanmoins des applications à l'échelle de l'exploitation, voire du département. De plus, ces mesures sont souvent coûteuses. Face à ces difficultés, nous proposons d'estimer différemment les rendements agricoles en utilisant seulement des données météorologiques (donc des données indirectes) : ces données sont accessibles à grande échelle (région, état ou continent), et sont moins coûteuses. Elles permettent ainsi des applications à l'échelle des régions, des états ou des continents.

3 Pourquoi prédire les rendements agricoles avec l'information météorologique ?

3.1 Des prévisions saisonnières (en cours d'année) [2]

La modélisation, de façon générale, recouvre deux aspects : d'une part, simuler un processus pour mieux le comprendre et d'autre part, prédire une variable de sortie. Les premiers modèles mis en place l'ont été à des fins d'investigation, comme bancs d'essai pour construire une théorie sur les processus étudiés. Une simulation fiable des rendements peut être utilisée de manière stratégique, pour analyser les conséquences possibles de stratégies de gestion (par exemple, la planification de l'utilisation des terres) et comme outil de gestion tactique (la date de semis, ou de récolte). Si les prédictions se font en temps réel, elles sont directement utilisables par les agriculteurs pour planifier la production. Tous ceux qui ont intérêt à connaître la production (commerciaux, planificateurs gouvernementaux, assureurs) s'en servent également [Boum94].

Application : aide à la gestion sur le court terme

A court terme (c'est-à-dire à une échelle intra-annuelle), les choix de culture et de technologie d'irrigation ont été faits et seuls les facteurs variables (eau, engrais, pesticides, etc.) peuvent être ajustés en fonction des réalisations du risque climatique ou des anticipations de l'agriculteur [Reyn09]. Dans ce contexte, les modèles d'impacts peuvent aider les cultivateurs à mieux gérer l'apport disponible en eau, et à anticiper d'éventuelles pénuries. Face à la sécheresse durable, si la quantité d'eau disponible ne peut être augmentée, il peut être par exemple préférable de sacrifier une partie des terres cultivées pour concentrer l'eau disponible sur une plus petite surface. Des modèles d'impacts fiables peuvent aider à prendre de telles décisions.

Les possibilités d'adaptation de l'agriculture au risque de sécheresse relèvent aussi de décisions collectives : restriction d'eau dans certaines régions par décret, et transferts d'eau entre régions ou entre usages [Amig06]. Les prévisions de rendements agricoles doivent donc être également faites à l'échelle du pays pour apporter des informations fiables et ponctuelles aux grandes organisations agricoles nationales (comme la Politique Agricole Commune (PAC) en Europe) et permettre des prises de décision rapides pendant la période de croissance.

Comme on le verra au chapitre suivant, les modèles d'impact dits "mécanistes" ou "de culture" sont a priori, davantage adaptés pour aider les cultivateurs dans la gestion quotidienne de leurs cultures, que les modèles de régression statistique : en effet, ils prennent en compte de façon plus directe, les relations plantes/nutriments et donc les effets de l'engrais, et des maladies (donc des pesticides). Cependant, les modèles de culture s'utilisent à petite échelle (de la parcelle [3] au département). Il est donc nécessaire de quantifier la fiabilité des modèles d'impact quant aux applications à court terme (durant la période de croissance).

2. Dans le domaine de la météorologie, l'expression "prévisions saisonnières" est souvent lié à l'utilisation de prévisions numériques du temps réalisées sur plusieurs mois. Ici nous utilisons une autre signification : dans ce manuscrit, on appelle "prévisions saisonnières du rendement" des prévisions du rendement réalisées pendant la saison de croissance de la plante, que cela soit au début, au milieu ou à la fin.

3. Une parcelle est une subdivision d'un champ d'un seul tenant et d'une même culture.

3.2 Des estimations en fin d'année

Quand les cultures ont atteint leur maturité et que la récolte approche, une estimation du rendement est alors faite. Ce procédé, qui a lieu en fin d'année, n'est pas une prévision au vrai sens du terme, mais plutôt une estimation. On utilisera dans la suite le terme de "monitoring".

Les estimations de production agricole sont utiles pour le commerce, les politiques de développement ou encore l'assistance humanitaire reliée à la sécurité alimentaire. Par exemple, les services agricoles du Centre Commun de Recherche européen (JRC - Joint Research Center) fournit à la PAC des données et des estimations de rendements agricoles. Ce service propose aussi au programme européen de sécurité alimentaire des évaluations et des alertes anticipées en cas de baisse potentielle de production dans les régions mondiales d'insécurité alimentaire. Toutes les informations que le JRC rassemble, aident à préparer des rapports sur l'équilibre alimentaire, qui sont utilisés pour des analyses de marchés, pour des décisions relatives à la gestion des stocks, des importations, des exportations, et des budgets en préparation.

Application aux assurances

Un moyen pour un agriculteur de se couvrir contre le risque météorologique est d'avoir recours à des systèmes d'assurance publique ou privée [Cobl00]. Les produits dérivés conçus pour se prémunir face aux variations de taux d'intérêt, de cours de change, d'indices boursiers ou encore de prix des matières premières ont été étendus au climat et il est donc désormais possible pour les cultivateurs, de couvrir leur exposition au risque météorologique par des options, des contrats à terme ou encore des swaps climatiques (Climpact-Metnex, ClimateSecure). Ces types de contrats utilisent des modèles d'impact qui estiment l'évolution des cultures et de leur rendement selon les aléas météorologiques observés.
Les événements météorologiques peuvent avoir des conséquences très étendues au sein d'une économie régionale. En effet, les pertes agricoles affectent non seulement les revenus des exploitants mais aussi les salaires de leurs employés ainsi que les réserves de nourriture. La sécheresse, qui affecte un grand nombre de personnes au même moment, met en évidence les insuffisances des méthodes traditionnelles de gestion des risques [Koch10]. Des défauts de paiement généralisés des prêts sont alors observés, ce qui dissuade progressivement les établissements financiers d'établir de nouveaux prêts. On comprend donc la nécessité de la mise en place de techniques de couvertures efficaces.

Les exploitations agricoles encourent deux types de risques principaux : les risques liés aux événements météorologiques menaçant les récoltes (risque de quantité) et les risques économiques qui menacent la valeur qu'ils peuvent tirer de leur activité (risque de prix).

L'assurance indicielle est un produit financier lié à un indice présentant une forte corrélation avec les rendements locaux. Les contrats sont rédigés de façon à protéger le contractant contre des risques ou des événements spécifiques (par exemple perte de rendement, sécheresse, ouragan, inondation) définis et consignés à l'échelle régionale

(par exemple dans une station météorologique locale). Les indemnisations sont déclenchées lorsque l'indice atteint une certaine valeur ou possède une certaine tendance et ne sont pas basées sur les rendements effectifs.

Les dérivés climatiques sont des produits dérivés financiers ayant comme sous-jacent un indice climatique (température, précipitations ou combinaison de différentes variables climatiques).

La construction d'indices climatiques devient donc nécessaire pour les assurances mais celle-ci est totalement dépendante des modèles d'impacts existants dont la fiabilité n'est pas toujours suffisante.

3.3 Des estimations à long terme : impact du changement climatique

Les modèles d'impact météorologique sur l'agriculture peuvent être utilisés pour estimer l'effet du changement climatique sur les rendements, ou simplement estimer l'évolution future d'une activité économique sous un climat futur. Elle repose essentiellement sur l'application de modèles de prévision agricole pré-existants ou non, à des variables d'entrée modifiées (comme l'augmentation de la température), ou plus récemment, à des données issues de projections climatiques futures. Cette tâche est très complexe : aux difficultés rencontrées pour construire des modèles de prévisions agricoles, s'ajoutent les problèmes d'extrapolation des pratiques agricoles, et les nombreuses incertitudes liées aux différents scénarios climatiques possibles, accompagnés de simulations changeant significativement d'un modèle à l'autre. La validation et la comparaison des études en sont d'autant plus difficiles.

La météorologie et l'agriculture sont deux systèmes très corrélés, tous deux ayant lieu à l'échelle mondiale. Le changement climatique affecte la moyenne et la variabilité des conditions météorologiques et la fréquence des événements extrêmes, qui déterminent une grande partie de la variabilité de la production et des rendements.

Les cultivateurs peuvent adopter différentes stratégies d'adaptation : changer de culture pour améliorer la résilience d'une certaine culture en changeant la variété, ajuster les dates de semis, changer l'utilisation d'engrais, ou encore agir sur l'irrigation. Certaines de ces mesures coûtent peu par rapport à d'autres, qui requièrent un investissement non négligeable pour une petite exploitation (irrigation...). Ainsi, si les cultivateurs adaptent leur comportement aux nouvelles conditions climatiques, cela pourrait largement diminuer l'impact sur leurs exploitations et le secteur agricole. Les études d'impact qui incorporent cette adaptation prédisent un impact du changement climatique bien moindre en présence d'adaptation, que ce soit dans les pays développés ou en voie de développement [Mend10, Mend99, Mend03, Mend09]. Cependant, l'évolution semble pouvoir varier d'un pays à l'autre, comme l'a fait remarquer le GIEC (Groupe Intergouvernemental d'Experts sur l'Évolution du Climat) en 2007 [GIEC07].

Sans même parler des risques potentiels liés au changement climatique, l'anticipation des états et des régions se doit d'être permanente. En effet, de nombreuses infrastructures - comme les barrages hydroélectriques, les canaux d'irrigation, ou encore les voies d'acheminement rapide de matières premières - demandent beaucoup d'années à être construites et doivent donc être décidées longtemps à l'avance. Pour ce faire, les agences régionales ont besoin de visibilité future sous forme d'estimations du climat mais aussi des réponses potentielles des activités socio-économiques. Elles ont besoin de savoir quel secteur industriel va connaitre des difficultés ; si des politiques de santés devront être renforcées ; si les déplacements de matières premières vont s'accroitre ;

quels secteurs va-t'il falloir attirer ou aider ? Toutes ces décisions reposent sur une demande d'information quant à l'évolution future et conjointe du climat et des activités socio-économiques.

Application à l'adaptation au changement climatique

L'agriculture est un secteur particulièrement exposé aux changements du climat, dont les effets sont susceptibles de s'accentuer à l'avenir. Étant donné sa contribution à l'économie des pays, à l'aménagement des territoires, et à la gestion de l'environnement, son adaptation est un défi majeur, tant pour éviter des difficultés financières que pour profiter de nouvelles opportunités. Malgré leurs incertitudes, les modèles d'impact fournissent des informations sur les évolutions probables futures. Ils peuvent donc servir de guide pour aider les agriculteurs, les régions ou les états à s'adapter et à anticiper le changement climatique. Le paragraphe suivant présente plusieurs adaptations possibles.

Devant les risques accrus de sécheresse, les états ou les régions peuvent s'engager à développer des technologies d'irrigation plus performantes (irrigation par goutte à goutte par exemple [Zilb03, Reyn09]). Un autre moyen à long terme d'atténuer les impacts de la sécheresse consiste à modifier les systèmes de culture, c'est-à-dire les assolements et les rotations culturales, au profit de variétés plus résistantes au stress hydrique. Selon [Amig06], la substitution du maïs irrigué par du sorgho irrigué permet une économie d'au moins 50% des volumes d'eau. Ces résultats obtenus en France en conditions expérimentales posent des problèmes de faisabilité à grande échelle (adaptation des filières, baisse des marges) qui ont peu été étudiés pour le moment, en tout cas dans le contexte français.

Dans la Creuse, le nombre de jours où la température moyenne journalière est supérieure à 25^oC est de l'ordre de 20 aujourd'hui, et serait de 27 à 28 environ d'ici 2040 en moyenne. Le blé, dont le seuil d'échaudage est de 25^oC, serait mis en difficulté pendant plus de jours qu'aujourd'hui en été, mais aussi en fin de printemps. Une piste d'adaptation est d'avancer la date de semis de façon à éviter les périodes les plus chaudes, et/ou choisir une semence dont le cycle végétatif est plus court, donc fauchée plus tôt, ou une semence mieux adaptée aux périodes de chaleurs estivales (entrant en dormance pendant les périodes mettant le plus en difficulté la plante) [Denh14]. De même, si l'on observe une augmentation des températures moyennes et des fréquences de sècheresse, une piste d'adaptation serait de profiter de cette opportunité en semant un peu plus tôt et d'utiliser le temps libéré sur la parcelle pour semer des cultures fourragères en dérobé.

Plus généralement, les changements de cultures constituent un des facteurs d'adaptation au changement climatique mis en exergue par le GIEC [GIEC07].

Différentes voies d'adaptation s'ouvrent dont on ne peut dire à l'avance lesquelles remporteront l'adhésion des acteurs. Car ces adaptations ne se feront qu'avec une forte mobilisation, non seulement de la part des agriculteurs mais aussi de tous les acteurs des filières (recherche, fournisseurs, coopératives de collecte, industriels, consommateurs, etc.). En matière de stratégies d'adaptation, tous ces acteurs seront confrontés à des choix d'autant plus complexes qu'aux facteurs techniques s'ajoutent des facteurs économiques, sociaux et organisationnels, qui joueront comme catalyseurs ou au contraire comme freins à la mise en oeuvre d'actions d'adaptation. C'est pourquoi les

simulations agroclimatiques, qui permettent d'appréhender les effets du changement climatique sur la croissance et la survie des cultures, ne seront pas suffisantes pour anticiper les défis à venir, mais demeurent essentielles [Vert13].

3.4 Impact en agriculture selon les échelles spatiales et temporelles

La Figure I.1, résume les différentes applications possibles des modèles d'impact météorologique en agriculture, selon l'échelle spatiale et selon que l'on souhaite des prévisions ou des estimations, donc des utilisations à court ou long terme. Les modèles d'impacts météorologiques en agriculture sont décrits en détail au chapitre II. On se concentre ici sur les applications possibles de ces modèles.

Les traders et les gouvernements utilisent souvent des estimations/prédictions de rendements pour prendre des décisions. En revanche, les agriculteurs utilisent plutôt des modèles prédisant des bilans hydriques (ou d'autres indicateurs climatiques), des modèles prédisant des stades de développement, des bilans azotés ou des indicateurs du niveau de risque de maladies.

Comme on peut le voir, les applications potentielles sont nombreuses et le choix des échelles spatiales et temporelles des différents modèles vont dépendre de l'application visée. Par exemple, il n'est aujourd'hui - à ma connaissance - pas encore possible de fournir des conseils de gestion aux cultivateurs pour une grande échelle spatiale (état ou continent). En revanche, l'aide à la stabilisation des prix du marché se fait en utilisant des modèles statistiques régionaux ou étatiques, et non (ou très rarement) des modèles mécanistes.

Échelles spatiales/utilisateurs	Applications à court terme	Applications à moyen terme	Applications à long terme
Mondial - agences internationales - multinationales	- cours des marchés agricoles - stabilité des prix - réduction de la spéculation - commerce international - produits financiers dérivés, assurance à large échelle	- traités internationaux - commerce international - cours des marchés	- traités internationaux (ex : ….)
États - gouvernements - secteur agroalimentaire	- gestion des importation/exportations - sécurité alimentaire - politique commerciale - développer des systèmes d'alertes	- politique agricole (soutient auprès des producteurs) - sécurité alimentaire du pays - politique commerciale	- construction d'infrastructures importantes : barrages, canaux pour l'irrigation - sécurité alimentaire - recherche nationale de nouvelles variétés
Régions - agences régionales - agro-industriels - usine transformation	- approvisionnement en matière première agricole - assurance récolte, analyse des risques - éviter les ruptures de stock	- assurance récolte	- planification économique et aménagement du territoire - choix des entreprises à attirer et à aider - services climatiques
Département - coopératives agricoles - fournisseurs - producteurs - semenciers	- mise en vente de produits plus adaptés en logistique - prix de vente, optimisation économique - gestion des stocks - demande en eau des cultivateurs/période de restriction	- gestion des stocks - gestion environnementale	- Semenciers : choix des cultures qui se développeront le mieux à un endroit
Exploitation - producteurs	Gestion des pratiques agricoles : - date de semis optimale - récolte au moment optimal - anticipation : stock en eau, multiplication d'insectes, infection fongique, taux de nutriments disponibles pour les cultures	- épargne ou non - anticipation : stock en eau, multiplication d'insectes, infection fongique, taux de nutriments disponibles pour les cultures	- Choix des cultures/mécanisation dans lesquelles investir - reconversions possibles
Champs - cultivateurs	- calendrier des pratiques agricoles (irrigation, pesticides, engrais) - mesures de protection contre le froid	- sélection des variétés - planification de l'utilisation des terrains	- sélection des variétés
	Prévision saisonnière	Estimation (fin d'année)	Prévision à long terme
	Pour optimiser	Pour mieux comprendre les dépendances à la météo et aider à la stabilité économique	Pour anticiper les 50 ou 100 prochaines années

Modèles de régression statistiques

Modèles mécanistes

FIGURE I.1 – Les utilisations potentielles des modèles d'impact météorologique en agriculture.

4 Objectifs et plan de la thèse

Cette thèse vise à quantifier l'impact de la météo (ou l'étude de la météo-sensibilité) sur les rendements agricoles. Le but étant de faire des prévisions et/ou des estimations à large échelle, l'utilisation des modèles mécanistes traditionnels n'est pas conseillée comme on le verra au Chapitre II. On se concentrera sur l'utilisation de modèles plus simples : des modèles de régression statistique. On souhaite couvrir une étendue temporelle aussi grande que possible (35 ans ici), ce qui n'est pas souvent le cas dans la littérature, et qui est pourtant primordial pour la qualité des modèles. Les États-Unis proposent une large base agronomique facilement accessible. De plus, le maïs est l'une des principales céréales dans ce pays donc de grande importance pour son économie. C'est donc sur cette céréale que vont porter les analyses de météo-sensibilité.
On souhaite aussi quantifier et comparer les facteurs climatiques qui influencent les rendements : pour ce faire une première analyse portera sur des variables météo simples, puis une seconde analyse prendra aussi en compte des indices agroclimatiques (Chapitre III).
On souhaite également quantifier le pourcentage de variabilité des rendements que la météo peut expliquer en utilisant des outils statistiques robustes, et comparer ces résultats aux études de la littérature.
Une étude similaire sera aussi traitée pour les rendements extrêmement faibles avec des outils statistiques adaptés. Enfin, on évaluera la réponse de nos modèles d'impact aux différents scénarios climatiques.

Les points novateurs de l'étude sont donc les suivants : une étude à grande échelle, traitée à l'aide de modèles statistiques multivariés, comprenant une comparaison d'un nombre important de variables météo et utilisant une base de données agricoles plus grande que la moyenne (35 ans). De plus, une grande importance est donnée à l'estimation et la quantification de la qualité des modèles statistiques. Le cas des rendements extrêmement faibles est aussi étudié.

Le prochain chapitre de cette thèse va se focaliser sur les différents modèles d'impact qui existent en agriculture, comparant leurs avantages, leur inconvénients, et leur type d'application.
Le troisième chapitre va quant à lui s'intéresser plus spécifiquement aux variables nécessaires pour réaliser de tels modèles. Seulement une partie de ces variables sera utilisée par la suite.
Le quatrième chapitre s'intéresse à la méthodologie, tant du point de vue mathématique (pour les modèles statistiques) que du point de vue de la modélisation. Il propose en particulier une description détaillée des modèles à effets mixtes linéaires, qui peut être négligée en première lecture.
Les premiers modèles d'impacts statistiques construits dans cette thèse sont développés au chapitre cinq. Dans ce chapitre, on cherche à comparer la capacité prédictive de différents modèles d'impact, et à savoir quelles régions sont davantage météo-sensibles. On analyse aussi les améliorations prédictives apportées par l'utilisation d'indices agroclimatiques.
Un autre aspect important, ignoré dans les précédents chapitres, concerne l'étude des rendements extrêmement faibles. C'est l'objet du sixième chapitre qui propose d'aborder cette question en utilisant un réseau de neurone.
Enfin, le dernier chapitre propose l'application d'un de ces modèles d'impact aux projections climatiques futures. L'idée est alors de mettre en évidence une tendance d'évolution des rendements jusqu'en 2060.

CHAPITRE II

Les modèles d'impact en agriculture
Comment prédire les rendements agricoles ?

Table des matières

« Tous les modèles sont des approximations. Essentiellement, tous les modèles sont faux, mais certains sont utiles. Cependant, la nature approximative du modèle doit toujours être prise en compte. »

George BOX, (1919-2013)
Statisticien [Box87]

Un modèle est une représentation simplifiée d'un phénomène. C'est donc une description limitée et orientée de la réalité. Il est impossible de décrire de manière absolue la réalité à des fins descriptives ou prédictives, sans même parler de l'intérêt d'une telle démarche. La citation ci-dessus de G. Box illustre bien qu'un modèle n'est pas conçu pour être vrai et pour représenter parfaitement la réalité mais pour répondre à un usage. Il existe une infinité de types de modèles (depuis le dessin, en passant par les modèles mathématiques, les modèles physiques, ou les modèles économiques) chacun d'eux pouvant être divisés en sous-modèles (par exemple dans les modèles mathématiques on trouve entre autre, les modèles géométriques, les modèles statistiques, et les modèles dynamiques).

Il semble qu'il n'y ait pas de classification standard des méthodes de prévisions agricoles [Makr98, Arms01]. Cependant, dans la littérature scientifique, deux grandes approches ont été testées pour étudier les relations entre la météo et le rendement des cultures [1] : (1) les modèles mécanistes, et (2) les modèles statistiques. Les modèles mécanistes sont utiles pour simuler avec beaucoup de détails les effets des caractéristiques environnementales sur la croissance des cultures [Hoog00, Jone03a]. Ils sont souvent spécifiques à un site particulier, et demandent beaucoup d'informations locales telles que les propriétés du sol. À l'inverse de cette approche *orientée-processus*, il est possible d'adopter une approche *orientée-données* [Kotl07a]. Ces seconds modèles sont calibrés sur des données historiques avec le moins d'information a priori possible. Cette deuxième méthode se focalise sur les interactions entre données, et requière moins d'information directe de la plante (comme la taille des épis, ou les indices de végétation). Ces modèles statistiques ne sont pas seulement utiles pour les estimations de rendement agricole : l'analyse des données historiques peut aussi améliorer la compréhension des impacts en jeu. **Pour les modèles statistiques, seule une part de la variabilité du rendement peut être expliquée par les données.**

Récemment, [Ferm11] ont publié un rapport détaillé décrivant et comparant les diverses méthodes de prévision et d'estimation de rendement. Les méthodes de prévisions peuvent être réparties dans plusieurs catégories selon la part des connaissances agronomiques nécessaires, des statistiques utilisées, des modèles et des données requises, et du degré de sophistication. Quand on regarde les analyses portant sur ce sujet [2], on remarque que bien des modèles débordent de la classification duale simpliste décrite ci-dessus. Ainsi, des solutions de compromis associent les deux approches (mécaniste et statistique). L'objectif consiste à utiliser plus efficacement toutes les sources de données disponibles, afin d'obtenir de meilleures prévisions. Les modèles peuvent varier de simples à complexes, qu'ils soient statistiques ou mécanistes. Les modèles simples sont souvent utilisés pour estimer le rendement sur de grande zone géographique et sont basés sur des informations statistiques reliées à la météo et aux rendements historiques, et incluent peu ou pas de détails quant aux processus sol-plante.

La section suivante rassemble les différents prédicteurs utilisés dans les modèles pour prédire les rendements agricoles. Les modèles mécanistes sont décrits à la Section 2 et ceux statistiques à la Section 3. Les différents cadres d'application de ces modèles sont analysés à la Section 4.

1. L'histoire de la modélisation et de l'estimation des rendements est un sujet traité en détails chez [Lian04] ou [Crai13].
2. [Geor10, Bass13, Word12, Euro94, Spin56b]

1 Les variables utilisées pour les prévisions agricoles

Cette section rassemble une liste non exhaustive de variables explicatives qui sont utilisées pour estimer ou prédire des rendements agricoles. Il faut remarquer que l'utilisation d'une variable se cantonne très rarement à une approche précise (mécaniste ou statistique). Un modèle statistique peut utiliser une variable informant sur le type de sol si elle est fiable et disponible ; et un modèle mécaniste pourra utiliser la température moyenne si elle intervient dans les relations physiques décrivant les processus en jeu. C'est pourquoi on liste ici un ensemble de variables prédictives qui peuvent être utilisées dans un cadre comme dans l'autre.

1.1 Variables météorologiques ou indices agroclimatiques

Les méthodes de prévision agricole utilisant essentiellement des variables météorologiques [3] sont principalement basées sur deux variables, température et précipitation, car ces deux variables sont très reliées aux besoins des cultures [Barn82] et peuvent être obtenues facilement à partir de stations météorologiques ou de mesures satellites [Whit11]. Ces deux variables peuvent être utilisées seules ou en combinaison, quotidiennement, mensuellement, ou saisonnièrement. Dans les régions agricoles non irriguées, les précipitations forment le facteur le plus important affectant la croissance et le rendement des cultures [Word12].

Pour fournir au modèle de prévision des informations davantage reliées à la croissance des cultures, des indices agroclimatiques ont été définis. Traditionnellement, les indices agroclimatiques sont calculés à partir de données météo brutes pour mieux représenter le lien entre météo et croissance de la plante, et pour faciliter la prise de décision en agriculture [Lepa12, Caub15]. Les indices agroclimatiques, couplés ou non avec des modèles de climat ou des méthodes de régionalisation, sont de plus en plus utilisés comme traceur du changement climatique.
Seules quelques études ont comparé plus de 10 indices agroclimatiques (tels que le FFP (Free Frost Period), le CHU (Cumulated Corn Heat Units), le SPEI (Standardized Precipitation-Evapotranspiration Index), ou l'humidité du sol) pour les prévisions de rendement agricole. L'humidité des sols (variable pouvant aider à prendre en compte l'effet de l'irrigation) est très peu utilisée, car difficile à estimer de façon fiable à grande échelle (entre autre).

Toutes les variables listées ci-dessus apportent une information sur la plante mais de façon indirecte : on les désignera donc dans cette thèse par "prédicteurs indirects", par opposition à des mesures in situ, sur la plante.

Dans cette thèse, une comparaison d'une quarantaine d'indices agroclimatiques sera réalisée en parallèle de l'utilisation de variables météorologiques standards (température et précipitation).

1.2 Variables issues de la télédétection

La télédétection est définie comme étant la science d'obtenir des informations sur un objet à travers l'analyse de données obtenues par un appareil qui n'est pas en contact avec l'objet [Lill15]. Un important développement depuis plusieurs décennies est la constante amélioration des données météo issues de l'observation spatiale.

3. Dans la littérature, elles sont référencées sous le nom de "modèles agro-météorologiques".

Les données obtenues par télédétection (de l'énergie électromagnétique) peuvent être enregistrées à partir de satellites ou d'avions. La chlorophylle n'absorbe pas toutes les longueurs d'onde de la lumière solaire ; elle absorbe les longueurs d'onde bleues et rouges tandis que la lumière verte est réfléchie [Camp96]. La réflexion du rayonnement visible est en grande partie fonction des pigments des feuilles, tandis que le proche-infrarouge NIR (Near-Infrared) est réfléchi par la structure interne des feuilles. Le rayonnement NIR traverse la première couche des feuilles, l'épiderme [Gaus69]. Quand il atteint le mésophylle et les cavités internes de la feuille, il est dispersé vers le haut (rayonnement réfléchi) et vers le bas (rayonnement transmis), comme le montre la Figure II.1.

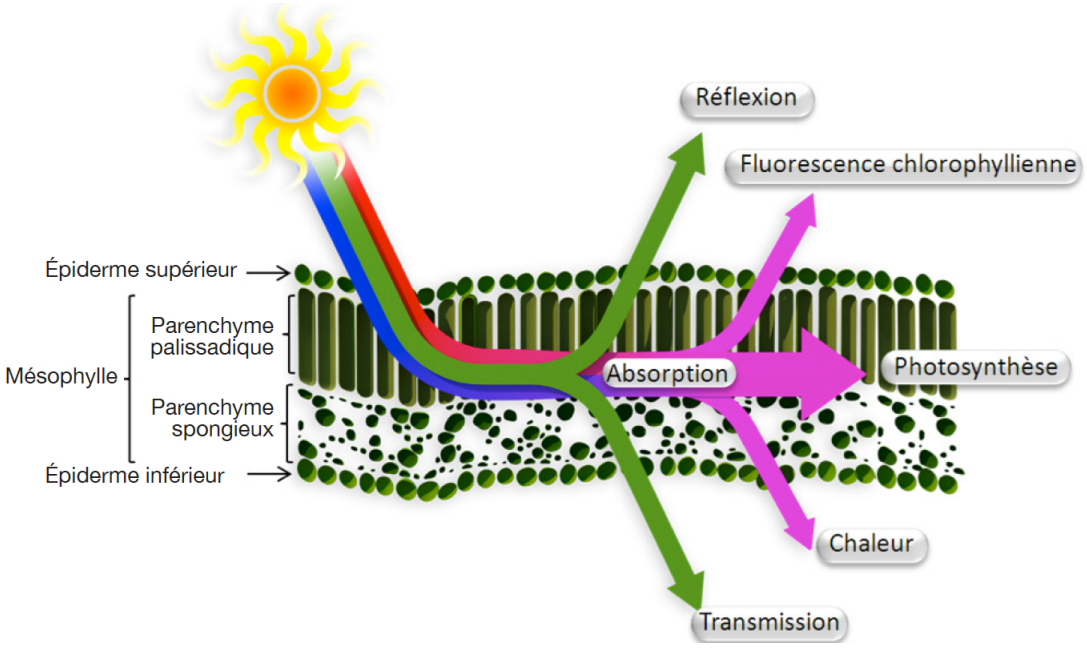


FIGURE II.1 – Représentation schématique du comportement du rayonnement solaire incident impactant la surface d'une feuille (tiré de [Abda16]).

Le comportement des rayonnements NIR dépend aussi de l'indice de surface foliaire LAI (Leaf Area Index), de la turgescence cellulaire, de l'épaisseur des feuilles, et de la quantité d'eau et d'air que contiennent les feuilles. La diminution relative de la réflectivité est plus importante dans le spectre visible que dans le NIR, à cause des effets de transmission du NIR à travers les feuilles et de l'absorption du rouge et du bleu par les pigments de chlorophylle. Grâce au mécanisme décrit ci-dessus, les cultures en bonne santé vont montrer de fortes valeurs de réflectivité dans le NIR et de faibles valeurs dans le spectre visible. Pendant la sénescence et pour les cultures sujettes au stress (par exemple, aux maladies, aux parasites, ou aux pénuries d'azote et d'eau) la chlorophylle rend possible l'expression d'autres pigments de la feuille tels que le carotène et le xanthophylle, ce qui cause un élargissement du pic de réflectivité dans les verts à 550nm, et une augmentation de la réflectivité de la lumière visible [Pint03]. En même temps, on observe une diminution de la réflectivité relative dans le NIR, comme conséquence d'une absorption moindre de la lumière visible dans les feuilles [Asne98]. La réflectivité du sol augmente de façon monotone, du visible jusqu'aux régions NIR du spectre électromagnétique, et sa pente varie en fonction du type de sol [Huet87]. Dans la partie visible du spectre, la réflectivité des feuilles est plus faible que celle du

sol, tandis que la réflectivité NIR des feuilles est plus forte que celle du sol. Ce comportement est utile pour l'utilisation de la réflectivité dans les applications agricoles et pour séparer visuellement les parcelles, des sols non cultivés [Baus93, Rond96]. Ces indicateurs ont pour certains le défaut d'être peu estimables donc peu utilisables dans un but de prévision à long terme [Lope15].

Une approche classique pour extraire des informations des propriétés de réflectivité mesurées par télédétection repose sur le calcul d'**indices de végétation**, comme l'Indice de Végétation par Différence Normalisé (NDVI). Le NDVI est construit à partir des canaux rouge (R) et proche infrarouge. L'indice de végétation normalisé met en valeur la différence entre la bande visible du rouge et celle du proche infrarouge :

$$NDVI = \frac{NIR - R}{NIR + R} .$$

Cet indice est élaboré pour être sensible à la vigueur et à la quantité de végétation. Cependant, cette approche a aussi ses limites (par exemple, la saturation du signal) et peut être contaminée par différentes sources, comme les illuminations, la présence de nuages, la géométrie des observations, la structure 3D des zones de végétation différente, ou la réflectivité du sol [Gobr97]. Des études ont montré que le NDVI cumulé était significativement corrélé avec les rendements agricoles [Grot93, Dora95]. Les méthodes statistiques pour prédire les rendements à partir de données télédétectées MODIS (Moderate Resolution Imaging Spectroradiometer) ont été présentées, par exemple par [Dora07] ou [Beck10].

Des approches plus complexes font appel aux modèles de transfert radiatif pour le couvert végétal pour en déduire des variables clés de végétation - comme le fAPAR (Fraction of Absorbed Photosynthetically Active Radiation) - à partir de la réflectivité du couvert végétal. Cela permet l'accès à des propriétés de végétation plus décontaminés des facteurs externes [Pint09]. En particulier, le fAPAR agit comme un indicateur intégré du status et de la santé de la végétation, et joue un rôle majeur dans la surveillance de la productivité végétale [Prin95].

Les informations dérivées des satellites peuvent aussi être reliées par régression aux paramètres physiologiques, ou directement au rendement [Rose94]. La plupart des applications utilisant cette approche ne permettent d'établir une relation que pour une région et une culture données. Les modèles de simulation de la croissance des cultures ou les modèles statistique utilisent de plus en plus les données météorologiques issues de la télédétection.

Si le rendement est bien lié à la surface des feuilles pour les pâturages et les prairies, tel n'est pas toujours le cas pour les céréales et encore moins pour les cultures permanentes [Meye94]. Les techniques de télédétection par satellite deviennent de plus en plus opérationnelles pour assurer des observations répétées avec une résolution suffisante au niveau du champ. Il n'en reste pas moins que l'étendue temporelle des données est parfois faible pour les bases de qualité (10 ans pour la base de données MODIS, et 30 ans pour la base issus de AVHRR [4]).

La durée des études proposées dans la littérature est souvent courte pour une bonne validation d'un modèle statistique. Par exemple, [Lope15] ont basé leur étude sur 13

4. https://phenology.cr.usgs.gov/ndvi_avhrr.php

ans de données, et [Kowa14] ont eux travaillé sur 11 ans. Les études récentes sont souvent limitées par les disponibilités des entrées du modèle comme le NDVI.

En résumé

— Les techniques de télédétection ont déjà été beaucoup utilisées pour la prévision des rendements sans résultats exceptionnels lorsqu'on les utilise seules. [Bass13] expliquent cela par le fait que, dans les pays développés, le système agricole est très stratifié, avec de petites exploitations, et par le fait que les images de télédétection ne parviennent pas encore à séparer les parcelles de cultures différentes. Cependant, les améliorations en reconnaissance d'image et la diminution du coût pour l'obtention d'images à haute résolution, vont probablement permettre à cette technique de se développer efficacement.

— Ces variables sont directement liées à la végétation contrairement aux variables météo, même si ce ne sont pas des mesures in situ. Elles peuvent donc, à ce titre, être utilisées dans des modèles d'impacts pour prédire le rendement.

— L'intégration de la télédétection dans les modèles mécanistes semble être une alternative intéressante pour la prévision de rendement. La télédétection peut aider à quantifier l'état des cultures à tout moment de la saison de croissance, et les modèles mécanistes peuvent décrire la croissance quotidienne de la plante [Maas88]. Les techniques de télédétection peuvent aussi aider à les calibrer.

1.3 Variables usuelles dans les modèles mécanistes, ou variables directes

Les modèles mécanistes ont besoin d'informations relatives à la gestion des cultures, au sol, et à l'atmosphère. La complexité des données d'entrée varie beaucoup, des données horaires, aux données quotidiennes, hebdomadaires, mensuelles, ou saisonnales [Nix84]. Cependant, les données quotidiennes sont les plus utilisées. [Hunt98] a définit une liste de données minimales MDS (Minimum Data Set) pour faire fonctionner un modèle de culture à un endroit précis (Tableau II.1).

Si plus de données sont disponibles, elles peuvent servir à paramétrer des fonctions plus complexes au sein du modèle mécaniste. Par exemple, le MDS est généralement utilisé pour calculer l'évapotranspiration potentielle grâce à la méthode de Priestley-Taylor. Si l'humidité et la vitesse du vent sont disponibles, alors la méthode de Penman-Monteith FAO-56 peut être préférable [Bass13]. La gestion et les conditions initiales peuvent être obtenues par des sondages envoyés aux cultivateurs, ou en utilisant des connaissances d'experts du domaine.

La méthode traditionnelle de prévision de rendements reste l'évaluation du statut des cultures par des experts du domaine (Inde, France (Arvalis), ou États-Unis). Les observations et les mesures sont faites tout au long de la saison de croissance : nombre de grains, nombre d'épis, leur taux de fertilité, le pourcentage de dégâts provoqués par les maladies ou les parasites, le pourcentage de graines infestées, etc. À partir des données ainsi obtenues, les rendements peuvent être estimés en utilisant des relations statistiques ou physiques, ou par la connaissance d'expertises locales. L'objectif des méthodes de prévisions de rendement est de fournir le plus tôt possible dans la période

Données reliées à la description du lieu	latitude et longitude, altitude, température annuelle moyenne, et aspect typographique du lieu et du climat
Météo	rayonnement incident global et quotidienne du sol, température minimale et maximale quotidienne, précipitation quotidienne
Sol	Type de sol, profondeur (divisé en n couches), texture du sol, carbone organique du sol, densité apparente, niveaux d'azote, pH
Conditions initiales du système	cultures précédentes, résidus laissés sur le sol (si présents), eau présente initialement dans le sol et quantité d'azote.
Gestion des champs et de la culture	nom et type de la culture, date de semis, espacement des rangées, nombre de plantes par mètre carré, quantité d'irrigation/azote apportée, dates d'irrigation/fertilisation, type d'engrais.

TABLEAU II.1 – La liste de données minimales MDS (Minimum Data Set) pour faire fonctionner un modèle mécaniste à un endroit précis [Hunt98].

de croissance, des estimations précises en considérant ou non, les effets de la météo et du climat.

Prédire les productions agricoles signifie aussi connaitre ou savoir prédire d'autres facteurs importants, comme le fait de savoir quantifier la surface plantée au début de la saison de croissance, mais aussi quantifier la surface récoltée. [Shar04] proposent une analyse de l'estimation des surfaces cultivées. Si nécessaire, un classement spatial des données est mis en place, tel que celui des découpes administratives européennes en provinces, régions, ou départements de NUTS (Nomenclature des Unités Territoriales Statistiques). En dehors de l'Europe (Russie, Kazachstan, Chine, Inde) on suit le système GAUL (Global Administrative Unit Layers). Les deux systèmes ont des régions organisées en différents niveaux administratifs. Le plus haut niveau est celui des pays. Les pays sont divisés en grandes régions (de la taille des régions françaises), et les régions sont divisées en zones plus petites (de la taille des départements français). Chez [Kowa14] on utilise aussi le découpage spatial en zones climatiques. Aux États-Unis, on utilise le découpage cantons/districts agricoles/états de la classification FIPS (Federal Information Processing Standards). Quant à [Agra01], leur étude est basée sur un découpage de l'Inde en zones agroclimatiques.

En résumé

Que ce soit le minimum, le maximum, la moyenne, en journalier, ou en mensuel, ou encore cumulée sous forme de degré-jours (de croissance, de mort, ou cumulés), ou de dépassement de seuil, la température est de loin la variable météo la plus utilisée[a] avec les précipitations (mensuelles, quotidiennes, cumulées ou non). L'effet de l'eau est aussi caractérisée par le déficit hydrique sous forme d'évapotranspiration (réelle, de référence ou potentielle), de différence entre précipitation et évapotranspiration, d'humidité du sol, de capacité au

champ, d'indice de stress hydrique, de déficit ou surplus eau. Dans les variables agroclimatiques, on rencontre aussi des dates (début et fin de croissance, dernier gel printanier, premier gel automnal) ou des nombres de jours (longueur de la saison sans gel) mais aussi des variables relatives aux effets du soleil (rayonnement, nombre d'heure d'ensoleillement). Les mesures ou comptages directement effectués sur les plantes sont souvent réalisés dans des parcelles test, soulevant la question d'un bon échantillonnage. Enfin, lorsqu'ils sont utilisés, les indices de végétation (fAPAR, LAI, NDVI) sont souvent peu accompagnés par d'autres prédicteurs, directs ou indirects.

Les données directes (taille des épis, ou NDVI) ont l'inconvénient de ne pas être prévisibles et donc non utilisables par un modèle dans un but de prévision à long terme. L'utilisation que l'on souhaite faire des modèles conditionne donc le choix des prédicteurs en entrée. De plus, l'obtention de telles données est souvent coûteuse (obtenues à l'aide de sondages auprès des cultivateurs), ce qui amène à questionner leur efficacité par rapport à des données indirectes, souvent moins coûteuses.

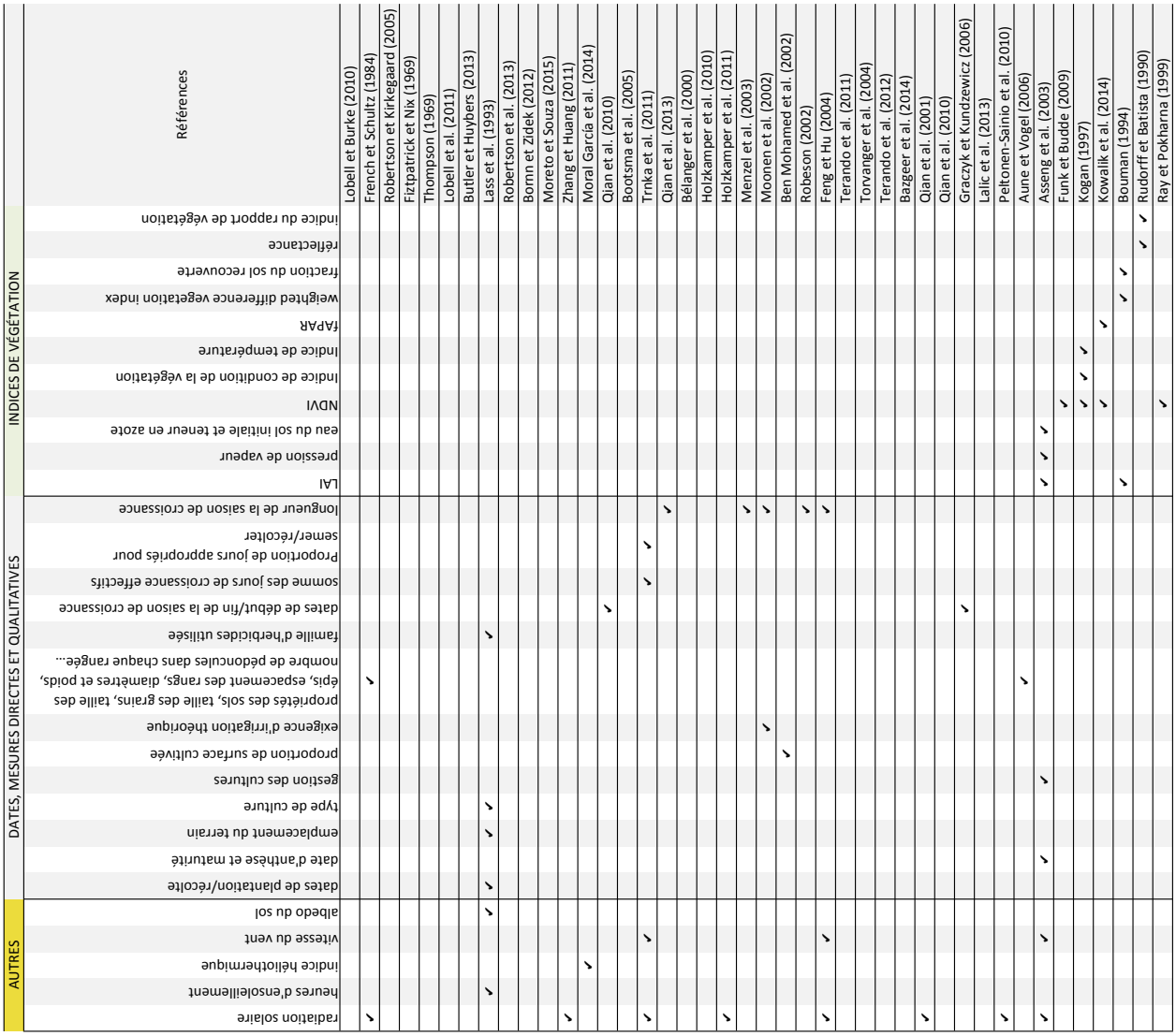
Une telle abondance de prédicteurs, soulève rapidement plusieurs questions. Lesquels sont les plus informatifs (selon l'application visée) ? Combien est-il nécessaire/raisonnable d'en garder ? Existe t'il une classification robuste de ces prédicteurs ? Dans la littérature, un grand nombre de ces prédicteurs ont parfois été utilisés, et souvent peu. Seulement, une ou deux études ont rassemblé un grand nombre de prédicteurs afin de comparer leur pouvoir informatif et leur complexité.

a. et la plus transformée

Le tableau des deux pages suivantes compare les données utilisées et le lieu d'étude de plus de quarante articles cherchant à estimer les rendements agricoles. Les années qui ont servi pour les données (ou leur nombre lorsque les données sont simulées) sont également mentionnées.

Références	Années	Localité	TEMPERATURE																				
			températures mensuelles	températures moyennes	températures maximales	températures minimales	température optimale	périodes de chaleur	indice de stress thermique	écarts de température	indice pour les nuits fraîches	degrés-jours de croissance	degrés-jours de croissance effectifs	unités thermiques analogiques cumulées entre plantation récolte	jours avec des températures inférieures à Tmin ou supérieures à Tmax / degrés-jours de mise à mort	heures cumulées d'exposition à un intervalle de température	unités thermiques de maïs	nombre de jours de gel	anomalie de température de surface de la mer	température du point de rosée	nombre de jours sans gel	date du dernier gel (mortel) printanier	date du premier gel (mortel) d'automne
Lobell et Burke (2010)	40	Affrique Sub-Sahari.		✔																			
French et Schultz (1984)	1964-1975	Sud de l'Australie		✔																			
Robertson et Kirkegaard (2005)	1991-2003	Australie																					
Fiztpatrick et Nix (1969)		Sud de l'Australie																					
Thompson (1969)	100	Corn Belt	✔																				
Lobell et al. (2011)	28	France, Brazil, China,	✔																				
Butler et Huybers (2013)	1981-2008	Moitiée Est des US										✔			✔								
Lass et al. (1993)	1979-1985	Idaho										✔		✔									
Robertson et al. (2013)	1978-2007	Canada		✔	✔	✔						✔				✔							
Bornn et Zidek (2012)	1976-2006	Canada																					
Moreto et Souza (2015)	1990-2010	Brésil		✔																			
Zhang et Huang (2011)	1980-2008	Chine		✔																			
Moral García et al. (2014)		Espagne		✔						✔	✔												
Qian et al. (2010)	1900-2007	Canada			✔	✔	✔	✔				✔	✔		✔		✔				✔	✔	✔
Bootsma et al. (2005)	1961-1990	Canada									✔		✔				✔						
Trnka et al. (2011)	1971–2000	Europe			✔	✔																✔	
Qian et al. (2013)	1961-1990	Canada			✔	✔							✔	✔			✔						
Bélanger et al. (2000)	1961-1990	Canada											✔				✔						
Holzkamper et al. (2010)		Suisse										✔			✔			✔					
Holzkamper et al. (2011)	1981-2009	Suisse			✔	✔						✔											
Menzel et al. (2003)	1951-2000	Europe		✔		✔																	
Moonen et al. (2002)	1878-1999	Italie		✔	✔	✔		✔													✔	✔	
Ben Mohamed et al. (2002)	1970-2000	Sahel			✔	✔													✔				
Robeson (2002)	1906–1997	Illinois				✔																✔	
Feng et Hu (2004)	1951–2000	Etats-Unis			✔	✔						✔									✔		
Terando et al. (2011)	1951–2010	États-Unis						✔	✔									✔					
Torvanger et al. (2004)	1958-2001	Norvège										✔											
Terando et al. (2012)	1961–2010	États-Unis						✔	✔									✔					
Bazgeer et al. (2014)	1985-2011	Iran								✔													
Qian et al. (2001)					✔	✔																	
Qian et al. (2010)	1961-1990	Canada										✔	✔				✔				✔	✔	✔
Graczyk et Kundzewicz (2006)	1951-2010	Pologne										✔									✔	✔	
Lalic et al. (2013)	1985–2005	Pays bas													✔								
Peltonen-Sainio et al. (2010)	1975–2008	Europe		✔	✔	✔																	
Aune et Vogel (2006)	5	États-Unis																					
Asseng et al. (2003)	1980-2010	Divers			✔	✔														✔			
Funk et Budde (2009)	2000-2008	Zimbabwe																					
Kogan (1997)	10	Divers																					
Kowalik et al. (2014)	1999–2009	Europe																					
Bouman (1994)	10	Europe																					
Rudorff et Batista (1990)	4	Brésil																					
Ray et Pokharna (1999)	10	Inde																					

Références	PRÉCIPITATION																					
	précipitations	précipitations quotidiennes	précipitations mensuelles	précipitations annuelles	précipitation avant/pendant la floraison	bilan hydrique	humidité du sol / humidité relative	indice de sécheresse	Indice de précipitation standardisé	évaporation	évapotranspiration	évapotranspiration réelle	évapotranspiration de référence	évapotranspiration potentielle	évapotranspiration potentielle (ou de référence) - précipitations	drainage / stockage d'eau dans le sol / ruissellement	indice de stress hydrique	Indice de sévérité de sécheressede Palmer (PDSI)	capacité au champs	moyenne quotidienne de l'humidité relative de l'air	nombre de jours avec déficits hydriques	longueur de la saison des pluies / saison sèche
Lobell et Burke (2010)	✔																					
French et Schultz (1984)	✔						✔			✔												
Robertson et Kirkegaard (2005)		✔	✔								✔					✔						
Fiztpatrick et Nix (1969)												✔										
Thompson (1969)			✔																			
Lobell et al. (2011)			✔																			
Butler et Huybers (2013)																						
Lass et al. (1993)							✔															
Robertson et al. (2013)		✔	✔																			
Bornn et Zidek (2012)												✔		✔			✔		✔			
Moreto et Souza (2015)	✔					✔										✔						
Zhang et Huang (2011)	✔																					
Moral García et al. (2014)			✔	✔	✔			✔														
Qian et al. (2010)		✔																				
Bootsma et al. (2005)						✔																
Trnka et al. (2011)																				✔	✔	
Qian et al. (2013)	✔						✔															
Bélanger et al. (2000)															✔							
Holzkamper et al. (2010)	✔	✔										✔		✔				✔				
Holzkamper et al. (2011)															✔							
Menzel et al. (2003)																						
Moonen et al. (2002)		✔					✔			✔			✔	✔								
Ben Mohamed et al. (2002)		✔	✔														✔				✔	✔
Robeson (2002)																						
Feng et Hu (2004)		✔					✔		✔													
Terando et al. (2011)																						
Torvanger et al. (2004)	✔																					
Terando et al. (2012)																						
Bazgeer et al. (2014)	✔																					
Qian et al. (2001)	✔																					
Qian et al. (2010)						✔																
Graczyk et Kundzewicz (2006)																						
Lalic et al. (2013)	✔											✔									✔	
Peltonen-Sainio et al. (2010)	✔												✔									
Aune et Vogel (2006)																						
Asseng et al. (2003)	✔						✔				✔											
Funk et Budde (2009)																						
Kogan (1997)																						
Kowalik et al. (2014)																						
Bouman (1994)																						
Rudorff et Batista (1990)											✔	✔										
Ray et Pokharna (1999)												✔										



1.4 Les bases de données disponibles

Cette section rassemble une liste non exhaustive des bases de données les plus connues concernant les données du sol, les données météorologiques ou climatiques, les données de télédétection (dont le GIS (Geographic Information System)), ainsi que les données agricoles [5].

Données du sol	• ISRIC – World Soil Information • FAO Soils Portal • European Soil Portal – Soil Data and Information Systems • EuDASM (European Digital Archive on Soil Maps of the World) • GEOSS (Group on Earth Observations) Portal • Harmonized World Soil Database V. 1.2 • Global Soil Map

5. Dernière date d'accès aux ressources : 20 octobre 2017

Données météo	• USA-NOAA • Environment Canada • European Centre for Medium-Range Weather Forecasts • Japan Meteorological Agency • NASA POWER (Prediction Of Worldwide Energy Resource Project)
Données satellites	• USGS Earth Explorer • GIMMS data set • SPOT Vegetation et VITO (Flemish Institute for Technology Research for SPOT Vegetation) • MODIS • GLCF (Global Land Cover Facility)
Données climatiques	• IPSL pour les simulations • The GOES Project • The Climate Change Knowledge Portal, The World Bank • NASA Global Change Master Directory • IRI/LDEO Climate Data Library • NARCCAP (North American Regional Climate Change Assessment Program)
Sources de données GIS	• Natural Earth • USGS Earth Explorer • NASA SEDAC (NASA Socioeconomic Data and Application Center) • UNEP Environmental Data Explorer
Données agricoles	• USDA Open Data Catalog • FAOSTAT • OECD.Stat • EUROSTAT

Le lecteur intéressé qui rechercherait d'autres bases de données (agricole, météo ou climatique) peut utiliser les lien hypertextes dans la version pdf du manuscrit pour plus de renseignements (échelles, résolutions spatiales et couvertures temporelles).

2 Modèles mécanistes [6]

2.1 Le principe

Les modèles mécanistes sont des représentations informatisées de la croissance des plantes, de leur développement et des rendements, le tout modélisé par des équations physiques fonctions des caractéristiques du sol, de la météo ou encore des pratiques agricoles [Hoog04] [7].

Avec l'émergence de l'informatique, les premiers modèles mécanistes sont nés de la physiologie végétale dans les années 60, dans l'étude des phénomènes d'interception

6. aussi appelés, modèles biophysiques, de croissance végétale, de culture, de simulation de culture, déterministes, ou encore traitant des processus en jeu (process-based)

7. La littérature confond parfois les modèles agronomiques (simulant diverses variables liés aux cultures et à leur environnement) et les modèles prédisant les rendements au sens strict (appelés modèles mécanistes ici). Cette confusion, provient probablement du fait que des modèles agronomiques peuvent aussi prédire des rendements.

de la lumière, la photosynthèse, l'absorption du dioxyde de carbone, la respiration, la production de biomasses par les organismes végétaux, ou encore la perte de dioxyde de carbone pendant la respiration. Bon nombre de ces modèles étaient très sophistiqués et se basaient principalement sur l'analyse de la croissance. Par la suite, la modélisation s'est intéressée au fonctionnement des stomates (niveau cellulaire), à la résistance à la diffusion ainsi qu'à l'interception de la lumière par les feuilles (niveau de l'organe) et à l'interception de la lumière par le couvert végétal (plante et culture). D'autres simulations plus réalistes sur les besoins hydriques et la consommation en eau ont ensuite été effectuées, l'assimilation des éléments minéraux venant généralement en dernier lieu. Les modèles mécanistes décrivent souvent des procédés végétaux qui changent rapidement sur des petites échelles de temps (par exemple, les phénomènes photosynthétiques ou de respiration changent rapidement au cours d'une journée, puisque le rayonnement et la température changent).

Le plus souvent [8], les modèles mécanistes sont à localité précise, au sens où ils prédisent la croissance et le développement d'une plante pour un espace restreint, plutôt que pour une région ou un pays [Schu00]. La plupart des variables utilisées sont aussi à localité précise : les informations à propos du sol sont par exemple, souvent obtenues à partir d'un profil de sol vertical. La gestion des cultures concerne le plus souvent les variables agricoles locales, à l'échelle du champs. Beaucoup d'études ont utilisé des données spatiales et géographiques GIS, dans leur méthodes pour prédire les rendements [Enge97, Hart99, Priy01].

[Jone98] ont indiqué que les modèles de croissance ont été développés pour un large panel de cultures et pour une variété d'applications, telle que la gestion de l'irrigation, des engrais, des parasites, la planification des terres ou la rotation des cultures. Des modèles plus complexes peuvent fournir des explications sur le système sol-plante-atmosphère et requièrent souvent une grande quantité de données en entrée, la plupart étant seulement disponible à petite échelle. Pourtant, selon [Bass13], les modèles mécanistes sont rarement utilisés pour répondre à des problèmes décisionnels. En revanche, ils sont souvent utilisés à des fins académiques pour une meilleure compréhension des procédés végétaux en jeu et de leurs interactions.

2.2 Avantages et difficultés

Les modèles mécanistes représentent une classe de modèles très vaste. Un modèle mécaniste à une structure représentant explicitement une compréhension des propriétés biologiques, chimiques et/ou physiques des procédés. Aussi, si le modèle échoue à l'extrapolation, on peut s'aider de l'architecture physique construite (et des hypothèses physiques faites) pour essayer de savoir d'où vient le problème.

On reproche souvent à ces modèles d'être lourds à paramétrer et à utiliser. Ils ont souvent besoin d'un nombre important de variables pour être calibrés, et sont donc difficilement applicables dès lors que ces entrées ne sont pas disponibles. D'après [Oles01], les modèles mécanistes - souvent spécifiques à un site particulier - ne sont pas très bien équipés pour gérer les extrapolations spatiales. A l'échelle du champ, il est relativement facile de déterminer les entrées nécessaires, mais pour des échelles plus grandes, cela devient problématique. Par exemple, il est difficile de déterminer ou d'obtenir les informations nécessaires à la gestion des cultures à l'échelle d'un pays, la

8. mais pas toujours, comme le montre l'exemple d'Orchidée page 27

constitution des sols, ou bien savoir quelle culture a été plantée, où et quand - ce qui est pourtant déterminant pour les prévisions de rendement durant la saison de croissance.

Cependant, de nouveaux modèles comme le modèle SALUS (System Approach to Land Use Sustainability) ont été développés pour dépasser ces problèmes. De même, [Bond07] et [Chal04] ont développé un modèle de simulation de culture qui peut tourner à l'échelle régionale avec moins de demandes en terme d'entrées et de données de calibration.

2.3 Étude bibliographique

Les modèles mécanistes sont nombreux à travers la monde. Dès le début des années 80, on trouve une abondante littérature sur les modèles mécanistes [Day85, Whis86, Loom79, Wisi87]. Par exemple, [Asse13] utilisent plus de 27 modèles de croissance du blé dans leur étude comparative. Certains des modèles sont très simples avec moins de 7 paramètres nécessaires pour décrire une culture particulière [Bond07]. D'autres, prennent en compte les processus détaillés de la photosynthèse à l'échelle de la feuille et nécessitent alors beaucoup de paramètres définis par l'utilisateur. Les modèles mécanistes ont été très utilisés pour évaluer les conséquences de la variabilité inter-annuelle du climat sur l'agriculture [Paz07, Seme07, Chal08]. Par exemple, certains modèles mécanistes sont censés capter les effets et l'échéance des cycles secs/humides sur la croissance des cultures, ce qui peut aider les cultivateurs dans la gestion de leur exploitation [Shin09]. Par exemple le "Yield prophet" est un modèle utilisable en ligne pour informer en temps réel les cultivateurs et les spécialistes du domaine à propos de l'évolution potentielle des cultures, évaluant les risques possibles, et faisant des propositions pertinentes aux cultivateurs pour la gestion de leur exploitation.

En Europe, les modèles mécanistes sont rassemblés au sein de grands projets comme le CGMS (Crop Growth Modelling System) du centre commun de recherche européen [Cant05, Baru07], le MCYFS de MARS (MARS Crop Yield Forecasting System) [Lope15], mais aussi le modèle de surface continentale ORCHIDEE (ORganizing Carbon and Hydrology In Dynamic Ecosystems Environment) de l'IPSL (Institut Pierre Simon Laplace). Certains modèles sont génériques et peuvent simuler diverses cultures (comme WOFOST [Diep89, Itte03], ou CropSyst). D'autres modèles sont prévus pour une culture spécifique (WARM pour le riz, CANEGRO pour la canne-à-sucre). Ailleurs, [Duch86] a utilisé le modèle CERES-Maize pour prédire les rendements de maïs pendant la saison de croissance pour une prédiction à grande échelle géographique. [Hodg87] se sont eux aussi servis du modèle CERES-Maize pour estimer les fluctuations annuelles de la production de maïs aux États-Unis (Corn-Belt) pour les années 1982-1985. [Hask95] ont examiné la performance du modèle de soja GLYCIM en modélisant les rendements régionaux de soja en Iowa à l'échelle des cantons, et de l'état sur une période de 20 ans. [Chip99] ont simulé les rendements de blé avec le modèle de blé CERES à l'échelle du district, rassemblant des centaines d'exploitations, différents sols, différents climats, et différentes pratiques agricoles [Hoog99].

Les paragraphes ci-dessous décrivent rapidement les modèles mécanistes les plus connus et les plus utilisés.

WOFOST (World Food Studies)
WOFOST simule la croissance quotidienne d'une plante. Le modèle calcule la lumière interceptée et la convertie en matière végétale (photosynthèse brute, et potentielle).

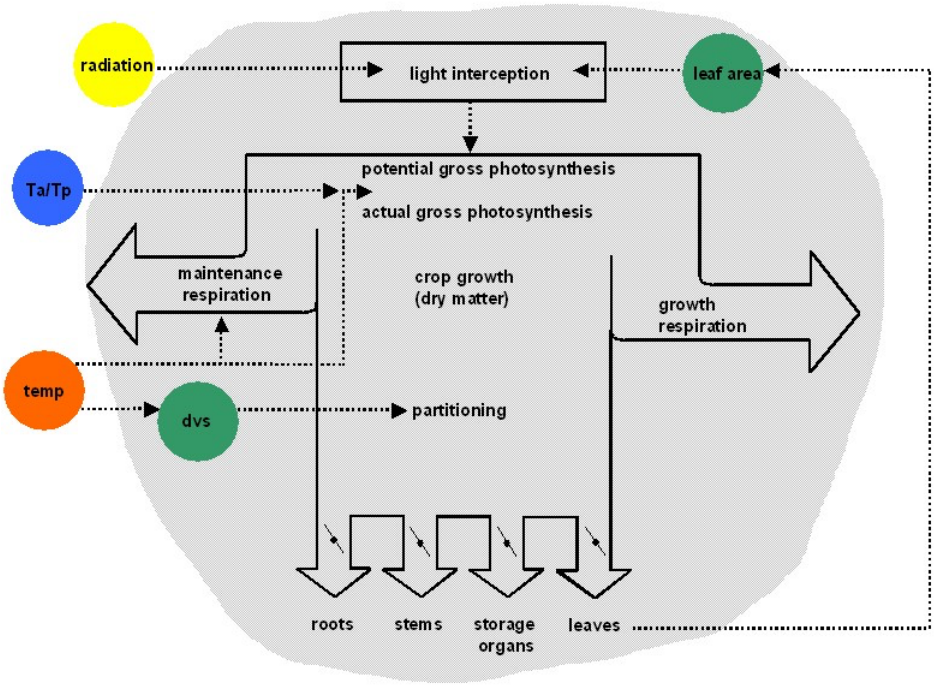


FIGURE II.2 – Flux quotidien de matière sèche dans le modèle WOFOST [9]

Quand il n'y a pas assez d'eau disponible, la matière végétale potentielle est réduite. Comme le montre la Figure II.2, après avoir retiré les coûts de maintenance (maintenance respiration), la nouvelle matière végétale formée est distribuée dans différents organes : racines, tiges, feuilles, et organes de stockage (graines/tubercules). Les parts reçues par les différents organes dépendent de l'âge de la plante. Pendant la transformation de la matière végétale aux organes, une partie de la matière est perdue dans les coûts de construction (respiration). Tôt dans la saison, la plupart de la matière végétale est investie dans les racines, les tiges et les feuilles. Plus les feuilles sont nombreuses, plus la quantité de lumière pouvant être interceptée augmente, donc la plante a une croissance rapide. Vers la fin de la saison, toute ou la plupart de la matière végétale est convertie en organes de stockage et les feuilles commencent à se désintégrer, réduisant ainsi la quantité de lumière interceptée : la croissance de la plante s'arrête donc. Les grains murissent.

WARM (Water Accounting Rice Model)
WARM est un modèle qui donne des résultats quotidiens, pour la simulation de la croissance et du développement des rizières. Le modèle prend en compte les principaux processus qui caractérisent ce système de culture. Il simule la croissance de la biomasse, l'évolution des feuilles, les événements phénologiques, l'effet de l'eau sur le profil thermique du sol, le développement des maladies, et les particularités hydrologiques des sols de rizières.

CropSyst (Cropping Systems Simulation Model)
CropSyst est un modèle simple à utiliser, et pouvant travailler sur plusieurs années et

9. Figure tirée de http://marswiki.jrc.ec.europa.eu/agri4castwiki/index.php/Crop_Simulation

plusieurs cultures, tout en étant un modèle donnant des résultats journaliers. Le modèle a été développé en tant qu'outil d'analyse pour étudier les effets de gestion des cultures sur la productivité et l'environnement. Le modèle simule la demande en eau du sol, la demande plante-sol en azote, le couvert végétal, la croissance des racines, la production de matière sèche, les rendements, la production, la décomposition des résidus, et l'érosion. Les options de gestion incluent : la sélection de la culture, la rotation des cultures, l'irrigation, la fertilisation en azote, le labourage, et la gestion des résidus.

Canegro
Canegro simule la croissance et le développement de la canne à sucre, à partir de données météo journalières et d'information sur la gestion de la culture. Il simule le développement du couvert végétal au niveau de la tige et des feuilles, l'absorption des rayonnements grâce à l'indice de surface foliaire (LAI), le bilan hydrique, l'accumulation de biomasse, et un découpage de la biomasse en ses différents constituants.

DSSAT (Decision Support System for Agrotechnology Transfer)
DSSAT est un ensemble de programmes informatiques qui simulent la croissance végétale [Jone03a]. Depuis les années 80, des agronomes issus de plus de 100 pays l'ont utilisé pour tester des pratiques agricoles [Thor08]. Il a aussi servi à évaluer les impacts potentiels du changement climatique sur l'agriculture, et à tester des méthodes adaptatives [ICAS14]. DSSAT est construit sous forme de plusieurs modules, avec différentes options possibles pour représenter des procédés de la croissance d'une plante, tels que l'évapotranspiration, ou l'accumulation de matière organique [Port09, Meng08, Gijs02]. DSSAT nécessite en entrée des paramètres relatifs aux caractéristiques du sol, à la météo, aux pratiques agricoles comme les engrais ou l'irrigation, et aux caractéristiques des variétés [Wu13].

LINGRA (diminutif de LINTUL-GRASS, pour Light INTerception and UtiLization simulator GRASS)
LINGRA est un modèle de croissance des prairies utilisé pour prédire le développement du ray-grass annuel en Europe sous des conditions d'apport en eau limité [Scha98]. Le modèle comporte des routines pour l'absorption et l'apport de lumière, les fréquences de labourage, l'aspect des feuilles, les contenus en eau du sol, et l'évaporation. En 1996, LINGRA a été inséré dans le projet CGMS (Crop Growth Monitoring System) pour prédire les rendements à travers l'Europe [Voss95, Boum96].

2.4 Un exemple à grande échelle : le modèle Orchidée

ORCHIDEE (Organising Carbon and Hydrology In Dynamic Ecosystems) est le schéma de surface du modèle de climat de l'IPSL [Krin05]. Il résout explicitement les bilans d'eau et d'énergie des surfaces continentales, ainsi que la phénologie et le cycle du carbone de la biosphère terrestre. Forcé par des jeux de données climatiques globaux ou régionaux, ORCHIDEE peut aussi être utilisé en interaction continue avec un modèle d'atmosphère pour étudier l'évolution du climat (passé, présent, futur). Ce modèle décrit la diversité de la végétation (cultures agricoles, herbacées, arbres tropicaux, tempérées et boréales ; décidues ou non) à l'aide de 3 modules (Figure II.3).

— Le premier est appelé SECHIBA - Schématisation des EChanges Hydriques à l'Interface entre la Biosphère et l'Atmosphère [Duco93, De R98]. Il simule les échanges biophysiques d'eau et d'énergie entre les surfaces continentales et l'atmosphère à

l'échelle de la demi-heure. SECHIBA résout également les mécanismes de photosynthèse, de respirations de croissance, et de transpiration de la végétation.

— La composante biogéochimique d'ORCHIDEE est le module STOMATE - Saclay Toulouse Orsay Model for the Analysis of Terrestrial Ecosystems [Viov97]. STOMATE calcule quotidiennement, des processus tel que la phénologie, l'allocation du carbone au sein de la plante, la décomposition de la litière et la respiration du sol. STOMATE fournit à SECHIBA la description physique de la végétation nécessaire pour calculer les flux (par exemple, l'indice foliaire LAI). En retour, il reçoit les facteurs environnementaux et climatiques qui affectent le développement de la végétation (par exemple, le stress hydrique et thermique). Le pas de temps de ce module est de un jour.

— Enfin, un module provenant du modèle LPJ [Sitc03] décrit la dynamique de la végétation naturelle potentielle, c'est à dire l'évolution à long terme d'un type de végétation par rapport à un autre. Il inclut l'apparition et la disparition de plantes en fonction de critères climatiques, la compétition pour la lumière, et le rôle des feux. Le pas de temps de ce module est de un an [Ngo-07].

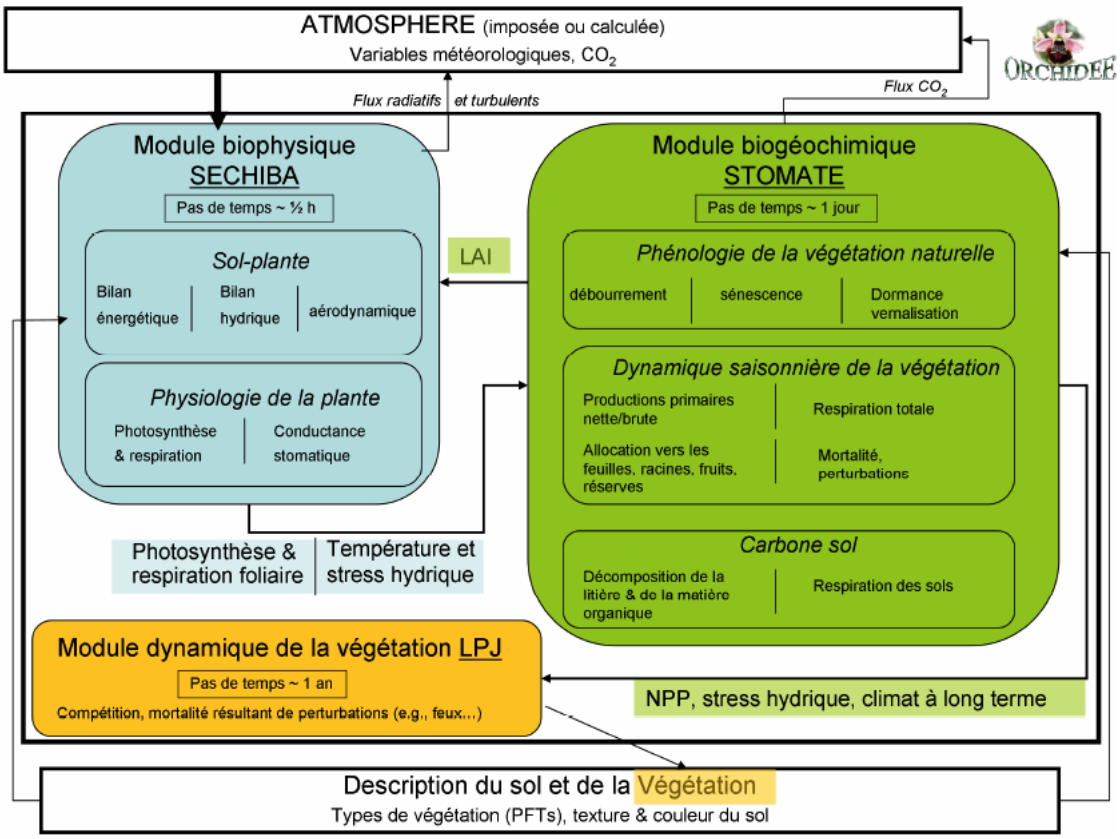


FIGURE II.3 – ORCHIDEE est le schéma de surface du modèle de climat de l'IPSL. Il résout explicitement les bilans d'eau et d'énergie des surfaces continentales, ainsi que la phénologie et le cycle du carbone de la biosphère terrestre (tiré de https://www6.inra.fr/basc/Recherche/Modeles/ORCHIDEE).

3 Modèles statistiques [10]

3.1 Le principe

Comme on a pu le voir dans les sections précédentes, le moyen le plus courant pour analyser les rendements est de construire un modèle sophistiqué, basé sur la connaissance des experts en agriculture, climatologie ou hydrologie, en prenant en compte beaucoup de variables différentes. Cette approche peut donner de bons résultats, mais reste relativement complexe, dépendant souvent de beaucoup d'hypothèses. Il existe une autre possibilité pour modéliser ce genre de phénomènes. À l'inverse de cette approche tournée vers les processus, il est possible de négliger la plupart des connaissances d'experts et d'essayer de construire un modèle basé sur la connaissance provenant des données. Le processus de modélisation commence avec les données et a pour but d'en extraire autant d'information que possible, avec le plus petit nombre d'hypothèses. Il s'agit donc de transformer les données observées d'une ou plusieurs expériences en un modèle probabiliste. Les modèles d'impact statistiques sont généralement basés sur une régression statistique [Sult09]. La calibration des modèles utilise une base de données qui inclut, en entrée, les variables météorologiques (plus toute autre information a priori disponible et pertinente) et en sortie le rendement ou la production. Les méthodes développées dans cette section sont parfois appelées méthodes empiriques ou statistiques descriptives, car les méthodes développées ont surtout pour but de résumer et de décrire les données empiriques (c'est-à-dire issues de l'expérience).

Les techniques statistiques utilisent des échantillons choisis de façon à ce qu'ils soient représentatifs de la variabilité observable des phénomènes [11]. Les données ainsi recueillies servent ensuite à élaborer des modèles permettant de calculer directement ou non le rendement ou la production à partir des mesures faites sur un nombre limité d'échantillons. L'exactitude des extrapolations réalisées par les modèles dépend alors étroitement de la méthode d'échantillonnage retenue, notamment du choix des emplacements, des situations et des paramètres étudiés. Les prévisions sont d'autant moins précises que les situations réelles diffèrent des situations de référence, pour lesquelles les modèles ont été élaborés : en règle générale, les anomalies importantes et inhabituelles, exigeant tout particulièrement le recueil d'informations et l'établissement de prévisions, correspondent également aux situations dans lesquelles les techniques statistiques donnent de moins bons résultats.

L'examen des rendements annuels successifs fait souvent apparaître l'existence d'une tendance temporelle générale : pour la construction du modèle de prévision, on sépare l'effet de cette tendance générale de l'effet des conditions météorologiques [Palm94]. En pratique, on étudie d'abord l'évolution sur le long terme des rendements en fonction du temps, indépendamment des conditions météorologiques. Ensuite, on s'efforce d'expliquer les fluctuations des résidus par rapport à la tendance, par les variables météorologiques. Par exemple, la solution retenue par [Palm94] pour modéliser la tendance générale est l'ajustement temporel du rendement y à l'équation $y = b_0 + b_1 t + b_2 t^2 + e$. On verra que ce n'est pas la seule solution. Le nombre de variables explicatives potentielles peut cependant être élevé ce qui s'avère préjudiciable pour les modèles [Kram04]. Pour réduire ce nombre de variables, on verra que plusieurs solutions statistiques existent (voir Chapitre IV page 123).

10. Aussi appelé "modèle de régression empirique", "modèle statistique empirique", "modèle agro-météorologiques" [Crai13] ou "modèle statistique descriptive", au sens où ils sont issus d'une expérience, d'une observation et non d'une théorie physique.

11. En statistique descriptive, une population est un ensemble fini d'objets (les individus ou unités statistiques) sur lesquels une étude se porte et dont les éléments répondent à une ou plusieurs caractéristiques communes. Un échantillon est un ensemble d'individus représentatifs d'une population.

3.2 Avantages et difficultés

Une telle approche a bien sûr des avantages et des inconvénients. Les avantages d'un modèle statistiques sont les suivants : des calculs faciles, un temps réduit pour faire tourner le modèle, et une demande en variables d'entrée limitée.

Le principal inconvénient est le fait que la méthode néglige certains faits bien connus par les experts, simplifiant fortement le système d'ensemble, menant parfois à de fausses conclusions. Comme tous les modèles, ils ne sont qu'une approximation du monde réel et beaucoup d'entre eux ne prennent pas en compte directement des facteurs importants tels que les types de graines, les maladies, les insectes, ou les effets du labourage [Jone01]. Comme les modèles statistiques ne prennent pas en compte les interactions plante-sol-atmosphère, ils sont donc limités dans l'information qu'ils peuvent transmettre (détermination des causes et des effets) aux cultivateurs (par exemple ajout ou non de fertilisants, retard dans la date de semis...). Ces modèles peuvent aussi être mal utilisés et ils sont limités dans l'information qu'ils peuvent fournir en dehors de l'intervalle de valeurs pour lequel le modèle a été paramétré. De plus, les sorties de ces modèles peuvent ne pas avoir de sens agronomiques, même s'ils sont corrects du point de vue statistique. Ces modèles ne sont pas faits pour prendre en compte le système complexe et continu sol-plante-atmosphère, ce qui peut s'avérer important lorsqu'on a affaire à des régions aux sols très différents. Par exemple, la réponse d'une culture à une certaine quantité de pluie sur un sol sableux est différente de celle d'une culture sur un sol argileux. Cependant, des modèles statistiques peuvent prendre en compte les effets de diverses catégories (par exemple, les types de sols) et appréhender leurs différences. La durée du stress hydrique subit pendant la saison de croissance (souvent ignoré dans ces modèles par manque de données), peut aussi influencer l'effet sur la plante : un stress thermique lors de la période de floraison réduira davantage le rendement qu'un stress thermique survenant pendant la phase végétative. L'inclusion de l'évapotranspiration et/ou de l'humidité initiale du sol comme entrées du modèle améliore le pouvoir prédictif du modèle, mais ne rend pas le modèle plus apte à répondre aux questions agronomiques des cultivateurs [Bass13].

Une difficulté très importante pour ces modèles vient du nombre limité de données dans des bases historiques (ou du grand nombre de valeurs manquantes). En agriculture un événement représente la production d'une année entière et il est généralement difficile d'obtenir des données pour plus d'une ou deux décennies. Il n'y a donc que 10 ou 20 échantillons disponibles pour calibrer les paramètres du modèle. Or ces bases de données doivent être suffisamment longues et riches pour documenter au mieux les relations complexes qui doivent être appréhendées par les modèles d'impact. Les modèles statistiques doivent avoir un niveau de complexité suffisant pour représenter des relations multivariées et non-linéaires de l'agronomie. Leur étalonnage avec une quantité limitée de données est un réel défi. Un vrai dilemme apparaît : la complexité des relations nécessite autant d'informations que possible dans les entrées du modèle d'impact, mais il n'y a pas suffisamment de données disponibles pour le calibrer. De ce dilemme résulte le problème de sur-apprentissage qui sera développé au Chapitre IV page 107 : l'étalonnage du modèle fonctionne très bien sur la base de données historique, mais le modèle n'est pas applicable sur de nouvelles données [Gema92, Hawk04, Bish96]. Ce problème apparaît très souvent en agriculture car les bases de données sont courtes mais les relations météo/rendement sont très complexes, exigeant de nombreux paramètres [Sult09]. Le modèle fonctionne très bien sur les données historiques utilisées pour calibrer le modèle, mais il n'est pas en mesure de

prévoir vraiment ce qui se passera sur de nouvelles données.

Une seconde difficulté provient du fait que les relations de causalité peuvent être altérées par d'autres facteurs que l'on n'a pas pris en compte. Par exemple, les effets de la précipitation peuvent être largement affaiblis/bruités par l'irrigation locale. Aussi, ce n'est pas évident de distinguer l'influence de la météo et celle d'autres facteurs importants comme l'irrigation, la fertilisation, la mécanisation, ou le type de sol. Puisque le but est de rechercher seulement l'impact de la météo sur les rendements agricoles, il est souvent nécessaire de limiter l'impact des facteurs non météorologiques pendant l'étape de pré-traitement [Kotl07a]. Ces facteurs non météorologiques ajoutent de l'hétérogénéité spatiale dans l'impact de la météo sur les cultures. Par exemple, [Schl09] ont montré comment les relations température/rendement pouvaient varier selon les différentes régions des États-Unis, et ont trouvé que les cultures du sud étaient moins sensibles aux chaleurs extrêmes. De même, [Cai12] ont montré que les relations météo/rendements de maïs aux États-Unis avaient une grande variabilité spatiale.

Les premiers modèles utilisés à grande échelle étaient des modèles statistiques. On regardait les séries temporelles des rendements moyens sur de grandes surfaces (départements ou régions) et sur plusieurs années pour reconnaitre des tendances temporelles générales dans les rendements agricoles [Thom69, Gage15]. Les écarts à la tendance sont très souvent régionalement corrélés aux données météo mensuelles moyennes de chaque année [Otto02].
Cependant, la prévision des rendements à grande échelle est un processus complexe qui demande à la fois des techniques statistiques appropriées, et de nombreuses données de qualité. Les données rassemblent parfois de très grandes surfaces mais aussi de très petites, aux cultures souvent différentes (comme c'est la cas en télédétection) et pouvant être irriguées ou non. Elles couvrent aussi des états aux politiques économiques, et au développement pouvant être différents. Les échantillons utilisés mélangent les types de sols, les taux de nuisibles, les maladies, ou les zones climatiques. Ainsi, les méthodes utilisées doivent permettre une très grande variabilité, même au sein d'une culture particulière.

Enfin, une des spécificités des modèles statistiques par rapport aux modèles mécanistes, est la nécessité de définir des hypothèses réalistes sur la distribution des résidus. Par exemple, on verra que dans la définition théorique des modèles linéaires mixtes, on suppose que les résidus sont indépendants d'un canton à l'autre, qu'ils sont indépendants d'une mesure à l'autre chez un même canton, et ont une distribution gaussienne. En pratique, ces hypothèses peuvent ne pas être vérifiées, mais il existe des généralisations des modèles statistiques qui permettent de définir des hypothèses sur les résidus qui soient plus réalistes (matrice de covariance non diagonale, par exemple). Souvent, cela complexifie davantage le modèle, ce qui n'est pas souhaitable.

3.3 Étude bibliographique

Les modèles statistiques ont beaucoup été utilisés pour étudier l'impact du changement climatique sur les rendements agricoles. La prévision anticipée des rendements agricoles repose principalement sur des modèles qui sont basés sur des régressions linaires multiples de rendements historiques et de moyennes mensuelles de température ou de précipitation [Thom69, Rudo90, Ray99, Shun04]. Le modèle statistique utilisé dans LACIE (Large Area Crop Inventory Experiment [MacD80]) était un modèle polynomial spécifique à une région, basé sur une régression linéaire multiple de la forme

$Y = \alpha X + \beta + \varepsilon$, où α et β sont deux paramètres, X est composée d'une liste de variables météorologiques, et ε représente l'erreur (partie non expliquée par les variables d'entrée). [Kand02] ont utilisé des modèles de régression multiple pour étudier les effets de la météo sur des rendements de curcuma en utilisant une base de données de 20 ans.

[Boat86] ont présenté un modèle statistique de prédiction des rendements en fonction de l'indice de stress hydrique CWSI (Crop Water Stress Index) : $y = 4290 + 3.97\ CWSI - 2.91\ CWSI^2$. CWSI peut être estimé à partir des températures journalières obtenues grâce à la télédétection. Bien d'autres indices de sécheresse peuvent aussi servir. Cette méthode reste un modèle statistique et ne donne pas d'information temporelle sur le déroulement des processus de croissance, ou l'apparition de maladies.
En Europe, Eurostats [12] a mis en place une plateforme de données agricoles pour les pays de la zone Euro facilement accessibles. Les données à l'échelle de l'état sont disponibles depuis 1960 environ, et ceux à l'échelle des régions (NUTS 2) depuis 2000. Des pays comme l'Ukraine, ou la Suisse n'en faisant pas partie, les données agricoles ne concernent pas ces pays, même s'ils peuvent être de grands producteurs. Les articles ont alors recours aux données des différents ministères de l'agriculture de chaque pays. De plus, les politiques agricoles sont encore plus diverses qu'aux États-Unis et la diversité des impacts s'en ressent [Lope15].

Aujourd'hui, les estimations de rendement utilisant comme entrée des prédicteurs agrométéorologiques dans une régression statistique sont une méthode plutôt classique [Aune06, Lobe09]. En général, un simple modèle statistique est construit, reliant par une équation les rendements historiques à des prédicteurs agrométéorologiques. Les publications des dix dernières années qui utilisent les modèles d'impact statistiques se classent en trois grands groupes :
(1) celles qui cherchent à comprendre la variabilité inter-annuelle des rendements ou de la production [Lobe09, Bris10, Neum10, Mako10, Lobe13b, Gras13, Itte13, Ben-14, Iizu14, Ben-16b], en reliant ou non les variations du rendement à la météo [Lobe07b, Tann08, Leng16, Lesk16],
(2) celles qui veulent démontrer que les variations du rendement sont reliées au changement climatique des 60 dernières années [Lobe11c, Iizu16],
(3) et celles, très nombreuses, qui analysent l'impact potentiel du changement climatique futur [Lobe07a, Schl09, Lobe10b, Lobe10a, Lobe13a, Iizu13, Lobe14b, Chal15, Lobe17].

Les études du premier groupe cherchent par exemple à mettre en évidence dans quelles régions du monde les rendements varient le plus/moins, ou encore quelles cultures enregistrent la plus grande stabilité/instabilité. Elles se questionnent aussi sur les facteurs à l'origine de ces variations (politique, pratiques agricoles, météo...), et sur la forme de la tendance moyenne représentant l'évolution du rendement ces 50 dernières années. Enfin, beaucoup se concentrent sur la question du "yield gap", c'est-à-dire de l'écart entre les rendements moyens et ceux potentiels, en quantifiant son importance et analysant les raisons de son existence. Celles qui utilisent spécifiquement les préditeurs météo, analysent soit les différents prédicteurs potentiels (efficacité, effet conjoint...), soit les situations extrêmes (de météo ou de perte de rendement).

Dans le second groupe, on trouve souvent des études qui mettent en parallèle les séries temporelles des prédicteurs météo et celles du rendement pour quantifier la part

12. http://ec.europa.eu/eurostat/web/agriculture/data/database

de la variabilité du rendement que l'on pourrait attribuer au changement climatique (à la différence du premier groupe, où les données météo - lorsqu'elles sont utilisées - ne sont pas vu comme des données climatiques).

Enfin, dans le troisième groupe, les études estiment selon leur modèle, quel serait le pourcentage de pertes (souvent) de rendement pour les prochaines décennies. Ces résultats sont très variables d'une étude à l'autre, mais toutes décrivent une perte non homogène à l'échelle du globe. Certaines analysent les évolutions de pratiques agricoles existantes ou non (comme l'agriculture intensive) sous le changement climatique, et beaucoup relient cela à l'impact sur la production alimentaire mondiale. Pour évaluer l'impact du changement climatique, ou juste l'augmentation de la température sur les rendements, la plupart des études construisent un modèle statistique (validé sur l'historique), et s'en servent pour regarder comment la sortie du modèle est modifiée lorsqu'un paramètre est modifié ou lorsqu'on augmente la valeur d'une entrée. Cela n'est pas sans risques, compte tenu des corrélations existant entre les différentes entrées du modèle. N'en changer qu'une n'est souvent pas réaliste. Enfin, devant les fortes variabilités et incertitudes (modèles, données, résultats des études...) des articles comparent les performances des différents modèles (statistiques et/ou mécanistes), et analysent l'effet sur les prévisions, des erreurs dans les bases de données climatiques.

En général, les résultats des modèles statistiques peuvent difficilement être extrapolés vers des régions et des époques trop différentes de celles qui ont servi dans la base de donnée, à cause de la variation des sols, des paysages, de la météo et surtout des corrélations reliant les différents prédicteurs [Zhan08]. Aussi, l'utilisation de ce type de modèles en tous lieux et temps doit s'effectuer avec précaution. Lorsqu'on simule le rendement avec un modèle statistique, les effets des changements dans la technologie agricole doivent être extrapolés de façon subjective lorsque l'ensemble des technologies n'est pas connu pour la nouvelle période considérée. Les modèles statistiques peuvent fournir beaucoup de connaissances sur les rendements passés et les influences historiques [Gage15, Lobe11c], mais leur utilisation sur le futur ou lors de modifications de données d'entrée doit se faire avec précaution.

3.4 Modèles fonctionnels ou semi-mécanistes : un intermédiaire entre ceux mécanistes et ceux statistiques

Les modèles dits "fonctionnels", sont d'un type intermédiaire : c'est une variante à la modélisation statistique classique à laquelle s'ajoute une expertise agronome [Palm93, Palm97]. Ces modèles utilisent beaucoup de données obtenues par télédétection, ce qui se justifie par le fait que la télédétection est censée pouvoir quantifier le statut des cultures à n'importe quel moment de la saison de croissance, ce qui est donc utile pour décrire la croissance des cultures jour après jour. De plus, des variables agricoles qui sont nécessaires pour faire tourner un modèle mécaniste peuvent être estimées grâce à la télédétection, soit à l'aide d'une relation statistique, soit à l'aide d'une relation physique qui simule les interactions entre les rayons du soleil et les différents composants de la végétation à travers l'utilisation de lois physiques [Dori07]. L'assimilation séquentielle de données satellites dans des modèles mécanistes de type "parcimonieux" est une méthode prometteuse pour un suivi efficace de la végétation [Dewa17, Batt16].

Un exemple semi-mécaniste en Europe : le projet MARS

Le centre commun de recherche européen JRC (Joint Reseach Center) dispose d'un programme de prévision et d'estimation de rendements. Ce programme est bien documenté par [Geno98, Geno06] et par [De W10], sur le site du JRC, ou encore chez [Crai13].

À la base, un modèle mécaniste : Le projet MCYFS (MARS[a] Crop Yield Forecasting System) du JRC vise à prédire des rendements de nombreuses cultures pour les pays européens, mais aussi d'autres régions du monde comme en Russie, Chine, Inde, Amérique du sud, ou Afrique. Le modèle agrométéorologique utilisé est le CGMS (Crop Growth Monitoring System), basé sur le modèle mécaniste WOFOST [Diep89] qui utilise les données du sol, les caractéristiques des cultures, et les conditions météorologiques pour simuler la croissance des cultures. Le système produit des rapports mensuels : (1) indications sur les conditions météorologiques, et (2) indications sur l'état des cultures (dont le stade de développement, l'indice de surface foliaire LAI, l'indice d'humidité des sols, la biomasse totale, et la biomasse des organes de stockages dans des conditions irriguées ou non). Les sorties sont données sous forme de cartes.

Les buts du CGMS :

— Un premier module concerne le traitement de données météo quotidiennes, le remplacement des valeurs manquantes, le calcul des prédicteurs dérivés comme le rayonnement solaire (à partir de la couverture nuageuse ou de la durée d'ensoleillement), l'évapotranspiration potentielle, l'interpolation sur une grille d'unité 50x50km, et la construction sous forme de cartes des conditions météorologiques pendant une période de 10 jours, un mois, ou une saison.

— Un module statistique relie les sorties du modèle par une régression et une fonction de tendance technologique calculée sur les rendements historiques. La régression analytique des années précédentes est seulement utilisée si cela donne des résultats satisfaisants en terme de significativité du coefficient de détermination, du coefficient de corrélation partielle, de la stabilité des coefficients de la régression, et de l'erreur. Sinon, on utilise seulement la fonction de tendance temporelle pour prédire le rendement, ou même le rendement de l'année précédente [Shun04].

Les données utilisées par le CGMS : La simulation des cultures prend en compte trois facteurs : la météo, le sol et les caractéristiques des cultures.

1. Les données météo sont calculées dans le module de suivi météorologique qui fournit chaque jour et pour chaque cellule de 25kmx25km un ensemble d'indicateurs météo. Deux versions de données météo sont utilisées en parallèle : celle basée sur des observations météo mesurées à terre, et celle basée sur des modèles météorologiques et calculée par une agence de prévisions. Sont disponibles pour une période suffisamment longue les données décadaires relatives à quatre caractéristiques : la température moyenne, la température maximale, la température minimale, et les précipitations.

2. Les paramètres de culture : chaque culture a ses propres paramètres qui influencent l'effet de la météo et des caractéristiques du sol. On peut regrouper les paramètres de culture en deux catégories. La première catégorie décrit le comportement de croissance de chaque plante, en regardant l'évolution selon les différents organes. La seconde décrit la variation temporelle et spatiale dans l'utilisation de la culture : par exemple, quelle variété est utilisée à quel endroit et quelles sont les dates moyennes de semis à cet endroit. Pour les cultures qui sont récoltées avant maturité, les dates moyennes de récoltes sont aussi précisées. Les paramètres de la seconde catégorie sont proposés à la même résolution que les données météo.
3. La carte du sol : le sol influence la quantité d'eau disponible pour la croissance de la plante et définie les limites physiques des racines (leur profondeur).

Des données pour la validation du système MCYFS : Les modules de simulation de culture du MCYFS ont étudié l'influence de la météo, en supposant l'influence de tous les autres facteurs constante. De plus, le MCYFS suppose les effets de la météo sur les rendements indépendants de la quantité de fertilisation [Supi99]. Cependant, [Cegl13] a noté que les résultats ne peuvent pas être concluants quand le rendement est co-déterminé par des facteurs non pris en compte dans l'analyse. Aussi, en fin de rapport du CGMS, on trouve d'autres sources comme des informations issues de télédétection. Ces images montrent un effet plus global de la météo, de l'humidité, et de la gestion des cultures. Le MCYFS suppose qu'au niveau régional, l'influence des facteurs individuels (non détectables avec la télédétection) se compense [Diep95].

Des rapports synthétisant l'évolution de la météo, des cultures, et les explications concernant leur évolutions (analyse agro-météorologique) sont disponibles en ligne[b] chaque mois, pour l'Europe (et ses régions voisines en Russie). Les premières estimations de rendements apparaissent dans le rapport de mars. Les rapports d'avril inclus également des informations pour semer puis jusqu'en octobre, des cartes de données obtenues par télédétection (fAPAR essentiellement).

a. Monitoring Agriculture through Remote Sensing
b. https://ec.europa.eu/jrc/en/mars/bulletins

4 Quel modèle pour quelle application ?

Comme cela a été décrit au Chapitre I page 6, les modèles d'impacts peuvent être utilisés pour trois applications : pour les prévisions saisonnières, pour les estimations en fin d'année (monitoring), ou pour les prévisions climatiques.

4.1 Modèles mécanistes versus statistiques

Les modèles mécanistes ne sont pas meilleurs que les modèles statistiques. Leurs utilisations et échelles spatiale/temporelle d'existence sont différentes. La plupart des modèles mécanistes s'utilisent à petite échelle (champs ou exploitations) et très peu à

Modèles mécanistes :	**Modèles statistiques :**
Un modèle qui a une structure représentant explicitement une compréhension des propriétés biologiques, chimiques et/ou physiques des procédés	Un modèle où la structure est déterminée par la relation observée entre les données expérimentales
Principes fondamentaux de la physique	Les relations et les tendances ne sont pas nécessairement physiquement pertinentes : interprétation parfois difficile
Cherchent à répondre à la question "comment"	Répond à la question "qui", et "quels effets"
Performant à l'échelle du champs, ou de la zone de calibration	Performants à l'échelle régionale ou nationale si bien paramétrés sur ces échelles
Met l'accent sur les causalités dans la succession des phénomènes	Donnent peu d'information sur les phénomènes physiques en jeu
Difficile et long à développer	Facile à implémenter, peu d'hypothèses, simplicité numérique
Si tous les paramètres sont issus d'experts on n'a pas besoin de beaucoup de données réelles	Ils peuvent incorporer des hypothèses mécaniques avec des paramètres estimés à partir des données et non des experts.
On a quand même besoin de données pour calibrer les paramètres non issus d'experts, et ces données ne sont pas toujours en nombre suffisant	Utile quand les phénomènes sous-jacents sont non connus ou difficiles à expliquer par des formules physiques, ou que l'on ne dispose pas de suffisamment de connaissances sur le sujet (car non expert).
Calibration sur des zones petites (échelle du champs ou de l'exploitation), rendant la généralisation hors de cette zone de calibration dangereuse	Ces modèles peuvent être utilisés pour développer des relations pour la prévision et la description des tendances, en restant vigilants.
Le nombre de paramètres augmente avec la complexité des phénomènes sous-jacents	
Si le modèle mécanique échoue à l'extrapolation, on peut s'aider de l'architecture physique construite (et des hypothèses physiques faites) pour essayer de savoir d'où vient le problème.	Critiqués pour leur manque de robustesse à l'extrapolation (application à des variables d'entrées hors des intervalles de l'apprentissage)
	Possibilité de faire des hypothèses statistiques, des tests, ou encore donner des intervalles de confiances (importance de la variabilité)
Les équations physiques doivent se comporter comme la nature l'a fixé.	Les données reflètent la façon dont la nature fonctionne, avec leur variabilité

grande échelle (mais il peut y avoir des exceptions, comme le modèle Orchidée, Section 2.4). Certains modèles mécanistes peuvent également utiliser des relations statistiques : il suffit d'avoir suffisamment de données et de contraintes pour pouvoir calibrer le modèle.
On remarque une certaine concurrence entre les deux écoles de pensée : les utilisateurs des modèles statistiques disent que, étant davantage empiriques, les modèles basés sur les données ont une meilleure valeur pratique en raison de la complexité infinie des phénomènes sous-jacents. Les utilisateurs des modèles mécanistes mettent l'accent sur un meilleur potentiel cognitif et prédictif des modèles basés sur la mécanique, qui sont capables de générer de nouvelles connaissances. En fait, il existe très peu de modèles purement statistiques ou mécaniques. Tous les modèles statistiques contiennent un élément mécaniste, même si cet élément est la liste des variables d'entrées, influençant la sortie du modèle [Nest99].

En modélisation mécaniste, le rendement des cultures est déterminé par des équations représentant les réponses physiologiques d'une culture aux variables environnementales [Jone03b, Keat03]. Cependant, ces modèles nécessitent un degré élevé de détail environnemental, de sorte que les données d'entrée peuvent être difficiles à acquérir dans des régions étendues et diverses du point de vue environnemental. Par exemple, les modèles mécanistes sont rares en écologie, car les informations physiologiques nécessaires sont rarement disponibles. En outre, les communautés naturelles sont influencées par des interactions inter-espèces complexes, la dynamique des populations et les voies de dispersion, ce qui rend les modèles mécanistes encore plus difficiles à développer [Este13]. Dans les modèles statistiques, les corrélations entre les distributions d'espèces et les prédicteurs climatiques sont utilisées pour évaluer l'adéquation de l'habitat. Bien que les variables prédictives soient souvent sélectionnées en fonction de la connaissance de la physiologie d'une espèce (donc issues de connaissances d'experts), elles peuvent également être choisies en fonction d'un meilleur ajustement empirique sans lien spécifique avec la physiologie.

En résumé

Le choix entre l'utilisation d'un modèle statistique ou d'un modèle mécaniste dépend donc de l'application visée.
Par exemple, si l'on s'intéresse à l'impact du climat futur, on ne pourra pas utiliser de données satellites (qui ne sont pas estimables) ou des informations issues de sondages sur le terrain, comme entrée de notre modèle. Si l'on cherche à faire des prévisions saisonnières en régional, un sondage peut être le plus efficace. Enfin, si l'on cherche à estimer le rendement en fin d'année (monitoring) sur une échelle spatiale trop grande pour utiliser des sondages, les données satellites peuvent être un bon compromis.

Le problème de sélection de modèle et des divers critères de qualités utilisés sera traité dans le Chapitre IV.

4.2 Échelles spatiales : du local au global, en passant par le régional

La modélisation des cultures à grande échelle provient de la nécessité de simuler les impacts de la météo sur les cultures par un procédé utilisant directement les sorties de modèles météorologiques. Le fondement de telles techniques combine les avantages des approches statistiques (faible quantité de variables d'entrées, une validation à grande échelle se passant ainsi des spécificités locales) avec ceux des modèles basés sur des processus à l'échelle du champs (validation sous différentes conditions environnementales) [Chal04]. L'estimation à grande échelle suppose une complexité appropriée : la communauté qui modélise les cultures est maintenant bien consciente des dangers associés aux modélisations dont le niveau de complexité n'est pas justifié face au degré d'incertitude et aux erreurs potentielles associées à la paramétrisation utilisée [Mont96]. Plus le nombre de processus simulés est grand, plus le nombre d'interactions potentielle entre eux, et le nombre de paramètres nécessaires pour la calibration sont grands, augmentant ainsi l'erreur potentielle. L'échelle spatiale et la complexité du modèle sont très liées, comme le soulignent [Chal08] ou [Tubi02].
De plus, pour améliorer la qualité des estimations, les paramètres doivent, si possible, être estimés à partir des observations, et le modèle testé sur de nouvelles données. Pour les études évaluant les effets du changement climatique, il faut faire attention à ne pas utiliser des paramètres observés pour le climat présent dans des situations où leurs valeurs pourraient avoir changé.

