Université de Montréal

Développement d'un modèle de classification probabiliste pour la cartographie du couvert nival dans les bassins versants d'Hydro-Québec à l'aide de données de micro-ondes passives

par

Mylène Teasdale

Département de mathématiques et de statistique
Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de
Maître ès sciences (M.Sc.)
en Statistique

8 septembre 2015

© Mylène Teasdale, 2015

Université de Montréal
Faculté des études supérieures

Ce mémoire intitulé

Développement d'un modèle de classification probabiliste pour la cartographie du couvert nival dans les bassins versants d'Hydro-Québec à l'aide de données de micro-ondes passives

présenté par

Mylène Teasdale

a été évalué par un jury composé des personnes suivantes :

Pierre Lafaye de Micheaux
(président-rapporteur)

Jean-François Angers
(directeur de recherche)

Mylène Bédard
(membre du jury)

Mémoire accepté le
1er septembre 2015

SOMMAIRE

Chaque jour, des décisions doivent être prises quant à la quantité d'hydroélectricité produite au Québec. Ces décisions reposent sur la prévision des apports en eau dans les bassins versants produite à l'aide de modèles hydrologiques. Ces modèles prennent en compte plusieurs facteurs, dont notamment la présence ou l'absence de neige au sol. Cette information est primordiale durant la fonte printanière pour anticiper les apports à venir, puisqu'entre 30 et 40% du volume de crue peut provenir de la fonte du couvert nival. Il est donc nécessaire pour les prévisionnistes de pouvoir suivre l'évolution du couvert de neige de façon quotidienne afin d'ajuster leurs prévisions selon le phénomène de fonte. Des méthodes pour cartographier la neige au sol sont actuellement utilisées à l'Institut de recherche d'Hydro-Québec (IREQ), mais elles présentent quelques lacunes.

Ce mémoire a pour objectif d'utiliser des données de télédétection en micro-ondes passives (le gradient de températures de brillance en position verticale (GTV)) à l'aide d'une approche statistique afin de produire des cartes neige/non-neige et d'en quantifier l'incertitude de classification. Pour ce faire, le GTV a été utilisé afin de calculer une probabilité de neige quotidienne via les mélanges de lois normales selon la statistique bayésienne. Par la suite, ces probabilités ont été modélisées à l'aide de la régression linéaire sur les logits et des cartographies du couvert nival ont été produites. Les résultats des modèles ont été validés qualitativement et quantitativement, puis leur intégration à Hydro-Québec a été discutée.

Mots clés : statistique bayésienne, mélanges de lois, modèle probabiliste, régression linéaire, logit, neige, télédétection, couvert nival, GTV.

SUMMARY

Every day, decisions must be made about the amount of hydroelectricity produced in Quebec. These decisions are based on the prediction of water inflow in watersheds based on hydrological models. These models take into account several factors, including the presence or absence of snow. This information is critical during the spring melt to anticipate future flows, since between 30 and 40 % of the flood volume may come from the melting of the snow cover. It is therefore necessary for forecasters to be able to monitor on a daily basis the snow cover to adjust their expectations about the melting phenomenon. Some methods to map snow on the ground are currently used at the Institut de recherche d'Hydro-Québec (IREQ), but they have some shortcomings.

This master thesis's main goal is to use remote sensing passive microwave data (the vertically polarized brightness temperature gradient ratio (GTV)) with a statistical approach to produce snow maps and to quantify the classification uncertainty. In order to do this, the GTV has been used to calculate a daily probability of snow via a Gaussian mixture model using Bayesian statistics. Subsequently, these probabilities were modeled using linear regression models on logits and snow cover maps were produced. The models results were validated qualitatively and quantitatively, and their integration at Hydro-Québec was discussed.

Keywords : Bayesian statistics, Gaussian mixture model, probabilistic model, linear regression, logit, snow, remote sensing, snow cover, GTV

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

REMERCIEMENTS

Mes premiers remerciements sont pour mon directeur de recherche, M. Jean-François Angers, qui m'a aidé et soutenu tout au long de ma maîtrise en s'assurant de suivre mes progrès semaine après semaine. Je tiens également à remercier Mme Danielle De Sève et M. Luc Perreault, chercheurs à l'Institut de recherche d'Hydro-Québec, pour m'avoir fourni un projet de recherche motivant et pour m'avoir guidé dans sa réalisation.

Je remercie particulièrement l'organisme MITACS de m'avoir accordé un financement, car cela m'a permis de me concentrer entièrement sur ma recherche sans souci financier.

Je tiens spécialement à remercier mon copain Michel. Je lui suis reconnaissante d'avoir été à mes côtés tout ce temps, d'avoir été patient et de m'avoir encouragée à chaque étape. Je remercie aussi ma famille et ma belle-famille pour m'avoir soutenue et avoir cru en moi.

INTRODUCTION

Hydro-Québec est un meneur nord-américain dans la production d'énergie. Plus particulièrement, il s'agit du principal producteur d'électricité au Canada et du principal producteur d'hydroélectricité au niveau mondial avec une puissance installée à 36 643 mégawatts (MW) (voir www.hydroquebec.com). Sa mission est de produire de l'électricité pour approvisionner le marché québécois et de commercialiser ses surplus sur les marchés de gros. Son parc de production comprend 61 centrales hydroélectriques et une centrale thermique. Les aménagements hydroélectriques comprennent 27 grands réservoirs, 668 barrages et 98 ouvrages régulateurs.

0.1. La situation d'Hydro-Québec et ses besoins

Une telle production d'énergie et une aussi grande quantité d'installations nécessitent que le tout soit géré adéquatement, la production devant s'adapter en tout temps aux fluctuations de la demande via des processus de planification qui consistent à élaborer des stratégies à long, moyen et court terme. En appui à la planification de la production, l'unité HOA (Hydrologie et obligation des affaires) d'Hydro-Québec produit quotidiennement des prévisions d'apports en eau des 200 prochains jours pour plus de 90 bassins versants. Ces prévisions sont essentielles, car elles permettent de prendre des décisions éclairées, efficaces et sécuritaires quant à la planification de la production hydroélectrique (voir De Sève *et al.*, 2008 ; Perreault, 2013). En effet, une meilleure stratégie de production est primordiale pour l'entreprise afin de répondre aux enjeux économiques sur le marché de l'énergie et à ceux en sécurité publique.

La prévision des apports est d'une grande importance et nécessite beaucoup de travail tout au long de l'année, mais un des grands défis est la prévision durant la période de crue printanière. À cette période de l'année, plusieurs facteurs causent la montée des apports en eau, dont les précipitations liquides, mais aussi la fonte du couvert nival qui représente entre 30% à 40% de la réserve hydraulique printanière (voir De Sève *et al.*, 2008, 2012ab). Il est donc primordial pour

Hydro-Québec de bien comprendre le comportement de la couverture de neige et de bien suivre son évolution, puisqu'il s'agit d'un intrant important pour les modèles hydrologiques. Pour ce faire, depuis plusieurs années, différentes activités de recherche sur cette thématique ont été démarrées, notamment en télédétection. Celles-ci sont dans la mouvance de l'intérêt mondial pour la recherche en cartographie de la neige via l'imagerie satellitaire.

0.2. Au niveau mondial

À l'échelle mondiale, la compréhension de l'évolution du couvert nival est d'intérêt depuis plus de 40 ans et plusieurs recherches et projets d'envergure ont été mis en oeuvre. Divers travaux ont été réalisés en utilisant différentes sources de données, majoritairement des données de télédétection provenant de capteurs optiques ou de micro-ondes. Un résumé de ces recherches est présenté, puis leur pertinence pour Hydro-Québec est discutée.

Dietz *et al.* (2012) et Roberge (2013) ont effectué une synthèse des principaux capteurs utilisés et des principales méthodes qui leurs sont appliquées (voir tableau 0.1), mais il existe aussi plusieurs autres techniques pour obtenir une cartographie du couvert nival. Roberge (2014) mentionne notamment celles basées sur les propriétés spectrales de la neige telles que : la classification spectrale supervisée (voir Qobilov *et al.*, 2001) et non supervisée (voir Slater *et al.*, 1999), les réseaux de neurones artificiels (voir Welch *et al.*, 1992 ; Simpson et MnIntire, 2001 ; Tedesco *et al.*, 2004 ; Evora *et al.*, 2008 ; Takala *et al.*, 2008), la modélisation de sous-pixels (voir Rosenthal et Dozier, 1996 ; Kaufman *et al.*, 2002 ; Vikhamar et Solberg, 2003) et les diverses techniques de segmentation qui sont globalement bonnes, mais qui ont une certaine limite (voir Roberge, 2013, 2014). Une des principales raisons expliquant les limitations des méthodes de segmentation est que des seuils fixes n'ont pas d'évolution temporelle. En effet, il ne s'agit que d'une valeur déterminée à l'avance permettant de décider si une mesure représente de la neige ou du sol. De plus, les conditions météorologiques, les propriétés de la neige, les particularités de certains environnements, etc. ont une influence sur le couvert nival, ce qu'un seuil fixe ne peut pas saisir aisément (voir Warren, 1982 ; Foster *et al.*, 1991 ; Walker et Goodison, 1993 ; Hall *et al.*, 1995, 1998 ; Kurvonen *et al.*, 1998 ; Hall *et al.*, 2001 ; Bitner *et al.*, 2002 ; Vikhamar et Solberg, 2003 ; Kelly *et al.*, 2003 ; Goïta *et al.*, 2003 ; Simic *et al.*, 2004 ; Foster *et al.*, 2005 ; Metsämäki *et al.*, 2005 ; Royer *et al.*, 2010 ; Hall *et al.*, 2012 ; Rittger *et al.*, 2013).

En utilisant ces diverses données et méthodes, certains produits opérationnels ont été développés (résumé basé sur Roberge (2013) et Roberge (2014)) :

TABLE 0.1. Synthèse des méthodes de cartographie automatique du couvert nival et leurs caractéristiques (Dietz *et al.* (2012), revue et augmentée par Roberge (2013)).

Capteur	Méthode(s)	Auteur(s)
MODIS	SNOWMAP, NDSI, SNOWFRAC, Image composite AQUA+TERRA, MODSCAG	Hall *et al.* (1995, 2002) ; Vikhamar et Solberg (2003) ; Klein et Barnett (2003) ; Salomonson et Appel (2004, 2006) ; Hall et Riggs (2007) ; Painter *et al.* (2009)
LANDSAT	NDSI, Arbres de décision, SNOWFRAC	Rosenthal et Dozier (1996) ; Vikhamar et Solberg (2003)
AVHRR	SCAMOD	Metsämäki *et al.* (2005)
SMMR	Gradient spectral	Chang *et al.* (1987)
SSM/I	Modèle d'inversion automatique d'émission de la neige	Pulliainen et Hallikainen (2001) ; Derksen *et al.* (2003a,b) ; Goïta *et al.* (2003)
AMSR-E	SWEMAP	Chang et Rango (2000)
GOES + SSM/I	Combinaison de produits	Romanov *et al.* (2000)
AVHRR + SSM/I	Combinaison de produits	Chokmani *et al.* (2009)
MODIS + AMSR-E	Combinaison de produits	Liang *et al.* (2008) ; Gao *et al.* (2010)
MODIS + AMSR-E + QSCAT	ANSA	Foster *et al.* (2011)

(1) un des produits les plus utilisés est la cartographie issue de l'algorithme SNOWMAP qui valorise les capteurs MODIS embarqués sur les satellites TERRA et AQUA qui sont disponibles à une fréquence quotidienne, hebdomadaire et mensuelle. L'algorithme SNOWMAP est une procédure

entièrement automatique qui permet de cartographier autant l'étendue du couvert nival que la sous-fraction de neige (pourcentage de couverture de neige présente dans un pixel) ;

(2) le produit IMS (Ice Mapping System) du NESDIS (National Environmental Satellite Data and Information Service) (voir Ramsay, 1998) utilisait initialement les capteurs AVHRR, GOES et METEOSAT. Des données de micro-ondes ont ensuite été ajoutées afin de bonifier le produit (voir Helfrich *et al.*, 2007) ;

(3) l'agence spatiale européenne (ESA) produit des cartes via le projet GLOBSNOW (voir Luojus *et al.*, 2010 ; Solberg *et al.*, 2010) et utilise le capteur optique ENVISAT-AATSR, le capteur de micro-ondes actives ERS-2 ATSR-2 (voir Metsämäki *et al.*, 2014) et les capteurs de micro-ondes passives SSM/I, SMMR et AMR-E (voir Luojus *et al.*, 2013) ;

(4) le NOHSRC (National Operational Hydrological Remote Sensing Center) produit une cartographie pour les États-Unis et le sud du Canada à l'aide des données des capteurs GOES et AVHRR (voir Bitner *et al.*, 2002) ;

(5) la cartographie produite avec la procédure GOES+SSM/I utilise les données du capteur géostationnaire GOES et celles du capteur de micro-ondes passives SSM/I (voir Romanov *et al.*, 2000).

Un des inconvénients pour Hydro-Québec d'utiliser les produits cités ci-dessus est qu'ils sont surtout développés pour de grands territoires. Puisque le Québec a un couvert nival bien particulier à cause de son climat et de ses zones végétatives, les produits utilisés à grande échelle ne sont pas nécessairement adaptés. Aussi, ces produits sont parfois développés dans une optique de recherche en changement climatique ou pour des analyses de caractérisation de la couverture de neige au niveau planétaire. Ils ne répondent donc pas aux besoins d'Hydro-Québec, qui recherche plutôt une méthode de cartographie qui est précise temporellement et spatialement avec une faible incertitude. Les chercheurs de l'Institut de recherche d'Hydro-Québec développent donc, depuis plusieurs années, différentes méthodes afin de cartographier quotidiennement la neige au sol pour le Québec. Ces méthodes sont décrites dans le premier chapitre. Par contre, un nouvel intérêt a récemment été accru pour la quantification de l'incertitude de la cartographie, ce que ces méthodes ne permettent pas nécessairement de faire.

Le but de ce mémoire de maîtrise, en collaboration avec l'Institut de recherche d'Hydro-Québec, est donc de développer, à l'aide de données de micro-ondes passives, un modèle statistique permettant de cartographier la présence/absence

de neige pour le Québec, tout en quantifiant l'incertitude associée à la prévision pour chacun des points de grille de la cartographie produite.

0.3. Contenu du mémoire

Dans le premier chapitre, le projet est mis en contexte par la présentation des particularités du domaine d'étude ainsi que des données et des méthodes de cartographie actuellement utilisées à l'Institut de recherche d'Hydro-Québec. Par la suite, le gradient de températures de brillance en polarisation verticale (GTV) est calculé. Il s'agit de la principale variable d'intérêt disponible sur tout le territoire et son comportement est étudié à l'aide de statistiques descriptives. Ensuite, les variables exogènes qui sont utilisées comme complément d'information au GTV et qui peuvent être pertinentes pour comprendre le comportement de la neige sont introduites.

Dans le deuxième chapitre, les principales méthodes statistiques utilisées pour développer un modèle de cartographie sont présentées. Tout d'abord, la théorie des mélanges de lois de probabilité est introduite dans un contexte de classification par l'explication des principaux concepts et de l'estimation des paramètres selon la statistique bayésienne. Ensuite, un rappel sur la régression linéaire multiple est effectué par la présentation des concepts de base du modèle, des différents outils pour évaluer la qualité d'un modèle et de la transformation logit.

Le but du troisième chapitre est de construire un modèle permettant de répondre à l'objectif, c'est-à-dire produire des cartographies de la répartition spatiale de la neige en intégrant son incertitude. L'approche utilisée met en relation des probabilités de neige calculées avec les mélanges de lois bayésiens et des variables exogènes à l'aide de la régression linéaire multiple sur les logits. Trois modèles différents sont ainsi obtenus et leurs cartographies respectives sont validées qualitativement.

Dans le quatrième chapitre, la qualité des cartographies produites par les modèles du chapitre 3 est évaluée quantitativement. Les résultats de la validation sont discutés selon certaines données de référence et différents outils de diagnostique pour l'année avec laquelle les modèles ont été construits.

Dans le dernier chapitre, l'intégration des modèles du chapitre 3 dans un contexte prévisionnel est discutée. Tout d'abord, la qualité des modèles est validée pour d'autres années afin de s'assurer qu'ils y offrent aussi des résultats satisfaisants. Par la suite, le choix d'un seul modèle parmi les trois disponibles est abordé et les modifications à lui apporter afin de pouvoir l'utiliser pour prévoir des probabilités de présence de neige pour des journées futures sont présentés.

Chapitre 1

DOMAINE D'ÉTUDE, DONNÉES ET CARTOGRAPHIES EXISTANTES

Ce chapitre a pour but de mettre le lecteur en contexte en présentant, dans un premier temps, le domaine d'étude ainsi que les données et les méthodes de cartographie utilisées pour le suivi du couvert nival à Hydro-Québec. Dans un deuxième temps, les variables utilisées pour mener à bien le projet sont introduites.

1.1. Particularités du domaine d'étude

Le domaine d'étude est le Québec. Il s'agit d'un vaste territoire de 1 667 712 km^2 pour lequel il y a une grande variabilité du climat (continental humide, subarctique, arctique et maritime de l'Est, voir www.gouv.qc.ca), de la température (-40 à 30 °C), de la topographie, de la végétation, etc. En fait, environ 45% de la surface est couverte de forêts décidues (forêts d'arbres dont les feuilles tombent en hiver et sont renouvelées chaque année), de forêts mixtes et de forêts de conifères. De plus, le territoire est caractérisé par d'importantes ressources hydriques, incluant plus de 500 000 lacs et 4 500 rivières (voir www.gouv.qc.ca). Autant de variété fait en sorte que la province est un domaine d'étude très hétérogène. Toutefois, certaines zones du Québec présentent des caractéristiques suffisamment homogènes et ainsi le Québec peut être divisé en quatre zones telles qu'illustrées aux figures 1.1 et 1.2.

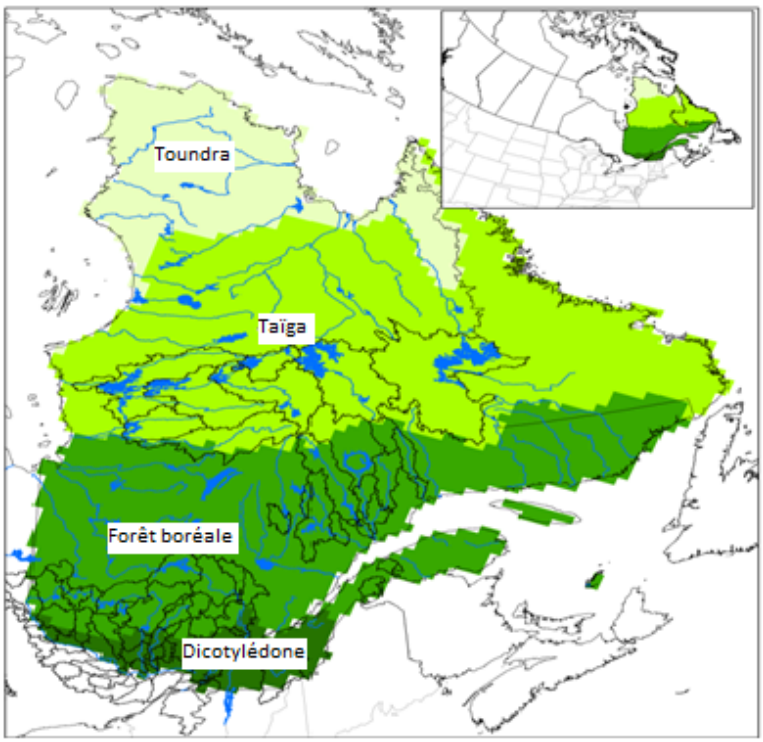

FIGURE 1.1. Domaine d'étude : le Québec, carte provenant de la base de données de l'Institut de recherche d'Hydro-Québec.

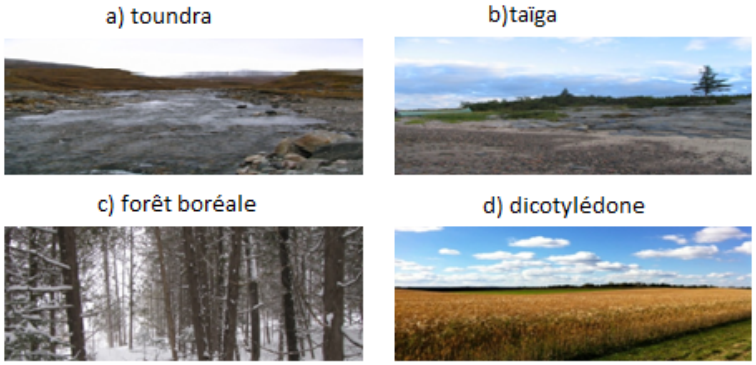

FIGURE 1.2. Illustration des différents types de végétation pour chacune des quatre zones, images fournies par l'Institut de recherche d'Hydro-Québec.

1.2. Forces et faiblesses des données et méthodes utilisées par l'Institut de recherche d'Hydro-Québec pour la cartographie du couvert nival

Comme il a été mentionné en introduction, l'objectif de cartographier le couvert nival du Québec n'est pas nouveau et plusieurs méthodes ont déjà été développées par les chercheurs de l'Institut de recherche d'Hydro-Québec et leurs

collaborateurs. Dans cette section, un bref survol des données qu'ils utilisent est effectué en spécifiant leurs forces et leurs faiblesses, puis les méthodes qui leurs sont associées sont décrites avec une présentation de leurs avantages et inconvénients.

1.2.1. Données de terrain *in situ* et méthodes associées

1.2.1.1. *Lignes et fosses à neige*

Il existe plusieurs techniques permettant de mesurer manuellement la neige au sol, telles que les lignes de neige et les fosses de neige qui permettent d'obtenir la valeur d'équivalent en eau de la neige (ÉEN). Les lignes de neige consistent à prendre des mesures à tous les 20 mètres le long d'un transect de 200 mètres. Ces mesures sont prises une fois par mois de janvier à mars, de façon bimensuelle d'avril à mai ou sur demande (voir De Sève, 1999). Les fosses de neige permettent également d'obtenir de l'information sur la neige, mais elles sont utilisées uniquement en mode recherche par l'Institut de recherche d'Hydro-Québec. Elles permettent de caractériser le couvert nival (description des grains, épaisseur et densité de la neige) verticalement (du bas vers le haut) le long d'une paroi pour un endroit précis du territoire.

La principale force de ces données est qu'elles sont précises pour l'endroit où elles ont été prises. Par contre, le Québec couvrant une très grande superficie (voir section 1.1), il devient difficile d'effectuer des mesures manuelles à la grandeur du territoire, autant pour des raisons économiques que pour l'accessibilité du territoire. Pour des raisons similaires, ces mesures ne sont disponibles que périodiquement, ce qui ne permet pas de suivre l'évolution du couvert nival en temps réel. La quantité d'information est donc assez limitée.

1.2.1.2. *Réseau automatique de hauteur de neige*

Afin d'avoir un suivi de la neige en temps réel, Hydro-Québec et ses partenaires ont aménagé un réseau de stations météorologiques, lesquelles sont munies de sonars de type SR50 qui fournissent notamment une mesure horaire de la hauteur de la neige. En complément à ce réseau, Hydro-Québec dispose d'un réseau de capteurs « Gamma Monitoring » (GMON) fournissant directement des valeurs d'équivalent en eau de la neige pour un pas de temps de six heures.

L'intérêt de ces données est qu'elles sont disponibles en temps quasi-réel (plusieurs fois par jour), ce qui permet de suivre l'évolution de la couverture de neige. Cependant, ces deux réseaux sont peu denses par rapport à la superficie totale du territoire québécois, car les capteurs sont onéreux et il n'est pas possible d'en

installer pour l'intégralité du Québec. Les figures 1.3 et 1.4 illustrent le positionnement des différents capteurs. Il est possible de remarquer qu'il y a une bonne partie du territoire québécois qui n'est pas couverte par les sonars SR50, notamment dans les Appalaches, au centre et au nord du Québec, et que les capteurs GMON sont presque uniquement situés dans le sud. La principale faiblesse des données des capteurs SR50 et GMON est donc leur limite au niveau spatial.

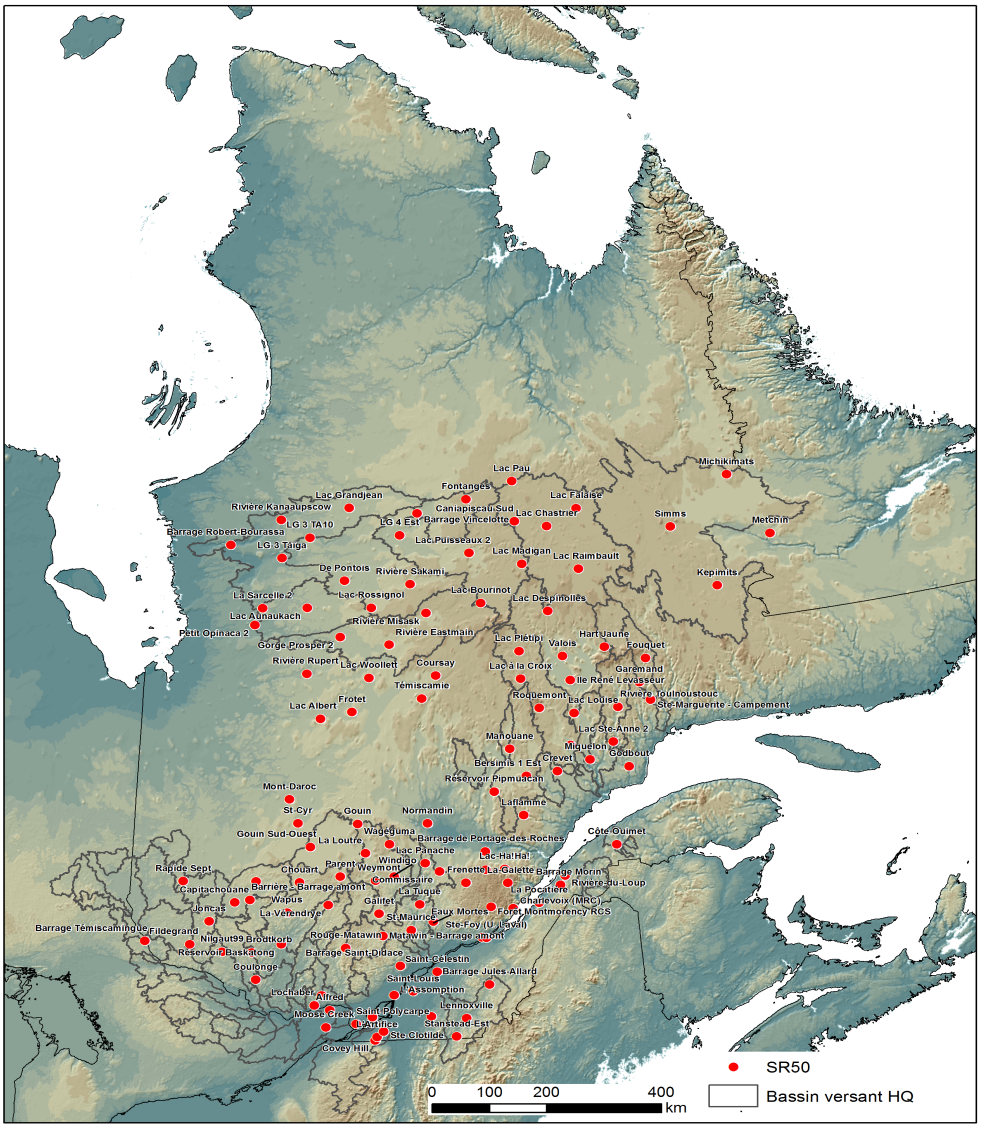

FIGURE 1.3. Location des sonars SR50 sur le territoire québécois, carte provenant de la base de données de l'Institut de recherche d'Hydro-Québec.

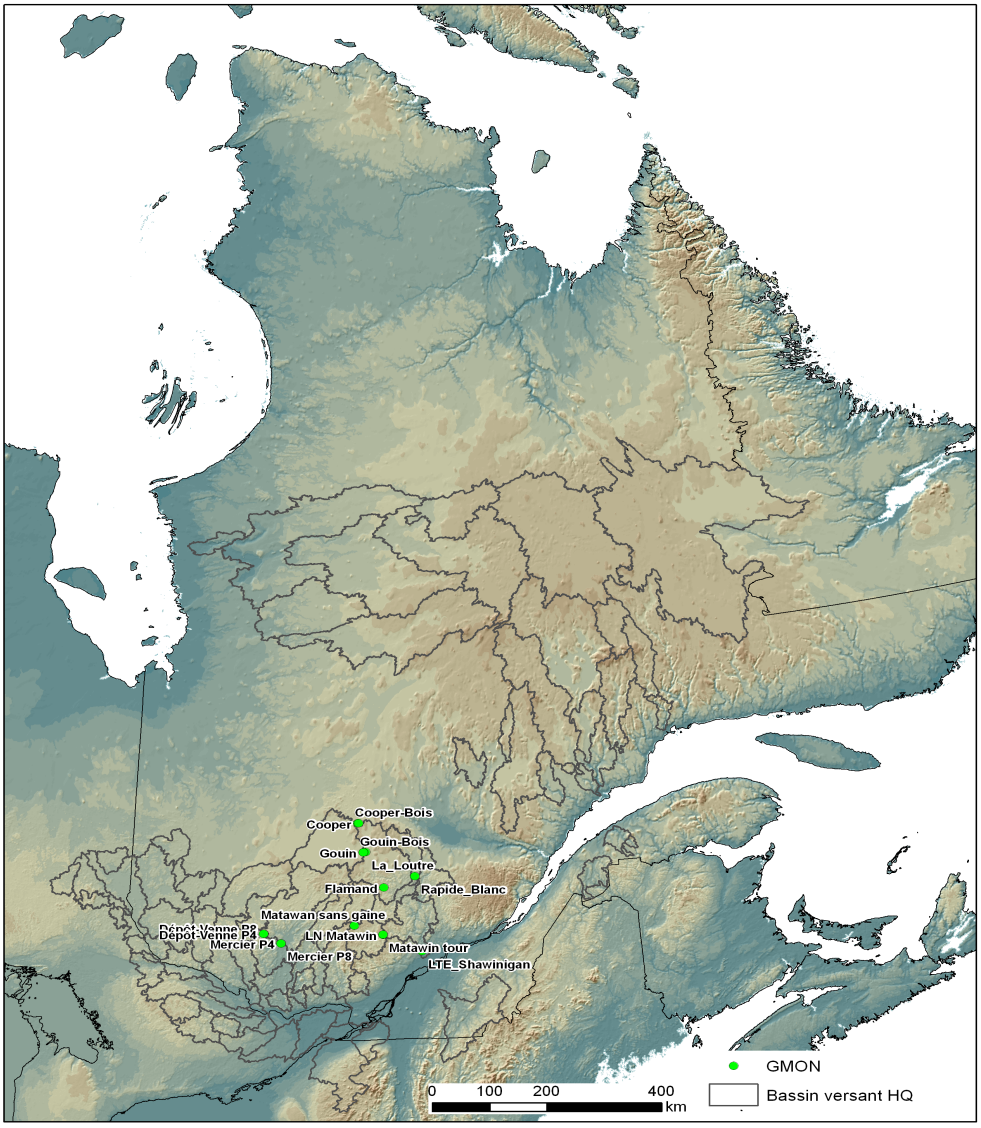

FIGURE 1.4. Location des capteurs GMON sur le territoire québécois, carte provenant de la base de données de l'Institut de recherche d'Hydro-Québec.

1.2.1.3. *Interpolation*

Dans le but de régler la problématique de couverture spatiale associée aux données *in situ* et ainsi avoir un portrait complet de la répartition spatiale de la neige pour le Québec, les chercheurs de l'Institut de recherche d'Hydro-Québec effectuent en mode recherche une interpolation des hauteurs de neige issues des SR50. De cette façon, ils obtiennent une information spatialisée de la hauteur de neige au sol pour la totalité du territoire québécois. Les SR50 fournissant des données en temps-quasi réel, l'information est donc complète au niveau spatio-temporel.

Il est également pertinent de mentionner que HOA réalise une interpolation des données de lignes de neige (densité et équivalent en eau de la neige (ÉEN)).

Un exemple de cartographie obtenue par l'interpolation des hauteurs de neige est illustré à la figure 1.5. Cette carte donne bel et bien une idée de la répartition de la neige et non-neige au Québec.

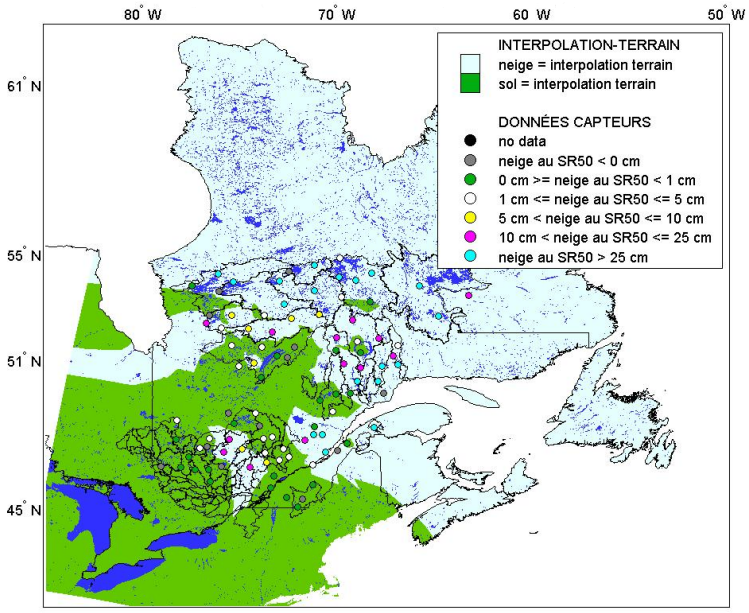

FIGURE 1.5. Exemple de cartographie de la neige suite à l'interpolation des hauteurs de neige pour le 31 mars 2012, carte provenant de la base de données de l'Institut de recherche d'Hydro-Québec.

La principale limitation de cette méthode est que les résultats de l'interpolation ne sont pas nécessairement toujours représentatifs car, bien que les mesures de neige soient assez fiables aux stations, celles-ci peuvent ne pas représenter la réalité des environs. En effet, les sonars peuvent avoir été installés dans un endroit où la végétation n'est pas la même que pour leur voisinage. Par exemple, si un sonar SR50 est situé dans une forêt très dense et qu'il indique la présence de neige, il est possible que, pour une zone dégagée, la neige ait déjà complètement disparu à cause de la dynamique de la cinétique de fonte qui est très différente dans les deux milieux. Cette cinétique de fonte peut également avoir été modifiée artificiellement par la présence d'une dalle de béton sur les sites de mesures, celle-ci pouvant provoquer une fonte prématurée de la neige.

1.2.2. Données de télédétection

Une façon d'obtenir un complément d'information aux données de terrain est l'utilisation de données de télédétection, dont plusieurs méthodes ont été présentées en introduction.

Deux types de données ont été privilégiées par l'Institut de recherche d'Hydro-Québec, soit les données de micro-ondes passives SSM/I issues des satellites américains DMSP (Defense Meteorological Satellite Program) et celles des capteurs optiques AVHRR qui sont embarqués sur le satellite NOAA (National Oceanic and Atmospheric Administration).

1.2.2.1. *Données SSM/I (Special Sensor Microwave/Imager)*

Les données SSM/I sont des températures de brillance (Tb) offertes en plusieurs fréquences (19, 22, 37 et 85 GHz), deux polarisations linéaires (verticale et horizontale) et qui sont obtenues plusieurs fois par jour (voir tableau 1.1).

TABLE 1.1. Caractéristiques des différentes fréquences de températures de brillance, tableau basé sur Hachem (2008).

	Polarisation						
	Verticale (V)				Horizontale (H)		
Fréquence (GHz)	19	22	37	85	19	37	85
Résolution au sol (Km×Km)	70×45	60×40	38×30	16×14	70×45	38×30	16×14
Échantillonnage spatial (Km)	25	25	25	12,5	25	25	12,5

Une température de brillance provient d'un signal émis initialement par le sol et qui est influencé par différentes composantes (atmosphère, neige, etc.) selon la fréquence du signal (voir figure 1.6). Les données de ces capteurs sont traitées pour une résolution de 25 km, ce qui représente 8900 données pour le domaine d'étude (entre 3 000 et 4 000 données lorsque les points d'eau sont ignorés). La couverture spatiale est donc plus intéressante que pour les données *in situ*.

L'intérêt principal des données SSM/I est qu'elles sont disponibles en temps quasi-réel et que certaines fréquences sont sensibles à la présence de neige. De plus, celles-ci sont indépendantes à l'éclairement solaire, transparentes aux nuages et offrent une vision synoptique du territoire. Toutefois, certaines limitations doivent être considérées, puisque les données SSM/I sont avant tout des températures de brillance et non pas directement un indicateur de neige. Elles peuvent donc être influencées par d'autres phénomènes (température, humidité, etc.). Un autre

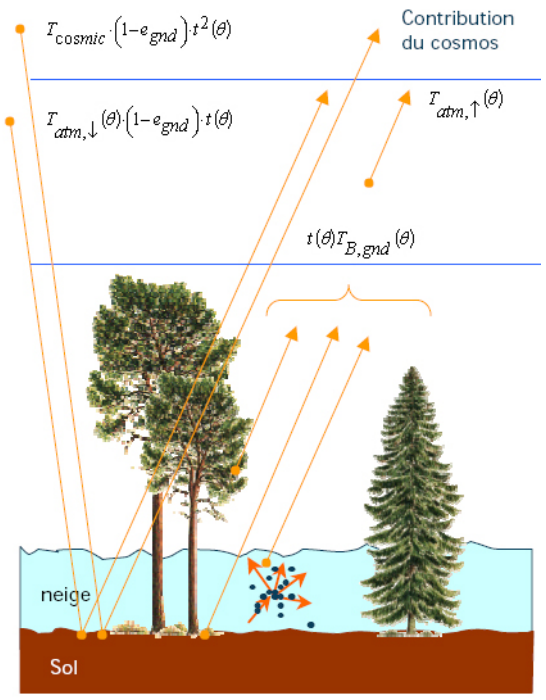

FIGURE 1.6. Schéma du signal pour les températures de brillance (les températures de brillance correspondent aux signaux émis par le sol et non à ceux provenant de l'atmosphère), image en provenance de la base de données de l'Institut de recherche d'Hydro-Québec.

désavantage des données SSM/I est que, bien qu'avoir une information aux points de grille de 25 km de résolution soit une amélioration au niveau de la couverture spatiale, il reste que la résolution est grossière et que certains points de grille peuvent être mixtes. À cette résolution, un même point de grille peut intégrer le signal de plusieurs types d'information, dont une partie couverte de neige et une autre qui ne l'est pas.

Le lecteur intéressé à une description plus approfondie des capteurs SSM/I et des données de micro-ondes passives peut consulter De Sève (1999).

1.2.2.2. *Données AVHRR (Advanced Very High Resolution Radiometer)*

Les capteurs AVHRR sont des radiomètres à miroir rotatif mesurant le rayonnement électromagnétique dans cinq régions spectrales allant du rouge visible à l'infrarouge thermique avec une résolution de 1 km au sol (voir Roberge, 2014). Un avantage des données AVHRR est qu'elles sont aussi disponibles plusieurs fois par jour, assurant ainsi des informations en temps quasi-réel. Aussi, la résolution spatiale de 1 km, qui est beaucoup plus fine que celles des données SSM/I, permet d'obtenir pour une seule image un total de 5 562 500 points de grille (en incluant

les points d'eau). Par contre, tout comme pour les données SSM/I, les données AVHRR peuvent aussi être influencées par des phénomènes exogènes, dont notamment la colonne atmosphérique. De plus, puisqu'il s'agit de données optiques, il est connu qu'elles sont sensibles à la présence de nuages (voir Chokmani *et al.*, 2006, 2009) et il y a donc une perte d'information sur le couvert nival pour les points de grille sous la couverture nuageuse.

1.2.2.3. *Algorithme de segmentation*

Une façon considérée par les chercheurs pour obtenir une cartographie du couvert nival à l'aide des températures de brillance est l'utilisation d'un algorithme de segmentation hiérarchique (voir Chokmani *et al.*, 2006, 2009). Cet algorithme consiste à appliquer une série de tests à des fréquences (respectivement régions spectrales). À chaque étape, le test consiste à comparer une des fréquences (respectivement régions spectrales) avec un seuil fixe ou à les comparer entre elles. Dès qu'un des tests échoue, le point de grille est considéré comme n'étant pas couvert de neige (sol). Toutefois, si tous les tests sont concluants, alors le point de grille est défini comme étant recouvert de neige. Pour les données AVHRR, certains tests peuvent aussi considérer un point de grille comme étant recouvert de nuages. Ces séries de tests sont illustrées sous forme hiérarchique aux figures 1.7 et 1.8 afin d'aider à mieux comprendre le fonctionnement de chaque algorithme. Il est à noter que les fréquences (respectivement régions spectrales) ne sont pas toutes utilisées dans l'algorithme. Par exemple, l'algorithme associé aux données SSM/I n'utilise pas la fréquence à 22 GHz.

Le principal avantage d'utiliser un algorithme de segmentation hiérarchique est que c'est une méthode simple, rapide et facilement implantable. Par contre, elle peut être sensible au choix des seuils, faisant en sorte que des tests échouent si une des données a été influencée par un phénomène exogène, provoquant ainsi de fausses détections.

Dans cette section, il a été vu que l'Institut de recherche d'Hydro-Québec a investigué différentes méthodes pour cartographier la présence/absence de neige sur le sol du Québec : les données de terrain avec interpolation et les données de télédétection de micro-ondes (SSM/I) et optique (AVHRR) avec des algorithmes de seuillage hiérarchique. D'autres travaux ont été réalisés pour essayer de bonifier ces techniques par la combinaison de produits (voir Chokmani *et al.*, 2009 ; Royer *et al.*, 2010 ; De Sève, 2014), pour explorer d'autres aspects du problème (voir Evora *et al.*, 2008 ; Vachon *et al.*, 2010) ou pour effectuer de la cartographie d'ensemble (voir De Sève *et al.*, 2012a ; De Sève, 2014 ; Roberge, 2014). Aussi, une évaluation de l'estimation de l'ÉEN via le jumelage des données SSM/I, un

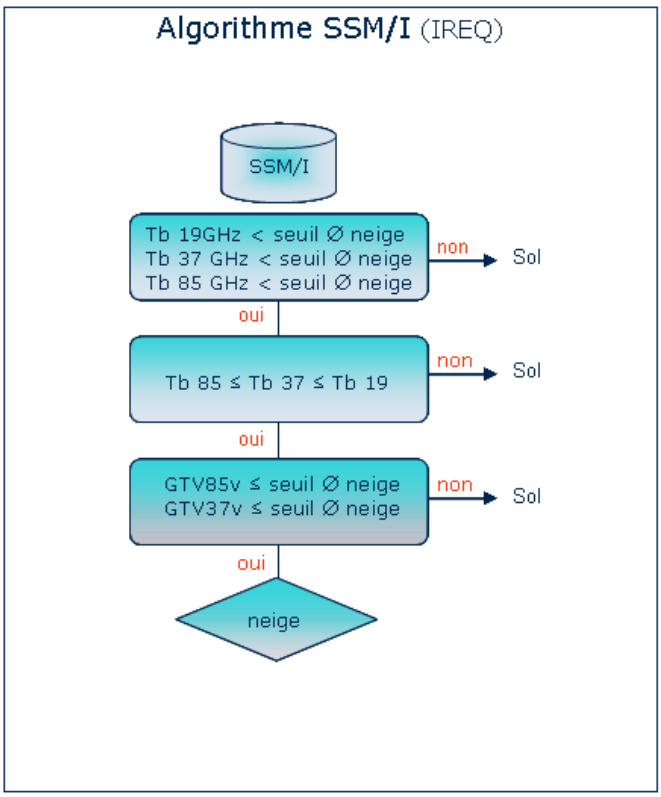

FIGURE 1.7. Représentation visuelle de l'algorithme utilisé par l'IREQ pour traiter les données SSM/I, image provenant de la base de données de l'Institut de recherche d'Hydro-Québec.

modèle de transfert radiatif et les données terrain est actuellement en évaluation chez HOA (voir Vachon *et al.*, 2015). Cependant, ces méthodes ne permettent pas toutes de quantifier l'incertitude associée à leur produit respectif. En parallèle, l'Institut de recherche d'Hydro-Québec et ses partenaires ont regardé la possibilité d'évaluer cette incertitude via une approche de cartographie d'ensemble en bruitant les seuils à l'intérieur d'une fourchette des seuils historiquement plausibles et en reproduisant 100 réalités différentes (voir De Sève *et al.*, 2012a). Par contre, cette technique n'utilise pas de méthode statistique.

Dans la section qui suit, les variables utilisées pour développer une nouvelle technique de cartographie pour laquelle il est possible de quantifier l'incertitude à l'aide de méthodes statistiques sont présentées.

1.3. VARIABLES À CONSIDÉRER

L'objectif de cette section est de présenter les variables utilisées pour le présent mémoire. Tout d'abord, la principale variable d'intérêt pour la modélisation est

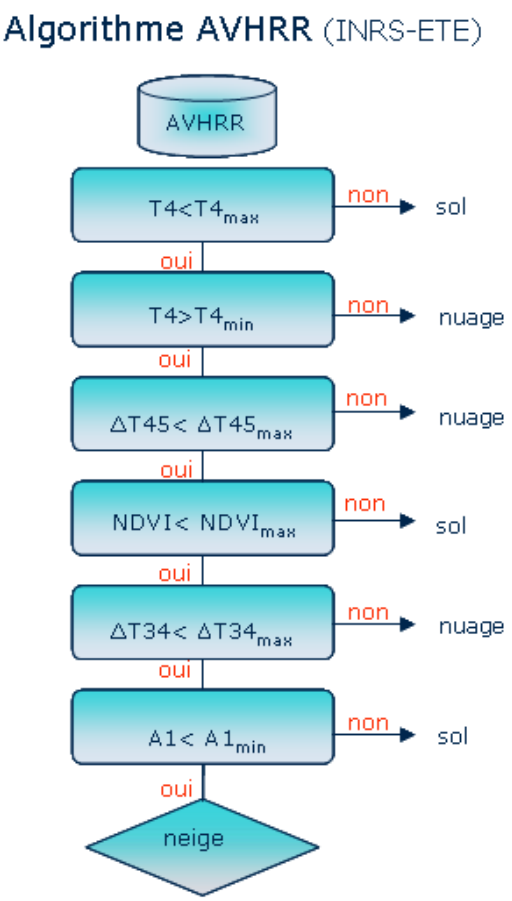

FIGURE 1.8. Représentation visuelle de l'algorithme utilisé par l'IREQ pour traiter les données AVHRR, image en provenance de la base de données de l'Institut de recherche d'Hydro-Québec et basé sur Chokmani *et al.* (2006).

définie, puis des variables exogènes pouvant fournir un complément d'information sont introduites.

Suite aux constats de la section précédente, les données SSM/I ont été privilégiées, car elles offrent une information assez complète au niveau spatio-temporel. Bien qu'il ait pu être préférable de choisir les données des capteurs AVHRR pour leur résolution plus fine, ces données n'ont pas été retenues à cause de leur sensibilité face à la présence de nuages. Cette sensibilité est peu pratique, car les prévisionnistes d'Hydro-Québec s'intéressent fortement à l'évolution du couvert nival en période de fonte des neiges et il serait impossible de suivre cette évolution dû à la forte présence de nuages au printemps.

1.3.1. Gradient de températures de brillance en polarisation verticale (GTV)

Les données SSM/I sont des valeurs de températures de brillance (Tb) disponibles pour plusieurs fréquences et deux polarisations linéaires (voir section 1.2.2.1). Pour la réalisation de ce projet, seules les températures de brillance aux fréquences à 19 et 37 GHz en polarisation verticale (V) ont été retenues (voir Chang *et al.*, 1987).

L'avantage d'utiliser la fréquence à 37 GHz est qu'il s'agit d'une fréquence qui est transparente aux composantes atmosphériques, mais aussi sensible à la présence de neige au sol (voir De Sève, 1999). En effet, une partie du signal émis initialement par le sol est diffusée dans le stock de neige (diffusion de volume) par les cristaux de neige (glace). Une partie du signal ne sera donc pas interceptée par le capteur SSM/I. Cette diffusion du signal dans la neige est ce qui est nommée le signal de la neige. Cette perte de signal est inversement proportionnelle à la hauteur de neige, donc la température de brillance en fréquence à 37 GHz prendra des valeurs plus faibles en présence d'une forte accumulation de neige et des valeurs plus fortes pour des faibles accumulations ou une absence de neige au sol. Précisons toutefois qu'à cette fréquence, le signal est aussi sensible à la structure du couvert nival (diamètre des cristaux de neige, présence de croûtes de glace) et à la présence d'eau liquide dans la neige (neige humide).

L'intérêt d'utiliser la température de brillance à 19 GHz est qu'elle permet d'obtenir de l'information complémentaire pour la cartographie de la neige. En effet, comme le signal n'est pas sensible à l'atmosphère et la neige, mais plutôt à d'autres phénomènes comme la température et les caractéristiques du sol, il permet d'isoler les fluctuations du signal qui leurs sont dues (voir De Sève, 1999).

Le choix de la polarisation verticale est justifié par le fait que celle-ci est la moins sensible à la réflexion de surface, donc à la structure du couvert nival (voir Ulaby *et al.*, 1986 et De Sève, 1999).

En connaissance du comportement des températures de brillance à 19 et 37 GHz en polarisation verticale, une transformation assez courante dans le domaine de la télédétection est la création du gradient de températures de brillance en polarisation verticale (GTV) (voir Chang *et al.*, 1987) :

$$GTV = \frac{Tb37V - Tb19V}{37 - 19} = \frac{Tb37V - Tb19V}{18}. \tag{1.3.1}$$

L'intérêt de cette transformation est qu'elle permet de soustraire les fluctuations dues aux autres phénomènes tels que la température et le sol de celles dues à la

neige. Il s'agit donc d'une variable dont le signal est affranchi des effets autres que ceux de la neige.

1.3.1.1. *Le GTV en théorie*

En théorie, lorsque la neige est homogène (même densité et aucune variation du diamètre et de la forme du grain), le GTV est une variable prenant ses valeurs dans $]-\infty, 0]$, puisqu'en l'absence de neige la valeur des Tb à 19V et 37V GHz sont identiques. Une valeur de zéro pour le GTV signifie donc l'absence de neige au sol. De plus, puisque la valeur du Tb37V diminue avec l'accumulation de neige, mais que la valeur du Tb19V ne change pas, une valeur négative du GTV est un indicateur de la présence de neige au sol. Une représentation graphique de ce comportement théorique est présentée à la figure 1.9.

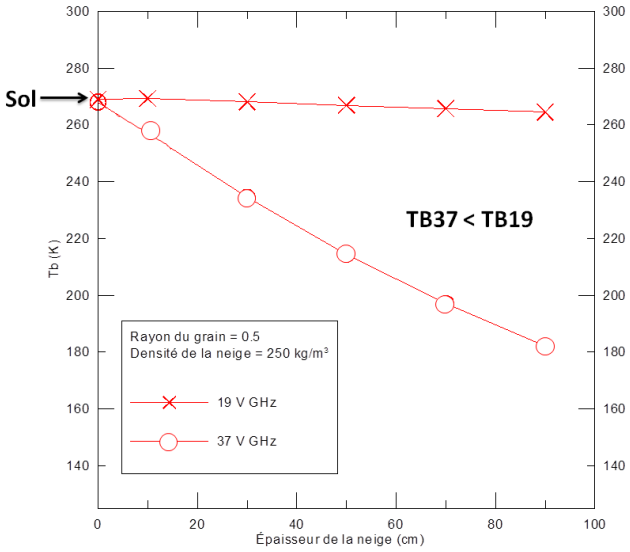

FIGURE 1.9. Comportement théorique du GTV lorsque la neige a un grain de rayon de 0,5 mm et une densité de 250kg/m^2, modification d'une image en provenance de la base de données de l'Institut de recherche d'Hydro-Québec.

.

1.3.1.2. *Le GTV dans des conditions réelles*

Bien que le GTV semble être un bon indicateur pour identifier la présence/absence de neige, l'utilisation d'un seuil valant zéro n'est valide que dans un cadre théorique. Dans des conditions réelles, la transformation n'est pas suffisante pour éliminer toutes les fluctuations qui ne sont pas dues à la neige. La figure 1.10 ci-dessous représente le véritable comportement du GTV pour le Québec, selon les différentes saisons de l'année 2011.

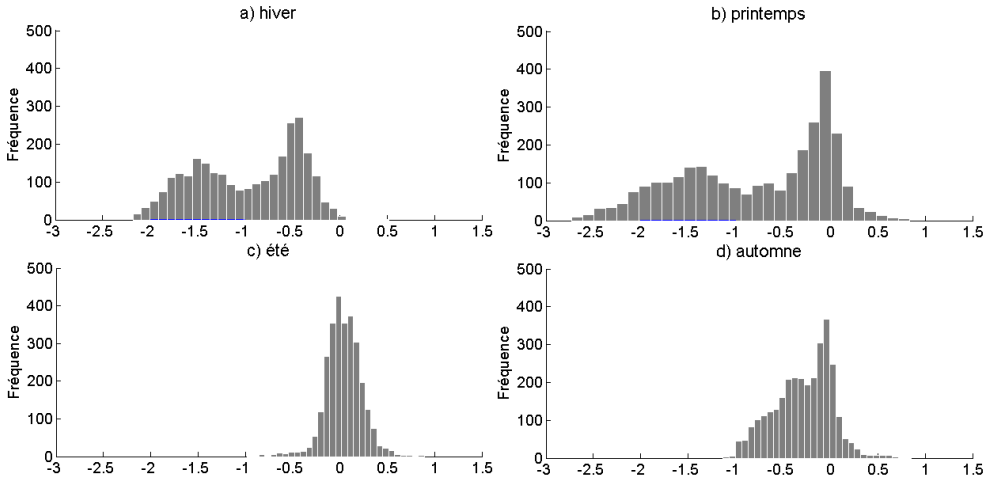

FIGURE 1.10. Histogramme du GTV selon les différentes saisons pour 2011, Québec.

.

À la vue de ces histogrammes, il est possible de constater que les valeurs du GTV s'étendent plutôt dans l'intervalle $[-3; 1, 5]$, où les valeurs fortement négatives représentent la neige et les valeurs avoisinantes de zéro représentent le sol, c'est-à-dire la non-neige. En effet, à la figure 1.10 a), il est possible de remarquer un comportement typique du GTV en hiver, lequel est caractérisé par des valeurs majoritairement négatives et peu près de zéro, indiquant de fortes épaisseurs de neige ou des grains de neige de grand diamètre. Aux figures 1.10 b) et d), il est possible de remarquer des valeurs à la fois négatives et près de zéro, ce qui illustre respectivement un couvert nival installé et le début de la fonte au printemps (b) et un sol sans neige avec un début d'accumulation à l'automne (d). Finalement, à la figure 1.10 c), l'absence de neige est caractérisée par des valeurs assez concentrées près de zéro.

La figure 1.11 présente une illustration du comportement spatialisé des valeurs printanières du GTV pour le domaine d'étude. Il est possible d'y observer le comportement identifié à l'aide des histogrammes, c'est-à-dire qu'il y a des valeurs très négatives (couleurs froides) et des valeurs proches de zéro (couleurs chaudes) lors de la fonte au printemps. La répartition des GTV est donc cohérente avec la réalité de la période de fonte, car il y a de fortes radiations en micro-onde au sud du Québec, lequel est déjà dégagé de neige, et de plus faibles au nord, qui est toujours couvert d'un manteau neigeux (il faut noter que la zone blanche triangulaire représente des données manquantes).

Le seuil théorique de zéro n'étant pas adéquat dans les conditions réelles pour discriminer la neige et le sol, l'Institut de recherche d'Hydro-Québec a exploré l'utilisation d'autres seuils (voir De Sève, 2011). Les travaux subséquents ont démontré que l'emploi d'un seuil fixe n'est pas complètement satisfaisant et que

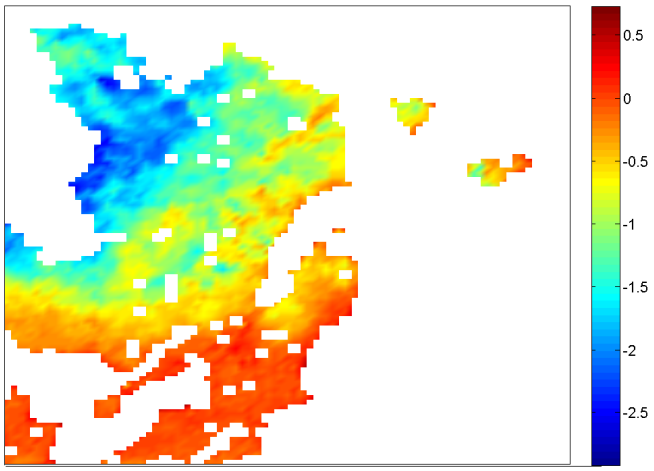

FIGURE 1.11. Valeurs printanières du GTV spatialisées pour le Québec, 19 avril 2011.

.

l'utilisation de seuils dynamiques pourrait être avantageuse (voir De Sève *et al.*, 2012a). Cela implique donc que les seuils soient variables dans le temps afin de s'adapter aux conditions environnementales changeantes. Ces seuils peuvent être assez ardus à choisir, puisqu'ils peuvent varier de jour en jour à l'intérieur d'une année, d'année en année, en plus d'une possible variation à l'intérieur du domaine d'étude.

Ayant souligné plusieurs problématiques liées à l'utilisation d'un seuil de neige fixe, il a été décidé que les modèles de cartographie ne devront pas dépendre de tels seuils.

1.3.2. Statistiques descriptives

Le GTV étant la variable d'intérêt pour la cartographie de la neige, il est important de comprendre son comportement avant de passer à l'étape de la modélisation. Pour ce faire, les statistiques considérées sont le minimum, le percentile 5%, la moyenne, la médiane, le percentile 95% et le maximum de tous les points de grille par jour julien. Le minimum et le maximum permettent de visualiser l'évolution de l'étendue des valeurs du GTV, la moyenne et la médiane permettent de voir l'évolution de la tendance centrale et les percentiles 5% et 95% nous informent sur la concentration des valeurs les plus éloignées de la moyenne et de la médiane. La figure 1.12 illustre ces statistiques descriptives pour les valeurs de GTV de l'année 2011 selon les zones du Québec présentées à la section 1.1. La constatation la plus intéressante est que l'année peut être divisée en quatre périodes qui sont décalées dans le temps. Ces périodes ne correspondent pas

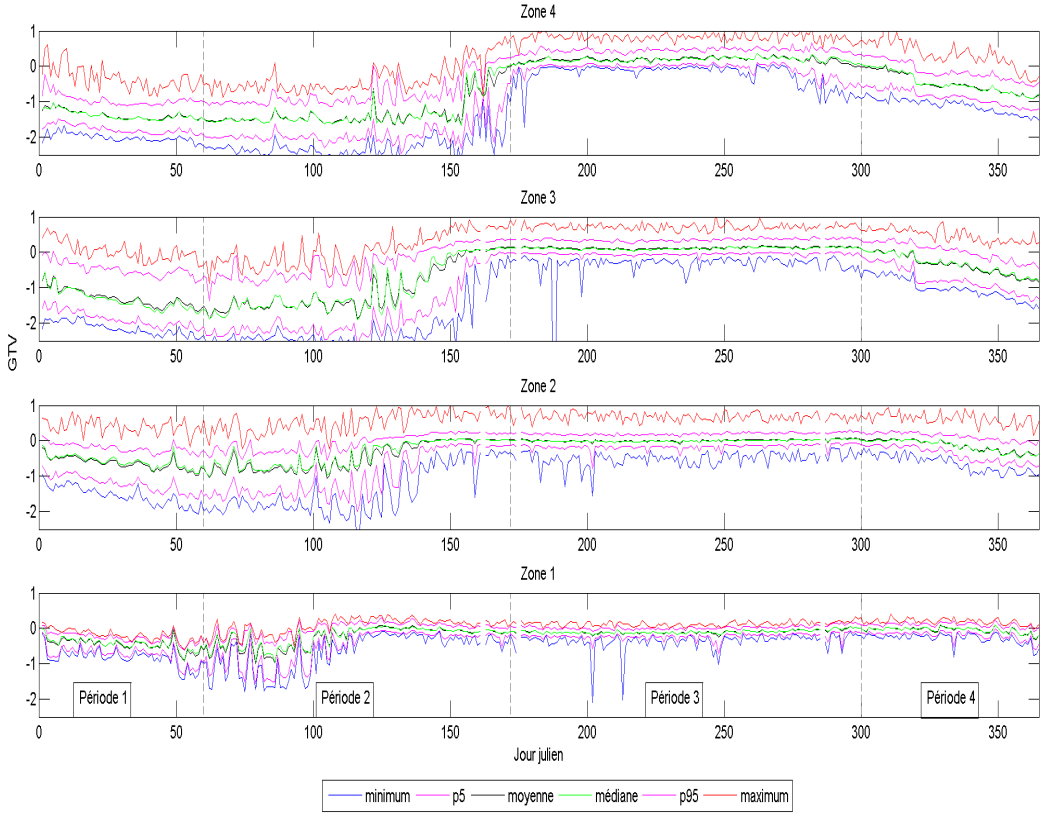

FIGURE 1.12. Minimum, p5%, moyenne, médiane, p95% et maximum du GTV par jour julien, 2011, Québec.

.

exactement aux saisons, mais plutôt à différentes étapes de l'évolution du couvert nival.

1.3.2.1. *Période 1 : période de neige*

La première période est caractérisée par la présence de neige sur l'intégralité du domaine d'étude. Les statistiques descriptives y sont relativement stables et la moyenne ainsi que la médiane sont sous le seuil de zéro. De plus, les percentiles 5% et 95% sont assez proches respectivement du minimum et du maximum, ce qui indique que le minimum et le maximum ne sont pas des valeurs extrêmes.

1.3.2.2. *Période 2 : fonte de la neige*

Nous attribuons à la deuxième période l'étiquette de période de fonte, laquelle comprend une alternance entre la fonte (neige humide) et le regel (croûte) et se termine par une fonte finale qui s'enclenche dès que la couverture nivale est isotherme. Une de ses caractéristiques est qu'il y a une augmentation de l'étendue,

donc de la variabilité du GTV. Cette forte variabilité peut s'expliquer par les épisodes de fonte (valeurs près de 0), de regel (valeurs très négatives) et au final par la disparition complète du couvert nival (valeurs près de 0). Aussi, pour la majorité des zones, les percentiles 5% et 95% sont plus éloignés respectivement du minimum et du maximum, ce qui indique que les valeurs minimales et maximales s'éloignent de la tendance générale des autres valeurs du GTV. De plus, la moyenne et la médiane augmentent graduellement, passant des valeurs négatives pour se stabiliser autour de zéro.

1.3.2.3. *Période 3 : période de sol*

La troisième période, qui contient en partie l'été, correspond à celle où il n'y a plus de neige au sol. Il s'agit de la période où les valeurs de GTV ont la plus faible étendue, mais où les percentiles 5% et 95% sont généralement les plus éloignés respectivement du minimum et du maximum en étant très rapprochés de la moyenne et de la médiane. Cela indique que les valeurs minimales et la maximales sont assez différentes de la tendance générale du GTV et, qu'outre ces valeurs, le GTV est très peu variable dû à l'absence de neige.

1.3.2.4. *Période 4 : accumulation de la neige*

La dernière période est celle que nous identifions comme étant l'accumulation de la neige. Elle débute lorsque les températures de l'air baissent, mais que la neige est toujours absente, et se termine lorsque le sol est entièrement recouvert de neige. Précisons que les accumulations ne sont pas nécessairement importantes. Ses caractéristiques sont similaires à celles de la deuxième période, mais les comportements sont inversés. En effet, la période est définie par une augmentation de l'étendue, mais les percentiles 5% et 95% se rapprochent respectivement du minimum et du maximum au lieu de s'en éloigner. De plus, la moyenne et la médiane diminuent pour se stabiliser avec des valeurs négatives au lieu d'augmenter pour se stabiliser à zéro.

1.3.2.5. *Différence entre les zones*

La figure 1.12 illustre la variation du GTV dans le temps et l'espace. Plus spécifiquement, elle s'exprime par le décalage dans le temps de chacune des périodes en allant du sud vers le nord. Par exemple, la fonte débute plus tôt au sud du Québec (zone 1 : environ jours juliens 75 à 115) qu'au nord (zone 4 : environ jours juliens 150 à 175).

1.3.3. Autres variables

Le GTV étant calculé à partir de données des capteurs SSM/I, il en présente sensiblement les mêmes avantages et limitations. Une force du GTV est donc sa résolution en 25 km et la disponibilité en temps quasi-réel des Tb à 19V et 37V GHz, ce qui assure une bonne couverture spatio-temporelle. Par contre, bien que le calcul du GTV annule certains effets externes par la soustraction des deux fréquences (voir équation (1.3.1)), les données sont quand même sensibles à des facteurs exogènes. C'est pourquoi il est primordial de considérer d'autres variables pouvant apporter de l'information supplémentaire telle que la température de l'air, les degrés-jours et les degrés-chauffes.

1.3.3.1. *Température de l'air*

Une des variables à considérer pour expliquer le comportement de la neige est la température de l'air, puisqu'il existe un lien entre cette variable et la présence/absence de neige et l'état du couvert nival. Une température froide pourrait donc être un indicateur de présence ou d'accumulation de neige et une température chaude pourrait être un indicateur d'absence ou de fonte de la neige.

L'utilisation de la température est également justifiée par le fait que les valeurs de GTV y sont sensibles. Par exemple, s'il y a un redoux en hiver, la neige va se comporter comme un corps noir et les températures de brillance seront très élevées, peu importe la fréquence. Les valeurs élevées du GTV indiqueraient donc l'absence de neige, alors que le couvert nival est encore important (voir De Sève, 1999).

Pour la réalisation de ce projet, nous avons utilisé une base de données de température de l'air interpolées aux stations météorologiques d'Environnement Canada pour une résolution de 25 km par des chercheurs de l'Institut de recherche d'Hydro-Québec. Celles-ci sont disponibles de façon périodique pour chaque point de grille (j mesures par jour). Étant donné que l'objectif est de cartographier le couvert nival de façon journalière et non pour un moment particulier de la journée, nous avons calculé une température quotidienne à partir des j mesures périodiques :

Définition 1.3.1. *Pour une journée t donnée, les variables $Tmin_t$, $Tmoy_t$ et $Tmax_t$ sont définies comme étant :*

Tmin$_t$: *minimum des j températures périodiques de la journée t*

Tmax$_t$: *maximum des j températures périodiques de la journée t*

Tmoy$_t$: *moyenne de $Tmin_t$ et $Tmax_t$*

1.3.3.2. *Degré-jour*

Une autre variable pertinente est le degré-jour, car elle donne de l'information sur le mûrissement de la neige. Il s'agit du nombre de degrés accumulés au-dessus d'un seuil sur un certain nombre de jours. Il est défini comme suit.

Définition 1.3.2. *Soit t la journée étudiée, s le seuil de fonte de la neige, $n = \{1, 2, 3\}$, la statistique pouvant être respectivement le minimum, la moyenne ou le maximum et soit $T_{t,n}$ la température de nature n pour le jour t. Le degré-jour calculé sur x jours est défini ainsi :*

$$\mathrm{DJ}_t(x; s; n) = \sum_{i=t-x+1}^{t} \mathbf{1}_{[s,+\infty[}(T_{i,n}) \times (T_{i,n} - s) \tag{1.3.2}$$

Dans le cadre du projet, c'est majoritairement la variable $\mathrm{DJ}_t(x; 0; 3)$ qui sera utilisée. Afin que la notation soit moins complexe, elle sera dénotée DJx.

1.3.3.3. *Degré-chauffe*

De façon similaire, la variable degré-chauffe peut amener une information pertinente. Il s'agit du nombre de degrés accumulés au-dessous d'un seuil sur un certain nombre de jours. Il est défini comme suit :

Définition 1.3.3. *Soit t la journée étudiée, s le seuil de fonte de la neige, $n = \{1, 2, 3\}$, la statistique pouvant être respectivement le minimum, la moyenne ou le maximum et soit $T_{t,n}$ la température de nature n pour le jour t. Le degré-chauffe calculé sur x jours est défini ainsi :*

$$\mathrm{DC}_t(x; s; n) = \sum_{i=t-x+1}^{t} \mathbf{1}_{]-\infty,s]}(T_{i,n}) \times (s - T_{i,n}) \tag{1.3.3}$$

Pour le présent mémoire, c'est majoritairement la variable $\mathrm{DC}_t(x; 0; 1)$ qui sera utilisée. Afin que la notation soit moins complexe, elle sera dénotée DCx.

1.4. Conclusion partielle

Dans ce chapitre, il a été question du contexte général dans lequel le projet s'inscrit. Le Québec a tout d'abord été présenté comme étant un domaine d'étude très hétérogène au niveau des conditions climatiques et végétales. Pour ces raisons, il a donc été divisé en quatre zones bioclimatiques plus homogènes.

Par la suite, une révision des principales données et méthodologies explorées par les chercheurs de l'Institut de recherche d'Hydro-Québec et leurs collaborateurs a été effectuée. Il a été présenté que les mesures manuelles ont des inconvénients au niveau spatio-temporel. Pour bonifier l'aspect temporel, les stations SR50 et GMON permettent d'obtenir des mesures de neige en temps quasi-réel, mais elles ne sont pas en quantité suffisante, malgré l'interpolation, pour satisfaire l'aspect spatial. Afin de répondre aux besoins spatiaux, les données de télédétection SSM/I et AVHRR sont satisfaisantes. De plus, leurs radiations sont très sensibles à la présence de neige. Par contre, ces données ont aussi certaines limites, dont la sensibilité à des variables exogènes. Pour ce qui de l'utilisation des données de télédétection pour obtenir de l'information neige/non-neige, les méthodes de segmentation hiérarchique employées sont globalement performantes, mais ne répondent pas aux objectifs qui ont été vus en introduction.

Finalement, il a été décidé de travailler avec les données SSM/I, car elles ne sont pas sensibles à la présence de nuages, ce qui a amené à introduire le GTV comme principale variable d'intérêt pour la réalisation de ce projet. Le chapitre se conclut sur la présentation de variables exogènes pouvant être utilisées en complément au GTV pour comprendre l'évolution du couvert nival, c'est-à-dire la température de l'air, les degrés-jours et les degrés-chauffes.

Chapitre 2

MÉTHODES STATISTIQUES

Dans ce chapitre, les principales méthodes statistiques utilisées dans le cadre de ce projet sont introduites par la présentation des différents éléments théoriques nécessaires à la compréhension des modèles de mélanges de lois et de la régression linéaire multiple tel que nous les avons utilisés pour atteindre les objectifs définis en introduction.

2.1. Mélange de lois de probabilité

Cette section a pour but de familiariser le lecteur avec les mélanges de lois de probabilité et de justifier leur utilisation dans le cadre de ce projet. Ensuite, l'algorithme de Gibbs auquel nous avons recours pour mettre en oeuvre l'estimation bayésienne des paramètres d'un mélange de lois est présenté et l'évaluation de la qualité d'un ajustement est abordée.

2.1.1. Définition

2.1.1.1. *Mélange de lois unidimensionnel à K composantes*

Un mélange de K distributions de probabilité est un modèle paramétrique basé sur l'hypothèse que les données d'un échantillon proviennent de différentes sous-populations (ou composantes). Ce mélange s'exprime comme une moyenne pondérée de densités de probabilité :

$$g(y_i|\boldsymbol{w}, \boldsymbol{\theta}) = \sum_{k=1}^{K} w_k f_k(y_i|\boldsymbol{\theta}_k),$$

où g représente la densité du mélange, y_i est une observation, $i = 1, \ldots, n$, n est le nombre d'observations, $\boldsymbol{\theta}_k$ est le vecteur des paramètres de la densité f_k et $\boldsymbol{w} = (w_1, w_2, \ldots, w_K)^T$ est le vecteur contenant les probabilités qu'une observation provienne de la distribution k.

Puisque les modèles de mélanges de lois normales sont beaucoup utilisés à l'Institut de recherche d'Hydro-Québec et que quelques essais ont permis d'établir qu'ils ajustent assez bien les données de ce projet, le reste de la section est adaptée pour l'utilisation de lois normales :

$$g(y_i|\boldsymbol{w},\boldsymbol{\theta}) = \sum_{k=1}^{K} w_k f_k(y_i|\boldsymbol{\theta}_k), \tag{2.1.1}$$

où $\boldsymbol{\theta}_k = (\zeta_k, \phi_k)^T$ et $f(\cdot|\boldsymbol{\theta}_k)$ représente la densité normale avec un paramètre de moyenne ζ_k et de précision ϕ_k (qui est l'inverse de la variance).