Les principes ci-dessus montrent comment les modèles d'impact à grande échelle se différencient de ceux à l'échelle du champ. Les modèles à grande échelle tentent d'être moins complexes, d'avoir moins de paramètres, et moins de paramètres - voire aucun - non estimés à partir des observations. Par exemple, le modèle de [Chal04] simule la croissance de la surface foliaire avec un paramètre spécifiant le taux maximal de variation de l'indice de surface foliaire, au lieu de simuler l'apparence de chaque feuille. Comme beaucoup d'autres modèles, les modèles d'impact à grande échelle doivent être utilisés de façon pertinente et raisonnée. Ils se concentrent sur l'influence de la météo, et leur fondement sur les relations observées, signifie qu'ils ne simulent pas les facteurs non-météorologiques influençant les rendements. Là où la variabilité des rendements est davantage influencée par ces facteurs non météorologiques, et aux endroits où la météo est d'une variabilité exceptionnelle [Baro05], les modèles d'impact météorologique à grande échelle ne se justifient plus. Cependant, il faut noter que le signal météorologique semble être prépondérant devant les autres facteurs à l'échelle régionale [Chal03, Bakk05].

4.3 Un exemple à grande échelle : le programme de l'USDA aux États-Unis

Le ministère de l'agriculture des États-Unis (USDA) comporte un service de statistique (NASS - National Agricultural Statistics Service) qui est chargé de fournir, des prévisions mensuelles et des estimations annuelles de rendements agricoles pour les cultures les plus répandues, sur l'ensemble des États-Unis. Les prévisions de NASS en maïs et soja sont réalisées en août, septembre, octobre, et novembre. L'estimation finale du rendement est donné en janvier après la fin des récoltes, se basant sur un sondage agricole global auprès des producteurs en décembre. Ce programme est abondamment documenté par [Abre08] et plus récemment par [Aune12].

Des avis d'experts et des observations in situ : Pour obtenir les données nécessaires aux prévisions mensuelles, NASS a recourt à deux grands sondages :
(1) Un sondage de rendement mensuel AYS (Agricultural Yield Survey) des producteurs est réalisé dans 32 états pour le maïs et dans 29 états pour le soja, avec un total d'environ 25000 producteurs interrogés (toutes cultures confondues) en août. On demande aux cultivateurs d'identifier la surface de maïs et de soja plantée, et de fournir une prévision du rendement final pour chaque culture. En fonction de ces réponses, des prévisions de rendements moyens sont réalisées pour chaque état interrogé et pour les États-Unis.
(2) Un sondage OYS (Objective Yield Survey) est réalisé ensuite dans 10 états pour le maïs et 11 états pour le soja. Le but du sondage OYS est de générer des prévisions de rendements en se basant sur des comptages et des mesures végétales.

Un échantillonnage sous contrôle : Les sondages sont généralement réalisés durant les 2 semaines qui précèdent d'une semaine la date de parution des prévisions. Pour le sondage AYS, un échantillon d'exploitations à interroger est déterminé parmi ceux qui ont répondu au sondage de juin concernant les surfaces cultivées. Même si les exploitations à interroger changent d'une année à l'autre, pour une année fixée, les mêmes exploitations sont interrogées chaque mois, d'août à novembre.
Les champs sélectionnés pour le sondage OYS (1920 pour le maïs et 1835 pour le soja) font aussi parti des exploitations ayant signalé la plantation (faite ou à venir) de maïs ou de soja dans le sondage de juin concernant les surfaces. On tire ensuite un échantillon aléatoire de champs avec une probabilité de sélection proportionnelle à la taille de la parcelle. Deux parcelles sont alors sélectionnées dans chaque champ. Les données collectées pour chaque parcelle de maïs pendant le cycle de prévision sont utilisées pour mesurer ou prédire le nombre d'épis et le poids des grains. Ces données contiennent la largeur du rang, le nombre de tiges par rang, le nombre de tiges avec épis par rang, le nombre d'épis en graine, la longueur des grains, le diamètre et le poids des épis, le poids des grains écalés, la quantité d'humidité, le poids total d'épis dans l'unité récoltée, le poids des épis échantillonnés en laboratoire, le poids des grains dans les épis échantillonnés, et la quantité d'humidité dans les grains des épis parvenus à maturité. Le rendement de maïs est prédit avec la prévision (ou la mesure, si la maturité le permet) du nombre d'épis, du poids des épis, et des pertes de récolte.

Le modèle : On utilise les données de plusieurs années pour construire des relations statistiques qui font le lien entre les mesures/comptages avant la récolte, et les mesures de rendement final après la récolte. [Voge99] ont publié un document détaillé sur les méthodes statistiques d'estimation de NASS [13].

Pour le maïs et le soja, la prévision du rendement moyen à l'échelle de l'état est calculé avec une simple moyenne des rendements pour chacun des champs sélectionnées dans le sondage OYS. De plus, une prévision à l'échelle de l'état est aussi faite en moyennant en premier les prévisions ou facteurs du rendement actuel (comme le comptage de tiges, le comptage d'épis, le poids des épis), puis en estimant le rendement moyen de l'état directement à partir de ces moyennes. Cette prévision se base sur une régression analytique des relations historiques (15 ans) entre les différents facteurs influençant le rendement, et le rendement moyen de l'état. Les rendements moyens des états sont associés pour donner la prévision du rendement à l'échelle des États-Unis.

13. http://www.usda.gov/nass/

Pour le détail des formules du modèle voir [Aune06].

Évaluations et performances du modèle : Les méthodes statistiques de sondage agricole sont très couteuses et prennent beaucoup de temps. Les informations agricoles doivent posséder plusieurs qualité : l'objectivité, la fiabilité, et être d'une couverture suffisante. Dans bien des pays producteurs, les statistiques agricoles ne rentrent pas dans ces critères [Shun04]. La section PECAD (Production Estimates and Crop Assessment Division) de l'USDA est chargée de fournir des informations sur l'état des cultures, et sur les rendements pour plusieurs cultures importantes à travers le monde [Reyn01]. Son principe repose sur un système de gestion de données appelé CCDRE (Crop Condition Data Retrieval and Evaluation). PECAD s'appuie sur plusieurs sources de données pour examiner les anomalies météorologiques qui affectent la production de cultures et la qualité des produits agricoles. Les deux principales sources de données météorologiques sont le WMO (World Meteorological Organization) avec plus de 6000 stations ; et les données météo de l'armée de l'air américaine AFMA (U.S. Air Force Weather Agency) qui combine les stations terrestres du WMO avec des images satellites. CCDRE contient plusieurs modèles mécanistes et plusieurs algorithmes de réduction de données qui sont exécutés sur les données d'entrées météorologiques.

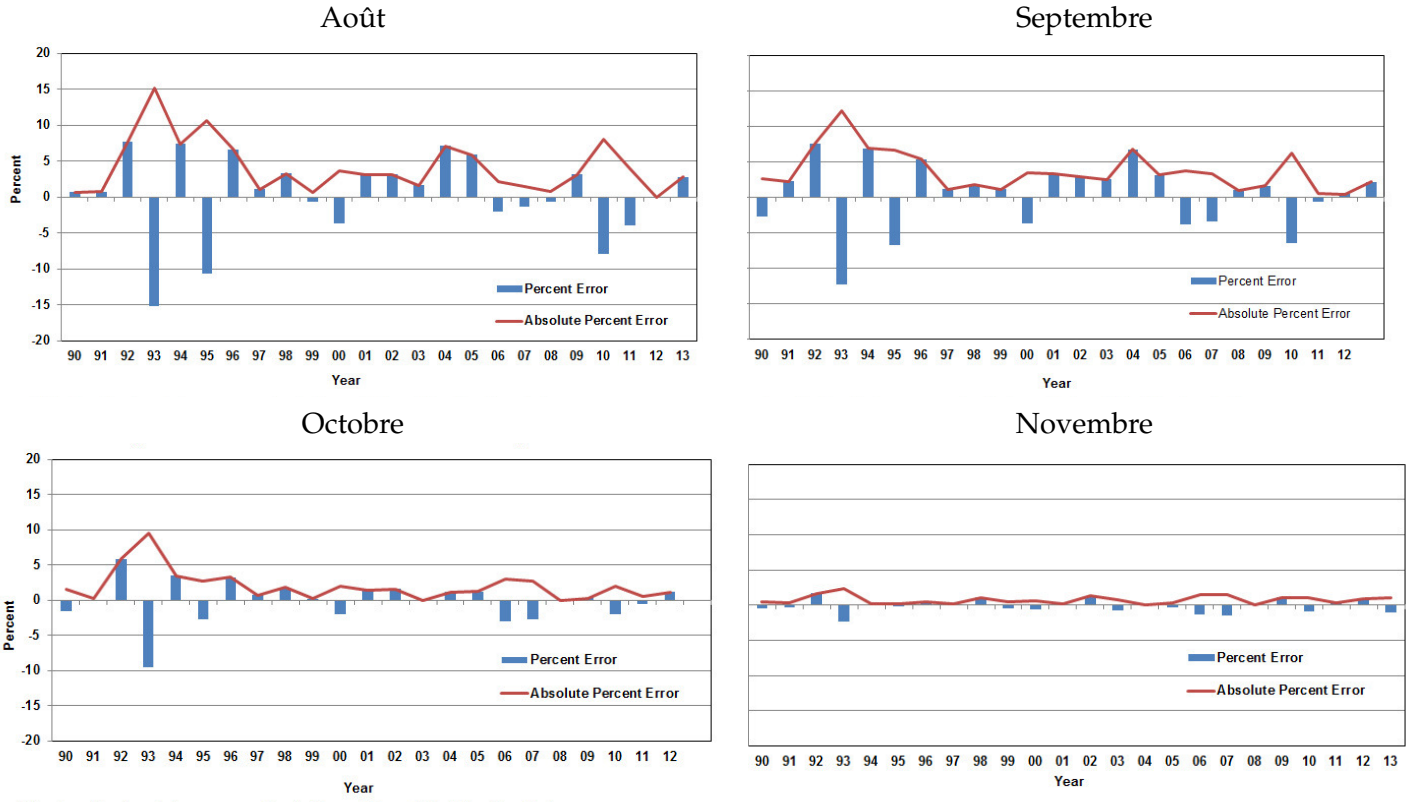


FIGURE II.4 – Erreurs commises par le modèle de l'USDA dans ses prévisions du mois d'août au mois de novembre de 1990 à 2013 (tiré de [Good14]).

La plus grande force du programme de NASS repose probablement sur son équipe de recenseurs entrainée (des milliers répartis dans tous les États-Unis) et très habituée

aux pratiques agricoles locales. Puisque la taille des échantillons et la superficie sont relativement faibles, la qualité des données est essentielle. NASS a fait des recherches sur beaucoup d'alternatives à leur modèle de rendement comme le bilan hydrique ag-met, ou les chaines de Markov, mais préfère encore l'approche par sondages, essentiellement basée sur des mesures et des comptages de plantes et de fruits.

NASS propose chaque année des rapports pour quantifier la qualité de ses prévisions [Irwi14]. Pour évaluer la précision historique des prévisions de rendement de l'USDA, les prévisions d'août, septembre, octobre et novembre sont comparées avec l'estimation du rendement final de janvier. Les différences entre les prévisions et les estimations finales (en pourcentage) de 1990 à 2013 sont présentées à la Figure II.4. Une erreur positive traduit une sous-estimation des rendements de la part de NASS, et une erreur négative traduit une sur-estimation. Les erreurs sont grandes pour certaines années, comme en 1993 ou en 1995. Ces erreurs ne sont pas surprenantes vu les évènements météorologiques inhabituels qui se sont produits ces années là. Cependant, en 2012, malgré la forte sécheresse et les difficultés à réaliser des mesures dans ces conditions, l'erreur de l'USDA est quasi nulle. Les prévisions de l'USDA se sont améliorées au cours du temps, au sens où les prévisions sont de plus en plus précises.

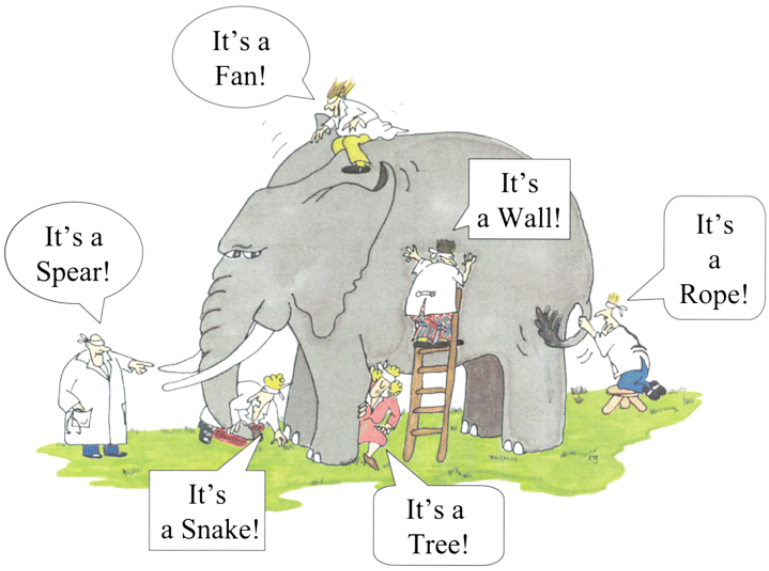
5 Conclusion

Comme on vient de le voir, la diversité des modèles d'impacts en agriculture est impressionnante et le choix du modèle dépend essentiellement des applications que l'on vise. Dans cette thèse, on se concentrera sur les modèles d'impact statistiques, et non mécanistes.
Les trois applications qui seront testées par la suite couvriront les prévisions saisonnières, les estimations en fin d'année (monitoring), et les prévisions climatiques.
Aussi, les modèles utiliseront des données indirectes simples (et non in situ, de sondages, ou satellites) qui auront les avantages d'être simulables (pour les prévisions climatiques), accessibles à l'échelle mondiale (pour pouvoir travailler à grande échelle), et peu couteuses. Cela sera donc des données météorologiques usuelles.
De plus, il nous semble important de présenter une méthodologie généralisable à d'autres cultures et à d'autres régions du monde.

CHAPITRE III

Analyse des données agricoles et météo aux États-Unis, ainsi que leurs interactions

Table des matières



1 La géographie et le climat des États-Unis

Les États-Unis ont une superficie de 9,6 millions de km^2 et se placent au quatrième rang mondial pour la superficie. Depuis l'indépendance à la fin du XVIIIe siècle, le pays est une république fédérale composée de 50 états qui disposent de pouvoirs importants (Figure III.1). Le relief des États-Unis a une organisation relativement simple qui se retrouve dans tout le continent américain.

FIGURE III.1 – Les 50 états des États-Unis.

Le fleuve Mississipi-Missouri est parmi les plus longs du monde en parcourant plus de 6 200 km. La situation des montagnes détermine en grande partie la répartition des climats (Figure III.2) : le climat de haute montagne est présent dans les Montagnes Rocheuses (qui possèdent plusieurs sommets à plus de 4 000 mètres d'altitude) ainsi qu'en Alaska. L'organisation méridienne du relief permet aux masses d'air polaire et tropical d'envahir le territoire américain, provoquant des phénomènes climatiques extrêmes : gel hivernal en Floride, canicules estivales dans le nord-est, tornades dans le centre (Figure III.2, haut). Les grands fleuves américains naissent des montagnes et permettent la production d'hydro-électricité et l'irrigation. Il est important d'avoir en tête ces informations pour comprendre, comparer et vérifier la répartition des variables météorologiques ou agroclimatiques qui seront utilisées dans les modèles par la suite.

Les 48 états continentaux sont compris dans la zone climatique tempérée. Dans la moitié est des États-Unis, les climats ont une répartition relativement simple, dictée par la latitude (Figure III.2, bas). Seule la chaîne des Appalaches, d'orientation Nord-Est/Sud-Ouest et qui s'étend sur plusieurs centaines de kilomètres, vient contrarier cette répartition zonale. Ces régions se caractérisent par un climat tempéré humide de façade orientale. Elles sont soumises à la circulation méridienne des masses d'air : des vagues de froid en hiver peuvent descendre jusqu'en Floride, et des vagues chaudes et humides remontent jusqu'à l'extrême nord du pays en provoquant des canicules puis les fameux "étés indiens". Le climat de la moitié orientale est humide : les précipitations sont supérieures à 700 mm par an. Cependant, on observe de grands écarts dans cet ensemble de plusieurs millions de km^2, entre la côte et l'intérieur, mais aussi entre le

nord et le sud. Les différences s'expliquent par la latitude, le degré de continentalité et la nature des courants marins.

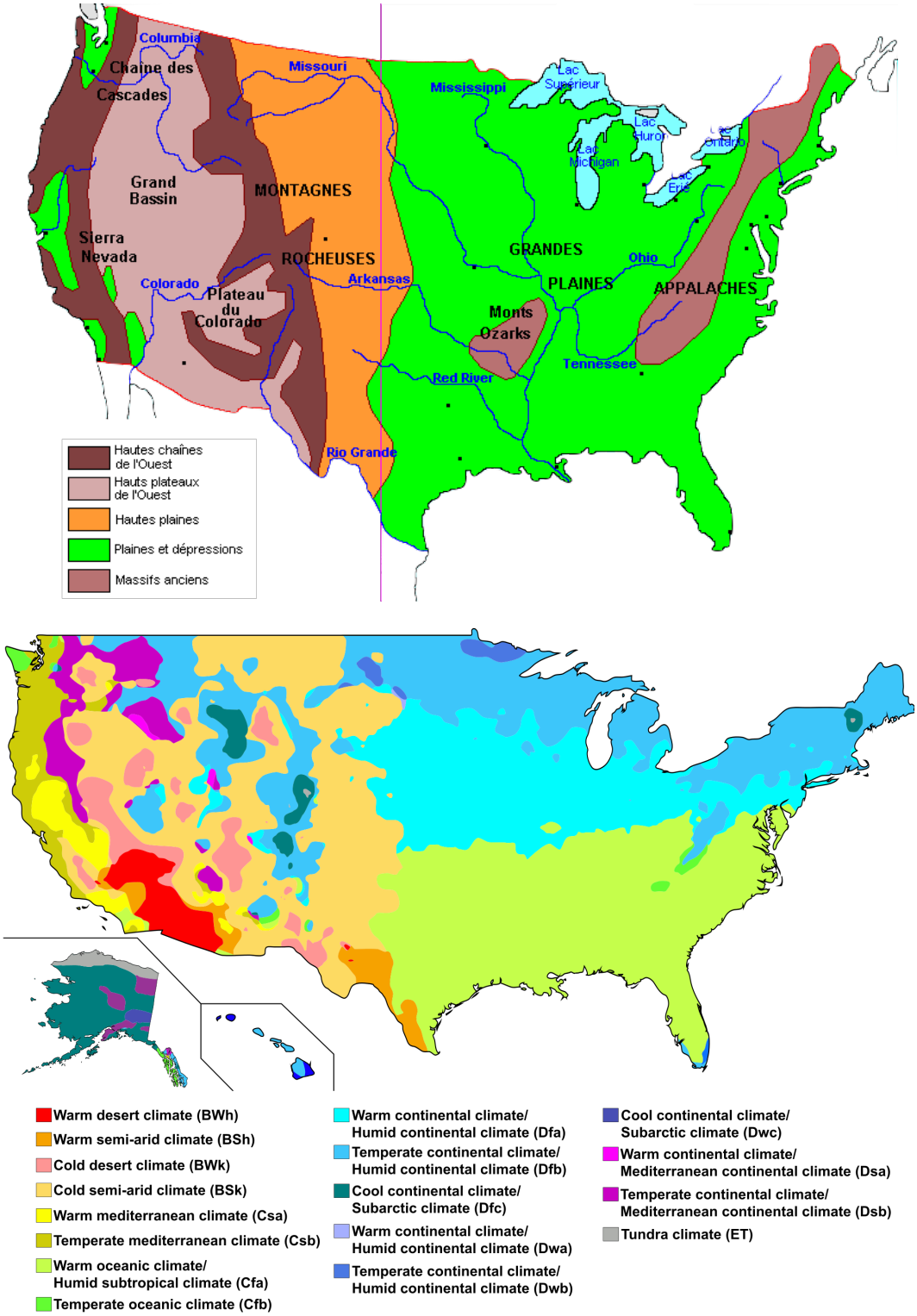


FIGURE III.2 – Haut : Carte des reliefs aux États-Unis [Houo11].
Bas : Les différents types de climats aux États-Unis : classification de Koppen [Peel07].

Sur les côtes du golfe du Mexique, le total annuel des précipitations est plus important que sur la côte atlantique à la même latitude (La Nouvelle-Orléans : 1 571 mm ; Jacksonville : 1 303 mm). Lorsque l'on remonte à l'intérieur des terres, dans le bassin du Mississippi, les précipitations annuelles diminuent et les températures se rafraîchissent, avec des gelées en hiver. La zone subtropicale est limitée approximativement par la rivière Ohio (38e parallèle nord) au niveau de Louisville (Kentucky) et de Cincinnati où la neige est courante en hiver. Plus au nord, on trouve un climat de type continental humide avec des hivers froids. Enfin, les montagnes Appalaches, qui culminent à plus de 2 000 mètres, perturbent le climat subtropical en Géorgie et en Caroline du Nord : elles augmentent les précipitations et refroidissent les températures.

Aux États-Unis, les 50 états fédéraux sont divisés en 300 districts agricoles environ, eux même divisés en environ 3000 cantons (Figure III.3). Le code FIPS (Federal Information Processing Standards) est un nombre à 5 chiffres qui désignent de façon unique les cantons des États-Unis. De gauche à droite, les deux premiers chiffres font référence au code de l'état dans lequel se trouve le canton, et les trois derniers, font référence au canton lui même. Généralement, les codes sont dans le même ordre que l'ordre alphabétique des cantons. Ce système de code est utilisé par l'EAS (Emergency Alert System), le NWR (NOAA Weather Radio) et l'USDA (US Department of Agriculture). On peut télécharger la liste de ces codes sur `www.schooldata.com/pdfs/US_FIPS_Codes.xls`. Si l'on veut regrouper les cantons par district, il peut être utile d'avoir le FIPS code à 7 chiffres, contenant le numéro du district entre celui de l'état et celui du canton. On peut télécharger ces codes plus complets sur `https://www.nass.usda.gov/Data_and_Statistics/County_Data_Files/Frequently_Asked_Questions/county_list.txt`. C'est cette liste qui est utilisée par l'USDA, dans leur base de données agricoles. Par exemple, le canton Pulaski dans l'état de l'Indiana (dont le code de l'état est le numéro 18) a pour code FIPS 18131 et pour code FIPS complet 1810131, puisqu'il se situe dans le district numéro 10.

D'après le recensement américain en 2000, la surface moyenne d'un canton était de 1610km^2, soit environ la même surface qu'un carré de 40km$\times$40km. Les cantons situés sur la moité ouest du pays sont beaucoup plus grands que ceux situés à l'est. Par exemple, la surface moyenne d'un canton de l'état de Géorgie est de 890km^2 (soit environ 30km$\times$30km), tandis que celle pour un canton de l'Utah est de 6290km^2 (80km$\times$80km).

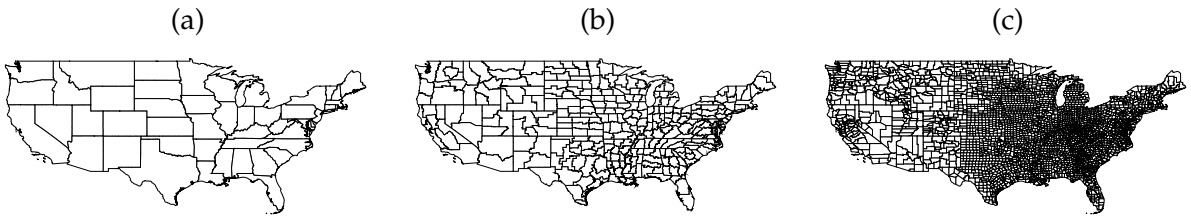


FIGURE III.3 – Découpage administratif des États-Unis, en états, districts et cantons : une hiérarchique spatiale qui sera utilisée dans les modèles d'impact.

2 Les données agricoles

Le maïs est une des cultures les plus importantes aux États-Unis, et cette thèse se concentre donc sur cette céréale car une base de données observées fournie est disponible. Les États-Unis produisent 359 millions de tonnes de maïs par an (année 2013-2014) soit environ 40% du maïs produits dans le monde. Les Américains eux-mêmes sont de très gros consommateurs, puisque 80% de leur propre production est absorbée par le marché intérieur. 86% des semences de maïs américain sont Roundup Ready en 2013, c'est-à-dire prêtes pour un traitement au Roundup (ou tout autre herbicide constitué de glyphosate).

Avec 20% de leur production à l'export, les États-Unis représentent 60% du marché mondial du maïs. Le maïs occupe la première place des productions de grains des États-Unis. Avec plus de 37 Mha semés en 2007, il est largement devant le soja (25,7 Mha) et le blé (24,5 Mha). Le maïs est principalement cultivé dans 5 grands bassins de production. La Corn Belt (principalement les états Iowa, est du Nebraska, Minnesota, Illinois, et Indiana) représente à elle seule plus de 60% de la production nationale. Vient ensuite la région des grandes plaines, avec un quart de la production ; elle est située à l'est de la barre montagneuse des Rocheuses, avec un climat continental sec. Ces différentes zones produisent du maïs dans des conditions très variables (Figure III.4) :
(1) La Corn Belt, au climat continental humide, est située dans le centre des États-Unis, sur des sols riches et profonds. La zone est particulièrement bien adaptée au maïs, qui y est majoritairement cultivé, en rotation avec le soja. Les précipitations étant abondantes (la quantité de pluie pendant les cinq mois de la saison de croissance est relativement constante, entre 43 et 48 cm [Neil90], l'irrigation n'est pas nécessaire, et les rendements en sec y sont les plus élevés du monde : 12 t/ha en 2016. Ainsi, seulement 11% de la surface en maïs sont irrigués, contre 50% en France.
(2) Dans les grandes plaines du Nord (Dakota du Nord et du Sud), le maïs, non irrigué et en rotation avec le soja, est en concurrence avec le blé de printemps HRS (Hard Red Spring). Le rendement y était de 9.3 t/ha en 2016.
(3) Les grandes plaines du centre (Nebraska, Kansas, Colorado) sont la deuxième région productrice, avec une grande partie du maïs irrigué à partir d'une nappe phréatique de 450 000 km^2 (la nappe Ogallala) qui s'étend sur 8 états. Le maïs est en concurrence avec le blé HRW (Hard Red Winter) et le soja, eux aussi irrigués. Le niveau de la nappe est néanmoins en baisse (-77 cm en moyenne entre 1980 et 1995), ce qui pourrait à terme poser des problèmes d'irrigation. En 2016, cette zone produisait 17% du maïs aux États-Unis et avait un rendement de 9.8 t/ha.
(4) La culture du maïs s'étend aussi le long de la vallée du Mississippi, où toutes les cultures sont irriguées. On y trouve du blé SRW (Soft Red Winter), du soja, du sorgho, et du coton [Leve08]. En 2016, cette région a fourni 3% du maïs des États-Unis avec un rendement moyen de 9.6 t/ha.
(5) Enfin, la côte Est produit environ 3.4% de maïs avec un rendement moyen de 6.2 t/ha [1].
L'irrigation est limitée.

Les statistiques de rendement de maïs (en boisseau par acre, ou tonne par hectare) ont été collecté par le ministère de l'Agriculture (USDA) des États-Unis, et plus précisément par son département de statistiques NASS. La plateforme en ligne Quickstat permet de télécharger facilement des données agricoles (production, surface, rendement) : https://quickstats.nass.usda.gov/. Un long historique est disponible

1. Les chiffres proviennent du ministère de l'agriculture américain, et du service national de la statistique agricole.

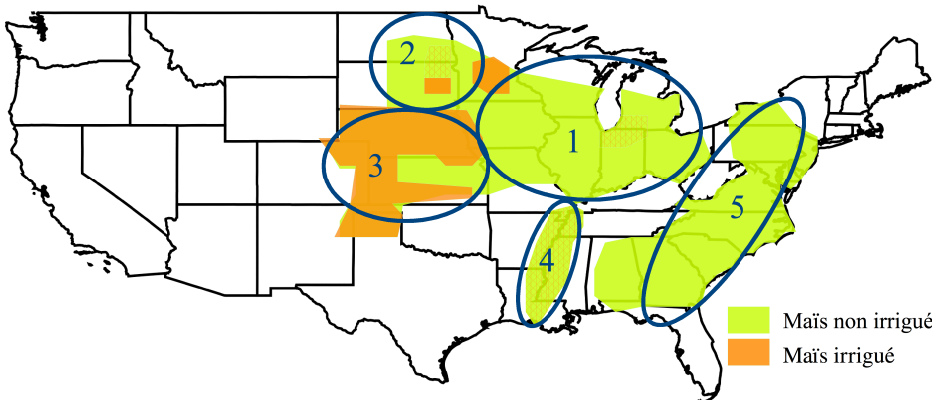


FIGURE III.4 – Principaux bassins de production de maïs aux États-Unis [Leve08]. Données issues de l'USDA, recensement agricole de 2002.

de 1910 à nos jours, mais les séries temporelles de rendement sont incomplètes pour beaucoup de cantons les 50 premières années. La Figure III.5 montre l'évolution du nombre de cantons disponibles de 1910 à 2013. Depuis 1910, un nombre croissant de cantons a été enregistré, avec un pic aux alentours de 1980 (avec plus de 2500 cantons renseignés). Depuis, le nombre de cantons renseignés décroit. Les données météorologiques qui seront présentées dans la section suivante ne sont disponibles qu'à partir de 1979. L'étude météo/rendement se concentrera donc sur les 35 années 1979-2013. Les rendements des années précédentes serviront tout de même pour la détermination de la tendance temporelle des rendements. La limitation aux 35 dernières années n'est pas une limitation gênante, vu les changements importants au cours du temps qui se sont produits en termes de pratiques agricoles. En effet, les données du début de XXème siècle sont trop éloignées de celles de la deuxième moitié du siècle pour qu'on puisse les rassembler et les traiter de façon similaire.

En fonction des années (de 1910 à 2013) les mêmes cantons ne sont pas renseignés : la Figure III.5 indique le nombre de cantons par an pour lesquels on dispose des données agricoles : seules les données de l'état du Nebraska sont renseignées de 1910 à 2013. Plus de 1000 cantons ont une série temporelle en rendement de plus de 70 années ; non consécutives parfois (Figure III.5, droite). De 1979 à 2013, c'est la moitié est des États-Unis qui est la mieux renseignée (Figure III.5, bas).

La Figure III.6 représente les séries temporelles de rendement de maïs pour quatre états géographiquement éloignés (Alabama, Texas, Minnesota et Iowa). Chaque ligne fait référence à la série temporelle d'un canton, et pour chaque état, les cantons d'un même district sont tracés avec la même couleur. Les cantons d'un même district ont des rendements aux variations très proches. La longueur des séries temporelles peut être courte - comme pour le Texas - ou longue - comme pour le Minnesota. Pour certains grands états, comme le Texas, les rendements des districts ne sont pas du même ordre de grandeur, tandis que pour d'autres - comme pour l'Iowa - les districts ont des rendements aux amplitudes similaires. Les variations des rendements d'une année sur l'autre peuvent être importantes, comme on peut le voir pour l'Iowa de 1975 à 1995. De 1979 à 2013, le rendements de maïs aux États-Unis est passé de 86 à 158 boisseaux par

acre [2] grâce aux améliorations génétiques et aux pratiques agricoles (pesticides, fertilisants).

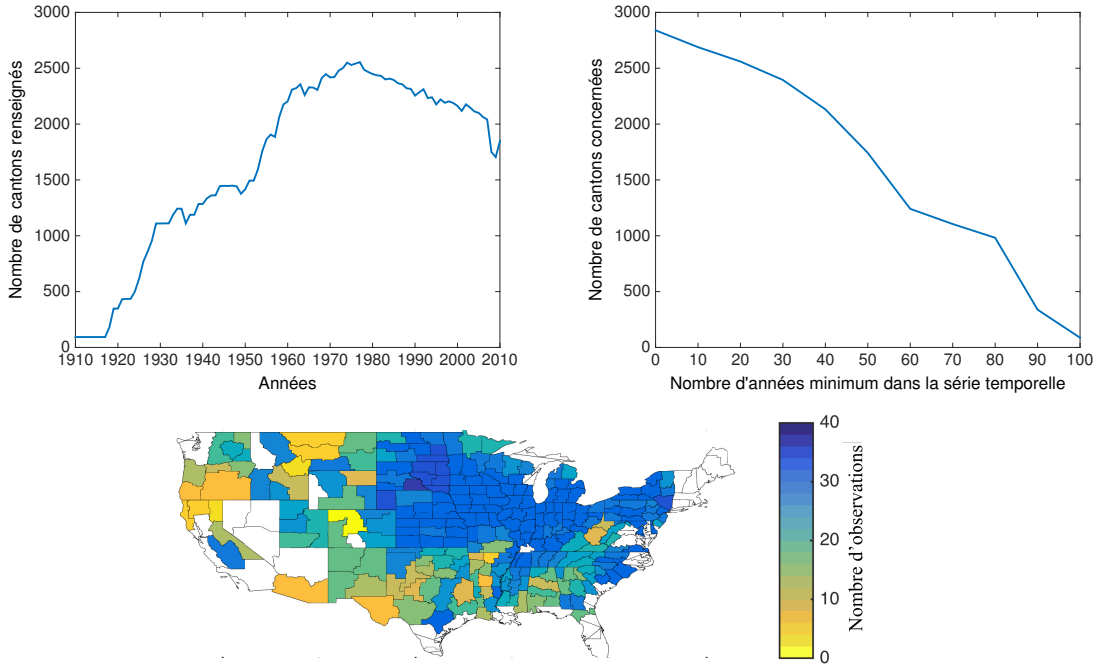


FIGURE III.5 – Gauche : Nombre de cantons pour lesquels on dispose des données de rendements en maïs en fonction des années.
Droite : Nombre de cantons pour lesquels la série temporelle des rendements est de longueur supérieure à un certain nombre d'années (de 1910 à 2013).
Bas : Nombre de données disponibles en moyenne par canton au sein de chaque district (de 1979 à 2013).

Les plus grandes séries temporelles commencent en 1914 et finissent en 2013. On dispose des données sur plus de 30 ans pour 2395 cantons, sur les 2849 répertoriés comme produisant du maïs. Pour moins du tiers de ces données, on sait s'il s'agit d'une culture irriguée ou non-irriguée. On remarque que même à l'échelle d'un état il existe des différences parfois très importantes entre les séries temporelles des rendements des différents districts. Cependant dans la plupart des cas, une évolution brutale du rendement d'un district pour une année est souvent aussi observée pour les autres districts (Figure III.6).

La production de maïs des États-Unis était de 2 milliards de boisseaux par an dans les années 30, lorsque les premiers maïs hybrides ont été commercialisés, et le rendement était alors de 24.2 boisseaux/acre (1518 kg/ha). La production est passée de 3 milliards de boisseaux par an dans les années 50, à 6 milliards de boisseaux par an dans les années 70, puis à plus de 9 milliards de boisseaux par an au début du XXIème siècle (Figure III.7). En 2005, le rendement des États-Unis atteint les 160.4 boisseaux/acre (10 059 kg/ha). Les augmentations de la pente du rendement au cours du temps sont attribuées à une meilleure faculté d'adaptation, au maïs hybride (fécondation croisée de l'ovule d'un maïs par du pollen d'une autre plante de la même espèce), à la mécanisation de la récolte, à une meilleure fertilité du sol (particulièrement dû à l'ajout d'azote), aux

2. Un boisseau/acre représente 62.7117 kg/ha (0.0627117 t/ha), ou un kg/ha représente 0.015946 bois/acre ([Tayl10], page 21)

variétés hybrides simples, à l'amélioration des cultures (en particulier l'augmentation de la densité des plants), et aux biotechnologies [Troy06].

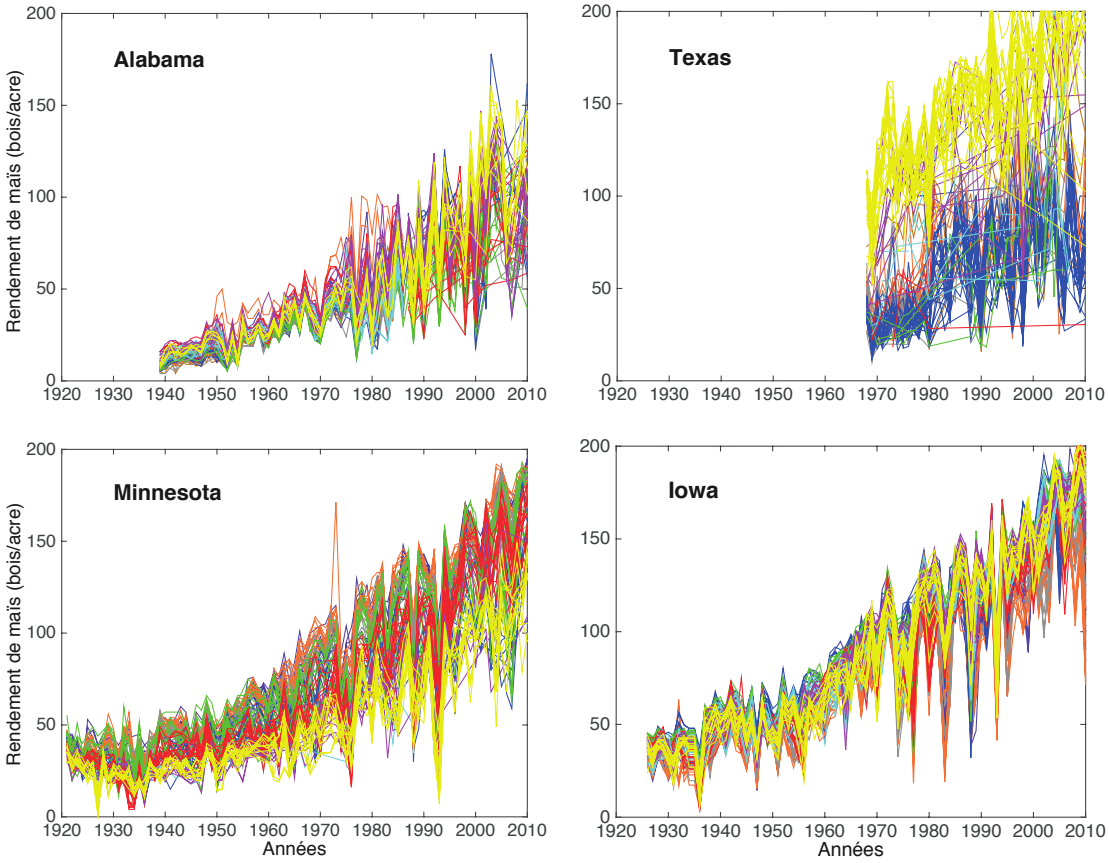


FIGURE III.6 – Série temporelle de production de maïs pour les cantons de quatre états spatialement éloignés : l'Alabama, le Texas, le Minnesota et l'Iowa. Les cantons d'un même district sont tracés avec une même couleur. Ces quatre états ont été choisis en fonction de la localisation, de l'importance des valeurs de rendement, et de la longueur des séries temporelles disponibles.

La répartition spatiale des rendements en 1975, 1990 et 2013 à l'échelle des cantons est représentée à la Figure III.8. Les cantons en blancs sont ceux pour lesquels l'année en question n'a pas été renseignée. L'est des États-Unis a un rendement de maïs bien plus important qu'à l'ouest, à cause du climat et de la topographie. Cette régularité traverse les frontières étatiques et s'étend sur plusieurs états voisins.

On dispose aussi des valeurs de la production (en boisseaux), et de la surface de terres récoltées (Figure III.9). Les cartes des productions sont bien moins régulières spatialement que celles des rendements, même si l'on retrouve les principaux bassins de production. Devant l'importante disparité de production des cantons, il devient nécessaire d'utiliser une échelle logarithmique pour pouvoir les comparer.

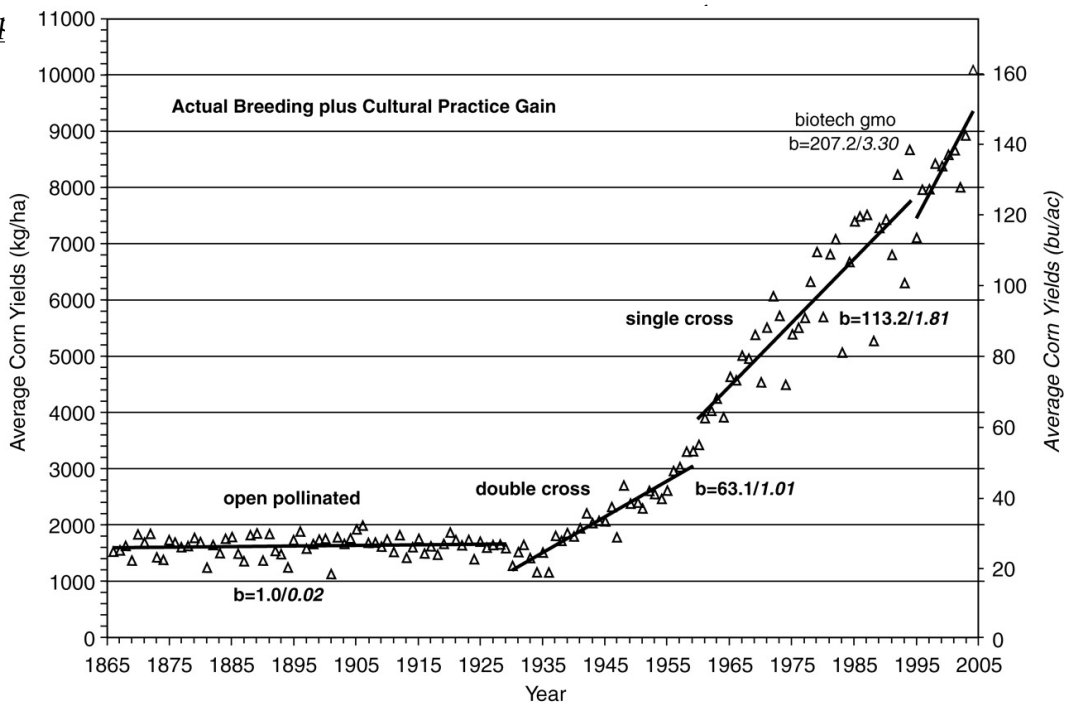


FIGURE III.7 – Rendements moyens de maïs aux États-Unis, et coïncidence avec l'utilisation de différentes variétés de maïs (USDA-NASS, 2005) [Troy06].

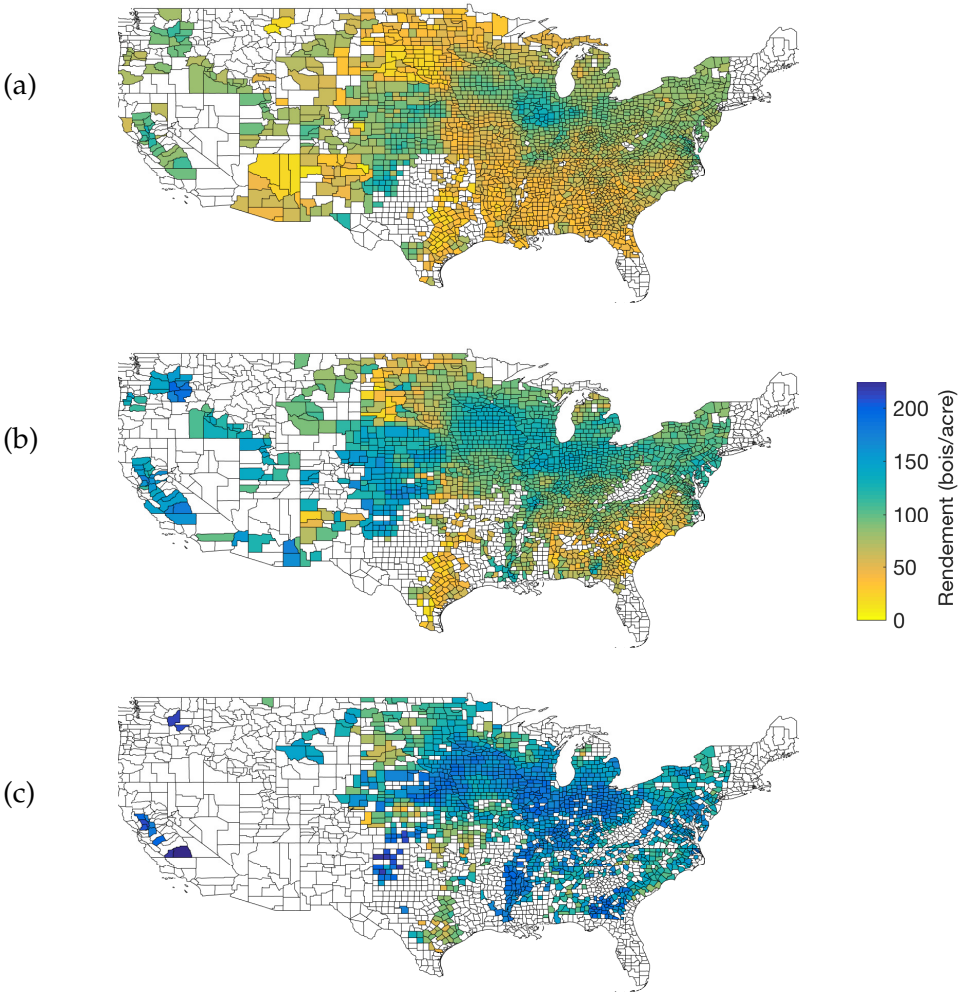


FIGURE III.8 – Rendement de maïs pour l'année 1975 (a), 1990 (b), et 2013 (c) par canton aux États-Unis, en boisseaux/acre. Les données pour les cantons en blancs ne sont pas connues.

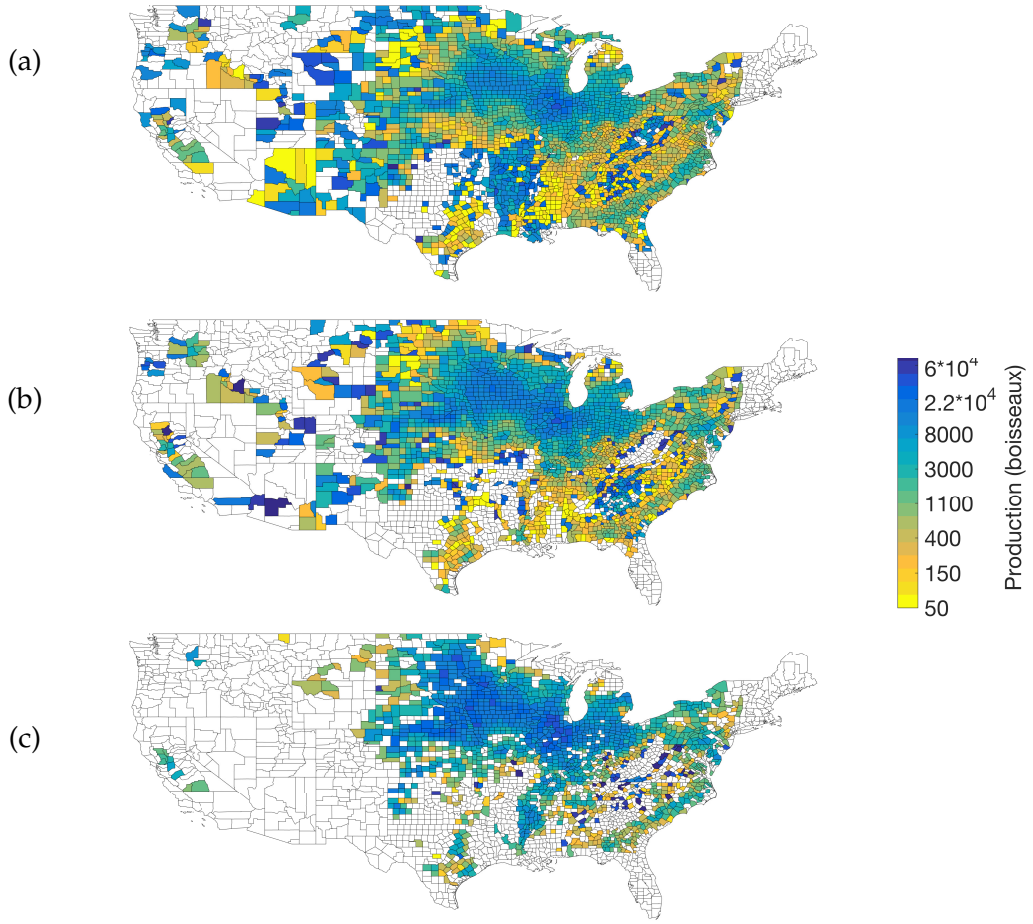


FIGURE III.9 – Production de maïs pour l'année 1975 (a), 1990 (b), et 2013 (c) par canton aux États-Unis, en boisseaux. Les données pour les cantons en blancs ne sont pas connues.

3 Les données météorologiques standard

Que sont les ré-analyses ?

La ré-analyse est une méthode utilisée pour obtenir un portrait exhaustif de l'état du système terrestre. Cette méthode consiste à combiner un modèle de prévision météorologique et des observations pour produire - généralement pour l'ensemble du globe - des archives d'un grand nombre de variables atmosphériques, océaniques ou de surfaces continentales sur des grilles à une résolution temporelle de quelques heures pour plusieurs décennies. On appelle les données ainsi produites des réanalyses. Pour mieux comprendre en quoi consistent les réanalyses, il est utile de s'attarder un peu sur la production des analyses.

Afin de produire des prévisions météorologiques de qualité, le modèle de prévision doit connaître l'état de l'atmosphère et de la surface au moment identifié comme étant le temps initial de la prévision à produire. Pour ce faire, des observations de diverses sources et régions du globe sont intégrées dans le modèle de prévision par le biais d'une procédure appelée "le cycle d'assimilation des données". Ces observations

proviennent de radiosondages aérologiques, de stations météorologiques de surface, de satellites, de radars, etc. L'assimilation de données observées repose sur des techniques mathématiques sophistiquées qui doivent composer avec le fait que les observations sont effectuées à des temps différents, qu'elles contiennent des erreurs ou sont incomplètes. Ces données peuvent aussi provenir de sources et d'instruments variés ayant diverses couvertures spatiales et temporelles. Elles peuvent même être inexistantes pour certaines variables du modèle, comme par exemple le contenu en eau du sol.

À l'instar des modèles de climat, les modèles de prévisions météorologiques sont aussi des modèles physiques basés sur les équations de la mécanique des fluides. Ces équations leur procurent la capacité de produire de nombreuses variables cohérentes entre elles dans l'espace et dans le temps. Une fois son cycle d'assimilation complété, le modèle de prévision établit le portrait le plus fidèle possible de l'atmosphère au temps initial pour plusieurs dizaines de variables cohérentes sur sa grille de calcul, incluant les variables qui ne sont pas mesurées : c'est ce que l'on appelle "une analyse". Autrement dit, les analyses comprennent deux types de variables : 1) celles pour lesquelles le modèle a pu tenir compte des observations et 2) celles qui sont de purs produits du modèle parce qu'aucune observation n'est disponible. Lorsque l'analyse est prête, le modèle l'utilise pour initialiser la simulation qui lui permettra de produire les prévisions météorologiques du temps à venir. Les analyses sont produites entre deux et quatre fois par jour et sont conservées par les différents centres de prévisions météorologiques dans le monde.

Les analyses météorologiques produites depuis plusieurs décennies ont une valeur importante car, de cette série de portraits de l'atmosphère réelle, il est possible d'en dériver les caractéristiques du climat réel pour un grand nombre de variables. Cette tâche en apparence simple peut s'avérer ardue car, à travers les années, des modèles de prévisions météorologiques de plus en plus performants se sont succédés, leur résolution horizontale et verticale s'est raffinée, certaines variables ont disparu tandis que de nouvelles ont été introduites. Ce manque d'homogénéité fait en sorte que les analyses sont des produits difficiles à utiliser, particulièrement pour analyser le climat sur de longues périodes.

Pour remédier à ce problème, certains centres de prévision ont choisi de revisiter leurs archives afin de produire des réanalyses. Pour ce faire, ces centres choisissent la version la plus récente de leur modèle de prévision météorologique et de leur algorithme d'assimilation de données, puis fixent ensuite une résolution horizontale et une résolution verticale qui demeureront uniformes pour toute la période revisitée. De nouvelles sources d'observations peuvent même être intégrées à la procédure ce qui permet d'améliorer la représentation du climat. Les réanalyses couvrent plusieurs décennies.

La production de réanalyses demande un effort important en termes de ressources financières et techniques qui ne sont à la portée que de quelques grands centres de prévision, les plus connus étant le NCEP (National Center for Environmental Prediction) ou l'ECMWF (European Center for Medium-Range Weather Forcasts). Même si les réanalyses intègrent des données observées et qu'elles offrent un portrait de la réalité, on peut remarquer des différences importantes selon les bases de données de réanalyses, surtout dans les régions ou les observations sont rares. De plus, les diverses réanalyses

sont issues de modèles de prévisions différents qui n'utilisent pas les mêmes schémas d'assimilation. Leurs résolutions tout comme leurs sélections d'observations varient également. Les variables pour lesquelles il n'existe pas d'observations présentent potentiellement les plus grandes disparités car elles sont déterminées uniquement par le modèle de prévision.

L'étude de [Chal05] a montré que les réanalyses de l'ECMWF sont bien adaptées à la prédictions des rendements agricoles.

Températures et précipitations.

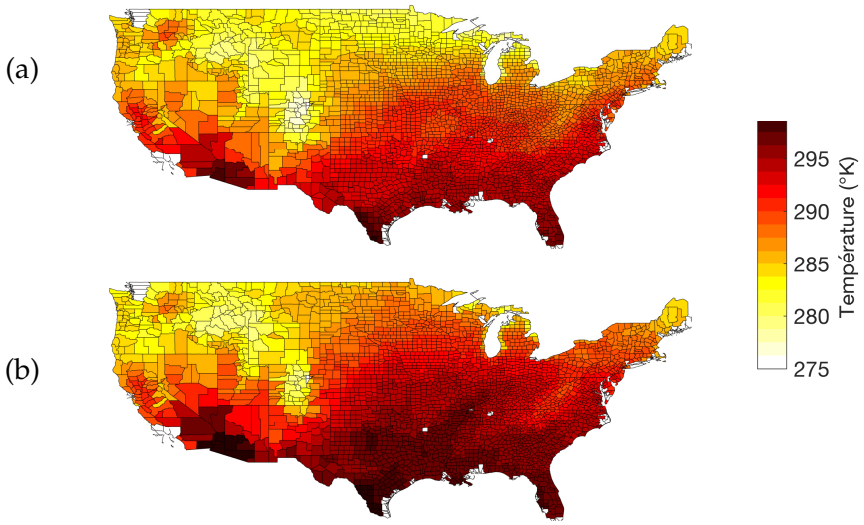


FIGURE III.10 – Températures moyennes de mai pour l'année 1979 (a) et 2012 (b) par canton aux États-Unis, en degrés Kelvin.

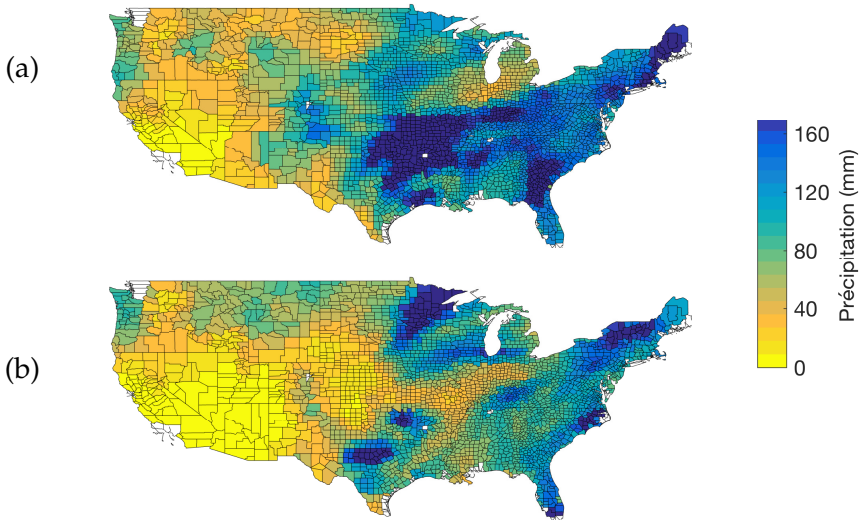


FIGURE III.11 – Cumuls de précipitations de mai pour l'année 1979 (a) et 2012 (b) par canton aux États-Unis, en mm.

Les données de température (moyenne mensuelle, minimum journalier, et maximum journalier à 2m de hauteur) et de précipitation (cumul mensuel) proviennent des ré-analyses ERA-Interim de l'ECMWF. Elles couvrent la période 1979-2013 [Uppa05]. Ces données météorologiques sont fournies sur une grille régulière de 75 km$\times$75 km. Les cartes des températures moyennes de mai en 1979 et en 2012 sont représentées par la Figure III.10. Les cumuls moyens de précipitation en mai sont aussi représentés à la Figure III.11. On observe de grandes variations spatiales : en général, le température décroît du sud au nord, tandis que les précipitations diminuent du sud au nord et de l'est à l'ouest. Les cantons qui présentent les plus hauts rendements, se trouvent dans des régions aux fortes précipitations mais aux températures modérées.

La Figure III.12 représente pour quatre états différents, les séries temporelles des températures moyennes mensuelles (une couleur par mois), de 1979 à 2013. Les températures des mois d'été sont semblables entre les 4 états, qu'ils soient au sud ou au nord. Le mois de juillet est toujours le mois le plus chaud, puis vient celui d'août. On n'observe pas clairement d'évolution croissante des températures de 1979 à 2010 quand on regarde ces températures moyennes mensuelles. En revanche, pour les mois du début d'année, les températures sont bien plus basses au nord, et les variations d'une année à l'autre sont plus importantes. Au sein de chaque état, les cantons ont bien sûr des températures aux variations très semblables.

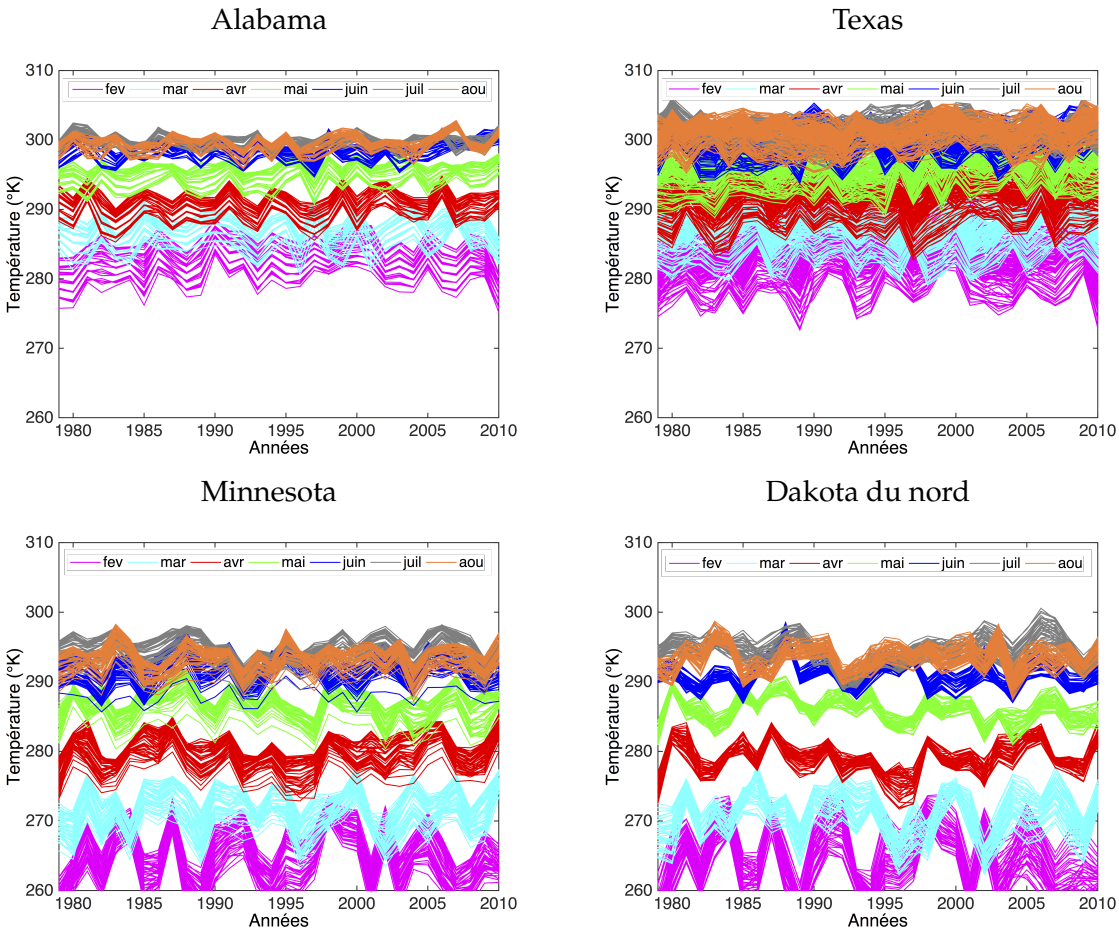


FIGURE III.12 – Séries temporelles des températures à deux mètres pour 4 états différents : Alabama, Texas, Minnesota et Dakota du nord. Chaque bloc de couleur rassemble les données d'un seul mois. Chaque ligne correspond aux données d'un canton particulier.

La Figure III.13 illustre les précipitations mensuelles moyennes (de 1979 à 2013) pour 4 états (Alabama, Texas, Minnesota et Dakota du nord). Chaque courbe correspond aux données d'un canton. On remarque que pour le Minnesota et le Dakota du nord, les précipitations sont moins importantes pendant les mois d'hiver que pendant le reste de l'année. Pour les deux autres états, les variations sont plus importantes d'un mois à l'autre, avec un pic de précipitation aux mois de mars et juillet pour l'Alabama, et aux mois de mai et septembre pour le Texas. La variabilité entre canton est souvent plus importante de mars à septembre. Les cantons du Texas ont des comportements particulièrement variables, comparé à ceux du Minnesota ou de l'Alabama. Ces comportements différents étaient aussi présents pour l'évolution des rendements à la Figure III.6.

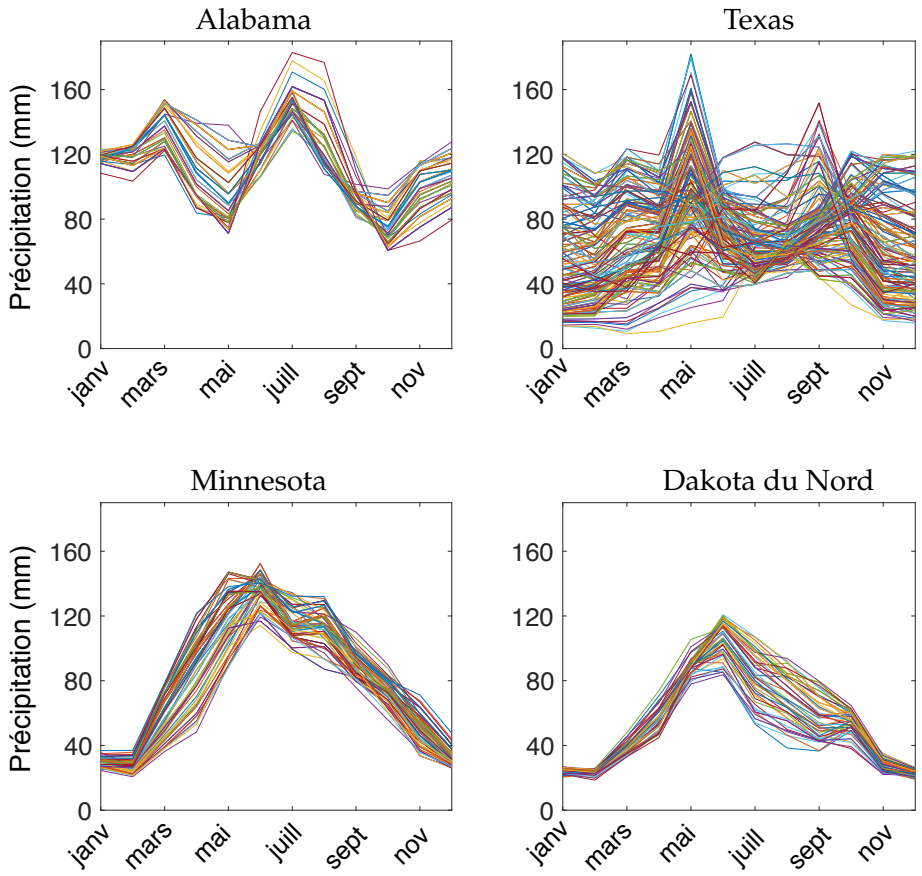


FIGURE III.13 – Moyennes des précipitations mensuelles sur l'étendue temporelle 1979-2013 pour 4 états différents : Alabama, Texas, Minnesota et Dakota du Nord. Chaque courbe correspond aux données d'un canton.

3.1 Irrigation

Selon des données recensées par l'USGS (U.S. Geological Survey) en 2005, la majorité des prélèvements d'eau à des fins d'irrigation (85 %) et des surfaces irriguées (74 %) était située dans 17 états de l'ouest. Ces 17 états sont localisés dans des zones où la précipitation annuelle moyenne vaut moins de 50 cm et est insuffisante pour satisfaire la récolte sans eau supplémentaire. La Californie, l'Idaho, le Colorado et le Montana représentent à eux seuls 49 % des prélèvements totaux d'eau pour l'irrigation et 64 % des prélèvement d'eau de surface. Quelques données d'irrigation sont disponibles à l'échelle du canton tous les cinq ans depuis 1985, auprès de l'USGS [USGC17].

Presque 90% de l'eau souterraine utilisée pour l'irrigation ont été prélevées dans 13 états et chacun de ces états a retiré plus de 3,7 millions de m^3 d'eau souterraine par jour pour l'irrigation en 2005. Parmi ces 13 états, l'eau souterraine était la principale source d'irrigation pour le Nebraska, l'Arkansas, le Texas, le Kansas, le Mississipi et Missouri. Cinq états - Californie, Nebraska, Texas, Arkansas, Idaho - concentrent 52 % des terres irriguées. Le Nebraska, le Texas et la Californie représentent 41 % de la superficie irriguée en utilisant des systèmes de micro-irrigation et d'aspersion.

L'impact de la précipitation est donc à nuancer face à l'apport considérable de l'irrigation dans certains états (Figure III.14) comme au Texas ou en Californie. Cependant,

en ce qui concerne les données agricoles pour le maïs, il s'avère qu'elles proviennent essentiellement de régions situées sur la moitié est des États-Unis, qui est celle où l'irrigation est la moins développée. Les 17 états qui irriguent le plus (ils sont situés à l'ouest de la bande noire de la Figure III.14) cumulent à eux seuls 93% de l'irrigation totale en surface, et 69% de l'irrigation souterraine totale. Ces régions là doivent être écartées pour obtenir un lien plus clair entre précipitation et rendement. Ainsi, tous les états situés à l'est du méridien -103°E (et quelques états isolés comme le Nebraska et le Texas) ne seront pas pris en compte dans cette étude. En outre, ces états produisent très peu de maïs.

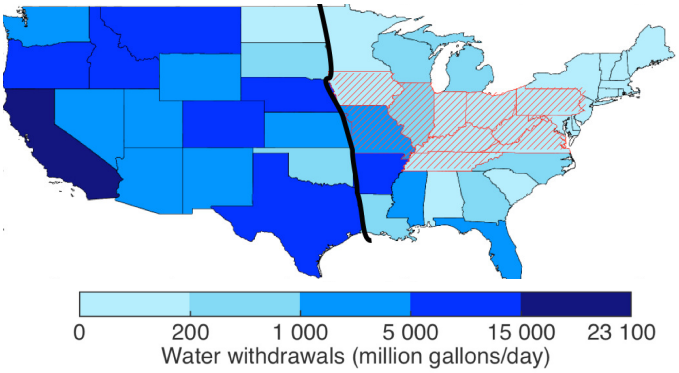


FIGURE III.14 – Prélèvements d'eau totaux pour l'irrigation, en millions de gallons par jour, par état, en 2010 [Maup14]. Les zones hachurées en rouge représentent les états les plus sensibles aux intempéries (comme on le verra par la suite). On voit aussi qu'ils dépendent peu de l'irrigation (http ://water.usgs.gov/edu/wuir.html).

Finalement, les 21 états considérés dans cette étude sont : Kansas, Oklahoma, Iowa, Missouri, Illinois, Indiana, Kentucky, New Jersey, Delaware, Maryland, Virginie, Ohio, Virginie occidentale, Pennsylvanie, Caroline de nord et du sud, Géorgie, Tennessee, Alabama, Louisiane, et Mississipi.

3.2 La projection des données météorologiques sur les cantons

On utilise l'organisation territoriale des États-Unis et sa subdivision en cantons, pour projeter les données météorologiques, de sa grille régulière d'unité 75 km$\times$75 km, vers des données à l'échelle du canton. Pour ce faire, on utilise pour chaque canton la technique du plus proche pixel météo.

Cette technique peut être améliorée en prenant en compte la part de la surface du pixel recouvrant le canton, ou en utilisant une interpolation bilinéaire. Cependant, dans notre cas, cela n'a pas beaucoup d'importance : les températures mensuelles et les précipitations ont une régularité spatiale qui dépassasse les frontières des cantons, et les pixels voisins ont des variables météorologiques aux valeurs très proches.

4 Les indices agroclimatiques

Les indices agroclimatiques sont en général construits à partir des données de températures et précipitations quotidiennes ou mensuelles pour obtenir des variables qui ont un meilleur sens agronomique. Les indices agroclimatiques sont censés mieux représenter l'impact de la météo sur les cultures. Les indices thermiques inclus des mesures cumulées (comme les degré-jours de croissance, l'unité thermique du maïs, les

degré-jours de chauffage, ou les degré-jours de refroidissement), des mesures seuils (par exemple, le nombre de jours où la température est supérieure à $30°C$), ou des dates (comme la date du premier gel automnal ou celle du dernier gel printanier, la longueur de la saison sans gel, la date de début de la saison de croissance, la date de fin de la saison de croissance, la longueur de la saison de croissance). Les indices relatifs à l'eau inclus les cumuls de précipitations sur différentes périodes temporelles (saison de croissance, mois d'été), l'évapotranspiration des cultures, ou le bilan hydrique. Les formules utilisées pour calculer ces indices sont résumées dans le Tableau III.1. La définition des acronymes utilisés dans la suite de l'étude pour désigner ces indices agroclimatiques, est également donnée. Plus de détails sont disponibles dans [Lepa12].

L'indice SPEI (Standardized Precipitation-Evapotranspiration Index) peut lui aussi être calculé, mais laborieusement. Un bon papier résumant le calcul de l'évapotranspiration par la méthode FAO-56 (Penman-Monteith) est celui de [Alle98]. Sinon, une base de données SPEI est disponible chez [Begu14] où l'évapotranspiration a aussi été calculée avec l'équation de Penman-Monteith. Plus d'informations sur le calcul du SPEI sont disponibles chez `http://sac.csic.es/spei/database.html`. Le calcul de l'évaporation potentielle est nécessaire pour obtenir l'indice SPEI : [Bare13] comparent différentes formules qui sont habituellement utilisées.

Comme le SPEI, l'humidité du sol (SM) est aussi une alternative aux données de précipitation et pourrait mieux représenter la quantité réelle d'eau dont dispose la plante. Les données simulées d'humidité du sol proviennent de l'ECMWF, sur une grille d'unité $75km \times 75km$. La variable utilisée dans cette étude est celle nommée "volumetric soil water layer 2 (7-28 cm)" en m^3/m^3 [Albe12, Bals12]. La seconde couche est celle qui est le plus en contact avec les racines du maïs.

Les variables météorologiques prennent des valeurs qui ne sont pas du même ordre de grandeur. Pour éviter que certaines variables soient considérées comme plus importantes que d'autres par le modèle du fait de leurs importantes valeurs, il est préférable de renormaliser chaque variable V dans l'intervalle $[-1, 1]$ par la formule $2\frac{V-\min(V)}{\max(V)-\min(V)} - 1$. Ainsi, nous considérons des anomalies d'indicateurs météorologiques. Dans la suite de ce manuscrit, "indicateur" ou "prédicteur" fera référence à ces variables météorologiques recentrées.

Ces données pourraient être complétées par d'autres variables prédictives potentielles, comme par exemple, des données d'ensoleillement, ou d'engrais. La faible disponibilité et la faible étendue temporelle de ces dernières sont l'une des principales limites à leur utilisation.

En ce qui concerne le type de sol, selon l'US General Soil Map [Soil98b], la moitié des régions concernées dans cette étude est des ultisols (sud-est des États-Unis), un quart sont des alfisols (est, midwest), et un quart sont des mollisols (ouest et nord-ouest de la région considérée) (Figure III.15). Les plaines inondables du Mississipi sont des vertisols. Quand on regarde la capacité de rétention d'eau de ces sols (en millimètres d'eau par mètre de sol), les vertisols ont une forte capacité (100-200mm), et les trois autres types de sol ont un potentiel modéré (25-100mm) selon [Soil98a]. Quasiment tous les cantons considérés ont un régime hygrométrique du sol dit "udic" (humid or subhumid climate), et appartiennent au biome "boréal humide. Comme on le verra par la suite, le type de sol ne fait pas partie des prédicteurs de nos modèles, et nos résultats ne montrent aucune corrélation claire entre la qualité des modèles, la répartition spatiale de cette qualité, et la classification des types de sols.

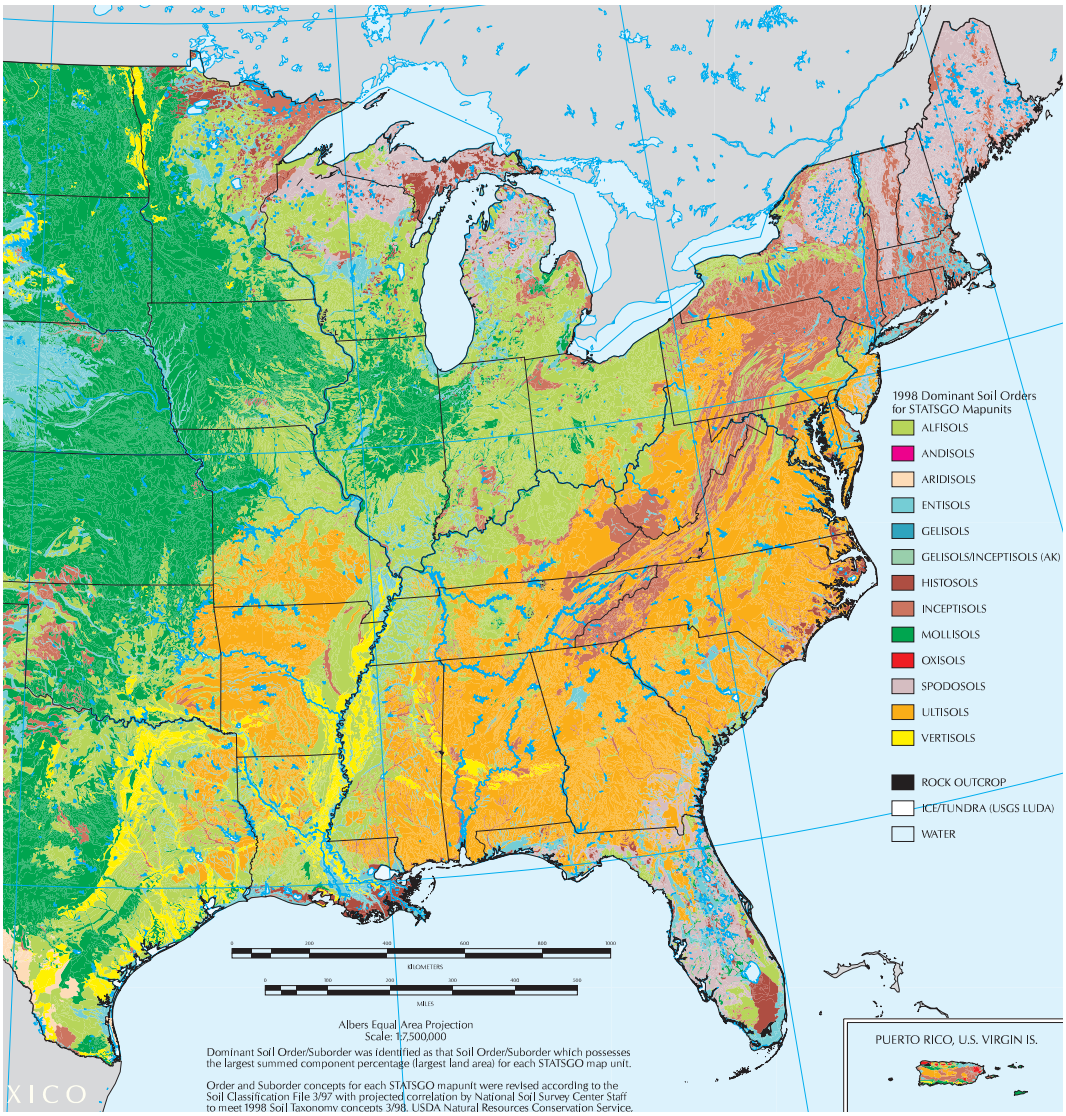


FIGURE III.15 – Carte du sol pour l'est des États-Unis (tiré de [USDA05]).

L'USDA propose aussi, par état, les dates les plus probables de semis et de récolte du maïs (Tableau III.2). Ces dates sont essentielles pour le calcul de la plupart des indices agroclimatiques.

Variables les plus courantes		Unités
P_j	Précipitation du jour j	mm
Pmois	Précipitations cumulées du mois "mois"	mm
$Tave_j$	Température moyenne du jour j $Tave_j = \frac{Tmin_j + Tmax}{2}$	^oC
Tbase	Seuil de température de croissance	^oC
T_{frost}	Seuil de température gélive	^oC
Tmois	Température moyenne du mois "mois"	^oC
$Tmin_j$	Température minimale du jour j	^oC
$Tmax_j$	Température maximale du jour j	^oC
Indices agroclimatiques (nom et formule)		
CDF5	Cumul des degrés-froid durant la période d'endurcissement (potentiel d'endurcissement) $CDF5 = \sum_{j=213}^{FPE-1} CDF5_j$ $CDF5_j = \begin{cases} 0 & si \quad j = 212 \\ \max\{0,\ CDF5_{j-1} + DF_j\} & sinon \end{cases}$ $FPE = \min\{j \mid Tmin_j \leqslant -10^oC \quad et \quad j > 212\}$ $DF_j = CD_j - DJ_j$ $CD_j = \begin{cases} 0 & si \quad Tave_j \geqslant 5^oC \\ \lvert Tave_j - 5\rvert & si \quad Tave_j < 5^oC \end{cases}$ $DJ_j = \begin{cases} 0 & si \quad Tave_j \leqslant 5^oC \\ Tave_j - 5 & si \quad Tave_j > 5^oC \end{cases}$	Jour
DGP	Date du dernier gel printanier $DGP_{T_{frost}} = \max\{j \mid Tmin_j \leqslant T_{frost}\}$ j=1, ..., 212	Jour julien
DJ	Cumul des degré-jours d'avril à octobre $DJ = \sum_{j=91}^{304} DJ_j$ $DJ_j = \max\{0, (Tave_j - Tbase)\}$	Degré-jours
DJO_{HIV}	Cumul des degré-jours au cours de la période froide (perte d'endurcissement) $DJO_{HIV} = \sum_{Dhiv}^{Fhiv} \max\{0, Tave_j\}$ $Dhiv = \min\{j \mid Tmin_j \leqslant -15^oC\}$ $Fhiv = \max\{j \mid Tmin_j \leqslant -15^oC\}$	Degré-jours
DJsc	Cumul des degré-jours durant la saison de croissance $DJsc = \sum_{j=DSC}^{FSC} DD_j$	Degré- jours
DSC	Date de début de la saison de croissance $DSC = \min\{j \mid TMMP5_j > 5.5^oC\}$ j=1, ..., 365	Jour
D_{UTM}	Date de début de cumul des unités thermiques du maïs $D_{UTM} = \min\{j \mid TMM5_j > 12.8^oC\}$	Jour

FSC	Date de fin de la saison de croissance $FSC = \max\{j \mid TMMP5_j > 5.5^oC\}$ j=1, ..., 365	Jour
F_{UTM}	Date de fin de cumul des unités thermiques du maïs $F_{UTM} = \min\{j \mid Tmin_j \leqslant -2^oC\}$ j=213, ..., 365	Jour
LSC	Longueur de la saison de croissance $LSC = FSC - DSC$	Jour
LSSG	Longueur de la saison sans gel $LSSG_{T_{frost}} = PGA_{T_{frost}} - DGP_{T_{frost}}$	Jour
OCA	Fréquence des températures supérieures à 30^oC $OCA_{TSUP30} = \sum_{j=1}^{365} TSUP30_j$ $TSUP30_j = \begin{cases} 1 & si \quad Tmax_j > 30^oC \\ 0 & si \quad Tmax_j \leqslant 30^oC \end{cases}$	Jour
PGA	Date du premier gel automnal $PGA_{T_{frost}} = \min\{j \mid Tmin_j \leqslant T_{frost}\}$ j=213, ..., 365	Jour julien
P_{sc}	Cumul des précipitations pendant la saison de croissance $P_{sc} = \sum_{j=DSC}^{FSC} P_j$	mm
SM	Humidité du sol entre 7 et 28 cm	m^3/m^3
TMA	Température minimale annuelle $TMA = \min\{Tmin_j\}$, j=1, ..., 365	^oC
$TMMP5_j$	Moyenne mobile pondérée des températures moyennes quotidiennes sur 5 jours 1/16($Tave_{j-4} + 4.Tave_{j-3} + 6.Tave_{j-2} + 4.Tave_{j-1} + Tave_j$)	
$TMM5_j$	moyenne mobile sur 5 jours de la température moyenne quotidienne 1/5($Tave_{j-4} + Tave_{j-3} + Tave_{j-2} + Tave_{j-1} + Tave_j$)	
UTM	Cumul des unités thermiques du maïs $UTM = \sum_{j=D_{UTM}}^{F_{UTM}} UTM_j$ $UTM_j = 1/2(Ymax_j + Ymin_j)$ $Ymax_j = \begin{cases} 0.33(Tmax_j - 10) - 0.084(Tmax_j - 10)^2 & si \quad Tmax_j > 30^oC \\ 0 & si \quad Tmax_j \leqslant 10^oC \end{cases}$ $Ymin_j = \begin{cases} 1.8(Tmin_j - 4.44) & si \quad Tmin_j > 4.44^oC \\ 0 & si \quad Tmin_j \leqslant 4.44^oC \end{cases}$	CHU

TABLEAU III.1 – Indices agroclimatiques utilisés dans cette thèse par ordre alphabétique. Plus de détails sont disponibles dans [Côt12]. Les dates habituelles de plantation et de récolte par état aux États-Unis sont disponibles dans [USDA10].

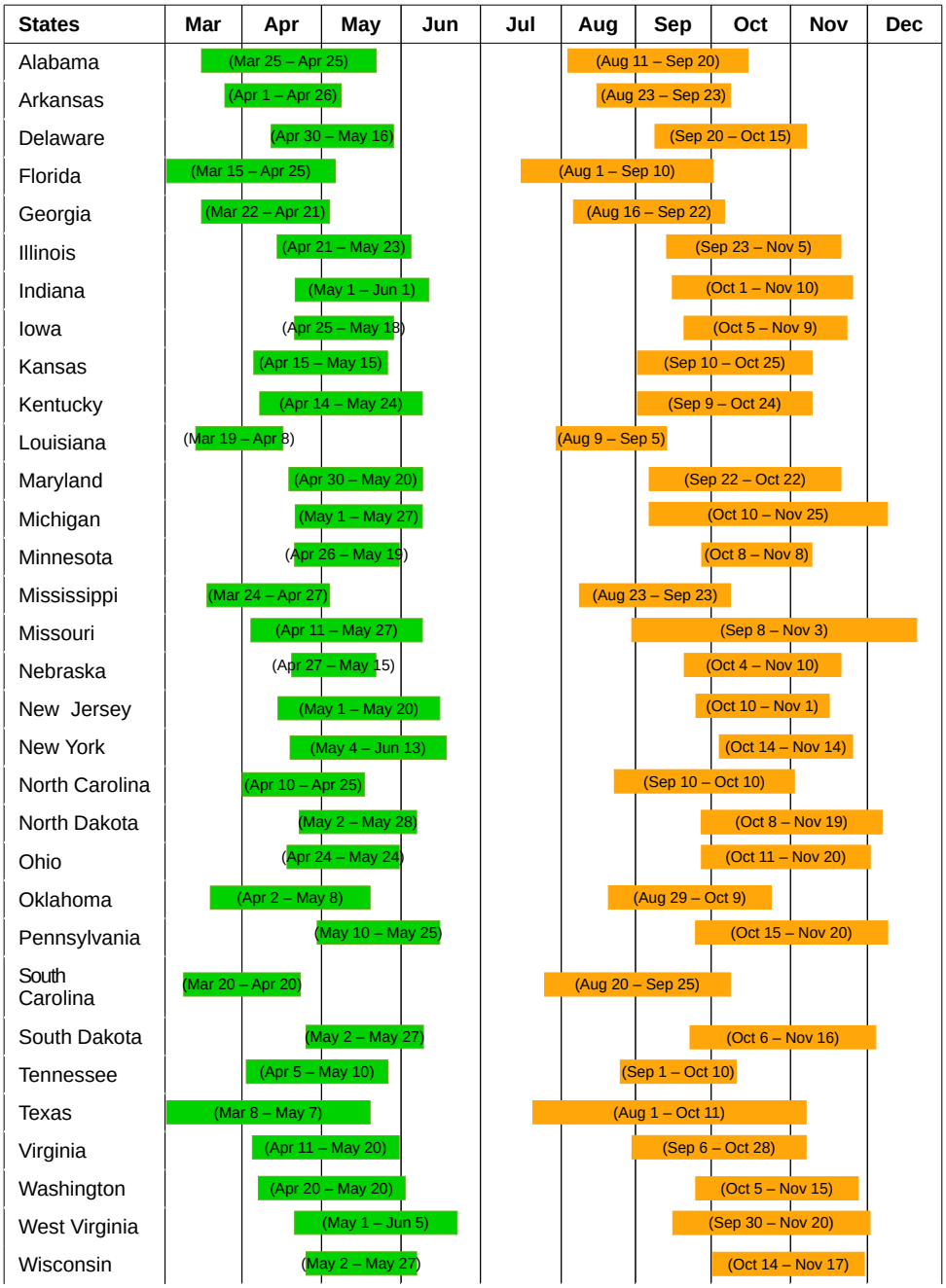


TABLEAU III.2 – Date de semis (en vert) et de récolte (en orange) en 2009 du maïs selon différents états des États-Unis. Entre parenthèse, les dates les plus rencontrées [USDA10].

5 Sensibilité du maïs à la météo

Avant toute analyse statistique poussée, il est important de bien analyser les données utilisées : l'utilisation de graphiques est alors primordial pour les confronter, les comparer, et connaitre leur variabilité. C'est aussi lors de cette pré-analyse que l'on initie l'étude de la capacité des variables météo à expliquer une part de la variabilité des rendements.

Une analyse de sensibilité peut être extrêmement détaillée. Ici, on se limitera à une étude graphique (représentation comparative ou non des données), et à une analyse de covariance, donc à une sensibilité globale. Cette analyse nous guidera dans la phase de sélection des variables météorologiques pour la prédiction des rendements agricoles.

5.1 Ce que dit l'industrie du maïs

Même si l'approche de cette thèse est une approche "orientée-données", et que l'on souhaite obtenir des informations à partir des données essentiellement, il peut être intéressant et important de comparer les résultats que nous donneront les données réelles avec les connaissances a priori d'experts du domaine. Dans cette sous-section, on liste les informations et les conseils qui sont transmis aux cultivateurs par l'industrie du maïs.

Selon l'industrie du maïs, le semis doit se faire le plus tôt possible, dès que la température de la terre dépasse 10 ^oC (entre fin-mars et mi-mai dans hémisphère nord) pour, (1) favoriser l'enracinement précoce des plantes, permettant une meilleure résistance à la sécheresse d'été et car les journées de mai sont plus efficaces en moyenne, sur la croissance du maïs que les journées de septembre, et (2) faire coïncider au maximum la période de floraison avec la période de plus forte luminosité (20 juin) pour obtenir une récolte précoce en automne.

Avec son système racinaire superficiel, le maïs nécessite une importante irrigation en zones à étés secs comme le Sud de l'Europe, l'Égypte, le Chili ou le Pérou. Selon le type de climat et de sol, les besoins varient de 800 à 1500 m^3 d'eau/ha/mois de la floraison jusqu'à la maturation des grains. Dans les zones plus humides, les grands producteurs de maïs dans le monde (États-Unis, Chine, Brésil, Argentine, Europe de l'Est) se passent d'irrigation (sauf pour leurs productions de semences). Dans les zones tempérées de l'hémisphère nord, le maïs est semé en avril-mai et fleurit en juillet-août. Les grains atteignent la maturité entre fin septembre et novembre selon les variétés. La récolte a lieu lorsque la plante jaunit et se dessèche. La plante entière peut également être récoltée et ensilée avant la maturité du grain (septembre).

Les professionnels du maïs utilisent le nombre de feuilles présentes sur la plante pour décider des actions à mener pendant sa croissance (Figure III.16). Ainsi, lorsque la plante a développé une première feuille complète (collerette bien apparente), c'est le stade V1 et il est conseillé de désherber. Au stade V8 (8e feuille complète), on recommande un apport en engrais pour une bonne fructification. Au stade V10, on démarre l'irrigation dans les zones où elle est nécessaire, etc. [Widh14] propose un document résumant les différentes étapes de croissance du maïs avec les jours estimés de l'année et les degré-jours de croissance.

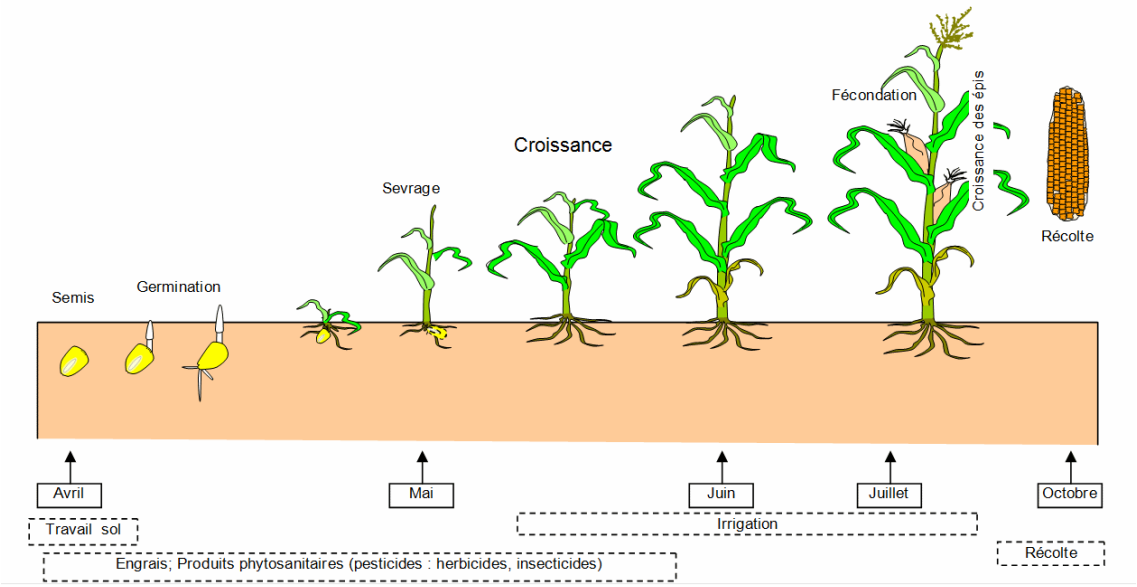


FIGURE III.16 – Les différentes phases de croissance du maïs et les traitements spécifiques en fonction des mois de la croissance (schéma d'Alain Gallien).

La courbe lisse noire de la Figure III.17 (courbe A) illustre le modèle moyen d'utilisation de l'eau à long terme par le maïs. Ce modèle montre les niveaux quotidiens typiques moyen (sur 10 ans) de l'Évapo-Transpiration (ET) de la culture tout au long de la période de croissance. La courbe avec beaucoup de variations (courbe B) illustre la fluctuation possible des valeurs quotidiennes de l'ET sur une année. Ainsi, les personnes responsables de l'irrigation doivent se familiariser avec la tendance à long terme, mais doivent aussi être en mesure de déterminer quelle était la valeur de l'ET les jours précédents. La connaissance de la tendance à long terme et des taux quotidiens réels d'utilisation en eaux des cultures sont essentiels pour déterminer l'irrigation et la quantité d'eau à fournir.

Comme on a pu le voir à la Figure III.2 (page 62), le semis du maïs ne se fait pas avant mars et la récolte pas après novembre. Il semble donc peu pertinent de considérer les mois de novembre à février comme pouvant influencer la croissance de la plante (sauf pour un impact à plus long terme).

Il apparaît que le maïs est une plante qui ne demande pas un apport en eau conséquent mais surtout à la bonne période de sa croissance (Figure III.17). La culture d'un hectare d'une plante comme le maïs nécessite en moyenne 6 000 m^3 d'eau au cours des 6 mois de culture soit environ 30 m^3 d'eau par jour et par hectare pendant la saison chaude et en l'absence de précipitations naturelles. A titre comparatif, selon [CNRS00], il faut en moyenne 454 litres d'eau pour produire un kilogramme de maïs contre, 590 pour un kg de pomme-de-terre ou de blé, et 900 pour un kg de soja.

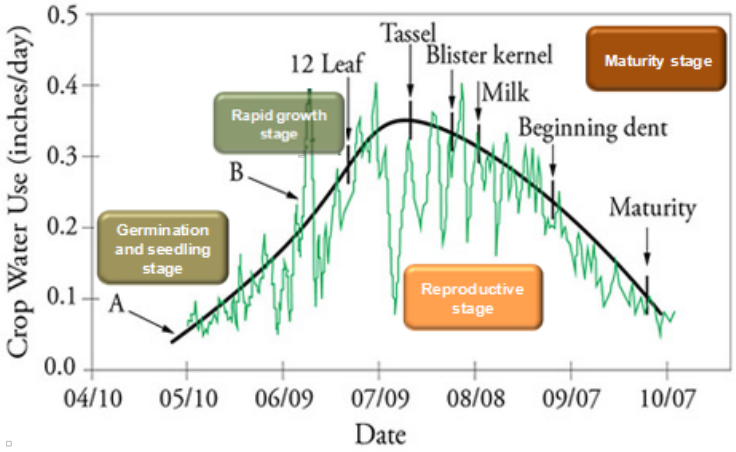


FIGURE III.17 – Besoin en eau du maïs pendant la période de croissance : moyenne journalière à long terme et utilisation annuelle de l'eau avec les étapes importantes de la croissance (tiré de [Kran08]).

5.2 Étude des corrélations entre rendement et variables météorologiques

Les différents prédicteurs que l'on peut regarder sont les températures moyennes mensuelle, des cumuls de précipitation (mensuels ou sur des périodes intelligemment choisies : saisonnière ou phase de croissance), ainsi que toute la liste des indices agroclimatiques qui ont été introduits page 60.

Cependant, dans cette première analyse de sensibilité, on se concentre sur des variables météorologique simples (T et P) facilement obtenu dans les modèles (et pouvant même être prédites pour le court, moyen ou long terme), et non sur les indices agroclimatiques, qui seront utilisés plus tard dans la thèse.

La Figure III.18 rend compte de la variabilité des variables Tjuillet et Pjuillet (anomalies de température moyenne et de précipitation en Juillet) selon les états. Pour Tjuillet, on note des différences de moyenne et de variance entre les états, avec un important nombre de valeurs extrêmes positives pour certains (Kansas, Missouri, Oklahoma). On observe pour le Dakota du Nord et du Sud (deux états au nord de la zone d'étude) une plus faible variabilité de Tjuillet et de Pjuillet avec beaucoup d'extrêmes négatifs pour Tjuillet. On verra que cette faible variabilité (des prédicteurs météorologiques ou du rendement) peu expliquer une moins bonne estimation statistique. Les disparités entre états devront être prises en compte dans les modèles. En ce qui concerne Pjuillet, quasiment tous les états enregistrent des moyennes négatives, poussées par des valeurs extrêmes positives.

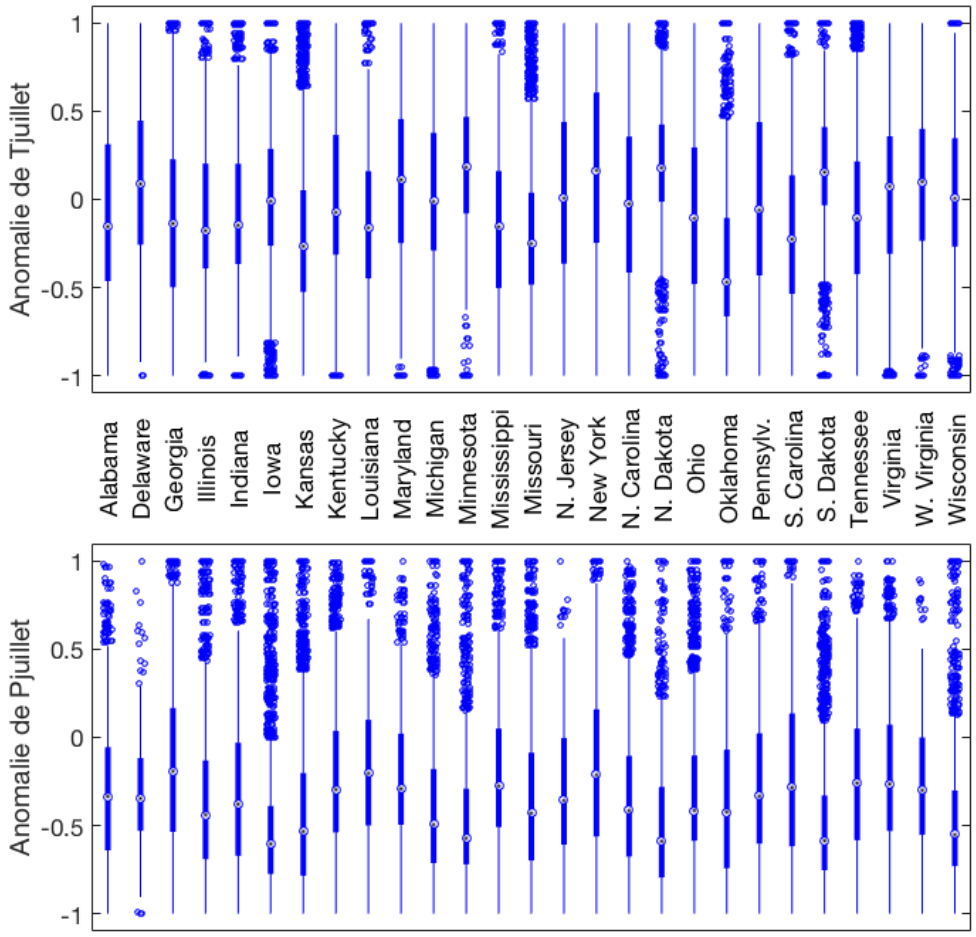


FIGURE III.18 – Box-plot des anomalies de température moyenne et de précipitation, en juillet, selon les états producteurs de maïs.

Sur la Figure III.19, on a classé les différents cantons des États-Unis (au nombre de 2370) par latitude croissance et on a indiqué par une couleur la température moyenne (de 2003 à 2013) pour les mois de février à septembre. On distingue ainsi l'évolution saisonnière moyenne des températures ainsi que l'effet potentiel de la latitude. De même, on a représenté sur la Figure III.20 les précipitations mensuelles moyennes en classant les cantons selon leur longitude. On aperçoit clairement les mois chauds d'été (juillet et août) où les températures sont entre 25^oC et 35^oC quelque soit les cantons. La décroissance des températures vers les hautes latitudes est réelle (voir Figure III.10 page 53) mais pas évidente sur ce graphe. En ce qui concerne les précipitations, les mois d'avril, mai et juin se distinguent des autres par leurs plus fortes précipitations. De plus, les précipitations semblent plus importantes pour les cantons situés dans la moitié est, ce qui est cohérent avec la Figure III.11 page 54.

Les Figures III.21 et III.22 représentent les corrélations entre les anomalies de rendement des séries de plus de 30 années et (1) les températures moyennes mensuelles, ou (2) les précipitations mensuelles.

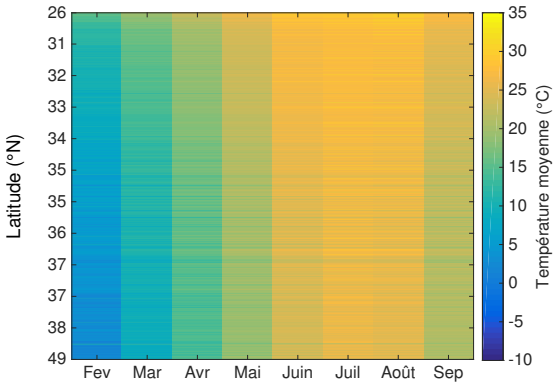


FIGURE III.19 – Températures moyennes mensuelles par cantons (rangés par latitude croissante) sur les 10 dernières années.

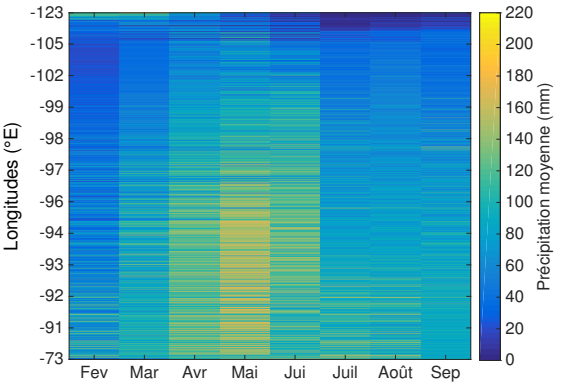


FIGURE III.20 – Cumuls de précipitations moyens mensuels par cantons (rangés par longitude croissante) sur les 10 dernières années.

On remarque sur la Figure III.21 que les corrélations des rendements avec la température sont globalement positifs pour les quatre mois de gauche (février à mai) et négatifs pour ceux de droite (juin à septembre). Il semblerait donc qu'une température supérieure à la moyenne soit favorable aux rendements lorsque cela a lieu pendant les mois de février à mai (en particulier pour le mois de mars), puis défavorable lorsque cela a lieu après le mois de juin (en particulier pour le mois de juillet où les corrélations prennent les valeurs les plus négatives). La Figure III.21 souligne une corrélation plus forte entre les anomalies de rendement et les températures pour les mois de mars-avril-juin-juillet-août. La température du mois de mai ne semble pas avoir un fort effet avec les anomalies de rendement en maïs.

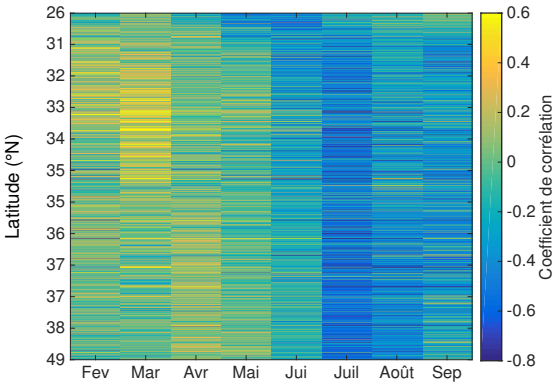


FIGURE III.21 – Corrélation entre les anomalies de rendement des séries de plus de 30 années[3], et les températures moyennes mensuelles. En colonne les températures des mois de février à septembre, en ligne les cantons concernés rangés par latitude croissante.

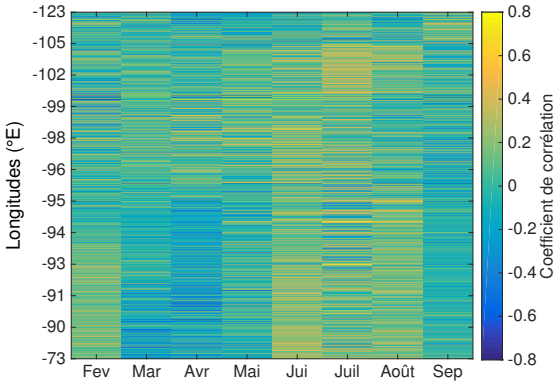


FIGURE III.22 – Corrélation entre les anomalies de rendement des séries de plus de 30 années, et les anomalies des cumuls de précipitation mensuels. En colonne les cumuls de précipitation de février à septembre, en ligne les cantons concernés rangés par longitude croissante.

3. Attention, c'est 30 ans de 1920 à 2013 et non de 1979 à 2013.

En ce qui concerne les températures à l'échelle des cantons, les représentations des corrélations mensuelles sur la carte des États-Unis soulignent une cohérence spatiale des corrélations. La partie ouest, peu renseignée, est souvent peu corrélée. Le mois de juillet semble être le plus fortement négativement corrélé à l'inverse d'avril pour les corrélations positives. On observe de plus une plus grande régularité lorsque l'on classe les cantons par latitudes croissantes : la latitude semble influencer faiblement la corrélation entre les anomalies de rendements et les températures moyennes mensuelles.

Concernant les précipitations, les observations sont très différentes : d'une part on remarque (Figure III.22) que les corrélations gardent des valeurs relativement faibles. Cela peut être dû au biais introduit par l'irrigation intensive dans certains états de l'ouest. La Figure III.22 souligne une corrélation plus élevée entre les anomalies de rendement et celles des cumuls de précipitation pour les mois de avril-juin-juillet-août. On retrouve ainsi les périodes critiques mentionnées par l'industrie du maïs à la Section 5.1 pour le stress hydrique. Lorsque l'on classe les cantons par longitudes, on observe un peu plus de structure pour les précipitations. Ceci n'est pas étonnant puisque le climat et le relief des États-Unis sont relativement bien liés avec la longitude (Figure III.2 page 44). Les cantons à l'est du méridien -95^oE rassemblent les corrélations les plus importantes ce qui reste cohérent avec l'irrigation faible dans ces régions. Les états à l'ouest du méridien -95^oE (c'est-à-dire à partir du Texas, Kansas, Nebraska, Dakota du Nord) sont ceux qui irriguent le plus leurs cultures. Cette rupture est visible sur la Figure III.22 pour les mois de mars, avril et juin.

Une analyse par composantes principales (ACP) a été menée en considérant l'ensemble des prédicteurs météorologiques potentiels. La Figure III.23 illustre la projection des poids accordés aux prédicteurs potentiels sur les deux premières composantes de la ACP : on observe une séparation par la première composante entre facteurs liés à la température (dont DJ ou températures mensuelles [4]) et ceux liés à l'humidité (dont SPEI, humidité du sol [5], ou précipitations mensuelles). La deuxième composante, quant à elle, semble partitionner les prédicteurs de façon saisonnière. On observe de plus une forte proximité des degré-jours mensuels (DJm) avec les températures du même mois. De même pour la proximité entre les SPEI et les précipitations de chaque mois.

4. Sauf mention du contraire, le mot "température" fera par la suite toujours référence à la température de l'air à 2m de hauteur.

5. Sauf mention du contraire, le mot "humidité" fera par la suite toujours référence à l'humidité du sol entre 7 et 28 cm (variable ayant pour acronyme SM)".

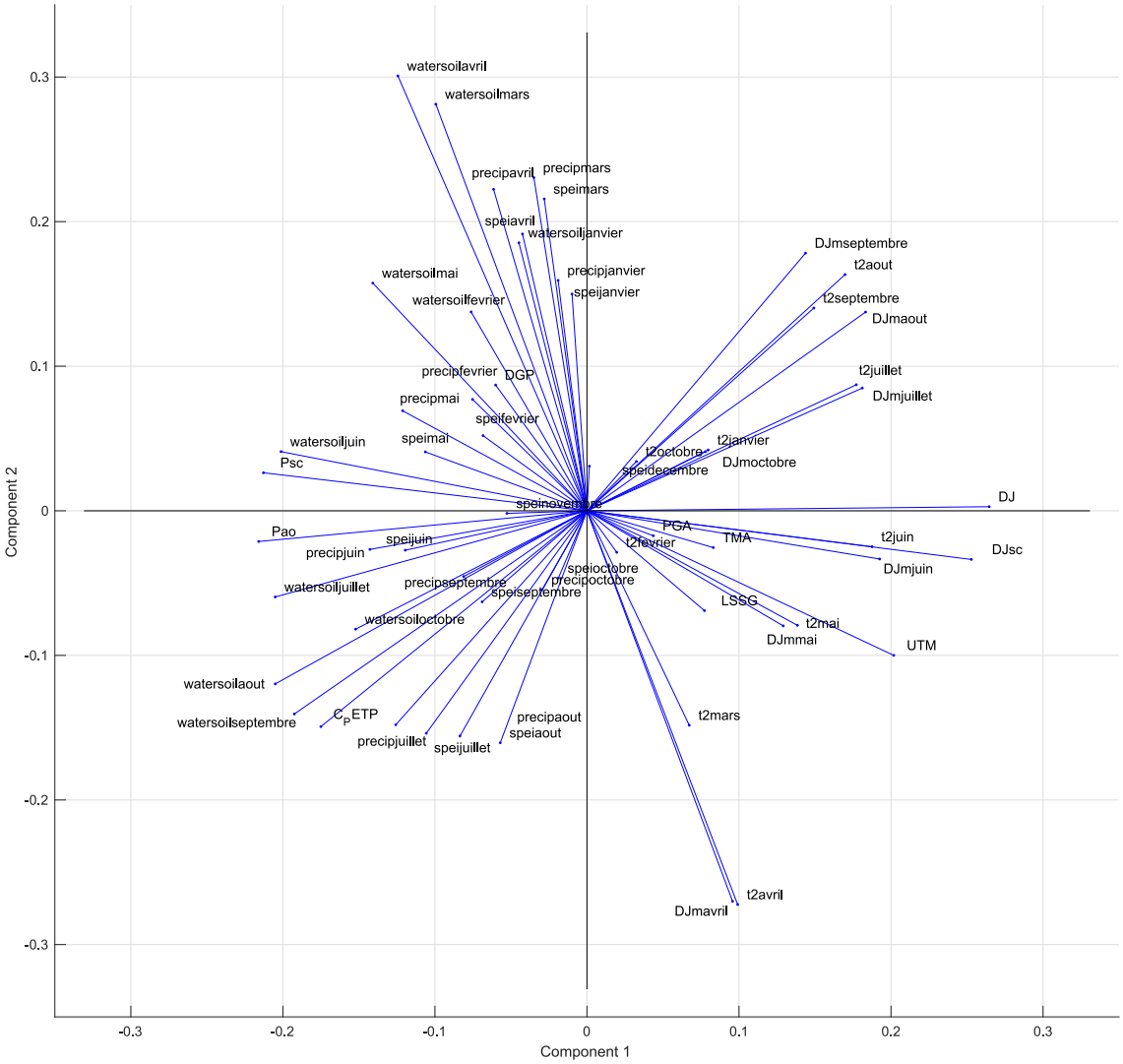


FIGURE III.23 – Projection poids accordés aux prédicteurs potentiels sur les deux premières composantes d'une ACP, considérant les zones climatiques 2 à 5.

5.3 Évolution globale des rendements en fonction des températures et des précipitations mensuelles

La Figure III.24, représente les anomalies de rendement en fonction des anomalies de températures pour les mois de mars à août, et pour six états spatialement éloignés et dont quatre font partie de la Corn Belt (Missouri, Indiana, Alabama, Illinois, Iowa, et Kansas). La densité des points est mise en évidence à l'aide de la barre de couleur. Dans cette sous-section, on tente à l'aide d'une relation polynomiale d'ordre 2 entre températures/précipitations mensuelles et rendements, de modéliser simplement la réponse du rendement aux variations d'une variable météorologique (température ou précipitation). Il s'agit du plus simple modèle que l'on puisse construire pour relier une seule variable météorologique aux rendements mais cela donne déjà une idée de la réponse du maïs aux variations des températures ou des précipitations [Thom86]. Cette figure permet simplement de voir quels sont les états (parmi les six regardés) pour lesquels le rendement de maïs est météo-sensible ou pas.

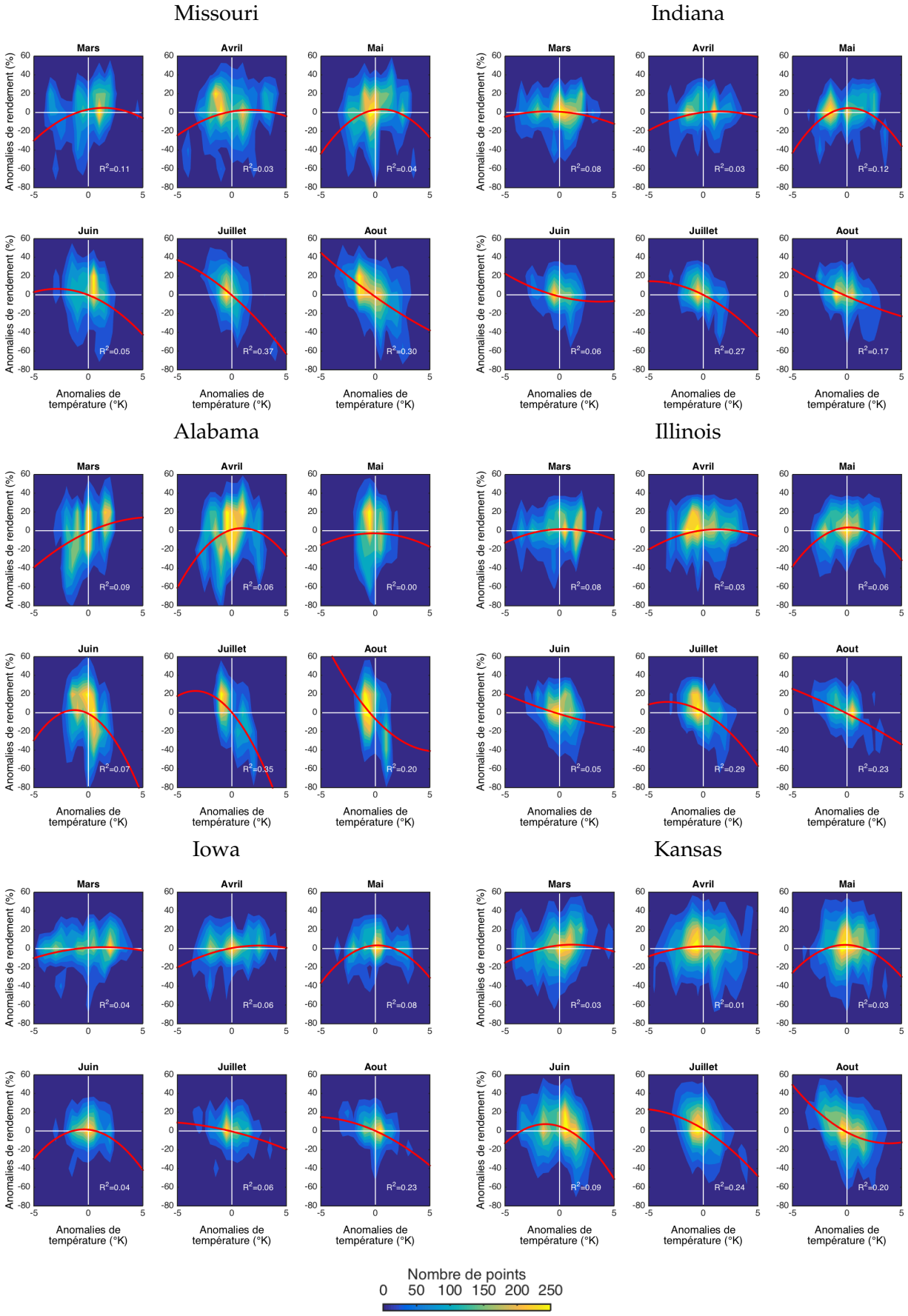


FIGURE III.24 – Anomalies de rendements en fonction d'anomalies de températures (exceptionnellement ici, il s'agit de l'écart par rapport à la température moyenne du canton considéré pour une meilleure comparaison avec les articles de Thompson) pour six états spatialement éloignés (Missouri, Indiana, Alabama, Illinois, Iowa, et Kansas) : la densité des points est représentée en couleur. Une régression polynomiale d'ordre 2 modélise cette relation en rouge.

Une régression polynomiale concave signifie que les valeurs extrêmes (positives ou négatives) des anomalies météorologiques concernées sont néfastes au rendement. À l'inverse, une régression polynomiale convexe souligne un impact positif des anomalies météorologiques extrêmes sur le rendement. Une régression plate souligne un lien peu marqué entre prédicteur météo et rendement.

Concernant les températures (Figure III.24), le mois de juillet reste encore le plus fortement corrélé : c'est le mois pour lequel le coefficient de détermination R^2 (qui mesure la qualité de la prédiction d'une régression linéaire) prend les plus fortes valeurs. Pour la plupart des états, on observe que le polynôme de régression est de dérivé négative pour les mois de juillet et d'août ce qui tendrait à une corrélation négative entre rendement et température de ces mois (tendance observée pour la plupart des états). Pour les mois de mars à mai, en revanche, on note la présence d'un maximum du polynôme (polynôme concave) proche de l'anomalie de température 0 (tendance souvent observée pour les autres états). Cela signifie que les anomalies de températures élevées (en positif ou en négatif) ont un impact négatif sur les rendements, l'optimum étant souvent atteint pour une anomalie de température positive mais faible. La concavité du polynôme varie d'un état à l'autre pour le mois de mai.

Ces résultats sont à comparer avec ceux qu'obtient [Thom86, Thom88] : dans ces articles, Thompson décrit aussi - sous la forme d'une régression polynomiale de degré 2 - la réponse du rendement aux anomalies de température et de précipitation pour les mois de juin à août : le polynôme y est concave et présente un maximum de faible valeur pour la température de juin. On retrouve cette concavité pour la plupart des états, et l'anomalie de température où le maximum de rendement est atteint, est parfois négative, parfois positive mais reste proche de zéro, ce à quoi l'on s'attend lorsque l'on trace une anomalie en fonction d'une autre). Pour le mois de juillet, Thompson décrit une réponse similaire, avec un maximum atteint pour des anomalies de température en juillet assez négative (environ -4) : on retrouve cette même tendance avec nos données. Enfin, pour le mois d'août, Thompson trouve un polynôme concave, dont le maximum est atteint aux alentours de -1. Les données dont nous disposons ne confirme pas cette tendance : on observe bien une pente négative aux alentours d'une anomalie de température nulle, mais nos polynômes de régression sont pour la plupart convexes, n'illustrant pas le fait que de faibles températures en août nuiraient au rendement (ce qui est commenté par Thompson). Nos valeurs de R^2 ne sont pas grandes, ce qui n'est pas étonnant puisque l'on ne regarde qu'un seul prédicteur mensuel et que l'on rassemble bien plus de données que Thompson : la récolte finale dépend de l'ensemble des mois la précédant.

En ce qui concerne les précipitations, l'irrigation semble prépondérante durant les mois de croissance (i.e. post-sevrage) soit de mai à août (Figure III.16). Dans les mois précédents, les précipitations naturelles semblent suffire donc on devrait observer plus de corrélation entre précipitation et rendement à partir de mai.

À l'échelle des états, pour une régression polynomiale de degré 2, entre cumuls de précipitation mensuels et rendements, le mois de juillet reste encore le plus fortement lié. Cette observation est détaillée à la Figure III.25 pour l'état de l'Indiana et de l'Alabama. Pour la plupart des états, on observe que l'information apportée par les précipitations est moins grande que pour les températures. La forme de la tendance (concave ou convexe) varie beaucoup d'un état à l'autre, quelque soit le mois considéré, ce qui montre que ce lien est moins robuste. Les données de précipitation, à elles

seules, semblent donc apporter beaucoup moins d'information que les données de températures.

Ici encore, ces résultats sont à comparer à ceux qu'obtiennent [Thom86, Thom88] : pour les états de la Corn Belt, Thompson a aussi noté une faible corrélation entre rendement et précipitation d'août (polynôme très plat). En revanche pour juillet il observe une droite de forte pente positive qui lie assez fortement les rendements aux anomalies de précipitations. Nous ne retrouvons pas exactement ces résultats dans notre étude.
Les zones sombres (en bleu foncé) ne contiennent aucun point : les réponses du rendement que l'on peut inférer dans ces zones grâce à ces régressions polynomiales ont donc peu de valeur et sont à utiliser avec précaution. Seules les zones à forte densité en points sont en mesure de fournir des résultats robustes de régression polynomiale. Cependant, les informations les plus intéressantes quant à la réponse du rendement concernent les situations hors de la "norme" (où l'anomalie météorologique est proche de zéro). L'analyse des courbes rouges demande donc de faire attention à ne pas surinterpréter l'extrapolation aux anomalies de température et de précipitation trop extrêmes (positivement ou négativement) lorsqu'elles sont peu renseignées (voir Chapitre IV page 107).

Il faut noter que l'on pourrait aussi utiliser des modèles plus complexes de régressions comme les LOESS (régression locale), les splines, les GAM (modèle additif généralisé) mais que ce n'est pas le but de cette pré-analyse : ici l'objectif n'est pas de déterminer avec précision la tendance d'évolution des prédicteurs météorologiques en fonction du rendement, mais bien de déterminer la force du signal entre ces variables, et de pourvoir comparer ces résultats à ceux de [Thom86, Thom88] qui ont eux aussi utilisé une régression quadratique.

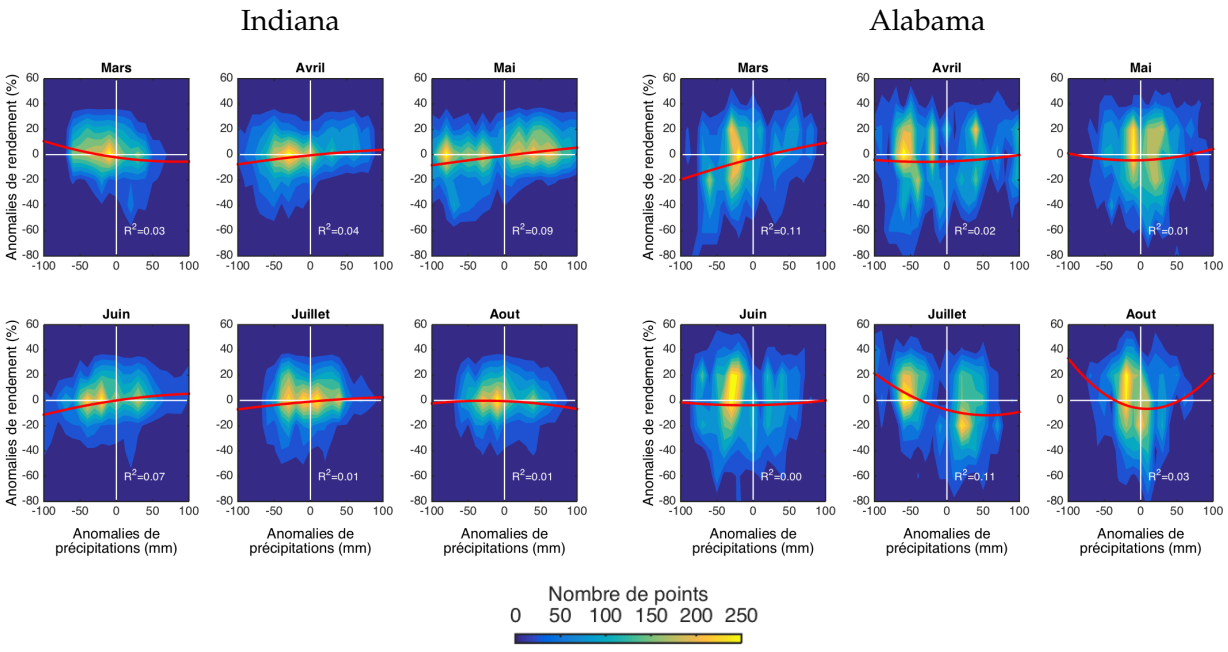


FIGURE III.25 – Anomalies de rendements en fonction d'anomalies de précipitations (exceptionnellement ici, il s'agit de l'écart par rapport à la précipitation moyenne du canton considéré pour une meilleure comparaison avec les articles de Thompson) pour les états de l'Indiana et de l'Alabama : la densité des points est représentée en couleur. Une régression polynomiale d'ordre 2 modélise cette relation en rouge.

En résumé

A travers (1) les informations tirées d'experts agronomes, (2) l'étude des corrélations, et (3) la simple modélisation polynomiale d'ordre 2 univariée, il ressort de cette analyse de sensibilité, que les prédicteurs importants concernant la température sont ceux des mois d'avril, juin, juillet, puis août, et les prédicteurs importants concernant les précipitations sont ceux des mois d'avril, mai, juin, juillet puis aout. On verra par la suite si l'étape de sélection des variables pour les modèles fournit des informations en phase avec cette première analyse de sensibilité très simple.

CHAPITRE IV

Méthodologie

Table des matières

On rappelle que l'objectif de cette thèse est de quantifier l'impact météorologique sur l'agriculture à l'aide de modèles statistiques permettant de prédire des rendements agricoles (ou assimilés) à partir de variables météorologiques observées ou prédites. Il s'agit donc de trouver :

— les bons prédicteurs (variables météorologiques pertinentes),
— le modèle approprié (linéaire, non linéaire, hiérarchique, etc) représentant la relation entre les variables météorologiques (entrées) et les rendements (sortie),
— les paramètres du modèle,
— une quantification des erreurs commises.

Ce chapitre agrège et reprend des résultats de la littérature [West07, Czad16, Wiki16, Hayk94]. Les deux premières sections de ce chapitre décrivent les différents modèles statistiques linéaires qui seront utilisés dans notre étude. La Section 3 est assez théorique et décrit en détail l'inférence statistique pour les modèles mixtes linéaires. Cette partie pourra être négligée en première lecture, et s'adresse aux personnes désirant connaître les détails mathématiques théoriques de l'estimation dans les modèles mixtes linéaires. La Section 4 décrit quant à elle un type de modèle non linéaire qui sera utilisé : les réseaux de neurones. Un tableau comparatif des différents modèles est proposé Section 5. La fin du chapitre présente les différentes étapes nécessaires à la modélisation statistique, comme la recherche d'un modèle de tendance temporelle pour les rendements agricoles de maïs, mais aussi la sélection des variables explicatives, ainsi que les outils et les difficultés liées à la validation/comparaison des modèles résultants.

1 La régression linéaire

Le plus simple des modèles statistiques - mais aussi le plus utilisé - est le modèle de régression linéaire multivariée.

1.1 Le modèle

Dans un modèle linéaire (LIN), la relation entre les observations de rendement y_j (ou assimilés [1]) d'un site fixé lors d'une année j, et les variables météorologiques X_{jk}

1. On verra qu'il s'agit plutôt d'anomalies de rendement.

($k = 1, \ldots, p$, où p est le nombre de variables d'entrée [2]), est formulée de la façon suivante :

$$\begin{aligned} y_j &= \beta_0 + \beta_1\phi_1(X_{j,1}) + \cdots + \beta_p\phi_p(X_{j,p}) + \boldsymbol{\varepsilon}_j \\ & \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad j = 1, \ldots, n. \end{aligned}$$

Théoriquement, les termes $\phi_1, \ldots, \phi_p$ peuvent être des fonctions non linéaires. L'appellation "linéaire" correspond au fait que les paramètres de la régression $(\beta_k)_{k=1,\ldots,p}$ interviennent de façon linéaire dans l'égalité ci-dessus. Dans cette thèse, toutes les fonctions $(\phi_k)_{k=1,\ldots,p}$ sont égales à la fonction identité. Les quantités $(\varepsilon_j)_{j=1,\ldots,n}$ sont des variables aléatoires qui représentent les erreurs faîtes par le modèle, c'est-à-dire la part de la cible y_j qui n'est pas expliquée par les entrées du modèle (qui sont des variables météorologiques ici). Dans cette étude, les n données à prédire proviennent de mesures répétées de plusieurs sites géographiques.

Dans cette thèse, ce type de modèle simple sera comparé à d'autres modèles plus complexes dans le but de prévoir des rendements agricoles. Dans la littérature traitant de ce sujet, les modèles linéaires sont très utilisés.

Le système décrit précédemment peut se réécrire sous la forme simplifiée :

$$Y = X\beta + \varepsilon, \tag{IV.1}$$

où

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^n \qquad X = \begin{pmatrix} 1 & X_{1,1} & \cdots & X_{1,p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n,1} & \cdots & X_{n,p} \end{pmatrix} \in \mathbb{R}^{n\times(p+1)}$$

$$\beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix} \in \mathbb{R}^{p+1} \qquad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} \in \mathbb{R}^n \quad \text{avec} \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n).$$

1.2 L'apprentissage (recherche des paramètres)

Dans le modèle de l'équation (IV.1), le vecteur aléatoire ε suit une loi gaussienne multivariée $\mathcal{N}(0, \sigma^2 I_n)$. L'estimation de β permet d'utiliser le modèle en prédiction mais celle de σ (l'écart-type) s'avère également utile pour l'analyse car elle quantifie la dispersion des erreurs $Y - X\beta$. L'estimateur classique dans le contexte de la régression linéaire est l'estimateur des moindres carrés. C'est aussi l'estimateur du maximum de vraisemblance. Il est obtenu en résolvant le problème de minimisation de l'erreur quadratique :

$$\min_{\beta\in\mathbb{R}^{p+1}} ||Y - X\beta||^2.$$

La solution est la projection orthogonale notée $X\hat{\beta}$ de Y sur l'espace $\{X\beta,\ \beta \in \mathbb{R}^{p+1}\}$. On a donc :

$$(X\beta)^T(Y - X\hat{\beta}) = 0, \quad \forall\, \beta \in \mathbb{R}^{p+1}\ ,$$

ce qui équivaut à

$$X^T(Y - X\hat{\beta}) = 0\ ,$$

2. Appelées aussi "prédicteurs" dans la suite du manuscrit.

ou encore

$$X^T Y = X^T X \hat{\beta}.$$

On en déduit alors l'estimateur des moindres carrés de β et celui de σ^2 par substitution [3]

$$\boxed{\hat{\beta} = \left(X^T X\right)^{-1} X^T Y.}$$

On dispose alors d'un estimateur biaisé $(\widehat{\sigma}_1^2)$ et non-biaisé $(\widehat{\sigma}_2^2)$ de σ^2 :

$$\begin{aligned} \widehat{\sigma}_1^2 &= \frac{1}{n-1}\sum_{j=1}^{n}(y_j - \overline{y})^2, \\ \widehat{\sigma}_2^2 &= \frac{1}{n-p-1}\sum_{j=1}^{n}\widehat{\varepsilon}_j^2 \quad \text{où} \quad \widehat{\varepsilon} = Y - X\widehat{\beta}. \end{aligned}$$

2 Modèles traitant les données groupées

La présence de groupes au sein d'observations est une configuration fréquente. Du point de vue de la modélisation, on distingue le cas où les prédicteurs sont des variables catégorielles (alors appelées "facteurs"), du cas où les prédicteurs sont des variables continues. Les Sous-sections 2.1 et 2.2 traitent de la situation où les groupes sont formés par des variables catégorielles qui sont aussi des prédicteurs : même si ce genre de situation ne sera pas rencontré dans cette thèse, il est intéressant de cerner les points communs et les différences avec notre situation. La Sous-section 2.3 et la suite, traiteront des situations où les prédicteurs sont des variables continues, et où les groupes sont des subdivisions internes constituantes des observations mais ne proviennent pas d'un facteur qui servirait en tant que prédicteur.

Comme pour beaucoup de modèles de régression, l'apprentissage (phase d'estimation des paramètres) peut se faire en utilisant l'ensemble des données, c'est-à-dire en rassemblant les données de tous les groupes (méthode appelée "pooling"), ou bien groupe par groupe, indépendamment (méthode appelé "no-pooling"). L'avantage du pooling est la disponibilité d'une grande quantité de données pour l'estimation des paramètres du modèle. En revanche, ce regroupement implique le mélange de groupes aux comportements différents. À l'inverse, le no-pooling est bien spécialisé à chaque groupe, mais peu de données sont disponibles pour l'estimation des paramètres. On verra que cela pose des problèmes lors de la phase d'apprentissage ainsi que sur la capacité du modèle à être vraiment utilisable en pratique (sur-apprentissage, voir Section 6.1).

2.1 Modèles à effets fixes pour une variable catégorielle

Supposons ici que l'on dispose de n observations $y \in \mathbb{R}^n$ et d'un facteur classant ces n observations en m groupes (par exemple, le facteur est le "type de sol" et présente m=3 niveaux : luvisol, phernosol et podzosol). On peut souhaiter connaître la diversité des groupes selon ce facteur, i.e. l'effet de ce facteur sur les observations y. Pour ce faire,

3. Si la matrice $X^T X$ n'est pas inversible, le problème n'a pas une unique solution et on considère alors la pseudo-inverse.

une modélisation habituelle pour relier ces n observations réparties en m groupes, avec n_i données par groupe ($i \in [\![1, m]\!]$) est

$$y_{ij} = \beta_i + e_{ij} = \mu + \alpha_i + e_{ij} \quad \text{avec} \quad \mu = \frac{1}{m}\sum_{i=1}^{m} \beta_i \quad et \quad \alpha_i = \beta_i - \mu \,, \tag{IV.2}$$

où y_{ij} est la $j^{\text{ème}}$ observation dans le groupe i, pour $i = 1, ..., m$ et $j = 1, ..., n_i$. Le μ dans l'équation (IV.2) représente la moyenne générale, α_i étant l'effet d'appartenir au groupe i. Les erreurs résiduelles e_{ij} sont considérées comme des variables aléatoires : dans ce modèle, elles ont une moyenne nulle ($\mathbb{E}(e_{ij}) = 0$), une variance homogène ($\text{var}(e_{ij}) = \sigma_e^2$ pour tout i et j), et sans covariance ($\text{cov}(e_{ij}; e_{i'j'}) = 0$ pour $i \neq i'$ et $j \neq j'$). Dans ce contexte, on regarde μ et les α_i comme des constantes fixes et inconnues, appelées effets fixes. C'est l'expression traditionnelle d'un modèle d'analyse de variance, et puisque seuls les e_{ij} dans l'équation (IV.2) sont des termes aléatoires, ce modèle est appelé **modèle à effets fixes**. On parle aussi de modèle standard d'analyse de variance (ANOVA à un facteur). On dispose alors de l'estimateur suivant de σ_e^2 :

$$\widehat{\sigma}_e^2 = \frac{1}{\left(\sum_{i=1}^{m} n_i\right) - m} \sum_{i=1}^{m}\sum_{j=1}^{n_i} (y_{ij} - \overline{y}_{i.})^2 \,, \qquad \text{où} \quad \overline{y}_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}.$$

L'estimateur de μ est donné par

$$\widehat{\mu} = \frac{1}{m}\sum_{i=1}^{m} \overline{y}_{i.} \,.$$

On remarque que si les groupes ne sont pas de même taille (on parle de "plan non équilibré") alors $\widehat{\mu} \neq \overline{y}_{..}$ où $\overline{y}_{..} = \frac{1}{n}\sum_{i,j} y_{ij}$. On en déduit un estimateur de α_i :

$$\widehat{\alpha}_i = \overline{y}_{i.} - \widehat{\mu}.$$

2.2 Modèles à effets aléatoires pour une variable catégorielle A : une volonté de séparer les effets de variance, des effets de moyenne

En statistique, le terme "effets" fait référence à la réponse du modèle lors de la perturbation d'une entrée. Les effets sont dits "fixes" s'ils sont constants pour l'ensemble de la population. L'estimation de ces effets est l'objectif traditionnel des modèles de régression.

Par comparaison, les effets aléatoires sont des variables aléatoires dépendantes des échantillons des données, c'est-à-dire que les effets varient d'un groupe à un autre. Un effet aléatoire est représenté dans le modèle sous la forme d'un terme additionnel dont la valeur dépend d'une densité de probabilité qui est estimée à partir des données. La densité de probabilité cherche à décrire les différences entre les groupes. Les effets aléatoires sont utiles quand les données tombent dans des groupes naturels, comme des catégories géographiques, des types de sols, ou encore des types d'utilisation du sol [Math15].

Dans la section précédente, nous considérions les α_i comme des constantes (c'est-à-dire des effets fixes) découlant du fait que les m groupes à partir desquels les données proviennent sont un ensemble de classes particulières qui ont été spécifiquement choisies pour l'étude. Par exemple, ils pourraient être des engrais dans une expérience agricole, ou des médicaments dans un essai clinique.
Contrairement à cela, il existe des situations dans lesquelles les classes n'ont pas été spécifiquement choisies mais peuvent être considérées comme un échantillon aléatoire d'une certaine population de classes. Les α_i sont alors des variables aléatoires : les α_i sont des valeurs réalisées de variables aléatoires non observables. Ils sont appelés **effets aléatoires**. L'équation du modèle est la même que dans le cas précédent (IV.2), sauf que les α_i sont différents. On suppose qu'ils sont aléatoires, de moyenne nulle, de variance uniforme, et sans covariance entre eux ni avec les termes d'erreurs.

$$y_{ij} = \mu + \alpha_i + e_{ij} \quad \text{avec} \quad \mathbb{E}(\alpha_i) = 0 \quad et \quad var(\alpha_i) = \sigma_\alpha^2 \quad \forall\, i,$$

$$cov(\alpha_i, \alpha_{i'}) = 0 \quad \forall\, i \neq i' \quad et \quad cov(\alpha_i, e_{i'j}) = 0 \quad \forall\, i,\ i' \ et\ j.$$

Ainsi, μ est le seul effet fixe. Les α_i ne sont pas des paramètres, et seule leur variance commune σ_α^2 est un paramètre. On appelle un tel modèle un **modèle à effets aléatoires**. En ce qui concerne les effets aléatoires, l'une des principales caractéristiques d'intérêt est leur variance σ_α^2.

À ce stade, pour les données catégorielles il faut souligner une dichotomie très importante des données : équilibrées ou non-équilibrées, i.e. que toutes les sous-classes ont le même nombre de données ou pas. On dit souvent que les données équilibrées forment un "plan" équilibré. Les estimations statistiques seront beaucoup plus facile dans le cas des plans équilibrés.

Il existe différentes méthodes d'estimation des composantes de la variance : estimation par maximum de vraisemblance, par maximum de vraisemblance restreint, ou encore par méthode des moments. C'est la méthode de type moments que nous développons ici. Elle consiste à écrire un système d'équations en égalant moments empiriques et moments théoriques (d'ordre 1) de certaines sommes de carrés. La solution de ce système fournit des valeurs pour les composantes de la variance et ces valeurs sont prises comme estimations de ces composantes.

Considérons donc les deux sommes de carrés définies ci-dessous.
► SS_A est la somme des carrés des écarts pris en compte par le modèle, c'est-à-dire par le facteur aléatoire A. On a :

$$SS_A = \sum_{i=1}^{m} n_i(\overline{y}_{i.} - \overline{y}_{..})^2 \underbrace{=}_{\text{si plan équilibré}} n_0 \sum_{i=1}^{m} (\overline{y}_{.i} - \overline{y}_{..})^2 \,.$$

Le calcul de l'espérance de SS_A, simple dans le cas équilibré, conduit à

$$\mathbb{E}[SS_A] = (m-1)\sigma_e^2 + \frac{1}{n}\left(n^2 - \sum_{i=1}^{m} n_i^2\right)\sigma_\alpha^2 \underbrace{=}_{\text{si plan équilibré}} (m-1)\left(\sigma_e^2 + n_0\sigma_\alpha^2\right).$$

Par ailleurs, les solutions du système que l'on va écrire feront intervenir le carré moyen relatif au facteur A, qui sera noté MS_A : $MS_A := \frac{1}{m-1} SS_A$ (SS_A est une quantité à $m-1$

degrés de liberté). Après calculs, on obtient :

$$\mathbb{E}[MS_A] = \sigma_e^2 + \frac{1}{n(m-1)}\left(n^2 - \sum_{i=1}^{m} n_i^2\right)\sigma_\alpha^2 \underbrace{=}_{\text{si plan équilibré}} \sigma_e^2 + n_0\sigma_\alpha^2.$$

► SSE est la somme des carrés des écarts résiduels (ceux associés aux erreurs du modèle, c'est-à-dire non pris en compte par le modèle) :

$$SSE = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(y_{ij} - \overline{y}_{i.}\right)^2.$$

L'espérance de SSE s'écrit :

$$\mathbb{E}[SSE] = (n-m)\sigma_e^2 .$$

On utilisera encore le carré moyen relatif aux erreurs, noté MSE : $MSE = \frac{SSE}{n-m}$ (SSE est à $n-m$ degrés de liberté). Son espérance vaut $\mathbb{E}(MSE) = \sigma_e^2$.
Par estimer σ_e et σ_α par la méthode des moments, on résout le système suivant (d'inconnues σ_e et σ_α) qui égale les espérances de MS_A et de MSE avec les valeurs empiriques de ces 2 carrés moyens (respectivement notées $MS_A(y)$ et $MSE(y)$) :

$$\begin{cases} MS_A(y) &= \sigma_e^2 + \frac{1}{n(m-1)}\left(n^2 - \sum_{i=1}^{m} n_i^2\right)\sigma_\alpha^2 \underbrace{=}_{\text{si plan équilibré}} \sigma_e^2 + n_0\sigma_\alpha^2 \\ MSE(y) &= \sigma_e^2. \end{cases}$$

Sa résolution fournit les estimations des composantes de la variance :

$$\begin{cases} \widehat{\sigma}_e^2 &= MSE(y) \\ \widehat{\sigma}_\alpha^2 &= \frac{n(m-1)(MS_A(y)-MSE(y))}{\left(n^2-\sum_{i=1}^{m} n_i^2\right)} \underbrace{=}_{\text{si plan équilibré}} \frac{MS_A(y)-MSE(y)}{n_0} \end{cases}$$

En pratique, on n'utilisera pas de tels modèles dans cette thèse mais plutôt leur généralisation directe : les modèles à effets mixtes.

2.3 Les modèles à effets mixtes [4]

Introduction

L'utilisation d'un modèle linéaire classique différent sur chaque groupe peut s'avérer dangereux pour les groupes dont on dispose de peu de données. Pour résoudre le problème de manque de données et obtenir tout de même des modèles performants, on peut agir sur la structure du modèle de l'équation (IV.1). C'est le cas des modèles à effets mixtes (ME).

Les modèles à effets mixtes sont basés sur l'idée que chaque donnée appartient à un groupe particulier, et que le modèle ME prend en compte les particularités de chaque groupe : souvent un comportement général pour le modèle d'impact existe mais il doit être ajusté d'un groupe à l'autre. Les modèles ME sont bien adaptés à ce type de comportement ; ils sont un cas particulier des modèles hiérarchiques bayésiens [Gelm03, Pinh09]. La calibration du modèle utilise toutes les données disponibles, cependant son comportement va être différent d'un groupe à l'autre. L'utilisation d'un

4. On traitera le cas des variables continues même si le cas des variables catégorielles existe aussi.

modèle ME permet de réduire le nombre de paramètres dans le modèle avec un effet de régularisation [5] qui augmente la qualité générale des modèles d'impact. Malgré leur pertinence pour spatialiser, les modèles ME sont rarement utilisés [Yang08]. On trouvera un exemple d'application dans [Aire12] où ils sont utilisés pour ajuster aux régions Britanniques un modèle de prédiction de consommation de sel de dessalage de route pour les hivers rigoureux.

Les modèles à effets mixtes traitent autant les effets fixes qu'aléatoires. L'ajout d'effets aléatoires améliore la fiabilité des estimations par apport à une estimation séparée de chaque groupe. Leur particularité (et leur force) résident dans la possibilité d'utiliser des corrélations entre sous-groupes de l'échantillon. En rajoutant des informations issues de classification par groupe, les modèles mixtes proposent un compromis entre

— ignorer les groupes, et rassembler toutes les données disponibles au sein d'un seul "grand ensemble" utilisé pour calibrer le modèle général,

— et réaliser un apprentissage différent pour chaque groupe avec des modèles différents.

En vérité, bien sûr, tous les modèles qui ont un μ et des termes d'erreurs sont un mélange d'effets fixes, μ, et de termes aléatoires, les erreurs. Mais le nom de "modèle mixte" est réservé aux modèles qui ont un mélange d'effets fixes et aléatoires autres que μ et les termes d'erreurs.

Les modèles mixtes sont utilisés lorsqu'on a affaire à des données groupées ou bien des données longitudinales. En agriculture, ces deux types de données sont très fréquents. Pour les données groupées, la réponse est mesurée pour chaque site (ou sujet) et chaque site appartient à un groupe de sites (par exemple, le poids de naissance des rats regroupés par litière). Pour les données longitudinales, la réponse est mesurée à plusieurs moments et le nombre de points temporels n'est pas grand (contrairement aux séries temporelles) : par exemple, la vente mensuelle d'un produit (12 mesures).

La plupart des expériences agricoles ont un mélange d'effets fixes et aléatoires : par exemple, des traitements (types d'engrais ou d'herbicides) sont fixes, mais la localité géographique ou les années peuvent être aléatoires. Les procédures ANOVA ou GLM (modèles linéaires généralisés) sont utilisées, mais elles sont des modèles à effets fixes (GLM peut donner un mauvais dénominateur pour le test F de Fisher ou une mauvaise estimation de l'erreur d'un traitement). GLM suppose de plus que tous les traitements ont des erreurs identiques et que ces erreurs ne sont pas corrélées (hypothèse iid).

La modélisation mixte peut aussi traiter l'analyse des données provenant de structures à plusieurs niveaux, ce qui est très fréquent en agriculture.

Modèles à effets mixtes, linéaires

Commençons par une description un peu formelle d'un modèle linéaire à effets mixtes. Un exemple simple relatif à notre application est donné juste après.

On considère un échantillon pouvant se répartir en m groupes distincts. Comme précédemment, on note y_{ij} la $j^{\text{ème}}$ observation du groupe i que l'on cherche à prédire. On note aussi x_{ij} la $j^{\text{ème}}$ valeur dans le groupe i d'une variable prédictive x (on ne considère qu'une seule variable prédictive dans cette première approche). Un modèle

5. voir Section 6.2

linéaire mixte qui présente une ordonnée à l'origine dépendant de chaque groupe, ainsi qu'une variable explicative continue, s'écrit sous la forme

$$y_{ij} = \beta_{0i} + \beta_1 x_{ij} + \varepsilon_{ij}, \qquad j = 1, .., n_i, \quad i = 1, ..., m,$$

$\varepsilon_{ij} \underset{iid}{\sim} \mathcal{N}(0, \sigma^2)$,
$\beta_{0i} = \beta_{00} + b_{0i}, \quad b_{0i} \underset{iid}{\sim} N(0, \sigma_0^2)$,
où b_{0i} et ε_{ij} sont indépendants les uns des autres. On peut regrouper les paramètres dépendants du groupe et ceux qui n'en dépendent pas. On obtient

$$y_{ij} = \underbrace{\beta_{00} + \beta_1 x_{ij}}_{\text{effets fixes}} + \underbrace{b_{0i}}_{\text{effets aléatoires}} + \varepsilon_{ij}.$$

Le paramètre relatif à la variable explicative x peut aussi dépendre du groupe et donc posséder un effet aléatoire. On obtient :

$$y_{ij} = \beta_{0i} + \beta_{1i} x_{ij} + \varepsilon_{ij}, \quad j = 1, ..., n_i, \quad i = 1, ..., m,$$

$\varepsilon_{ij} \underset{iid}{\sim} N(0, \sigma^2)$,
$\beta_{0i} = \beta_{00} + b_{0i}, \quad b_{0i} \underset{iid}{\sim} N(0, \sigma_0^2)$,
$\beta_{1i} = \beta_{10} + b_{1i}, \quad b_{1i} \underset{iid}{\sim} N(0, \sigma_1^2)$,
ou bien $\quad b_i = \begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix} \underset{iid}{\sim} \mathcal{N}\left(0, \begin{pmatrix} \sigma_0^2 & 0 \\ 0 & \sigma_1^2 \end{pmatrix}\right)$.
À ce stade, les effets aléatoires ne sont pas corrélés mais s'il apparait pertinent qu'ils le soient, on supposera que les effets aléatoires b_i vérifient $b_i = \begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix} \underset{iid}{\sim} \mathcal{N}(0, \Psi(\theta))$ où Ψ est une matrice symétrique 2×2 semi-définie positive, paramétrée par un vecteur de variance θ.
On peut encore regrouper les paramètres dépendants du groupe et ceux qui n'en dépendent pas. On obtient :

$$y_{ij} = \underbrace{\beta_{00} + \beta_{10} x_{ij}}_{\text{effets fixes}} + \underbrace{b_{0i} + b_{1i} x_{ij}}_{\text{effets aléatoires}} + \varepsilon_{ij}, \quad j = 1, ..., n_i, \quad i = 1, ..., m.$$

Le modèle proposé ici est très simple car ne propose qu'une seule variable explicative. En général, on dispose de plusieurs variables explicatives et il n'est pas obligé d'attribuer à chacune d'elle un effet fixe et un effet aléatoire. Aussi, il est commun de distinguer par leur notation, les variables explicatives qui disposent d'un effet fixe seulement (elles seront regroupées dans une matrice notée X, appelée matrice de design pour les effets fixes), et celles qui disposent aussi d'un effet aléatoire (regroupées dans la matrice notée Z, appelée matrice de design des effets aléatoires).

Le modèle exprimant la réponse y peut donc s'écrire

$$y_{ij} = \begin{bmatrix} 1 & x_{ij} \end{bmatrix} \begin{pmatrix} \beta_{00} \\ \beta_{10} \end{pmatrix} + \begin{bmatrix} 1 & x_{ij} \end{bmatrix} \begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix} + \varepsilon_{ij}, \qquad j = 1, ..., n_i, \quad i = 1, ..., m,$$

ce qui correspond à la formalisation usuelle :

$$y = X\beta + Zb + \varepsilon \ ,$$

où X est la matrice de design pour les effets fixes, et Z la matrice de design pour les effets aléatoires [Gelm07, Pinh09]. ε représente le bruit, ou la part non expliquée de la réponse par les variables d'entrée [Hox02, Hari08]. β est le vecteur des effets fixes qui sont constants pour l'ensemble des sites [Lins88].

Voici un exemple simple sur notre application à l'agriculture :
On souhaite réaliser la régression linéaire des anomalies de rendements y contre deux entrées météorologiques (Tjuillet et Pjuillet) avec un modèle à effets mixtes. Après analyse [6], il apparait inutile de considérer un effet aléatoire pour l'ordonnée à l'origine. En revanche, il semble pertinent d'accorder un effet aléatoire aux deux variables explicatives Tjuillet et Pjuillet. Le modèle s'écrit donc,
pour $i = 1, ..., 3000$, et $j = 1979, ..., 2013$,

$$y_{ij} = \beta_0 + \beta_{1i} T_{juillet}(i,j) + \beta_{2i} P_{juillet}(i,j) + \varepsilon_{ij} \ ,$$

avec
y_{ij} l'anomalie de rendement du canton i à l'année j
$\varepsilon_{ij} \underset{iid}{\sim} \mathcal{N}\left(0, \sigma^2\right)$
$\beta_{1i} = \beta_{Tjuillet} + \underbrace{b_{Tjuillet,\,i}}_{\underset{iid}{\sim}\mathcal{N}\left(0,\sigma^2_{T_{juillet}}\right)}$ $\qquad$ $\beta_{2i} = \beta_{Pjuillet} + \underbrace{b_{Pjuillet,\,i}}_{\underset{iid}{\sim}\mathcal{N}\left(0,\sigma^2_{P_{juillet}}\right)}$
d'où $b_m = \begin{pmatrix} b_{Tjuillet,\,i} \\ b_{Pjuillet,\,i} \end{pmatrix} \underset{iid}{\sim} \mathcal{N}\left(0, \begin{pmatrix} \sigma^2_{T_{juillet}} & 0 \\ 0 & \sigma^2_{P_{juillet}} \end{pmatrix}\right)$, si on choisi des effets aléatoires non corrélés.
Ce modèle demande donc l'estimation de 6 paramètres : β_0, $\beta_{Tjuillet}$, $\beta_{Pjuillet}$, σ^2, $\sigma^2_{Tjuillet}$, et $\sigma^2_{Tjuillet}$. [7]

Comme dans la suite de ce manuscrit, l'analyse des autocorrélations temporelles des résidus d'un même canton a montré qu'elles étaient négligeables. Les corrélogrammes (partiels ou non) ne montrent aucune saisonnalité.

Modèles à effets mixtes, non linéaires

Il peut s'avérer utile de modéliser une relation prédicteurs/cible par une fonction non linéaire. Les modèles mixtes non linéaires sont donc une extension des modèles mixtes linéaires. On suppose encore ici que les données utilisées pour calibrer le modèle peuvent être regroupées dans m groupes distincts $i = 1, \ldots, m$. Pour prendre en compte les groupes dans le modèle d'impact, la réponse pour l'échantillon j du groupe i s'écrit :

$$\begin{aligned} y_{ij} &= f(\phi_i, X_{ij}) + g(\phi_i, X_{ij})\varepsilon_{ij} \\ &\qquad \text{pour} \qquad j = 1, ..., n_i, \quad i = 1, ..., m, \end{aligned} \tag{IV.3}$$

6. [Khur98] proposent un livre très complet sur les différents tests possibles pour les modèles mixtes. Pour savoir si les effets fixes contribuent à expliquer la variable dépendante, on peut par exemple analyser les tableaux de résultats des tests de type III des effets aléatoires, qui permettent de vérifier leur significativité en utilisant un test F de Fisher. Comme dans le modèle linéaire classique, l'étude des écart-types ainsi que des intervalles de confiance servent aussi à diagnostiquer leur significativité.

7. Il faut remarquer que le nombre de groupes m dans la "population" n'intervient pas ici, car aucun "vrai" paramètre ne dépend du groupe considéré.

où le scalaire y_{ij} est la cible du modèle (rendements de maïs), X_{ij} est le vecteur des variables d'entrées [8] (variables météorologiques dans notre cas), et ϕ est le vecteur des paramètres du modèle. On suppose que les erreurs ε_{ij} sont indépendantes et identiquement distribuées avec $\mathbb{E}[\varepsilon_{ij}] = 0$ et $\text{var}(\varepsilon_{ij}) = 1$. Souvent, on suppose que les ε_{ij} suivent une loi normale. La fonction f spécifie la forme du modèle c'est-à-dire le modèle structurel, et la fonction g désigne le modèle d'erreur (modèle d'erreur résiduelle, souvent prise comme constante), c'est-à-dire la variabilité intra-groupe [Lins90].

Dans l'équation (IV.3), les paramètres ϕ varient d'un groupe à l'autre et les $\boldsymbol{\phi}_i$ sont une combinaison d'effets fixes et d'effets aléatoires :

$$\phi_i = F\beta + Rb_i,$$

où les effets aléatoires b_i suivent souvent une loi normale multivariée, de moyenne nulle et de matrice de covariance Ψ. La matrice de covariance Ψ des effets aléatoires décrit la variabilité inter-groupe. Le premier rôle de la matrice Ψ est d'exprimer le taux de ressemblance entre les groupes. Quand les coefficients de Ψ sont faibles, tous les sites se comportent de façon similaire, et quand ils sont grands, chaque groupe se comporte indépendamment des autres.

Un effet aléatoire a une unique variance pour tous les groupes : si cette variance est grande, cela signifie que la variable dont elle quantifie l'effet prend des valeurs bien distinctes d'un groupe à l'autre. Si elle est faible, c'est le contraire : les groupes se ressemblent quand on regarde la variable. Il faut remarquer que les effets aléatoires b_i ne sont pas simulés (c'est un abus de langage, lorsque ce terme est employé) mais calculés avec une relation matricielle déterministe une fois que tous les paramètres du modèle sont estimés : lorsque les "vrais" paramètres du modèles sont estimés, on peut alors donner une valeur aux effets aléatoires (qui sont des variables dites latentes ou cachées) : cette valeur utilise l'estimation de leur variance mais aussi les données groupe à groupe pour proposer une valeur calculée matriciellement et non tirée selon une loi de probabilité. S'il en était autrement, on ne pourrait pas s'assurer que la valeur d'une $\mathcal{N}(0, \sigma_b^2)$ donne une valeur qui s'adapte à chaque groupe, avec une seule variance pour tous les groupes. En réalité, pour chaque prédicteur possédant un effet aléatoire, on calcule autant d'effets aléatoires qu'il y a de groupes, et leurs valeurs n'ont rien d'aléatoires [9].

L'estimation des effets fixes β et la covariance des effets aléatoires Ψ fournit une description de la "population" qui ne suppose pas que les paramètres ϕ_i soient constants d'un groupe à l'autre. Le calcul des effets aléatoires b_i donne une description des spécificités de chaque groupe (variabilité inter-groupes). Les matrices F et R - appelées ici aussi "matrice de design" - expriment les paramètres comme une combinaison linaire des effets fixes et des effets aléatoires respectivement. Dans cette thèse, les modèles mixtes non linéaires utilisent des matrices de design égales à la matrice identité.

Si f est une application linéaire des paramètres $\boldsymbol{\phi}_i$, on obtient un modèle à effets mixtes linéaire [Lins88]. Dans cette thèse, on a utilisé un modèle mixte *non-linéaire* (avec une fonction logistique) pour identifier la tendance temporelle des rendements

8. Cette notation ne doit pas être confondue avec la notation matricielle utilisée page 76. On insiste sur le fait qu'une virgule est utilisée pour désigner le coefficient d'une matrice. Ainsi la matrice de design de la page 76 est $X = (X_{i,k})_{i=1\ldots n, k=1\ldots p} \in \mathbb{R}^{n\times p}$.

9. Ceci était aussi le cas dans les modèles précédents qui possédaient des effets aléatoires.

(Section 2), et plusieurs modèles mixtes *linéaires* pour l'estimation des anomalies de rendement (Chapitre V).

Les modèles mixtes linéaires pour l'estimation des anomalies de rendement sont construits avec des effets fixes et des effets aléatoires non corrélés pour chaque variable d'entrée. Les groupes utilisés sont des catégories administratives et géographiques (État fédéraux, ou districts agricoles, ou cantons), comme décrit à la fin du Chapitre III.

Remarque sur le modèle d'erreur résiduel :
Le choix du modèle d'erreur résiduel g est très libre est permet de rendre compte des différentes hypothèses faites sur la distribution des erreurs. Notons $f_i = f(\phi_i, X_i)$. Les modèles d'erreurs les plus courants sont :

— le modèle d'erreur constant : $g = a$ d'où $y_i = f_i + a\varepsilon_i$

— le modèle d'erreur proportionnel : $g = b.f$ d'où $y_i = f_i + b.f_i\varepsilon_i$

— le modèle d'erreur mixte : $g = a + b.f$ d'où $y_i = f_i + (a + b.f_i)\varepsilon_i$

— le modèle d'erreur exponentiel : ici le modèle s'écrit $\log(y_i) = \log(f_i) + a\varepsilon_i$ donc $g = a$ et $y_i = f_i e^{a\varepsilon_i}$.

Exemple de régression non linéaire avec un modèle à effets mixtes :
On se propose de modéliser l'évolution des rendements agricoles de maïs en fonction du temps, pour les différents cantons des États-Unis. On choisit d'utiliser une fonction logistique pour relier les années j, au rendement du canton i pour l'année j, y_{ij}. Le modèle s'écrit donc :

$$y_{ij} = C_i + \frac{K_i}{1 - P_i e^{-R_i j}} + \varepsilon_{ij} \qquad j = 1979, ..., 2013, \quad i = 1, ..., 3000$$

avec
$\varepsilon_{ij} \underset{iid}{\sim} \mathcal{N}(0, \sigma^2)$
$C_i = C_0 + c_i \quad et \quad c_i \underset{iid}{\sim} \mathcal{N}(0, \sigma_c^2)$
$K_i = K_0 + k_i \quad et \quad k_i \underset{iid}{\sim} \mathcal{N}(0, \sigma_k^2)$
$P_i = P_0 + p_i \quad et \quad p_i \underset{iid}{\sim} \mathcal{N}(0, \sigma_p^2)$
$R_i = R_0 + r_i \quad et \quad r_i \underset{iid}{\sim} \mathcal{N}(0, \sigma_r^2).$

On commence donc par attribuer à chaque paramètre un effet aléatoire (mais on verra que cela n'est peut être pas nécessaire) en supposant des effets aléatoires non corrélés. L'analyse de la matrice de covariance (estimée) des effets aléatoires nous montre une variance négligeable pour l'effet aléatoire de R. Il vaut donc mieux l'enlever du modèle pour limiter sa complexité. La log-vraisemblance variant peu, et les critères AIC ou BIC étant plus faibles, cela renforce la décision de supprimer l'effet aléatoire du paramètre R.

Le modèle s'écrit donc :

$$y_{ij} = C_i + \frac{K_i}{1 - P_i e^{-Rj}} + \varepsilon_{ij} \qquad j = 1979, ..., 2013, \quad i = 1, ..., 3000$$

avec
$\varepsilon_{ij} \underset{iid}{\sim} \mathcal{N}(0, \sigma^2)$

$C_i = C_0 + c_i \quad et \quad c_i \underset{iid}{\sim} \mathcal{N}(0, \sigma_c^2)$
$K_i = K_0 + k_i \quad et \quad k_i \underset{iid}{\sim} \mathcal{N}(0, \sigma_k^2)$
$P_i = P_0 + p_i \quad et \quad p_i \underset{iid}{\sim} \mathcal{N}(0, \sigma_p^2)$

De plus, l'hypothèse d'une matrice de covariance des effets aléatoires diagonale doit être vérifiée. Pour cela on transforme la matrice de covariance estimée en une matrice de corrélation. On peut alors connaître la corrélation que partagent les effets aléatoires. On décide alors pour quels effets aléatoires il est pertinent d'imposer une corrélation non nulle : l'analyse montre qu'il vaut mieux imposer une corrélation nulle entre l'effet aléatoire de K et celui de P. Il y a donc 5 paramètres à estimer pour les effets aléatoires, 4 pour les effets fixes, et 1 pour le terme d'erreur, soit 10 paramètres en tout dans le modèle.

Les résultats de ce modèle de tendance sont décrits en détails à la Section 7.1.

3 Inférence statistique pour les modèles mixtes linéaires

On se concentre ici sur l'inférence statistique pour les modèles mixtes linéaires à un niveau de groupe, car se seront eux qui serviront par la suite. Le but de cette section est de présenter leurs aspects théoriques et mathématiques qui permettent l'estimation des paramètres, mais aussi leur évaluation et leurs limites. [Pinh09] ainsi que [West07] sont de bons livres sur le sujet.

On observe donc n données longitudinales y, provenant de m groupes distincts. Même si on verra qu'aucun "vrai" paramètre du modèle ne dépend des groupes, il est d'usage de séparer les observations et les prédicteurs selon les groupes dans les écritures vectorielles ou matricielles. Le modèle linéaire à effets mixte exprime le vecteur réponse $y_{i.} = (y_{ij})_j$ de dimension n_i, pour le groupe i, par

$$\begin{gathered} y_{i.} = X_{i.}\beta + Z_{i.}b_i + \varepsilon_{i.}, \quad i = 1, ..., m, \\ b_i \sim \mathcal{N}_q(0, \Psi), \quad \varepsilon_{i.} \sim \mathcal{N}_{n_i}(0, \sigma^2 I), \end{gathered}$$

où β est le vecteur des effets fixes de dimension p, b_i est désormais le vecteur des effets aléatoires de dimension q, $X_{i.}$ (de taille $n_i \times p$) et $Z_{i.}$ (de taille $n_i \times q$) sont les matrices de régression connues pour les effets fixes et aléatoires respectivement. Les colonnes de $Z_{i.}$ sont généralement un sous-ensemble des colonnes de $X_{i.}$. $\varepsilon_{i.}$ est le vecteur d'erreur de dimension n_i, de distribution normale. L'hypothèse $Var(\varepsilon_i) = \sigma^2 I$ peut être assouplie et généralisée. On suppose que les effets aléatoires b_i et les erreurs intra-groupes $\varepsilon_{i.}$ sont indépendants entre les groupes mais aussi au sein d'un même groupe.

Puisque la distribution des effets aléatoires $\boldsymbol{b_i}$ est supposée gaussienne de moyenne nulle, elle est complètement caractérisée par sa matrice de variance-covariance Ψ. Cette matrice doit être symétrique, positive semi-définie ; c'est-à-dire que toutes ses valeurs propres doivent être positives ou nulles. Souvent on fait l'hypothèse plus forte que toutes les valeurs propres sont strictement positives. Cette restriction est possible car un modèle indéfini peut toujours être ré-exprimé en tant que modèle défini-positif de dimension inférieure.

3.1 Formulation matricielle du modèle mixte linéaire

Rappelons que $n = \sum_{i=1}^{m} n_i$, p le nombre d'effets fixes [10] et q le nombre d'effets aléatoires. On pose ensuite :

$$y = \begin{pmatrix} y_{1.} \\ \vdots \\ y_{m.} \end{pmatrix} \in \mathbb{R}^n, \qquad X = \begin{pmatrix} X_{1.} \\ \vdots \\ X_{m.} \end{pmatrix} \in \mathbb{R}^{n \times p} \qquad \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \in \mathbb{R}^p,$$

$$Z = \begin{pmatrix} Z_{1.} & 0_{n_1 \times q} & \cdots & 0_{n_1 \times q} \\ 0_{n_2 \times q} & Z_{2.} & & \\ \vdots & & \ddots & \\ 0_{n_m \times q} & & & Z_{m.} \end{pmatrix} \in \mathbb{R}^{n \times (mq)}, \qquad b = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix} \in \mathbb{R}^{mq},$$

$$\varepsilon = \begin{pmatrix} \varepsilon_{1.} \\ \vdots \\ \varepsilon_{m.} \end{pmatrix} \in \mathbb{R}^n \qquad G = \begin{pmatrix} \Psi & & \\ & \ddots & \\ & & \Psi \end{pmatrix} \in \mathbb{R}^{mq \times mq}$$

$$R = \begin{pmatrix} R_1 & & 0 \\ & \ddots & \\ 0 & & R_m \end{pmatrix} \underbrace{=}_{souvent} \sigma^2 \begin{pmatrix} I_{n_1} & & 0 \\ & \ddots & \\ 0 & & I_{n_m} \end{pmatrix} \in \mathbb{R}^{n \times n}.$$

Par la suite, on se place dans le cas classique où $R = \sigma^2 I_n$. La formulation matricielle du modèle mixte linéaire s'écrit

$$y = X\beta + Zb + \varepsilon \,,$$

$$\text{avec} \quad \begin{pmatrix} b \\ \varepsilon \end{pmatrix} \sim \mathcal{N}_{mq+n} \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix} ; \begin{pmatrix} G & 0_{mq \times n} \\ 0_{mq \times n} & R \end{pmatrix} \right].$$

ce qui s'exprime sous forme d'un modèle hiérarchique à deux niveaux :

$$\left.\begin{array}{rcl} y|b & \sim & \mathcal{N}_n(X\beta + Zb, R) \\ b & \sim & \mathcal{N}_{mq}(0, G) \end{array}\right\} \text{Modèle hiérarchique à 2 niveaux} \tag{IV.4}$$

Ainsi le modèle se re-écrit $y = X\beta + \varepsilon^*$ où $\varepsilon^* = Zb + \varepsilon \sim \mathcal{N}_n(0, \Sigma)$ avec $\Sigma = ZGZ^T + R$. Le modèle marginal s'écrit :

$$\left.\begin{array}{rcl} y & = & X\beta + \varepsilon^* \\ \varepsilon^* & \sim & \mathcal{N}_n(0, \Sigma) \end{array}\right\} \text{Modèle marginal} \tag{IV.5}$$

On rappelle que, comme Ψ est une matrice symétrique semi-définie positive, paramétrée par un vecteur de variance θ, on peut aussi utiliser ce paramètre pour décrire la matrice Σ : $\Sigma(\theta) = ZG(\theta)Z^T + R(\theta)$.

Si l'on s'intéresse à l'estimation de β et au calcul de b on doit utiliser le modèle hiérarchique à deux niveaux de l'équation (IV.4). Si l'on s'intéresse uniquement à l'estimation de β on peut utiliser le modèle linéaire ordinaire de l'équation (IV.5) (modèle marginal).

10. ordonnée à l'origine comprise

3.2 Vocabulaire sur les vraisemblances

La fonction de vraisemblance, notée $L(x_1, ..., x_n, \theta_1, ..., \theta_k)$ est une fonction de probabilités conditionnelles qui décrit les valeurs x_i d'une loi statistique en fonction des paramètres θ_j supposés connus. Si on suppose que les x_i sont indépendants et identiquement distribués entre eux, elle s'exprime à partir de la fonction de densité $f(x, \theta)$ par $L(x_1, ..., x_n, \theta) = \prod_{i=1}^{n} f(x_i, \theta)$. La vraisemblance dépend habituellement de nombreux paramètres. Selon l'application, nous nous intéressons souvent qu'à un sous-ensemble de ces paramètres.

Désignons par θ les paramètres qui nous intéressent, et par β les paramètres qui ne sont pas d'intérêt principal. La manière standard d'aborder le problème d'estimation est de maximiser la vraisemblance afin d'obtenir des estimations de θ et de β. Cependant, comme l'intérêt principal réside dans θ, les vraisemblances partielles, apparentes et marginales, proposent d'autres façons d'estimer θ sans estimer β. Pour faire la différence, notons L(θ, β | données) la vraisemblance habituelle.

Maximum de vraisemblance	Trouver θ et β qui maximisent L(θ, β | données).
Vraisemblance partielle	Si l'on peut écrire la vraisemblance comme : L(θ, β | données) = $L_1(\theta \mid \text{données})\ L_2(\beta \mid \text{données})$. Alors on maximise simplement $L_1(\theta \mid \text{données})$.
Vraisemblance apparente	Si l'on peut exprimer β comme une fonction de θ alors on remplace β par la fonction correspondante. Ainsi, si β = g(θ), alors on maximise $L(\theta, \text{g}(\theta) \mid \text{données})$
Vraisemblance marginale	On intègre l'équation de vraisemblance selon β en exploitant le fait que l'on peut identifier la distribution de probabilité de β conditionnellement à θ.

Plusieurs méthodes d'estimation des paramètres ont été utilisées pour les modèles à effets mixtes linéaires. On se concentre ici sur deux méthodes générales : la méthode du maximum de vraisemblance (ML) et la méthode du maximum de vraisemblance restreint (REML).

3.3 Estimation avec structure de covariance connue

Dans cette section, nous considérons un cas particulier d'estimation ML pour les modèles mixtes linaires, où le vecteur paramètre de covariance θ, et par conséquent la matrice Σ, sont connues. Bien que cette situation ne se produise pas en pratique, cela a des implications informatiques importantes. C'est pourquoi on le présente séparément. Comme nous supposons que θ est connu, les seuls paramètres que nous devons estimer sont les effets fixes β.

Estimation de β quand Ψ et σ^2 (donc Σ), sont connues

En utilisant (IV.5), nous obtenons l'estimateur du maximum de vraisemblance (ou l'estimateur pondéré des moindres carrés) de β suivant :

$$\boxed{\widehat{\beta} = \left(X^T \Sigma^{-1} X\right)^{-1} X^T \Sigma^{-1} y.} \tag{IV.6}$$

En effet, comme $y = X\beta + \varepsilon^*$ avec $\varepsilon^* \sim \mathcal{N}_n(0, \Sigma)$, avec Σ connue, on peut écrire $\Sigma = \Sigma^{1/2}\left(\Sigma^{1/2}\right)^T$. D'où

$$\begin{aligned} \Sigma^{-1/2}y &= \Sigma^{-1/2}X\beta + \underbrace{\Sigma^{-1/2}\varepsilon^*}_{\sim \mathcal{N}_n(0, \underbrace{\Sigma^{-1/2}\Sigma(\Sigma^{-1/2})^T}_{I_n})}. \end{aligned} \tag{IV.7}$$

L'estimateur du maximum de vraisemblance dans (IV.7) correspond au modèle linéaire classique, mais avec l'addition de la matrice non diagonale de pondération $\Sigma^{-1/2}$. Donc l'estimateur des moindres carrés de β dans (IV.7) donne :

$$\begin{aligned} \widehat{\beta} &= \left(X^T(\Sigma^{-1/2})^T\Sigma^{-1/2}X\right)^{-1} X^T(\Sigma^{-1/2})^T\Sigma^{-1/2}y \\ &= \left(X^T\Sigma^{-1}X\right)^{-1} X^T\Sigma^{-1}y. \end{aligned}$$

L'estimation $\widehat{\beta}$ a la propriété statistique souhaitable d'être le meilleur estimateur linéaire sans biais (BLUE) de β. La formule close de l'équation (IV.6) définit également une relation fonctionnelle entre le paramètre de covariance θ, et la valeur de β qui maximise $L(\beta, \theta)$. Nous utilisons cette relation dans la section suivante pour déterminer les paramètres des effets fixes β à partir de la log-vraisemblance, et les écrire comme une fonction de θ seulement.

Calcul des effets aléatoires b_i

Les effets aléatoires b_i ne sont pas des paramètres et ne sont donc pas estimés par le modèle. Ce sont des variables cachées dont on a pourtant besoin pour faire fonctionner le modèle sur de nouvelles données dans un but prédictif.
Du modèle hiérarchique à deux niveaux (IV.4) il vient $y \sim \mathcal{N}_n(X\beta, \Sigma)$ et $b \sim \mathcal{N}_{mq}(0, G)$. Or,

$$\begin{aligned} Cov(y, b) &= Cov(X\beta + Zb + \varepsilon, b) \\ &= \underbrace{Cov(X\beta, b)}_{=0} + Z\underbrace{Var(b)}_{G} + \underbrace{Cov(\varepsilon, b)}_{=0} \\ &= ZG\,. \end{aligned}$$

Donc

$$\begin{pmatrix} y \\ b \end{pmatrix} \sim \mathcal{N}_{n+mq}\left[\begin{pmatrix} X\beta \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma & ZG \\ GZ^T & G \end{pmatrix}\right].$$

De plus, on rappelle que si $\begin{pmatrix} U \\ T \end{pmatrix} \sim \mathcal{N}_p\left[\begin{pmatrix} \mu_U \\ \mu_T \end{pmatrix}, \begin{pmatrix} \Sigma_U & \Sigma_{UT} \\ \Sigma_{TU} & \Sigma_T \end{pmatrix}\right]$ alors $T|U \sim \mathcal{N}(\mu_{T|U}, \Sigma_{T|U})$ avec $\mu_{T|U} = \mu_T + \Sigma_{TU}\Sigma_U^{-1}(U - \mu_U)$ et $\Sigma_{T|U} = \Sigma_T - \Sigma_{TU}\Sigma_U^{-1}\Sigma_{UT}$.

Donc pour notre application, on a $\mathbb{E}[b|y] = 0 + GZ^T\Sigma^{-1}(y - X\beta) = GZ^T\Sigma^{-1}(y - X\beta)$ qui est le meilleur prédicteur (et non estimateur) linéaire sans biais (BLUP) de b. Ainsi,

le meilleur BLUP empirique de b est

$$\boxed{\widehat{b} = GZ^T\Sigma^{-1}(y - X\widehat{\beta}).} \tag{IV.8}$$

Soit, pour le groupe $i \in [\![1; m]\!]$

$$\widehat{b}_i = \Psi Z_i^T \Sigma_i^{-1}(y_i - X_i\widehat{\beta}) \quad \in \mathbb{R}^q \, .$$

Maximisation jointe de la vraisemblance de (y, b) par rapport à (β, b)

Désignons par f les fonctions densité. D'après le modèle hiérarchique à 2 niveaux (IV.4), on a

$$\begin{aligned} f(y,b) &= f(y|b).f(b) \\ &\propto \exp\left(-\frac{1}{2}(y - X\beta - Zb)^T R^{-1}(y - X\beta - Zb)\right) \exp\left(-\frac{1}{2}b^T G^{-1} b\right) \\ \text{d'où } \ln f(y,b) &= -\frac{1}{2}(y - X\beta - Zb)^T R^{-1}(y - X\beta - Zb) \qquad \underbrace{-\frac{1}{2}b^T G^{-1} b}_{\text{terme de pénalisation pour } b} \\ &\quad + \text{ constante indep. de } (\beta, b). \end{aligned}$$

Donc pour maximiser $\ln f(y, b)$ par rapport à β et b, il suffit de minimiser

$$\begin{aligned} Q(\beta,b) &:= (y - X\beta - Zb)^T R^{-1}(y - X\beta - Zb) + b^T G^{-1} b \\ &= b^T R^{-1} b - 2\beta^T X^T R^{-1} y + 2\beta^T X^T R^{-1} Zb - 2b^T Z^T R^{-1} y \\ &\quad + \beta^T X^T R^{-1} X\beta + b^T Z^T R^{-1} Zb + b^T G^{-1} b \, , \end{aligned}$$

d'où

$$\begin{aligned} \frac{\partial}{\partial\beta} Q(\beta,b) &= -2X^T R^{-1} y + 2X^T R^{-1} Zy + 2X^T R^{-1} X\beta \overset{But}{=} 0 \\ \frac{\partial}{\partial b} Q(\beta,b) &= -2Z^T R^{-1} X\beta - 2Z^T R^{-1} y + 2Z^T R^{-1} Zb + 2G^{-1} b \overset{But}{=} 0 \end{aligned}$$

$$\Leftrightarrow \left\{ \begin{array}{rcl} X^T R^{-1} X\widetilde{\beta} + X^T R^{-1} Z\widetilde{b} &=& X^T R^{-1} y \\ Z^T R^{-1} X\widetilde{\beta} + (Z^T R^{-1} Z + G^{-1})\widetilde{b} &=& Z^T R^{-1} y \, , \end{array} \right.$$

ce qui équivaut à

$$\boxed{\begin{pmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & Z^T R^{-1} Z + G^{-1} \end{pmatrix} \begin{pmatrix} \widetilde{\beta} \\ \widetilde{b} \end{pmatrix} = \begin{pmatrix} X^T R^{-1} y \\ Z^T R^{-1} y \end{pmatrix}.}$$

Ce sont les équations d'Henderson ou les équations dites "du modèle mixte". Un de leurs avantages est le suivant : les matrices en jeu sont d'ordre beaucoup plus petit que n (qui est l'ordre de Σ et qu'il faut inverser pour trouver $\widehat{\beta} = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}y$).

De plus, quand $R = \sigma^2 I_n$ et $\Psi = diag(\sigma_1^2, ..., \sigma_q^2)$ alors ce système équivaut à

$$\boxed{\begin{pmatrix} X^T X & X^T Z \\ Z^T X & Z^T Z + diag\left(\frac{\sigma^2}{\sigma_k^2}\right) \end{pmatrix} \begin{pmatrix} \widetilde{\beta} \\ \widetilde{b} \end{pmatrix} = \begin{pmatrix} X^T y \\ Z^T y \end{pmatrix}.}$$

On peut montrer que $\widetilde{\beta} = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}y$ et $\widetilde{b} = GZ^T\Sigma^{-1}(y - X\widetilde{\beta})$ qui ont été trouvés en (IV.6) et (IV.8), sont solutions du système précédent.

3.4 Estimation quand la structure de covariance est inconnue

Dans cette section, on se concentre sur l'estimation des paramètres de covariance θ, et le paramètre des effets fixes β. On suppose donc le modèle marginal décrit à l'équation (IV.5) :

$$y = X\beta + \varepsilon^*, \quad \varepsilon^* \sim \mathcal{N}_n(0, \Sigma),$$

avec $\Sigma = ZGZ^T + R$, et où G et R sont entièrement déterminées par un paramètre de variance θ inconnu, i.e. on écrit

$$\Sigma(\theta) = ZG(\theta)Z^T + R(\theta).$$

Estimation ML pour le modèle marginal étendu

$$y = X\beta + \varepsilon^*, \quad \varepsilon^* \sim \mathcal{N}_n(0, \Sigma(\theta)) \quad \text{avec} \quad \Sigma(\theta) = ZG(\theta)Z^T + R(\theta)$$

La log-vraisemblance pour (β, θ) s'écrit

$$l(\beta, \theta) = -\frac{1}{2}\left[\ln|\Sigma(\theta)| + (y - X\beta)^T\Sigma(\theta)^{-1}(y - X\beta)\right] + \text{const. ind. de } \beta, \theta \qquad \text{(IV.9)}$$

Si l'on maximise l'équation (IV.9) par rapport à β, en considérant θ fixe, on obtient

$$\widetilde{\beta}(\theta) := \left(X^T\Sigma(\theta)^{-1}X\right)^{-1}X^T\Sigma(\theta)^{-1}y.$$

Tout d'abord, pour obtenir des estimations pour les paramètres de covariance dans θ, on construit la log-vraisemblance apparente $l_{p(ML)}(\theta)$. La fonction $l_{p(ML)}(\theta)$ provient de $l(\beta, \theta)$ en remplaçant les paramètres β par l'expression définissant $\widetilde{\beta}$. On obtient la log vraisemblance apparente :

$$\begin{aligned} l_{p(ML)}(\theta) &:= l(\widetilde{\beta}(\theta), \theta) \\ &= -\frac{1}{2}\left[\ln|\Sigma(\theta)| + (y - X\widetilde{\beta}(\theta))^T\Sigma(\theta)^{-1}(y - X\widetilde{\beta}(\theta))\right]. \end{aligned}$$

En maximisant $l_{p(ML)}(\theta)$ par rapport à θ, on obtient l'estimateur du maximum de vraisemblance $\widehat{\theta}_{ML}$.

En général, la maximisation de $l_{p(ML)}(\theta)$, par rapport à θ est un exemple d'optimisation non linéaire, avec des contraintes d'inégalités imposées pour θ afin que les exigences de matrice définie-positive sur Ψ et R soient satisfaites.

Il n'y a pas de formule explicite pour le θ optimal, donc l'estimation de θ se fait en effectuant des algorithmes itératifs jusqu'à la convergence. Lorsque les estimations ML des paramètres de covariance dans θ (et par conséquent, les estimations des variances et covariances Ψ et R) sont obtenues par un processus itératif, on peut alors calculer $\widehat{\beta}$. Cela peut se faire sans processus itératif, en utilisant la définition de Σ et l'équation (IV.6).

Puisque nous avons remplacé Σ par son estimation $\widehat{\Sigma}$, on dit que $\widehat{\beta}$ est le meilleur estimateur empirique linéaire sans biais (BLUE) de β.

Les estimations ML de θ sont biaisées car elles ne prennent pas en compte la perte des degrés de liberté qui résulte de l'estimation des paramètres à effets fixes β (voir [Verb00] pour une discussion sur le biais dans les estimations ML de θ dans le contexte de modèle mixte linéaire). Une méthode alternative au maximum de vraisemblance, connue sous le nom d'estimation REML est souvent utilisée pour éliminer le biais dans les estimations ML des paramètres de covariance. Nous abordons l'estimation REML dans la sous-section suivante.

Maximum de vraisemblance restreint (REML)

L'estimation REML est une autre façon d'estimer les paramètres de covariance θ. L'estimation REML (parfois appelée "residual maximum likelihood estimation") a été introduite au début des années 1970 par [Patt71] comme méthode d'estimation des composantes de la variance dans le contexte de plans incomplets et/ou non équilibrés.

REML est souvent préférée à l'estimation ML car elle propose des estimations sans biais des paramètres de covariance en tenant compte de la perte des degrés de liberté qui résulte de l'estimation des effets fixes β.

Nous utilisons ici la log-vraisemblance marginale pour l'estimation de θ

$$l_R(\theta) := \ln\left(\int L(\beta,\theta)d\beta\right).$$

$$\int L(\beta,\theta)d\beta = \int \frac{1}{(2\pi)^{n/2}}|\Sigma(\theta)|^{-1/2} \times \exp\left[-\frac{1}{2}(y-X\beta)^T\Sigma(\theta)^{-1}(y-X\beta)\right]d\beta.$$

On considère :

$$\begin{aligned}(y-X\beta)^T\Sigma(\theta)^{-1}(y-X\beta) &= \beta^T \underbrace{X^T\Sigma(\theta)^{-1}X}_{A(\theta)}\beta - 2y^T\Sigma(\theta)^{-1}X\beta + y^T\Sigma(\theta)^{-1}y \\ &= (\beta - B(\theta)y)^T A(\theta)(\beta - B(\theta)y) + y^T\Sigma(\theta)^{-1} - y^T B(\theta)^T A(\theta)B(\theta)y\,,\end{aligned}$$

où $B(\theta) := A(\theta)^{-1}X^T\Sigma(\theta)^{-1}$.
On remarque que $B(\theta)^T A(\theta) = \Sigma(\theta)^{-1}XA(\theta)^{-1}A(\theta) = \Sigma(\theta)^{-1}X$.
Aussi, on obtient

$$\begin{aligned}\int L(\beta,\theta)d\beta &= \frac{|\Sigma(\theta)|^{-1/2}}{(2\pi)^{n/2}}\exp\left[-\frac{1}{2}y^T[\Sigma(\theta)^{-1} - B(\theta)^T A(\theta)B(\theta)]y\right] \\ &\quad \cdot \underbrace{\int \exp\left[-\frac{1}{2}(\beta) - B(\theta)y^T A(\theta)(\beta - B(\theta)y)\right]d\beta}_{\frac{(2\pi)^{p/2}}{|A(\theta)^{-1}|^{-1/2}} \quad (de\ variance\ A(\theta)^{-1})}. \qquad \text{(IV.10)}\end{aligned}$$

et

$$\begin{aligned} & (y - X\widetilde{\beta}(\theta))^T \Sigma(\theta)^{-1} (y - X\widetilde{\beta}(\theta)) \\ = \; & y^T \Sigma(\theta)^{-1} y - 2y^T \Sigma(\theta)^{-1} X\widetilde{\beta}(\theta) + \widetilde{\beta}(\theta)^T \underbrace{X^T \Sigma(\theta)^{-1} X}_{A(\theta)} \widetilde{\beta}(\theta) \\ = \; & y^T \Sigma(\theta)^{-1} y - 2y^T \Sigma(\theta)^{-1} XB(\theta) y + y^T B(\theta)^T A(\theta) B(\theta) y \\ = \; & y^T \Sigma(\theta)^{-1} y - y^T B(\theta)^T A(\theta) B(\theta) y. \end{aligned}$$

On utilise désormais

$$\widetilde{\beta} = (X^T \Sigma(\theta)^{-1} X)^{-1} X^T \Sigma(\theta)^{-1} y = A(\theta)^{-1} X^T \Sigma(\theta)^{-1} y = B(\theta) y \; ,$$

et

$$B(\theta)^T A(\theta) B(\theta) = \Sigma(\theta)^{-1} X A(\theta)^{-1} A(\theta) B(\theta) = \Sigma(\theta)^{-1} X B(\theta).$$

On peut alors re-écrire l'équation (IV.10) :

$$\begin{aligned} \int L(\beta, \theta) d\beta \; = \; & \frac{|\Sigma(\theta)|^{-1/2}}{(2\pi)^{n/2}} \exp\left[-\frac{1}{2} (y - X\widetilde{\beta}(\theta))^T \Sigma(\theta)^{-1} (y - X\widetilde{\beta}(\theta)) \right] \\ & \times \frac{(2\pi)^{n/2}}{|A(\theta)^{-1}|^{-1/2}} \; , \end{aligned}$$

d'où

$$\begin{aligned} l_R(\theta) \; = \; & \ln \left(\int L(\beta, \theta) d\beta \right) \\ = \; & -\frac{1}{2} \left[\ln |\Sigma(\theta)| + (y - X\widetilde{\beta}(\theta))^T \Sigma(\theta)^{-1} (y - X\widetilde{\beta}(\theta)) \right] - \frac{1}{2} \ln |A(\theta)| + C \\ = \; & l_p(\theta) - \frac{1}{2} \ln |A(\theta)| + C. \end{aligned}$$

Ainsi, la ML restreinte (REML) de θ est donnée par $\widetilde{\theta}_{REML}$ qui maximise

$$l_R(\theta) = l_p(\theta) - \frac{1}{2} \ln |X^T \Sigma(\theta)^{-1} X| \; .$$

En résumé : estimation pour les modèles mixtes linéaires avec covariance inconnue

Considérant le modèle linéaire mixte suivant,

$$y = X\beta + Zb + \varepsilon, \quad \begin{pmatrix} b \\ \varepsilon \end{pmatrix} \sim \mathcal{N}_{mq+n}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} G(\theta) & 0_{mq\times n} \\ 0_{n\times mq} & R(\theta) \end{pmatrix} \right],$$

avec $\Sigma(\theta) = ZG(\theta)Z^T + R(\theta)$,
on estime le paramètre (vecteur) de covariance θ par,

— soit $\widehat{\theta}_{ML}$ qui maximise

$$l_p(\theta) = -\frac{1}{2}\left[\ln|\Sigma(\theta)| + (y - X\widetilde{\beta}(\theta))^T\Sigma(\theta)^{-1}(y - X\widetilde{\beta}(\theta))\right]$$

où $\widetilde{\beta} = (X^T\Sigma(\theta)^{-1}X)^{-1}X^T\Sigma(\theta)^{-1}y$

— soit $\widehat{\theta}_{REML}$ qui maximise $\quad l_R(\theta) = l_p(\theta) - \frac{1}{2}\ln|X^T\Sigma(\theta)^{-1}X|$.
Les effets fixes β et les effets aléatoires b sont estimés par

$$\begin{aligned} \widehat{\beta} &= \left(X^T\widehat{\Sigma}^{-1}X\right)^{-1}X^T\widehat{\Sigma}^{-1}y \\ \widehat{b} &= \widehat{G}Z^T\widehat{\Sigma}^{-1}(y - X\widehat{\beta}) \qquad \text{où} \quad \widehat{\Sigma} = \Sigma(\widehat{\theta}_{ML}) \ \textit{ou} \ \Sigma(\widehat{\theta}_{REML}). \end{aligned}$$

Cas particulier, si

$$\begin{aligned} \Sigma = ZGZ^T + R &= \begin{pmatrix} Z_1\Psi Z_1^T + R_1 & & 0 \\ & \ddots & \\ 0 & & Z_m\Psi Z_m^T + R_m \end{pmatrix} \qquad \text{(diagonale par blocs)} \\ &= \begin{pmatrix} \Sigma_1 & & \\ & \ddots & \\ & & \Sigma_m \end{pmatrix} \qquad \text{où} \quad \Sigma_i = Z_i\Psi Z_i^T + R_i. \end{aligned}$$

On définit $\widehat{\Sigma}_i := Z_i\Psi(\widehat{\theta})Z_i^T + R_i(\widehat{\theta})$, où $\widehat{\theta} = \widehat{\theta}_{ML}$ ou $\widehat{\theta}_{REML}$. Alors

$$\begin{aligned} \widehat{\beta} &= (X^T\widehat{\Sigma}^{-1}X)^{-1}X^T\widehat{\Sigma}^{-1}y \\ &= \left(\sum_{i=1}^m X_i^T\widehat{\Sigma}_i^{-1}X_i\right)^{-1}\sum_{i=1}^m X_i^T\widehat{\Sigma}_i^{-1}y_i \qquad \in \mathbb{R}^p \,, \end{aligned}$$

et $\widehat{b} = \widehat{G}Z^T\widehat{\Sigma}^{-1}(y - X\widehat{\beta})$ a pour composantes

$$\widehat{b}_i = \Psi(\widehat{\theta})Z_i^T\widehat{\Sigma}_i^{-1}(y_i - X_i\widehat{\beta}) \qquad \in \mathbb{R}^q.$$

Contrairement à l'estimation ML, la méthode REML ne fournit pas de formule pour les estimations. Au lieu de cela, nous utilisons l'équation (IV.6) de l'estimation ML pour estimer les paramètres des effets fixes et leurs erreurs. Même si nous utilisons

les mêmes formules des paramètres à effets fixes de l'équation (IV.6) pour l'estimation REML et ML, il est important de noter que les $\widehat{\beta}$ résultants de l'estimation REML et ML sont différents, car la matrice Σ est différente dans chaque cas.

3.5 Intervalles de confiance et tests d'hypothèses

Puisque $y \sim \mathcal{N}(X\beta, \Sigma(\theta))$, une approximation de la covariance de $\widehat{\beta} = \left(X^T\Sigma^{-1}(\widehat{\theta})X\right)^{-1} X^T\Sigma^{-1}(\widehat{\theta})y$ est donnée par

$$A(\widehat{\theta}) := \left(X^T\Sigma^{-1}(\widehat{\theta})X\right)^{-1}.$$

Remarque : on suppose ici que $\Sigma(\widehat{\theta})$ est fixe et ne dépend pas de y. Ainsi $\widehat{\sigma}_j := \left(X^T\Sigma^{-1}(\widehat{\theta})X\right)^{-1}_{jj}$ sont considérés comme les estimateurs de $Var(\widehat{\beta}_j)$. Donc

$$\left[\widehat{\beta}_j - z_{1-\alpha/2}\sqrt{\left(X^T\Sigma^{-1}(\widehat{\theta})X\right)^{-1}_{jj}} \; ; \; \widehat{\beta}_j + z_{1-\alpha/2}\sqrt{\left(X^T\Sigma^{-1}(\widehat{\theta})X\right)^{-1}_{jj}}\right]$$

est un intervalle de confiance approximatif à $100(1-\alpha)\%$ pour β_j, avec $z_{1-\alpha/2}$ le quantile d'ordre $1-\frac{\alpha}{2}$ d'une loi normale $\mathcal{N}(0,1)$. On s'attend à ce que $\left(X^T\Sigma^{-1}(\widehat{\theta})X\right)^{-1}_{jj}$ sous-estime $Var(\widehat{\beta}_j)$ puisque la variation de θ n'est pas prise en compte. Une analyse bayésienne complète utilisant des méthodes MCMC (Monte-Carlo par Chaines de Markov) est préférable à ces approximations [West07].

Sous l'hypothèse que $\widehat{\beta}$ est asymptotiquement normal, de moyenne β et de matrice de covariance $A(\theta)$, l'hypothèse classique de test peut être formulée, i.e.

— $H_0 : \beta_j = 0 \quad versus \quad H_1 : \beta_j \neq 0$

$$\text{On rejette} \quad H_0 \Leftrightarrow |t_j| = \left|\frac{\widehat{\beta}_j}{\widehat{\sigma}_j}\right| > z_{1-\alpha/2}.$$

— $H_0 : C\beta = d \quad versus \quad H_1 : C\beta \neq d \quad$ où $\text{rang}(C) = r$.

On rejette $\quad H_0 \Leftrightarrow W := (C\widehat{\beta}-d)^T(C^TA(\widehat{\theta})C)^{-1}(C\widehat{\beta}-d) > \chi^2_{1-\alpha,r} \qquad$ (Test de Wald),

ou

On rejette $\quad H_0 \Leftrightarrow -2\left[l(\widehat{\beta},\widehat{b} - l(\widehat{\beta}_R),\widehat{b}_R)\right] > \chi^2_{1-\alpha,r} \qquad$ (Test du rapport de vraisemblance),

où $\widehat{\beta}$, $\widehat{b}$ sont les estimateurs issus du modèle non-restreint,
$\widehat{\beta}_R$, $\widehat{b}_R$ sont les estimateurs issus du modèle restreint $(C\beta = d)$.

3.6 Algorithmes pour l'optimisation de la vraisemblance

Après avoir défini les méthodes d'estimation ML et REML, nous présentons brièvement les algorithmes utilisés en pratique pour effectuer l'estimation d'un modèle mixte linéaire. Comme on a pu le voir, la principale difficulté de calcul dans l'analyse de ces modèles est l'estimation des paramètres de covariance. On utilise l'optimisation numérique itérative des log-vraisemblances introduites à la Sous-section 3.4 page 91 pour l'estimation ML et page 92 pour l'estimation REML, sous réserve des contraintes

imposées aux paramètres pour assurer le caractère défini-positif des matrices Ψ et R. Les algorithmes itératifs les plus couramment utilisés pour ce problème d'optimisation dans le contexte d'effets mixtes sont l'algorithme EM (Expectation-Maximization), l'algorithme de Newton-Raphson (N-R) (méthode la plus répandue) et l'algorithme "Fisher-scoring". Plus de détails sont disponibles dans [West07].

4 Les réseaux de neurones

Les réseaux de neurones (NN) sont une méthode statistique permettant la mise en place de modèles paramétriques non linéaires [Hayk94, Bish96, Haga14]. Les réseaux de neurones sont extrêmement efficaces mais ils nécessitent un nombre de données suffisant.

Un réseau de neurones est l'association des processeurs élémentaires en un graphe plus ou moins complexe. Les différents types de réseaux se distinguent par : l'organisation du graphe (i.e. leur architecture : graphe en couches, graphe complet...), leur niveau de complexité (nombre de neurones, présence de boucles de rétroaction...), leur type de neurones, leurs fonctions d'activation et leur objectif (apprentissage supervisé ou non, optimisation, systèmes dynamiques).

4.1 Le neurone artificiel

À l'image d'un neurone biologique qui comporte des synapses (connexion avec les autres neurones), des dendrites ("entrée du neurone"), un axone ("sortie du neurone") et un noyau (activant la sortie en fonction des stimulations en entrée), un neurone artificiel (Figure IV.1) est un modèle caractérisé par : (1) des entrées $x_1, ..., x_{\eta_{inputs}}$, (2) un vecteur de poids W, (3) une fonction d'activation $f : \mathbb{R} \to \mathbb{R}$, et (4) une sortie y_η.

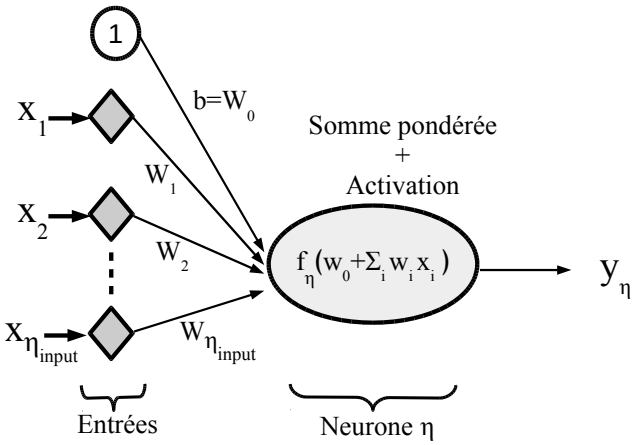


FIGURE IV.1 – Représentation schématique d'un neurone formel.

La fonction d'activation f s'applique à une combinaison affine des signaux d'entrée, dont les coefficients sont un vecteur de poids $W = [w_0, ..., w_{\eta_{inputs}}] \in \mathbb{R}^{\eta_{inputs}+1}$ associé à chaque entrée. On appelle b le biais (ou "offset", ou encore "seuil") du neurone : le biais est une entrée qui émet un signal d'intensité constante 1. Les valeurs des poids sont estimées lors de l'apprentissage.

$$y_\eta = f\left(b + \sum_{i=0}^{\eta_{inputs}} w_i x_i\right).$$

Il existe diverses fonctions d'activation possibles - souvent de type sigmoïde - dont les principales sont :

— la fonction linéaire : f est l'identité,

— la fonction logistique simple : $f(x) = 1/(1+\exp(-x))$,

— la fonction seuil : $f(x) = \mathbf{1}_{[0,+\infty[}(x)$,

— la fonction radiale : $f(x) = \frac{1}{\sqrt{2\pi}}\exp(-x^2/2)$.

La fonction seuil, la fonction arc-tangente ou la fonction logistique, font partie des fonctions sigmoïdes. La fonction d'activation doit être continue si l'on veut faire de l'apprentissage en minimisant la fonction de coût, car il faut que cette fonction de coût soit différentiable par rapport aux paramètres W du réseau. On pourra alors utiliser une technique d'optimisation pour la minimisation, par exemple une descente de gradient.

De loin les plus utilisées, les fonctions linéaires ou sigmoïdes sont bien adaptées aux algorithmes d'apprentissage comportant une rétro-propagation du gradient car elles sont différentiables. La fonction seuil, pourtant plus conforme à la réalité biologique, est peu utilisée car non différentiable.

4.2 Le perceptron multi-couches (PMC)

Dans cette thèse, on considère la possibilité d'utiliser un modèle non-linéaire pour prédire les rendements (ou assimilés). Les réseaux de neurones artificiels sont de bons candidats. Ici, on considère seulement les réseaux de neurones acycliques qui sont calibrés lors d'un apprentissage supervisé [Bish96]. Ces types de réseaux de neurones possèdent seulement des connections depuis la couche des variables d'entrée vers celle de sortie.

On définit une couche comme un ensemble de neurones sans connexion entre eux. Le perceptron multi-couches est un réseau construit à partir de couches successives et où l'apprentissage suit la règle du delta (ou règle d'apprentissage de Widrow-Hoff) : l'apprentissage se fait à travers la valeurs des erreurs commises par le modèle c'est-à-dire la différence entre les valeurs désirées et les valeurs proposées par le modèle.

Le perceptron multi-couches est un modèle paramétrique non linéaire qui calcule une sortie (multivariée ou non) lorsqu'on lui présente une entrée (multivariée) [Hert91]. Comme tous les réseaux de neurones, il est constitué de plusieurs "processeurs" interconnectés qui effectuent des calculs locaux : les neurones. Dans le PMC, les neurones sont organisés en couches successives de neurones indépendants. Traditionnellement, on ne compte pas la couche d'entrée dans le nombre de couches du réseau. Toutes les couches, sauf les couches d'entrée et de sortie, sont appelées couches "cachées". Sur une couche, chaque neurone est relié à tous les neurones de la couche précédente par une "liaison synaptique" à laquelle est associé un poids synaptique w_{ij}. On notera donc w_{ij} le poids d'un neurone j d'une certaine couche, s'appliquant aux sorties du neurone i de la couche précédente. À architecture fixée (nombre de couches cachées, nombre de neurones par couches, et fonction d'activation choisis), toute l'information du réseau

est incluse dans ces poids synaptiques. Le neurone du perceptron effectue successivement deux opérations [McCu43] :

— le neurone j calcule la somme pondérée " h_j" (appelée entrée totale) de ses entrées x_i, plus un biais b_j :

$$h_j = b_j + \sum_{i=1}^{p} w_{ij} x_i \ ,$$

avec p le nombre d'entrées du neurone j,

— et applique à cette entrée totale h_j, une fonction d'activation (ou de seuil) f :

$$y_j = f(h_j).$$

Remarque : Le biais b_j de sortie du neurone j est aussi un paramètre du réseau qui doit être déterminé. Il peut être assimilé à un poids synaptique relié à un neurone de sortie 1, constante. Dans la suite du texte, on omettra parfois ce neurone de biais par souci de concision dans les notations.