2.1.1.2. *Introduction de la variable latente* $\boldsymbol{z}_i$

Étant donné qu'il a été supposé que les données proviennent de différentes sous-populations, il peut être utile de conserver l'information sur l'origine de chacune d'entre elles. Pour ce faire, il faut introduire la variable latente :

$$\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iK}) \sim Mult_K(1|w_1, \ldots, w_K),$$

où K est le nombre de sous-populations distinctes, $\sum_{k=1}^{K} w_k = 1$, $i = 1, \ldots, n$ et

$$z_{ik} = \begin{cases} 1 & \text{si l'observation } y_i \text{ provient de la sous-population } k; \\ 0 & \text{sinon,} \end{cases}$$

pour $k = 1, \ldots, K$.

La notation $Mult_K(1|w_1, \ldots, w_K)$ représente une loi multinomiale de dimension K et de paramètres $(w_1, \ldots, w_K)$. Dans le présent contexte, un seul des z_{ik} peut valoir 1, où $P(z_{ik} = 1) = w_k$.

2.1.1.3. *Densité a priori*

Pour estimer les paramètres du modèle, nous considérons la perspective bayésienne. Il faut donc établir des lois *a priori* pour les paramètres $\boldsymbol{w}$ et $\boldsymbol{\theta}_k = (\zeta_k, \phi_k)$. Afin de faciliter les calculs, il a été décidé d'utiliser des densités *a priori* conjuguées.

Puisque les $\boldsymbol{z}_i$ suivent une loi multinomiale de paramètres $(1|w_1, \ldots, w_K)$, la densité *a priori* conjuguée pour $\boldsymbol{w}$ est une loi de Dirichlet (voir Robert, 2007) :

$$\boldsymbol{w} \sim Dirichelet(\beta_1, \ldots, \beta_K),$$

où les $\beta_k, k = 1, \ldots, K$, sont des hyperparamètres qui devront être spécifiés ultérieurement.

Il a été supposé à la section 2.1.1.1 que les observations y_i proviennent de lois

normales, donc

$$y_i|\boldsymbol{z}_i = k, \boldsymbol{w}, \boldsymbol{\theta}_k \sim \mathcal{N}(\zeta_k, \phi_k).$$

Les densités conjuguées à utiliser pour les paramètres ζ_k et ϕ_k sont respectivement une loi normale et une loi gamma (voir Evin *et al.*, 2011) :

$$\zeta_k|\phi_k \sim \mathcal{N}\left(\mu_\pi, n_\pi\phi_k\right);$$
$$\phi_k \sim \mathcal{G}\left(\alpha_\pi, \gamma_\pi\right),$$

où μ_π, n_π, α_π et γ_π sont des hyperparamètres qui devront être spécifiés ultérieurement.

2.1.1.4. *Modèle hiérarchique*

L'ajout des variables latentes a introduit une structure hiérarchique au modèle. Il s'agit d'une structure à deux niveaux, où le premier niveau correspond au mécanisme qui génère l'appartenance des observations aux différentes sous-populations et où le second niveau correspond au mécanisme qui génère les observations selon leur sous-population respective (voir Robert, 2007). Cette structure hiérarchique peut être illustrée à l'aide du diagramme directionnel à la figure 2.1, où les quantités dans les carrés sont connues et celles dans les cercles sont inconnues et doivent être éventuellement estimées.

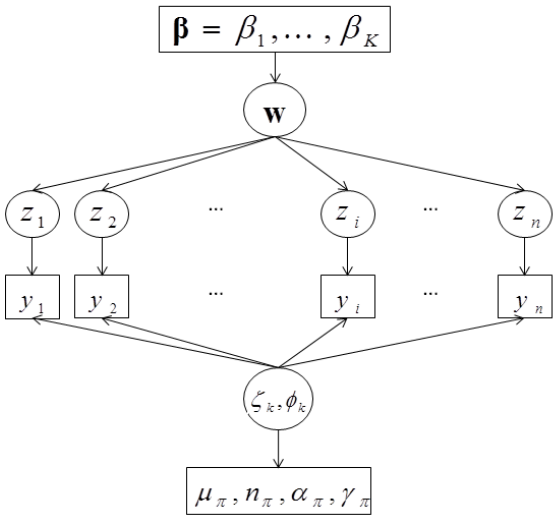

FIGURE 2.1. Diagramme directionnel illustrant la structure hiérarchique du modèle de mélange de lois (diagramme tiré de Perreault et Merleau, 2014).

2.1.2. Pourquoi les mélanges de lois ?

L'utilisation des mélanges de lois est justifiée tout d'abord par le fait qu'il s'agit d'un modèle qui est flexible. En effet, comme il est possible de le voir à la figure 2.2 pour les histogrammes du GTV vus précédemment (section 1.3.1.2), les mélanges de lois permettent d'ajuster plusieurs types et formes de distributions différentes.

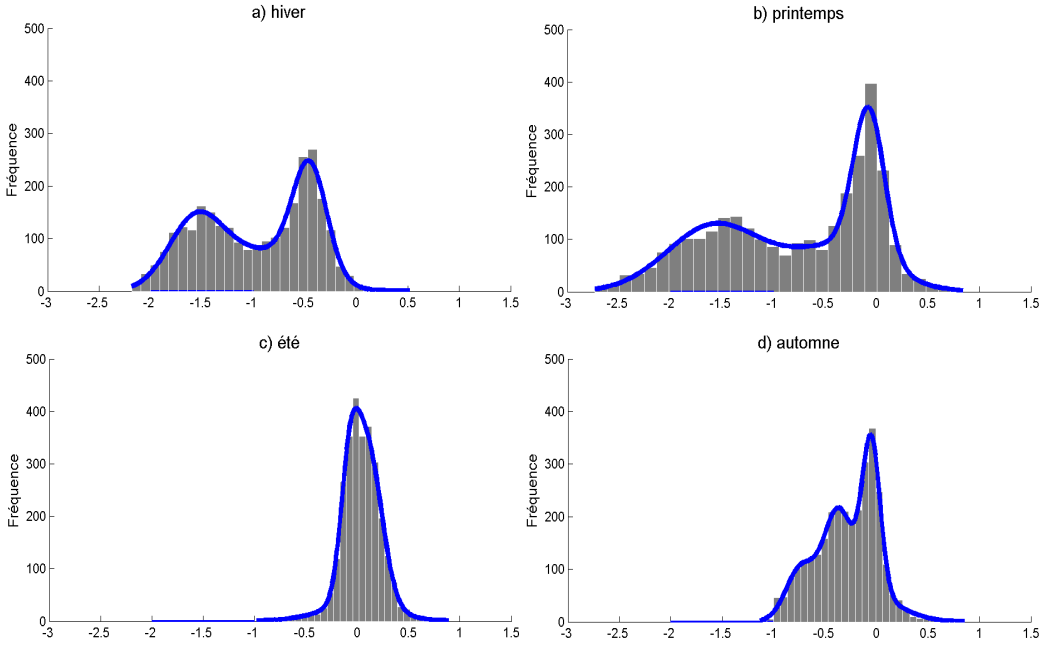

FIGURE 2.2. Histogramme du GTV selon les différentes saisons avec ajustement d'un mélange de lois normales, Québec, 2011

Aussi, l'utilisation des mélanges de lois s'inscrit bien dans un contexte où il est suspecté que des observations proviennent de sous-populations différentes, mais qu'il n'est pas possible d'identifier de laquelle elles proviennent individuellement (par exemple, neige et non-neige supposent que les observations proviennent de deux sous-populations). Ce modèle peut donc être utilisé afin de classifier les observations. En fait, il s'agit d'une méthode de classification plus riche que les méthodes descriptives usuelles, car elle permet l'intégration d'hypothèses paramétriques sur les observations et ainsi il est possible de quantifier les incertitudes. De plus, puisque les mélanges de lois sont utilisés ici avec l'approche bayésienne, cela permet également d'intégrer de l'information connue à l'aide des lois *a priori*, par exemple le savoir des experts du domaine d'application.

2.1.3. Algorithme de Gibbs

Comme il a été mentionné à la fin de la section 2.1.1, plusieurs quantités doivent être estimées lors de l'ajustement d'un modèle de mélange de lois. Dû à la complexité du modèle, il faut faire appel à une méthode de calcul numérique pour obtenir une approximation des quantités voulues. Comme nous avons opté pour l'estimation bayésienne avec des lois *a priori* conjuguées, il est naturel d'utiliser le sous-échantillonnage de Gibbs pour approcher les lois *a posteriori* et en déterminer les estimations.

2.1.3.1. *Densités conditionnelles*

Pour l'utilisation de la méthode de Gibbs, il est nécessaire de connaître la densité *a posteriori* conditionnelle complète de toutes les quantités inconnues impliquées dans le modèle. La densité *a posteriori* conditionnelle complète d'une variable est la densité *a posteriori* de cette variable conditionnellement à toutes les quantités restantes dans le modèle. Evin *et al.* (2011) ont montré que les densités d'intérêts pour un mélange de lois défini comme en (2.1.1) sont :

$$\begin{aligned}
\boldsymbol{w}|\boldsymbol{z},\boldsymbol{y} &\sim Dirichelet(\beta_1^*,\ldots,\beta_k^*,\ldots,\beta_K^*);\\
\boldsymbol{z}_i|\boldsymbol{w},y_i,\boldsymbol{\theta} &\sim Mult_K(w_1^*,\ldots,w_K^*);\\
\zeta_k|\phi_k,\boldsymbol{y},\boldsymbol{z} &\sim \mathcal{N}\left(\mu_k^*,n_k^*\phi_k\right);\\
\phi_k|\zeta_k,\boldsymbol{y},\boldsymbol{z} &\sim \mathcal{G}\left(\alpha_k^*,\gamma_k^*\right),
\end{aligned}$$

où, pour $k=1,\ldots,K$,

$$\begin{aligned}
{\beta_k}^* &= \beta_k + m_k(\boldsymbol{z});\\
m_k(\boldsymbol{z}) &= \sum_{i=1}^{n} z_{ik};\\
{w_k}^* &= \frac{w_k\mathcal{N}(y_i;\zeta_k,\phi_k)}{\sum_{s=1}^{K} w_s\mathcal{N}(y_i;\zeta_s,\phi_s)};\\
n_k &= \sum_{i=1}^{n} \mathbf{1}_{(\boldsymbol{z}_i=k)}\\
\mu_k^* &= \frac{n_\pi\mu_\pi + n_k\widehat{\mu}_k}{n_\pi+n_k};\\
\widehat{\mu}_k &= \frac{1}{n_k}\sum_{i:\boldsymbol{z}_i=k} y_i;\\
n_k^* &= n_\pi + n_k;\\
\alpha_k^* &= \alpha_\pi + \frac{n_k+1}{2};
\end{aligned}$$

$$\gamma_k^* = \gamma_\pi + \frac{1}{2}\left\{\sum_{i:\boldsymbol{z}_i=k}(y_i - \widehat{\mu}_k)^2 + n_k(\zeta_k - \widehat{\mu}_k)^2 + n_\pi(\zeta_k - \mu_\pi)^2\right\}.$$

où $\mathcal{N}(y_i; \zeta_k, \phi_k)$ est la densité au point y_i.

2.1.3.2. *Algorithme*

L'idée derrière l'algorithme de Gibbs est de simuler des valeurs pour les quantités inconnues $\boldsymbol{z}_i, w_k, \zeta_k$ et $\phi_k, i = 1, \ldots, n, k = 1, \ldots, K$ en fonction de leur densité *a posteriori* conditionnelle complète. Si l'algorithme converge, les valeurs ainsi obtenues sont alors des réalisations des lois *a posteriori*.

Tout d'abord, il faut fixer des valeurs initiales aux quantités inconnues :

– Initialisation : Commencer avec des valeurs arbitraires ou, mieux encore, tirées dans les lois *a priori*

$$\boldsymbol{z}_i{}^{(0)}, w_k{}^{(0)}, \zeta_k{}^{(0)} \text{ et } \phi_k{}^{(0)}, i = 1, \ldots, n, k = 1, \ldots, K.$$

Par la suite, pour un nombre N d'itérations, il faut effectuer les étapes suivantes :

– Itération r :

Sachant $\boldsymbol{z}_i{}^{(r-1)}, \zeta_k{}^{(r-1)}$ et $\phi_k{}^{(r-1)}, i = 1, \ldots, n, k = 1, \ldots, K,$ il faut générer

$$\begin{aligned}
w_k{}^{(r)} &\text{ à partir de } \pi\left(\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{z}^{(r-1)}\right);\\
\left(\boldsymbol{z}_1{}^{(r)}, \ldots, \boldsymbol{z}_n{}^{(r)}\right) &\text{ à partir de } \pi\left(\boldsymbol{z}_1|\boldsymbol{w}^{(r)}, y_1, \boldsymbol{\theta}_1^{(r-1)}\right), \ldots, \pi\left(\boldsymbol{z}_n|\boldsymbol{w}^{(r)}, y_n, \boldsymbol{\theta}_n^{(r-1)}\right);\\
\left(\phi_1{}^{(r)}, \ldots, \phi_K{}^{(r)}\right) &\text{ à partir de } \pi\left(\phi_1|\boldsymbol{y}, \boldsymbol{z}^{(r)}, \zeta_1^{(r-1)}\right), \ldots, \pi\left(\phi_K|\boldsymbol{y}, \boldsymbol{z}^{(r)}, \zeta_n^{(r-1)}\right);\\
\left(\zeta_1{}^{(r)}, \ldots, \zeta_K{}^{(r)}\right) &\text{ à partir de } \pi\left(\zeta_1|\boldsymbol{y}, \boldsymbol{z}^{(r)}, \phi_1{}^{(r)}\right), \ldots, \pi\left(\zeta_K|\boldsymbol{y}, \boldsymbol{z}^{(r)}, \phi_K{}^{(r)}\right),
\end{aligned}$$

où le choix de N doit être fait de façon à tenir compte du temps de convergence des itérations :

$$N = T + R,$$

où T est le temps de chauffe, c'est-à-dire le nombre d'itérations servant à atteindre la convergence, et R est le nombre de réplicats, c'est-à-dire le nombre de simulations conservées suite à la convergence pour l'inférence sur les quantités à estimer (voir Robert, 1998).

2.1.3.3. *Probabilités d'appartenance à chacune des sous-populations*

Il a été mentionné à la section 2.1.2 qu'un avantage des mélanges de lois est qu'il est possible de quantifier les incertitudes sur la classification. Cette quantification est illustrée par le calcul de la probabilité d'appartenance à chacune des

sous-populations à l'aide des R derniers réplicats des variables $\boldsymbol{z}_i$. En effet, il est possible de montrer (voir Casella et George, 1992) qu'effectuer la moyenne

$$\hat{P}(y_i \text{ appartienne à la sous-population } k) = \frac{\sum_{s=T+1}^{T+R} \mathbf{1}(z_{ik}{}^{(s)} = 1)}{R} \tag{2.1.2}$$

donne une bonne estimation de la probabilité que l'observation y_i appartienne à la sous-population k.

2.1.4. Comparaison de différents ajustements

Une considération importante lors de l'utilisation d'un modèle de mélange de lois est de pouvoir comparer la qualité de différents ajustements. Pour ce projet, nous avons décidé d'utiliser le critère de Schwarz, car il permet un compromis entre la qualité de l'ajustement et le nombre de paramètres (voir Ghosh *et al.*, 2007) :

$$\textit{Critère de Schwarz} = \log(L) - \frac{(p+1)}{2} \times \log(n), \tag{2.1.3}$$

où $\log(L)$ est le logarithme de la vraisemblance maximisée, p est le nombre de régresseurs du modèle et n est le nombre d'observations. Selon ce critère, le modèle qui a le meilleur ajustement pour les données est celui qui a la plus **forte** valeur (voir Schwarz, 1978 ; Merleau et Bibeau, 2013). Il permet donc une utilisation qui est intuitive, car il faut le maximiser tout comme la vraisemblance.

2.2. Régression linéaire multiple

Cette section décrit brièvement la régression linéaire multiple, car il s'agit de la méthode utilisée pour étudier le lien entre une variable réponse $\boldsymbol{y}$ (par exemple, le GTV) et des variables explicatives $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ (par exemple, la température, les degrés-jours et les degrés-chauffes). Pour plus de détails, le lecteur peut consulter Weisberg (2005) ainsi que Montgomery *et al.* (2006).

2.2.1. Rappel de la base

2.2.1.1. *Définition et estimation des paramètres*

L'idée générale de la régression linéaire multiple est d'exprimer la variable réponse comme une combinaison linéaire des variables explicatives de la façon suivante :

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.2.1}$$

où $\boldsymbol{y}$ est un vecteur de dimension n, $\boldsymbol{X} = (\mathbf{1}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$ est une matrice de dimension $n \times (p+1)$, $\mathbf{1}$ est un vecteur de dimension n contenant uniquement la valeur 1, $\boldsymbol{\beta}$ est un vecteur de coefficients de dimension $(p+1)$, $\boldsymbol{\varepsilon}$ est un vecteur

d'erreurs de dimension n, n étant le nombre d'observations et p le nombre de variables explicatives.

Évidemment, la valeur des coefficients β_j, $j = 0, \ldots, p$ est inconnue et doit être estimée par $\widehat{\boldsymbol{\beta}}$ à partir d'un jeu de données. Une méthode fréquemment utilisée pour résoudre ce problème est celle des moindres carrés ordinaires (voir Weisberg, 2005 ; Montgomery *et al.*, 2006), qui stipule que $\widehat{\boldsymbol{\beta}}$ est le vecteur des coefficients qui minimise la fonction suivante :

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} \varepsilon_i^2 = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) . \tag{2.2.2}$$

La résolution de ce problème d'optimisation pour (2.2.2) permet d'obtenir une estimation des coefficients de régression :

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}.$$

L'intérêt d'utiliser la méthode des moindres carrés ordinaires pour l'estimation est qu'elle permet, sous certaines conditions, d'obtenir le meilleur estimateur linéaire sans biais, communément référé sous le nom d'estimateur « BLUE » (Best Linear Unbiaised Estimator) (voir Montgomery *et al.*, 2006). Les conditions pour l'obtention d'un tel estimateur sont énoncées au théorème 2.2.1, connu sous le nom de théorème de Gauss-Markov.

Théorème 2.2.1. *Théorème de Gauss-Markov*
Parmi tous les estimateurs linéaires sans biais, l'estimateur utilisant la méthode des moindres carrés ordinaires est celui qui a la variance la plus faible si les termes d'erreur :

(a) ont une espérance nulle ;

(b) ont une variance constante ;

(c) ne sont pas corrélés.

Dans le but de pouvoir faire de l'inférence statistique, une hypothèse supplémentaire doit être ajoutée. En effet, il faut supposer que les termes d'erreurs sont distribués selon une loi normale, indépendamment et identiquement.

2.2.1.2. *Analyse de la variance*

Suite à l'estimation des coefficients de régression, il faut déterminer si la régression est statistiquement significative. Le but est donc de confronter les hypothèses nulle (H_0) et alternative (H_1) suivantes :

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$$
$$H_1 : \text{au moins un } \beta_j \neq 0, j = 1, \ldots, p.$$

Pour ce faire, il faut effectuer des tests de Fisher (tests F) en calculant les quantités indiquées dans le tableau 2.1.

TABLE 2.1. Tableau d'ANOVA pour la régression linéaire multiple.

Source	Degrés de liberté	Somme des carrés	Carré moyen	Statistique F
Régression	p	SS_{reg}	$MSE = \frac{SS_{reg}}{p}$	$F_{obs} = \frac{MSE}{\widehat{\sigma}^2}$
Résidus	$n-p-1$	SS_{res}	$\widehat{\sigma}^2 = \frac{SS_{res}}{n-p-1}$	
Total	$n-1$	SS_{tot}		

Dans le tableau 2.1, les différents termes signifient :

$$SS_{reg} = \sum_{i=1}^{n} (\widehat{y}_i - \overline{y}_i)^2; \tag{2.2.3}$$
$$SS_{res} = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2;$$
$$SS_{tot} = \sum_{i=1}^{n} (y_i - \overline{y}_i)^2,$$

où $\widehat{\sigma}^2$ correspond à une estimation de la variance des termes d'erreurs et F_{obs} suit une loi de Fisher à p et $(n-p-1)$ degrés de liberté ($F_{obs} \sim F(p, n-p-1)$).

Nous rejetons H_0 à un seuil de α choisi, c'est-à-dire que la régression est considérée significative, lorsque la valeur de la statistique F_{obs} est plus grande qu'une valeur critique F_α, qui correspond à la valeur du $100 \times (1-\alpha)^{\text{e}}$ quantile de la loi de Fisher avec p et $(n-p-1)$ degrés de liberté. La même conclusion peut être obtenue en calculant la valeur-p (« p-value ») :

$$\text{valeur-p} = P\left(F(p, n-p-1) > F_{obs}\right). \tag{2.2.4}$$

La régression est considérée comme étant statistiquement significative (rejet de H_0) si la valeur-p est plus petite que le seuil de signification α choisi.

2.2.1.3. *Importance de chacune des variables explicatives*

De façon parallèle au test d'hypothèses pour vérifier la régression, il est intéressant de tester individuellement si chacune des variables explicatives est utile pour expliquer la variable réponse.

Une première façon d'aborder le problème est de tester si la somme de carrés supplémentaire associée à la variable explicative $\boldsymbol{x}_j$ est significative (voir Montgomery *et al.*, 2006). Pour ce faire, il faut confronter le modèle complet (H_0) avec un modèle réduit sans la variable explicative $\boldsymbol{x}_j$ (H_1) :

$$H_0 : \boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{x}_1 + \ldots + \beta_p \boldsymbol{x}_p + \varepsilon \text{ (modèle complet)}$$
$$H_1 : \boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{x}_1 + \ldots \beta_{j-1} \boldsymbol{x}_{j-1} + \beta_{j+1} \boldsymbol{x}_{j+1} + \ldots + \beta_p \boldsymbol{x}_p + \varepsilon \text{ (modèle réduit).}$$

La quantité permettant de vérifier l'intérêt de la variable explicative est la somme de carrés supplémentaire associée à $\boldsymbol{x}_j$:

$$SS_{reg}(\beta_j | \beta_1, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_p) = SS_{reg}^{H_0} - SS_{reg}^{H_1},$$

où $SS_{reg}^{H_s}$ est la somme des carrés due à la régression pour le modèle spécifié sous H_s (voir équation (2.2.3)). Cette quantité ayant 1 degré de liberté, la statistique pour le test est donc :

$$F_j = \frac{SS_{reg}(\beta_j | \beta_1, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_p)/1}{\hat{\sigma}^2} \sim F_{1,n-p-1}.$$

La variable $\boldsymbol{x}_j$ est statistiquement significative au seuil de α pour expliquer la variable réponse si la statistique F_j est plus grande que la valeur critique F_α ou si la valeur-p

$$\text{valeur-p} = P\left(F(1, n-p-1) > F_j\right)$$

est plus petite que α.

Une autre approche pour tester l'importance de la variable $\boldsymbol{x}_j$ est de confronter les hypothèses nulle (H_0) et alternative (H_1) suivantes (voir Weisberg, 2005) :

$$H_0 : \beta_j = 0$$
$$H_1 : \beta_j \neq 0.$$

Pour ce faire, la statistique utilisée est la suivante :

$$t_j = \frac{\widehat{\beta}_j}{SD(\widehat{\beta}_j)} \sim t(1), \tag{2.2.5}$$

où $SD(\widehat{\beta}_j)$ est l'écart-type de $\widehat{\beta}_j$ et $t(1)$ représente une loi de Student avec un degré de liberté. La variable est statistiquement utile à un seuil α pour expliquer la variable réponse si la valeur $|t_j|$ est plus grande que la valeur critique $t_{\frac{\alpha}{2}}$ ou si la valeur-p

$$\text{valeur-p} = P\left(t(1) > |t_j|\right)$$

est plus petite que $\alpha/2$.

Il est à noter que les deux approches présentées ci-dessus sont équivalentes. Le choix de l'approche n'a donc pas d'impact sur la signification d'une variable explicative. En effet, il est possible de montrer que

$$t_j^2 = F_j.$$

2.2.1.4. *Valeurs prédites et valeurs ajustées*

Lorsqu'un modèle de régression est entièrement construit pour un certain jeu de données, il est aisément possible de calculer les valeurs ajustées pour les valeurs actuelles des variables explicatives et les valeurs prédites pour de nouvelles valeurs de ces mêmes variables :

$$\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}};$$
$$\boldsymbol{y}^* = \boldsymbol{X}^*\widehat{\boldsymbol{\beta}},$$

où $\widehat{\boldsymbol{y}}$ est le vecteur des valeurs ajustées pour la matrice $\boldsymbol{X}$ contenant les valeurs originales des variables explicatives et $\boldsymbol{y}^*$ est le vecteur des valeurs prédites pour la matrice $\boldsymbol{X}^*$ contenant de nouvelles valeurs pour ces mêmes variables.

2.2.2. Évaluation du modèle

Avant de déterminer si un modèle est adéquat pour l'ajustement de la variable réponse, il faut être en mesure d'évaluer la qualité de l'ajustement.

2.2.2.1. *Coefficient de détermination*

Une mesure couramment utilisée afin de quantifier la qualité d'un modèle est le coefficient de détermination :

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}.$$

Intuitivement, le R^2 représente la quantité (%) de la variation de la variable réponse qui est expliquée linéairement par les variables explicatives. Il se situe toujours entre 0 et 1. La valeur 0 représente le pire ajustement, c'est-à-dire qu'il y a 0% de la variation de la variable réponse qui est expliquée linéairement par les variables explicatives. La valeur 1 représente l'ajustement parfait, c'est-à-dire que 100% de la variation de la variable réponse est expliquée par les variables explicatives.

Cependant, puisque le R^2 a tendance à augmenter avec le nombre de variables explicatives dans le modèle, il ne peut pas être utilisé pour comparer des modèles contenant un nombre différent de variables explicatives. Pour remédier à cet inconvénient et ainsi pouvoir comparer entre eux des modèles contenant différents

nombres de variables explicatives, il est conseillé d'utiliser l'ajustement suivant (voir Montgomery *et al.*, 2006) :

$$R^2 \text{ ajusté} = 1 - (1 - R^2)\left(\frac{n-1}{n-p-1}\right), \tag{2.2.6}$$

où p est le nombre de variables explicatives dans le modèle et n est le nombre d'observations.

2.2.2.2. *Multicolinéarité*

L'utilisation de plusieurs variables explicatives lors de la modélisation d'une régression linéaire multiple peut engendrer un problème de multicolinéarité, c'est-à-dire qu'il est possible que les résultats soient faussés par la présence d'une relation quasi-linéaire entre certaines variables explicatives. Parmi les différents problèmes possibles, il se peut que le modèle obtenu n'offre que des prévisions de faible qualité, que les coefficients de régression soient trop sensibles aux données avec lesquelles ils ont été obtenus ou que leur variance soit inutilement trop élevée (voir Montgomery *et al.*, 2006).

Afin d'éviter qu'un modèle soit affecté par un tel problème, il faut définir des indicateurs permettant de juger du degré de multicolinéarité et permettant d'indiquer quelles sont les variables explicatives qui induisent ce problème. Un indicateur couramment utilisé est le facteur d'inflation de la variance, couramment référé comme le « VIF » (Variance Inflation Factor) :

$$\text{VIF}_j = \frac{1}{1 - {R_j}^2}, j = 1, \ldots, p, \tag{2.2.7}$$

où ${R_j}^2$ est le coefficient de détermination de la régression obtenue si la variable réponse est $\boldsymbol{x}_j$ et que les variables explicatives sont $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{j-1}, \boldsymbol{x}_{j+1}, \ldots, \boldsymbol{x}_p$. Aussi, il peut être prouvé que :

$$\text{VIF}_j = C_{jj},$$

où C_{jj} est le j^{e} élément de la diagonale de la matrice $\boldsymbol{C} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}$ (voir Montgomery *et al.*, 2006).

Dans le but de pouvoir interpréter une valeur du facteur d'inflation de la variance, il est pertinent de remarquer que si une variable $\boldsymbol{x}_j$ a une relation trop forte avec les autres prédicteurs, alors la valeur de ${R_j}^2$ sera élevée. Du fait même, la valeur VIF_j sera trop élevée et la variable $\boldsymbol{x}_j$ devra être retirée du modèle. Montgomery *et al.* (2006) suggère qu'une valeur est trop élevée lorsqu'elle est supérieure à dix.

2.2.2.3. *Diagnostics des résidus*

Afin de pouvoir affirmer qu'un modèle de régression est bien ajusté aux données et afin de pouvoir faire de l'inférence statistique, il faut que les hypothèses du théorème de Gauss-Markov (théorème 2.2.1) soient respectées et que les termes d'erreur soient distribués selon une loi normale. Cependant, puisque les ε_i ne sont pas connus, il faut vérifier ces conditions à partir des résidus de régression. Pour ce faire, deux graphiques sont utilisés.

Le premier graphique représente les résidus standardisés en ordonnée en fonction des valeurs ajustées en abscisse, où les résidus standardisés sont calculés ainsi :

$$\text{resstand}_i = \frac{res_i}{\sqrt{\text{MSE}(1 - h_{ii})}},$$

où res_i est le résidu de l'observation i ($\boldsymbol{res} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$), MSE est défini tel que dans le tableau 2.1 et h_{ii} est le i^{e} élément de la diagonale de la matrice $\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T$. Ce graphique permet de vérifier les hypothèses que les termes d'erreur sont d'espérance 0 et de variance constante. La première hypothèse se vérifie par des résidus standardisés répartis autour de 0 en ordonnée et la seconde se vérifie par des résidus standardisés uniformément distribués. Un exemple pour lequel ces hypothèses sont respectées est illustré à la figure 2.3. En effet, il est

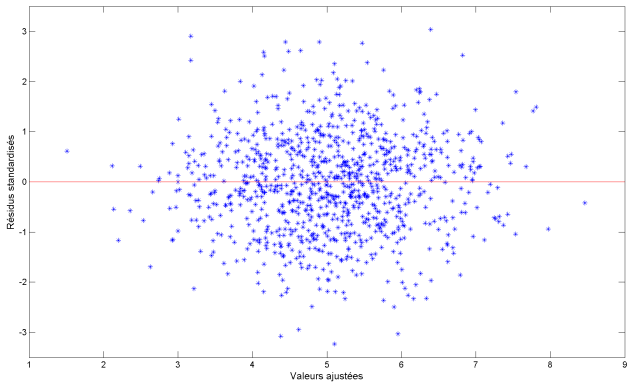

FIGURE 2.3. Exemple de graphique des résidus standardisés en fonction des valeurs ajustées où les résidus standardisés sont uniformément distribués autour de 0 en ordonnée.

possible de constater qu'ici les résidus standardisés sont bel et bien uniformément distribués autour de 0 en ordonnée. La figure 2.4 représente des exemples pour lesquels les résidus standardisés ne sont pas distribués uniformément autour de 0 en ordonnée ou n'ont pas une variance constante.

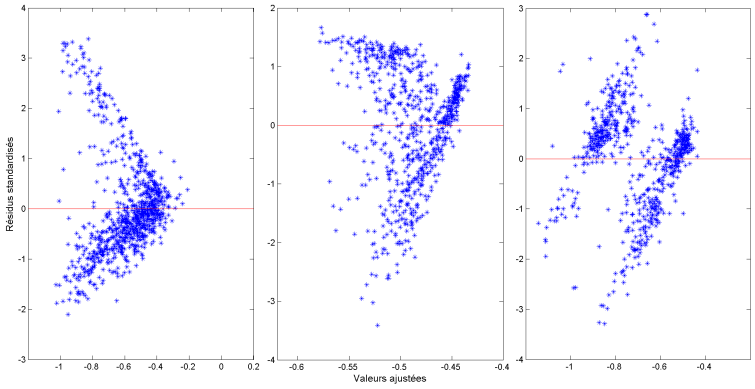

FIGURE 2.4. Exemples de graphiques des résidus standardisés en fonction des valeurs ajustées où les résidus standardisés ne sont pas uniformément distribués autour de 0 en ordonnée ou n'ont pas une variance constante.

Le second graphique est le diagramme quantile-quantile des résidus standardisés. Il permet de vérifier l'hypothèse de normalité des résidus standardisés. Pour ce faire, il faut que le graphique représente la droite identité, c'est-à-dire que les quantiles observés correspondent aux quantiles théoriques de la loi normale. Un exemple de diagramme quantile-quantile où les résidus standardisés sont distribués selon une loi normale est illustré à la figure 2.5. La figure 2.6 représente des

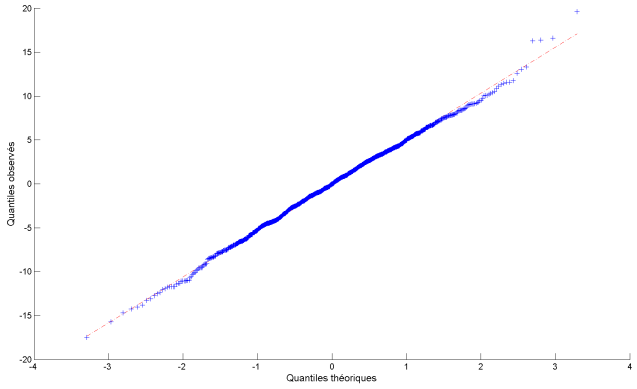

FIGURE 2.5. Exemple de diagramme quantile-quantile où les quantiles observés correspondent aux quantiles théoriques de la loi normale.

exemples pour lesquels l'hypothèse de normalité n'est pas respectée.

2.2.3. Transformation des variables

Dans le cadre de ce projet, une des variables étudiées est définie comme étant une proportion, donc ayant des valeurs comprises dans l'intervalle (0,1). Une

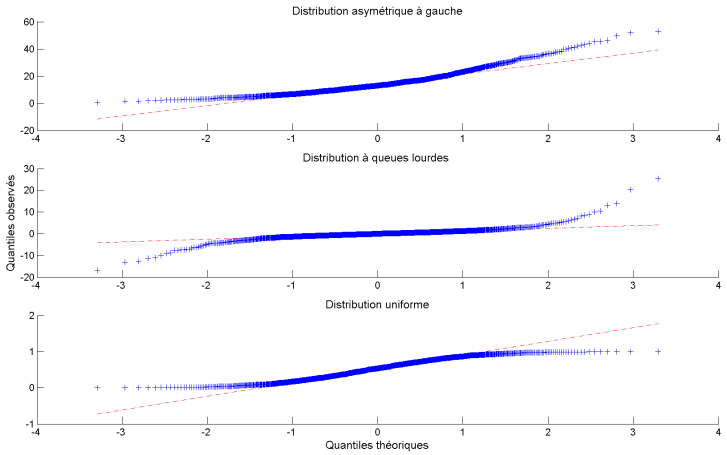

FIGURE 2.6. Exemples de diagrammes quantile-quantile où les quantiles observés ne correspondent pas aux quantiles théoriques de la loi normale.

telle variable ne peut pas être utilisée comme variable réponse $\boldsymbol{y}$ dans un modèle de régression linéaire sans subir de transformation, car son domaine est borné et donc ne vérifie pas l'hypothèse de normalité. Pour remédier à ce problème, il faut transformer la variable réponse $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^T$ afin d'obtenir une nouvelle variable réponse $\boldsymbol{z} = (z_1, z_2, \ldots, z_n)^T$ ayant des valeurs dans $\mathbb{R}$. Ici, la transformation utilisée est la transformation logit, c'est-à-dire :

$$z_i = \log\left(\frac{y_i}{1 - y_i}\right), i = 1, \ldots, n. \tag{2.2.8}$$

Une considération importante lorsque le modèle de régression est entièrement construit est de pouvoir effectuer une transformation inverse sur les valeurs ajustées ou prédites afin de les ramener dans l'échelle originale. Pour la transformation (2.2.8), la transformation inverse lors du calcul des valeurs ajustées $\hat{y}_i$ et prédites ${y_i}^*$ est :

$$\begin{aligned} \hat{y}_i &= \frac{1}{1 + e^{-\widehat{z_i}}}; \\ {y_i}^* &= \frac{1}{1 + e^{-z_i^*}}. \end{aligned} \tag{2.2.9}$$

2.2.4. Sélection de variables

Il a été vu à la section 2.2.1.3 qu'il est possible de tester individuellement si chacune des variables explicatives est significative pour expliquer la variable réponse. Toutefois, peu importe l'issue de chacun des tests, toutes les variables sont incluses dans la régression même si elles ne s'avèrent pas nécessaires.

Le but de cette section est de présenter la sélection de variables à rebours, qui est la méthode utilisée dans le présent mémoire afin de tenir compte du fait que certaines variables ne sont pas statistiquement utiles pour expliquer la variable réponse.

2.2.4.1. *Définition de la corrélation partielle*

Avant de décrire la méthode, la définition de la corrélation partielle est introduite, car elle sera utile pour la compréhension de certaines étapes :

Définition 2.2.1. *Corrélation partielle*
La corrélation partielle entre une variable explicative $\boldsymbol{x}$ *et une variable réponse* $\boldsymbol{y}$ *lorsque le modèle contient déjà les* m *variables* $\boldsymbol{Z} = (\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_m)$*, notée* $\rho_{\boldsymbol{xy \cdot Z}}$ *est le coefficient de corrélation linéaire entre les résidus respectifs de la régression de* $\boldsymbol{x}$ *selon* $\boldsymbol{Z}$ *et de* $\boldsymbol{y}$ *selon* $\boldsymbol{Z}$*. Elle est calculée de la façon suivante :*

$$\rho_{\boldsymbol{xy \cdot Z}} = \frac{\rho_{\boldsymbol{xy}} - \rho_{\boldsymbol{xZ}}\rho_{\boldsymbol{yZ}}}{\sqrt{\left(1 - \rho^2_{\boldsymbol{xZ}}\right)\left(1 - \rho^2_{\boldsymbol{yZ}}\right)}},$$

où $\rho_{\boldsymbol{AB}}$ *est le coefficient de corrélation linéaire entre les variables* $\boldsymbol{A}$ *et* $\boldsymbol{B}$*.*

2.2.4.2. *Sélection à rebours*

La sélection à rebours, communément connue sous le nom de « backward elimination », consiste à inclure toutes les variables explicatives dans le modèle puis de retirer une à une toutes les variables considérées comme étant non utiles pour expliquer la variable réponse (voir Weisberg, 1985).

Pour commencer, une régression linéaire multiple incluant toutes les variables explicatives est effectuée. Par la suite, des variables explicatives sont retirées une à une selon un critère parmi les suivants :

(a) la variable retirée est celle qui a la plus faible valeur, en valeur absolue, de statistique t ;

(b) la variable retirée est celle dont le retrait cause la plus faible variation du R^2 ;

(c) la variable retirée est celle dont la corrélation partielle avec la variable réponse est la plus faible lorsque les autres variables explicatives dans le modèle sont prises en considération.

Après avoir retiré un certain nombre de variables, il faut arrêter le processus selon un des critères suivants :

(a) le processus est arrêté lorsque le modèle contient un nombre p^*, déterminé à l'avance, de variables explicatives ;

(b) le processus est arrêté lorsque la valeur absolue de la statistique t de la prochaine variable à retirer est plus grande qu'un seuil t_{out}, où t_{out} est déterminé comme étant le $100 \times \left(1 - \frac{\alpha^*}{2}\right)^{\text{e}}$ quantile d'une loi de Student à un degré de liberté. Généralement, $\alpha^* = 0{,}10$ (voir Draper et Smith, 1998).

Les variables qui ont été retirées du modèle sont celles considérées comme étant non utiles pour expliquer la variable réponse.

Une méthode alternative à la sélection à rebours pourrait être de tester toutes les combinaisons de variables explicatives possibles, car cela permet de considérer un plus grand éventail de modèles. Toutefois, cette alternative n'a pas été retenue dans le cadre de ce projet à cause de l'utilisation du logiciel Matlab, car il aurait fallu programmer chacun des modèles manuellement.

2.3. Conclusion partielle

Dans ce chapitre, nous avons abordé la théorie reliée aux méthodes statistiques qui ont été utilisées pour ce projet, soit les mélanges de lois et la régression linéaire multiple. Dans les chapitres qui suivent, ces méthodes sont appliquées sur les données de télédétection afin d'obtenir un modèle permettant de répondre à l'objectif énoncé en introduction.

Chapitre 3

MODÉLISATION DE LA PROBABILITÉ DE NEIGE À L'AIDE DE LA RÉGRESSION LINÉAIRE SUR LES LOGITS

Le but de ce chapitre est de répondre à l'objectif principal énoncé en introduction, soit développer un modèle permettant de cartographier le couvert nival et de quantifier l'incertitude de la classification. Les données utilisées sont les valeurs de GTV de l'année 2011.

D'autres travaux ont été réalisés en amont de celui qui est présenté ici, mais ils ne sont pas tous documentés dans le présent mémoire, car ils ne permettaient pas d'atteindre le but voulu. Ils se trouvent cependant aux annexes A et B pour le lecteur intéressé. À l'annexe A.1, le potentiel du GTV (voir équation (1.3.1)) a été exploré afin de déterminer s'il pouvait discriminer la neige et le sol. Il a été conclu que l'utilisation du GTV seul ne permettait pas une classification neige/non-neige cohérente avec la réalité. À la suite de ce constat, la relation entre le GTV et différentes variables explicatives a été étudiée via la régression linéaire à l'annexe A.2. Bien que le GTV n'ait pas pu être modélisé de façon satisfaisante, il a été conclu qu'il était pertinent d'utiliser des variables exogènes afin de compléter l'information disponible.

À l'annexe B, il est question d'une approche considérée pour traiter l'information contenue dans le GTV avec des variables explicatives. Pour ce faire, une variable de seuils dynamiques a été créée puis modélisée à l'aide de la régression linéaire multiple. Les cartographies neige/non-neige obtenues n'étant pas cohérentes avec la réalité, cette démarche n'a pas été retenue.

La méthodologie présentée dans le présent chapitre permet d'obtenir des résultats cohérents avec la réalité en plus de répondre à l'objectif principal. Il s'agit d'une régression linéaire multiple qui met en relation les logits des probabilités de neige p_t avec diverses variables exogènes.

3.1. Calcul de la probabilité de neige journalière p_t

L'idée de calculer une probabilité de neige quotidienne à partir des valeurs de GTV vient du fait qu'il est possible, en théorie, de séparer les valeurs de GTV en neige/non-neige selon un seuil (voir section 1.3.1.1). En ajustant un mélange de deux lois normales (une composante de neige et une composante de non-neige) sur les données de GTV d'une journée t, il est possible de calculer la valeur de la fonction de répartition pour ce seuil et cette valeur peut être interprétée comme la probabilité de neige de la journée t. Ayant vu à la section 1.3.1.2, que le seuil 0 n'est pas cohérent avec la réalité, un seuil plus réaliste a été estimé à $s = -0{,}15$ par les spécialistes de l'Institut de recherche d'Hydro-Québec (voir figure 3.1).

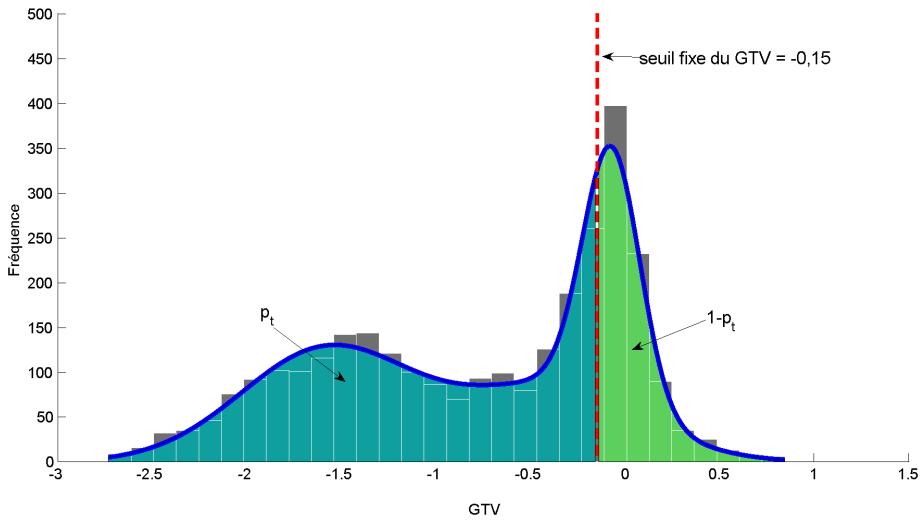

Figure 3.1. Représentation du calcul de p_t pour une journée t au printemps.

L'algorithme utilisé pour calculer la probabilité de neige journalière peut être exprimé en plusieurs étapes que voici :

(a) poser $s = -0{,}15$ le seuil permettant de séparer les valeurs de GTV en état neige ou non-neige ;

(b) ajuster un mélange de deux lois normales à $\boldsymbol{y}_t$ pour chaque jour t de l'année 2011. L'utilisation des mélanges de lois est justifiée par la flexibilité du modèle, tandis que le choix de deux sous-populations est expliqué par la volonté d'obtenir une classe de neige et une classe de non-neige ;

(c) calculer :

$p_t = P(u_t < s)$ en l'estimant par :

$$\widehat{p}_t = \widehat{w}_1 \times \Phi\left(\sqrt{\widehat{\phi}_{1t}} \times \left(s - \widehat{\zeta}_{1t}\right)\right) + \widehat{w}_2 \times \Phi\left(\sqrt{\widehat{\phi}_{2t}} \times \left(s - \widehat{\zeta}_{2t}\right)\right), \qquad (3.1.1)$$

où $t = 1, \ldots, 365$, $\boldsymbol{y}_t$ représente les valeurs de GTV de la journée t, u_t est une variable aléatoire, Φ représente la fonction de répartition d'une loi normale, $\widehat{w}_1$, $\widehat{w}_2$, $\widehat{\zeta}_{1t}$, $\widehat{\phi}_{1t}$, $\widehat{\zeta}_{2t}$ et $\widehat{\phi}_{2t}$ ($\widehat{w_1} + \widehat{w_2} = 1$) sont les estimations des paramètres du mélange de deux lois normales obtenues en moyennant les paramètres des sous-échantillons de Gibbs. La valeur de p_t peut être interprétée comme la probabilité qu'il y ait de la neige à la journée t.

3.2. Considérations techniques pour les mélanges de lois bayésiens

Il a été mentionné à la section 2.1 que certaines valeurs nécessaires pour appliquer les mélanges de lois bayésiens doivent être spécifiées *a priori*. La présente section a donc pour but d'expliquer comment ces valeurs ont été choisies.

3.2.1. Détermination des valeurs des hyperparamètres

Tout d'abord, les valeurs des hyperparamètres des lois *a priori* doivent être spécifiées avant de débuter l'algorithme (voir section 2.1.1.3). Pour la loi *a priori* du paramètre $\boldsymbol{w}$, il faut choisir des valeurs pour $\beta_1, \ldots, \beta_K$ selon les connaissances sur le sujet. Ne disposant pas d'information sur celui-ci, il a été décidé d'utiliser une loi *a priori* non informative, c'est-à-dire que les β_k sont tous égaux à 1. Ainsi, la densité de $\boldsymbol{w}$ est toujours uniforme sur le simplexe de dimension K.

Les codes disponibles pour l'application des mélanges de lois ont été développés à l'Institut de recherche d'Hydro-Québec et s'appuient sur les travaux de Evin *et al.* (2011). Pour les lois *a priori* des paramètres ζ_k et $\phi_k, k = 1, \ldots, K$, les valeurs des hyperparamètres suggérées sont résumées dans le tableau 3.1, où $\overline{y}_t$ et $s^2_{y_t}$ sont respectivement la moyenne et la variance échantillonnale des valeurs de GTV de la journée t.

Table 3.1. Tableau des valeurs pour les hyperparamètres des lois *a priori*, telles que recommandées dans Evin *et al.* (2011).

Hyperparamètre	Valeur
μ_π	$\overline{y}_t$
n_π	1
α_π	$\frac{n_\pi}{2} = \frac{1}{2}$
γ_π	$n_\pi \times \frac{s^2_{y_t}}{2} = \frac{s^2_{y_t}}{2}$

3.2.2. Choix de R et T pour la convergence de l'algorithme de Gibbs

L'estimation des mélanges de lois dans un cadre bayésien nécessite des calculs numériques par sous-échantillonnage de Gibbs. Il faut donc étudier la convergence de la méthode afin de déterminer les valeurs de T et de R qui doivent être utilisées pour s'assurer que l'algorithme converge. Dans le cadre de ce projet, il a été décidé que $T = R$. De plus, il est considéré qu'il y a convergence lorsque les diagrammes en boîte ne varient presque plus lorsque la valeur de T et R augmente.

Pour effectuer l'étude de convergence, quelques journées de l'année 2011 ayant des distributions de GTV différentes ont été sélectionnées et les mélanges de lois à une, deux et trois sous-populations ont été ajustés 100 fois pour $T = R = \{500, 1\ 000, 2\ 000, 5\ 000, 10\ 000, 15\ 000, 20\ 000\}$ aux valeurs pour l'ensemble des points de grille de la journée (un ajustement par jour qui utilise les i points de grille). Pour chaque ajustement, les paramètres ont été estimés en moyennant les valeurs pour chacun de leurs sous-échantillons de Gibbs respectifs. Cette étude est réalisée pour les mélanges d'une, deux et trois lois normales, même si le calcul des probabilités de neige n'utilise que le mélange de deux normales. Cela permet d'obtenir une information plus complète et de mieux comprendre la sensibilité de la convergence associée au choix du nombre de sous-populations. Pour chacun des ajustements, la valeur du critère de Schwarz (voir équation (2.1.3)) a été calculée. La variabilité des valeurs du critère a été ensuite étudiée en fonction du nombre d'itérations à l'aide de diagrammes en boîte.

L'examen des diagrammes en boîte pour l'ajustement des mélanges sur l'ensemble des points de grille du 5 mai 2011 (voir figure 3.2) permet de constater que la convergence est plus rapide lorsqu'il y a moins de sous-populations. En effet, les 100 valeurs se stabilisent plus rapidement pour une sous-population que pour deux et trois sous-populations. La convergence semble atteinte entre 10 000 et 20 000 itérations ($T = R$ entre 5 000 et 10 000) lorsqu'il y a une sous-population, entre 30 000 et 40 000 itérations ($T = R$ entre 15 000 et 20 000) lorsqu'il y a deux sous-populations et à environ 40 000 itérations ($T = R = 20\ 000$) lorsqu'il y a trois sous-populations. Les diagrammes en boîte présentés aux figures C.1 à C.4 montrent que les résultats sont similaires pour les autres journées étudiées. Il semble donc que $T = R = 20\ 000$ soit nécessaire afin d'assurer la convergence de l'algorithme peu importe le type de distribution du GTV. Les distributions des GTV des journées qui ont été sélectionnées pour l'étude de convergence sont disponibles en annexe à la figure C.5.

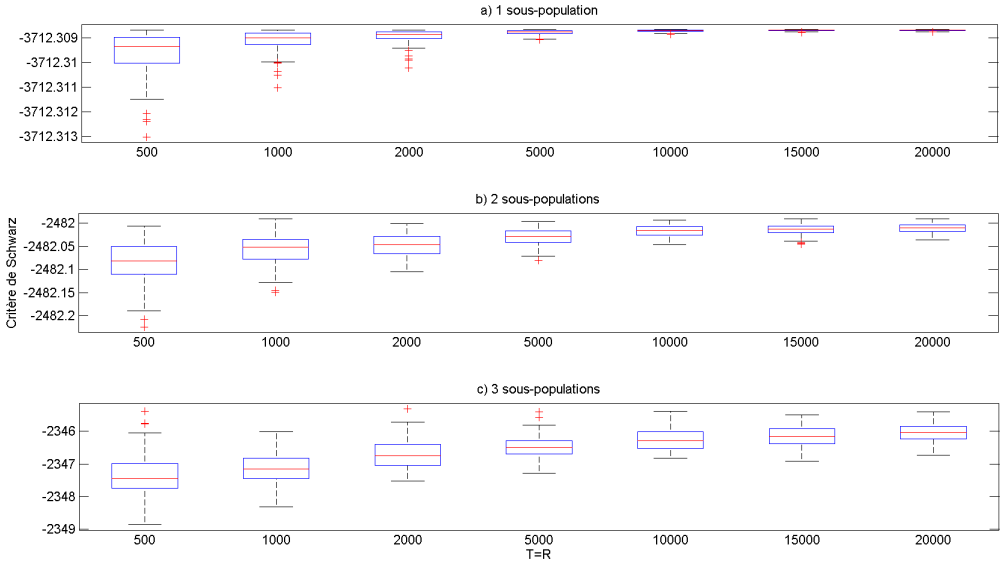

FIGURE 3.2. Diagrammes en boîte des 100 valeurs du critère de Schwarz pour chacune des valeurs de $T = R$, a) une sous-population, b) deux sous-populations, c) trois sous-populations, Québec, 5 mai 2011.

.

3.3. MODÉLISATION POUR LA ZONE 2

Avant de procéder à la modélisation sur l'ensemble du territoire, nous nous sommes concentrés sur la zone 2, car il s'agit de la zone du Québec qui intéresse le plus les prévisionnistes d'Hydro-Québec. Dans l'éventualité où les résultats seraient concluants pour cette zone, une méthodologie applicable pour l'ensemble du Québec sera développée.

Les résultats qui sont présentés dans cette section utilisent la variable p_t calculée avec le seuil fixe de GTV $s = -0{,}15$ (voir section 3.1) proposé par les chercheurs de l'Institut de recherche d'Hydro-Québec. D'autres seuils ont été considérés (-0,18, -0,21, -0,24, -0,27 et -0,30), mais la qualité des résultats n'étant pas supérieure, la valeur $s = -0{,}15$ a été retenue.

3.3.1. Préparation de la variable réponse

Ne pouvant pas modéliser une variable définie comme étant une probabilité ou une proportion avec une régression linéaire sans effectuer de transformation (voir section 2.2.3), la variable réponse étudiée dans la présente section correspond à

la transformation logit de la variable p_t (voir équation (2.2.8)), notée lp_t :

$$lp_t = \log\left(\frac{p_t}{1-p_t}\right).$$

Dans le cadre de ce projet, cette transformation occasionne cependant un problème, car les valeurs de lp_t deviennent très grandes ou très petites lorsque p_t tend vers 0 ou 1. Pour remédier à cette situation, nous avons décidé de tronquer les valeurs de p_t à trois décimales avant de les transformer. La transformation (2.2.8) a donc été appliquée à la variable p_t^* définie comme suit :

$$p_t^* = \max\left(0{,}001; \min(0{,}999, p_t)\right).$$

De cette façon, la nouvelle variable réponse est contrôlée afin d'éviter d'avoir des valeurs trop extrêmes. De plus, il n'y pas vraiment d'information perdue en effectuant la troncature car, dans le cadre de ce projet, remplacer une probabilité de neige inférieure à 0,001 par 0,001 signifie quand même qu'il est fort improbable qu'il y ait de la neige et vice-versa pour une probabilité supérieure à 0,999.

3.3.2. Variables explicatives à considérer

À cause de la nature des données, il est important de mentionner que l'hypothèse de non-corrélation/indépendance des observations n'est pas respectée. En effet, la présence/absence de neige dépend de la réalité neige/non-neige des journées précédentes. Une solution à cette dépendance dans les données aurait pû être de considérer les séries chronologiques. Toutefois, nous avons plutôt décidé de considérer les observations comme étant indépendantes et nous avons opté pour un modèle de régression linéaire multiple où les réalités neige/non-neige des journées précédentes sont incluses comme variables explicatives pour un délai allant de 1 à 7 jours.

Le modèle de régression linéaire multiple (voir équation (2.2.1)) est donc composé de la variable $\boldsymbol{y}$ (de longueur 365) qui représente la valeur du logit de la probabilité de neige journalière et de la matrice $\boldsymbol{X}$ (de dimension 365 par 9) qui contient un vecteur de 1, des vecteurs avec de l'information sur les probabilités de neige des sept derniers jours (lp_{t-1} à lp_{t-7}) et un vecteur correspondant à la température pour chaque journée :

$$\begin{aligned} E[y_t] = \beta_0 + \beta_1 lp_{t-1} + \beta_2 lp_{t-2} + \beta_3 lp_{t-3} + \beta_4 lp_{t-4} + \beta_5 lp_{t-5} \\ + \beta_6 lp_{t-6} + \beta_7 lp_{t-7} + \beta_T \text{Temp}_t \text{ (modèle complet).} \end{aligned}$$

En supposant que la sélection de variables ne retient que trois variables explicatives, le but est d'obtenir un modèle qui s'exprime de la façon suivante pour la journée t :

$$E[y_t] = \beta_0 + \beta_A lp_{A,t} + \beta_B lp_{B,t} + \beta_T \text{Temp}_t, \tag{3.3.1}$$

où l'on suppose que lp_A et lp_B sont les délais conservés par la sélection de variables et que Temp est la variable représentant la température. Un délai allant jusqu'à sept jours est utilisé, mais il n'a pas été dépassé, car les experts de l'Institut de recherche d'Hydro-Québec n'ont pas jugé pertinent d'utiliser des journées trop éloignées de la journée actuelle.

Pour ce qui est du choix de la variable donnant les informations sur la température journalière, les variables Tmin, Tmoy et Tmax sont disponibles quotidiennement à l'intérieur de chacun des points de grille (voir section 1.3.3.1). Toutefois, l'algorithme proposé à la section 3.1 ne calcule que les probabilités de neige de façon journalière et non au point de grille, c'est-à-dire qu'il ne fourni que la probabilité qu'il y ait de la neige pour la journée t pour la zone étudiée et non une probabilité individuelle pour chaque point de grille. Une statistique permettant de résumer l'information de tous ces points à une valeur unique pour la journée doit donc être utilisée. Pour ce faire, le minimum, la moyenne et le maximum des variables Tmin, Tmoy et Tmax ont été considérés (pour un total de neuf variables différentes).

Puisque ces variables donnent toutes de l'information semblable et qu'elles sont très corrélées (voir figure 3.3), elles ne peuvent donc pas être toutes utilisées dans le modèle de régression. En effet, si l'ensemble des neuf variables de température est utilisé, un problème de multicolinéarité serait inévitable. Dans ces conditions, une seule variable de température doit être conservée. Ce choix est effectué en calculant la corrélation entre chacune des variables de température à l'intérieur de la zone 2 et les probabilités de neige de cette même zone, permettant ainsi de voir quelle est la variable la plus informative. Voulant éventuellement que la régression soit appliquée pour l'ensemble du Québec, les probabilités de neige pour les zones 1, 3 et 4 ainsi que les corrélations avec leurs variables de température respectives ont été calculées. Le calcul du coefficient de corrélation linéaire est approprié, car chaque variable de température a bel et bien un lien linéaire avec la variable lp_t de sa zone respective, sauf la zone 1 où la relation est linéaire par partie (voir figures D.1 à D.4). Pour cette zone, si une partie de la relation entre les variables est négative, nous la désignons comme étant la partie descendante ($\searrow$). Si une partie de la relation est positive, alors elle est nommée la partie ascendante ($\nearrow$).

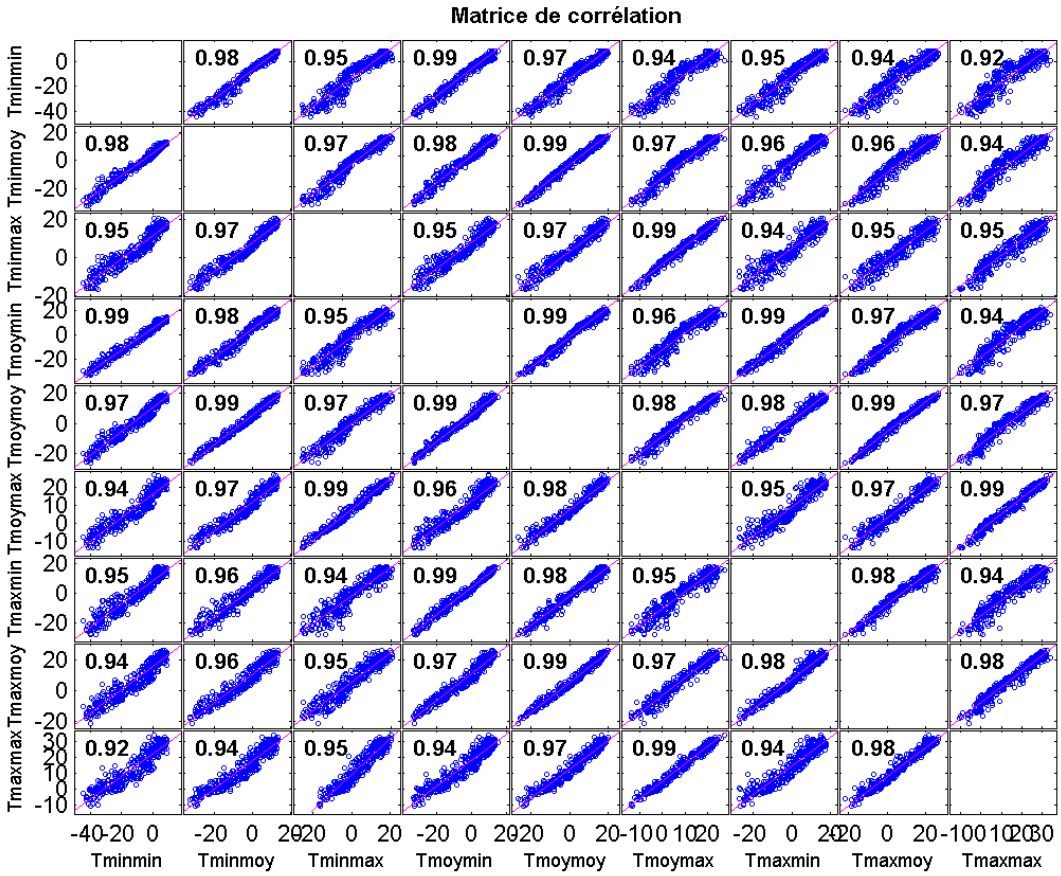

FIGURE 3.3. Graphique des nuages de points et coefficient de corrélation linéaire entre chacune des variables de température, 2011, Québec.

L'examen du tableau 3.2 permet de constater que la variable Tminmin affiche presque systématiquement la corrélation la plus élevée avec lp_t, peu importe la zone, sauf pour la partie ascendante de la zone 1. Cette variable est donc conservée comme information sur la température pour la suite du chapitre.

Tout comme les données de températures Tmin, Tmoy et Tmax, les valeurs des variables DJx et DCx sont disponibles de façon journalière à l'intérieur de chaque point de grille. Il faut donc encore une fois que l'information de tous les points de grille soit résumée afin d'avoir une seule valeur de degré-jour et de degré-chauffe par jour. Pour ce faire, le maximum est utilisé pour DJx et le minimum est utilisé pour DCx.

Afin de choisir le nombre de jours x sur lesquels DJx et DCx doivent être calculés, quelques valeurs de x parmi $x = 1, 2, 3, 4, 5, 6, 7$ sont confrontées. Pour ce faire, il faut vérifier à l'aide de nuages de points et de coefficients de corrélation

TABLE 3.2. Corrélation entre les variables de température de chaque zone et lp_t pour ces mêmes zones.

Variables	Description	Zone 1 ↘	Zone 1 ↗	Zone 2	Zone 3	Zone 4
Tminmin	Minimum des points de grille pour Tmin	-0,7601	0,6121	-0,8347	-0,8976	-0,8707
Tminmoy	Moyenne des points de grille pour Tmin	-0,7200	0,6647	-0,8139	-0,8825	-0,8622
Tminmax	Maximum des points de grille pour Tmin	-0,6596	0,7001	-0,7969	-0,8326	-0,8201
Tmoymin	Minimum des points de grille pour Tmoy	-0,7336	0,6394	-0,8000	-0,8803	-0,8646
Tmoymoy	Moyenne des points de grille pour Tmoy	-0,7484	0,5453	-0,7866	-0,8607	-0,8577
Tmoymax	Maximum des points de grille pour Tmoy	-0,7395	0,4692	-0,7765	-0,8276	-0,8020
Tmaxmin	Minimum des points de grille pour Tmax	-0,6879	0,4083	-0,7490	-0,8417	-0,8395
Tmaxmoy	Moyenne des points de grille pour Tmax	-0,6896	0,5210	-0,7435	-0,8239	-0,8439
Tmaxmax	Maximum des points de grille pour Tmax	-0,6829	0,4539	-0,7374	-0,7977	-0,7758

linéaire s'il y a au moins une de ces valeurs pour laquelle les variables DJx et DCx respectives semblent pertinentes pour expliquer la variable réponse (voir figures D.5 à D.8). Malgré que les coefficients de corrélation soient tous assez élevés, les relations ne sont pas tout à fait linéaires, mais plutôt linéaires par parties (relation linéaire soit positive ou négative suivie d'une relation nulle/effet de plateau). Cet effet de plateau illustre une absence de relation entre les variables, ce qui justifie que les variables DJx et DCx ne soient pas conservées.

3.3.3. Sélection de variables explicatives pour expliquer la variable réponse lp_t

Parmi les variables considérées à la section précédente, seules quelques-unes peuvent être pertinentes pour expliquer la variable réponse lp_t. La méthode de sélection de variables utilisée pour construire un modèle permettant d'expliquer la variable lp_t est la sélection à rebours. Le critère de retrait considéré est d'éliminer la variable ayant la plus grande valeur-p. Pour s'assurer d'un modèle parcimonieux, c'est-à-dire qui compte le moins de variables explicatives possible, nous nous sommes limités à $p^* = 2$, $p^* = 3$ ou $p^* = 4$ variables explicatives dans le modèle final, mais en imposant Tminmin si elle est significative (valeur-p du test-t inférieure à 0,05). Tminmin est systématiquement conservée tant qu'elle est significative, car il s'agit de la seule variable explicative apportant de l'information sur la journée courante.

Les tableaux 3.3 à 3.5 présentent les trois dernières étapes de la sélection de variables. Les premières étapes sont disponibles aux tableaux E.1 à E.4.

À la suite de cet exercice, trois modèles potentiels de la forme (3.3.1) sont obtenus :

$$\begin{aligned}
\widehat{lp}_t &= -0{,}6542 + 0{,}7491\ lp_{t-1} - 0{,}0508\ \text{Tminmin}\ (p^* = 2);\\
\widehat{lp}_t &= -0{,}5461 + 0{,}4639\ lp_{t-1} + 0{,}3455\ lp_{t-6} - 0{,}0429\ \text{Tminmin}\ (p^* = 3);\\
\widehat{lp}_t &= -0{,}5093 + 0{,}4208\ lp_{t-1} + 0{,}1895\ lp_{t-5} + 0{,}2185\ lp_{t-6}\\
&\quad - 0{,}0401\ \text{Tminmin}\ (p^* = 4).
\end{aligned}$$

Toutefois, avant de les utiliser pour construire le modèle de cartographie de la neige, les hypothèses de la régression linéaire doivent être vérifiées. La section qui suit est consacrée à cette validation.

3.3.4. Qualité du modèle et diagnostics des résidus

Le tableau 3.6, montre que les trois modèles sont significatifs (valeurs-p inférieures à 0,0001, voir équation (2.2.4)), donc qu'il y a au moins une variable qui est utile pour expliquer la variation de lp_t. De plus, ils expliquent assez bien la variation totale (R^2 ajusté respectivement 0,892, 0,911 et 0,916 pour les modèles à deux, trois et quatre variables explicatives (voir équation (2.2.6)) et il n'y a pas d'importants problèmes de multicolinéarité selon Montgomery *et al.* (2006) (VIF inférieurs à dix (voir équation (2.2.7)).

L'examen du graphique des résidus en fonction des valeurs ajustées pour chacun des modèles (voir figure 3.4) permet de constater qu'ils ne sont pas tout à

TABLE 3.3. Étape 5 de la sélection de variables à rebours pour la régression linéaire sur lp_t calculés de la zone 2 pour l'année 2011, Québec. Variable lp_{t-5} à retirer.

Étape 5 ($p^* = 4$)				
Variables	Estimation	Erreur standard	Statistique t	Valeur-p
origine	-0,5093	0,0894	-5,6968	2,6857e-08
lp_{t-1}	0,4208	0,0420	10,0153	8,2149e-21
lp_{t-5}	0,1895	0,0464	4,0839	5,5523e-05
lp_{t-6}	0,2185	0,0495	4,4093	1,4013e-05
Tminmin	-0,0401	0,0060	-6,7321	7,3311e-11

TABLE 3.4. Étape 6 de la sélection de variables à rebours pour la régression linéaire sur lp_t calculés de la zone 2 pour l'année 2011, Québec. Variable lp_{t-6} à retirer.

Étape 6 ($p^* = 3$)				
Variables	Estimation	Erreur standard	Statistique t	Valeur-p
origine	-0,5461	0,0903	-6,0449	3,9433e-09
lp_{t-1}	0,4639	0,0419	11,0708	1,5945e-24
lp_{t-6}	0,3455	0,0387	8,9179	3,0069e-17
Tminmin	-0,0429	0,0060	-7,1547	5,2129e-12

TABLE 3.5. Étape 7 de la sélection de variables à rebours pour la régression linéaire sur lp_t calculés de la zone 2 pour l'année 2011, Québec. Modèle le plus parcimonieux.

Étape 7 ($p^* = 2$)				
Variables	Estimation	Erreur standard	Statistique t	Valeur-p
origine	-0,6542	0,0984	-6,6483	1,1611e-10
lp_{t-1}	0,7491	0,0296	25,3511	8,4710e-81
Tminmin	-0,0508	0,0065	-7,8371	5,7789e-14

fait distribués uniformément autour de 0 en ordonnée. Toutefois, ils sont quand même globalement assez centrés en 0 et ils ne présentent pas de forte fluctuation d'étendue. Les hypothèses d'espérance nulle et de variance constante semblent donc respectées.

Les graphiques quantile-quantile des résidus standardisés (voir figure 3.5) montrent que les quantiles observés ne correspondent pas toujours aux quantiles théoriques. En effet, les quantiles observés prennent des valeurs trop fortement négatives et trop fortement positives par rapport aux quantiles théoriques de la loi normale. L'hypothèse de normalité n'est donc pas respectée.

Par contre, il est connu que les estimations ne sont pas affectées par la non-normalité (voir section 2.2.1.1). De plus, suffisamment d'observations sont disponibles (voir Ratcliffe, 1968) et les résidus sont assez symétriques (voir Geary,

TABLE 3.6. Valeur-p de la régression de lp_t, R^2 ajusté et facteur d'inflation de la variance (VIF) par variable pour chacun des modèles (p^*=2, 3 et 4), Québec zone 2, 2011.

Modèle	Valeur-p	R^2 ajusté	Variable	VIF
2	<0,0001	0,892	lp_{t-1}	2,7943
			Tminmin	2,7943
3	<0,0001	0,911	lp_{t-1}	6,8795
			lp_{t-6}	5,8222
			Tminmin	2,8663
4	<0,0001	0,916	lp_{t-1}	7,1620
			lp_{t-5}	8,7068
			lp_{t-6}	9,9051
			Tminmin	2,9468

1947 et figure 3.6), ce qui fait en sorte que les tests-t vérifiant si les coefficients de régression sont significatifs (voir (2.2.5)) soient tout de même interprétables.

3.3.5. Production des cartographies de neige et non-neige

Puisque les modèles semblent adéquats, ils peuvent être utilisés pour cartographier la probabilité de neige d'une journée t pour chacun des points de grille du Québec. Pour ce faire, chacun des modèles est utilisé pour chaque point de grille i afin d'estimer $lp_{t,i}$. Ensuite, la transformation inverse de l'équation (2.2.9) est appliquée afin d'obtenir $p_{t,i}$, qui est la probabilité de neige pour la journée t dans le point de grille i. Par exemple, pour le modèle avec deux variables explicatives :

$$\widehat{lp}_{t,i} = -0{,}6542 + 0{,}7491\ lp_{t-1,i} - 0{,}0508\ \mathrm{Tmin}_{t,i};$$
$$\widehat{p}_{t,i} = \frac{1}{1 + e^{-\widehat{lp}_{t,i}}}.$$

Évidemment, puisque les informations sont disponibles au niveau des points de grille, il aurait été préférable d'ajuster un modèle pour chacun d'entre eux et non pour la journée. Toutefois, une telle approche était trop complexe pour répondre au besoin d'Hydro-Québec et au niveau opérationnel. Il a donc été privilégié de trouver un modèle journalier et de l'appliquer à chacun des points de grille, même si ce n'est pas la méthode optimale à utiliser.