La couche d'entrée, possède un neurone par entrée et doit lire les signaux entrants tandis que la couche de sortie fournit le résultat du système. Le perceptron peut posséder une ou plusieurs couches cachées qui participent à l'estimation des paramètres du modèle.

En résumé, un perceptron multi-couches réalise une transformation des entrées sous la forme :

$$y = \phi(x_1, ..., x_m; W, b),$$

où W est l'ensemble des poids w_{ijl} (avec l parcourant le nombre de couches, j le nombre de neurone sur cette couche, et i le nombre des entrées de chacun des neurones).
La couche d'entrée ($l = 0$) n'est pas paramétrée et sert juste à distribuer les entrées aux neurones de la couche suivante. On appelle ϕ la fonction de transfert.

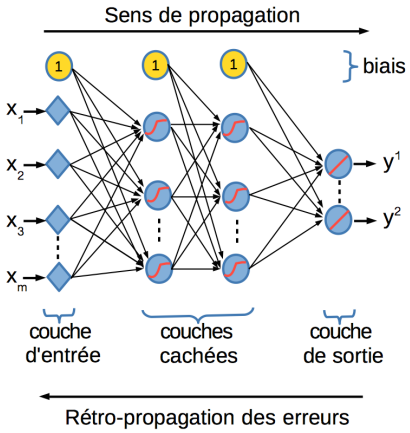


FIGURE IV.2 – Schéma d'un perceptron multi-couches (ici, une couche d'entrée, deux couches cachées et une couche de sortie).

De même b est la matrice des biais : b_{jl} représente le biais du $j^{\text{ème}}$ neurone de la couche l. La plupart du temps, la dernière couche est constituée de neurones avec une fonction d'activation linéaire, tandis que les neurones des couches cachées sont munies d'une fonction sigmoïde.

De cette façon, dans le cadre d'une régression avec un perceptron à une couche cachée, possédant m_1 neurones (biais non compris), et un neurone de sortie ; la fonction

de transfert ϕ s'écrit (combinaison affine des sorties de la couche cachée) :

$$\begin{aligned} y = \phi(x; W, b) &= b_{12} + \sum_{j=1}^{m_1} w_{j,1,2}.y_{j1}, \\ \text{avec } y_{j1} &= f(b_{j1} + \sum_{i=1}^{m_0} w_{ij1}.x_i) \qquad \text{pour } \; j = 1, ..., m_1, \end{aligned}$$

où y_{jl} représente la sortie du $j^{\text{ème}}$ neurone de la couche l.

Un réseau acyclique à deux couches, avec des fonctions d'activation logistiques sur les neurones de la couche cachée et une fonction linéaire sur le neurone de sortie, est utilisé au Chapitre V pour l'estimation des anomalies de rendement. L'anomalie de rendement de maïs y est donc modélisée par l'équation suivante :

$$y = \sum_{j=1}^{n_{neuron}} w_j \cdot \text{logsig}\left(\sum_{i=1}^{n_{\text{inputs}}} x_i w_{ij} + b_j \right) + b_{output},$$

où comme précédemment, les x_i sont les variables d'entrée météorologiques, les b_j sont les différents biais du réseau de neurone, n_{inputs} est le nombre d'entrées, et n_{neuron} est le nombre de neurones sur la couche cachée. $logsig$ désigne la fonction d'activation logistique $x \mapsto 1/(1+e^{-x})$. Ce modèle demande donc d'estimer $n_{inputs} + 3$ paramètres (les poids).

Les PMC sont souvent appelés réseaux à propagation avant (feedforward network) car la mise-à-jour des neurones se fait couche par couche, de la couche d'entrée à la couche de sortie. La mise-à-jour des neurones dans une couche est synchrone (i.e. en même temps). Les inter-relations entre neurones spécifient des compétitions et des coopérations entre neurones. Des architectures plus complexes (boucles, récurrences, connexions latérales, éliminations de certains poids), spécifiant des inter-relations plus complexes peuvent aussi être utilisées dans certains cas.

4.3 La descente de gradient

Le problème d'apprentissage classique consiste à trouver la valeur optimale des paramètres qui minimise les erreurs que commet le réseau de neurones sur l'ensemble d'apprentissage : c'est donc une fonction appelée "fonction de coût" à p variables $F : \mathbb{R}^p \to \mathbb{R}$ que l'on cherche à minimiser. Le gradient d'une telle fonction - noté $\nabla F(x)$ - est un vecteur de taille p dont les composantes sont les p dérivées partielles par rapport à chacune des variables. Par définition, $\nabla F(x)$ indique la direction vers laquelle F croit le plus vite au voisinage de x et sa norme nous indique à quel point F varie localement (Figure IV.3). Ainsi si l'on cherche à faire décroitre la fonction de coût, il faut se déplacer en sens contraire.

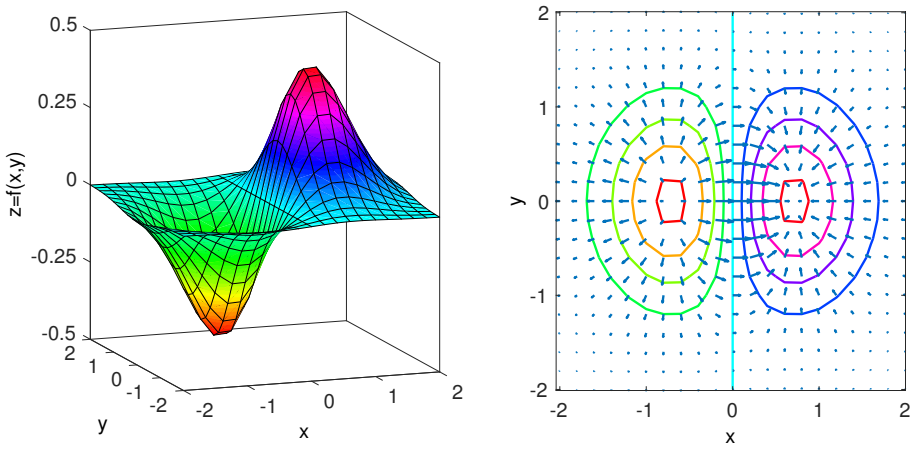


FIGURE IV.3 – Exemple d'une surface représentative d'une fonction $F : \mathbb{R}^2 \longrightarrow \mathbb{R}$. Quelques courbes de niveaux sont aussi représentées avec les gradients. Ici il s'agit de la fonction $F(x, y) = xe^{-x^2-y^2}$.

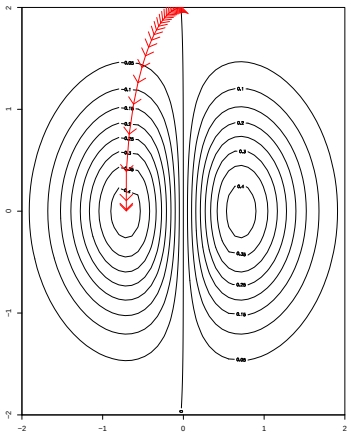
Les algorithmes de descente de gradient sont utilisés pour trouver numériquement le minimum d'une fonction (ou du moins s'en approcher le plus près possible). Minimiser une fonction $F : \mathbb{R}^p \longrightarrow \mathbb{R}$ de classe $\mathcal{C}^1$, revient à trouver $\bar{x} \in \mathbb{R}^p$ tel que

$$F(\bar{x}) = \min_{x \in \mathbb{R}^p} F(x) \ .$$

Toutes ces méthodes sont des méthodes itératives : elles débutent par le choix d'un point initial x_0 considéré comme une première ébauche de la solution (étape d'initialisation), puis procèdent par itérations au cours desquelles elles déterminent une succession de solutions approximatives raffinées qui se rapprochent graduellement de la solution cherchée $\bar{x}$. Les points générés sont appelés des itérés.

Un moyen de calculer $\bar{x}$ est donc de résoudre $\nabla F(x) = 0$. Mais cela ne garantit pas qu'il s'agisse d'un minimiseur et la résolution peut être complexe. On utilise plutôt des algorithmes dits "de descente", dans lesquels F décroit à chaque étape (Figure IV.4).

FIGURE IV.4 – Descente de gradient à pas constant pour la fonction $F(x, y) = xe^{-x^2-y^2}$, avec un pas égal à 1, un point initial égal à $(0; 2)$, et un seuil de tolérance $\varepsilon = 0.001$. La convergence est atteinte après 29 itérations.



Méthode de gradient :

— Initialisation
  — choisir $x_0 \in \mathbb{R}^p$ et poser $k = 0$
  — choisir le pas $\tau > 0$
  — fixer un seuil de convergence (ou critère d'arrêt) $\varepsilon > 0$
— Itérations (boucles sur k)
  — calculer $\nabla F(x_k)$
  — choisir comme direction de descente $-\nabla F(x_k)$
  — déterminer x_{k+1} selon la formule $x_{k+1} = x_k - \tau \nabla F(x_k)$
  — test de convergence : $||x_{k+1} - x_k|| < \varepsilon$ ou $|F(x_{k+1}) - F(x_k)| < \varepsilon$
— Fin de l'algorithme lorsqu'un test de convergence est vérifié, ou lorsque le nombre maximum d'itérations est dépassé.

4.4 La rétro-propagation du gradient

L'algorithme de rétro-propagation du gradient est utilisé dans le perceptron multi-couche lors de l'apprentissage. On présente au réseau des entrées et on lui demande de modifier sa pondération de telle sorte que l'estimation s'approche de la sortie désirée.

Pour minimiser la fonction de coût (i.e. minimiser l'erreur) l'algorithme calcule son gradient qui fournit une indication sur la direction et l'amplitude du changement à effectuer sur les poids W [11]. L'algorithme de rétro-propagation du gradient ne garantit pas d'aboutir au minimum global de la fonction de coût, mais celui-ci est beaucoup plus rapide que les autres méthodes permettant de trouver le minimum global d'une fonction.

L'algorithme consiste dans un premier temps à propager vers l'avant les entrées jusqu'à obtenir une sortie calculée par le réseau (forward pass). La seconde étape compare la sortie calculée à la sortie désirée pour chaque observation. On modifie alors les poids de telle sorte qu'à l'itération suivante, l'erreur commise entre la sortie calculée et la sortie désirée soit minimisée. La présence de couches cachées ne permet pas un accès direct aux erreurs commises par les neurones cachés. C'est pourquoi on rétro-propage l'erreur commise vers l'arrière jusqu'à la couche d'entrée tout en modifiant les pondérations (backward pass). On répète ce processus sur tous les exemples jusqu'à ce que l'on obtienne une erreur de sortie considérée comme acceptable.

Notations

Les notations utilisées ici sont celles proposées par [Hayk94] :

— les indices i, j, k font référence à des neurones différents : si on considère que le signal se propage de gauche à droite dans le réseau, le neurone j fera partie d'une couche à droite du neurone i et le neurone k fera partie d'une couche à droite de celle contenant le neurone j (le neurone j est alors dans une couche cachée).
— à l'itération n, la $n^{\text{ème}}$ observation d'apprentissage est présentée au réseau

11. En réalité, pour des raisons de complexité on utilise une estimation du gradient calculée sur une seule observation. C'est plutôt une méthode de gradient stochastique.

— $e_j(n)$ représente l'erreur du signal en sortie du neurone j à la $n^{\text{ème}}$ itération

— $\mathcal{E}(n)$ représente la somme des erreurs au carré à la $n^{\text{ème}}$ itération (i.e. pour la $n^{\text{ème}}$ observation). On note $\mathcal{E}_{av}$, la moyenne des $\mathcal{E}(n)_n$ (i.e. la moyenne des erreurs sur l'ensemble total d'apprentissage).

— pour k un neurone d'une couche de sortie, $d_k(n)$ représente la valeur attendue pour le neurone k i.e. notre observation. On l'utilise donc pour calculer $e_k(n)$.

— on note $f_j(.)$ la fonction d'activation du neurone j

— $w_{ij}(n)$ représente le poids synaptique reliant la sortie du neurone i avec l'entrée du neurone j à l'itération n. La correction apportée à ce poids à l'itération n est noté $\Delta w_{ij}(n)$.

— b_j représente le biais appliqué au neurone j. Son effet est toujours caractérisé par un poids synaptique $w_{0j} = b_j$ relié à une entrée constante égale à $+1$.

— $\nu_j(n)$ représente la somme pondérée par les poids (y compris le biais) des entrées synaptiques. C'est le signal qui sera appliqué à la fonction d'activation associée au neurone j.

— $y_j(n)$ représente le résultat en sortie du neurone j, après application de la fonction d'activation, à l'itération n : $y_j(n) = f_j(\nu_j(n))$

— on note $x_i(n)$ le $i^{\text{ème}}$ élément d'un vecteur d'entrée

— on note $o_k(n)$ le $k^{\text{ème}}$ élément de la sortie globale

— on note τ, le taux-d'apprentissage (paramètre)

— m_l désigne le nombre de neurones de la couche l, avec $l = 0, 1, ..., L$. L est donc la profondeur du réseau. Ainsi, m_0 désigne la taille de la couche d'entrée, m_1 celle de la première couche cachée et m_L celle de la couche de sortie.

Optimisation et règle du delta

Considérons tout d'abord un neurone k appartenant à la couche de sortie.
A l'itération n, l'erreur en sortie du neurone k est définie par :

$$e_k(n) = d_k(n) - y_k(n) \ . \tag{IV.11}$$

On définie l'erreur instantanée du neurone k par $\frac{1}{2}e_k^2(n)$ et l'erreur totale instantanée $\mathcal{E}(n)$ comme la somme des $\frac{1}{2}e_k^2(n)$ sur l'ensemble des neurones de la couche de sortie.

$$\mathcal{E}(n) = \frac{1}{2}\sum_{k=1}^{m_L} e_k^2(n) \ . \tag{IV.12}$$

Soit N le nombre total d'échantillons contenus dans l'ensemble d'apprentissage. Par définition de $\mathcal{E}_{av}$ on a

$$\mathcal{E}_{av} = \frac{1}{N}\sum_{n=1}^{N} \mathcal{E}(n) \ . \tag{IV.13}$$

Ainsi $\mathcal{E}(n)$ et donc $\mathcal{E}_{av}$ dépendent des paramètres (poids synaptiques et biais) du réseau. Étant donné un ensemble d'apprentissage, $\mathcal{E}_{av}$ représente la fonction de coût (mesure de la performance de l'apprentissage) : l'objectif va donc être de minimiser $\mathcal{E}_{av}$ pour ajuster au mieux les paramètres du réseau. Il va donc falloir calculer le gradient de $\mathcal{E}(n)$, ou du moins ses dérivées partielles par rapports aux poids synaptiques.

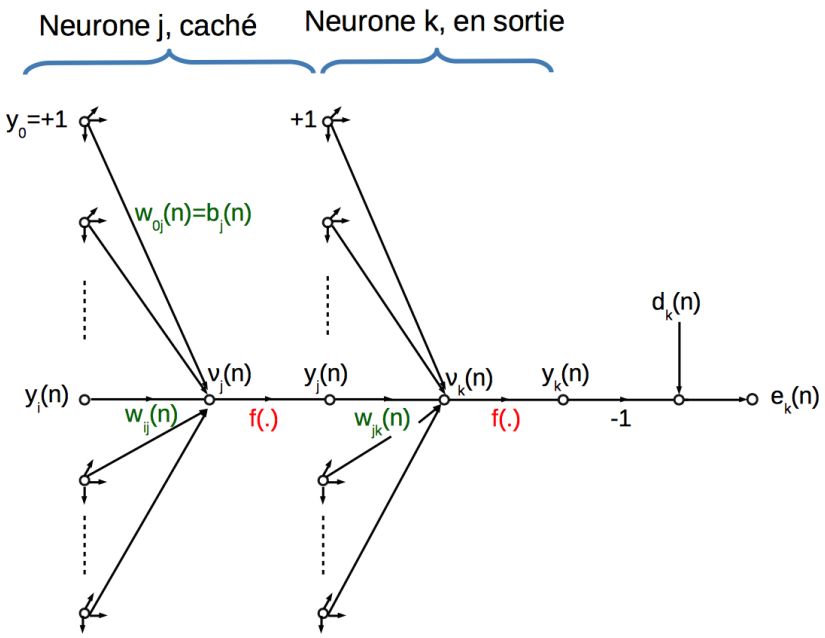


FIGURE IV.5 – Propagation du signal d'un neurone j d'une couche cachée à un neurone k de la couche de sortie (inspirée de [Hayk94]).

L'information qui parvient en entrée de la fonction d'activation d'un neurone k vaut :

$$\nu_k(n) = \sum_{j=0}^{m_{L-1}} w_{jk}(n) y_j(n), \tag{IV.14}$$

où m_{L-1} est le nombre total de neurones présents dans la couche précédente (cachée) ; biais non compris : le poids w_{0k} correspond à l'entrée fixe $y_0 = +1$ et est égal au biais b_k appliqué au neurone k. Ainsi, le signal en sortie du neurone k à l'itération n vaut

$$y_k(n) = f_k(\nu_k(n)). \tag{IV.15}$$

L'algorithme de rétro-propagation du gradient consiste à corriger les poids w_{jk} en ajoutant le terme $\Delta w_{jk}(n)$, lui même proportionnel à la dérivée partielle $\frac{\partial \mathcal{E}(n)}{\partial w_{jk}(n)}$. En utilisant la méthode de dérivée en chaine, on obtient l'expression :

$$\frac{\partial \mathcal{E}(n)}{\partial w_{jk}(n)} = \frac{\partial \mathcal{E}(n)}{\partial e_k(n)} \frac{\partial e_k(n)}{\partial y_k(n)} \frac{\partial y_k(n)}{\partial \nu_k(n)} \frac{\partial \nu_k(n)}{\partial w_{jk}(n)}. \tag{IV.16}$$

Or,

* $\frac{\partial \mathcal{E}(n)}{\partial e_k(n)} = e_k(n)$, en dérivant l'équation (IV.12) par rapport à $e_k(n)$
* $\frac{\partial e_k(n)}{\partial y_k(n)} = -1$, en dérivant l'équation (IV.11) par apport à $y_k(n)$
* $\frac{\partial y_k(n)}{\partial \nu_k(n)} = f_k'(\nu_k(n))$, en dérivant l'équation (IV.15) par apport à $\nu_k(n)$
* $\frac{\partial \nu_k(n)}{\partial w_{jk}(n)} = y_k(n)$, en dérivant l'équation (IV.14) par apport à $w_{jk}(n)$.

Ainsi, on peut réécrire l'équation (IV.16) de la façon suivante :

$$\frac{\partial \mathcal{E}(n)}{\partial w_{jk}(n)} = -e_k(n) f_k'(\nu_k(n)) y_j(n). \tag{IV.17}$$

Le terme correctionnel ajouté au poids $w_{jk}(n)$ est défini en utilisant la règle du delta, comme on a pu la voir pour la descente de gradient à la Section 4.3 :

$$\Delta w_{jk}(n) = -\tau \frac{\partial \mathcal{E}(n)}{\partial w_{jk}(n)}, \tag{IV.18}$$

où τ est le taux d'apprentissage de l'algorithme de rétro-propagation.
En introduisant l'équation (IV.17) dans (IV.18) on obtient

$$\Delta w_{jk}(n) = \tau \delta_k(n) y_j(n), \tag{IV.19}$$

avec le gradient local $\delta_k(n)$ défini par $\delta_k(n) = -\frac{\partial \mathcal{E}(n)}{\partial \nu_k(n)}$ ou encore

$$\begin{aligned} \delta_k(n) &= \frac{\partial \mathcal{E}(n)}{\partial e_k(n)} \frac{\partial e_k(n)}{\partial y_k(n)} \frac{\partial y_k(n)}{\partial \nu_k(n)} \\ &= e_k(n) f_k'(\nu_k(n)). \end{aligned} \tag{IV.20}$$

La valeur de $\delta_k(n)$ est proportionnelle à : (1) la différence entre la valeur désirée et la valeur calculée par le réseau (en effet, s'il n'y a pas de différence on ne doit pas corriger), et (2) à la dérivée de la fonction d'activation (plus la pente est forte, plus on doit corriger).
Le nouveau poids se calcule alors en partant de l'ancien poids et en ajoutant une correction proportionnelle d'un côté à l'erreur, et de l'autre à la valeur reçue de la couche intermédiaire.

On considère désormais un neurone j appartenant à une couche cachée.

Même si les neurones des couches cachées ne sont pas directement accessibles, ils ont aussi une part de responsabilité dans la valeur des erreurs en sortie. Il faut cependant que le neurone intermédiaire qui soit le plus responsable des erreurs soit le plus corrigé. On ne dispose cependant pas de valeur désirée en sortie d'un neurone lorsque celui-ci appartient à une couche cachée. C'est pourquoi il va falloir déterminer les erreurs cachées de façon récursive à partir des erreurs des neurones auxquels le neurone caché est connecté en sortie.
Par analogie avec l'équation (IV.20), on définit le gradient local $\delta_j(n)$ d'un neurone caché j par

$$\begin{aligned} \delta_j(n) &= -\frac{\partial \mathcal{E}(n)}{\partial y_j(n)} \frac{\partial y_j(n)}{\partial \nu_j(n)} \\ &= -\frac{\partial \mathcal{E}(n)}{\partial y_j(n)} f_j'(\nu_j(n)). \end{aligned} \tag{IV.21}$$

Pour calculer la dérivée partielle $\frac{\partial \mathcal{E}(n)}{\partial y_j(n)}$ on différentie tout d'abord l'équation (IV.12) par rapport à $y_j(n)$:

$$\frac{\partial \mathcal{E}(n)}{\partial y_j(n)} = \sum_{k=1}^{m_L} e_k(n) \frac{\partial e_k(n)}{\partial y_j(n)}, \tag{IV.22}$$

puis la règle de dérivation en chaîne pour $\frac{\partial e_k(n)}{\partial y_j(n)}$ (elle permet de faire le lien entre la sortie $y_j(n)$ du neurone caché j et la somme pondérée $\nu_k(n)$ en entrée des neurones de

la couche de sortie).

$$\frac{\partial \mathcal{E}(n)}{\partial y_j(n)} = \sum_{k=1}^{m_L} e_k(n) \frac{\partial e_k(n)}{\partial \nu_k(n)} \frac{\partial \nu_k(n)}{\partial y_j(n)}. \tag{IV.23}$$

Comme $e_k(n) = d_k(n) - y_k(n) = d_k(n) - f_k(\nu_k(n))$ on a

$$\frac{\partial e_k(n)}{\partial \nu_k(n)} = -f'_k(\nu_k(n)), \tag{IV.24}$$

avec $\nu_k(n) = \sum_{j=0}^{m} w_{jk}(n) y_j(n)$ (m est le nombre total de neurones - biais exclu - en entrée du neurone de sortie k). En différentiant cette équation de $\nu_k(n)$ par rapport à $y_j(n)$, on obtient $\frac{\partial \nu_k(n)}{\partial y_j(n)} = w_{jk}(n)$. Il suffit d'utiliser cette équation et l'équation (IV.24) pour remplacer dans (IV.23) :

$$\begin{aligned} \frac{\partial \mathcal{E}(n)}{\partial y_j(n)} &= -\sum_{k=1}^{m_L} e_k(n) f'_k(\nu_k(n)) w_{jk}(n) \\ &= -\sum_{k=1}^{m_L} \delta_k(n) w_{jk}(n), \end{aligned} \tag{IV.25}$$

où on a utilisé l'expression du gradient local $\delta_k(n)$ donnée à l'équation (IV.20). Finalement en regroupant (IV.21) et (IV.25) on obtient la formule de rétro-propagation pour le gradient local $\delta_j(n)$ d'un neurone caché j :

$$\Delta w_{ij}(n) = \tau \frac{\partial \mathcal{E}(n)}{\partial w_{ij}(n)} = \tau \frac{\partial \mathcal{E}(n)}{\partial y_j(n)} \frac{\partial y_j(n)}{\partial w_{ij}(n)} = \tau \delta_j(n) y_i(n),$$

avec

$$\delta_j(n) = f'_j(\nu_j(n)) \sum_{k=1}^{m_L} \delta_k(n) w_{jk}(n). \tag{IV.26}$$

Cette erreur est aussi proportionnelle à la dérivée de la fonction d'activation, et à la somme pondérée des erreurs des neurones de la couche suivante auxquels le neurone j est relié. Ainsi, le neurone de la couche suivante ayant le plus gros poids et la plus grosse erreur interviendra le plus dans la correction.

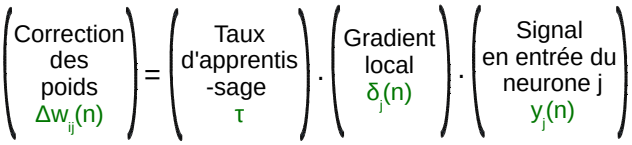


FIGURE IV.6 – Rétro-propagation du signal des erreurs de la couche de sortie vers un neurone de la couche cachée antérieur.

Dans l'équation (IV.26) le facteur $f'_j(\nu_j(n))$ dépend seulement de la fonction d'activation associée au neurone caché j tandis que dans la somme, $\delta_k(n)$ requiert la connaissance des erreurs $e_k(n)$, pour tous les neurones k présents dans la couche suivante - i.e. celle juste à droite de celle comportant le neurone caché j - et directement connectés au neurone j. $w_{jk}(n)$ sont les poids synaptiques associés à ces connections (Figure IV.6).

En résumé

— le terme de correction apporté aux poids synaptiques reliant le neurone i au neurone j est défini selon la règle du delta (Figure IV.6)

— l'expression du gradient local $\delta_j(n)$ dépend de la position du neurone j dans le réseau : caché ou en sortie

* si le neurone j est en sortie, $\delta_j(n)$ est égal au produit de la dérivée $f'_j(\nu_j(n))$ par l'erreur $e_j(n)$ (tous les deux étant associés au neurone j) (Eq IV.20)
* si le neurone j est dans une couche cachée, $\delta_j(n)$ est égal au produit de la dérivée $f'_j(\nu_j(n))$ par la somme pondérée des $(\delta_k(n))_k$ eux mêmes calculés pour les neurones de la couche suivante (cachée ou de sortie) qui sont connectés au neurone j (Eq IV.26).

L'algorithme

L'algorithme de rétro-propagation du gradient se décline comme suit :

Initialisation
Tirer les poids w_{ijk} selon une loi uniforme sur $[0, 1]$
Normaliser les données d'apprentissage dans $[0, 1]$
while $\mathcal{E}_{av} > errmax$ ou $Niter < Itermax$ **do**
. Ranger la base d'apprentissage dans un ordre aléatoire
. **for** chaque élément $n = 1, ..., N$ de la base **do**
. Calculer $e_k(n) = d_k(n) - y_k(n)$ en propageant les entrées vers l'avant.
. Mise à jour des poids de la couche de sortie par : $\Delta w_{jk}(n) = \tau\delta_k(n)y_j(n)$ avec $\delta_k(n) = e_k(n)f'_k(\nu_k(n))$
. L'erreur est "rétro-propagée" dans les différentes couches afin d'affecter à chaque entrée une responsabilité dans l'erreur globale par les équations de rétro-propagation :
. Mise à jour des poids des couches cachées par : $\Delta w_{ij}(n) = \tau\delta_j(n)y_i(n)$ avec $\delta_j(n) = f'_j(\nu_j(n)) \sum_k \delta_k(n)w_{jk}(n)$
. **end for**
end while

À la fin de chaque retro-propagation (i.e. dans les itérations de la boucle "for" - par exemple pour $n = 1$), ce sont les poids précédemment corrigés qui servent pour la propagation suivante (par exemple pour $n = 2$).

5 Conclusion sur les différents modèles utilisés dans cette thèse

Le Tableau IV.1 résume les avantages et les inconvénients des trois modèles d'impact décrits précédemment (LIN, ME et NN). Par nature, un modèle est imparfait et toute modélisation statistique demande des choix et des compromis. En comparant ces modèles sur les mêmes bases de données, cette thèse quantifie à quel point la qualité de l'apprentissage est affectée par leurs inconvénients, ou améliorée par leurs avantages.

Dans cette étude, la qualité des modèles se trouve essentiellement limitée par la qualité des données d'apprentissage et en particulier, leur nombre.

Modèle	**Avantages**	**Inconvénients**
LIN pooling	- plus de données disponibles - simple	- mélange de populations aux comportements différents - pas de prise en compte des groupes (géographiques ici)
LIN no-pooling	- spécialisé localement - simple	- très peu de données, - risques de sur-apprentissage
ME	- plus de données disponibles - prise en compte des spécificités locales - partage de l'information entre groupe - modèle linéaire ou non-linéaire	- implémentation plus complexe - effort de calcul plus important
NN	- simple à implémenter - modèle non-linéaire	- interprétation moins aisée de l'apprentissage - nécessité d'être régularisé

TABLEAU IV.1 – Avantages et inconvénients des modèles linéaires (LIN), à effets mixtes (ME), et des réseaux de neurones (NN).

Il existe bien entendu des dizaines d'autres approches statistiques qui peuvent être considérées pour modéliser l'impact de la météo sur les rendements de maïs (par exemple, les arbres de régression) : cependant, au vu des résultats développés dans les chapitres suivants, on verra que ce n'est pas le choix de la méthode statistiques qui influence le plus la qualité des résultats mais bien le contenu en information présent dans les prédicteurs météorologiques qui demeure le principal facteur limitant.

Dans cette thèse, la programmation a été réalisée à l'aide du logiciel Matlab (package *fitlme* pour les modèles mixtes, ou *nnet* pour les réseaux de neurones) mais d'autres logiciels pourraient aussi servir comme le logiciel R (package *nlme* ou *neuralnet*).

6 Qualité et validation des modèles

La validation des modèles est d'une grande importance dans la construction des modèles d'impacts car les bases de données sont souvent très limitées, et car la capacité du modèle à généraliser doit être quantifiée pour connaître son efficacité. Bien que de plus en plus considérée, cette étape essentielle n'est pas toujours suffisamment considérée dans la littérature, souvent par manque de connaissances en apprentissage statistique.

6.1 Sur-apprentissage et dilemme biais-variance

Un modèle trop simpliste fait des erreurs importantes car son manque de complexité ne lui permet pas de représenter correctement la diversité de la cible (Figure IV.7) : sa sortie varie peu (faible variance), mais est souvent loin des données à estimer (fort biais) [Sess06]. Un modèle très complexe (avec un nombre important de paramètres par exemple) peut aussi commettre des erreurs importantes car même s'il est capable d'apprendre très précisément l'ensemble d'apprentissage (faible biais), sa trop forte capacité à coller aux données d'apprentissage lui donne un comportement très variable, donc une forte variance [Burn02]. C'est le cas du modèle rouge à la Figure IV.8.

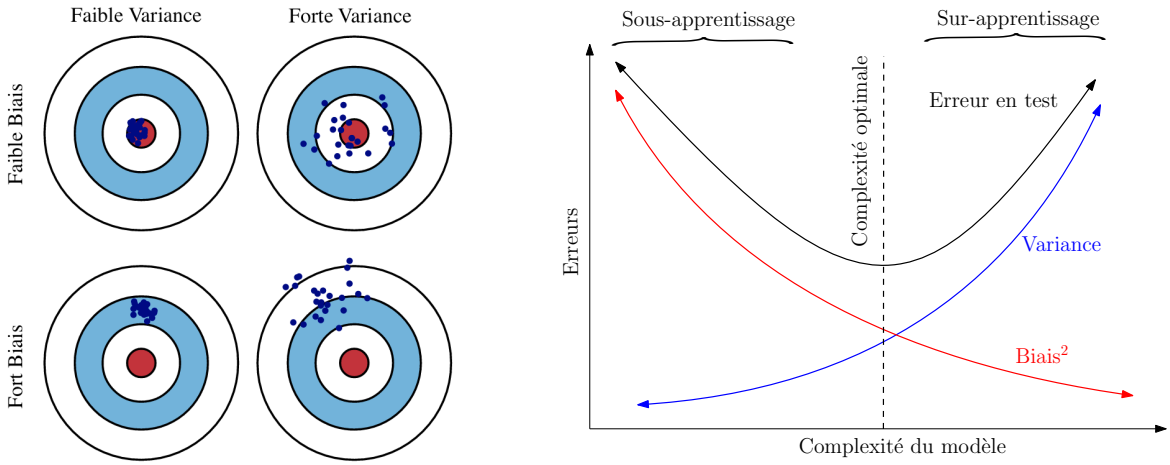


FIGURE IV.7 – Gauche : Illustration graphique du biais et de la variance.
Droite : Contribution du biais et de la variance à l'erreur totale en fonction de la complexité du modèle (inspirée de [Hast01]).

La Figure IV.8 représente un jeu de données où les croix bleues sont les données qui ont servi pour l'apprentissage et les croix rouges sont celles qui serviront pour le test. Pour réaliser la régression de y par rapport à x, on se propose de tester deux modèles, un polynôme de degré 9 (en rouge) et un polynôme de degré 2 (en noir). L'apprentissage par le polynôme de degré 9 est parfait sur l'ensemble d'entrainement car la courbe de régression passe par tous les points et l'erreur d'apprentissage est donc nulle. En revanche, lorsque l'on étudie la capacité de ce modèle à généraliser sur un ensemble de test (croix rouges) on se rend compte que les erreurs sont énormes. En effet, la complexité de ce modèle face au faible nombre de données est telle que les libertés prises par ce modèle pour coller aux données d'apprentissage sont très grandes : c'est le problème du sur-apprentissage. Lorsque l'on regarde la régression fournie par le polynôme d'ordre 2, on a certes des erreurs plus élevées en apprentissage mais beaucoup plus faible en test (et comparable aux erreurs en apprentissage). Ce modèle bien moins complexe réalise un meilleur compromis entre biais et variance.

On dit qu'un modèle sur-apprend lorsqu'il donne de bons résultats sur les données d'apprentissage mais commet des erreurs importantes sur des données nouvelles (par exemple, sur les données de test). Le modèle ne parvient pas à généraliser [Hawk04, Bish96]. Le sur-apprentissage survient quand un modèle est trop complexe, par exemple lorsqu'il a trop de paramètres et pas assez de données pour les estimer correctement. La qualité de généralisation d'un modèle fait référence à sa capacité à donner une bonne estimation sur de nouvelles données.

La réduction de la complexité des modèles passe aussi bien sûr par un nombre restreint de variables d'entrées : la sélection des variables est donc une étape très importante. Elle sera développée à la Section 7.2.

Formulation du compromis biais/variance

L'erreur due au biais est la différence entre la prédiction moyenne du modèle et la valeur correcte que l'on souhaite prédire. Il faut imaginer que l'on dispose de plusieurs ensembles d'apprentissages différents et que l'on regarde les différentes valeurs proposées par le modèle pour ces différents ensembles. Ainsi, on peut calculer une prédiction

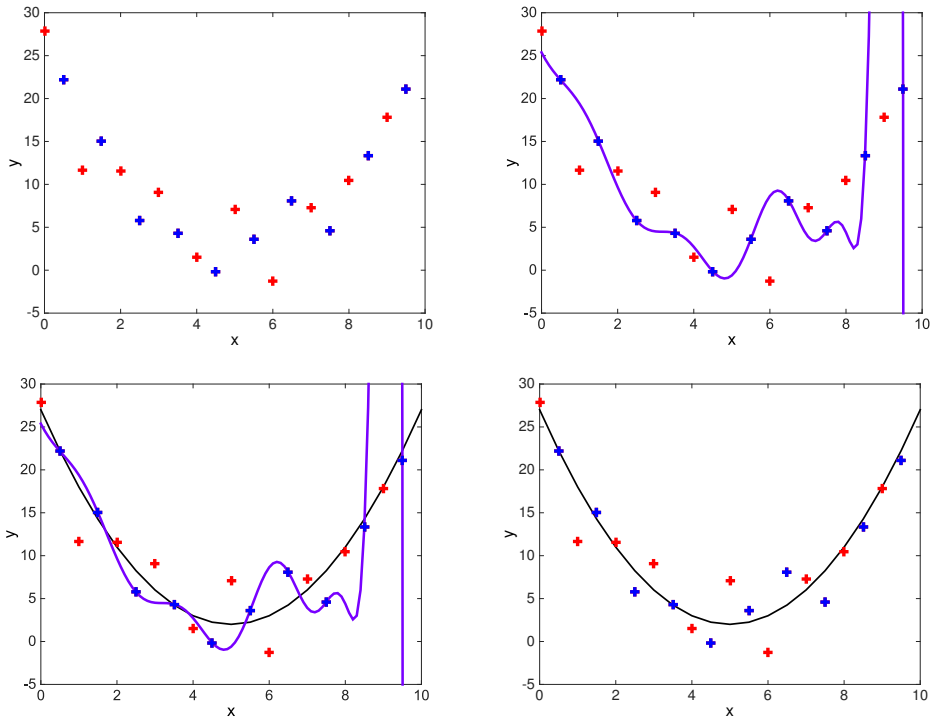


FIGURE IV.8 – Différentes complexité de modèles de régression. Les points bleus sont les données d'apprentissage et ceux en rouge sont celles du test. La courbe rouge est la régression polynomiale de degré 9, et la courbe en noir est la régression polynomiale de degré 2.

moyenne de la valeur désirée. Le biais mesure à quel point la prédiction moyenne sera éloignée de la valeur correcte.
L'erreur due à la variance est la variabilité des prédictions du modèle pour une valeur donnée que l'on souhaite prédire. De la même façon que pour le biais, imaginons que l'on dispose de plusieurs prédictions pour cette valeur correcte, issues de plusieurs ensembles d'apprentissage fournis au modèle. La variance quantifie à quel point les différentes prédictions varient quand on change les ensembles d'apprentissage et donc les réalisations du modèle.

L'approche la plus courante pour l'inférence est la minimisation de l'erreur quadratique moyenne. Supposons que nous avons un ensemble d'apprentissage constitué d'un ensemble de points x_1, ... , x_n et de valeurs réelles y_i associées à chaque point x_i. Nous supposons qu'il existe une relation fonctionnelle bruitée $y_i = f(x_i) + \varepsilon$, où le bruit ε, a une moyenne nulle et une variance σ^2. Trouver une fonction $\hat{f}$ qui se généralise à des points extérieurs à l'ensemble d'apprentissage peut être fait avec l'un des nombreux algorithmes utilisés pour l'apprentissage supervisé. Selon la fonction $\hat{f}$ que nous choisissons, son erreur attendue sur un échantillon test x peut se décomposer comme suit

$$\mathbb{E}\left[(y - \hat{f}(x))^2\right] = \underbrace{\left(\mathbb{E}\left[\hat{f}(x)\right] - f(x)\right)^2}_{Biais[\hat{f}(x)]^2} + \underbrace{\mathbb{E}\left[(\hat{f}(x) - \mathbb{E}[\hat{f}(x)])^2\right]}_{Var\left[\hat{f}(x)\right]} + \sigma^2.$$

où l'espérence porte sur l'aléatoire de $\widehat{f}$. Les trois termes sont :
- le biais au carré de la méthode d'apprentissage, qui peut être vue comme l'erreur due aux hypothèses simplifiées de la méthode utilisée. Par exemple, approcher une fonction non linéaire à l'aide d'une méthode linéaire va produire des erreurs d'estimation dues à cette hypothèse ;
- la variance de la méthode d'apprentissage, ou plus intuitivement, de combien la méthode d'apprentissage $\hat{f}(x)$ se déplace autour de sa moyenne ;
- l'erreur irréductible σ^2.

Étant donné que tous les trois termes sont positifs, cela constitue une limite inférieure sur l'erreur attendue sur des échantillons test.

Estimation de l'erreur de généralisation

L'estimation de l'erreur de généralisation requiert l'utilisation de bases de test indépendantes de la base d'apprentissage. Dans un modèle de régression (linéaire ou non), on vise durant l'apprentissage à obtenir l'écart le plus faible entre les sorties du modèle de régression et les observations réelles. Cette différence faible sur la base d'apprentissage peut être intéressante en tant que telle, néanmoins dans la plupart des cas, l'intérêt principal d'un modèle est sa capacité de généralisation, c'est-à-dire sa capacité à fournir des résultats satisfaisants sur des données indépendantes de celles de la base d'apprentissage.

Comme on vient de le voir, posséder un nombre de données suffisant est une condition nécessaire pour une bonne capacité de généralisation. La capacité à traiter des données non connues de la base d'apprentissage (faculté de généralisation), permet de faire de l'interpolation ou de l'extrapolation [Frie94]. Pour estimer la qualité d'un modèle, on fait généralement appel à plusieurs critères quantifiant cette qualité. Les critères les plus connus sont résumés à la Section 6.3.

Éviter le sur-apprentissage

Des techniques de régularisation peuvent être utilisées pour réduire le problème de sur-apprentissage. Elles peuvent agir sur la représentation des données, sur l'algorithme d'apprentissage, ou encore sur la structure même du modèle [Hast01]. On peut par exemple réduire le nombre d'entrées en analysant les corrélations (voir Section 7.2). On peut aussi garder toutes les entrées disponibles mais utiliser un terme de pénalisation : par exemple pour la régression ridge (pénalisation L^2, comme c'est le cas avec le weight-decay chez les réseaux de neurones), ou lasso (pénalisation L^1), ou élastic-net (une combinaison linéaire de la pénalisation ridge et lasso) [Hast01].
Les techniques d'arrêt anticipé, permettent aussi de limiter le sur-apprentissage lorsque l'apprentissage se fait de façon itérative, comme chez les réseaux de neurones : plus le nombre d'itérations en phase d'apprentissage est grand, plus les erreurs en apprentissage sont faibles, mais plus le modèle est sensible au sur-apprentissage. L'arrêt anticipé consiste à stopper l'apprentissage dès que l'erreur en validation augmente sur plusieurs itérations (Section 6.2).

6.2 Le contrôle de la complexité des réseaux de neurones

Pour un réseau de neurones, l'utilisateur doit effectuer de nombreux choix :

1. Les variables d'entrée et celle(s) de sortie. D'éventuelles transformations de ces variables sont parfois nécessaires.

2. L'architecture du réseau, donc le nombre de paramètres
   — Le nombre de couches cachées (souvent une ou deux) correspond à une aptitude à traiter des problèmes de non-linéarité
   — Le nombre de neurones par couche cachée : d'après le théorème d'approximation universelle une seule couche cachée suffit pour approcher une fonction continue d'un compact de $\mathbb{R}^p$ dans $\mathbb{R}^q$ avec une précision arbitraire, à condition d'avoir suffisamment de neurones dans cette couche. Il faut donc optimiser le nombre de neurones tout en veillant à la complexité du modèle et au problème du sur-apprentissage.

   Ces deux choix conditionnent le nombre de paramètres (les poids) à estimer et donc **la complexité du modèle**. Ils participent à la recherche d'un bon compromis biais-variance.
3. Le nombre maximum d'itérations, l'erreur maximale tolérée, et un éventuel terme de **régularisation** (decay). Augmenter ces critères participe à trouver le bon compromis entre apprentissage et généralisation.
   Une stratégie consiste à considérer un échantillon indépendant de l'ensemble de validation - appelé "ensemble de test" - et à arrêter l'apprentissage lorsque l'erreur sur cet échantillon commence à se dégrader tandis que l'erreur sur l'échantillon d'apprentissage ne peut que continuer à décroître (Figure IV.9). On parle d'arrêt anticipé.

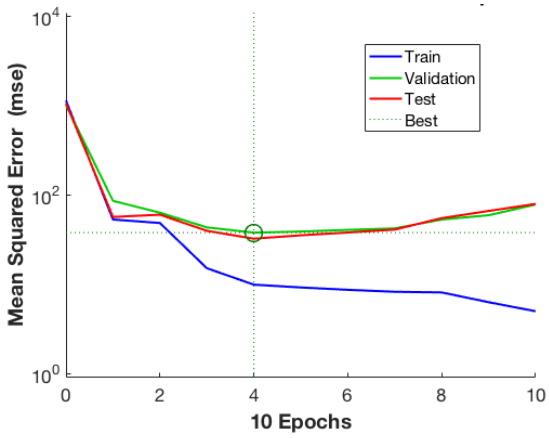


FIGURE IV.9 – Performance d'un réseau (ici mesurée à l'aide des moindres carrés) pour les échantillons d'apprentissage, de validation, et de test à chaque itération de l'apprentissage (données simulées).

4. Le taux d'apprentissage τ : τ peut être choisi comme constant (déterminé par l'utilisateur) ou variable au cours des itérations : il semble raisonnable de penser que, grand au début pour aller vite, ce taux décroisse ensuite au cours des itérations pour aboutir à un réglage plus précis au fur et à mesure que l'algorithme s'approche de la solution. De plus, sa décroissance peut être linaire, géométrique voire exponentielle [Robb51, Bott04, Beng12].

En pratique, on ne peut pas régler ces paramètres en même temps : il faut veiller au sur-apprentissage en limitant le nombre de neurones, ou la durée d'apprentissage ou encore, augmenter le coefficient de pénalisation de la norme des paramètres. Pour cela il faut choisir une méthode d'estimation de l'erreur : échantillons validation/test, ou validation croisée, ou bootstrap.

Comme en régression *ridge*, ou *lasso*, on peut introduire un terme de pénalisation dans la définition de la fonction de coût à minimiser pour éviter le problème du sur-apprentissage. Celui-ci devient alors $\mathcal{E}_{av}(W)+\gamma||W||^2$. On doit alors régler le coefficient de pénalisation γ (decay) : plus celui-ci est important, moins les paramètres (poids) peuvent prendre de trop grandes valeurs, ce qui contribue à réduire le risque de sur-apprentissage. Un γ petit pénalise peu le risque, et l'important est donc davantage de rester proche des données d'apprentissage dans l'estimation. Une stratégie simple (et efficace si la taille de l'échantillon ne permet pas de prendre en compte un ensemble de validation) consiste à introduire un nombre plutôt élevé de neurones puis à optimiser le seul coefficient de pénalisation (decay) par validation croisée.

Le Tableau IV.2 résume les différentes méthodes pour réduire la complexité d'un réseau de neurone.

	Agir sur le modèle	Modifier l'objectif	Modifier l'algorithme
Les modèles complexes tendent à sur-apprendre	On limite leur existence par la nature du modèle.	On les pénalise dans la formulation de l'objectif.	On empêche l'algorithme d'explorer des modèles trop complexes.
Action	Limiter la complexité du modèle, son nombre de paramètres	Ajouter un terme de pénalité	Arrêt anticipé
Paramètres à ajuster	Nombre de couches et de neurones par couche	Ajout d'une pénalité L_1, L_2 à la fonction de coût	Nombre d'époques
Choix des paramètres	Validation croisée ou ensemble de test		Moindres carrés (Figure IV.9)

TABLEAU IV.2 – Leviers pour réduire le risque de sur-apprentissage pour un réseau de neurones.

6.3 Les critères de validation des modèles

Que ce soit pour réduire la dimension en sélectionnant certaines variables explicatives, ou pour comparer des modèles entre-eux, on a besoin de critères mesurant la qualité des modèles. Ils sont nombreux et aucun n'est parfait [Duve16]. On en liste seulement quelques uns ici.

Considérons n données d'anomalies de rendement $y_1, \dots, y_n$, chacune associée à sa prédiction $\widehat{y}_1, \dots, \widehat{y}_n$ par un modèle de régression. Soit $\overline{y}$ la moyenne des données observées.

Le coefficient de détermination R^2 est la fraction des variances du modèle sur les variances observées :

$$R^2 = \frac{SS_{Reg}}{SS_{Tot}} = \frac{\sum\limits_i (\widehat{y}_i - \overline{y})^2}{\sum\limits_i (y_i - \overline{y})^2}.$$

Il représente le pourcentage de variance des anomalies de rendement qui a été expliqué par le modèle. SS_{Tot} est la somme totale des erreurs au carré (proportionnel à la

variance des données), et SS_{Reg} est la somme des erreurs de régression au carré. R^2 est un critère très utilisé pour comparer différents modèles de régression. Plus ce critère est proche de 1, et plus les prévisions sont proches des observations, donc meilleur est le modèle. En pratique, l'utilisation de R^2 doit se faire avec prudence. D'une part, il est important de vérifier sa valeur sur des ensembles d'apprentissage mais aussi sur des ensembles de test. D'autres part, le R^2 est particulièrement sensible au nombre d'échantillon, surtout dans les ensembles de tests qui sont souvent plus petits. L'inconvénient majeur du R^2 est sa tendance à augmenter de façon monotone avec l'introduction de nouvelles variables même si celles-ci sont peu corrélées avec la variable cible, et ne sont donc pas pertinentes. Pour comparer un sous-modèle de $p-1$ variables d'un modèle avec $(p-1+r)$ variables, on peut utiliser le test de Fisher partiel : il nous dit si l'introduction des variables supplémentaires augmente significativement le R^2 ou non. Pour remédier à ce problème, il existe le R^2 ajusté R^2_{ajust}. Le R^2 ajusté est toujours inférieur au R^2.

$$R^2_{ajust} = 1 - \frac{(1-R^2)(n-1)}{n-k-1},$$

où n est le nombre d'échantillons, et k est le nombre de variables explicatives indépendantes (excepté le terme constant).

Une seconde mesure de qualité est le **coefficient de corrélation linéaire** $Corr$ entre les $(y_i)_{i=1:n}$ observés, et les $(\hat{y}_i)_{i=1:n}$ prédits. Pour un modèle linéaire, $Corr$ est exactement la racine de R^2 et ne donne donc pas d'avantage d'informations. Si $A = (y_i)_{i=1:n}$, $\hat{A} = (\hat{y}_i)_{i=1:n}$, et σ_A l'écart-type de A,

$$Corr = \frac{cov\left(A, \hat{A}\right)}{\sigma_A \sigma_{\hat{A}}} = \sqrt{R^2}\,.$$

Une troisième mesure de qualité est **l'erreur quadratique moyenne (ou MSE pour "Mean Squared Error")**. Elle mesure l'écart des prévisions aux observations. Beaucoup de techniques de régression sont basées sur une minimisation de la MSE. Si on utilise les mêmes notations que ci-dessus, on a :

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2\,.$$

Parfois, c'est **la racine du MSE, le RMSE** qui est considérée [Desc80]. Sans apporter davantage d'information, le RMSE a l'avantage de fournir une mesure qui a la même unité que la cible. On a

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}\,.$$

Enfin, lorsque l'on estime un modèle statistique, il est possible d'augmenter la vraisemblance du modèle en ajoutant un paramètre. **Le critère d'information d'Akaike (AIC)** permet de pénaliser les modèles en fonction du nombre de paramètres afin de satisfaire le critère de parcimonie. On choisit alors le modèle avec le critère d'information d'Akaike le plus faible :

$$AIC = 2k - 2\ln(L),$$

où k est le nombre de paramètres du modèle à estimer (paramètres libres) et L est le maximum de la vraisemblance du modèle. Le critère d'information AIC s'applique aux modèles estimés par une méthode du maximum de vraisemblance : les analyses de variance, les régressions linéaires multiples, les régressions logistiques peuvent rentrer dans ce cadre. Le critère AIC représente donc un compromis entre le biais (diminuant avec le nombre de paramètres libres), et la parcimonie (volonté de décrire les données avec le plus petit nombre de paramètres possible). Cependant, la vraisemblance dépendant considérablement de la taille de l'échantillon, le critère AIC ne permet pas de comparer deux modèles dont les ensembles d'apprentissage sont de taille très différente.

Le critère MSE et ceux issus de la log-vraisemblance (dont AIC) sont des statistiques différentes : AIC est basé sur la vraisemblance/probabilité plutôt que sur les moindres carrés. R^2 peut être considéré comme un multiple du MSE, en le comparant à la variance de la réponse, puisque $1 - R^2 = \frac{MSE}{var(y)}$.

La valeur des critères de qualité formulés ci-dessus dépend des échantillons utilisés. L'application de leur formule sur deux échantillons différents ne donnera pas forcément le même résultat. Aussi, pour prendre en compte cette variabilité, il est nécessaire d'avoir recourt à des techniques de validation croisée ou de boostrap. C'est ce qui sera fait dans la suite de ce manuscrit. Un critère de qualité appelé $MSEP$ (mean squared error of prediction) a été mis en place pour faire la différence entre un MSE calculé sur un seul échantillon et plusieurs MSE calculés sur divers échantillons dont la valeur est moyennée à la fin [Wall13]. Par simplification, on continuera dans la suite de ce manuscrit à appeler les critères $Corr$, MSE ou R^2 même si ceux-ci devraient plutôt s'appeler $CorrP$, $MSEP$ et R^2P.

6.4 La méthode des runs d'ensembles

Les méthodes décrites précédemment, visent à améliorer la capacité de généralisation du modèle, pour un apprentissage donné. Cependant, les critères de qualités sont souvent sensibles à la taille de la base utilisée, ce qui entraine une estimation pas toujours robuste de la qualité des modèles. Il est donc nécessaire d'avoir recours à des techniques statistiques supplémentaires pour obtenir des estimations plus réalistes et plus robustes.

La méthode dite "des runs d'ensemble", permet également d'améliorer la généralisation, mais en exploitant l'ensemble des résultats d'une série d'apprentissages effectués sur des données différentes. Par exemple, considérons que nous possédons en tout N_{total} données de rendement. Laissons alors de côté N_{test} données qui constitueront la base de test. Il faut alors séparer les $N_{total} - N_{test}$ données restantes en une base d'apprentissage et une base de validation. L'idée est alors de pratiquer un grand nombre de fois le leave-one-out. Notons N_{lo} ce nombre. Ceci permet non seulement d'augmenter artificiellement la taille de la base de validation mais aussi d'obtenir plusieurs jeux de coefficients. Plusieurs manières de procéder sont alors envisageables. Afin de simplifier le formalisme, nous considérons dans la suite le cas d'un modèle linéaire à une variable.

Coefficients moyennés

On effectue N_{lo} leave-one-out. Ceci est donc équivalent à considérer N_{lo} bases d'apprentissage.

On choisit alors comme jeu de paramètres :

On a donc :

$$\begin{aligned} Y_1 &= a_1 X_1 + b_1 \\ Y_2 &= a_2 X_2 + b_2 \\ &\vdots \\ Y_{N_{lo}} &= a_{N_{lo}} X_{N_{lo}} + b_{N_{lo}} \end{aligned}$$

$$a_{moy} = \frac{1}{N_{lo}} \sum_{i=1}^{N_{lo}} a_i,$$

$$b_{moy} = \frac{1}{N_{lo}} \sum_{i=1}^{N_{lo}} b_i.$$

La prévision sur la base de test est alors :

$$Y = a_{moy} X + b_{moy}.$$

Cela revient donc à régulariser l'indice puisque l'on considère la moyenne et donc le barycentre affecté de poids unité, des paramètres. Les paramètres finaux sont donc nécessairement compris entre le minimum et le maximum. Il s'agit en quelque-sorte d'une contrainte similaire à la régularisation de Tikhonov. Supposons, par exemple, que le jeu de coefficient (a_1 ; b_1) apprenne une caractéristique d'une année particulière correspondant à du bruit (et donc en aucun cas à quelque-chose de généralisable). Les jeux de paramètres issus d'apprentissages ne prenant pas cette année particulière en compte ne présenteront pas ce défaut. Moyenné avec ceux-ci, l'impact du sur-apprentissage dû à (a_1 ; b_1) sera alors limité [Koch10].

Runs d'ensemble

Le principe est de considérer les N_{lo} jeux de paramètres obtenus par leave-one-out et de les appliquer un-à-un sur la base de test. On obtient ainsi un nombre N_{lo} de prédictions sur cette base de test. La prévision est alors la prévision moyenne : $Y = \frac{1}{N_{lo}} \sum_{i=1}^{N_{lo}} Y_i$. Or on sait que $Y_i = a_i X + b_i$, d'où

$$\begin{aligned} Y &= \frac{1}{N_{lo}} \sum_{i=1}^{N_{lo}} a_i X + b_i \\ &= \left(\frac{1}{N_{lo}} \sum_{i=1}^{N_{lo}} a_i \right) X + \left(\frac{1}{N_{lo}} \sum_{i=1}^{N_{lo}} b_i \right) \\ &= a_{moy} X + b_{moy}. \end{aligned}$$

On aboutit donc au fait que, dans le cadre d'un modèle linéaire, prévision moyenne et prévision obtenue en utilisant les paramètres moyennés sont identiques. L'approche par runs d'ensemble s'avère tout de même plus riche puisqu'elle fournit la dispersion des résultats.
Dans cette thèse, nous avons utilisé la méthode des runs d'ensemble pour pouvoir fournir des prévisions sur l'ensemble de la zone géographique étudiée.

6.5 Des variables d'entrées corrélées spatialement

Une validation croisée est souvent appliquée en apprentissage statistique pour vérifier le sur-apprentissage et pour assurer la capacité prédictive d'un modèle. Pour cela, une séparation de l'ensemble global des données doit être fait de façon aléatoire, avec par exemple 80% des données qui serviront pour l'apprentissage et 20% des données restantes qui serviront pour la validation du modèle. Cette étape est réalisée sur une centaine d'ensembles apprentissage/test différents. Ces 100 runs permettent d'avoir une estimation des mesures de qualité (R^2, RMSE...) et aussi de leur variabilité selon

les ensembles de test (incertitudes associées) caractérisée par les écart-types de ces mesures de qualité.

Cependant, il faut être très prudent lorsqu'on a affaire à des données spatio-temporelles car les données issues de pixels proches sont souvent très corrélées. Comme remarqué par [Le R13] ou par [Arau05] (dans un contexte d'écologie), la méthode de validation croisée en présence de données spatio-temporelle doit être adaptée. Sinon, on obtient des résultats qui ne sont pas fiables et donc de fausses conclusions sur la qualité des modèles de régression. Pour la validation croisée, les ensembles d'apprentissage et de validation se doivent d'être indépendants [Arlo10]. Or cette hypothèse critique n'est pas toujours vérifiée, surtout dans les études de climat où la plupart des données (températures moyennes mensuelles, et précipitations) sont spatialement corrélées (même à l'échelle des états). Cela complique la comparaison des études et de leur qualité.

Pour résoudre ce problème, on se doit d'utiliser des données issues d'années différentes de celle présentes dans l'ensemble d'apprentissage, pour l'ensemble de test. Ainsi, pour définir un ensemble d'apprentissage, on choisit aléatoirement 28 années (environ 80% des 35 années disponibles), et les données des 7 années restantes constitueront l'ensemble de test.

7 Modèle de tendance, et sélection de variables

Comme nous l'avons signalé dans l'introduction générale, la méthode de prévision des rendements retenue dans cette thèse est basée sur l'estimation de deux composantes : la première est liée à la tendance temporelle des rendements agricoles, et l'autre à l'effet des conditions météorologiques sur les rendements. En pratique, nous étudions d'abord l'évolution des rendements en fonction du temps (évolution lente et non liée à la météo à long terme), indépendamment des conditions météorologiques. Nous nous efforçons ensuite d'expliquer les fluctuations des résidus par rapport à la tendance, par les variables météorologiques.

7.1 L'identification de la tendance temporelle

La production de maïs est influencée par les facteurs météorologiques mais aussi - et de façon non moins négligeable - par d'autres facteurs tels que la fertilisation industrielle, l'irrigation, le type de sol, ou encore la mécanisation. Puisque l'on cherche à estimer l'impact seul des facteurs météorologiques sur les rendements de maïs il nous faut séparer au mieux l'impact météorologique des autres impacts.

On note y_{it} le rendement du canton i pour l'année t.
Comme dans [Kotl07b], l'identification des facteurs non météorologiques est basée sur l'hypothèse qu'ils n'affectent les rendements que sur des variations à long terme, à l'inverse des facteurs météorologiques qui eux, agissent davantage à court terme. La météo d'une année influe sur le rendement de l'année. L'approche qui consiste à définir une tendance - c'est-à-dire une fonction qui représente l'évolution à long terme de la grandeur étudiée, et qui traduit l'aspect général de la série - pour se concentrer sur l'impact météorologique seul, s'appuie sur la même hypothèse.

Pour cela on définit pour chaque série temporelle de rendements y_i une tendance $\widetilde{y}_i$ et l'on considère les anomalies des rendements c'est-à-dire l'écart des valeurs de rendements par rapport à la tendance $y_{it} - \widetilde{y}_{it}$. De plus, de façon à éviter que la valeur de la tendance induise un biais entre les différentes anomalies, on considère que l'impact

météorologique dépend aussi de la valeur de la tendance en définissant finalement l'impact météorologique par le pourcentage de variation suivant :

$$\frac{y_t - \widetilde{y}_t}{\widetilde{y}_t}.$$

Dans les études qui considèrent aussi des tendances de rendement (mais qui ne considèrent pas forcément des anomalies comme calculé ici) les tendances les plus utilisées sont : (1) une tendance linéaire [Kotl07b], ou linéaire avec un plateau de stagnation [Lin12], (2) polynomiale (degré 5 et 15 chez [Kotl07b]), et (3) logarithmique [Thom86, Thom88]. [Mero13] ont utilisé un modèle à effets aléatoires pour déterminer leur tendance, et [Kowa14] une régression des moindres carrés partiels (PLS). A notre connaissance, les modèles à effets mixtes n'ont jamais été utilisés pour l'identification d'une tendance temporelle des rendements.

[Palm93] ont comparé 10 modèles de tendance pour les séries temporelles de rendement nationaux de plusieurs cultures en Europe (pour des séries temporelles allant de 11 à 29 années).

(1)	$y = a + b_1 t,$	(2)	$y = a + b_1 t + b_2 t^2$
(3)	$y = a + b_1/t,$	(4)	$y = a + b_1 \log t,$
(5)	$y = a + b_1 t + b_2 t^2 + b_3/t + b_4 \log t,$	(6)	$\log y = a + b_1 t,$
(7)	$\log y = a + b_1 t + b_2 t^2,$	(8)	$\log y = a + b_1/t,$
(9)	$\log y = a + b_1 \log t,$	(10)	$\log y = a + b_1 t + b_2 t^2 + b_3/t + b_4 \log t.$

La comparaison des résultats montre que de tous les modèles testés, c'est le modèle quadratique (2) qui donne les meilleurs résultats, et peu de différences sont observées avec le modèle linéaire (1). [Swan79] se sont basés sur 27 années de rendement (de maïs et de soja aux États-Unis) et ont testé plusieurs tendances dont une tendance linéaire du temps et plusieurs non linéaires ($y = a + bt^2$, $y = a + b\sqrt{t}$, $y = a + b_1 t + b_2 t^2$, $y = a + b_1 t + b_2\sqrt{t}$). [Saka78] propose une tendance linéaire sur 10 ans (1940 à 1950) puis un plateau jusqu'en 1972 pour l'étude du blé en Afrique du Sud. Il conclut que les effets de la technologie ne sont pas forcément indépendants de la météo, et face au manque de variables quantifiant cette technologie il considère que la tendance temporelle technologique dépend de la variation des résidus (même considération dans [Agra82] pour les rendements de riz dans un district de l'Inde). Selon [Houg90] et [Smit75], la météo affecte le choix des pratiques agricoles, comme la surface plantée, le calendrier et la durée des opérations, l'utilisation de produits phytosanitaires, etc. Ils pensent donc que la tendance temporelle doit être analysée en même temps que les variables explicatives météorologiques. [Swan79] ont montré que la tendance était sous-estimée lorsque les données météorologiques n'étaient pas analysées en même temps que la tendance. Des résultats similaires sont donnés pour le millet au Botswana dans [Voss90]. Ces considérations ne seront pas prises en compte dans cette thèse : en effet, le fait de ne pas avoir la même tendance au sud ou au nord des États-Unis (par exemple) prend quand même en compte la dépendance de la tendance aux conditions météorologiques. L'objectif ici n'est pas d'expliquer/modéliser le lien tendance-météo. La plus longue série temporelle possible est considérée afin de trouver le meilleur ajustement pour décrire l'évolution à long terme.
Une étude future pourra tester une identification de la tendance temporelle différente,

mais au vu des résultats sur la tendance (que l'on considère comme bons) on ne s'attend pas a voir de grandes modifications.

Le choix de la tendance dépend bien sûr du comportement de la culture étudiée, de la région considérée (développement économique plus ou moins rapide, effet technologique plus ou moins important), et de la longueur des séries temporelles dont on dispose. Plus cette longueur est courte et moins on a d'information, donc plus il est facile de se tromper et difficile de dégager une forme d'évolution moyenne des rendements au cours du temps.

Dans cette thèse, où l'on dispose de séries temporelle de plus de 100 ans, plusieurs tendances ont été testées : (1) linéaire, (2) linéaire par morceaux, (3) polynomiale de degré 2 et 3, et (4) logistique.

Tendance linéaire par morceaux, continue

L'analyse graphique des séries temporelle du rendement de 1910 à 2013 suffit à rejeter l'utilisation d'une tendance linéaire. Le choix d'une tendance linéaire par morceaux peut se justifier par le fait que la plupart des courbes possèdent des points de rupture. Ce sont des points où la série change complètement allure. Ils sont normalement explicables (Figure IV.10), et imposent une analyse séparée de la série, par morceaux.

Les rendements de grain historiques nous fournissent un aperçu de la tendance suivie par le rendement de maïs à l'échelle des États-Unis (Figure IV.10). De 1866 à environ 1936, les rendements de maïs aux États-Unis étaient assez constants et ont atteint en moyenne environ 26 bu/acre pendant cette période de 70 ans. Curieusement, les données historiques ne montrent aucun changement notable de la productivité pendant toute cette période. L'adoption du maïs hybride "double cross" par des agriculteurs après les années "Dust Bowl" a abouti à la première amélioration significative de la productivité et a permis un taux annuel d'amélioration de rendement d'environ 0.8 bu/ac/an de 1937 à 1955 environ.

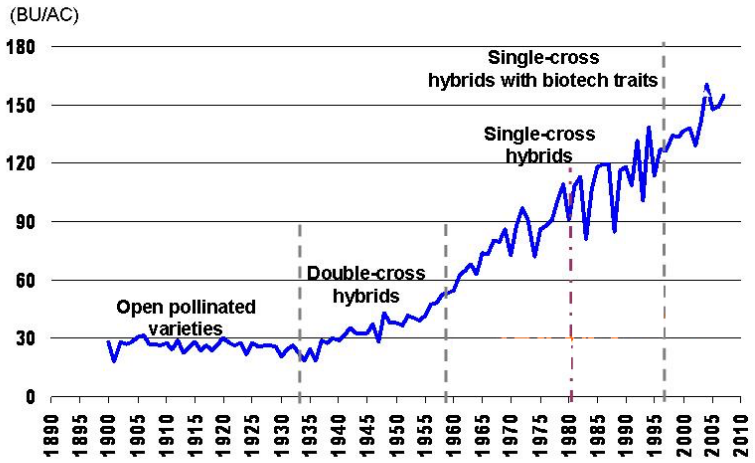
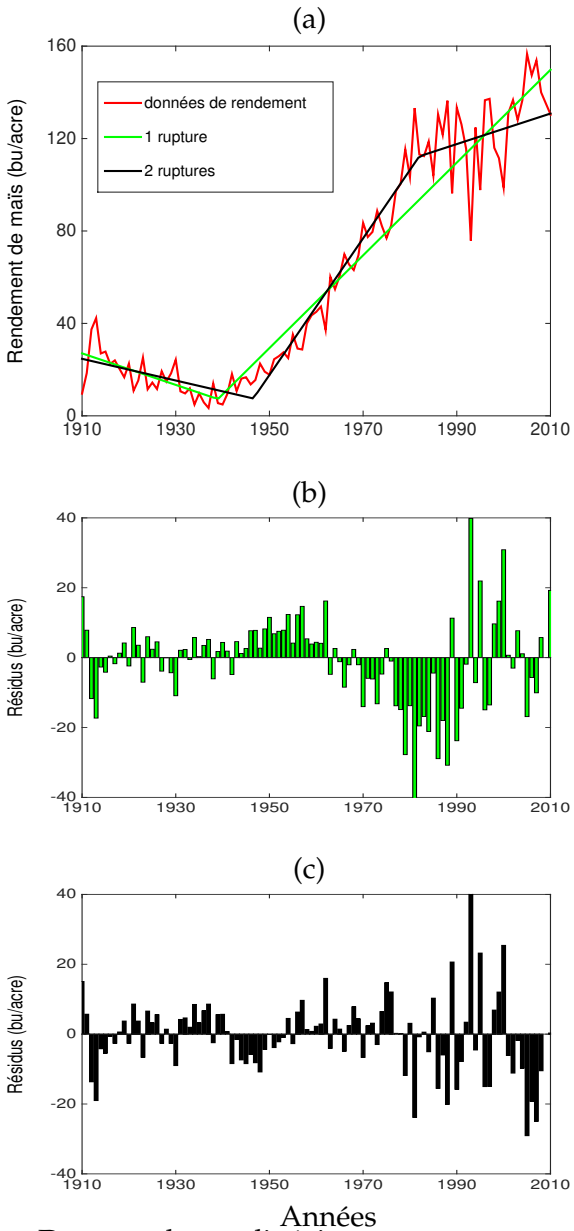


FIGURE IV.10 – Évolution historique du rendement de maïs national aux États-Unis, de 1900 à 2010 (Source : USDA, IRRI, FAO).

Un deuxième changement significatif dans le taux annuel d'amélioration de rendement est arrivé au milieu des années 1960 par l'utilisation en parallèle de maïs génétique, d'engrais (N), de pesticides chimiques et de mécanisation. Au milieu des années 60, la technique de croisement double a été développée, avec l'adoption de la loi sur la Protection des Variétés (Plant Variety Protection Act). Ces deux événements ont permis des gains de rendement impressionnants. Quand le maïs biotechnologique a fait son apparition au milieu des années 1990, les rendements ont encore augmentés. Cependant, les données ne montrent l'apparition d'une nouvelle tendance des rendements lors de l'apparition du maïs hybride transgénique au milieu des années 1990.



Le choix des années de rupture n'est pas simple. Si l'on observe la répartition des années de rupture pour les droites linéaires par morceaux qui minimisent les moindres carrés, on ne peut rien conclure quant à la définition d'une année fixe pour le premier et pour le second point de rupture, pour l'ensemble des profils. Hormis le fait que la première année de rupture semble avoir lieu avant 1960 et la seconde après 1980, il n'y a pas d'information précise sur leur position.

FIGURE IV.11 – Résultat de la tendance linéaire par morceaux sur le canton 33 de l'état du Nebraska. En (a) la superposition des données avec les régressions linéaires : 1 point de rupture (en vert) ou 2 points de rupture (en noir). Les figures suivantes représentent les écarts entre les données et la tendance linéaire par morceaux (1 point de rupture en (b) et 2 points de rupture en (c)). La norme des résidus vaut 144 bu/acre en (b) et 115 bu/acre en (c) : la tendance avec 2 points de rupture est donc de meilleure qualité ici.

Deux tendances linéaires par morceaux ont été testées avec un ou deux points de rupture (Figure IV.11). Leur avantage est avant tout leur simplicité de mise en œuvre. La diversité des formes dans les séries temporelles des rendements est telle que la tendance avec seulement une rupture est peu envisageable. Celle avec deux ruptures semble plus appropriée et forme un bon compromis entre le souhait de moyenner et

celui de ne pas trop coller aux données. C'est un modèle à 4 paramètres (l'hypothèse de continuité évite d'estimer deux paramètres supplémentaires).

Tendance polynomiale

Deux tendances polynomiales ont été testées : un polynôme de degré 2 et un de degré 3 (Figure IV.12). Tout comme pour une tendance linéaire, la régression quadratique joue le rôle d'une tendance "très simple" voire trop simple.

L'inconvénient majeur de l'utilisation d'un polynôme de degré 2 ou 3 est leur évolution non réaliste en dehors de l'étendue temporelle de la série : pour la plupart des séries, cette tendance estime un rendement décroissant pour les années antérieures à 1920 ce qui semble peu probable, tout comme une croissance/décroissance extrêmement rapide pour les années à venir. De plus, il est illusoire de penser qu'un polynôme de degré 15 puisse être utilisé : le problème de sur-apprentissage se pose très rapidement vu le peu de données disponibles dans les séries temporelles. Avec 14 années, il existe un unique polynôme de degré 15 passant par tous les points, mais inutilisable en pratique.

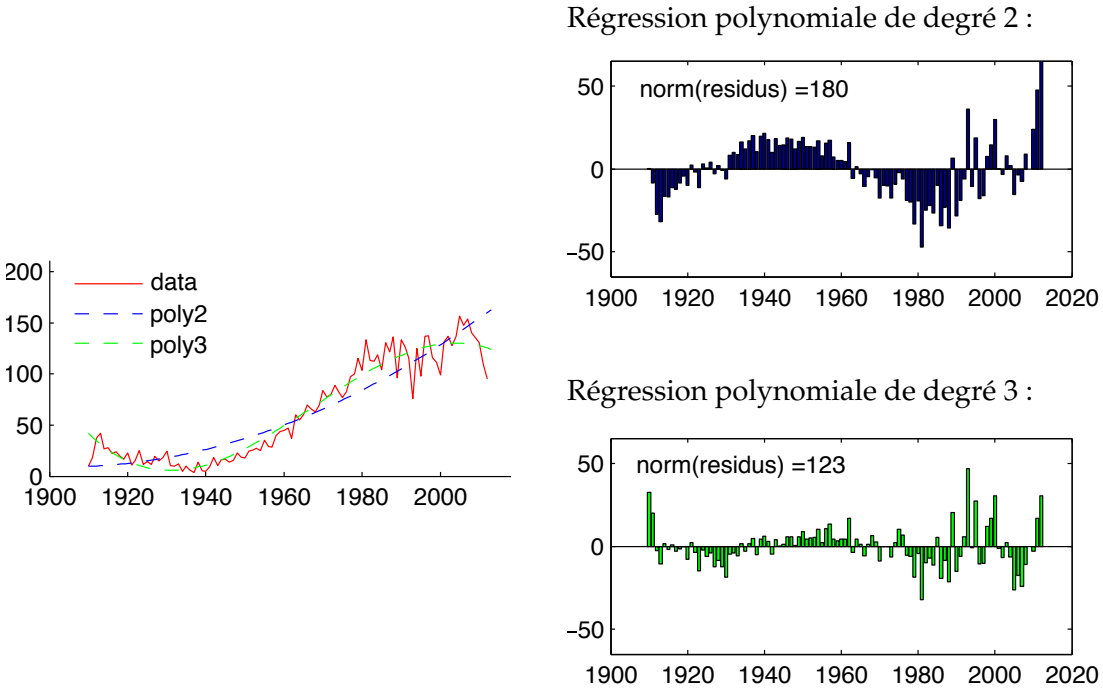


FIGURE IV.12 – Tendance polynomiale de degré 2 ou 3 sur le canton 33 de l'état du Nebraska.

Tendance logistique obtenue avec les modèles à effets mixtes

Le rendement de maïs aux États-Unis a fortement augmenté de 1920 à 2013. Pour modéliser la tendance temporelle du rendement, nous avons utilisé un modèle à effets mixtes non-linéaire avec une fonction logistique définie à l'aide de quatre paramètres $\widetilde{y}_t = c + \frac{K}{1-pe^{-rt}}$ (Section 2.3 page 85). Les valeurs initiales de ces quatre paramètres sont choisies à l'aide de la relation $\text{logit}((\widetilde{y}_t - c)/K) = \ln\left(\frac{1}{pe^{-rt}}\right) = rt - \ln(p)$.

Le choix de ces valeurs initiales est un problème sensible et conditionne l'efficacité du ME non-linéaire. Le paramètre c désigne la valeur minimale que la sigmoïde peut

atteindre et celui K sa valeur maximale (présence d'une asymptote horizontale quand t devient très grand). L'ensemble $c = 20$, K= 150, $\rho = e^{119}$, et $r = 6/100$ convient.

La Figure IV.13 illustre l'identification de la tendance pour quatre cantons différents en Alabama, Nebraska, Texas et Dakota du nord. Ces exemples diffèrent par la localité du canton considéré (différents états), par la longueur de la série temporelle et par la forme globale de celle-ci (maximum, minimum, amplitude). Les anomalies de rendement qui en résultent sont aussi représentées. Malgré les différences de longueur des séries temporelles, le modèle non-linéaire ME semble bien adapté pour cette tâche : il utilise l'ensemble des données des États-Unis et se spécialise pourtant bien à chaque canton. Le partage de l'information entre les cantons qui ont beaucoup de données (comme en Alabama) et ceux qui en ont peu (comme au Texas), donne une régression de bonne qualité pour l'ensemble des cantons. Une telle relation fonctionnelle entre rendement et années s'adapte bien à toute valeur maximale, minimale, ou encore à la diversité de croissance. Les anomalies qui en sont déduites sont souvent comprises entre -0.5 et 0.5, ce qui signifie que la variation relative du rendement par rapport à la tendance est comprise entre -50% et $+50\%$ de la tendance.

L'utilisation d'une fonction logistique implique la présence d'un palier (asymptote) quand $t \longrightarrow \infty$. C'est une hypothèse qui peut s'avérer gênante sur le long terme pour un usage d'extrapolation concernant les 50 ou 100 prochaines années. Cependant, si seules les anomalies de rendements nous intéressent, alors l'usage de la tendance n'est pas nécessaire pour donner des estimations des anomalies de rendement des 50 ou 100 prochaines années.

De plus, il faut rappeler que la régression se fait bien en utilisant l'ensemble des données et non canton par canton, à la différence du modèle linéaire par morceaux. Cette régression logistique ne demande pas de choisir deux années de rupture de pente ce qui est un net avantage. En revanche, malgré les apparences, le nombre de paramètres à estimer est plus grand que pour le "pooling" linéaire : en effet, comme cela est décrit à la Section 2.3 page 85, r est un effet fixe mais c, K, et ρ possèdent tous les trois des effets aléatoires dont il faut estimer la variance. Ce modèle de tendance logistique nécessite donc 10 paramètres à estimer.

En résumé :

Un modèle de tendance ne doit pas être trop complexe au risque de prendre en compte dans la tendance des situations qui n'en font pas partie. Ici le nombre de paramètre à estimer pour réaliser les tendances des 3000 cantons est faible (dix). De plus, l'un des inconvénients des méthodes paramétriques est qu'elles peuvent être sensibles aux valeurs aberrantes or ici, il n'y en a quasiment pas. De plus, une méthode non-paramétrique ne permettrait pas aussi bien d'apprendre à l'échelle des États-Unis (donc avec le maximum de données) tout en prenant en compte les spécificités locales comme le fait le modèle mixte utilisé ici.

Dans la littérature, on trouve d'autres approches pour l'identification de la tendance : des méthodes non-paramétriques (souvent basée sur l'estimation de quantiles conditionnels), des méthodes d'estimation par noyaux, la méthode de la constante locale, ou encore une méthode d'estimation par noyau produit ont été utilisées [Duch95, Poir00, Gann02].

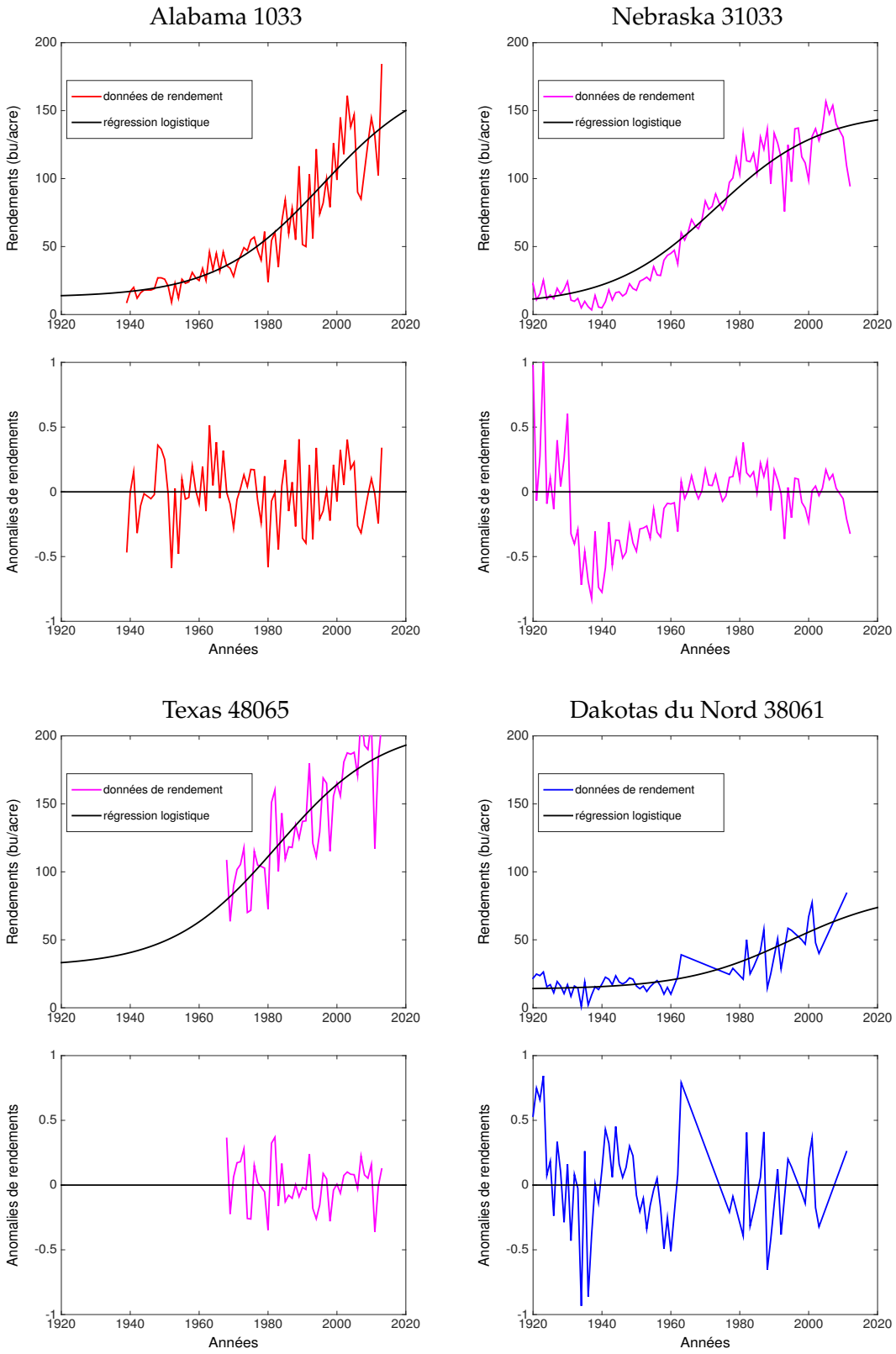


FIGURE IV.13 – Quatre exemples de régression logistique issus d'un modèle ME des séries temporelles de rendement. Par couleur, le graphique du haut représente la série temporelle d'un canton avec la courbe logistique estimée par la régression ; le graphique du bas représente les anomalies déduites à partir du graphique du haut. Ces quatre états se distinguent par leurs différentes localités géographiques, par l'importance des valeurs de rendements, ainsi que par la durée des séries temporelles dont on dispose.

7.2 La sélection des variables explicatives

La complexité des relations météo/rendement nécessite autant d'information que possible dans les entrées du modèle d'impact. Cependant, la quantité limitée de donnée pousse à réduire la complexité du modèle pour éviter le sur-apprentissage. Une sélection des variables explicatives est donc nécessaire pour limiter le nombre d'entrées, et donc limiter le nombre de paramètres du modèle, et donc sa complexité. C'est ce que l'on appelle le principe de parcimonie [Craw11].

Nous voulons obtenir la meilleure combinaison de prédicteurs de manière statistique. Or cette procédure s'avère relativement coûteuse en temps de calcul car elle teste de nombreuses combinaisons de variables explicatives (sur des bases de validation et de test). Il est évidemment impossible de tester l'ensemble des modèles envisageables ainsi que des valeurs des paramètres de régularisation sur chaque combinaison de prédicteurs possible. Il convient donc de séparer le choix du modèle et des paramètres de régularisation d'une part et la recherche des meilleurs prédicteurs d'autre part [Koch10].

Pour sélectionner les variables explicatives, on utilise une méthode itérative qui fournie une hiérarchie des variables explicatives potentielles (une centaine de variables météorologiques ou assimilées). On choisi tout d'abords un critère de qualité, par exemple COR, AIC, ou une p-value pour quantifier le pertinence de garder une variable comme entrée du modèle.

La première variable sélectionnée est celle qui propose la meilleure valeur du critère d'information (par exemple, la plus grande corrélation avec les anomalies de rendement). La seconde variable est choisie parmi toutes celle disponibles - excepté la première variable sélectionnée - comme étant celle qui, combinée avec la première, définit un modèle linéaire à deux entrées qui propose la meilleure valeur du critère de qualité. Ce procédé est répété jusqu'à avoir sélectionné le nombre de variables souhaité. En théorie, cette méthode identifie en premier les variables qui sont le plus en lien avec les anomalies de rendement, mais assure aussi la sélection des variables les plus complémentaires. Par exemple, deux variables qui se ressembleraient beaucoup ne seraient pas toutes deux sélectionnées. Cette méthode s'appelle "méthode ascendante" (forward selection).

Suite à cette hiérarchisation des variables explicatives, une méthode de validation croisée peut servir pour décider du nombre optimal total d'entrées à considérer dans le modèle : si l'ajout d'une $k^{\text{ème}}$ variable n'améliore plus la qualité de généralisation du modèle, on décide de ne pas la considérer en tant qu'entrée du modèle et on ne considère donc que les $k-1$ variables explicatives précédentes. Si besoin, la matrice des corrélations des différentes variables explicatives peut servir à décider à quelle variable on s'arrête (si la dernière visée est trop corrélée avec les précédentes, on peut décider de ne pas la considérer).

La sélection des variables successives ne doit pas se faire qu'avec l'aide d'un seul critère de qualité. Il est important de réaliser en parallèle une sélection de variable dictée par d'autres mesures de qualité comme le RMSE ou le critère AIC. En effet, ces différents critères ne sont pas sensibles aux mêmes effets et sont parfois complémentaires. Si les sélections finales proposées par les différents critères sont proches, cela garanti une certaine robustesse dans le choix des variables du modèle. Dans ce genre de méthode de sélection de variables, le choix d'une variable est très dépendant des

variables choisies précédemment. Il faut donc veiller à ce que la liste des variables disponibles soit de bonne qualité pour éviter une divergence très rapide de qualité. Par exemple, une variable qui présenterait beaucoup de données manquantes, peut faire échouer la méthode de sélection dans le traitement informatique si le logiciel utilisé ne prend en compte que les échantillons sans données manquantes.

D'autres méthodes existantes de sélection de variables sont décrites ci-dessous.

— Méthode descendante (Backward Elimination)
Elle est considérée comme la plus simple des procédures de sélection. On commence avec toutes les variables explicatives prises comme entrées du modèle. On enlève les variables qui possèdent la plus forte p-valeur du t-test (Student) [12], plus grande qu'un seuil α_{crit}. On fait à nouveau tourner le modèle et on reprend l'étape précédente. On s'arrête lorsque toutes les p-valeurs sont inférieures à α_{crit}. Souvent α_{crit} est proche de 15-20%. Le principal inconvénient réside dans le fait que l'on commence avec des modèles qui prennent en compte l'ensemble des entrées, ce qui peut s'avérer très lourd en calcul et donc couteux en temps.

— Méthode ascendante avec élimination possible (Stepwise selection)
C'est une amélioration de la méthode ascendante : à chaque étape, on réexamine toutes les variables introduites précédemment dans le modèle. En effet, une variable considérée comme la plus significative à une étape de l'algorithme peut à une étape ultérieure devenir non significative, en raison de ses corrélations avec d'autres variables introduites. Cette procédure propose après l'introduction d'une nouvelle variable dans le modèle de (1) réexaminer les tests de Student pour chaque variable explicative anciennement admise dans le modèle, et (2) après réexamen, si des variables ne sont plus significatives, retirer du modèle la moins significative d'entre elles. Le processus continue jusqu'à ce que plus aucune variable ne puisse être introduite ni retirée du modèle.

— Sélection de modèle par pénalisation : ici on garde toutes les variables d'entrée mais des paramètres de pénalisation forcent les coefficients relatifs à de nombreuses entrées à prendre de petite valeur. Les algorithmes ridge, lasso [Tibs96], ou elastic net [Zou05] sont utilisés pour cela. L'approche par modèle pénalisé permet de régler le problème de la multi-colinéarité entre les variables dans les situations où toutes les variables sont gardées.

12. Le test de Student est souvent utilisé pour tester la nullité d'un coefficient dans le cadre d'une régression linéaire, et est donc utilisé pour savoir si on a raison ou pas de rajouter une variable dans un modèle.

CHAPITRE V

Prévision et estimation des rendements agricoles

Table des matières

Dans ce chapitre, on estime en fin d'année, ou on prédit en milieu d'année, les anomalies de rendements en maïs en fonction de variables météorologiques. Pour ce faire, on commence par réaliser cette étude en utilisant seulement des variables météo facilement accessibles et non transformées, comme les températures et les précipitations (Section 1). Ce début de chapitre se concentre d'abord sur le développement de modèles d'impact météorologique pour le rendement des cultures, et sur l'évaluation statistique fiable de la sensibilité à la météo. Nous comparons cinq modèles statistiques. Une seconde analyse utilisant des données météorologiques plus ambitieuses (comme

les indices agroclimatiques) fait suite à cette première (Section 2). Les deux cas d'étude sont comparés à la Section 3.

Ce chapitre de thèse a fait l'objet de deux publications dans les journaux internationaux suivants : *The Journal of Applied Meteorology and Climatology* et *Agricultural and Forest Meteorology*.

La Figure V.1 schématise l'ensemble des étapes nécessaire à la modélisation statistique des anomalies de rendement agricole en fonction de variables météorologiques.

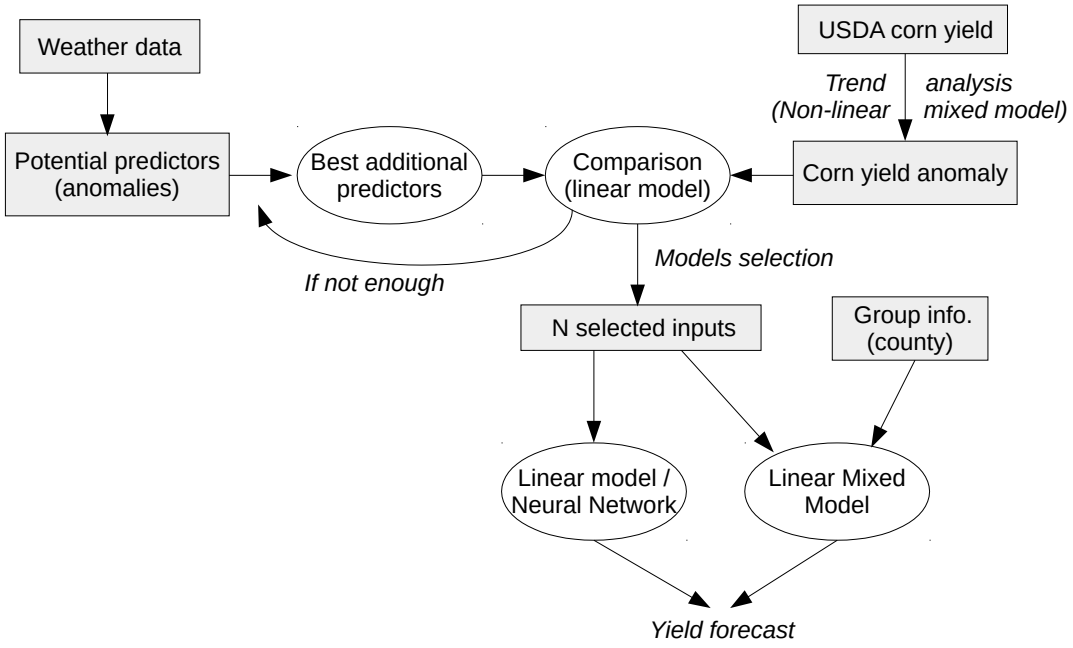


FIGURE V.1 – Schéma des bases de données et des méthodes utilisées dans ce chapitre.

1 Modélisation à l'aide de variables météorologiques non transformées

Dans cette première étude, on utilise simplement des variables météorologiques non transformées, c'est à dire des températures et des précipitations mensuelles. On dispose de trois grands types de modèles : un modèle linéaire, un modèle à effet mixtes linéaire, et un réseau de neurones. Le modèle à effets mixtes se divise en trois sous modèles, puisque l'on souhaite tester et comparer la classification par état fédéral, par district, puis par canton.

1.1 Prise en compte des décompositions spatiales et régularisation du réseau de neurones

Prise en compte des groupes spatiaux - Dans cette section, nous proposons deux directions pour étudier l'hétérogénéité spatiale des trois modèles d'impact météorologique :

— Tout d'abord, une analyse des avantages fournis par les modèles linéaires à effets mixtes (ME) : comme on a pu le voir au Chapitre III, aux États-Unis, les cinquante états sont divisés en environ 300 districts ; eux-mêmes divisés en environ 3000 cantons. Chaque donnée agricole se réfère à un canton spécifique, mais les données peuvent être rassemblées par district ou par état. Les modèles mixtes peuvent rendre compte de cette hiérarchie spatiale qui classe les données par groupes. Ici, le terme "groupe" se réfère à une classification spatiale (les différents cantons, les districts ou les états). Les modèles ME sont basés sur l'idée que chaque information appartient à un groupe particulier, et le modèle prend en compte les particularités de chacun de ces groupes [Fahr94].

— Ensuite, l'échelle spatiale utilisée par le modèle d'impact (canton, district, état ou nation) est optimisée pour divers modèles statistiques (linéaire, réseau neuronal ou effets mixtes) afin d'analyser leurs avantages et leurs limites.

En accord avec les résultats de l'analyse de sensibilité des données à priori, il faut choisir quelles sont les entrées qu'il est significatif de garder parmi les prédicteurs potentiels. La méthode utilisée ici va consister à choisir les variables prédictives de façon itérative. Une description détaillée du processus a été donnée au chapitre IV, page 123.

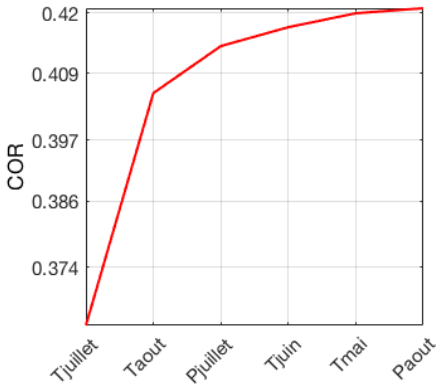


FIGURE V.2 – Les entrées sélectionnées selon le plus grand pourcentage de variance expliquée par le modèle. Évolution du pourcentage de variance expliquée au fur et à mesure que les entrées sont choisies.

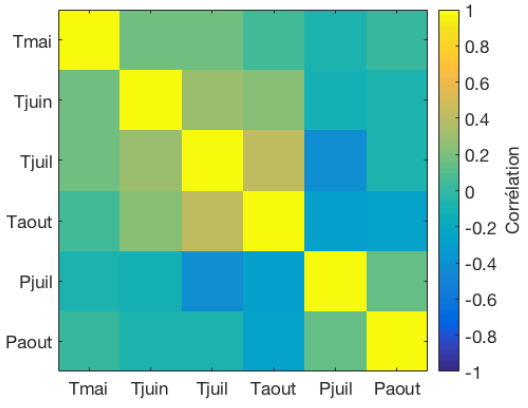


FIGURE V.3 – Matrice des corrélations entre les six premières entrées.

Remarque : nous avons aussi réalisé la sélection de variables en autorisant les carrés des variables et les multiplications 2 à 2. Les résultats ont montré une faible amélioration (à la sixième variable sélectionnée, la corrélation avait gagné 3 points). Aussi nous avons préféré utiliser seulement des variables météorologiques brutes.

Tous les modèles de cette section ont les mêmes six entrées (anomalies météorologiques) : **Tmai, Tjuin, Tjuillet, Taoût, Pjuilllet, Paoût** (Figure V.2). Cela rend les comparaisons des résultats plus faciles et plus légitimes. Par conception, les corrélations

entre ces six entrées sont faibles, mais pas nulles (Figure V.3).
[Thom86] a aussi utilisé un modèle statistique pour déterminer l'impact du changement climatique et de la variabilité météorologique sur la production de maïs dans cinq états du midwest américain. Il a trouvé que les précipitations de pré-saison (juin), la température de juin, et les températures et précipitations de juillet et août sont étroitement corrélées avec les variations des rendements du maïs en dehors de la tendance. Un modèle linéaire et un réseau de neurones sont testés (1) pour chaque canton, (2) pour chaque district, (3) pour chaque état et (4) pour l'ensemble des États-Unis. Un modèle ME est construit uniquement aux niveaux national, de l'état ou du district (pas au niveau du canton, car les données du canton ne peuvent être subdivisées en sous-classe, donc cela reviendrait à faire un modèle linéaire classique).

Pour faire des comparaisons significatives entre les modèles, les évaluations doivent être effectuées à la même échelle spatiale. Par conséquent, même si le modèle est entrainé à une échelle spatiale plus élevée, la validation sera toujours effectuée à la même échelle spatiale pour être comparé (par exemple à l'échelle du canton). Cela signifie qu'un modèle à échelle spatiale plus élevée (par exemple au niveau de l'état) sera appliqué au niveau du canton, puis que les comparaisons seront faites entre les modèles à cette échelle spatiale.

Régularisation du réseau de neurones - Pour éviter d'utiliser un trop grand nombre de paramètres, les réseau de neurones doivent être régularisés (Section 6.2 page 110). Tout d'abord, un terme de pénalisation - appelé "weight decay" - peut être rajouté dans le critère de qualité qui est minimisé pendant l'apprentissage. Ensuite, le nombre de paramètres doit également être réduit en limitant le nombre de neurones des couches cachées. Le choix optimal des paramètres de régularisation (weight decay et nombre de neurones) repose sur les critères de qualité utilisés pour valider les modèles. Supposons que l'on dispose d'un ensemble d'apprentissage et d'un ensemble de test fixés : plusieurs réseaux de neurones sont construits/entrainés avec un weight decay d allant de 0 à 1/4 et un nombre de neurones N, allant de 1 à 20 (Figure V.4).

Pour chaque couple (N, d), des critères de qualité tels que le R^2 ou le MSE sont calculés sur les ensembles d'apprentissage et de test. On répète cette opération 100 fois sur d'autres bases d'apprentissage et de test pour disposer par Monte-Carlo, de résultats plus fiables et pour connaître leurs sensibilités aux données test.

Selon les résultats de régularisation, aucun des réseaux de neurone avec plus d'un neurone sur sa couche cachée, ne fournit de meilleurs résultats qu'un réseau avec seulement un neurone. Un réseau avec un neurone sur sa couche cachée représente une régression logistique traditionnelle (c'est-à-dire que le modèle résulte en une combinaison linéaire pondérée des entrées suivies d'une fonction logistique). Cela n'est pas une surprise : les relations météo/rendement sont simples, proches de linéaires et la non-linéarité permet d'obtenir des effets de saturation pour certaines configurations d'entrée. Un effet de saturation apparaît lorsque certaines entrées n'ont aucun impact sur la sortie une fois qu'elles atteignent un certain seuil.

À l'échelle nationale (Figure V.4), le weight-decay optimal est de 0, alors qu'à l'échelle des états fédéraux, il est de 1/128, et à l'échelle du district ou du canton il est de 1/32 (non montré). Un weight decay plus important pour l'apprentissage à l'échelle du district ou du canton est cohérent, car le modèle dispose alors de moins de données pour l'apprentissage à l'échelle du canton ou du district : il est alors nécessaire de pénaliser

davantage la complexité du modèle pour éviter le sur-apprentissage. Quand l'apprentissage se fait à l'échelle des états, moins de données sont fournies au modèle - en comparaison avec un apprentissage national - pour estimer le même nombre de paramètres. Donc, moins d'information sont fournies au réseau de neurones et le risque de sur-apprentissage est plus important. On s'attend donc à ce que le weight decay choisi par le processus de validation croisée soit plus élevé que pour un apprentissage national.