Une problématique associée à l'utilisation de cette approche est qu'il faut connaître la probabilité de neige des premiers jours pour débuter les calculs. En effet, pour calculer $lp_{t,i}$, il faut que $lp_{t-1,i}$, $lp_{t-5,i}$ et $lp_{t-6,i}$ aient déjà été calculées (selon le modèle utilisé). Il faut donc initialiser tous les points de grille des journées $t = 1, 2, 3, 4, 5, 6$, afin que tous les délais soient disponibles à partir de $t = 7$. Pour ce faire, la valeur $lp_{t,i} = \log\left(\frac{0{,}999}{1-0{,}999}\right) \simeq 6{,}9068$ leur est attribuée. Cette décision

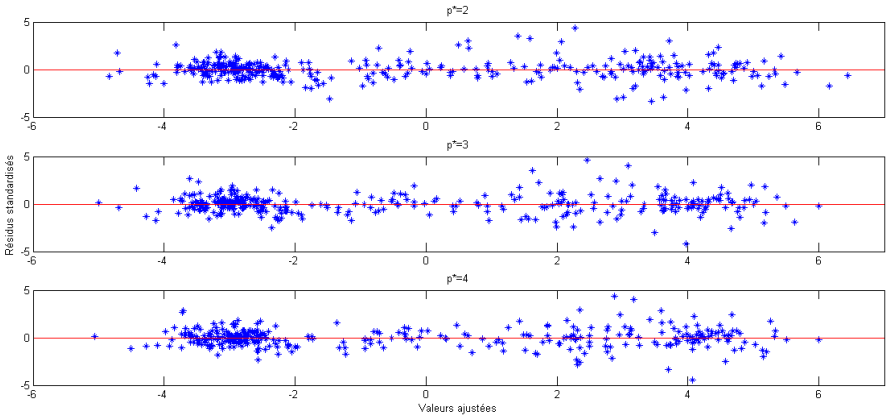

FIGURE 3.4. Graphique des résidus standardisés en fonction de valeurs ajustées pour la régression de lp_t avec deux, trois et quatre variables explicatives, zone 2 du Québec, 2011.

.

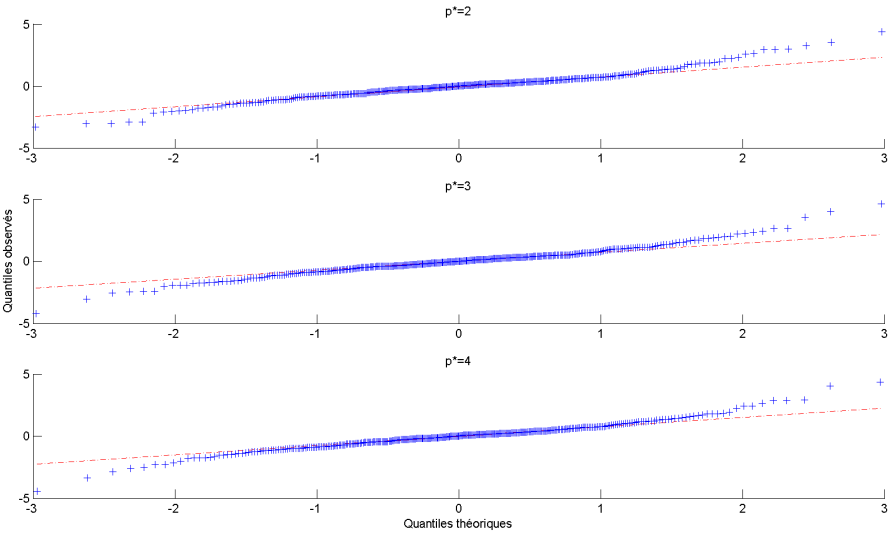

FIGURE 3.5. Diagramme quantile-quantile des résidus standardisés pour la régression de lp_t avec deux, trois et quatre variables explicatives, zone 2 du Québec, 2011.

.

est logique, car il est connu qu'il y a de la neige partout sur le Québec pour les journées du 1er au 6 janvier.

Les cartographies produites selon cette méthode affichent un comportement qui semble cohérent avec la réalité, peu importe le modèle (voir figure 3.7) :

– l'année commence bel et bien avec une forte probabilité de neige partout sur la zone d'intérêt (14 février) ;

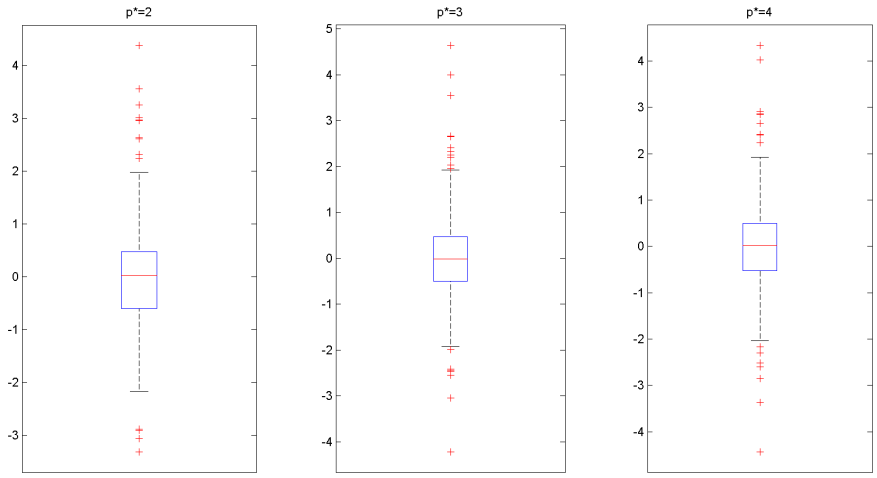

FIGURE 3.6. Diagrammes en boîte des résidus standardisés pour la régression des lp_t avec deux, trois et quatre variables explicatives, zone 2 du Québec, 2011.

– la fonte de la neige est bien amorcée et elle est caractérisée par la diminution de la probabilité de neige du sud vers le nord (15 avril) ;
– l'été est représenté par la journée du 19 juillet, où la probabilité de neige est très faible ou même nulle sur toute la zone ;
– le retour de la neige est défini par une augmentation de la probabilité de neige du nord vers le sud (10 décembre) ;
– de façon générale, l'évolution journalière des cartographies est réaliste, car les variations d'une journée à l'autre sont raisonnables et représentent bien le phénomène étudié (voir figure 3.8).

Puisque les modèles de régression à deux, trois et quatre variables explicatives semblent adéquats pour modéliser la variable lp_t et que les cartographies obtenues sont assez satisfaisantes, nous allons donc procéder au même exercice pour les autres zones.

3.4. MODÉLISATION DANS LES AUTRES ZONES

Pour modéliser la variable lp_t dans chacune des trois autres zones, la méthodologie utilisée pour la zone 2 est appliquée.

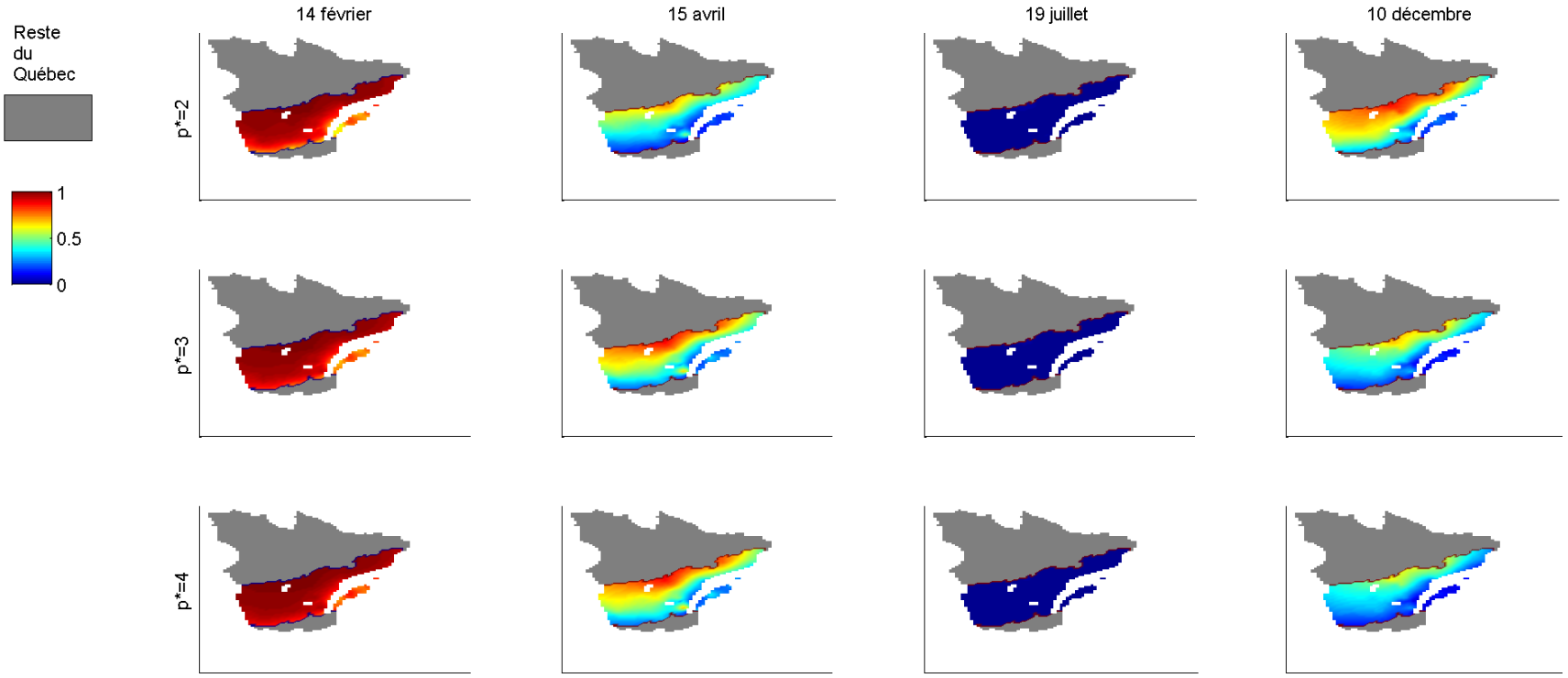

FIGURE 3.7. Exemple de cartographies de la probabilité de neige par point de grille pour zone 2 du Québec pour l'année 2011 où la cartographie semble cohérente avec la réalité.

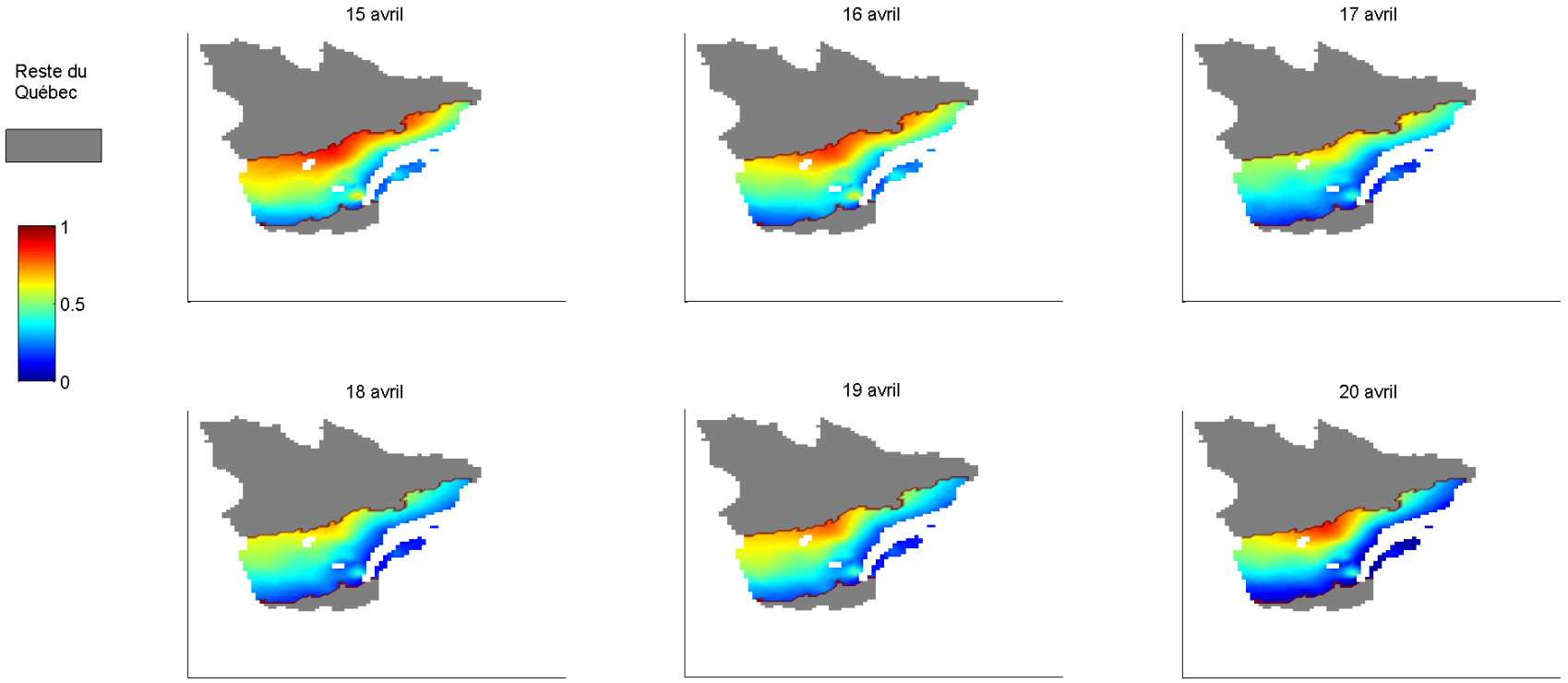

FIGURE 3.8. Exemple de cartographies de la probabilité de neige par point de grille pour zone 2 du Québec pour l'année 2011 où l'évolution d'une journée à l'autre semble réaliste, modèle à quatre variables explicatives.

3.4.1. Sélection des modèles

Les étapes de la sélection de variables pour chaque zone sont disponibles en annexe aux tableaux E.5 à E.25.

Comme il est possible d'observer aux tableaux 3.7 à 3.9, chacun des modèles contient les variables lp_{t-1} et Tminmin, incluant également ceux pour la zone 2 (tableau 3.6). Cela signifie que l'information sur la probabilité de neige de la veille ainsi que sur la température de la journée courante est importante partout au Québec pour expliquer la probabilité de neige de la journée actuelle. Toutefois, pour ce qui est des autres variables donnant de l'information sur la probabilité de neige des journées précédentes, ce ne sont pas toujours les mêmes délais qui sont le plus important selon la zone. Cela implique que les modèles ont des variables explicatives qui diffèrent d'une zone à l'autre.

3.4.2. Qualité des modèles et diagnostics des résidus

La prochaine étape est de vérifier la qualité des différents modèles. Toujours dans les mêmes tableaux, il est possible de remarquer que tous les modèles ont une valeur-p inférieure à 0,05, peu importe la zone. De plus, les coefficients de détermination ajustés (R^2 ajusté) sont tous suffisamment élevés pour permettre d'affirmer que chaque modèle explique assez bien la variation de la variable réponse. Toutefois, pour les zones 3 et 4, les modèles à trois et quatre variables explicatives ont un problème de multicolinéarité, car certains VIF sont supérieurs à dix.

Les figures E.1 à E.9 affichent des conclusions similaires à celles énoncées pour la zone 2, c'est-à-dire que les résidus standardisés sont approximativement uniformément répartis autour de 0 en ordonnée, mais que l'hypothèse de normalité ne semble pas respectée. Toutefois, il y a encore un grand nombre d'observations et les résidus sont symétriques, donc l'inférence sur chacune des variables explicatives est considérée valide.

3.4.3. Production des cartographies de neige et non-neige

Les diagnostics de chacun des modèles étant assez satisfaisants malgré le problème de multicolinéarité pour les zones 3 et 4, les cartographies des probabilités de neige sont produites. Pour ce faire, la méthodologie de la zone 2 est utilisée, mais en appliquant à un point de grille un modèle associé à la zone à laquelle il appartient. Afin de tenir compte du fait que le nombre de variables explicatives optimal peut différer d'une zone à l'autre, toutes les combinaisons possibles ont été cartographiées.

TABLE 3.7. Valeur-p de la régression de lp_t, R^2 ajusté et facteur d'inflation de la variance (VIF) par variable pour chacun des modèles (p^*=2, 3 et 4), Québec zone 1, 2011.

Modèle	Valeur-p	R^2 ajusté	Variable	VIF
2	<0,0001	0,735	lp_{t-1}	1,4889
			Tminmin	1,4889
3	<0,0001	0,795	lp_{t-1}	2,7157
			lp_{t-6}	2,5257
			Tminmin	1,5242
4	<0,0001	0,811	lp_{t-1}	3,1995
			lp_{t-3}	3,2188
			lp_{t-6}	3,1430
			Tminmin	1,5438

TABLE 3.8. Valeur-p de la régression de lp_t, R^2 ajusté et facteur d'inflation de la variance (VIF) par variable pour chacun des modèles (p^*=2, 3 et 4), Québec zone 3, 2011.

Modèle	Valeur-p	R^2 ajusté	Variable	VIF
2	<0,0001	0,951	lp_{t-1}	4,6349
			Tminmin	4,6349
3	<0,0001	0,957	lp_{t-1}	13,3691
			lp_{t-7}	10,1835
			Tminmin	4,6272
4	<0,0001	0,958	lp_{t-1}	16,7351
			lp_{t-3}	18,1281
			lp_{t-7}	13,8318
			Tminmin	4,8017

TABLE 3.9. Valeur-p de la régression de lp_t, R^2 ajusté et facteur d'inflation de la variance (VIF) par variable pour chacun des modèles (p^*=2, 3 et 4), Québec zone 4, 2011.

Modèle	Valeur-p	R^2 ajusté	Variable	VIF
2	<0,0001	0,956	lp_{t-1}	3,9125
			Tminmin	3,9125
3	<0,0001	0,964	lp_{t-1}	17,7747
			lp_{t-3}	16,9265
			Tminmin	3,9937
4	<0,0001	0,965	lp_{t-1}	21,3013
			lp_{t-3}	23,9043
			lp_{t-5}	20,7399
			Tminmin	3,9895

À la suite d'un examen visuel de toutes les cartographies produites, il a été conclu que les résultats obtenus n'étaient pas tout à fait satisfaisants. En effet, bien que les cartographies obtenues indiquent la probabilité de neige pour chacun

des points de grille du Québec, le comportement de celles-ci n'est pas considéré comme étant réaliste, peu importe le nombre de variables explicatives utilisées. Par exemple, un premier problème est l'effet de coupure entre chacune des zones dû à l'utilisation de différents modèles (voir figure 3.9). En effet, cela rend l'interprétation plus ardue, puisqu'il y a une discontinuité entre les probabilités sous la coupure et celles directement au-dessus. Aussi, une autre problématique est due au fait que certaines cartographies (voir la même figure) affichent des probabilités de neige quasi-nulles au sud de la zone 2, alors qu'elles sont encore assez fortes dans la zone 1. Un tel comportement n'est pas cohérent avec la réalité.

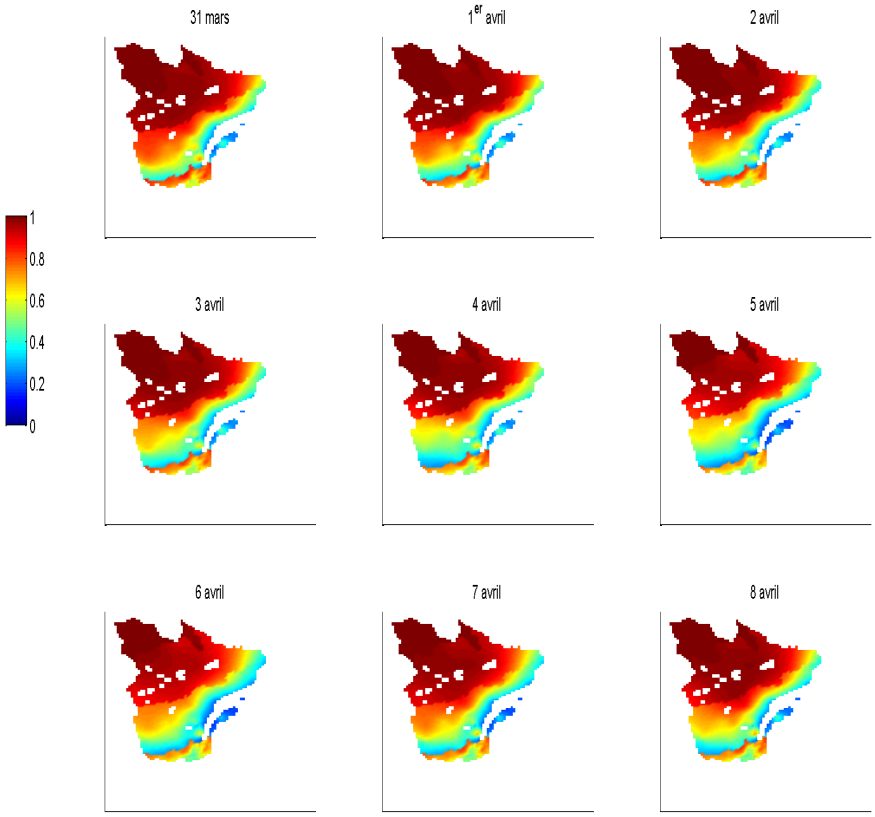

FIGURE 3.9. Exemple de cartographies pour le Québec pour l'année 2011 avec un modèle par zone avec $p^* = 3$ où il y a des coupures entre les zones et où les probabilités de neige dans la zone 1 sont fortes, mais quasi-nulles au sud de la zone 2.

3.5. Québec en entier

Il a été conclu à la section précédente que les résultats n'étaient pas réalistes lorsqu'un modèle différent était utilisé selon la zone du Québec. C'est pourquoi la méthodologie a été modifiée afin qu'elle permette d'obtenir un seul modèle pour tout le Québec, sans perdre l'information quant aux différentes zones.

3.5.1. Construction du modèle

Afin de développer un tel modèle, la même variable réponse et les mêmes variables explicatives ont été conservées, mais en ajoutant de l'information sur les zones. Cette information peut être introduite indirectement par le calcul des variables ou par l'ajout de trois variables dichotomiques qui permettent de déterminer de quelle zone provient une donnée :

$$\text{ind\_z}j = \begin{cases} 1 & \text{si la donnée provient de la zone j;} \\ 0 & \text{sinon,} \end{cases}$$

où $j = 2, 3, 4$. Si les trois variables sont nulles, alors la donnée provient de la zone 1.

Les différentes étapes de la sélection de variables à rebours (voir tableaux E.26 à E.30 pour les premières étapes et les tableaux 3.10 à 3.12 pour les autres étapes) permettent de constater que les variables dichotomiques sont éliminées dès le début de la procédure, donc qu'elles n'étaient pas considérées statistiquement utiles pour expliquer la variable lp_t, car l'information est déjà présente indirectement dans les autres variables. Aussi, les variables lp_{t-1} et Tminmin, sont encore une fois présentes dans tous les modèles ($p^* = 2, 3, 4$). Pour le modèle à trois et quatre variables explicatives, lp_{t-6} et lp_{t-3} s'ajoutent.

Table 3.10. Étape 6 de la sélection de variables à rebours pour la régression linéaire sur lp_t calculés pour l'année 2011, Québec. Variable lp_{t-3} à retirer.

Étape 6 ($p^* = 4$)				
Variables	Estimation	Erreur standard	Statistique t	Valeur-p
origine	-0,3339	0,0514	-6,4986	1,1353e-10
lp_{t-1}	0,4632	0,0220	21,0901	1,2089e-85
lp_{t-3}	0,1794	0,0220	8,1496	8,2070e-16
lp_{t-6}	0,2556	0,0213	12,0257	9,9707e-32
Tminmin	-0,0347	0,0038	-9,2192	1,1049e-19

TABLE 3.11. Étape 7 de la sélection de variables à rebours pour la régression linéaire sur lp_t calculés pour l'année 2011, Québec. Variable lp_{t-6} à retirer.

Étape 7 ($p^* = 3$)				
Variables	Estimation	Erreur standard	Statistique t	Valeur-p
origine	-0,3948	0,0528	-7,4723	1,3951e-13
lp_{t-1}	0,5394	0,0204	26,4339	7,0640e-125
lp_{t-6}	0,3383	0,0192	17,6110	9,1167e-63
Tminmin	-0,0395	0,0039	-10,2602	7,4940e-24

TABLE 3.12. Étape 8 de la sélection de variables à rebours pour la régression linéaire sur lp_t calculés pour l'année 2011, Québec. Modèle le plus parcimonieux. 0* signifie que la valeur est trop près de 0 pour le logiciel Matlab, donc il affiche uniquement la valeur 0.

Étape 8 ($p^* = 2$)				
Variables	Estimation	Erreur standard	Statistique t	Valeur-p
origine	-0,4780	0,0577	-8,2773	2,9145e-16
lp_{t-1}	0,8293	0,0132	62,7121	0*
Tminmin	-0,0476	0,0042	-11,3551	1,2117e-28

3.5.2. Qualité du modèle et diagnostics des résidus

Comme pour toutes les autres méthodologies impliquant la régression linéaire multiple, il faut vérifier la qualité des modèles obtenus ainsi que le comportement de leurs résidus.

Le tableau 3.13 montre que les trois modèles ont une valeur-p inférieure à 0,05, que les coefficients de détermination ajustés sont très élevés (R^2 ajusté $>$ 0,9) et que les VIF sont tous inférieurs à dix, peu importe le modèle.

TABLE 3.13. Valeur-p de la régression de lp_t, R^2 ajusté et facteur d'inflation de la variance (VIF) par variable pour chacun des modèles (p^*=2, 3 et 4), Québec, 2011.

Modèle	Valeur-p	R^2 ajusté	Variable	VIF
2	<0,0001	0,903	lp_{t-1}	2,5140
			Tminmin	2,5140
3	<0,0001	0,920	lp_{t-1}	7,2246
			lp_{t-6}	6,3822
			Tminmin	2,5511
4	<0,0001	0,926	lp_{t-1}	8,9235
			lp_{t-3}	8,9661
			lp_{t-6}	8,3355
			Tminmin	2,5934

Pour ce qui est des diagnostics des résidus, la figure 3.10 permet de constater que ceux-ci semblent assez bien répartis autour de 0 en ordonnée et que leur variance semble constante (les points agglutinés sous forme de droite dans les coins inférieurs gauches et supérieurs droits des nuages de points sont dus à la transformation (3.3.1)). De plus, même s'ils ne semblent pas respecter l'hypothèse de normalité (voir figure 3.11), les résidus sont symétriques (voir figure 3.12) et le nombre d'observations est grand. L'inférence sur les variables explicatives est donc valide.

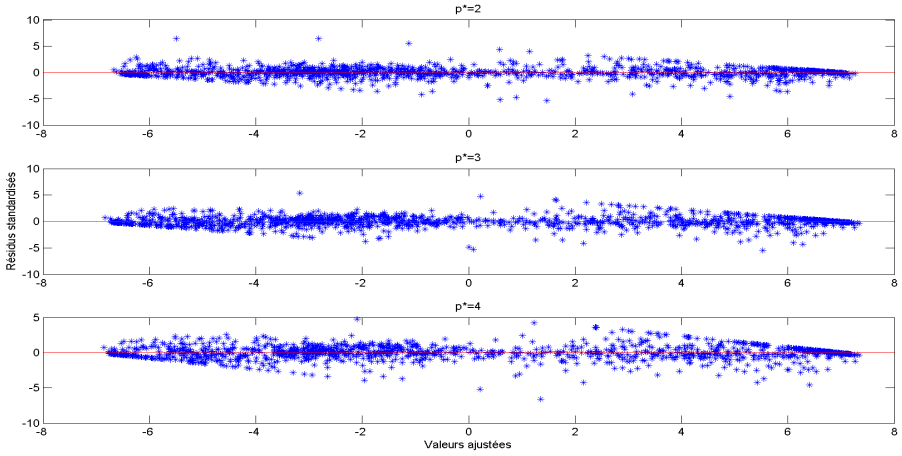

FIGURE 3.10. Graphique des résidus standardisés en fonction de valeurs ajustées pour la régression de lp_t avec deux, trois et quatre variables explicatives, Québec, 2011.

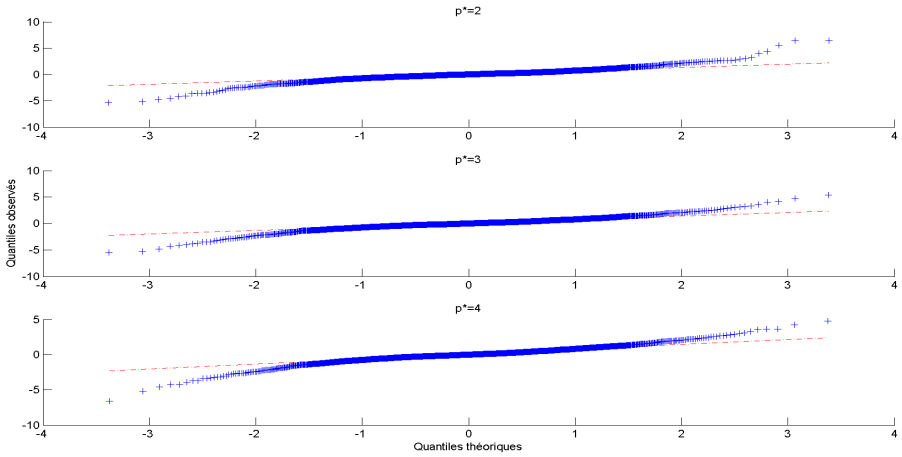

FIGURE 3.11. Diagramme quantile-quantile des résidus standardisés pour la régression de lp_t avec deux, trois et quatre variables explicatives, Québec, 2011.

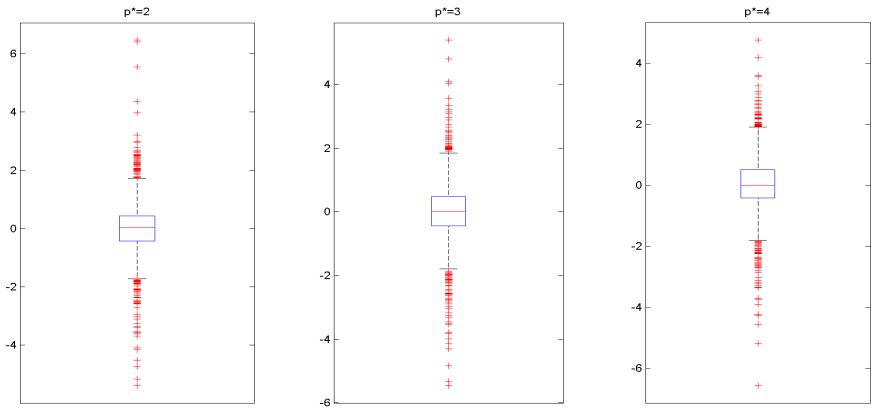

FIGURE 3.12. Diagrammes en boîte des résidus standardisés pour la régression des lp_t avec deux, trois et quatre variables explicatives, Québec, 2011.

3.5.3. Production des cartographies de neige et non-neige

Puisque les trois modèles potentiels semblent adéquats, leurs cartographies respectives peuvent être produites en procédant de la même façon qu'à la section 3.3.5, mais pour le Québec en entier et non seulement la zone 2.

En examinant les différentes cartographies produites, il est possible de remarquer qu'elles semblent assez cohérentes avec la réalité (voir figures 3.13 et 3.14) :

- il y a bel et bien de la neige en hiver (14 février), car les probabilités de neige sont assez fortes ;
- il y a une fonte au printemps (15 avril), car les probabilités de neige sont de plus en plus faibles du sud vers le nord ;
- il n'y a pas de neige en été (19 juillet), car les probabilités de neige sont quasi-nulles ;
- il y a une accumulation de neige à l'automne (10 décembre), car les probabilités de neige sont de plus en plus élevées du nord vers le sud ;
- de façon générale, l'évolution journalière des probabilités de neige est raisonnable, car il n'y a pas de transition brusque et il est possible de voir les cartographies évoluer peu à peu comme le fait généralement le couvert nival. Aussi, les cartographies n'ont pas de discontinuité, ce qui est un avantage par rapport à la méthode étudiée à la section 3.4.

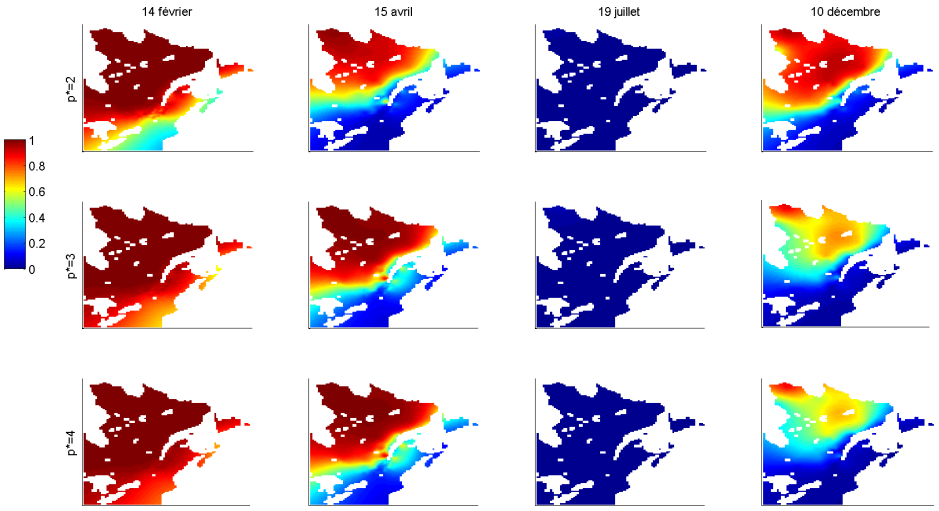

FIGURE 3.13. Exemple de cartographies de la probabilité de neige par point de grille pour le Québec pour l'année 2011 où la cartographie semble cohérente avec la réalité.

.

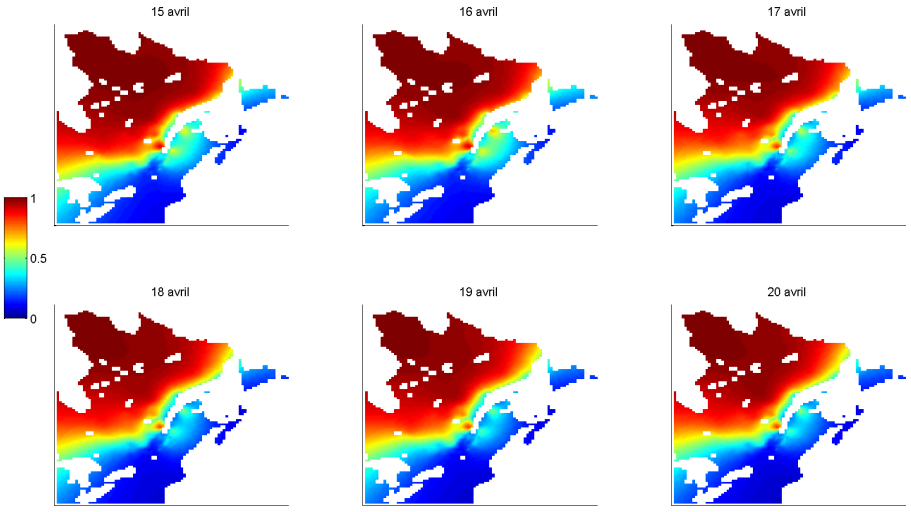

FIGURE 3.14. Exemple de cartographies de la probabilité de neige par point de grille pour le Québec pour l'année 2011 où l'évolution d'une journée à l'autre semble réaliste.

3.6. Conclusion partielle

Dans ce chapitre, les logits des probabilités de neige ont été modélisées à l'aide de la régression linéaire multiple. Tout d'abord, le travail a été effectué uniquement dans la zone 2 et il a été conclu que la méthode produisait des cartographies de qualité satisfaisante. L'expérience a été répétée pour les zones 1, 3 et 4 et les cartographies pour l'ensemble du territoire québécois ont été produites en jumelant les résultats pour chacune des zones. Celles-ci sont toutefois peu réalistes, car il y a des discontinuités entre les zones ainsi que quelques incohérences telles que la fonte de la zone 2 qui précède celle de la zone 1.

Une alternative considérée pour remédier à la situation a été d'obtenir un seul modèle pour le Québec en entier et non un modèle différent pour chacune des zones. Cette nouvelle approche a permis d'obtenir des cartographies du Québec ayant un comportement cohérent avec la réalité, et cela sans discontinuité entre les zones. Les résultats étant satisfaisants, nous pouvons donc passer à l'étape de validation qui consiste à comparer les cartographies obtenues avec des données de terrain et d'autres méthodes de cartographies.

Chapitre 4

ÉTUDE DE LA QUALITÉ DES CARTOGRAPHIES PRODUITES

Dans ce chapitre, la qualité des cartographies produites par les modèles à deux, trois et quatre variables explicatives construits au chapitre précédent est quantifiée. Jusqu'à présent, un simple examen visuel qui tend à démontrer que les résultats sont satisfaisants a été effectué. Pour confirmer l'efficacité de l'approche, il faut maintenant comparer les produits qui ont été développés avec des modèles de référence.

Tout d'abord, les données ainsi que les outils qui ont été utilisés pour effectuer cette validation sont présentés. Par la suite, les résultats pour l'année 2011 sont analysés afin d'évaluer si les modèles produisent de bons résultats pour l'année ayant servie à leur construction.

4.1. Données de référence et outils pour la validation

Dans cette section, les données de référence avec lesquelles les modèles ont été comparés sont présentées. Il s'agit des valeurs ponctuelles des capteurs SR50 et GMON (voir section 1.2.1.2) ainsi que des cartes de neige/non-neige produites avec l'algorithme de segmentation hiérarchique des données SSM/I (voir section 1.2.2.1). Pour simplifier l'appellation, ces cartes de neige/non-neige seront référées comme étant les cartes SSM/I.

Pour vérifier la similarité entre les cartographies qui sont produites par les modèles développés, trois outils sont utilisés : le score de Brier, le pourcentage de concordance et le critère d'erreurs d'omission et de commission.

4.1.1. SR50 et GMON

Les avantages d'utiliser les valeurs ponctuelles des capteurs SR50 et GMON sont qu'elles sont disponibles en temps réel et qu'elles représentent bien la réalité

à la station. Un bon modèle devra donc au moins reproduire la réalité des capteurs à l'intérieur de leur point de grille respectif. Puisque la réalité qui nous intéresse est la présence/absence de neige, les valeurs supérieures à 0 cm représentent théoriquement la présence de neige et les autres le sol sans neige. Toutefois, la réalité d'un capteur à son point de grille respectif n'est pas exactement la même qu'à la station (voir section 1.2.1). Pour résoudre le problème, un seuillage à 1 ou 2 cm pour les SR50 et 1 cm pour les GMON a été suggéré par les chercheurs de l'Institut de recherche d'Hydro-Québec. Afin de vérifier la sensibilité des algorithmes aux choix des seuils, des valeurs de 1 à 5 cm ont été systématiquement analysées.

Afin de pouvoir comparer les sorties des modèles aux données de terrain, les cartes de probabilités de neige ont été croisées avec ces données. Dans l'ensemble, l'information d'une station correspond à un seul point de grille de la carte de probabilités, mais ce n'est pas toujours le cas pour les capteurs GMON pour lesquels il arrive qu'un point de grille contienne plus d'un capteur et que leur mesure respective se contredise (un capteur au sud du point de grille peut indiquer que la neige est fondue alors qu'un capteur au nord de ce même point de grille peut indiquer que la neige est encore présente). Afin de tenir compte de cette information multiple, une validation différente a été effectuée pour chaque combinaison permettant de prendre un seul capteur par point de grille. Par exemple, si un même point de grille contient les capteurs A et B et qu'un autre point de grille contient les capteurs C et D, quatre validations sont réalisées (AC, BC, AD et BD).

4.1.2. Cartes SSM/I

L'utilité des cartes SSM/I est qu'elles offrent la possibilité de comparer un à un tous les points de grille et non seulement ceux contenant un capteur SR50 ou GMON. De plus, elles ont la même résolution que les données de GTV utilisées.

Puisqu'une carte SSM/I donne directement l'information sur la présence ou l'absence de neige, il n'y a pas de manipulation à effectuer. Par contre, il ne faut pas négliger le fait que les cartes SSM/I sont le résultat d'un autre modèle, donc qu'elles ne sont pas une représentation exacte de la réalité. Les modèles développés dans ce projet sont considérés satisfaisants si les cartographies obtenues sont semblables aux cartes SSM/I, sans nécessairement être identiques.

4.1.3. Score de Brier

Le score de Brier permet d'évaluer la qualité d'une prévision probabiliste d'une variable ne pouvant prendre qu'un nombre limité de valeurs (voir Brier, 1950 et Wilks, 2011). Dans le cadre de ce projet, l'utilisation de cet outil est

donc appropriée, car nous souhaitons évaluer la qualité des probabilités de neige produites par les différents modèles en les comparant avec les réalités neige/non-neige correspondantes.

Dans le contexte de la présente étude, le score de Brier (SB) est défini comme suit :

$$\mathrm{SB} = \frac{1}{n}\sum_{i=1}^{n}\left(p_i - r_i\right)^2, \tag{4.1.1}$$

où n est le nombre de points de grille ayant un capteur, p_i est la probabilité de neige estimée au point de grille i et

$$r_i = \begin{cases} 1 & \text{si neige;} \\ 0 & \text{si non-neige.} \end{cases}$$

Le score de Brier prend ses valeurs dans l'intervalle [0, 1]. Plus le score de Brier est faible, meilleure est la prévision. Dans le contexte du projet, un score de Brier est bon si les probabilités de neige sont faibles lorsque la réalité est l'absence de neige et qu'elles sont fortes si la réalité est la présence de neige.

4.1.4. Proportion de concordance

Le calcul de la proportion de concordance (PC) entre deux cartographies est un indicateur intuitif de la qualité d'un modèle. En effet, il calcule la proportion de cas pour lesquels les diagnostics de neige ou non-neige sont identiques. Cet indicateur est défini comme suit :

$$\mathrm{PC} = \frac{\mathrm{OO} + \mathrm{NN}}{n}, \tag{4.1.2}$$

où OO (respectivement NN) est le nombre de cas où les valeurs au point de grille de notre modèle et de la carte de référence indiquent simultanément une présence (respectivement absence) de neige et n est le nombre total de points de grille qui ont été comparés. Un modèle est considéré satisfaisant si la proportion de concordance est élevée, car cela signifie que les probabilités estimées aux points de grille sont très souvent en accord avec la réalité proposée par le produit de référence. Notons ici que l'interprétation des résultats est différente par rapport au score de Brier, car 0 est la meilleure valeur pour le score de Brier tandis que 1 est la meilleure valeur pour la proportion de concordance.

Dans le cadre de ce projet, le modèle développé n'indique pas directement la présence/absence de neige, mais plutôt des probabilités de neige. Afin de pouvoir calculer une proportion de concordance, il faut donc transformer les probabilités en réalisations neige/non-neige en utilisant neuf points de césure variant de 0,1 à

0,9 par bonds de 0,1. Si la probabilité de neige est inférieure au point de césure, alors il ne s'agit pas de neige, sinon il s'agit de neige.

4.1.5. Erreurs d'omission et de commission

Le critère d'erreurs d'omission/commission (OC) est un outil utilisé par l'Institut de recherche d'Hydro-Québec pour évaluer le comportement des cartographies par rapport à des données de référence. Il indique si les réalisations neige/non-neige produites par un modèle sont bonnes ou non pour chaque journée pour un point de grille :

$$\mathrm{OC} = \begin{cases} 1 & \text{erreur de commission;} \\ 0 & \text{pas d'erreur ;} \\ -1 & \text{erreur d'omission.} \end{cases} \tag{4.1.3}$$

Une erreur de commission est commise lorsque le modèle indique de la neige alors que la référence n'en indique pas. Au contraire, il y a une erreur d'omission si le modèle n'indique pas de neige, mais que la référence indique qu'il y en a.

4.2. Présentation des résultats de la validation pour 2011

Le but de cette section est de vérifier si les modèles à deux, trois et quatre variables explicatives ($p^* = 2$, 3 ou 4) construits avec les données de l'année 2011 produisent des cartographies représentant bien la réalité de cette même année. Pour ce faire, les scores de Brier (voir équation (4.1.1)) et les proportions de concordance (voir équation (4.1.2)) sont calculés pour les données SR50, GMON et SSM/I. Par la suite, le critère de l'omission/commission (voir équation (4.1.3)) est calculé pour quelques capteurs SR50 et pour l'ensemble des capteurs GMON.

4.2.1. Validation à l'aide du score de Brier

Le premier outil de validation utilisé est le score de Brier. Les valeurs de ce score ont été évaluées en comparant les probabilités de présence de neige produites par nos modèles avec les réalisations neige/non-neige des différentes données de référence.

4.2.1.1. *Scores de Brier avec les données des capteurs SR50*

Tout d'abord, les scores de Brier pour les trois modèles (voir section 3.5) ont été calculés en utilisant les réalisations neige/non-neige des capteurs SR50 pour les seuils de neige allant de 1 à 5 cm. L'examen de la figure 4.1 montre que, peu importe le seuil ou le modèle utilisé, les scores de Brier sont toujours très près

de 0 pour les périodes de neige et de sol. Cela signifie qu'ils fournissent pour ces périodes une probabilité de neige près de 1 lorsqu'il y a de la neige et une probabilité proche de 0 lorsqu'il n'y a pas de neige. Les modèles à deux, trois et quatre variables explicatives semblent donc adéquats lorsqu'il faut prévoir la présence ou l'absence de neige sur l'intégralité du domaine d'étude. Toutefois, comme l'indiquent les scores de Brier plus élevés, la qualité des sorties cartographiques n'est pas aussi satisfaisante pour les périodes de fonte et d'accumulation. Cette augmentation du score de Brier peut être expliquée de deux façons. Premièrement, cela peut s'expliquer par l'incertitude d'un modèle par rapport à la réalité (référé comme un problème d'incertitude). En effet, dès que les probabilités de neige se rapprochent de 0,5, c'est-à-dire lorsque le modèle ne sait pas s'il s'agit de neige ou de sol, la probabilité est évidemment éloignée de la réalité du capteur (une probabilité de 0,5 est autant loin de la réalisation 0 (non-neige) que 1 (neige)). D'autre part, l'augmentation du score peut être due au fait que le modèle ne prévoit pas correctement l'état du sol (référé comme un problème de mauvaise réalité ou de cohérence). En effet, si le modèle fournit une forte probabilité de neige alors que le capteur SR50 n'en indique pas ou vice-versa, alors le score de Brier augmente, et cette augmentation est plus importante que pour un problème d'incertitude (voir exemple au tableau 4.1). Puisque les scores de Brier qui sont différents de 0 lors de la période de fonte sont majoritairement inférieurs à 0,5, il semble raisonnable de supposer qu'il s'agit majoritairement d'un problème d'incertitude. Cependant, les valeurs du score sont plus près de 1 pour les modèles à trois et quatre variables explicatives lors de la période d'accumulation de neige, ce qui semble signifier que les résultats de ces modèles ne sont pas satisfaisants pour cette période de l'année dû à un problème de cohérence avec la réalité.

TABLE 4.1. Exemples de scores de Brier obtenus en cas d'incertitude ou de mauvaise réalité.

Réalité	Incertitude		Mauvaise réalité	
	Probabilité	Contribution	Probabilité	Contribution
0	0,4	$(0{,}4-0)^2 = 0{,}16$	0,8	$(0{,}8-0)^2 = 0{,}64$
0	0,5	$(0{,}5-0)^2 = 0{,}25$	0,9	$(0{,}9-0)^2 = 0{,}81$
0	0,6	$(0{,}6-0)^2 = 0{,}36$	1,0	$(1{,}0-0)^2 = 1{,}00$
1	0,4	$(0{,}4-1)^2 = 0{,}36$	0,0	$(0{,}0-1)^2 = 1{,}00$
1	0,5	$(0{,}5-1)^2 = 0{,}25$	0,1	$(0{,}1-1)^2 = 0{,}81$
1	0,6	$(0{,}6-1)^2 = 0{,}16$	0,2	$(0{,}2-1)^2 = 0{,}64$
Score de Brier	0,2567		0,8167	

Toujours à la figure 4.1, il est également possible de remarquer que les trois modèles semblent fournir des cartographies de qualité équivalente lorsqu'il n'y a que de la neige ou du sol. Par contre, les modèles à trois et quatre variables explicatives produisent des cartographies de qualité supérieure au modèle à deux variables explicatives en période de fonte, alors que c'est plutôt l'inverse qui est observé lors de l'accumulation de neige au sol. Aussi, les modèles à trois et quatre variables explicatives semblent assez équivalents, peu importe la période de l'année.

La figure 4.2 montre que les scores de Brier sont plus faibles pour les grandes valeurs du seuil. De plus, les scores de Brier suivent tous le même comportement temporel, peu importe le modèle considéré. La qualité des modèles n'est donc pas trop sensible au choix du seuil et il est possible de poursuivre cette étude de la validation avec un seul d'entre eux. Ayant vu à la section précédente que les seuils suggérés par les chercheurs de l'Institut de recherche d'Hydro-Québec étaient de 1 ou 2 cm et sachant que les scores de Brier sont plus satisfaisants pour de fortes valeurs du seuil, nous poursuivons donc avec le seuil à 2 cm.

4.2.1.2. *Scores de Brier avec les données des capteurs GMON*

Suite aux calculs effectués avec les données des capteurs SR50, les scores de Brier sont calculés pour les trois modèles en utilisant les réalisations neige/non-neige des capteurs GMON selon les seuils de neige variant de 1 à 5 cm.

Les capteurs GMON qui sont disponibles pour l'année 2011 sont énumérés au tableau 4.2, où les numéros d'identification et les noms des capteurs sont fournis par l'Institut de recherche d'Hydro-Québec. Comme mentionné à la section précédente, certains capteurs se situent dans un même point de grille (même bloc dans le tableau). Si nous voulons n'en sélectionner qu'un seul par point de grille, cela représente un total de 384 combinaisons possibles et effectuer une validation pour toutes ces combinaisons serait très long et fastidieux. Par contre, en s'intéressant aux dates couvertes par chacun des capteurs, il est possible de constater qu'ils ne couvrent pas tous l'année entière. Il a donc été possible d'en regrouper quelques-uns, ce qui a permis de réduire le nombre de combinaisons possibles à 32 (voir tableau 4.3).

Étant donné que les paires de capteurs 242 et 243, 245 et 246 ainsi que 247 et 249 ne couvrent que la fin de l'année et donc pas la période la plus intéressante qui est la fonte, un seul capteur par paire est conservé. Les capteurs 242, 245 et 247 ont été choisis, ce qui diminue le total à quatre combinaisons possibles (voir tableau 4.4 pour la liste réduite de capteurs et tableau 4.5 pour le détail des quatre combinaisons).

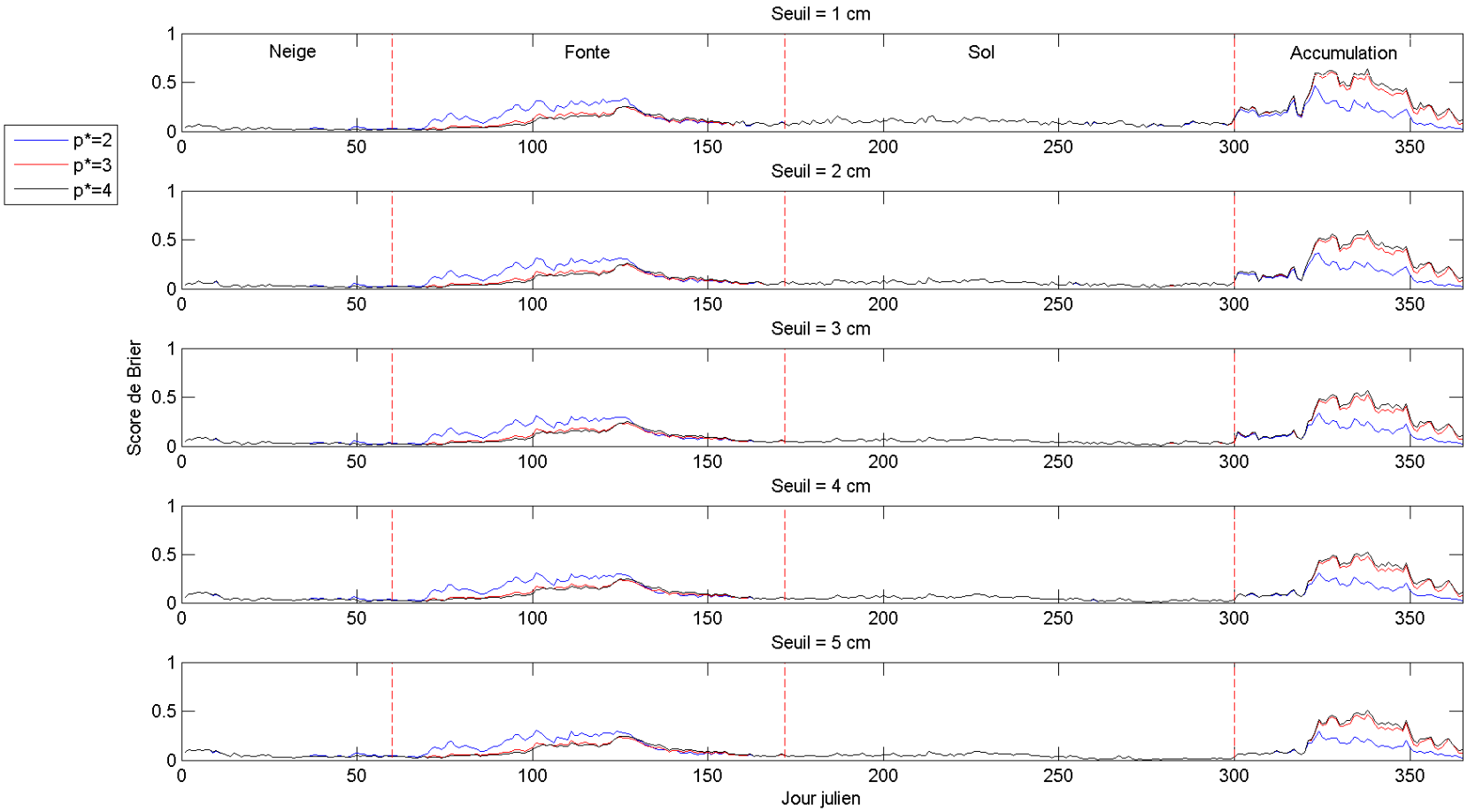

FIGURE 4.1. Scores de Brier pour les modèles à deux, trois et quatre variables explicatives calculés pour les réalités neige/non-neige des données des capteurs SR50 selon les seuils de 1 à 5 cm, représentation par seuil, Québec, 2011.

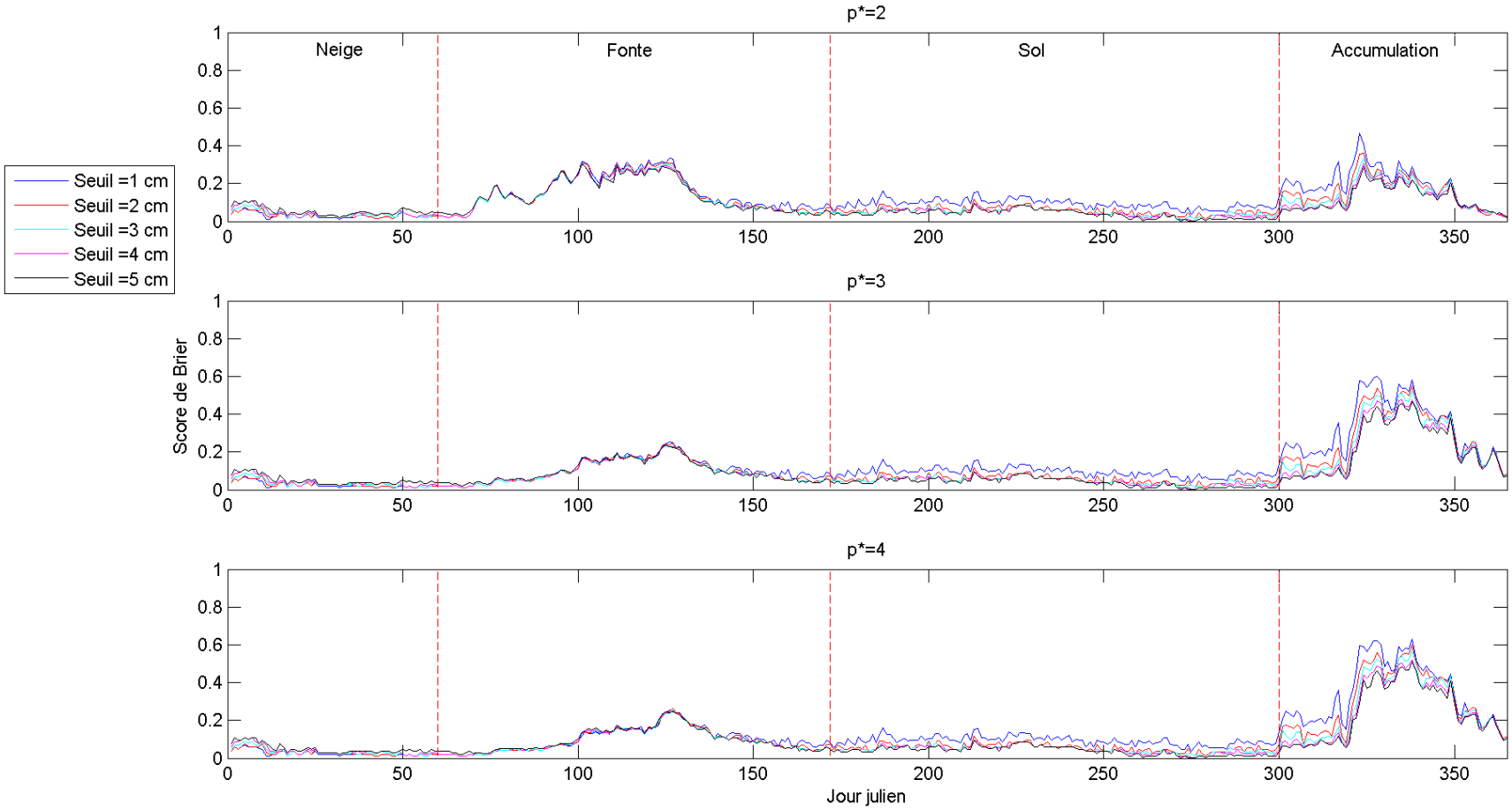

FIGURE 4.2. Scores de Brier pour les modèles à deux, trois et quatre variables explicatives calculés pour les réalités neige/non-neige des capteurs SR50 selon les seuils de 1 à 5 cm, représentation par modèle, Québec, 2011.

TABLE 4.2. Liste des capteurs GMON selon le numéro d'identification, leur nom et les dates couvertes pour l'année 2011.

Numéro d'identification	Nom du GMON	Dates couvertes
201	GOUINRES	30/11-31/12
202	GOUINBO	30/11-31/12
255	BARRAGEGOUIN	01/01-14/06
258	BARRAGEGOUINBOIS	01/01-14/06
203	COOPER	30/11-31/12
205	COOPERBO	30/11-31/12
251	CHEMINCOOPER	01/01-14/06
259	CHEMINCOOPERBOIS	01/01-14/06
204	RAPBLANC	30/11-31/12
253	RAPIDEBLANC	01/01-15/06
241	NECOPASTIC	01/01-31/12
242	BARRAGEMERCIERP8	13/09-31/12
243	BARRAGEMERCIERP4	14/09-31/12
244	MATAWAN	01/10-31/12
245	DEPOTVENNEP8	16/09-31/12
246	DEPOTVENNEP4	15/09-31/12
247	MATAWINTOUR	30/08-31/12
249	MATAWINLN	31/08-31/12
252	MATAWIN	01/01-15/06
248	FLAMAND	01/01-31/12
250	SHAW	01/01-31/12

Les résultats de la figure 4.3 indiquent des conclusions semblables à celles émises pour les données des capteurs SR50, c'est-à-dire que les cartographies sont satisfaisantes lors des périodes de neige et de sol, mais qu'elles le sont moins durant les périodes de fonte et d'accumulation de neige. Lors de la période de fonte, les scores de Brier sont majoritairement inférieurs à 0,5, ce qui semble indiquer un problème d'incertitude et non un problème de cohérence avec la réalité. Cependant, les modèles à trois et quatre variables explicatives ont des valeurs du score plus près de 1 lors de l'accumulation de neige, nous informant qu'ils sont moins performants pour cette période.

En ce qui concerne la comparaison des différents modèles entre eux, il est possible de remarquer que ceux à trois et quatre variables explicatives sont encore une fois assez équivalents. De plus, ils sont plus satisfaisants que le modèle à deux variables explicatives pour la période de fonte. Cependant, les conclusions diffèrent quelque peu de celles obtenues avec les capteurs SR50 pour la période d'accumulation. En effet, le modèle à deux variables explicatives est bel et bien le plus performant pour les seuils de 1 à 3 cm, mais il affiche de moins bons résultats pour les seuils de 4 et 5 cm.

TABLE 4.3. Liste des capteurs GMON jumelés et les dates couvertes pour l'année 2011.

Numéro d'identification	Dates couvertes
201+255	01/01-14/06 et 30/11-31/12
202+258	01/01-14/06 et 30/11-31/12
203 + 251	01/01-14/06 et 30/11-31/12
205+259	01/01-14/06 et 30/11-31/12
204+253	01/01-15/06 et 30/11-31/12
241	01/01-31/12
242	13/09-31/12
243	14/09-31/12
244	01/10-31/12
245	16/09-31/12
246	15/09-31/12
247+252	01/01-15/06 et 30/08-31/12
249+252	01/01-15/06 et 31/08-31/12
248	01/01-31/12
250	01/01-31/12

TABLE 4.4. Liste réduite des capteurs GMON jumelés et dates couvertes pour l'année 2011.

Numéro d'identification	Dates couvertes
201+255	01/01-14/06 et 30/11-31/12
202+258	01/01-14/06 et 30/11-31/12
203 + 251	01/01-14/06 et 30/11-31/12
205+259	01/01-14/06 et 30/11-31/12
204+253	01/01-15/06 et 30/11-31/12
241	01/01-31/12
242	13/09-31/12
244	01/10-31/12
245	16/09-31/12
247+252	01/01-15/06 et 30/08-31/12
248	01/01-31/12
250	01/01-31/12

L'étude de la figure 4.4 permet de remarquer que les scores de Brier sont plus faibles lorsque la valeur du seuil augmente, sauf pour la fin de l'année pour le modèle à deux variables explicatives. De plus, l'écart entre les scores des différents seuils est considérable lors de la période d'accumulation. Cela signifie que le choix du seuil pour séparer les données GMON en neige/non-neige a une influence importante sur l'évaluation de la qualité des cartographies produites par chacun des modèles. Les conclusions étant toutefois semblables durant la période de fonte,

TABLE 4.5. Composition des différentes combinaisons des capteurs GMON disponibles pour l'année 2011.

Combinaison	Capteurs			
1	201+255	203+251	204+253	241
	242	244	245	247+252
2	201+255	205+259	204+253	241
	242	244	245	247+252
3	202+258	203+251	204+253	241
	242	244	245	247+252
4	202+258	205+259	204+253	241
	242	244	245	247+252

nous poursuivons la validation avec le seuil de 1 cm recommandé par les chercheurs de l'Institut de recherche d'Hydro-Québec.

Les conclusions obtenues étant semblables, peu importe la combinaison de capteurs utilisée (voir annexe F.1), nous ne considérons pour la suite que la combinaison 1.

4.2.1.3. *Scores de Brier avec les données SSM/I*

Pour conclure la validation de l'année 2011 à l'aide du score de Brier, ce dernier est calculé pour chacun des modèles par rapport aux données SSM/I.

Lorsque la figure 4.5 est observée, il est possible de remarquer que les valeurs sont assez proches de 0 pour la période de sol, mais qu'elles sont plus élevées pour le reste de l'année. Puisqu'elles sont majoritairement entre 0 et 0,3, cela semble indiquer plus certainement un problème d'incertitude qu'un problème de cohérence avec la réalité. Pour la comparaison des différents modèles, les conclusions sont assez semblables à celles qui avait été observées avec les capteurs SR50, c'est-à-dire que :

- les trois modèles sont assez équivalents lors des périodes de neige et de sol ;
- ceux à trois et quatre variables explicatives sont meilleurs lors de la fonte ;
- celui à deux variables explicatives est plus performant lors de l'accumulation de neige.

4.2.2. Validation à l'aide des proportions de concordance

Afin de continuer la validation des cartographies produites par les modèles à deux, trois et quatre variables explicatives, les proportions de concordance (voir équation (4.1.2)) entre les réalisations neige/non-neige produites par les trois différents modèles selon différents points de césure et les données de référence sont

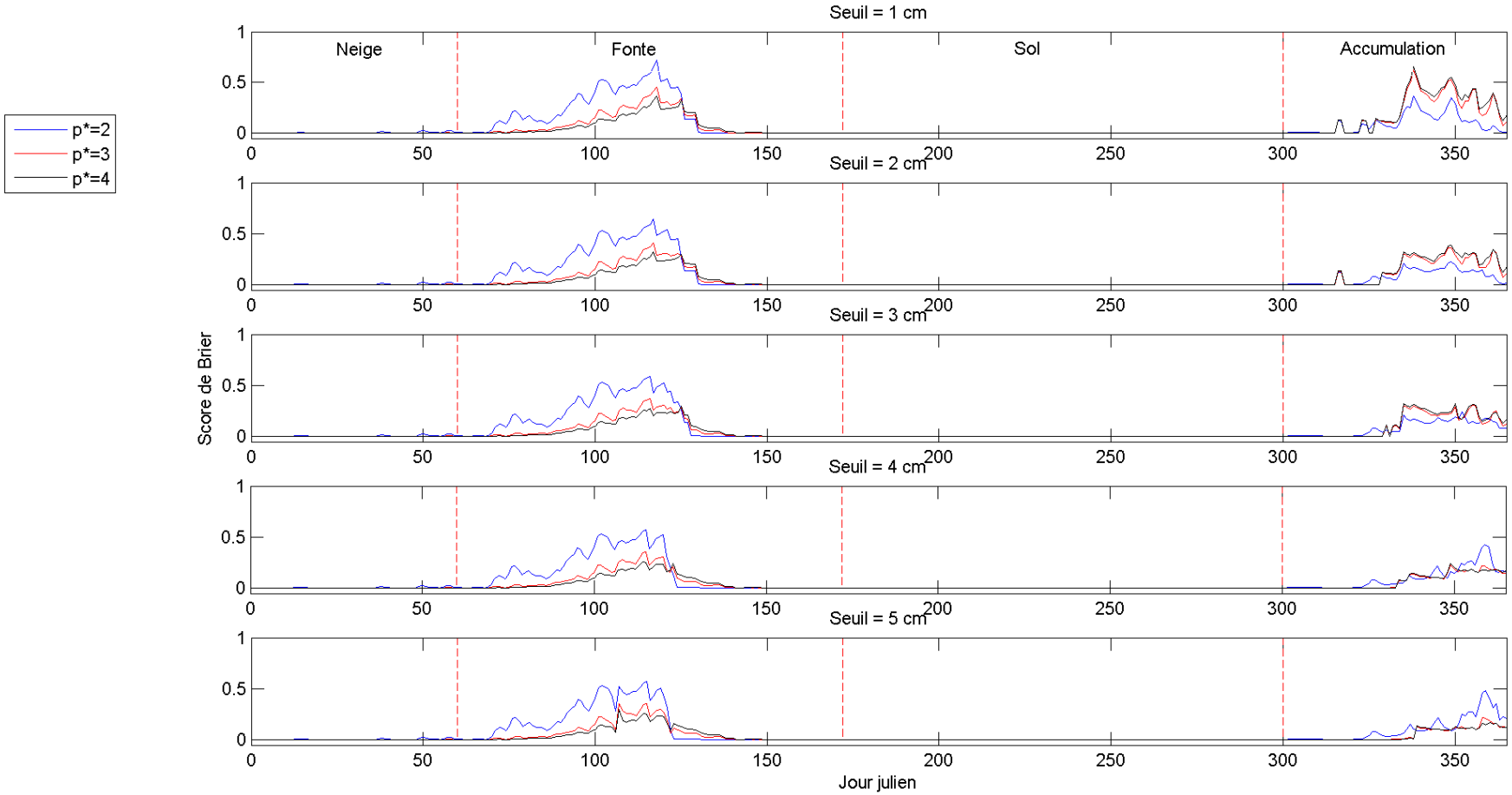

FIGURE 4.3. Scores de Brier pour les modèles à deux, trois et quatre variables explicatives calculés pour les réalités neige/non-neige de la première combinaison de capteurs GMON selon les seuils de 1 à 5 cm, représentation par seuil, Québec, 2011.

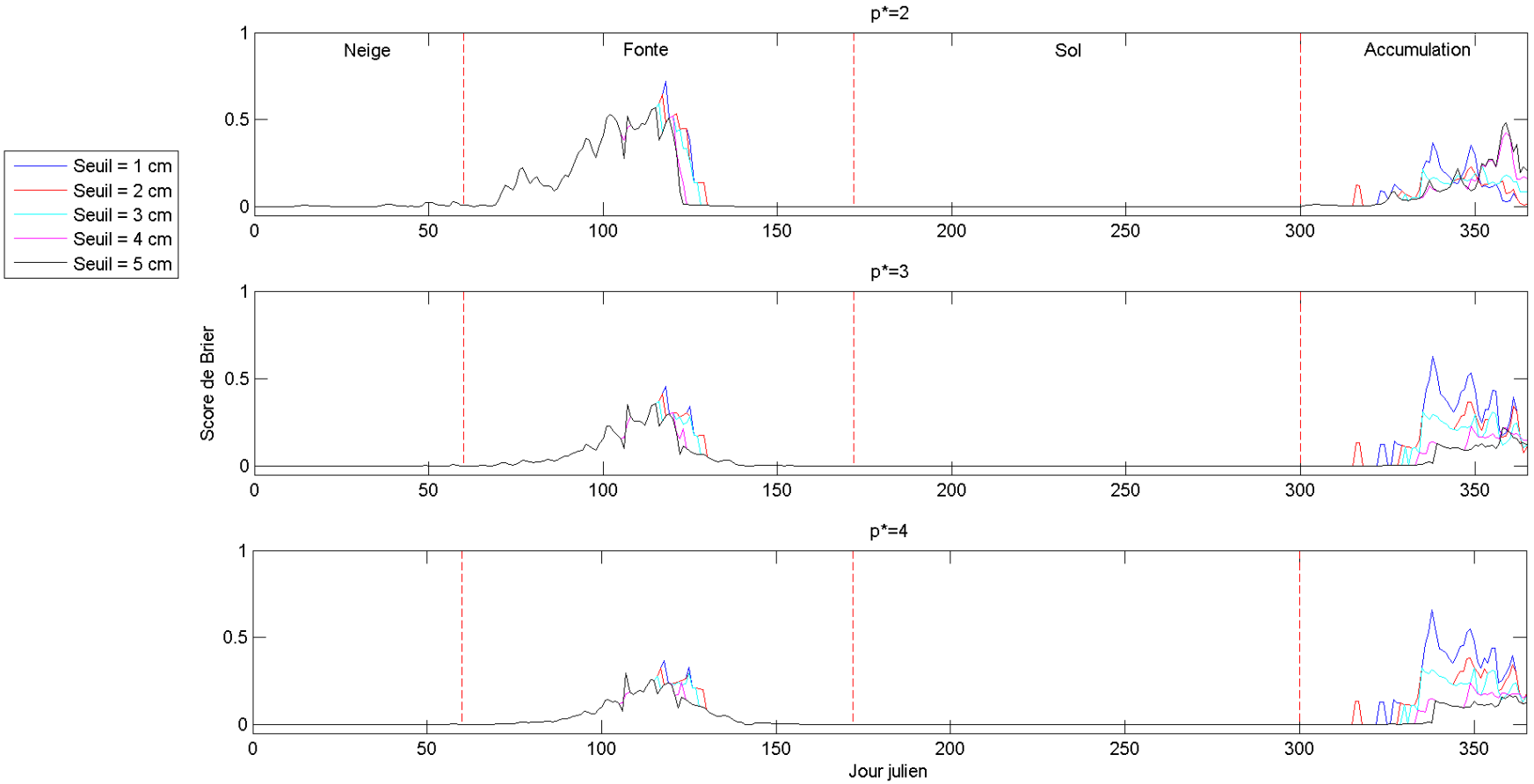

FIGURE 4.4. Scores de Brier pour les modèles à deux, trois et quatre variables explicatives calculés pour les réalités neige/non-neige de la première combinaison de capteurs GMON selon les seuils de 1 à 5 cm, représentation par modèle, Québec, 2011.