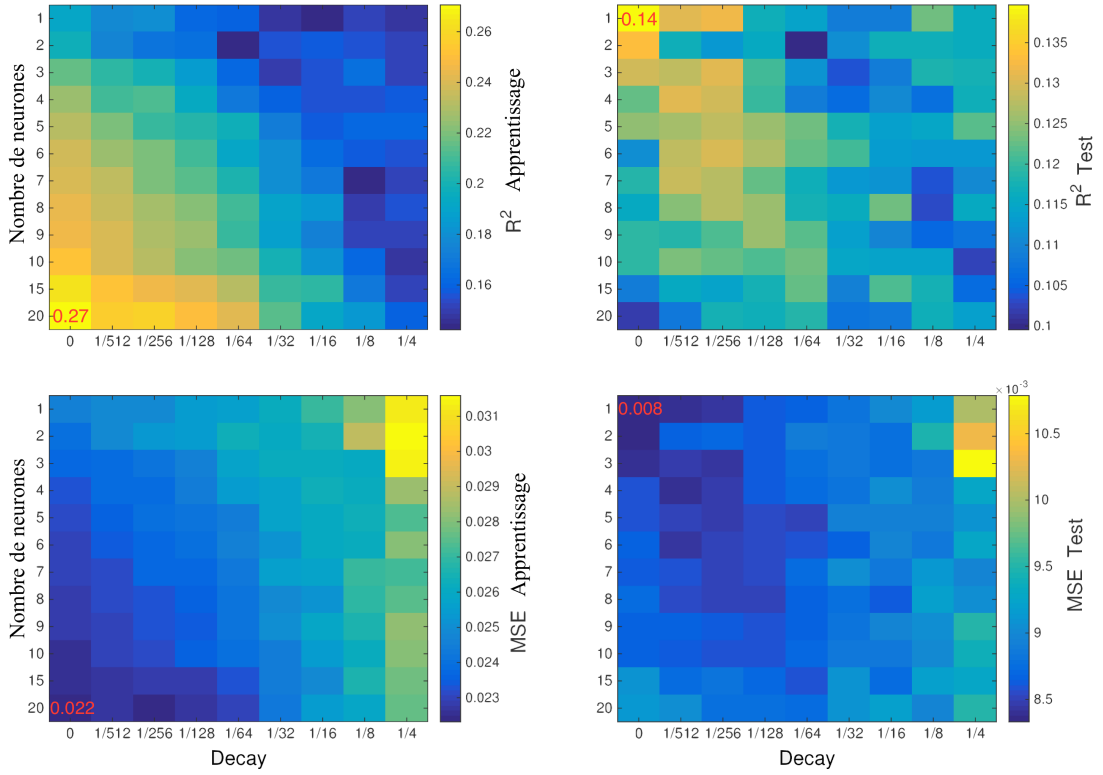


FIGURE V.4 – Statistiques de la validation croisée pour la régularisation du réseau de neurones en utilisant les données à l'échelle des États-Unis : R^2 (en haut) et MSE (en bas) pour l'apprentissage (à gauche) et la validation (à droite). Le nombre rouge souligne la valeur obtenue pour le couple (N, d) optimal.

Après avoir examiné les configurations expérimentales, une attention particulière est accordée à l'évaluation de la capacité prédictive des modèles. Les modèles d'impact sont comparés à la Section 1.2 et une étude détaillée est effectuée sur l'état d'Illinois. Un modèle d'impact est finalement testé pour les prévisions de production de maïs et les prévisions saisonnières.

1.2 La comparaison des modèles d'impact

On décrit ici les notations importantes. Le rendement du maïs (en grain) pour l'année t est noté $y(t)$, et la valeur de la tendance à long terme $\bar{y}(t)$. Cette tendance représente l'évolution lente du rendement de maïs (principalement des changements dans les pratiques agricoles). Le détail du calcul de la tendance a été donné à la Section 7.1 page 120. $y(t)$ et $\bar{y}(t)$ sont tous les deux exprimés en boisseau/acre qui est une unité

courante aux États-Unis. L'anomalie relative du rendement de maïs $a(t)$ est alors définie comme une variation en pourcentage autour de la tendance :

$$a(t) = \frac{y(t) - \bar{y}(t)}{\bar{y}(t)}$$

Si $a(t) > 0$, le rendement de cette année est supérieur à une année standard, et si $a(t) < 0$, le rendement est plus bas. Par exemple, une anomalie $a(t) = 0.25$ signifie que le rendement du maïs pour l'année t est 25% supérieur à la tendance annuelle.

Apprentissage sur les États-Unis		Total			Régions centrales		
Modèle	Valid. sur	R^2	Corr	MSE	R^2	Corr	MSE
LIN	États-Unis	15	37	42	-	-	-
	états	22	42	44	31	53	33
	district	25	44	45	34	54	31
	canton	27	**41**	43	32	**48**	27
NN	États-Unis	16	38	43	-	-	-
	état	17	38	43	17	38	41
	district	19	39	42	18	38	41
	canton	30	**48**	43	30	**48**	42
ME-état	États-Unis	16	40	44	-	-	-
	état	22	41	44	28	50	33
	district	24	43	46	31	52	32
	canton	32	**50**	45	38	**57**	33
ME-district	États-Unis	16	39	43	-	-	-
	état	19	38	44	25	47	33
	district	22	41	45	28	48	32
	canton	30	**48**	45	35	**53**	33
ME-canton	États-Unis	21	45	40	-	-	-
	état	23	42	41	33	54	30
	district	25	43	43	34	55	29
	canton	34	**51**	42	41	**59**	30

TABLEAU V.1 – Critères de qualité en généralisation (R^2 (%), Corr (%) et MSE $\times 10^3$ pour les cinq modèles d'impact formés à l'échelle des États-Unis (pooling) : un modèle linéaire (LIN), un réseau de neurones (NN), un modèle à effets mixtes avec classification par (1) état fédéral (*ME-état*), (2) district (*ME-district*), et (3) canton (*ME-canton*). Les résultats sont fournis pour deux domaines spatiaux : les statistiques "Total" incluent - après le tri lié à l'irrigation - tous les états où les données de rendement sont disponibles, et les "Régions centrales" (Missouri, Illinois, Indiana, Ohio, Pennsylvanie , Virginie, Virginie-Occidentale, Kentucky et Tennessee). Les modèles d'impact sont évalués à différentes échelles spatiales : au niveau du canton, du district, de l'état fédéral ou des États-Unis. Les situations des cellules en gras sont représentées graphiquement à la Figure V.5.

Le Tableau V.1 représente les statistiques en généralisation [1] de R^2, Corr et MSE, pour les cinq modèles d'impact formés à l'échelle des États-Unis (pooling complet) : un modèle linéaire (LIN), un réseau de neurones (NN), un modèle à effets mixtes avec classification (1) par état fédéral (*ME-état*), (2) par district (*ME-district*), et (3) par canton (*ME-canton*). Les résultats sont fournis pour deux domaines spatiaux : les statistiques "Total" incluent - après le tri lié à l'irrigation - tous les états où les données sur le rendement de maïs sont disponibles, et les "Régions centrales" qui regroupent les régions

1. Sauf si précisé, tous les critères de qualité dans ce manuscrit sont estimés sur des données indépendantes des données d'apprentissage.

où le maïs semble le plus sensible au climat (Missouri, Illinois, Indiana, Ohio, Pennsylvanie , Virginie, Virginie-Occidentale, Kentucky et Tennessee). Même s'ils sont entrainés à l'échelle des États-Unis, les modèles d'impact peuvent être évalués à différentes échelles spatiales : les statistiques peuvent évaluer le modèle au niveau du canton, du district, de l'état fédéral ou, des États-Unis. Pour comparer les différents modèles, il faut utiliser la même échelle spatiale de validation. Nous nous concentrerons ici sur l'évaluation au niveau du canton (i.e. la dernière ligne pour chaque modèle). Afin de faciliter la discussion, les cellules à caractères gras font ressortir les quantités qui seront les plus souvent commentées.

Les modèles ME sont meilleurs que le modèle LIN (Corr est d'environ 50% par rapport à 41%). Cela peut s'expliquer par le fait que les modèles linéaires ME sont équivalents à LIN, sauf qu'une information supplémentaire leur est fournie (c'est-à-dire la classe de localisation). Quand on réalise un pooling à une grande échelle spatiale, comme au niveau de l'état fédéral, l'utilisation de l'information d'appartenance aux districts ou aux cantons aide le modèle ME à s'adapter aux conditions locales. Lorsqu'un apprentissage à petite échelle est réalisé, le ME peut bénéficier de plus d'échantillons à grande échelle, où le modèle LIN ne peut utiliser qu'un jeu de données limité, ce qui le rend sensible aux problèmes de sur-apprentissage (Section 6.1 page 107). Le ME est également toujours meilleur ou équivalent au réseau de neurone NN. Le réseau de neurones bien régularisé fournit un résultat similaire à celui de la régression linéaire mais pas meilleur. Ceci est logique puisque le cas linéaire est compris dans le cas non-linéaire. Néanmoins, il ne peut faire mieux ici du fait du manque de données. Les avantages de non-linéarité du NN par rapport à la linéarité des modèles ME ne semble pas pertinents pour cette application de modélisation du rendement de maïs : la relation entre le rendement de maïs et la météo est assez simple pour être modélisée par un modèle linéaire. Cette supposition doit néanmoins être modérée, puisque le NN a de meilleures statistiques que le LIN (Corr de 48% par rapport à 41%) montrant que la non-linéarité du NN permet de s'adapter un peu mieux aux conditions locales. Cette dépendance à l'état du NN est presque aussi bonne que les modèles ME, même si aucune information de classe spatiale n'est fournie au NN (uniquement des informations météorologiques). Cela signifie que lorsque aucune information de classe n'est disponible, le NN est un candidat très intéressant, mais lorsque les informations de classe sont disponibles, un modèle ME peut bénéficier de cette information supplémentaire. Les modèles ME fournissent une corrélation d'environ 50% pour le domaine "Total", mais ce nombre atteint 59% pour les "régions centrales" où la relation météo/rendement est plus forte. Cela signifie que 41% ($R^2 = 41\%$) de la variabilité du rendement de maïs s'explique par les informations météorologiques (moyennes mensuelles). La variance restante s'explique par d'autres facteurs (différentes pratiques agricoles, entre-autre) ou une information plus détaillée sur la météo.

Sur le Tableau V.2, l'apprentissage se fait indépendamment sur chaque état fédéral : la quantité de données pour l'apprentissage de chaque modèle d'impact est donc réduite par rapport au Tableau V.1. Les deux modèles LIN et NN se sont améliorés, montrant que la spécialisation par état aide à décrire les conditions locales des états. Dans le même temps, le nombre de données disponibles au niveau de l'état semble être suffisant pour éviter les problèmes de sur-apprentissage. Les modèles ME ne s'améliorent pas lorsqu'ils sont entrainés au niveau de l'état. Cela était à prévoir parce que les conditions locales sont déjà considérées dans les modèles ME du Tableau V.1 à travers la classe spatiale "état", "district" ou "canton". Un commentaire général est également que

tous les modèles (LIN, NN et ME) semblent être de qualité similaire, ce qui signifie que la focalisation des modèles d'impact sur chaque état est le meilleur compromis entre l'accent mis sur les conditions locales et l'apport suffisant de données pour l'apprentissage.

1 apprentissage par état		Total			Régions centrales		
Modèle	Valid. sur	R^2	Corr	MSE	R^2	Corr	MSE
LIN	état	20	40	44	26	47	32
	district	23	42	45	28	48	31
	canton	32	**50**	44	36	**54**	32
NN	état	20	39	45	25	46	36
	district	20	39	46	26	47	35
	canton	32	**50**	45	36	**53**	35
ME-district	état	21	40	44	27	49	32
	district	23	42	46	30	50	31
	canton	31	**49**	45	37	**56**	32
ME-canton	état	20	40	44	28	49	33
	district	23	41	46	30	50	32
	canton	31	**49**	45	38	**56**	33

TABLEAU V.2 – Similaire au Tableau V.1 mais pour un apprentissage à l'échelle des états fédéraux.

Quand on regroupe toutes les données par district ou par canton, les résultats ne sont pas meilleurs (Tableaux V.3 et V.4). La qualité de généralisation se dégrade lorsque la résolution de l'apprentissage augmente. Ainsi, pour LIN ou NN, l'état fédéral comme échelle spatiale est un bon compromis, permettant de se spécialiser davantage localement que lorsqu'on regarde l'ensemble des États-Unis, tout en conservant suffisamment de données pour calibrer le modèle LIN ou NN. Le modèle d'impact ME-canton est toujours celui qui fournit les meilleurs résultats.

1 apprentissage par district		Test			Régions centrales		
Modèle	Valid. sur	R^2	Corr	MSE	R^2	Corr	MSE
LIN	district	23	42	50	29	49	33
	canton	30	48	46	35	54	34
NN	district	21	39	51	26	45	37
	canton	31	49	49	35	52	38
ME-canton	district	22	41	92	28	47	34
	canton	30	48	47	35	53	35

TABLEAU V.3 – Critères de qualité en généralisation lorsque des données sont rassemblées par district agricole pour l'apprentissage : R^2 (%), Corr (%), MSE$\times 10^3$.

1 apprentissage par canton		Test			Régions centrales		
Modèle	Valid. sur	R^2	Corr	MSE	R^2	Corr	MSE
LIN	canton	28	46	51	33	51	39
NN	canton	27	45	51	31	49	42

TABLEAU V.4 – Critères de qualité en généralisation lorsque des données sont rassemblées par canton pour l'apprentissage : R^2 (%), Corr (%), MSE$\times 10^3$.

En se basant sur les résultats des Tableaux V.1 et V.2, quel modèle devrait être utilisé ? Pour répondre à cette question, un autre point doit être pris en considération. Comme mentionné dans le Chapitre IV, de nombreux efforts ont été faits dans cette étude afin d'éviter le sur-apprentissage du modèle.

Cependant, plus le modèle est localisé (des États-Unis jusqu'aux cantons, en passant par les états fédéraux ou les districts), plus il y a de différences entre les résultats de test et ceux d'apprentissage. Par exemple, pour l'apprentissage à l'échelle des états fédéraux, pour le modèle LIN, $Corr_{app.} = 52\%$ et $Corr_{test} = 40\%$, mais pour l'apprentissage à l'échelle des cantons $Corr_{app.} = 68$ % et $Corr_{test} = 46\%$. La quantité de données disponibles est réduite lorsqu'on se concentre localement, de sorte que le dilemme biais-variance devient plus serré, et même si l'on utilise des techniques de régularisation, le sur-apprentissage devient préoccupant. Par conséquent, il est recommandé d'utiliser un modèle entrainé avec autant de données que possible, même au niveau des États-Unis. Les modèles ME avec leurs informations supplémentaires sur l'emplacement spatial ont un véritable avantage, en particulier pour les applications où la taille des jeux de données est limitée et où les informations spatiales peuvent être utilisées pour considérer l'ensemble des données [Aire12].

Comme mentionné précédemment, les meilleures statistiques sont obtenues sur les "Régions centrales" (Missouri, Illinois, Indiana, Ohio, Pennsylvanie, Virginie, Virginie-Occidentale, Kentucky et Tennessee), avec des valeurs plus élevées de R^2 et $Corr$: *ME-canton* atteint 59% (entrainé, toujours, à l'échelle nationale) et environ 52% pour les modèles LIN et NN (lorsqu'ils sont entrainé à l'échelle des états fédéraux). En outre, une diminution substantielle du MSE de 10 points pour presque tous les modèles montre comment ces états sont plus directement touchés par la météo. Dans ces états, la météo a un impact plus important/direct sur le rendement du maïs puisque notre modèle peut extraire cette information. Les "Régions centrales" représentent les deux zones climatiques 4A et 5A de la carte des zones climatiques américaines de [The 06] : ces zones ont un été chaud avec un climat mixte-humide pour la zone 4A et un climat froid-humide pour la zone 5A. Les États des "Régions centrales" ont produit près de 40% de maïs aux États-Unis en 2014. Comme le montre la Figure III.14, les états des "Régions centrales" (sauf le Missouri) dépendent le moins de l'irrigation avec moins de 230 gallons par jour en 2010 selon [Maup14].

La Figure V.5 représente la distribution spatiale du critère Corr. La cohérence spatiale est bonne : les cantons voisins ont été estimés dans une qualité similaire puisque leurs corrélations entre les valeurs observées et les prévisions sont proches. Cette observation n'est pas surprenante car les cantons voisins ont des données météorologiques très similaires (en raison de la régularité des données météorologiques mensuelles). Pour les modèles LIN et ME, la partie centrale des États-Unis est toujours mieux estimée que les régions à basse ou haute latitude. Les régions du nord-est fournissent beaucoup de données mais nos modèles ne parviennent pas à généraliser correctement sur ces régions. Les résultats pour NN sont moins bons en moyenne mais beaucoup plus homogènes spatialement que ME ou LIN, où certaines zones sont nettement mieux apprises que d'autres. Cette homogénéité montre la régularisation efficace du modèle NN : il n'y a pas de canton très bien appris proche d'un canton mal appris, ce qui aurait souligné un problème de sur-apprentissage. Ces résultats sont dus aux paramètres choisis du NN pour fournir la meilleure généralisation (en moyenne), pour un canton donné, ce qui entraîne une plus grande homogénéité. ME et LIN donnent de meilleurs résultats sur les "Régions centrales" même si LIN ne parvient pas à bien généraliser en dehors de cette zone.

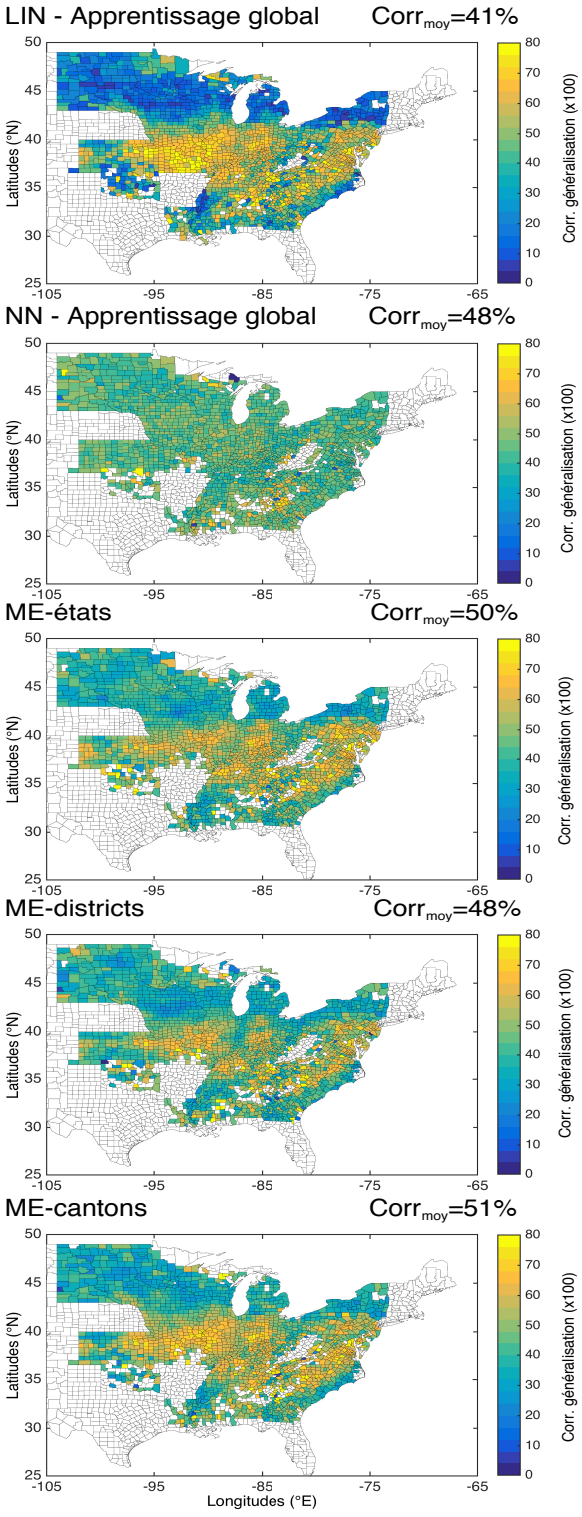


FIGURE V.5 – Valeurs des corrélations des cantons $Corr$ entre les anomalies de rendement observées et prédites en généralisation lorsque l'apprentissage se fait à l'échelle des États-Unis et la validation se fait sur les cantons, pour cinq modèles de régression : LIN, NN, ME-état, ME-district et ME-par canton (Tableau V.1). Les valeurs moyennes des cantons sont indiquées au dessus de chaque sous-figure.

Cette différence entre LIN et ME - qui sont tout deux des régressions linéaires - souligne l'amélioration générée par l'information de groupe, en particulier pour la classification par canton. Même si ME offre de meilleurs résultats dans les régions du nord-est que LIN, sa performance n'est pas aussi forte que dans les "Régions centrales". Une sensibilité plus élevée et plus claire aux conditions météorologiques pour les "Régions centrales" peut expliquer cette différence.

1.3 Focus sur l'état d'Illinois

Nous nous concentrons ici sur l'état d'Illinois (un état de la Corn-Belt) en utilisant le modèle d'impact ME avec classification par canton (ME-canton). Les prévisions des anomalies de rendement en fonction des observations sont représentées à la Figure V.6. Pour cet état, $R^2_{test} = 40\%$, $Corr_{test} = 58\,\%$ et $MSE_{test} = 26 \times 10^{-3}$. Les anomalies négatives importantes (inférieures à $-0,4$) ne sont pas bien prédites par le modèle d'impact, et il existe apparemment un effet de saturation pour ces valeurs extrêmement faibles. Les valeurs élevées sont bien identifiées par le modèle d'impact, ce qui est satisfaisant, mais les valeurs faibles sont amorties.

Ce problème peut avoir plusieurs raisons : (1) ces extrêmes ne sont pas bien représentés (en nombre) dans l'ensemble d'apprentissage, ce qui provoque toujours des difficulté lors de l'apprentissage d'un modèle statistique [Cole01], (2) les informations pertinentes pour prédire ces extrêmes ne sont peut être pas utilisées/disponibles dans notre modèle (e.g. maladies, insectes, etc.), (3) notre ME modèle utilise un modèle simple linéaire qui pourrait ne pas être bien adapté au comportement des évènements extrêmes. Toutefois, une relation non-linéaire requerrait un nombre d'évènements extrêmes dans notre base d'apprentissage encore plus grand. Le prochain chapitre portera sur les prévisions de ces cas extrêmes, en utilisant un modèle plus complexe (Chapitre VI).

Le rendement du maïs prédit $y(t)$ est estimé en utilisant la tendance du rendement de maïs $\bar{y}(t)$ (Chapitre IV page 122) et la prévision des anomalies de rendement $a(t)$: $y(t) = \bar{y}(t)\,(1 + a(t))$. Afin d'évaluer la qualité de généralisation du modèle d'impact, la Figure V.7 représente les séries temporelles des observations de rendement (noir) et les prévisions (gris) pour deux cantons du district 60 d'Illinois[2]. Les prévisions sont faites sur des données d'ensembles de tests pour tester la qualité de généralisation : chaque année de prévision dans la série temporelle est le résultat d'une prévision qui a été formée à partir d'un ensemble de données d'apprentissage qui ne comprenait pas cette année particulière (méthode des runs d'ensemble, page 114).

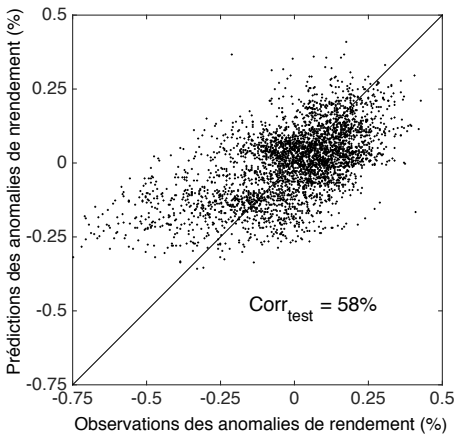


FIGURE V.6 – Anomalies de rendement prédites *versus* observées pour l'Illinois, en utilisant le modèle d'impact ME-canton.

2. On rappelle que le numéro 60 fait référence à la classification FIPS utilisée aux États-Unis pour le découpage administratif.

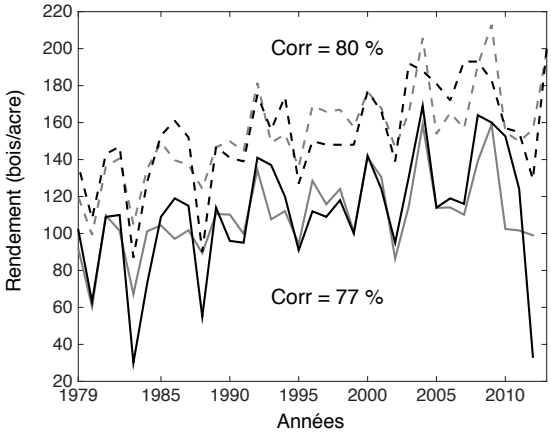


FIGURE V.7 – Séries temporelles des rendements observés (en noir) et prédits (en gris) pour deux cantons (lignes continues et pointillées) du district 60 d'Illinois. Les prévisions sont faites en utilisant le modèle d'impact ME-canton. Les corrélations entre les séries temporelles observées et prédites sont également indiquées.

Même si les événements extrêmes - en particulier les valeurs à faible rendement (par exemple pour les années 1983 ou 1988) - ne sont pas bien estimés par le modèle d'impact (comme on l'a vu à la Figure V.6), la prévision semble être satisfaisante avec une corrélation de 80%. Il est surprenant d'obtenir une corrélation aussi élevée, étant donné que l'anomalie de rendement pour l'état de l'Illinois affiche une corrélation d'environ 60% (Figure V.6), mais cette corrélation accrue résulte de l'utilisation de la tendance qui est ici connue sans incertitude. En outre, les séries temporelles et les courbes de tendance sont, par construction, fortement corrélées. Cela explique également pourquoi les prévisions de production de maïs s'ajustent si bien aux différents cantons dans le même état : la tendance a été obtenue à la Section 7.1 page 120 pour chaque canton à l'aide d'un modèle non-linéaire global à effets mixtes.

1.4 Exploitation du modèle d'impact ME-canton

Estimation de la production de maïs en fin d'année (monitoring)

Le modèle d'impact ME-canton a été développé dans les sections précédentes pour quantifier et analyser les liens entre la météo et les rendements de maïs. Il peut également être utilisé pour évaluer les effets de la météo sur la production de maïs $p(t) = y(t) \times s(t)$, où $y(t)$ est la prévision du rendement du maïs de l'année t (Section 1.3) et $s(t)$ la superficie récoltée (en acres). La Figure V.8 représente la production de maïs qui en résulte pour quatre cantons éloignés du district 60 d'Illinois. La régression est assez impressionnante, et les anomalies interannuelles sont bien prises en compte, sauf pour quelques années (par exemple, en 2008), et les différences entre cantons sont bien représentées.

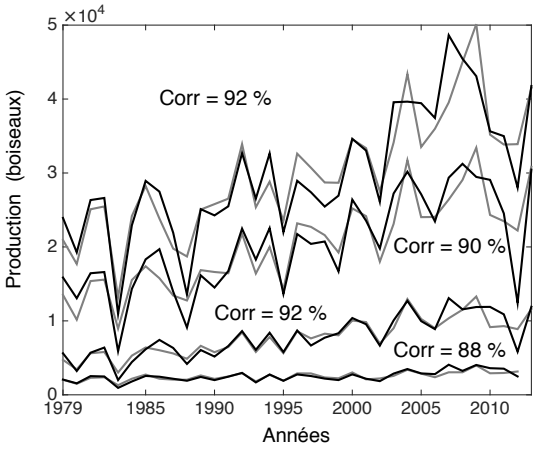


FIGURE V.8 – Production de maïs (en boisseaux) observée (en noir) et prédite (en gris) pour quatre cantons du district 60 de l'Illinois. La prévision utilise le modèle d'impact ME-canton. Les corrélations entre les séries temporelles observées et celles prédites sont également indiquées.

Par rapport aux prévisions de rendement de la Figure V.7, en plus de l'anomalie de rendement et de l'information de tendance, ce chiffre utilise également la série temporelle des surfaces récoltées $s(t)$. Cela améliore évidemment la qualité de la prévision.

Des Figures V.6 à V.7, la corrélation entre les valeurs observées et celles prédites est passée de 60% (pour la prévision des anomalies de rendement) à 80% (pour la prévision des rendements), soulignant la quantité d'information fournie par la tendance temporelle. De même, entre la Figure V.7 et la Figure V.8, la corrélation entre les valeurs observées et celles prédites est passée de 80% (pour la prévision des rendements) à 90% (pour la prévision de la production), soulignant la quantité d'information fournie par les données de surfaces récoltées.

La Figure V.9 représente la capacité de prévision de la production de maïs pour les trois modèles mixtes ME considérés dans cette étude, et utilisant l'ensemble des données des États-Unis : classification par canton, par district ou par état fédéral. Les résultats sont représentés pour le canton Montgomery en Illinois, puis pour le district 60 de l'Illinois, et enfin pour l'état d'Illinois. On constate que la sensibilité de la qualité du modèle ME à l'échelle spatiale de regroupement (canton, district ou état) est limitée puisqu'ils donnent tous des résultats similaires (confirmant ce qui a été vu dans la Section 1.2). En outre, les statistiques au niveau du canton, du district ou de l'état sont très similaires, soit environ 90% de la corrélation, de sorte que les modèles peuvent être utilisés adéquatement aux trois échelles spatiales selon les besoins, pour prédire les productions de maïs.

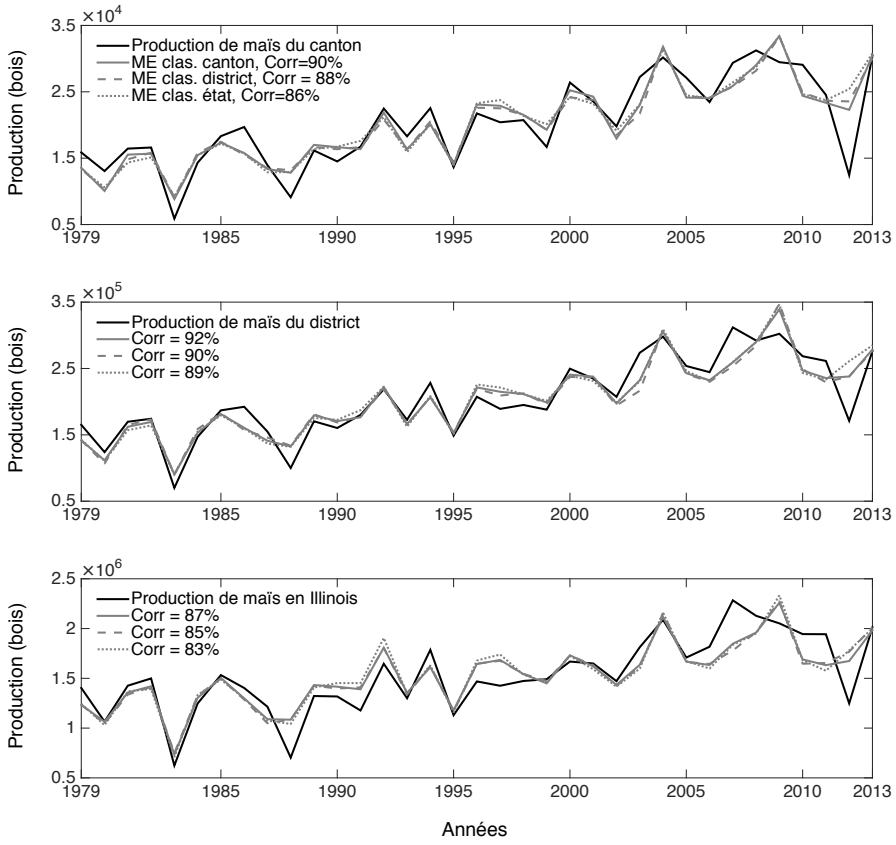


FIGURE V.9 – Prévision de la production de maïs pour les trois classifications géographiques du modèle ME, entrainé à l'échelle des États-Unis (classification par canton, par district, et par état fédéral) pour le canton Montgomery (17135) en Illinois (en haut), pour le district 60 d'Illinois (au centre), et pour l'état d'Illinois (en bas).

Prévisions saisonnières

Dans cette section, nous quantifions les informations fournies par les entrées du modèle du mois de mai à celui d'août sur le modèle d'impact ME-canton : parmi les six entrées sélectionnées page 127 (Tmai, Tjuin, Tjuillet, Taoût, Pjuillet et Paoût) le modèle ME-canton est entrainé en premier avec seulement les entrées disponibles (1) jusqu'en mai (donc seulement Tmai). Un autre apprentissage s'effectue avec les entrées disponibles (2) jusqu'en juin (Tmai et Tjuin) ; (3) jusqu'en juillet (Tmai, Tjuin, Tjuillet et Pjuillet) et enfin (4) jusqu'en août (les six entrées sélectionnées).

De mai à août, la capacité prédictive de ce modèle augmente de $Corr = 1\%$ à $Corr = 74\%$ pour le district 60 d'Illinois (courbe en gris clair en haut de la Figure V.10). La météo en mai n'influence pas le rendement du maïs (du moins le modèle ne perçoit pas cette influence). Au contraire, l'ajout de juin et, plus important encore, l'ajout des variables météorologiques de juillet (Tjuin, Tjuillet et Pjuillet) améliorent fortement la capacité prédictive du modèle.

La Figure V.10 (milieu) illustre comment la prévision des anomalies de rendement du district est améliorée lors de l'ajout progressif des entrées météo selon le mois limite (jusqu'en juin, puis jusqu'en juillet et enfin jusqu'en août) : la prévision des anomalies du district avec seulement l'entrée Tmai (la ligne continue grise) est à peu près égale à zéro pour chaque année et très loin de la ligne noire (les anomalies de rendement observées). Cela signifie que lorsqu'on utilise l'information disponible seulement en mai, aucune information sur le rendement annuel n'est fournie par le modèle ME-canton. Lorsque les informations météorologiques de juillet et d'août sont ajoutées, la capacité de prévision augmente, soulignant l'importance des derniers mois pour la culture. Malheureusement, les prévisions saisonnières ne semblent pas être vraiment possibles dans cette application, car la prévision des rendements n'est raisonnable qu'à partir d'août, ce qui est relativement tard. Cependant, il faut noter qu'aucune information sur l'état de la récolte n'est utilisée pour cette prévision saisonnière (comme la biomasse de la plante au moment de l'estimation), qui est souvent une information importante pour ce type d'application saisonnière. La Figure V.10 (en bas), illustre les mêmes quatre configurations d'entrée, mais pour la prévision de la production. Les trois courbes grise en pointillées (prévisions de juin, juillet, puis août) sont beaucoup plus proches l'une de l'autre que les anomalies de rendement. L'amélioration apportée par les données météorologiques des mois successifs est souvent moins importante pour la production de maïs. Cela s'explique par l'utilisation de la même tendance et des données de surface recoltée pour les quatre configurations d'entrées.

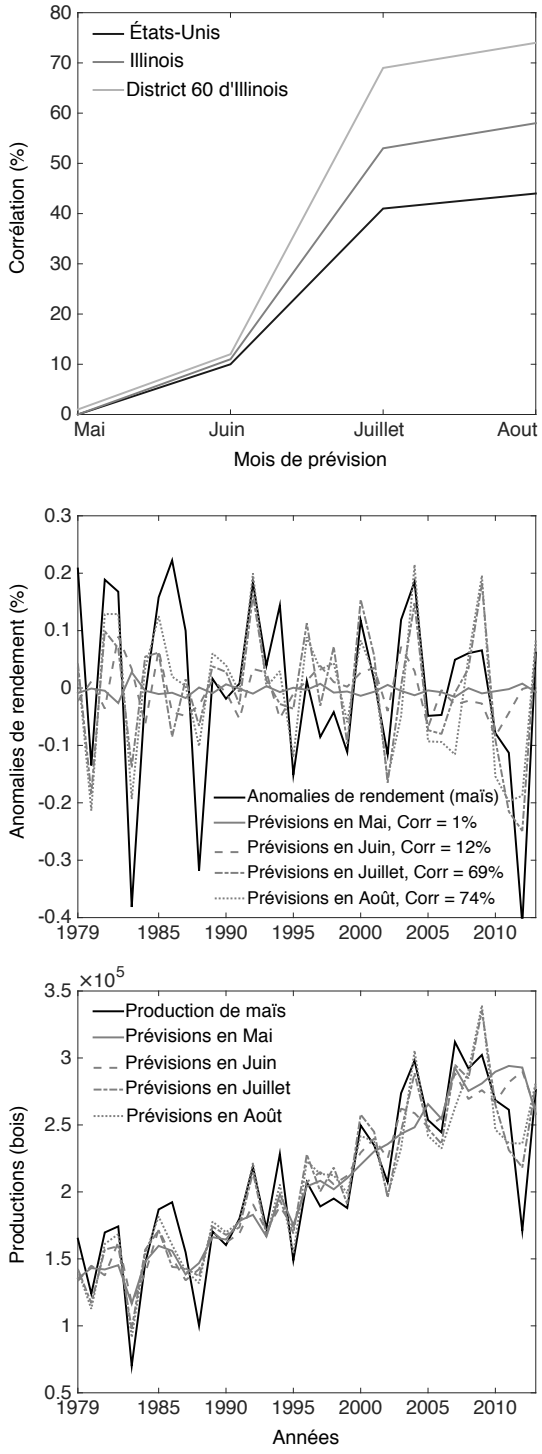


FIGURE V.10 – Illustration de la prévision saisonnière. Haut : corrélation entre les valeurs observées et celles prédites du rendement de maïs par le modèle ME-canton (à l'échelle des États-Unis, de l'état et du district), selon le mois de la prévision.
Milieu : Séries temporelle des prévisions d'anomalies de rendement.
Bas : Séries temporelle des prévisions de production selon le mois de prévision pour le district 60 d'Illinois.

1.5 Comparaison avec le modèle de l'USDA

Une prévision statistique du rendement de maïs à l'échelle des États-Unis est fournie par le ministère de l'agriculture des États-Unis (USDA). Comme on a pu le voir au Chapitre II page 38, l'USDA utilise deux importants sondages pour la prévision et l'estimation des rendements [Aune06]. Dans le premier sondage, des agriculteurs sélectionnés aléatoirement sont invités à déclarer leur rendement final ou, leur meilleure évaluation du rendement potentiel en fonction des conditions actuelles pendant la période de prévision. Le second sondage, appelé "Objective Yield Surveys", utilise des comptages de plantes et des mesures de fruits. Le modèle de l'USDA n'utilise pas de données météorologiques et est basé sur des informations directement liées à la plante ou au fruit donc davantage lié au rendement. La Figure V.11 compare les prévisions de notre modèle ME-canton avec les résultats donnés par le rapport de septembre de l'USDA [Irwi14]. Le modèle ME n'utilise pas les données météorologiques après le mois d'août, de sorte que les deux modèles peuvent être comparés.

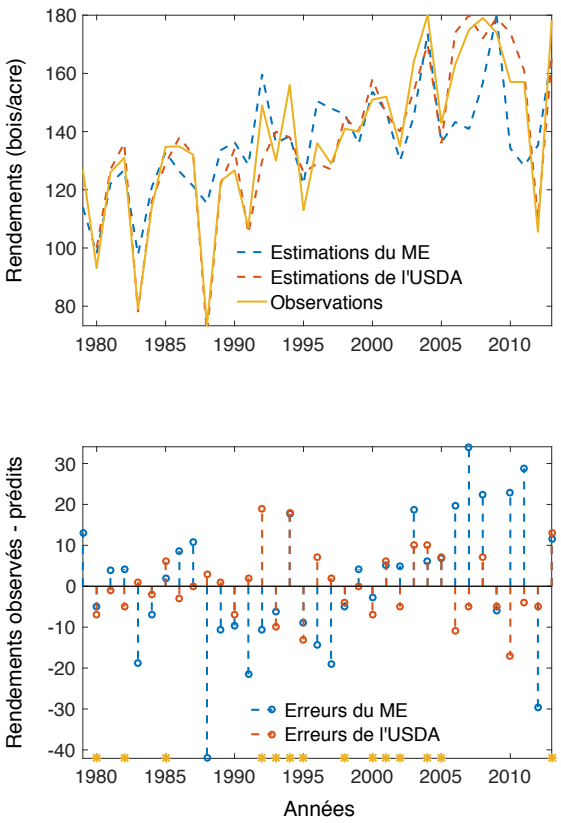


FIGURE V.11 – Comparaison des prévisions du modèle de l'USDA (rapport de septembre) avec celles du modèle ME-canton dans l'état d'Illinois.
Haut : séries temporelles des prévisions et observations du rendement de maïs en Illinois. Bas : erreurs de prévision ME et USDA. Les points oranges soulignent les années mieux estimées par ME que par le modèle de l'USDA.

Les résultats de l'USDA sont meilleurs, en particulier pour les valeurs de rendement extrêmes. Cependant, le ME fournit de meilleurs résultats que le modèle de l'USDA 14 fois sur 35, soit 41% des 35 années de l'étude (Figure V.11, en bas). Les meilleurs résultats du modèle de l'USDA étaient attendus car ils utilisent des intrants directement liés à la production de maïs, alors que les données météorologiques influencent plus indirectement la production de maïs. Nos résultats sont attendus, lorsqu'il s'agit d'uniquement utiliser des informations météorologiques : en effet, selon [Deep15], la variation météorologique explique un tiers de la variabilité globale des rendements des cultures.

Toutefois, le coût des deux enquêtes nécessaires pour le modèle de l'USDA (en particulier, le sondage "Objective Yield survey") est beaucoup plus élevé que celui de la collecte mensuelle des températures et des précipitations. Les méthodes de statistiques agricoles prennent beaucoup de temps et sont très coûteuses [Lian04]. De plus, en utilisant des entrées beaucoup plus simples, nos modèles peuvent effectuer des prévisions passées et futures sans aucune enquête.

1.6 Premières conclusions

Cette première étude présente une méthodologie pour construire des modèles d'impact météorologique en agriculture. On discute de l'évaluation de la sensibilité à la météo et des prévisions saisonnières. Dans le contexte agricole, la construction de tels modèles d'impact subit quelques difficultés : dans la plupart des cas, il n'existe pas suffisamment de données pour représenter pleinement les relations biophysiques complexes. En outre, des informations importantes peuvent être manquantes (telles que les propriétés du sol ou les pratiques agricoles) pour effectuer une bonne prévision.

Dans cette première étude, nous nous sommes concentrés sur la question du surapprentissage et sur l'évaluation fiable du modèle d'impact dans ce contexte de faible nombre de données. Ces préoccupations ne sont pas toujours suffisamment détaillées dans la littérature, ce qui conduit parfois à une surestimation de ce qui peut être obtenu avec ce type de modèles (pour les modèles d'impact biophysique ou statistique). Trois types de modèles statistiques ont été considérés : un modèle linéaire, un réseau de neurones, et les modèles linéaires à effets mixtes. L'échelle spatiale d'apprentissage du modèle est un aspect important, le modèle pouvant être entrainé au niveau du canton, du district, de l'état fédéral ou du pays. Selon les résultats, même si les différents modèles semblent proches, le choix semble dépendre de la résolution spatiale choisie. Si les données agricoles sont disponibles à la résolution la plus fine (Tableau V.1), le modèle ME avec classification par canton, est une bonne solution. Il fournit les meilleurs résultats sur les régions les plus sensibles à la météo et les plus productives en maïs ("Régions centrales").

Si les données sont disponibles uniquement au niveau de l'état (Tableau V.2), les modèles ME deviennent équivalents au modèle LIN : moins de données sont disponibles pour l'apprentissage et plus de paramètres doivent être estimés pour le modèle ME que pour le LIN. À cette résolution (état), même si le modèle ME avec classification par canton donne les meilleurs résultats pour les régions centrales, le modèle LIN est une solution simple et efficace en moyenne. La meilleure qualité du modèle NN réside dans son homogénéité spatiale et même s'il ne donne pas les meilleurs taux de généralisation, il fournit de meilleurs résultats que les modèles LIN ou ME sur certains états (Figure V.5).

Ainsi, nos résultats montrent que pour cette application particulière, l'état fédéral comme échelle spatiale d'apprentissage est un bon compromis : elle permet de se spécialiser aux conditions locales tout en conservant suffisamment de données pour étalonner le modèle linéaire ou le réseau de neurones. Même si la non-linéarité du réseau de neurone autorise en théorie le modèle à se spécialiser aux conditions locales, l'information de groupe utilisée dans les modèles ME est une information plus directe que ce que le NN semble pouvoir déduire uniquement des entrées météorologiques. Notre modèle ME avec classification par canton peut prédire les anomalies de rendement de maïs du canton avec une corrélation de 50% (en moyenne) entre les valeurs prédites et celles observées. Dans les régions plus sensibles à la météo (Missouri, Illinois, Indiana, Ohio, Pennsylvanie, Virginie, West-Virginia, Kentucky et Tennessee), cette corrélation s'élève à 60% : cela signifie qu'environ 40% de la variance du rendement s'explique par les entrées météorologiques mensuelles.

Comme indiqué dans [Deep15], ces valeurs sont cohérentes avec la simplicité des modèles considérés (en particulier, les informations limitées apportées par les prédicteurs météorologiques moyens mensuels). Une étude plus détaillée sur l'Illinois montre que le modèle ME-canton donne une prévision satisfaisante avec une corrélation observation/prédiction entre 70% et 80% au niveau du canton. Les prévisions saisonnières

montrent que, compte tenu des entrées disponibles en juin seulement, le modèle MEcanton donne une corrélation pour les anomalies de rendement de seulement 12%. Cette valeur augmente à 70% lorsqu'on considère les entrées des mois de juin et juillet.

Les perspectives sont nombreuses pour améliorer ces modèles d'impact météorologique : l'étude qui suit quantifie les avantages apportés par l'utilisation d'indices agroclimatiques au lieu de simples données météorologiques mensuelles.

L'USDA fournit des données régulières sur les dates habituelles de plantation et de récolte pour les différents états des États-Unis [USDA10]. Dans certains états, le maïs est planté avant mai. À ce stade de la thèse, les performances des modèles ne nous permettent pas de considérer les prévisions saisonnières de rendements de maïs à partir de mai. Cependant, les résultats pour les prévisions saisonnières sont cohérents avec la simplicité des modèles, et en particulier le faible nombre d'entrées. Des améliorations prévisionnelles sont attendues grâce à l'enrichissement de la liste des prédicteurs potentiels.

2 Modélisation à l'aide d'indices agroclimatiques également

L'idée de la première analyse (Section 1 page 126) était de comparer différents modèles statistiques sur une base de prédicteurs simples pour prédire des anomalies de rendement. L'idée de cette nouvelle section n'est pas de garder à nouveau cette diversité de modèles mais de sélectionner celui qui nous parait le plus approprié et de regarder - pour ce modèle fixé - l'apport d'information fourni par les indices agroclimatiques pour la prévision des anomalies de rendement.

Il nous faut donc choisir un seul modèle parmi ceux qui ont été testés à la Section 1. On décide après les comparaisons faites précédemment, de se concentrer sur le modèle linaire à effets mixtes, avec une classification spatiale selon les cantons, et entrainé avec l'ensemble des données des États-Unis.

2.1 Utilisation des indices agroclimatiques dans la littérature

Afin de fournir au modèle de prévision, des informations plus directement liées au rendement des cultures, certains indices agroclimatiques (c'est-à-dire les indices basés sur la météo et l'agriculture) sont ajoutés en entrées. Traditionnellement, les indices agroclimatiques sont obtenus à partir des données météorologiques directes afin de mieux représenter le lien entre la météo et la croissance des cultures et pour faciliter la prise de décision en agriculture [Lepa12, Caub15]. Par exemple, [Lass93], [Robe13] ou [Born12] ont utilisé les degrés-jours de croissance ; [Butl13] a utilisé les degrés-jours de croissance et de mort ; [More15], [Zhan11] ont utilisé l'évapotranspiration ; et [Mora14] a utilisé divers indices agroclimatiques pour évaluer l'adéquation du vin à une région.

Certaines caractéristiques agricoles sont parfois utilisées (par exemple, les propriétés du sol, la taille des grains) [Aune06, Asse13] comme entrés des modèles statistiques. Les indices agroclimatiques, en particulier les degrés-jours de croissance ou certains types d'indices de stress hydrique (par exemple, Précipitation-Évaporation), ont longtemps été utilisés pour les prédictions de rendement des cultures [Lass93, Hol11]. L'utilisation de données météorologiques directes n'est devenue courante qu'au cours des dernières décennies, lorsque les données météorologiques quadrillées sont devenues un intrant majeur pour la prévision de rendement à l'échelle régionale [Zhan11, More15]. Les indices agroclimatiques, couplés aux modèles climatiques ou aux méthodes de régionalisation, sont de plus en plus utilisés comme traceurs du changement climatique : [Qian10, Bela00, Boot05, Trnk11, Qian13] utilisent des degrés-jours de croissance effectifs et des unités thermiques de culture, et [Boot05] utilisent des degrés-jour de croissance effectifs, des déficits hydriques et des unités thermiques de culture. [Hol10] et [Hol11] ont évalué l'adéquation du climat pour l'agriculture en fonction des indices agroclimatiques.

Certaines études rapportent les changements dans la tendance des indices agroclimatiques aux changements dans la phénologie des cultures pour des régions d'Eurasie [Menz03, Moon02, Crop05, Seme06], ou pour les pays du Sahel [Ben 02], ou dans la majeure partie de l'Amérique du Nord [Robe02, Feng04, Tera11] (ce dernier, avec les jours de gel, et l'indice de stress thermique).

[Torv04] ont analysé la relation importante entre les rendements de pommes de terre/ orge/avoine/blé, et la température/précipitation en Norvège. Selon [Tera12], les indices sélectionnés (jours de gel, temps thermique, indice de stress thermique) sont importants pour comprendre l'impact potentiel des changements climatiques anthropiques futurs sur la production agricole. D'autres études ont examiné la dépendance de

l'agriculture à la variabilité climatique en analysant les séries temporelles de plusieurs indices climatiques [Bazg14].

[Qian10] ont analysé un ensemble d'indices agroclimatiques représentant les conditions climatiques canadiennes pour la production de fourrage. Quant a [Qian01], ils ont utilisé les précipitations, la température maximale, la température minimale et le rayonnement solaire. [Bela00] ont utilisé le déficit hydrique (précipitation - évapotranspiration), les degrés-jours de croissance effectifs et les unités thermiques du maïs (UTM) afin de déterminer l'impact du climat sur l'agriculture au Québec. Ces indices sont considérés comme les variables climatiques les plus utilisées pour déterminer l'adéquation d'un terrain à la culture de céréales, selon [Agro95]. [Grac16] ont examiné les durées de la saison de croissance et la saison sans gel, la date du dernier gel printanier, et celle du premier gel automnal, ainsi que les sommes annuelles des degrés-jours de croissance pour trois valeurs seuil de température en Pologne.

Seules quelques études [Trnk11, Lali13, Pelt10] ont testé plus de 10 indices agroclimatiques (tels que la durée de la période sans gel, les unités thermiques du mais UTM, le SPEI (indice de Précipitation-Evapotranspiration standardisé) ou l'humidité du sol - voir Tableau III.1 page 60 - en vue de prédire le rendement des cultures. Nous proposons ici d'analyser plus de 50 indices agroclimatiques différents.

Le but de cette nouvelle analyse est : **(1)** de proposer une approche qui identifie les indices agroclimatiques les plus importants pour la prévision des rendements agricoles, **(2)** d'identifier quelles régions américaines sont les plus sensibles à la météo pour le rendement de maïs, et surtout, **(3)** de mesurer l'impact de l'utilisation des indices agroclimatiques au lieu des informations météorologiques non-transformées pour la prévision du rendement. Une analyse de sensibilité est présentée à la Section 2.3 et les améliorations des prévisions de rendement à la Section 3.

2.2 Les données agricoles par bassins de production

Une description des principaux bassins de production de maïs aux États-Unis a déjà était faite au Chapitre III page 2. Par la suite, "zone 1" désignera la Corn-Belt, "zone 2" les grandes plaines du nord, "zone 3" les plaines du centre, "zone 4" la vallée du Mississipi, et enfin "zone 5" la côte est (Figure V.12).

Dans cette nouvelle étude, les 28 états considérés sont : Illinois, Indiana, Iowa, Kentucky, Michigan, Minnesota, Missouri, Ohio et Wisconsin (zone 1) ; Dakota du Nord et Dakota du Sud (zone 2) ; Kansas, Nebraska et Oklahoma (zone 3) ; Arkansas, la Louisiane, le Mississipi et le Tennessee (zone 4) ; New Jersey, Delaware, Maryland, Virginie, Virginie-Occidentale, Pennsylvanie, Caroline du Nord, Caroline du Sud, Géorgie, Alabama (zone 5). L'analyse a été réalisée au niveau de chaque zone de production, et aussi au niveau national (c'est-à-dire, en considérant les 28 états énumérés ci-dessus).

La Figure V.13 illustre la dispersion naturelle et la variabilité des anomalies de rendement pour les cinq zones de production. Dans la zone 1, il y a une dispersion plus faible (suggérant une sensibilité moins élevée aux conditions météorologiques et une faible variabilité inter-annuelle), mais de nombreuses valeurs extrêmes négatives (ce qui suggère une sensibilité plus élevée aux conditions météorologiques défavorables). Les zones 2, 3 et 5 présentent une variabilité naturelle très similaire. La zone 4 présente une faible dispersion et quelques valeurs extrêmes.

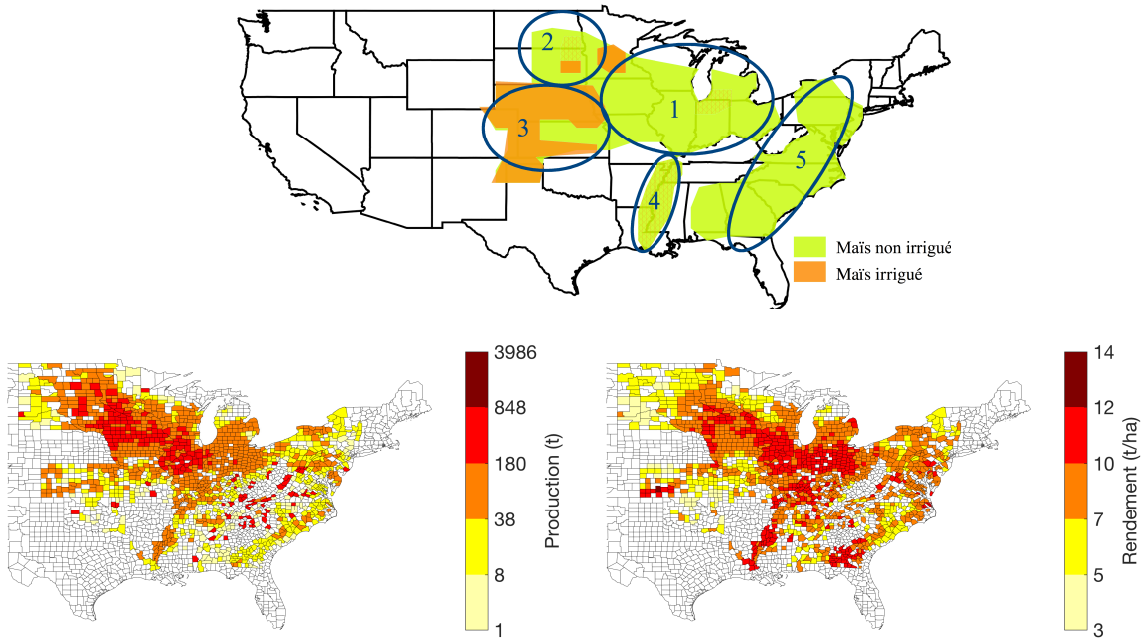


FIGURE V.12 – Haut : principaux bassins de production de maïs aux États-Unis [Leve08] (Source : USDA, recensement agricole de 2002) : (1) la Corn-Belt, (2) les grandes plaines du nord, (3) les plaines du centre, (4) la vallée du Mississipi, et (5) la côte est. Les couleurs vertes et orange indiquent respectivement des zones non-irriguées et irriguées. En bas : production de maïs (gauche) et rendement (droite) en 2013. Les cantons en blancs ne sont pas disponibles pour cette année.

2.3 Météo-sensibilité des différents bassins de production

Avant de modéliser l'impact de la météo sur le rendement du maïs, une analyse de météo-sensibilité par bassin de production peut être effectuée pour mieux comprendre notre application.

Classification des prédicteurs en fonction des zones de production

Cinquante-huit prédicteurs météorologiques potentiels sont disponibles sur notre base de données. Cette section compare et quantifie les liens entre les différents prédicteurs météorologiques possibles et les anomalies de rendement selon les bassins de production. Les meilleurs prédicteurs sont donc classés pour chaque zone. Pour quantifier la robustesse de cette sélection, nous considérons plusieurs critères de qualité : (1) la corrélation avec les anomalies de rendement Corr, (2) le RMSE et (3) le critère AIC. Le Tableau V.5 résume les six premiers prédicteurs sélectionnés (de gauche à droite) selon le critère COR. Des tableaux similaires sont donnés pour RMSE et AIC (Tableaux V.6 et V.7).

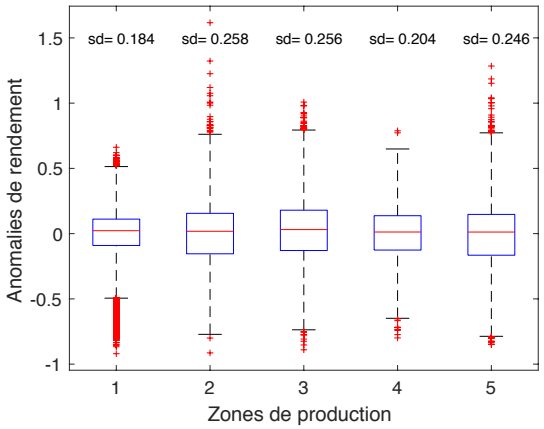


FIGURE V.13 – Boites à moustaches, d'anomalies de rendement (en %) pour les cinq zones. Elles résument la variabilité des anomalies de rendement pour les cinq zones de production. L'écart-type des statistiques est également fourni.

Zone	⟶ Sélection Successive des Prédicteurs ⟶					
1	DJaoût	SPEIjuillet	OCA	SMavril	SPEIjuin	(UTM)
	0.38	0.47	0.50	0.52	0.54	0.55
2	OCA	Tseptembre	SPEIjuillet	Tjulillet	P-ET	DJaoût
	0.28	0.33	0.36	0.40	0.41	0.42
3	Tjuillet	DJaoût	SPEIjuin	DJoctobre	Toctobre	(OCA)
	0.44	0.46	0.49	0.50	0.52	0.53
4	DJjuillet	Paoût	SPEIjuin	TCD	Tavril	(Tseptembre)
	0.50	0.53	0.57	0.59	0.61	0.62
5	SMjuillet	SPEIjuin	SPEIjuillet	Tjuillet	SMavril	DJaoût
	0.42	0.50	0.52	0.55	0.57	0.58

TABLEAU V.5 – Les six premiers prédicteurs sélectionnés (successivement) pour les cinq zones de production, selon le critère de corrélation. Les chiffres indiquent la corrélation entre les anomalies de rendement et la sortie d'un modèle linéaire construit avec le prédicteur situé au dessus du chiffre, et les prédicteurs précédents. Les prédicteurs qui n'apparaissent pas entre parenthèses seront utilisés comme entrées de modèle, plus loin dans le chapitre. La procédure de sélection est la même que celle utilisée dans le début du chapitre : on rappelle qu'elle est décrite en détail au Chapitre IV, page 123. Voir Tableau III.1 page 60 pour la signification des abréviations.

Zone	⟶ Sélection Successive des Prédicteurs ⟶									
1	DJaoût	SPEIjuillet	OCA	SMavril	SPEIjuin	UTM	DJjuillet	SPEIaoût	DJmars	DJsc
2	OCA	Tseptembre	SPEIjuillet	Tjuillet	P-ET	DJaoût	Paug-oct	DJavril	SMavril	Pmars
3	Tjuillet	DJaoût	SPEIjuin	Toctobre	DJoctobre	OCA	Tseptembre	LSC	Poctobre	Taoût
4	DJjuillet	Paoût	SPEIjuin	Paug-oct	DJseptembre	TMA	Toctobre	DGP	SMaoût	Pavril
5	SMjuillet	Tjuillet	SPEIjuin	SPEIjuillet	SMavril	DJaoût	Tseptembre	FFP	Pavril	SPEIaoût

TABLEAU V.6 – Les dix premiers prédicteurs météorologiques successivement sélectionnés pour les cinq zones, selon le critère AIC.

Zone	⟶ Sélection Successive des Prédicteurs ⟶									
1	DJaoût	SPEIjuillet	OCA	SMavril	SPEIjuin	UTM	DJjuillet	SPEIaoût	DJmars	DJsc
2	OCA	Tseptembre	SPEIjuillet	Tjuillet	P-ET	DJaoût	Paug-oct	DJavril	SMavril	Pmars
3	Tjuillet	DJaoût	SPEIjuin	Toctobre	DJoctobre	OCA	LSC	Tseptembre	Poctobre	Taoût
4	DJjuillet	TCD	SPEIjuin	Paoût	Tavril	Tseptembre	Paug-oct	Toctobre	TMA	SMaoût
5	SMjuillet	SPEIjuin	SPEIjuillet	Tjuillet	SMavril	DJaoût	B-UTM	DGP	TMA	Pavril

TABLEAU V.7 – Les dix premiers prédicteurs météorologiques successivement sélectionnés pour les cinq zones, selon le critère RMSE.

Pour les zones 1, 2 et 3, les trois critères de qualité (Corr, RMSE et AIC) ont sélectionné les mêmes 10 premiers prédicteurs. Dans le cas des zones 4 et 5, les variables sélectionnées sont très similaires, mais elles ne sont parfois pas choisies dans le même ordre (on sait que ces procédures de sélection peuvent légèrement varier en fonction des critères de qualité, en particulier pour l'ordre des prédicteurs). Cependant, ces fortes similitudes confirment la robustesse de ces résultats.

Dans les paragraphes suivants, nous interprétons la sélection des prédicteurs en termes agronomiques. L'information fournie par cette classification est de deux types : d'une part, elle donne des informations sur les mois d'importance et, d'autre part, elle indique qui de la température ou de l'apport en eau revêt une plus grande importance pour le rendement.

La sélection pour la zone 1 souligne l'importance de DJaoût (variable corrélée négativement avec les anomalies du rendement), et ensuite de l'eau en juillet. En général, en raison de la demande croissante en eau de la biomasse pendant la période de croissance, le plus bas volume en eau du sol est atteint en août. C'est une zone à irrigation limitée, de sorte que l'approvisionnement naturel en eau est important pendant la période de chaleur. SPEIjuillet est sélectionné comme le deuxième indicateur météorologique le plus important (positivement corrélé). Le maïs est très sensible à l'humidité en juillet et en août, car c'est la période de formation des grains. Pendant le stress hydrique, surtout à faible humidité relative, les températures élevées peuvent déshydrater les soies et endommager ou tuer le pollen. La pollinisation ne sera pas affectée par des températures élevées s'il y a suffisamment d'humidité dans le sol car la libération du pollen survient habituellement pendant les heures du matin.

Le stress hydrique pendant la pollinisation est donc un facteur aggravant la réduction de rendement du maïs. Un stress sévère en matière d'humidité entraînera un ralentissement de la production des soies et une réduction de la pollinisation en raison d'un manque de pollen viable et d'une réduction de la synchronisation entre l'apparition des soies et la pollinisation. Sous un stress sévère, certaines plantes ne formeront aucune soie, ou des soies apparaîtront après la fin de la production de pollen ; résultant en épis mal développés. Le stress hydrique pendant cette étape peut entraîner une baisse du rendement de 50% [Wiat10]. En outre, c'est une zone où le nombre de jours avec $T > 30^oC$ est faible. Cependant, OCA est sélectionné comme 3ème prédicteur, ce qui semble avoir un impact significatif. C'est une zone où le nombre de jours avec $T > 30^oC$ est souvent bas. Cependant, la fréquence des températures supérieures à 30^oC (OCA) est choisie comme troisième prédicteur, ce qui semble avoir un impact significatif sur le rendement. Ainsi, les données montrent que le stress thermique est également important, comme souligné dans [Schl09]. De plus, il est difficile d'indiquer qui de la contrainte thermique ou hydrique est la plus importante pour la zone 1, car ces deux contraintes interagissent les unes avec les autres.

La zone 2 est située au nord, où le nombre de jours avec $T > 30^oC$ est faible. Pour cette zone aussi, le paramètre OCA semble expliquer la variabilité des anomalies de rendement (corrélation négative, mais faible par rapport aux autres corrélations des premiers prédicteurs sélectionnés dans les autres zones). Les deux états de la zone 2 sont parmi ceux ayant les dates de récolte les plus tardives (8oct-19nov en moyenne). La température de septembre (corrélée positivement) peut donc influencer le rendement. Ensuite, l'importance de l'eau en juillet est soulignée avec la sélection de SPEI (et non la précipitation) et de la température en juillet (le mois où le maïs grandit le plus). Le SPEI et l'humidité du sol (SM) apparaissent souvent avant la précipitation :

pour la zone 4, il semble que l'évapotranspiration soit moins limitante. Ceci est logique puisque ces indices agroclimatiques représentent mieux l'eau disponible pour la plante que la précipitation.

Pour la zone 3, les deux prédicteurs les plus importants sont liés à la température, ce qui est cohérent pour une zone qui demande une irrigation modérée (les régions fortement irriguées ont été retirées de notre analyse). Tjuillet est sélectionné en premier (anti-corrélé) : juillet est la période de floraison du maïs. Une température élevée en juillet affecte la photosynthèse qui devient moins efficace : les enzymes impliquées dans les réactions chimiques de la photosynthèse sont sensibles à la température et sont modifiées à haute température [Berr80]. [Craf02] montre que les températures nécessaires pour dénaturer les enzymes sont supérieures à 45^oC (une valeur rare pour Tjuillet) mais à des températures supérieures à 20^oC, le taux de photosynthèse diminue parce que les enzymes ne fonctionnent pas aussi efficacement. En outre, [Craf02] a montré que des températures plus basses altèrent la photosynthèse, mais cet impact ne semble pas prédominant lors de l'analyse des données. Le stress thermique entraîne des dommages morphologiques et réduit donc le rendement [Voll05]. Le stress thermique réduit la photosynthèse car il s'agit de la partie fonctionnelle de la plante la plus thermo-sensible. Le rendement du maïs peut également être réduit en raison des températures élevées de l'air (30^oC et plus) pendant la pollinisation [Sanc14]. Ici aussi, bien que l'analyse des données semble accorder plus d'importance au stress thermique qu'au stress hydrique (parce que les prédicteurs liés à la chaleur sont plus nombreux que ceux liés à l'eau), il est difficile d'affirmer la prédominance de l'un sur l'autre tellement leur impact sont étroitement liés. Les températures élevées n'affectent pas la pollinisation, mais elles semblent nuire pendant la phase de pollinisation. Les mois de juillet et d'août sont encore sélectionnés pour cette zone. Selon les zones, T ou DJ est choisi, mais il n'y a pas de prédominance générale de T sur DJ (pour juillet, c'est plutôt T qui est sélectionné, alors que pour août c'est DJ).
La sélection pour la zone 4 souligne l'importance de la température en juillet (avec une corrélation élevée de -0,5), puis des variables liées à l'approvisionnement en eau pour les mois de juin et d'août.

Enfin pour la zone 5, trois variables d'eau en juin et en juillet sont sélectionnées en premier (ce qui est cohérent parce que c'est une zone à faible irrigation) et les trois sont négativement corrélées. SPEIjuin est sélectionné comme deuxième indicateur météorologique (corrélée positivement). Le maïs grandit activement en juin et l'approvisionnement en eau est essentiel. Lorsque le stress hydrique survient en juin, la disponibilité et l'utilisation des nutriments sont réduites et les plantes affaiblies par le stress sont plus sensibles aux maladies et aux dommages causés par les insectes. La température en juillet semble également être importante. La sélection de SPEIjuillet après celle de SMjuillet est curieuse car il s'agit de deux variables très similaires (corrélation de 0,48). Cependant, la p-valeur pour le test de nullité du paramètre lié à SPEIjuillet est très faible $(3,2475 \times 10^{-263})$, donc la sélection de SPEIjuillet est significative.

Ces résultats montrent que les variations de l'eau et de la température au cours des mois d'été ont un impact important sur le rendement du maïs, avec un impact positif de l'approvisionnement en eau et un impact négatif de la chaleur. La faible information de certains indices agroclimatiques pourrait être surprenante (par exemple PGA ou UTM). Il semble que leurs valeurs soient très similaires aux États-Unis, de sorte qu'ils ne forment pas une discrimination suffisante entre les différents cantons. En outre, leur

série temporelle ne montre pas de variations importantes d'une année à l'autre. Même s'ils sont biologiquement liés au rendement du maïs, ils n'expliquent pas les variations mensuelles et inter-annuelles de l'anomalie de rendement. Par conséquent, beaucoup de ces indices ne semblent pas appropriés pour notre application, mais les conclusions pourraient être différentes pour d'autres régions du monde ou pour d'autres cultures.

Il convient toutefois de noter que l'impact thermique pourrait également dépendre de la capacité du sol à stocker l'eau (qui est une information spécifique au lieu). Pour un sol avec une faible capacité de stockage, la chaleur peut avoir un effet important sur la culture : la sécheresse et le stress thermique sont fréquents (mais pas toujours) en combinaison, en particulier en juillet et en août, ce qui conduit à des réponses plus fortes au cours de ces mois. Dans la littérature, de nombreux articles ont étudié ce point : [Kebe12] ont montré que l'effet combiné de la chaleur et du stress sévère dû au déficit d'humidité avait un impact beaucoup plus important pour le maïs non irrigué que l'effet indépendant. Des études similaires ont analysé les effets indépendants et combinés de la haute température et du stress hydrique sur le rendement des plantes [Grig11, Fabi08, Pras11].

Comparaison de la météo-sensibilité des différentes zones

Nous voulons maintenant comparer la météo-sensibilité des différentes zones de production en utilisant la classification faite à la Section 2.3. Comme les zones sont de taille différente, elles ne contiennent pas le même nombre d'échantillons. Le critère AIC est fortement influencé par le nombre d'échantillons, de sorte qu'il ne peut pas être utilisé pour comparer les différentes zones. Nous nous concentrons sur le critère RMSE et le critère Corr (le R^2 ajusté, moins influencé par la taille de l'échantillon - et donc mieux adapté aux comparaisons de modèles - fournit exactement le même classement que Corr pour les cinq zones).

Pour comparer les cinq zones, l'idée est d'utiliser un modèle (plus sophistiqué que le modèle linéaire simpliste utilisé pour classer les prédicteurs) avec les prédicteurs sélectionnés pour les différentes zones afin d'estimer le rendement de fin d'année. Les modèles à effets mixtes linéaires sont de bons candidats, comme on a pu le voir en première partie de ce chapitre. La première étape de cette approche de modélisation est de savoir combien de prédicteurs on conserve pour chaque zone. Cette étape est effectuée par validation croisée (Figure V.14). Lorsque la corrélation en généralisation n'augmente plus de manière significative, ou que le RMSE ne diminue plus de manière significative quand on considère un autre prédicteur, on arrête l'ajout de prédicteurs à la liste des entrées du modèle. Il est important de suivre le principe de parcimonie pour réduire le risque de sur-apprentissage. La Figure V.14 montre l'évolution des critères Corr (bleu) et RMSE (orange) dans les échantillons de test (c'est-à-dire, les statistiques de généralisation) lors de l'augmentation du nombre de prédicteurs. La validation croisée indique que les cinq premiers prédicteurs sont nécessaires pour les zones 1, 3 et 4. Pour les zones 2 et 5, les six premiers prédicteurs sont requis.

Le Tableau V.5 résume quels prédicteurs ont été conservés pour chaque zone (ce sont ceux qui n'apparaissent pas entre parenthèses). Lorsque la décision indiquée par Corr n'est pas exactement la même que celle indiquée par RMSE, c'est la matrice des corrélations croisées des prédicteurs qui aide à prendre la décision finale. Par exemple, dans la zone 3, la 6ème entrée OCA étant relativement corrélée avec la 1ère et la 2ème entrée Tjuilllet et DJaoût (Corr = 0.78 pour Tjuillet et 0.54 pour DJaoût), il a été décidé de

ne pas la conserver. De même pour la zone 5, l'entrée 7 (B-UTM) est relativement corrélée avec les entrées précédentes, de sorte que la sélection a été interrompue à l'entrée 6. Par conséquent, les entrées sélectionnées sont relativement peu corrélées. Finalement, nous disposons d'un modèle mixte par zone, qui utilise les entrées du Tableau V.5.

Afin de comparer les différentes zones, il est préférable de diviser le RMSE par l'écart-type de chaque zone. À la Figure V.14 (en bas à droite), les courbes de Corr pour les cinq zones sont superposées pour mieux les comparer. Étant donné que les prédicteurs diffèrent pour chaque zone, il n'y a pas d'axe commun en abscisse. De la zone la moins météo-sensible à la plus météo-sensible, on obtient le classement :

$$\overbrace{Zone\ 2 < Zone\ 3 \simeq Zone\ 1}^{\text{les 3 bassins les plus productifs}} < \underbrace{Zone\ 4 < Zone\ 5}_{\text{les moins productifs}}$$

La sous-figure avec les courbes superposées de RMSE divisé par l'écart-type de chaque zone donne des résultats similaires (Figure V.14, en bas à gauche). Les zones qui produisent le plus sont les moins sensibles à la météo, ou celles qui produisent le moins sont les plus sensibles aux variations météorologiques. Ce constat n'est pas surprenant : en plus de la pertinence du sol pour la culture du maïs, les producteurs ont sélectionné des zones qui ne sont pas trop affectées par les conditions météorologiques, ou encore les zones qui ont la possibilité d'être irriguées. Ces régions sont les plus productives.

	Apprentissage		Test	
Zones	Corr	RMSE	Corr	RMSE
1	0.62 $\pm$0.029	0.15 $\pm$0.004	0.45 $\pm$0.158	0.16 $\pm$0.026
2	0.50 $\pm$0.019	0.22 $\pm$0.005	0.32 $\pm$0.153	0.24 $\pm$0.033
3	0.58 $\pm$0.022	0.21 $\pm$0.005	0.45 $\pm$0.177	0.23 $\pm$0.025
4	0.64 $\pm$0.021	0.16 $\pm$0.002	0.53 $\pm$0.134	0.16 $\pm$0.013
5	0.69 $\pm$0.011	0.18 $\pm$0.002	0.54 $\pm$0.095	0.20 $\pm$0.009

TABLEAU V.8 – Corrélation Corr et RMSE en apprentissage et en test, entre la sortie du modèle et les observations d'anomalies de rendement selon les zones de production. Les écarts types sont également fournis, quantifiant la sensibilité de Corr et de RMSE aux 200 ensembles d'apprentissage/validation différents.

Le Tableau V.8 résume les critères Corr et RMSE en apprentissage et en test, entre la sortie du modèle et les observations d'anomalies de rendement, selon les zones de production. Les zones 2 et 3 sont les moins sensibles aux variations climatiques et les moins bien prédites par le modèle à effets mixtes. Les écarts-types sont également fournis : pour les RMSE, les écarts types sont faibles, ce qui illustre la stabilité des estimations du RMSE. En revanche, ils sont plus élevés pour Corr, en test, montrant la difficulté d'obtenir une estimation robuste pour ce critère de qualité.

Le diagramme des quantiles-conditionnels montrent la qualité de la prévision pour les cinq zones, et sont donnés à la Figure V.15. La courbe moyenne du nuage de points et quatre quantiles illustrent la dispersion des erreurs. La Figure V.15 représente la distribution conjointe des prévisions et des observations d'anomalies de rendement. La distribution conditionnelle de l'observation donnée pour chacune des prévisions

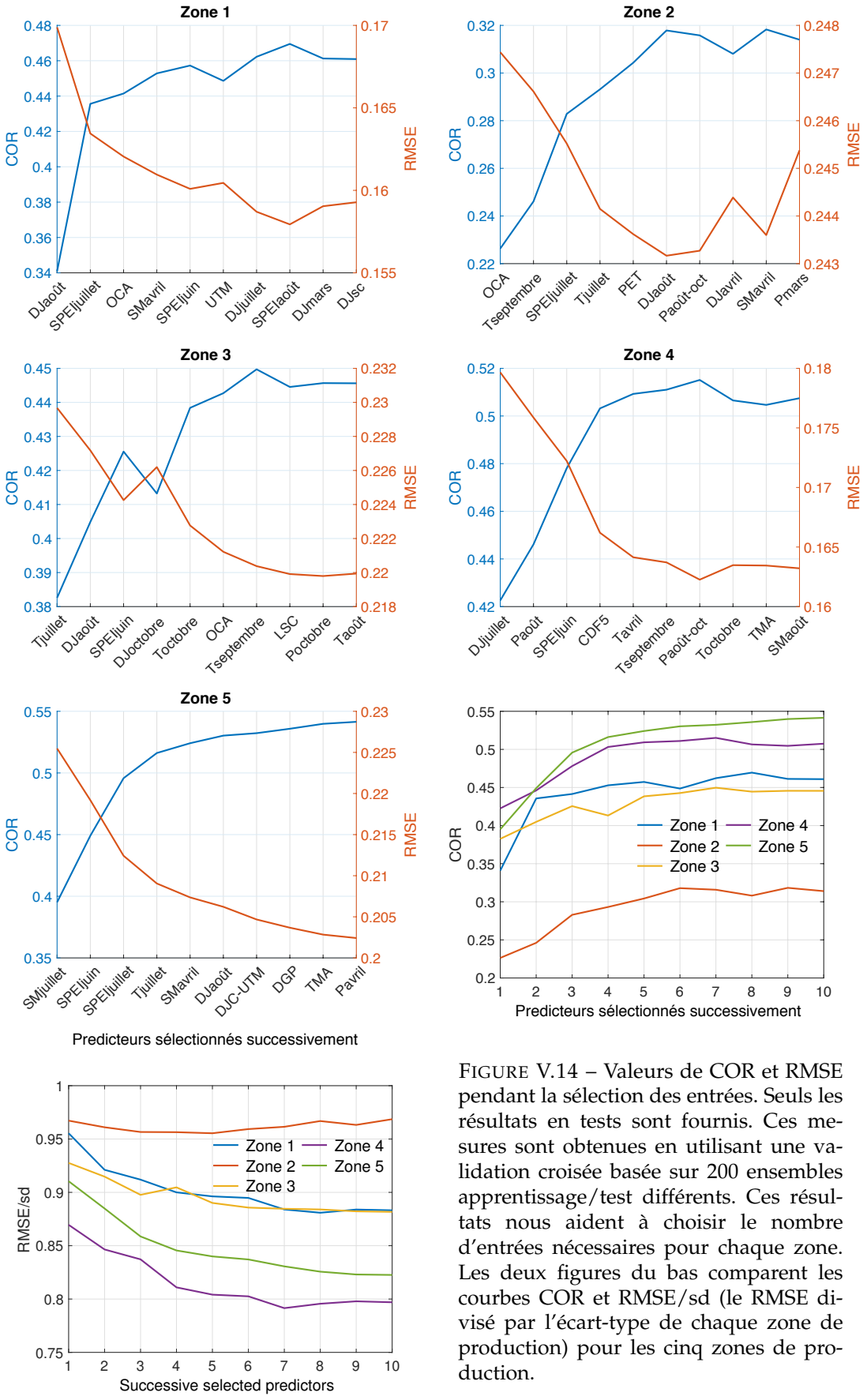


FIGURE V.14 – Valeurs de COR et RMSE pendant la sélection des entrées. Seuls les résultats en tests sont fournis. Ces mesures sont obtenues en utilisant une validation croisée basée sur 200 ensembles apprentissage/test différents. Ces résultats nous aident à choisir le nombre d'entrées nécessaires pour chaque zone. Les deux figures du bas comparent les courbes COR et RMSE/sd (le RMSE divisé par l'écart-type de chaque zone de production) pour les cinq zones de production.

est représentée en termes de quantiles sélectionnés, en comparaison avec la ligne diagonale $1:1$ représentant une prévision parfaite. Ici, on peut voir que les prévisions présentent une faible proportion de surestimation (les médianes conditionnelles des anomalies observées sont presque toujours plus petites que les prévisions). Les prévisions sont globalement non biaisées, sauf pour les queues de distribution gauches des zones 1 et 2. Les histogrammes dans la partie inférieure des sous-figures représentent la distribution des prévisions. On peut voir que les prévisions dans les zones 1, 2 et 3 sont plus raffinées, avec plus d'anomalies extrêmes estimées, surtout sur la gauche des queues de distribution.

La corrélation des anomalies de rendement peut parfois être trompeuse : si une zone de production est moins sensible à la température, la variabilité naturelle peut être faible et la corrélation sera faible (faible sensibilité à la météo) même si le modèle exploite bien les informations météorologiques. La zone 2 est la moins corrélée, mais nous avons vu à la Figure V.13 que c'est la zone avec la plus grande variabilité, en particulier avec des extrêmes positifs difficiles à expliquer. Il est très possible que la faible corrélation dans la zone 2 soit provoquée par les données extrêmes. La zone 1 semble bien prédite en moyenne : malgré une variabilité réduite du rendement de maïs (Figure V.13), elle présente de nombreux extrêmes négatifs probablement liés à des conditions météorologiques défavorables. Ce classement dépend clairement de la liste des prédicteurs météo utilisés dans cette étude. Un autre modèle pourrait être sélectionné si une autre liste de prédicteurs était utilisée [Burn02].

3 Améliorations apportées par l'utilisation d'indices agroclimatiques

L'utilisation des indices agroclimatiques comme entrées peut être compliquée car ils sont parfois difficiles à obtenir sur une grande étendue spatiale et temporelle. En particulier, il peut être difficile de les obtenir en tant que sortie de modèle climatique. Il est donc légitime de se demander (et de quantifier) si l'utilisation de ces indices apporte une réelle amélioration par rapport aux variables météorologiques non transformées. Dans cette section, le modèle de rendement n'est plus séparé selon les zones de production. Il sera ainsi plus facile de mesurer l'impact de l'utilisation des indices agroclimatiques. En outre, cela aidera à réduire les problèmes de sur-apprentissage car plus de données seront disponibles pour le modèle général, ce qui donnera des résultats encore plus robustes.

Dans cette section, deux configurations seront comparées afin de mesurer l'impact de l'utilisation des indices agroclimatiques en plus des informations météorologiques non transformées :
• la "Configuration (W)" fera référence aux indicateurs météorologiques non-transformés, à savoir les températures et les précipitations mensuelles. Il utilisera le modèle à effets mixtes linéaire construit à la section précédente (avec classification selon les cantons). Les entrées de ce modèle sont fournies au Tableau V.9 (première ligne).
• la "Configuration (I)" fera référence aux indices agroclimatiques, en plus des indicateurs météorologiques non transformés. Un autre modèle mixte linéaire formé à l'échelle nationale est également utilisé ici. Les entrées de ce deuxième modèle sont également fournies au Tableau V.9 (deuxième ligne). Les entrées sont dans l'ordre de leur sélection pour les deux configurations (W) et (I).

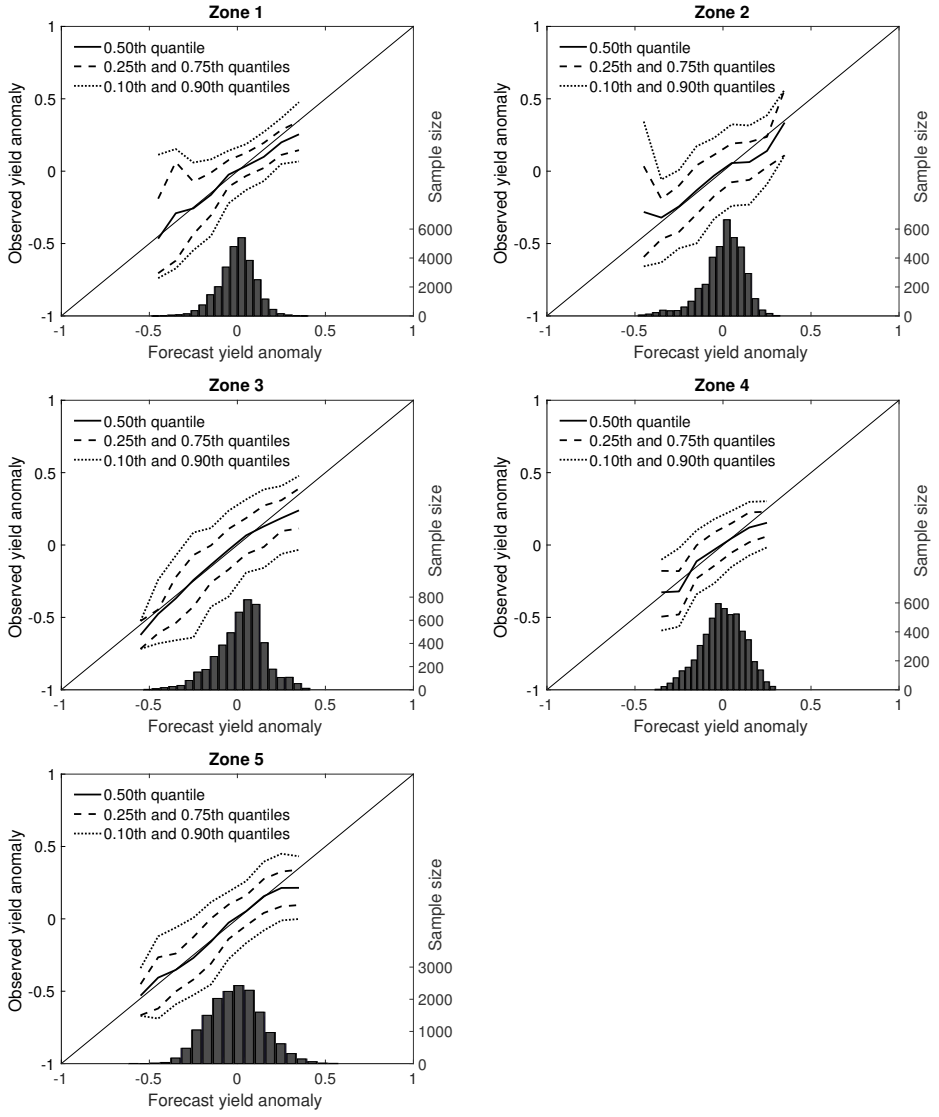


FIGURE V.15 – Tracés des quantiles conditionnels des observations et des prévisions d'anomalies de rendement pour chaque zone. Ils représentent la distribution conjointe des prévisions et des observations des anomalies de rendement. La distribution conditionnelle de l'observation donnée pour chacune des prévisions est représentée en termes de quantiles sélectionnés, par comparaison à la ligne diagonale $1:1$ représentant une prévision parfaite. Les histogrammes dans les parties inférieures des sous-figures représentent la distribution des prévisions.

Tjuillet et T/DJaoût ont été sélectionnés pour les deux configurations. En juillet, SPEIjuillet a été sélectionné pour (I) au lieu de Pjuillet pour (W) ce qui n'est pas surprenant puisque la variable SPEI est davantage liée à l'eau réellement disponible pour la culture. En juin, le bilan hydrique semble être plus informatif que la température moyenne mensuelle.

(W)	Tjuillet	Taoût	Pjuillet	Tjuin	Tmai	Paoût
COR_{cum}	0.40	0.42	0.43	0.44	0.44	0.45
(I)	Tjuillet	SPEIjuillet	SPEIjuin	DJaoût	DJavril	
COR_{cum}	0.40	0.45	0.50	0.51	0.53	

TABLEAU V.9 – Les prédicteurs météorologiques sélectionnés (écrits en suivant l'ordre de sélection) pour les deux configurations d'entrées : avec les données météorologiques directes seulement (W), ou lors de l'ajout d'indices agroclimatiques (I) à la liste des prédicteurs potentiels. Nous donnons également la corrélation COR_{cum} entre les observations d'anomalie du rendement et la prévision des modèles mixtes successifs.

Deux types d'applications sont encore considérés ici : (1) le mode "*monitoring*" nécessite la connaissance complète des indicateurs météorologiques (c'est-à-dire, de tous les prédicteurs jusqu'à la récolte). L'estimation du rendement des cultures peut alors être effectuée uniquement une fois la saison de croissance terminée. Ce mode vise à analyser la sensibilité du rendement final à la météo, obtenant ainsi une évaluation indépendante du rendement à la fin de la saison. Rappelons que les applications d'une telle modélisation sont nombreuses : aider à la gestion des stocks, guider dans le choix d'une assurance-récolte (applications à l'échelle de l'exploitation ou à l'échelle coopérative), consolider les politiques agricoles et commerciales et assurer la sécurité alimentaire dans le pays (applications à l'échelle du pays) ; aider à la gestion des prix du marché et du commerce international (applications mondiales). (2) Le mode "*prévisions saisonnières*" a pour but d'estimer le rendement final de la culture pendant la saison de croissance, avant la récolte. L'application de ce mode est simple : aider les commerçants et les grandes coopératives dans la gestion des stocks.

3.1 Amélioration en mode monitoring

Nous analysons d'abord le mode monitoring à l'échelle des États-Unis, puis nous nous concentrons sur un district particulier.

À l'échelle des États-Unis

Nous comparons ici la capacité de prévision pour les configurations (W) et (I) à l'échelle américaine. Le Tableau V.10 fournit la corrélation et le RMSE entre les anomalies observées et prédites du rendement (calculées sur des échantillons de test, pour tester la capacité de généralisation du modèle). Trente cinq années sont disponibles dans la base de données, et chaque année contient un nombre similaire de données. Pour définir un ensemble d'apprentissage 28 ans sont choisis au hasard (80% des 35 années disponibles) et les 7 années restantes sont utilisées pour l'ensemble de tests. Plusieurs mesures de qualité sont enregistrées pour cet ensemble de tests. Cette méthode est répétée 100 fois pour analyser le modèle sur différents ensembles de tests et obtenir des résultats plus robustes. Les valeurs finales des critères de qualité sont la moyenne des critères pour ces 100 expériences.
Le pourcentage de variance du rendement expliqué par les entrées météorologiques

passe de 20 % (Corr = 0,45) à 28 % (Corr = 0,53) de la configuration (W) à la configuration (I), ce qui montre l'impact significatif de l'utilisation des indices agroclimatiques en plus des informations météorologiques plus directes. L'utilisation d'indices agroclimatiques bien choisis au lieu de données météorologiques simples permet d'atteindre ces nombres.

	Config. (W) données météo directes	Config. (I) indices agroclimatiques
Corr	0.45	0.53
RMSE	0.197	0.181

TABLEAU V.10 – Corrélation Corr et RMSE (sur des ensembles de tests) entre anomalies observées et prédites du rendement de maïs, pour les deux configurations (W) et (I), à l'échelle des États-Unis (la moitié est des États-Unis). Ces résultats proviennent d'une validation croisée adaptée et des processus Monte-Carlo utilisant des ensembles apprentissage/test rassemblant 60 000 observations.

Le Tableau V.10 fournit des statistiques à l'échelle des États-Unis, mais il est possible d'analyser l'hétérogénéité spatiale de la météo-sensibilité : la Figure V.16 représente les cartes des corrélations entre l'anomalie de rendement observée et prédite, pour les deux configurations (W) et (I). De (W) à (I), la corrélation augmente pour 185 districts (sur les 210), avec une augmentation moyenne des corrélations de 0,12. Les régions qui bénéficient le plus de (I) sont situées à l'est, au centre et au nord-est. Les districts du sud et du Kansas ne se sont pas améliorés. Les indices agroclimatiques utilisés dans (I) et choisis à l'échelle des États-Unis ne semblent pas adaptés à cet état. Au contraire, les districts qui sont déjà bien prédits avec (W) (ceux ayant une corrélation supérieure à 50%) bénéficient de manière significative des indices agroclimatiques de (I). En général, l'amélioration est plus faible pour les districts moins bien estimés.

À l'échelle du district

Nous nous concentrons ici sur l'état de Virginie pour voir l'amélioration de (W) à (I) en termes de séries temporelles. D'après la première étude (première partie de ce chapitre), cet état est l'un des états les plus sensibles à la météo pour le rendement de maïs. Dans la Figure V.18, la prévision de l'anomalie de rendement (à gauche), et la prévision du rendement (à droite) sont représentées pour les deux configurations (W) et (I). Les variations globales sont bien retranscrites pour les deux configurations, mais des erreurs plus importantes pour les amplitudes extrêmes sont bien réduites dans la configuration (I), par exemple en 1980, 1983 ou en 2012. Pour cet état, les indices agroclimatiques améliorent considérablement la qualité des prévisions. La corrélation entre les anomalies de rendement observées et prédites augmente de 0,13.

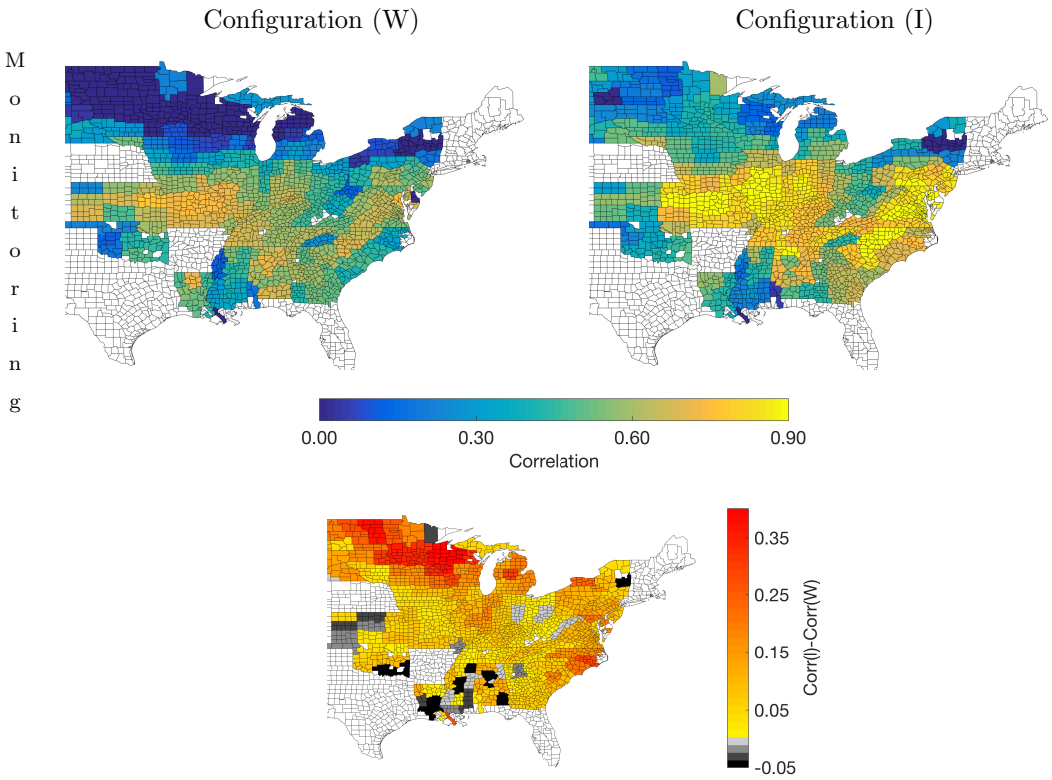


FIGURE V.16 – Haut : Cartes des corrélations Corr (entre anomalie du rendement observées et prédites, à l'échelle du district) pour les configurations (W) et (I).
Bas : Différence entre Corr(I) et Corr(W).

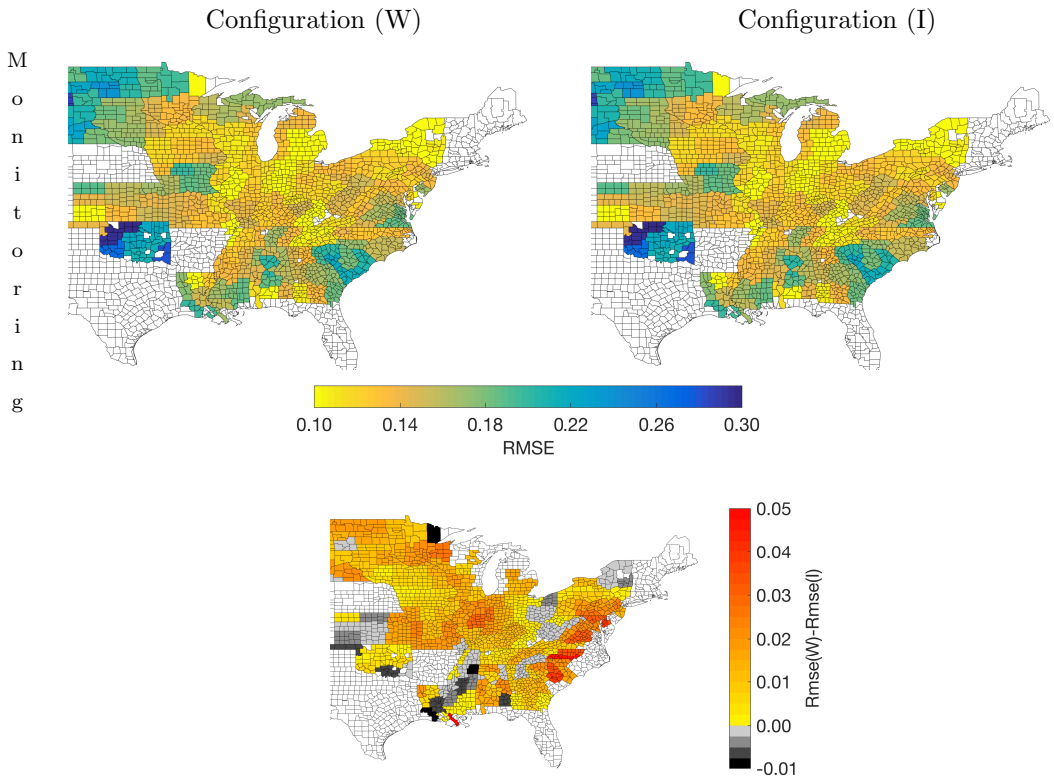


FIGURE V.17 – Haut : Cartes des RMSE (entre anomalies de rendement en maïs et prévisions météorologiques, à l'échelle du district) pour les configurations (W) et (I).
Bas : Différence entre RMSE(W) et RMSE(I).

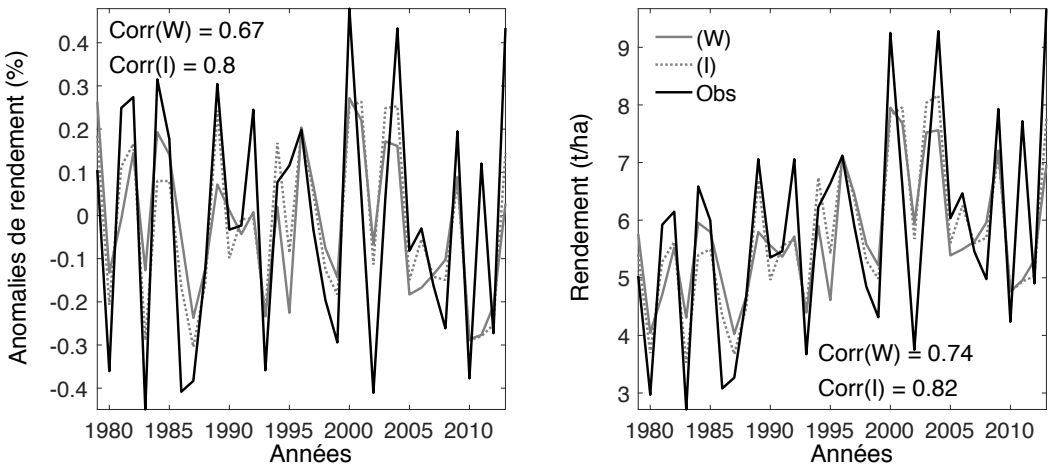


FIGURE V.18 – Mode monitoring sur l'état de Virginie (district appelé "district central") pour la prévision des anomalies de rendement (à gauche) et des rendements (à droite). Les corrélations entre les séries temporelles observées et prédites (dans les échantillons test) pour les configurations (W) et (I) sont également fournies.

3.2 Amélioration des prévisions saisonnières

Pour l'application des prévisions saisonnières de nouvelles sélections d'entrées ont été faites de mai à août et pour les deux configurations (W) et (I) : la sélection s'effectue parmi tous les prédicteurs disponibles avant ou pendant le mois de prévision afin de comparer équitablement les résultats de prévision de mai à août. Par exemple, pour la prédiction de juin, tous les prédicteurs météorologiques de janvier à juin peuvent être sélectionnés en entrée du modèle prévisionnel de juin. Les différentes entrées (obtenues après une validation croisée, comme précédemment) sont indiquées au Tableau V.11.

	Mois	⟶ Sélection successive des entrées ⟶					
(W)	Mai	Tavril					
	Juin	Pjuin	Tjuin	Tavril			
	Juillet	Tjuillet	Pjuin	Tavril	Pjuillet	Tjuin	
	Août	Tjuillet	Taoût	Pjuillet	Tjuin	Tmai	Paoût
(I)	Mai	DJavril	SPEImai				
	Juin	SPEIjuin	Tjuin	DJavril			
	Juillet	Tjuillet	SPEIjuillet	SPEIjuin	DJavril		
	Août	Tjuillet	SPEIjuillet	SPEIjuin	DJaoût	DJavril	

TABLEAU V.11 – Liste des entrées (dans leur ordre de sélection) pour les différents mois de prévision saisonnière, et pour les deux configurations (W) et (I).

À l'échelle des États-Unis

Les cartes de la Figure V.19, représentent les corrélations entre les anomalies de rendement observées et prédites, de mai à août et pour les deux configurations (W) et (I). Les cartes dans la colonne de droite illustrent les districts dont les prévisions se sont améliorées ou se sont aggravées, avec l'ajout d'indices agroclimatiques. Il y a une amélioration nette des prévisions saisonnières des anomalies du rendement, pour mai et juillet. Il existe donc également pour ce mode là, un avantage clair à considérer la configuration (I) par rapport à (W) : fin juillet, la corrélation moyenne (entre tous les districts) entre les observations et les prévisions est de $0,41$ pour (W) et de $0,50$ pour (I).

Les résultats montrent une augmentation monotone de la qualité des prévisions saisonnières lors de l'approche de la récolte. L'amélioration en juin est moins significative : en juin, la corrélation moyenne entre les observations et les prévisions est de $0,23$ pour (W) et de $0,30$ pour (I). (I) fournit des informations presque partout, en particulier dans le nord des États-Unis, où (W) semble être inefficace. Nous obtenons avec (I) une qualité de prévision légèrement plus homogène. La configuration (I) est toujours meilleure que (W), quel que soit le mois. Aucune information n'est fournie en mai par (W), alors que la connaissance n'est pas nulle pour (I). De plus amples informations peuvent déjà être fournies en juin pour le sud de la côte est. Enfin, pour les deux configurations, la plupart des informations sont fournies fin juillet. Les cartes pour le RMSE donnent des résultats similaires (Figure V.20).