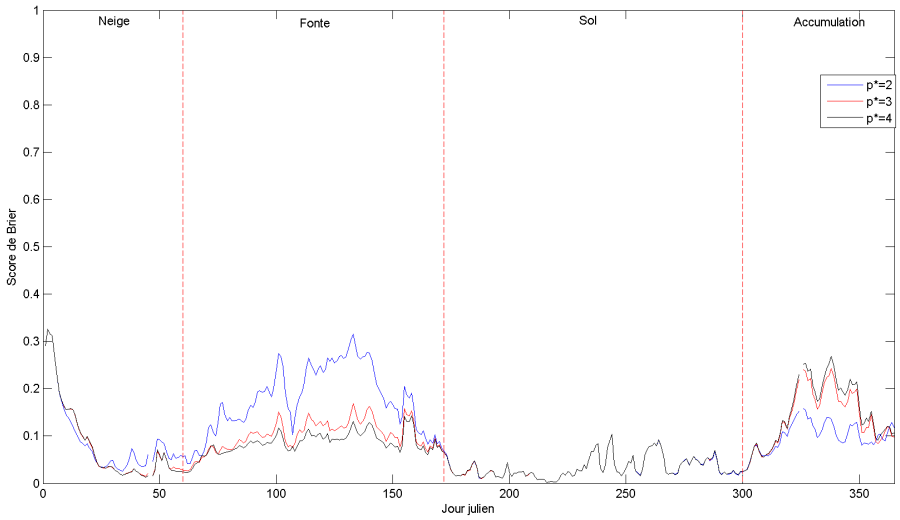

FIGURE 4.5. Scores de Brier pour les modèles à deux, trois et quatre variables explicatives calculés pour les réalités neige/non-neige des données SSM/I, Québec, 2011.

calculées. Suite aux décisions des sections précédentes, les réalisations neige/non-neige pour les capteurs SR50 sont calculées avec le seuil de 2 cm et celles pour les capteurs GMON sont calculées avec le seuil à 1 cm pour la combinaison 1. Les cartes SSM/I représentant déjà des réalisations neige/non-neige, il n'a pas été nécessaire de choisir un seuil.

4.2.2.1. *Proportions de concordance avec les données des capteurs SR50*

Tout d'abord, les proportions de concordance pour chacun des trois modèles ont été calculées avec les points de césure 0,1 à 0,9 et les réalisations neige/non-neige des capteurs SR50 selon le seuil de 2 cm.

L'examen de la figure 4.6 montre que, peu importe le modèle, la proportion de concordance vaut 1 (ou proche de 1) lors de la période de neige et de sol. Cela signifie que, peu importe le point de césure utilisé, les réalisations neige/non-neige des modèles correspondent presque parfaitement à celles des capteurs SR50 en utilisant le seuil de 2 cm. Ce résultat est cohérent avec ceux obtenus à l'aide du score de Brier (voir figures 4.1 et 4.2). De plus, cette proportion diminue pendant la fonte et l'accumulation de neige selon le point de césure utilisé, ce qui est cohérent avec les valeurs des scores de Brier pour les mêmes périodes.

En observant attentivement les résultats pour chacun des points de césure, il est possible de constater que les proportions de concordance sont généralement

plus stables ou plus élevées pour les points de césure 0,3, 0,4 et 0,5 (voir figure 4.7). Pour la comparaison des modèles, la figure 4.8 affiche que les modèles à trois et quatre variables explicatives ont généralement les plus fortes proportions de concordance à la fonte, alors que celui à deux variables explicatives est généralement celui avec les plus fortes proportions de concordance lors de l'accumulation de neige. De plus, pour les périodes de neige et de sol, les trois modèles sont équivalents. En effet, les proportions de concordance y sont toutes très près de 1.

4.2.2.2. *Proportions de concordance avec les données des capteurs GMON*

Comme pour les données des capteurs SR50, les proportions de concordance sont calculées pour chacun des trois modèles selon les points de césure 0,1 à 0,9 avec les réalisations neige/non-neige des capteurs GMON obtenu avec un seuil de 1 cm.

L'observation de la figure 4.9 permet d'affirmer que les conclusions obtenues sont assez semblables à celles émises en utilisant les données des capteurs SR50. En effet, les proportions de concordance sont très fortes pendant les périodes de neige et de sol, mais elles diminuent considérablement pendant la fonte et l'accumulation de neige. Ces comportements sont cohérents avec ceux des scores de Brier pour les mêmes périodes (voir figures 4.3 et 4.4). De plus, cette diminution des proportions de concordance est également modérée par le choix du point de césure. Ici, une étude de chacun de ces points permet de déterminer que les proportions sont encore une fois généralement plus stables ou plus élevées pour les points 0,3, 0,4 et 0,5 (voir figure 4.10).

En examinant la figure 4.11, il est possible d'observer, encore ici, que les modèles sont tous équivalents (proportions près de 1) lors de périodes de neige et de sol. Cependant, cette fois-ci, les modèles à trois et quatre variables explicatives ont généralement des proportions de concordance plus élevées que celles du modèle à deux variables explicatives autant lors de la fonte que lors de l'accumulation de neige.

4.2.2.3. *Proportions de concordance avec les données SSM/I*

Les proportions de concordance sont finalement calculées pour chacun des trois modèles pour les points de césure 0,1 à 0,9 avec les données SSM/I.

La figure 4.12, montre que le comportement général par période est semblable à celui qui avait été constaté pour les données des capteurs SR50 et GMON. De plus, ces résultats sont cohérents avec les comportements des scores de Brier (voir figure 4.5). Aussi, les points de césures les plus satisfaisants sont encore une fois

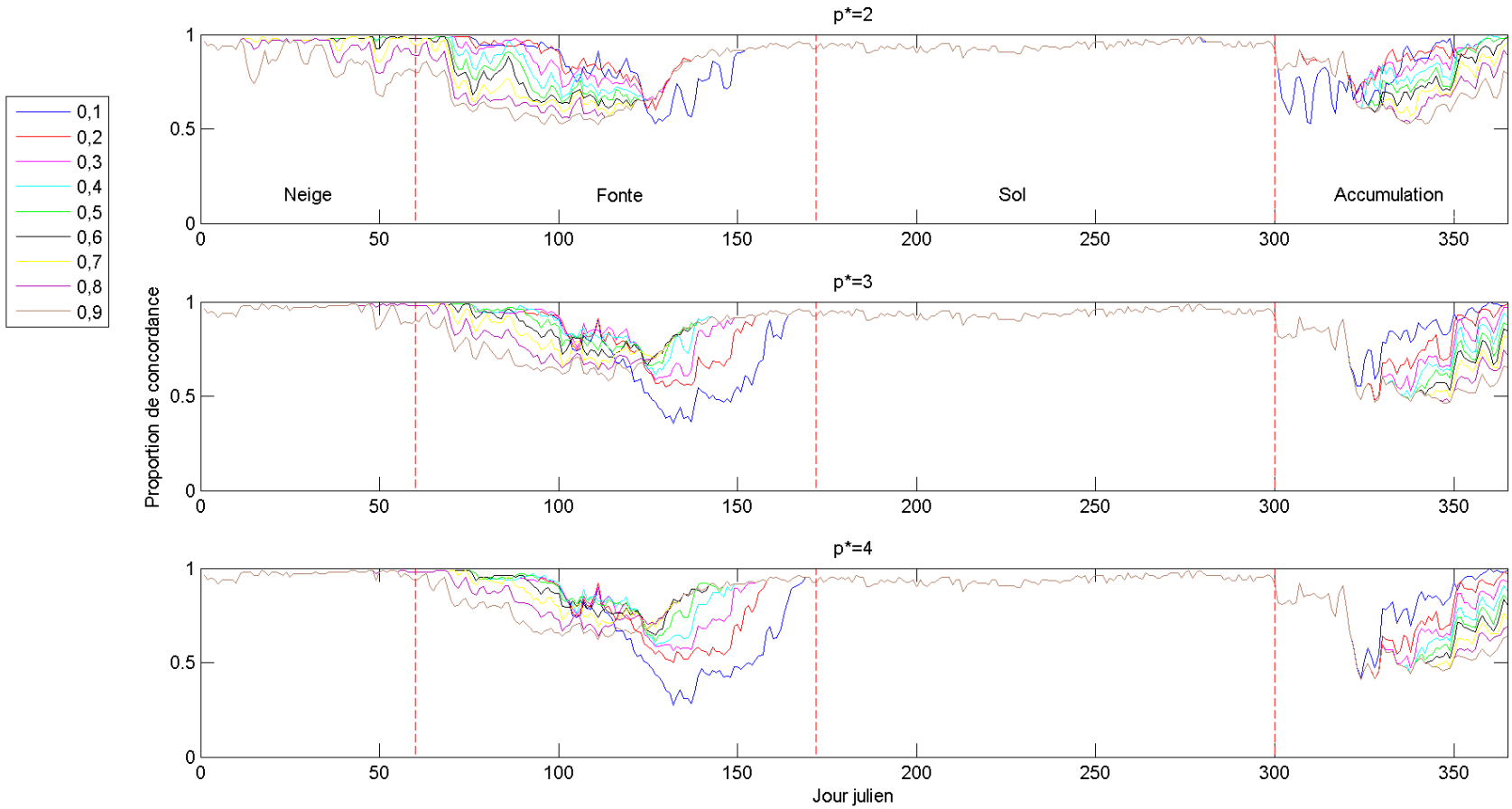

FIGURE 4.6. Proportions de concordance entre les réalités neige/non-neige des modèles à deux, trois et quatre variables explicatives selon les points de césure 0,1 à 0,9 calculés avec les réalités neige/non-neige des données des capteurs SR50 selon le seuil de 2 cm, représentation par modèle, Québec, 2011.

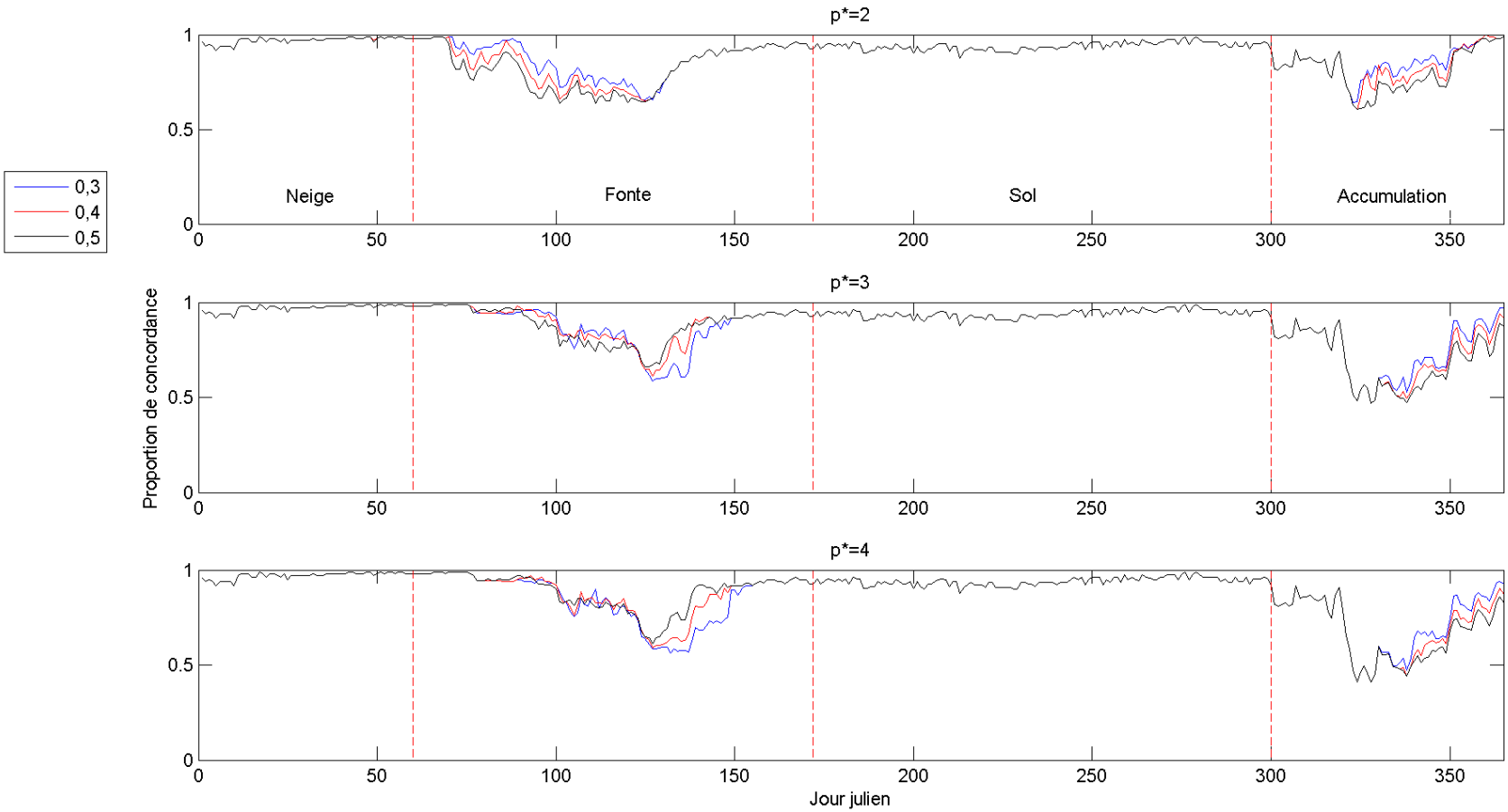

FIGURE 4.7. Proportions de concordance entre les réalités neige/non-neige des modèles à deux, trois et quatre variables explicatives selon les points de césure 0,3, 0,4 et 0,5 calculés avec les réalités neige/non-neige des données des capteurs SR50 selon le seuil de 2 cm, représentation par modèle, Québec, 2011.

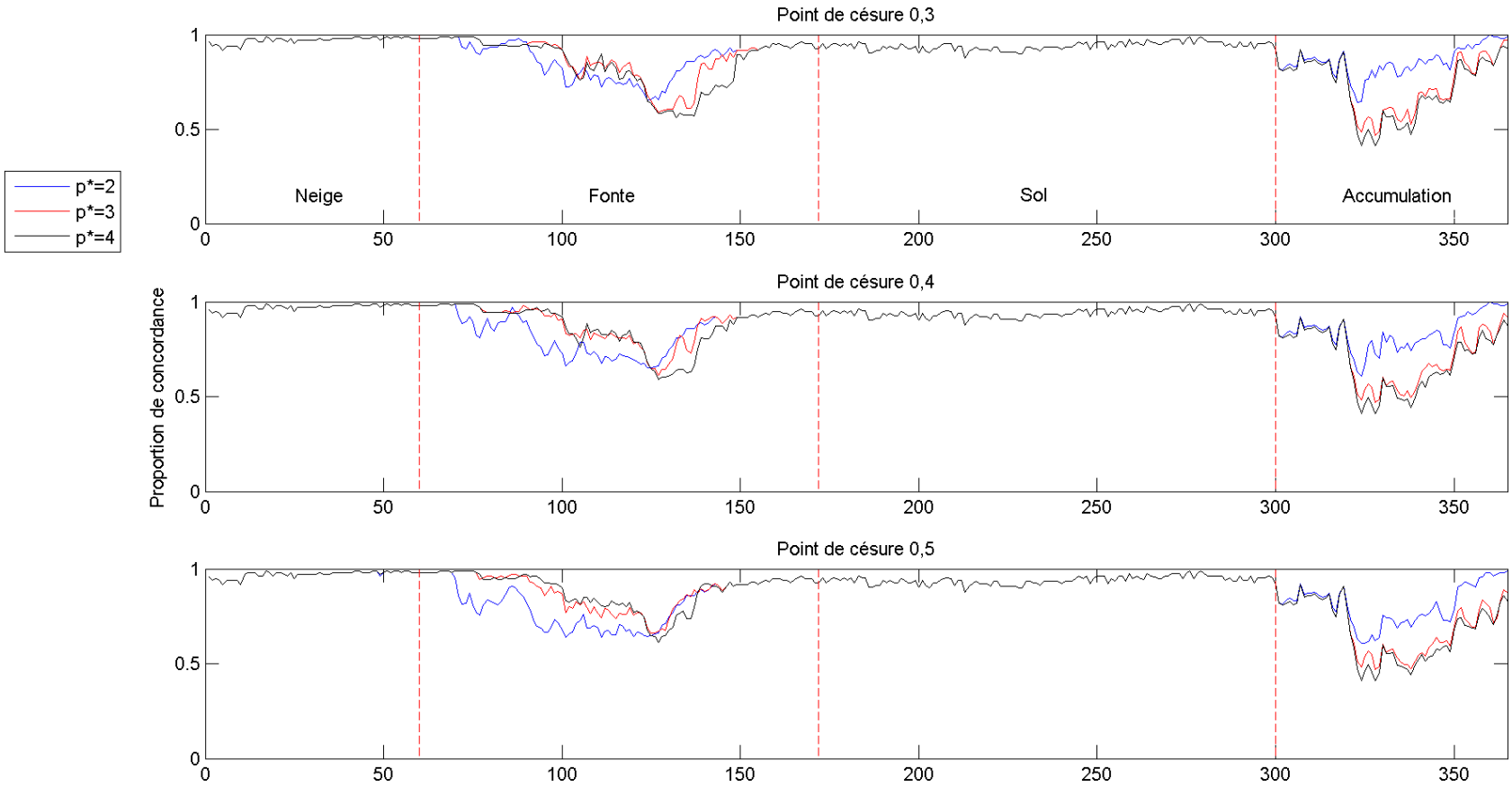

FIGURE 4.8. Proportions de concordance entre les réalités neige/non-neige des modèles à deux, trois et quatre variables explicatives selon les points de césure 0,3, 0,4 et 0,5 calculés avec les réalités neige/non-neige des données des capteurs SR50 selon le seuil de 2 cm, représentation par point de césure, Québec, 2011.

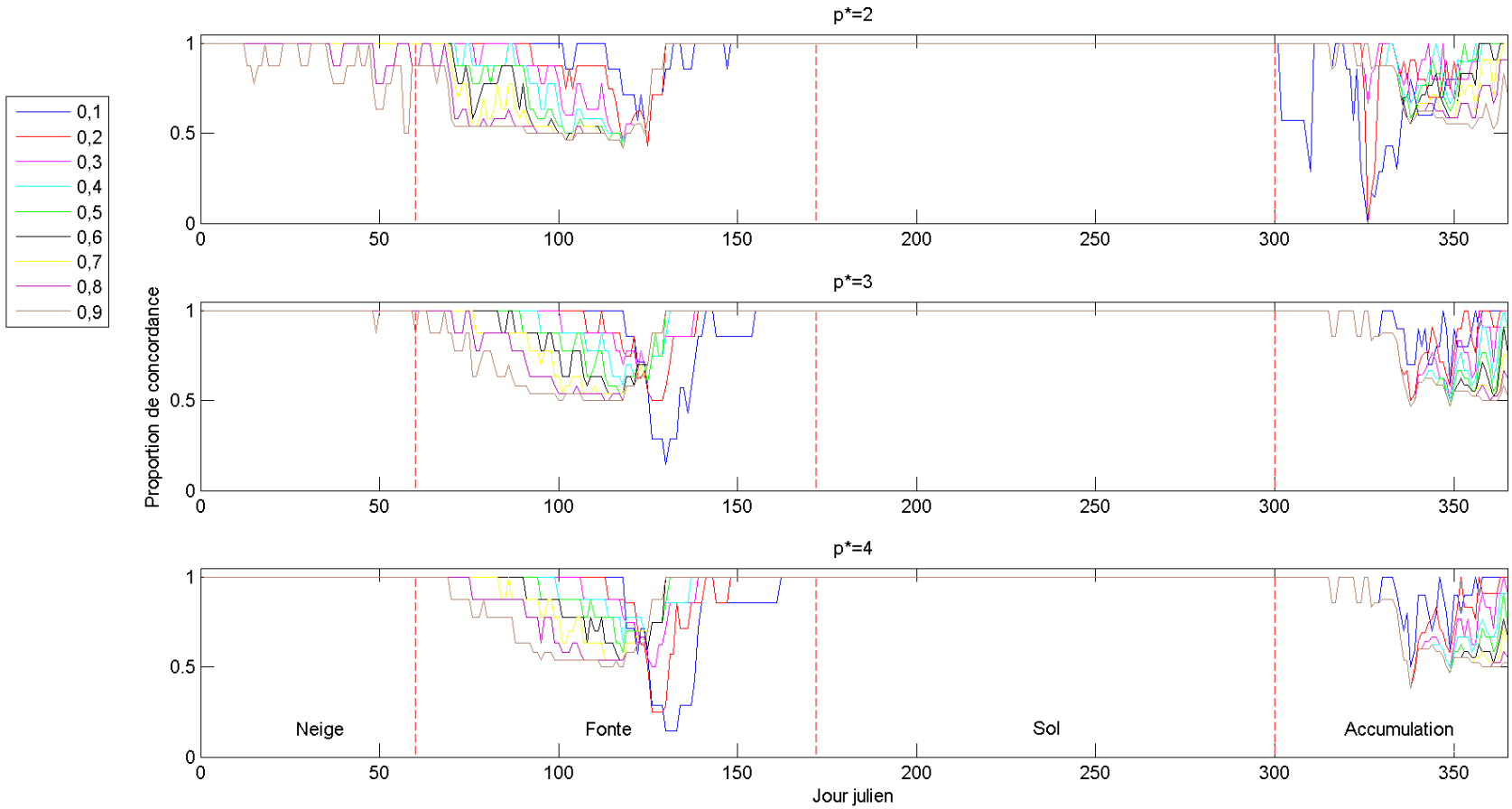

FIGURE 4.9. Proportions de concordance entre les réalités neige/non-neige des modèles à deux, trois et quatre variables explicatives selon les points de césure 0,1 à 0,9 calculés avec les réalités neige/non-neige de la première combinaison de données des capteurs GMON selon le seuil de 1 cm, représentation par modèle, Québec, 2011.

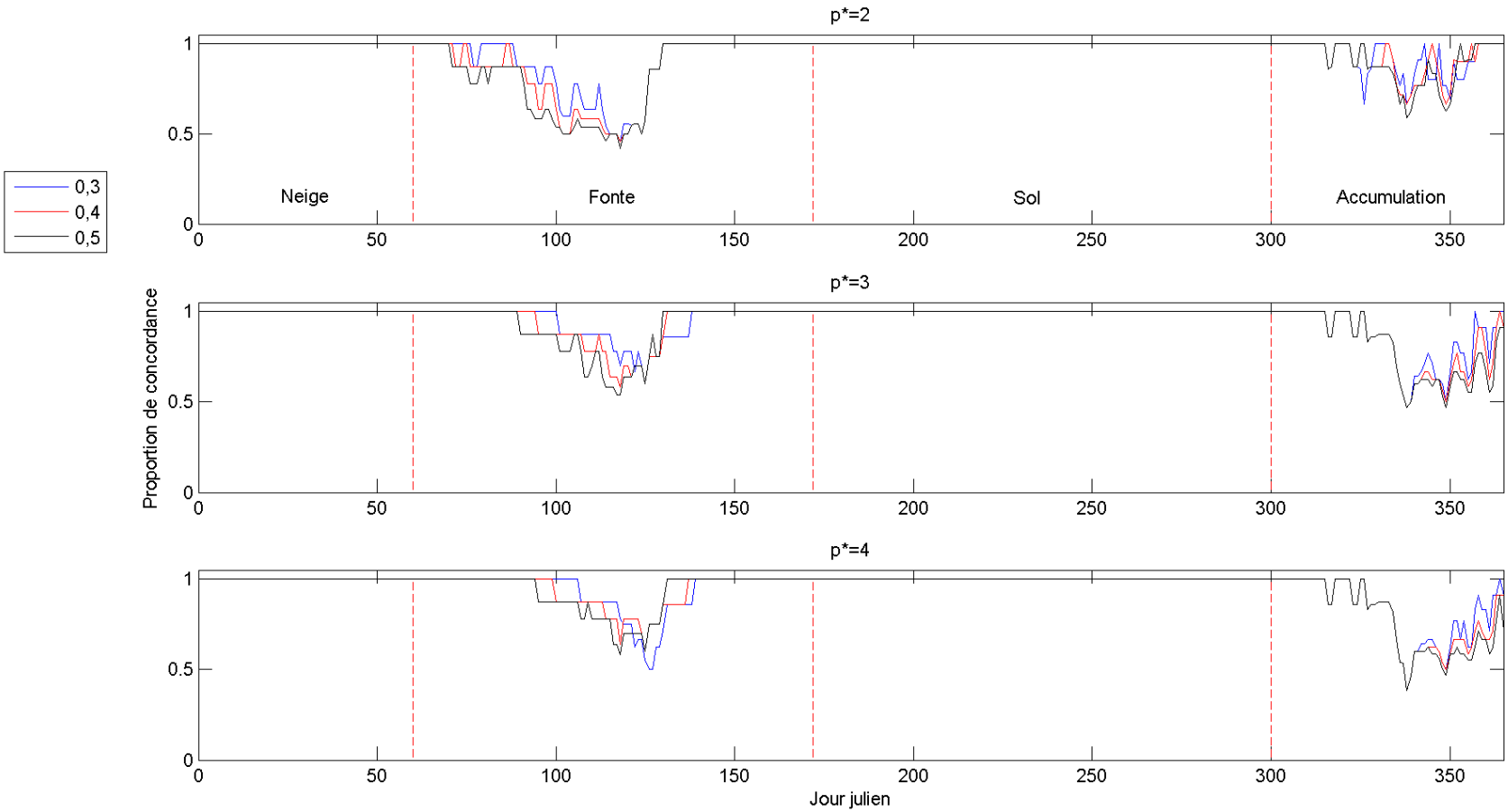

FIGURE 4.10. Proportions de concordance entre les réalités neige/non-neige des modèles à deux, trois et quatre variables explicatives selon les points de césure 0,3, 0,4 et 0,5 calculés avec les réalités neige/non-neige de la première combinaison de données des capteurs GMON selon le seuil de 1 cm, représentation par modèle, Québec, 2011.

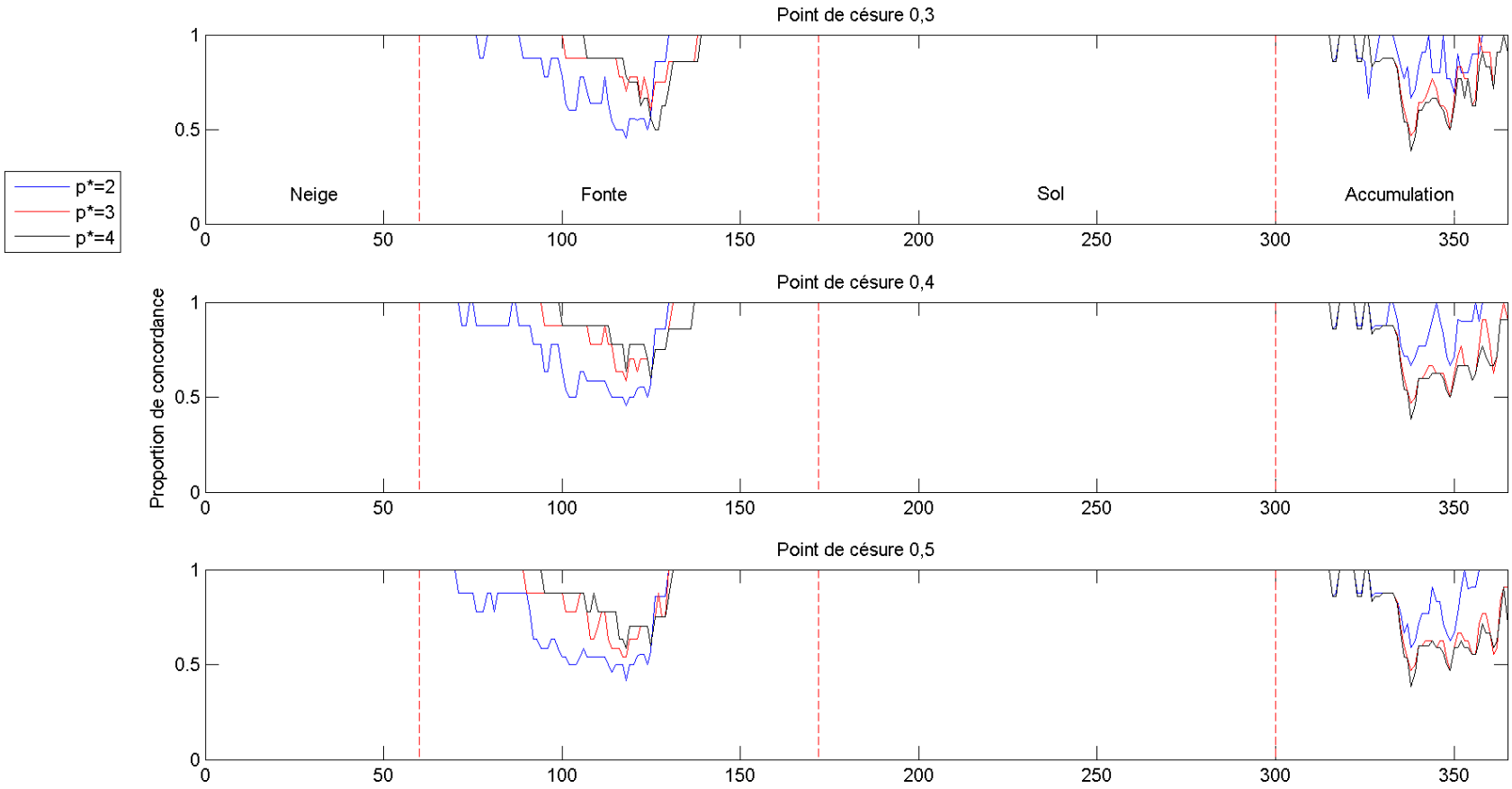

FIGURE 4.11. Proportions de concordance entre les réalités neige/non-neige des modèles à deux, trois et quatre variables explicatives selon les points de césure 0,3, 0,4 et 0,5 calculés avec les réalités neige/non-neige de la première combinaison de données des capteurs GMON selon le seuil de 1 cm, représentation par point de césure, Québec, 2011.

0,3, 0,4 et 0,5 (voir figure 4.13), même si 0,2 semble aussi être un choix judicieux.

4.2.3. Validation à l'aide du critère d'erreurs d'omission/commission

Le dernier outil utilisé pour vérifier la qualité des résultats est le critère d'erreurs d'omission/commission. Ayant conclu à la section précédente que les points de césure permettant d'obtenir les résultats les plus satisfaisants sont 0,3, 0,4 et 0,5, ceux-ci sont conservés pour le calcul des réalités neige/non-neige des modèles à deux, trois et quatre variables explicatives. Quant aux données des capteurs SR50 et GMON, les seuils de 2 cm et 1 cm respectivement sont encore considérés.

Dans le cadre de ce projet, il est important de comprendre que les différents types d'erreur (omission et commission) ont une interprétation différente selon la période où elles se produisent. Lors de la fonte, une erreur d'omission signifie que le modèle fait fondre la neige trop vite alors qu'une erreur de commission indique que le modèle ne fait pas fondre la neige assez rapidement. Au contraire, pendant l'accumulation de neige, une erreur d'omission signifie que le modèle ne saisit pas assez tôt l'apparition de la neige, tandis qu'une erreur de commission indique que le modèle saisit l'apparition de la neige de façon précoce. Il n'est toutefois pas possible de déterminer quelle erreur est la moins dommageable. Par exemple, pour la fonte, une erreur d'omission est problématique. En effet, le modèle ne prévoit alors plus de neige, ce qui signifie, à tort, plus aucun apport en eau dû à la neige. Cet apport en eau est donc inutilisé et cela peut avoir des répercussions au niveau économique (perte de production, donc perte de ventes) et au niveau de la sécurité publique (par exemple, des inondations). De plus, une erreur de commission est dommageable, car le modèle ne saisit pas que la neige fond. Il y a donc des apports en eau, alors que ce n'était pas anticipé. Les équipements ne sont pas nécessairement préparés pour gérer une telle situation.

Pour le présent mémoire, le critère de l'omission/commission est appliqué dans une optique temporelle. Pour ce faire, chaque point de grille est suivi en comparant les réalités neige/non-neige quotidiennes produites par chacun des modèles avec une des références (SR50 ou GMON). Des vecteurs contenant soit des 0, 1, ou -1 sont ainsi obtenus, ce qui permet d'identifier quelles sont les périodes de l'année où les réalisations neige/non-neige sont satisfaisantes et les erreurs commises lorsque les réalisations prévues par un modèle sont erronées. Afin de résumer l'information, la moyenne de la valeur absolue des résultats est calculée pour la valeur à chacun des points de grille. De cette façon, il est possible d'identifier, à l'échelle globale du domaine d'étude, quelles sont les périodes où il y a le plus d'erreurs. L'usage de la valeur absolue a pour but d'éviter que

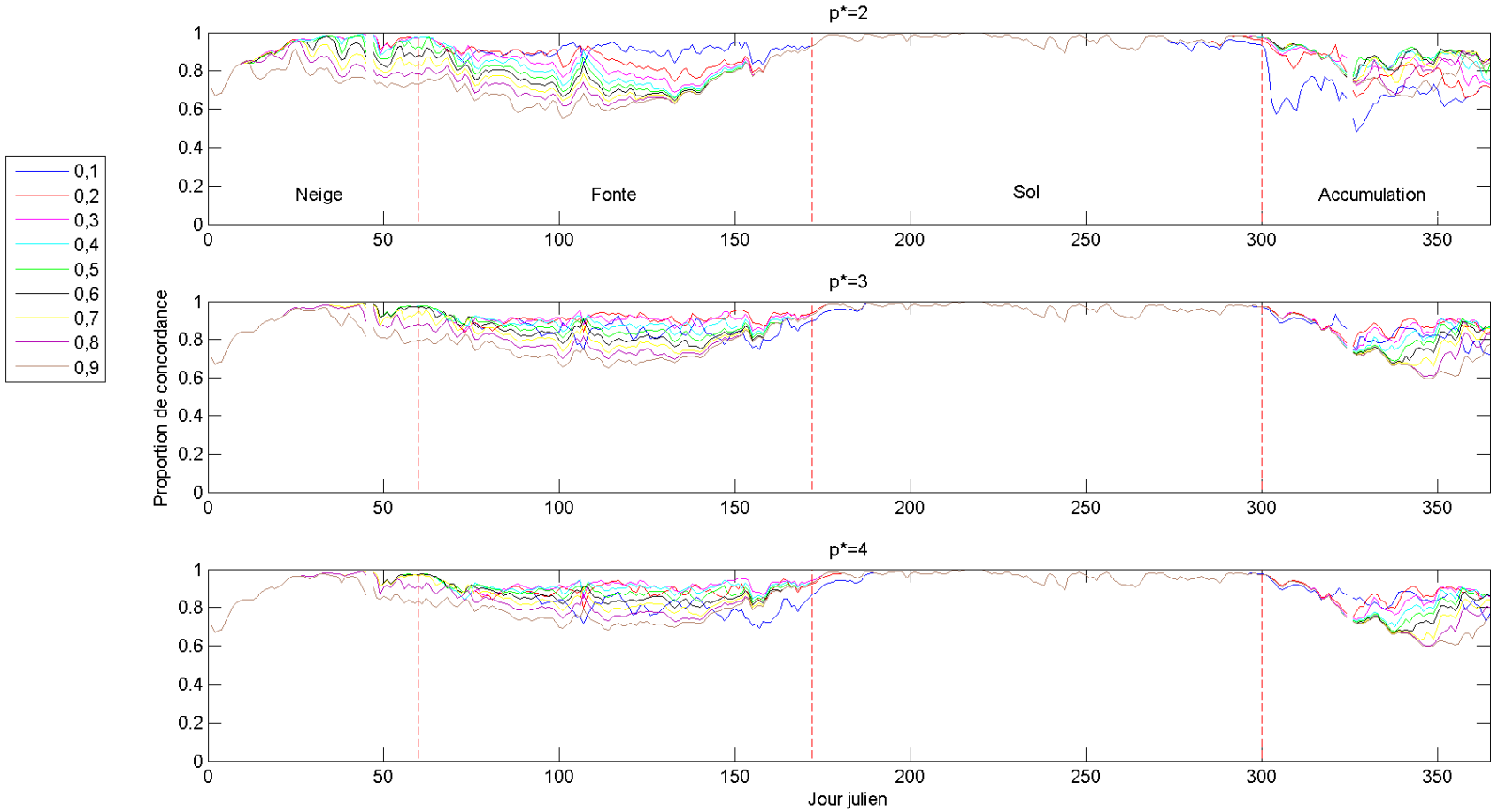

FIGURE 4.12. Proportions de concordance entre les réalités neige/non-neige des modèles à deux, trois et quatre variables explicatives selon les points de césure 0,1 à 0,9 calculés avec les réalités neige/non-neige des données SSM/I, représentation par modèle, Québec, 2011.

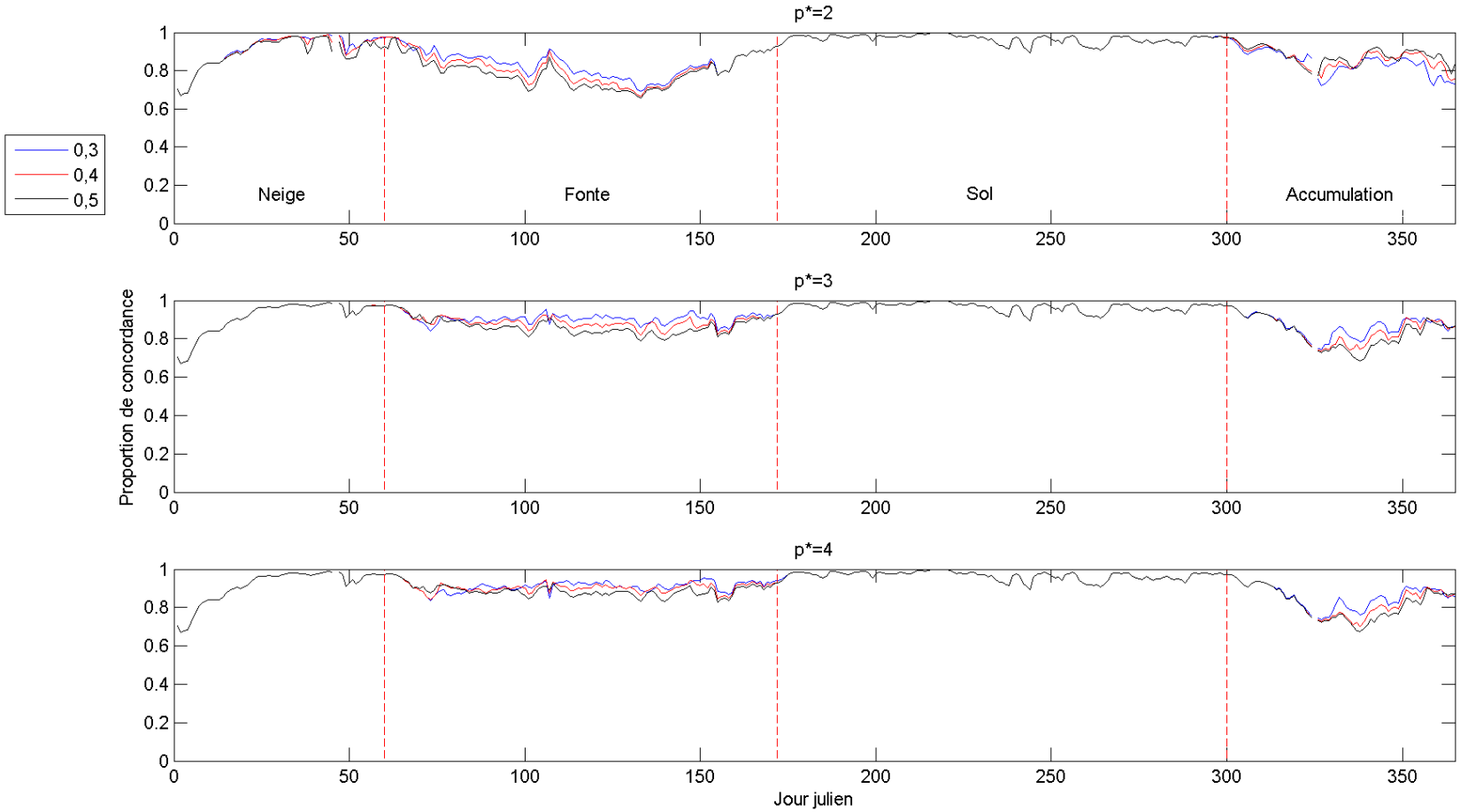

FIGURE 4.13. Proportions de concordance entre les réalités neige/non-neige des modèles à deux, trois et quatre variables explicatives selon les points de césure 0,3, 0,4 et 0,5 calculés avec les réalités neige/non-neige des données SSM/I, représentation par modèle, Québec, 2011.

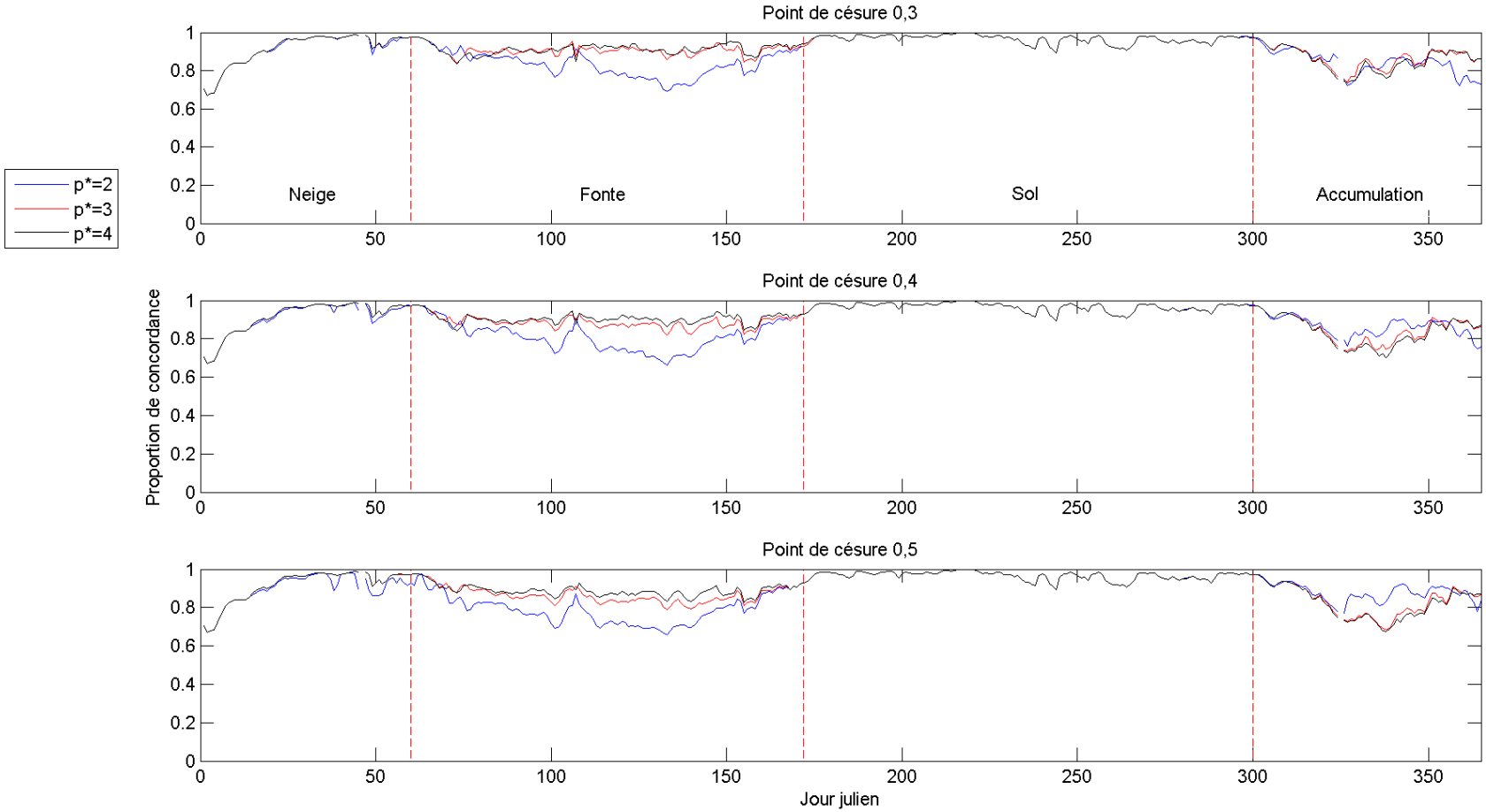

FIGURE 4.14. Proportions de concordance entre les réalités neige/non-neige des modèles à deux, trois et quatre variables explicatives selon les points de césure 0,3, 0,4 et 0,5 calculés avec les réalités neige/non-neige des données SSM/I, représentation par point de césure, Québec, 2011.

les deux types d'erreur s'annulent lorsqu'ils sont en égale quantité. En effet, s'il y a autant d'erreurs d'omission (-1) que de commission (1), la moyenne vaut 0 et cela indiquerait à tort qu'il n'y a pas d'erreur. Toutefois, cela ne permet pas d'identifier le type d'erreur commise de façon individuelle. La moyenne des résultats sans valeur absolue est donc aussi calculée :

- si la moyenne est négative, il est possible de constater qu'il y a des erreurs et qu'il s'agit majoritairement d'omission ;
- si la moyenne est positive, il s'agit majoritairement d'erreurs de commission ;
- si la moyenne vaut 0 et que la moyenne des valeurs absolues vaut également 0, alors il n'y a pas d'erreur ;
- si la moyenne vaut 0, mais que la moyenne des valeurs absolues est supérieure à 0, alors il y a des erreurs et qu'elles sont autant d'omission que de commission.

Le critère de l'omission/commission pourrait également être utilisé dans une optique spatiale, car l'information est disponible pour chacun des points de grille. Il serait donc possible d'observer comment se comportent les modèles selon différents groupes de capteurs situés à différents endroits dans le Québec. Ce type de validation n'est toutefois pas effectué dans le présent mémoire, car le choix des groupes à évaluer dépend du lecteur. En effet, il pourrait éventuellement vouloir appliquer les modèles développés au chapitre 3 sur seulement certaines parties du Québec. Dans ce cas, il observerait uniquement les capteurs pour ces endroits précis. Aussi, ce lecteur pourrait vouloir vérifier les résultats pour uniquement quelques points de grille particuliers qui sont liés directement à sa problématique. Nous nous contenterons ici de couvrir tous les capteurs GMON et un sous-ensemble de capteurs SR50 qui sont répartis dans le Québec en entier.

4.2.3.1. *Omission/commission avec les données des capteurs SR50*

Étant donné qu'il y a près d'une centaine de capteurs SR50, il est fastidieux d'étudier le critère de l'omission/commission des points de grille en utilisant tous les capteurs. C'est pourquoi une sélection d'une quinzaine des capteurs pertinents et/ou représentatifs a été effectuée par certains employés d'Hydro-Québec et de son Institut de recherche pour lesquels l'évolution temporelle des réalisations neige/non-neige est étudiée.

L'examen des figures 4.15 à 4.20 permet d'observer que les réalisations neige/non-neige ne comportent généralement pas d'erreur pour les périodes de neige et de sol, sauf pour quelques cas d'omission. En effet, les valeurs du critère de l'omission/commission sont pratiquement toutes nulles lorsque chacun des points de grille est étudié individuellement et les moyennes (calculées avec et sans la

valeur absolue) sont proches de 0 lorsque l'information est résumée. Ces résultats sont cohérents avec les scores de Brier peu élevés et les proportions de concordance proche 1 durant ces périodes.

Pour les périodes de fonte et d'accumulation, il y a des erreurs et il semble y en avoir autant, peu importe le nombre de variables explicatives. Par contre, le type d'erreur varie selon le nombre de variables explicatives. Pour le modèle à deux variables explicatives il s'agit surtout d'erreurs d'omission, peu importe la période, car il y a plusieurs valeurs de -1 pour les graphiques par point de grille et une moyenne négative pour le résumé de l'information. Pour les modèles à trois et quatre variables explicatives, il s'agit plutôt d'erreurs d'omission pour la période d'accumulation. Pour la période de fonte, la moyenne des valeurs absolues est grande, ce qui signifie qu'il y a aussi des erreurs pour cette période. De plus, il y a autant des valeurs de -1 et 1 pour les graphiques par point de grille et la moyenne est proche de 0, ce qui signifie que les erreurs sont autant d'omission que de commission.

Des conclusions similaires sont obtenues en utilisant les points de césure 0,4 et 0,5 (voir figures à l'annexe F.2).

4.2.3.2. *Omission/commission avec les données des capteurs GMON*

Puisque le nombre de capteurs GMON est largement inférieur au nombre de capteurs SR50, il n'est pas nécessaire d'effectuer une sélection de ceux qui sont les plus pertinents ou les plus représentatifs. En effet, en tenant compte du fait que certains capteurs peuvent être pairés (voir tableau 4.3), il n'y a que quinze points de grille/capteurs pour lesquels il est possible d'étudier l'évolution temporelle des réalisations neige/non-neige.

Les figures 4.21 à 4.26 affichent qu'il n'y a pas d'erreur pour les périodes de neige et de sol, mais qu'il y en a durant les périodes de fonte et d'accumulation de neige. Pour la période de fonte, les moyennes s'approchent de 0 et il y a de moins en moins de valeurs de -1 et 1 sur les graphiques par point de grille lorsque le nombre de variables explicatives augmente. Cela signifie que le nombre d'erreurs diminue, c'est-à-dire qu'il y a moins de points de grille pour lesquels le modèle n'est pas en accord avec les capteurs GMON. Puisque la moyenne des erreurs est négative et qu'il y a davantage de valeurs -1 que 1 dans les graphiques par point de grille, cela signifie que les erreurs produites pour cette période sont majoritairement de l'omission. Par contre, la moyenne des résultats se rapproche de 0 plus rapidement que celle des valeurs absolues lorsque p^* augmente, ce qui indique que, même s'il y a davantage d'erreurs d'omission, le nombre d'erreurs de

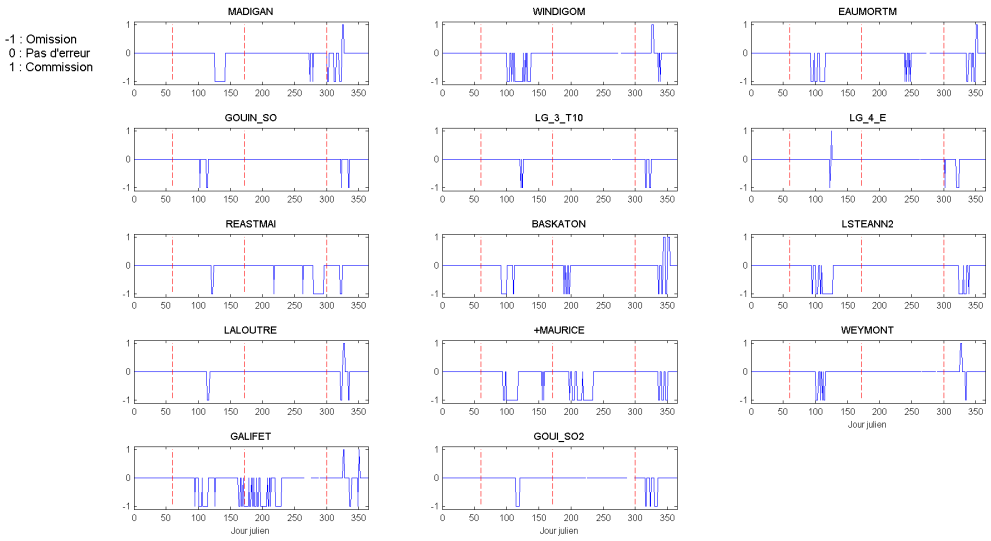

FIGURE 4.15. Omission/commission pour les réalisations neige/non-neige du modèle à deux variables explicatives avec le point de césure 0,3 avec les réalisations des capteurs SR50 pour le seuil de 2 cm, Québec, 2011.

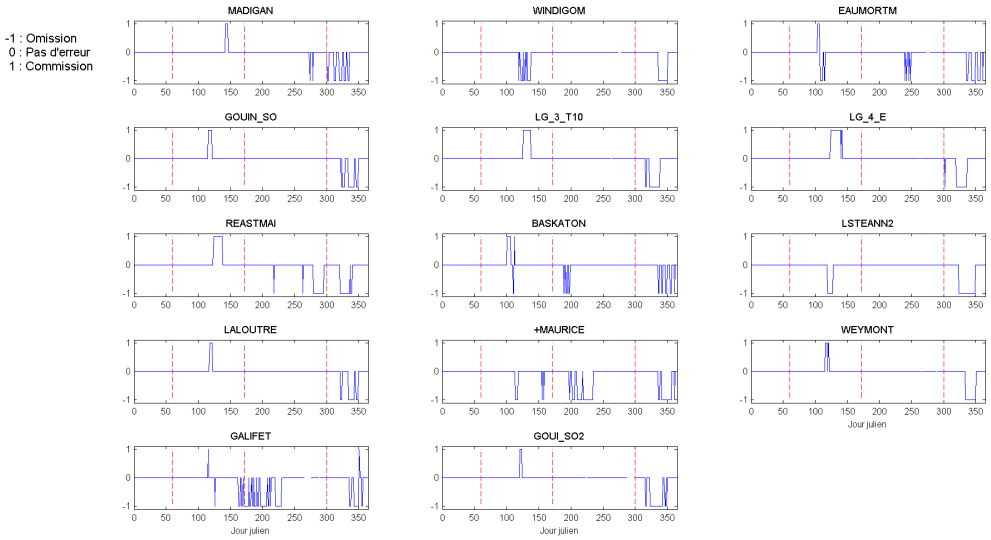

FIGURE 4.16. Omission/commission pour les réalisations neige/non-neige du modèle à trois variables explicatives avec le point de césure 0,3 avec les réalisations des capteurs SR50 pour le seuil de 2 cm, Québec, 2011.

chaque type semble vouloir s'équilibrer lorsque le nombre de variables explicatives augmente.

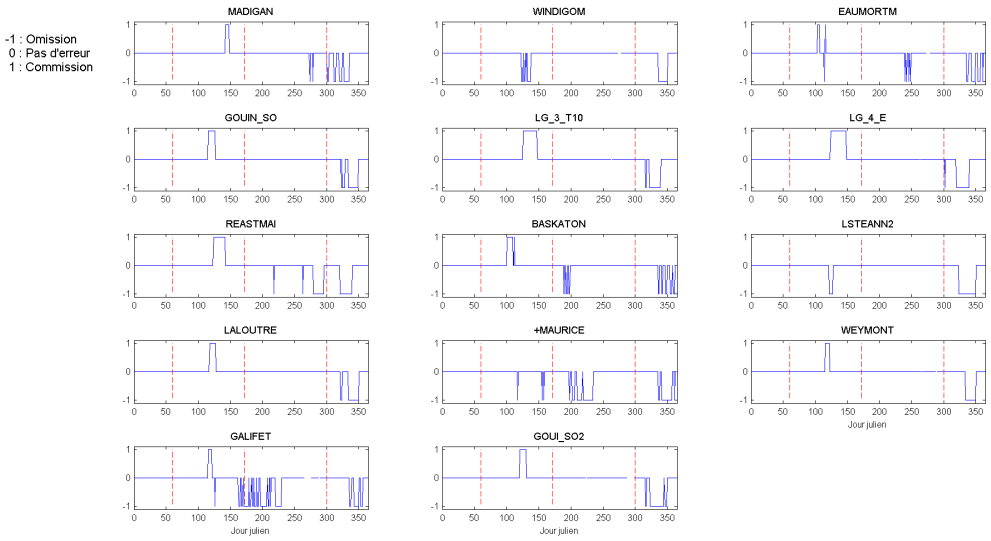

FIGURE 4.17. Omission/commission pour les réalisations neige/non-neige du modèle à quatre variables explicatives avec le point de césure 0,3 avec les réalisations des capteurs SR50 pour le seuil de 2 cm, Québec, 2011.

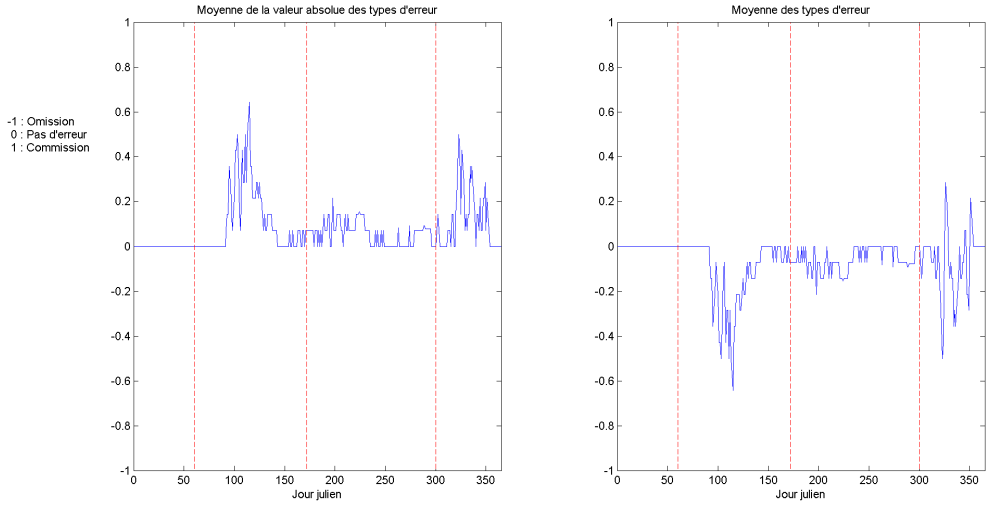

FIGURE 4.18. Moyenne sur les différents points de grille de l'omission/commission pour les réalisations neige/non-neige du modèle à deux variables explicatives selon le point de césure 0,3 avec les réalisations des capteurs SR50, seuil 2 cm, Québec, 2011.

Pour la période d'accumulation, la moyenne augmente avec le nombre de variables explicatives et il y a de plus en plus de valeurs de -1 et de 1 pour les graphiques par point de grille. Il y a donc davantage de points de grille pour lesquels le modèle n'est pas en accord avec les capteurs GMON. Pour ce qui est

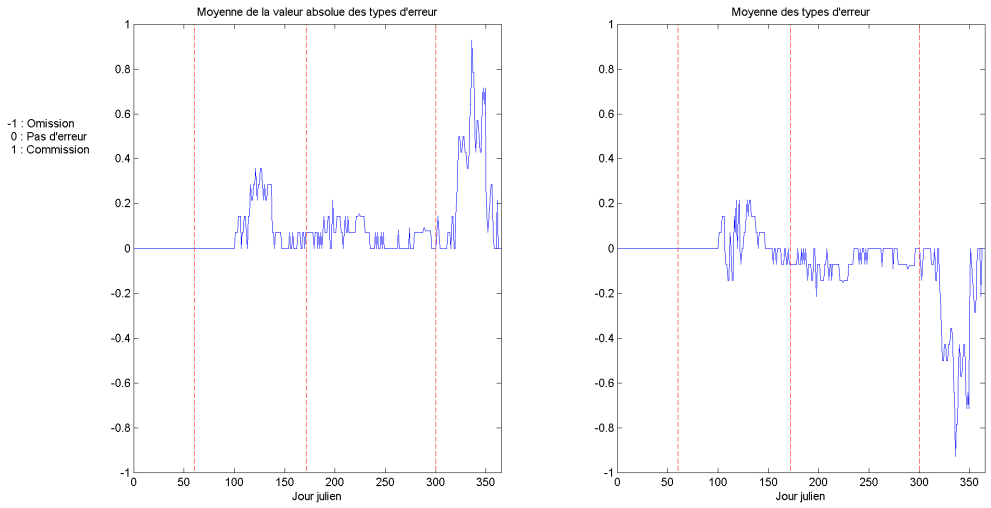

FIGURE 4.19. Moyenne sur les différents points de grille de l'omission/commission pour les réalisations neige/non-neige du modèle à trois variables explicatives selon le point de césure 0,3 avec les réalisations des capteurs SR50, seuil 2 cm, Québec, 2011.

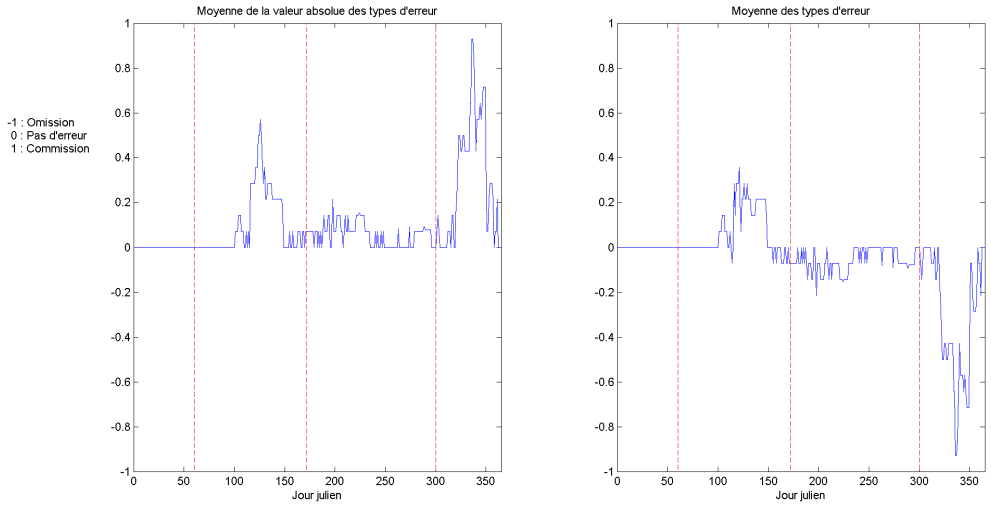

FIGURE 4.20. Moyenne sur les différents points de grille de l'omission/commission pour les réalisations neige/non-neige du modèle à quatre variables explicatives selon le point de césure 0,3 avec les réalisations des capteurs SR50, seuil 2 cm, Québec, 2011.

du type d'erreurs commises lors de cette période, le comportement est semblable à celui de la période de fonte, mais dans le sens inverse. En effet, il semble y avoir un certain équilibre entre le nombre d'erreurs de chaque type pour le modèle à deux variables explicatives, même s'il y a plus d'erreurs d'omission puisque la

moyenne est proche de 0, mais négative. Cependant, le nombre d'erreurs d'omission augmente avec le nombre de variables explicatives, car la moyenne est de plus en plus proche de -1.

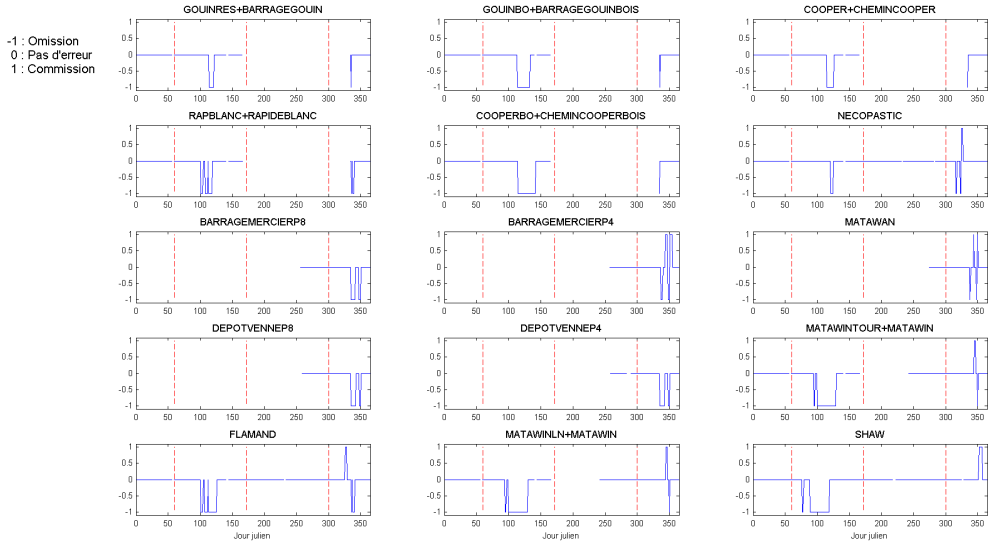

FIGURE 4.21. Omission/commission pour les réalisations neige/non-neige du modèle à deux variables explicatives avec le point de césure 0,3 avec les réalisations des capteurs GMON pour le seuil de 1 cm, Québec, 2011. Les espaces vides dans les graphiques sont dus à des valeurs manquantes dans les données des GMON

.

Des conclusions similaires sont obtenues avec les points de césure 0,4 et 0,5 (voir figures à l'annexe F.3).

4.3. CONCLUSION PARTIELLE

Dans ce chapitre, les données et les outils utilisés pour effectuer la validation ainsi que les principaux résultats de la validation pour l'année 2011 ont été présentés. Trois critères ont été considérés : le score de Brier, la proportion de concordance et le critère d'erreurs d'omission/commission. Nous avons toutefois favorisé les résultats du score de Brier, car celui-ci est particulièrement adapté au contexte du projet.

– pour le score de Brier, les modèles à deux, trois et quatre variables explicatives ont été évalués en fonction des données SR50 et GMON pour des seuils de 1 à 5 cm et en fonction des données SSM/I. Il a été conclu que le comportement général des résultats est semblable, peu importe le seuil

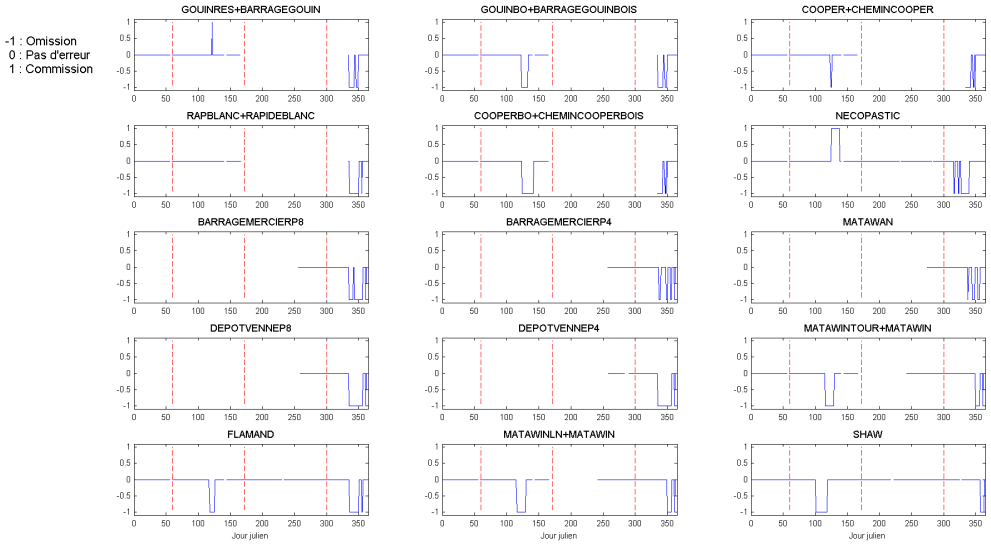

FIGURE 4.22. Omission/commission pour les réalisations neige/non-neige du modèle à trois variables explicatives avec le point de césure 0,3 avec les réalisations des capteurs GMON pour le seuil de 1 cm, Québec, 2011.

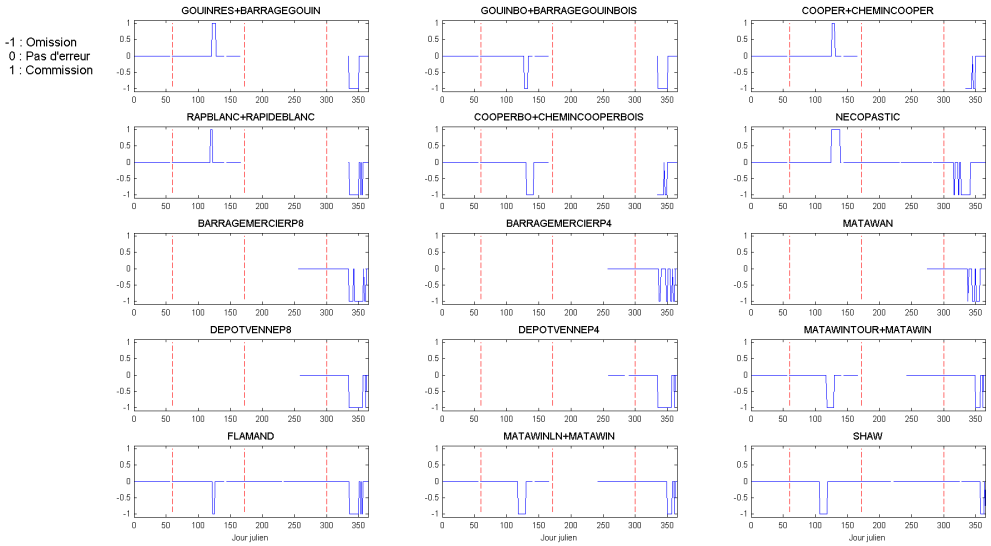

FIGURE 4.23. Omission/commission pour les réalisations neige/non-neige du modèle à quatre variables explicatives avec le point de césure 0,3 avec les réalisations des capteurs GMON pour le seuil de 1 cm, Québec, 2011.

utilisé pour les capteurs et peu importe la combinaison de capteurs GMON utilisée ;

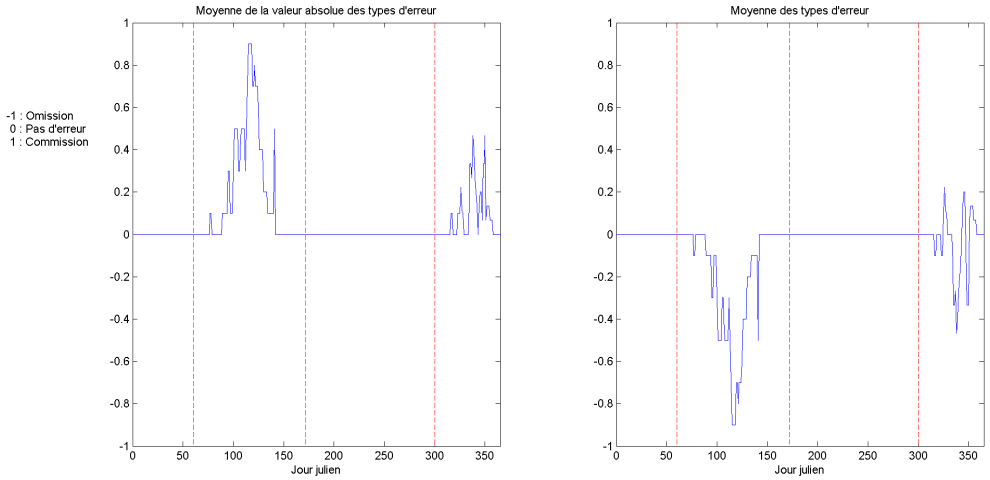

FIGURE 4.24. Moyenne sur les différents points de grille de l'omission/commission pour les réalisations neige/non-neige du modèle à deux variables explicatives selon le point de césure 0,3 avec les réalisations des capteurs GMON, seuil 1 cm, Québec, 2011.

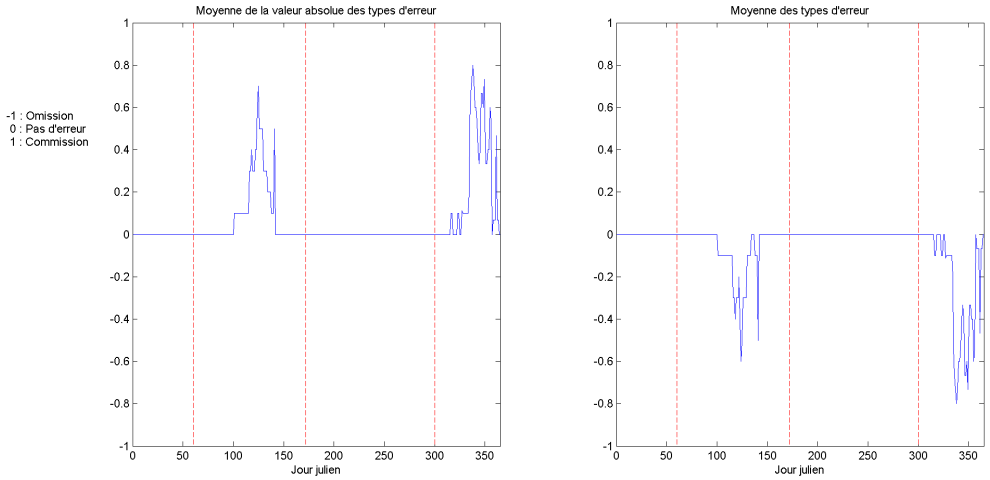

FIGURE 4.25. Moyenne sur les différents points de grille de l'omission/commission pour les réalisations neige/non-neige du modèle à trois variables explicatives selon le point de césure 0,3 avec les réalisations des capteurs GMON, seuil 1 cm, Québec, 2011.