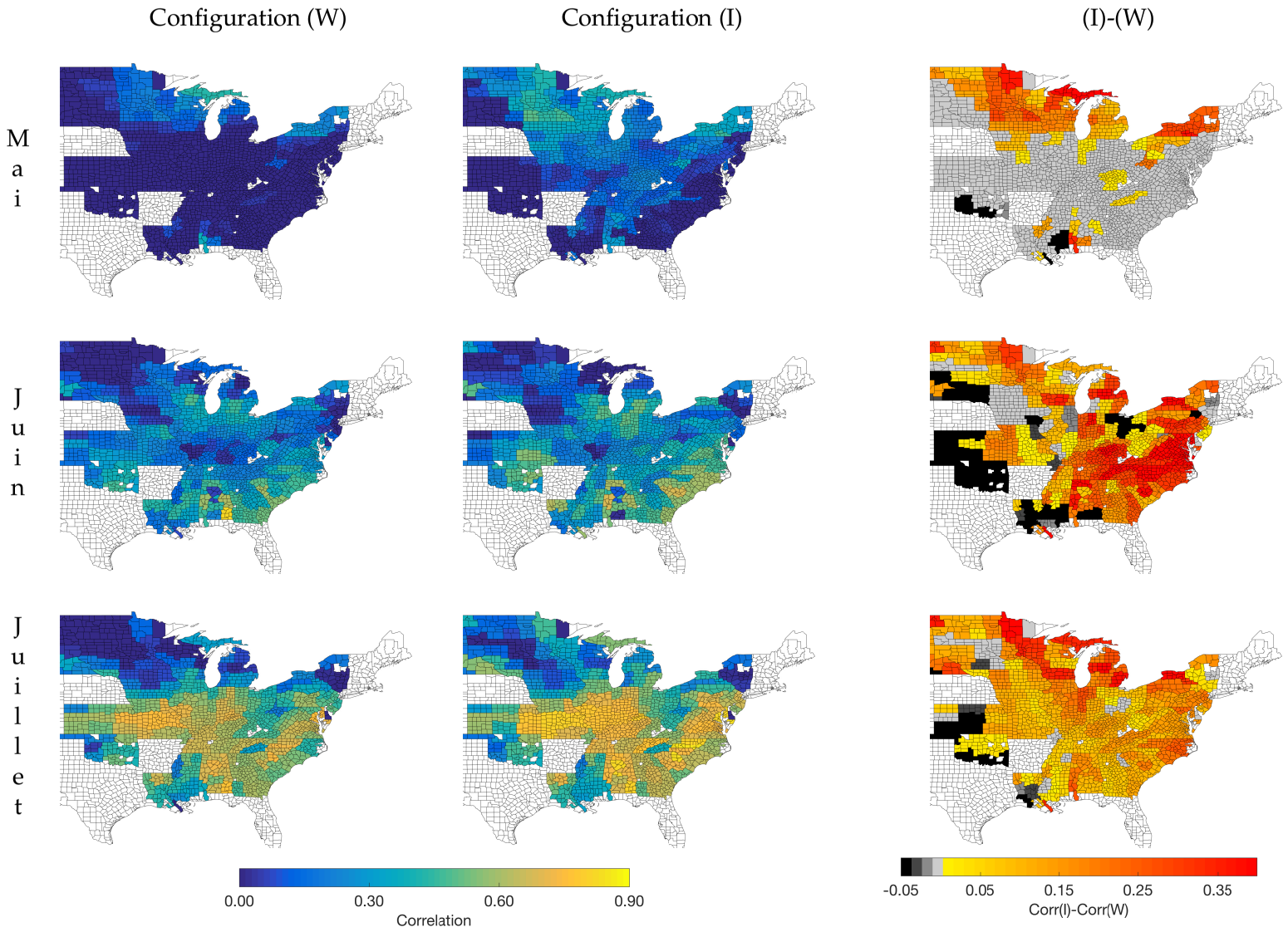


FIGURE V.19 – Les cartes des corrélations Corr entre les anomalies de rendement observées et prédites des prévisions saisonnières, de mai à août et pour les deux configurations (W) et (I). Les cartes pour août ont déjà été montrées à la Figure V.16 étant donné que pour ce mois, on obtient la même sélection des entrées que celle du mode monitoring. Les cartes dans la colonne de droite illustrent la différence de corrélation entre la configuration (I) et celle (W).

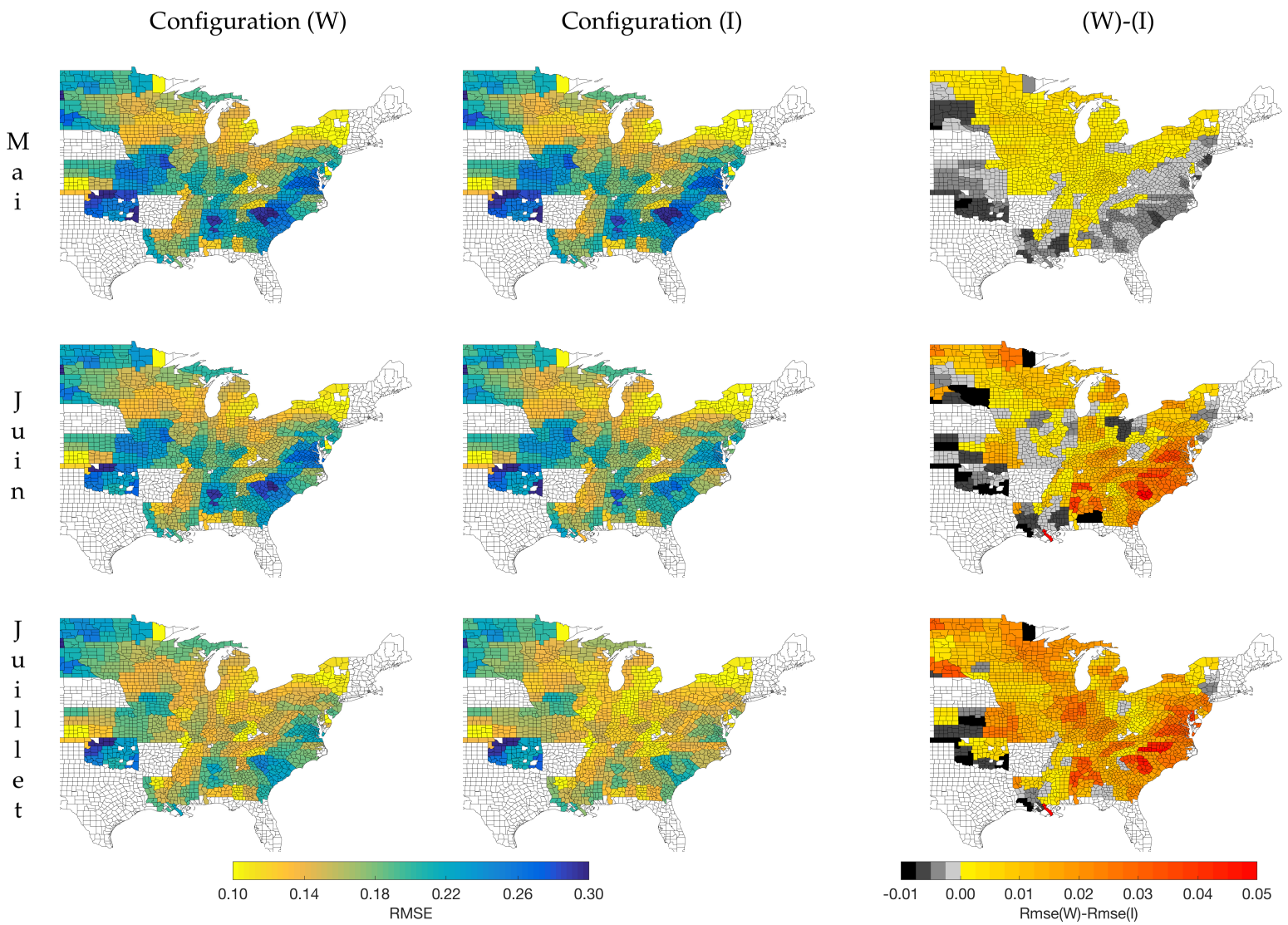


FIGURE V.20 – Cartes du RMSE entre les anomalies de rendement observées et prédites, de la prévision saisonnière de mai à août, pour les deux configurations (W) et (I). Les cartes pour août ont déjà été montrées à la Figure V.17. Les cartes de la colonne de droite illustrent la différence de RMSE entre la configuration (W) et celle (I).

À l'échelle du district

Dans cette section nous quantifions les informations fournies par les entrées météorologiques de mai à août, pour prédire le rendement de maïs en utilisant un modèle mixte, dans l'état de Virginie. Le Tableau V.12 résume la capacité du modèle à prédire les anomalies de rendement en Virginie (en termes de corrélation entre les anomalies de rendement (ou les rendements) observées et prédites) en utilisant les entrées météorologiques disponibles de mai à août comme à la section précédente.

	Anomalies de rendement				Rendement			
	Mai	Juin	Juillet	Août	Mai	Juin	Juillet	Août
(W)	0.00	0.35	0.67	0.67	0.38	0.53	0.72	0.73
(I)	0.10	0.52	0.76	0.80	0.39	0.63	0.79	0.82
USDA	-	-	-	-	na	na	0.90	0.95

TABLEAU V.12 – Gauche : corrélations entre les séries temporelles d'anomalies de rendement observées et prédites pour l'état de Virginie (district central). Droite : corrélations entre les séries temporelles de rendement observées et prédites. Des statistiques sont fournies pour les configurations (W) et (I) selon le mois de prévision saisonnière, de mai à août. Les résultats pour le modèle de l'USDA (rapport d'août et de septembre) sont également fournis pour comparer.

Comme indiqué au Tableau V.12, la configuration (W) fournit moins d'informations avant juin (Corr $\leqslant$ 0,35) alors qu'une corrélation de 0,52 peut être obtenue avec la configuration (I) fin juin. Ces chiffres soulignent l'importance de l'information fournie par SPEIjuin dans (I) : les deux autres entrées (Tjuin et Tavril pour (W), ou Tjuin et DJavril pour (I) - Tableau V.11) sont très similaires. L'amélioration de juin à juillet, ou de juillet à août, est très impressionnante sur la configuration (W) : environ 0,50 de juin à juillet, et moins de juillet à août.

En juillet, 91 % de la valeur maximale (celle observée en août) est déjà atteinte pour la configuration (W) et 95 % pour la configuration (I). Ainsi, presque toutes les informations sont déjà disponibles à la fin du mois de juillet. Encore une fois, les variations météorologiques en juillet semblent être très importantes pour la croissance du maïs et, avec les indices agroclimatiques il est possible d'obtenir des informations exploitables fin juillet. En ce qui concerne les prévisions de rendement (Tableau V.12, à droite), les différences entre (W) et (I) sont moins importantes en mai que pour les anomalies de rendement, parce que les deux configurations utilisent la même tendance pour calculer le rendement à partir des anomalies. En utilisant les entrées (I), le modèle mixte montre qu'en Virginie 62% des variations de l'anomalie du rendement s'expliquent en juillet par les variations météorologiques (Corr = 0,79).

Le modèle de l'USDA n'utilise pas de données météorologiques : il est basé sur des informations sur le rendement direct. Il est donc naturel que sa performance soit meilleure que le modèle basé sur le temps. Avec moins de 5 prédicteurs météorologiques, le modèle mixte (I) proposé ici est beaucoup plus simple que le modèle de l'USDA, et moins coûteux à mettre en place. En outre, notre modèle pourrait être étendu à d'autres parties du monde où il serait difficile d'organiser les deux enquêtes in situ nécessaires pour le modèle de l'USDA. Il est également intéressant de noter qu'il est possible d'avoir des informations pertinentes en juin ou en juillet lorsque les sondages de l'USDA ne sont disponibles que mi-août.

Ces résultats sont illustrés plus en détails à la Figure V.21 : les séries temporelles de l'anomalie de rendement prédites en Virginie pour les configurations (W) et (I) sont comparées, lorsque nous utilisons les entrées sélectionnées pour mai (en haut à gauche), pour juin (en haut à droite), pour juillet (en bas à gauche) et enfin pour août - c'est-à-dire les mêmes entrées que pour le mode monitoring. Les corrélations entre les séries temporelles observées et prédites sont également fournies. En juin, aucune information sur les anomalies de rendement n'est fournie par (W), et peu par (I). Dès juin, même si les anomalies de rendement et de production sont loin d'être bien estimées, les variations des prédictions sont correctes. En juillet, les amplitudes négatives des prédictions sont bien meilleures pour (I) que pour (W), surtout pour les années 1983 et 1989. Seules de petites améliorations sont visibles de juillet à août.

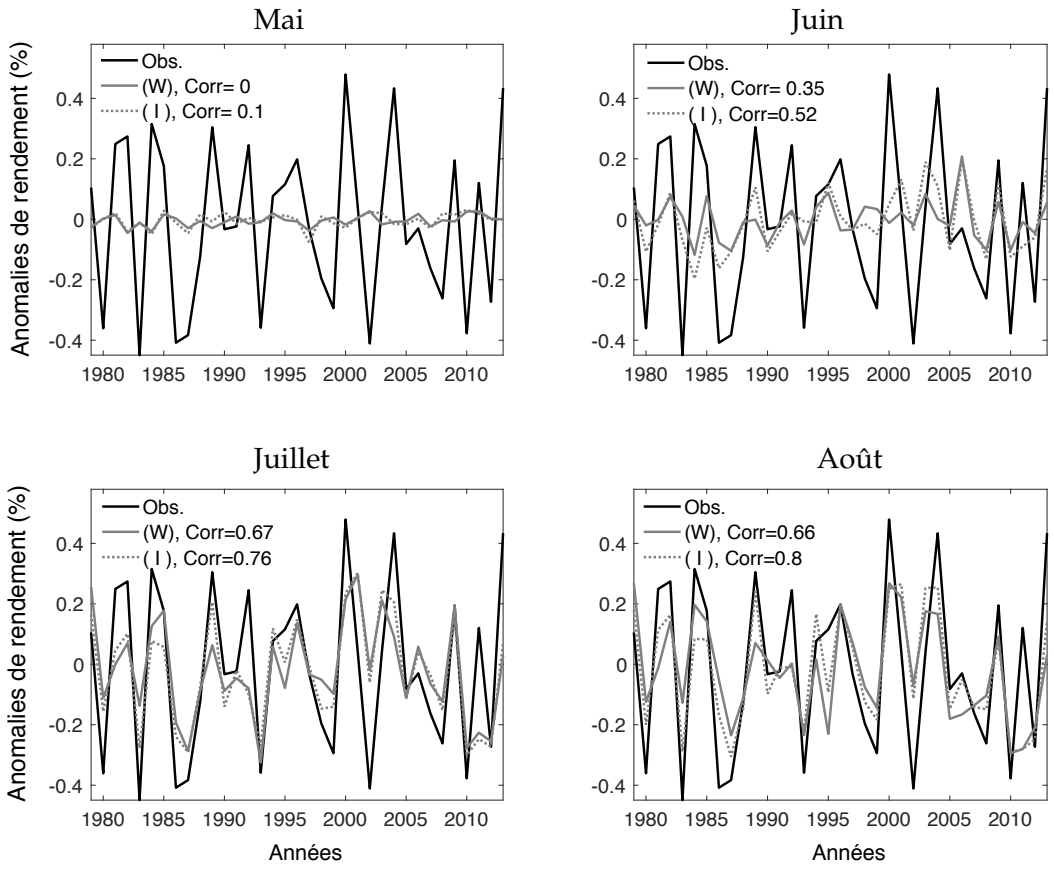


FIGURE V.21 – Prévisions saisonnières de l'anomalie du rendement en Virginie. Pour chaque mois de prévision, les deux configurations d'entrées (W) et (I) sont comparées.

4 Discussion

Stress thermique et disponibilité en eau - De nombreuses études ont analysé l'impact du stress thermique et du déficit hydrique sur le rendement en maïs et ont également souligné l'importance de ces prédicteurs. [Kaus16] a analysé les effets du stress thermique sur diverses cultures vivrières et leurs réponses. [Stan97] et [Agui07] ont étudié l'influence de la sécheresse sur plusieurs hybrides de maïs. [Ça14], [Ston97] et [Solt13] ont déterminé l'effet du stress hydrique sur le maïs. Les études de [Aslm13] ont

montré que le stress hydrique avait des effets délétères sur les semis, la croissance végétative, la photosynthèse, la croissance des racines, la pollinisation et la formation de grains dans les cultures de maïs. En outre, [Hawk13] a trouvé une réduction significative du rendement en maïs français pour chaque jour avec une température maximale supérieure à 32^oC. Des études similaires se réfèrent au maïs africain [Lobe11a].

[Dery14] a quantifié à l'échelle mondiale, les impacts du stress thermique extrême sur le rendement du maïs, et [Lobe14a] suggèrent que les changements agronomiques tendent à améliorer la tolérance à la sécheresse des plantes mais pas à diminuer la sensibilité à la sécheresse. De plus, [Bish16] a montré que la dépendance des cultures vis-à-vis des pollinisateurs peut être modifiée par le stress thermique et suggère que la pollinisation par les insectes pourrait devenir plus importante dans la production agricole. La plupart de ces études analysant l'impact du stress thermique ou du déficit hydrique sur le rendement en maïs soutiennent notre sélection de prédicteurs météorologiques. Il convient de noter, cependant, que l'impact de la chaleur peut également dépendre de la capacité de stockage en eau du sol (c'est-à-dire une information spécifique au lieu). Pour un sol avec une faible capacité de stockage de chaleur, la chaleur peut avoir un effet encore plus important sur la culture.

Interaction entre prédicteurs météorologiques - La sécheresse et le stress dû à la chaleur se produisent souvent (mais pas toujours) en combinaison, surtout en juillet/août, ce qui entraîne des réactions plus fortes au cours de ces mois. Dans la littérature, de nombreux articles ont étudié ce point : [Kebe12] a montré que l'effet combiné de la chaleur et du stress hydrique sévère avait un impact beaucoup plus important pour le maïs non irrigué que l'effet indépendant. Des études similaires ont analysé les effets indépendants et combinés de la haute température et du stress hydrique sur le rendement des plantes [Grig11, Fabi08, Pras11].

D'autres paramètres influencent le rendement des cultures, tels que le type de sol, les engrais, le type de graines, les maladies, etc. Une partie de cette variabilité est prise en compte dans la tendance à long terme modélisée par un modèle à effets mixtes. Contrairement au type de sol ou à la variété des cultures, les conditions météorologiques ont un impact à court terme sur le rendement des cultures et expliquent une partie de la variance interannuelle. De telles différences entre les cantons sont déjà prises en compte grâce aux informations de groupe dans le modèle à effets mixtes, mais ces informations supplémentaires pourraient probablement améliorer nos résultats. Cela pourrait faire l'objet d'une prochaine étude.

Sensibilité des zones de production - La seconde partie de ce chapitre est venue confirmer les résultats de météo-sensibilité issus de la première partie, en localisant les zones de production de maïs les plus sensibles aux conditions météorologiques. Les régions du sud et de l'est sont les plus sensibles : là, la météo représente plus de 28% de la variabilité du rendement du maïs. Ce sont les régions qui produisent le moins de maïs. Dans d'autres régions, moins de 20% de la variabilité du rendement est liée aux conditions météorologiques. Ces chiffres peuvent sembler faibles, mais ils restent cohérents avec les récentes recherches sur ce sujet : [Deep15] a montré qu'à l'échelle mondiale, l'information météorologique pourrait expliquer 1/3 de la variabilité du rendement.

En conclusion

Ce chapitre se concentre sur les variations annuelles de rendement liées aux conditions météorologiques. Le climat étant à la fois une ressource et une contrainte pour l'agriculture, il est important de bien comprendre son impact sur le rendement des cultures. Les informations météorologiques peuvent être exploitées pour développer des modèles d'impact qui permettent d'estimer ou de prévoir le rendement des cultures. Nous avons introduit une méthodologie statistique rigoureuse afin de classer les informations météorologiques les plus efficaces et de quantifier leur impact sur les prévisions de rendements. L'analyse de sensibilité a montré que les informations sur la température et l'eau sont nécessaires. Notre méthodologie permet de mesurer la sensibilité climatique des régions agricoles : ce chapitre montre, en utilisant une classification des prédicteurs météorologiques, que la zone 2 (nord) et la zone 3 (ouest) sont moins sensibles aux conditions météorologiques. Ceci est confirmé par la qualité de prédiction des modèles d'impact.

Il y a des variations d'une zone de production à l'autre, mais les similitudes sont suffisamment importantes pour faire un modèle général dans la dernière partie de ce chapitre, permettant des résultats plus robustes, compte tenu des données limitées disponibles pour paramétrer plusieurs modèles locaux. Les modèles statistiques permettent d'estimer une part importante de la variabilité du rendement des cultures (64% en Virginie), surtout si les entrées du modèle sont optimisées, en utilisant des indices agroclimatiques. Il est démontré que les indices agroclimatiques ne sont pas plus efficaces que les informations météorologiques plus indirectes pour la prévision de rendement. Les indices agroclimatiques sophistiqués ne semblent pas très instructifs : des prédicteurs simples fonctionnent bien. L'indice SPEI n'est pas trop sophistiqué, ni trop simple comme les précipitations ; il semble que ce soit un bon compromis, fournissant des informations plus directes pour la plante mais de manière simple et efficace. Ce type de modèle peut être utilisé pour exploiter des informations météorologiques indirectes afin de prédire le rendement en maïs comme outil de suivi, et en mode prévisionnel saisonnier (pendant la saison de croissance, et à partir de fin juillet).

5 Perspectives

Ce travail peut être étendu de plusieurs façons. Premièrement, il semble concevable d'envisager un modèle statistique différent pour chaque zone de production si des séries temporelles plus longues étaient disponibles. La base de données NOAA ESRL pourrait fournir un ensemble de données nécessaires sur une période plus large (de 1948 à aujourd'hui).
Deuxièmement, le modèle de régression effectue une prédiction satisfaisante des valeurs autour de la moyenne. Mais pour les années les plus extrêmes (hautes ou basses), il est toujours sous-prédictif. Les prévisions de ces pertes de rendement extrêmes sont très particulières : elles nécessitent un ensemble d'outils statistiques différent, et sera l'objet du chapitre suivant en utilisant une classification par réseau de neurones.
En outre, il est très possible que la faible corrélation dans certaines zones de production soit provoquée par les données extrêmes. Une méthode de régression robuste a été appliquée dans les études de prévision de rendement pour traiter les problèmes de

valeurs extrêmes [Rous05, New114].

Enfin, la prévision climatique (pour les 20 à 50 prochaines années) est une extension possible de ce travail. Améliorer notre compréhension de l'influence du climat sur la production agricole est nécessaire pour faire face aux changements futurs prévus de la température et des précipitations. Notre travail permet la sélection d'indices agro-climatiques qui doivent être obtenus à partir des simulations de modèles climatiques globaux (GCM), afin de les utiliser comme entrées de nos modèles d'impact. Cela permettra d'améliorer nos connaissances sur l'impact du climat sur l'agriculture, de développer des modèles d'impact efficaces, et ainsi de permettre une meilleure gestion des risques climatiques. Un dernier chapitre se concentrera donc sur l'utilisation du modèle ME-canton sur les cycles climatiques pour analyser l'impact du changement climatique sur l'agriculture telle que pratiquée aujourd'hui, et pour faciliter le développement de stratégies d'adaptation. Dans ce cas, la tendance à long terme devient importante. Cette tendance à long terme du rendement du maïs n'est pas le résultat du changement climatique, mais plutôt le résultat de changements dans la pratique agricole (fertilisation, outils, pratiques d'irrigation, génétique des semences, etc.). Ainsi, un rendement de tendance externe pour les prochaines décennies est nécessaire pour tout type de prévision climatique, quel que soit le modèle (statistique ou mécaniste). Pour l'agriculture, nous pourrions définir plusieurs scenarios avec une augmentation de rendement au cours des prochaines décennies, ou nous pourrions utiliser un scénario où nous supposons une pratique agricole constante. Notre objectif n'est évidemment pas de quantifier ce que le rendement du maïs sera dans 50 ans, mais plutôt de voir si le changement climatique sera bénéfique/préjudiciable à l'agriculture actuelle.

La seconde partie de ce chapitre montre également qu'il est important d'accroître les interactions et les collaborations entre les scientifiques de l'agronomie, du climat et des statistiques. Les agronomes et les agro-météorologues peuvent expliquer pourquoi, quand et comment les conditions météorologiques ont une incidence sur la croissance de la plante. À l'inverse, les outils statistiques sont également utiles aux agronomes car ils peuvent confirmer ou réfuter leurs hypothèses, sans utiliser d'hypothèses *a priori*, mais en exploitant simplement les données historiques d'une manière statistique et rigoureuse.

CHAPITRE VI

Estimation des valeurs extrêmes de rendement

Table des matières

« CHICAGO (27 août 2013) : l'année dernière, les conditions météorologiques extrêmes ont obligé le programme fédéral d'assurance-récolte à verser 17,3 milliards de dollars pour les pertes de récoltes. Les paiements versés aux agriculteurs pour couvrir les pertes causées par la sécheresse, la chaleur et le vent chaud, représentaient à eux seuls 80% de toutes les pertes agricoles. Les états les plus durement touchés étaient ceux du Midwest et des Grandes Plaines. Selon le conseil de défense des ressources naturelles, une grande partie de ces pertes auraient pu être évitées en gérant mieux la distribution des eaux.»

Natural Resources Defense Council
[NRDC13]

Ce chapitre de thèse a fait l'objet d'une publication dans le journal international *Earth and Space Sciences*.

1 Introduction

1.1 État de l'art sur les extrêmes

Avec une augmentation possible de la fréquence des conditions météorologiques anormales au cours des prochaines décennies, il est de plus en plus important pour les systèmes de prévision de rendement de prédire avec précision l'impact des anomalies météorologiques et agricoles. Les agriculteurs ont besoin d'estimations précises à des fins d'assurance-récolte, de planification, de commercialisation, de livraison, ou de stockage. Plusieurs acteurs (producteurs, planificateurs, centres de distribution, assurances, décideurs, ou commerçants de matières premières) sont sévèrement impactés par des rendements anormalement bas, et il est donc crucial d'anticiper ces événements [Schl09].

Les faibles rendements peuvent également être liés aux maladies [Oerk06] ou à l'instabilité politique [Maco00]. [Cerd17] a quantifié les pertes de rendement du café dues aux ravageurs et aux maladies et a identifié les prédicteurs les plus importants du rendements et de ses pertes. L'impact de la maladie ou l'instabilité politique sont présents dans les anomalies de rendement, mais nous nous concentrons ici sur les données météorologiques qui ne peuvent pas expliquer de telles anomalies. Nous rappelons qu'un modèle basé sur des informations météorologiques n'est capable de prédire qu'un pourcentage de la variance des anomalies de rendement. Notons cependant que les impacts météorologiques peuvent également induire des invasions de ravageurs et des maladies par des effets indirects [Prak15].

Comme cela a été vu dans les chapitres précédents, les prédicteurs climatiques utilisés pour les prévisions de rendement sont souvent choisis comme de simples variables météorologiques telles que la température et les précipitations. Ils sont ensuite appliqués sur des modèles mécanistes ou statistiques. Par exemple, [Kotl07a] a utilisé une régression non paramétrique avec une seule variable météo (température ou précipitation) pour montrer que les variables météorologiques des mois d'été (juin-août) sont les facteurs les plus importants expliquant la production américaine de maïs. Une longue liste d'indices agroclimatiques a été analysée à la place ou en complément de prédicteurs météorologiques simples, le plus souvent à l'échelle régionale : température moyenne mensuelle, rayonnement solaire, précipitations cumulatives et évapotranspiration [Pelt10, Mathss]. Dans [Goua15], chaque modèle est une combinaison linéaire de 5-7 variables climatiques, au niveau départemental. [Land00] ont construit un modèle parcimonieux dans lequel chaque paramètre reflétait un effet climatique connu sur le système culture-environnement du Royaume-Uni pour permettre une interprétation mécaniste. [Bann11] a identifié des relations entre les années de rendement bas et élevé du blé et les températures maximales et minimales de l'air avec une simple analyse de corrélation au niveau régional. Enfin, une analyse multivariée dans [Gobi10] a montré des relations statistiques significatives entre le rendement et le déficit de pression de vapeur, la température, la durée de la saison de croissance, et la sécheresse ; ces prédicteurs ont été inclus dans un modèle mécaniste régional. Malgré toutes ces publications sur la prédiction des rendements en général, peu d'entre elles se sont concentrées sur

des événements agricoles extrêmes (pertes de rendement importantes par exemple).

De nombreux articles se concentrent sur l'impact des événements météorologiques extrêmes sur les équilibres alimentaires nationaux [Brow08] ; les vagues de chaleur étant plus fréquentes ces dernières années avec des conséquences importantes sur le rendement des cultures [Boye13, Fink04]. Cependant, les phénomènes météorologiques extrêmes ne provoquent pas toujours des rendements extrêmes : ils dépendent de l'occurrence, de la sensibilité des cultures et de l'adaptation des agriculteurs (par exemple : irrigation, ajustement des dates de semis et de récolte et sélection de cultures adaptées) [Oles11, Veld10]. Ainsi, de nombreuses études ont considéré des événements agricoles [Lal12, Anam02, Tria11] mais en se concentrant sur les conditions météorologiques extrêmes, et non sur les rendements extrêmement faibles. Par exemple, [Powe16] a montré, avec des techniques économétriques, que "le nombre de jours avec des températures extrêmement élevées [...] a considérablement augmenté depuis le début des années 1900".

[Moth11] a fait un résumé des inondations, des sécheresses et des ouragans qui ont eu un impact sur la production alimentaire. [Fink04] (en Europe) et [Boye13] (aux Etats-Unis) ont analysé les phénomènes météorologiques extrêmes, mais pas les rendements extrêmes. D'autres études visent à améliorer la capacité prédictive des modèles mécanistes dans des conditions météorologiques extrêmes (principalement des températures élevées/basses et un déficit/excès en eau) [Bell14]. Enfin, [Siva05] ont commenté la prévision d'événements extrêmes, mais sans fournir un prédicteur d'entrée au modèle qui soit capable d'estimer les probabilités de rendement extrêmement faible.

Peu d'études se concentrent sur les rendements extrêmes en utilisant des indices agroclimatiques (ils ont généralement l'intention de prédire la variabilité interannuelle). [Ciai05] a analysé les sécheresses et les chaleurs extrêmes pour montrer que la réduction de la productivité en Europe de l'est et de l'ouest peut être expliquée par le déficit pluviométrique et la chaleur extrême de l'été, respectivement. [Waha13] ont analysé dans quelle mesure les changements dans la production agricole historique dans les principales régions productrices peuvent être attribués à la survenue d'événements météorologiques extrêmes. [McCo86] et [Deba17] ont déterminé les pertes de rendement à partir des polluants atmosphériques (ozone, par exemple) et [Verg15] ont évalué la fiabilité des indices de végétation pour prédire le rendement et évaluer la sévérité des maladies. Enfin, en utilisant la densité de mauvaises herbes ou leur surface foliaire relative, [Edal11] ont évalué la qualité d'ajustement de différents modèles de perte de rendement, au cours des saisons de croissance en 2008 et 2009.

Plusieurs études qui ont tenté de prédire les rendements à partir de prédicteurs météorologiques ont mis en évidence la difficulté de prédire avec précision les rendements extrêmement faibles [Math16]. Dans les études agricoles, l'une des premières difficultés est la faible quantité de données et cette difficulté est exacerbée lorsqu'on examine des données extrêmes. Un soin particulier doit alors être apporté à la méthodologie pour assurer l'obtention de modèles extrêmes robustes. Cela nécessite de résoudre plusieurs difficultés telles que : la quantité très limitée de données disponibles pour calibrer correctement un modèle statistique, la nécessité de faire des choix méthodologiques ad hoc avec des conséquences potentiellement importantes sur les résultats, l'information limitée que peuvent apporter les différents prédicteurs météorologiques, ou le besoin de définir des diagnostics fiables pour mesurer de manière réaliste la qualité des modèles.

Récemment, [Ben-16a] ont fourni une étude très intéressante et bien documentée sur les rendements de maïs et de blé en France et en Espagne, en comparant la capacité des prédicteurs (agro-) météorologiques et des simulations de modèles (WOFOST) à prédire les pertes de rendement extrêmes. [Ben-16a] quantifient en détail la sensibilité des résultats aux différents choix méthodologiques (tels que les seuils ou la définition des extrêmes). Dans leur étude, une analyse de sensibilité a été réalisée indépendamment pour chaque prédicteur ; une limite qu'ils proposent d'explorer dans leurs perspectives. En fait, [Kebe12] ont montré que l'effet combiné de la chaleur et du stress hydrique s'avère avoir un impact beaucoup plus important pour le maïs non irrigué que prit indépendamment. De même, [Schl09] ont exploré les interactions entre les résultats de températures et de précipitations et ont montré qu'une plus grande précipitation atténue partiellement les dommages causés par des températures extrêmement élevées.

Ce chapitre est donc une extension des études de [Ben-16a] et [Math16]. Il utilise un modèle statistique non linéaire qui a l'intention de détecter les pertes de rendement extrêmes de maïs dans l'est des États-Unis à partir d'information météorologique seulement. Nous comparons et mesurons de façon systématique la capacité d'une variété de prédicteurs météorologiques à détecter les rendements extrêmes de maïs. Ces prédicteurs englobent des variables météorologiques brutes, des indices d'approvisionnement en eau, mais aussi des indices agroclimatiques calculés sur un mois, une année ou une saison de croissance afin d'évaluer quel type d'information est le plus lié aux rendements extrêmes. Notre objectif est de fournir une méthodologie qui : (1) identifie les prédicteurs météorologiques qui font le mieux la distinction entre les pertes de rendements extrêmes et les rendements habituels (en considérant comme extrême, un événement dont la probabilité d'occurrence est inférieure à 5%), (2) analyse en termes agricoles, ce classement des prédicteurs agroclimatiques et leur combinaison, et (3) avec un modèle de classification simple mais précis, utilise la meilleure combinaison de prédicteurs pour estimer la probabilité d'avoir une perte de rendement extrême.

Les aspects méthodologiques sont résumés dans la Section 2 et les principaux résultats sont présentés dans la Section 3. L'influence des choix méthodologiques sur les résultats est analysée à la Section 4. Enfin, la Section 5 présente nos conclusions et perspectives pour ce travail.

1.2 Localisation de l'étude

Pour analyser les événements de rendement extrêmes, nous nous intéressons aux 12 états les plus sensibles aux conditions météorologiques (voir Figure VI.1) comme indiqué au Chapitre précédent.

La Figure VI.2 représente la série temporelle du rendement en maïs pour cinq états spatialement éloignés (Illinois, New Jersey, Caroline du Sud, Tennessee, et Alabama), de 1979 à 2013. Le rendement en maïs ne varie pas de la même façon pour ces cinq états même si certains états sont corrélés (voir la matrice de corrélation dans la même figure). Il est donc nécessaire que les modèles statistiques soient ajustés géographiquement. Les sécheresses américaines importantes sont représentées par une flèche verticale noire. Le service de sécheresse américain (US Drought Monitor) fournit une base de données nationale pour suivre la sévérité des sécheresses aux États-Unis. Pour la caractérisation à long terme de la sécheresse, l'indice de sévérité de sécheresse de Palmer

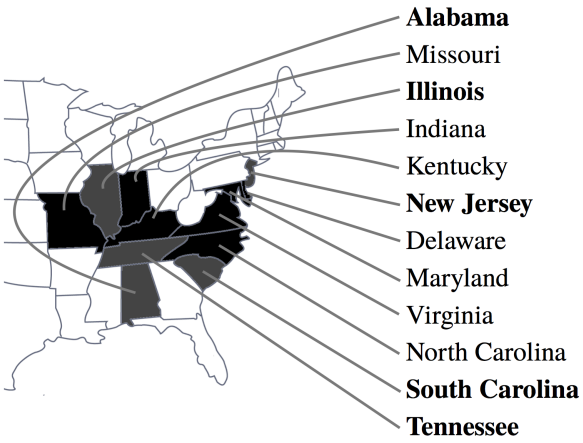


FIGURE VI.1 – Répartition spatiale des 12 états considérés. Les noms en caractères gras indiquent les états qui seront discutés par la suite.

(PDSI) [Palm65] est l'indice de sécheresse le plus couramment utilisé [1]. Les pertes de rendement extrêmes correspondent souvent à un épisode de sécheresse.

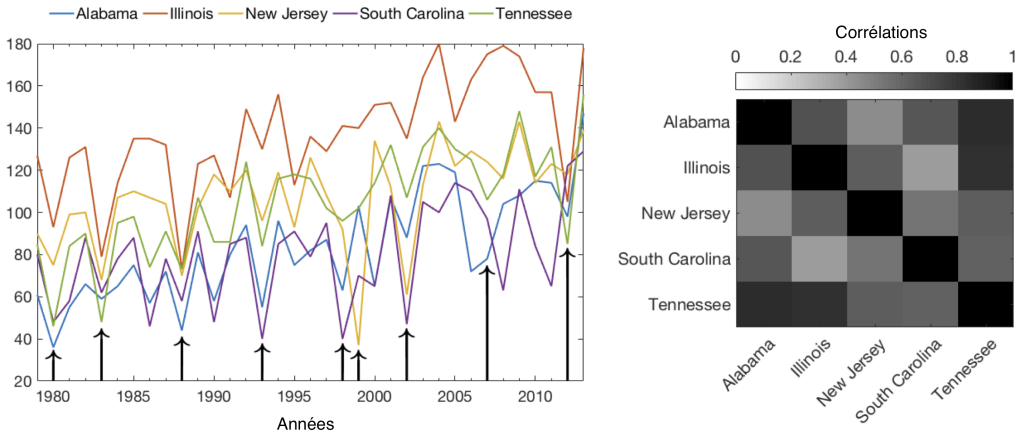


FIGURE VI.2 – À gauche : séries temporelles du rendement de maïs pour cinq états (à l'est, au sud, au nord-ouest et au centre de la région considérée). Chaque flèche verticale indique une sécheresse importante (à l'exception de 1983 qui était un événement «El Niño»). À droite : les corrélations des anomalies de rendement entre 1979 et 2013 entre les cinq états (Alabama, Illinois, New Jersey, Caroline du Sud, Tennessee).

2 Méthodologie pour la détection des événements extrêmes

Cette section détaille la méthodologie utilisée pour définir la perte de rendement extrême (Section 2.1) et les outils statistiques utilisés pour les prédire (Sections 2.2 à 2.5).

1. Un ensemble de données mondiale haute résolution de 50 ans de PDSI (1.0-degree, résolution mensuelle - http://hydrology.princeton.edu/data.pdsi.php) peut être trouvé chez [Shef06, Shef12].

2.1 Définition de la perte de rendement extrême

Dans ce chapitre, on considère également les anomalies de rendement calculées au Chapitre IV.

La Figure VI.3 illustre la distribution des anomalies de rendement. Cet histogramme prend en compte les 12 états énumérés ci-dessus, et les 35 années 1979-2013. La plupart des valeurs (82 %) sont comprises entre $-0,3$ et $0,3$. De plus, 46% des anomalies sont négatives. Les valeurs positives les plus élevées (supérieures à $0,5$) sont extrêmement rares, car il est rare d'obtenir un rendement 50% plus haut que la tendance attendue, pour une année particulière.

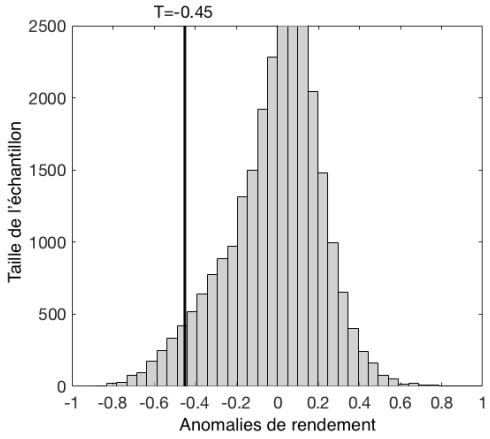


FIGURE VI.3 – Distribution des anomalies de rendement en maïs pour les 12 états considérés dans la Figure VI.1. La ligne verticale noire représente le seuil T choisi ici pour les anomalies extrêmes de rendement (gauche) et non extrême (droite) (5% des anomalies sont ici considérées comme extrêmes).

Ici, nous nous concentrons uniquement sur les pertes de rendement extrêmes (anomalies négatives extrêmes). Afin de définir ce qu'est une perte de rendement extrême pour notre application, nous devons choisir un bon seuil T en dessous duquel toutes les anomalies correspondront à une perte de rendement extrême. Un bon seuil d'anomalie de rendement pour les extrêmes ne doit pas être trop élevé pour ne prendre que des cas extrêmes, et pas trop bas pour obtenir suffisamment d'observations extrêmes. Nous considérons comme extrême un événement dont la probabilité d'existence est inférieure à 5% (ceci est un choix *ad hoc*, discuté à la Section 4 [Bomm14]). En termes d'anomalies, ce seuil correspond à la valeur $-0,45$. À la Section 4, les résultats obtenus avec ce seuil sont comparés à ceux obtenus avec d'autres seuils pour évaluer la robustesse de nos conclusions.

2.2 Travailler avec un jeu de données non équilibré

Par définition, le nombre d'extrêmes (5%) est significativement plus faible que les cas non-extrêmes. Notre approche utilise un modèle de classification pour faire la distinction entre les cas extrêmes et les cas non-extrêmes. Quand elle est calibrée sur un ensemble de données non-équilibré, la classification donne plus d'importance aux cas non-extrêmes. Afin d'obtenir de bonnes performances pour les cas extrêmes, la méthodologie de classification doit être adaptée. Différentes techniques de ré-échantillonnage

ont été proposées dans la littérature [More16]. Nous choisissons ici une méthode de sur-échantillonnage simple qui consiste à dupliquer les éléments d'une classe minoritaire un certain nombre de fois jusqu'à ce que les deux classes aient un nombre d'échantillons similaire.

2.3 Apprentissage et validation

Comme on a pu le voir dans les chapitres précédents, l'ensemble de données est divisé en un ensemble d'apprentissage et un ensemble de validation. L'ensemble d'apprentissage est utilisé pour estimer la valeur des paramètres du modèle et l'ensemble de validation est utilisé pour évaluer la force et l'utilité de la relation prédictive après l'estimation des paramètres du modèle. Cette section met l'accent sur deux facteurs à considérer lors de la construction des ensembles d'apprentissage et de validation : les corrélations spatiales et la distribution des extrêmes. Les difficultés liées aux corrélations spatiales ont été vues au Chapitre IV page 115.

Le fait que les deux classes considérées aient une taille similaire (grâce à la méthode de sur-échantillonnage) n'offre pas une plus grande diversité de cas extrêmes. Les ensembles d'apprentissage et de validation doivent inclure un nombre suffisant d'années extrêmes pour obtenir des résultats satisfaisants, c'est-à-dire une meilleure estimation possible des cas extrêmes, tout en limitant les erreurs de généralisation (dans l'ensemble de validation). Parmi les 35 années disponibles, quatre d'entre elles (1980, 1983, 1988 et 2012) contiennent un très grand nombre d'anomalies extrêmes. Ces années correspondent à une année de grande sécheresse (voir Figure VI.2).

Six années (1991, 1999, 2002, 2003, 2005 et 2011) contiennent un petit nombre d'anomalies extrêmes. Deux seulement correspondent à une année de sécheresse. Une analyse de validation croisée montre que les meilleurs résultats sont obtenus lorsque l'ensemble d'apprentissage comprend 2 ou 3 des quatre «grandes années extrêmes» et au moins trois des 6 «petites années extrêmes». Les autres années extrêmes sont ensuite utilisées dans l'ensemble de validation.

Ainsi, pour déterminer les ensembles d'apprentissage et de validation, nous utilisons une méthode de validation croisée, adaptée à la fois (1) à la non-corrélation de deux années successives et (2) à la distribution des années comprenant un grand nombre d'extrêmes dans l'ensemble d'apprentissage.

2.4 Critère de classification

Comme le montre le chapitre précédent, une régression peut être utilisée pour estimer le rendement, mais la prévision des pertes de rendement extrêmes est très difficile car il y a moins de données pour calibrer le modèle statistique. Aussi, nous préférons dans ce chapitre nous concentrer sur la capacité d'un modèle statistique à prévoir la probabilité qu'une année soit extrême. C'est une demande moins ambitieuse.

Dans ce chapitre, nous nous concentrons sur la détection des pertes de rendement de maïs extrêmes et un modèle spécifique doit être développé. L'objectif est d'estimer la probabilité qu'une année soit extrême ou non. Nous désignons par $\mathbb{P}(extrême \mid météo)$ la probabilité conditionnelle de faire face à une perte de rendement extrême compte tenu des conditions météorologiques. Dans cette section, nous présentons certains critères qui peuvent être utilisés pour évaluer la performance d'un prédicteur météorologique et aider à comparer les différents classificateurs.

Une anomalie de rendement inférieure à un seuil donné T (à définir) sera considérée comme extrême (Figure VI.3). Afin de travailler avec des classes et non des anomalies, on attribue la valeur 1 aux anomalies de rendement extrêmes, et 0 à celles non-extrêmes. Le Tableau VI.1 définit toutes les mesures couramment utilisées pour quantifier la qualité d'un classificateur lorsque l'on considère deux classes. Vrais Positifs (VP) représente le nombre de positifs (c'est-à-dire extrêmes) bien classés comme positifs. Faux Négatif (FN) se réfère au nombre de positifs incorrectement classés comme négatifs. Faux Positifs (FP) représente le nombre de négatifs (c'est-à-dire non-extrêmes) classés comme positifs (donc, mal classés). Enfin, Vrais Négatifs (VN) se réfère au nombre de négatifs classés comme négatifs.
Le Taux de Vrais Positifs (TVP), le Taux de Faux Négatifs (TFN), le Taux de Faux Positifs (TFP) et le Taux de Vrais Négatifs (TVN) se réfèrent à un pourcentage. TVP représente la proportion de positifs bien classés, et TFN représente la proportion de positifs mal classés. De même, TFP est le taux d'événements négatifs classés à tort comme positifs, et TVN mesure la proportion de négatifs qui sont correctement identifiés comme tels [Fawc06]. L'objectif d'une bonne classification est d'obtenir un TVP et un TVN élevés tout en limitant les erreurs (TFN et TFP).

2.5 Modèle de classification

Un modèle statistique est d'abord entrainé pour prédire l'état d'un rendement (perte extrême ou non) compte tenu de plusieurs variables météorologiques. Le résultat de ce modèle est un indice de sévérité des pertes de rendement semblable à une estimation de la probabilité pour le canton de faire face à une perte de rendement extrême.

L'indice de sévérité des pertes de rendement est modélisé en utilisant un réseau de neurones (NN) avec une couche cachée de dix neurones et un neurone de sortie. Toutes les fonctions d'activation sont des tangentes hyperboliques. Le réseau est entraîné en utilisant des variables météo comme entrées et deux sorties binaires (extrêmes ou non) : le réseau prédit alors une valeur réelle entre 0 et 1 qui est semblable à l'estimation de la probabilité d'être extrême ou non [Bish96]. Ce réseau de neurones est entrainé en utilisant l'algorithme de rétropropagation de Levenberg-Marquardt sur un échantillon équilibré de 50 000 observations, au niveau du canton (50% des cas proposant une perte de rendement extrême). Le réseau renvoie une valeur réelle entre 0 et 1, mais nous avons en fait besoin d'une valeur binaire pour décider s'il s'agit d'une année extrême (1) ou non-extrême (0). Il est donc nécessaire de définir un seuil de décision C sur la sortie du réseau qui discriminera entre les états 0 et 1. Par exemple, si $C = 0.5$, toutes les sorties du réseau inférieures à 0.5 seront classées comme "non-extrême" et tous les résultats supérieurs à 0.5 seront classés comme "extrêmes". Les valeurs des mesures de qualité décrites dans le Tableau VI.1 dépendront de ce seuil choisi. La Figure VI.4 synthétise les différentes étapes et jeux de données utilisés dans cette étude.

2.6 Courbes ROC (Receiver Operating Characteristic)

La courbe ROC est un outil classique d'évaluation des classificateurs [Fawc04, Brow06]. Dans une courbe ROC, le TVP et le TFP sont tracés, en fonction du choix du seuil de décision C. Chaque point de la courbe ROC est associé à un seuil de décision particulier discriminant les deux groupes. Une discrimination parfaite (pas de chevauchement dans les distributions des deux groupes) fourni une courbe ROC qui passe par le coin supérieur gauche [Zwei93]. La précision est mesurée par l'aire sous la courbe ROC. Une aire égale à 1 représente une classification parfaite ; une aire égale à 0.5 représente

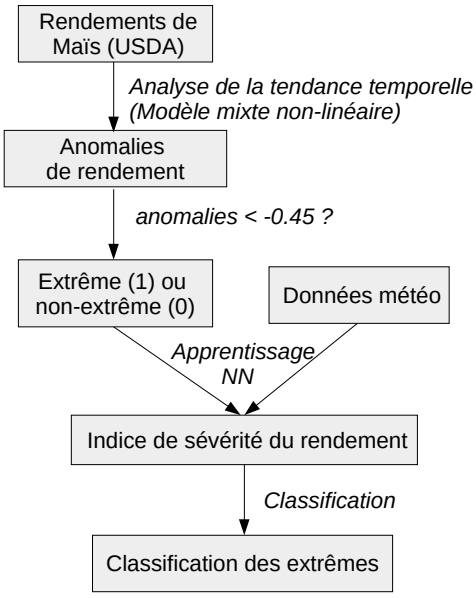


FIGURE VI.4 – Schéma des jeux de données et des méthodes utilisés dans l'étude.

une classification inutile (pas meilleure qu'une décision arbitraire). Cette aire mesure la qualité de la discrimination, c'est-à-dire la possibilité de classer correctement les cas extrêmes par rapport aux cas non-extrêmes.

La Figure VI.5 (droite) représente pour notre application, et de manière plus explicite, les taux TVP et TFP lorsque la valeur du seuil de décision varie (axe des abscisses). Lorsqu'un seuil de décision élevé est sélectionné, le TFP est bas mais le TVP est bas aussi. Lorsqu'un seuil de décision bas est sélectionné, le TVP est élevé mais le TFP est également élevé. Un compromis est donc nécessaire. Habituellement, le seuil de décision C choisi est celui qui minimise la somme des Faux Négatifs et des Faux Positifs (FN + FP) c'est-à-dire le nombre d'événements mal classés, mais d'autres mesures de qualité peuvent également être utilisées pour le choisir. En effet, en plus de la minimisation de FN + FP, nous voulons également éviter un TFP trop élevé. Ainsi, nous pouvons définir une valeur maximale de TFP qui ne doit pas être dépassée. Par exemple, nous pouvons choisir d'avoir TFP<15%. Ainsi, deux conditions doivent être vérifiées pour choisir le meilleur seuil de décision C (voir Figure VI.5, à droite) :

$$\begin{array}{rl} \text{Minimisation de} & \text{FN+FP} \\ \text{Sous la contrainte} & \text{TFP<15\%} \end{array}$$

Dans la Figure VI.5, à droite, le seuil de décision minimisant FN + FP est mis en évidence par une ligne verticale noire, tandis que le seuil réalisant aussi la contrainte TFP<15 % est indiqué par la ligne verticale verte. Pour ce classificateur, deux résultats différents sont obtenus en utilisant ces deux valeurs seuil : le TVP est certainement meilleur (84% contre 75%) quand on minimise seulement FN + FP (intersection entre la courbe bleue et la droite verticale noire) mais le TFP est plus grand (21% contre 15%), ce qui n'est pas une bonne chose (intersection entre la courbe orange et la droite verticale noire). Cela montre qu'au lieu d'utiliser des mesures de qualité simples, il est parfois préférable de les combiner avec des contraintes supplémentaires. L'impact de ces contraintes sur les résultats est testé à la Section 3.3.

		Prédits 1	Prédits 0
Observés	1	VP	FN
	0	FP	VN

$$\text{TVP} = \frac{\sum \text{Vrais Positifs}}{\sum \text{Positifs}} \qquad \text{TFN} = \frac{\sum \text{Faux Négatifs}}{\sum \text{Positifs}}$$

$$\text{TFP} = \frac{\sum \text{Faux Positifs}}{\sum \text{Négatifs}} \qquad \text{TVN} = \frac{\sum \text{Vrais Négatifs}}{\sum \text{Négatifs}}$$

TABLEAU VI.1 – Gauche : matrice de confusion d'un problème de classification avec deux classes, 0 (négatif) et 1 (positif). Chaque colonne représente l'état dans une classe prédite alors que chaque ligne représente l'état des classes réelles. Chaque mesure définie ici peut être utilisée pour quantifier la performance d'un classificateur. Les acronymes sont les suivants : VP (Vrais Positifs), VN (Vrais Négatifs), FP (Faux Positifs) et FN (Faux Négatifs).
Droite : mesures de qualité TVP (Taux de Vrais Positifs), TVN (Taux de Vrais Négatifs), TFP (Taux de Faux Positifs) et TFN (Taux de Faux Négatifs). Pour plus de détails, voir Section 2.4.

2.7 Probabilités conditionnelles

Le lien entre la météo et le rendement extrême peut être analysé de deux manières symétriques. Premièrement, compte tenu des informations extrêmes/non-extrêmes, que peut-on dire des conditions météorologiques ? Ce premier point de vue est lié à la probabilité conditionnelle $\mathbb{P}(météo \mid extrême)$ d'observer certains événements climatiques pour une année de rendement extrêmement faible. Deuxièmement, étant donné l'information météorologique, que peut-on dire des probabilités extrêmes/non-extrêmes ? Ce second point de vue est lié à la probabilité conditionnelle $\mathbb{P}(extrême \mid météo)$ d'observer des rendements extrêmement bas en fin d'année, compte tenu d'un état météorologique pour cette année. Les probabilités conditionnelles $\mathbb{P}(météo \mid extrême)$ et $\mathbb{P}(extrême \mid météo)$ sont différentes et ne nous renseignent pas sur la même chose. La comparaison de ces deux probabilités conditionnelles est donnée à la Section 3.1.

3 Résultats

La Section 3.1 identifie comment les prédicteurs météo sont distribués lors des années extrêmes, et comment les années extrêmes sont distribuées selon les prédicteurs météo. La Section 3.2 présente un classement qui compare les prédicteurs météo en fonction de leur performance à discriminer les extrêmes des non-extrêmes. Enfin, la Section 3.3 montre les résultats de la classification par un réseau de neurones.

3.1 Probabilité jointe des prédicteurs météo et des anomalies extrêmes

La Figure VI.6 représente la distribution de quatre prédicteurs météorologiques importants : Tjuillet (température moyenne en juillet), Taoût (similaire en août), SPEIjuin (indice de précipitation-évapotranspiration en juin) et SPEIjuillet (similaire en juillet). Ces prédicteurs ont été sélectionnés au chapitre précédent comme étant ceux qui estiment le mieux les rendements de maïs. Par ailleurs, le choix de T et de SPEI pour les mois de juin, juillet et août est soutenu par l'analyse de [Kotl07a], comme mentionné dans l'introduction de ce chapitre.

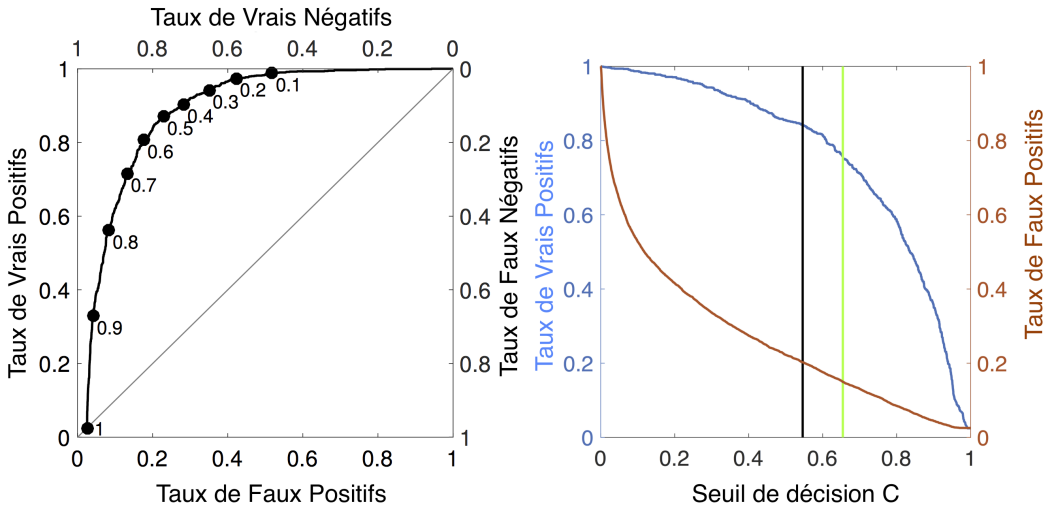


FIGURE VI.5 – Choix du seuil de décision C.
Gauche : une courbe ROC régulière dans notre application avec les quatre mesures de qualité différentes. Les points noirs et les nombres représentent différentes valeurs du seuil de décision.
Droite : les variations du TVP (axe à gauche) ou du TFP (axe à droite) en fonction de la valeur du seuil de décision (axe des abscisses). Le seuil de décision qui minimise la somme des Faux Négatifs et des Faux Positifs (FN + FP) (c'est-à-dire le nombre d'événements mal classés), est indiqué par la droite verticale noire, tandis que celui qui réalise également la contrainte TFP<15 % est indiqué par la droite verticale verte.

Dans ces graphiques, les données ont été séparées entre celles qui sont liées à une perte de rendement extrême et celles qui ne le sont pas. Les variables météorologiques ayant la plus petite intersection des deux distributions seront les meilleurs prédicteurs pour la détection des extrêmes. Dans la Figure VI.6, les droites noires verticales délimitent la valeur optimale (appelée ρ par la suite) du prédicteur météorologique discriminant les événements extrêmes et non-extrêmes. Pour Tjuillet, la valeur optimale est $\rho = 0.05$; pour SPEIjuillet -0.35 ; pour SPEIjuin -0.15 ; et pour Taoût -0.05. Ce seuil est successivement calculé en maximisant le TVP quand on classe comme "extrêmes" les données avec Tjuillet> ρ (ou Taoût> ρ, ou SPEIjuin <ρ, ou SPEIjuillet <ρ). Ces résultats dépendent du seuil pris pour définir les anomalies extrêmes (voir Section 2.1).

La Figure VI.7 représente les diagrammes de dispersion de quatre prédicteurs météorologiques (SPEIjuin, SPEIjuillet, Tjuillet, Taoût) qui ont été sélectionnés grâce à l'analyse des histogrammes de la Figure VI.6. Les points rouges sont pour les cas extrêmes et les points verts pour les cas non-extrêmes. La valeur inscrite sur ces quatre sous-figures correspond aux seuils identifiés à la Figure VI.6.

Comment les prédicteurs météo sont-ils distribués lors des années extrêmes ? Les diagrammes de dispersion montrent que les données extrêmes ne sont pas réparties aléatoirement. Pour les années extrêmes, nous notons que Tjuillet $\gg 0$, SPEIjuillet $\ll 0$, et T-SPEI $\gg 0$ en mois d'été, c'est-à-dire que l'écart entre température et SPEI est très prononcé) (Figure VI.7). Cependant, l'inverse n'est pas vrai : dans beaucoup de cas, ces trois éléments sont vérifiés, bien qu'il ne se réfère pas à un rendement extrême. La probabilité $\mathbb{P}(w \mid e)$ est la probabilité d'observer une condition météorologique (par exemple une variable météorologique w inférieure à un certain seuil) dans le cas d'une année extrême. Par exemple, sur QI, nous pouvons voir que lorsque nous observons que 64 % des événements extrêmes enregistrent Tjuillet > 0.05 et SPEIjuillet $<$ - 0.35.

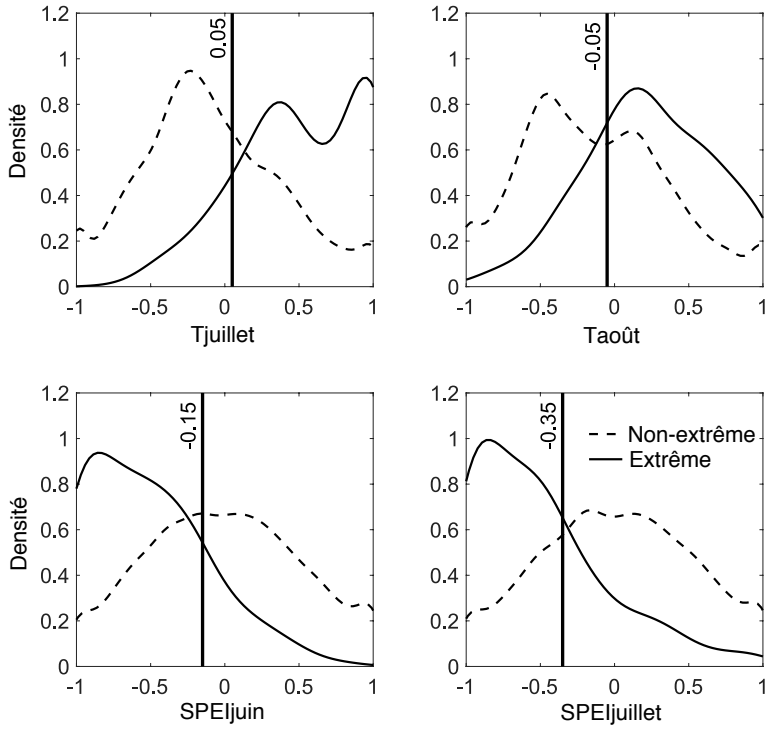


FIGURE VI.6 – Histogramme de quatre anomalies importantes de prédicteurs météorologiques (Tjuillet, Taoût, SPEIjuin et SPEIjuillet), après la séparation des données en événements extrêmes (courbe continue) et non-extrêmes (courbe en pointillée). La droite noire verticale délimite la valeur optimale (ρ) qui distingue les cas extrêmes des cas non-extrêmes pour les quatre variables météorologiques (classification uni-variée).

En d'autres termes, $\mathbb{P}(\text{Tjuillet} > 0.05 \quad \text{et} \quad \text{SPEIjuillet} < -0.35 \mid \text{extrême}) = 0.64$. Cela signifie que cet état météorologique particulier est particulièrement lié aux pertes extrêmes de rendement en maïs.

Comment les années extrêmes sont-elles réparties selon les prédicteurs météo ? La probabilité d'observer Tjuillet> 0.05 et SPEIjuillet < -0.35 lors d'une année extrême est égale à 64 %. Inversement, la probabilité que la perte de rendement soit extrême si Tjuillet> 0.05 et SPEIjuillet < -0.35 peut être très différente. La probabilité $\mathbb{P}(e \mid w)$, fait référence à la proportion de données extrêmes (en rouge) lorsqu'on sait qu'elles se trouvent dans un sous-carré spécifique. Par exemple, nous pouvons voir que 20% des données avec Tjuillet> 0.05 et SPEIjuillet < -0.35 représentent des pertes de rendement extrêmes (sous-figure QI, en bas à droite).
On dispose d'une relation très simple entre $\mathbb{P}(w \mid e)$ et $\mathbb{P}(e \mid w)$: $\mathbb{P}(w \mid e) = \frac{\mathbb{P}(e|w)\mathbb{P}(w)}{\mathbb{P}(e)}$.

3.2 Classement de variables

Dans ce paragraphe, nous identifions les prédicteurs météorologiques qui expliquent le mieux les pertes de rendement extrêmes afin de distinguer les cas extrêmes des cas non-extrêmes. Selon la Figure VI.6, les prédicteurs météorologiques qui révèlent le plus d'informations sur les pertes de rendement extrêmes sont Tjuillet, SPEIjuillet et SPEIjuin. Toutes les variables météorologiques ou agro-météorologiques n'ont pas la même capacité à prédire les fluctuations du rendement de maïs. Certains d'entre eux donnent des informations sur le rendement final car ils représentent un phénomène biologique et interviennent à un moment critique dans le développement de la plante.

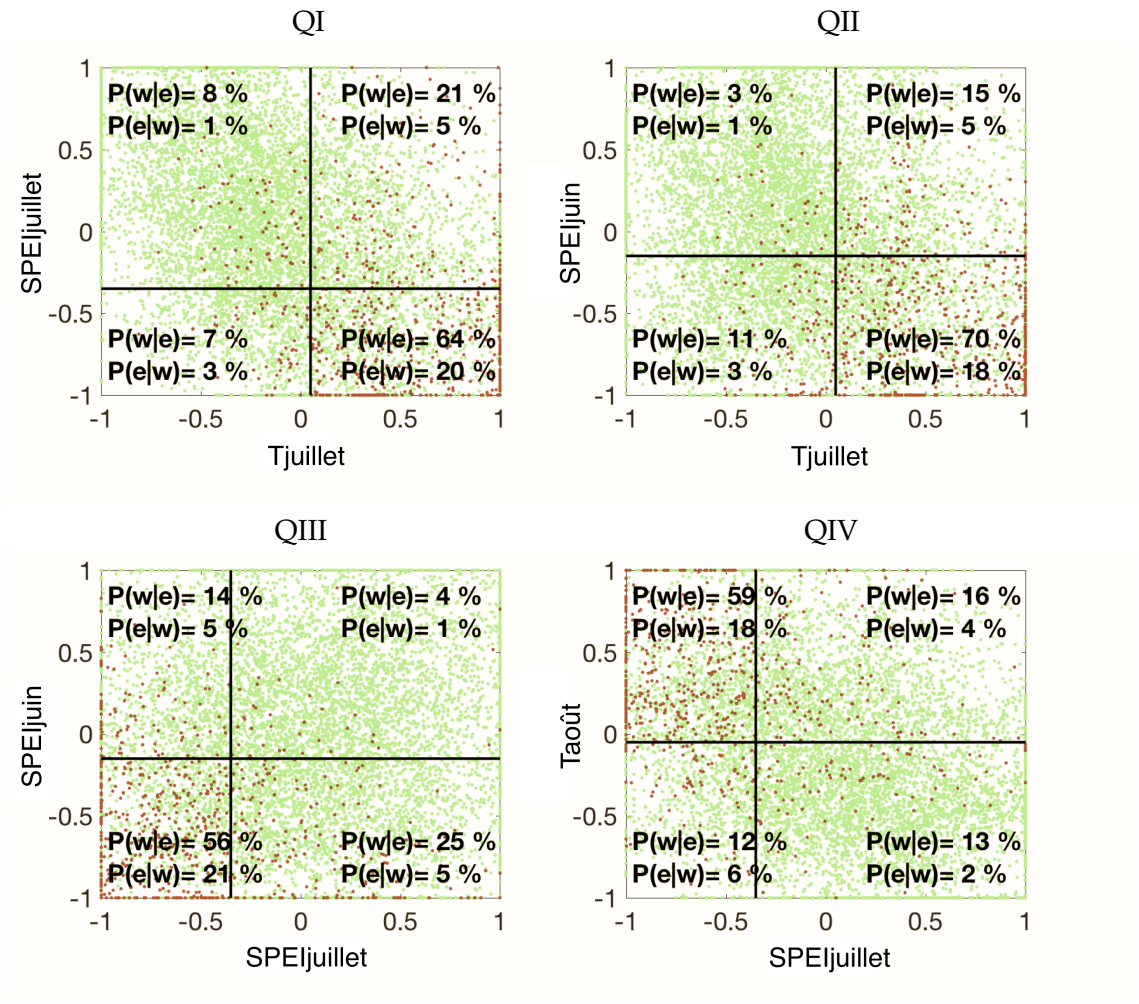


FIGURE VI.7 – Scatter-plot de quatre prédicteurs météo : SPEIjuin, SPEIjuillet, Tjuillet, Taoût. Les points rouges sont pour les cas extrêmes et verts pour les cas non-extrêmes. Les délimitations horizontales et verticales ont été définies à la Figure VI.6. Comme indiqué à la Section 2.1, 5% des anomalies de rendement sont des anomalies extrêmes.

Ainsi, comme le montrent [Kotl07a] et le chapitre précédent, l'information varie fortement d'un prédicteur météorologique à un autre : les données montrent que seul un petit nombre de prédicteurs fourni des informations solides sur le rendement du maïs, et beaucoup de prédicteurs fournissent très peu d'information.

Comme mentionné dans la Section 2.4, nous désignons par 1 les anomalies extrêmes de perte de rendement et par 0 les autres. Notre procédure pour sélectionner les meilleurs prédicteurs utilise un réseau de neurones simple. Un processus d'induction vers l'avant est adopté : 50 ensembles d'apprentissage différents sont choisis ; pour chacun d'eux, le prédicteur qui obtient le TVP le plus élevé en généralisation est sélectionné. Nous sélectionnons comme premier prédicteur l'indicateur météorologique le plus fréquemment sélectionné parmi les 50 validations différentes (il s'agit de **SPEIjuillet**). Considérant SPEIjuillet comme première entrée du classificateur, la seconde entrée est choisie de façon similaire parmi tous les prédicteurs, comme étant celle qui, combinée avec SPEIjuillet, définit un classificateur à 2 entrées qui donne le plus grand TVP en généralisation. Ce processus est répété jusqu'à ce qu'un nombre souhaité d'entrées soit sélectionné. Les deuxième et troisième entrées sélectionnées sont **Tjuillet** et

SPEIjuin.

Nous avons testé 59 variables météorologiques ou indices agroclimatiques (Tableau III.1). Le processus de sélection nous permet de les classer en fonction de leur capacité à discriminer les rendements extrêmes des non-extrêmes. Une fois ce classement effectué, puisqu'il n'est pas souhaitable de conserver trop d'entrées, une validation croisée est utilisée pour savoir combien nous devons en garder : les quatre premières entrées (Tjuillet, SPEIjuillet, SPEIjuin et Taoût) sont suffisantes car la classification ne s'améliore pas en ajoutant une cinquième variable.

En gardant à l'esprit la contrainte TFP<15 %, le Taux de Vrais Positifs passe de 50% (avec seulement l'entrée SPEIjuillet) à 63% avec les deux premières entrées sélectionnées (SPEIjuillet et Tjuillet). Après l'ajout de SPEIjuin, le TVP atteint 70 $\%$. Le TVP se stabilise à 71% avec la quatrième entrée (Taoût).

3.3 Modèle de classification des pertes de rendement extrêmes

Les réseaux de neurones sont des modèles habituels pour la classification. Dans cette étude, nous utilisons la fonction *patternnet* du logiciel Matlab pour construire un classificateur avec dix neurones sur sa couche cachée.

Comme expliqué à la Section 2.4, la mesure de qualité utilisée pour définir le seuil de décision permet un compromis entre la minimisation habituelle de FN + PF et une contrainte subjective sur le TFP (par exemple TFP<15%). Le Tableau VI.2 donne le TVP obtenu avec une contrainte sur le TFP variant de 5% à 40% : dans chaque cas, afin de quantifier l'impact des ensembles d'apprentissage et de validation sur les résultats, 200 réseaux de neurones sont entrainés avec un ensemble d'apprentissage et de validation différents.

Contrainte sur TFP	$\overline{\text{TVP}}$	sd_{TVP}	$\overline{Cutoff_{opt}}$	$sd_{Cutoff_{opt}}$
5%	44%	0.04	0.81	0.09
10%	60%	0.09	0.72	0.07
15%	71%	0.08	0.60	0.09
20%	78%	0.07	0.55	0.11
25%	83%	0.08	0.51	0.12
30%	85%	0.07	0.52	0.13
35%	86%	0.08	0.52	0.13
40%	87%	0.08	0.53	0.14

TABLEAU VI.2 – Le TVP moyen en validation lorsque la contrainte sur le TFP varie de 5 % à 40 %. Dans chaque cas, 200 réseaux de neurones sont entrainés avec un ensemble d'apprentissage et de validation différents. La valeur du seuil de décision moyen et l'écart-type sont également donnés. Le seuil de décision optimal minimise FN + FP sous la contrainte sur TFP.

Considérant la contrainte TFP<15 %, notre classificateur peut détecter 71% des pertes de rendement extrêmes. En d'autres termes, TFN=100-71 % =29 % des années extrêmes sont mal classées. Lorsque nous limitons la contrainte sur le TFP, la proportion d'années extrêmes bien détectées (TVP) augmente mais de plus en plus d'années sont considérées comme extrêmes alors qu'elles ne le sont pas (le TFP augmente aussi). L'écart-type des résultats est assez stable lorsque la contrainte sur le TFP varie. Il permet de vérifier la robustesse des résultats et la qualité de généralisation du classificateur. La Figure VI.8 quantifie la sensibilité du TVP et du TFP aux ensembles d'apprentissage et de validation. Le classificateur est exécuté avec 200 ensembles d'apprentissage et de

validation différents. La valeur moyenne du TVP et du TFP est illustrée et un bandeau de confiance est donné avec les 2^{eme} et 8^{eme} déciles. Ces résultats sont satisfaisants : la sensibilité du TVP à l'ensemble d'apprentissage n'est pas négligeable mais considérée comme naturelle pour cette étude. Le TVP varie plus avec l'ensemble d'apprentissage que le TFP : l'écart-type moyen pour le TVP est de 0.08 et 0.05 pour le TFP. L'écart-type le plus important pour le TVP est de 0.15 (obtenu pour un seuil de décision égal à 0.86), et l'écart-type le plus important pour le TFP est de 0.13 (obtenu pour un seuil de 0.06) (Figure VI.8).

Le Tableau VI.3 montre que les valeurs des différentes mesures de qualité (TVP, TFP...) ne changent que légèrement en se concentrant sur la base de données originale (et non sur celle sur-échantillonnée).

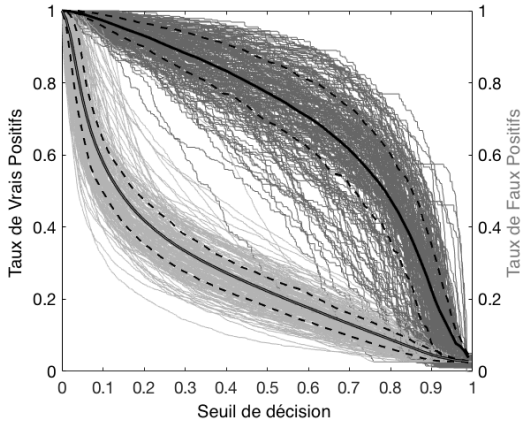


FIGURE VI.8 – Variabilité des courbes du TVP et du TFP pour 200 ensembles d'apprentissage et de validation différents. Les courbes noires en gras représentent la valeur moyenne du TFP (ou du TVP) par rapport à la valeur du seuil de décision. Les courbes noires en pointillés représentent les 2^{eme} et 8^{eme} déciles des deux ensembles de courbes grisées.

	Nombre de 0	FP	TFP	Nombre de 1	VP	TVP
Avec sur-échantillonnage	24306	3640	15	24111	18848	78
Sans sur-échantillonnage	24306	3640	15	1301	992	76

TABLEAU VI.3 – Mesure de qualité avec ou sans sur-échantillonnage. Le seuil de décision est égal à 0.60.

La Figure VI.9 illustre les résultats pour trois années : l'une se référant à une sécheresse non-extrême (2013), l'autre à une sécheresse modérée (2000) et un troisième à une sécheresse extrême (1983). Les observations d'anomalies de rendement sont comparées à la sortie du classificateur (dans $[0, 1]$) et à la sortie du classificateur après application du seuil de décision C (dans $\{0, 1\}$). Le seuil de décision optimal pour ce classificateur est $C = 0.60$, et est obtenu lorsque toutes les données sont rassemblées dans l'ensemble d'apprentissage. La sortie du classificateur peut être considérée comme un indice de sévérité des pertes de rendement, montrant quelles régions sont les plus touchées par les conditions météorologiques et les plus sujettes aux pertes de rendement. Plus la sortie du classificateur est proche de 1, plus la probabilité d'une perte de rendement extrême

est élevée. Inversement, plus la sortie du classificateur est proche de 0, plus la probabilité d'une perte de rendement extrême est faible. Avant d'appliquer le seuil de décision, la carte des sorties du classificateur contient beaucoup d'informations.

La cohérence spatiale de la sortie du classificateur est bonne. La corrélation spatiale entre les anomalies de rendement et la sortie du classificateur pour les deux années extrêmes est assez élevée (-0.51 pour 1983 et -0.59 pour 2000). Les cantons dont la sortie du réseau de neurones est très proche du seuil optimal sont ceux dont la classification "0/1" est la plus ambiguë. La qualité du classificateur dépend de la variabilité du seuil de décision optimal C entre les régions spatiales et entre les années. Clairement, pour les deux années de sécheresse (1983 et 2000), le seuil estimé globalement est trop petit et les résultats seraient meilleurs (TFP serait plus faible, c'est-à-dire que moins de cantons non-extrêmes seraient prédits comme extrêmes) si le seuil de décision était plus élevé (environ 0.70). Certaines années ont un TFP supérieur à 15%, même si globalement (pour toute la base de données), il y a une contrainte avec TFP $< 15\%$ qui est bien respectée. Il convient de noter que, pour cette application, il est plus important d'attribuer l'étiquette "extrême" à un rendement (même si c'est faux) que de manquer un cas extrême réel.

En résumé

Le réseau de neurones utilisé comme classificateur prédit bien les années avec un nombre faible ou élevé d'extrêmes. La localisation spatiale de la prédiction est bonne mais pourrait être améliorée : les cantons proches des cantons extrêmes sont souvent prédits comme extrêmes alors qu'ils ne le sont pas. C'est normal car ils ont le même climat, mais ce même climat n'a pas les mêmes effets. Il y a probablement un manque d'informations spécifiques au site. Les mêmes conditions météorologiques ne mèneront pas toujours à des pertes de rendement extrêmes dans deux états fédéraux voisins (politiques agricoles différentes, ressources financières différentes...) : le seuil pourrait être optimisé pour chaque état fédéral. Le comportement des séries temporelles des anomalies de rendement et des sorties du réseau de neurones est très similaire : leurs corrélations temporelles sont représentées à la Figure VI.10. Notons que dans les cinq premières figures, l'opposé de la sortie du réseau est représentée, et non le réseau lui-même. Les variations opposées sont donc bien mieux représentées : ces séries temporelles sont clairement anti-corrélées. La carte des États-Unis à la Figure VI.10 rassemble la corrélation temporelle entre les anomalies de rendement et les sorties du classificateur pour tous les cantons. Dans la plupart des régions considérées (60%), la corrélation est inférieure à -0.6, illustrant la qualité de la sortie du classificateur comme un prédicteur de la variabilité du rendement, même si elle a été conçue pour la classification (extrême/non-extrême), et non pour la prédiction des valeurs de rendement.

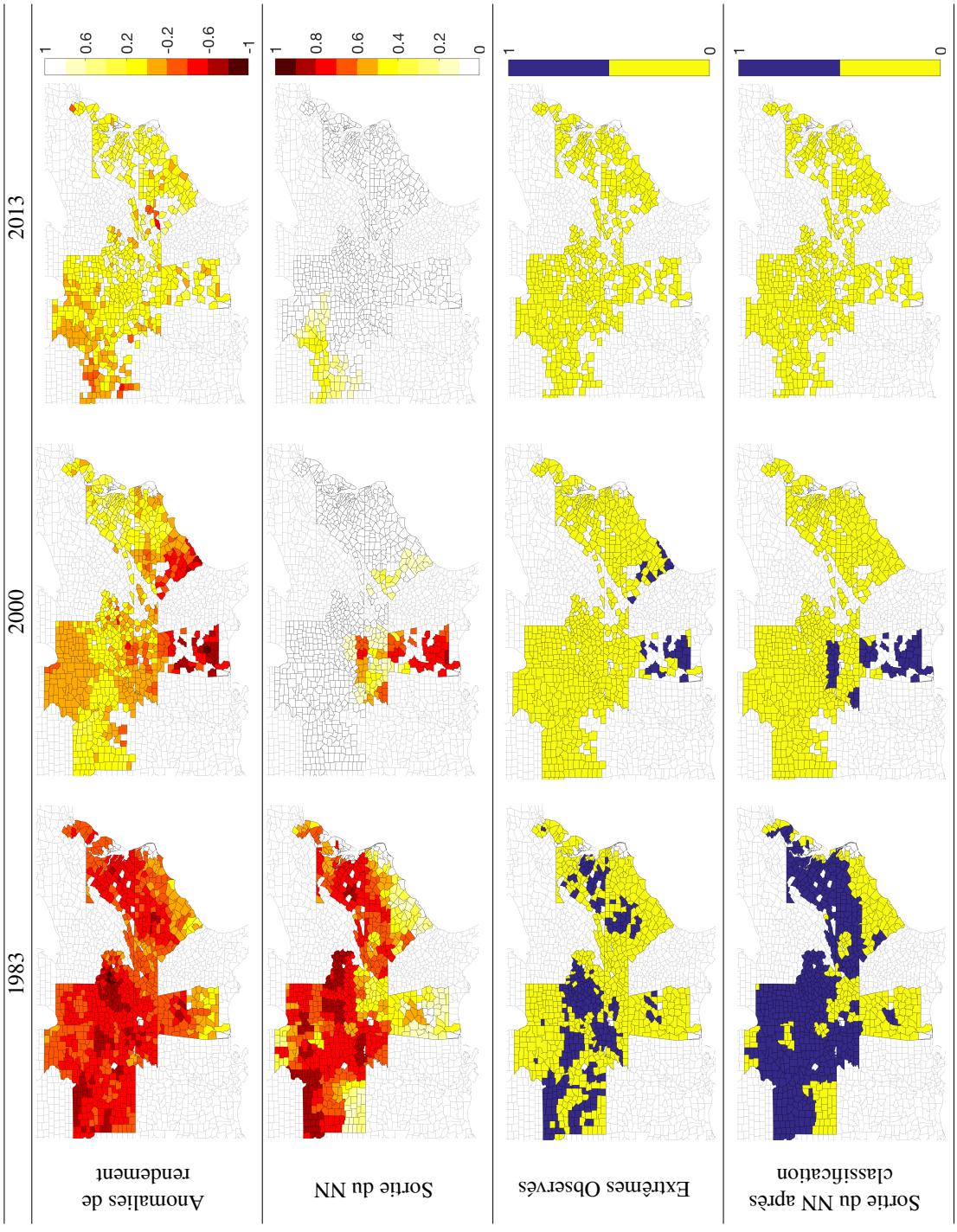


FIGURE VI.9 – De haut en bas : cartes des anomalies du rendement en maïs de l'est des États-Unis pour les années 1983, 2000 et 2013 ; Sortie du réseau de neurones (NN) pour les mêmes années ; observation pour les mêmes années ; classification pour ce réseau (NN).

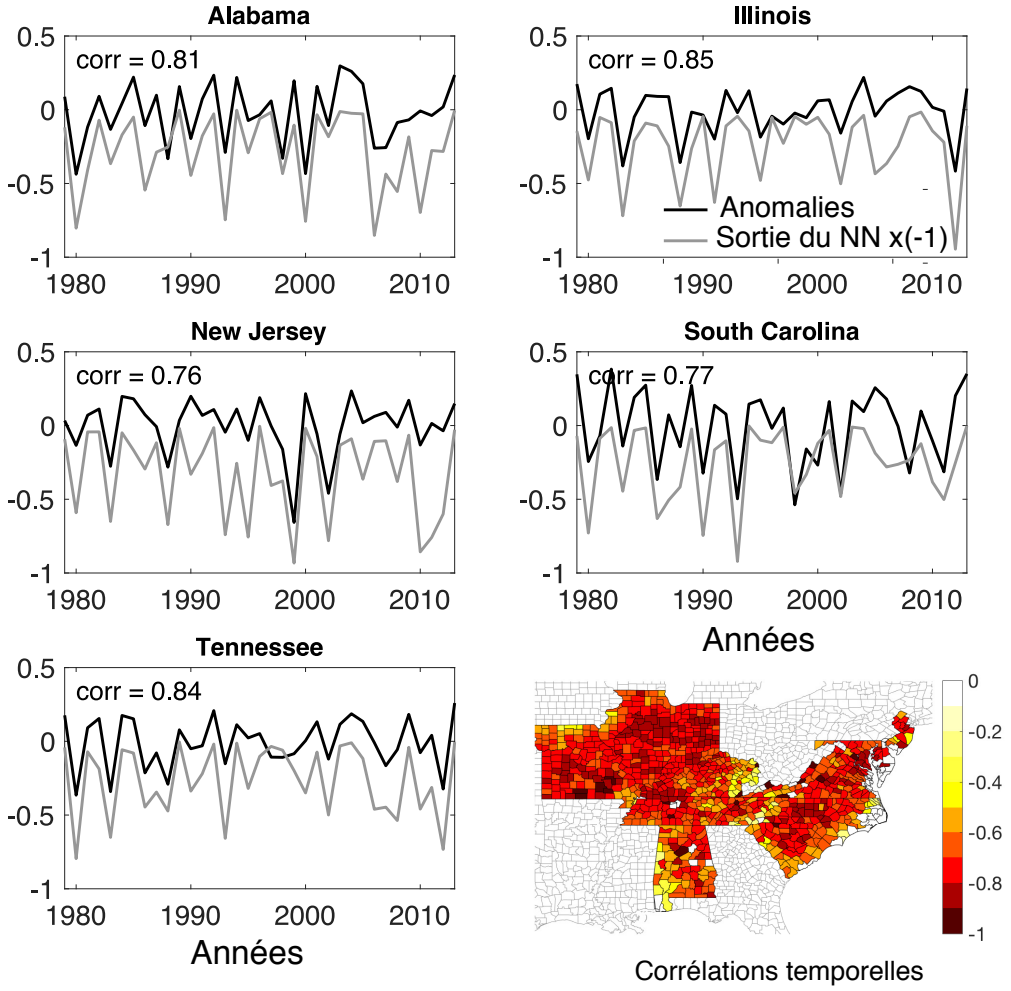


FIGURE VI.10 – Séries temporelles des anomalies de rendement et de NNoutput $\times(-1)$ (l'opposé de la sortie du réseau de neurones) pour cinq cantons dans plusieurs états distants (choisis à la Figure VI.1). En bas à droite : corrélation temporelle entre les séries temporelles des anomalies de rendement et la série temporelle de la sortie du réseau, à l'échelle du canton. Les séries temporelles de la sortie du réseau de neurones (NN) et de l'anomalie de rendement sont négativement corrélées, car elles varient de façon opposée.

Le nuage de points de la Figure VI.11 montre combien les anomalies de rendement (toutes recueillies) sont liées aux réponses du réseau et à leur variabilité : plus l'anomalie de rendement est basse, plus la sortie du réseau est élevée (en moyenne). Le R^2 du scatter-plot est égal à 0.41, ce qui illustre un lien élevé entre les anomalies de rendement et la sortie du réseau (41% de la variabilité des anomalies de rendement peut être expliqué par la sortie du réseau).

La procédure du réseau de neurone peut faire l'objet d'analyses post-apprentissage afin d'estimer la réponse du modèle aux variations des entrées. Par exemple, le comportement du réseau peut être observé sur une base de données régulière, en utilisant deux variables d'entrée prédominantes, telles que SPEIjuillet et Tjuillet (Figure VI.12) : des anomalies positives de température de juillet associées à des anomalies négatives de SPEI conduisent à la classification "extrême". Dans la Figure VI.12 (à gauche), la valeur du seuil de décision C = 0.60 est soulignée par une courbe noire. Tous les cas

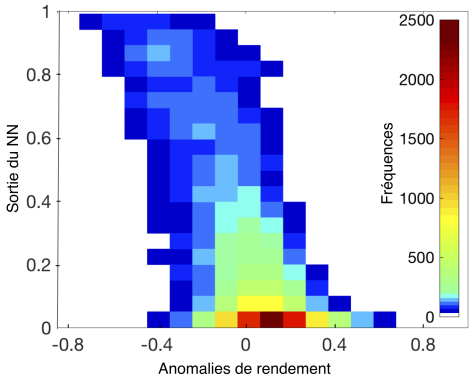


FIGURE VI.11 – Scatter-plot des anomalies de rendement par rapport aux sorties du réseau (NN). La barre de couleurs indique la densité de l'échantillon.

en-dessous de cette courbe sont classés par le réseau comme non-extrêmes, et tous les cas au-dessus de cette courbe sont classés comme extrêmes. En regardant la partie gauche/supérieure de la figure, la ligne de séparation entre les cas extrêmes et non-extrêmes est assez linéaire (même si une régression linéaire ou un arbre de décision binaire ne donnerait pas de résultats similaires pour ces cas). Le coin supérieur droit n'est pas très fiable car très peu d'échantillons sont disponibles pour ce domaine (Figure VI.12, droite). On peut voir dans la Figure VI.12 (à droite), que plus d'échantillons avec Tjuillet élevé et SPEIjuillet faible seraient extrêmement précieux pour améliorer la calibration de notre classification des cas extrêmes.

La Figure VI.12 (gauche) montre pour une même valeur de SPEIjuillet (axe horizontal) que la probabilité d'extrêmes n'est pas la même pour toutes les valeurs de Tjuillet (axe vertical) ; et inversement, pour une valeur spécifique de Tjuillet, la probabilité d'extrêmes peut ne pas être la même pour toutes les valeurs de SPEIjuillet. Cela signifie qu'il existe des interactions entre ces deux prédicteurs pour estimer la probabilité des extrêmes, et que le réseau est capable d'attraper ce comportement. Les séparateurs du réseau sont clairement multivariés.

Selon cette analyse post-apprentissage, les cas avec Tjuillet $\ll 0$ sont toujours classés comme non-extrêmes (pour toutes les valeurs de SPEIjuillet), ce qui pourrait montrer les limites de la classification. Il peut être surprenant de voir que si Tjuillet $\ll 0$, le maïs réagit de la même manière, quelle que soit la valeur de SPEIjuillet (si SPEIjuillet $\ll 0$ ou SPEIjuillet $\gg 0$). Cependant, [Cutf86] a montré que pour la croissance des racines du maïs au cours de l'émergence, la sensibilité à la teneur en eau diminue avec la diminution de la température. Lorsque les températures sont très inférieures à la moyenne, le taux de croissance racinaire évolue très peu, quelle que soit l'eau relative disponible.

La Figure VI.13 illustre la distribution des réponses du réseau de neurones pour toutes les anomalies de rendement de la base de données. La courbe jaune illustre la distribution des anomalies de rendement considérées comme "extrêmes" avec un degré élevé de certitude (NNoutput $\geqslant 0.7$). La courbe bleue illustre la distribution des anomalies de rendement considérées comme "non-extrêmes" avec un degré élevé de certitude (NNoutput $\leqslant 0.5$). Enfin, la courbe rouge illustre la distribution des anomalies de rendement considérées comme "extrêmes" ou "non-extrêmes" avec peu de certitude ($0.5 <$ NNoutput < 0.7 : la sortie du réseau est proche du seuil de décision $C = 0.60$).

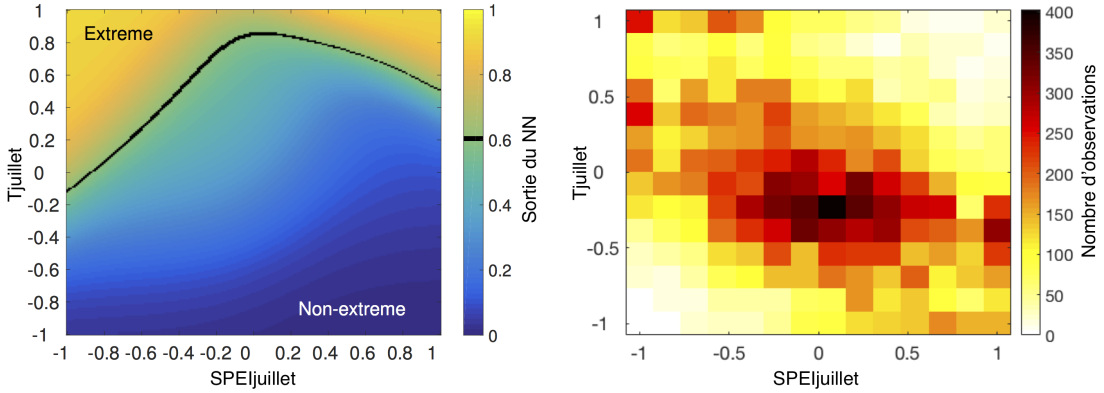


FIGURE VI.12 – Comportement du classificateur (avec les deux entrées les plus importantes, Tjuillet et SPEIjuillet) sur une grille régulière de ces deux entrées. Le seuil de décision optimal C pour classifier les situations extrêmes et non-extrêmes est de 0.60 et est représenté par une courbe discriminante noire. La figure de droite représente le nombre d'échantillons $\{Tjuillet,\ SPEIjuillet\}$ dans les observations.

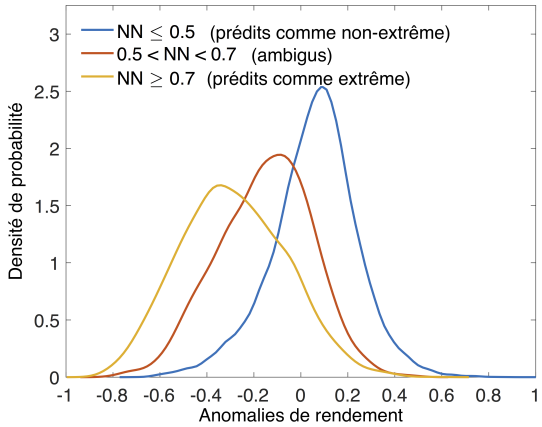


FIGURE VI.13 – Distribution de probabilités des anomalies de rendement pour les échantillons classés comme (1) "extrêmes" avec une probabilité élevée (courbe jaune), (2) comme "non-extrêmes" avec une probabilité élevée (courbe bleue) , et (3) comme "ambigus" (courbe rouge).

Les histogrammes jaune et bleu ne se croisent pas trop, ce qui signifie que le classificateur fait un bon travail en séparant les cas extrêmes des cas non-extrêmes. La valeur moyenne de la distribution jaune (quand NNoutput $\geqslant 0.7$) est proche de -0.4, alors que la valeur moyenne de la distribution bleue (quand NNoutput $\leqslant 0.5$) est proche de 0.1. Il est satisfaisant d'obtenir une valeur moyenne de la distribution rouge (quand $0,5 <$ NNoutput $< 0,7$), proche de -0.1 : les trois moyennes de ces distributions sont bien séparées. En ce sens, la sortie du réseau donne de bons résultats en tant qu'indice de gravité des pertes de rendement.

Le modèle statistique décrit ici semble donner des informations quantitatives sur le rendement. La valeur de la sortie du réseau nous informe sur deux choses. D'une part, il nous informe sur la précision et sur l'assurance du réseau quant à sa décision : des rendements très inférieurs à 0.5 sont clairement non extrêmes, et ceux beaucoup plus forts que 0.7 sont clairement extrêmes. D'autre part, la sortie du réseau nous informe sur la probabilité de faire face à une perte de rendement extrême : si la sortie est proche de zéro, alors avec une forte probabilité, la sortie finale sera correcte, ou meilleure que les autres années. La Figure VI.13 montre que presque aucune des pertes extrêmes de rendement n'est mal classée par le réseau. Le modèle se comporte comme un indice de gravité des pertes de rendement.

4 Analyse de sensibilité aux choix méthodologiques

Dans cette section, nous analysons la sensibilité de nos résultats aux principaux choix méthodologiques. Le premier choix concerne le seuil T utilisé pour définir les cas extrêmes et non-extrêmes. Dans cet article, il est fixé à -0.45. Ici, nous analysons la sensibilité lorsque ce seuil varie de -0.30 à -0.60.
Le deuxième choix méthodologique est l'existence ou non d'un écart entre échantillons extrêmes et non-extrêmes : en d'autres termes, faut-il supprimer les données dont l'anomalie de rendement est très proche du seuil T fixé ci-dessus (par exemple, avec une distance de séparation de 0.05) ? Dans cette section, nous analysons la sensibilité des résultats pour une distance de séparation prenant successivement les valeurs 0, 0.025, 0.05, 0.075, puis 0.1.

Enfin, la qualité du classificateur dépend de la contrainte subjective sur le TFP : dans cette section, les résultats sont comparés lorsque cette contrainte varie de 5% à 40% par pas de 5%. Dans les sections précédentes, il est fixé à 15% (TFP$\leqslant$ 15%).

Ces paramètres impactent les résultats à trois niveaux pour notre étude : les valeurs des probabilités conditionnelles $\mathbb{P}(météo|extrême)$ et $\mathbb{P}(extrême|météo)$ (Section 4.1), les prédicteurs sélectionnés (Section 4.2), et la qualité de la classification (Section 4.3).

4.1 Sensibilité des probabilités conditionnelles au seuil T utilisé pour la définition des extrêmes, et à la distance de séparation

Dans cette section, nous analysons la sensibilité des nuages de points et des probabilités conditionnelles $\mathbb{P}(météo|extrême)$ et $\mathbb{P}(extrême|météo)$ (Section 3.1) au choix du seuil T utilisé pour définir les anomalies extrêmes, et à la distance de séparation entre la classe extrême et la classe non-extrême. Nous nous concentrons sur les deux indicateurs les plus importants : Tjuillet et SPEIjuillet. Nous rappelons que nous travaillons avec des anomalies météorologiques. La Figure VI.14 analyse quatre probabilités conditionnelles d'observer des conditions météorologiques, sachant quelles données se réfèrent à une perte de rendement extrême : $\mathbb{P}(Tjuillet < 0.05, SPEIjuillet > -0.35|extrême)$, $\mathbb{P}(Tjuillet > 0.05, SPEIjuillet > -0.35|extrême)$, $\mathbb{P}(Tjuillet < 0.05,$

$SPEIjuillet < -0.35|extrême)$, et $\mathbb{P}(Tjuillet > 0.05, SPEIjuillet < -0.35|extrême)$. La Figure VI.15 se concentre quant à elle sur les quatre probabilités conditionnelles d'observer des rendements extrêmes connaissant les conditions météorologiques : $\mathbb{P}(extrême|Tjuillet < 0.05, SPEIjuillet > -0.35)$, $\mathbb{P}(extrême|Tjuillet > 0.05, SPEIjuillet > -0.35)$, $\mathbb{P}(extrême|Tjuillet < 0.05, SPEIjuillet < -0.35)$, et $\mathbb{P}(extrême|Tjuillet > 0.05, SPEIjuillet < -0.35)$.

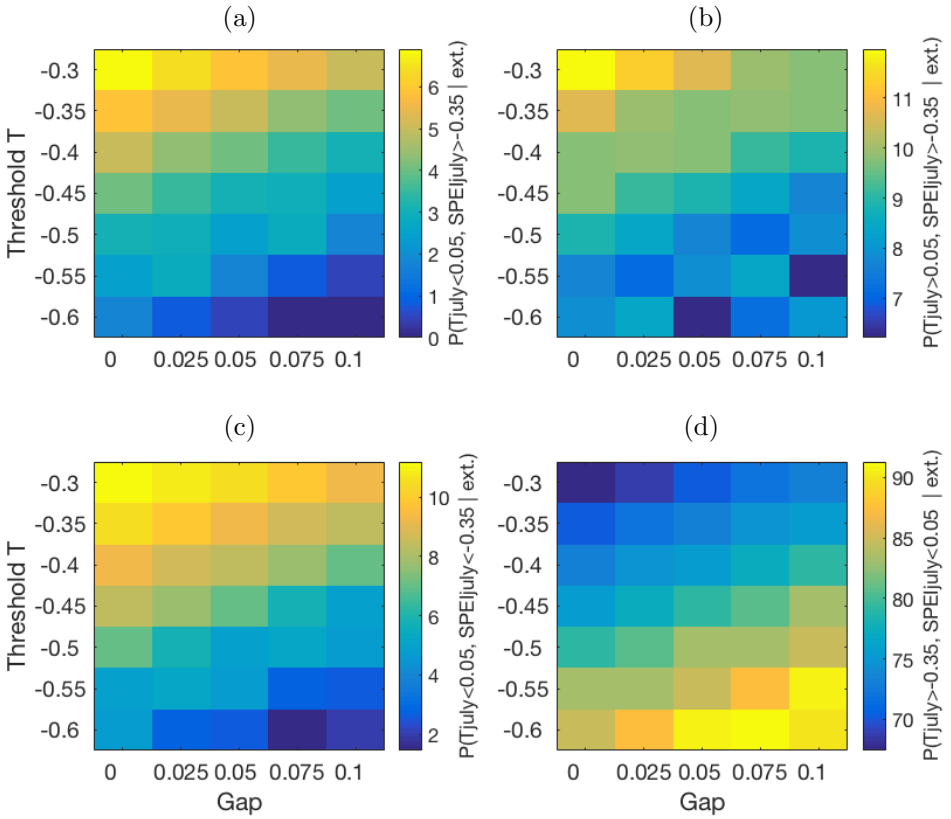


FIGURE VI.14 – Les quatre probabilités conditionnelles $\mathbb{P}(Tjuillet, SPEIjuillet|extrême)$ (en %) lorsque le seuil T varie de -0.3 à -0.6 et la distance de séparation (Gap) varie de 0 (pas de séparation) à 0.1.
(a) fait référence à $\mathbb{P}(Tjuillet < 0.05, SPEIjuillet > -0.35|extrême)$, (b) fait référence à $\mathbb{P}(Tjuillet > 0.05, SPEIjuillet > -0.35|extrême)$, (c) fait référence à $\mathbb{P}(Tjuillet < 0.05, SPEIjuillet < -0.35|extrême)$, et (d) fait référence à $\mathbb{P}(Tjuillet > 0.05, SPEIjuillet < -0.35|extrême)$.

La Figure VI.14 représente les quatre probabilités conditionnelles $\mathbb{P}(météo|extrême)$ listées ci-dessus lorsque le seuil T varie de -0.3 à -0.6, et la distance de séparation varie de 0 (sans écart) à 0.1. En (a), (b) et (c), pour un seuil T fixe, plus l'écart est élevé, plus la probabilité d'observer les conditions météorologiques est faible : plus les deux classes (extrêmes et non-extrêmes) sont distinctes, moins les conditions météorologiques (concernant Tjuillet et SPEIjuillet dans (a), (b) et (c)), caractérisent des pertes de rendement extrêmes. Ces trois probabilités conditionnelles sont faibles, comparées à $\mathbb{P}(Tjuillet > 0.05, SPEIjuillet < -0.35|extrême)$ sur la Figure VI.14, (d).
En (d), pour un seuil fixe, plus l'écart est élevé, plus la probabilité d'observer $Tjuillet > 0.05$ et $SPEIjuillet < -0.35$ pour des pertes de rendement extrêmes est élevée : les pertes de rendement très extrêmes sont presque toujours liées à la combinaison de

$Tjuillet > 0.05$ et $SPEIjuillet < -0.35$. Pour un écart fixe, plus le seuil est bas (plus la classe des extrêmes est grande), plus la probabilité d'observer $Tjuillet > 0.05$ et $SPEIjuillet < -0.35$ est faible. L'information est en train de se diluer lorsqu'on ajoute à la catégorie des extrêmes, des données qui ne sont pas vraiment considérées comme extrêmes.

Pour toutes les sous-figures de la Figure VI.14, étant donné un seuil T fixé, la variation relative de la probabilité conditionnelle pour un écart variant de 0 à 0.1 n'est pas très importante. Donc, la distance de séparation a une petite influence.

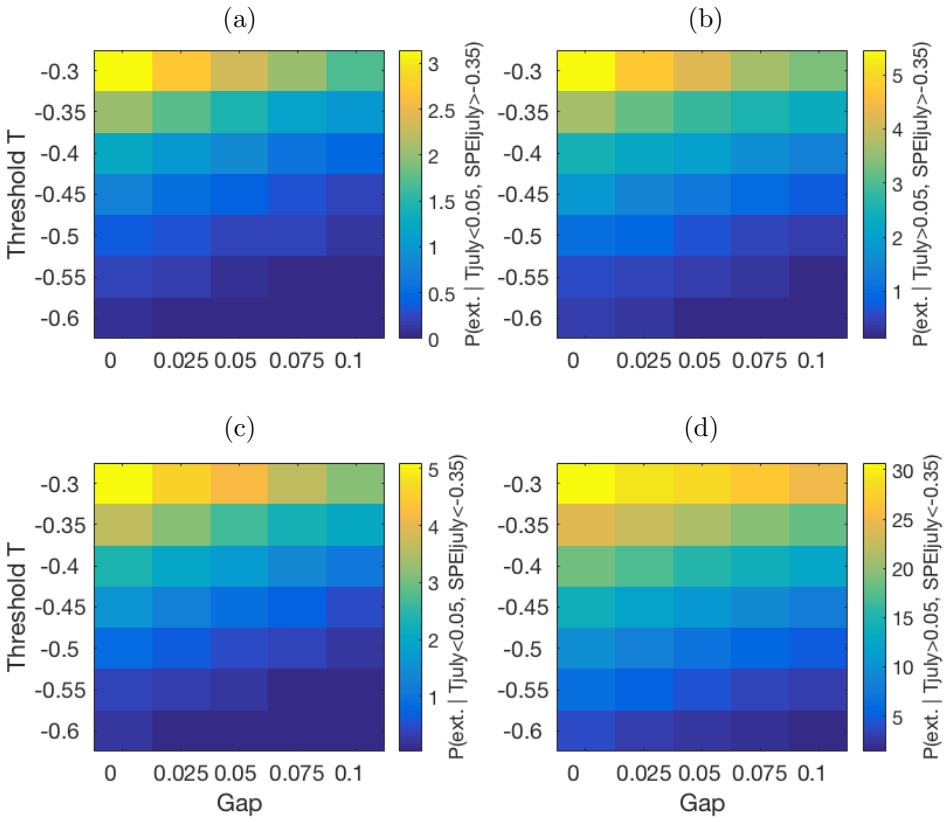


FIGURE VI.15 – Les quatre probabilités conditionnelles $\mathbb{P}(extrême|Tjuillet, SPEIjuillet)$ (en %) quand le seuil T varie de -0.3 à -0.6, et que la distance de séparation (Gap) varie de 0 (pas de séparation) à 0.1.
(a) fait référence à $\mathbb{P}(extrême|Tjuillet < 0.05, SPEIjuillet > -0.35)$, (b) fait référence à $\mathbb{P}(extrême|Tjuillet > 0.05, SPEIjuillet > -0.35)$, (c) fait référence à $\mathbb{P}(extrême|Tjuillet < 0.05, SPEIjuillet < -0.35)$, et (d) fait référence à $\mathbb{P}(extrême|Tjuillet > 0.05, SPEIjuillet < -0.35)$.

La Figure VI.15 représente les quatre probabilités conditionnelles $\mathbb{P}(extrême|météo)$ listées ci-dessus quand le seuil varie de -0.3 à -0.6 et la distance de séparation varie de 0 (sans écart) à 0.1. Un comportement similaire peut être observé en (a), (b), (c) et (d) mais avec des amplitudes différentes. La Figure VI.15 (d) propose la plus haute probabilité $\mathbb{P}(extrême|météo)$: $\mathbb{P}(extrême|Tjuillet > 0.05, SPEIjuillet < -0.35)$ est la probabilité conditionnelle la plus pertinente ce qui n'est pas surprenant. Plus le seuil T est proche de zéro (plus la classe extrême est grande), plus la probabilité $\mathbb{P}(extrême|Tjuillet > 0.05, SPEIjuillet < -0.35)$ est élevée (les nouveaux échantillons extrêmes se réfèrent aux conditions météorologiques $Tjuillet > 0.05$ et $SPEIjuillet < -0.35$). Ici aussi,

pour toutes les sous-figures de la Figure VI.15, et étant donné un seuil, la variation relative de la probabilité conditionnelle pour un écart variant de 0 à 0.1 n'est pas très importante.

Par la suite, comme l'impact d'une distance de séparation sur les probabilités conditionnelles est secondaire, nous le considérons comme négligeable et ce paramètre sera fixé à 0.

4.2 Sensibilité de la sélection des entrées au seuil T de définition des extrêmes

Seuil T	Proportion d'extrêmes	Entrées sélectionnées successivement 1^{ere}	2^{eme}	3^{eme}	4^{eme}
-0.30	11%	DJjuillet	SPEIjuillet	SPEIjuin	DJaoût
-0.35	8%	SPEIjuin	SPEIjuillet	Tjuillet	DJaoût
-0.40	6%	SPEIjuillet	SPEIjuin	Tjuillet	DJaoût
-0.45	5%	SPEIjuillet	Tjuillet	SPEIjuin	Taoût
-0.50	3%	Tjuillet	SPEIjuillet	SPEIjuin	LSF
-0.55	2%	Tjuillet	SMjuillet	SPEIjuin	LSF
-0.60	1%	DJjuillet	SMjuillet	SPEIavril	DJjuin

TABLEAU VI.4 – Les quatre premières entrées sélectionnées lorsque le seuil T diminue de -0.3 à -0.6. Chaque résultat provient de 50 ensembles d'apprentissage et de validation différents (entrée prédominante). SM se réfère au prédicteur météo "humidité du sol" (7-28 cm de sol), et LSF à "Last Spring Frost".

Le Tableau VI.4 définit l'influence du seuil T sur la sélection des entrées. La proportion d'extrême varie (de façon assez linéaire) de 1% à 11% lorsque le seuil T varie de -0.6 à -0.3. Le Tableau VI.4 montre une grande cohérence pour les trois premières entrées sélectionnées : même si le classement n'est pas le même, (1) la chaleur en juillet (Tjuillet, ou DJjuillet), (2) l'humidité en juillet (SPEIjuillet ou SMjuillet), et (3) le bilan hydrique en juin (SPEIjuin), sont toujours sélectionnés (quelle que soit la valeur du seuil T, excepté -0.6). Quand, moins d'un pour cent des données est considéré comme extrême, SPEIavril est choisi comme troisième entrée, au lieu de SPEIjuin (cas T=-0.6). La quatrième entrée sélectionnée est celle qui varie le plus : la chaleur du mois d'août est choisie pour les seuils T inférieurs à -0.5. Après cette valeur, le dernier gel printanier (LSF) remplace Taoût. La fiabilité de la quatrième entrée sélectionnée pour $T = -0.60$ soulève quelques questions, étant donné la faible taille du groupe extrême.
En outre, en analysant à quelle fréquence cette quatrième entrée est choisie, on est souvent confronté à une faible fréquence ce qui signifie que la sélection de la quatrième entrée dépend fortement des échantillons d'apprentissage et de test.

Dans ce chapitre, la quatrième entrée est la température moyenne en août (Taoût). Taoût et DDaoût ont des distributions très similaires. Sachant que la chaleur en août est sélectionnée comme quatrième entrée 4 fois sur 7, nous sommes assez confiants pour la robustesse de notre sélection d'entrées.

4.3 Sensibilité du classificateur au seuil T pour la définition des extrêmes, et à la contrainte sur le TFP

La Figure VI.16 définit l'influence du seuil T et de la contrainte TFP sur le TVP moyen. Pour un seuil fixe T, plus la contrainte sur le TFP est faible (c'est-à-dire, TFP $\leqslant \alpha$ avec une valeur α élevée), plus le TVP est élevé. Ceci est cohérent, car plus les

données non-extrêmes sont mal classées, plus le nombre d'échantillons classés comme extrêmes est élevé, et donc plus le pourcentage d'extrêmes bien classés est élevé.

Pour une contrainte fixe, il y a un effet de saturation pour les valeurs de TVP lorsque TFP $\leqslant \alpha$ avec $\alpha \geqslant 30\%$ (partie droite du tableau). Même si nous sommes moins exigeants sur le nombre de faux positifs, le nombre de vrais positifs ne s'améliore pas. Pour les contraintes $\alpha \leqslant 25\%$, plus le seuil T est petit, meilleur est le TVP. En effet, lorsque T est élevé (proche de -0.3), beaucoup d'anomalies négatives sont décrétées comme extrêmes (sans toujours se référer à une situation météorologique extrême) et le groupe extrême est plus hétérogène, rassemblant une variété de situations météorologiques. Inversement, si le seuil T est petit (proche de -0.6) seules les anomalies de rendement très négatives sont considérées comme extrêmes et le groupe extrême contient moins de bruit. Il semblerait que les quatre entrées sélectionnées (Tjuillet, SPEIjuillet, SPEIjuin, et Taoût) expliquent mieux les anomalies très négatives, que lorsque ces anomalies sont mêlées à des anomalies négatives modérées.

Dans ce chapitre, nous avons choisi de ne pas dépasser un TFP de 15%, et de considérer 5% des anomalies comme extrêmes (ce qui correspond au seuil $T = -0.45$). L'analyse de sensibilité globale montre ici que, sans changer la contrainte sur TFP, la taille du groupe extrême doit être réduite par deux ou par trois pour obtenir un TVP significativement meilleur. Cependant, nous savons que les résultats sont moins fiables lorsque la taille du groupe extrême est réduite.
En gardant un seuil T égal à -0.45, le TVP est significativement meilleur lorsque la contrainte sur le TFP est relaxée à TFP $\leqslant 20\%$. Cependant, il faut s'attendre à ce que encore plus d'anomalies soient classées comme extrêmes bien qu'elles ne le soient pas.

	$\overline{TVP}$ (%)	Contrainte sur TFP (%)							
		5	10	15	20	25	30	35	40
Seuil T	-0.30	39	57	69	77	80	83	83	83
	-0.35	41	58	70	78	83	85	86	86
	-0.40	43	59	70	78	84	86	87	87
	-0.45	44	60	71	78	83	85	86	87
	-0.50	46	61	73	79	82	86	86	86
	-0.55	52	68	75	80	84	85	86	86
	-0.60	55	69	78	83	85	85	86	86

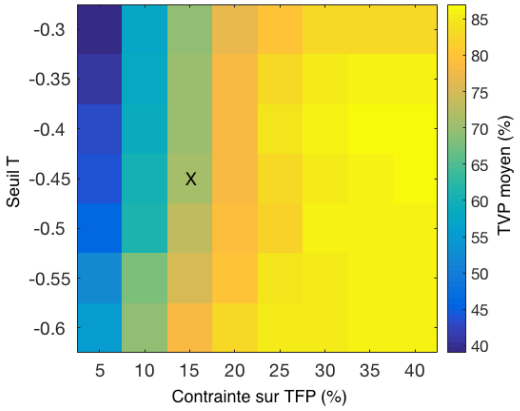


FIGURE VI.16 – Gauche : Moyenne du TVP (%) dans l'ensemble de validation lorsque la contrainte sur le TFP varie de 5% à 40% et le seuil de définition extrême T varie de -0.30 à -0.60. Dans chaque cas, 200 réseaux de neurones sont exécutés avec un ensemble d'apprentissage et de validation différents. Pour toutes les configurations, les quatre entrées du réseau de neurones sont SPEIjuillet, Tjuillet, SPEIjuin et Taoût.
Droite : représentation colorée du tableau de gauche. La croix fait référence à la configuration des résultats des sections précédentes : $T = -0.45$ et TFP $\leqslant 15\%$.

Conclusion sur les choix méthodologiques

Plusieurs choix méthodologiques inhérents ont été faits tout au long de l'étude et peuvent avoir une influence sur les conclusions de ce chapitre. Cette section teste cette sensibilité et ce paragraphe donne un aperçu des principales conclusions.

Pour un seuil T fixé, $\mathbb{P}(w|e)$ et $\mathbb{P}(e|w)$ ne changent pas de façon significative si les anomalies proches du seuil ne sont pas prises en compte. De plus, seul les deux probabilités $\mathbb{P}(Tjuillet > 0.05,\ SPEIjuillet < -0.35|extrême)$ et $\mathbb{P}(extrême|Tjuillet > 0.05, SPEIjuillet < -0.35)$ ont des valeurs suffisamment élevées pour observer des variations substantielles lors de la variation du seuil : plus le seuil T est proche de zéro ($> -0,4$), plus la classe extrême est grande, moins l'événement $\{Tjuillet > 0.05, SPEIjuillet < -0.35\ \}$ caractérise les rendements extrêmes (et vice versa, avec un seuil inférieur à -0.50).
Le choix du seuil T pour la définition des extrêmes n'affecte pas la sélection des trois premières entrées. Cependant, la quatrième entrée sélectionnée est plus sensible au seuil T : les prédicteurs météo liés à la chaleur en août ne sont choisis que pour $T > -0.5$. Lorsque moins de 2% des données sont considérées comme extrêmes ($T \leqslant -0.55$), les résultats de la quatrième entrée sélectionnée sont différents.

Enfin, avec un seuil T proche de -0.45, les résultats du réseau de neurones ne pourraient guère être meilleurs sans libérer la contrainte sur le TFP, ce qui ne serait pas profitable aux résultats puisqu'il y a encore trop de Faux Positifs. En gardant la contrainte $TFP < 15\%$, la taille du groupe extrême devrait être réduite par deux par trois pour obtenir un meilleur TVP. Comme pour toutes les études extrêmes, le faible nombre d'échantillons extrêmes est déterminant pour la qualité des résultats.

Selon cette analyse de sensibilité aux choix méthodologiques, nous sommes assez confiants quant à la robustesse de l'étude.

5 Discussion et conclusion du chapitre

Ce chapitre introduit un classificateur utilisant un réseau de neurones pour prévoir la probabilité de perte de rendement extrême en agriculture, en particulier pour la production de maïs aux États-Unis. Sans hypothèses agronomiques, dans une approche axée sur les données, notre modèle de classification statistique atteint une bonne précision. Le classificateur identifie bien les années avec beaucoup, modérément, ou peu de pertes de rendement extrêmes et fournit une estimation de la probabilité extrême qui peut être utilisée comme un bon indice de gravité des pertes de rendement. Le modèle est général et peut être utilisé tel quel, pour tout état de l'est des États-Unis. La relation météo-rendement est complexe et incomplète (d'autres facteurs manquants influencent également le rendement). Par construction, les réseaux de neurones sont bien adaptés pour prendre en compte les interactions entre les entrées. Ceci est une caractéristique importante ; le réseau est capable de trouver et d'exploiter la combinaison

simultanée de chaleur élevée et de faible humidité qui est dévastatrice pour les rendements des cultures. De plus, une analyse de sensibilité a déterminé si, et dans quelle mesure, les choix inhérents à notre méthodologie ont un impact sur nos résultats. Cela permet d'évaluer la robustesse de notre méthodologie. Il a été montré que les paramètres utilisés dans le modèle (tels que le seuil T utilisé pour définir les extrêmes ou la contrainte sur les Faux Positifs), n'ont pas d'impact significatif sur les résultats.

D'autres approches statistiques auraient pu être traitées pour modéliser l'impact de la météo sur les rendements extrêmement faibles : nous avons par exemple réalisé une étude à l'aide d'arbres de décision, mais les résultats étant plus sensibles au choix de la base d'apprentissage, nous avons préféré utiliser des réseaux de neurones qui étaient plus stables. De plus, la sortie du réseau de neurone (un nombre entre 0 et 1) est plus informative qu'une classification binaire $0/1$. On notera encore une fois que pour cette application, les résultats des modèles de prédiction ne sont pas limités par la sophistication des modèles statistiques, mais par le contenu en information des prédicteurs et de la taille de la base d'apprentissage. Il nous semble donc inutile de tester tout l'éventail des modèles disponibles.

Nos conclusions sont cohérentes avec les observations agronomiques sur les étapes critiques de la vie des plants de maïs [Ritc93, Dard06]. Le comportement du réseau est logique en termes de croissance des cultures puisque juillet est la période de floraison du maïs. Une température élevée en juillet affecte la photosynthèse qui devient moins efficace. Le rendement en maïs peut aussi être réduit en raison de la température élevée de l'air (35^oC et plus) pendant la pollinisation. Le maïs est très sensible à l'humidité en juillet et en août parce que c'est la période de formation des grains. Cependant, la pollinisation ne sera pas affectée par des températures élevées s'il y a assez d'humidité dans le sol, car le pollen se produit habituellement pendant les heures du matin. Le stress hydrique au cours de cette étape peut entraîner jusqu'à 50% de réduction du rendement [Wiat10].

L'approche proposée diffère des approches de modélisation du rendement des cultures, en particulier pour la modélisation du rendement extrême, sur les points suivants. Elle ne nécessite pas de connaissances spécialisées a priori et nos conclusions sont tirées des données. Nous nous concentrons sur le rendement observé du maïs dans les états les plus sensibles à la météorologie, couvrant trois zones climatiques contrastées (c'est-à-dire chaud-humide, mixte-humide et froide [Baec10]), de 1979 à 2013. L'analyse des prédicteurs météo se fait au départ indépendamment, puis de façon multivariée dans un modèle de classification. Les performances de prédiction du modèle sont évaluées par un processus de validation croisée. Notre méthodologie est très générale et peut facilement être appliquée à d'autres cultures ou à d'autres pays.
Il est intéressant de lister les différences avec l'étude de [Ben-16a]. Nos contributions principales sont les suivantes : (1) nous avons développé une sélection de prédicteurs multivariés et hiérarchiques où Ben-Ari effectue une analyse de sensibilité indépendamment pour chaque prédicteur ; (2) nous utilisons un classificateur multivarié et non linéaire (un réseau de neurones) au lieu d'une technique de seuil univariée, point important parce que certains prédicteurs interagissent pour agir sur le rendement en

maïs (par exemple chaleur et eau) ; (3) notre étude est une application nouvelle sur l'ensemble de l'est des États-Unis ; et (4) nous avons effectué une analyse spatiale complète de la sensibilité aux intempéries. Certaines de ces innovations ont été mentionnées dans la section "Room for improvement" de [Ben-16a].

Il peut être surprenant que les principaux prédicteurs pour la détection du rendement très négatif soient des prédicteurs simples tels que la température moyenne mensuelle et l'indice SPEI. Cela a été noté aussi dans [Ben-16a] qui a noté "qu'il est surprenant que des indicateurs simples [...] fonctionnent aussi bien que (et parfois même surperforment) des indicateurs composites et une simulation de modèle de culture". [Kotl07a] montre que les attributs météorologiques les plus importants pour le maïs sont ceux liés aux mois d'été. En particulier, son analyse confirme que la croissance des cultures dépend fortement de la quantité d'eau en juillet. L'absence de mois de croissance précoce peut s'expliquer par le fait que les prédicteurs d'avril/mai fournissent moins d'information sur le développement futur de la plante et sont moins précis quant à l'évolution et la capacité de la plante à se remettre des conditions climatiques difficiles. [Kotl07a] montre aussi que l'agrégation des conditions météorologiques mensuelles en saison a diminué le gain d'information. L'utilisation d'une résolution temporelle plus faible n'améliorerait pas les résultats. Cependant, [Ben-16a] semble montrer qu'il y a un gain d'information lorsque l'on considère des données sur des périodes statiques d'évolution de la plante. La combinaison de variables ne semble pas non plus pertinente, et le réseau de neurone est davantage capable de le faire seul. S'il y a trop de liberté dans les entrées du modèle, mais pas assez de données pour calibrer le modèle, cela nuit à la qualité de l'estimateur. Il est nécessaire de trouver une complexité (modèle et information d'entrée) compatible avec le nombre de données disponibles pour calibrer tout cela. C'est donc une forme de régularisation. Si nous avions plus de données historiques décrivant un large éventail de situations météorologiques, avec des résultats de rendement très variables, nous pourrions utiliser des choses plus complexes dans notre modèle. Mais ce n'est pas le cas.
Ce qui est également surprenant, comme mentionné par [Ben-16a], c'est que même les modèles agronomiques ne sont pas meilleurs ou même moins bon que ces simples modèles. Soit cela signifie que ces modèles ne sont pas suffisamment complets (les variables qui influencent les rendements ne sont pas pris en compte), soit cela suggère qu'il y a aussi un problème de calibration du modèle avec des données historiques.

Le but de notre modélisation statistique n'est pas d'estimer un rendement mais plutôt de définir un indice de sévérité de rendement qui indiquerait des pertes de rendement extrêmes. En outre, il est extrêmement difficile de prédire avec précision en juillet ou en aval les pertes extrêmes. Les causes des pertes de rendement extrêmes sont nombreuses et peuvent survenir juste avant la récolte. En outre, la faible proportion de cas extrêmes dans la base de données n'est pas suffisante pour capturer avec précision les valeurs de rendement pour divers modèles météorologiques. Une probabilité exige (et fournit également) moins d'informations. C'est pourquoi, nous recherchons une probabilité de cas extrêmes plutôt qu'une stricte prévision du rendement.

Notre modèle peut être utilisé de plusieurs façons. Les informations fournies par le classificateur sont précieuses pour de nombreuses personnes : les agriculteurs pourraient éviter un déficit ponctuel en fin d'année, en anticipant la signature d'un contrat d'assurance. Les compagnies d'assurance pourraient couvrir leurs risques contre des

demandes d'indemnisation importantes et imprévues. Cependant, les offres d'assurance sont encore limitées, comme en Europe [Biel09]. Les organisations non gouvernementales pourraient anticiper un manque possible de nourriture dans une région spécifique. Les fabricants de produits alimentaires pourraient éviter de manquer de stocks de matières premières, anticipant leur approvisionnement d'une autre région [Iizu13]. Les magasins d'alimentation pourraient également adapter l'approvisionnement alimentaire en cherchant un autre fabricant. Puisque le réseau de neurones estime bien les années avec de nombreux extrêmes à partir de simples informations météorologiques, il peut être utilisé dans des simulations climatiques pour étudier l'impact du changement climatique dans l'agriculture. Le modèle fournit également un indice de gravité des pertes de rendement qui peut être utilisé pour obtenir des alertes précoces.

La classification de la perte de rendement extrême est satisfaisante mais pas parfaite. Cela était à prévoir : la prévision des cas extrêmes est difficile notamment à cause de la rareté des échantillons disponibles pour calibrer les modèles statistiques. Une plage importante d'anomalies modérément négatives (entre les années normales et extrêmes) peut être qualifiée d'ambiguë : la sortie du réseau de neurone est proche de la valeur C du seuil de décision.

Il existe de nombreuses façons d'améliorer la précision de notre classificateur extrême, et nous en mentionnons ici deux. (1) Le nombre de rendements non extrêmes classés comme "extrêmes" est plus élevé que prévu : la contrainte sur le Taux de Faux Positif (TFP<15%) à l'échelle des US n'est pas totalement satisfaisante. Une solution partielle à ce problème consisterait à spécialiser le classificateur sur les conditions locales, en adaptant par exemple la valeur C du seuil de décision pour chaque état (en supposant que suffisamment de cas extrêmes soient disponibles dans l'historique pour calibrer un modèle pour chaque état ou district). (2) La classification pourrait également être améliorée en ajoutant d'autres prédicteurs. Cependant, il est difficile de trouver de bons prédicteurs qui influent sur le rendement des cultures, disponibles depuis plus de 30 ans. Par exemple, certaines études utilisent les indices de végétation [Pand10, Tan13, Xiji13] tels que l'indice de végétation par différence normalisée (NDVI) ou la fraction de rayonnement photo-synthétiquement actif (FPAR). Potentiellement, nous pourrions également utiliser des informations non liées aux conditions météorologiques, telles que les propriétés du sol ou la topographie si celles-ci sont disponibles en grande quantité.

CHAPITRE VII

Impact à long terme : les prévisions climatiques

Table des matières

1 Introduction

1.1 Changement climatique et impacts inégaux

La météorologie et l'agriculture sont très corrélées. Au delà de la variabilité naturelle qui contrôle une partie de la variance des rendements agricoles, le changement climatique va affecter sur le long terme l'agriculture à l'échelle mondiale. En effet, le changement climatique affecte la moyenne et la variabilité des conditions météorologiques et la fréquence des événements extrêmes, qui déterminent en grande partie la variabilité de la production et des rendements [Desc07]. [Gorn10] passent en revue des études récentes sur les productions agricoles qui peuvent être affectées par le changement climatique. En utilisant plusieurs modèles de climat, ils présentent les changements météorologiques attendus pour illustrer les principales incertitudes.

Le changement climatique affecte déjà l'agriculture avec des effets inégalement répartis dans le monde. Selon [Chen16], le réchauffement climatique a couté environ 820 million de dollars aux producteurs de maïs et de soja en Chine, ces 10 dernières années. Il semble que les effets du changement climatique soient négatifs pour les pays aux faibles latitudes, et mitigés pour les pays aux latitudes plus élevées [Port14, Chal09]. Le changement climatique va probablement accroitre le risque d'insécurité alimentaire, surtout dans les pays en voie de développement [HLPE15].

[Anto12] passent en revue la littérature scientifique traitant des impacts de la météo sur la variabilité des rendements et étudient les besoins en terme d'assurance agricole, et l'efficacité de différentes stratégies des exploitations agricoles, en utilisant des données australiennes, canadiennes et espagnoles. Ils résument les effets que peut avoir le changement climatique sur les cultures, face à l'augmentation des températures [Brya08, Adam90], la modification du régime des pluies [Warr86], l'augmentation des évènements extrêmes dans certaines régions [Thom75, McQu81, Rami73], et l'augmentation de la dispersion géographique de certains parasites et de certaines maladies [Bere89].

Les quatrième et cinquième rapports du GIEC (Groupe d'Experts Intergouvernemental sur l'Évolution du Climat) servent de référence pour les projections du changement climatique [GIEC07, GIEC14]. Ils fournissent 50 années de projections résultant d'une importante synthèse d'études et de modèles de la littérature scientifique. Plusieurs laboratoires (l'IPSL et NARCCAP (North American Regional Climate Change Assessment Program) entre autres) proposent des bases de données climatiques pour les 50 prochaines années voir plus, et sous différents scénarios climatiques. Cependant, de grandes incertitudes demeurent concernant l'étendue et la distribution spatiale de ces changements.

Les analyses suggèrent que les effets vont dépendre des climats locaux, de la façon dont le climat évolue localement, et de conditions supplémentaires locales telles que les caractéristiques du sol. Selon [Anto12], une grande partie de l'impact du changement climatique sur l'agriculture viendra des effets des événements extrêmes. Quelques pays développés au climat tempéré vont tirer profit du changement climatique. D'autres peuvent penser à remplacer leurs activités actuelles par des activités plus adaptées, ce qui réduirait les préjudices. D'autres pays seront beaucoup moins chanceux et pourront souffrir d'une large augmentation des températures, d'une diminution des précipitations, ou ne pas être capables de s'adapter [Mend08]. Viennent s'ajouter à cela des facteurs indirects comme une augmentation potentielle des invasions d'insectes et des pathogènes, une modification des caractéristiques de sols et/ou un changement dans les besoins en eau pour l'irrigation. Ces facteurs qui sont pour certains

positifs et pour d'autres négatifs pour l'agriculture s'entrechoquent, annulant les effets de l'un et de l'autre et rendant l'estimation de l'impact très difficile à prédire [Brkl98, Desc07, DaSi09].

Concernant les États-Unis, le rapport [CCSP08] fournit une estimation de l'effet du changement climatique sur l'agriculture, sur les capacités des sols, les apports disponibles en eau, et sur la biodiversité. Il aborde la capacité du pays à identifier, observer et gérer les facteurs influençant l'agriculture, pour évaluer l'importance de ces facteurs et la façon dont ils vont évoluer dans le futur. Ce rapport identifie les changements observés sur les ressources disponibles et analyse si ces changements peuvent être attribués - entièrement ou partiellement - au changement climatique. L'horizon temporel utilisé dans ce rapport décrit généralement le passé récent jusqu'à la période 2030-2050, même si des résultats à plus long terme (2100) sont parfois considérés. Plus récemment, le rapport "US National Climate Assessment " résume dans [USGC14] l'impact présent et futur du changement climatique sur les États-Unis.

1.2 Les scénarios RCP (Representative Concentration Pathway)

Les scénarios RCP (pour Representative Concentration Pathway) sont quatre scénarios relatifs à l'évolution de la concentration en gaz à effet de serre au cours du XXI^e siècle, établis par le GIEC pour son cinquième rapport [GIEC14]. Chacun des quatre scénarios RCP - qui ont pour but de modéliser le climat futur - est basé sur une hypothèse différente concernant la quantité de gaz à effet de serre qui sera émise dans les années à venir (période 2000-2100) [Vuur11]. Le climat probable résultant du niveau d'émission choisi comme hypothèse de travail est résumé par chaque scénario RCP. Les noms des quatre scénarios proviennent de la gamme de forçage radiatif obtenue pour l'année 2100 : le scénario RCP 2.6 correspond à un forçage de +2,6 W/m^2, le scénario RCP 4.5 à +4,5 W/m^2, et de même pour les scénarios RCP 6 et RCP 8.5 [Moss08].

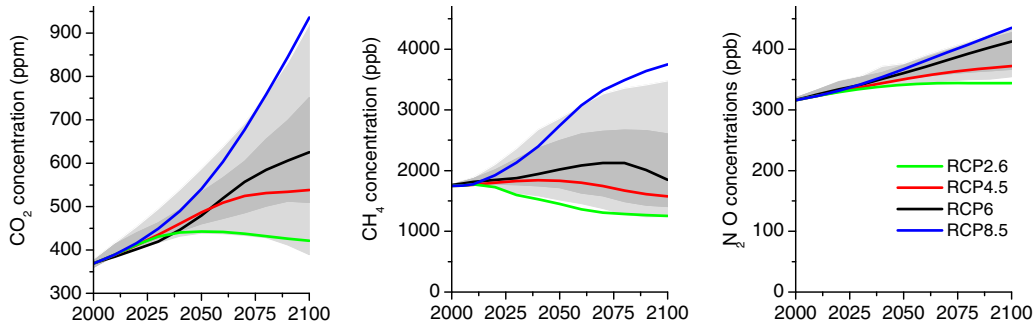


FIGURE VII.1 – Tendances des concentrations de gaz à effet de serre. La zone grise indique les 98e et 90e percentiles (gris clair/foncé) de la récente étude de [Clar10] (Figure tirée de [Vuur11])

La Figure VII.1 compare l'évolution de la concentration des gaz à effets de serre selon le scénario choisi. Ces quatre trajectoires correspondent chacune à une concentration atmosphérique en CO_2, qui aura un impact sur l'effet de serre, et donc sur le climat. Le scénario RCP 2.6 est le plus optimiste, car celui-ci intègre des effets de politiques de réduction des émissions et vise une limitation de l'augmentation de la température à 2^oC. La synthèse des différentes publications présentées dans le 5^e rapport du GIEC indique que, si l'on prend comme référence la période 1986-2005, le scénario

RCP 8.5 conduit à un réchauffement global en 2100 compris entre 2.6^oC et 4.8^oC. Pour le RCP 2.6, l'estimation la plus probable se situe entre 0.3^oC et 1.7^oC. Ce scénario est le seul qui permette de maintenir la température en dessous du seuil des 2^oC par rapport à la période pré-industrielle.

Dans ce chapitre, nous nous concentrerons sur les scénarios RCP 4.5 (qui prévoit une stabilisation des émissions à l'horizon 2100) et le RCP 8.5 (fortes émissions) puisque le scénario 2.6 - le moins contraignant en terme de forçage radiatif - est devenu irréaliste [Raft17].

1.3 Impacts potentiels sur les cultures et adaptation

Les impacts positifs du changement climatique sur l'agriculture que l'on peut espérer sont une augmentation de la productivité en raison des températures plus chaudes, la prolongation de la saison de croissance, l'augmentation de la productivité découlant de l'augmentation de CO_2 atmosphérique, et l'accélération des taux de maturation.
A l'inverse, les impacts négatifs auxquels il faut s'attendre sont une augmentation des invasions d'insectes, des dommages causés aux cultures par la chaleur extrême, des problèmes de planification découlant du manque de fiabilité des prévisions, une augmentation de l'érosion, un accroissement de la fréquence des maladies, une baisse d'efficacité des herbicides et des pesticides, et une augmentation du stress hydrique et des sécheresses.

Adaptation à cours terme - Le principe de base est qu'il faut en premier lieu travailler sur la baisse des besoins en eau, pour ensuite, dans un deuxième temps, travailler à l'augmentation de la ressource. Or aujourd'hui d'après [Denh14], la plupart des efforts visent à augmenter les ressources. Par exemple, augmenter la culture du sorgho dans le sud-ouest de la France permet de baisser les besoins d'irrigation.

Adaptation à long terme en continu - Sans même parler des risques potentiels liés au changement climatique, l'anticipation des états et des régions se doit d'être permanente. En effet, de nombreuses infrastructures - comme les barrages hydroélectriques, les canaux d'irrigation, ou encore les voies d'acheminement rapide de matières premières - demandent beaucoup d'années à être construites et de forts investissements, et doivent donc être décidées longtemps à l'avance. Pour ce faire, les agences régionales ont besoin de visibilité future sous forme d'estimations du climat mais aussi des réponses potentielles des activités socio-économiques. Elles ont besoin de savoir quels secteurs industriels vont connaitre des difficultés ; si des politiques de santés devront être renforcées ; si les déplacements de matières premières vont être nécessaires ; quels secteurs va t'il falloir attirer ou aider, etc. Toutes ces décisions reposent sur une information quant à l'évolution future et conjointe du climat et des activités socio-économiques.

Mesures politiques et fortes incertitudes - La réalité auxquelles les politiques doivent faire face est complexe et le risque porté à la production agricole sur le long terme est sujet à de fortes incertitudes quant aux véritables impacts du changement climatique selon les localités et le comportement des cultivateurs dans ce contexte d'incertitude. La volonté des assurances de fournir aux cultivateurs une garantie est aussi affectée par ces incertitudes et ces attentes. Les décisions prisent par les cultivateurs pour gérer les risques agricoles, les décideurs politiques, et les compagnies d'assurances vont prendre en compte ces incertitudes sur le climat futur. Les estimations actuelles des impacts du changement climatique sont généralement caractérisées par une large incertitude

qui résulte de la connaissance limitée que nous avons des procédés physiques, biologiques, socio-économiques, et de leurs interactions. Ces limitations freinent les efforts accomplis pour anticiper et s'adapter au changement climatique. Réduire ces incertitudes grâce à une meilleure compréhension des contributions de chaque facteur en jeu est déterminant mais il est peu probable que de telles incertitudes puissent être entièrement résolues. Il est donc important de prendre en compte ces incertitudes quand on analyse l'impact du changement climatique sur la variabilité de la production, tout comme les cultivateurs sont amenés à le faire quand ils analysent les risques de gestion agricole.

1.4 Simulations et projections climatiques de l'IPSL

Pour notre étude, nous utilisons les données climatiques brutes de simulations globales CMIP 5 (Coupled Model Intercomparison Project Phase 5) de l'IPSL avec une résolution spatiale de $2.5^o \times 1.25^o$ [IPSL15]. Les données sont disponibles à travers la plateforme ESGF (Earth System Grid Federation) qui stocke et distribue des bases de données mondiales issues de multiples simulations de modèles climatiques couplés océan-atmosphère. Plusieurs scénarios climatiques sont disponibles, dont le RCP 4.5 et le RCP 8.5 de 2006 à 2100. Une base de données pour des estimations passées non-forcées s'étale de 1850 à 2005 (historique simulé).

Le modèle intégré de climat de l'IPSL est complexe et repose sur une prise en compte conjointe des différentes parties du système climatique et des différents processus qui le régissent. Il contient par exemple un modèle d'atmosphère (LMDZ), un modèle d'océan, de glace de mer et de biogéochimie marine (NEMO), un modèle de surfaces continentales (ORCHIDEE), et un modèle de chimie troposphérique et stratosphérique (INCA-REPROBUS).

En résumé

Dans ce chapitre, le but n'est pas d'utiliser les modèles d'impacts construits tout au long de cette thèse pour quantifier précisément l'impact du changement climatique sur l'évolution à long terme des rendements agricoles. On cherchera plutôt à décrire qualitativement une tendance globale des anomalies de rendements ainsi que leur répartition géographique. On utilisera pour cela le modèle à effets mixtes linéaire à six entrées météorologiques simples (Tmai, Tjuin, Tjuillet, Taoût, Pjuillet et Paoût) du Chapitre V page 153.
Ce modèle nous permettra de comparer l'évolution des anomalies de rendement estimées jusqu'en 2060 selon les scénarios RCP 4.5 (prévoyant la stabilisation des émissions de gaz à effet de serre à l'horizon 2100) et RCP 8.5 (prévoyant de fortes émissions, sans stabilisation).

2 Calibration des projections climatiques

2.1 Méthodologie

Pourquoi faire des calibrations ?

Les prévisions climatiques ne sont pas un reflet exact de ce que sera la météo dans 50 ou 100 ans. Elles sont construites en tenant compte de contraintes spatiales (forçage de la température de surface, pour l'historique, par exemple) pour suivre à long terme un scénario climatique que se sont fixé les modélisateurs. Il n'y a donc pas de raison que les observations réelles soient identiques aux simulations climatiques. On observe très souvent un décalage entre la série temporelle des observations météo et celle des projections futures.

Dans une simulation climatique (et contrairement aux simulation pour la prévision météorologique) on n'utilise aucune mesure du temps qu'il fait ou qu'il a fait dans le passé (sauf pour initier la simulation). Vu le caractère chaotique du climat, il n'y a aucune chance que le temps qu'il fait un jour donné, soit le même que celui simulé par le modèle. Ce qu'il faut comparer, ce sont les propriétés statistiques du temps simulé. L'IPSL a fait plusieurs simulations (r1i1p1, r2i1p1...), que l'on peut comparer et dont on peut voir qu'elles produisent des résultats assez différents jour à jour, ou mois par mois, mais dont les propriétés statistiques sont bien comparables.

L'ajustement des simulations est donc très important pour les études d'impact au changement climatique, surtout si elles reposent sur des variables qui dépendent de seuils absolus (par exemple, nombre de jours de gel, ou jours de fortes précipitations) [Liu14, Teut12].

Notations

Lorsque l'on s'apprête à calibrer des bases de données, nous disposons de trois ensembles de données :

— une base d'observations (parcourant les années 1979-2013 dans notre cas),

— une base de simulations historiques (même année que pour les observations) issue du modèle climatique,

— et une base de simulations futures (2006-2100 dans notre cas).

Dans cette thèse, nous nous concentrerons sur deux scénarios climatiques, l'un lié au RCP 4.5 et l'autre lié au RCP 8.5. Une correction des simulations historiques vers les observations va être calculée, puis sera appliquée aux données futures.

Un résumé des méthodes les plus utilisées pour effectuer des calibrations est donné dans [Teut12]. Dans ce chapitre, on va se concentrer sur deux méthodes assez différentes : l'ajustement linéaire (appelée "linear scaling" en anglais), et la méthode quantile-quantile (ou "CDF matching").
La première méthode permet aux deux séries temporelles (observation et historique simulé) d'avoir la même moyenne et la même variance. L'approche quantile-quantile peut être considérée comme une technique non-linéaire pour éliminer les différences systématiques entre observations et historiques simulés. Grâce à cette méthode, les simulations futures sont remises à l'échelle de telle sorte que leur fonction de répartition se rapproche au mieux de celle des observations.

Par la suite, nous désignons pour chaque canton,

— $T_{obs}(m)$ et $P_{obs}(m)$ la température moyenne et les précipitations observées du mois m (années 1979-2013)

— $T_{hist}(m)$ et $P_{hist}(m)$ la température moyenne et les précipitations historiques simulées du mois m (années 1979-2013)

— $T_{fut}(m)$ et $P_{fut}(m)$ la température moyenne et les précipitations futures du mois m (années 2006-2100)

— $T^{cor}_{hist}(m)$ et $P^{cor}_{hist}(m)$ la température $T_{hist}(m)$ et les précipitations $P_{hist}(m)$ après calibration

— $T^{cor}_{fut}(m)$ et $P^{cor}_{fut}(m)$ la température $T_{fut}(m)$ et les précipitations $P_{fut}(m)$ après calibration

Les barres horizontales représentent les moyennes temporelles des variables, calculée pour chaque canton sur les années disponibles.

Ajustement Linéaire

L'ajustement linéaire est basé sur les différences entre les observation et les simulations historiques [Lend07]. Les précipitations sont corrigées avec un facteur basé sur le rapport entre les moyennes temporelles observées et celles historiques simulées :

$$\begin{aligned} P^{cor}_{hist}(m) &= P_{hist} \times a \\ P^{cor}_{fut}(m) &= P_{fut} \times a \end{aligned} \qquad \text{avec } a = \frac{\overline{P_{obs}(m)}}{\overline{P_{hist}(m)}}$$

La température est corrigée à l'aide d'un terme additif basé sur la différence entre les moyennes temporelles des données observées et de celles historiques simulées :

$$\begin{aligned} T^{cor}_{hist}(m) &= T_{hist} + b \\ T^{cor}_{fut}(m) &= T_{fut} + b \end{aligned} \qquad \text{avec } b = \overline{T_{obs}(m)} - \overline{T_{hist}(m)}$$

Les facteurs de correction et les termes ajoutés sont supposés rester invariables dans le temps. Par définition, les simulations climatiques historiques corrigées ont mêmes valeurs moyennes que les observations.

Approche quantile-quantile par lois connues

L'idée de la méthode quantile-quantile[1] est de corriger la fonction de répartition des données climatiques pour se rapprocher de la fonction de répartition des données observées. Pour cela, une première méthode est d'utiliser des lois de probabilités adaptées pour corriger les fonctions des répartitions.

La loi Gamma, de paramètres α et β est souvent considérée comme appropriée pour la distribution des précipitations. Plusieurs études ont montré que cette distribution était efficace pour l'analyse des précipitations dans la méthode quantile-quantile [Bloc09, Ines06, Pian10, Watt03]. Pour corriger le biais des précipitations, il faut donc calculer pour chaque canton, les paramètres α et β de la loi Gamma, puis corriger les

1. Aussi appelée en anglais "probability mapping" [Bloc09, Ines06] "quantile–quantile mapping" [Boe07, Dequ07, John11], "histogram equalization" [Roja11, Senn09], ou "distribution mapping" [Teut12]

données simulées par :

$$P_{hist}^{cor}(m) = F_{\Gamma}^{-1}(F_{\Gamma}(P_{hist}(m)|\alpha_{hist,m},\ \beta_{hist,m})|\alpha_{obs,m},\ \beta_{obs,m})$$

$$P_{fut}^{cor}(m) = F_{\Gamma}^{-1}(F_{\Gamma}(P_{fut}(m)|\alpha_{hist,m},\ \beta_{hist,m})|\alpha_{obs,m},\ \beta_{obs,m})$$

où F_{Γ} représente la fonction de répartition de la loi gamma, et F_{Γ}^{-1} son inverse. Pour l'analyse des températures, la loi normale de paramètres μ (la moyenne) et σ (la variance) est préférée [Scho90, Thom54]. Pour corriger le biais des températures, il faut donc calculer pour chaque canton, les paramètres μ et σ d'une loi normale, puis corriger les données simulées par :

$$T_{hist}^{cor}(m) = F_{N}^{-1}(F_{N}(T_{hist}(m)|\mu_{hist,m},\ \sigma_{hist,m})|\mu_{obs,m},\ \sigma_{obs,m})$$

$$T_{fut}^{cor}(m) = F_{N}^{-1}(F_{N}(T_{fut}(m)|\mu_{hist,m},\ \sigma_{hist,m})|\mu_{obs,m},\ \sigma_{obs,m})$$

où F_N représente la fonction de répartition de la loi normale, et F_N^{-1} son inverse.

Approche quantile-quantile empirique

L'utilisation de lois connues à des limites, surtout lorsque les biais sont importants et que la distribution des données ne s'applique pas parfaitement aux lois voulues (présence d'extrêmes plus importante, etc.).

Une seconde méthode pour corriger les fonctions de répartitions est de créer une fonction de transfert qui déplace la fonction de répartition des simulations vers celles des observations. On a alors recourt à une correction des données plus empirique : on calcule pour cela les fonctions de répartition des données historiques-simulées et observées. Le but est alors de construire une correspondance bijective entre les valeurs historiques et celles observées pour les corriger, comme indiqué à la Figure VII.2.

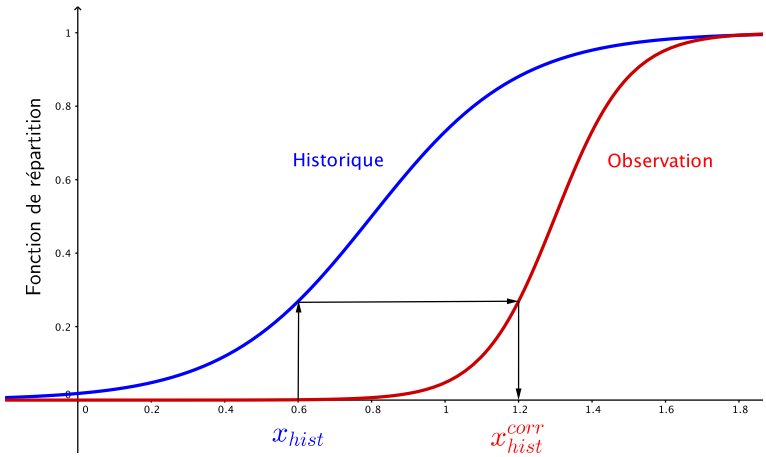


FIGURE VII.2 – Principe de calibration dans la méthode quantile-quantile. On a représenté en rouge la fonction de répartition d'un prédicteur météo observé pour un canton, et en bleu la fonction de répartition des données historiques. On corrige chaque x_{hist} de la base historique, par la valeur du prédicteur qui donne la même probabilité dans la base des observations. Sur le schéma, il s'agit de x_{hist}^{corr}.

On réalise donc une régression non-linéaire (souvent une sigmoïde) des deux fonctions de répartition empirique et on calcule la fonction réciproque de celle des observations pour obtenir la donnée corrigée (x_{hist}^{corr} à la Figure VII.2).

En résumé

Le tableau ci-dessous compare l'ajustement linéaire à la méthode de calibration par quantile-quantile (empirique ou non).

Méthode	Description	Commentaires	
Ajustement linéaire	Ajuste les projections climatiques avec des valeurs de correction basées sur les différences entre valeurs moyennes observées et simulations historiques	Correction identique pour tous les événements d'un même canton	Fréquence et intensités non corrigées séparément
Approche quantile quantile	Fait correspondre la fonction de répartition des simulations historiques-simulées à celle des observations	Evénements ajustés non linéairement	Fréquence et intensités corrigées séparément

2.2 Comparaison des méthodes de calibration

Nous allons comparer les résultats de calibration pour (1) l'ajustement linéaire, (2) l'approche quantile-quantile par lois connues, et (3) l'approche quantile-quantile empirique.

Par simplicité, nous nous concentrons sur le mois de juillet. Nous regardons donc les corrections apportées à la température moyenne et au cumul de précipitations en juillet. De plus, les illustrations suivantes concernent le canton de l'Alabama dont le FIPS code est 1001.

Résultat de l'ajustement linéaire

L'ajustement linéaire est très simple à utiliser et donne de bons résultats (Figure VII.3). Les séries temporelles futures de température ont seulement subit une translation vers le bas quand la moyenne des observations est plus faible que celle historique, et vers le haut dans le cas contraire (Figure VII.3, bas). La tendance est clairement à l'augmentation des températures de juillet pour la plupart des cantons des États-Unis. Malgré les valeurs bien plus élevées prises par les températures simulées du scénario RCP 8.5, la correction de calibration est identique à celle du scénario RCP 4.5. Les deux scénarios, très semblables jusqu'en 2040, divergent le plus souvent à l'horizon 2050. Les précipitations futures ont des valeurs et une variabilité similaires à celles observées et présentent un comportement stationnaire (Figure VII.3, haut). Le facteur de correction peut réduire ou augmenter les amplitudes.

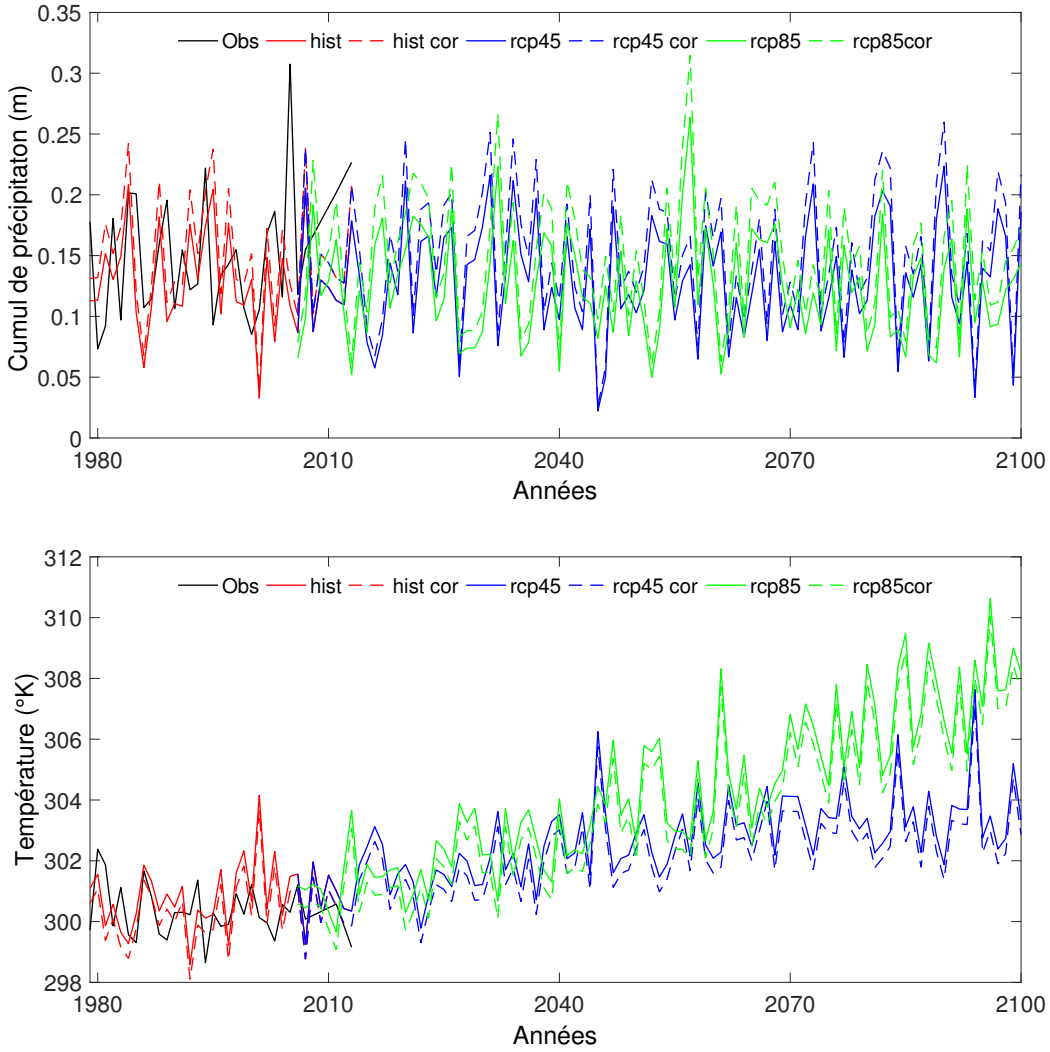


FIGURE VII.3 – Séries temporelles des précipitations (haut) et de la température moyenne (bas) en juillet, pour le canton 1001 de l'Alabama. Les données calibrées par ajustement linéaire (historique-simulées "hist cor", du scénario RCP 4.5 "rcp45 cor", et du scénario RCP 8.5 "rcp85 cor") dont les courbes sont en pointillés, sont à comparer aux données brutes (courbes pleines).

Résultat de l'approche quantile-quantile avec des lois connues

La Figure VII.4 illustre les résultats de la méthode quantile-quantile utilisant la loi normale pour les températures. Les résultats confirment que cette loi est bien adaptée. Les corrections pour les précipitations, très similaires à celles obtenues par la méthode d'ajustement linéaire, ne sont pas montrées ici. Les limites de cette méthode sont assez claires : lorsqu'une donnée simulée prend une valeur très éloignée d'une loi normale/gamma illustrant la distribution des observations, les lois classiques (qui n'ont pas des queues de distribution très longues) ne sont plus adaptées pour retourner une valeur calibrée correcte. On observe ce problème pour les températures puisque les valeurs simulées sortent largement de l'ensemble des valeurs atteintes dans les observations.

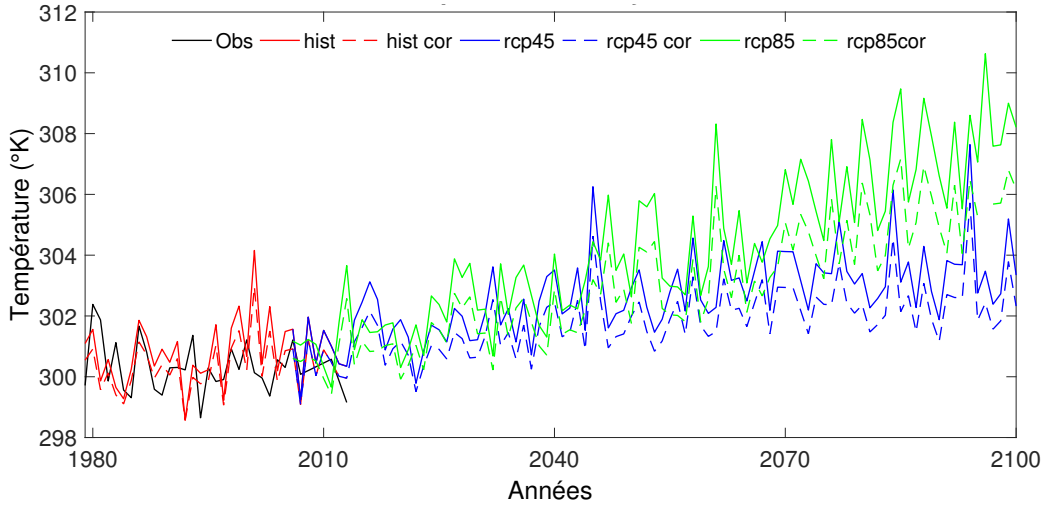


FIGURE VII.4 – Séries temporelles de la température moyenne en juillet, pour le canton 1001 de l'Alabama. Les données corrigées (historique-simulée "hist cor", du scénario RCP 4.5 "rcp45 cor", et du scénario RCP 8.5 "rcp85 cor") par la méthode quantile-quantile à l'aide de la loi normale (courbes en pointillés) sont à comparer aux données brutes (courbes pleines).

On observe cependant des différences entre les séries temporelles calibrées par cette méthode et celles issues de l'ajustement linéaire (Figure VII.3) : les températures calibrées sont plus éloignées des données d'origine avec la méthode quantile-quantile donc la correction est plus forte. L'amplitude des données simulées est conservée. De plus, les températures corrigées par quantile-quantile prennent des valeurs moins grandes (par exemple entre 2080 et 2100) que par la méthode d'ajustement linéaire.
La Figure VII.5 montre comment la fonction de répartition des données historiques simulées (courbe rouge pleine) a été corrigée (courbe rouge en pointillés) pour se rapprocher de la courbe des observations (en noir). Ainsi, que ce soit en terme de distribution ou en terme de série temporelle, cette méthode semble aussi donner des résultats satisfaisants.

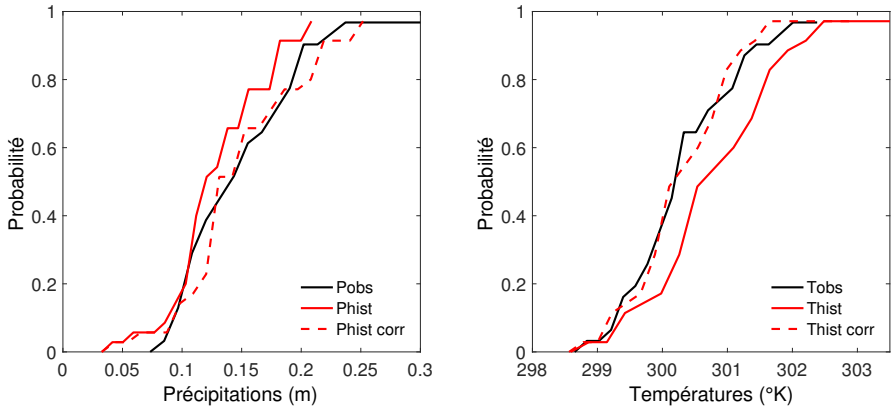


FIGURE VII.5 – Fonction de répartition de Pjuillet (gauche) et Tjuillet (droite) pour les données observées "Obs", historiques-simulées "hist", et historiques-simulées corrigées "hist cor" par la méthode de calibration quantile-quantile de lois connues. Il s'agit des données du canton 1001 de l'Alabama et des données des années 1979-2013.

Résultat de l'approche quantile-quantile empirique

La difficulté de cette méthode repose sur la bonne estimation des fonctions de répartition empiriques, qui fait souvent appel à des régressions polynomiales ou sigmoïdes. Sans suffisamment de données, ces régressions peuvent être difficiles et donner des résultats très éloignés de ce qui est attendu, en dehors des valeurs d'apprentissage. La difficulté provient des valeurs simulées inférieures (respectivement supérieures) au minimum (respectivement maximum) des valeurs observées. Ici, 35 données (années 1979-2013) sont disponibles pour estimer chaque fonction de répartition ce qui semble être suffisant.

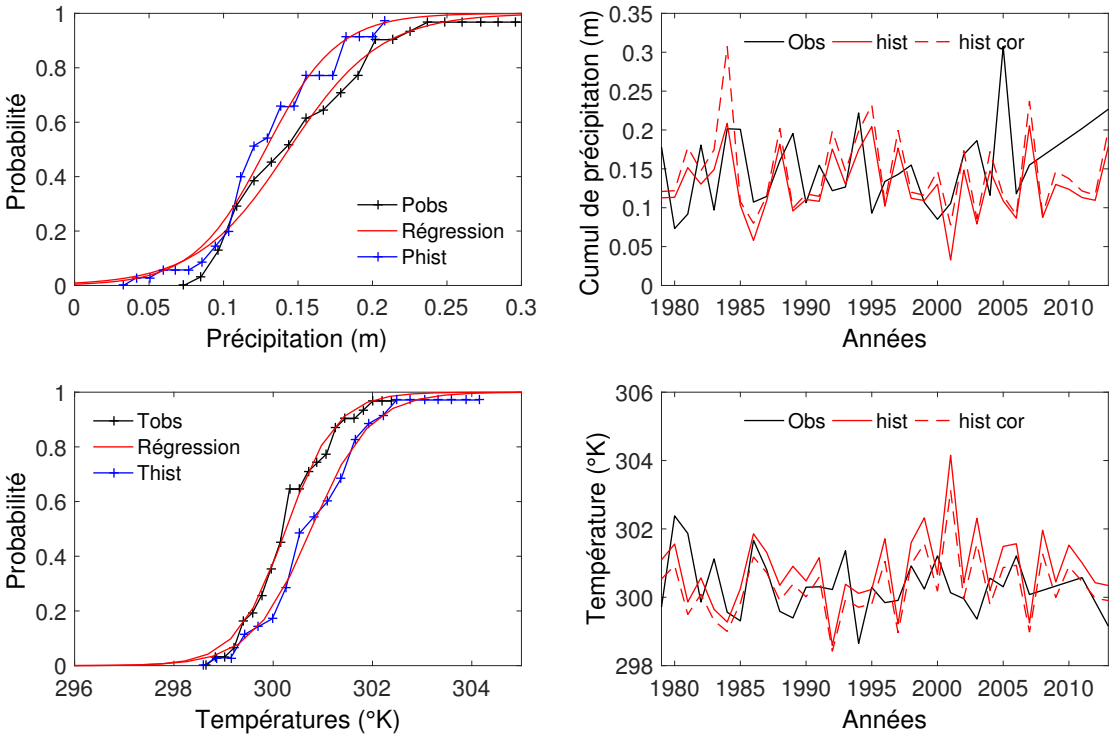


FIGURE VII.6 – Gauche : fonctions de répartition empiriques des données observées (Pobs en haut et Tobs en bas) et historiques simulées (Phist en haut et Thist en bas) ainsi que la courbe de leur régression (en rouge). Les croix représentent la probabilité d'obtenir des valeurs équitablement éloignées.
Droite : séries temporelles des observations "Obs", de l'historique-simulée "hist", et de l'historique-simulé corrigé "hist cor" par la méthode quantile-quantile empirique : Pjuillet en haut et Tjuillet en bas.

La Figure VII.6 illustre les fonctions de répartition empiriques des données observées et historiques-simulées, ainsi que la courbe de leur régression (en rouge). Une régression sigmoïde $\left(y = \frac{1}{1+c_1 e^{-xc_2}}\right)$ a été utilisée pour l'estimation des probabilités empiriques. On observe un léger biais entre les fonctions de répartition observées et historiques-simulées qu'il est nécessaire de corriger.

Les résultats sont très semblables à la méthode quantile-quantile utilisant les lois

Gamma et normale. D'une complexité plus grande, la méthode quantile-quantile empirique permet pourtant (si une bonne régression des fonctions de répartition est possible) de fournir plus facilement une valeur corrigée, même si celle-ci déborde de l'ensemble des valeurs possibles observées. Par soucis de concision, la méthode quantile-quantile avec des lois connues n'est pas dessinée vu qu'elle donne des résultats très semblables à celle empirique.

En résumé

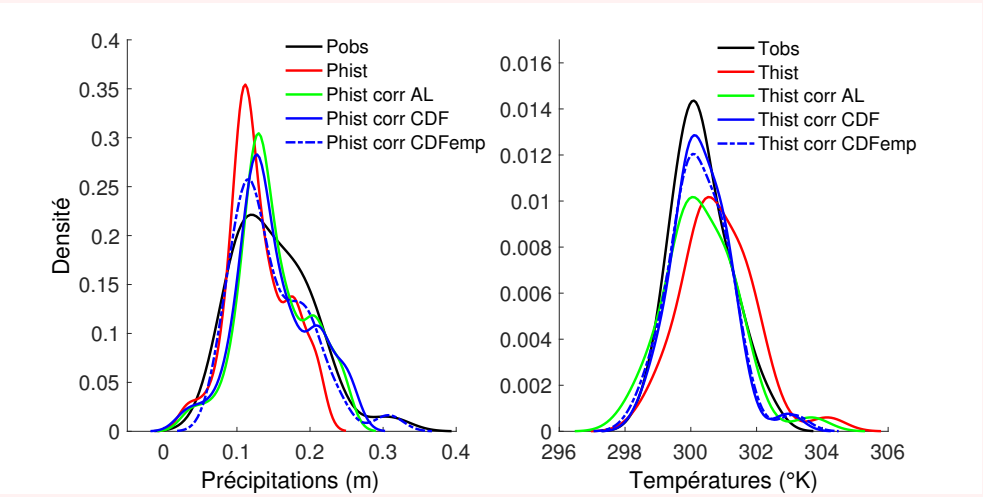


FIGURE VII.7 – Gauche : densités de probabilité des précipitations observées de juillet "Pobs", des simulations historiques "Phist", et des trois calibrations : par ajustement linéaire "Phist cor AL", par approche quantile-quantile de lois connues "Phist cor CDF", et par approche quantile-quantile empirique "Phist cor CDFemp".
Droite : densités de probabilité des températures observées de juillet "Tobs", des simulations historiques "Thist", et des trois calibrations : par ajustement linéaire "Thist cor LS", par approche quantile-quantile de lois connues "Thist cor CDF", et par approche quantile-quantile empirique "Thist cor CDFemp". Il s'agit des données du canton 1001 de l'Alabama.

Pour comparer les trois méthodes de calibration utilisées, on a superposé à la Figure VII.7 les densités des observations et des trois calibrations selon la méthode utilisée. Les précipitations observées présentent davantage de grandes valeurs (>25cm) que celles simulées. Les méthodes de calibration corrigent bien ces différences, surtout pour les méthodes quantile-quantiles. A l'inverse, les températures observées ne présentent pas de valeurs supérieures à 304^oK, ce qui est pourtant le cas pour les températures simulées : les deux méthodes quantile-quantiles sont plus efficaces que l'ajustement linéaire pour réduire ces différences. Que cela soit pour les précipitations ou la température, la méthode d'ajustement linéaire donne de moins bons résultats que celles quantile-quantiles : les courbes de densités en vert sont plus éloignées que celles en bleues de la densité observée (courbe noire). Les deux méthodes quantile-quantiles sont très proches : l'utilisation de lois connues semble former de nouvelles densités moins dispersées.

Face à ces conclusions, nous décidons de poursuivre la suite de l'étude en utilisant la calibration par méthode quantile-quantile empirique.

2.3 Obtention des anomalies météo pour les données climatiques

La Figure VII.8 illustre les anomalies météo futures pour les précipitations et la température de juillet. On peut comparer les évolutions relatives au scénario RCP 8.5 (courbes en pointillées) à celles relatives au scénario RCP 4.5 : pour les années postérieures à 2050, les anomalies de températures prennent de très grandes valeurs pour le scénario RCP 8.5, tandis que celui RCP 4.5 semble se stabiliser. En revanche, comme les précipitations futures sont très proches de celles observées, les anomalies futures de précipitation sont semblables à celles observées. Du point de vue de la modélisation, cela signifie que l'ensemble des possibles pour les précipitations n'a pas changé tandis que celui des températures a beaucoup évolué, ce qui peut être problématique pour la prédiction du modèle d'impact : les nouvelles situations météo proposées ne se rencontrent pas forcément dans la base d'apprentissage. **Aussi, il apparaît plus rigoureux de ne pas considérer les anomalies climatiques qui s'éloignent trop des valeurs sur lesquelles le modèle a été entraîné : les années postérieures à 2060 ne seront donc pas prises en compte.**

3 Résultats du modèle d'impact sur les projections climatiques

3.1 Modèle d'impact utilisé

Comme précisé en introduction de ce chapitre, nous utilisons le modèle linéaire à effets mixtes construit au Chapitre Chapitre V page 153 avec six entrées météorologiques simples : Tmai, Tjuin, Tjuillet, Taoût, Pjuillet et Paoût. Le Chapitre V a montré que ce modèle donnait des résultats satisfaisants en mode monitoring sur l'ensemble des États-Unis en pouvant expliquer 20% de la variabilité des rendements et jusqu'à 45% sur les états les plus sensibles comme la Virginie. Il s'agit d'un modèle conçu et calibré pour estimer la valeur des anomalies de rendement et non pour estimer la fréquence des rendements extrêmes.

3.2 Évolution temporelle des rendements

On observe que les prévisions des rendements issus des deux scénarios divergent après 2050 pour l'ensemble des états. Cela n'est pas étonnant vu que les données climatiques des deux scénarios, puis leurs anomalies, sont assez proches jusqu'en 2050 (Figure VII.8).

Devant les trop fortes anomalies climatiques après 2060, nous avons choisi d'estimer les anomalies de rendements jusqu'en 2060 seulement. Au delà, le climat est trop différent de celui utilisé pour calibrer le modèle d'impact. Le modèle d'impact statistique peut être utilisé sur des données extrêmes mais uniquement si de tels extrêmes ont pu être rencontrés dans la base de données observées. Sinon, on rentre dans le domaine des extrapolations.

La Figure VII.9 représente l'évolution temporelle estimée des anomalies jusqu'en 2060 pour un canton de l'Alabama (a), Indiana (b), Missouri (c), Caroline du sud (d), Dakota du sud (e), et Virginie (f). Les évolutions liées au scénario RCP 4.5 (traits pleins)

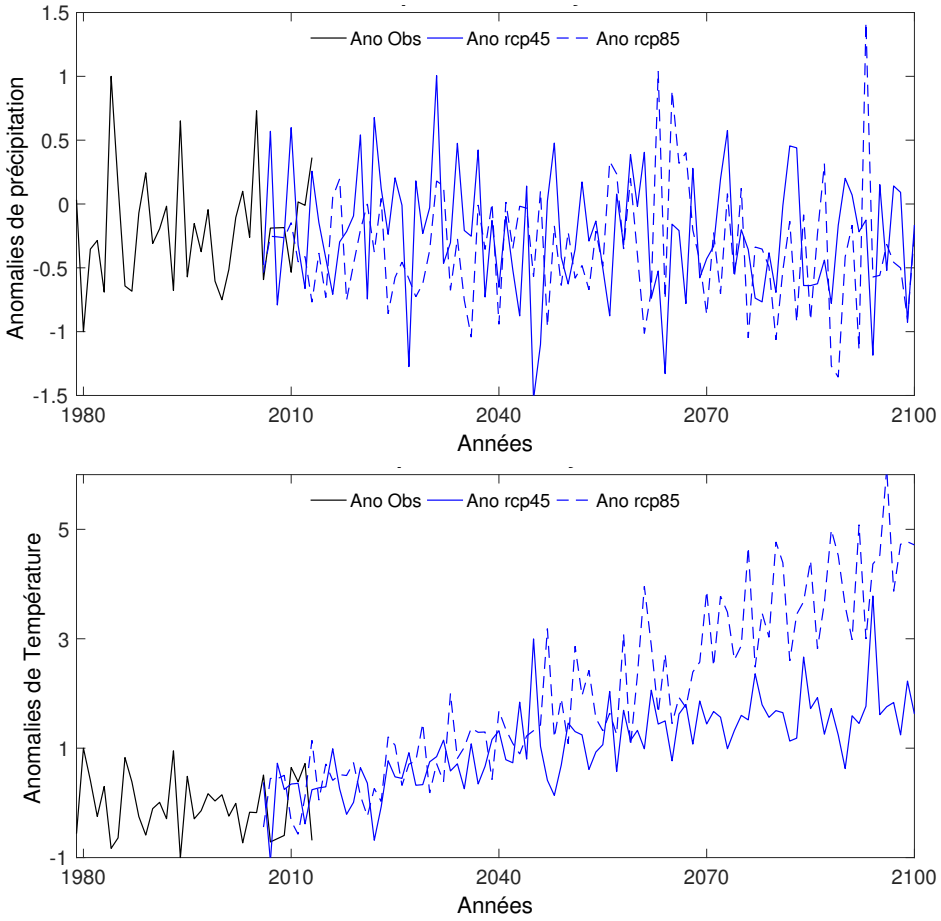


FIGURE VII.8 – Séries temporelles des anomalies obtenues pour les données climatiques : précipitation de juillet pour le canton 1001 de l'Alabama (haut), et température moyenne de juillet pour ce même canton (bas). La méthode de calibration utilisée est l'approche quantile-quantile empirique. La courbe en pointillés fait référence au scénario climatique RCP 8.5, et celles en trait plein au scénario RCP 4.5.

sont à comparer à celles du scénario RCP 8.5 (traits pointillés). On rappelle qu'une anomalie de -0.5 signifie que le rendement de maïs est 50% inférieur à la tendance annuelle. On observe des disparités d'évolution des anomalies de rendement entre les états des États-Unis même si tous ont une tendance négative (Figure VII.9, courbes de régression linéaire rouges). On remarque aussi que pour certains cantons, la courbe du scénario RCP 4.5 reste proche de celle du RCP 8.5 (exemples c) et e)) tandis que pour d'autre, elle s'écarte largement, et assez tôt (exemples d) et a)). De plus, les amplitudes des anomalies de rendement sont très diverses : très fortes pour (a) et (f) et faibles pour (b). Le changement climatique semble donc affecter à la fois les moyennes mais aussi les variances des anomalies de rendement.

Pour avoir une vision globale de la tendance future des anomalies de rendement, une régression linéaire des séries temporelles futures des anomalies de rendement (tous cantons confondus) a été réalisée par état. Le Tableau VII.1 nous renseigne, pour chaque état de l'étude, sur l'évolution des anomalies de rendement de 2010 à 2060 pour les deux scénarios climatiques RCP 4.5 et RCP 8.5 selon cette régression linéaire. Les deux dernières colonnes renseignent sur la rapidité de l'évolution puisqu'il s'agit du pourcentage de variation de l'anomalie des rendements par décade. On observe une grande

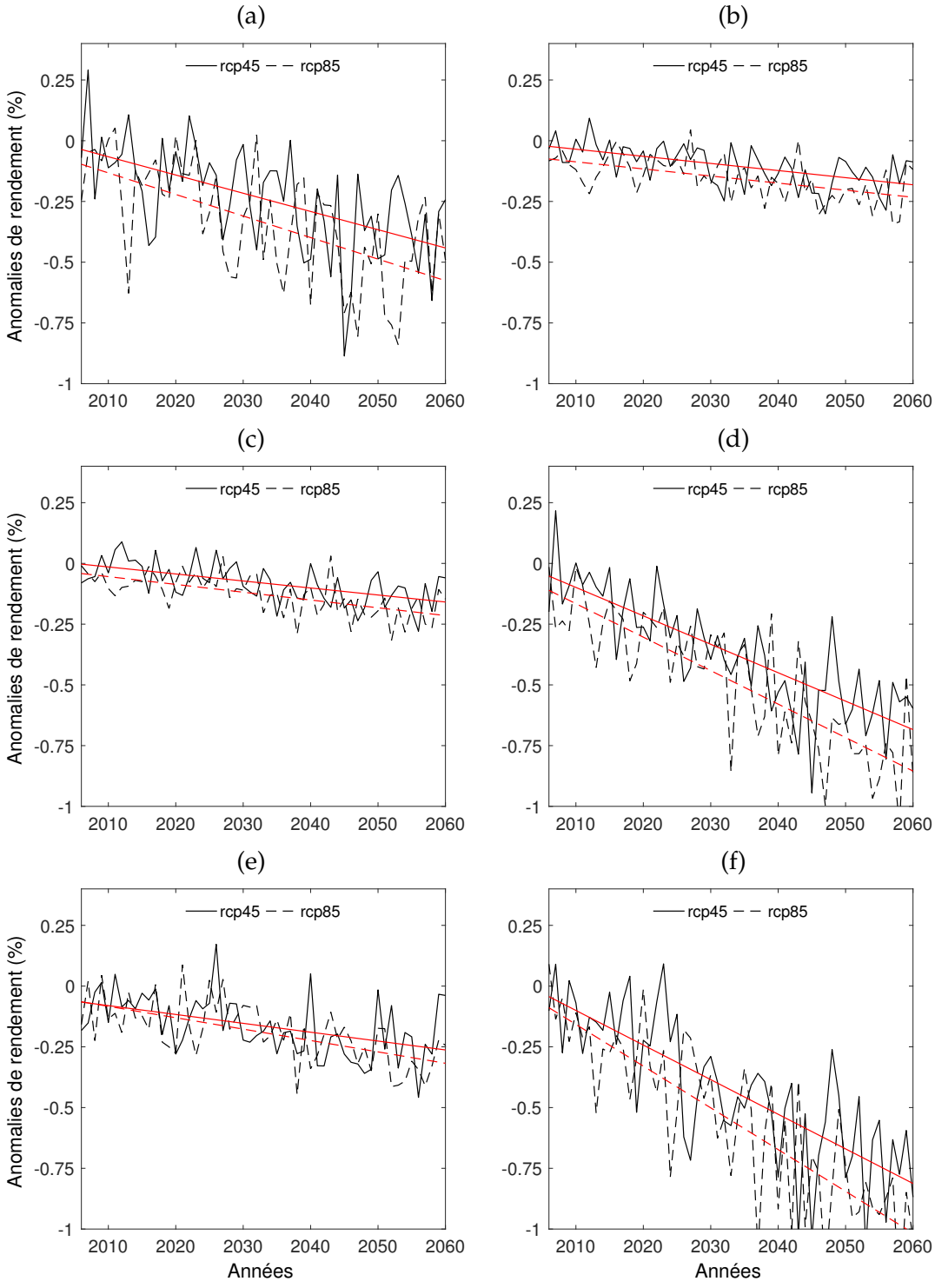


FIGURE VII.9 – Évolution temporelle estimée jusqu'en 2060 pour un canton de l'Alabama (a), Indiana (b), Missouri (c), Caroline du sud (d), Dakota du sud (e), et Virginie (f). Ces états sont éloignés les uns des autres. Les courbes pleines font référence au scénario RCP 4.5 et celles en pointillé au scénario RCP 8.5. Ces estimations sont obtenues avec une calibration par méthode quantile-quantile empirique. Les droites rouges sont les régressions linéaires des séries temporelles pour chaque scénario.

diversité dans les évolutions futures. Certains états qui ont des valeurs comparables en 2010 (comme le Minnesota et la Caroline du Sud) n'évoluent plus de la même façon en 2060 : le Minnesota présente une anomalie de rendement de seulement -6% en 2060 selon le scénario RCP 4.5 alors que la Caroline du Sud subit -46%.

État	En 2010		En 2060		% de variation par décade	
	RCP 4.5	RCP 8.5	RCP 4.5	RCP 8.5	RCP 4.5	RCP 8.5
Alabama	-4	-10	-41	-51	-7.3	-8.1
Caroline du Nord	-6	-12	-50	-60	-8.6	-9.7
Caroline du Sud	-2	-9	-46	-58	-8.8	-9.7
Dakota du Nord	-3	-1	-9	-14	-1.0	-2.6
Dakota du Sud	-3	-3	-15	-20	-2.4	-3.3
Delaware	-10	-11	-42	-52	-6.3	-8.3
Géorgie	-3	-9	-42	-51	-7.8	-8.5
Illinois	-1	-9	-22	-31	-4.1	-4.4
Indiana	-1	-8	-23	-32	-4.2	-4.9
Iowa	-3	-5	-15	-19	-2.3	-2.9
Kansas	2	-6	-25	-29	-5.2	-4.6
Kentucky	-1	-8	-25	-37	-4.8	-5.6
Louisiane	-10	-13	-39	-46	-5.8	-6.6
Maryland	-11	-12	-50	-64	-7.8	-10.4
Michigan	-8	-7	-18	-21	-2.1	-2.9
Minnesota	-2	0	-6	-10	-0.8	-1.9
Mississipi	-3	-7	-25	-33	-4.5	-5.1
Missouri	0	-11	-27	-35	-5.3	-4.7
New Jersey	-11	-10	-49	-64	-7.5	-10.7
New York	-4	-3	-14	-16	-1.9	-2.5
Ohio	-5	-10	-30	-39	-4.9	-5.7
Oklahoma	-2	-11	-31	-33	-5.7	-4.4
Pennsylvanie	-9	-8	-35	-48	-5.2	-8.0
Tennessee	-1	-10	-30	-41	-5.7	-6.2
Virginie	-11	-15	-59	-73	-9.6	-11.7
Virginie occid.	-16	-18	-46	-58	-6.0	-7.9
Wisconsin	-3	-3	-10	-14	-1.3	-2.1

TABLEAU VII.1 – Pour les deux scénarios RCP 4.5 et RCP 8.5, et pour chaque état de l'étude, on a réalisé une régression linéaire des séries temporelles futures des anomalies de rendement (tous cantons confondus). Les colonnes 2 et 3 donnent en % la valeur de l'anomalie de rendement en 2010 et les colonnes 4 et 5, la valeur en 2060 selon cette régression. Les deux dernières colonnes fournissent le pourcentage de variation de l'anomalie par décade. 70% du maïs a été produit par les états en caractères gras en 2016.

Les états pour lesquels les anomalies de rendement vont chuter le plus rapidement sont la Caroline du Nord/Sud, le Maryland, le New Jersey, la Virginie et la Virginie occidentale : ces états présentent en 2060 des anomalies inférieures à -46% sous le scénario RCP 4.5 et -58% sous le scénario RCP 8.5. Ce ne sont pas les états les plus productifs en maïs, mais ce sont ceux qui seront les plus rapidement touchés, demandant une rapide mise en place de stratégies d'adaptation.

À l'inverse, des états du nord tels que le Dakota du Nord, le Minnesota, New York ou le Wisconsin seront beaucoup moins touchés avec moins de -1,9%/décade de variation des anomalies sous le scénario RCP 4.5 et moins de -2,6%/décade sous le scénario RCP 8.5. Les états les plus productifs (ceux de la Corn Belt) enregistrent une anomalie en 2060 de -23% en moyenne sous le scénario RCP 4.5, et -30% sous le scénario RCP 8.5.

3.3 Distribution spatiale de l'évolution des rendements

La Figure VII.10 illustre les anomalies de rendement moyennes par décade, à l'échelle des cantons. De 2020 à 2030 (en haut à gauche), on observe des anomalies négatives mais faibles (<-0.15) pour la plupart des cantons, voire très proche de zéro pour certains. La Corn-Belt et la côte est présentent les plus faibles anomalies.

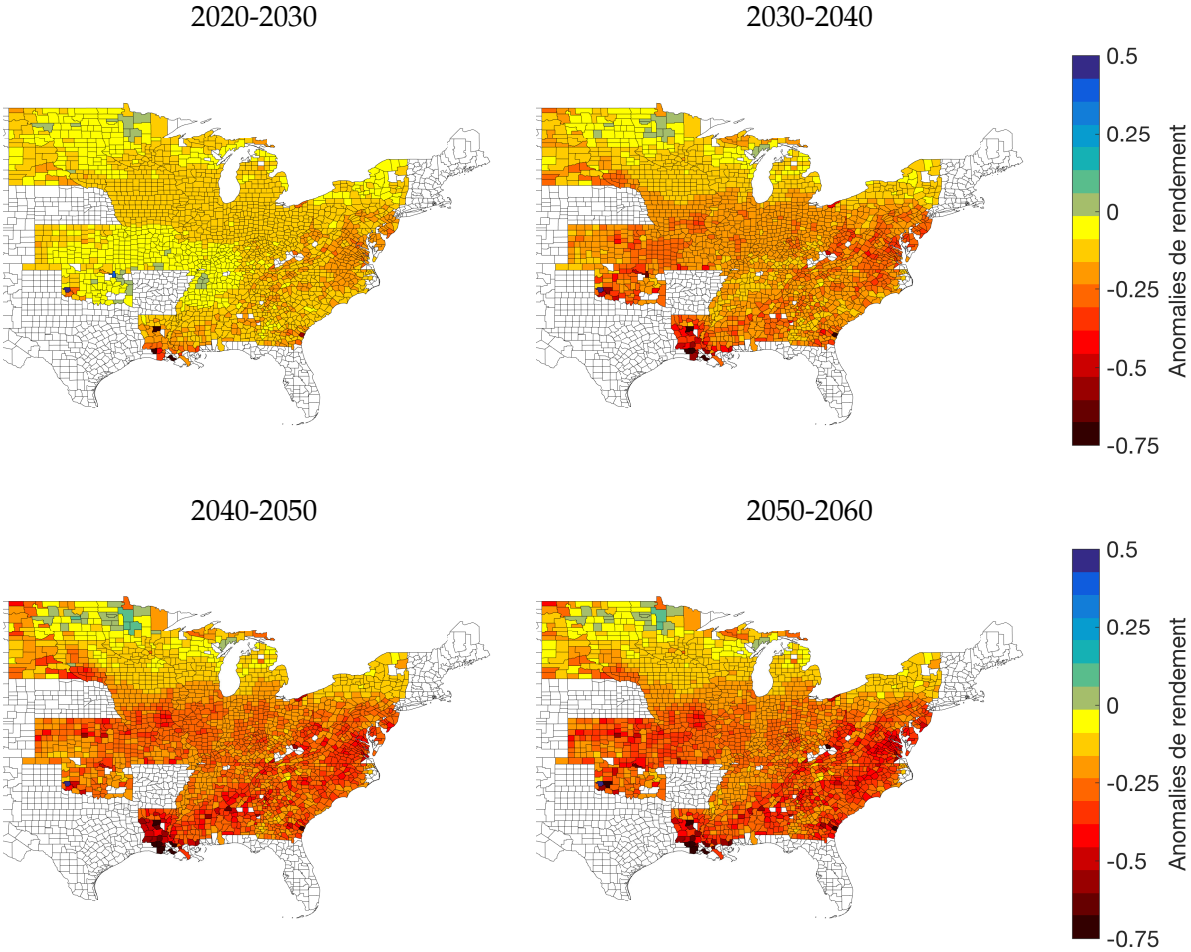


FIGURE VII.10 – Carte des anomalies de rendement moyennes (en %) par décade 2020-2030, 2030-2040, 2040-2050 et 2050-2060 à l'échelle du canton, pour le scénario RCP 4.5.

Concernant la décade 2030-2040, les zones qui avaient déjà de faibles anomalies voient leurs valeurs diminuer, surtout pour les états situés au sud ouest de la zone étudiée.

Les deux scénarios climatiques fournissent des valeurs de températures et de précipitations proches jusqu'en 2060, ce qui conduit à une évolution des rendements similaires. Les cartes pour le scénario RCP 8.5 ne sont donc pas montrées ici.

Entre 2040 et 2050, la majorité des cantons situés au nord garde des valeurs supérieures à -0.25 (diminution des rendements de 25% par rapport à la tendance) mais

les autres ont une anomalie moyenne inférieure à -0.35. On retrouve des résultats similaires pour la période 2050-2060. Il ressort de ces prévisions que, sous l'hypothèse d'une pratique agricole constante, le climat estimé en 2050 va toucher en priorité les cantons situés au sud des grands lacs, avec de fortes pertes attendues sur la côte est, le golfe du Mexique et l'ouest de la Corn-Belt. Seuls quelques cantons au nord des États-Unis semblent tirer profit de ce changement climatique avec des anomalies de rendement positives mais ces valeurs ne sont pas significatives.

4 Discussion

4.1 Comparaison avec d'autres études

Pour savoir si les modèles statistiques étaient efficaces pour saisir l'impact du changement climatique sur les cultures, [Lobe10b] ont comparé les prévisions de trois types de modèles statistiques (séries temporelles, séries longitudinales [2] et modèles transversaux [3]) à la sortie du modèle CERES-Maize auparavant simulé pour prendre en compte un réchauffement de 2^oC et une réduction de 20% des précipitations (même si celui ci ne représente qu'une partie des nombreux processus qui affectent les rendements). Ils ont montré que les modèles statistiques reposant sur des informations provenant de plusieurs sites (ce qui est notre cas), à savoir les modèles longitudinaux ou transversaux, étaient meilleurs pour prédire les réponses aux changements de température et de précipitations.

[Lobe17] ont comparé les sensibilités du maïs aux changements de température, de précipitations, de dioxyde de carbone (CO_2) et d'ozone sous différents scénarios climatiques en utilisant un modèle mécaniste et un modèle statistique. Ils n'ont pas trouvé de différences systématiques entre les sensibilités prédites par le modèle mécaniste et celles prédites par le modèle statistique concernant un scénario de réchauffement de 2^oC. Pour les précipitations, les comparaisons sont moins robustes, entre-autre parce que les changements de précipitations sont rarement le facteur dominant pour prédire les impacts. De plus, les modèles mécanistes tendent à inclure les effets des augmentations de CO_2 qui accompagnent le réchauffement, contrairement aux modèles statistiques. Les auteurs soulignent le besoin d'intégrer les effets du CO_2 dans les études statistiques, et précisent qu'à court terme, les deux approches bien menées sont susceptibles de fournir des estimations similaires des impacts du réchauffement climatique ; les modèles statistiques nécessitant généralement moins de ressources pour produire des estimations robustes.
De leur côté, [Urba15] ont mis au point un modèle statistique pour le maïs aux États-Unis (Iowa, Indiana, et Illinois) qui comprenait un terme pour la demande de pression de vapeur, qui varie avec la température et l'humidité. Ils ont montré que les estimations des anomalies de rendement qui excluent les effets du CO_2 sont bien plus négatives que celles qui les prennent en compte : par exemple, ils montrent qu'en 2050 aux États-Unis, les anomalies de rendement de maïs seront de -10% au lieu de -20% sous le scénario RCP 4.5 si l'impact du CO_2 est pris en compte dans le modèle. Ils soulignent que (1) les modèles statistiques incluant des interactions chaleur/humidité, conduisent

2. Un ensemble de données longitudinales possède une dimension à la fois transversale et temporelle, et (souvent) toutes les unités de section sont observées pendant toute la période.

3. Les données transversales sont observées à un moment donné pour plusieurs cantons. L'intérêt réside dans la modélisation de la distinction des cantons, et l'hétérogénéité entre cantons.

à une amélioration significative du modèle et à une variabilité de rendement significativement plus élevée que lors de la prise en compte de chaque facteur indépendamment ; (2) l'inclusion des avantages du CO_2 élevé réduit de manière significative les impacts, en particulier pour la variabilité du rendement ; et (3) les dommages nets causés par le changement climatique et le CO_2 deviennent plus importants pour le scénario RCP 8.5 dans la seconde moitié du XXIe siècle (ce que nous observons aussi).
En plus du CO_2, la réorganisation des cultures peut aussi être prise en compte : [Leng17] ont montré que le rendement en maïs devrait diminuer de 20 à 40% d'ici 2050 si l'on tient compte des changements dans la distribution spatiale des cultures, soit 6 à 12% de moins que les estimations avec un modèle de culture fixe.

À l'aide du modèle mécaniste DSSAT, [Robe15] a lui aussi estimé les rendements pour ces 50 prochaines années. Même si les changements qu'il propose sur la réduction des rendements sont plus faibles, on retrouve les mêmes structures spatiales avec une plus forte baisse pour les états au sud des grands lacs, et une évolution optimiste pour ceux aux mêmes latitudes. Ces observations sont aussi faites dans [Sout00] en utilisant des données agricoles simulées par le modèle mécaniste CERES-Maize.
Comme [Robe15], [Mull14] ont utilisé le modèle mécaniste DSSAT et non un modèle statistique. [Mull14] ont quantifié les changements relatifs de la production du maïs non-irrigué projetés par les modèles mécanistes DSSAT et LPJmL (un modèle mécaniste mondial, gérant la croissance des plantes de façon dynamique) pour la base de données climatiques issue du modèle IPSL-CM5A-LR pour le scénario d'émission RCP 8.5 [4]. Leurs résultats avec le modèle DSSAT semblent proches des nôtres alors que ceux avec le modèle mécaniste LPJmL donnent des changements relatifs quasiment deux fois moins importants. Pour les deux modèles, les états au centre des États-Unis sont les plus touchés, ce que nous observons aussi. [Adam90] et [Takl13] ont eux aussi des résultats spatiaux qui dépendent du modèle mécaniste utilisé. Comme pour la plupart des études sur le sujet, les résultats montrent que l'impact dépend de la région, avec des états qui subissent bien plus le changement climatique que d'autres [Kim15].
De façon similaire, [Schl09] ont estimé le lien entre la météo et les rendements de maïs aux États-Unis en utilisant le modèle climatique Hadley III [5]. Ils ont montré que les rendements moyens devraient diminuer de 20-25% d'ici 2050 sous le scénario B1 (équivalent du RCP 4.5) et diminuer de 30-35% sous le scénario pessimiste A1Fl (équivalent du RCP 8.5).

Enfin, en regardant l'impact du changement climatique sur le cycle hydrologique du maïs, [John15] ont aussi estimé les anomalies de rendement pour les états du centre-est américain à l'aide du modèle CERES-maize et du modèle de circulation générale atmosphérique ECHAM5. Selon eux, le modèle prévoit globalement une augmentation des rendements (ce que l'on n'observe pas) avec cependant, une grande variabilité entre les cantons : des baisses de rendement de l'ordre de -10% à -20% en anomalies sont attendues en Ohio, Illinois, Missouri ainsi qu'au sud du Minnesota, sous le scénario A2 du GIEC de 2007 (un peu moins pessimiste que le scénario RCP 8.5). Les cartes des anomalies de [John15] sont bien plus in-homogènes que les nôtres : ils présentent souvent deux cantons voisins avec pour l'un, des anomalies de +20% et pour l'autre -20% ce qui souligne une forte variance du modèle et donc possiblement un problème

4. Dans ce chapitre, nous avons utilisé la base IPSL-CM5A-MR (Moyenne Résolution).
5. www. metoffice.com/research/hadleycentre/

de calibration ou d'extrapolation.

En résumé

Ces études montrent que les prévisions obtenues sembles cohérentes avec ce qui a été fait dans la littérature (même si cela a été traité le plus souvent à l'aide de modèles mécanistes), y compris en terme de variabilité spatiale. Les modulations sur l'amplitude des projections peuvent être dues (1) à l'utilisation de variables climatiques différentes, (2) à l'utilisation de modèles d'impact différents, (3) au choix des scénarios climatiques, et (4) à la prise en compte ou non d'une évolution dans les pratiques agricoles.

4.2 Retour sur les choix méthodologiques

Tendance à long terme - Nos projections futures sont basées sur l'hypothèse d'une pratique agricole constante à celle qui a donnée lieu à la tendance temporelle des rendements sur l'historique. L'importance des pertes de rendements est aussi à relativiser puisque la modélisation ne prend pas en compte les différentes adaptations que l'Homme va mettre en place d'ici les 10 ou 30 prochaines années, ce qui est peu réaliste. Cependant, ces adaptations peuvent être incluses dans une tendance à long terme, facilement applicable sur nos anomalies de rendement. De plus, nous savons que le changement climatique va faire évoluer la fréquence des maladies et leurs dispersions, ce qui n'est pas pris en compte dans le modèle. Ainsi, les valeurs des projections des anomalies de rendement sont peu susceptibles de représenter précisément l'évolution future des impacts. Plutôt que de les interpréter comme représentant l'ampleur attendue des impacts du changement climatique, il faut les voir comme une tendance à long terme pour une hypothèse donnée. Ces résultats peuvent nous informer sur la répartition spatiale des impacts à venir et nous aider à savoir quelles régions seront plus touchées que les autres, et si ces changements sont importants ou pas.

Scénarios climatiques - Nous avons choisi de tester deux scénarios climatiques, l'un étant très pessimiste (le RCP 8.5), l'autre plutôt optimiste (le RCP 4.5). Le scénario intermédiaire (RCP 6) n'a pas été traité pour ne pas alourdir les comparaisons. Au vu des différences observées dans les estimations entre RCP 4.5 et RCP 8.5, il ressort que la prise en compte du scénario intermédiaire aurait été superflue.

Période étudiée - Plusieurs études, comme celle de [Urba15], montrent aussi que les deux scénarios se distinguent davantage entre 2050 et 2100 mais nous avons procédé à une restriction de l'application du modèle d'impact aux années 2010-2060 puisque les estimations climatiques de l'IPSL fournissent un climat trop éloigné de celui qui a servit à calibrer le modèle statistique. Il est donc plus rigoureux de ne garder que des entrées relativement similaires aux données d'apprentissage.

Un modèle unique - Certaines études testent leur modèle (statistique ou mécaniste) sur plusieurs simulations climatiques. Ici, nous faisons le choix de n'en tester qu'un seul, le modèle à effets mixtes développé au Chapitre V. L'accent a été mis sur la quantification de la qualité de généralisation de ce modèle au Chapitre V. L'application sur plusieurs modèles climatiques nous renseignerait sur la dispersion des modèles (et

non sur l'incertitude autour des estimations), mais cela n'a pas été traité ici.

Nous avons enfin choisi de comparer trois méthodes de calibrations différentes. Ces méthodes sont toutes univariée (chaque variable climatique est corrigée séparément). Or, des études ont montré que la correction quantile-quantile univariée peut affecter les indices climatiques et les modèles d'impact qui nécessitent plus d'une variable climatique en entrée, en raison du défaut de cohérence entre les variables climatiques. Récemment, des algorithmes de correction de biais multivariés ont été proposés afin de réduire ces inconvénients [Vrac15]. Ces nouvelles méthodes pourraient être testées dans une prochaine étude.

4.3 Quelles stratégies d'adaptation et d'atténuation ?

L'évolution des anomalies de rendements agricoles que l'on observe est principalement due à une augmentation des températures puisque les estimations des précipitations sont stables. Face à de telles évolutions potentielles du rendement, des pistes de solutions existent pour essayer de contrecarrer, de réduire ou de s'adapter aux changement climatique [Denh14]. Elles visent principalement à promouvoir une agriculture efficiente en eau. L'adaptation est définie comme "les initiatives et les mesures visant à réduire la vulnérabilité des systèmes naturels et humains face aux effets réels ou prévus du changement climatique" [GIEC07]. Des exemples d'adaptation de l'agriculture aux impacts du changement climatique peuvent inclure des ajustements des dates de plantation, des variétés de cultures, des systèmes de drainage et des régimes de gestion des terres pour maintenir les rendements et la fertilité du sol. Dans le Midwest américain, on recommande par exemple des pratiques telles que la minimisation du travail du sol et l'utilisation de cultures de couverture pour protéger les sols de l'érosion [Lal11]. Pour répondre à la diminution du confort hydrique et à l'augmentation des risques d'échaudage, il est possible de jouer sur la date des semis ou le choix d'une variété précoce. Le maïs peut aussi être intégré au sein d'une rotation céréalière ou relocalisé géographiquement. Diversifier les rotations de cultures, intégrer le bétail aux systèmes de production végétale, améliorer la qualité des sols, minimiser les flux de nutriments et de pesticides hors exploitation et autres pratiques généralement associées à l'agriculture durable augmentent également la résilience du système agricole aux impacts du changement climatique sur la productivité [East10]. Les stratégies d'adaptation sont connues et ont des avantages tangibles au niveau des exploitants agricoles [Arbu15].
Le maintient du rendement peut aussi passer par l'utilisation de nouvelles variétés et la mobilisation de davantage d'eau pour l'irrigation, sous réserve de disponibilité. On peut aussi diversifier l'assolement face à la contrainte hydrique et réserver l'irrigation aux cultures rémunératrices. L'adaptation comprend aussi l'abandon de la production de maïs dans certaines régions devenues trop contraignantes pour la plante, et la réorientation vers des cultures économes en eau ou des usages non agricoles. Ces adaptations touchent aussi les politiques agricoles régionales qui devront proposer des solutions pour gérer les ressources naturelles de manière durable et optimiser le stockage en eau, mais aussi l'enseignement des nouvelles façons de produire aux futurs exploitants.

L'adaptation et l'atténuation ("mitigation" en anglais) du changement climatique en l'agriculture sont des concepts proches mais différents. Alors que l'adaptation fait

depuis longtemps partie intégrante de l'agriculture, l'atténuation du changement climatique est relativement nouvelle. L'atténuation est définie comme "le changement technologique et la substitution qui réduisent les intrants de ressources et les émissions par unité de production" [GIEC07], et donc principalement axée sur la réduction des émissions de gaz à effet de serre et/ou l'augmentation de la séquestration et du stockage du carbone. Les avantages potentiels des mesures d'atténuation sont incertains [Walt12]. En agriculture, les principales stratégies d'atténuation ont été axées sur des réponses collectives, comme les taxes sur les émissions. L'atténuation en agriculture se justifie sous l'hypothèse que la production humaine de gaz à effet de serre est le principal moteur du changement climatique. Ainsi, les attitudes et la volonté de soutenir les mesures d'atténuation sont influencées par les croyances sur le changement climatique et le rôle des activités humaines.