– pour la proportion de concordance, les calculs ont été effectués entre les réalités neige/non-neige des modèles (points de césure 0,1 à 0,9) par rapport aux réalités neige/non-neige des données des capteurs (seuil de 2 cm pour les capteurs SR50 et seuil de 1 cm avec la combinaison 1 de capteurs GMON)

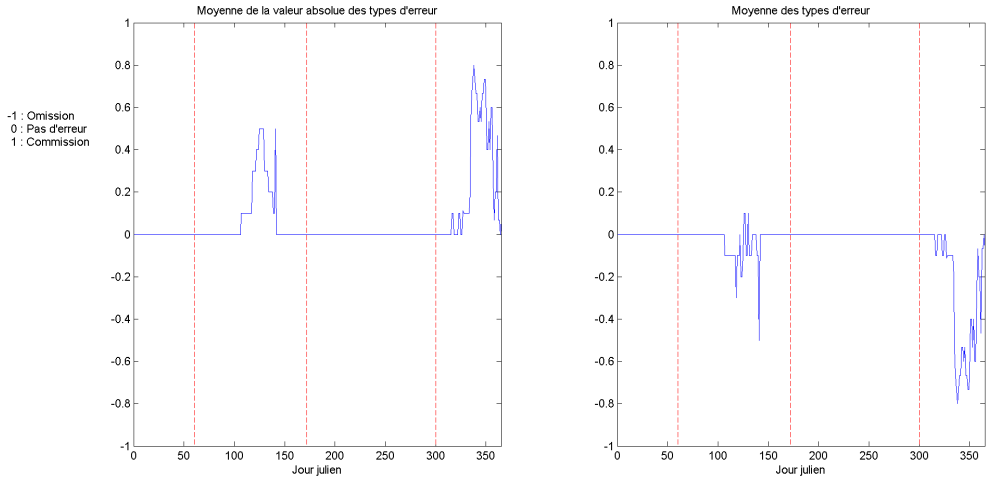

FIGURE 4.26. Moyenne sur les différents points de grille de l'omission/commission pour les réalisations neige/non-neige du modèle à quatre variables explicatives selon le point de césure 0,3 avec les réalisations des capteurs GMON, seuil 1 cm, Québec, 2011.

et des données SSM/I. Il a été constaté que les résultats étaient de meilleure qualité lorsque les points de césure utilisés étaient 0,3, 0,4 et 0,5 ;

– pour le critère d'erreurs d'omission/commission, les calculs ont été réalisés avec les réalités neige/non-neige des modèles selon les points de césure 0,3, 0,4 et 0,5 pour une sélection des capteurs SR50 (seuil de 2 cm) et pour tous les capteurs GMON (seuil de 1 cm).

L'utilisation de chacun des outils diagnostiques a montré que les trois modèles fournissent des cartographies assez fidèles à la réalité pour les périodes de neige et de sol, mais que les résultats sont un peu moins bons lors de la fonte et de l'accumulation de neige. Finalement, les modèles à trois et quatre variables explicatives sont plus performants que celui ne considérant que deux variables explicatives.

Chapitre 5

UTILISATION DES MODÈLES DANS UN CONTEXTE PRÉVISIONNEL

Dans ce chapitre, il est question de l'application des modèles dans un contexte prévisionnel. Puisque les probabilités de présence de neige pour l'année 2011 ont été modélisées de façon satisfaisante, nous désirons nous servir de ces modèles afin de prévoir les probabilités de neige pour des journées futures.

Dans un premier temps, nous vérifions si les modèles estimés à partir des données de 2011 permettent d'obtenir de bons résultats pour la cartographie d'autres années. Par la suite, nous abordons la problématique de la sélection d'un seul modèle. Finalement, les considérations pratiques d'implantation sont discutées afin de pouvoir prédire les probabilités de neige pour des journées futures. Il est à noter que les annexes G à M citées dans ce chapitre ne sont pas incluses dans le présent mémoire, car elles sont trop volumineuses. Elles sont uniquement disponibles sur demande et à l'Institut de recherche d'Hydro-Québec.

5.1. Généralisation du modèle sur d'autres années

Dans cette section, les modèles estimés aves les données de 2011 sont utilisés pour calculer les estimations des probabilités de neige pour les années 2005 à 2010 et 2012. La qualité de ces estimations est ensuite évaluée à l'aide des outils présentés à la section 4.1.

Il est à noter que les données des capteurs SR50 et les données SSM/I sont disponibles pour toutes les années de la validation. Toutefois, nous ne disposons que des données des capteurs GMON pour les années 2009, 2010 et 2012. De plus, la validation est effectuée avec le seuil de 2 cm pour les SR50 ainsi qu'avec la combinaison 1 de chaque année et le seuil de 1 cm pour les GMON pour les mêmes raisons qu'au chapitre précédent. Aussi, le modèle à quatre variables explicatives a été utilisé pour la comparaison entre les années, car c'est celui qui fournit les

meilleurs résultats selon le score de Brier et les proportions de concordance lors de la période de fonte.

5.1.1. Validation à l'aide du score de Brier

Les résultats de la validation pour chaque année sont présentés ici selon le score de Brier. En fait, l'intérêt de cette validation est de vérifier si le comportement des scores de Brier calculés sur les données des capteurs SR50, celles des capteurs GMON et les données SSM/I est le même pour toutes les années.

Suite à l'étude des graphiques correspondant pour chaque année (le lecteur peut consulter les annexes G.1.1, H.1.1, I.1.1, J.1.1, K.2.1, L.2.1, M.2.1, K.2.2, L.2.2, M.2.2, G.1.2, H.1.2, I.1.2, J.1.2, K.2.3, L.2.3 et M.2.3 si nécessaire), il est possible de constater que les conclusions sont assez similaires à celles qui avaient été obtenues pour l'année 2011. En effet, les modèles à deux, trois et quatre variables explicatives sont assez satisfaisants pendant les périodes de neige et de sol (scores de Brier près de 0), sauf pour le début de quelques années pour les données SSM/I. De plus, ils sont un peu moins satisfaisants dans les périodes de fonte et d'accumulation dû possiblement à un problème d'incertitude pendant la fonte et de mauvaise réalité durant l'accumulation. Aussi, les scores de Brier se comportent similairement, peu importe le seuil choisi pour les données des capteurs SR50 et peu importe la combinaison (voir le détail des combinaisons aux annexes K.1, L.1 et M.1) ou le seuil pour les données des capteurs GMON. Également, les modèles à trois et quatre variables explicatives semblent, encore une fois, être les plus performants pendant la période de fonte, tandis que celui à deux variables explicatives semble plus satisfaisant pendant la période d'accumulation de neige.

Les figures 5.1 à 5.3 montrent un comportement des scores de Brier similaire d'une année à l'autre. En effet, ils sont stables durant les périodes de neige et de sol et l'évolution pendant les périodes de fonte et d'accumulation de neige est semblable. Le modèle à quatre variables explicatives est satisfaisant, peu importe l'année, sauf pour la période d'accumulation de neige, même si le modèle fournit de meilleures prévisions pour certaines années.

5.1.2. Validation à l'aide des proportions de concordance

Dans cette section, il est vérifié si le comportement des proportions de concordance est similaire d'une année à l'autre. L'examen des graphiques correspondant pour chaque année (le lecteur peut consulter les annexes G.2.1, H.2.1, I.2.1, J.2.1, K.3.1, L.3.1, M.3.1, K.3.2, L.3.2, M.3.2, G.2.2, H.2.2, I.2.2, J.2.2, K.3.3, L.3.3 et M.3.3 si nécessaire) montre également pour cet outil que les conclusions sont similaires aux résultats de 2011. Effectivement, les modèles à deux, trois et quatre

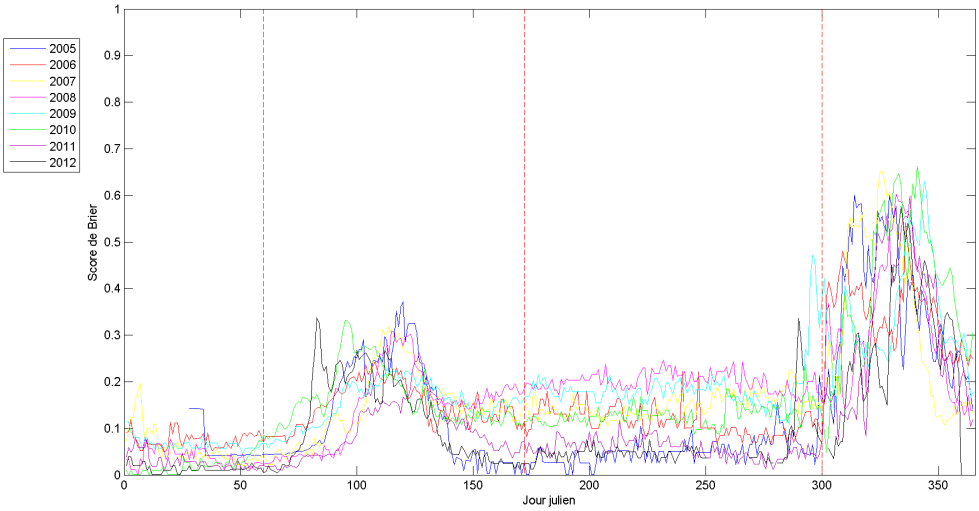

FIGURE 5.1. Comparaison des scores de Brier pour les données des capteurs SR50 des années 2005 à 2012, seuil à 2 cm, modèles à quatre variables explicatives, Québec.

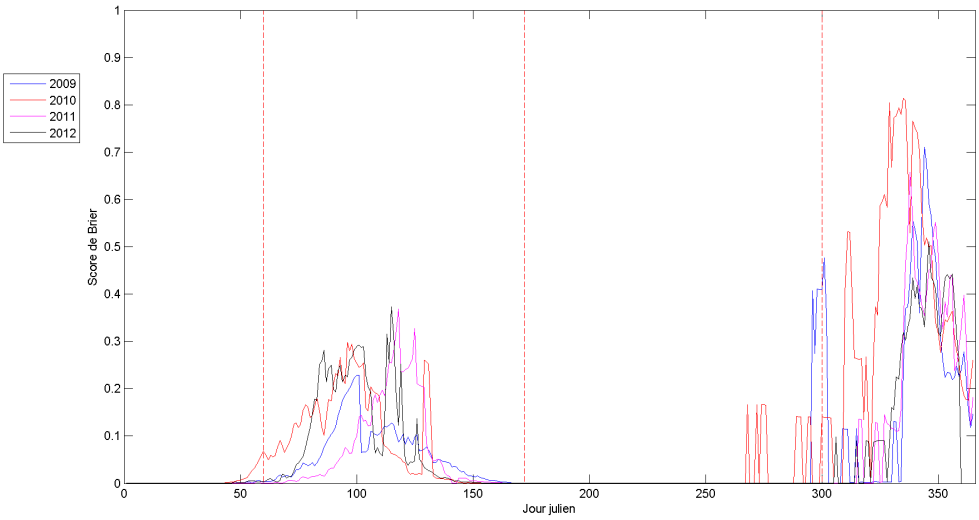

FIGURE 5.2. Comparaison des scores de Brier pour les données des capteurs GMON des années 2009 à 2012, seuil à 1 cm, combinaison 1, modèles à quatre variables explicatives, Québec.

variables explicatives sont tous généralement satisfaisants pendant les périodes de neige et de sol (proportions de concordance proches de 1), sauf pour le début de quelques années pour les données SSM/I. Aussi, les proportions de concordance

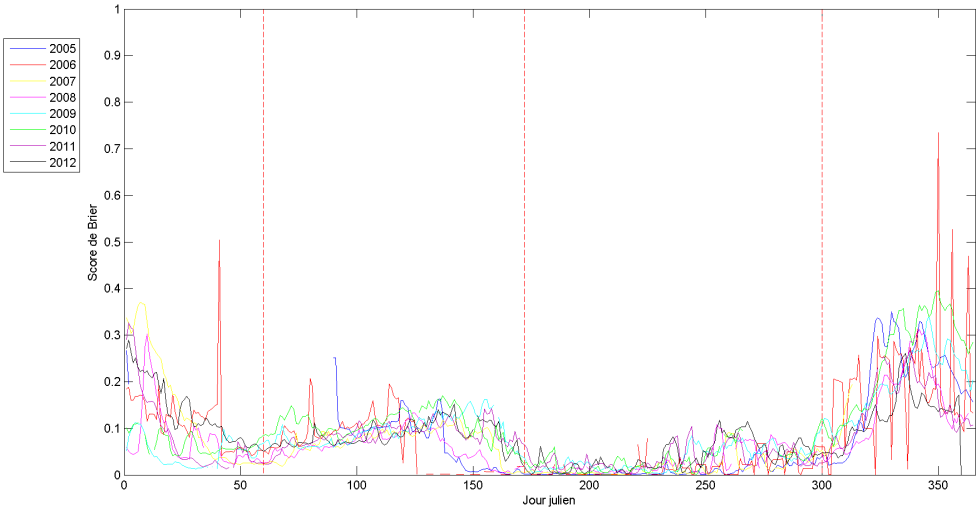

FIGURE 5.3. Comparaison des scores de Brier pour les données SSM/I des années 2005 à 2012, modèles à quatre variables explicatives, Québec.

diminuent durant les périodes de fonte et d'accumulation de neige et ces variations sont similaires, peu importe le seuil ou la combinaison utilisée pour les SR50 et les GMON. Également, les modèles à trois et quatre variables explicatives sont, encore une fois, généralement plus performants que celui à deux variables explicatives durant la fonte et celui à deux variables explicatives est plus satisfaisant durant l'accumulation de neige.

En comparant les années entre elles aux figures 5.4, 5.5 et 5.6, il est possible de constater que le comportement des proportions de concordance est semblable pour chaque année. Effectivement, les valeurs sont toutes très satisfaisantes durant les périodes de neige et de sol et elles sont toutes similaires durant les périodes de fonte et d'accumulation, même si les résultats de certaines années semblent meilleurs selon la période observée.

5.1.3. Validation à l'aide des erreurs d'omission/commission

Les dernières validations sont effectuées à l'aide du critère des erreurs d'omission/commission pour les différents modèles (points de césure 0,3, 0,4 et 0,5) selon les données des capteurs SR50 (seuil de 2 cm) et des capteurs GMON (seuil de 1 cm). En étudiant les figures correspondantes (voir les annexes G.3.1, H.3.1, I.3.1, J.3.1, K.4.1, L.4.1 et M.4.1 pour la validation avec les SR50 ainsi que les annexes K.4.2, L.4.2 et M.4.2 pour la validation avec les GMON, si nécessaire), il est remarqué que les conclusions sont similaires à celles qui avaient été obtenues pour

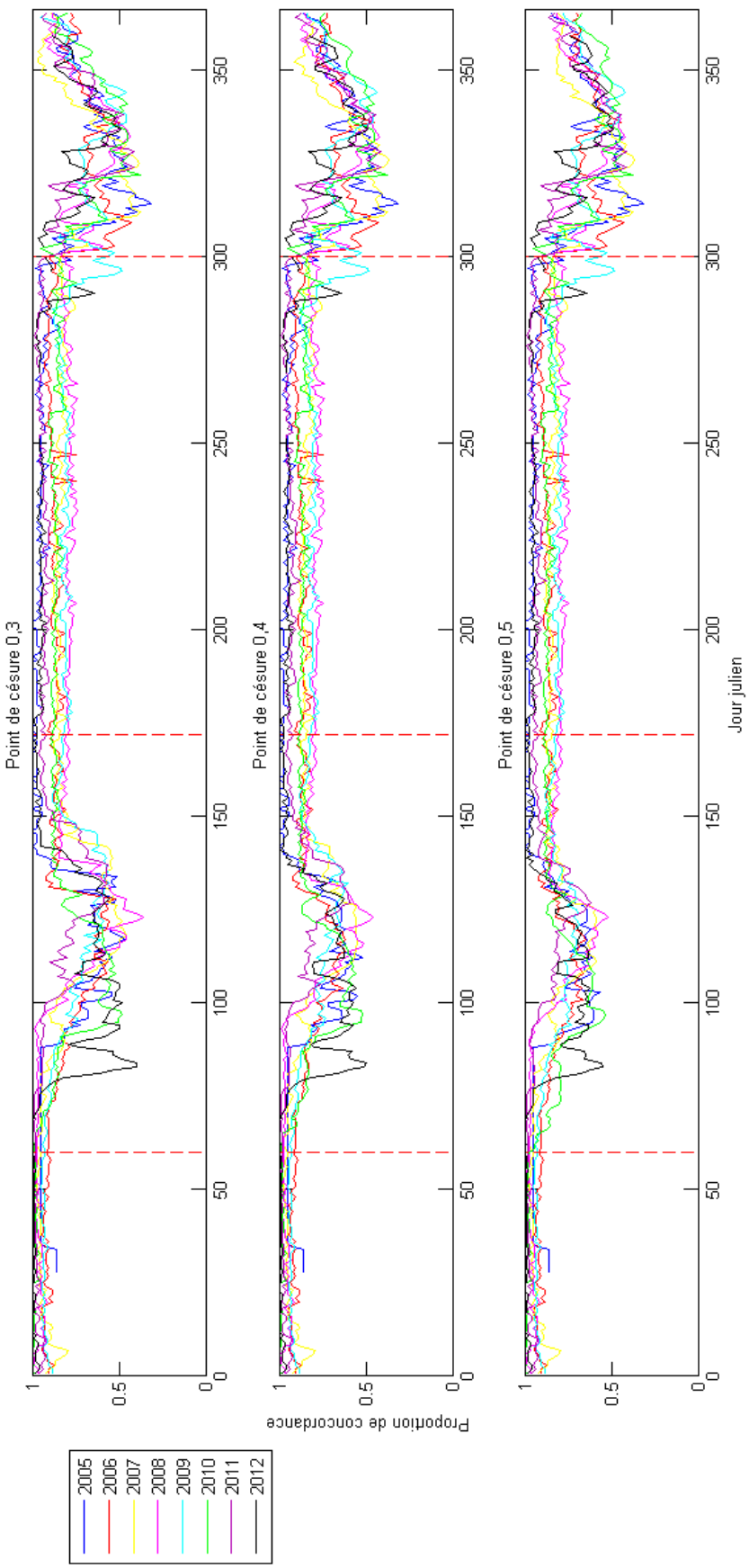

FIGURE 5.4. Comparaison des proportions de concordance pour les données des capteurs SR50 des années 2005 à 2012, seuil à 2 cm, point de césure 0,3, 0,4 et 0,5, modèles à quatre variables explicatives, Québec.

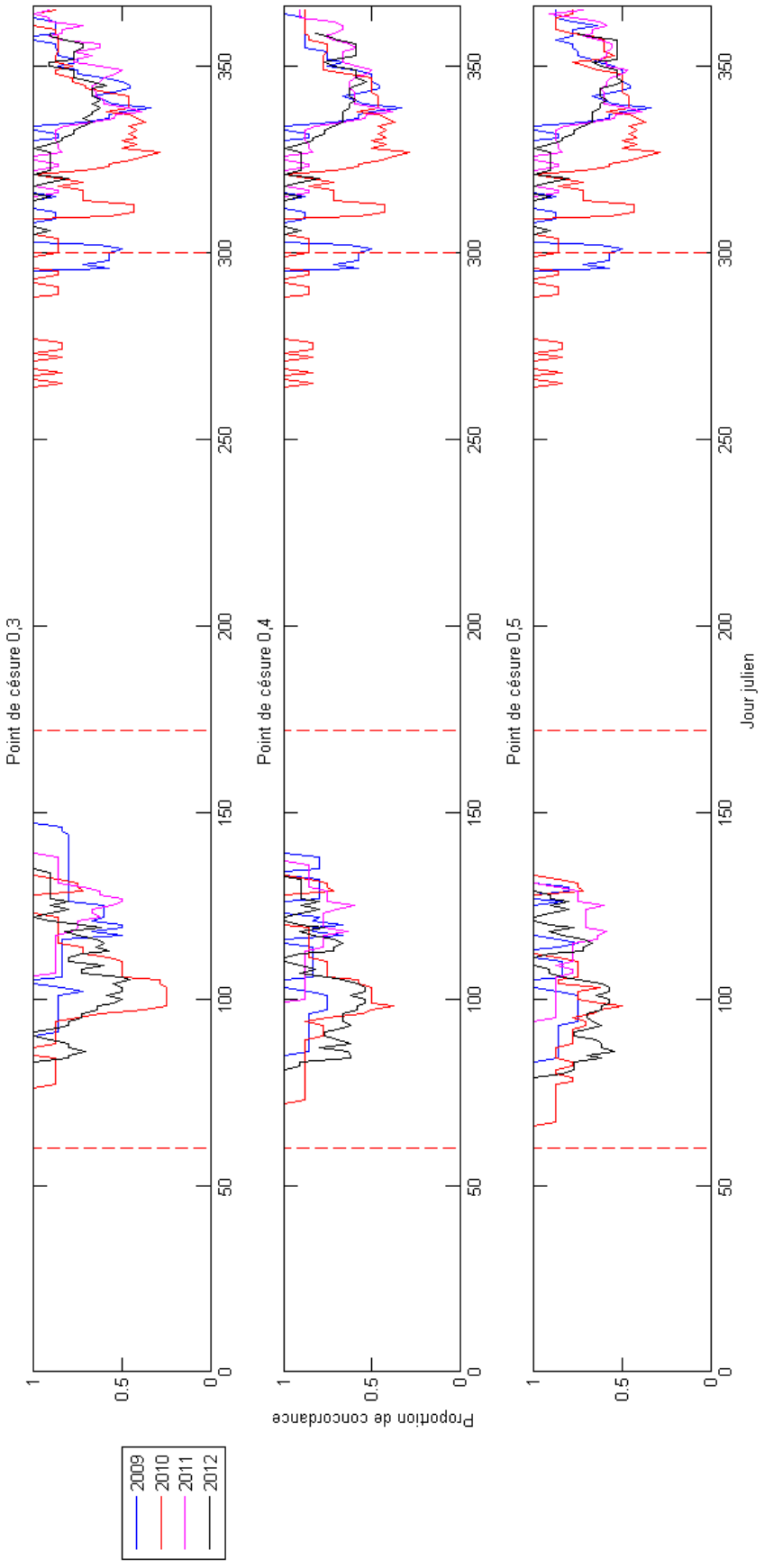

FIGURE 5.5. Comparaison des proportions de concordance pour les données des capteurs GMON des années 2009 à 2012, seuil à 1 cm, combinaison 1, point de césure 0,3, 0,4 et 0,5, modèles à quatre variables explicatives, Québec.

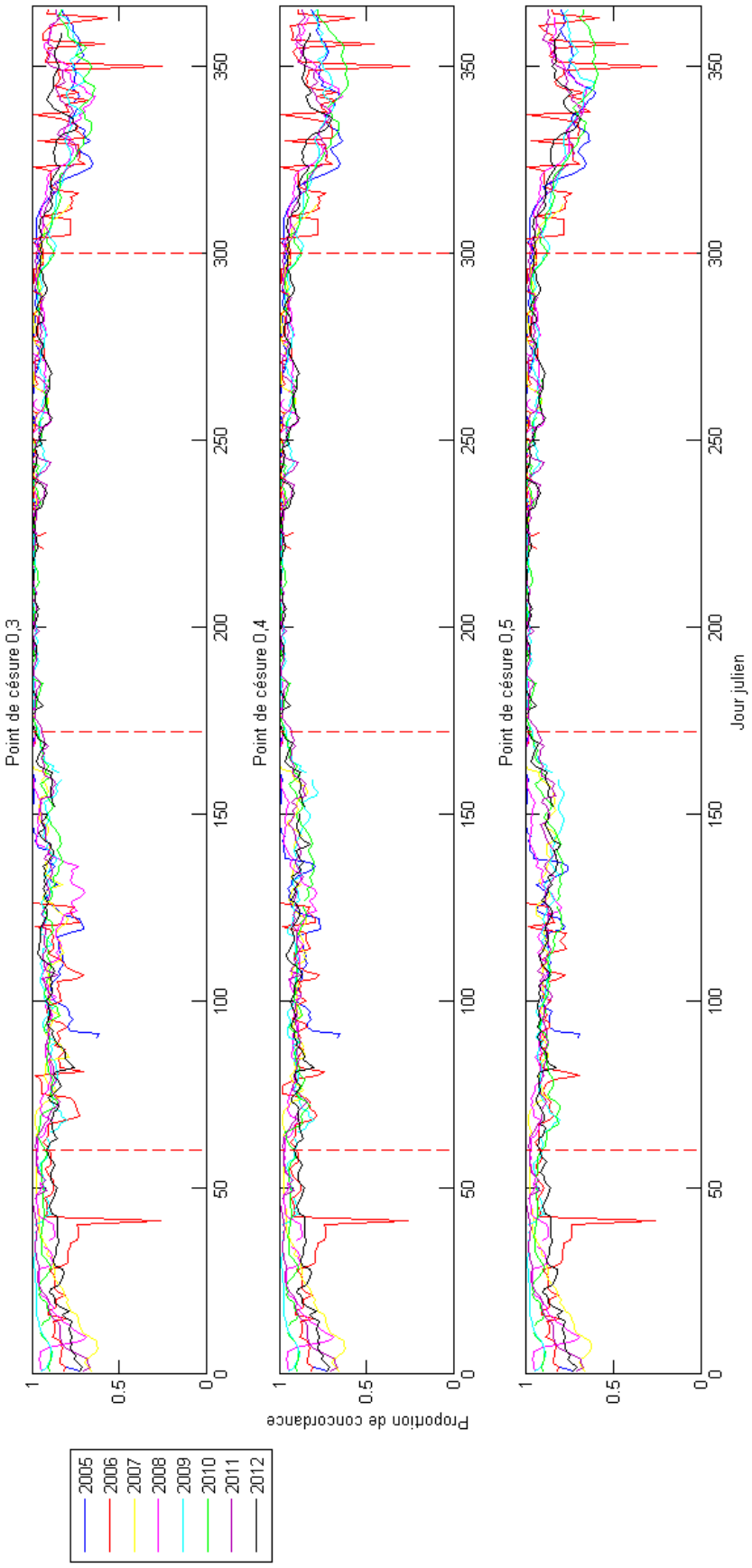

FIGURE 5.6. Comparaison des proportions de concordance pour les données SSM/I des années 2005 à 2012, point de césure 0,3, 0,4 et 0,5, modèles à quatre variables explicatives, Québec.

l'année 2011. Selon ce critère avec les SR50, peu importe le point de césure ou le nombre de variables explicatives, il n'y a presque pas d'erreurs pour les périodes de neige et de sol. Pour la période de fonte, il y a des erreurs d'omission et de commission, tandis que, durant la période d'accumulation, il s'agit surtout d'erreurs d'omission. Avec les GMON, l'examen des résultats montre qu'il n'y a pas d'erreur durant les périodes de neige et de sol, mais qu'il y en a pour les périodes de fonte et d'accumulation. Pour la période de fonte, il s'agit principalement d'erreurs de d'omission, mais il y a de plus en plus d'erreurs de commission lorsque le nombre de variables explicatives augmente. Lors de la période d'accumulation, des erreurs des deux types surviennent, mais le nombre d'erreurs d'omission augmente avec le nombre de variables explicatives. De façon générale, il y a moins d'erreurs (tous les types confondus) lorsque le nombre de variables explicatives est grand.

5.2. Choix du modèle à conserver

Dans l'optique où une prévision doit être effectuée quotidiennement aux points de grille, le choix d'un seul modèle devient une priorité. Pour ce faire, le choix d'un seul des modèles parmi les trois qui ont été proposés dans le chapitre 3 est discuté. Par la suite, le moyennage de modèles est présenté comme étant une alternative permettant de combiner l'information de plusieurs modèles, et ainsi de prendre en compte leur incertitude.

5.2.1. Choix d'un seul modèle parmi ceux proposés

Tout d'abord, à la lumière des résultats obtenus au chapitre 4 et à la section 5.1, il est possible d'affirmer que le modèle à deux variables explicatives n'est pas le plus optimal par rapport à ceux à trois et quatre variables explicatives lorsque nous évaluons la qualité de chaque modèle pour l'ensemble du territoire québécois. En effet, même si les différents outils semblent indiquer qu'il est le plus satisfaisant lors de l'accumulation de neige, il est peu performant pendant la fonte, qui est une période critique pour la prévision hydrologique d'Hydro-Québec. Toutefois, les modèles à trois et quatre variables explicatives semblent assez similaires, il est donc plus ardu de choisir entre les deux modèles.

Puisque ces conclusions ont été obtenues en calculant les valeurs des outils pour le domaine d'étude en entier, il se peut que cela cache des comportements différents selon les zones du Québec. Autrement dit, il est possible que le nombre de variables explicatives nécessaire pour effectuer une bonne prévision diffère selon la zone étudiée. Afin de vérifier cette hypothèse, les scores de Brier sont calculés

individuellement pour chacune des zones pour l'année 2011 avec les données des capteurs SR50, celles des capteurs GMON et les données SSM/I pour chaque zone et pour chaque modèle. Les figures 5.7, 5.8 et 5.9 montrent que les résultats sont similaires à ceux obtenus pour le Québec en entier. En effet, pour la majorité des données et des zones, le modèle à deux variables explicatives est le moins satisfaisant durant la période de fonte, mais il est le plus performant pendant l'accumulation de neige. De plus, la qualité des modèles à trois et quatre variables explicatives est encore une fois semblable, donc il est difficile de faire un choix.

La période d'intérêt pour Hydro-Québec étant la fonte, il est assez évident que le modèle à deux variables explicatives n'est pas celui qui doit être conservé. Au niveau statistique, les résultats indiquent que les valeurs du score de Brier et des autres outils sont généralement meilleures pour le modèle à quatre variables explicatives. Toutefois, les différences avec le modèle à trois variables explicatives sont faibles, ce qui peut amener à s'interroger sur la pertinence d'ajouter une quatrième variable explicative.

5.2.2. Moyennage de modèles

À la lumière des résultats de validation, les modèles à trois et quatre variables explicatives produisent de bons résultats pour la période de fonte tandis que celui à deux variables explicatives est meilleur pour la période d'accumulation de neige. À la section précédente, différentes avenues pour le choix d'un seul des modèles ont été discutées. Dans celle-ci, une alternative permettant de combiner l'information de différents modèles est décrite. Le but de cette approche est de pouvoir conserver les meilleurs résultats autant pour la fonte que pour l'accumulation de neige.

Pour ce faire, un moyennage de modèles est effectué de façon journalière par le calcul d'un poids $\omega_{t,k}$ attribué au modèle k pour la journée t. Par exemple, s'il n'y a que deux modèles à moyenner (M_1 et M_2), le modèle moyenné M est créé comme suit :

$$\mathrm{M}_t = \omega_{t,1}\mathrm{M}_{t,1} + \omega_{t,2}\mathrm{M}_{t,2}, \tag{5.2.1}$$

où $\omega_{t,1} + \omega_{t,2} = 1$ pour $t = 1, \ldots, 365$.

Le calcul des poids peut se faire de différentes façons. Perreault *et al.* (1995) proposent de calculer les poids en utilisant la somme des carrés des résidus de chaque modèle de régression de la façon suivante dans le cas où il n'y a que deux modèles considérés :

$$\omega_r = \frac{1}{1+\lambda}; \tag{5.2.2}$$

$$\omega_s = \frac{\lambda}{1+\lambda}, \tag{5.2.3}$$

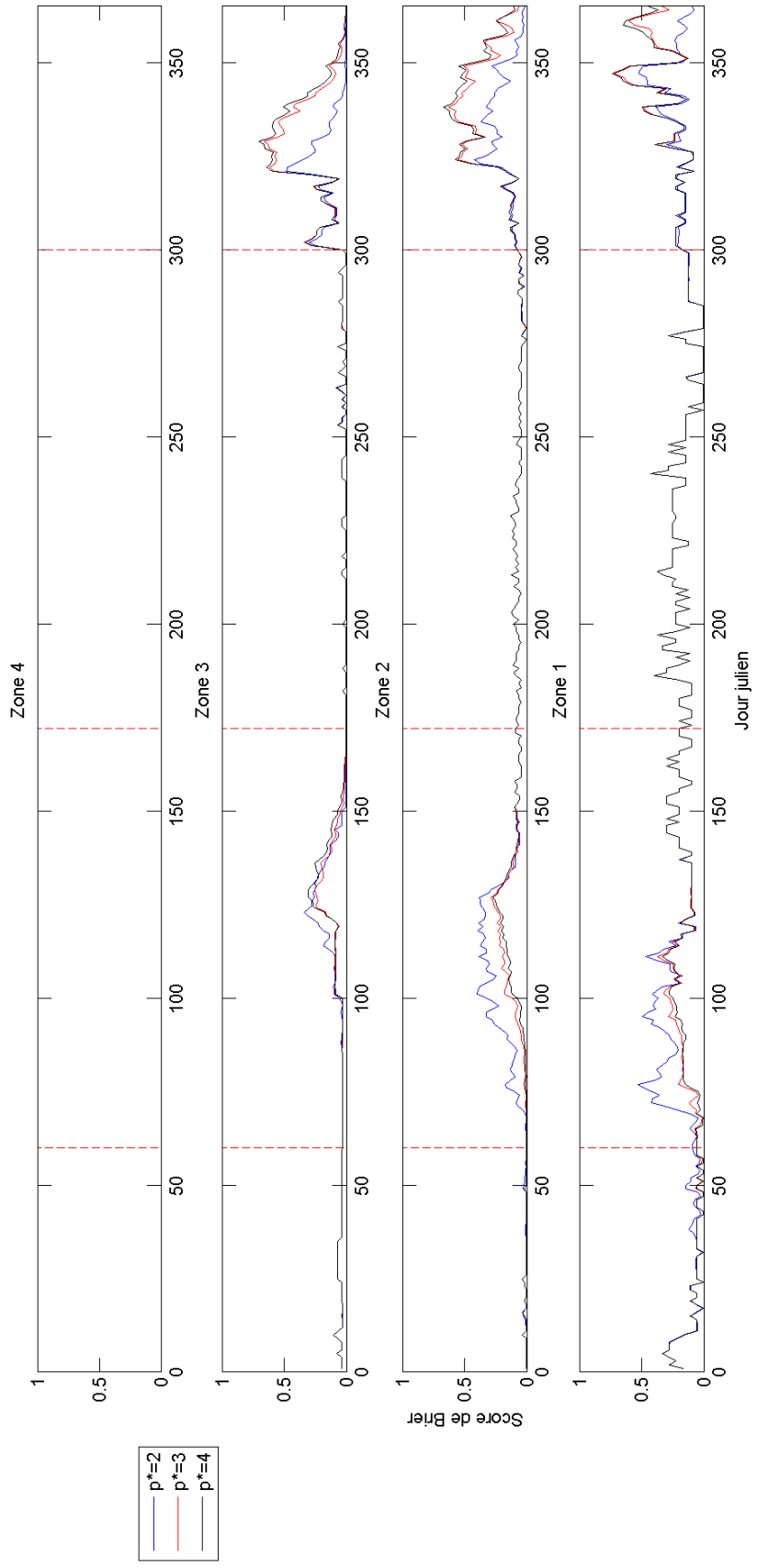

FIGURE 5.7. Score de Brier pour les données des capteurs SR50 selon la zone bioclimatique, seuil à 2 cm, Québec, 2011. Il n'y a pas d'information dans la zone 4, car il n'y a pas de capteurs SR50.

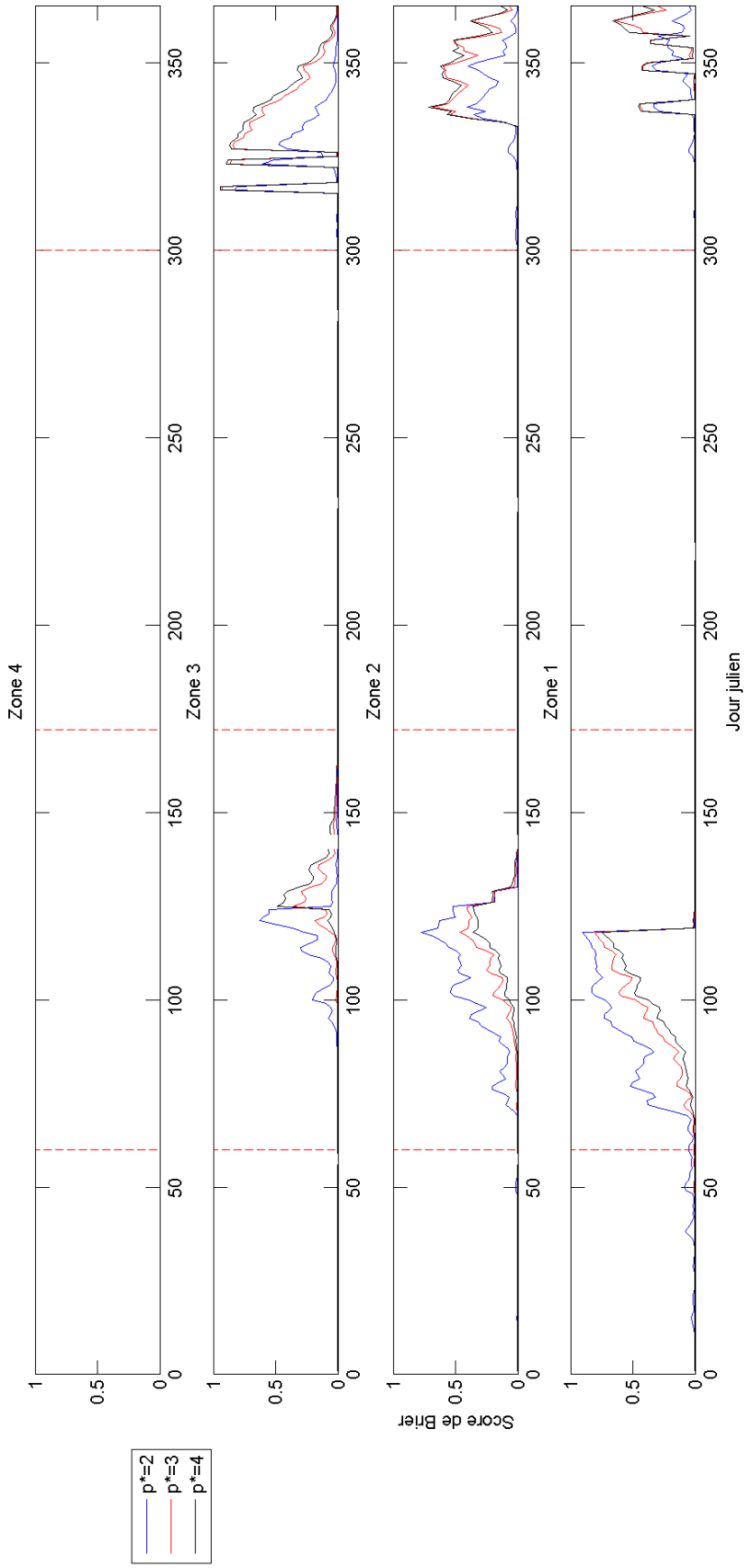

FIGURE 5.8. Score de Brier pour les données des capteurs GMON selon la zone bioclimatique, seuil à 1 cm, Québec, 2011. Il n'y a pas d'information dans la zone 4, car il n'y a pas de capteurs GMON.

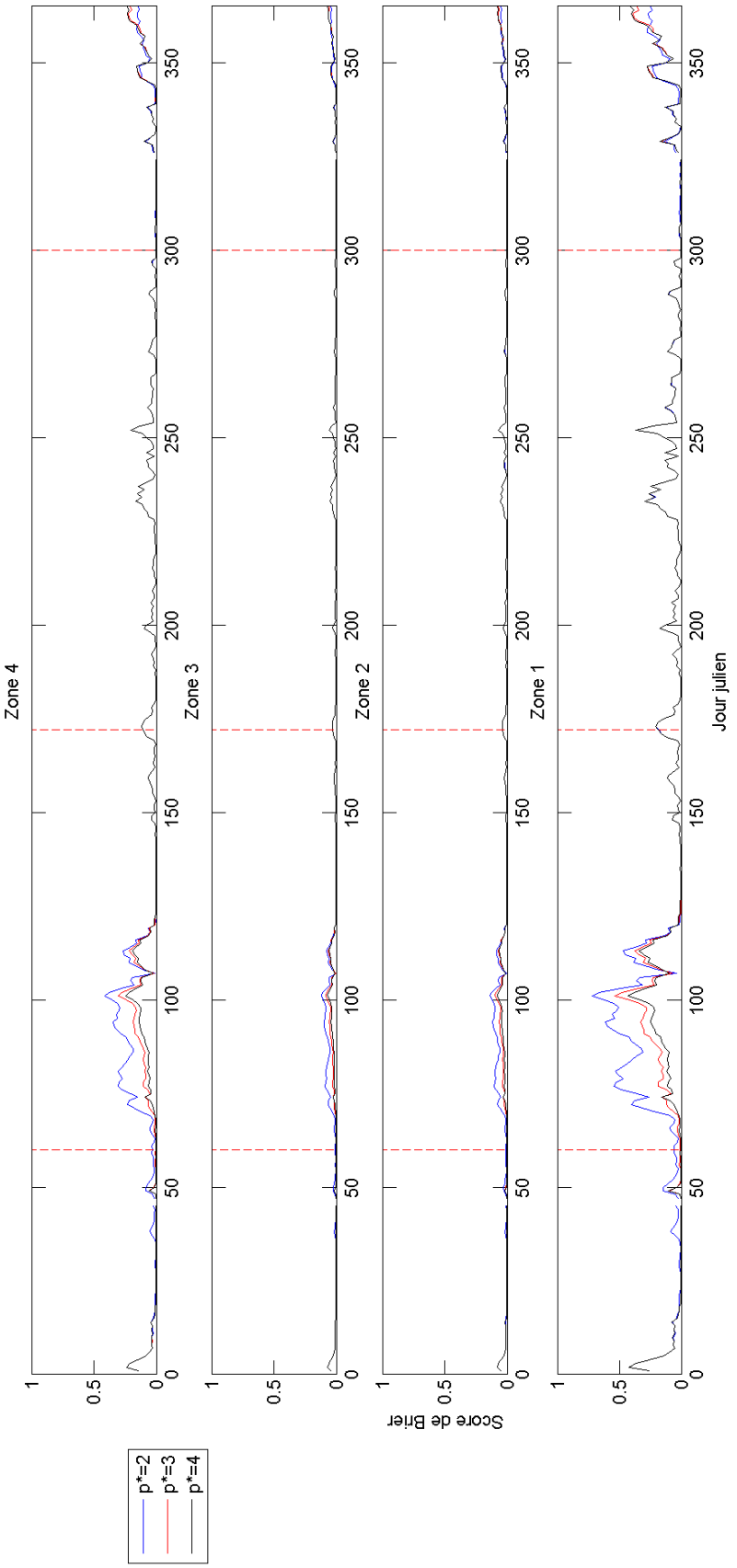

FIGURE 5.9. Score de Brier pour les données des capteurs SSM/I selon la zone bioclimatique, Québec, 2011.

où $\lambda = \frac{SSE_r}{SSE_s}$ et SSE_k est la somme des carrés des résidus du modèle k. Toutefois, même si nous disposons bel et bien de modèles de régression, et donc des sommes de carrés des résidus, cette approche ne permet pas de calculer des poids quotidiens. Également, il semble plus pertinent pour le projet de calculer les poids selon les outils de validation utilisés, car nous souhaitons que le moyennage reflète les conclusions obtenues au chapitre 4 et à la section 5.1.

Puisque le score de Brier est l'outil le plus approprié dans notre contexte (voir section 4.1.3), il a été décidé de présenter une méthodologie de moyennage basé sur celui-ci. Pour ce faire, le modèle à deux variables explicatives est utilisé, car il produit les meilleurs résultats pendant l'accumulation, et celui à trois variables explicatives est retenu, car il produit de bons résultats pendant la fonte. Les autres combinaisons possibles pour le moyennage de modèles ne sont pas présentées dans cette section, car la méthodologie est introduite à titre exploratoire.

En s'inspirant des équations (5.2.2) et (5.2.3), le calcul des poids est fait de la façon suivante :

$$\omega_{t,2} = \frac{1}{1+\lambda_t};$$
$$\omega_{t,3} = \frac{\lambda_t}{1+\lambda_t},$$

où $\omega_{t,j}$ est le poids du jour t du modèle à j variables explicatives, $\lambda_t = \frac{\mathrm{SB7}_{3,t}}{\mathrm{SB7}_{2,t}}$ et $\mathrm{SB7}_{j,t} = \frac{1}{7}\sum_{i=1}^{7} \mathrm{SB}_{j,t-i}$ est la moyenne des scores de Brier du modèle j pour les sept jours précédents. Le choix d'effectuer cette moyenne sur sept jours est justifié de la même façon qu'à la section 3.3.2, c'est-à-dire par le fait qu'il s'agit d'un choix qui permet de tenir compte d'un nombre suffisant de jours précédents, sans trop s'éloigner de la journée d'émission de la prévision. Cette valeur est cependant exploratoire et pourrait être modifiée.

Une fois que les poids pour chacun des modèles sont calculés, le modèle moyenné est obtenu à l'aide de l'équation (5.2.1). Par la suite, les scores de Brier associés à ce nouveau modèle sont calculés en utilisant les données de capteurs SR50 (seuil 2 cm), celles des capteurs GMON (seuil 1 cm et combinaison 1) et les données SSM/I.

La figure 5.10 montre que le modèle obtenu par moyennage représente bel et bien un compromis entre les modèles à deux et trois variables explicatives. En effet, ses valeurs du score de Brier sont comprises entre celles des deux autres modèles. En procédant de la sorte, un gain a donc été obtenu par rapport au modèle à trois variables explicatives pour la période d'accumulation de neige, car les performances du modèle sont désormais plus similaires à celles du modèle à

deux variables explicatives. Toutefois, il y a également une perte de performance par rapport à ce même modèle pendant la période de fonte.

Pour des travaux futurs qui pourraient être exécutés par une équipe de l'Institut de recherche d'Hydro-Québec ou par un autre étudiant, il serait intéressant de continuer l'exploration du moyennage de modèles. Dans le cas où la méthodologie présentée dans cette section serait utilisée, il serait important de la rendre plus rigoureuse. En effet, les résultats n'ont été présentés que pour le moyennage des modèles à deux et trois variables explicatives, mais une autre combinaison de deux modèles aurait pu être utilisée. Aussi, la méthodologie aurait pu être adaptée pour combiner les trois modèles. De plus, dans l'optique de n'obtenir qu'un seul modèle, il faudrait calculer les poids selon un seul type de données (SR50, GMON ou SSM/I) ou il faudrait trouver une façon de tous les considérer dans le calcul d'une seule série de poids. Enfin, il pourrait être pertinent d'élaborer d'autres méthodologies et d'en comparer les résultats. Effectivement, de nombreux articles scientifiques assez récents traitent du problème de moyennage de modèle. Citons notamment Hoeting *et al.* (1999), Duan *et al.* (2007) et Diks et Vrugt (2010).

5.3. Implantation future du modèle dans un contexte prévisionnel

Rappelons que le but de ce projet est de développer une méthodologie permettant de produire une cartographie des probabilités de présence de neige pour l'ensemble du Québec. Cette cartographie doit pouvoir être produite de façon quotidienne pour des journées futures, puisque l'information sur le couvert nival est prise en considération lors de la production des prévisions hydrologiques. Dans cette section, certains aspects de l'implantation future de notre approche sont donc discutés.

5.3.1. Comment utiliser les modèles afin de produire les prévisions ?

Les modèles à deux, trois et quatre variables explicatives qui ont été développés et validés aux chapitres 3 et 4 atteignent bien les objectifs visés. En effet, le modèle de régression linéaire sur les logits appliqué de façon journalière permet d'obtenir des estimations des probabilités de neige aux points de grille. Par contre, ces modèles ont été construits et validés pour les années 2005 à 2012, donc pour des journées pour lesquelles les valeurs sont déjà observées pour toutes variables explicatives (probabilités de neige des journées précédentes et températures minimales par point de grille). Ces valeurs ne sont pas nécessairement disponibles

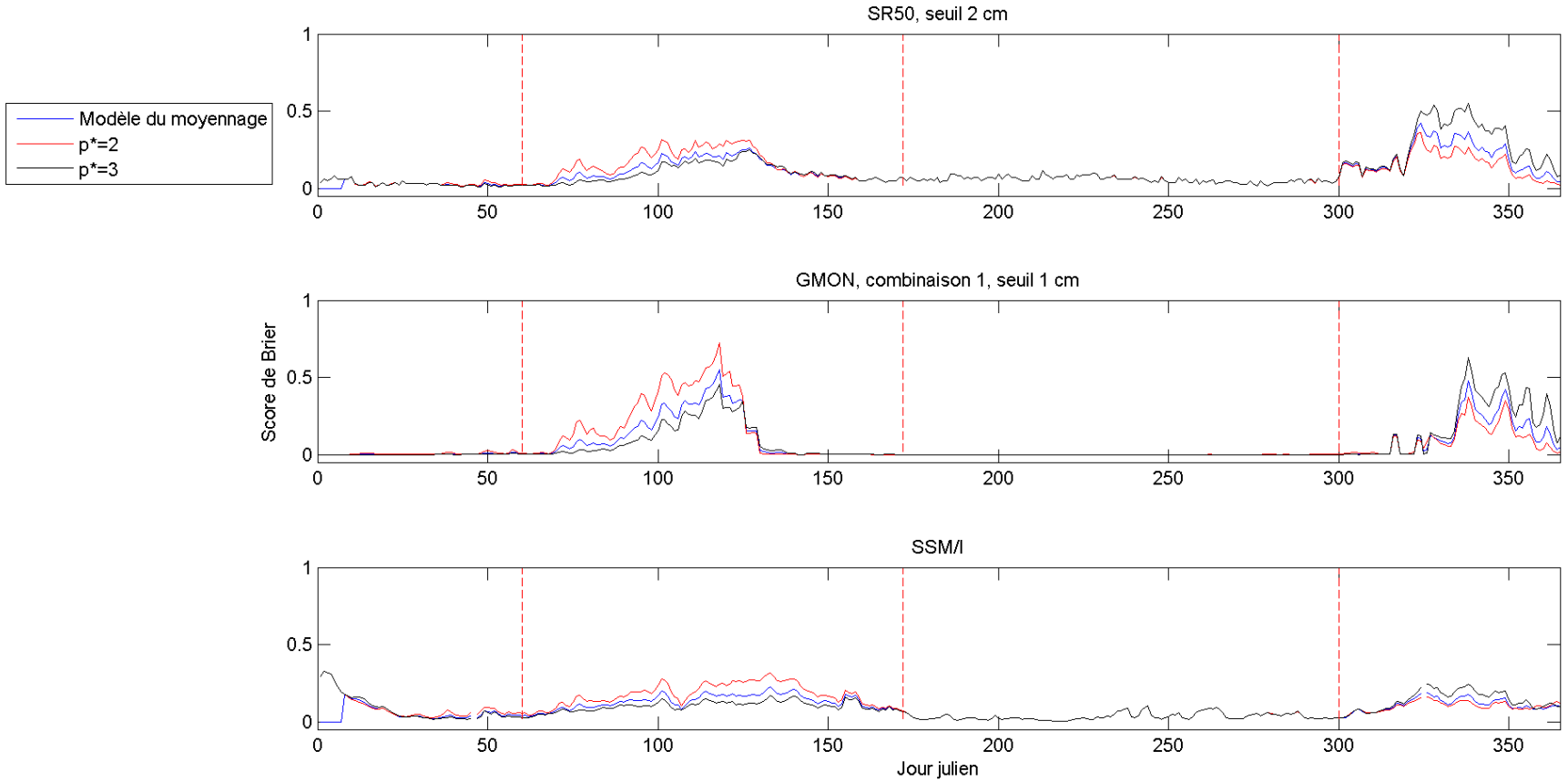

FIGURE 5.10. Comparaison des scores de Brier pour les modèles à deux et trois variables explicatives et pour le modèle obtenu par moyennage. Données des capteurs SR50 (seuil à 2 cm), des capteurs GMON (seuil à 1 cm et combinaison 1) et SSM/I, Québec, 2011.

pour des journées futures. Afin de pouvoir utiliser ces modèles dans un cadre prévisionnel, nous recommandons donc de procéder de la façon suivante :

(1) pour le début de chaque année, initialiser les valeurs des probabilités de neige des sept premiers jours selon les recommandations de la section 3.3.5. La réinitialisation des valeurs au début de chaque année est due au fait que les modèles qui offrent de bons résultats pendant la fonte fournissent généralement de moins bonnes prévisions durant l'accumulation de neige. Toutefois, si la qualité de ces prévisions devait s'améliorer, par exemple par l'utilisation du moyennage, la réinitialisation pourrait ne plus être nécessaire ;

(2) pour prévoir la probabilité de neige du point de grille i pour la journée future t, il faut disposer des valeurs des prévisions des probabilités de neige de ce point de grille pour la journée $t-1$, $t-3$ ou $t-6$, selon le modèle utilisé. Si la probabilité de neige est prédite pour le lendemain, les prévisions des journées antérieures ont déjà toutes été produites, donc elles sont directement utilisables. Par contre, s'il faut prévoir à plus long terme, il faut procéder différemment. Une solution pourrait être d'itérer le processus de prévision jusqu'à ce que les valeurs pour les journées précédentes soient toutes disponibles ;

(3) pour prévoir la probabilité de neige du point de grille i pour la journée future t, la température minimale doit être disponible. Cependant, il s'agit d'une journée future et la température n'a pas encore été observée. Afin d'obtenir une estimation pour cette valeur, une solution possible serait d'utiliser les prévisions de températures d'un des fournisseurs de prévisions météorologiques d'Hydro-Québec ;

(4) lorsque toutes les données nécessaires au modèle sélectionné sont disponibles, il faut appliquer la transformation (2.2.8) puis utiliser la formule de régression associée au modèle. Par la suite, il faut appliquer la transformation inverse (2.2.9) aux valeurs obtenues, ce qui permet d'obtenir les probabilités de neige par point de grille.

5.3.2. Recommandations pour de futurs travaux

Pour de futurs travaux, nous suggérons d'effectuer quelques études afin de vérifier la qualité des résultats dans le cadre prévisionnel.

Au niveau de l'initialisation des premières valeurs, il pourrait tout d'abord être pertinent de vérifier la sensibilité des prévisions par rapport aux valeurs initiales. De plus, il serait intéressant d'effectuer le même exercice en variant le moment

où la réinitialisation annuelle est effectuée (par exemple, en février au lieu d'en janvier).

Pour le choix des données météorologiques, nous recommandons de refaire les étapes de validation des différents modèles, mais en calculant les probabilités de neige pour 2005 à 2012 avec les prévisions de températures. Cela permettrait de vérifier si les modèles sont toujours satisfaisants lorsque les données de températures ne sont pas des observations réelles. Dans le cas où il ne serait pas possible d'obtenir des prévisions pour les températures minimum ou que celles-ci ne soient pas de bonnes qualités, il peut également être intéressant de vérifier la qualité des prévisions en utilisant des températures moyennes ou maximales. Enfin, une autre alternative d'intérêt pourrait être l'utilisation de prévisions d'ensembles de températures, car cela intègre de l'information quant à l'incertitude associée à ces prévisions.

5.4. Conclusion partielle

Dans ce chapitre, les différents points dont il fallait tenir compte afin de pouvoir éventuellement utiliser un des modèles dans un contexte prévisionnel ont été explorés. Dans un premier temps, il a été vérifié que les trois modèles produisent des prévisions de bonne qualité pour les années 2005 à 2010 et 2012. Par la suite, l'obtention d'un seul modèle à été discutée et deux solutions ont été proposées : effectuer un choix parmi les trois modèles développés ou utiliser une méthode de moyennage. Aucune décision quant au modèle à conserver n'est effectuée dans ce mémoire, car il a été jugé que ce choix appartient à ceux qui l'utiliseront. Finalement, les étapes à suivre et quelques recommandations de travaux supplémentaires à effectuer afin de mettre en opération le modèle choisi ont été suggérées.

CONCLUSION

Ce mémoire portait sur l'utilisation de données de micro-ondes passives afin produire une cartographie du couvert nival avec une quantification de l'incertitude via des méthodes statistiques.

Dans le premier chapitre, le domaine d'étude ainsi que les données et les outils actuellement utilisés par l'Institut de recherche d'Hydro-Québec et leurs collaborateurs afin de produire une cartographie de la neige/non-neige ont été présentés. Malgré la précision des données de terrain, leur limite spatio-temporelle représentait un inconvénient pour la cartographie de la neige au sol. Les données des capteurs SR50 et GMON ont donc été considérées pour compléter l'information temporelle. Par contre, même avec l'utilisation d'une méthode d'interpolation, l'aspect spatial n'était pas satisfaisant. Une solution au problème spatial a été d'utiliser des données de télédétection (SSM/I et AVHRR) et de produire une cartographie du couvert nival avec un algorithme de segmentation. Cette méthode ne permettant pas de quantifier l'incertitude et les données AVHRR étant sensibles à la présence de nuage, il a été décidé d'utiliser les données SSM/I afin de développer une nouvelle méthodologie pour produire des cartes neige/non-neige. La principale variable utilisée, le GTV, a donc été dérivée et son comportement a été étudié. Nous avons remarqué qu'elle prenait des valeurs distinctes selon la période de l'année en question, ce qui a été jugé comme une bonne caractéristique pour permettre de discriminer la neige et le sol. Aussi, la pertinence d'utiliser des données relatives à la température de l'air dans un contexte de neige a été discutée et il a été décidé d'utiliser différentes variables explicatives dérivées : Tmin, Tmoy, Tmax, DJx et DCx.

Dans le deuxième chapitre, les différents modèles statistiques utilisés dans le cadre de ce mémoire ont été décrits, soit les mélanges de lois et la régression linéaire multiple. Pour les mélanges de lois, il a été question de la théorie de base ainsi que des méthodes d'estimation des paramètres selon une approche bayésienne. Pour la régression linéaire multiple, un rappel de la théorie de base a été effectué, suivi des différentes façons d'évaluer la qualité d'un modèle. Par ailleurs,

il a été expliqué qu'il était possible d'adapter la méthode pour une variable de probabilité sur l'intervalle [0,1] en effectuant une transformation logit. Enfin, la méthode de sélection de variables qui a été utilisée pour le projet, la sélection à rebours, a été brièvement décrite.

Le but du troisième chapitre était le développement d'un modèle permettant de produire une cartographie quotidienne de la neige/non-neige avec une quantification de l'incertitude. Des expériences antérieures présentées en annexe ont montré que le GTV tel quel ne permettait pas de bien discriminer la neige et le sol et qu'il fallait intégrer de l'information exogène. Pour ce faire, une nouvelle variable représentant la probabilité de neige journalière a été calculée, puis elle a été modélisée via la régression linéaire sur les logits en utilisant des délais allant jusqu'à sept jours et le minimum des températures minimum de la journée comme variables explicatives. L'exercice a tout d'abord été réalisé sur la zone 2 du Québec et, les résultats étant satisfaisants, il a été répété pour le Québec en entier. Trois modèles ont été retenus et validés qualitativement.

Dans le quatrième chapitre, la qualité des trois modèles conservés au chapitre précédent est examinée quantitativement en comparant les cartographies obtenues pour chacun des modèles avec des cartographies de référence. Afin d'effectuer ces comparaisons, trois outils diagnostiques ont été considérés : le score de Brier, la proportion de concordance et le critère des erreurs d'omission/commission. Les données de référence utilisées sont les données des capteurs SR50 et GMON et les données SSM/I classifiées selon un algorithme de segmentation hiérarchique. De façon générale, les résultats ont été jugés satisfaisants, même s'ils sont légèrement moins performants durant les périodes de fonte et d'accumulation selon le modèle considéré. Aussi, il a été conclu que les trois modèles sont de qualité équivalente durant les périodes de neige et de sol, mais que ceux à trois et quatre variables explicatives sont meilleurs que celui à deux variables explicatives durant la période de fonte et moins performants durant la période d'accumulation.

Dans le dernier chapitre, l'utilisation d'un modèle dans un contexte prévisionnel a été mise en perspective. La première étape a été de vérifier si les trois modèles offraient des résultats satisfaisants pour diverses années autres que celle ayant servi à leur construction. Cette validation a été effectuée sur les années 2005 à 2010 et 2012 et il a été conclu que les cartographies obtenues étaient de qualité similaire à celles de 2011. Par la suite, la sélection d'un seul modèle a été discutée et plusieurs pistes de réflexion ont été suggérées, notamment l'utilisation de moyennage de modèles. Finalement, l'implantation future du modèle conservé a été abordée en expliquant comment il pouvait être adapté dans un contexte

prévisionnel et en proposant quelques travaux supplémentaires afin d'en vérifier les performances en mode opérationnel.

L'originalité de ce projet de recherche est que la cartographie du couvert nival est abordée différemment par rapport aux méthodes vues en introduction et au premier chapitre. En effet, une approche statistique a été utilisée en substitut des méthodes usuelles. Aussi, la méthodologie se démarque par l'utilisation de variables explicatives afin de compléter l'information fournie par le GTV. Enfin, un grand atout des modèles qui ont été développés dans le présent mémoire est qu'ils fournissent des probabilités de présence de neige, ce qui intègre une incertitude quant à la prévision de la réalité neige/non-neige.

Lors de la réalisation de ces travaux, plusieurs questions ont émergé qui pourraient faire l'objet de travaux futurs :

- est-ce que la variable p_t est construite de la façon la plus optimale ?
- est-ce que d'autres variables exogènes sont disponibles et pourraient être considérées ?
- est-ce qu'il serait bénéfique de diviser le Québec en plus de quatre zones bioclimatiques ?
- existe-t-il d'autres méthodes de cartographies avec lesquelles les résultats pourraient être comparés ?
- quelles autres validations seraient pertinentes afin d'assurer la performance du modèle choisi ?
- etc.

Évidemment, d'autres avenues pourraient être explorées selon les questions soulevées par les lecteurs suite à la lecture de ce mémoire.

Bibliographie

Bitner, D., T. Carroll, D. Cline et P. Romanov (2002). An assessment of the differences between three satellite snow cover mapping techniques. *Hydrological Processes* **16**(18), 3723–3733.

Brier, G. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review* **78**(1), 1–3.

Burden, R. et J. Faires (2010). *Numerical Analysis*. Cengage Learning.

Casella, G. et E. George (1992). Explaining the gibbs sampler. *The American Statistician* **46**(3), 167–174.

Chang, A., J. Foster et D. Hall (1987). Nimbus-7 SMMR derived global snow cover parameters. *Annals of glaciology* **9**(9), 39–44.

Chang, A. et A. Rango (2000). Algorithm theoretical basis document (ATBD) for the AMSR-E snow water equivalent algorithm. *NASA/GSFC, Nov.*

Chokmani, K., M. Bernier, L. Pâquet, K. Goïta, A. Royer, F. Comtois-Boutet, M. Turcotte, Y. Zhang, L. Forcier et A. Massalabi (2009). *Développement d'algorithmes pour le suivi par satellite de la couverture de neige au sol à l'échelle du bassin versant : Rapport final*. INRS-ETE.

Chokmani, K., M. Bernier et M. Slivitzky (2006). Suivi spatio-temporel du couvert nival du Québec à l'aide des données NOAA-AVHRR. *Revue des sciences de l'eau/Journal of Water Science* **19**(3), 163–179.

De Sève, D. (1999). *Développement d'un algorithme pour cartographier l'équivalent en eau de la neige au sol (EEN) dans un environnement de taïga à partir des données de micro-ondes passives du capteur SSM/I.* Thèse, Université du Québec, Institut national de la recherche scientifique.

De Seve, D., F. Vachon et Y. Choquette (2012a). A dynamic algorithm for mapping of snow cover using SSMI data. In *Geoscience and Remote Sensing Symposium (IGARSS), 2012 IEEE International*, pp. 4875–4878. IEEE.

De Sève, D. (2011). Cartographie neige/non-neige : résultats préliminaires.

De Sève, D. (2014). Cartographie neige/non-neige : La démarche, les limites et les pistes de solutions proposées.

De Sève, D., L. Perreault, F. Vachon, F. Guay et Y. Choquette (2012b). Assessment of dynamic probabilistic methods for mapping snow cover in québec canada. In *EGU General Assembly Conference Abstracts*, Volume 14, pp. 9052.

De Sève, D., F. Vachon, M. Massala-Kivoua, D. Tapsoba et E. N. (2008). Prototype de fusion automatique de données de télédétection in situ pour l'estimation de l'équivalent en eau de la neige (ÉEN) pour le bassin versant de la rivière la grande.

Derksen, C., A. Walker et B. Goodison (2003a). A comparison of 18 winter seasons of in situ and passive microwave-derived snow water equivalent estimates in Western Canada. *Remote Sensing of Environment* **88**(3), 271–282.

Derksen, C., A. Walker, E. LeDrew et B. Goodison (2003b). Combining SMMR and SSM/I data for time series analysis of central North American snow water equivalent. *Journal of hydrometeorology* **4**(2), 304–316.

Dietz, A., C. Kuenzer, U. Gessner et S. Dech (2012). Remote sensing of snow–a review of available methods. *International Journal of Remote Sensing* **33**(13), 4094–4134.

Diks, C. et J. Vrugt (2010). Comparison of point forecast accuracy of model averaging methods in hydrologic applications. *Stochastic Environmental Research and Risk Assessment* **24**(6), 809–820.

Draper, N. et H. Smith (1998). Applied regression analysis. *Wiley series in probability and statistics Show all parts in this series*.

Duan, Q., N. Ajami, X. Gao et S. Sorooshian (2007). Multi-model ensemble hydrologic prediction using Bayesian model averaging. *Advances in Water Resources* **30**(5), 1371–1386.

Evin, G., J. Merleau et L. Perreault (2011). Two-component mixtures of normal, gamma, and Gumbel distributions for hydrological applications. *Water Resources Research* **47**(8).

Evora, N., D. Tapsoba et D. De Seve (2008). Combining artificial neural network models, geostatistics, and passive microwave data for snow water equivalent retrieval and mapping. *Geoscience and Remote Sensing, IEEE Transactions on* **46**(7), 1925–1939.

Foster, J., A. Chang, D. Hall et A. Rango (1991). Derivation of snow water equivalent in boreal forests using microwave radiometry. *Arctic* **44**, 147–152.

Foster, J., D. Hall, J. Eylander, G. Riggs, S. Nghiem, M. Tedesco, E. Kim, P. Montesano, R. Kelly, K. Casey *et al.* (2011). A blended global snow product using visible, passive microwave and scatterometer satellite data. *International journal of remote sensing* **32**(5), 1371–1395.

Foster, J., C. Sun, J. Walker, R. Kelly, A. Chang, J. Dong et H. Powell (2005). Quantifying the uncertainty in passive microwave snow water equivalent observations. *Remote Sensing of environment* **94**(2), 187–203.

Gamerman, D. et H. Lopes (2006). *Markov Chain Monte Carlo : Stochastic Simulation for Bayesian Inference, Second Edition*. Chapman & Hall/CRC Texts in Statistical Science.

Gao, Y., H. Xie, N. Lu, T. Yao et T. Liang (2010). Toward advanced daily cloud-free snow cover and snow water equivalent products from Terra–Aqua MODIS and Aqua AMSR-E measurements. *Journal of Hydrology* **385**(1), 23–35.

Geary, R. (1947). Testing for normality. *Biometrika* **34**(3/4), 209–242.

Gelman, A., J. Carlin, H. Stern et D. Rubin (2003). *Bayesian Data Analysis, Second Edition*. Chapman & Hall/CRC Texts in Statistical Science.

Ghosh, J., M. Delampady et T. Samanta (2007). *An introduction to Bayesian analysis : theory and methods*. Springer.

Goïta, K., A. Walker et B. Goodison (2003). Algorithm development for the estimation of snow water equivalent in the boreal forest using passive microwave data. *International Journal of Remote Sensing* **24**(5), 1097–1102.

Hachem, S. (2008). *Cartographie des températures de surface, des indices de gel et de dégel et de la répartition spatiale du pergélisol à l'aide du Moderate Resolution Imaging Spectroradiometer (MODIS)*. Thèse.

Hall, D., J. Foster, S. Kumar, J. Chien et G. Riggs (2012). Improving the accuracy of the AFWA-NASA (ANSA) blended snow-cover product over the Lower Great Lakes region. *Hydrology and Earth System Sciences Discussions* **9**(1), 1141–1161.

Hall, D., J. Foster, V. Salomonson, A. Klein et J. Chien (2001). Development of a technique to assess snow-cover mapping errors from space. *Geoscience and Remote Sensing, IEEE Transactions on* **39**(2), 432–438.

Hall, D., J. Foster, D. Verbyla, A. Klein et C. Benson (1998). Assessment of snow-cover mapping accuracy in a variety of vegetation-cover densities in central alaska. *Remote sensing of Environment* **66**(2), 129–137.

Hall, D. et G. Riggs (2007). Accuracy assessment of the MODIS snow products. *Hydrological Processes* **21**(12), 1534–1547.

Hall, D., G. Riggs et V. Salomonson (1995). Development of methods for mapping global snow cover using moderate resolution imaging spectroradiometer data. *Remote sensing of Environment* **54**(2), 127–140.

Hall, D., G. Riggs, V. Salomonson, N. DiGirolamo et K. Bayr (2002). MODIS snow-cover products. *Remote sensing of Environment* **83**(1), 181–194.

Helfrich, S., D. McNamara, B. Ramsay, T. Baldwin et T. Kasheta (2007). Enhancements to, and forthcoming developments in the Interactive Multisensor Snow and Ice Mapping System (IMS). *Hydrological processes* **21**(12), 1576–1586.

Hoeting, J., D. Madigan, A. Raftery et C. Volinsky (1999). Bayesian model averaging : a tutorial (with comments by M. Clyde, David Draper and E. I. George), and a rejoinder by the authors. *Statistical Science* **14**(4), 382–417.

Kaufman, Y., R. Kleidman, D. Hall, J. Martins et J. Barton (2002). Remote sensing of subpixel snow cover using 0.66 and 2.1 μm channels. *Geophysical Research Letters* **29**(16), 28–1.

Kelly, R., A. Chang, L. Tsang et J. Foster (2003). A prototype AMSR-E global snow area and snow depth algorithm. *Geoscience and Remote Sensing, IEEE Transactions on* **41**(2), 230–242.

Klein, A. et A. Barnett (2003). Validation of daily MODIS snow cover maps of the upper rio grande river basin for the 2000–2001 snow year. *Remote Sensing of Environment* **86**(2), 162–176.

Kurvonen, L., J. Pulliainen et M. Hallikainen (1998). Monitoring of boreal forests with multitemporal special sensor microwave imager data. *Radio Science* **33**(3), 731–744.

Liang, T., X. Zhang, H. Xie, C. Wu, Q. Feng, X. Huang et Q. Chen (2008). Toward improved daily snow cover mapping with advanced combination of MODIS and AMSR-E measurements. *Remote Sensing of Environment* **112**(10), 3750–3761.

Luojus, K., J. Pulliainen, M. Takala, C. Derksen, H. Rott, T. Nagler, R. Solberg, A. Wiesmann, S. Metsämäki, E. Malnes *et al.* (2010). Investigating the feasibility of the GlobSnow snow water equivalent data for climate research purposes. In *Geoscience and Remote Sensing Symposium (IGARSS), 2010 IEEE International*. IEEE.

Luojus, K., J. Pulliainen, M. Takala, J. Lemmetyinen, M. Kangwa, T. Smolander et C. Derksen (2013). Algorithm Theoretical Basis Document - SWE-algorithm.

Merleau, J. et L. Bibeau (2013). Modélisation statistique de données multidimensionnelles avec des applications pour les études de crue statistique. Rapport de recherche IREQ-2013-0127.

Metsämäki, S., S. Anttila, H. Markus et J. Vepsäläinen (2005). A feasible method for fractional snow cover mapping in boreal zone based on a reflectance model. *Remote sensing of Environment* **95**(1), 77–95.

Metsämäki, S., M. Salminen, J. Pulliainen, K. Luojus, T. Nagler, G. Bippus, R. Solberg, A. Salberg, O. Due Trier et A. Wiesmann (2014). Algorithm theoretical basis document - SE-algorithm.

Montgomery, D., E. Peck et G. Vining (2006). *Introduction to Linear Regression Analysis*. Wiley Series in Probability and Statistics. Wiley.

Painter, T., K. Rittger, C. McKenzie, P. Slaughter, R. Davis et J. Dozier (2009). Retrieval of subpixel snow covered area, grain size, and albedo from MODIS. *Remote Sensing of Environment* **113**(4), 868–879.

Perreault, L. (2013). Vérification des prévisions hydrologiques probabilistes - version 2. Rapport de recherche IREQ-2013-019.

Perreault, L., B. Bobée, R. Roy et L. Mathier (1995). La combinaison de modèles appliquée à la validation en temps réel des apports naturels aux réservoirs hydriques. *Canadian Journal of Civil Engineering* **22**(5), 934–944.