Dans une étude récente, [Arbu15] ont montré que les agriculteurs de l'Iowa, semblent être prêts à s'adapter aux conditions climatiques changeantes. Cependant, les agriculteurs qui croient que le changement climatique se produit et est principalement anthropiques sont plus susceptibles de favoriser les mesures gouvernementales sur les gaz à effet de serre, mais ils ne représentent qu'une petite fraction des agriculteurs de l'étude [Arbu15]. De plus, plus de la moitié ne pensent pas que le changement climatique se produit, ou sont incertains quant à son existence, ou encore croient que cela est principalement dû à des causes naturelles.

5 Conclusion sur ce chapitre

Les modèles statistiques représentent un outil très utile quoique imparfait pour estimer l'impact du changement climatique. Des études antérieures ont projeté que le changement climatique aura un impact négatif sur les rendements moyens du maïs aux États-Unis, tout en augmentant la variabilité du rendement. Nous observons les mêmes évolutions.

La base de donnée CMIP5 le l'IPSL suggère une fréquence plus élevée des années pendant lesquelles les cultures seront confrontées à de fortes chaleurs et sous des conditions limitées en eau. L'évolution des rendements ces 50 prochaines années provient ainsi de l'impact important de l'augmentation des températures, sans augmentation des précipitations. La comparaison de plusieurs méthodes de calibration a montré que l'approche par quantile quantile empirique était bien adaptée à notre étude et fournissait de bons résultats. Le modèle statistique utilisé dans ce chapitre souligne que les états qui seront les plus touchés sont localisés au sud des grands lacs, en particulier ceux de la côté est et du sud des États-Unis. Pour ces états, on s'attend, s'il n'y a pas de stratégie d'adaptation mise en place, à enregistrer des anomalies de rendement de l'ordre de -40% sous le scénario RCP 4.5 et de -50% sous le scénario RCP 8.5. Les états du nord seront peu touchés. Les différences entre les estimations futures des anomalies de rendement sous les deux scénarios climatiques étudiés sont souvent plus faibles que la variabilité inter-annuelle. Cependant, les risques semblent s'accroitre, au delà de 2060 où le changement climatique est encore plus prononcé d'un scénario à l'autre. Pour les deux scénarios, il semble urgent de proposer et de réaliser des solutions d'adaptation concrètes pour les 2/3 des états de l'est américain.

Comme souligné par [Lobe10a], il existe trois écueils importants à l'utilisation de modèles statistiques pour simuler les réponses au changement climatique. Le premier concerne le problème d'extrapolation du modèle au-delà de la plage de données d'étalonnage. En particulier, les températures peuvent facilement dépasser l'année la plus chaude dans la base des observations. Pour éviter cet écueil, il est recommandé d'utiliser les modèles statistiques uniquement sur des valeurs d'entrées relativement proches de celles présentes dans l'historique. La deuxième mise en garde implique l'hypothèse de stationnarité. À mesure que les variétés de cultures et les systèmes de gestion changent, la réponse des rendements aux variations météorologiques peut également changer. Comme pour l'extrapolation, l'hypothèse de stationnarité devient plus discutable à mesure que l'horizon temporel des projections s'étend dans le futur. Une troisième mise en garde concerne l'utilisation de modèles qui supposent implicitement qu'aucune adaptation ne se produira. Cependant, même si cette adaptation n'est pas directement prise en compte, l'objectif des projections reste le même : identifier les plus grandes menaces pour l'agriculture si nous ne nous adaptons pas [Lobe11b]. De plus, les projections peuvent être comparées à celles issus de modèles prenant en compte cette adaptation, fournissant ainsi une mesure de l'impact potentiel de l'adaptation.

L'analyse des incertitudes est une question très intéressante que nous traiterons très probablement en perspective de cette thèse. Elle n'a pas été traitée par manque de temps : pour être menée rigoureusement elle implique de rassembler les multiples incertitudes. Les plus grandes incertitudes proviennent des données d'entrées : elles sont limitées concernant les ré-analyses, mais bien plus grandes pour les données climatiques. Une estimation de ces incertitudes est possible en comparant plusieurs modèles climatiques. Une autre source d'incertitude provient de notre modèle statistique (intervalles de confiances pour la valeurs des paramètres par exemple, ou sensibilité à la base d'apprentissage). Enfin, l'évolution de la pratique agricole est peu facilement quantifiable, et sa modélisation est difficile.

Conclusions et perspectives

L'agriculture est un domaine important surtout dans un contexte de changement climatique. Pour estimer les rendements agricoles on dispose de modèles mécanistes et de modèles statistiques. Des études ont montré que les modèles statistiques donnent des résultats aussi bons que les modèles mécanistes, et sont plus simples d'utilisation. Dans cette thèse, la construction de modèles statistiques sert à prédire les rendements de maïs aux États-Unis. On distingue trois modes d'application : les prévisions en cours d'année (i.e. les prévisions saisonnières), les estimations en fin d'année et les prévisions à long terme. Ici un seul modèle permet de gérer ces trois applications, soulignant l'intérêt d'un tel modèle.

La météo sensibilité - De nombreux paramètres influencent le rendement des cultures, que cela soit le type de sol, l'engrais, le type de graines, les maladies, l'apport en eau, l'ensoleillement etc. Dans cette thèse, on a fait le choix de se concentrer sur l'impact de la météo uniquement, en construisant des modèles statistiques simples qui prennent en entrées, essentiellement des données météorologiques. Les interactions et les processus qui déterminent les rendements agricoles lorsque la météo varie sont nombreux et complexes. Cela fait intervenir les variables météorologiques (température, précipitation...), mais aussi les interactions liées au stress hydrique (humidité du sol), les caractéristiques du sol, tout comme un grand nombre d'autres variables environnementales.

Tenir compte des difficultés simultanément - Les principales difficultés proviennent du manque de données agricoles puisque l'on dispose d'un seul échantillon par an. La météorologie n'explique qu'une partie de la variabilité des rendements, et l'on ne dispose pas d'information sur les parties restantes. L'utilisation de modèles simples avec des prédicteurs simples est donc nécessaire. Lors de la construction des modèles statistiques, il a fallu veiller simultanément au risque de sur-apprentissage, en adaptant les méthodes de sélection de modèles aux corrélations spatiales et temporelles de nos données, tout en respectant les principes de parcimonie notamment dans le nombre d'entrées prises en compte, et dans la simplicité des modèles statistiques. Il a aussi fallu veiller à la robustesse des critères de qualité pour pouvoir obtenir des résultats fiables.

Les modèles d'impacts construits - L'estimation des anomalies de rendement par des modèles linéaires à effets mixtes et des réseaux de neurones a montré qu'il était important de prendre en compte les spécificités locales (par l'information de groupe à travers les modèles à effets mixtes) pour la tendance à long-terme des rendements mais aussi pour une meilleure estimation des rendements à l'échelle des États-Unis. Il est aussi ressorti qu'une fois les entrées optimisées, les modèles linéaires, mixtes, et neuronaux donnaient des résultats proches, ce qui prouve qu'une représentation linéaire (ou quasi linéaire, avec les modèles linéaires mixtes) du lien météo/rendement est suffisante.

Les prévisions saisonnières et les estimations en fin d'année - Avec seulement des températures et les précipitations, le meilleur modèle construit permet d'expliquer en fin d'année 20% de la variabilité du rendement, et l'optimisation sur les prédicteurs améliore ce score à 28%. Certains états, comme la Virginie, font partie des mieux prédits avec un score de 64%. En cours d'année, les prévisions saisonnières semblent acceptables à partir du mois de juillet : à travers un modèle linéaire mixte, la météo explique 25% de la variabilité des rendements à l'échelle des États-Unis (58% en Virginie). Ces résultats sont cohérents avec ceux de récentes études sur le sujet. Les régions les moins bien prédites sont celles situées au nord et au sud de la moitié-est des États-Unis : dans ces régions là, le lien entre météo et rendement est faible, soulignant l'importance de facteurs non pris en compte dans la liste de nos prédicteurs potentiels. La thèse propose ainsi une méthode fiable pour mesurer la météo-sensibilité.

Le cas des rendements extrêmement faibles - C'est sans doute le cas le plus préoccupant pour les exploitants, mais ce n'est pourtant pas le plus aisé à traiter en termes statistiques. Les difficultés liées au manque de données s'amplifient et il n'est pas envisageable pour le moment de chercher à estimer la valeur du rendement. Prédire la probabilité d'observer un rendement très faible en fin d'année est un objectif moins ambitieux mais plus réaliste. Devant la faible robustesse des arbres de décisions aux différents ensembles d'apprentissage, nous avons choisi de réaliser ces estimations à l'aide d'un réseau de neurones, comme classificateur. 71% des rendements très faibles sont bien détectés, avec 15% des rendements normaux classés comme très faibles. Ces scores sont plutôt bons face à la simplicité du classificateur (4 entrées). La réponse du réseau peut servir d'indice de sévérité à une perte de rendement extrême.

Les prédicteurs météorologiques - Sur la cinquantaine de prédicteurs météorologiques testés dans cette thèse, il s'est avéré que peu d'entre eux étaient suffisamment informatifs quant au lien météo/rendement pour être sélectionnés comme entrées dans un modèle d'impact à grande échelle. En particuliers, les prédicteurs définis comme des comptages de jours, ou des dates de début ou de fin de période ne semblent pas adaptés à ce genre de modèles. Les prédicteurs plus simples, comme les températures, les précipitations ou les indices de bilan hydrique (SPEI) restent les plus efficaces. L'optimisation des prédicteurs n'améliore que peu les estimations/prévisions par rapport à des prédicteurs simples, ce qui est un bon point pour les prévisions climatiques de rendements. Les prévisions saisonnières profitent davantage de cette optimisation que les estimations en fin d'année. Ce sont les prédicteurs relatifs aux mois d'été qui sont les plus informatifs.

Impact à long terme - La prédiction des anomalies de rendement pour les 50 prochaines années, tout comme l'interprétation des résultats nécessitent une prudence particulière, face (1) aux incertitudes liées aux prévisions climatiques, (2) aux facteurs multiples non pris en compte par nos modèles (facteurs qui seront pourtant influencé par le changement climatique), et (3) à l'absence d'information sur la façon dont les agriculteurs vont pouvoir s'adapter. A l'aide de six variables météorologiques simples (Tmai, Tjuin, Tjuillet, Taoût, Pjuillet et Paoût), notre modèle statistique nous informe sur la rapidité d'évolution du rendement pour les différents états et permet de localiser ceux qui seront le plus touchés. Les états pour lesquels les anomalies de rendement vont chuter le plus rapidement présentent en 2060 des anomalies inférieures à -46% sous le scénario RCP 4.5 et -58% sous le scénario RCP 8.5. À l'inverse, les états du nord seront beaucoup moins touchés avec moins de -1,9% de variation des anomalies par décade sous le scénario RCP 4.5 et moins de -2,6% sous le scénario RCP 8.5.

Perspectives - Les possibilités de poursuites de cette thèse sont nombreuses.

— *Méthodologie :* Les modèles statistiques construits dans cette thèse pourraient prendre en entrées des prédicteurs non-météorologiques comme le type de sol, pour augmenter leur qualité. De plus, l'utilisation des données satellites en combinaison avec les meilleurs prédicteurs sélectionnés dans cette thèse améliorerait sans doute la qualité de prévision des modèles pour les estimations en fin d'année. Enfin, l'étude des rendements extrêmement faibles n'a pas été traitée pour les 50 prochaines années. Construire ces projections pourrait faire l'objet d'une future étude.

— *Applications :* L'étude de faisabilité d'un indice d'assurance à partir des modèles d'impact pourrait trouver sa place dans un secteur qui est amené à se développer. Devant l'importance des échelles et de la géographie lorsqu'on souhaite déterminer les rendements agricoles, les modèles mixtes ont montré que le challenge des études futures, sur la réponse des cultures aux variations météorologiques, pourrait reposer sur un bon équilibre entre généralité et spécificité, entre petite échelle et grande échelle.

Les méthodes statistiques développées ici pourraient aussi être exploitées sur d'autres continents (Europe, Afrique) et sur d'autres cultures. Pour cela, la disponibilité suffisante en données agricoles est primordiale. On pourrait par exemple considérer le blé en Europe. Cette culture étant sensible à l'excès d'eau dans le sol, il pourrait être nécessaire d'intégrer des données satellites à la liste des prédicteurs potentiels, en plus de l'humidité du sol déjà présente.

Il serait aussi intéressant de considérer, en plus de cette étude de cas, un second exemple disposant de données moins abondantes ; cela permettrait d'évaluer la performance des modèles statistiques dans des conditions moins favorables en terme

de disponibilité en données. Le manque de données est un réel problème et cela peut conditionner le fait qu'on puisse utiliser un modèle statistique ou pas. Cela pourrait aussi se faire en dégradant la base de données utilisée ici, par exemple en rajoutant du bruit ou en ne considérant qu'un sous-ensemble des données.

De plus, pour fournir davantage d'information pendant la saison de croissance, il serait intéressant de tester la méthodologie avec une résolution temporelle plus fine (hebdomadaire ou quotidienne) même si cela demandera de comparer encore plus de prédicteurs. Ces données existent déjà mais nous ne les avons pas utilisées ici. Cela pourrait nettement améliorer le modèle sur les extrêmes, qui sont les faits d'évènements plus ponctuels et cela pourrait permettre de séparer les effets du jour et de la nuit.

Le but d'un modèle étant d'être utilisé, il serait enfin souhaitable de rendre accessible les sortie des modèles aux différents utilisateurs (grandes coopératives, collectivités territoriales ou des ministères). Cela pourrait se faire par le développement d'une plateforme en ligne avec l'aide d'experts du domaine, pour fournir des services climatiques ou météorologiques ajustés aux utilisateurs potentiels.

Bibliographie

[Abda16] Abdallah F., Philippe W., and Goffart J. Utilisation de la fluorescence chlorophyllienne pour l'évaluation du statut azoté des cultures (synthèse bibliographique). *Biotechnologie, Agronomie, Société et Environnement*, 20(1) :83–93, 2016.

[Abre08] Abreu D. and Riberas Z. General overview of the nass objective yield and objective measurement programs. Technical report, RDD Research Report Number RDD-08-10 USDA NASS Fairfax VA., 2008.

[Adam90] Adams R., Rosenzweig C., Pearl R., Ritchie J., McCarl B., Glyer D., Curry B., Jones J., Boote K., and Allen H. Global climate change and US agriculture. *Nature*, 345 :219–224, 1990.

[Adam98] Adams R., Hurd B., Lenhart S., and Leary N. Effects of global climate change on agriculture : an interpretative review. *Climate Research*, 11 :19–30, 1998.

[Agra82] Agrawal R. and Jain R. Composite model for forecasting rice yields. *Indian Journal of Agricultural Sciences*, 52 :177–181, 1982.

[Agra01] Agrawal R., Jain R., and Mehta S. Yield forecast based on weather variables and agricultural inputs on agroclimatic zone basis. *Indian Journal of Agricultural Sciences*, 7(71) :487–490, 2001.

[Agro95] Agronomic Interpretation Working Group. Systèmes de classification des terres selon leurs aptitudes pour les cultures : 1. la production des céréales de printemps. W.W. pettapiece (dir.), Bulletin technique 1995-6F. Technical report, Centre de recherche sur les terres et les ressources biologiques, Agriculture et Agroalimentaire Canada., 1995.

[Agui07] Aguilar M., Borjas F., and Espinosa M. Agronomic response of maize to limited levels of water under furrow irrigation in southern Spain. *Spanish Journal of Agricultural Research*, 5(4), 2007.

[Aire12] Aires F. Using random and mixed-effects to built impact models when the available historical record is short. *Journal of Applied Meteorology and Climatology*, 51 :1994–2004, 2012.

[Albe12] Albergel C., de Rosnay P., Balsamo G., Isaksen L., and Sabater J. Soil moisture analyses at ECMWF : Evaluation using global ground-based in situ observations. *Journal of Hydrometeorology*, 13(5) :1442–1460, 2012.

[Alle98] Allen R., Pereira L., Raes D., and Smith M. Crop evapotranspiration - guidelines for computing crop water requirements. Technical report, FAO Irrigation and drainage paper 56, 1998. www.climasouth.eu/sites/default/files/FAO%2056.pdf.

[Alli13] Allianz A. R. T. G. The weather business : How companies can protect against increasing weather volatility. Technical report, Allianz Global Corporate and Specialty, 2013.

[Amig06] Amigues J., Debaeke P., Itier B., Lemaire G., Seguin B., Tardieu F., and Thomas A. Sécheresse et agriculture. Réduire la vulnérabilité de l'agriculture à un risque accru de manque d'eau. Technical report, Expertise scientifique collective, synthèse du rapport, INRA (France), 2006.

[Anam02] Anaman K. *Assessing the economic and social impacts of extreme events on agriculture and the use of meteorological information to reduce adverse impacts*, chapter 4, pages 52–69. Secretariat of the World Meteorological Organization - Geneva - Switzerland, 2002.

[Anto12] Anton J., Kimura S., Lankoski J., and Cattaneo A. A comparative study of risk management in agriculture under climate change. In *OECD Food, Agriculture and Fisheries Papers*. OECD Publishing, Paris, 2012.

[Arau05] Araujo M., Pearson R., Thuiller W., and Erhard M. Validation of species–climate impact models under climate change. *Global Change Biology*, 11 :1504–1513, 2005.

[Arbu15] Arbuckle J., Morton L., and Hobbs J. Understanding farmer perspectives on climate change adaptation and mitigation : The roles of trust in sources of climate information, climate change beliefs, and perceived risk. *Environment and Behavior*, 47(2) :205–234, 2015.

[Arlo10] Arlot S. and Celisse A. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4 :40–79, 2010.

[Arms01] Armstrong J. *Principles of Forecasting : A Handbook for Researchers and Practitioners*. Boston, Kluwer Academic Publishers, 2001.

[Aslm13] Aslma M., Zamir M., Afzal I., Yaseen M., Mubeen M., and Shoaib A. Drought stress, its effect on maize production and development of drought tolerance through potassium application. *Cercetari Agronomice in Moldova*, 2013.

[Asne98] Asner G. Biophysical and biochemical sources of variability in canopy reflectance. *Remote Sensing of Environment*, 64 :234–253, 1998.

[Asse13] Asseng S., Ewert F., Rosenzweig C., Jones J., et al. Uncertainty in simulating wheat yields under climate change. *Nature Climate Change*, 3 :827–832, 2013.

[Aune06] Aune D. and Vogel F. The yield forecasting program of NASS, by the statistical methods branch and estimates division. *NASS, USDA, Washington, D.C., NASS Staff Report No. SMB 06-01.*, 2006.

[Aune12] Aune D. and Vogel F. The yield forecasting program of NASS, by the statistical methods branch and estimates division. *NASS, USDA, Washington, D.C., NASS Staff Report No. SMB 06-01.*, 2012.

[Baec10] Baechler M., Williamson J., Gilbride T., Cole P., Hefty M., and Love P. Guide to determining climate regions by county, building america best practices series, volume 7.1. Technical report, US Departement of Energy, 2010. http://apps1.eere.energy.gov/buildings/publications/pdfs/building_america/ba_climateguide_7_1.pdf.

[Bakk05] Bakker M., Govers G., Ewert F., Rounsevell M., and Jones R. Variability in regional wheat yields as a function of climate, soil and economic variables : assessing the risk of confounding. *Agriculture, Ecosystems and Environment*, 110 :195–209, 2005.

[Bals12] Balsamo G., Albergel C., Beljaars A., et al. Era-interim/land : A global land-surface reanalysis based on ERA-Interim meteorological forcing. Technical report, ERA Report Series, 2012.

[Bann11] Bannayan M. and Sanjani S. Weather conditions associated with irrigated crops in an arid and semi arid environment. *Agricultural and Forest Meteorology*, 151(12) :1589–1598, 2011.

[Bare13] Barella-Ortiz A., Polcher J., Tuzet A., and Laval K. Potential evaporation computation through an unstressed surface energy balance and its sensitivity to climate change effect. *Hydrology and Earth System Sciences Discussions*, 10 :10340–, 2013.

[Barn82] Barnett L. and Thompson D. The use of large-area spectral data in wheat yield estimation. *Remote Sensing of Environment*, 12(6) :509–518, 1982.

[Baro05] Baron C., Sultan B., Balme M., Sarr B., Teaore S., Lebel T., Janicot S., and Dingkuh M. From GCM grid cell to agricultural plot : scale issues affecting modelling of climate impacts. *Philosophical Transactions of the Royal Society B : Biological Sciences*, 360 :2095–2108, 2005.

[Baru07] Baruth B., Genovese G., and Leo O. *CGMSVersion 9.2 – User Manual and Technical Documentation.*, 2007.

[Bass13] Basso B., Cammarano D., and Carfagna E. Review of crop yield forecasting methods and early warning systems. Technical report, Michigan State University, USA, 2013.

[Batt16] Battude M., Al Bitar A., Morin D., Cros J., Huc M., Marais C., Le Dantec V., and Demarez V. Estimating maize biomass and yield over large areas using high spatial and temporal resolution Sentinel-2 like remote sensing data. *Remote Sensing of Environment*, 184 :668 – 681, 2016.

[Baus93] Bausch W. Soil background effects on reflectance based crop coefficients for corn. *Remote Sensing of Environment*, 46(2) :213–222, 1993.

[Bazg14] Bazgeer S., Fadavi G., and Hossainy S. Statistical modelling for cotton yield estimation using agricultural climate indices (a case study of Gharakhil district in Mazandaran province, Iran). *Research Journal of Environmental Sciences*, 8(2) :109–116, 2014.

[Beck10] Becker–Reshef I., Vermote E., Lindeman M., and Justice C. A generalized regression-based model for forecasting winter wheat yields in kansas and ukraine using modis data. *Remote Sensing of Environment*, 114 :1312–1323, 2010.

[Begu14] Beguería S., Vicente-Serrano S., Reig F., and Latorre B. Standardized precipitation evapotranspiration index (SPEI) revisited : parameter fitting, evapotranspiration models, tools, datasets and drought monitoring. *International Journal of Climatology*, 34(10) :3001–3023, 2014. http://sac.csic.es/spei/database.html.

[Bela00] Belanger G. and Bootsma A. Impact des changements climatiques sur l'agriculture au Québec. In *65ème Congrés de l'Ordre des Agronomes du Québec.*, 2000.

[Bell14] Bellocchi G., Villalobos F., Donatelli M., Christensen O., Rojas O., Confalonieri R., Athanasiadis I., Carpusca I., and Stöckle C. Extending existing models to capture vegetation response to extreme weather events : the modextreme project. In Daniel P.A. Nigel W.T. Q. and Rizzoli A. E., editors, *International Environmental Modelling and Software Society, 7th Intl. Congress on Env. Modelling and Software*, 2014.

[Ben 02] Ben Mohamed A., van Duivenbooden N., and Abdoussallam S. Impact of climate change on agricultural production in the Sahel – part 1. Methodological approach and case study for millet in Niger. *Climatic Change*, 2002.

[Ben-14] Ben-Ari T. and Makowski D. Decomposing global crop yield variability. *Environmental Research Letters*, 9(11), 2014.

[Ben-16a] Ben-Ari T., Adrian J., Klein T., Calanca P., Van der Veldec M., and Makowski D. Identifying indicators for extreme wheat and maize yield losses. *Agricultural and Forest Meteorology*, 2016.

[Ben-16b] Ben-Ari T. and Makowski D. Analysis of the trade-off between high crop yield and low yield instability at the global scale. *Environmental Research Letters*, 11(10), 2016.

[Beng12] Bengio Y. Practical recommendations for gradient-based training of deep architectures. *ArXiv e-prints*, 2012.

[Bere89] Beresford R. and Fullerton R. Effects of climate change on plant diseases. Technical report, DSIR Plant Division Submission to Climate Change Impacts Working Group, May, 1989 (Wellington, New Zealand : Department of Industrial and Scientific Research), 1989.

[Berr80] Berry J. and Bjorkman O. Photosynthetic response and adaptation to temperature in higher plants. *Annual Review of Plant Physiology*, 31 :491–543, 1980.

[Biel09] Bielza Diaz-Caneja M., Conte C., Catenaro R., Gallego Pinilla F., Dittmann C., and Stroblmair J. Agricultural insurance schemes, executive summary. Report, European Commission- Joint Research Centre, 2009.

[Bish96] Bishop C. *Neural networks for pattern recognition.* Clarendon Press - Oxford., 1996.

[Bish16] Bishop J., Jones H., Lukac M., and Potts S. Insect pollination reduces yield loss following heat stress in faba bean (Vicia faba L.). *Agriculture, Ecosystems and Environment*, 220 :89–96, 2016.

[Bloc09] Block P., Souza Filho F., Sun L., and Kwon H. A streamflow forecasting framework using multiple climate and hydrological models. *Journal of the American Water Resources Association*, 45(4) :828–843, 2009.

[Boat86] Boatwright G. and Whitehead V. Early warning and crop condition assessment research. *IEEE Transactions on Geoscience and Remote Sensing*, 24 :54–64, 1986.

[Boe07] Boe J., Terray L., Habets F., and Martin E. Statistical and dynamical downscaling of the Seine basin climate for hydro-meteorological studies. *International Journal of Climatology*, 27(12) :1643–1655, 2007.

[Bomm14] Bommier E. Peaks-over-threshold modelling of environmental data. Report, U.U.D.M. Project Report 2014, Uppsala University, 2014.

[Bond07] Bondeau A., Smith P., Zaehle S., Schaphoff S., Lucht W., Cramer W., and Gerten D. Modelling the role of agriculture for the 20th century global terrestrial carbon balance. *Global Change Biology*, 13(3) :679–706, 2007.

[Boot05] Bootsma A., Gameda S., and McKenney D. Impacts of potential climate change on selected agroclimatic indices in Atlantic Canada. *Canadian Journal of Soil Science*, 85(2) :329–343, 2005.

[Born12] Bornn L. and Zidek J. Efficient stabilization of crop yield prediction in the Canadian prairies. *Agricultural and Forest Meteorology*, 152 :223–232, 2012.

[Bott04] Bottou L. and LeCun Y. Large-scale on-line learning. In *NIPS'2003*, 2004.

[Boum94] Bouman B. Yield prediction by crop modeling and remote sensing. In *Proceedings of the Yield Forecasting Seminar, Villefranche 24-27 October, Eurostat-JRC-DGVI-FAO*, 1994.

[Boum96] Bouman B., Schapendonk A., Stol W., and van Kraalingen D. *Quantitative Approaches in Systems Analysis*, chapter Description of LINGRA, a model approach to evaluate potential productivities of grass-lands in diVerent European climate regions. DLO Research Institute for Agrobiology and Soil Fertility and C.T. De Wit Graduate School for Production Ecology, Wageningen, 1996.

[Box87] Box G. and N.R. D. *Empirical Model-Building and Response Surfaces.* Wiley, New York, 1987.

[Boye13] Boyer J., Byrneb P., Cassmand K., et al. The US drought of 2012 in perspective : a call to action. *Global Food Security*, 2(3) :139–143, 2013.

[Bris10] Brisson N., Gate P., Gouache D., Charmet G., Oury F., and Huard F. Why are wheat yields stagnating in europe ? A comprehensive data analysis for France. *Field Crops Research*, 119(1) :201 – 212, 2010.

[Brkl98] Brklacich M., Smit B., Bryant C., Veenhof B., and Beauchesne A. The Canada country study : Climate impacts and adaptation. *Toronto : Environnement Canada*, 1998.

[Brow06] Brown C. and Davis H. Receiver operating characteristic curves and related decision measures : a tutorial. *Chemometrics and Intelligent Laboratory Systems*, 80 :24–38, 2006.

[Brow08] Brown M. and Funk C. Food security under climate change. *Science*, 2008.

[Brya08] Bryant C., Singh B., Thomassin P., Baker l., Desroches S., Savoie M., Delusca K., Doyon M., and Seyoum E. Evaluation of agricultural adaptation processes and adaptive capacity to climate change and variability. Technical report, Project number A1332 submitted to Natural Resources Canada, Climate Change Impacts and Adaptation Program, Université de Montréal and McGill University Research Team, 2008.

[Burn02] Burnham K. and Anderson D. *Model selection and multimodel inference : A practical information-theoretic approach.* Springer, 2002.

[Butl13] Butler E. and Huybers P. Adaptation of US maize to temperature variations. *Nature Climate Change*, 3 :68–72, 2013.

[Cai12] Cai R., Yu D., and Oppenheimer M. Estimating the effects of weather variations on corn yields using geographically weighted panel regression. *Journal of Agricultural and Resource Economics*, 39(2) :230–252, 2012.

[Camp96] Campbell J. *Introduction to Remote Sensing, second edition.* The Guilford Press. New York., 1996.

[Cant05] Cantelaube P. and Terres J. Seasonal weather forecasts for crop yield modelling in Europe. *Tellus*, 57 :476–487, 2005.

[Caub15] Caubel J., Cortázar-Atauri I., Launay M., Noblet-Ducoudré N., Huard F., Bertuzzi P., and Graux A. Broadening the scope for ecoclimatic indicators to assess crop climate suitability according to ecophysiological, technical and quality criteria. *Agricultural and Forest Meteorology*, 207, 2015.

[Ça14] Çakir R. Effect of water stress at different development stages on vegetative and reproductive growth of corn. *Field Crops Research*, 89(1) :1–16, 2014.

[CCSP08] CCSP. The effects of climate change on agriculture, land resources, water resources, and biodiversity in the United States. Technical report, A Report by the U.S. Climate Change Science Program and the Subcommittee on Global Change Research. Washington, DC., USA, 2008. Backlund, P., A. Janetos, D. Schimel, J. Hatfield, K. Boote, P. Fay, L. Hahn, C. Izaurralde, B.A. Kimball, T. Mader, J. Morgan, D. Ort, W. Polley, A. Thomson, D. Wolfe, M. Ryan, S. Archer, R. Birdsey, C. Dahm, L. Heath, J. Hicke, D. Hollinger, T. Huxman, G. Okin, R. Oren, J. Randerson, W. Schlesinger, D. Lettenmaier, D. Major, L. Poff, S. Running, L. Hansen, D. Inouye, B.P. Kelly, L Meyerson, B. Peterson, and R. Shaw.

[Cegl13] Ceglar A. MCYFS - MARS Crop Yield Forecasting System, CGMS - Crop monitoring in Europe. Technical report, Monitoring Agricultural Resources Unit - Joint Research Centre, 2013. http://ies-webarchive-ext.jrc.it/mars/mars/content/download/3224/16230/file/02_Ceglar_JRC.pdf.

[Cerd17] Cerda R., Avelino J., Gary C., Tixier P., Lechevallier E., and Allinne C. Primary and secondary yield losses caused by pests and diseases : Assessment and modeling in coffee. *Gent D, ed. PLoS ONE.*, 12(1), 2017.

[Chal03] Challinor A., Slingo J., Wheeler T., Craufurd P., and Grimes D. Toward a combined seasonal weather and crop productivity forecasting system : determination of the working spatial scale. *Journal of Applied Meteorology*, 42 :175–192, 2003.

[Chal04] Challinor A., Wheeler T., Slingo J., Craufurd P., and Grimes D. Design and optimisation of a large-area process-based model for annual crops. *Agricultural and Forest Meteorology*, 124(1–2) :99–120, 2004.

[Chal05] Challinor A., Wheeler T., Slingo J., Craufurd P., and Grimes D. Simulation of crop yields using era-40 : Limits to skill and nonstationarity in weather–yield relationships. *Journal of Applied Meteorology*, 44 :516–531, 2005.

[Chal08] Challinor A. and Wheeler T. Crop yield reduction in the tropics under climate change : Processes and uncertainties. *Agricultural and Forest Meteorology*, 148(3) :343–356, 2008.

[Chal09] Challinor A., Ewert F., Arnold S., Simelton E., and Fraser E. Crops and climate change : progress, trends, and challenges in simulating impacts and informing adaptation. *Journal of Experimental Botany*, 60(10) :2775–2789, 2009.

[Chal15] Challinor A., Parkes B., and Ramirez-Villegas J. Crop yield response to climate change varies with cropping intensity. *Global Change Biology*, 21(4) :1679–1688, 2015.

[Chen16] Chen S., Chen X., and Xu J. Impacts of climate change on agriculture : Evidence from China. *Journal of Environmental Economics and Management*, 76 :105–124, 2016. http://www.sciencedirect.com/science/article/pii/S0095069615000066.

[Chip99] Chipanshi A., Ripley E., and Lawford R. Large-scale simulation of wheat yields in a semi-arid environment using a crop-growth model. *Agricultural Systems*, 59 :57–66, 1999.

[Ciai05] Ciais P. et al. Europe-wide reduction in primary productivity caused by the heat and drought in 2003. *Nature*, 437 :529–533, 2005.

[Clar10] Clarke L., Edmonds J., Krey V., Richels R., Rose S., and Tavoni M. International climate policy architectures : overview of the EMF 22 international scenarios. *Energy Economics*, page S64–S81, 2010.

[CNRS00] CNRS. Dossier scientifique sur l'eau : usages - cultures. http://www.cnrs.fr/cw/dossiers/doseau/decouv/usages/consoIndus.html, 2000.

[Cobl00] Coble K., Heifner R., and Zuniga M. Implications of crop yield and revenue insurance for producer hedging. *Journal of Agricultural and Resource Economics*, 25(2) :432–442, 2000.

[Cole01] Coles S. *An Introduction to Statistical Modeling of Extreme Values*. Springer, 2001.

[Craf02] Crafts-Brandner S. and Salvucci M. Sensitivity of photosynthesis in a C4 plant, maize, to heat stress. *Plant Physiology*, 129(4) :1773–1780, 2002.

[Crai13] Craig M. and Atkinson D. A literature review of crop area estimation. Technical report, FAO, 2013. http://www.fao.org/fileadmin/templates/ess/documents/meetings_and_workshops/GS_SAC_2013/Improving_methods_for_crops_estimates/Crop_Area_Estimation_Lit_review.pdf.

[Craw11] Crawley M. *Statistics : An Introduction using R*. Wiley, 2011.

[Crop05] Crops and Climate Change C. S., editors. *Effects of observed climate fluctuation on wheat flowering as simulated by the European crop growth monitoring system (CGMS).*, 2005.

[Cutf86] Cutforth H., Shaykewich C., and Cho C. Effect of soil water and temperature on corn (zea mays l.) root growth during emergence. *Canadian Journal of Soil Science*, 66(1) :51–58, 1986.

[Czad16] Czado C. Lecture 10 : Linear Mixed Models, 2016. https://www.google.fr/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0ahUKEwje9-z-jZnYAhURGOwKHXUHAUcQFggtMAA&url=http%3A%2F%2Fwww2.stat.duke.edu%2F~sayan%2FSta613%2F2017%2Flec%2FLMM.pdf&usg=AOvVaw0mik3SKhlW6ak7DaOivf5S, [En ligne ; Page disponible le 20 décembre 2017].

[Côt12] Côté H. and Ouranos. Calcul des indices agroclimatiques, Atlas agroclimatique du Québec. http://www.agrometeo.org/atlas/display_guide/Formules_atlas_agroclim_2012_v8.pdf, 2012. Accessed : 2016-07-30, www.agrometeo.org/atlas/display_guide/Formules_atlas_agroclim_2012_v10.docx.

[Dard06] Dardy H. and Lauer J. Plant physiology : Critical stages in the life of a corn plant. corn.agronomy.wisc.edu/Management/pdfs/CriticalStages.pdf, 2006.

[DaSi09] DaSilva L. L'impact économique des changements climatiques sur l'agriculture canadienne. Master's thesis, HEC Montréal, 2009.

[Day85] Day W. and Arkin R. *Proc. of NATO Advanced Research Workshop*, chapter Wheat growth and modelling. Plenum Press New York, 1985.

[De R98] De Rosnay P. and Polcher J. Modeling root water uptake in a complex land surface scheme coupled to a GCM. *Hydrology and Earth System Sciences*, 2 :239–256, 1998.

[De W10] De Wit A., Baruth B., Boogaard H., Diepen C., Kraalingen D., Micale F., Roller J., Supit I., and Wijngaart R. Using ERA-INTERIM for regional crop yield forecasting in Europe. *Climate Research*, 44 :41–53, 2010.

[Deba17] Debaje S. Estimated crop yield losses due to surface ozone exposure and economic damage in India. *Environmental Science and Pollution Research International*, 2017.

[Deep15] Deepak K., James S., Graham K., and Paul C. Climate variation explains a third of global crop yield variability. *Nature Communications*, 2015.

[Denh14] Denhartigh C. Adaptation de l'agriculture aux changements climatiques - recueil d'expériences territoriales. Technical report, Réseau Action Climat – France, 2014.

[Dequ07] Dequé M., Rowell D., Luthi D., Giorgi F., Christensen J., Rockel B., Jacob D., Kjellstrom E., De Castro M., and van den Hurk B. An intercomparison of regional climate simulations for Europe : assessing uncertainties in model projections. *Climate Change*, 81 :53–70, 2007.

[Dery14] Deryng D., Conway D., Ramankutty N., Price J., and Warren R. Global crop yield response to extreme heat stress under multiple climate change futures. *Environmental Research Letters*, 9(3) :034011, 2014.

[Desc80] Deschamps B. and Mehta D. Predictive ability and descriptive validity of earnings forecasting models. *The Journal of Finance*, 35, 1980.

[Desc07] Deschênes O. and Greenstone M. The economic impacts of climate change : Evidence from agricultural profits and random fluctuations in weather. *American Economic Review*, 97(1) :354–385, 2007.

[Dewa17] Dewaele H., Munier S., Albergel C., Planque C., Laanaia N., Carrer D., and Calvet J. Parameter optimisation for a better representation of drought by LSMs : inverse modelling vs. sequential data assimilation. *Hydrology and Earth System Sciences*, 21(9) :4861–4878, 2017.

[Diep89] Diepen C., Wolf J., and Keulen H. Wofost : a simulation model of crop production. *Soil Use Manage.*, 5 :16–24, 1989.

[Diep95] Diepen (van) C. and van der Wal T. *Workshop for Central and Eastern Europe on agrometeorological models : theory and applications in the MARS project, 21-25 November 1994, Ispra, Italy. EUR 16008 EN*, chapter Crop growth monitoring and yield forecasting at regional and national scale. Office for Off. Pub. of the EU, Luxembourg, 1995.

[Dora95] Doraiswamy P. and Cook P. Spring wheat yield assessment using NOAA AVHRR data. *Canadian Journal of Remote Sensing*, 21 :43–51, 1995.

[Dora07] Doraiswamy P., Akhmedov B., Beard L., Stern A., and Mueller R. Operational prediction of crop yields using modis data and products. in : Proceedings of the international archives of photogrammetry. *Remote Sensing and Spatial Information Sciences Special Publications*, 114 :1312–1323, 2007.

[Dori07] Dorigo W., Zurita-Milla R., De Wit A., Brazile J., Singh R., and Schaepman M. A review on reflective remote sensing and data assimilation techniques for enhanced agroecosystem modelling. *International Journal of Applied Earth Observation and Geoinformation*, 9(2) :165–193, 2007.

[Duch86] Duchon C. Corn yield prediction using climatology. *Journal of Climate and Applied Meteorology*, 25(5) :581–590, 1986.

[Duch95] Ducharme G., Gannoun A., Guertin M., and Jéquier J. Reference values obtained by kernel-based estimation of quantile regression. *Biometrics*, 51 :1105–1116, 1995.

[Duco93] Ducoudré N., Laval K., and Perrier A. Sechiba, a new set of parameterizations of the hydrologic exchanges at the land-atmosphere interface within the lmd atmospheric general circulation model. *Journal of Climate*, 6 :248–273, 1993.

[Duve16] Duveiller G., Fasbender D., and Meroni M. Revisiting the concept of a symmetric index of agrement for continous dataset. Technical report, Nature Scientific report, 2016.

[East10] Easterling W. *Guidelines for adapting agriculture to climate change. Handbook of Climate Change and Agroecosystems : Impacts, Adaptation, and Mitigation, ICP Series in Climate Change Impacts, Adaptation, and Mitigation – Vol. 1*. Imperial College Press, 2010.

[Edal11] Edalat M., Ghadiri H., Hamzehzarghani H., and Kazemeini S. Prediction of corn yield loss due to different redroot pigweed density and irrigation level using empirical models. *Australian Journal of Crop Science*, 5 :187–196, 02 2011.

[Enge97] Engel T., Hoogenboom G., Jones J., and Wilkens P. W. AEGIS-win : A computer program for the application of crop simulation models across geographical areas. *Agronomy Journal*, 89(6) :919, 1997.

[Este13] Estes L., Bradley B., Beukes H., Hole D., Lau M., Oppenheimer M., Schulze R., Tadross M., and Turner W. Comparing mechanistic and empirical model projections of crop suitability and productivity : implications for ecological forecasting. *Global Ecology and Biogeography*, 22 :1007–1018, 2013.

[Euro94] Eurostat. Méthodes de prévision de rendements agricoles. Technical report, Eurostat, FAO, 1994. Actes du séminaire de Villefranche-sur-Mer, du 24 au 27 octobre 1994.

[Fabi08] Fabián A., Jäger K., and Barnabás B. Effects of drought and combined drought and heat stress on germination ability and seminal root growth of wheat (Triticum aestivum L) seedlings. *Acta Biologica Szegediensis*, 52(1) :157–159, 2008.

[Fahr94] Fahrmeir L. K. and Tutz G. *Multivariate Statistical Modelling Based on Generalized Linear models*. Springer-Verlag, New York, NY, 1994.

[Fawc04] Fawcett T. ROC graphs : Notes and practical considerations for researchers. *Pattern Recognition Letters.*, 27 :882–891, 2004.

[Fawc06] Fawcett T. An introduction to roc analysis. *Pattern Recognition Letters*, 27(8) :861–874, 2006.

[Feng04] Feng S. and Hu Q. Changes in agro-meteorological indicators in the contiguous United-States : 1951–2000. *Theoretical and Applied Climatology*, 78(4) :247–264, 2004.

[Ferm11] Fermont A. and Benson T. Estimating yield of food crops grown by smallholder farmers : A review in the Uganda context. Technical report, International Food Policy Research Institute, IFPRI Discussion Paper 01097, 2011.

[Fink04] Fink A., Brucher T., Kruger A., Leckebusch G., Pinto J., and Ulbrich U. The 2003 european summer heatwaves and drought and synoptic diagnosis and impacts. *Weather*, 59 :209–216, 2004.

[Frie94] Friedman J. *An overview of predictive learning and function approximation. An overview of predictive learning and function approximation.* NATA ASI Seris, 1994.

[Gage15] Gage S., Doll J., and Safir G. *The Ecology of Agricultural Landscapes : Long-Term Research on the Path to Sustainability.*, chapter A crop stress index to predict climatic effects on row-crop agriculture in the U.S. North Central Region. Oxford University Press New York, USA., 2015.

[Gann02] Gannoun A., Girard S., Guinot C., and Saracco J. Trois méthodes non paramétriques pour l'estimation de courbes de référence - application à l'analyse de propriétés biophysiques de la peau. *Revue de Statistique Appliquée*, 50(1) :65–89, 2002.

[Gaus69] Gausman H., Allen W., and Cardenas R. Reflectance of cotton leaves and their structure. *Remote Sens.Environ.*, 1(1) :19–22, 1969.

[Gelm03] Gelman A., Carlin J., Stern H., and Rubin D. *Bayesian data analysis*. CRC texts in statistical science, 2003.

[Gelm07] Gelman A. and J. H. *Data Analysis Using Regression and Multilevel/Hierarchical Models.* New York, NY : Cambridge University Press, 2007.

[Gema92] Geman S., Bienenstock E., and Doursat R. Neural networks and the bias-variance dilema. *Neural Computing and Applications*, 4 :1–58, 1992.

[Geno98] Genovese G. *Agro-meteorological Applications for Regional Crop Monitoring and Production Assessment*, chapter The methodology, the results and the evaluation of the MARS crop yield forecasting system. Luxembourg, Office for Official Publications of the European Communities., 1998.

[Geno06] Genovese G., Fritz S., and Bettio M. A comparison and evaluation of performances among crop yield forecasting models based on remote sensing : Results from the geoland observatory of food monitoring. Technical report, EC - Joint Research Centre, IPSC, Agriculture., 2006.

[Geor10] George A. and Hanuschak S. Timely and accurate crop yield forecasting and estimation, history and initial gap analysis. Technical report, FAO, 2010.

[GIEC07] GIEC. Climate change impacts, adaptation and vulnerability. in : Fourth assessment report of the intergovernmental panel on climate change. *Cambridge University Press*, 2, 2007.

[GIEC14] GIEC. Changements climatiques 2014 : Rapport de synthèse. contribution des groupes de travail i, ii et iii au cinquième rapport d'évaluation du groupe d'experts intergouvernemental sur l'évolution du climat. Technical report, Sous la direction de l'équipe de rédaction principale, R.K. Pachauri et L.A. Meyer, GIEC, Genève, Suisse, 2014.

[Gijs02] Gijsman A., Jagtap S., and Jones J. Wading through a swamp of complete confusion : How to choose a method for estimating soil water retention parameters for crop models. *European Journal of Agronomy*, 18 :77–106, 2002.

[Gobi10] Gobin A. Modelling climate impacts on crop yields in belgium. *Climate Research*, 44(1) :55, 2010.

[Gobr97] Gobron N., Pinty B., and Verstraete M. Theoretical limits to the estimation of the leaf area index on the basis of visible and nearinfrared remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 35 :1438–1445, 1997.

[Gomm03] Gommes R. The FAO crop monitoring and forecasting approach. Technical report, FAO, 2003. Crop and Rangeland Monitoring in Eastern Africa for Early Warning and Food Security (D. Rijks, F. Rembold, T. Nègre, R. Gommes and M. Cherlet, eds). Proceedings of a JRC/FAO International Workshop, Nairobi, 28–30 January 2003. Rome, FAO).

[Good14] Good D. and Irwin S. Are USDA corn yield forecasts getting better or worse over time ? *Farmdoc daily (4) :166, Department of Agricultural and Consumer Economics University of Illinois*, 2014.

[Gorn10] Gornall J., Betts R., Burke E., Clark R., Camp J., Willett K., and Wiltshire A. Implications of climate change for agricultural productivity in the early twenty-first century. *Philosophical Transactions of the Royal Society B : Biological Sciences*, 7(365) :2973–2989, 2010.

[Goua15] Gouache D., Bouchon A., Jouanneau E., and Le Bris X. Agrometeorological analysis and prediction of wheat yield at the departmental level in France. *Agricultural and Forest Meteorology*, 209–210 :1–10, 2015.

[Grac16] Graczyk D. and Kundzewicz Z. Changes of temperature-related agroclimatic indices in Poland. *Theoretical and Applied Climatology*, 124(1–2) :401–410, 2016.

[Gras13] Grassini P., Eskridge K., and Cassman K. Distinguishing between yield advances and yield plateaus in historical crop production trends. *Nature Communications*, 4(2918) :1679–1688, 2013.

[Gree01] Greenough G., McGeehin M., Bernard S., Trtanj J., Riad J., and Engelberg D. The potential impacts of climate variability and change on health impacts of extreme weather events in the United States. *Climate Change*, 109(2) :191–198, 2001.

[Grig11] Grigorova B., Vasena I., Demirevska K., and Feller U. Combined drought and heat stress in wheat : changes in some heat shock proteins. *Biologia Plantarum*, 55(1) :105–111, 2011.

[Grot93] Groten S. NDVI – crop monitoring and early yield assessment of Burkina Faso. *International Journal of Remote Sensing*, 14 :1495–1515, 1993.

[Haga14] Hagan M. T., Demuth H. B., Beale M. H., and De Jesús O. *Neural Network Design*. Hagan, 2014.

[Hari08] Hariharan S. and Rogers J. *Estimation Procedures for Hierarchical Linear Models. Multilevel Modeling of Educational Data*. Charlotte, 2008.

[Hart99] Hartkamp A. D., Whit J., and Hoogenboom G. Interfacing geographic information systems with agronomic modeling : A review. *Agronomy Journal*, 91 :761–772, 1999.

[Hask95] Haskett J., Pachepsky Y., and Acock B. Estimation of soybean yields at county and state levels using GLYCIM : A case study for Iowa. *Agronomy Journal*, 87 :926,931, 1995.

[Hast01] Hastie T., Tibshirani R., and Friedman J. *The Elements of Statistical Learning*. Springer Series in Statistics Springer New York Inc., 2001.

[Hawk04] Hawkings D. The problem of overfitting. *Journal of Chemical Information and Modeling*, 44 :12–12, 2004.

[Hawk13] Hawkins E., Fricker T., Challinor A., Ferro C., Ho C., and Osborne T. Increasing influence of heat stress on french maize yields from the 1960s to the 2030s. *Global Change Biology*, 19 :937–947, 2013.

[Hayk94] Haykin. *Neural network a comprehensive foundation*. MacMillan Publishing Company, 2 edition, 1994.

[Hert91] Hertz J., Kroght A., and Palmer R. *Introduction to the theory of neural computation.* Cambridge Press., 1991.

[HLPE15] HLPE. Report on water for food security and nutrition extract from the report : Summary and recommendations. Technical report, HLPE, 2015.

[Hodg87] Hodges T., Botner B., Sakamoto C., and Hays H. Using the CERES-maize model to estimate production for the US Corn Belt. *Agricultural and Forest Meteorology*, 40(4) :293–303, 1987.

[Hol10] *Evaluating Climate Suitability for Agriculture in Switzerland. Modelling for Environment's Sake, International Congress on Environmental Modelling and Software, Fifth Biennal Meeting International Environmental Modelling. July 5 - 8 2010, Ottawa, Ontario, Canada.*, 2010.

[Hol11] *Analyzing climate effects on agriculture in time and space. 1st Conference on Spatial Statistics*, 2011.

[Hoog99] Hoogenboom G., Wilkens P., and Tsuji G. *Dssat v3*. University of Hawaii, 1999.

[Hoog00] Hoogenboom H. Contribution of agrometeorology to the simulation of crop production and its applications. *Agricultural and Forest Meteorology*, 103 :137–157, 2000.

[Hoog04] Hoogenboom G., J.W. W., and Messina C. From genome to crop : integration through simulation modelling. *Field Crops Research*, 90(1) :145–163, 2004.

[Houg90] Hough M. Agrometeorological aspects of crops in the United-Kingdom and Ireland. a review for sugar beet, oilseed rape, peas, wheat, barley, oats, potatoes, apples and pears. *Office for Official Publications of the EU, Luxembourg*, page 310, 1990.

[Houo11] Houot A. Carte des reliefs des États-unis. http://lamericaclub.blogspot.fr/2011/02/les-reliefs-des-etats-unis.html, 2011.

[Hox02] Hox J. *Multilevel Analysis, Techniques and Applications.* Lawrence Erlbaum Associates, 2002.

[Huet87] Huete A. Soil-dependent response in a developing crop canopy. *Agronomy Journal*, 79(1) :61–68, 1987.

[Hunt98] Hunt L.A.and Boote K. *Understanding options for agricultural production*, chapter Data for model operation, calibration, and evaluation., pages 9 – 39. Kluwer Academic Publisher, Great Britain, 1998.

[ICAS14] ICASA. Decision support system for agrotechnology transfer (DSSAT) developed under the international consortium for agricultural systems applications (ICASA). Technical report, Compendium on methods and tools to evaluate impacts of, and vulnerability and adaptation to, climate change. UNFCCC Nairobi Work Programme on impacts, vulnerability and adaptation to climate change., 2014.

[Iizu13] Iizumi T., Sakuma H., Yokozawa M., Luo J., Challinor A., Brown M., Sakurai G., and Yamagata T. Prediction of seasonal climate-induced variations in global food production. *Nature Climate Change*, 3 :904–908, 2013.

[Iizu14] Iizumi T., Yokozawa M., Sakurai G., Travasso M., Romanenkov V., Oettli P., Newby T., Ishigooka Y., and Furuya J. Historical changes in global yields : major cereal and legume crops from 1982 to 2006. *Global Ecology and Biogeography*, 23(3) :346–357, 2014.

[Iizu16] Iizumi T. and Ramankutty N. Changes in yield variability of major crops for 1981–2010 explained by climate change. *Environmental Research Letters*, 11(3) :1679–1688, 2016.

[Ines06] Ines A. and Hansen J. Bias correction of daily gcm rainfall for crop simulation studies. *Agricultural and Forest Meteorology*, 138(1-4) :44–53, 2006.

[IPSL15] IPSL. CMIP5 - data access, https://pcmdi.llnl.gov/mips/cmip5/data-portal.html. https://pcmdi.llnl.gov/mips/cmip5/data-portal.html, 2015.

[Irwi14] Irwin S., Good D., and Sanders D. Are usda corn yield forecasts getting better or worse over time ? *farmdoc daily (4) :166, Department of Agricultural and Consumer Economics, University of Illinois at Urbana-Champaign*, 2014. http://www.farmdoc.illinois.edu/marketing/agmas/reports/06_01/AgMAS06_01.pdf.

[Itte03] Ittersum (van) M., Leffelaar P., van Keulen H., Kropff M., Bastiaans L., and Goudriaan J. On approaches and applications of the wageningen crop models. *European Journal of Agronomy*, 18 :201–234, 2003.

[Itte13] Ittersum M. and Cassman K. Yield gap analysis - rationale, methods and applications - introduction to the special issue. *Field Crops Research*, 143(Supplement C) :1–3, 2013.

[Jews05] Jewson S. and Brix A. *Weather derivative valuation - The meteorological, statistical, financial and mathematical foundations*. Cambridge University Press, 2005.

[John11] Johnson F. and Sharma A. Accounting for interannual variability : a comparison of options for water resources climate change impact assessments. *Water Resources Research*, 47(4), 2011.

[John15] Johnston R., Sandefur H., Bandekar P., Matlock M., Haggard B., and Thoma G. Predicting changes in yield and water use in the production of corn in the United States under climate change scenarios. *Ecological Engineering*, 82(Supplement C) :555 – 565, 2015.

[Jone98] Jones J. and Luyten J. *Agricultural Systems, Modeling and Simulation*, chapter Simulation of biological processes. R.M. Peart and R.B. Curry, 1998.

[Jone01] Jones J., Keating B., and Porter C. Approaches to modular model development. *Agricultural Systems*, 70 :421–443, 2001.

[Jone03a] Jones J., Hoogenboom G., Porter C., Boote K., Batchelor W., Hunt L., Wilkens P., Singh U., Gijsman A., and Ritchie J. The DSSAT cropping system model. *European Journal of Agronomy*, 18 :235–265, 2003.

[Jone03b] Jones P. and Thornton P. The potential impacts of climate change on maize production in Africa and latin America in 2055. *Global Environmental Change*, 13 :51–59, 2003.

[Kand02] Kandiannan K., Chandaragiri K., Sankaran N., Balasubramanian T., and Kailasam C. Crop-weather model for turmeric yield forecasting for Coimbatore district, Tamil Nadu, India. *Agricultural and Forest Meteorology*, 112 :133–137, 2002.

[Kaus16] Kaushal N., Bhandari K., Siddique K., and Nayyar H. Food crops face rising temperatures : An overview of responses, adaptive mechanisms, and approaches to improve heat tolerance. In *Cogent Food and Agriculture*, 2016.

[Kayl91] Kaylen M. and Koroma S. Trend, weather variables, and the distribution of US corn yields. *ERAE*, 13 :249–258, 1991.

[Keat03] Keating B., Carberry P., Hammer G., et al. An overview of apsim, a model designed for farming systems simulation. *European Journal of Agronomy*, 18 :267–288, 2003.

[Kebe12] Kebede H., Fisher D., D.K., and Young L. Determination of moisture deficit and heat stress tolerance in corn using physiological measurements and a low-cost microcontroller-based monitoring system. *Journal of Agronomy and Crop Science*, 2012.

[Khur98] Khuri A., Mathew T., and Sinha B. *Statistical Tests for Mixed Linear Models*. Wiley, 1998.

[Kim15] Kim S., Kim J., Walko R., Myoung B., Stack D., and Kafatos M. Climate change impacts on maize-yield potential in the southwestern United States. *Procedia Environmental Sciences*, 29(Supplement C) :279–280, 2015. Agriculture and Climate Change - Adapting Crops to Increased Uncertainty (AGRI 2015).

[Koch10] Koch E. Etude de faisabilité d'une assurance rendement basée sur indice climatique. Master's thesis, Université Paris-Dauphine, France, 2010.

[Kotl07a] Kotlowski W. Qualitative models of climate variations impact on crop yields. Technical report, IIASA Interim Report - IIASA. Laxenburg. Austria. IR-07-034, 2007.

[Kotl07b] Kotlowski W. Qualitative models of climate variations impact on crop yields. Technical report, IIASA, 2007.

[Kowa14] Kowalik W., Dabrowska-Zielinska K., Meroni M., Raczka T., and De Wit A. Yield estimation using spot-vegetation products : A case study of wheat in european countries. *International Journal of Applied Earth Observation and Geoinformation*, 32 :228–239, 2014.

[Kram04] Kramer M. Automatic model selection in the mixed models framework. In *Annual Conference on Applied Statistics in Agriculture*. New Prairie Press, 2004. http://newprairiepress.org/agstatconference/2004/proceedings/9.

[Kran08] Kranz W., Irmak S., Van Donk S., Yonts C., and Martin D. Irrigation management for corn. Technical report, University of Nebraska NebGuide G1850, 2008.

[Krin05] Krinner G., Viovy N., de Noblet-Ducoudré N., Ogée J., Polcher J., Friedlingstein P., Ciais P., Sitch S., and Prentice I. A dynamic global vegetation model for studies of the coupled atmosphere-biosphere system. *Global Biogeochemical Cycles*, 19, 2005.

[Lal11] Lal R., Delgado J., Groffman P., Millar N., Dell C., and Rotz A. Management to mitigate and adapt to climate change. *Journal of Soil and Water Conservation*, 66 :276–285, 2011.

[Lal12] Lal L., Delgado J., Gulliford J., Nielsen D., Rice C., and Van Pelt R. Adapting agriculture to drought and extreme events. *Journal of Soil and Water Conservation*, 67(6) :162A–166A, 2012.

[Lali13] Lalic B., Eitzinger J., Mihailovic D., Thaler S., and Jancic M. Climate change impacts on winter wheat yield change–which climatic parameters are crucial in Pannonian lowland. *Journal of Agricultural Science*, 151(6) :757–774, 2013.

[Land00] Landau S., Mitchell R., Barnett V., Colls J., Craigon J., and Payne R. A parsimonious, multiple-regression model of wheat yield response to environment. *Agricultural and Forest Meteorology*, 101(2–3) :151–166, 2000.

[Lass93] Lass L., Callihan R., and Everson D. Forecasting the harvest date and yield of sweet corn by complex regression models. *Journal of the American Society for Horticultural Science*, 118 :450–455, 1993.

[Lazo11] Lazo J., Lawson M., Larsen P., and Waldman D. U.S. economic sensitivity to weather variability. *American Metorological Society*, 2011.

[Le R13] Le Rest K. *Méthodes statistiques pour la modélisation des facteurs influençant la distribution et l'abondance de populations : application aux rapaces diurnes nichant en France.* Biologie de l'environnement, des populations, écologie, Université de Poitiers, Poitiers, 2013.

[Leck02] Leckebusch G., Ulbrich U., and Speth P. Identification of extreme events under climate change conditions over Europe and the northwest-atlantic region : spatial patterns and time series characteristics. *Geophysical Research Abstracts*, 4, 2002.

[Lend07] Lenderink G., Buishand A., and Van Deursen W. Estimates of future discharges of the river rhine using two scenario methodologies : direct versus delta approach. *Hydrology and Earth System Sciences*, 11(3) :1145–1159, 2007.

[Leng16] Leng G., Zhang X., Huang M., Asrar G., and Leung L. The role of climate covariability on crop yields in the conterminous United-States. *Scientific Reports*, 6(33160) :346–357, 2016.

[Leng17] Leng G. and Huang M. Crop yield response to climate change varies with crop spatial distribution pattern. *Scientific Reports*, 7(1) :1463, 2017.

[Lepa12] Lepage M., Bourgeois G., and Bélanger G. Indices agrométéorologiques pour l'aide à la décision dans un contexte de climat variable et en évolution. *Centre de référence en agriculture et agroalimentaire du Québec (CRAAQ), Québec, QC, Canada.*, 2012.

[Lesk16] Lesk C., Rowhani P., and Ramankutty N. Influence of extreme weather disasters on global crop production. *Nature*, 529 :84–87, 2016.

[Leve08] Leveau V. and Peigney S. Zoom sur la production de maïs des États-unis. *Perspectives agricoles*, 347, 2008.

[Lewa99] Lewandrowski J. and Schimmelpfennig D. Economic implications of climate change for US agriculture : assessing recent evidence. *Land Economics*, 75 :39–57, 1999.

[Lian04] Liang S. *Quantitative Remote Sensing of Land Surfaces*. Wiley, 2004.

[Lill15] Lillesand T. and Kiefer R. *Remote Sensing and Image Interpretation*. Wiley, 2015.

[Lin12] Lin M. and Huybers P. Reckoning wheat yield trends. *Environmental Research Letters*, 7, 2012.

[Lins88] Linsdtrom M. and Bates D. Newton raphson and EM algorithm for linear mixed-effects models for repeated-me asures data. *Journal of the American Stastitical Association*, 83 :1014–1022, 1988.

[Lins90] Linsdtrom M. and Bates D. Nonlinear mixed effects models for repaeted measures data. *Biometrics*, 46 :673–687, 1990.

[Liu14] Liu M., Rajagopalan K., Chung S., Jiang X., Harrison J., Nergui T., Guenther A., Miller C., Reyes J., and Tague C. What is the importance of climate model bias when projecting the impacts of climate change on land surface processes ? *Biogeosciences*, 11(10) :2601–2622, 2014.

[Lobe07a] Lobell D. and Field C. Global scale climate–crop yield relationships and the impacts of recent warming. *Environmental Research Letters*, 2(1), 2007.

[Lobe07b] Lobell D., Cahill K., and Field C. Historical effects of temperature and precipitation on California crop yields. *Climate Change*, 81 :187–203, 2007.

[Lobe09] Lobell D., Cassman K., and Field C. Crop yield gaps : their importance, magnitudes, and causes. *Annual Review of Environment and Resources*, 34(1) :179–204, 2009.

[Lobe10a] Lobell D. and Burke M. *Climate Change and Food Security : Adapting Agriculture to a Warmer World*. Springer, 2010.

[Lobe10b] Lobell D. and Burke M. On the use of statistical models to predict crop yield responses to climate change. *Agricultural and Forest Meteorology*, 150 :1443–1452, 2010.

[Lobe11a] Lobell D., Bänziger M., Magorokosho C., and Vivek B. Nonlinear heat effects on African maize as evidenced by historical yield trials. *Nature Climate Change*, 1 :42–45, 2011.

[Lobe11b] Lobell D., Burke M., Tebaldi C., Mastrandrea M., Falcon W., and Naylor R. Prioritizing climate change adaptation needs for food security in 2030. *Science*, 319(5863) :607–610, 2011.

[Lobe11c] Lobell D., Schlenker W., and Costa-Roberts J. Climate trends and global crop production since 1980. *Science*, 333(6042) :616–620, 2011.

[Lobe13a] Lobell D. Errors in climate datasets and their effects on statistical crop models. *Agricultural and Forest Meteorology*, 170 :58–66, 2013.

[Lobe13b] Lobell D. The use of satellite data for crop yield gap analysis. *Field Crops Research*, 143 :56–64, 2013.

[Lobe14a] Lobell D., Roberts M., Schlenker W., Braun N., Little B., Rejesus R., and Hammer G. Greater sensitivity to drought accompanies maize yield increase in the U.S. Midwest. *Science*, 2014.

[Lobe14b] Lobell D.B. andCassman K. and Field C. Getting caught with our plants down : the risks of a global crop yield slowdown from climate trends in the next two decades. *Environmental Research Letters*, 9(7), 2014.

[Lobe17] Lobell D. and Asseng S. Comparing estimates of climate change impacts from processbased and statistical crop models. *Environmental Research Letters*, 12(1) :58–66, 2017.

[Loom79] Loomis R., Rabbinge R., and Ng E. Explanatory models in crop physiology. *Annual Review of Plant Physiology*, 30(1) :339–367, 1979.

[Lope15] Lopez-Lozano R., Duveillera G., Seguini L., Meroni M., García-Condado S., Hooker J., Leo O., and Baruth B. Towards regional grain yield forecasting with 1 km-resolution eo biophysical products : Strengths and limitations at pan-european level. *Agricultural and Forest Meteorology*, 206 :12–32, 2015.

[Maas88] Maas S. Using satellite data to improve model estimates of crop yield. *Agronomy Journal*, 80(4) :655–662, 1988.

[MacD80] MacDonald R. and Hall F. Global crop forecasting. *Science*, 208 :670–679, 1980.

[Maco00] Macours K. and Swinnen J. Causes of output decline in economic transition : The case of central and eastern european agriculture. *The Journal of Agricultural Science*, 28(1) :172–206, 2000.

[Mako10] Makowski D. and Mittinty M. Comparison of scoring systems for invasive pests using ROC analysis and Monte Carlo simulations. *Risk Analysis*, 30(6) :906–15, 2010.

[Makr98] Makridadis S., Wheelwright S., and Hyndman R. *Forecasting, Methods and Applications.* New York, John Wiley and Sons, 1998.

[Mart04] Marteau D., Carle J., Fourneaux S., Holz R., and Moreno M. *La gestion du risque climatique.* Coll. Gestion, 2004.

[Math15] Mathieu J., Hatté C., Balesdent J., and Parent E. Deep soil carbon dynamics are driven more by soil type than by climate : a worldwide meta-analysis of radiocarbon profiles. *Global Change Biology*, 21(11), 2015.

[Math16] Mathieu J. and Aires F. Statistical weather impact models : an application of neural network and mixed-effects forn corn production over the United-States. *Journal of applied Meteorology and Climatology*, 55(11) :2509–2527, 2016.

[Mathss] Mathieu J. and Aires F. Impact of agro-climatic indices to improve crop yield forecasting. *Agricultural and forest Meteorology*, "review process".

[Maup14] Maupin M., Kenny J., Hutson S., Lovelace J., Barber N., and Linsey K. Estimated use of water in the United-States in 2010 : U.S. Geological Survey Circular 1405. Technical report, USGS, 2014. http ://dx.doi.org/10.3133/cir1405.

[McCo86] McCool P., Musselman R., Teso R., and Oshima R. Determining crop yield losses from air pollutants. *California Agriculture*, 40(7), 1986.

[McCu43] McCulloch W. and Pitts W. A logical calculus of the ideas immanent in nervous activity. *Bulletin of mathematical biophysics*, 5 :115–133, 1943.

[McQu81] McQuigg J. *Food-Climate Interactions*, chapter Climate variability and crop yield in high and low temperature regions. D. Reidel, Dordrecht, The Netherlands., 1981.

[Mend99] Mendelsohn R. and Dinar A. Climate change, agriculture, and developing countries : Does adaptation matter ? *World Bank Research Observer*, 1999.

[Mend03] Mendelsohn R. and Dinar A. Climate, water, and agriculture. *Land Economics*, 79(3) :328–341, 2003.

[Mend08] Mendelsohn R. The impact of climate change on agriculture in developing countries. *Journal of Natural Resources Policy Research*, 1(1) :5–19, 2008. http://dx.doi.org/10.1080/19390450802495882.

[Mend09] Mendelsohn R. and Dinar A. *Climate Change and Agriculture : An Economic Analysis of Global Impacts, Adaptation, and Distributional Effects.* Edward Elgar Publishing, England, 2009.

[Mend10] Mendelsohn R. Agriculture and economic adaptation to agriculture. Technical report, OECD, Paris, 2010.

[Meng08] Meng L. and Quiring S. A comparison of soil moisture models using soil climate analysis network observations. *Journal of Hydrometeorology*, 9 :641–659, 2008.

[Menz03] Menzel A., Jakobi G., Ahas R., Scheifinger H., and Estrella N. Variations of the climatological growing season (1951–2000) in Germany compared with other countries. *International Journal of Climatology*, 23 :793–812, 2003.

[Mero13] Meroni M., Marinho E., Sghaier N., Verstrate M., and Leo O. Remote sensing based yield estimation in a stochastic framework — case study of durum wheat in tunisia. *Remote Sensing*, 2013.

[Meye94] Meyer-Roux J. Introduction à la prévision des rendements. In *Méthodes de prévision de rendements agricoles. Actes du séminaire de Villefranche-sur-Mer, du 24 au 27 octobre 1994*, pages 13–15, 1994.

[Mont96] Monteith J. The quest for balance in crop modeling. *Agronomy Journal*, 88 :695–697, 1996.

[Moon02] Moonen A., Ercoli L., Mariotti M., and Masoni A. Climate change in Italy indicated by agrometeorological indices over 122 years. *Agricultural and Forest Meteorology*, 111(1) :13–27, 2002.

[Mora14] Moral García F. J., Rebollo F. J., Paniagua L. L., and García A. The integration of bioclimatic indices in an objective probabilistic model for establishing and mapping viticulture suitability in a region. In *EGU General Assembly Conference Abstracts*, 2014.

[More15] Moreto V. and Souza R. Agrometeorological models for groundnut crop yield forecasting in the Jaboticabal, São Paulo State region, Brazil. *Acta Scientiarum Agronomy*, 37(4) :403–410, 2015.

[More16] More A. Survey of resampling techniques for improving classification performance in unbalanced datasets. *ArXiv e-prints*, August 2016.

[Moss08] Moss R., Babiker M., Brinkman S., Calvo E., et al. Towards new scenarios for analysis of emissions, climate change, impacts, and response strategies. Technical report, Intergovernmental Panel on Climate Change, 2008.

[Moth11] Motha R. *The Impact of Extreme Weather Events on Agriculture in the United-States*, chapter 30, pages 397–407. UNL Faculty. 1311, 2011.

[Mull14] Muller C. and Robertson R. Projecting future crop productivity for global economic modeling. *Agricultural Economics*, 45(1) :37–50, 2014.

[Neil90] Neild R. and Newman J. Growing season characteristics and requirements in the corn belt. ., 1990.

[Nest99] Nestorov I., Rowland M., Hadjitodorov S., and Petrov I. Empirical versus mechanistic modelling : Comparison of an artificial neural network to a mechanistically based model for quantitative structure pharmacokinetic relationships of a homologous series of barbiturates. *AAPS Pharmsci*, 4, 1999.

[Neum10] Neumann K., Verburg P., Stehfest E., and Muller C. The yield gap of global grain production : A spatial analysis. *Agricultural Systems*, 103(5) :316–326, 2010.

[Newl14] Newlands N., Zamar D., Kouadio L., Zhang Y., Chipanshi A., Potgieter A., Toure S., and Hill H. An integrated model for improved seasonal forecasting of agricultural crop yield under environmental uncertainty. *Frontiers in Environmental Science*, 2 :17, 2014.

[Ngo-07] Ngo-Duc T., Laval K., Ramillien G., Polcher J., , and Cazenave A. Validation of the land water storage simulated by Organising Carbon and Hydrology In Dynamic Ecosystems (orchidee) with Gravity Recovery And Climate Experiment (grace) data. *Water Resources Research*, 43, 2007.

[Nix84] Nix H. Minimum data sets for agrotechnology transfer. In *Proceedings of the International Symposium of Minimum Data Sets for Agrotechnology Transfer, 21-26 March 1983, ICRISAT Center, India, Patancheru, India.*, 1984.

[NRDC13] NRDC. Soil matters : How the federal crop insurance program should be reformed. Technical report, Natural Resources Defense Council, 2013. https://www.nrdc.org/media/2013/130827.

[Oerk06] Oerke E. Crop losses to pests. *The Journal of Agricultural Science*, 144(1) :31–43, 2006.

[Oles01] Olesen J., Bocher P., and Jensen T. Comparison of scales of climate and soil data for aggregating simulated yields of winter wheat in Denmark. *Agriculture, Ecosystems and Environment*, 82(1–3) :213–228, 2001.

[Oles11] Olesen J., Trnka M., Kersebaum K., Skjelvag A., Seguin B., Peltonen-Sainio P., Rossi F., Kozyra J., and Micale F. Impacts and adaptation of european crop production systems to climate change. *European Journal of Agronomy*, 34(2) :96–112, 2011.

[Otto02] Otto C. *Effects of climate change and variability on agricultural production systems*. Boston, Kluwer Academic Publishers, 2002.

[Palm65] Palmer W. Meteorological drought. *Research Paper No. 45, US Weather Bureau, Washington, DC.*, 1965.

[Palm93] Palm R. and Dagnelie P. Tendance génerale et effets du climat dans la prévision des rendements agricoles des différent pays des C.E. *Office for Official Publications of the EU, Luxembourg*, 22 :128, 1993.

[Palm94] Palm R. Modèles agrométéorologiques : régression et analyse de la tendance. In *Méthode de prévision de rendements agricoles. Actes du séminaire de Villefranche-sur-Mer, du 24 au 27 octobre 1994*, pages 67–76, 1994.

[Palm97] Palm R. Les modèles de prévision statistique : cas du modèle eurostat-agromet. Technical report, Luxembourg Office for Official Publications of the European Communities, 1997. In : Estimation de la production agricole à une échelle régionale (B. Tychon and V. Tonnard, eds).

[Pand10] Panda S., Ames D., and Panigrahi S. Application of vegetation indices for agricultural crop yield prediction using neural network techniques. *Remote Sensing*, 2(3) :673–696, 2010.

[Patt71] Patterson H. and Thompson R. Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58 :545–554, 1971.

[Paz07] Paz J., Fraisse C., Hatch L., Garcia A., Guerra L., Uryasev O., Bellow J., Jones J., and Hoogenboom G. Development of an enso-based irrigation decision support tool for peanut production in the southeastern US. *Computers and Electronics in Agriculture*, 55(1) :28–35, 2007.

[Peel07] Peel M. C., Finlayson B. L., and McMahon T. A. Updated world map of the Koppen-Geiger climate classification. *Hydrology and Earth System Sciences*, 11(5) :1633–1644, 2007.

[Pelt10] Peltonen-Sainio P., Jauhiainen L., Trnka M., et al. Coincidence of variation in yield and climate in Europe. *Agriculture, Ecosystems and Environment*, 139(4) :483 – 489, 2010.

[Pian10] Piani C., Haerter J., and Coppola E. Statistical bias correction for daily precipitation in regional climate models over Europe. *Theoretical and Applied Climatology*, 99(1) :187–192, 2010.

[Pinh09] Pinheiro J. and Bates D. *Mixed-effects models in S and S-plus. Statistics and Computing*. Springer, 2009.

[Pint03] Pinter P., Ritchie J., Hatfield J., and Hart G. The agricultural research service's remote sensing program : An example of interagency collaboration. *Photogrammetric Engineering and Remote Sensing*, 69(6) :615–618, 2003.

[Pint09] Pinty B., Lavergne T., Widlowsky J.-L., Gobron N., and Verstraete M. On the need to observe vegetation canopies in the near-infrared to estimate visible light absorption. *Remote Sensing of Environment*, 113 :10–23, 2009.

[Poir00] Poiraud-Casanova S. *Estimation non paramétrique des quantiles conditionnels.* Thèse de doctorat, Université Toulouse 1., 2000.

[Port09] Porter C., Jones J., Adiku S., Gijsman A., Gargiulo O., and Naab J. Modeling organic carbon and carbon-mediated soil processes in DSSAT v4.5. *Operational Research*, 10 :247–278, 2009.

[Port14] Porter J. et al. *Food security and food production systems*, chapter Executive summary. IPCC 2014, 2014.

[Powe16] Powell J. and Reinhard S. Measuring the effects of extreme weather events on yields. *Weather and Climate Extremes*, 12 :69 – 79, 2016.

[Prak15] Prakash A., Rao J., Mukherjee A. K., Jeyaveeran B., Pokhare S., Adak T., S M., and PR S. *Climate change impact on crop pests*. Applied Zoologists Research Association (AZRA), 04 2015.

[Pras11] Prasad P., Pisipati S., Momčilović I., and Ristic Z. Independent and combined effects of high temperature and drought stress during grain filling on plant yield and chloroplast EF-Tu expression in spring wheat. *Journal of Agronomy and Crop Science*, 197 :430–441, 2011.

[Prin95] Prince S. and Goward S. Global primary production : a remote sensing approach. *Journal of Biogeography*, 22 :815–835, 1995.

[Priy01] Priya S. and Shibasaki R. National spatial crop yield simulation using GIS-based crop production model. *Ecological Modelling*, 136(2–3) :113–129, 2001.

[Qian01] Qian B., Hayhoe H., and Gameda S. Developing daily climate scenarios for agricultural impact studies. Technical report, Eastern Cereal and Oilseed Research Centre, Agriculture and Agri-Food Canada, Ottawa, 2001.

[Qian10] Qian B., Zhang X., Chen K., Feng Y., and O'Brien T. Observed long-term trends for agroclimatic conditions in Canada. *Journal of Applied Meteorology and Climatology*, 49(4) :604–618, 2010.

[Qian13] Qian B., De Jong R., Gameda S., Huffman T., Neilsen D., Desjardins R., Wang H., and McConkey B. Impact of climate change scenarios on canadian agroclimatic indices. *Canadian Journal of Soil Science*, 93(2) :243–259, 2013.

[Raft17] Raftery A., Zimmer A., Frierson D., Startz R., and Liu P. Less than 2°C warming by 2100 unlikely. *Nature Climate Change*, 7 :637–641, 2017.

[Rami73] Ramirez J. and Bauer A. Small grains response to growing degree units. *Agronomy Abstracts*, 1973.

[Raub05] Rauber R., Walsh J., and Charlevoix D. *Severe and hazardous weather : an introduction to highimpact meteorology*. Kendall Hunt Pub Co, 2005.

[Ray99] Ray S. and Pokharna S. Cotton yield estimation using agrometeorological model and satellite-derived spectral profile. *International Journal of Remote Sensing*, 20(14) :2693–2702, 1999.

[Reyn01] Reynolds C. Input data sources, climate normals, crop models and data extraction routines utilized by pecad. In *paper presented at 3rd Int, Conf. Geospatial Information in Agriculture and Forestry, Denver, CO.*, 2001.

[Reyn09] Reynaud A. Adaptation à court et à long terme de l'agriculture au risque de sécheresse : une approche par couplage de modèles biophysiques et économiques. *Revue d'Etudes en Agriculture et Environnement - Review of agricultural and environmental studies, INRA Editions*, 90(2) :121–154, 2009.

[Ritc93] Ritchie S., Hanway J., and Benson G. How a corn plant develops. Report, Spec. Rep. 48 (revised). Iowa State Univ. of Sc. and Technol. Coop. Ext. Serv., Ames, 1993. www.publications.iowa.gov/18027.

[Robb51] Robbins H. and Monro S. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3) :400–407, 1951.

[Robe02] Robeson S. Increasing growing-season length in Illinois during the 20th century. *Climatic Change*, 52(1–2) :219–238, 2002.

[Robe13] Robertson S., Jeffrey S., Unterschultz J., and Boxall P. Estimating yield response to temperature and identifying critical temperatures for annual crops in the canadian prairie region. *Canadian Journal of Plant Science*, 93(6) :1237–1247, 2013.

[Robe15] Robertson R. *DSSAT-estimated climate change impacts featured - IFPRI*, chapter 2. National Geographic Magazine, November 2015. https://www.nationalgeographic.com/climate-change/how-to-live-with-it/crops.html.

[Roja11] Rojas R., Feyen L., Dosio A., and Bavera D. Improving pan-european hydrological simulation of extreme events through statistical bias correction of RCM-driven climate simulations. *Hydrology and Earth System Sciences*, 15(8) :2599–2620, 2011.

[Rond96] Rondeaux G., Steven M., and Baret F. Optimization of soil-adjusted vegetation index. *Remote Sensing of Environment*, 55(2) :95–107, 1996.

[Rose94] Rosema A. Use of meteosat for crop yield forecasting. In *Proceedings of the Yield Forecasting Seminar, Villefranche 24-27 October, Eurostat-JRC-DGVI-FAO*, 1994.

[Rous05] Rousseeuw P. and Leroy A. *Robust Regression and Outlier Detection*. John Wiley & Sons., 2005.

[Rudo90] Rudorff B. and Batista G. Yield estimation of sugarcane based on agrometeorological- spectral models. *Remote Sensing of Environment*, 33(3) :183–192, 1990.

[Saka78] Sakamoto S. The Z-index as a variable for crop yield estimation. *Agricultural Meteorology*, 19 :305–313, 1978.

[Sanc14] Sanchez B., Rasmussen A., and Porter J. Temperatures and the growth and development of maize and rice : a review. *Global Change Biology*, 20(2) :408–417, 2014.

[Scha98] Schapendonk A., Stol W., van Kraalingen D., and Bouman B. LINGRA, a sink/source model to simulate grassland productivity in Europe. *European Journal of Agronomy*, 9(2–3) :87–100, 1998. http://www.sciencedirect.com/science/article/pii/S1161030198000276.

[Schi96] Schimmelpfennig D. Uncertainty in economic models of climate-change impacts. *Climate Change*, 33 :213–234, 1996.

[Schl09] Schlenker W. and Roberts M. Nonlinear temperature effects indicate severe damages to U.S. crop yields under climate change. *Proceedings of the National Academy of Sciences*, 106(15) :594–598, 2009.

[Scho90] Schoenau G. and Kehrig R. A method for calculating degree-days to any base temperature. *Energy Build*, 14(4) :299–302, 1990.

[Schu00] Schulze R. Transcending scales of space and time in impact studies of climate and climate change on agrohydrological responses. *Agriculture, Ecosystems and Environment*, 82(1–3) :185–212, 2000.

[Seme06] Semenov S., Yasukevich V., and Gelver E. Identification of climatogenic changes. Technical report, Moscow, Publishing Centre "Meteorology and Hydrology", 2006.

[Seme07] Semenov M. and Doblas-Reyes F. Utility of dynamical seasonal forecasts in predicting crop yield. *Climate Research*, 34 :71–81, 2007.

[Senn09] Sennikovs J. and Bethers U. *Statistical downscaling method of regional climate model results for hydrological modelling.*, chapter 3, page 3962–3968. Anderssen, R.S., Braddock, R.D., Newham, L.T.H. (Eds.), 2009.

[Sess06] Sessions D. and Stevans N. Investigating omitted variable bias in regression parameter estimation : A genetic algorithm approach. *Computational Statistics and Data Analysis*, 50 :2835–2854, 2006.

[Shar04] Sharma S., Srivastava A., and Sud D. Small area crop estimation methodology for crop yield estimates at gram panchayat level. *Journal of the Indian Society of Agricultural Statistics*, 57 :26–37, 2004. http://www.isas.org.in/jsp/volume/vol57/sdsharma.pdf.

[Shef06] Sheffield J., Goteti G., and Wood E. Development of a 50-yr high-resolution global dataset of meteorological forcings for land surface modeling. *Journal of Climate*, 19(13) :3088–3111, 2006. http://hydrology.princeton.edu/data.pdsi.php.

[Shef12] Sheffield J., Wood E., and Roderick M. Little change in global drought over the past 60 years. *Nature*, 491 :435–438, 2012.

[Shin09] Shin D., Baigorria G., Lim Y., Cocke S., LaRow T., O'Brien J., and Jones J. Assessing crop yield simulations with various seasonal climate data. Technical report, Science and Technology Infusion Climate Bulletin. NOAA's National Weather Service. 7th NOAA Annual Climate Prediction Application Science Workshop. Norman. OK. 24–27 October 2009., 2009.

[Shun04] Shunlin L. *Quantitative Remote Sensing of Land Surfaces*. Wiley, 2004.

[Sitc03] Sitch S., Smith B., Prentice I., Arneth A., Bondeau A., Cramer W., and Venevsky S. Evaluation of ecosystem dynamics, plant geography and terrestrial carbon cycling in the lpj dynamic global vegetation model. *Global Change Biology*, 2 :161–185, 2003.

[Siva05] Sivakumar M., Motha R., and Das H. *Natural disasters and extreme events in agriculture : Impacts and mitigation*. Springer International, 01 2005.

[Smit75] Smith L. Methods in agricultural meteorology. *Developments in Atmospheric Science*, 1975.

[Soil98a] Soil Survey Staff. Assessment of water holding capacity of soils map. Technical report, Soil map and soil climate map, USDA-NRCS, Soil Science Division, World Soil Resources, Washington D.C., 1998. https://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/use/?cid=nrcs142p2_054022.

[Soil98b] Soil Survey Staff. The US general soil map (STATSGO2). Technical report, Natural Resources Conservation Service, United States Department of Agriculture. Web Soil Survey. Available online at http ://websoilsurvey.nrcs.usda.gov/., 1998. https://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/survey/class/maps/?cid=nrcs142p2_053589.

[Solt13] Soltani A., Waismoradi A., Heidari M., and Rahmati H. Effect of water deficit stress and nitrogen on yield and compatibility metabolites on two medium maturity corn cultivars. *International Journal of Agriculture and Crop Sciences*, 5(7) :737–740, 2013.

[Sout00] Southworth J., Randolph J., Habeck M., Doering O., Pfeifer R., Rao D., and Johnston J. Consequences of future climate change and changing climate variability on maize yields in the midwestern United-States. *Agriculture, Ecosystems and Environment*, 82(1) :139 – 158, 2000.

[Spin56a] Spinks G. Uses and method of crop forecasting. *Review of marketing and agricultural economics*, 24, 1956. http://ageconsearch.umn.edu/bitstream/8933/1/24010018.pdf.

[Spin56b] Spinks G. Uses and methods of crops forecasting. *Review of marketing and agricultural economics*, 1956.

[Stan97] Stan I. and Näescu V. Maize response to water deficit. *Romanian Agricultural Research*, 7 :77–90, 1997.

[Star00] Starr M. The effects of weather on retail sales. *Finance and economics discussion series*, 8, 2000.

[Ston97] Stone P., Wilson D., and Gillespie R. Water deficit effects on growth, water use and yield of sweet corn. *Proceedings Agronomy Society of N.Z.*, 1997.

[Sult09] Sultan B., Bella-Medjo M., Berg A., Quirion P., and Janicot S. Multi-scales and multi-sites analyses of the roleof rainfallin cotton yields in west africa. *International Journal of Climatology*, 30 :58–71, 2009.

[Supi99] Supit I. An exploratory study to improve predictive capacity of the crop growth monitoring system as applied by the european commission. Technical report, Treebook No. 4 ISBN 90-80-443-5-9. Treemail Publishers. Heelsum. The Netherlands, 1999.

[Swan79] Swanson E. and Nyankori J. Influence of weather and technology on corn and soybean yield trends. *Agricultural Meteorology*, 20 :327–342, 1979.

[Takl13] Takle E., Gustafson D., Beachy R., Nelson G., Mason-D'Croz D., and Palazzo A. US food security and climate change : Agricultural futures. *Economics*, 34(Special 7) :1–41, 2013.

[Tan13] Tan C., Samanta A., Jin X., Tong L., Ma C., Guo W., Knyazikhin Y., and Myneni R. Using hyperspectral vegetation indices to estimate the fraction of photosynthetically active radiation absorbed by corn canopies. *International Journal of Remote Sensing*, 34(24) :8789–8802, 2013.

[Tann08] Tannura M. A., Irwin S. H., and Good D. L. *Weather, Technology, and Corn and Soybean Yields in the US Corn Belt*. Marketing and Outlook Res. Rep. No. 2008-01, 2008.

[Tayl97] Taylor S. and Carlson R. Weather and yield trends. In *ICM conference*, pages 175–188, 1997.

[Tayl10] Taylor E. and Todey D. *Weather Effects on Crop Yields*. Iowa State University, 2010.

[Tera11] Terando A., Easterling W., Easterling D., and Keller K. Observed and modeled twentieth-century spatial and temporal patterns of selected agro-climate indices in North America. *Journal of Climate*, 25(2) :473–490, 2011.

[Tera12] Terando A., Keller K., and Easterling W. Probabilistic projections of agro-climate indices in North America. *Journal of Geophysical Research*, 117(D08115) :1–16, 2012.

[Teut12] Teutschbein C. and Seibert J. Bias correction of regional climate model simulations for hydrological climate-change impact studies : Review and evaluation of different methods. *Journal of Hydrology*, 456-457 :12–29, 2012.

[The 06] The international code council, editor. *The International Energy Conservation Code*. ICC, 2006.

[Thom54] Thom H. The rational relationship between heating degree days and temperature. *Monthly Weather Review*, 82(1) :1–6, 1954.

[Thom69] Thompson L. Weather and technology in the production of corn in the u.s. corn belt. *Agronomy Journal*, 61(3) :453–456, 1969.

[Thom75] Thompson L. Weather variability, climatic change, and grain production. *Science*, 188(4188) :535–541, 1975.

[Thom86] Thompson L. Climatic change, weather variability and corn production. *Journal of Agronomy*, 78 :649–653, 1986.

[Thom88] Thompson L. Effects of changes in climate and weather variability on the yield of corn and soybean. *Journal of Agronomy Production*, 1 :20027, 1988.

[Thor08] Thorp K., Dejonge K., Kaleita A., Batchelor W., and Paz J. Methodology for the use of dssat models for precision agriculture decision support. *Computers and Electronics in Agriculture*, 64 :276–285, 2008.

[Tibs96] Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 58 :267–288, 1996.

[Torv04] Torvanger A., Twena M., and Romstad B. Climate change impacts on agricultural productivity in Norway. Technical report, CICERO Working Paper, 2004.

[Tria11] Triantafyllou A. and Sarris A. *Extreme Volatility in Agricultural Commodity MArkets and Implications for Food Security*, chapter 9. Palgrave, 2011.

[Trnk11] Trnka N. et al. Agroclimatic conditions in Europe under climate change. *Global Change Biology*, 17 :2298–2318, 2011.

[Troy06] Troyer A. Adaptedness and heterosis in corn and mule hybrids. *Crop Science*, 46 :528–543, 2006.

[Tubi02] Tubiello F. and Ewert F. Simulating the effects of elevated CO_2 on crops : approaches and applications for climate change. *European Journal of Agronomy*, 18 :57–74, 2002.

[Uppa05] Uppala S. et al. The ERA-40 Re-Analysis. *Quarterly Journal of the Royal Meteorological Society*, 131 :2961–3012, 2005.

[Urba15] Urban D. and Sheffield J .and Lobell D. The impacts of future climate and carbon dioxide changes on the average and variability of US maize yields under two emission scenarios. *Environmental Research Letters*, 10(4) :045003, 2015.

[USDA05] USDA. Global soil regions map. https://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/use/worldsoils/?cid=nrcs142p2_054013, 2005.

[USDA10] USDA. Field crops usual planting and harvesting dates. Technical report, USDA, National Agricultural Statistics Service, October 2010. http://usda.mannlib.cornell.edu/usda/current/planting/planting-10-29-2010.pdf.

[USGC14] USGCRP A. Agriculture. Climate change impacts in the US : the third national climate assessment. Technical report, US Global Change Research Program, 2014. Hatfield, J., G. Takle, R. Grotjahn, P. Holden, R. C. Izaurralde, T. Mader, E. Marshall, and D. Liverman.

[USGC17] USGC. Water-use data available from USGS, 2017. https://water.usgs.gov/watuse/data/.

[Veld10] Velde (van der) M., Wriedt G., and Bouraoui F. Estimating irrigation use and effects on maize yield during the 2003 heatwave in France. *Agriculture, Ecosystems and Environment*, 2010.

[Verb00] Verbeke G. and Molenberghs G. *Linear Mixed Models for Longitudinal Data.* New York, NY : Springer-Verlag., 2000.

[Verg15] Vergara-Diaz O., Kefauver S., Elazab A., Nieto-Taladriz M., and Araus J. Grain yield losses in yellow-rusted durum wheat estimated using digital and conventional parameters under field conditions. *The Crop Journal*, 3(3) :200 – 210, 2015. Special Issue : Breeding to Optimize Agriculture in a Changing World.

[Vert13] Vert J., Schaller N., and Villien C. Agriculture forêt climat : vers des stratégies d'adaptation,. Technical report, Centre d'études et de prospective, Ministère de l'Agriculture, de l'Agroalimentaire et de la Forêt, 2013.

[Viov97] Viovy N. Interannuality and CO_2 sensitivity of the SECHIBA-BGC coupled SVAT-BGC model. *Physics and Chemistry of The Earth*, 21 :489–497, 1997.

[Voge99] Vogel F. and Bange G. Understanding USDA crop forecasts. Miscellaneous Publication No. 1554. Washington, DC : USDA, 1999.

[Voll05] Vollenweider P. and Günthardt-Goerg M. Diagnosis of abiotic and biotic stress factors using the visible symptoms in foliage. *Environmental Pollution*, 137(3) :455–465, 2005.

[Voss90] Vossen P. Rainfall and agricultural production in Botswana. *Afrika Focus*, 1990.

[Voss95] Vossen P. and Rijks D. Early crop yield assessment of the eu countries : The system implemented by the joint research centre. Technical report, Publications of the EC, Luxembourg, 1995.

[Vrac15] Vrac M. and Friederichs P. Multivariate—intervariable, spatial, and temporal—bias correction. *Journal of Climate*, 28(1) :218–237, 2015.

[Vuur11] Vuuren (van) D., Edmonds J., Kainuma M., et al. Special issue : The representative concentration pathways : an overview. *Climatic Change*, 109(1-2), 2011.

[Waha13] Waha K., Coumou D., and Müller C. Extreme weather events and agriculture. From Climate Reconstructions to Climate Predictions., 2013.

[Wall13] Wallach D., Makowski D., Jones J., and Brun F. *Working with Dynamic Crop Models : Methods, Tools, and Examples for agriculture and Environment*. Academic Press, 2013.

[Walt12] Walthal C., Hatfield J., Backlund P., Lengnick L., Marshall E., Walsh M., and Ziska L. Climate change and agriculture in the United States : Effects and adaptation. USDA technical bulletin 1935, Washington, DC : U.S. Department of Agriculture, 2012.

[Warr86] Warrick R., Gifford R., and Parry M. *The Greenhouse Effect, Climatic Change and Ecosystems*, chapter CO_2, Climatic Change and Agriculture. SCOPE 29, John Wiley and Sons, Chichester, 1986.

[Watt03] Watterson I. Simulated changes due to global warming in daily precipitation means and extremes and their interpretation using the gamma distribution. *Journal of Geophysical Research*, 108(D13) :4379, 2003.

[West07] West B., Welch K., and Galecki A. *Linear Mixed Models : a practical guide using statistical software.* Chapman-Hall/CRC., 2007.

[Whis86] Whistler F., Acock B., Baker D., Fye R., Hodges H., Lambert J., Lemmon H., McKinion J., and Reddy V. Crop simulation models in agronomic systems. *Advances in Agronomy.*, 40 :141–208, 1986.

[Whit11] White J., Hoogenboom G., Wilkens P., Stackhouse P., and Hoel J. Evaluation of satellite-based, modeled-derived daily solar radiation data for the continental United-States. *Agronomy Journal*, 148(10) :1574–1584, 2011.

[Wiat10] Wiatrak P. Environmental conditions affecting corn growth. *Retrieved from URL, Clemson University*, 2010. http://www.clemson.edu/extension/rowcrops/corn/guide/environmental_conditions.html.

[Widh14] Widhalm M. Corn growth stages with estimated calendar days and growing-degree units, by R.G. Hall (SDSU), Sep 2014.

[Wiki16] Wikistat. Anova : analyse de variance univariée, 2016. http://wikistat.fr/pdf/st-m-modmixt3-anova.pdf, [En ligne ; Page disponible le 20 décembre 2017].

[Wisi87] Wisiol K. and Hesketh J. *Plant growth modeling for resource management.* CRC Press. Boca Raton Florida USA., 1987.

[Word12] Word Meteorological Organisation. *Guide to Agricultural Meteorological Practices (GAMP).* WMO edition, 2012. http://www.wamis.org/agm/gamp/GAMP_Chap06.pdf.

[Wu13] Wu C., Anlauf R., and Ma Y. Application of the DSSAT model to simulate wheat growth in eastern China. *Journal of Agricultural Science*, 5, 2013.

[Xiji13] Xijie L. *Remote Sensing, Normalized Difference Vegetation Index (NDVI), and crop yield forecasting*. PhD thesis, Graduate College of the University of Illinois at Urbana-Champaign, 2013.

[Yang08] Yang R. Why is mixed analysis underutilized ? *Canadian Journal of Plant Sciences*, 88 :563–567, 2008.

[Yu09] Yu G., Schwartz Z., and Walsh J. A weather resolving index for assessing the impact of climate change on tourism related climate resources. *Climatic Change*, 95(3–4) :551–573, 2009.

[Zhan08] Zhang Y. Improved methods in statistical and first principles modeling for batch process control and monitoring. Master's thesis, University of Texas at Austin, 2008.

[Zhan11] Zhang T. and Huang Y. Impacts of climate change and inter-annual variability on cereal crops in China from 1980 to 2008. *J. Sci. Food Agr.*, 2011.

[Zilb03] Zilberman D., Dinar A., MacDougall N., Khanna M., Brown C., and Castillo F. Individual and institutional responses to drought : The case of californian agriculture. Working Paper, Dept. of Agr. and Res. Econ., University of California, Berkeley., 2003.

[Zou05] Zou H. and Hastie T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 67 :301–320, 2005.

[Zwei93] Zweig M. and Campbell G. Receiver-operating characteristic (ROC) plots : a fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, 39(8) :1589, 1993.

Résumé

En agriculture, la météo est le principal facteur de variabilité d'une année sur l'autre. Cette thèse vise à construire des modèles statistiques à grande échelle qui estiment l'impact des conditions météorologiques sur les rendements agricoles. Le peu de données agricoles disponibles impose de construire des modèles simples avec peu de prédicteurs, et d'adapter les méthodes de sélection de modèles pour éviter le sur-apprentissage. Une grande attention a été portée sur la validation des modèles statistiques. Des réseaux de neurones et modèles à effets mixtes (montrant l'importance des spécificités locales) ont été comparés.

Les estimations du rendement de maïs aux États-Unis en fin d'année ont montré que les informations de températures et de précipitations expliquent en moyenne 28% de la variabilité du rendement. Dans plusieurs états davantage météo-sensibles, ce score passe à près de 70%. Ces résultats sont cohérents avec de récentes études sur le sujet.
Les prévisions du rendement au milieu de la saison de croissance du maïs sont possibles à partir de juillet : dès juillet, les informations météorologiques utilisées expliquent en moyenne 25% de la variabilité du rendement final aux États-Unis et près de 60% dans les états plus météo-sensibles comme la Virginie. Les régions du nord et du sud-est des États-Unis sont les moins bien prédites. Le rendements extrêmement faibles ont nécessité une méthode particulière de classification : avec seulement 4 prédicteurs météorologiques, 71% des rendements très faibles sont bien détectés en moyenne.
L'impact du changement climatique sur les rendements jusqu'en 2060 a aussi été étudié : le modèle construit nous informe sur la rapidité d'évolution des rendements dans les différents cantons des États-Unis et localisent ceux qui seront le plus impactés. Pour les états les plus touchés (au sud et sur la côte Est), et à pratique agricole constante, le modèle prévoit des rendements près de deux fois plus faibles que ceux habituels, en 2060 sous le scénario RCP 4.5 du GIEC. Les états du nord seraient peu touchés.
Les modèles statistiques construits peuvent aider à la gestion sur le cours terme (prévisions saisonnières) ou servent à quantifier la qualité des récoltes avant que ne soient faits les sondages post-récolte comme une aide à la surveillance (estimation en fin d'année). Les estimations pour les 50 prochaines années participent à anticiper les conséquences du changement climatique sur les rendements agricoles, pour définir des stratégies d'adaptation ou d'atténuation. La méthodologie utilisée dans cette thèse se généralise aisément à d'autres cultures et à d'autres régions du monde.

Mots Clés

Modélisation statistique, modèles à effets mixtes, réseaux de neurones, régression et classification, application à la prédiction des rendements agricoles, modèle à large échelle

Abstract

In agriculture, weather is the main factor of variability between two consecutive years. This thesis aims to build large-scale statistical models that estimate the impact of weather conditions on agricultural yields. The scarcity of available agricultural data makes it necessary to construct simple models with few predictors, and to adapt model selection methods to avoid overfitting. Careful validation of statistical models is a major concern of this thesis. Neural networks and mixed effects models are compared, showing the importance of local specificities. Estimates of US corn yield at the end of the year show that temperature and precipitation information account for an average of 28% of yield variability. In several more weather-sensitive states, this score increases to nearly 70%. These results are consistent with recent studies on the subject.
Mid-season maize crop yield forecasts are possible from July: as of July, the meteorological information available accounts for an average of 25% of the variability in final yield in the United States and close to 60% in more weather-sensitive states like Virginia. The northern and southeastern regions of the United States are the least well predicted.

Predicting years for which extremely low yields are encountered is an important task. We use a specific method of classification, and show that with only 4 weather predictors, 71% of the very low yields are well detected on average.
The impact of climate change on yields up to 2060 is also studied: the model we build provides information on the speed of evolution of yields in different counties of the United States. This highlights areas that will be most affected. For the most affected states (south and east coast), and with constant agricultural practice, the model predicts yields nearly divided by two in 2060, under the IPCC RCP 4.5 scenario. The northern states would be less affected.

The statistical models we build can help for management on the short-term (seasonal forecasts) or to quantify the quality of the harvests before post-harvest surveys, as an aid to the monitoring (estimate at the end of the year). Estimations for the next 50 years help to anticipate the consequences of climate change on agricultural yields, and to define adaptation or mitigation strategies. The methodology used in this thesis is easily generalized to other cultures and other regions of the world.

Keywords

Statistical modeling, mixed effects models, neural networks, regression and classification, application to crop yield prediction, large scale model