Perreault, L. et J. Merleau (2014). Mélanges de lois normales, gamma et gumbel : application à des séries de pointes et volumes de crues printanières. Rapport de recherche (IREQ).

Pulliainen, J. et M. Hallikainen (2001). Retrieval of regional snow water equivalent from space-borne passive microwave observations. *Remote sensing of environment* **75**(1), 76–85.

Qobilov, T., F. Pertziger, L. Vasilina et M. Baumgartner (2001). Operational technology for snow-cover mapping in the Central Asian mountains using NOAA-AVHRR data. *IAHS PUBLICATION*, 76–80.

Ramsay, B. (1998). The interactive multisensor snow and ice mapping system. *Hydrological Processes* **12**(10), 1537–1546.

Ratcliffe, J. (1968). The effect on the t distribution of non-normality in the sampled population. *Applied Statistics* **17**(1), 42–48.

Rittger, K., T. Painter et J. Dozier (2013). Assessment of methods for mapping snow cover from MODIS. *Advances in Water Resources* **51**, 367–380.

Roberge, S. (2013). Développement d'un algorithme satellitaire à seuils évolutifs pour le suivi de l'étendue spatiale du couvert nival au Québec-Labrador et son adaptation à l'estimation d'ensemble. Mémoire de maîtrise, Université du Québec, INRS-Eau Terre et Environnement.

Roberge, S. (2014). Étude de la variabilité spatiotemporelle de l'étendue spatiale du couvert nival au Québec-Labrador à partir des données du capteur optique NOAA-AVHRR. Examen doctoral, Université du Québec, INRS-Eau Terre et Environnement.

Robert, C. (1998). *Discretization and MCMC convergence assessment*, Volume 135. Springer Science & Business Media.

Robert, C. (2007). *The Bayesian Choice : From Decision-Theoretic Foundations to Computational Implementation*. Springer Texts in Statistics. Springer.

Romanov, P., G. Gutman et I. Csiszar (2000). Automated monitoring of snow cover over North America with multispectral satellite data. *Journal of Applied Meteorology* **39**(11), 1866–1880.

Rosenthal, W. et J. Dozier (1996). Automated mapping of montane snow cover at subpixel resolution from the Landsat Thematic Mapper. *Water Resources Research* **32**(1), 115–130.

Royer, R., K. Goïta, J. Kohn et D. De Sève (2010). Monitoring dry, wet, and no-snow conditions from microwave satellite observations. *Geoscience and Remote Sensing Letters, IEEE* **7**(4), 670–674.

Salomonson, V. et I. Appel (2004). Estimating fractional snow cover from MODIS using the normalized difference snow index. *Remote sensing of environment* **89**(3), 351–360.

Salomonson, V. et I. Appel (2006). Development of the Aqua MODIS NDSI fractional snow cover algorithm and validation results. *Geoscience and Remote Sensing, IEEE Transactions on* **44**(7), 1747–1756.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics* **6**(2), 461–464.

Simic, A., R. Fernandes, R. Brown, P. Romanov et W. Park (2004). Validation of VEGETATION, MODIS, and GOES+ SSM/I snow-cover products over canada based on surface snow depth observations. *Hydrological Processes* **18**(6), 1089–1104.

Simpson, J. et T. McIntire (2001). A recurrent neural network classifier for improved retrievals of areal extent of snow cover. *IEEE Transactions on Geoscience and Remote Sensing* **39**(10), 2135–2147.

Slater, M., D. Sloggett, W. Rees et A. Steel (1999). Potential operational multi-satellite sensor mapping of snow cover in maritime sub-polar regions. *International Journal of Remote Sensing* **20**(15-16), 3019–3030.

Solberg, R., B. Wangensteen, J. Amlien, H. Koren, S. Metsämäki, T. Nagler, K. Luojus et J. Pulliainen (2010). A new global snow extent product based on ATSR-2 and AATSR. In *Geoscience and Remote Sensing Symposium (IGARSS), 2010 IEEE International*, pp. 1780–1783. IEEE.

Takala, M., J. Pulliainen, M. Huttunen et M. Hallikainen (2008). Detecting the onset of snow-melt using SSM/I data and the self-organizing map. *International Journal of Remote Sensing* **29**(3), 755–766.

Tedesco, M., J. Pulliainen, M. Takala, M. Hallikainen et P. Pampaloni (2004). Artificial neural network-based techniques for the retrieval of SWE and snow depth from SSM/I data. *Remote sensing of Environment* **90**(1), 76–85.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *the Annals of Statistics* **22**(4), 1701–1728.

Ulaby, F., R. Moore et A. Fung (1986). *Microwave remote sensing : Active and passive. Volume 3 - From theory to applications*.

Vachon, F., D. De Sève, Y. Choquette et F. Guay (2015). SWE monitoring during the winter and spring melt by combining microwaves remote sensing data, modeling and ground data. In *Geoscience and Remote Sensing Symposium (IGARSS), 2015 IEEE International*, pp. (à venir).

Vachon, F., K. Goïta, D. De Sève et R. Royer (2010). Inversion of a snow emission model calibrated with in situ data for snow water equivalent monitoring. *IEEE Transactions on Geoscience and Remote Sensing* **48**(1), 59–71.

Vikhamar, D. et R. Solberg (2003). Subpixel mapping of snow cover in forests by optical remote sensing. *Remote Sensing of Environment* **84**(1), 69–82.

Walker, A. et B. Goodison (1993). Discrimination of a wet snowcover using passive microwave satellite data. *Annals of Glaciology* **17**, 307–311.

Warren, S. (1982). Optical properties of snow. *Reviews of Geophysics* **20**(1), 67–89.

Weigel, A., M. Liniger et C. Appenzeller (2007). The discrete brier and ranked probability skill scores. *Monthly Weather Review* **135**(1), 118–124.

Weisberg, S. (2005). *Applied Linear Regression*. Wiley Series in Probability and Statistics. Wiley.

Welch, R., S. Sengupta, A. Goroch, P. Rabindra, N. Rangaraj et M. Navar (1992). Polar cloud and surface classification using AVHRR imagery : An intercomparison of methods. *Journal of Applied Meteorology* **31**(5), 405–420.

Wilks, D. (2011). *Statistical Methods in the Atmospheric Sciences*. Academic Press. Academic Press.

Annexe A

MODÉLISATION DU GTV PAR MÉLANGE DE LOIS NORMALES ET PAR RÉGRESSION LINÉAIRE

Dans cette annexe, le potentiel de discrimination neige/non neige du GTV utilisé seul est étudié à l'aide des mélanges de lois et de la régression linéaire.

A.1. Classification du GTV en neige/non-neige à l'aide du mélange de lois normales

Cette section tente d'établir le potentiel du GTV pour l'identification de la neige et de la non-neige à l'aide des mélanges de lois normales. Cet exercice permet de classifier les valeurs de GTV en présence/absence de neige au sol et de vérifier si ces classifications sont cohérentes avec la réalité.

A.1.1. Application à l'ensemble du domaine

La classification a d'abord été appliquée aux données GTV sur l'ensemble du domaine d'étude, peu importe les différentes zones bioclimatiques.

A.1.1.1. *Choix du nombre de composantes*

Comme mentionné à la section 2.1.4, l'utilisation des mélanges de lois nécessite d'effectuer un choix quant au nombre de composantes dans le mélange. Afin de vérifier quel est le nombre de composantes idéal suggéré par les données, une sélection de modèles a été effectuée à l'aide du critère de Schwarz (voir équation (2.1.3)) pour les mélanges à une, deux et trois composantes.

Puisque les données de GTV sont disponibles quotidiennement, la sélection de modèles a été réalisée pour chacune des journées de l'année 2011. Le graphique présenté à la figure A.1 illustre, pour chaque jour, le nombre de composantes optimal selon le critère de Schwarz. Pour l'année entière, c'est majoritairement

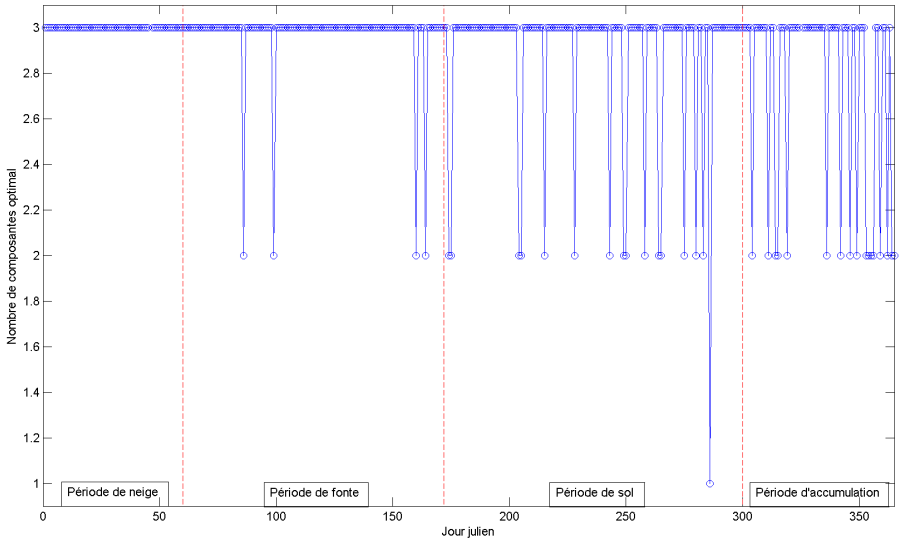

FIGURE A.1. Graphique du nombre de composantes optimal selon le critère de Schwarz pour le mélange de lois normales pour les données de 2011, Québec.

le mélange à trois composantes qui est considéré optimal (choisi dans 90% des cas). Toutefois, ces résultats sont étonnants. En effet, n'ayant que de la neige ou de la non-neige pendant les périodes de neige et de sol, les mélanges à une seule composante auraient dû y être privilégiés. Malgré cela, il est majoritairement optimal, au sens du critère de Schwarz, d'y utiliser un mélange à deux (8% des cas) ou trois composantes (90% des cas). Déjà à cette étape, il est possible de conclure que la classification ne donnera pas les résultats espérés lorsqu'il est censé n'y avoir que de la neige ou que du sol. Par contre, les résultats semblent raisonnables pour la période de fonte, qui est celle qui intéresse majoritairement les prévisionnistes d'Hydro-Québec. Effectivement, il est considéré logique qu'il puisse y avoir réellement deux (4% des cas pour la période) ou trois composantes (96% des cas pour la période), c'est-à-dire neige, non-neige et une composante de transition. C'est pourquoi les résultats de la classification en deux et trois composantes sont tout de même étudiés, mais uniquement pour cette période.

A.1.1.2. *Classification en deux composantes*

Lorsque le mélange de deux normales est appliqué pour la période de fonte, une probabilité de neige est obtenue pour chaque point de grille selon l'équation (2.1.2). Ces probabilités suivent un code de couleur sur une échelle de 0 à 1, où une faible probabilité de neige est bleue et une forte probabilité de neige est rouge. Ces valeurs sont ensuite utilisées pour construire une carte de probabilité

d'appartenance à la composante de neige, ce qui permet d'illustrer la répartition spatiale des probabilités de neige sur le territoire québécois. Par exemple, la figure A.2 montre une carte obtenue pour le 13 avril 2011.

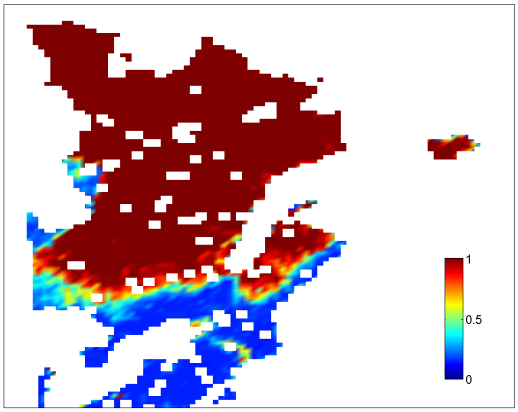

FIGURE A.2. Cartographie de la probabilité de neige obtenue avec le mélange de deux normales, 13 avril 2011, Québec. Les points de grille blanc correspondent à de l'eau ou à une absence de données SSM/I.

Sans nécessairement quantifier immédiatement la qualité de la cartographie obtenue, il est possible de constater que celle-ci semble satisfaisante. En effet, les régions de fortes et de faibles probabilités de neige sont cohérentes avec ce qui est réellement observé à ce temps de l'année. De plus, lorsque la cartographie est comparée avec la carte produite par l'algorithme de segmentation des données SSM/I (ensuite référé comme carte SSM/I) pour cette journée à la figure A.3 a), il est possible de voir que la séparation neige/non-neige de la carte de référence est assez similaire à la séparation forte/faible probabilité de neige de la carte produite par le mélange de lois. De plus, il est possible de constater que les classes obtenues sont cohérentes avec les valeurs de GTV de cette journée. En effet, lorsque la cartographie produite par l'utilisation du mélange de deux lois normales est comparée avec celle des valeurs de GTV à la figure A.3 b), il est possible de voir que la région correspondante à la faible probabilité de neige correspond à la région en orange et rouge de la carte des valeurs de GTV et que la région de forte probabilité de neige correspond au reste du Québec.

Toutefois, la cartographie n'est pas toujours satisfaisante. Effectivement, pour le 20 avril 2011, même si les classes obtenues par la classification avec mélange de deux normales (figure A.4) sont cohérentes avec la dispersion des valeurs de GTV dans le Québec (figure A.5 b)), la séparation forte/faible probabilité de neige est

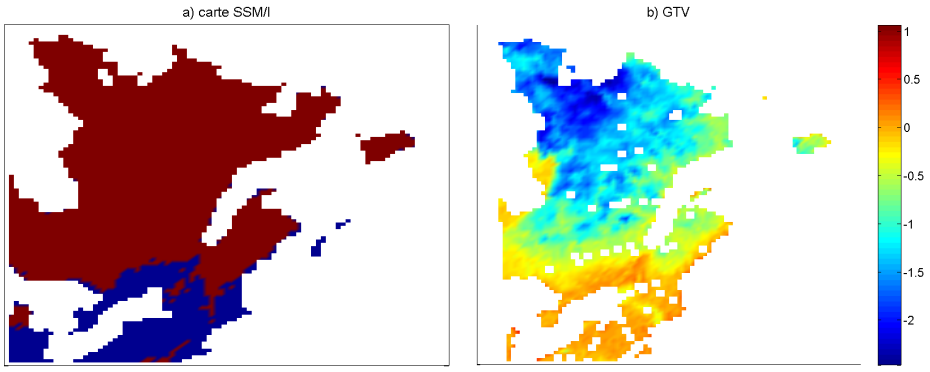

FIGURE A.3. a) carte SSM/I (bleu = non-neige, rouge = neige) et b) valeurs de GTV, 13 avril 2011, Québec

très différente de la séparation neige/non-neige de la carte SSM/I (figure A.5 a)). De plus, cette séparation n'est pas cohérente avec ce qui est généralement observé à ce temps de l'année.

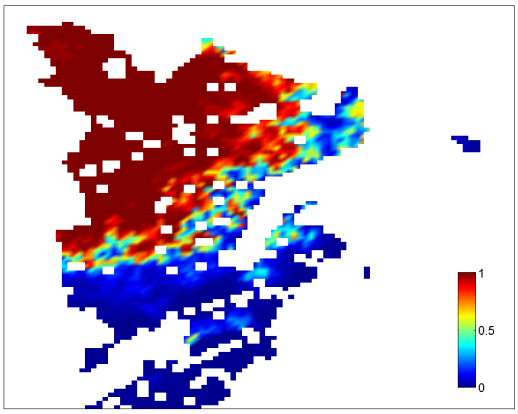

FIGURE A.4. Cartographie de la probabilité de neige obtenue avec le mélange de deux normales, 20 avril 2011, Québec

Puisque cette situation est assez fréquente (environ 67% des journées de la période de fonte), il a été décidé que la classification des données de GTV pour l'ensemble de territoire à l'aide des mélanges de deux lois normales ne satisfait pas les objectifs.

A.1.1.3. *Classification en trois composantes*

Suite aux résultats jugés insatisfaisants de l'utilisation des mélanges de deux lois normales, il a été décidé d'ajuster un mélange de trois normales pour chaque

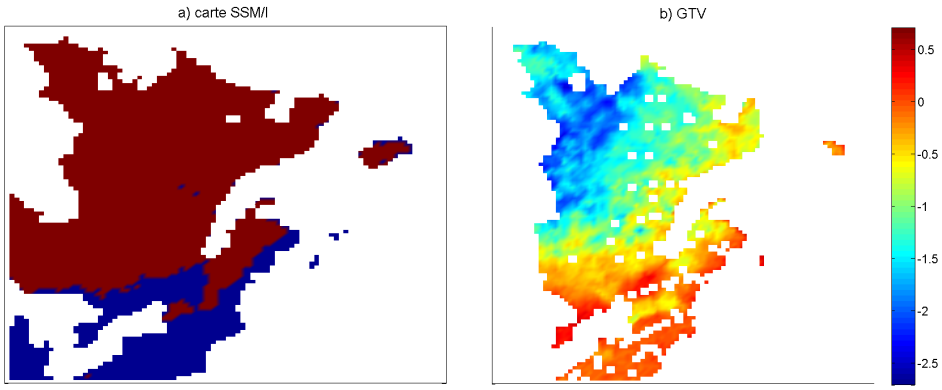

FIGURE A.5. a) carte SSM/I (bleu = non-neige, rouge = neige) et b) valeurs de GTV, 20 avril 2011, Québec

jour de la même période. Le but de cet exercice est de permettre l'identification d'une zone de transition entre la neige et le sol (non-neige) et ainsi avoir une séparation forte/faible probabilité de neige plus fidèle à la réalité.

Lorsque le mélange de trois normales est appliqué à la journée du 20 avril 2011, il est possible de constater que, cette fois-ci, en plus que les classes fournies par le mélange de lois (figure A.6) soient cohérentes avec les valeurs de GTV de la journée (figure A.5 b)), la séparation est similaire à celle fournie par la carte SSM/I (figure A.5 a)). En effet, la région de forte probabilité d'être une zone de transition semble bien séparer ce qui est du sol (au sud de la zone de transition) et ce qui est certainement de la neige (au nord de la zone de transition). La nature de la zone de transition n'est pas connue, mais il est supposé qu'il puisse s'agir de neige humide ou d'une forte végétation. De plus, la séparation neige/non-neige semble cohérente avec la réalité observée généralement à cette période de l'année.

Même si la spécification d'un mélange à trois composantes semble être une bonne solution dans certains cas, il y a des jours pour lesquels la troisième composante ne représente pas une zone de transition et donc pour lesquels la séparation neige/non-neige n'est toujours pas réaliste. Par exemple, c'est le cas pour le 24 avril 2011. En effet, même s'il est possible d'observer aux figures A.7 et A.8 b) que les classes obtenues avec le mélange de trois normales sont cohérentes avec les données GTV de la journée, il est aussi possible d'observer que la séparation neige/non-neige n'est pas similaire à celle de la carte SSM/I (figure A.8 a)). De plus, cette séparation n'est pas cohérente avec la réalité généralement observée durant cette période de l'année.

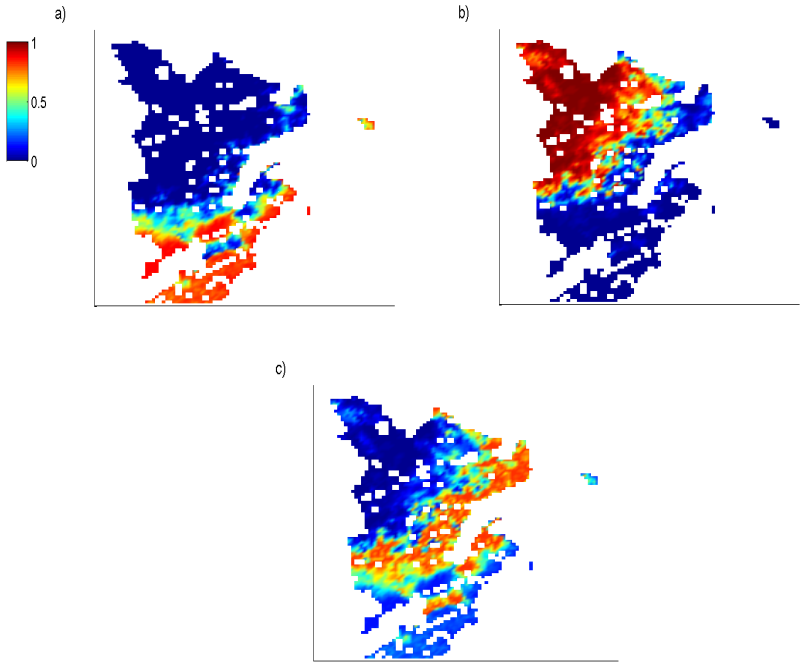

FIGURE A.6. Cartographie de la probabilité d'appartenir à chacune des composantes du mélange de trois normales, a) probabilité d'appartenir à la composante 1 (neige), b) probabilité d'appartenir à la composante 2 (non-neige), c) probabilité d'appartenir à la composante 3 (transition), 20 avril 2011, Québec

Puisque cela se produit pour environ 50% des jours de la période de fonte, l'application des mélanges de trois normales directement aux données de GTV pour l'ensemble du territoire est également considéré comme n'étant pas satisfaisant pour atteindre les objectifs.

A.1.2. Application à la zone 2

Jusqu'à présent, les mélanges de lois normales ont été appliqués aux valeurs de GTV pour l'intégralité du domaine d'étude. Cependant, il a été vu à la section 1.1 et à la section 1.3.2 que le Québec peut être divisé en quatre zones dans lesquelles les valeurs de GTV se comportent différemment. C'est pourquoi il peut être pertinent d'effectuer une classification différente pour chaque zone. Puisque la zone 2 est celle qui semble la plus intéressante pour les prévisionnistes d'Hydro-Québec, l'application présentée précédemment est reprise ici, mais uniquement pour les valeurs de GTV appartenant à celle-ci.

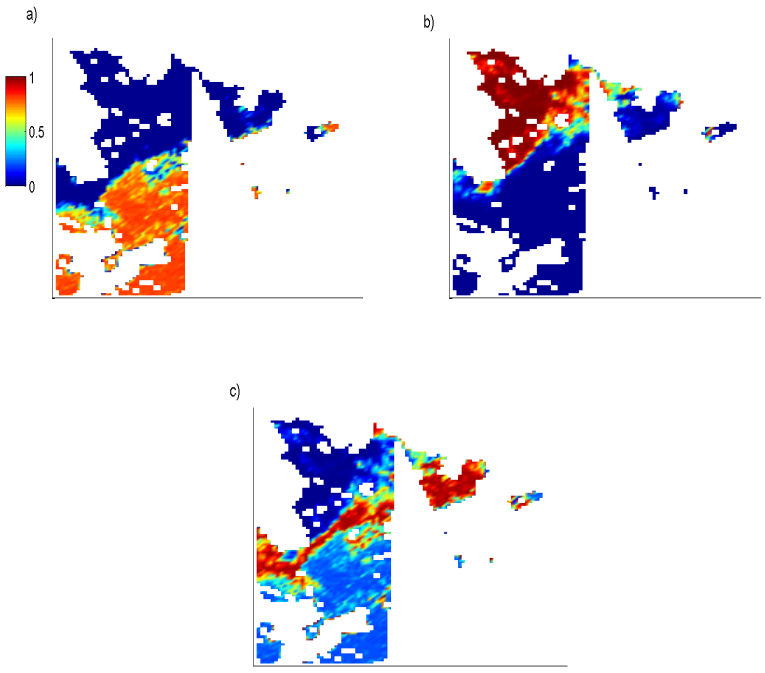

FIGURE A.7. Cartographie de la probabilité d'appartenir à chacune des composantes du mélange de trois normales, a) probabilité d'appartenir à la composante 1 (non-neige), b) probabilité d'appartenir à la composante 2 (neige), c) probabilité d'appartenir à la composante 3 (transition), 24 avril 2011, Québec

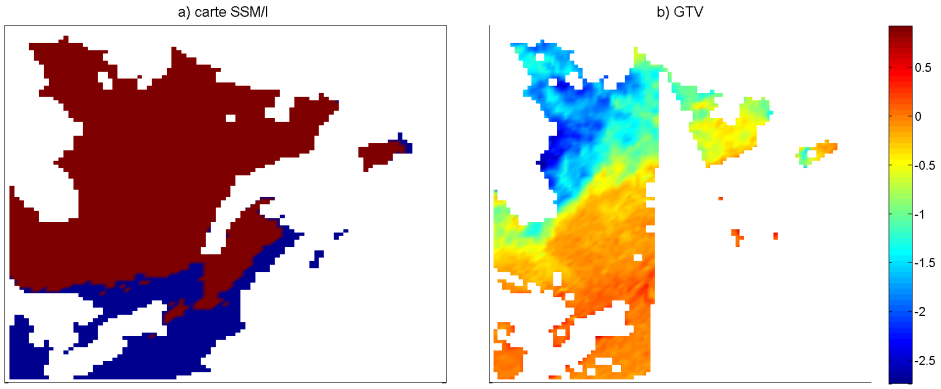

FIGURE A.8. a) carte SSM/I (bleu = non-neige, rouge = neige) et b) valeurs de GTV, 24 avril 2011, Québec

A.1.2.1. *Choix du nombre de composantes*

Tout comme le cas incluant le Québec en entier, il faut tout d'abord effectuer une sélection de modèles afin de déterminer quel est le nombre de composantes

optimal pour le mélange de lois. Le graphique présenté à la figure A.9 illustre ce nombre selon le critère de Schwarz pour chaque jour de l'année 2011 pour la zone 2 du Québec.

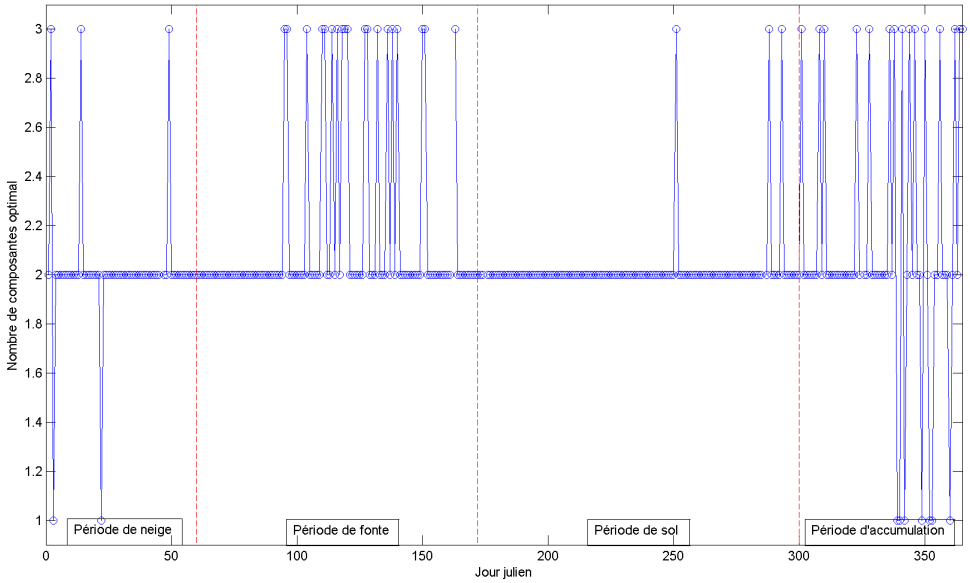

FIGURE A.9. Graphique du nombre de composantes optimal selon le critère de Schwarz pour le mélange de lois pour les données de 2011, Québec zone 2.

Dans ce cas-ci, c'est majoritairement le mélange de deux normales qui est optimal (choisi dans 86% de tous les jours) pour l'année entière, sauf pour quelques exceptions. Encore une fois, le résultat obtenu n'est pas celui attendu, c'est-à-dire que ce n'est pas le mélange à une seule composante qui est retenu pour les périodes de neige et de sol seulement. En effet, pour ces périodes, c'est plutôt le mélange à deux composantes qui est choisi dans 96% des cas et celui à trois composantes qui est choisi dans 3% des cas. La classification est donc encore uniquement effectuée pour les données de la période de fonte, car l'hypothèse de deux ou trois sous-populations est réaliste (choisi dans respectivement 83% et 17% des cas pour la période de fonte).

A.1.3. Classification en deux et trois composantes

Comme il a été vu aux sections A.1.1.2 et A.1.1.3, ni la classification à deux ni trois composantes n'offraient des résultats satisfaisants, car il y avait trop de cartographies pour lesquelles les résultats ne correspondaient pas du tout à ce

qui était espéré. Les mêmes conclusions sont également obtenues en ne s'intéressant qu'à la zone 2. En effet, environ 70% et 85% des cartographies produites respectivement par les mélanges de deux et trois normales pour la période de fonte n'offrent pas une séparation neige/non-neige qui est similaire à la réalité généralement observée ou aux cartes SSM/I.

A.1.4. Conclusion partielle

Suite à tous ces essais, il est clair que le GTV seul ne permet pas de discriminer la neige et la non-neige de façon satisfaisante avec une méthode de classification comme les mélanges de lois standards. En effet, autant pour les mélanges à deux ou trois composantes et autant pour le domaine d'étude en entier que pour seulement la zone 2, il y a un pourcentage trop élevé de cas pour lesquels la séparation obtenue n'est pas similaire à celle de leur carte SSM/I respective ou à la séparation généralement observée à cette période de l'année.

L'exercice n'a pas été répété pour les zones restantes (1, 3 et 4). En effet, il n'a pas été jugé pertinent de s'intéresser aux autres zones pour un souci d'homogénéité.

Ayant observé que les classifications obtenues étaient toujours cohérentes avec les valeurs de GTV des journées respectives, peu importe si le produit concordait ou non avec la carte de référence, le fait que les résultats soient jugés insatisfaisants n'est pas dû à la performance du modèle de mélange de lois. Il semble plutôt que le problème soit dû au fait que les classes obtenues ne représentent pas ce qui est recherché, c'est-à-dire la neige et le sol.

Les raisons pouvant expliquer les mauvaises séparations en neige/non-neige ne sont pas connues, mais des hypothèses ont été formulées. Il est possible que les mélanges de lois doivent être utilisés différemment, c'est-à-dire utilisés conjointement avec une autre méthode statistique ou avec d'autres données. Aussi, l'information contenue dans la variable GTV n'est possiblement pas suffisante. Il faut donc ajouter des variables explicatives telles la température, les degrés-jours et les degrés-chauffes.

A.2. Étude de la relation entre le GTV et des variables exogènes via la régression linéaire

Ayant conclu à la section précédente que le GTV seul ne permet pas d'identifier la neige et le sol, la présente section étudie la relation entre le GTV et des variables exogènes à l'aide de la régression linéaire.

L'idée ici est de supposer que les valeurs de GTV journalières peuvent être exprimées comme une combinaison linéaire de certaines variables explicatives $(\boldsymbol{x}_{t,j}, t = 1, \ldots, 365)$, où les coefficients $\beta_{tj}, j = 1, \ldots, p$ et les coefficients de détermination quotidiens R_t^2 (voir équation (2.2.2.1)) changent de jour en jour et où p est le nombre de variables explicatives dans le modèle. Ainsi, il serait possible de déterminer quelles variables peuvent expliquer le comportement du GTV et si elles l'expliquent toujours aussi bien tout au long de l'année. Il est donc supposé que :

$$\mathbf{GTV}_t \sim \mathcal{N}(\boldsymbol{X}_t^T \boldsymbol{\beta}_t, \sigma_t^2)$$

où $\mathbf{GTV}_t$, $\boldsymbol{X}_t$, $\boldsymbol{\beta}_t$ et σ_t^2 sont respectivement les valeurs de GTV, les valeurs des variables explicatives, les coefficients de régression et la variance (supposée constante pour une journée donnée) associés à la journée $t = 1, \ldots, 365$.

Les variables exogènes considérées sont les valeurs de GTV par point de grille de la journée précédente ($\mathbf{GTV}_{t-1}$), la température minimale, moyenne et maximale par point de grille de la journée t ($\mathbf{Tmin}_t$, $\mathbf{Tmoy}_t$ et $\mathbf{Tmax}_t$), le degré-jour par point de grille de la journée t calculés sur cinq jours ($\mathbf{DJ5}_t$) et le degré-chauffe par point de grille de la journée t calculés aussi sur cinq jours ($\mathbf{DC5}_t$) (voir section 1.3.3). Puisqu'il est ardu d'effectuer une sélection de variables avec ces variables pour chacune des journées (il y aurait 365 sélections de modèles à effectuer), une première approche a été d'effectuer une régression linéaire simple pour chacune des variables explicatives pour chaque jour. Le seuil de signification utilisé est $\alpha = 0,05$ (5%).

A.2.1. GTV de la veille ($\mathbf{GTV}_{t-1}$)

La première variable explicative considérée est la valeur de GTV par point de grille de la journée $t - 1$:

$$\mathbf{GTV}_t = \beta_{0,t} + \beta_{1,t}\mathbf{GTV}_{t-1} + \boldsymbol{\varepsilon}_t.$$

La figure A.10 illustre les variations du coefficient de régression $\beta_{1,t}$ pour $\mathbf{GTV}_{t-1}$, du R^2 et de la valeur-p du test-t pour le coefficient de $\mathbf{GTV}_{t-1}$ qui ont été calculés avec le modèle sur l'ensemble des points de grille. Tout d'abord, il est possible de remarquer que le coefficient de régression pour la variable du $\mathbf{GTV}_{t-1}$ semble relativement constant avec des valeurs variant majoritairement entre 0,5 et 1, ce qui implique qu'une augmentation/diminution d'une unité du GTV de la veille occasionne une augmentation/diminution d'environ une demie à une unité du GTV de la journée actuelle. Aussi, le GTV de la veille explique relativement bien le GTV de la journée actuelle pour les périodes de neige et d'accumulation de neige (c'est-à-dire dès qu'il y a généralement un peu de neige) étant donné que

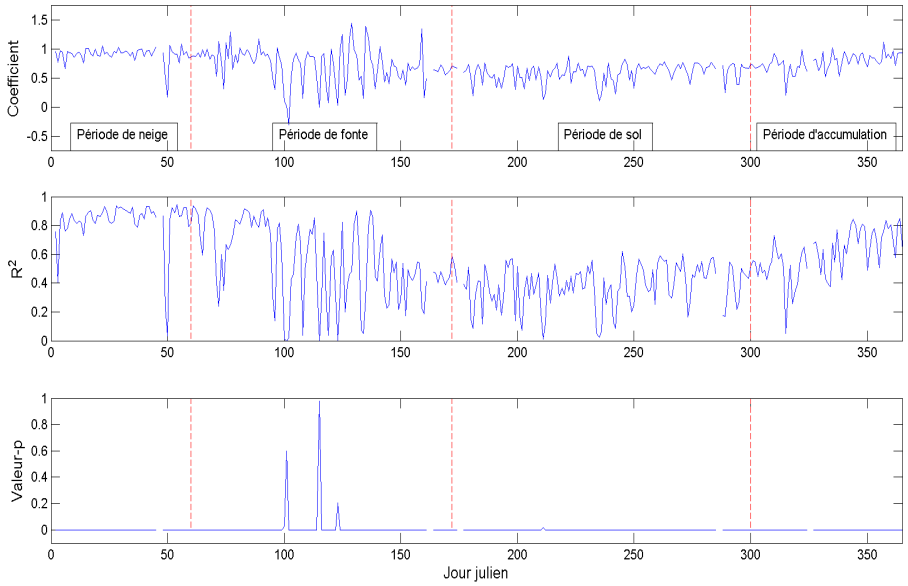

FIGURE A.10. Graphique du coefficient de régression, du R^2 et de la valeur-p pour la variable $\mathbf{GTV}_{t-1}$ par jour julien, zone 2 du Québec, 2011.

la valeur du R^2 y est souvent entre 0,5 et 1 et que le coefficient de régression est pratiquement toujours statistiquement significatif (à 5%). Toutefois, la relation est moins forte lorsqu'il y a moins de neige sur la zone, c'est-à-dire lors des périodes de fonte de neige et de sol, puisque le R^2 y est plus faible que pour le reste de l'année. La relation entre les valeurs de GTV de la veille et celles de la journée actuelle est donc plus pertinente lorsqu'il y a de la neige que lorsqu'il n'y en a pas.

Même si cette variable explicative semble apporter de l'information pertinente à propos du GTV, il n'est pas approprié de l'utiliser pour construire un modèle de régression linéaire simple. En effet, les hypothèses nécessaires ne sont pas toujours respectées. L'examen des graphiques des résidus standardisés en fonction des valeurs ajustées et les diagrammes quantile-quantile des résidus standardisés pour chaque jour montre que les résidus ne sont pas toujours uniformément distribués autour de 0 en ordonnée, qu'ils n'ont pas toujours une variance constante et qu'ils ne suivent pas toujours une distribution approximativement normale. Des exemples de ces comportements sont illustrés à la figure A.11.

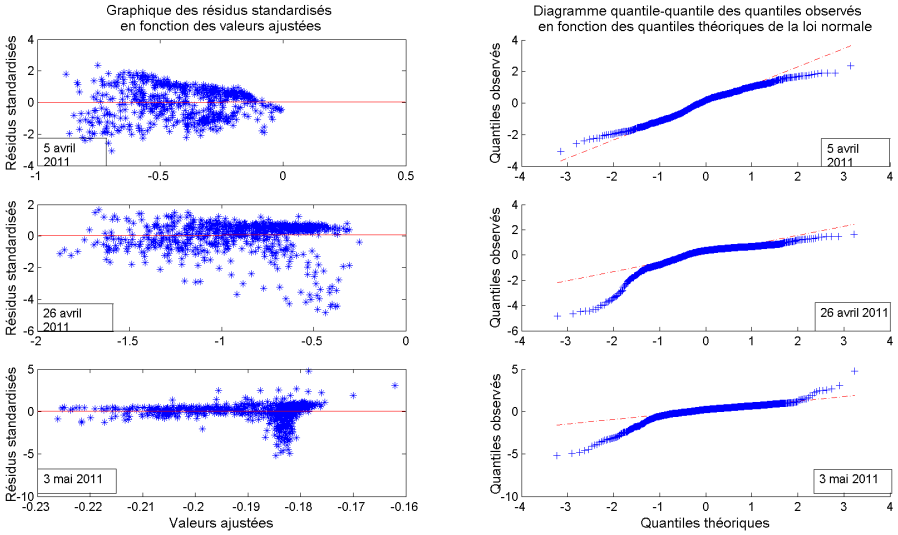

FIGURE A.11. Exemples de jour pour la variable $\mathbf{GTV}_{t-1}$ où le graphique des résidus standardisés en fonction des valeurs ajustées et le diagramme quantile-quantile illustrent un diagnostic des résidus non satisfaisant, zone 2 du Québec, 2011.

A.2.2. Température minimum, moyenne et maximum de la journée t ($\mathbf{Tmin}_t$, $\mathbf{Tmoy}_t$ et $\mathbf{Tmax}_t$), degrés-jours et degrés-chauffes sur 5 jours ($\mathbf{DJ5}_t$ et $\mathbf{DC5}_t$)

Les prochaines variables considérées par point de grille sont la température minimum, moyenne et maximum de la journée t ($\mathbf{Tmin}_t$, $\mathbf{Tmoy}_t$ et $\mathbf{Tmax}_t$) ainsi que les degrés-jours et les degrés-chauffes calculés sur cinq jours. Le choix d'utiliser $x = 5$ pour le calcul de (1.3.2) et (1.3.3) a été effectué par les experts de l'Institut de recherche d'Hydro-Québec. Selon la qualité des résultats pour le calcul avec cinq jours, il pourrait potentiellement être pertinent d'essayer un nombre de jours différents.

Les figures A.12 et A.13 illustrent les variations des coefficients de régression respectivement pour $\mathbf{Tmin}_t$, $\mathbf{Tmoy}_t$ et $\mathbf{Tmax}_t$, $\mathbf{DJ5}_t$, $\mathbf{DC5}_t$, du R^2 et de la valeur-p du test-t pour chacun des coefficients qui ont été calculés avec leur modèle respectif. Cette fois-ci, il est possible de constater que les coefficients de régression pour les différentes variables sont moins constants que ceux pour la variable $\mathbf{GTV}_{t-1}$, que les variables sont peu utiles pour expliquer le GTV de la journée t (R^2 majoritairement entre 0 et 0,5) et que les coefficients sont plus souvent non significatifs (à 5%). De plus, le diagnostic des résidus n'est pas satisfaisant puisqu'il y a plusieurs journées pour lesquelles les résidus standardisés

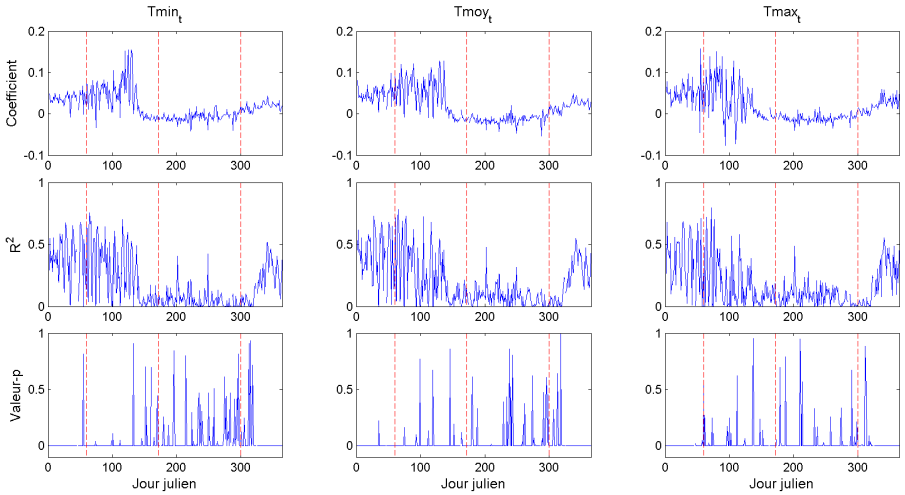

FIGURE A.12. Graphique du coefficient de régression, du R^2 et de la valeur-p du test-t pour les variables $\mathbf{Tmin}_t$, $\mathbf{Tmoy}_t$ et $\mathbf{Tmax}_t$ par jour julien, zone 2 du Québec, 2011.

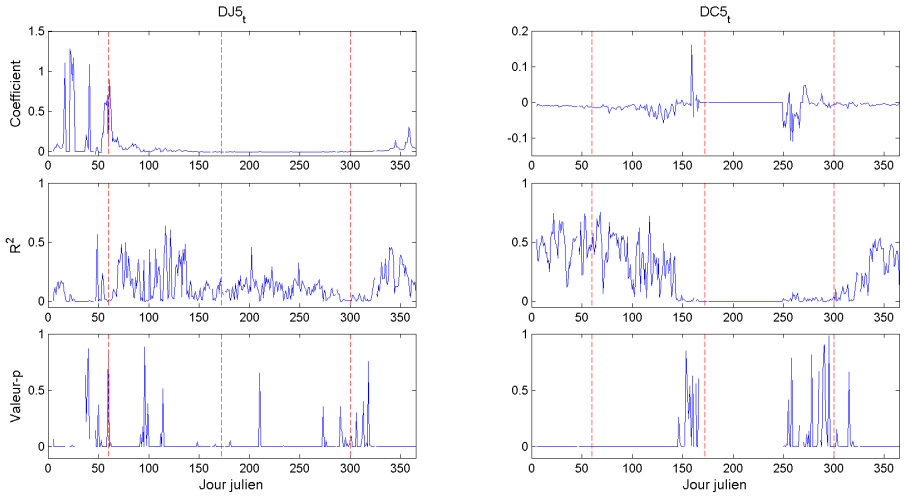

FIGURE A.13. Graphique du coefficient de régression, du R^2 et de la valeur-p pour les variables $\mathbf{DJ5}_t$ et $\mathbf{DC5}_t$ par jour julien, zone 2 du Québec, 2011.

en fonction des valeurs ajustées ne sont pas uniformément distribués autour de 0 en ordonnée et les résidus standardisés ne sont pas distribués approximativement selon la loi normale (voir exemples aux figures A.14 à A.18).

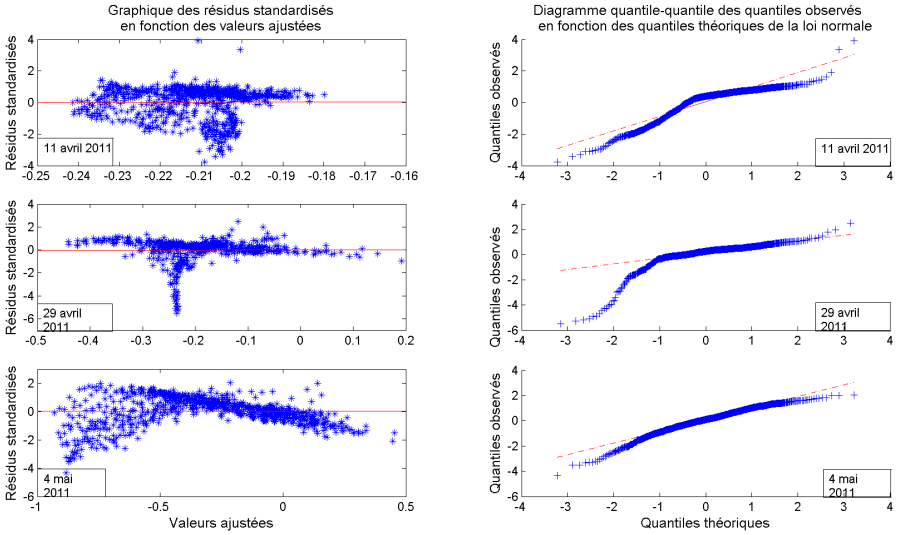

FIGURE A.14. Exemples de jour pour la variable $\mathbf{Tmin}_t$ où le graphique des résidus standardisés en fonction des valeurs ajustées et le diagramme quantile-quantile illustrent un diagnostic des résidus non satisfaisant, zone 2 du Québec, 2011.

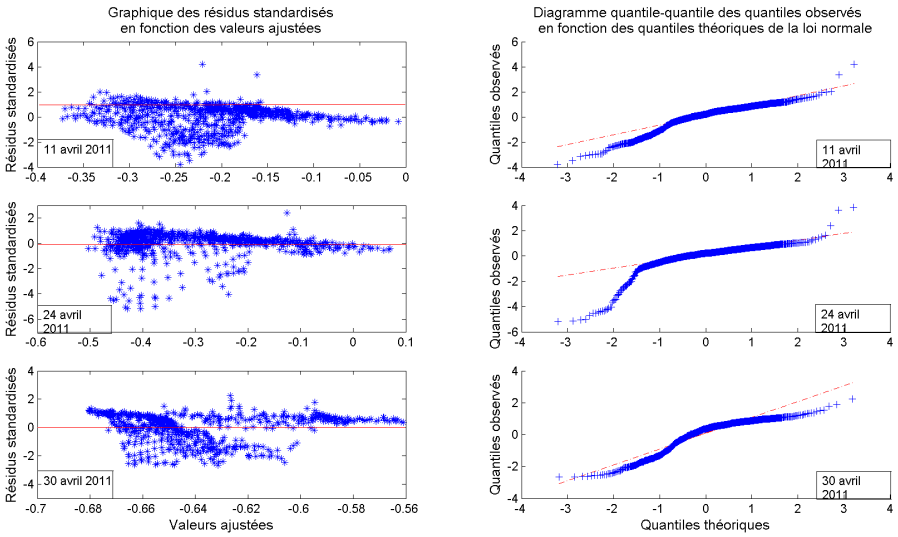

FIGURE A.15. Exemples de jour pour la variable $\mathbf{Tmoy}_t$ où le graphique des résidus standardisés en fonction des valeurs ajustées et le diagramme quantile-quantile illustrent un diagnostic des résidus non satisfaisant, zone 2 du Québec, 2011.

Suite à ces différents essais, il est conclu que la modélisation du GTV à l'aide de la régression linéaire n'aidera pas à la compréhension de son comportement. En effet, même si certaines variables étaient significatives et que certains R^2 étaient

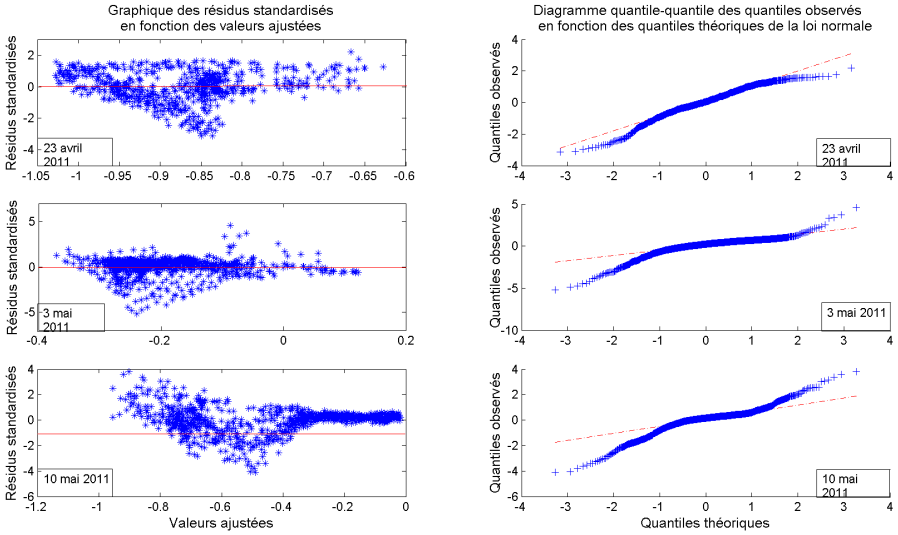

FIGURE A.16. Exemples de jour pour la variable $\mathbf{Tmax}_t$ où le graphique des résidus standardisés en fonction des valeurs ajustées et le diagramme quantile-quantile illustrent un diagnostic des résidus non satisfaisant, zone 2 du Québec, 2011.

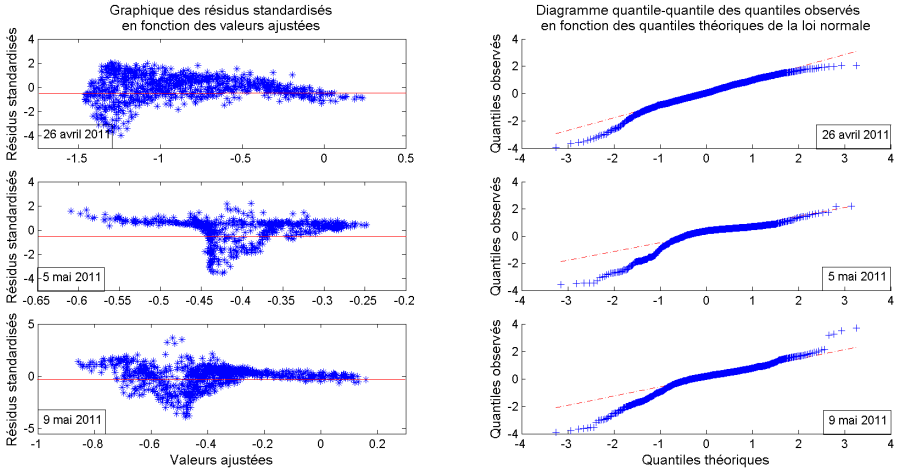

FIGURE A.17. Exemples de jour pour la variable $\mathbf{DJ5}_t$ où le graphique des résidus standardisés en fonction des valeurs ajustées et le diagramme quantile-quantile illustrent un diagnostic des résidus non satisfaisant, zone 2 du Québec, 2011.

suffisamment élevés (par exemple pour la variable $\mathbf{GTV}_{t-1}$), les diagnostics de résidus n'étaient pas tous suffisamment satisfaisants afin que les hypothèses nécessaires au modèle linéaire soient respectées. Aussi, dû aux résultats non concluants pour les variables de degrés-jours et de degrés-chauffes calculées sur cinq jours,

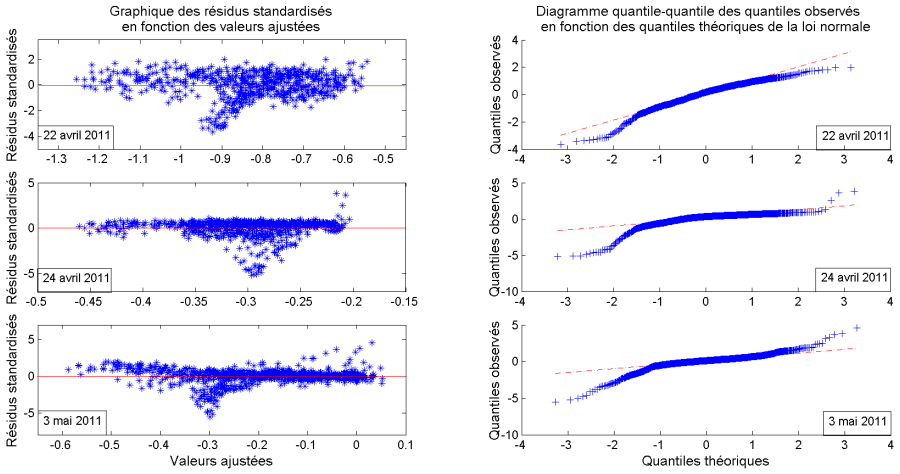

FIGURE A.18. Exemples de jour pour la variable $\mathbf{DC5}_t$ où le graphique des résidus standardisés en fonction des valeurs ajustées et le diagramme quantile-quantile illustrent un diagnostic des résidus non satisfaisant, zone 2 du Québec, 2011.

d'autres valeurs de x (nombre de jours) n'ont pas été considérées. Les variables qui ont été étudiées dans cette section ne seront donc pas utilisées pour expliquer le GTV. Par conséquent, des modèles avec plus d'une variable explicative n'ont pas été jugés pertinents pour l'étude.

A.3. CONCLUSION PARTIELLE

Dans cette annexe, le potentiel du GTV à discriminer la neige et la non-neige a été étudié et son comportement a été examiné. Pour l'étude du potentiel de discrimination, les mélanges de lois ont été utilisés sur la période de fonte afin de classifier les valeurs de GTV en neige/non-neige en quantifiant l'incertitude de classification à l'aide de probabilités d'appartenance aux différentes composantes (neige, non-neige et transition). Les cartographies obtenues à partir de cette approche de classification ne sont pas assez satisfaisantes, donc le GTV seul ne permet pas de bien identifier la neige et le sol.

La classification basée uniquement sur le GTV n'ayant pas donné les résultats escomptés, l'ajout de variables explicatives a été considéré via la régression linéaire. Le comportement quotidien des valeurs de GTV a donc été étudié en effectuant des régressions linéaires simples selon différentes variables explicatives. Certaines variables étaient plus pertinentes que d'autres, par exemple les valeurs de GTV de la veille. Cependant, peu importe celle considérée, les diagnostics des

résidus n'étaient pas satisfaisants quant aux hypothèses à respecter pour l'utilisation d'un modèle de régression linéaire.

En conclusion, les valeurs de GTV ne donnent pas les résultats voulus lorsqu'elles sont utilisées seules, mais il est possible que l'utilisation de variables exogènes puisse être un complément d'information pertinent pour combler les lacunes du GTV.

Annexe B

MODÉLISATION DE SEUILS DYNAMIQUES À L'AIDE DE LA RÉGRESSION LINÉAIRE MULTIPLE

Ayant vu à l'annexe A que le GTV ne pouvait pas être utilisé tel quel et seul afin de discriminer la neige/non-neige, la présente annexe présente une approche modélisant des seuils dynamiques (calculés à partir des valeurs du GTV) avec des variables exogènes via la régression linéaire multiple.

B.1. Calcul du seuil dynamique quotidien

L'algorithme développé pour le calcul du seuil dynamique nécessite la recherche d'un zéro approximatif pour une fonction $f(\cdot)$ par calcul numérique. La méthode utilisée est *regula falsi*, aussi connu sous le nom de méthode de la fausse position, et est définie comme suit (voir Burden et Faires, 2010) :

(a) initialiser deux valeurs, a et b avec $f(a) \times f(b) < 0$, qui délimitent l'intervalle dans lequel se trouve le 0 ;

(b) calculer :

$$c = a - \frac{a-b}{f(a)-f(b)} \times f(a),$$

et $f(c)$;

(c) si $f(a) \times f(c) < 0$, poser $b = a$. Sinon, b conserve sa valeur ;

(d) poser $a = c$;

(e) répéter jusqu'à ce que le critère d'arrêt $|c-a| < \delta$ soit respecté. La dernière valeur de c est la valeur approximative du zéro recherché pour la fonction $f(\cdot)$.

L'algorithme utilisé pour obtenir des seuils dynamiques dépend des valeurs p_t calculées à la section 3.1 (voir équation (3.1.1)) et peut être exprimé en plusieurs étapes que voici :

(1) ayant p_t, calculer pour chaque valeur de t :

$$p_{t2} = \min(p_t, 1 - p_t),$$

où p_{t2} est la probabilité de la composante la moins probable. Elle devrait être proche de 0 durant les périodes de neige et de sol ;

(2) faire la moyenne :

$$m = \frac{\sum_{t=1}^{365} p_{t2}\mathbf{1}(p_{t2} \text{ n'est pas une valeur manquante})}{n},$$

où

$$n = \sum_{t=1}^{365} \mathbf{1}(p_{t2} \text{ n'est pas une valeur manquante}).$$

L'intérêt de faire la moyenne de p_{t_2} est d'approximer la valeur de la probabilité à partir de laquelle une composante est considérée absente. Autrement dit, si la probabilité de neige (respectivement la probabilité de non-neige) est inférieure à m, alors il n'y a pas de neige (respectivement non-neige) ;

(3) trouver q_t tel que $P(y_{ti} > q_t) = m$ en résolvant avec la méthode *regula falsi* l'équation

$$f(s) = \widehat{w}_1 \times \Phi\left(\sqrt{\widehat{\phi}_{1t}} \times \left(s - \widehat{\zeta}_{1t}\right)\right) + \widehat{w}_2 \times \Phi\left(\sqrt{\widehat{\phi}_{2t}} \times \left(s - \widehat{\zeta}_{2t}\right)\right) - m,$$

en utilisant respectivement le minimum et le maximum des valeurs de GTV de la journée t pour les valeurs de a et b ainsi que la valeur $\delta = 1 \times 10^{-10}$ comme critère d'arrêt. Le zéro approximatif trouvé à la dernière étape représente le seuil q_t pour la journée t. La variable q_t représente le seuil journalier pour lequel la probabilité de neige est considérée nulle selon le seuil m.

B.2. Comparaison des seuils pour la zone 2 avec des seuils journaliers observés historiquement

Tout d'abord, l'algorithme décrit à la section précédente est utilisé pour calculer la probabilité de neige journalière dans la zone 2 du Québec. Si les résultats sont concluants, la méthodologie sera répétée pour les autres zones. Afin de vérifier si les seuils dynamiques calculés semblent réalistes, ils ont été comparés avec

des seuils de neige journaliers historiques calculés par les experts de l'Institut de recherche d'Hydro-Québec.

La figure B.1 illustre la comparaison entre les valeurs calculées et les valeurs

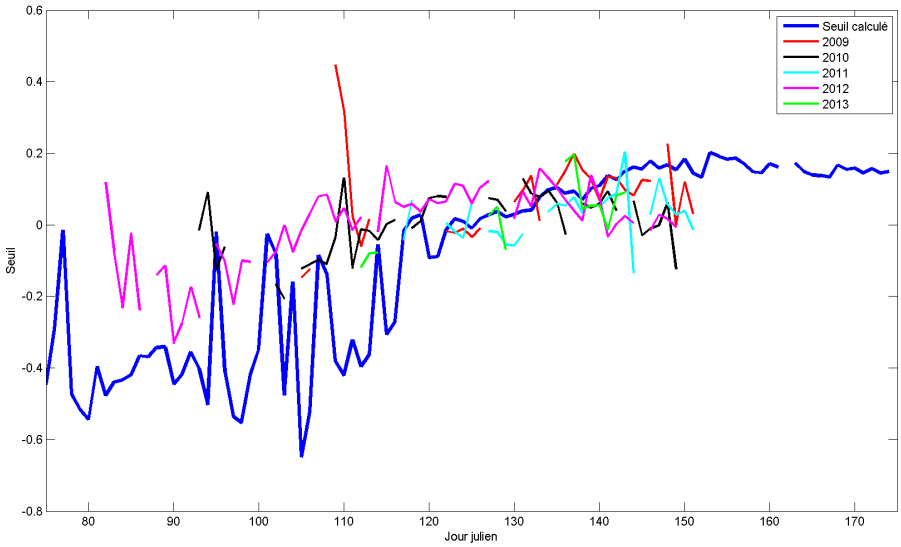

FIGURE B.1. Comparaison par jour julien des seuils calculés selon l'algorithme proposé pour la zone 2 du Québec pour l'année 2011 avec les seuils historiques pour 2009 à 2013.

ponctuelles historiques des jours juliens entre 80 et 155 pour les années 2009 à 2013. Il est possible de remarquer que les seuils calculés allant jusqu'au jour julien 130 sont plus faibles que les seuils historiques, puis ils sont plus élevés par la suite. Cela peut porter à croire qu'ils ne sont pas similaires. Cependant, lorsque l'écart-type du calcul des valeurs historiques (calculés par les chercheurs l'Institut de recherche d'Hydro-Québec) est intégré comme incertitude à l'aide d'un intervalle ayant une demi-longueur de deux écarts-types (figure B.2), il est possible de constater que les valeurs calculées sont souvent à l'intérieur des intervalles. Les seuils qui ont été calculés semblent donc cohérents avec les seuils historiques lorsque leur incertitude est prise en compte.

B.3. MODÉLISATION DES SEUILS CALCULÉS À L'AIDE DE VARIABLES EXPLICATIVES

Maintenant que des seuils dynamiques ont été calculés pour chaque jour de l'année 2011, il peut être intéressant de modéliser ces seuils en utilisant des variables explicatives à l'aide d'une régression linéaire multiple afin d'obtenir un

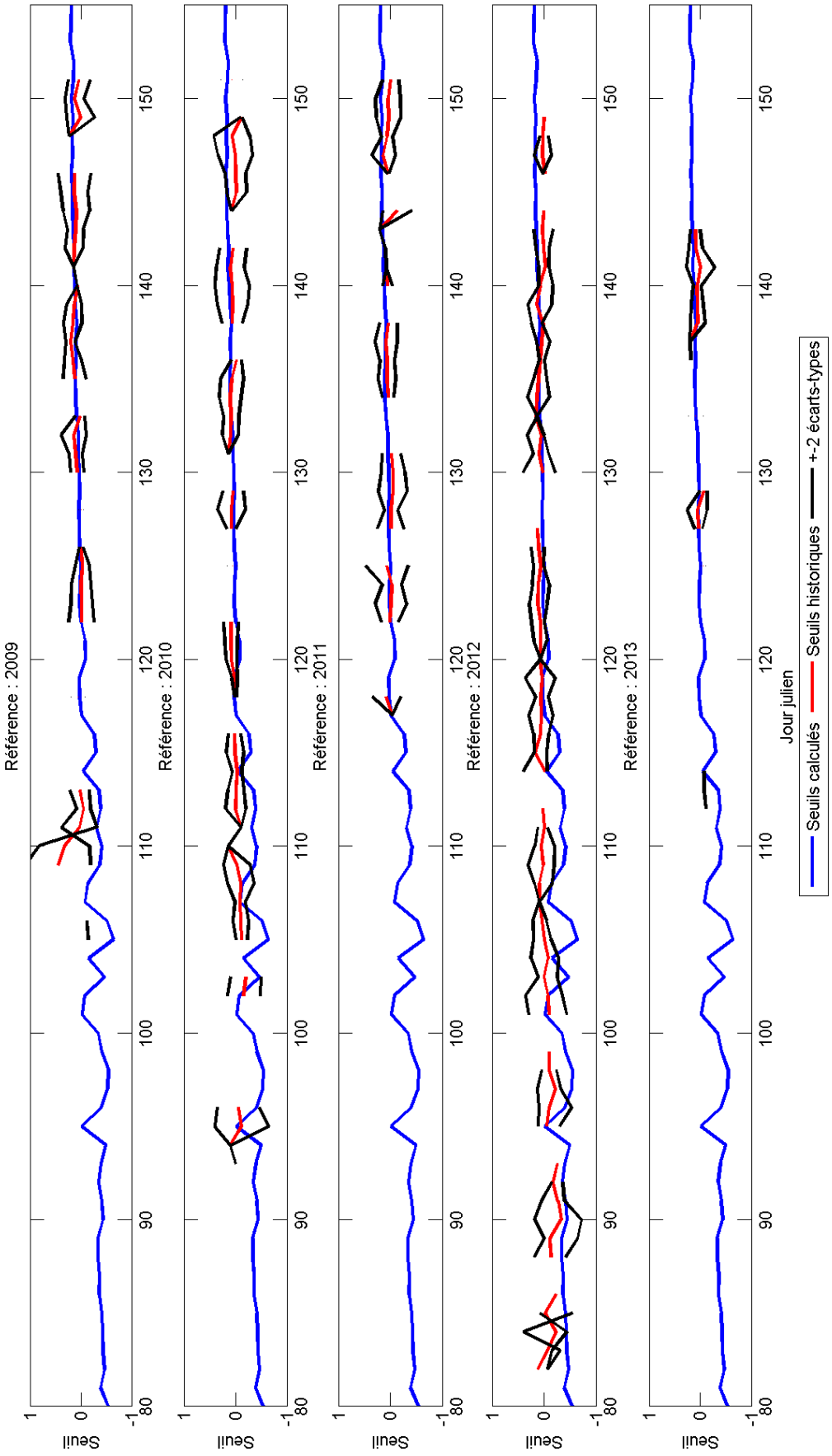

FIGURE B.2. Comparaison des seuils calculés pour la zone 2 du Québec pour l'année 2011 avec les seuils historiques et un intervalle de + ou - 2 écarts-types pour 2009 à 2013.

modèle de la forme (3.3.1). En effet, cela permettrait de tenir compte des facteurs exogènes. De plus, il serait éventuellement possible de calculer des seuils dynamiques pour différentes années, en faisant l'hypothèse que les mêmes facteurs exogènes puissent être utilisés.

B.3.1. Variables à considérer

De façon similaire à la modélisation de lp_t effectuée au chapitre 3, l'information des journées précédentes allant jusqu'à un délai de sept jours est utilisée comme variables explicatives. Pour cette section, il s'agit des seuils $\text{seuil}_{t-1}, \ldots, \text{seuil}_{t-7}$ calculées uniquement pour la zone 2. Aussi, l'information sur la température est intégrée en utilisant les variables Tmin, Tmoy ou Tmax.

Puisque les variables Tmin, Tmoy et Tmax fournissent de l'information journalière par point de grille, alors que la variable réponse seuil_t est une information journalière pour l'ensemble de la zone 2 (voir section 3.3.2), il a été décidé de résumer l'information à l'intérieur des points de grille pour une journée à l'aide de leur minimum, leur moyenne et leur maximum (Tminmin, Tminmoy, Tminmax, Tmoymin, Tmoymoy, Tmoymax, Tmaxmin, Tmaxmoy, Tmaxmax). Ces neuf variables étant très corrélées entre elles (voir figure 3.3) et pouvant donc induire un problème de multicolinéarité, il a fallu n'en conserver qu'une seule. Cet exercice a été effectué ici en calculant le coefficient de corrélation linéaire entre chacune des variables de température et la variable seuil_t pour les zones 1 à 4 (voir section 3.3.2).

Le contenu du tableau B.1, permet de conclure que :

- la variable Tminmin est celle qui est la plus corrélée aux seuils dynamiques des zones 1, 2 et 3 ;
- la variable Tminmoy est celle qui est la plus corrélée dans la zone 4 ;
- la variable Tminmin est celle qui est la deuxième plus corrélée avec les seuils calculés de la zone 4.

Du coup, la variable Tminmin a été retenue comme variable de température.

B.3.2. Sélection de variables explicatives

Comme pour le chapitre 3, la sélection de variables est réalisée avec la méthode à rebours jusqu'à l'obtention de $p^* = 2, 3, 4$ variables dans les modèles. Nous imposons la variable Tminmin lorsqu'elle est significative (valeur-p du test-t inférieure à 0,05). Les tableaux B.2 à B.8 montrent les différentes étapes de la sélection de modèle jusqu'à l'obtention des trois modèles potentiels qui seront confrontés lors de leur comparaison avec les cartes neige/non-neige de référence.

TABLE B.1. Corrélation entre les variables de température de chaque zone et les seuils calculés pour ces mêmes zones.

Variables	Description	Zone 1	Zone 2	Zone 3	Zone 4
Tminmin	Minimum sur les points de grille pour Tmin	0,6230	0,8004	0,8557	0,8455
Tminmoy	Moyenne sur les points de grille pour Tmin	0,6073	0,7853	0,8398	0,8478
Tminmax	Maximum sur les points de grille pour Tmin	0,5808	0,7721	0,7846	0,8116
Tmoymin	Minimum sur les points de grille pour Tmoy	0,5720	0,7554	0,8259	0,8316
Tmoymoy	Moyenne sur les points de grille pour Tmoy	0,5721	0,7482	0,8026	0,8329
Tmoymax	Maximum sur les points de grille pour Tmoy	0,5708	0,7489	0,7693	0,7754
Tmaxmin	Minimum sur les points de grille pour Tmax	0,5061	0,6945	0,7868	0,8046
Tmaxmoy	Moyenne sur les points de grille pour Tmax	0,5243	0,6960	0,7507	0,8086
Tmaxmax	Maximum sur les points de grille pour Tmax	0,5428	0,7036	0,7257	0,7294

À la suite de la sélection de variables, les différents modèles possibles sont :

$$\widehat{\text{seuil}}_t = 0{,}0358 + 0{,}7295\ \text{seuil}_{t-1} + 0{,}0038\ \text{Tminmin}\ (p^* = 2);$$

$$\widehat{\text{seuil}}_t = 0{,}0315 + 0{,}4456\ \text{seuil}_{t-1} + 0{,}3525\ \text{seuil}_{t-6} + 0{,}0033\ \text{Tminmin}\ (p^* = 3);$$

$$\widehat{\text{seuil}}_t = 0{,}0299 + 0{,}4164\ \text{seuil}_{t-1} + 0{,}1486\ \text{seuil}_{t-5} + 0{,}2526\ \text{seuil}_{t-6} + 0{,}0031\ Tminmin\ (p^* = 4).$$

B.3.3. Diagnostics des différents modèles

Disposant désormais de trois modèles potentiels pour expliquer les seuils calculés, il faut vérifier s'ils sont adéquats.

TABLE B.2. Étape 1 de la sélection de variables à rebours pour la régression linéaire sur les seuils calculés de la zone 2 pour l'année 2011, Québec. Variable seuil$_{t-4}$ à retirer.

Étape 1				
Variables	Estimation	Erreur standard	Statistique t	Valeur-p
origine	0,0297	0,0061	4,8490	<0,0001
seuil$_{t-1}$	0,3791	0,0552	6,8618	<0,0001
seuil$_{t-2}$	-0,0287	0,0581	-0,4945	0,6213
seuil$_{t-3}$	0,0536	0,0575	0,9312	0,3525
seuil$_{t-4}$	-0,0112	0,0574	-0,1954	0,8452
seuil$_{t-5}$	0,1502	0,0557	2,6967	0,0074
seuil$_{t-6}$	0,2066	0,0559	3,6928	0,0003
seuil$_{t-7}$	0,0792	0,0516	1,5340	0,1261
Tminmin	0,0031	0,0004	6,9761	<0,0001

TABLE B.3. Étape 2 de la sélection de variables à rebours pour la régression linéaire sur les seuils calculés de la zone 2 pour l'année 2011, Québec. Variable seuil$_{t-2}$ à retirer.

Étape 2				
Variables	Estimation	Erreur standard	Statistique t	Valeur-p
origine	0,0293	0,0061	4,8468	<0,0001
seuil$_{t-1}$	0,3708	0,0531	6,9770	<0,0001
seuil$_{t-2}$	-0,0250	0,0576	-0,4344	0,6643
seuil$_{t-3}$	0,0467	0,0501	0,9317	0,3522
seuil$_{t-5}$	0,1470	0,0489	3,0072	0,0028
seuil$_{t-6}$	0,2112	0,0554	3,8145	0,0002
seuil$_{t-7}$	0,0825	0,0511	1,6135	0,1076
Tminmin	0,0031	0,0004	6,9400	<0,0001

TABLE B.4. Étape 3 de la sélection de variables à rebours pour la régression linéaire sur les seuils calculés de la zone 2 pour l'année 2011, Québec. Variable seuil$_{t-3}$ à retirer.

Étape 3				
Variables	Estimation	Erreur standard	Statistique t	Valeur-p
origine	0,0296	0,0060	4,9586	<0,0001
seuil$_{t-1}$	0,3615	0,0472	7,6615	<0,0001
seuil$_{t-3}$	0,0353	0,0425	0,8309	0,4067
seuil$_{t-5}$	0,1464	0,0485	3,0206	0,0027
seuil$_{t-6}$	0,2091	0,0548	3,8175	0,0002
seuil$_{t-7}$	0,0784	0,0501	1,5638	0,1188
Tminmin	0,0031	0,0004	7,0442	<0,0001

Le tableau B.9 donne la valeur-p de chacune des régressions (voir équation (2.2.4)), leur R^2 ajusté (voir équation (2.2.6)) ainsi que le facteur d'inflation de

TABLE B.5. Étape 4 de la sélection de variables à rebours pour la régression linéaire sur les seuils calculés de la zone 2 pour l'année 2011, Québec. Variable seuil_{t-7} à retirer.

Étape 4				
Variables	Estimation	Erreur standard	Statistique t	Valeur-p
origine	0,0306	0,0061	5,0194	<0,0001
seuil_{t-1}	0,3914	0,0468	8,3712	<0,0001
seuil_{t-5}	0,1434	0,0475	3,0192	0,0027
seuil_{t-6}	0,2214	0,0564	3,9239	0,0001
seuil_{t-7}	0,0609	0,0507	1,2022	0,2302
Tminmin	0,0031	0,0004	6,9760	<0,0001

TABLE B.6. Étape 5 de la sélection de variables à rebours pour la régression linéaire sur les seuils calculés de la zone 2 pour l'année 2011, Québec. Variable seuil_{t-5} à retirer.

Étape 5 ($p^* = 4$)				
Variables	Estimation	Erreur standard	Statistique t	Valeur-p
origine	0,0299	0,0060	4,9655	1,0964e-06
seuil_{t-1}	0,4164	0,0419	9,9425	1,4452e-20
seuil_{t-5}	0,1486	0,0470	3,1615	0,0017
seuil_{t-6}	0,2526	0,0494	5,1166	5,2660e-07
Tminmin	0,0031	0,0004	6,9197	2,3307e-11

TABLE B.7. Étape 6 de la sélection de variables à rebours pour la régression linéaire sur les seuils calculés de la zone 2 pour l'année 2011, Québec. Variable seuil_{t-6} à retirer.

Étape 6 ($p^* = 3$)				
Variables	Estimation	Erreur standard	Statistique t	Valeur-p
origine	0,0315	0,0060	5,2548	2,6256e-07
seuil_{t-1}	0,4456	0,0410	10,8600	8,8780e-24
seuil_{t-6}	0,3525	0,0378	9,3208	1,5202e-18
Tminmin	0,0033	0,0004	7,4596	7,3469e-13

TABLE B.8. Étape 7 de la sélection de variables à rebours pour la régression linéaire sur les seuils calculés de la zone 2 pour l'année 2011, Québec. Modèle le plus parcimonieux.

Étape 7 ($p^* = 2$)				
Variables	Estimation	Erreur standard	Statistique t	Valeur-p
origine	0,0358	0,0066	5,4097	1,1824e-07
seuil_{t-1}	0,7295	0,0308	23,6840	2,5687e-74
Tminmin	0,0038	0,0005	7,9090	3,5507e-14

la variance (VIF) (voir équation (2.2.7)) pour chacune des variables explicatives.

TABLE B.9. Valeur-p de la régression des seuils, R^2 ajusté et facteur d'inflation de la variance (VIF) par variable pour chacun des modèles (p^*=2, 3 et 4).

Modèle	Valeur-p	R^2 ajusté	Variable	VIF
2	<0,0001	0,862	seuil_{t-1}	2,3807
			Tminmin	2,3807
3	<0,0001	0,888	seuil_{t-1}	5,2354
			seuil_{t-6}	4,4176
			Tminmin	2,4214
4	<0,0001	0,890	seuil_{t-1}	5,4659
			seuil_{t-5}	6,8046
			seuil_{t-6}	7,5238
			Tminmin	2,4866

Il est possible de constater que toutes les valeurs-p sont inférieures au seuil de signification fixé à 5%, donc chacune des régressions explique de façon significative les seuils calculés. Aussi, le R^2 est assez élevé (plus grand que 0,8), donc la variation des variables explicatives explique une forte proportion de la variation des seuils calculés. Aussi, les facteurs d'inflation de la variance sont inférieurs à dix, ce qui signifie qu'il n'y a pas de problème de multicolinéarité selon Montgomery *et al.* (2006) (voir section 2.2.2.2).

Par contre, la figure B.3 permet de constater que les résidus ne sont pas uniformément distribués autour de 0 en ordonnée, peu importe le modèle. En effet, bien que les résidus semblent centrés en zéro en ordonnée, il semble que les résidus n'aient pas une variance constante. De plus, la figure B.4 montre que les quantiles observés ne correspondent pas entièrement aux quantiles théoriques de la loi normale. Il semble donc que ces trois modèles de régression ne vérifient pas certaines hypothèses de la régression. Malgré cela, les cartographies binaires de neige/non-neige pour chacun des modèles sont produites afin de voir si elles sont cohérentes avec la réalité. Dans le cas d'une telle éventualité, les modèles de régression pourraient être améliorés afin de respecter les hypothèses de la régression. Aussi, il faudrait développer une façon de quantifier l'incertitude des cartographies. Dans le cas contraire, il serait préférable d'utiliser une approche différente.

B.4. PRODUCTION DES CARTOGRAPHIES DE NEIGE ET NON-NEIGE

Afin de pouvoir cartographier la neige quotidiennement, il faut calculer les seuils journaliers selon chacun des modèles. Pour ce faire, deux façons distinctes (similaires à celle du chapitre 3) peuvent être employées. La première façon

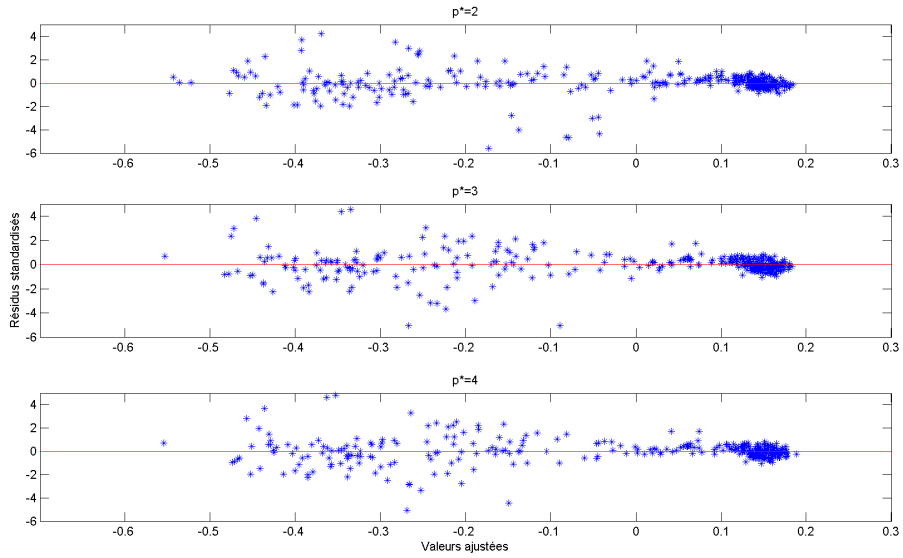

FIGURE B.3. Graphique des résidus standardisés en fonction de valeurs ajustées pour la régression des seuils calculés avec deux, trois et quatre variables explicatives, zone 2 du Québec, 2011.

.

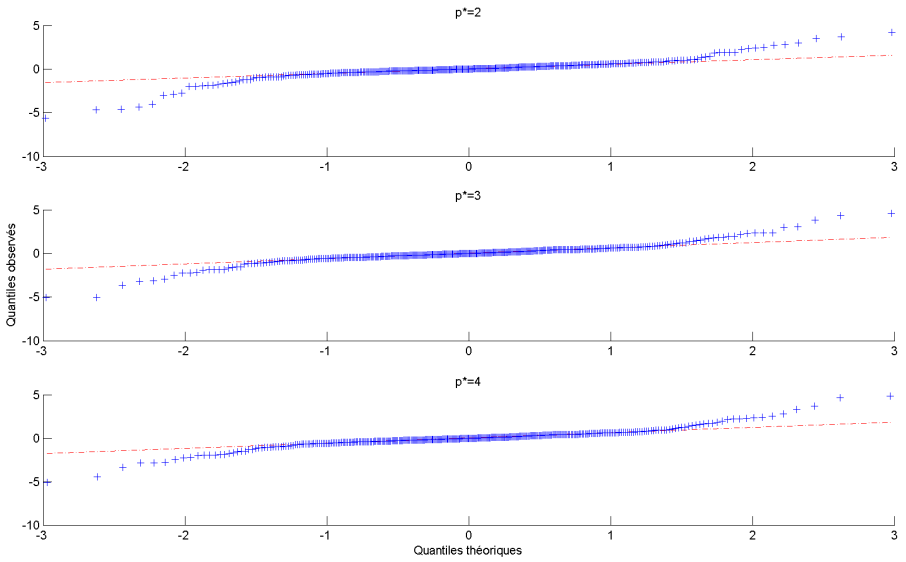

FIGURE B.4. Diagramme quantile-quantile des résidus standardisés pour la régression des seuils calculés avec deux, trois et quatre variables explicatives, zone 2 du Québec, 2011.

.

consiste à calculer le seuil de la journée t selon les seuils calculés des journées

précédentes et selon la valeur du minimum de Tmin pour cette journée (Tminmin). Par exemple, pour le modèle avec deux variables explicatives, le seuil de la journée t est calculé de la façon suivante :

$$\widehat{\text{seuil}}_t = 0{,}0358 + 0{,}7295 \times \text{seuil}_{t-1} + 0{,}0038 \times \text{Tminmin}_t, \tag{B.4.1}$$

où les coefficients sont ceux qui ont été calculés à l'étape 7 de la sélection de variables au tableau B.8. La variable seuil_t est le seuil pour la journée t, seuil_{t-1} est le seuil pour la journée $t-1$ et Tminmin_t est le minimum des Tmin pour tous les points de grille de la journée t. Cette façon de faire fournit un seuil pour la journée t et les données GTV de cette journée peuvent être transformées en neige/non-neige en les comparant à ce seuil : si elles sont inférieures au seuil il s'agit de neige, sinon il s'agit de sol.

Bien que ce ne soit pas la façon la plus optimale de procéder (voir section 3.3.5), la seconde approche consiste à calculer le seuil de la journée t individuellement pour chaque point de grille. Pour ce faire, l'idée est de leur appliquer un par un le modèle (B.4.1). De cette façon, le fait que les seuils peuvent changer selon l'emplacement dans le Québec peut être pris en considération. Chacun des seuils est calculé selon les valeurs des seuils des journées précédentes à l'intérieur du même point de grille et selon la valeur de Tmin pour ce point de grille. Par exemple, si le modèle avec deux variables explicatives est considéré, le seuil du point de grille i pour la journée t correspond à :

$$\widehat{\text{seuil}}_{t,i} = 0{,}0358 + 0{,}7295 \times \text{seuil}_{t-1,i} + 0{,}0038 \times \text{Tmin}_{t,i},$$

où $\text{seuil}_{t,i}$ est la valeur du seuil pour le point de grille i et la journée t, $\text{seuil}_{t-1,i}$ est le seuil du même point de grille i à la journée $t-1$ et $\text{Tmin}_{t,i}$ est la valeur de Tmin dans le point de grille i et à la journée t. Cette approche fournit donc un seuil par point de grille par jour. Les cartes de neige/non-neige sont obtenues en confrontant la valeur du GTV d'un point de grille pour la journée t avec la valeur calculée du seuil pour la même journée et le même point de grille : si la valeur de GTV d'un point de grille est inférieure au seuil qui lui est associé, alors il s'agit de neige, sinon c'est du sol.

Une légère problématique associée à ces deux approches est qu'il faut connaître la valeur de certains seuils pour débuter les calculs (voir section 3.3.5). En effet, pour calculer seuil_t (respectivement $\text{seuil}_{t,i}$) il faut que seuil_{t-1} (respectivement $\text{seuil}_{t-1,i}$), seuil_{t-5} (respectivement $\text{seuil}_{t-5,i}$) et seuil_{t-6} (respectivement $\text{seuil}_{t-6,i}$) aient déjà été calculés selon le modèle utilisé. Pour la première approche, les valeurs seuil_1 à seuil_6 ont été initialisées avec la valeur calculée pour ces journées selon l'algorithme de la section B.1. Ainsi, la valeur des seuils pour les autres

journées de l'année peut être calculée de façon récursive. Pour la seconde méthode, il faut initialiser tous les points de grilles des journées $t = 1, 2, 3, 4, 5, 6$. Pour ce faire, la même valeur que pour le seuil calculé selon l'algorithme de la section B.1 a été attribuée à chaque point de grille de la journée t.

L'examen des figures B.5 et B.6 montre que les cartographies réalisées pour chaque modèle et selon les deux différentes approches n'ont pas une répartition neige/non-neige qui est cohérente avec la réalité. En effet, la quantité de neige est souvent très forte peu importe la période de l'année, alors qu'elle serait supposée diminuer durant la période de fonte (représentée ici par le 30 avril) pour augmenter par la suite durant la période d'accumulation de neige (représentée ici par le 10 décembre).

De plus, les cartographies ont des comportements étranges à quelques reprises. Il y a notamment des grands écarts entre la quantité de neige d'une journée à une autre. Par exemple, aux figures B.7 et B.8, il y a une journée pour laquelle il y a beaucoup de sol, alors que toutes les journées n'en ont pas.

B.5. Conclusion partielle

Dans cette annexe, des seuils dynamiques ont été modélisés afin de catégoriser les valeurs de GTV en neige/non-neige. Tout d'abord, le calcul d'une série de seuils pour l'année 2011 a été présenté et il a été conclu qu'elle était assez similaire à des séries annuelles de seuils dynamiques calculés par des experts de l'Institut de recherche d'Hydro-Québec. Ces seuils ont donc été modélisés à l'aide de la régression linéaire multiple afin de développer un modèle tenant compte de variables exogènes. Trois modèles parcimonieux ont été considérés, mais les diagnostics de leurs résidus ne satisfaisaient pas les hypothèses pour l'utilisation d'un modèle de régression linéaire. Malgré cela, les cartographies neige/non-neige ont été produites de deux façons différentes, mais la répartition de neige/non-neige n'était pas représentative de la réalité. Cette approche ne permet donc pas de répondre à l'objectif de discriminer les valeurs de GTV en neige et non-neige.

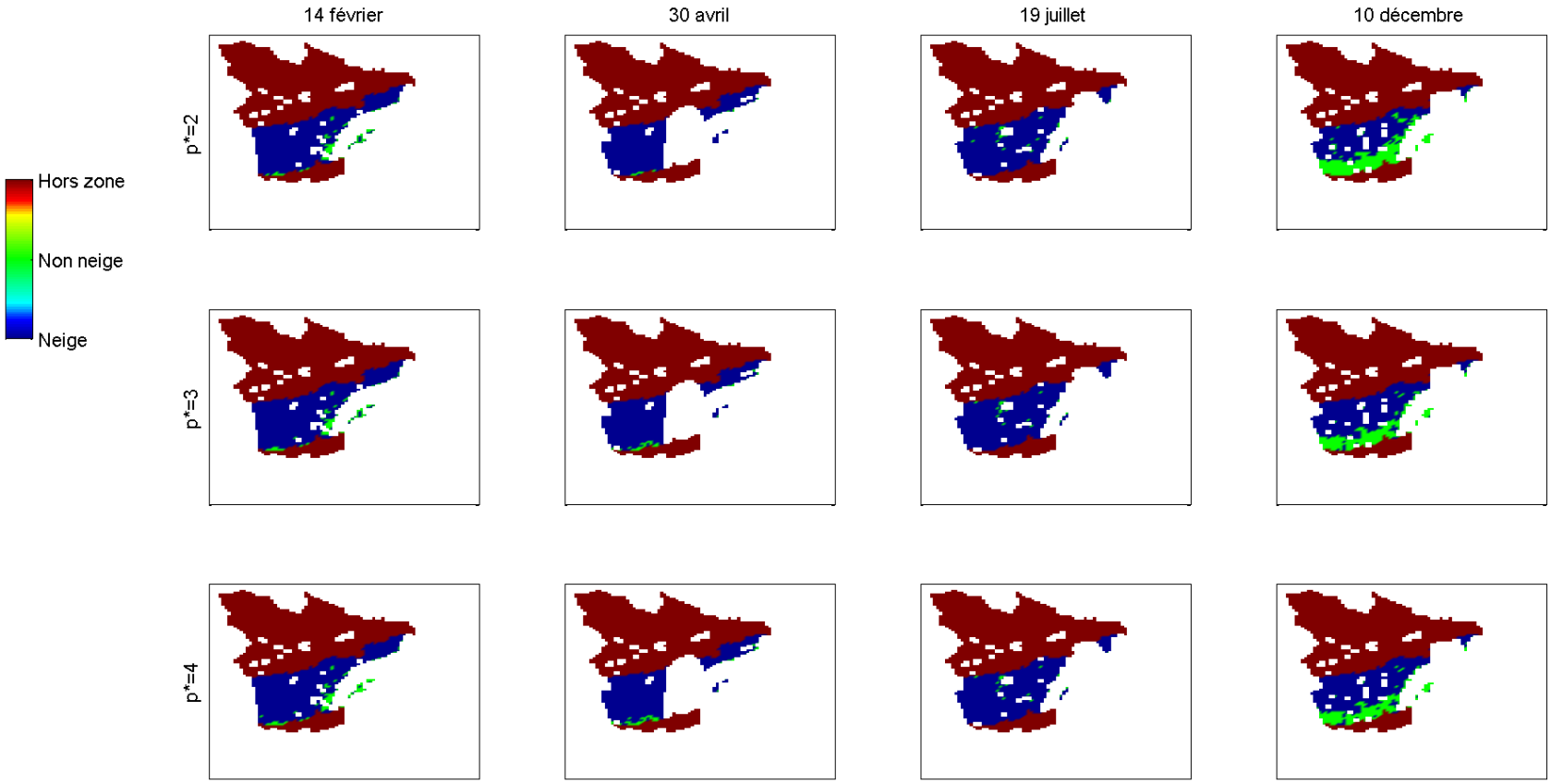

FIGURE B.5. Exemple de cartographies avec un seuil par jour pour deux, trois et quatre variables explicatives où la méthode indique presque toujours de la neige, peu importe le moment de l'année, zone 2 du Québec, 2011.

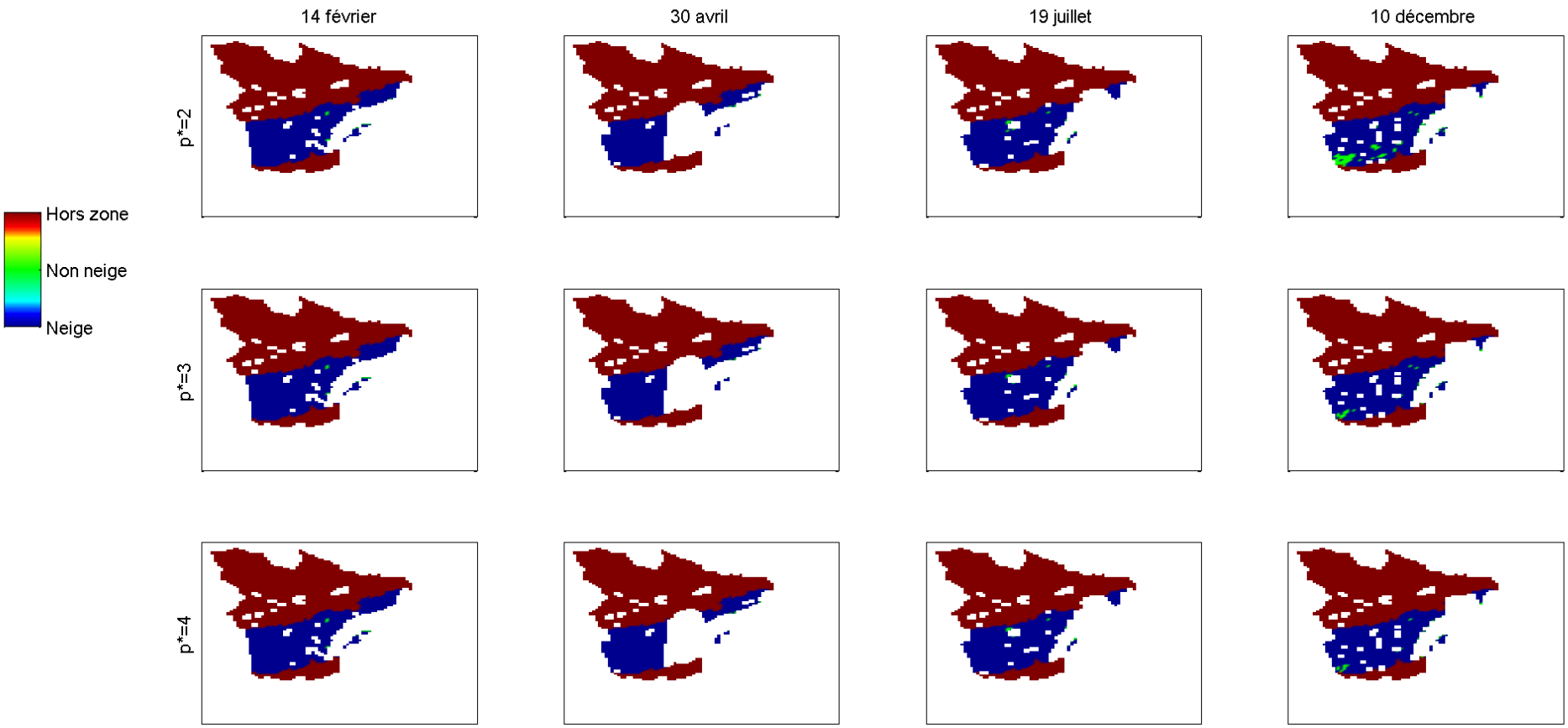

FIGURE B.6. Exemple de cartographies avec un seuil par point de grille pour deux, trois et quatre variables explicatives où la méthode indique de la neige, peu importe le moment de l'année, zone 2 du Québec, 2011.

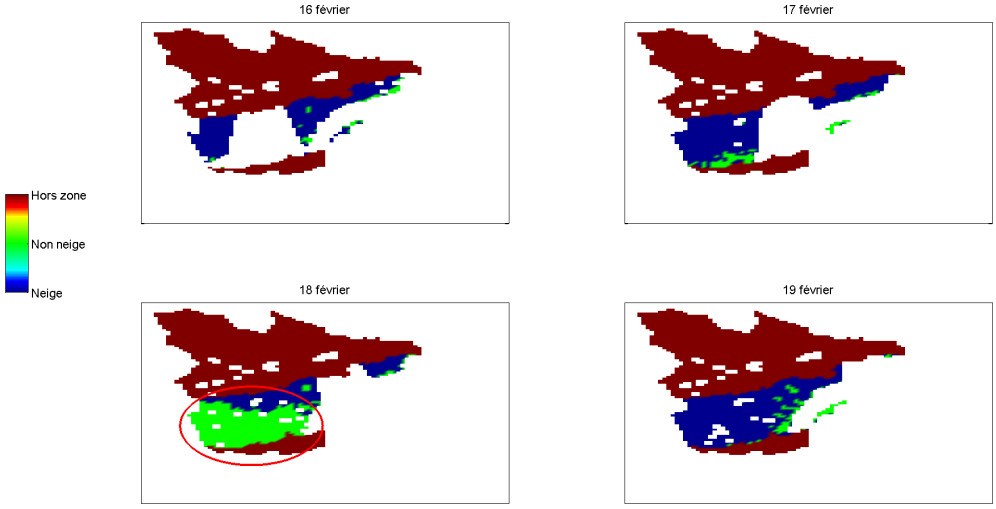

FIGURE B.7. Exemple de cartographies avec un seuil par jour pour deux, trois et quatre variables explicatives où la répartition neige/non-neige se comporte étrangement, zone 2 du Québec, 2011.

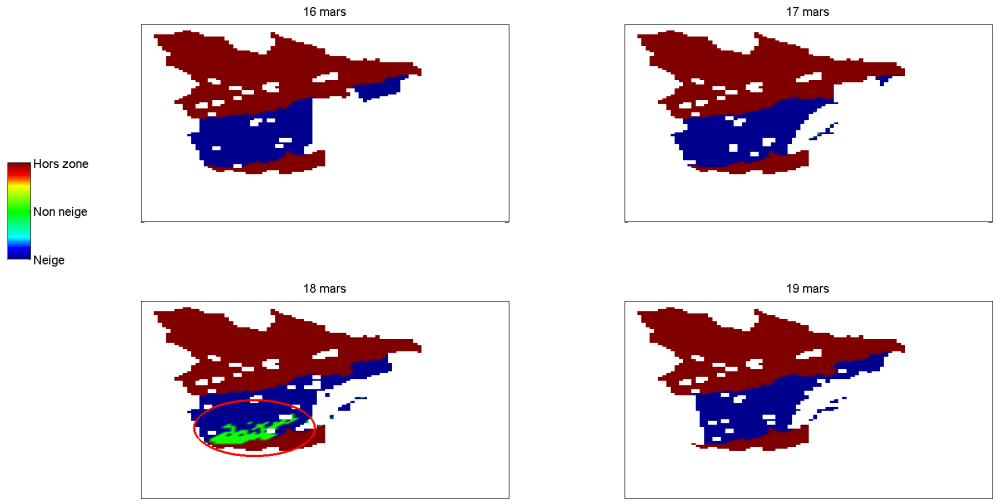

FIGURE B.8. Exemple de cartographies avec un seuil par point de grille pour deux, trois et quatre variables explicatives où remarquons un comportement étrange de la répartition neige/non-neige, zone 2 du Québec, 2011.

Annexe C

ÉTUDE DE CONVERGENCE POUR LES MÉLANGES DE LOIS NORMALES À UNE, DEUX OU TROIS SOUS-POPULATIONS

C.1. Diagrammes en boîtes des autres journées ayant servi dans l'étude de convergence

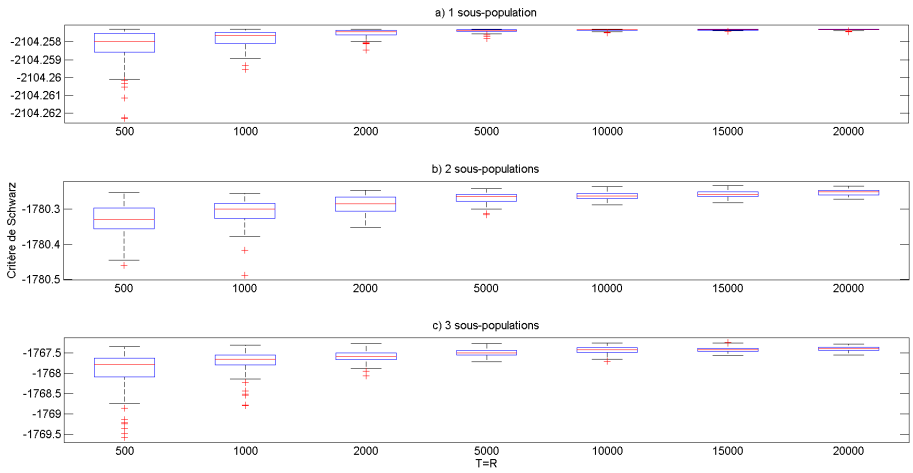

Figure C.1. Diagrammes en boîte des 100 valeurs du critère de Schwarz pour chacune des valeurs de $T = R$ et pour une à trois sous-populations, Qc, 7 janvier 2011.

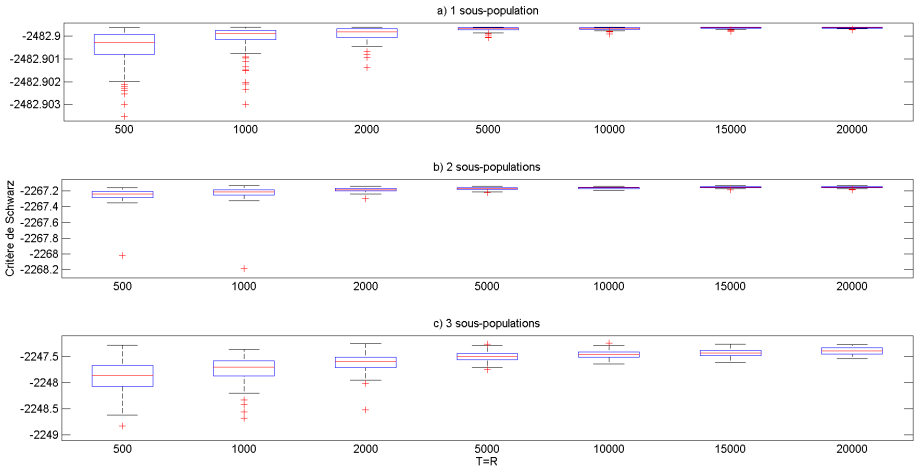

FIGURE C.2. Diagrammes en boîte des 100 valeurs du critère de Schwarz pour chacune des valeurs de $T = R$ et pour une à trois sous-populations, Qc, 21 février 2011.

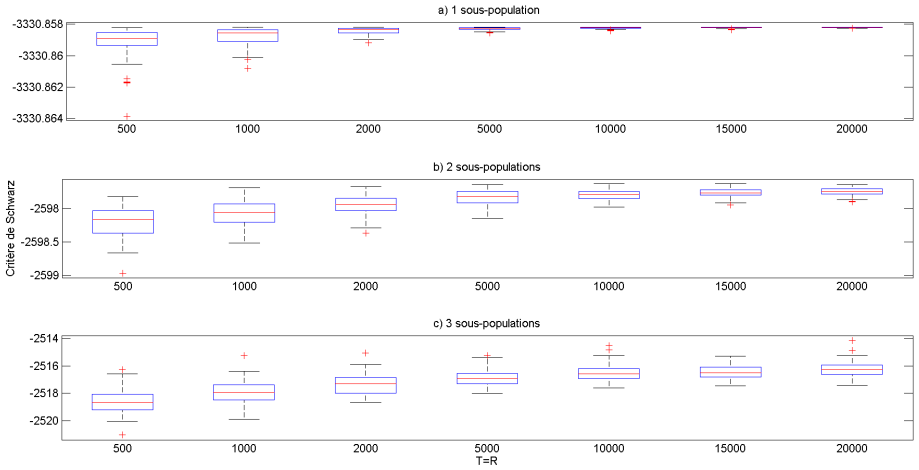

FIGURE C.3. Diagrammes en boîte des 100 valeurs du critère de Schwarz pour chacune des valeurs de $T = R$ et pour une à trois sous-populations, Qc, 13 mars 2011.

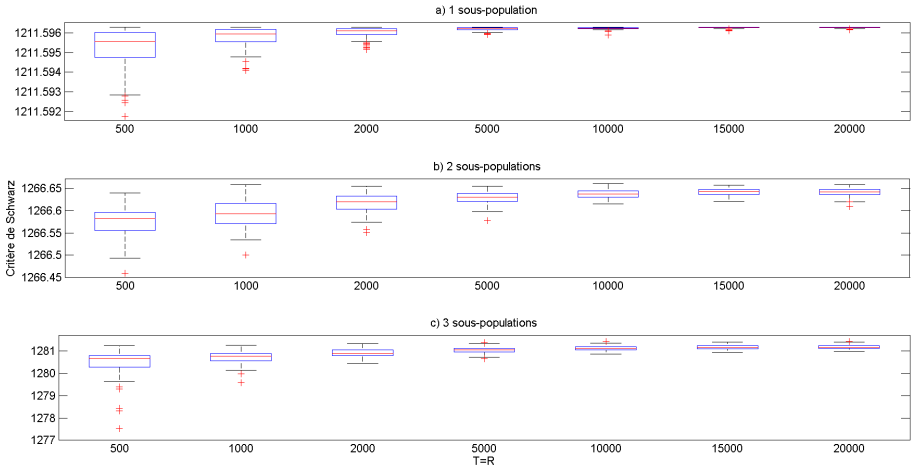

FIGURE C.4. Diagrammes en boîte des 100 valeurs du critère de Schwarz pour chacune des valeurs de $T = R$ et pour une à trois sous-populations, Qc, 27 juin 2011.

C.2. HISTOGRAMME DES VALEURS DU GTV DES JOURNÉES UTILISÉES POUR L'ÉTUDE DE CONVERGENCE

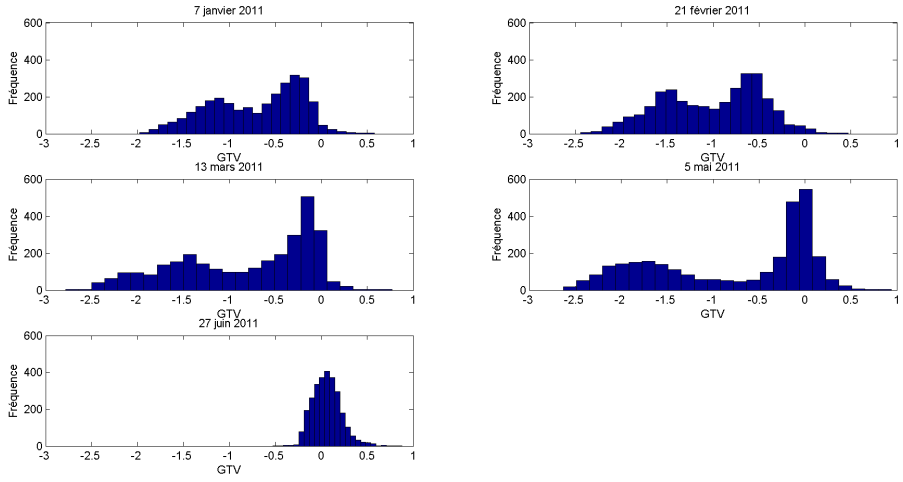

FIGURE C.5. Histogramme des valeurs de GTV des journées ayant servies à l'étude de convergence, Qc, 2011.

Annexe D

ÉTUDE DE LA RELATION ENTRE LES VARIABLES DE TEMPÉRATURES, DJX, DCX ET LP_T

D.1. Lien avec les variables de température

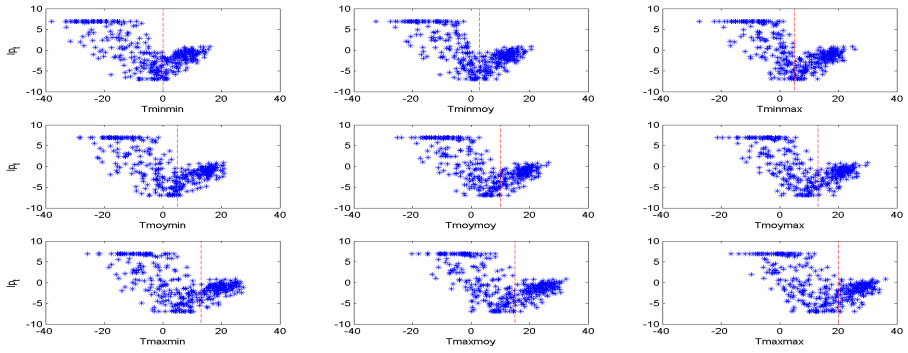

FIGURE D.1. Graphique de lp_t en fonction des différentes variables de température, zone 1 du Québec, 2011.

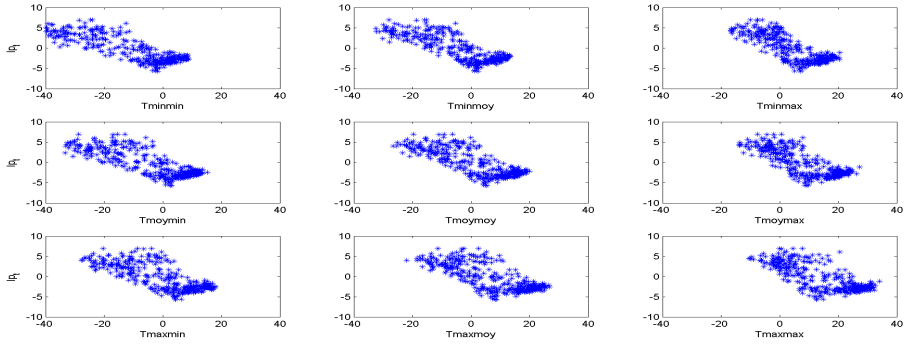

FIGURE D.2. Graphique de lp_t en fonction des différentes variables de température, zone 2 du Québec, 2011.

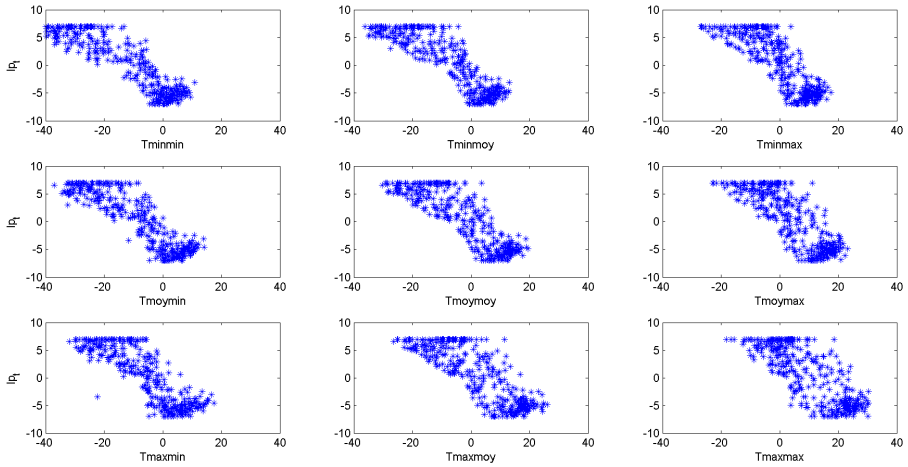

FIGURE D.3. Graphique de lp_t en fonction des différentes variables de température, zone 3 du Québec, 2011.

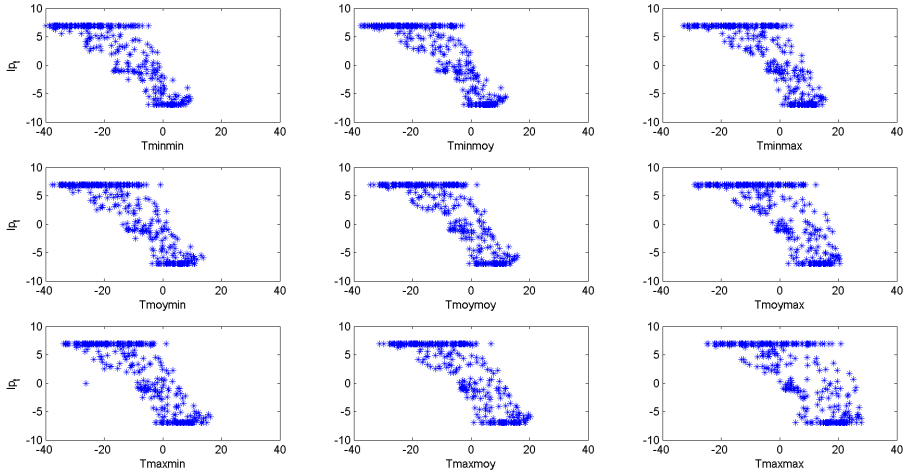

FIGURE D.4. Graphique de lp_t en fonction des différentes variables de température, zone 4 du Québec, 2011.

D.2. Lien avec les variables de degrés-jours et degrés-chauffes

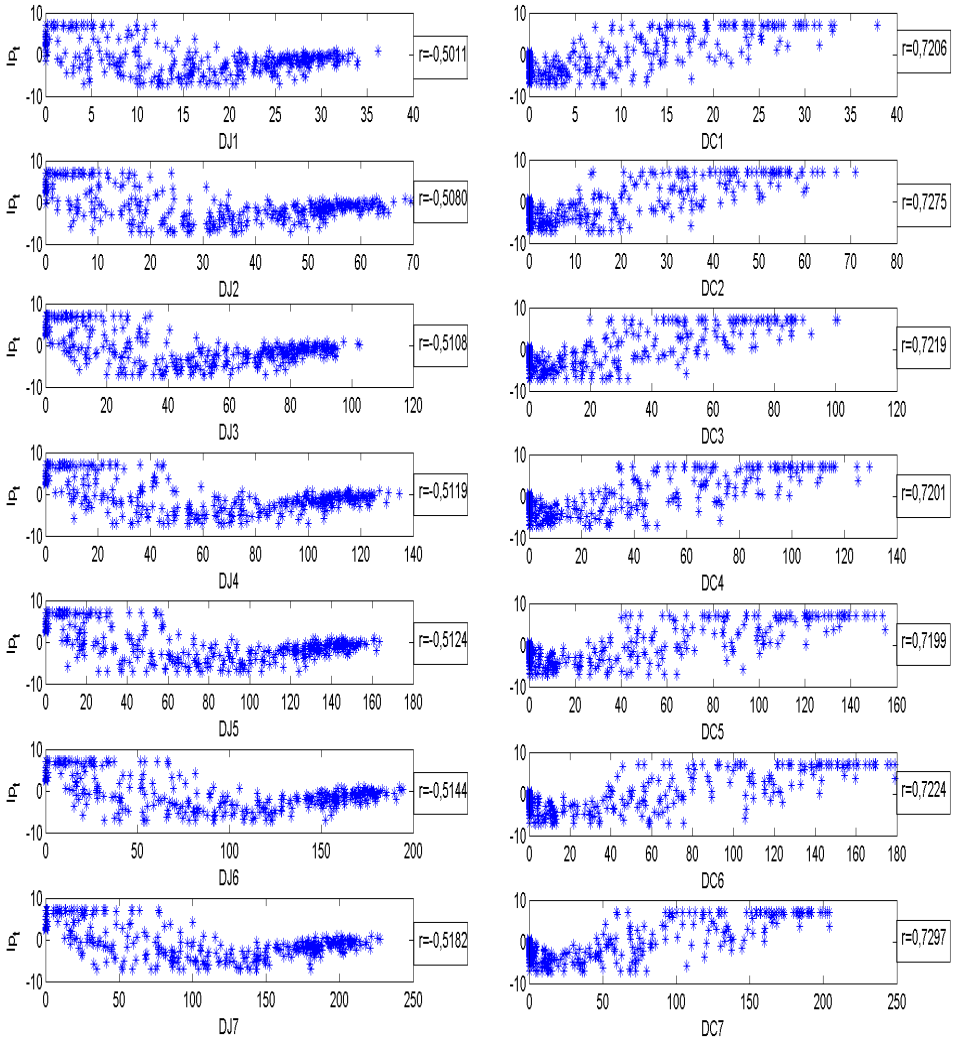

FIGURE D.5. Graphique de lp_t en fonction des degrés-jours et degrés-chauffes pour un à sept jours, zone 1 du Québec, 2011.

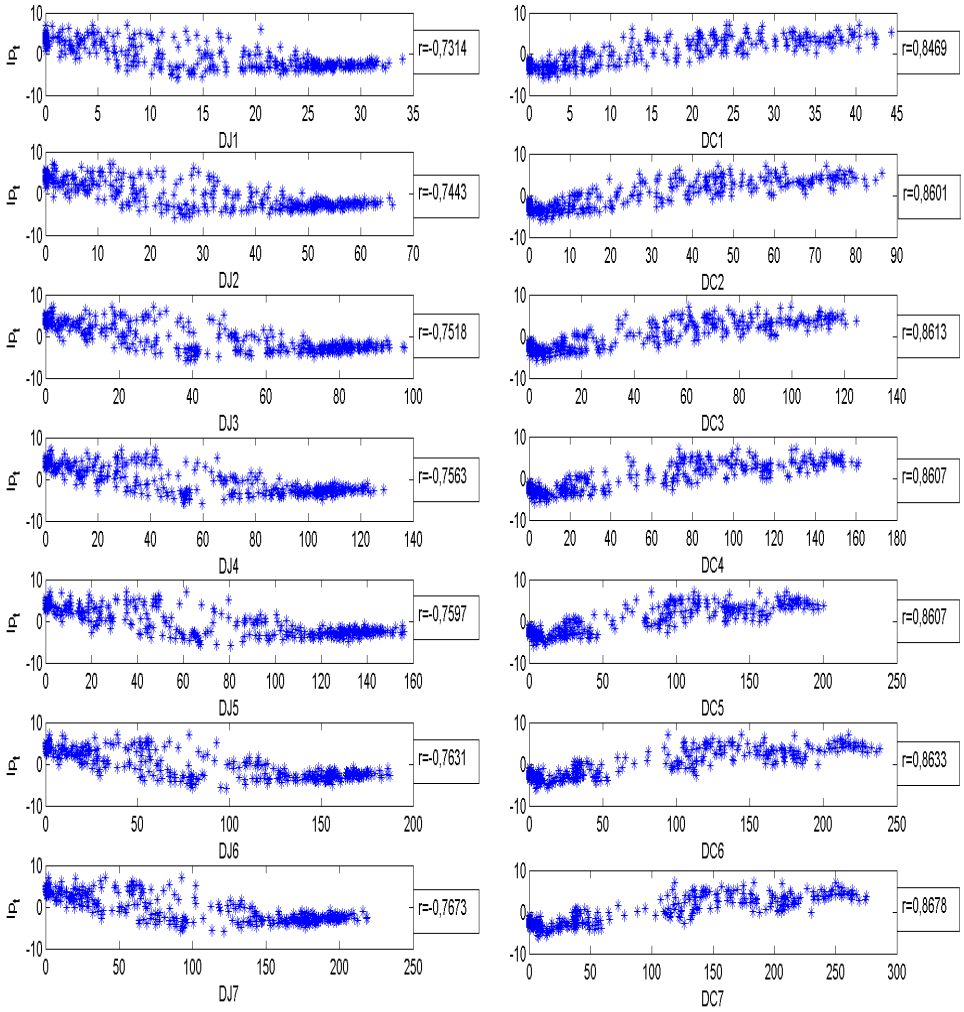

FIGURE D.6. Graphique de lp_t en fonction des degrés-jours et degrés-chauffes pour un à sept jours, zone 2 du Québec, 2011.

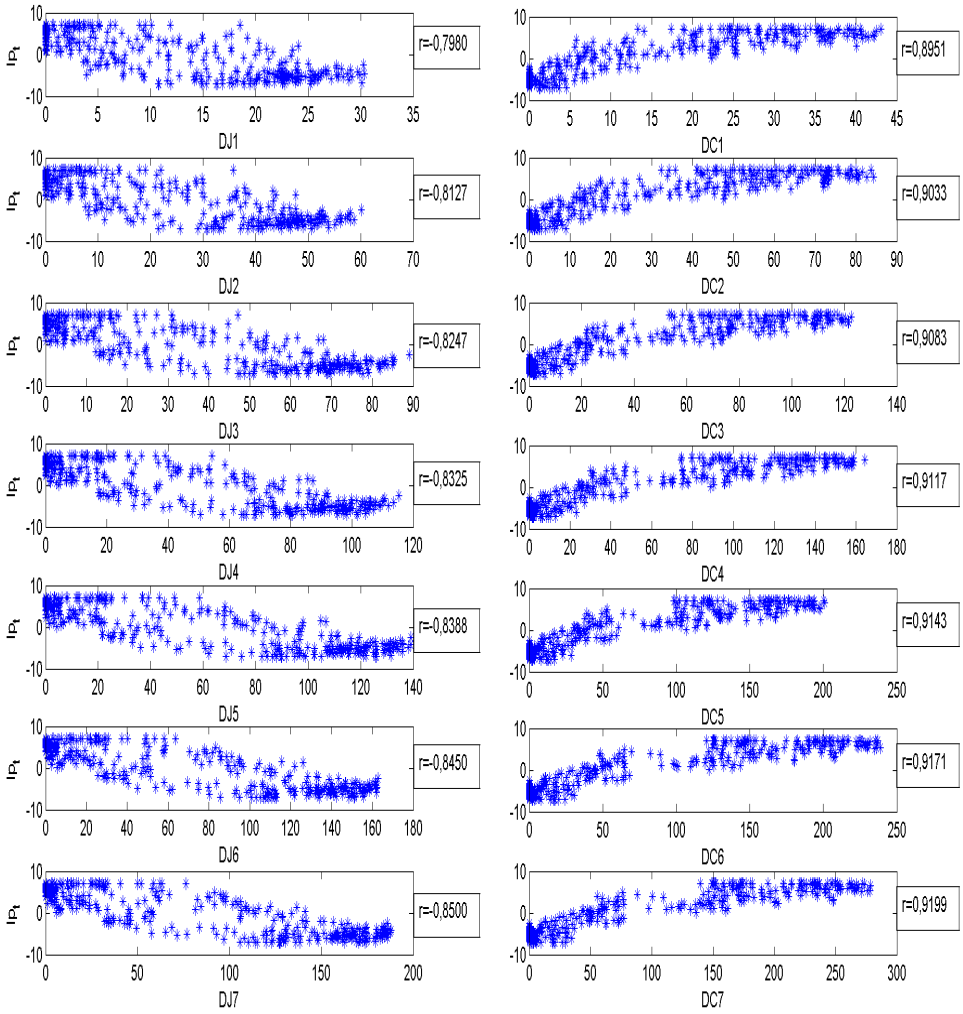

FIGURE D.7. Graphique de lp_t en fonction des degrés-jours et degrés-chauffes pour un à sept jours, zone 3 du Québec, 2011.

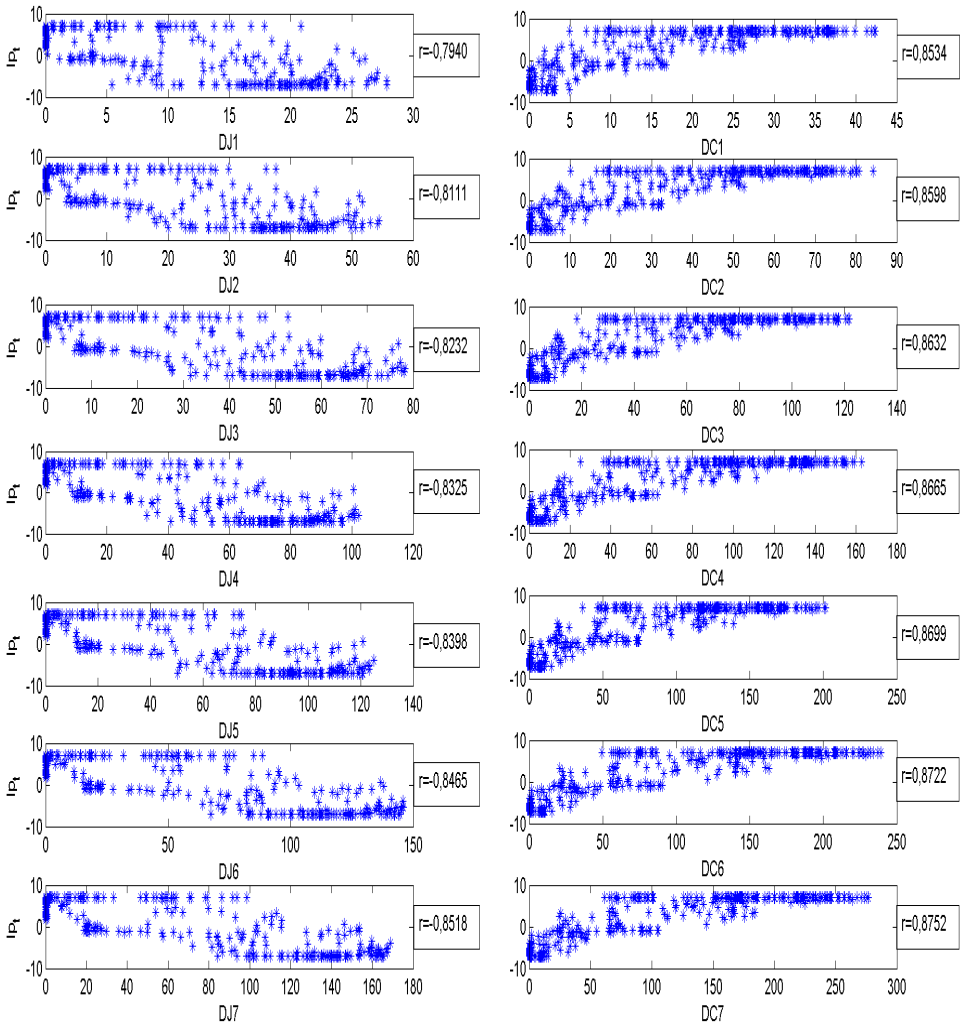

FIGURE D.8. Graphique de lp_t en fonction des degrés-jours et degrés-chauffes pour un à sept jours, zone 4 du Québec, 2011.

Annexe E

SÉLECTION DE VARIABLES ET DIAGNOSTICS DES RÉSIDUS

E.1. Sélection de variables à rebours

E.1.1. Zone 2

Table E.1. Étape 1 de la sélection de variables à rebours pour la régression linéaire sur lp_t calculés de la zone 2 pour l'année 2011, Québec. Variable lp_{t-4} à retirer.

Étape 1				
Variables	Estimation	Erreur standard	Statistique t	Valeur-p
origine	-0,4921	0,0882	-5,5816	<0,0001
lp_{t-1}	0,4276	0,0548	7,8063	<0,0001
lp_{t-2}	-0,0709	0,0588	-1,2046	0,2293
lp_{t-3}	0,0817	0,0582	1,4042	0,1613
lp_{t-4}	-0,0340	0,0582	-0,5849	0,5591
lp_{t-5}	0,1758	0,0560	3,1356	0,0019
lp_{t-6}	0,1684	0,0555	3,0331	0,0026
lp_{t-7}	0,0892	0,0505	1,7649	0,0786
Tminmin	-0,0404	0,0059	-6,8548	<0,0001

TABLE E.2. Étape 2 de la sélection de variables à rebours pour la régression linéaire sur lp_t calculés de la zone 2 pour l'année 2011, Québec. Variable lp_{t-2} à retirer.

Étape 2				
Variables	Estimation	Erreur standard	Statistique t	Valeur-p
origine	-0,4819	0,0878	-5,4914	<0,0001
lp_{t-1}	0,4196	0,0524	8,0043	<0,0001
lp_{t-2}	-0,0708	0,0571	-1,2403	0,2158
lp_{t-3}	0,0656	0,0497	1,3201	0,1878
lp_{t-5}	0,1647	0,0484	3,4046	0,0007
lp_{t-6}	0,1740	0,0552	3,1535	0,0018
lp_{t-7}	0,0901	0,0499	1,8066	0,0718
Tminmin	-0,0395	0,0058	-6,7482	<0,0001

TABLE E.3. Étape 3 de la sélection de variables à rebours pour la régression linéaire sur lp_t calculés de la zone 2 pour l'année 2011, Québec. Variable lp_{t-3} à retirer.

Étape 3				
Variables	Estimation	Erreur standard	Statistique t	Valeur-p
origine	-0,4789	0,0878	-5,4521	<0,0001
lp_{t-1}	0,3884	0,0472	8,2297	<0,0001
lp_{t-3}	0,0302	0,0425	0,7094	0,4786
lp_{t-5}	0,1595	0,0482	3,3074	0,0010
lp_{t-6}	0,1831	0,0551	3,3256	0,0010
lp_{t-7}	0,0781	0,0492	1,5875	0,1134
Tminmin	-0,0396	0,0059	-6,7505	<0,0001

TABLE E.4. Étape 4 de la sélection de variables à rebours pour la régression linéaire sur lp_t calculés de la zone 2 pour l'année 2011, Québec. Variable lp_{t-7} à retirer.

Étape 4				
Variables	Estimation	Erreur standard	Statistique t	Valeur-p
origine	-0,5164	0,0900	-5,7387	<0,0001
lp_{t-1}	0,3962	0,0467	8,4856	<0,0001
lp_{t-5}	0,1757	0,0472	3,7208	0,0002
lp_{t-6}	0,1816	0,0565	3,2122	0,0014
lp_{t-7}	0,0765	0,0495	1,5444	0,1235
Tminmin	-0,0407	0,0060	-6,7757	<0,0001

E.1.2. Zone 1

TABLE E.5. Étape 1 de la sélection de variables à rebours pour la régression linéaire sur lp_t calculés de la zone 1 pour l'année 2011, Québec. Variable lp_{t-5} à retirer.

Étape 1				
Variables	Estimation	Erreur standard	Statistique t	Valeur-p
origine	-0,2100	0,1122	-1,8715	0,0622
lp_{t-1}	0,4237	0,0559	7,5832	<0,0001
lp_{t-2}	-0,1012	0,0578	-1,7508	0,0810
lp_{t-3}	0,1646	0,0585	2,8156	0,0052
lp_{t-4}	0,0762	0,0587	1,2970	0,1956
lp_{t-5}	-0,0338	0,0563	-0,6011	0,5482
lp_{t-6}	0,3134	0,0557	5,6311	<0,0001
lp_{t-7}	0,0470	0,0516	0,9120	0,3625
Tminmin	-0,0368	0,0098	-3,7498	0,0002

TABLE E.6. Étape 2 de la sélection de variables à rebours pour la régression linéaire sur lp_t calculés de la zone 1 pour l'année 2011, Québec. Variable lp_{t-7} à retirer.

Étape 2				
Variables	Estimation	Erreur standard	Statistique t	Valeur-p
origine	-0,2112	0,1106	-1,9093	0,0571
lp_{t-1}	0,4330	0,0543	7,9761	<0,0001
lp_{t-2}	-0,1106	0,0550	-2,0097	0,0453
lp_{t-3}	0,1637	0,0570	2,8711	0,0044
lp_{t-4}	0,0574	0,0514	1,1172	0,2647
lp_{t-6}	0,2995	0,0490	6,1118	<0,0001
lp_{t-7}	0,0516	0,0509	1,0125	0,3121
Tminmin	-0,0363	0,0097	-3,7328	0,0002

TABLE E.7. Étape 3 de la sélection de variables à rebours pour la régression linéaire sur lp_t calculés de la zone 1 pour l'année 2011, Québec. Variable lp_{t-4} à retirer.

Étape 3				
Variables	Estimation	Erreur standard	Statistique t	Valeur-p
origine	-0,1958	0,1096	-1,7871	0,0749
lp_{t-1}	0,4553	0,0503	9,0442	<0,0001
lp_{t-2}	-0,1148	0,0548	-2,0960	0,0369
lp_{t-3}	0,1703	0,0558	3,0516	0,0025
lp_{t-4}	0,0550	0,0494	1,1146	0,2659
lp_{t-6}	0,3269	0,0428	7,6406	<0,0001
Tminmin	-0,0353	0,0096	-3,6593	0,0003

TABLE E.8. Étape 4 de la sélection de variables à rebours pour la régression linéaire sur lp_t calculés de la zone 1 pour l'année 2011, Québec. Variable lp_{t-2} à retirer.

Étape 4				
Variables	Estimation	Erreur standard	Statistique t	Valeur-p
origine	-0,2007	0,1100	-1,8235	0,0691
lp_{t-1}	0,4534	0,0479	9,4599	<0,0001
lp_{t-2}	-0,1184	0,0548	-2,1631	0,0313
lp_{t-3}	0,2047	0,0488	4,1910	<0,0001
lp_{t-6}	0,3559	0,0422	8,4265	<0,0001
Tminmin	-0,0359	0,0097	-3,6995	0,0003

TABLE E.9. Étape 5 de la sélection de variables à rebours pour la régression linéaire sur lp_t calculés de la zone 1 pour l'année 2011, Québec. Variable lp_{t-3} à retirer.

Étape 5 ($p^* = 4$)				
Variables	Estimation	Erreur standard	Statistique t	Valeur-p
origine	-0,2058	0,1092	-1,8850	0,0603
lp_{t-1}	0,4078	0,0424	9,6176	1,7493e-19
lp_{t-3}	0,1483	0,0423	3,5081	0,0005
lp_{t-6}	0,3409	0,0418	8,1525	7,3207e-15
Tminmin	-0,0371	0,0096	-3,8455	0,0001

TABLE E.10. Étape 6 de la sélection de variables à rebours pour la régression linéaire sur lp_t calculés de la zone 1 pour l'année 2011, Québec. Variable lp_{t-6} à retirer.

Étape 6 ($p^* = 3$)				
Variables	Estimation	Erreur standard	Statistique t	Valeur-p
origine	-0,2655	0,1123	-2,3655	0,0186
lp_{t-1}	0,4600	0,0403	11,4238	8,6597e-26
lp_{t-6}	0,4012	0,0386	10,3821	4,1509e-22
Tminmin	-0,0419	0,0099	-4,2183	3,1624e-05

TABLE E.11. Étape 7 de la sélection de variables à rebours pour la régression linéaire sur lp_t calculés de la zone 1 pour l'année 2011, Québec. Modèle le plus parcimonieux.

Étape 7 ($p^* = 2$)				
Variables	Estimation	Erreur standard	Statistique t	Valeur-p
origine	-0,3857	0,1262	-3,0554	0,0024
lp_{t-1}	0,7418	0,0337	21,9840	1,3661e-67
Tminmin	-0,0582	0,0111	-5,2405	2,7953e-07

E.1.3. Zone 3

TABLE E.12. Étape 1 de la sélection de variables à rebours pour la régression linéaire sur lp_t calculés de la zone 3 pour l'année 2011, Québec. Variable lp_{t-2} à retirer.

Étape 1				
Variables	Estimation	Erreur standard	Statistique t	Valeur-p
origine	-0,5305	0,1143	-4,6412	<0,0001
lp_{t-1}	0,4525	0,0544	8,3113	<0,0001
lp_{t-2}	0,0100	0,0596	0,1681	0,8667
lp_{t-3}	0,1012	0,0596	1,6983	0,0904
lp_{t-4}	0,0940	0,0592	1,5859	0,1138
lp_{t-5}	0,0802	0,0587	1,3676	0,1724
lp_{t-6}	-0,0114	0,0589	-0,1931	0,8470
lp_{t-7}	0,1404	0,0517	2,7141	0,0070
Tminmin	-0,0463	0,0080	-5,7743	<0,0001

TABLE E.13. Étape 2 de la sélection de variables à rebours pour la régression linéaire sur lp_t calculés de la zone 3 pour l'année 2011, Québec. Variable lp_{t-6} à retirer.

Étape 2				
Variables	Estimation	Erreur standard	Statistique t	Valeur-p
origine	-0,5322	0,1129	-4,7133	<0,0001
lp_{t-1}	0,4581	0,0481	9,5183	<0,0001
lp_{t-3}	0,1046	0,0541	1,9359	0,0538
lp_{t-4}	0,0940	0,0587	1,6009	0,1104
lp_{t-5}	0,0803	0,0579	1,3865	0,1666
lp_{t-6}	-0,0086	0,0582	-0,1486	0,8820
lp_{t-7}	0,1392	0,0511	2,7238	0,0068
Tminmin	-0,0462	0,0079	-5,8432	<0,0001

TABLE E.14. Étape 3 de la sélection de variables à rebours pour la régression linéaire sur lp_t calculés de la zone 3 pour l'année 2011, Québec. Variable lp_{t-5} à retirer.

Étape 3				
Variables	Estimation	Erreur standard	Statistique t	Valeur-p
origine	-0,5471	0,1138	-4,8059	<0,0001
lp_{t-1}	0,4656	0,0481	9,6794	<0,0001
lp_{t-3}	0,1192	0,0538	2,2167	0,0273
lp_{t-4}	0,0853	0,0592	1,4420	0,1503
lp_{t-5}	0,0698	0,0532	1,3126	0,1902
lp_{t-7}	0,1275	0,0454	2,8106	0,0052
Tminmin	-0,0463	0,0079	-5,8333	<0,0001

TABLE E.15. Étape 4 de la sélection de variables à rebours pour la régression linéaire sur lp_t calculés de la zone 3 pour l'année 2011, Québec. Variable lp_{t-4} à retirer.

Étape 4				
Variables	Estimation	Erreur standard	Statistique t	Valeur-p
origine	-0,5551	0,1130	-4,9104	<0,0001
lp_{t-1}	0,4833	0,0461	10,4811	<0,0001
lp_{t-3}	0,1227	0,0534	2,2962	0,0223
lp_{t-4}	0,1179	0,0534	2,2089	0,0279
lp_{t-7}	0,1430	0,0433	3,2996	0,0011
Tminmin	-0,0465	0,0079	-5,8960	<0,0001

TABLE E.16. Étape 5 de la sélection de variables à rebours pour la régression linéaire sur lp_t calculés de la zone 3 pour l'année 2011, Québec. Variable lp_{t-3} à retirer.

Étape 5 ($p^* = 4$)				
Variables	Estimation	Erreur standard	Statistique t	Valeur-p
origine	-0,5704	0,1134	-5,0284	8,0534e-07
lp_{t-1}	0,5032	0,0450	11,1798	6,8179e-25
lp_{t-3}	0,1805	0,0469	3,8460	0,0001
lp_{t-7}	0,1767	0,0412	4,2854	2,3837e-05
Tminmin	-0,0476	0,0079	-6,0245	4,4436e-09

TABLE E.17. Étape 6 de la sélection de variables à rebours pour la régression linéaire sur lp_t calculés de la zone 3 pour l'année 2011, Québec. Variable lp_{t-7} à retirer.

Étape 6 ($p^* = 3$)				
Variables	Estimation	Erreur standard	Statistique t	Valeur-p
origine	-0,6605	0,1139	-5,8013	1,4997e-08
lp_{t-1}	0,5826	0,0409	14,2531	1,6637e-36
lp_{t-7}	0,2581	0,0359	7,1979	3,9041e-12
Tminmin	-0,0541	0,0079	-6,8583	3,2727e-11

TABLE E.18. Étape 7 de la sélection de variables à rebours pour la régression linéaire sur lp_t calculés de la zone 3 pour l'année 2011, Québec. Modèle le plus parcimonieux

Étape 7 ($p^* = 2$)				
Variables	Estimation	Erreur standard	Statistique t	Valeur-p
origine	-0,6926	0,1211	-5,7175	2,3326e-08
lp_{t-1}	0,8208	0,0255	32,1798	3,4144e-106
Tminmin	-0,0562	0,0084	-6,7175	7,5826e-11

E.1.4. Zone 4

TABLE E.19. Étape 1 de la sélection de variables à rebours pour la régression linéaire sur lp_t calculés de la zone 4 pour l'année 2011, Québec. Variable lp_{t-4} à retirer.

Étape 1				
Variables	Estimation	Erreur standard	Statistique t	Valeur-p
origine	-0,3255	0,1100	-2,9582	0,0033
lp_{t-1}	0,4622	0,0546	8,4707	<0,0001
lp_{t-2}	0,0759	0,0612	1,2407	0,2156
lp_{t-3}	0,2783	0,0608	4,5793	<0,0001
lp_{t-4}	0,0170	0,0622	0,2732	0,7849
lp_{t-5}	0,1203	0,0601	2,0016	0,0462
lp_{t-6}	-0,1381	0,0598	-2,3095	0,0215
lp_{t-7}	0,1076	0,0538	2,0005	0,0463
Tminmin	-0,0323	0,0083	-3,8822	0,0001

TABLE E.20. Étape 2 de la sélection de variables à rebours pour la régression linéaire sur lp_t calculés de la zone 4 pour l'année 2011, Québec. Variable lp_{t-2} à retirer.

Étape 2				
Variables	Estimation	Erreur standard	Statistique t	Valeur-p
origine	-0,3278	0,1096	-2,9902	0,0030
lp_{t-1}	0,4659	0,0527	8,8425	<0,0001
lp_{t-2}	0,0757	0,0607	1,2465	0,2135
lp_{t-3}	0,2733	0,0573	4,7695	<0,0001
lp_{t-5}	0,1266	0,0570	2,2196	0,0271
lp_{t-6}	-0,1160	0,0586	-1,9812	0,0484
lp_{t-7}	0,0965	0,0514	1,8781	0,0613
Tminmin	-0,0332	0,0083	-4,0069	0,0001

TABLE E.21. Étape 3 de la sélection de variables à rebours pour la régression linéaire sur lp_t calculés de la zone 4 pour l'année 2011, Québec. Variable lp_{t-7} à retirer.

Étape 3				
Variables	Estimation	Erreur standard	Statistique t	Valeur-p
origine	-0,3755	0,1092	-3,4373	0,0007
lp_{t-1}	0,4865	0,0479	10,1464	<0,0001
lp_{t-3}	0,2937	0,0543	5,4112	<0,0001
lp_{t-5}	0,1512	0,0565	2,6765	0,0078
lp_{t-6}	-0,1158	0,0589	-1,9672	0,0500
lp_{t-7}	0,0980	0,0511	1,9170	0,0561
Tminmin	-0,0370	0,0082	-4,4962	<0,0001

TABLE E.22. Étape 4 de la sélection de variables à rebours pour la régression linéaire sur lp_t calculés de la zone 4 pour l'année 2011, Québec. Variable lp_{t-6} à retirer.

Étape 4				
Variables	Estimation	Erreur standard	Statistique t	Valeur-p
origine	-0,3679	0,1090	-3,3757	0,0008
lp_{t-1}	0,4988	0,0477	10,4642	<0,0001
lp_{t-3}	0,3082	0,0524	5,8815	<0,0001
lp_{t-5}	0,1676	0,0549	3,0532	0,0024
lp_{t-6}	-0,0606	0,0530	-1,1440	0,2534
Tminmin	-0,0360	0,0082	-4,3872	<0,0001

TABLE E.23. Étape 5 de la sélection de variables à rebours pour la régression linéaire sur lp_t calculés de la zone 4 pour l'année 2011, Québec. Variable lp_{t-5} à retirer.

Étape 5 ($p^* = 4$)				
Variables	Estimation	Erreur standard	Statistique t	Valeur-p
origine	-0,3851	0,1087	-3,5445	0,0004
lp_{t-1}	0,4855	0,0470	10,3373	5,7178e-22
lp_{t-3}	0,2832	0,0497	5,6951	2,6644e-08
lp_{t-5}	0,1433	0,0462	3,1020	0,0021
Tminmin	-0,0369	0,0082	-4,4976	9,4336e-06

TABLE E.24. Étape 6 de la sélection de variables à rebours pour la régression linéaire sur lp_t calculés de la zone 4 pour l'année 2011, Québec. Variable lp_{t-3} à retirer.

Étape 6 ($p^* = 3$)				
Variables	Estimation	Erreur standard	Statistique t	Valeur-p
origine	-0,3807	0,1091	-3,4905	0,0005
lp_{t-1}	0,5430	0,0432	12,5580	4,7769e-30
lp_{t-3}	0,3670	0,0421	8,7170	1,2210e-16
Tminmin	-0,0365	0,0082	-4,4338	1,2462e-05

TABLE E.25. Étape 7 de la sélection de variables à rebours pour la régression linéaire sur lp_t calculés de la zone 4 pour l'année 2011, Québec.

Étape 7 ($p^* = 2$)				
Variables	Estimation	Erreur standard	Statistique t	Valeur-p
origine	-0,4805	0,1188	-4,0437	6,4790e-05
lp_{t-1}	0,8766	0,0222	39,5487	5,4145e-131
Tminmin	-0,0467	0,0089	-5,2326	2,8907e-07

E.1.5. Québec complet avec variables indicatrices pour les zones

TABLE E.26. Étape 1 de la sélection de variables à rebours pour la régression linéaire sur lp_t calculés pour l'année 2011, Québec. Variable ind_z4 à retirer (dès qu'une des indicatrices est éliminée, il faut toutes les retirer).

Étape 1				
Variables	Estimation	Erreur standard	Statistique t	Valeur-p
origine	-0,2163	0,0735	-2,9426	0,0033
lp_{t-1}	0,4381	0,0274	16,0038	<0,0001
lp_{t-2}	-0,0519	0,0295	-1,7587	0,0789
lp_{t-3}	0,1618	0,0297	5,4537	<0,0001
lp_{t-4}	0,0473	0,0298	1,5849	0,1132
lp_{t-5}	0,0465	0,0288	1,6137	0,1068
lp_{t-6}	0,1721	0,0286	6,0233	<0,0001
lp_{t-7}	0,0795	0,0257	3,0999	0,0020
Tminmin	-0,0379	0,0040	-9,4893	<0,0001
ind_z2	-0,2408	0,1027	-2,3453	0,0192
ind_z3	-0,2077	0,1026	-2,0232	0,0433
ind_z4	-0,1517	0,1009	-1,5028	0,1331

TABLE E.27. Étape 2 de la sélection de variables à rebours pour la régression linéaire sur lp_t calculés pour l'année 2011, Québec. Variable lp_{t-5} à retirer.

Étape 2				
Variables	Estimation	Erreur standard	Statistique t	Valeur-p
origine	-0,3361	0,0522	-6,4398	<0,0001
lp_{t-1}	0,4423	0,0274	16,1698	<0,0001
lp_{t-2}	-0,0511	0,0296	-1,7279	0,0842
lp_{t-3}	0,1632	0,0297	5,4945	<0,0001
lp_{t-4}	0,0487	0,0298	1,6320	0,1029
lp_{t-5}	0,0466	0,0288	1,6159	0,1064
lp_{t-6}	0,1726	0,0286	6,0344	<0,0001
lp_{t-7}	0,0777	0,0257	3,0284	0,0025
Tminmin	-0,0349	0,0038	-9,1668	<0,0001

TABLE E.28. Étape 3 de la sélection de variables à rebours pour la régression linéaire sur lp_t calculés pour l'année 2011, Québec. Variable lp_{t-2} à retirer.

Étape 3				
Variables	Estimation	Erreur standard	Statistique t	Valeur-p
origine	-0,3372	0,0518	-6,5157	<0,0001
lp_{t-1}	0,4524	0,0269	16,8242	<0,0001
lp_{t-2}	-0,0433	0,0285	-1,5199	0,1288
lp_{t-3}	0,1533	0,0292	5,2462	<0,0001
lp_{t-4}	0,0702	0,0265	2,6439	0,0083
lp_{t-6}	0,1927	0,0256	7,5383	<0,0001
lp_{t-7}	0,0755	0,0255	2,9580	0,0032
Tminmin	-0,0346	0,0038	-9,1466	<0,0001

TABLE E.29. Étape 4 de la sélection de variables à rebours pour la régression linéaire sur lp_t calculés pour l'année 2011, Québec. Variable lp_{t-4} à retirer.

Étape 4				
Variables	Estimation	Erreur standard	Statistique t	Valeur-p
origine	-0,3405	0,0514	-6,6285	<0,0001
lp_{t-1}	0,4361	0,0239	18,2678	<0,0001
lp_{t-3}	0,1323	0,0263	5,0365	<0,0001
lp_{t-4}	0,0701	0,0265	2,6513	0,0081
lp_{t-6}	0,1897	0,0254	7,4761	<0,0001
lp_{t-7}	0,0712	0,0255	2,7901	0,0053
Tminmin	-0,0351	0,0038	-9,3376	<0,0001

TABLE E.30. Étape 5 de la sélection de variables à rebours pour la régression linéaire sur lp_t calculés pour l'année 2011, Québec. Variable lp_{t-7} à retirer.

Étape 5				
Variables	Estimation	Erreur standard	Statistique t	Valeur-p
origine	-0,3450	0,0516	-6,6928	<0,0001
lp_{t-1}	0,4378	0,0232	18,9013	<0,0001
lp_{t-3}	0,1675	0,0226	7,4122	<0,0001
lp_{t-6}	0,2100	0,0252	8,3261	<0,0001
lp_{t-7}	0,0832	0,0253	3,2885	0,0010
Tminmin	-0,0355	0,0038	-9,4145	<0,0001

E.2. Diagnostics des résidus

E.2.1. Zone 1

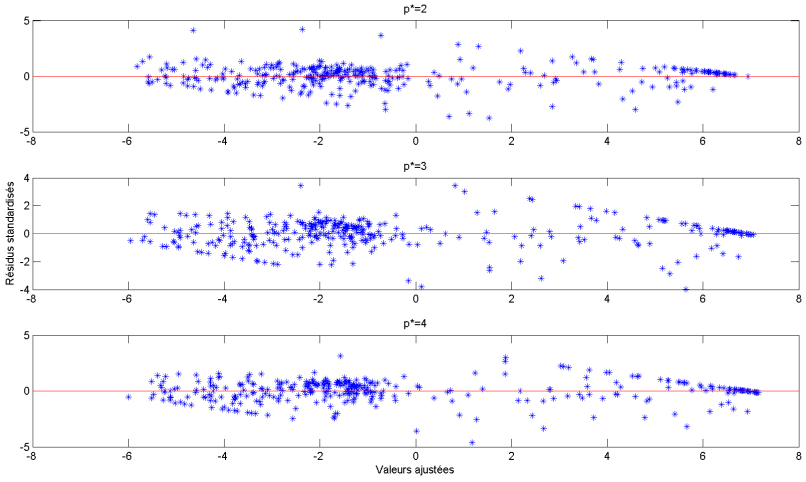

Figure E.1. Graphique des résidus standardisés en fonction de valeurs ajustées pour la régression de lp_t avec deux, trois et quatre variables explicatives, zone 1 du Québec, 2011.

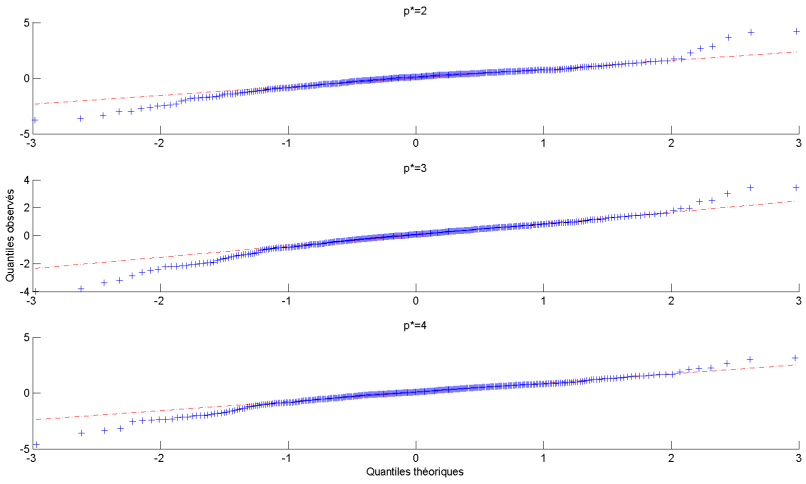

Figure E.2. Diagramme quantile-quantile des résidus standardisés pour la régression de lp_t avec deux, trois et quatre variables explicatives, zone 1 du Québec, 2011.

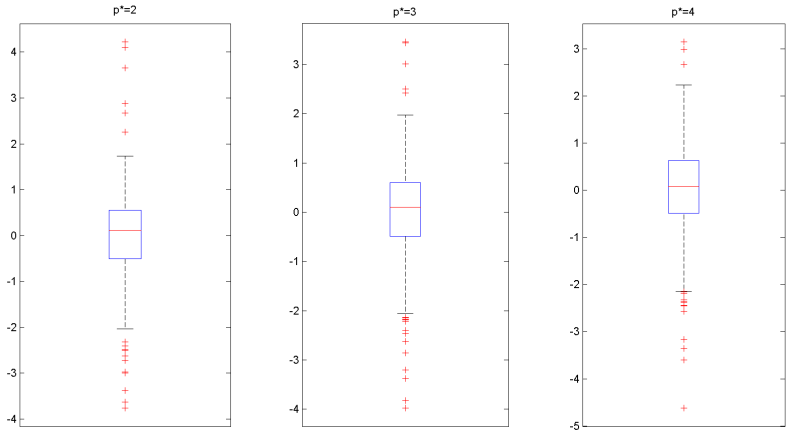

FIGURE E.3. Diagrammes en boîte des résidus standardisés pour la régression des lp_t avec deux, trois et quatre variables explicatives, zone 1 du Québec, 2011.

E.2.2. Zone 3

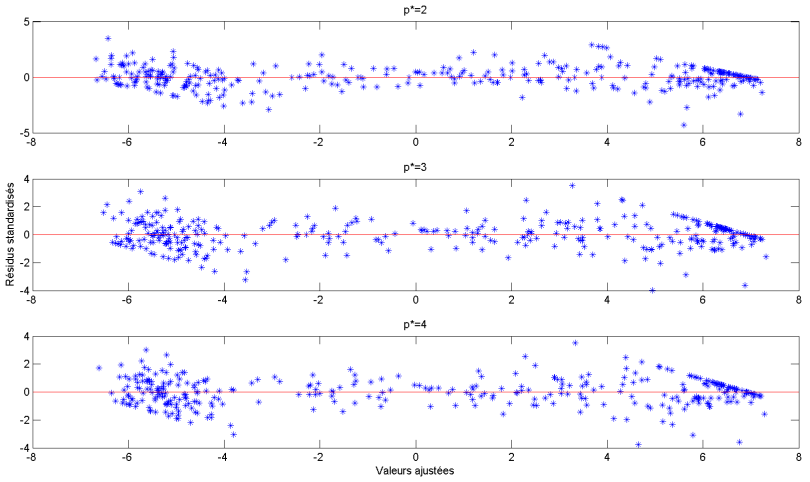

FIGURE E.4. Graphique des résidus standardisés en fonction de valeurs ajustées pour la régression de lp_t avec deux, trois et quatre variables explicatives, zone 3 du Québec, 2011.

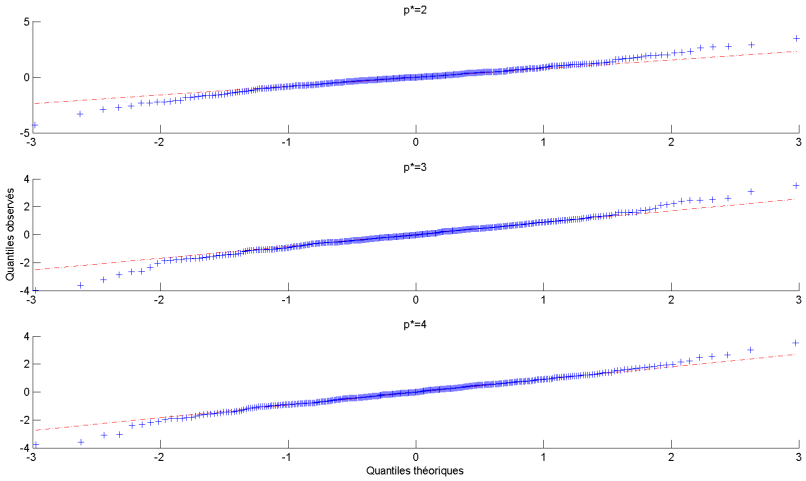

FIGURE E.5. Diagramme quantile-quantile des résidus standardisés pour la régression de lp_t avec deux, trois et quatre variables explicatives, zone 3 du Québec, 2011.

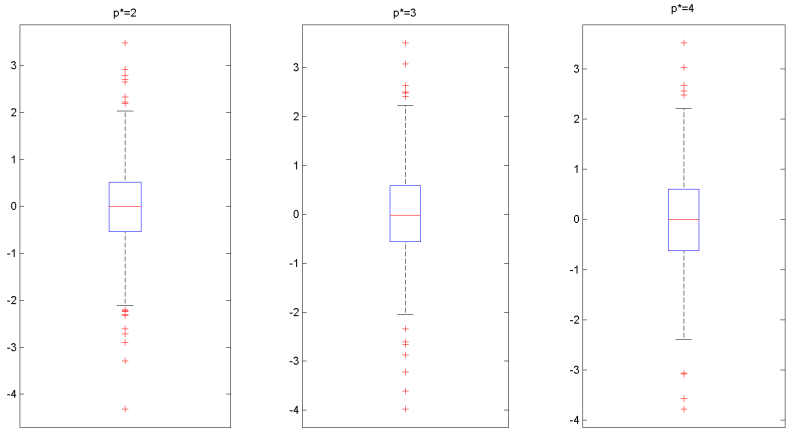

FIGURE E.6. Diagrammes en boîte des résidus standardisés pour la régression des lp_t avec deux, trois et quatre variables explicatives, zone 3 du Québec, 2011.

E.2.3. Zone 4

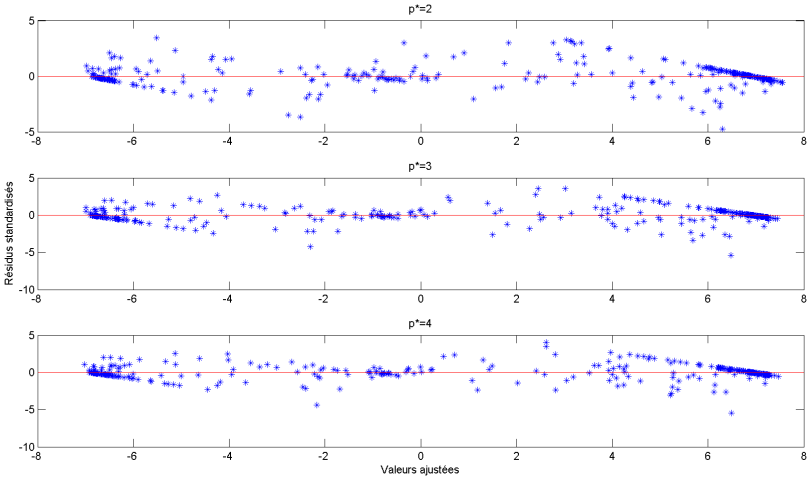

FIGURE E.7. Graphique des résidus standardisés en fonction de valeurs ajustées pour la régression de lp_t avec deux, trois et quatre variables explicatives, zone 4 du Québec, 2011.

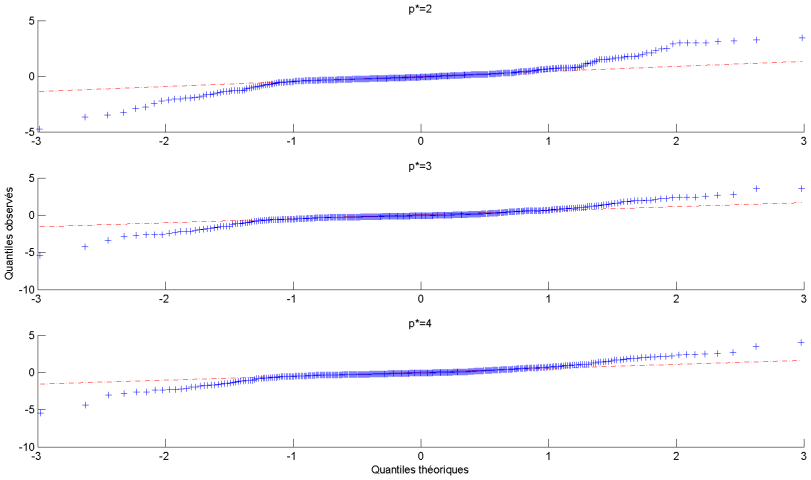

FIGURE E.8. Diagramme quantile-quantile des résidus standardisés pour la régression de lp_t avec deux, trois et quatre variables explicatives, zone 4 du Québec, 2011.

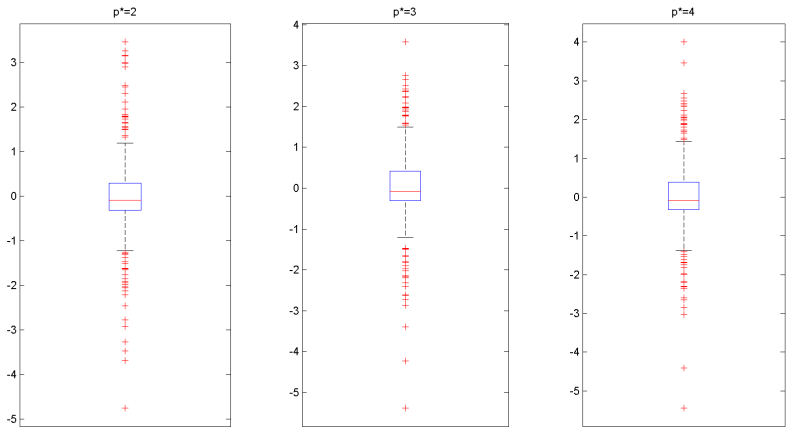

FIGURE E.9. Diagrammes en boîte des résidus standardisés pour la régression des lp_t avec deux, trois et quatre variables explicatives, zone 4 du Québec, 2011.

Annexe F

VALIDATION SUR 2011

F.1. Scores de Brier avec les données de capteurs GMON pour les combinaisons deux, trois et quatre

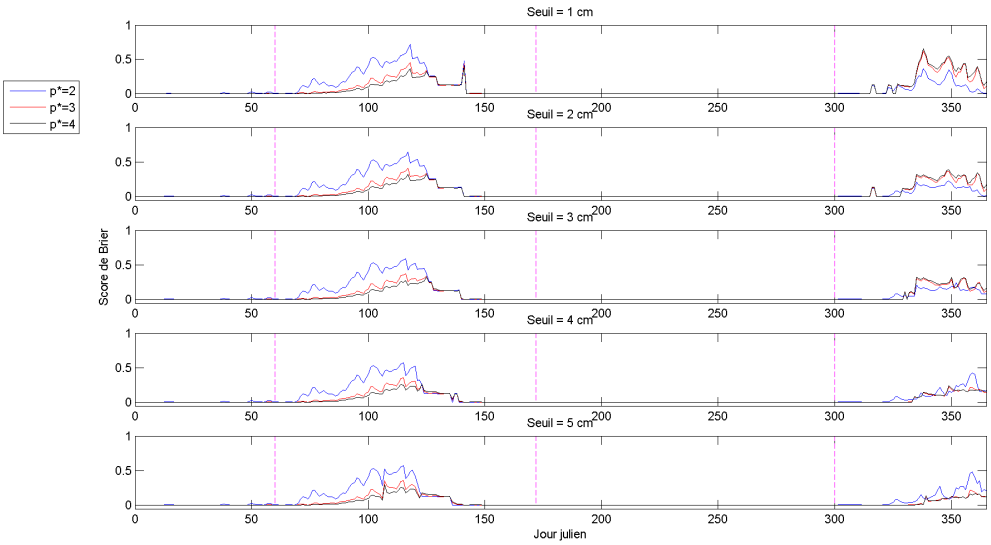

Figure F.1. Scores de Brier pour les modèles à deux, trois et quatre variables explicatives calculés pour les réalités neige/non-neige de la deuxième combinaison de capteurs GMON selon les seuils de 1 à 5 cm, représentation par seuil, Québec, 2011.

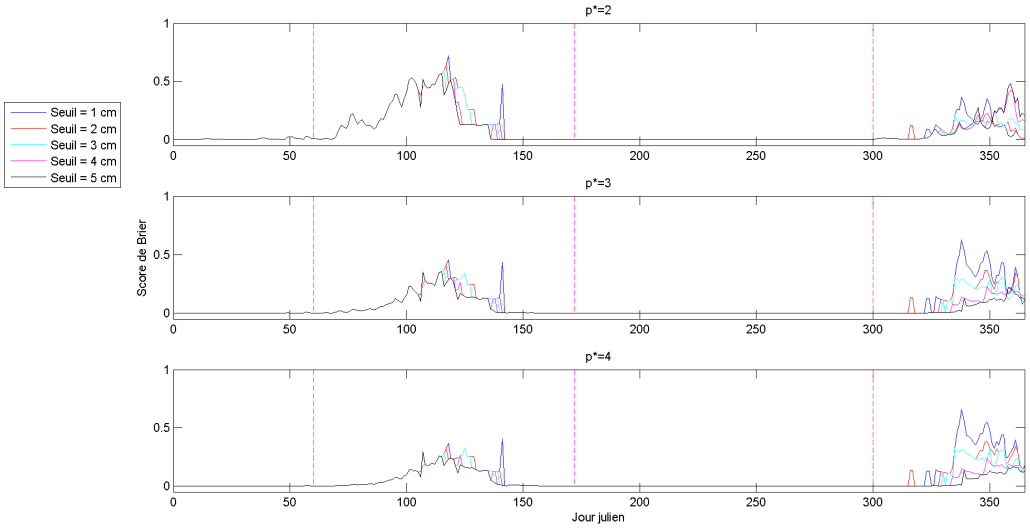

FIGURE F.2. Scores de Brier pour les modèles à deux, trois et quatre variables explicatives calculés pour les réalités neige/non-neige de la deuxième combinaison de capteurs GMON selon les seuils de 1 à 5 cm, représentation par modèle, Québec, 2011.

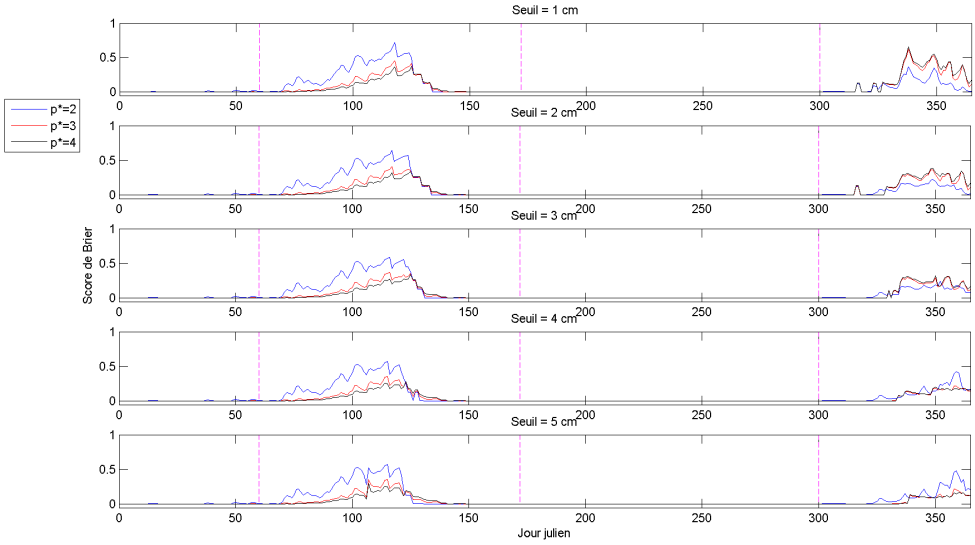

FIGURE F.3. Scores de Brier pour les modèles à deux, trois et quatre variables explicatives calculés pour les réalités neige/non-neige de la troisième combinaison de capteurs GMON selon les seuils de 1 à 5 cm, représentation par seuil, Québec, 2011.

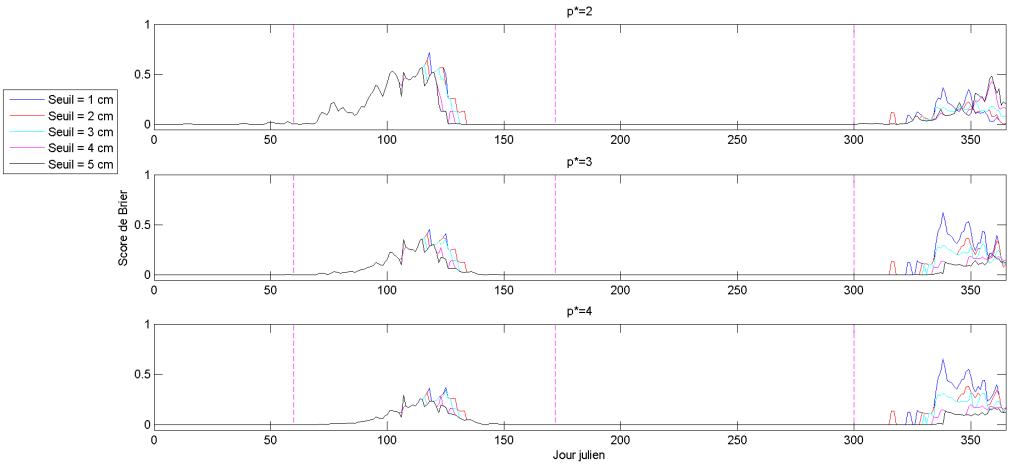

FIGURE F.4. Scores de Brier pour les modèles à deux, trois et quatre variables explicatives calculés pour les réalités neige/non-neige de la troisième combinaison de capteurs GMON selon les seuils de 1 à 5 cm, représentation par modèle, Québec, 2011.

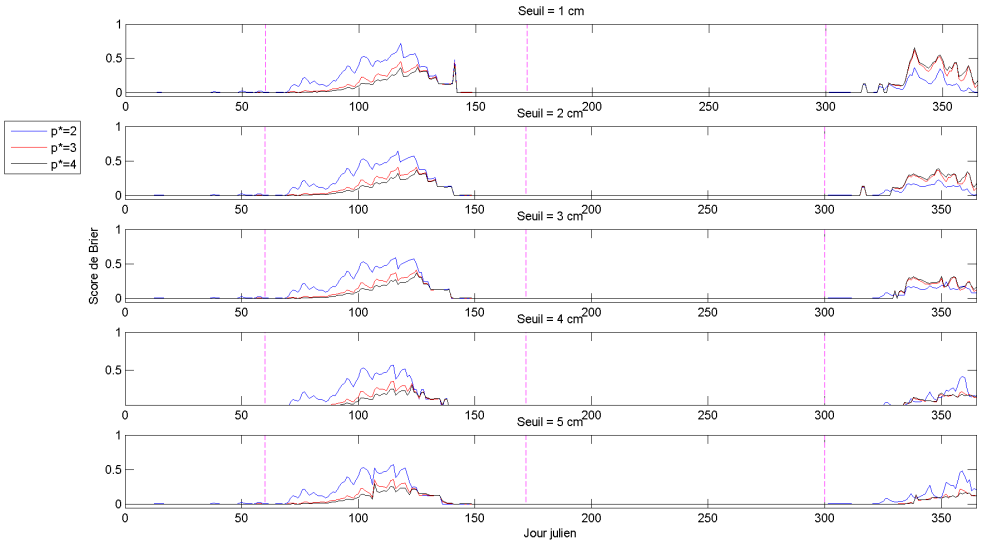

FIGURE F.5. Scores de Brier pour les modèles à deux, trois et quatre variables explicatives calculés pour les réalités neige/non-neige de la quatrième combinaison de capteurs GMON selon les seuils de 1 à 5 cm, représentation par seuil, Québec, 2011.

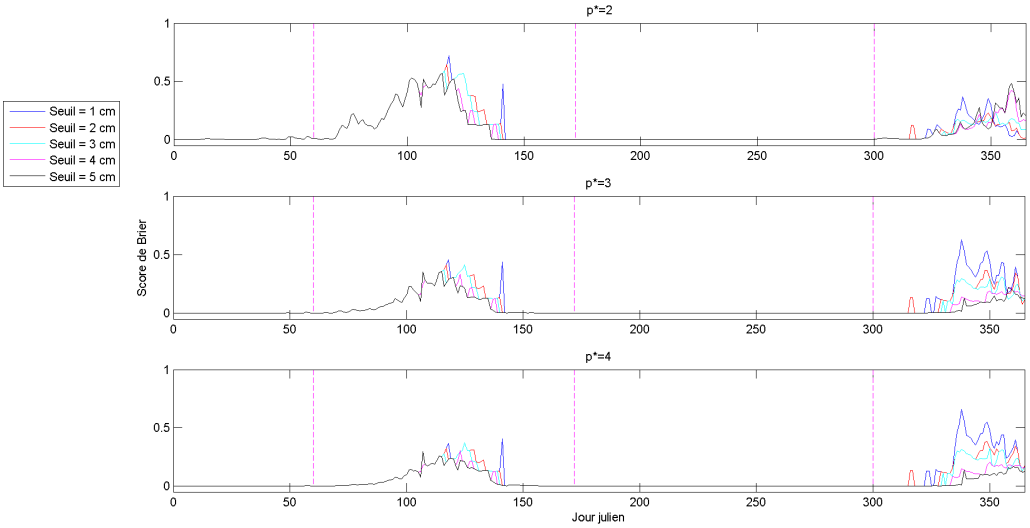

FIGURE F.6. Scores de Brier pour les modèles à deux, trois et quatre variables explicatives calculés pour les réalités neige/non-neige de la quatrième combinaison de capteurs GMON selon les seuils de 1 à 5 cm, représentation par modèle, Québec, 2011.

F.2. Omission/commission avec les données des capteurs SR50 pour les points de césure 0,4 et 0,5

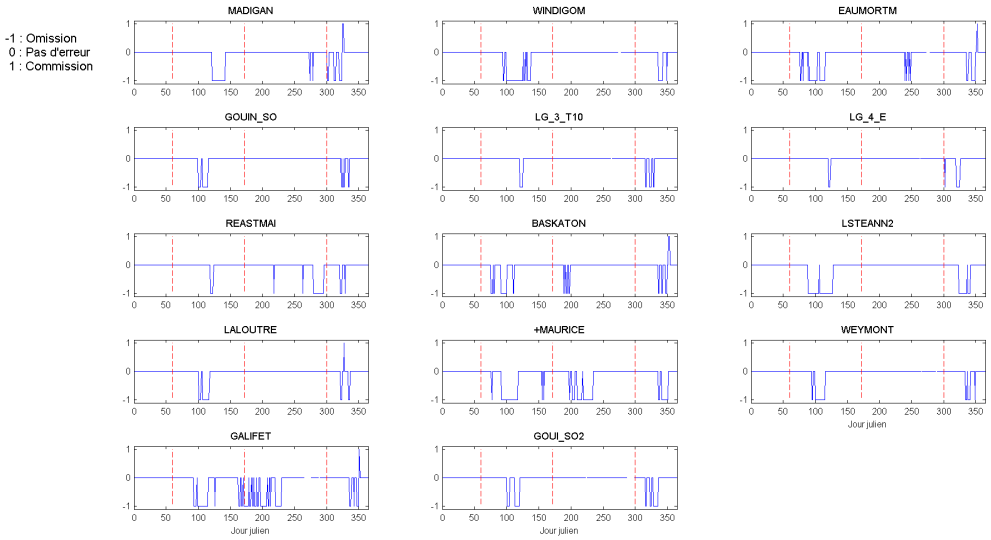

Figure F.7. Omission/commission pour les réalisations neige/non-neige du modèle à deux variables explicatives avec le point de césure 0,4 avec les réalisations des capteurs SR50 pour le seuil de 2 cm, Québec, 2011.

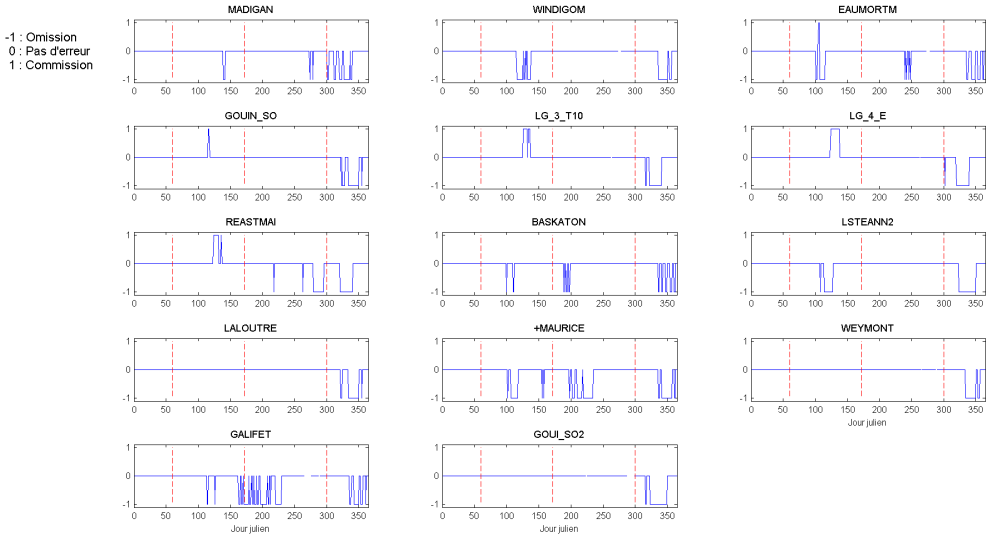

Figure F.8. Omission/commission pour les réalisations neige/non-neige du modèle à trois variables explicatives avec le point de césure 0,4 avec les réalisations des capteurs SR50 pour le seuil de 2 cm, Québec, 2011.

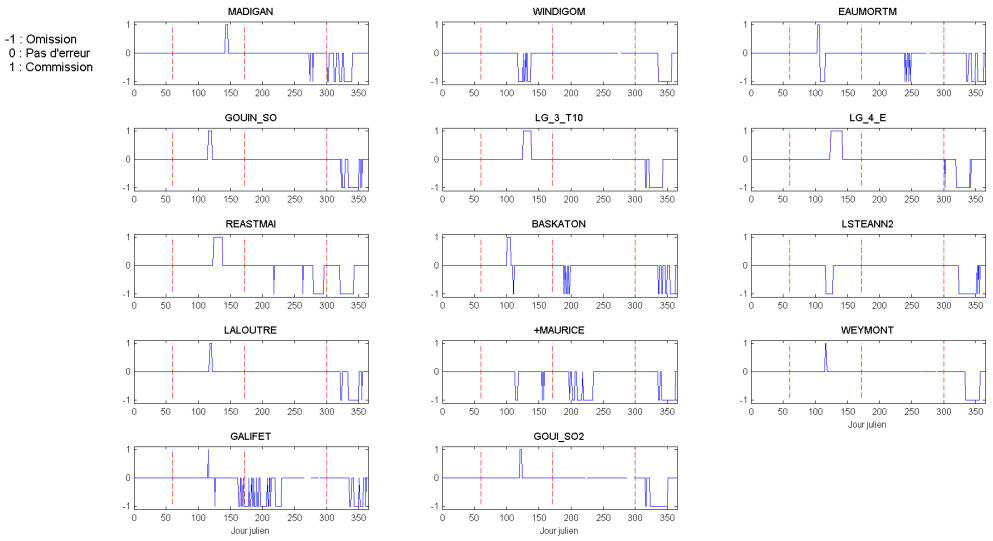

FIGURE F.9. Omission/commission pour les réalisations neige/non-neige du modèle à quatre variables explicatives avec le point de césure 0,4 avec les réalisations des capteurs SR50 pour le seuil de 2 cm, Québec, 2011.

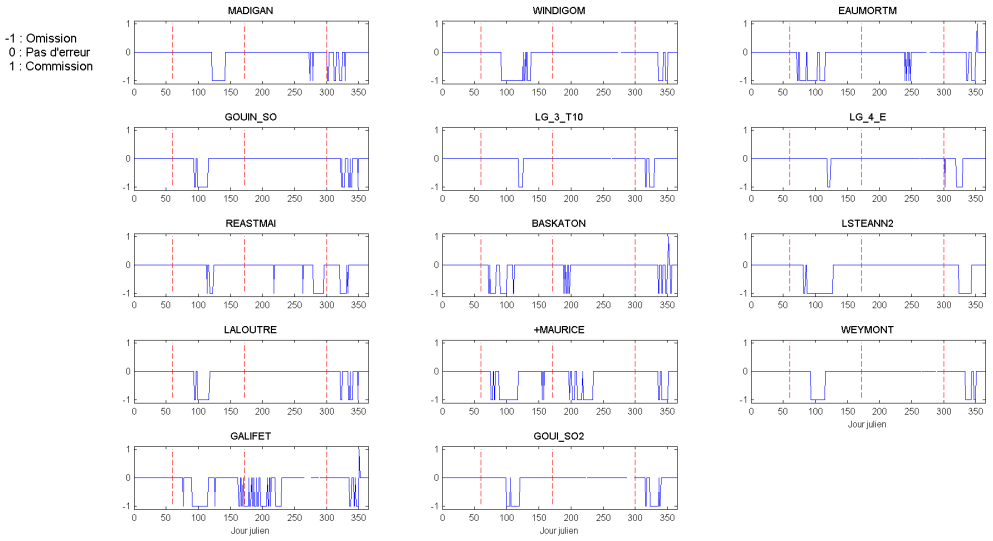

FIGURE F.10. Omission/commission pour les réalisations neige/non-neige du modèle à deux variables explicatives avec le point de césure 0,5 avec les réalisations des capteurs SR50 pour le seuil de 2 cm, Québec, 2011.

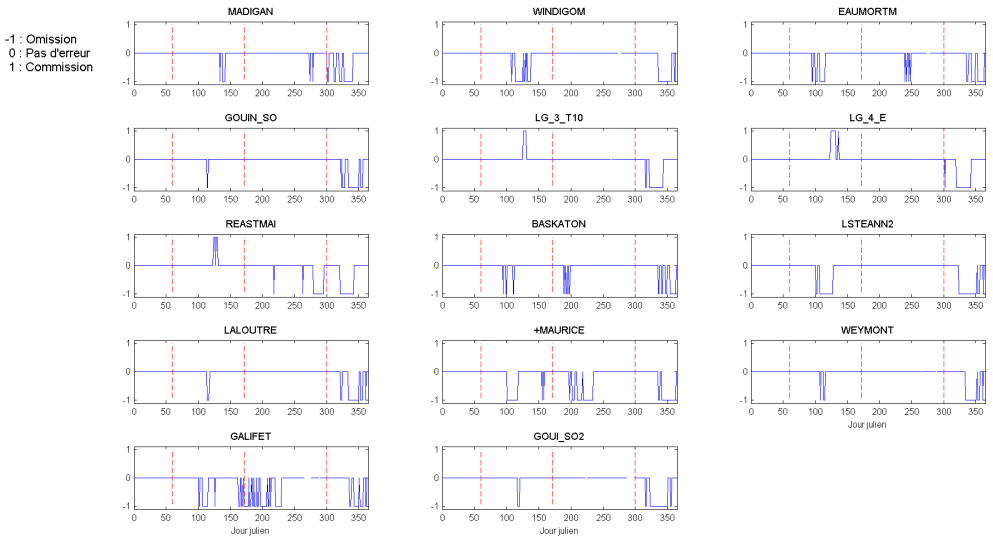

FIGURE F.11. Omission/commission pour les réalisations neige/non-neige du modèle à trois variables explicatives avec le point de césure 0,5 avec les réalisations des capteurs SR50 pour le seuil de 2 cm, Québec, 2011.

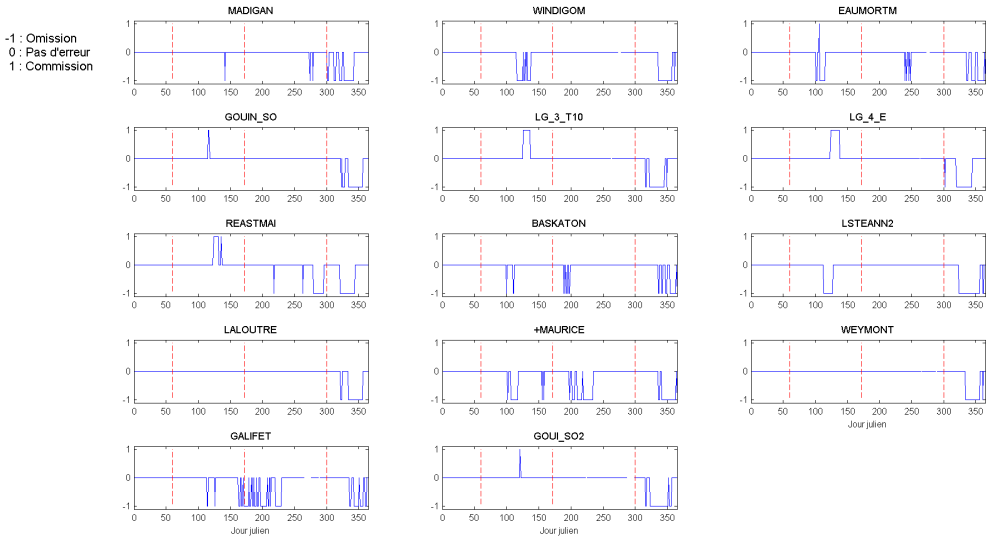

FIGURE F.12. Omission/commission pour les réalisations neige/non-neige du modèle à quatre variables explicatives avec le point de césure 0,5 avec les réalisations des capteurs SR50 pour le seuil de 2 cm, Québec, 2011.

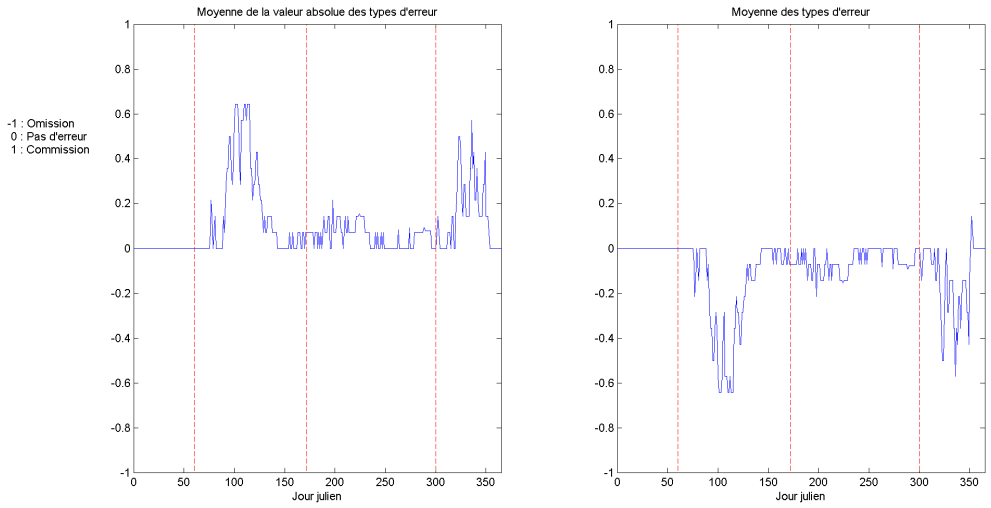

FIGURE F.13. Moyenne sur les différents points de grille de l'omission/commission pour les réalisations neige/non-neige du modèle à deux variables explicatives selon le point de césure 0,4 avec les réalisations des capteurs SR50, seuil 2 cm, Québec, 2011.

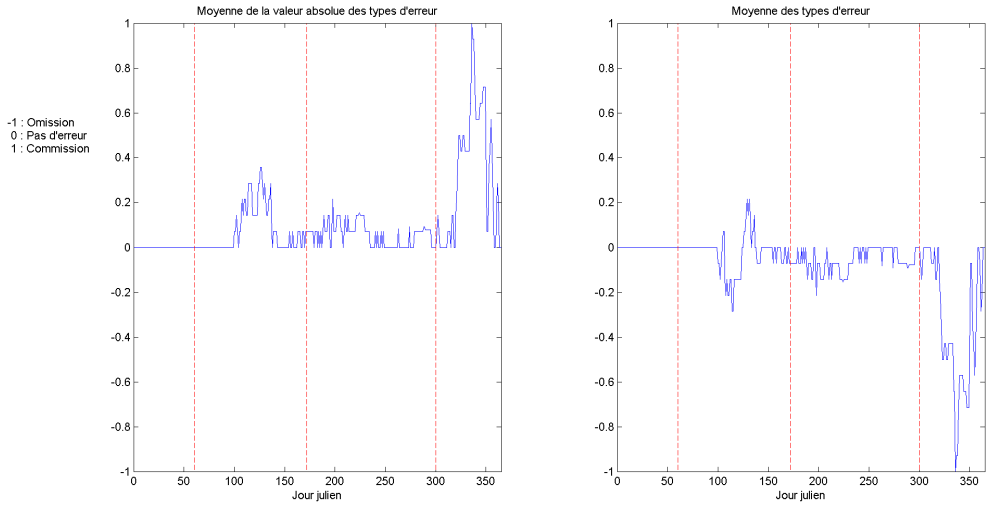

FIGURE F.14. Moyenne sur les différents points de grille de l'omission/commission pour les réalisations neige/non-neige du modèle à trois variables explicatives selon le point de césure 0,4 avec les réalisations des capteurs SR50, seuil 2 cm, Québec, 2011.

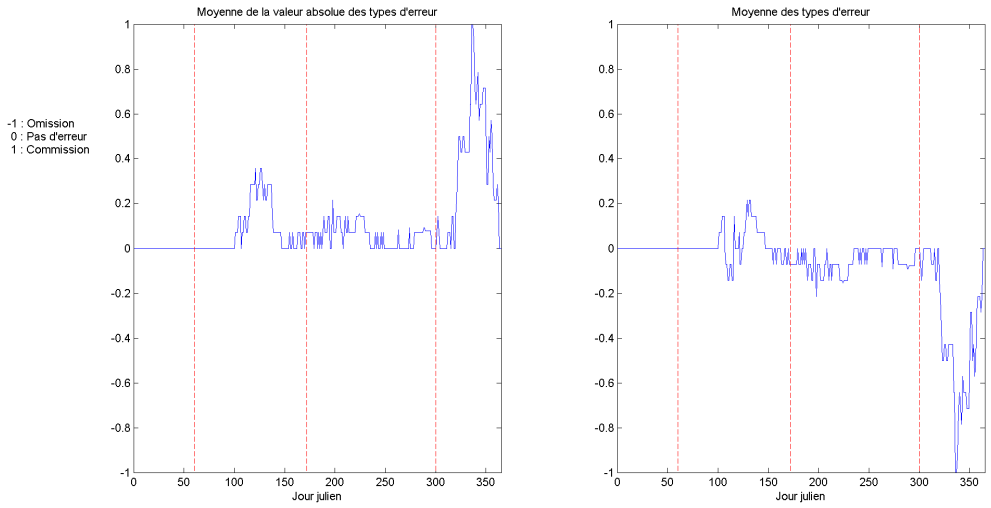

FIGURE F.15. Moyenne sur les différents points de grille de l'omission/commission pour les réalisations neige/non-neige du modèle à quatre variables explicatives selon le point de césure 0,4 avec les réalisations des capteurs SR50, seuil 2 cm, Québec, 2011.

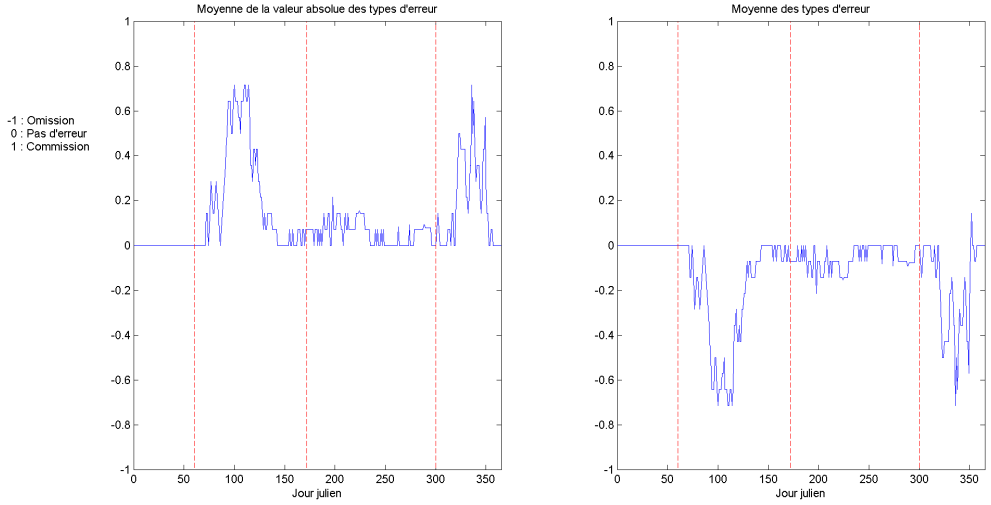

FIGURE F.16. Moyenne sur les différents points de grille de l'omission/commission pour les réalisations neige/non-neige du modèle à deux variables explicatives selon le point de césure 0,5 avec les réalisations des capteurs SR50, seuil 2 cm, Québec, 2011.

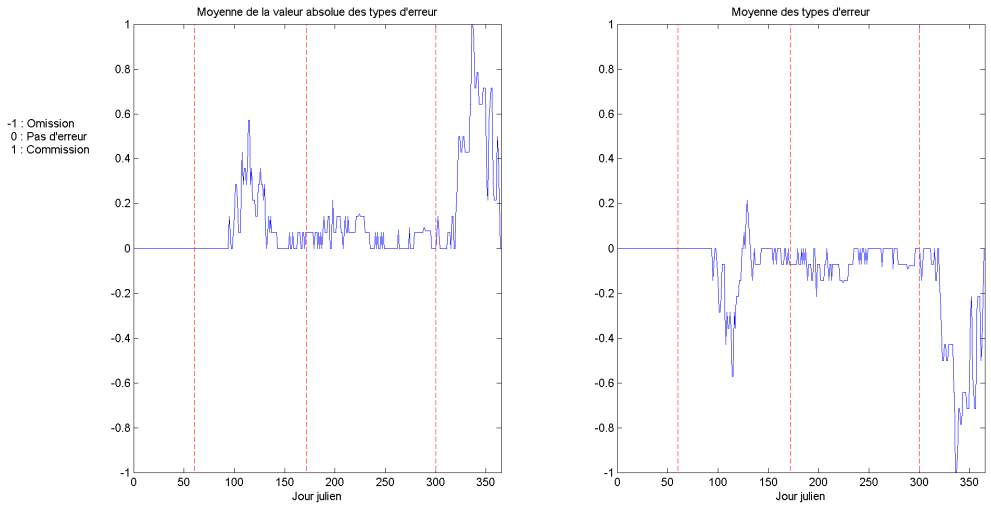

FIGURE F.17. Moyenne sur les différents points de grille de l'omission/commission pour les réalisations neige/non-neige du modèle à trois variables explicatives selon le point de césure 0,5 avec les réalisations des capteurs SR50, seuil 2 cm, Québec, 2011.

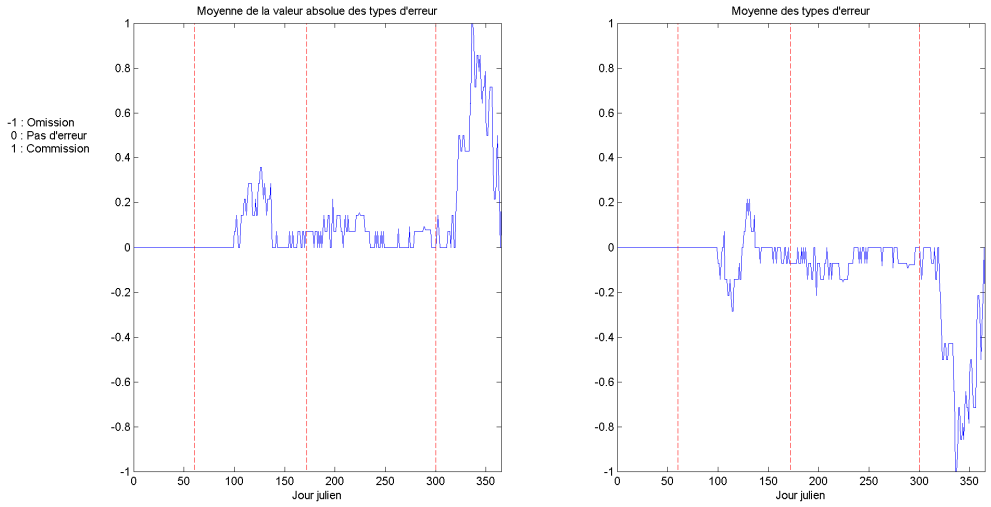

FIGURE F.18. Moyenne sur les différents points de grille de l'omission/commission pour les réalisations neige/non-neige du modèle à quatre variables explicatives selon le point de césure 0,5 avec les réalisations des capteurs SR50, seuil 2 cm, Québec, 2011.

F.3. Omission/commission avec les données des capteurs GMON pour les points de césure 0,4 et 0,5

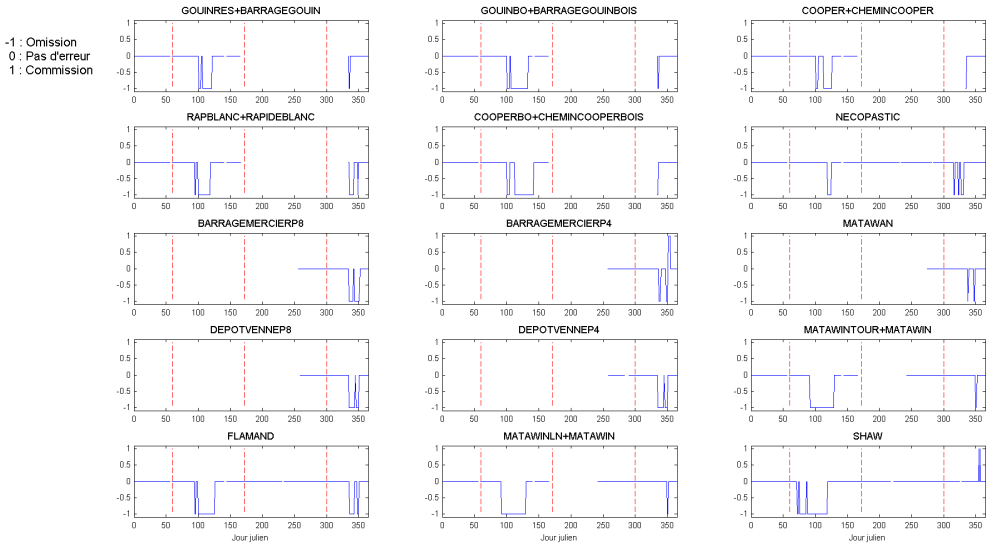

FIGURE F.19. Omission/commission pour les réalisations neige/non-neige du modèle à deux variables explicatives avec le point de césure 0,4 avec les réalisations des capteurs GMON pour le seuil de 1 cm, Québec, 2011.

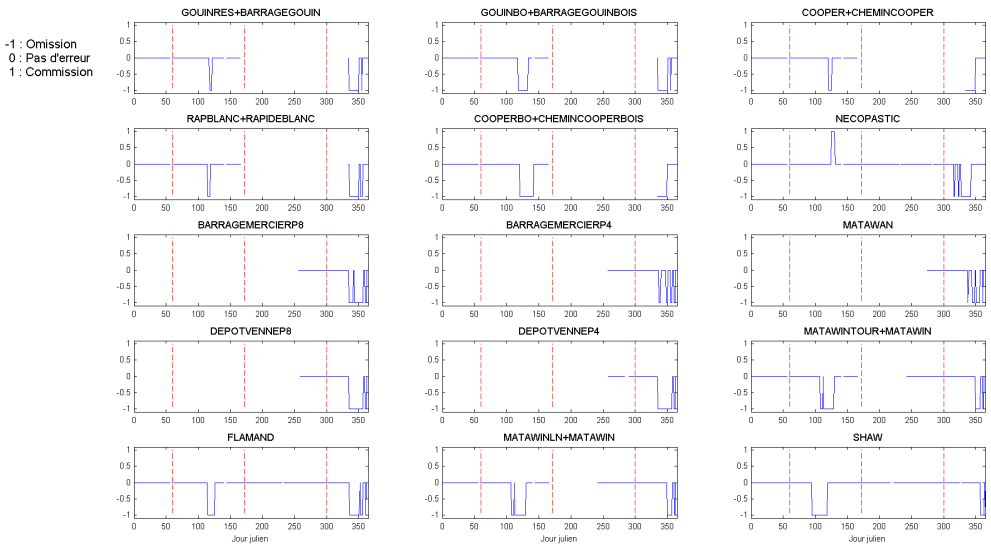

FIGURE F.20. Omission/commission pour les réalisations neige/non-neige du modèle à trois variables explicatives avec le point de césure 0,4 avec les réalisations des capteurs GMON pour le seuil de 1 cm, Québec, 2011.

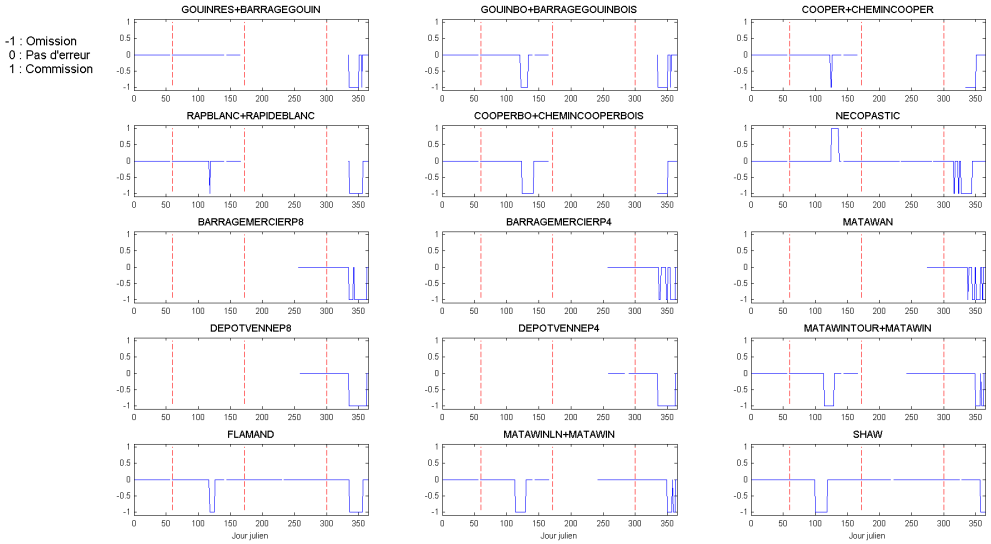

FIGURE F.21. Omission/commission pour les réalisations neige/non-neige du modèle à quatre variables explicatives avec le point de césure 0,4 avec les réalisations des capteurs GMON pour le seuil de 1 cm, Québec, 2011.

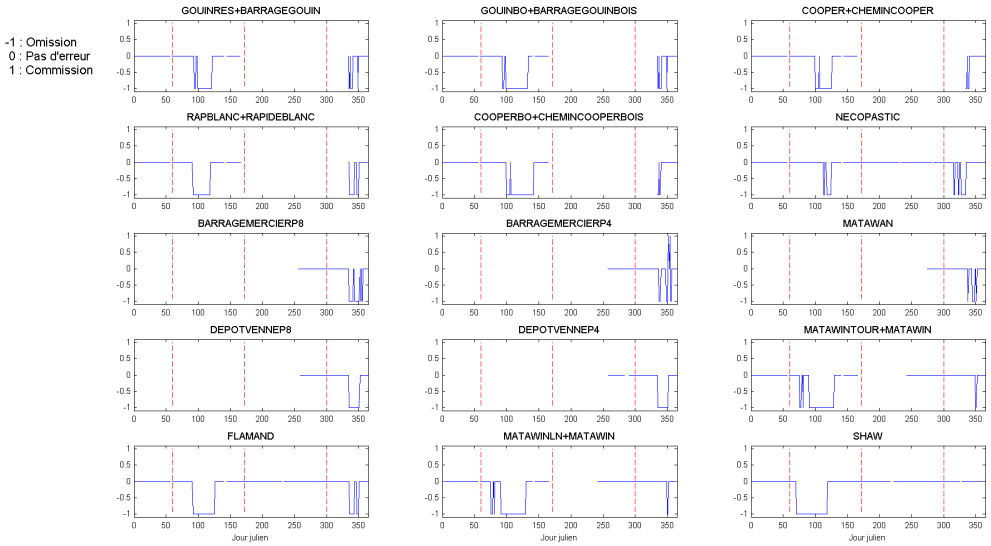

FIGURE F.22. Omission/commission pour les réalisations neige/non-neige du modèle à deux variables explicatives avec le point de césure 0,5 avec les réalisations des capteurs GMON pour le seuil de 1 cm, Québec, 2011.

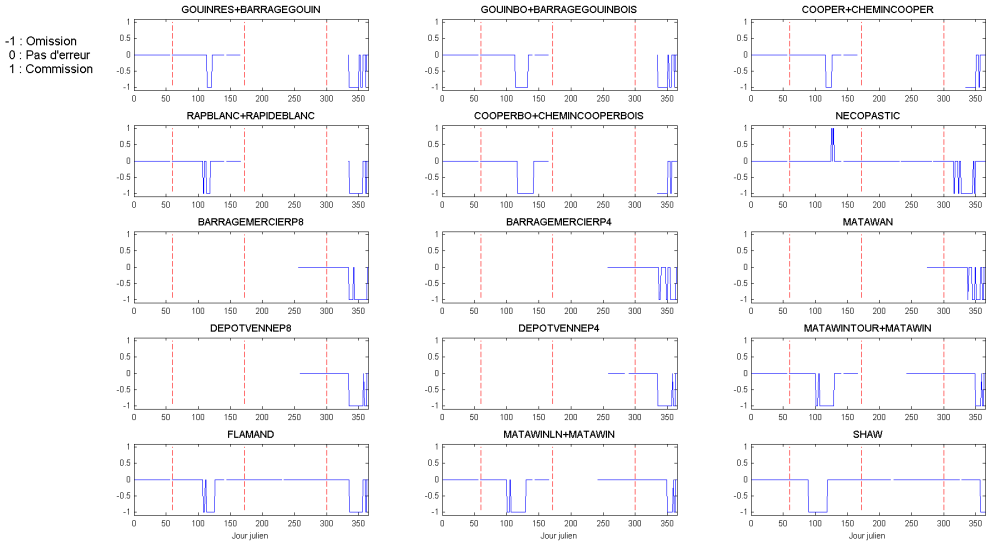

FIGURE F.23. Omission/commission pour les réalisations neige/non-neige du modèle à trois variables explicatives avec le point de césure 0,5 avec les réalisations des capteurs GMON pour le seuil de 1 cm, Québec, 2011.

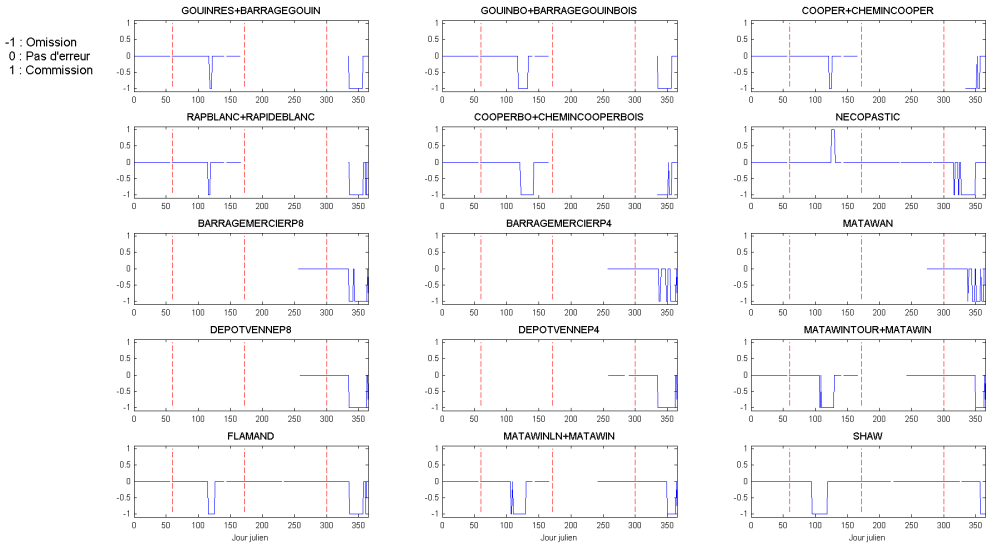

FIGURE F.24. Omission/commission pour les réalisations neige/non-neige du modèle à quatre variables explicatives avec le point de césure 0,5 avec les réalisations des capteurs GMON pour le seuil de 1 cm, Québec, 2011.

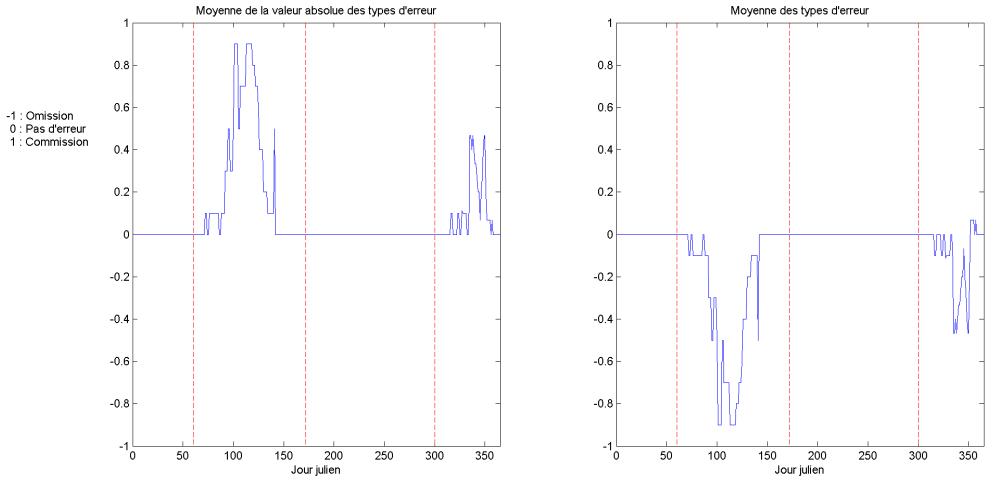

FIGURE F.25. Moyenne sur les différents points de grille de l'omission/commission pour les réalisations neige/non-neige du modèle à deux variables explicatives selon le point de césure 0,4 avec les réalisations des capteurs GMON, seuil 1 cm, Québec, 2011.

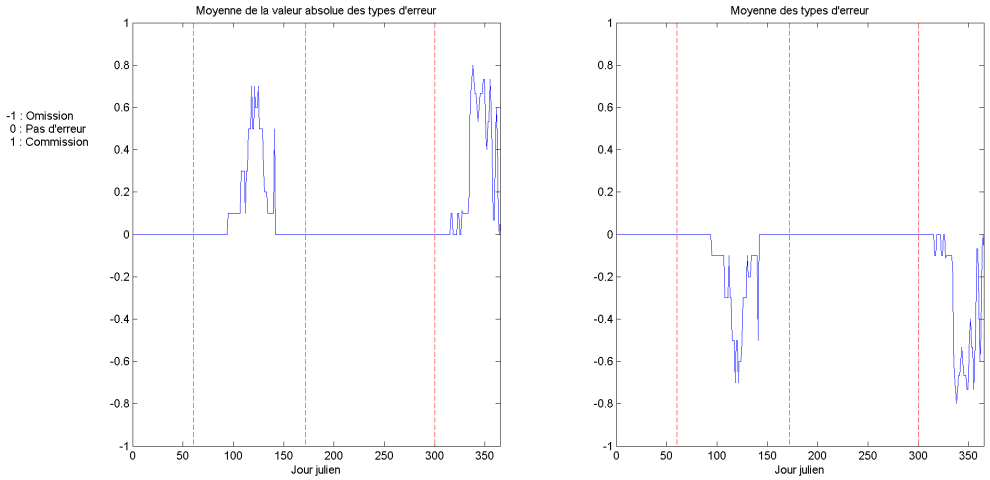

FIGURE F.26. Moyenne sur les différents points de grille de l'omission/commission pour les réalisations neige/non-neige du modèle à trois variables explicatives selon le point de césure 0,4 avec les réalisations des capteurs GMON, seuil 1 cm, Québec, 2011.

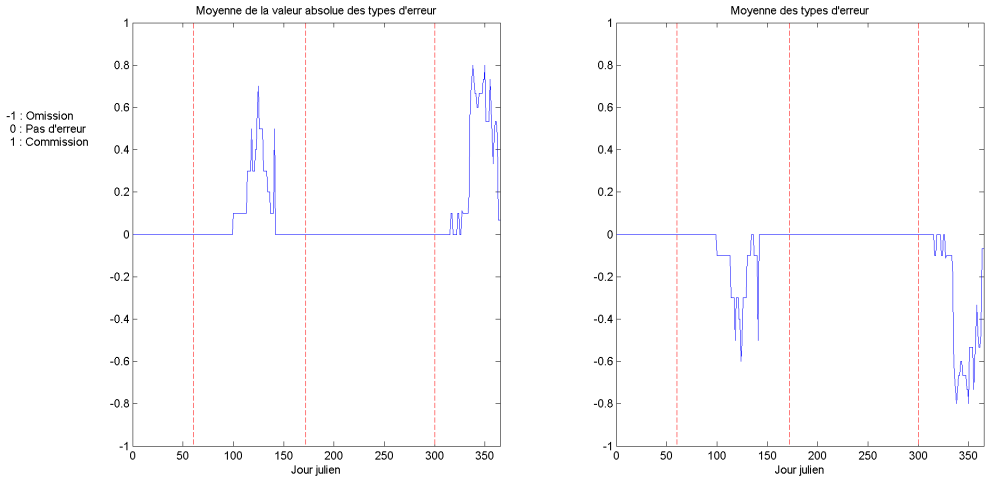

FIGURE F.27. Moyenne sur les différents points de grille de l'omission/commission pour les réalisations neige/non-neige du modèle à quatre variables explicatives selon le point de césure 0,4 avec les réalisations des capteurs GMON, seuil 1 cm, Québec, 2011.

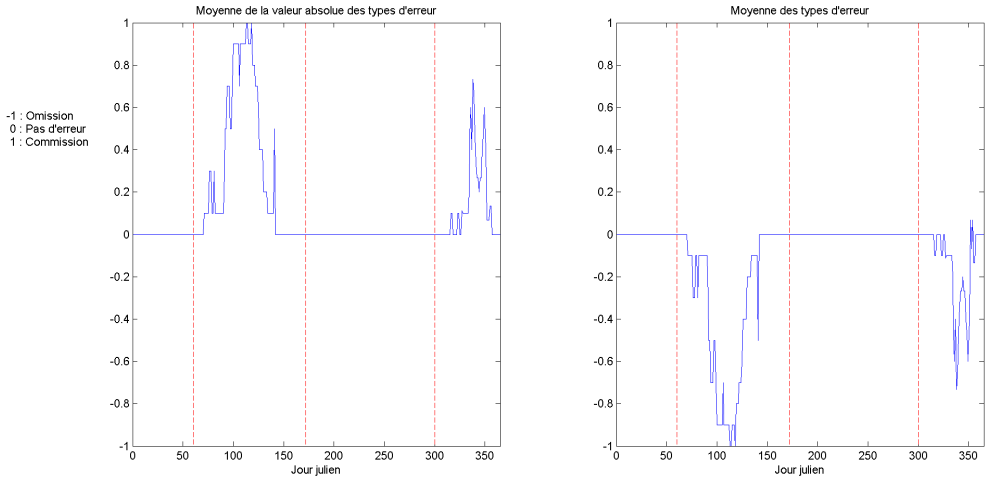

FIGURE F.28. Moyenne sur les différents points de grille de l'omission/commission pour les réalisations neige/non-neige du modèle à deux variables explicatives selon le point de césure 0,5 avec les réalisations des capteurs GMON, seuil 1 cm, Québec, 2011.

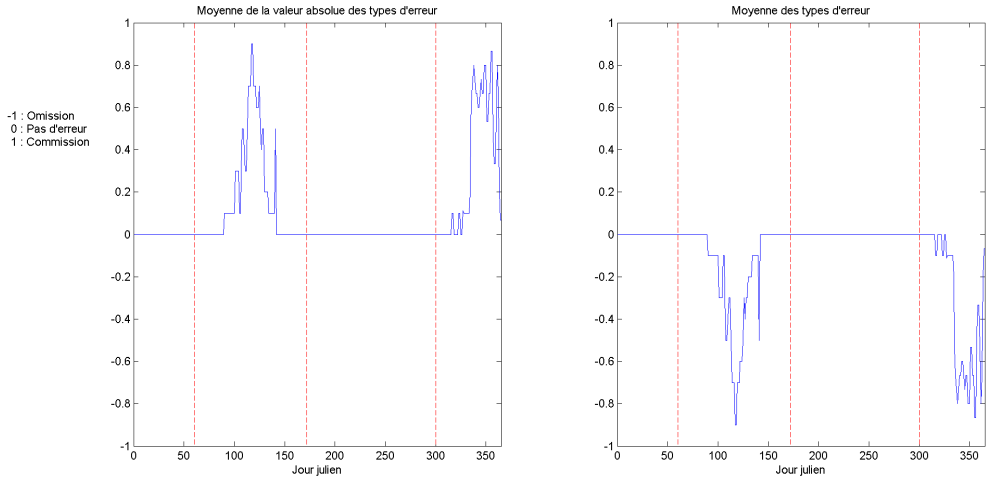

FIGURE F.29. Moyenne sur les différents points de grille de l'omission/commission pour les réalisations neige/non-neige du modèle à trois variables explicatives selon le point de césure 0,5 avec les réalisations des capteurs GMON, seuil 1 cm, Québec, 2011.

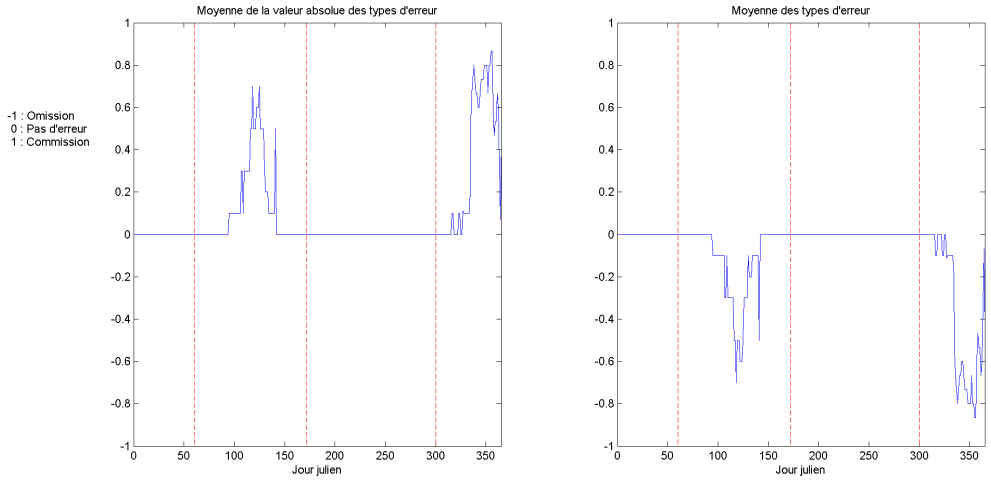

FIGURE F.30. Moyenne sur les différents points de grille de l'omission/commission pour les réalisations neige/non-neige du modèle à quatre variables explicatives selon le point de césure 0,5 avec les réalisations des capteurs GMON, seuil 1 cm, Québec, 2011.